Atlas of nonribosomal peptide and polyketide biosynthetic pathways reveals common occurrence of nonmodular enzymes

Hao Wang[a,1], David P. Fewer[a], Liisa Holm[b], Leo Rouhiainen[a], and Kaarina Sivonen[a]

[a]Division of Microbiology and Biotechnology, Department of Food and Environmental Sciences and [b]Institute of Biotechnology and Department of Biosciences, Viikki Biocenter, University of Helsinki, FIN-00014, Helsinki, Finland

Edited by Susan S. Golden, University of California, San Diego, La Jolla, CA, and approved May 12, 2014 (received for review January 27, 2014)

Nonribosomal peptides and polyketides are a diverse group of natural products with complex chemical structures and enormous pharmaceutical potential. They are synthesized on modular nonribosomal peptide synthetase (NRPS) and polyketide synthase (PKS) enzyme complexes by a conserved thiotemplate mechanism. Here, we report the widespread occurrence of NRPS and PKS genetic machinery across the three domains of life with the discovery of 3,339 gene clusters from 991 organisms, by examining a total of 2,699 genomes. These gene clusters display extraordinarily diverse organizations, and a total of 1,147 hybrid NRPS/PKS clusters were found. Surprisingly, 10% of bacterial gene clusters lacked modular organization, and instead catalytic domains were mostly encoded as separate proteins. The finding of common occurrence of nonmodular NRPS differs substantially from the current classification. Sequence analysis indicates that the evolution of NRPS machineries was driven by a combination of common descent and horizontal gene transfer. We identified related siderophore NRPS gene clusters that encoded modular and nonmodular NRPS enzymes organized in a gradient. A higher frequency of the NRPS and PKS gene clusters was detected from bacteria compared with archaea or eukarya. They commonly occurred in the phyla of *Proteobacteria*, *Actinobacteria*, *Firmicutes*, and *Cyanobacteria* in bacteria and the phylum of *Ascomycota* in fungi. The majority of these NRPS and PKS gene clusters have unknown end products highlighting the power of genome mining in identifying novel genetic machinery for the biosynthesis of secondary metabolites.

biosynthetic gene cluster | data mining | distribution | bioactive compound

Nonribosomal peptides and polyketides are two diverse families of natural products with a broad range of biological activities and pharmacological properties (1). They include toxins, siderophores, pigments, antibiotics, cytostatics, and immunosuppressants (2, 3). Nonribosomal peptide and polyketide natural products have remarkably diverse structures and can be linear or cyclic or have branched structures (4). They can be further reengineered to produce complex products with exotic chemical structures and biological activities (5).

Nonribosomal peptides and polyketides are synthesized on large nonribosomal peptide synthetase (NRPS) and polyketide synthase (PKS) enzyme complexes, respectively. PKSs are currently classified into three types that differ in their organization of catalytic domains (6). Type I PKSs are large multidomain enzymes using a modular strategy, with each module being comprised of catalytic domains responsible for recognition, activation, and condensation of acyl-CoA (7). The catalytic sites of type II and type III PKSs are organized into separate proteins (6, 8). NRPSs are usually defined as modular multidomain enzymes (7). However, a nonmodular NRPS enzyme, a stand-alone peptidyl carrier protein (BlmI) from the bleomycin gene cluster, has been reported (9). Nonmodular NRPS enzymes are found in well-known siderophore biosynthetic pathways: e.g., EntE (adenylation), VibH (condensation), and VibE (adenylation) in enterobactin and vibriobactin clusters, respectively (10). The condensation, adenylation, and acyl carrier domains for brucebactin biosynthesis in *Brucella abortus* strain 2308 are encoded as fully separated proteins (11); thus, these NRPSs could be considered to have type II architecture. The arrangement of modules within the NRPS and type I PKS enzymes often determines the number and order of the monomer constituents of the product (12), despite deviations in module iteration (13, 14) and skipping (15). A growing number of gene clusters encoding both NRPSs and type I PKSs have been identified for biosynthesis of complex natural products (16).

The bulk of natural products in clinical use today come from a handful of bacterial and fungal lineages (17–21). However, genomics studies imply that the ability to make these compounds is much more widespread (22–24). To gain insights into the occurrence and distribution of the ability to produce nonribosomal peptides and polyketides, we undertook a systematic genome-mining study. Here, we show the widespread occurrence of NRPS and PKS genetic machinery across the three domains of life with the discovery of 3,339 gene clusters from 991 organisms, by examining a total of 2,699 genomes. Our data mining further revealed that more than half of the NRPS and type I PKS enzymes have a nonmodular composition. A total of 314 gene clusters that are comprised mostly of these nonmodular enzymes were discovered in noncanonical organizations, which deviate from the present definition.

Results

Widespread Distribution of NRPSs and Type I PKSs. Our survey demonstrated the widespread distribution of NRPS and type I

Significance

This study demonstrates the widespread distribution of nonribosomal peptide synthetase and modular polyketide synthase biosynthetic pathways across the three domains of life, by cataloging a total of 3,339 gene clusters from 2,699 genomes. Our analysis suggests that noncanonical nonmodular biosynthetic enzymes are common in bacteria. *Proteobacteria*, *Actinobacteria*, *Firmicutes*, and *Cyanobacteria* in bacteria and *Ascomycota* in fungi contained higher number of these gene clusters and are likely to produce a wide variety of nonribosomal peptide and polyketide types of natural products. The data generated here provide a basis for the exploration of nonribosomal peptide and polyketide biosynthetic capacity and present a compelling wealth of new information for natural product discovery.

