An Integrated Metabolomic and Genomic Mining Workflow to Uncover the Biosynthetic Potential of Bacteria

Månsson, Maria; Vynne, Nikolaj Grønnegaard; Klitgaard, Andreas; Rasmussen, Jane Lind Nybo; Melchiorsen, Jette; D. Nguyen, Don; Sanchez, Laura M.; Ziemert, Nadine; Dorrestein, Pieter C.; Andersen, Mikael Rørdam; Gram, Lone

Published in:
mSystems

Link to article, DOI:
10.1128/mSystems.00028-15

Publication date:
2016

Document Version
Publisher's PDF, also known as Version of record

AMERICAN SOCIETY FOR MICROBIOLOGY | mSystems™

RESEARCH ARTICLE
Novel Systems Biology Techniques

An Integrated Metabolomic and Genomic Mining Workflow To Uncover the Biosynthetic Potential of Bacteria

Maria Maansson,[a]* Nikolaj G. Vynne,[a]* Andreas Klitgaard,[a]* Jane L. Nybo,[a] Jette Melchiorsen,[a] Don D. Nguyen,[b] Laura M. Sanchez,[d,e] Nadine Ziemert,[c,d] Pieter C. Dorrestein,[c,e,f] Mikael R. Andersen,[a] Lone Gram[a]

Department of Systems Biology, Technical University of Denmark, Kgs. Lyngby, Denmark[a]; Department of Chemistry and Biochemistry, University of California San Diego, La Jolla, California, USA[b]; Center for Marine Biotechnology and Biomedicine, Scripps Institution of Oceanography, University of California San Diego, La Jolla, California, USA[c]; Interfaculty Institute of Microbiology and Infection Medicine, University of Tübingen, Tübingen, Germany[d]; Collaborative Mass Spectrometry Innovation Center, University of California at San Diego, La Jolla, California, USA[e]; Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California at San Diego, La Jolla, California, USA[f]

* Present address: Maria Maansson, Chr. Hansen A/S, Hoersholm, Denmark; Nikolaj G. Vynne, Novo Nordisk A/S, Hillerød, Denmark; Andreas Klitgaard, Evolva A/S, Copenhagen, Denmark.

ABSTRACT Microorganisms are a rich source of bioactives; however, chemical identification is a major bottleneck. Strategies that can prioritize the most prolific microbial strains and novel compounds are of great interest. Here, we present an integrated approach to evaluate the biosynthetic richness in bacteria and mine the associated chemical diversity. Thirteen strains closely related to *Pseudoalteromonas luteoviolacea* isolated from all over the Earth were analyzed using an untargeted metabolomics strategy, and metabolomic profiles were correlated with whole-genome sequences of the strains. We found considerable diversity: only 2% of the chemical features and 7% of the biosynthetic genes were common to all strains, while 30% of all features and 24% of the genes were unique to single strains. The list of chemical features was reduced to 50 discriminating features using a genetic algorithm and support vector machines. Features were dereplicated by tandem mass spectrometry (MS/MS) networking to identify molecular families of the same biosynthetic origin, and the associated pathways were probed using comparative genomics. Most of the discriminating features were related to antibacterial compounds, including the thiomarinols that were reported from *P. luteoviolacea* here for the first time. By comparative genomics, we identified the biosynthetic cluster responsible for the production of the antibiotic indolmycin, which could not be predicted with standard methods. In conclusion, we present an efficient, integrative strategy for elucidating the chemical richness of a given set of bacteria and link the chemistry to biosynthetic genes.

IMPORTANCE We here combine chemical analysis and genomics to probe for new bioactive secondary metabolites based on their pattern of distribution within bacterial species. We demonstrate the usefulness of this combined approach in a group of marine Gram-negative bacteria closely related to *Pseudoalteromonas luteoviolacea*, which is a species known to produce a broad spectrum of chemicals. The approach allowed us to identify new antibiotics and their associated biosynthetic pathways. Combining chemical analysis and genetics is an efficient "mining" workflow for identifying diverse pharmaceutical candidates in a broad range of microorganisms and therefore of great use in bioprospecting.

KEYWORDS: *Pseudoalteromonas*, comparative genomics, natural products, untargeted metabolomics

Received 12 December 2015 **Accepted** 1 April 2016 **Published** 3 May 2016

Citation Maansson M, Vynne NG, Klitgaard A, Nybo JL, Melchiorsen J, Nguyen DD, Sanchez LM, Ziemert N, Dorrestein PC, Andersen MR, Gram L. 2016. An integrated metabolomic and genomic mining workflow to uncover the biosynthetic potential of bacteria. mSystems 1(3):e00028-15. doi:10.1128/mSystems.00028-15.

