Title

The natural product domain seeker NaPDoS: a phylogeny based bioinformatic tool to classify secondary metabolite gene diversity.

Journal

PloS one, 7(3)

ISSN

1932-6203

Authors

Ziemert, Nadine
Podell, Sheila
Penn, Kevin
et al.

Publication Date

2012

DOI

10.1371/journal.pone.0034064

Peer reviewed

OPEN ACCESS Freely available online

PLoS one

The Natural Product Domain Seeker NaPDoS: A Phylogeny Based Bioinformatic Tool to Classify Secondary Metabolite Gene Diversity

Nadine Ziemert[1], Sheila Podell[3], Kevin Penn[1], Jonathan H. Badger[2], Eric Allen[3], Paul R. Jensen[1]*

1 Center for Marine Biotechnology and Biomedicine, Scripps Institution of Oceanography, University of California San Diego, La Jolla, California, United States of America, **2** Microbial and Environmental Genomics, J. Craig Venter Institute, San Diego, California, United States of America, **3** Marine Biology Research Division, Scripps Institution of Oceanography, University of California San Diego, La Jolla, California, United States of America

Abstract

New bioinformatic tools are needed to analyze the growing volume of DNA sequence data. This is especially true in the case of secondary metabolite biosynthesis, where the highly repetitive nature of the associated genes creates major challenges for accurate sequence assembly and analysis. Here we introduce the web tool Natural Product Domain Seeker (NaPDoS), which provides an automated method to assess the secondary metabolite biosynthetic gene diversity and novelty of strains or environments. NaPDoS analyses are based on the phylogenetic relationships of sequence tags derived from polyketide synthase (PKS) and non-ribosomal peptide synthetase (NRPS) genes, respectively. The sequence tags correspond to PKS-derived ketosynthase domains and NRPS-derived condensation domains and are compared to an internal database of experimentally characterized biosynthetic genes. NaPDoS provides a rapid mechanism to extract and classify ketosynthase and condensation domains from PCR products, genomes, and metagenomic datasets. Close database matches provide a mechanism to infer the generalized structures of secondary metabolites while new phylogenetic lineages provide targets for the discovery of new enzyme architectures or mechanisms of secondary metabolite assembly. Here we outline the main features of NaPDoS and test it on four draft genome sequences and two metagenomic datasets. The results provide a rapid method to assess secondary metabolite biosynthetic gene diversity and richness in organisms or environments and a mechanism to identify genes that may be associated with uncharacterized biochemistry.

Citation: Ziemert N, Podell S, Penn K, Badger JH, Allen E, et al. (2012) The Natural Product Domain Seeker NaPDoS: A Phylogeny Based Bioinformatic Tool to Classify Secondary Metabolite Gene Diversity. PLoS ONE 7(3): e34064. doi:10.1371/journal.pone.0034064

Editor: Valerie de Crécy-Lagard, University of Florida, United States of America

Received November 26, 2011; **Accepted** February 26, 2012; **Published** March 29, 2012

Funding: This work was supported by the National Institutes of Health grant numbers RO1GM086261 to PRJ, UO1TW0007401 to PRJ, and RO1GM085770 to PRJ, and a grant from the German Research Foundation (DFG 1325/1-1) to NZ. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: pjensen@ucsd.edu

Introduction

Genome sequencing has revealed that the secondary metabolite potential of even well studied bacteria has been severely underestimated [1,2]. This revelation has led to an explosion of interest in genome mining as an approach to natural product discovery [3,4,5,6,7,8]. Considering that natural products remain one of the primary sources of therapeutic agents [9,10], sequence analysis provides opportunities to identify strains with the greatest genetic potential to yield novel secondary metabolites prior to chemical analysis and thus increase the rate and efficiency with which new drug leads are discovered. In addition, community or metagenomic analyses can be used to identify environments with the greatest secondary metabolite potential and to address ecological questions related to secondary metabolism. To capitalize on these opportunities, it is critical that new bioinformatics tools be developed to handle the massive influx of sequence data that is being generated from next generation sequencing technologies [11].

Polyketide synthases (PKSs) and non-ribosomal peptide synthetases (NRPSs) are large enzyme families that account for many clinically important pharmaceutical agents. These enzymes employ complimentary strategies to sequentially construct a diverse array of natural products from relatively simple carboxylic acid and amino acid building blocks using an assembly line process [12,13]. The molecular architectures of PKS and NRPS genes have been reviewed in detail and minimally consist of activation (AT or A), thiolation (ACP or PCP), and condensation (KS or C) domains, respectively [14,15,16,17,18]. These genes are among the largest found in microbial genomes and can include highly repetitive modules that create considerable challenges to accurate assembly and subsequent bioinformatic analysis [8].

When the challenges associated with PKS and NRPS gene assembly can be overcome, a number of effective bioinformatics tools have been developed for domain parsing [19,20] and domain string analysis [21,22]. In cases of modular type I PKSs and NRPSs where domain strings follow the "co-linearity rule" such that substrates are incorporated and processed according to the precise domain organization observed in the pathway, bioinformatics has been used to make accurate structural predictions about the metabolic products of those pathways [23]. However, the increasing number of exceptions to co-linearity, such as module skipping and stuttering [24], create limitations for precise, sequence-based structure prediction. The bioinformatic tools

currently available for secondary metabolism have been reviewed [25,26] and are complemented by the recent release of antiSMASH, which has the capacity to accurately identify and provide detailed sequence analysis of gene clusters associated with all known secondary metabolite chemical classes [27]. While all of these tools have useful applications, NaPDoS employs a phylogeny based classification system that can be used to quantify and distinguish KS and C domain types from a variety of datasets including the incomplete genome assemblies typically obtained using next generation sequencing technologies. These specific domains were selected because they are highly conserved and have proven to be among the most informative in a phylogenetic context [28,29].

Phylogenomics provides a useful approach to infer gene function based on phylogenetic relationships as opposed to sequence similarities [30,31]. While the evolutionary histories of PKS and NRPS genes are largely uninformative due to their size and complexity, KS and C domain phylogenies reveal highly supported clustering patterns. These patterns have been used to distinguish type II PKSs associated with spore pigment and antibiotic biosynthesis [32], type I modular and hybrid PKSs [33], and subsequently to identify many different PKSs types [34]. KS phylogeny has also been used to predict pathway associations [26,35] and, in some cases, the secondary metabolic products of those pathways [28,36,37]. Phylogenetics has also been used to successfully identify PKS sequences from complex metagenomic datasets [38]. Likewise, C domain phylogeny clearly delineates functional subtypes as opposed to species relationships [39] and has been used to identify new functional classes, such as the "starter" C domain [29]. Taken together, the established phylogenetic relationships of KS and C domains provide an effective framework within which to assess secondary metabolite gene richness and diversity and to identify new functional classes that may be associated with uncharacterized biosynthetic mechanisms.