Author contributions: H.W. and D.P.F. designed research; H.W. performed research; K.S. contributed new reagents/analytic tools; H.W., D.P.F., and L.H. analyzed data; and H.W., D.P.F., L.R., and K.S. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. E-mail: wang.hao@helsinki.fi.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1401734111/-/DCSupplemental.

MICROBIOLOGY

Table 1. Summary of NRPS/PKS gene clusters and genomes analyzed in this study

	Gene cluster				Genome	
Domain	No.	No. of proteins	No. of domains	Cumulative size, Mb	No.	Cumulative size, Gb
Bacteria	2,976	15,889	56,269	95.88	2,478	8.69
Archaea	3	6	27	0.03	160	0.38
Eukarya	360	699	3,304	6.44	61	6.65
Total	3,339	16,594	59,600	102.35	2,699	15.72

PKS biosynthetic pathways in all three domains of life (Table 1 and Fig. 1). A total of 3,339 NRPS, PKS, and hybrid NRPS/PKS gene clusters, which amount to 102.35 Mb in size, were discovered by mining of 15.72 Gb of genomic sequences from 2,699 organisms (Table 1 and Datasets S1 and S2). The majority of the gene clusters (2,976, 89%) were detected in bacteria. They were less frequent in eukarya and rare in archaea.

NRPS and PKS gene clusters show a nonuniform distribution in bacteria (Fig. 1 and Table S1). They are common in the phyla *Proteobacteria*, *Actinobacteria*, *Firmicutes*, and *Cyanobacteria*. However, the gene clusters are noticeably absent in some lineages with small genomes, such as the phyla of *Tenericutes* and *Spirochaetes*, which have many sequenced genomes of intracellular pathogens. In addition, we showed that *Chlamydiae*, *Deinococcus-Thermus*, *Chlorobi*, *Verrucomicrobia*, *Nitrospirae*, and *Elusimicrobia* could be potential producers of nonribosomal peptides and polyketides. There appeared a 2-Mb threshold for genomes with the gene clusters (Fig. S1). Genome size seems not to be the only requirement because some bacteria with large genomes (>8 Mb) still lack these gene clusters (Fig. S1). We found that the average numbers of clusters are highly correlated ($P < 0.00001$) with the average genome sizes among strains in the orders of *Proteobacteria*, *Actinobacteria*, *Firmicutes*, and *Cyanobacteria* (Fig. S2). NRPS and PKS gene clusters were found throughout the eukaryotic lineages. Fungi, known producers of nonribosomal peptides and polyketides, were a rich source of NRPSs and type I PKSs. A total of 307 gene clusters were located from 12 strains in the phylum *Ascomycota* (Table S2). However, no NRPS or PKS gene clusters were found in the other two *Microsporidia* and three *Basidiomycota* strains (Fig. 1). NRPS and PKS gene clusters were also distributed sporadically in many other lineages of protist, plant, and animal (Fig. 1). Both NRPSs and PKSs were found from phyla of *Amoebozoa*, *Dinoflagellata*, *Apicomplexa*, *Nematoda*, *Arthropoda*, and *Mollusca*. In addition, we also located NRPS gene clusters in phyla of alga *Stramenopiles* and *Chlorophyta*, as well as PKSs from slime-molds and metazoan phyla of *Annelida* and *Chordata*. Archaea appears to rarely contain these biosynthetic pathways in the 128 genomes analyzed in this study. Only three NRPS gene clusters were found in two strains, which belong to classes *Methanobacteria* and *Methanomicrobia*, respectively (Fig. 1). Meanwhile, modular PKSs were absent from these sequenced archaean genomes.

Overall, despite biases in the number of genomes available for each lineage, NRPS and type I PKS gene clusters were more frequent in the phyla of *Proteobacteria*, *Firmicutes*, *Actinobacteria*, and *Cyanobacteria* in bacteria (Table S1) and *Ascomycota* in fungi (Table S2). The NRPS and PKS gene clusters identified in this study can be accessed from an associated web database (http://npgc.biocenter.helsinki.fi).

Gene-Cluster Composition Reveals High Diversity of NRPS and PKS Biosynthetic Machineries and Their Close Relationship. A total of 3,339 gene clusters were detected in this genome mining study. By examining the composition of the biosynthetic and tailoring enzymes of these gene clusters, we collectively obtained 59,600 domains (Table 1), among which the core biosynthetic domains accounted for the majority (Fig. S3).

These gene clusters were classified into NRPS, PKS, and hybrid types, according to the presence of core domains of NRPS, PKS, or both systems (Fig. 2). One third (1,147, 34.4%) of gene clusters belonged to the hybrid type and encoded 462 hybrid proteins that contain both NRPS and PKS core domains (Fig. 2 and Dataset S3). The hybrid clusters tend to be larger and possess more domains, compared with stand-alone NRPS and PKS gene clusters (Fig. S4). Therefore, they may produce more complex products than NRPS and PKS clusters on their own. Moreover, nearly all domain types are presented in hybrid gene clusters, especially in bacteria where domains from hybrid gene clusters are more frequently observed than in fungi (Fig. S3). For example, a total of 113 bacterial gene clusters were found with free-standing acyltransferases (AT), which iteratively load the monomers *in trans* for other modular PKSs lacking an AT domain (25). They are almost all the hybrid type, illustrated by the fact that their docking and dehydratase domains are entirely from hybrid clusters (Fig. S3). Among the 298 AT-less PKS module-containing enzymes found from these gene clusters, 117 are hybrid enzymes with NRPS modules. Interestingly, we also found three enzymes in which the normal PKS module and the AT-less module are fused