Editor Peter J. Turnbaugh, G. W. Hooper Research Foundation

Address correspondence to Mikael R. Andersen, mr@bio.dtu.dk (for questions on bioinformatics analyses), or Lone Gram, gram@bio.dtu.dk.

Analysing and comparing the genes and chemical components from oceanic bacteria reveals a potential for discovery of new antibiotics

Microorganisms have remarkable biosynthetic capabilities and can produce secondary metabolites with high structural complexity and important biological activities. Microorganisms in particular have been a rich source of antibiotics (1, 2) and have served as scaffolds for many other types of drugs. Chemical identification of microbial metabolites is a major bottleneck, and tools that can aid in the prioritization of the most prolific microbial strains and attractive compounds are of great interest.

The search for novel chemical diversity can be done "upstream," at the genome level, or "downstream," at the metabolite level. Historically, the approach has been to identify target molecules; however, with the availability of genomes at low costs, genome mining has become highly attractive (3–6). Genome mining analyses are greatly aided by several *in silico* prediction tools (7), such as antiSMASH (8, 9) and NaPDoS (10) for secondary metabolite pathway identification. Several studies have explored the general genomic capabilities within a group of related bacteria (11–16), but only a few studies have explored the overall biosynthetic potential and pathway diversity (17–21). Ziemert et al. (18) compared 75 genomes from three closely related *Salinispora* species and predicted 124 distinct biosynthetic pathways, which by far exceeds the 13 currently known compound classes from these bacteria. The study underlined the discovery potential in looking at multiple strains within a limited phylogenetic space, as a third of the predicted pathways were found only in a single strain.

A large potential is found by combining genome mining with the significant advances in analytical methods for compound identification. Building on the versatility, accuracy, and high sensitivity that liquid chromatography-mass spectrometry (LC-MS) platforms have achieved, sophisticated algorithms and software suites have been developed for untargeted metabolomics (22–26). The core of these programs is, first, feature detection (or peak picking), i.e., the identification of all signals caused by true ions (27), and, second, peak alignment, matching identical features across a batch of samples. Today, many programs consider not only the parent mass and the retention time (RT) but also the isotopic pattern, ion adducts, charge states, and potential fragments (27), which greatly improves the confidence in these feature detection algorithms (28). These high-quality data can be combined with multivariate analysis tools, which not only aids analysis and interpretation but also forms a perfect basis for integration with genomic information. Recently, molecular networking has been introduced as a powerful tool in small-molecule genome mining (21, 29, 30). It builds on an algorithm (31, 32) capable of comparing characteristic fragmentation patterns, thus highlighting molecular families with the same structural features and potentially the same biosynthetic origin. This enables the study and comparison of a high number of samples, at the same time aiding dereplication and tentative structural identification or classification (33).

Here, we present an integrated diversity mining approach that links genes, pathways, and chemical features at the very first stage of the discovery process using a combination of publicly available prediction tools and machine learning algorithms. We use genomic data to interrogate the chemical data and vice versa to get an overview of the biosynthetic capabilities of a group of related organisms and identify unique strains and compounds suitable for further chemical characterization. We demonstrate our approach on a unique group of marine bacterial strains all closely related to *Pseudoalteromonas luteoviolacea* based on 16S rRNA gene sequence similarity (34, 35). Previous studies in our lab have shown that it is a highly chemically prolific and diverse species with strains producing a cocktail of the antibiotics violacein and either pentabromopseudilin or indolmycin (36). We use the integrated approach to evaluate the promise of continued sampling and discovery efforts within this species as demonstrated by the finding of an additional group of antibiotics, the thiomarinols.

RESULTS

Thirteen closely related strains previously identified as *P. luteoviolacea* by gene sequence similarity (36) were analyzed for their genomic potential and ability to produce