Here we introduce the web tool Natural Product Domain Seeker (NaPDoS), which extracts and rapidly classifies KS and C domains from a wide range of sequence data. The results can be used to assess the potential for PKS and NRPS secondary metabolite biosynthesis in organisms or environments and to identify new phylogenetic lineages, which can subsequently be investigated as a source of new mechanistic biochemistry. We tested NaPDoS on four draft bacterial genome sequences and two metagenomic datasets. The results reveal a remarkable level of secondary metabolite gene diversity among closely related strains and provide a mechanism to assess secondary metabolism from poorly assembled genomic data.

Materials and Methods

Reference database

KS and C domains were extracted from select PKS and NRPS genes associated with experimentally characterized biosynthetic pathways using the online program NRPS-PKS (http://www.nii.res.in/searchall.html) [19,21]. The pathways selected include representatives of the currently known enzyme architectures and functions associated with type I and II PKSs and NRPSs and thus this database is not meant to be comprehensive. The biochemical function and enzyme architecture of each domain was manually confirmed by analysis of the associated domain string and secondary metabolic product. Based on these results, each sequence was preliminarily assigned to a domain class. The compound produced by the associated pathway, the literature reference including PubMed ID, and the gene accession number was also recorded for each domain.

Sequence alignment and phylogeny

The amino acid sequences of all reference KS and C domains were aligned using either MUSCLE [40] or ClustalX (version 1.83) [41] with the BLOSUM 62 protein weight matrix. The alignments were manually adjusted using Mesquite [42]. Maximum likelihood, parsimony, and neighbor-joining phylogenetic trees were constructed using the "a la carte" mode at the Phylogeny.fr website (http://www.phylogeny.fr/) [43]. Final maximum likelihood trees were constructed from the reference data set with the program PHYML [44]. Final domain classifications were made based on the phylogenetic relationships observed in these trees.

NaPDoS and Webportal

The NaPDoS web portal identifies candidate KS and C domains through a combination of hidden markov model (HMM) searches and the basic local alignment search tool (BLAST) algorithm [45] optimized for query input type as shown in Figure 1. PCR products or coding sequences (CDS) in nucleotide or amino acid format are analyzed directly by local BLASTX or BLASTP searches against the manually curated reference database of experimentally verified KS and C domains described above. This BLAST-based approach proved more effective than HMM models in detecting the target domains from short query sequences. Genomic sequences (including contigs, incomplete drafts, or complete genomes) and metagenomic nucleotide data sets are first pre-screened to obtain rough coordinates for KS and C domains using the KS domain HMM developed by Yadav and co-workers [21] and the PFAM C domain model PF00668 [46]. The resulting candidate domains are then subjected to BLAST analyses using the same manually curated reference database as described above.

BLAST results are linked to a back-end MySQL relational database via CGI-scripting to retrieve and report domain classification and related pathway information. Query sequences are trimmed according to their BLAST match coordinates by a custom Perl script then aligned to each other and their database matches using MUSCLE [40]. Trimmed sequences can be downloaded along with best BLAST matches in FASTA or MSF aligned format. Finally, trimmed and aligned candidate KS and C sequences plus BLAST matches can be inserted into a phylogenetic tree generated from the reference database using FastTree to estimate maximum likelihood [44]. Newick format output from FastTree is converted to SVG format graphic images using the Newick-Utilities program [47]. NaPDoS does not employ any stand-alone software that was created specifically for its operation but instead employs pre-existing and publically available programs as described above.

Draft genomes and metagenomes

Draft genome sequences of *S. arenicola* strain CNH-643 (accession number PRJNA84391), *S. arenicola* strain CNT-088 (accession number PRJNA84269), *"S. pacifica"* strain CNS-143 (accession number PRJNA84389), and *"S. pacifica"* strain CNT-133 (accession number PRJNA84271) were obtained at 8× coverage at the J. Craig Venter Institute using 454 GS FLX pyrosequencing and 0.5× Sanger sequencing as previously described [48] based on an estimated genome size of 5.6 Mb. The sequence data were assembled using the Newbler Assembler with the mapping option [49]. *S. arenicola* strains were mapped onto the complete *S. arenicola* strain CNS-205 genome and the *S.*