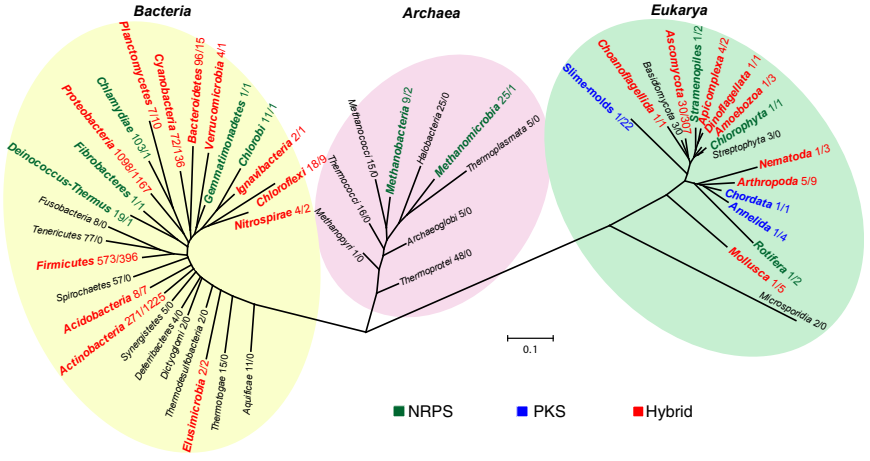

Fig. 1. The widespread distribution of NRPSs and PKSs across the three domains of life. The phylogenetic analysis is based on 16S or 18S rRNA genes from selected organisms (Table S3) for representative phyla in bacteria and eukarya, and classes in archaea. The midpoint tree was constructed by PhyML 3.0 using the GTR substitution model and with 100 bootstrap replications for each branch. The lineages containing both NRPSs and PKSs, or hybrid NRPS-PKS enzymes are indicated in red, the ones containing only NRPSs are indicated in green, and those containing only PKSs are in blue. The numbers of examined genomes and discovered gene clusters for each phylum or class are next to the taxon name and separated by a slash. The biosynthetic pathways of NRPS and modular PKS not only were found densely distributed among bacterial phyla and fungi, but also were found in animals, plants, and protists in eukarya and archaean strains.

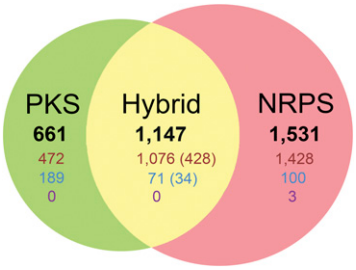

Fig. 2. A Venn diagram of PKS, NRPS, and hybrid gene-cluster numbers. The gene-cluster numbers of bacteria, archaea, and eukarya are shown in red, purple, and blue, respectively. The values in parentheses represent the numbers of hybrid enzymes that contain both NRPS and PKS core domains.

together; one of them (SBI_00671) even has a middle NRPS module residing between normal and AT-less PKS modules.

The NRPS/PKS gene clusters displayed highly diverse organizational architectures, based on their domain and protein composition (Fig. S5). The number of biosynthetic enzymes and domains of these gene clusters was mostly fewer than 10 and 20, respectively. On the other hand, some hybrid clusters coded for over 100 domains (Fig. S4). Based on our data, the largest PKS (MULP_065, 17,019 aa) has nine modules constituting 47 domains whereas the largest NRPS (plu2670, 16,367 aa) was found with 15 modules consisting of 46 domains, and the longest hybrid enzyme (MXAN_3779, 14,274 aa) possesses 39 domains fused into 1 PKS and 11 NRPS modules. These enzymes are in a highly modular structure and may represent the upper limits of the NRPS and PKS modularity level.

Evolutionary Analysis of NRPS Biosynthetic Enzymes. Phylogenetic analysis was conducted based on the sequences of the most abundant condensation domain type $^{L}C_{L}$, which catalyzes the formation of a peptide bond between two adjacent L-amino acids (Fig. S6). The condensation domains from archaea, bacteria, fungi, and other eukaryotic organisms did not form monophyletic clades and were instead mixed. A heat map showing similarities between L-amino acid condensation domains among the major lineages was constructed based on 4,087 $^{L}C_{L}$ domain sequences (Fig. 3). Homologies could be evidently observed within the lineages of *Actinobacteria*, *Cyanobacteria*, *Firmicutes*, *Deltaproteobacteria*,and *Gammaproteobacteria*. The high levels of similarity (indicated by the regions in blue and red in Fig. 3) clearly demonstrate events of inner lineage duplication. Moreover, the data also suggest horizontal transfers among these lineages: for example, between *Cyanobacteria* and *Proteobacteria*.

Common Occurrence of Nonmodular Biosynthetic Enzymes in Bacteria. There are a total of 15,889 biosynthetic and tailoring enzymes identified from the 2,976 bacterial gene clusters (Table 1). Surprisingly, nonmodular enzymes showed an extraordinary abundance (8,906) compared with others with multiple domains (Fig. 4). Among these nonmodular enzymes, 4,012 are from NRPS gene clusters, 1,086 from PKS, and 3,808 from hybrid clusters. Nearly every domain type was found as a nonmodular enzyme (Fig. S7). Some modification domains of aminotransferase and polyketide cyclase were found almost exclusively as separate proteins. Once again, many of the nonmodular enzymes are also present in hybrid clusters, such as the AT proteins from trans-AT clusters (Fig. S7).