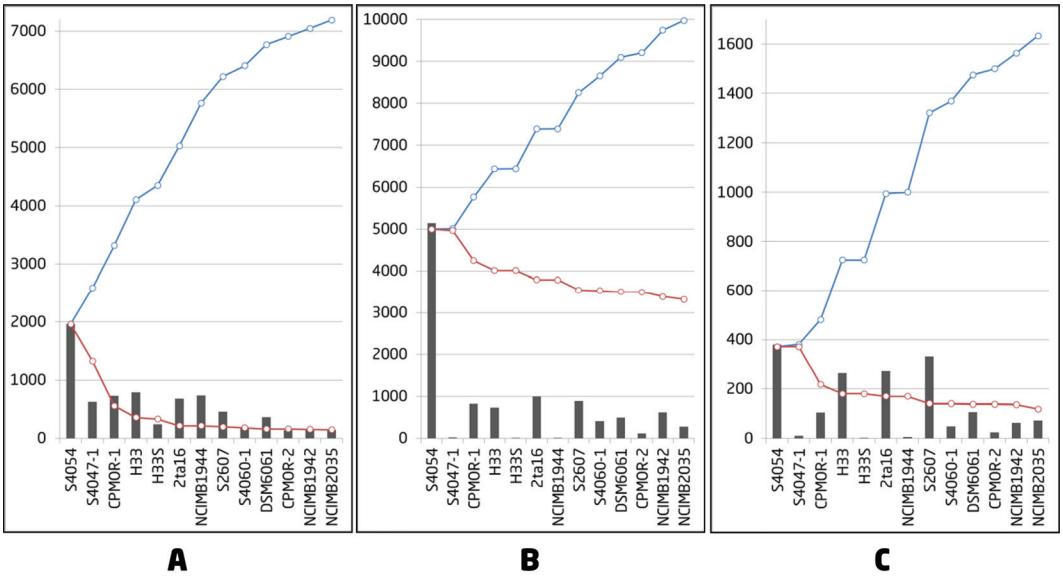


FIG 1 Pan- and core-metabolome and genome plots of 13 *P. luteoviolacea* strains. (A) The pan-metabolome curve (blue) connects the cumulative number of molecular features detected (positive and negative mode merged). The core-metabolome curve (red) connects the conserved number of features. The bars show the number of new molecular features detected in each extract (medium components excluded). (B) Pan-genome (blue) and core-genome (red) curves for all predicted genes. (C) Pan-genome (blue) and core-genome (red) curves for genes predicted to be involved in secondary metabolism.

secondary metabolites. The bacteria were cultivated on a complex medium known to support production of secondary metabolites (37) and extracted sequentially by ethyl acetate and butanol to obtain broad metabolite coverage. To obtain a global, unbiased view of the metabolites produced, molecular features were detected by LC-electrospray ionization (ESI)–high-resolution MS (HRMS) in an untargeted metabolomics experiment. On average, more than ~2,000 molecular features were detected in each strain. Merging of ESI^+/ESI^- data resulted in a total of 7,190 features from the 13 strains (excluding medium components), with more features detected in positive mode (6,736) than negative mode (2,151). To facilitate comparison to genomic data, the features were represented as pan- and core plots commonly used for comparative microbial genomics (38, 39). Here, core-metabolome features are shared between all strains, while the pan-metabolome represents the total repertoire of features detected within the collection (Fig. 1A).

Surprisingly, only 2% of the features were shared between all the strains. In contrast, 30% of all features were unique to single strains. As the number and detection of features in each strain change with the chosen threshold for feature filtering, the pan- and core plots were also made based on the 2,000 and 500 most intense features, respectively (see Fig. S1 in the supplemental material). Here, the same trend was observed with 6 to 10% core features and 20% unique features. Thus, regardless of feature filtering settings, the overall pattern of diversity is the same.

To link the chemical diversity to the genomic diversity in these closely related strains, we analyzed the 13 genomes by different comparative approaches. The average genome size was approximately 6 Mb with approximately 5,100 putative protein-encoding genes per strain (see Table S1 in the supplemental material). The corresponding pan- and core-genomic analysis was performed using CMG-biotools (39) (Fig. 1B). A total of 9,979 protein-encoding genes were predicted in the pan-genome, including 3,322 genes (33%) conserved between all strains; thus, on average, the core genome constituted ~65% for each strain. Of the accessory genome, 23% of the total genes (2,329) could be found only in a single strain (singletons/unique genes). Considering only genes predicted to be involved in secondary metabolism, the diversity was even higher (Fig. 1C). On average, 8.6% of the total genes were predicted to be allocated to