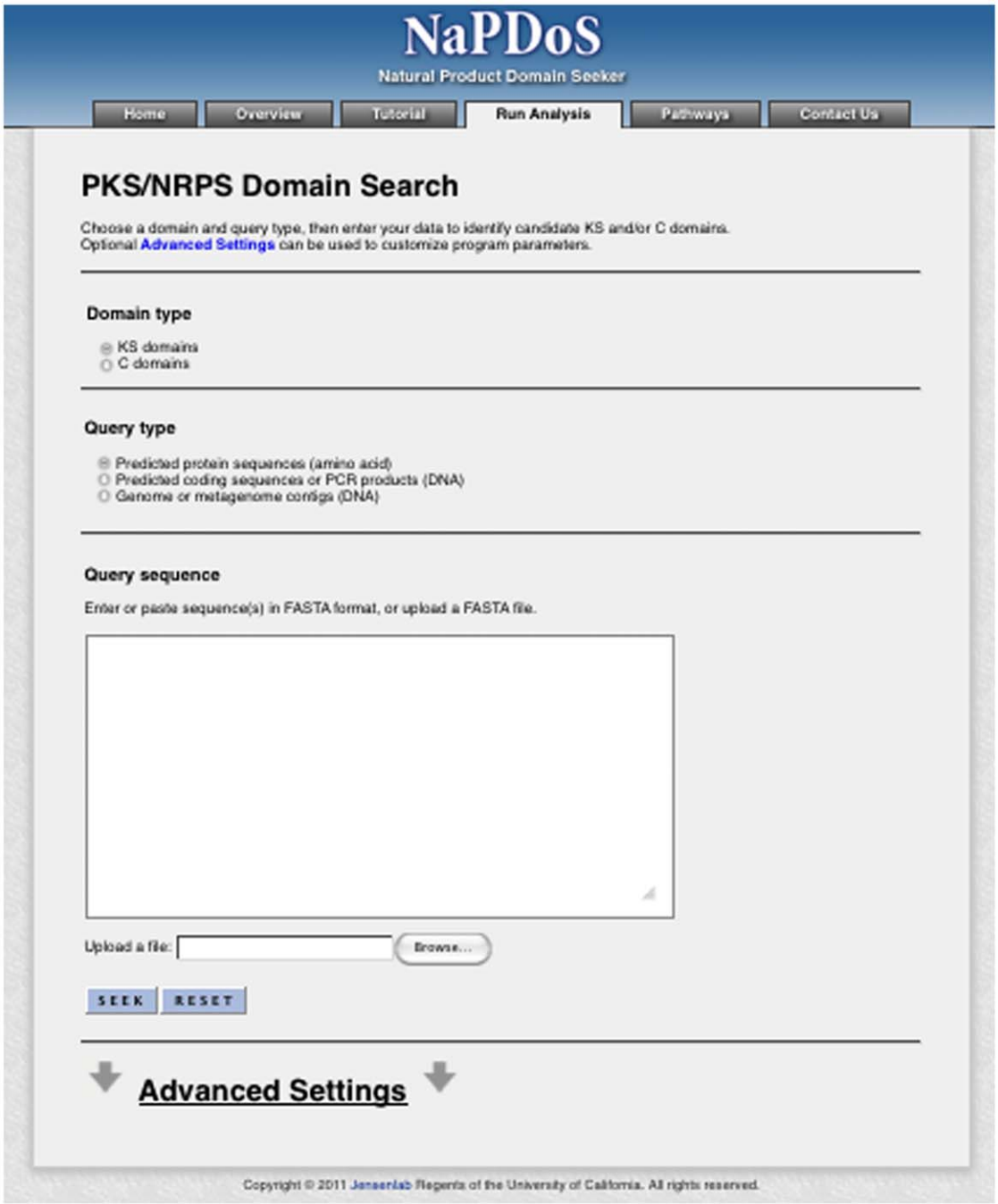

Figure 1. NaPDoS bioinformatic pipeline. The web interface to this pipeline is divided 3 consecutive steps. Nucleic acid sequences are translated into predicted amino acids and genomic sequences are screened using Hidden Markov Models (HMM). For protein and small nucleic acid sequences a BLAST search is performed against curated reference database examples to identify matches to known PKS/NRPS pathways. Selected candidate sequences plus the BLAST results are trimmed and inserted into a manually curated reference alignment, keeping the original reference alignment intact. This alignment is used to build a tree.
doi:10.1371/journal.pone.0034064.g001

pacifica strains were mapped to the complete *S. tropica* CNB-440 genome [50] while any unmapped sequence data was assembled de novo. The four draft *Salinispora* genomes were mined for KS and C domains using NaPDoS with default settings. The metagenomic datasets (whale fall, AAFZ00000000, AAFY00000000, AAGA00000000 and Minnesota farm soil, AAFX00000000, [51]) were mined using default HMM settings (e^{-5}) and the resulting sequences further subjected to a loose BLAST analysis with an e-value cut-off of 1 to obtain more precise coordinates and assign initial domain classifications.

Results

The Natural Product Domain Seeker (NaPDoS)

The publically available web tool NaPDoS (http://npdomainseeker.ucsd.edu/) was created to detect and classify

KS and C domains in nucleotide and amino acid sequence data. The query data can be PCR amplicons, genes, contigs, genomes, or metagenomes. The current query size limits are <30 MB and <50,000 individual sequences. The website provides a detailed tutorial on the use of this tool, which is implemented using a web interface (Figure 2) that follows the bioinformatic pipeline shown in Figure 1. Query sequences are BLASTed against the reference database, which currently contains 459 KS and 190 C domains derived from 66 PKS, 20 NRPS, 8 PKS/NRPS hybrid, and 5 fatty-acid synthase (FAS) biosynthetic pathways. These sequences can be downloaded from the website and encompass all major classes of type I and II KS and C domains currently described in the literature [12,29,52,53]. This manually curated database will be updated periodically as new modular architectures and biochemical features are discovered for each domain type.

The primary output for all analyses includes the query identification, best database match, percent identity, alignment length, e-value, and product and classification of the biosynthetic pathway associated with the best match. KS and C domain sequences derived from the input data can then be output in raw format or aligned with the best BLAST matches. A NaPDoS independent BLAST of the output domain sequence(s) against the NCBI nr database is also highly recommended to check for matches that do not occur in the reference database.

To generate a final classification for each domain sequence, it is highly recommended to construct a phylogenetic tree, especially in cases where the percent sequence identity to the top database match is low. If that option is chosen, a profile alignment is generated in which the query sequences are incorporated into a carefully curated reference alignment generated from the sequences in the reference database. This alignment is then used to create a phylogenetic tree, which needs to be manually interpreted to establish a final classification for each sequence. Interpreting sequences in the context of a phylogenetic tree is particularly important given that the NaPDoS pipeline is intentionally set to low stringency in an effort to detect all possible KS and C domains. Thus, homologs not involved in secondary metabolism such KSs associated with fatty acid biosynthesis are regularly detected. These sequences can be readily classified based on the phylogenetic tree.

Domain classification

KS and C domain phylogenies form the basis of the NaPDoS classification system (Figure 3). KS domains clade based on biochemical function and enzyme architecture, which are described in Table 1. In some cases, e.g. enediynes, these clades are also predictive of structural motifs associated with the secondary metabolites produced. The KS phylogeny clearly delineates type I and II PKSs (Figure 3A). The shared ancestry reported between type II PKS and FAS sequences [54] is clearly maintained in this tree. The vast majority of the reference sequences fall into the type I PKS clade. This clade can be further resolved into seven classes, which are not always monophyletic. This polyphyly reflects the complex evolutionary histories of the different classes such as the *trans*-AT KSs, which evolved by extensive HGT and exploit considerably greater modular architectures than the *cis*-AT group, which has largely evolved by gene duplication [55]. However, all of these lineages are highly