Most of the nonmodular enzymes are found together with multidomain enzymes in gene clusters. However, one fifth of them (1,803) occurred in close genomic vicinity with each other and form into gene clusters (Fig. 5). Our data mining located a total of 314 (10.6%) bacterial NRPS and PKS gene clusters that are comprised mostly of nonmodular enzymes (Dataset S4; also indicated by the blue circle in Fig. S5). These gene clusters are structurally similar to type II PKSs (26). The majority (260,

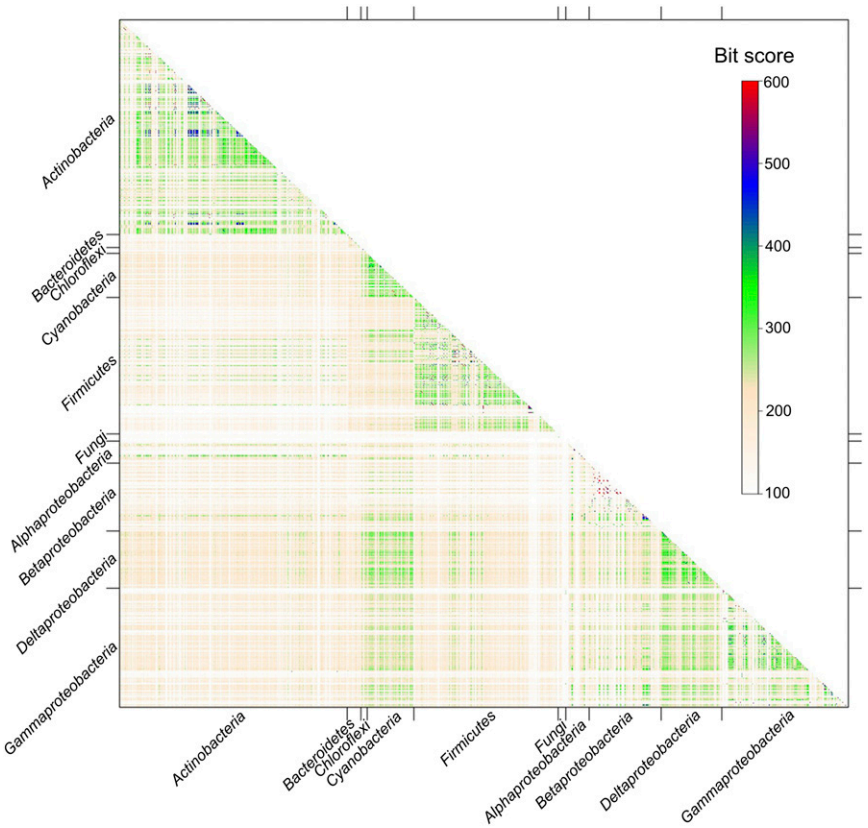

Fig. 3. A heat map showing sequence similarities between L-amino acid condensation domains among the major lineages. Sequence similarities are measured by the bit scores of the reciprocal bla2seq alignments and indicated in color as the scheme shows. Bit scores are shown in red if they equal 600 or more and in white if they equal 100 or less. The self-hits in the diagonal line were omitted for clarity.

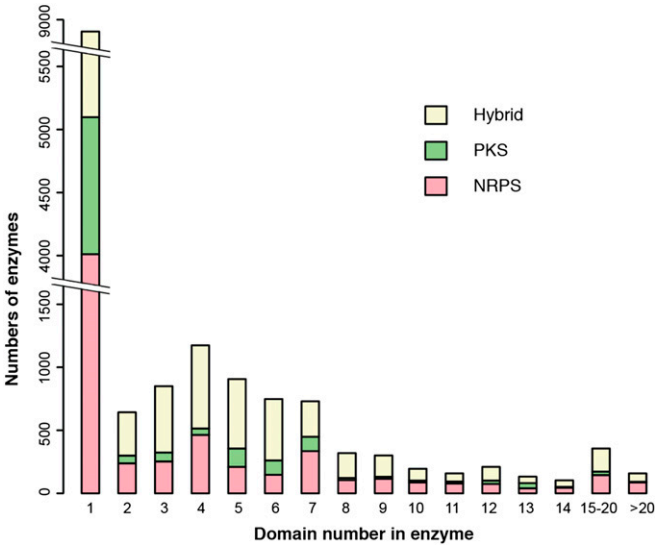

Fig. 4. Distribution of bacterial NRPS/PKS enzymes ranked according to their domain numbers. A total of 15,889 enzymes were found in 2,976 gene clusters in bacteria. These enzymes were grouped according to the number of their domains. The nonmodular enzymes showed high levels of abundance (8,906) compared with other multidomain enzymes. Enzymes from NRPS gene clusters are indicated in red, PKS gene clusters in green, and hybrid gene clusters in yellow.

82.8%) of them are the NRPS type. There are also 37 hybrid clusters, in which NRPS and type I PKS nonmodular enzymes were present together (Fig. 5), in addition to 17 PKS clusters. Although these type II-like gene clusters were found in multiple lineages of bacteria (Dataset S4), many of the NRPS clusters are from proteobacteria and appear to be related to siderophore biosynthesis. For example, a total of 16 putative brucebactin gene clusters, in which the condensation, adenylation, and carrier domains are presented as separated proteins, were found from sequenced *Brucella* genomes.

Phylogenetic Analysis of the Siderophore Biosynthesis NRPS Machineries with Gradient-Domain Organizations. Our analysis showed that the domains of NRPS and type I PKS display gradient organizations among gene clusters, from fully dissociated nonmodular enzymes to a highly modular mode in massive multidomain enzymes. For example, the iron-chelating siderophore acinetobactin biosynthesis gene clusters in *Acinetobacter baumannii* were found with many nonmodular enzymes and grouped as type II-like gene clusters in this study (Dataset S4). The mono- and didomain enzymes of these gene clusters can be mapped to the six-domain VibF in the vibriobactin biosynthesis gene cluster (27). Recently, the well-conserved tandem heterocyclization domains were also found in the siderophore serratiochelin gene clusters (28). These siderophores share one or more 2,3-dihydroxypheny-5-methyloxazoline-acyl groups, which is synthesized by the domains of modular VibF in *Vibrio cholerae* (29), and also likely by similar domains but organized as SchF1, -2, and -3 in *Serratia* (28) and BsaD, -A, and -B in *A. baumannii* (27) (Fig. 6). Although the biosynthesis of parabactin in *Paracoccus denitrificans* is still unknown, its 2-hydroxypheny-5-methyloxazoline-acyl group appears to be synthesized by the shown modular enzyme (Fig. 6). Gene clusters with similar domains were also detected from strains of *Pseudomonas*, *Marinomonas*, *Halomonas*, and *Agrobacterium*, which may also produce siderophores with similar structures (Fig. 6). The first condensation domain C1 of VibF was missing from some gene clusters, likely due to its dispensability in biosynthesis (29).