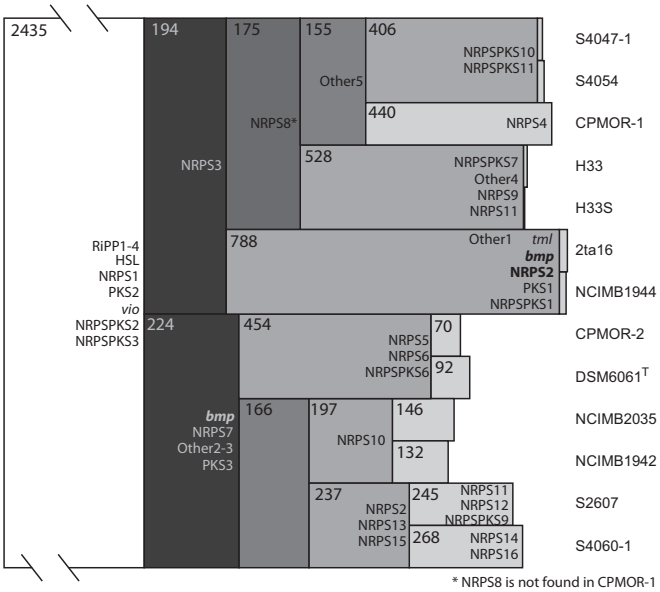


FIG 2 Icicle plot of shared genes for groups of species with OBUs overlaid. The numbers in the boxes show the number of mutual 1:1 orthologs found in the species to the right of that box. The areas of the individual boxes are proportional to the number of genes.

secondary metabolism (see Table S1), which is extremely high compared to other sequenced strains belonging to *Pseudoalteromonas* (40, 41). Similar to the total pan-genome, 24% (386) of the genes putatively involved in secondary metabolism were found in only a single strain; however, only 7% (119) were shared between all 13 strains. Thus, we see approximately a 5-fold-higher genetic diversity in secondary metabolism compared to the full pan-genome.

The high number of unique genes and molecular features suggests an open pan-genome/metabolome (38) in which there is a continuous increase in diversity with continued sampling, which is very attractive for discovery purposes. Both sets of data suggest that 90% of the diversity/genomic potential for secondary metabolism can be covered with 10 strains but that each new strain holds promise for new compounds and biosynthetic pathways.

Pan-genomic diversity and pathway mapping suggest a highly dynamic accessory genome. To determine the potential evolutionary relationship between the strains and associated pathways, a pan-genomic map was generated illustrating shared orthologs between groups of species (Fig. 2).

The method uses a conservative BLAST-based nongreedy pairing of genes, which results in 2,435 genes found to be present as 1:1 orthologs in all strains, which is slightly fewer than the 3,388 genes found in the method illustrated in Fig. 1. In general, we observed two main clades based on shared genes, one consisting of six strains and the other of seven. Each clade has 190 to 220 genes unique for that clade. The method also further reflects the genetic diversity of each strain, as illustrated in Fig. 1B and C. Based on the shared orthologs, we generated presence/absence patterns for all genes showing in which other strains that gene has orthologs, a useful starting point for data correlation.

For genetic analysis of biosynthetic pathways in multiple strains, pathways were predicted using antiSMASH across the 13 strains and grouped into 37 operational biosynthetic units (OBUs) (18) (see Table S2 in the supplemental material). OBU presences were compared to the pan-genomic map (Fig. 2) to trace biosynthetic pathways. Only 10 pathways were conserved in all strains, including a glycosylated lantipeptide (RiPP1) and two bacteriocins (RiPP2 and RiPP3). All strains maintained essential pathways likely responsible for production of siderophores (NRPS1 putative

catechol-based siderophore) and homoserine lactones (different variations). The violacein pathway *vio* is also conserved in all strains (consistent with the purple phenotype of the pseudoalteromonads), in addition to an unassigned type III polyketide synthase (PKS) and a hybrid nonribosomal peptide synthetase (NRPS)-PKS pathway. Interestingly, the majority of clusters follow the strain lineage suggested by Fig. 2, suggesting that many of the pathways have been introduced and retained based on a competitive advantage of those clusters. More than 50% of the predicted pathways are restricted to one or two strains, suggesting that many pathways are introduced highly dynamically (in evolutionary scale). Whether gene gain or gene loss is responsible for the patchy distribution for most of these OBUs is unclear and was not part of this study. However, evolutionary studies in other organisms have proven that horizontal gene transfer is an important part of the evolution of secondary metabolite clusters (18, 42–45). The exact mechanism of the transfer is not known. No significant amount of transposases or other mobile elements has been found within or in the direct vicinity of the clusters.

Key discriminative metabolites are revealed through feature prioritization and dereplication of the pan-metabolome by SVM and molecular networking. To explore the diversity within the pan-metabolome and prioritize chemical features for more detailed structural analysis, a two-pronged approach was used: multivariate analysis based on machine learning algorithms and comparative analyses based on the pattern of conservation generated from the pan-genomic diversity map. A classifier based on a combination of a genetic algorithm (GA) and support vector machine (SVM) (46, 47) was used as a feature selection method to filter the most important features from the complex data set, starting with the 500 most intense features and reducing it to the 50 most significant features to distinguish all 13 strains (see Table S3 in the supplemental material). In addition, extracts from all strains were analyzed with LC-ESI-MS/MS to generate a molecular network (see Fig. S2a for full details) (30). The candidates identified by multivariate and comparative analyses were correlated with the molecular network (29, 33) for dereplication and connection of molecular features that likely belong to the same structural class and thus biosynthetic pathway. For example, the *vio* pathway (48) was found in all 13 strains, and the antibiotic violacein was a discriminating core feature (see Table S3). In the molecular network, violacein was found to belong to a molecular family of a minimum of five related analogues (see Fig. S2b) likely associated with the *vio* pathway, including proviolacein and oxyviolacein, as well as a novel analogue with two extra hydroxyl groups.