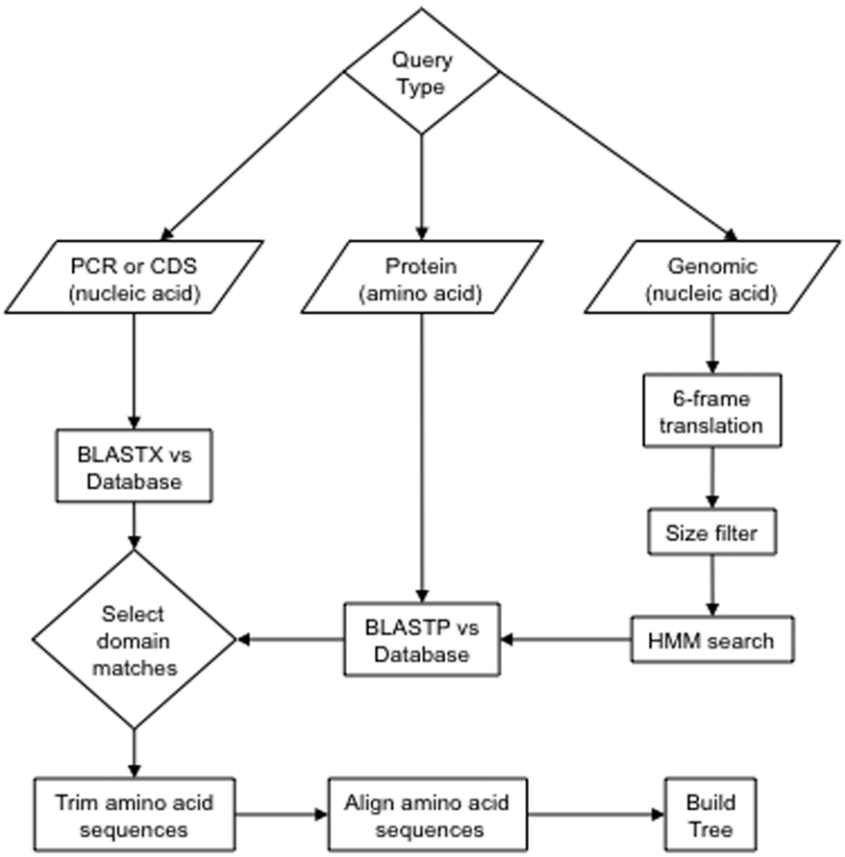

Figure 2. Screen shot of the NaPDoS webpage.
doi:10.1371/journal.pone.0034064.g002

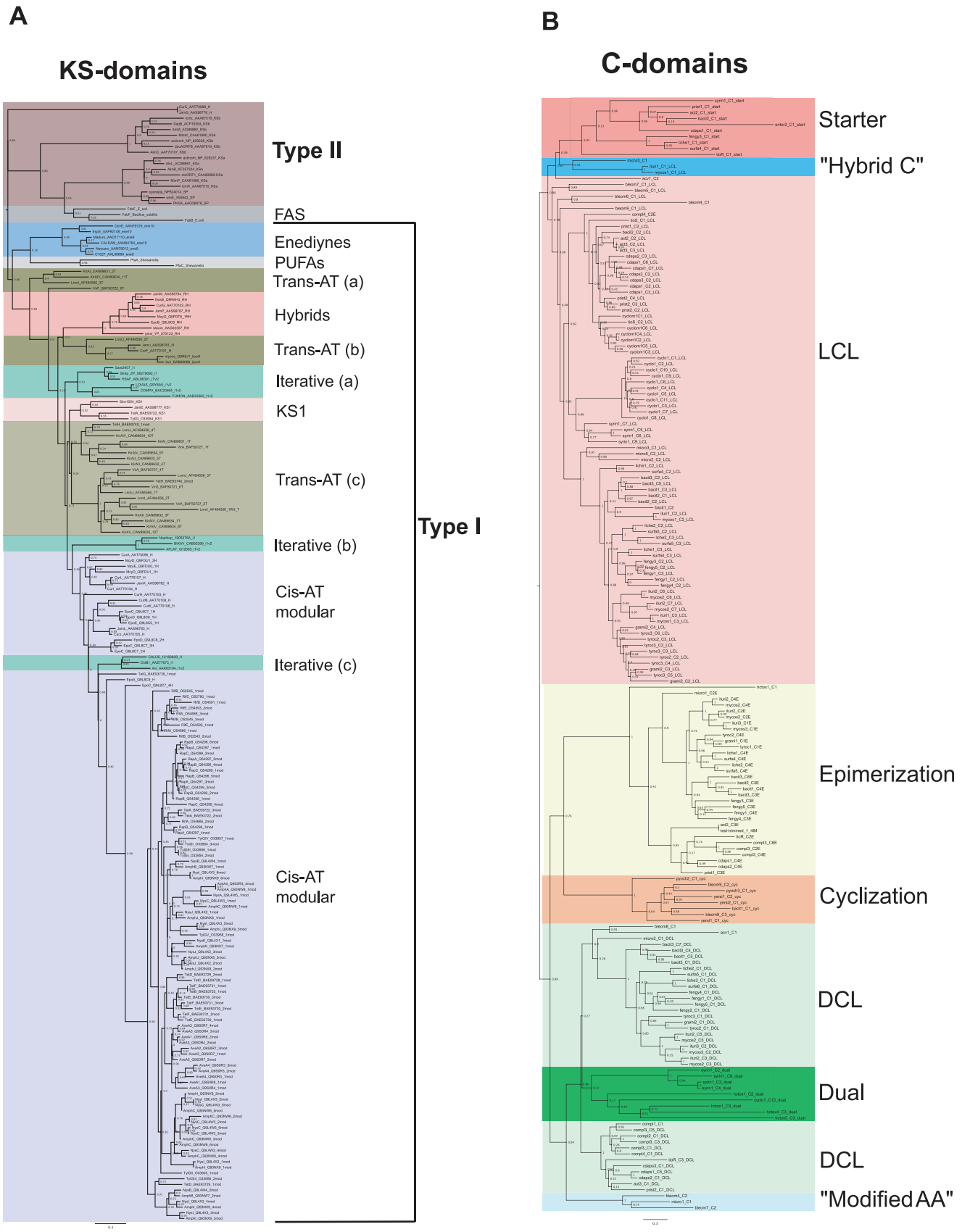

Figure 3. Phylogeny based domain classification. A) KS domain phylogeny. Polyphyletic groups are distinguished by letters. B) C domain phylogeny.
doi:10.1371/journal.pone.0034064.g003

Table 1. KS domain classification.