A neighbor-joining tree was constructed based on the C2 domains of VibF in *V. cholerae* and their homologs from other siderophore gene clusters (Fig. 6), which possess similar domains that, however, arranged in a gradient of organizations, from fully modular enzymes to nonmodular separate proteins (Fig. 6). The phylogeny of the C2 domains is congruent with the gradient-domain organizations of these siderophore biosynthetic pathways.

Discussion

Nonribosomal peptides and polyketides account for a significant portion of known natural products, which are the major source of drug candidates (30). Over 80% of these natural products were derived from fungi and actinomycetes (31). Recent studies showed the rich sources of nonribosomal peptides and polyketides in bacterial lineages of myxobacteria (18), pseudomonads (32), cyanobacteria (17), and streptomycetes (33). Genomic analysis further revealed the presence of NRPSs and PKSs in archaea (22), metazoa (23), and dinoflagellates (24). In this systematic study, we have demonstrated that NRPS and type I PKS biosynthetic pathways are widely distributed in bacteria and found sporadically in archaea and eukarya (Fig. 1 and Table 1). We have shown the high frequency of these gene clusters in bacteria and fungi and their spreading into specific archaean classes and many eukaryotic phyla. The accumulation of NRPS and PKS gene clusters in some bacteria phyla may partly reflect the biased distribution of sequenced genomes toward pathogens and environmentally important strains (19). Nevertheless, there is no doubt that this distribution will be inevitably expanded, because of the exponential increasing of genome data.

The archaean and eukaryotic NRPSs and PKSs appeared to be acquired from bacteria via horizontal gene transfer according to sequence and phylogenetic analysis in this study (Fig. S6) and others (23, 34). However, more evidence is needed to prove whether they are recently spread from bacteria or are remnants of ancient pathways after genome reduction. On the other hand, sequence analysis indicates that the evolution of nonribosomal

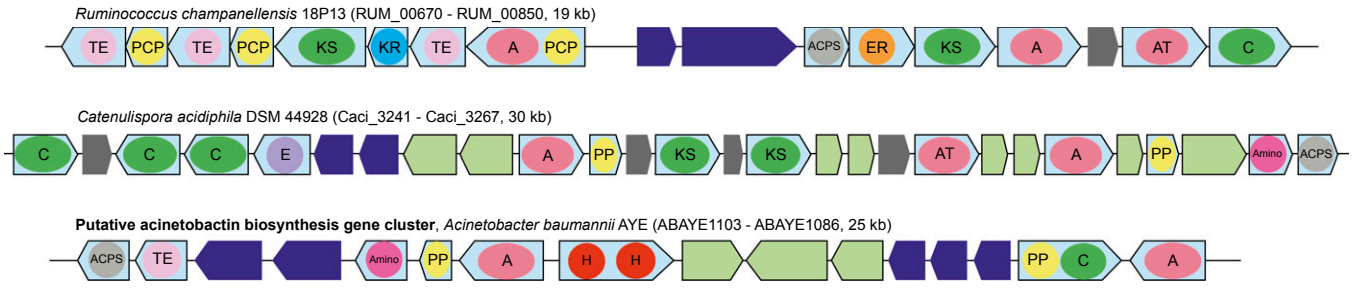

Fig. 5. Examples of gene clusters comprised of nonmodular enzymes. A total of 314 gene clusters formed mostly by nonmodular enzymes were found in bacteria. Examples of two hybrid gene clusters and a putative acinetobactin biosynthesis gene cluster are shown. The domains are indicated by abbreviations as adenylation (A), peptidyl carrier domain (PCP), condensation (C), acyltransferase (AT), acyl carrier or peptidyl carrier domain (PP), ketosynthase (KS), thioesterase (TE), epimerization (E), heterocyclization (H), ketoreductase (KR), enoylreductase (ER), aminotransferase (Amino), and 4′-phosphopantetheinyl transferase (ACPS). The ABC transport system proteins are indicated in blue, other tailoring enzymes in light green, and hypothetical proteins in gray.