Type	Class	Description	Product (example)
I	Enediyne	Iteratively acting, builds typical 9 or 10 membered enedyines.	Enediyne (calicheamicin)
	Trans-AT	Module lacks cognate AT domain; this activity is provided instead by a discrete protein encoded in *trans*.	Polyketide/macrolide (leinamycin)
	Cis-AT	Multi-domain module that includes AT domain.	Polyketide/macrolide (erythromycin)
	Hybrid	Catalyzes a condensation reaction between an amino acid and an acyl extender unit in a NRPS/PKS pathway.	Peptide-polyketide (microcystin)
	Iterative	Domain is used multiple times in a cyclic fashion.	Polycyclic polyketide (aflatoxin)
	PUFA	Produces long chain fatty acids containing more than one double bond.	Polyunsaturated fatty acid (omega-3-fatty-acid)
	KS1	Occurs in the first module of multimodular genes, includes typical starter KSs (KSQ) as well as KS domains that incorporate unusual precursors.	Polyketide, peptide-polyketide (salinosporamide)
II	Type II	Each domain occurs on a discrete protein.	Aromatic polyketide (actinorhodin)
	FAS	Involved in fatty acid biosynthesis (eg., FabB and FabF from bacteria).	Fatty acid (palmitic acid)

doi:10.1371/journal.pone.0034064.t001

supported in the tree (likelihood values 0.7–1.0) and largely agree with previous phylogenetic studies [29,56].

In the case of C domains, the sequences generally clade based on substrate specificity, i.e. the stereochemistry of the amino acids incorporated and the subsequent tailoring reactions they perform (Figure 3B). Eight clades are identified in the tree of which six are functionally characterized. The characterized clades are comprised of LCL domains, which catalyze a peptide bond between two L-amino acids, DCL domains, which link an L-amino acid to a growing peptide ending with a D-amino acid, starter C domains, which acylate the first amino acid with a β-hydroxy-carboxylic acid, cyclization domains, which catalyze both peptide bond formation and the subsequent cyclization of cysteine, serine or threonine residues, epimerization (E) domains, which switch the chirality of the last amino acid in the growing peptide, and dual E/C domains, which catalyze both epimerization and condensation reactions. These six functional classes are well supported in the tree and largely monophyletic. Two experimentally uncharacterized clades are identified in the tree, one of which has been conditionally assigned the name "modified AA" (Figure 3B). This clade contains domains from the bleomycin and microcystin pathways. Although the biochemical function of these domains has not been experimentally defined, they appear to be involved in the modification of the incorporated amino acid, for example the dehydration of serine to dehydroalanine [57,58]. C domains in the second functionally uncharacterized clade have been conditionally assigned the name "hybrid C". The three sequences in this clade (micro5, ituri1, and mycos1) are each located downstream of an aminotransferase domain and appear to be involved in the condensation of an amino acid to an aminated polyketide resulting in a hybrid PKS/NRPS secondary metabolite. The phylogenetic relationships of the KS and C domains in the reference dataset form the basis of the NaPDoS classification system and provide a framework within which new clades and biochemical functions can be discovered.

Genome analyses

As a positive control, NaPDoS was used to analyze the genome sequence of *Streptomyces avermitilis* strain MA-4680. This analysis revealed 67 KS and 15 C domains (Table S1), which encompass all of the PKS, NRPS, and hybrid PKS/NRPS gene clusters that were reported to contain these domains [59]. NaPDoS also correctly identified all of the KS and C domains in the complete genome sequences of *S. tropica* (strain CNB-440) and *S. arenicola* (strain CNS-205) [50]. NaPDoS was then tested on four draft *Salinispora* genome sequences. These low coverage drafts were generated using pre-Titanium 454 technology (avreage read length 244 bp) and yielded poor assemblies and a large number of contigs (Table S2). There was no evidence that any biosynthetic gene clusters had been completely assembled based on the analysis of flanking regions and comparisons with pathways that appeared common with the CNB-440 and CNS-205 sequences [50]. Nonethe-less, NaPDoS successfully detected 18–36 KS domains and 5–14 C domains in the un-annotated FASTA files generated for each of the four draft genomes (Table S2). More than half (56%) of these sequences showed no significant BLAST matches to domains associated with biochemically characterized biosynthetic genes and thus could not be linked to specific secondary metabolic products. More significantly, 8 KS and 9 C domains detected in the four draft sequences were not observed in the two closed *Salinispora* genomes (Table S3). These sequences (KS7-14 and C5-13) cover a broad range of domain classes and indicate considerable new biosynthetic potential among a group of closely related strains. Two C domains fell into the "Modified AA" clade, which has yet to be experimentally characterized. Given that the upstream A domain specifies serine in both cases, it can be predicted that this domain results in the incorporation of dehydrated serine (ie., dehyroalanine) into the non-ribosomal peptide. This hypothesis has not yet been tested, but is supported by the reference sequences in this clade, which perform similar dehydration reactions.

Interestingly, two KS domains with close matches (89% and 94%) to those associated with the biosynthesis of salinosporamide A [60] were observed in "*S. pacifica*" strain CNT-133. This was unexpected given that compounds in this series had previously been reported exclusively from *S. tropica* [61]. This observation subsequently led to the discovery of a new compound in the salinosporamide series [5] and a rare window into pathway evolution in two closely related bacterial species [37]. Furthermore, a KS domain that shares close sequence identity with domains involved in the biosynthesis of tylactone in *Streptomyces fradiae* [62] was detected in strain CNH-643 (Table S3).

Subsequent chemical studies revealed the production of several new tylactone derivatives by this strain (unpublished data). The same four draft genome sequences were also analyzed using antiSMASH [27], a sophisticated pipeline that can make structure predictions for a diverse range of secondary metabolic pathways. While antiSMASH worked well on the two complete *Salinispora* genomes, NaPDoS consistently detected more KS domains in the draft genomes (Table S4). While this is not surprising given that NaPDoS is specifically designed for this purpose, it nonetheless highlights the value of the sequence tag approach when working with draft genome sequences that contain many unassembled contigs.

Metagenomic analyses

NaPDoS was further tested on metagenomic data sets generated from a Minnesota farm soil and whale fall [51]. While the numbers of KS domains detected in both datasets are similar (Table S5), removing redundant sequences reveals a higher diversity of KS domains in the soil sample. The majority of the whale fall KS domains were classified as FASs suggesting they are associated with fatty acid biosynthesis. In contrast, nearly half of the KS domains detected in the Minnesota farm soil appear to be involved in secondary metabolite biosynthesis. These results are in agreement with a previous study in which these datasets were manually screened for type I PKSs [38]. All of the sequences shared <70% identity to the reference database or NCBI BLAST matches associated with experimentally characterized pathways and thus no predictions could be made about the structures of the potential secondary metabolic products. None-the-less, the majority of the KS domains detected could be rapidly classified by NaPDoS. The incorporation of these domains into a phylogenetic tree containing the reference sequences led to the reclassification of some and the prediction that others are functionally distinct from KS domains (Tables S6 and S7). These sequences were likely detected due to the low stringency at which the NaPDoS BLAST analyses were performed on the meta-data and is a positive indication that the KS analysis was comprehensive. The reclassification of some sequences emphasizes the importance of incorporating phylogeny into the analyses.

Discussion

Rapid advances in DNA sequencing technologies are providing unprecedented opportunities to incorporate DNA sequence data into the natural product discovery process. The effective use of this information requires bioinformatic tools that can rapidly analyze large datasets in the context of a wide array of complex biosynthetic paradigms. While a number of excellent bioinformatic tools targeting secondary metabolism have been developed [22,25,27], they are largely predicated on accurate gene and operon assembly, something that has often proven challenging to obtain given the modular and highly repetitive nature of many genes involved in secondary metabolism [8]. This challenge can become especially problematic in the case of metagenomic analyses of complex microbial communities.

The Natural Product Domain Seeker (NaPDoS) is a web-based bioinformatic tool that was developed to detect and classify KS and C domains from a wide variety of sequence data. The use of domain sequence tags as proxies for the biosynthetic genes in which they reside is based on the well established and highly informative phylogentic relationships they maintain. These relationships form the foundation of the NaPDoS classification system and provide a rapid mechanism to delineate secondary metabolite biosynthetic gene richness and diversity within a genome or environmental sample. Short sequence tags (e.g. 600 bp) can be effectively analyzed using NAPDoS and thus minimum coverage, next generation sequence assemblies are well suited for this tool. The resulting estimates of biosynthetic potential can be used to guide more extensive sequencing efforts or targeted operon assembly. In cases where query sequences closely match domains derived from experimentally characterized biosynthetic pathways (eg., >90% sequence identity), it has proven possible to make accurate predictions about the structural class of the secondary metabolite(s) produced [37,63]. The low stringency of the HMM searches and the ability to adjust the internal BLAST parameters provides opportunities to detect more highly divergent KS and C domains associated with secondary metabolism as well as domains that are not associated with secondary metabolism (e.g. fatty acid biosynthesis) and thus all results should be carefully scrutinized. As the number of experimentally characterized biosynthetic pathways increases, this approach will provide an increasingly effective method to "de-replicate", i.e. to identify strains that have the greatest potential to produce known compounds.

There is ample evidence that the mechanistic diversity of polyketide and non-ribosomal peptide assembly is considerably greater than that currently recognized [14,64], and thus it can be expected that the NaPDoS classification system will need to evolve as new phylogenetic lineages are linked to specific biochemical functions and enzyme architectures. There is considerable preliminary evidence that the classes defined here will be further delineated once more experimentally characterized sequence data is obtained. For example, the current KS1 clade includes traditional starter KSs (KSQ) as well as domains from the salinosporamide (stro1024) and jamaicamide (JamE) pathways, which are involved in the incorporation of unusual extender units [8,65]. Likewise, the Type II clade includes deeply branching KS domains derived from CurC and JamG that are predicted to be involved in decarboxylation as opposed to condensation reactions [65,66]. A third example is the Iterative (a) class, which include traditional iterative KSs as well as those involved in the biosynthesis of polycyclic tetramate macrolactams [67]. Finally, the *trans*-AT (b) clade is comprised of KS sequenced derived from what appears to be an evolutionarily independent lineage of *trans*-AT sequences as well as genes associated with *beta*-branching [28,55]. Despite the potential oversimplification of the current classification system, it provides a useful method to estimate the numbers and functional types of biosynthetic genes present in complex data sets.

Despite poor assembly, a large number and diversity of KS and C domains were detected among the four draft *Salinispora* genome sequences. Seventeen of these domains were not observed in either of the two complete *Salinispora* genomes providing evidence of the considerable biosynthetic variability that may occur among closely related strains. In addition, two C domains fell into the "Modified AA" clade, a lineage whose biochemical function has yet to be experimentally characterized. While the metagenomic datasets revealed similar total numbers of KS domains, the classification of those domains revealed dramatic differences in functional types. Analyses such as these provide insight into the potential significance of secondary chemistry in mediating population and community dynamics while at the same time identifying environments that can be prioritized for secondary metabolite discovery efforts.

Traditional natural product discovery paradigms have become increasingly inefficient [68] and are rapidly moving towards approaches that capitalize on access to DNA sequence data [69]. NaPDoS is a publically available bioinformatic tool that capitalizes

on the well-established phylogenetic relationships of KS and C domains. It provides a rapid method to make informed interpretations of secondary metabolism based on small sequence tags extracted from a variety of data types including poorly assembled, next generation datasets. A major application of NaPDoS is the exploration of sequence space and the identification of new domain lineages, which have a high probability of being associated with new mechanisms of secondary metabolite biosynthesis. Prioritizing these lineages for experimental characterization will facilitate the discovery of new biochemistry and represents a rationale approach to secondary metabolite discovery.

At present, NaPDoS is optimized for the identification and classification of bacterial PKS and NRPS genes. Nonetheless, it is possible for NaPDoS to identify eukaryotic KS and C domains given their shared evolutionary history with prokaryotic homologs. The results obtained for non-bacterial sequences should be interpreted with caution however, as the reference database has not been adequately populated with these sequences to provide a robust classification system. Future plans include the expansion of NaPDoS to include additional eukaryotic sequences and subgroups within the FAS and PUFA lineages, the later of which were recently shown to cluster phylogenetically based on functional type [70]. Additional goals are to include type III PKSs, which were originally found in plants but are now known to occur in a wide range of bacteria [71]. These PKSs are distantly related to types I and II and thus will require a separate alignment and analysis pipeline. The inclusion of additional secondary metabolite families, such as terpenoids, alkaloids, and ribosomal peptides, is also conceivable.

Supporting Information

Table S1 NaPDoS derived KS and C domains from the *S. avermitilis* MA-4680 genome.
(DOC)

Table S2 NaPDoS results for six *Salinispora* genomes.
(DOC)

Table S3 KS and C domains detected in four draft *Salinispora* genomes.
(DOC)

Table S4 NaPDoS and antiSMASH-derived KS and C domains.
(DOCX)

Table S5 NaPDoS KS results for metagenomic data sets.
(DOC)

Table S6 KS domains detected in the whale fall metagenomic data set.
(DOC)

Table S7 KS domains detected in the Minnesota farm soil data set.
(DOC)

Author Contributions

Conceived and designed the experiments: NZ SP KP EA PJ. Performed the experiments: NZ SP KP. Analyzed the data: NZ SP KP JB EA PJ. Contributed reagents/materials/analysis tools: NZ SP KP JB EA PJ. Wrote the paper: NZ PJ.