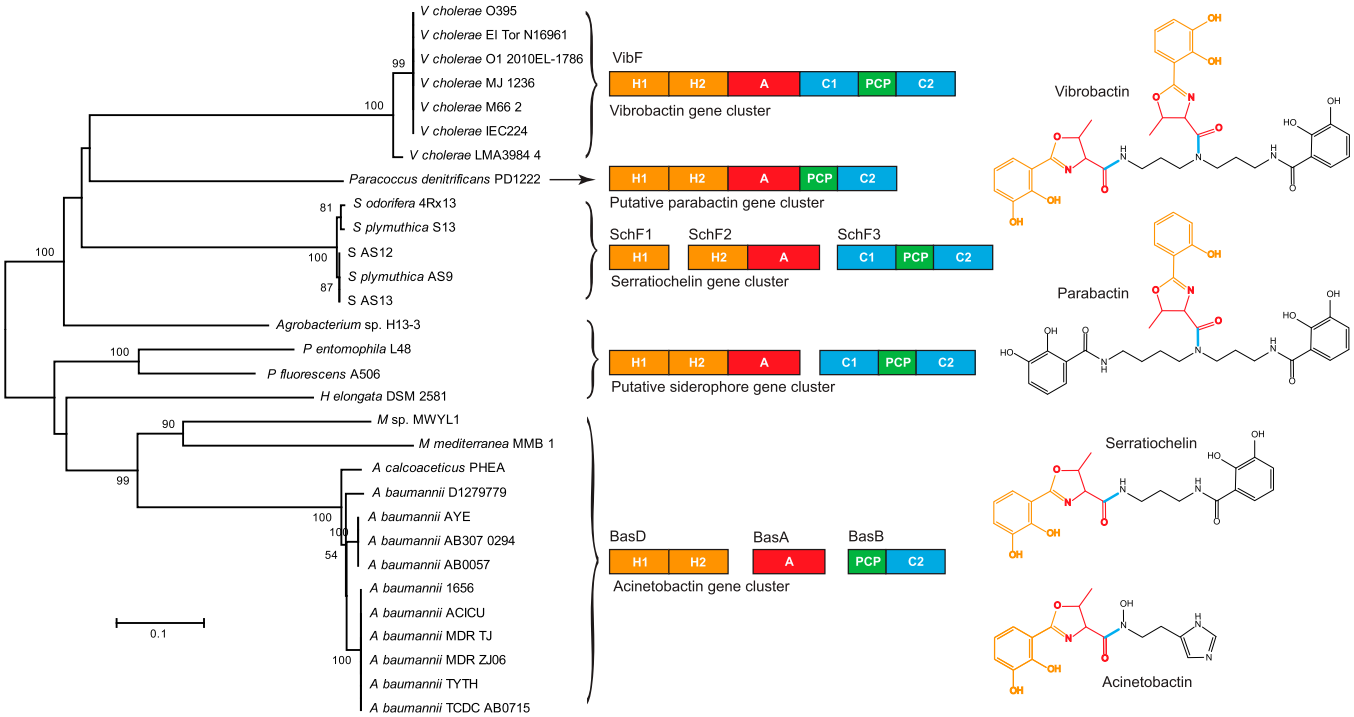

Fig. 6. Phylogenetic analysis of condensation domains C2 from siderophore biosynthetic enzymes with gradient-domain organizations. The neighbor-joining tree was constructed by MEGA5.1 (46) using Poisson model and with 100 bootstrap replications for each branch. These siderophore biosynthetic pathways share an unusual pair of tandem heterocyclization domains and others with similar composition but in gradient organizations, which are congruent with the phylogeny of the C2 domains. These domains are responsible for the biosynthesis of the 2,3-dihydroxypheny-5-methyloxazoline-acyl group, which is synthesized by the double heterocyclization (H) domains from a 2,3-dihydroxybenzoate and an L-threonine that is activated by the adenylation (A) domain, and fused by the condensation (C) domain C2 with other substrates. The activated L-threonine and the derived group are both tethered by the peptidyl carrier (PCP) domain.

machineries is consistent with the phylogenetic classifications based on ribosomal sequences (Fig. 3). In addition, the plasmid-borne gene clusters (Table S1) might be associated with horizontal gene transfer events in bacteria.

A previous study analyzing 141 strains implied that bacterial strains with PKSs usually have genomes larger than 2 Mb (26). Based on a much larger dataset, our analysis conclusively indicates that this threshold applies to both NRPSs and PKSs. We therefore suggest that bacterial strains with genomes fewer than 2 Mb may not be able to sustain these secondary metabolic pathways, perhaps due to the extra metabolic burdens. *Myxococcales* (proteobacteria), *Actinomycetales* (actinobacteria), *Pleurocapsales*, and *Nostocales* (cyanobacteria) are rich sources of NRPS and PKS gene clusters (Fig. S2); thus, new genomes from these orders would most likely provide novel natural products.

NRPSs and type I PKSs produce distinct groups of natural products with different enzyme systems, despite their common assembly line-like mechanism (7). Surprisingly, a high number of hybrid NRPS-PKS gene clusters were found in this study (Fig. 2), and all NRPS/PKS domain types are presented in hybrid gene clusters (Fig. S3). These analyses demonstrated their close relationship, which might be an outcome of long-term convergent interaction of the two systems and might yield a higher magnitude of structural diversity for the derived hybrid peptide-polyketide compounds that could be exploited as potential biological agents and drug leads (35, 36).

By definition, type I PKSs are modular enzyme systems with multidomain organization (7) whereas type II and III PKSs are nonmodular enzymes (6, 8). In contrast, NRPSs are defined only in a modular synthetic scheme (7) although there are occurrences of many NRPSs with a single domain (9, 10). In this study, we revealed that such nonmodular enzymes of NRPS and type I PKS are abundantly present in bacteria (Fig. 4), including both core and tailoring domains of the two systems (Fig. S7). Moreover, our analysis showed that 18% (260 out of 1,428) of NRPS gene clusters in bacteria were found in a type II-like organization (Dataset S4). Unlike NRPS-like enzymes (37, 38), these type II-like NRPS clusters have a complete core domain set, such as the identified putative brucebactin gene clusters that contain one fully dissociated NRPS module. Previously, reclassification of PKSs had been proposed (39), due to the discoveries of transition states between type I and II PKSs, the trans-AT PKS systems (25), and that between type II and III PKSs shown to be independent of the acyl carrier protein (40). The type II-like PKS clusters found in this study also appeared as transition states between type I and type II PKSs. They are more similar to type II PKSs than the trans-AT PKS systems, which have only a discrete AT enzyme. Note that NRPS/PKS hybrids also occurred in these clusters (Fig. 5). Consequently, the structural diversity of these enzyme systems presented here suggests that reclassification of NRPSs and PKSs is needed.