References

1. Bentley SD, Chater KF, Cerdeno-Tarraga AM, Challis GL, Thomson NR, et al. (2002) Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). Nature 417: 141–147.
2. Ikeda H, Ishikawa J, Hanamoto A, Shinose M, Kikuchi H, et al. (2003) Complete genome sequence and comparative analysis of the industrial microorganism *Streptomyces avermitilis*. Nat Biotech 21: 526–531.
3. Challis GL (2008) Genome mining for novel natural product discovery. J Med Chem 51: 2618–2628.
4. Lautru S, Deeth RJ, Bailey LM, Challis GL (2005) Discovery of a new peptide natural product by *Streptomyces coelicolor* genome mining. Nat Chem Biol 1: 265–269.
5. Eustáquio AS, Nam S-J, Penn K, Lechner A, Wilson MC, et al. (2011) The discovery of salinosporamide K from the marine bacterium "*Salinispora pacifica*" by genome mining gives insight into pathway evolution. ChemBioChem 12: 61–64.
6. Hornung A, Bertazzo M, Dziarnowski A, Schneider K, Welzel K, et al. (2007) A genomic screening approach to the structure-guided identification of drug candidates from natural sources. ChemBioChem 8: 757–766.
7. Winter JM, Behnken S, Hertweck C (2011) Genomics-inspired discovery of natural products. Curr Opin Chem Biol 15: 22–31.
8. Udwary DW, Zeigler L, Asolkar RN, Singan V, Lapidus A, et al. (2007) Genome sequencing reveals complex secondary metabolome in the marine actinomycete *Salinispora tropica*. Proc Natl Acad Sci 104: 10376–10381.
9. Baker DD, Chu M, Oza U, Rajgarhia V (2007) The value of natural products to future pharmaceutical discovery. Nat Prod Rep 24: 1225–1244.
10. Newman DJ, Cragg GM (2007) Natural products as sources of new drugs over the last 25 years. Journal of Natural Products 70: 461–477.
11. McPherson JD (2009) Next-generation gap. Nat Methods 6: S2–5.
12. Hertweck C (2009) The biosynthetic logic of polyketide diversity. Angew Chem Int Ed Engl 48: 4688–4716.
13. Finking R, Marahiel MA (2004) Biosynthesis of non-ribosomal peptides. Annual Review of Microbiology 58: 453–488.
14. Shen B (2003) Polyketide biosynthesis beyond the type I, II and III polyketide synthase paradigms. Curr Opin Chem Biol 7: 285–295.
15. Weissman KJ (2004) Polyketide biosynthesis: understanding and exploiting modularity. Philosophical Transactions of the Royal Society of London Series A, Mathematical, Physical and Engineering Sciences 362: 2671–2690.
16. Lautru S, Challis GL (2004) Substrate recognition by nonribosomal peptide synthetase multi-enzymes. Microbiology 150: 1629–1636.
17. Sieber SA, Marahiel MA (2005) Molecular mechanisms underlying nonribosomal peptide synthesis: Approaches to new antibiotics. Chemical Reviews 105: 715–738.
18. Fischbach MA, Walsh CT (2006) Assembly-line enzymology for polyketide and nonribosomal peptide antibiotics: logic, machinery, and mechanisms. Chem Rev 106: 3468–3496.
19. Ansari MZ, Yadav G, Gokhale RS, Mohanty D (2004) NRPS-PKS: a knowledge-based resource for analysis of NRPS/PKS megasynthases. Nucleic Acids Res 32: W405–413.
20. Rausch C, Weber T, Kohlbacher O, Wohlleben W, Huson DH (2005) Specificity prediction of adenylation domains in nonribosomal peptide synthetases (NRPS) using transductive support vector machines (TSVMs). Nucleic Acids Res 33: 5799–5808.
21. Yadav G, Gokhale RS, Mohanty D (2009) Towards prediction of metabolic products of polyketide synthases: an in silico analysis. PLoS Comput Biol 5: e1000351.
22. Starcevic A, Zucko J, Simunkovic J, Long PF, Cullum J, et al. (2008) ClustScan: an integrated program package for the semi-automatic annotation of modular biosynthetic gene clusters and in silico prediction of novel chemical structures. Nucleic Acids Research 36: 6882–6892.
23. McAlpine JB, Bachmann BO, Piraee M, Tremblay S, Alarco AM, et al. (2005) Microbial genomics as a guide to drug discovery and structural elucidation: ECO-02301, a novel antifungal agent, as an example. Journal of Natural Products 68: 493–496.
24. Moss SJ, Martin CJ, Wilkinson B (2004) Loss of co-linearity by modular polyketide synthases: a mechanism for the evolution of chemical diversity. Natural Product Reports 21: 575–593.
25. Bachmann BO, Ravel J (2009) Methods for in silico prediction of microbial polyketide and nonribosomal peptide biosynthetic pathways from DNA sequence data. Methods Enzymol 458: 181–217.
26. Jenke-Kodama H, Dittmann E (2009) Bioinformatic perspectives on NRPS/PKS megasynthases: advances and challenges. Nat Prod Rep 26: 874–883.
27. Medema MH, Blin K, Cimermancic P, de Jager V, Zakrzewski P, et al. (2011) antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Nucleic Acids Research 39: W339–W346.
28. Nguyen T, Ishida K, Jenke-Kodama H, Dittmann E, Gurgui C, et al. (2008) Exploiting the mosaic structure of trans-acyltransferase polyketide synthases for natural product discovery and pathway dissection. Nat Biotechnol 26: 225–233.