In conclusion, we have demonstrated the power of genome mining in studying natural product biosynthesis by showing the widespread distribution of NRPS/PKS gene clusters and by the finding of previously unidentified pathways. Our analysis showed that these various domain architectures from nonmodular to modular organization are evolutionarily related, indicating their similar enzymology. These type II-like gene clusters further expand the spectrum of the known NRPS and PKS machineries, which could be used in the future for natural product research.

Materials and Methods

Datasets. A total of 2,699 genomes across three domains of life (Dataset S1) were downloaded from the National Center for Biotechnology Information genome database (ftp://ftp.ncbi.nlm.nih.gov/genomes/). To obtain reliable information, only completed genomes of bacteria, archaea, fungi, and protist were studied. Repeatedly sequenced genomes were detected by examining the strain codes, and only the newest versions were analyzed. For higher

eukaryotic organisms, partial genomes were used because there are no complete genomes available.

Gene-Cluster Finding. Two stand-alone tools for secondary metabolite biosynthetic enzyme detection were used: 2metDB (41) and antiSMASH (42) version 2.0.2. FASTA format protein sequences were used as input of 2metDB whereas GenBank format files were queried by antiSMASH. Gene-cluster organizations were determined by antiSMASH with a 20-kb genomic span. Their results were combined, clusters found by antiSMASH were compared with the hits of 2metDB output, and the clusters predicted by both tools were counted. The gene clusters were determined as the NRPS type only if the adenylation or condensation domains were present, as the PKS type only if the ketosynthase or acyltransferase domains were found, and the hybrid type if they contained both the PKS and NRPS core domains. The type II-like gene clusters were defined as at least half of the biosynthetic enzymes are non-modular enzymes and the ratio of detected domain number versus protein number was no more than 1.5. We created a web database (http://npgc.biocenter.helsinki.fi) to deposit all identified gene clusters, which can be queried by taxonomic groups, organism names, or user-defined criteria.

Tree Building. The rRNA sequences of representative species from the three domains of life (Table S3) were aligned by MUSCLE (43) with default parameters, and nonconserved positions were removed by Gblocks (44) (0.91b) with stringent criteria. The midpoint tree was constructed by PhyML 3.0 (45) using the GTR substitution model and with 100 bootstrap replications for each branch. The condensation domain sequences were aligned by MUSCLE (43) with default parameters. The neighbor-joining tree was constructed by MEGA5.1 (46) using Poisson model and with 100 bootstrap replications for each branch.

Heat-Map Construction. A homologous map was constructed according to the bit scores of the reciprocal bla2seq results of the $^{L}C_{L}$ domain sequences, whose lengths were examined, and the ones larger or smaller than the average size over 10% were excluded.

ACKNOWLEDGMENTS. We thank Dr. Jouni Jokela for helpful comments on the manuscript. This work was financially supported by Academy of Finland Grants 258827 and 118637 (to K.S.) and by the Viikki Doctoral Programme in Molecular Biosciences in years 2010–2012 (H.W.).

1. Cane DE, Walsh CT, Khosla C (1998) Harnessing the biosynthetic code: Combinations, permutations, and mutations. *Science* 282(5386):63–68.
2. Finking R, Marahiel MA (2004) Biosynthesis of nonribosomal peptides1. *Annu Rev Microbiol* 58:453–488.
3. Weissman KJ, Leadlay PF (2005) Combinatorial biosynthesis of reduced polyketides. *Nat Rev Microbiol* 3(12):925–936.
4. Kopp F, Marahiel MA (2007) Where chemistry meets biology: the chemoenzymatic synthesis of nonribosomal peptides and polyketides. *Curr Opin Biotechnol* 18(6): 513–520.
5. Walsh CT (2008) The chemical versatility of natural-product assembly lines. *Acc Chem Res* 41(1):4–10.
6. Shen B (2003) Polyketide biosynthesis beyond the type I, II and III polyketide synthase paradigms. *Curr Opin Chem Biol* 7(2):285–295.
7. Fischbach MA, Walsh CT (2006) Assembly-line enzymology for polyketide and nonribosomal peptide antibiotics: Logic, machinery, and mechanisms. *Chem Rev* 106(8): 3468–3496.
8. Yu D, Xu F, Zeng J, Zhan J (2012) Type III polyketide synthases in natural product biosynthesis. *IUBMB Life* 64(4):285–295.
9. Du L, Shen B (1999) Identification and characterization of a type II peptidyl carrier protein from the bleomycin producer *Streptomyces verticillus* ATCC 15003. *Chem Biol* 6(8):507–517.
10. Crosa JH, Walsh CT (2002) Genetics and assembly line enzymology of siderophore biosynthesis in bacteria. *Microbiol Mol Biol Rev* 66(2):223–249.
11. Bellaire BH, et al. (2003) Genetic organization and iron-responsive regulation of the *Brucella abortus* 2,3-dihydroxybenzoic acid biosynthesis operon, a cluster of genes required for wild-type virulence in pregnant cattle. *Infect Immun* 71(4):1794–1803.
12. Marahiel MA, Stachelhaus T, Mootz HD (1997) Modular peptide synthetases involved in nonribosomal peptide synthesis. *Chem Rev* 97(7):2651–2674.
13. Mootz HD, Schwarzer D, Marahiel MA (2002) Ways of assembling complex natural products on modular nonribosomal peptide synthetases. *ChemBioChem* 3(6):490–504.
14. Fisch KM (2013) Biosynthesis of natural products by microbial iterative hybrid PKS-NRPS. *RSC Adv* 3(40):18228–18247.
15. Wenzel SC, Müller R (2005) Formation of novel secondary metabolites by bacterial multimodular assembly lines: deviations from textbook biosynthetic logic. *Curr Opin Chem Biol* 9(5):447–458.
16. Du L, Shen B (2001) Biosynthesis of hybrid peptide-polyketide natural products. *Curr Opin Drug Discov Devel* 4(2):215–228.
17. Welker M, von Döhren H (2006) Cyanobacterial peptides - nature's own combinatorial biosynthesis. *FEMS Microbiol Rev* 30(4):530–563.
18. Wenzel SC, Müller R (2007) Myxobacterial natural product assembly lines: fascinating examples of curious biochemistry. *Nat Prod Rep* 24(6):1211–1224.
19. Donadio S, Monciardini P, Sosio M (2007) Polyketide synthases and nonribosomal peptide synthetases: the emerging view from bacterial genomics. *Nat Prod Rep* 24(5): 1073–1109.
20. Doroghazi JR, Metcalf WW (2013) Comparative genomics of actinomycetes with a focus on natural product biosynthetic genes. *BMC Genomics* 14:611.
21. Bushley KE, et al. (2013) The genome of *tolypocladium inflatum*: evolution, organization, and expression of the cyclosporin biosynthetic gene cluster. *PLoS Genet* 9(6): e1003496.
22. Leahy SC, et al. (2010) The genome sequence of the rumen methanogen *Methanobrevibacter ruminantium* reveals new possibilities for controlling ruminant methane emissions. *PLoS ONE* 5(1):e8926.
23. Gladyshev EA, Meselson M, Arkhipova IR (2008) Massive horizontal gene transfer in bdelloid rotifers. *Science* 320(5880):1210–1213.
24. López-Legentil S, Song B, DeTure M, Baden DG (2010) Characterization and localization of a hybrid non-ribosomal peptide synthetase and polyketide synthase gene from the toxic dinoflagellate *Karenia brevis*. *Mar Biotechnol (NY)* 12(1):32–41.
25. Cheng YQ, Tang GL, Shen B (2003) Type I polyketide synthase requiring a discrete acyltransferase for polyketide biosynthesis. *Proc Natl Acad Sci USA* 100(6):3149–3154.
26. Jenke-Kodama H, Sandmann A, Müller R, Dittmann E (2005) Evolutionary implications of bacterial polyketide synthases. *Mol Biol Evol* 22(10):2027–2039.
27. Mihara K, et al. (2004) Identification and transcriptional organization of a gene cluster involved in biosynthesis and transport of acinetobactin, a siderophore produced by *Acinetobacter baumannii* ATCC 19606^{T}. *Microbiology* 150:2587–2597.
28. Seyedsayamdost MR, et al. (2012) Mixing and matching siderophore clusters: structure and biosynthesis of serratiochelins from *Serratia* sp. V4. *J Am Chem Soc* 134(33): 13550–13553.
29. Marshall CG, Hillson NJ, Walsh CT (2002) Catalytic mapping of the vibriobactin biosynthetic enzyme VibF. *Biochemistry* 41(1):244–250.
30. Walsh CT (2004) Polyketide and nonribosomal peptide antibiotics: modularity and versatility. *Science* 303(5665):1805–1810.
31. Donadio S, et al. (2005) Sources of polyketides and non-ribosomal peptides. *Biocombinatorial Approaches for Drug Finding*, eds Wohlleben W, Spellig T, Müller-Tiemann B (Springer, Heidelberg), pp 19–41.
32. Gross H, Loper JE (2009) Genomics of secondary metabolite production by *Pseudomonas* spp. *Nat Prod Rep* 26(11):1408–1446.
33. Weber T, Welzel K, Pelzer S, Vente A, Wohlleben W (2003) Exploiting the genetic potential of polyketide producing streptomycetes. *J Biotechnol* 106(2-3):221–232.
34. Lawrence DP, Kroken S, Pryor BM, Arnold AE (2011) Interkingdom gene transfer of a hybrid NPS/PKS from bacteria to filamentous Ascomycota. *PLoS ONE* 6(11):e28231.
35. Cane DE, Walsh CT (1999) The parallel and convergent universes of polyketide synthases and nonribosomal peptide synthetases. *Chem Biol* 6(12):R319–R325.
36. Du L, Sánchez C, Shen B (2001) Hybrid peptide-polyketide natural products: biosynthesis and prospects toward engineering novel molecules. *Metab Eng* 3(1):78–95.
37. Balskus EP, Walsh CT (2010) The genetic and molecular basis for sunscreen biosynthesis in cyanobacteria. *Science* 329(5999):1653–1656.
38. Forseth RR, et al. (2013) Homologous NRPS-like gene clusters mediate redundant small-molecule biosynthesis in *Aspergillus flavus*. *Angew Chem Int Ed Engl* 52(5): 1590–1594.
39. Müller R (2004) Don't classify polyketide synthases. *Chem Biol* 11(1):4–6.
40. Kwon HJ, Smith WC, Xiang L, Shen B (2001) Cloning and heterologous expression of the macrotetrolide biosynthetic gene cluster revealed a novel polyketide synthase that lacks an acyl carrier protein. *J Am Chem Soc* 123(14):3385–3386.
41. Bachmann BO, Ravel J (2009) Chapter 8. Methods for *in silico* prediction of microbial polyketide and nonribosomal peptide biosynthetic pathways from DNA sequence data. *Methods Enzymol* 458:181–217.
42. Medema MH, et al. (2011) antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Res* 39(Web Server issue):W339–W346.
43. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32(5):1792–1797.
44. Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* 17(4):540–552.
45. Guindon S, et al. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* 59(3):307–321.
46. Tamura K, et al. (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* 28(10):2731–2739.