29. Rausch C, Hoof I, Weber T, Wohlleben W, Huson DH (2007) Phylogenetic analysis of condensation domains in NRPS sheds light on their functional evolution. BMC Evol Biol 7: 78.
30. Eisen JA (1998) A phylogenomic study of the MutS family of proteins. Nucleic Acids Res 26: 4291–4300.
31. Eisen JA, Fraser CM (2003) Phylogenomics: intersection of evolution and genomics. Science 300: 1706–1707.
32. Metsa-Ketela M, Salo V, Halo L, Hautala A, Hakala J, et al. (1999) An efficient approach for screening minimal PKS genes from Streptomyces. FEMS Microbiol Lett 180: 1–6.
33. Moffitt MC, Neilan BA (2003) Evolutionary affiliations within the superfamily of ketosynthases reflect complex pathway associations. J Mol Evol 56: 446–457.
34. Jenke-Kodama H, Sandmann A, Müller R, Dittmann E (2005) Evolutionary implications of bacterial polyketide synthases. Mol Biol Evol 22: 2027–2039.
35. Ginolhac A, Jarrin C, Robe P, Perriere G, Vogel T, et al. (2005) Type I polyketide synthases may have evolved through horizontal gene transfer. J Mol Evol 60: 716–725.
36. Gontang EA, Fenical W, Jensen PR (2007) Phylogenetic diversity of gram-positive bacteria cultured from marine sediments. Appl Environ Microbiol 73: 3272–3282.
37. Freel KC, Nam S-J, Fenical W, Jensen PR (2011) Evolution of secondary metabolite genes in three closely related marine actinomycete species. Appl Environ Microbiol 77: 7261–7270.
38. Foerstner KU, Doerks T, Creevey CJ, Doerks A, Bork P (2008) A computational screen for type I polyketide synthases in metagenomics shotgun data. PLoS ONE 3: e3515.
39. Roongsawang N, Lim SP, Washio K, Takano K, Kanaya S, et al. (2005) Phylogenetic analysis of condensation domains in the nonribosomal peptide synthetases. FEMS Microbiol Lett 252: 143–151.
40. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792–1797.
41. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 25: 4876–4882.
42. Maddison WP, Maddison DR (2009) Mesquite: a modular system for evolutionary analysis. Version 2.71 ed.
43. Dereeper A, Guignon V, Blanc G, Audic S, Buffet S, et al. (2008) Phylogeny.fr: robust phylogenetic analysis for the non-specialist. Nucleic Acids Res 36: W465–469.
44. Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Systematic Biol 52: 696–704.
45. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. Journal of Molecular Biology 215: 403–410.
46. Finn RD, Mistry J, Tate J, Coggill P, Heger A, et al. (2009) The Pfam protein families database. Nucleic Acids Res 38: D211–222.
47. Junier T, Zdobnov EM. The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. Bioinformatics 26: 1669–1670.
48. Goldberg SMD, Johnson J, Busam D, Feldblyum T, Ferriera S, et al. (2006) A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. Proceedings of the National Academy of Sciences 103: 11240–11245.
49. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437: 376–380.
50. Penn K, Jenkins C, Nett M, Udwary DW, Gontang EA, et al. (2009) Genomic islands link secondary metabolism to functional adaptation in marine Actinobacteria. ISME J 3: 1193–1203.
51. Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, et al. (2005) Comparative metagenomics of microbial communities. Science 308: 554–557.
52. Jenke-Kodama H, Dittmann E (2009) Evolution of metabolic diversity: insights from microbial polyketide synthases. Phytochemistry 70: 1858–1866.
53. Ridley CP, Lee HY, Khosla C (2008) Evolution of polyketide synthases in bacteria. Proc Natl Acad Sci U S A 105: 4595–4600.
54. Jenke-Kodama H, Sandmann A, Muller R, Dittmann E (2005) Evolutionary implications of bacterial polyketide synthases. Mol Biol Evol 22: 2027–2039.
55. Piel J (2010) Biosynthesis of polyketides by trans-AT polyketide synthases. Natural Product Reports 27: 996–1047.
56. Jenke-Kodama H, Dittmann E (2005) Combinatorial polyketide biosynthesis at higher stage. Mol Syst Biol 1: 2005–0025.
57. Tillett D, Dittmann E, Erhard M, von Dohren H, Borner T, et al. (2000) Structural organization of microcystin biosynthesis in Microcystis aeruginosa PCC7806: an integrated peptide-polyketide synthetase system. Chem Biol 7: 753–764.
58. Du L, Sanchez C, Chen M, Edwards DJ, Shen B (2000) The biosynthetic gene cluster for the antitumor drug bleomycin from Streptomyces verticillus ATCC15003 supporting functional interactions between nonribosomal peptide synthetases and a polyketide synthase. Chem Biol 7: 623–642.
59. Nett M, Gulder TA, Kale AJ, Hughes CC, Moore BS (2009) Function-oriented biosynthesis of beta-lactone proteasome inhibitors in Salinispora tropica. J Med Chem 52: 6163–6167.
60. Eustaquio AS, McGlinchey RP, Liu Y, Hazzard C, Beer LL, et al. (2009) Biosynthesis of the salinosporamide A polyketide synthase substrate chloroethylmalonyl-coenzyme A from S-adenosyl-L-methionine. Proc Nat Acad Sci 106: 12295–12300.
61. Jensen PR, Williams PG, Oh DC, Zeigler L, Fenical W (2007) Species-specific secondary metabolite production in marine actinomycetes of the genus *Salinispora*. Appl Environ Microbiol 73: 1146–1152.
62. Cundliffe E, Bate N, Butler A, Fish S, Gandecha A, et al. (2001) The tylosin-biosynthetic genes of Streptomyces fradiae. Antonie Van Leeuwenhoek 79: 229–234.
63. Gontang E, GaudÍncio S, Fenical W, Jensen P (2010) Sequence-based analysis of secondary-metabolite biosynthesis in marine actinobacteria. Appl Environ Microbiol 76: 2487–2499.
64. Wenzel SC, Muller R (2005) Formation of novel secondary metabolites by bacterial multimodular assembly lines: deviations from textbook biosynthetic logic. Curr Opin Chem Biol 9: 447–458.
65. Edwards DJ, Marques BL, Nogle LM, McPhail K, Goeger DE, et al. (2004) Structure and biosynthesis of the Jamaicamides, new mixed polyketide-polypeptide neurotoxins from the marine cyanobacterium Lyngbya majuscula. Chemistry & Biology 11: 817–833.
66. Chang Z, Sitachitta N, Rossi JV, Roberts MA, Flatt PM, et al. (2004) Biosynthetic pathway and gene cluster analysis of Curacin A, an antitubulin natural product from the tropical marine cyanobacterium *Lyngbya majuscula*. Journal of Natural Products 67: 1356–1367.
67. Blodgett JAV, Oh D-C, Cao S, Currie CR, Kolter R, et al. (2010) Common biosynthetic origins for polycyclic tetramate macrolactams from phylogenetically diverse bacteria. Proceedings of the National Academy of Sciences 107: 11692–11697.
68. Li JW, Vederas JC (2009) Drug discovery and natural products: end of an era or an endless frontier? Science 325: 161–165.
69. Davies J (2011) How to discover new antibiotics: harvesting the parvome. Curr Opin Chem Biol 15: 5–10.
70. Shulse CN, Allen EE (2011) Widespread occurrence of secondary lipid biosynthesis potential in microbial lineages. PLoS One 6: e20146.
71. Moore BS, Hopke JN (2001) Discovery of a new bacterial polyketide biosynthetic pathway. Chembiochem 2: 35–38.