Some strains have lost the ability to produce polyhalogenated compounds. The discriminating features do not necessarily reflect the same groupings as the genomic analyses. Therefore, they can be used as a tag for identifying the corresponding biosynthetic pathway through correlation with genomic presence/absence patterns. On the list of descriptive features generated using the SVM (see Table S3 in the supplemental material), there are six highly halogenated features that all seem to be restricted to seven strains: CPMOR-2/DSM6061^T, S2607/S4060-1, NCIMB1944/2ta16, and CPMOR-1. To investigate whether halogenation in general is unique to those strains, a list of features with a high mass defect was made, resulting in more than 40 halogenated compounds (see Table S4) restricted to the seven strains. Most of them had no match to known compounds, but many match the structural scaffolds of polyhalogenated phenols and pyrroles or hybrids thereof (49) and have expected antibacterial activity (50).

No pathway predicted by antiSMASH had a halogenase incorporated; thus, the pattern of presence in these seven strains was used to probe for associated clusters. Indeed, we found an intact group of 11 genes (including two brominases) conserved in the seven abovementioned strains (see Fig. S3a in the supplemental material). The recently characterized *bmp* pathway corresponds to these genes (*bmp1* to *bmp10*) (49) and is responsible for the production of polybrominated phenols/pyrroles in strain 2ta16 and a putative multidrug transporter (tentatively named *bmp11*). Surprisingly, all 11 genes were also found in NCIMB1942/NCIMB2035, where no halogenated compounds were detected. Incidentally, in both genomes, the cluster is divided across two

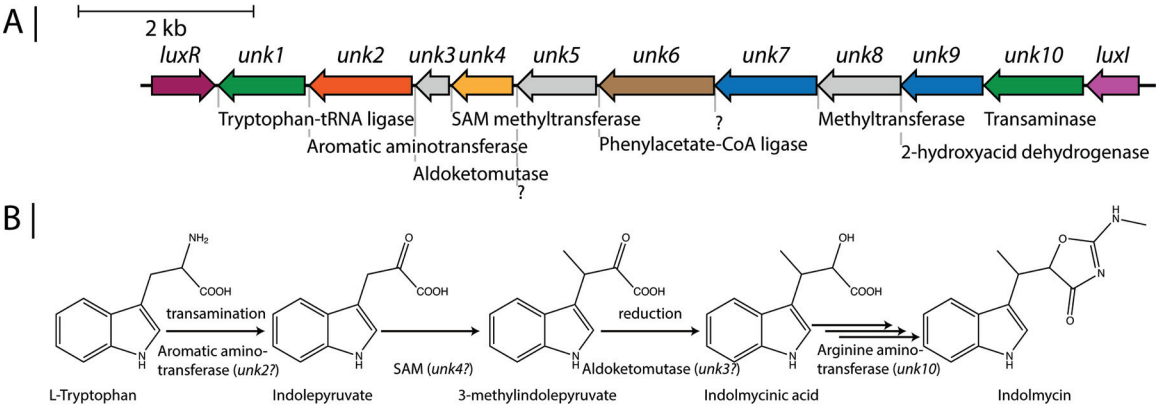


FIG 3 Putative biosynthetic cluster (A) and proposed biosynthetic scheme (B) (51) for indolmycin. CoA, coenzyme A.

contigs with the break point being in *bmp1* in both genomes. Should this be an actual physical division of the contig, or an inserted unsequenceable repeat sequence, it could provide an explanation for the lack of halogenated compounds. However, sequencing of the *bmp1* gene in NCIMB2035 revealed a 1-kb insert in the thioesterase (TE) domain of the gene, likely explaining the lack of compounds (J. Busch, V. Agarwal, A. A. El Gamal, B. S. Moore, G. W. Rouse, L. Gram, and P. R. Jensen, unpublished data). Also, *bmp1*, *bmp2*, a part of *bmp7*, and *bmp8* to *bmp11* were found in S4047-1/S4054, which suggests that a common ancestor had an intact *bmp* pathway.

Two of the discriminative features found in the seven strains are two isomeric dimeric bromophenol-bromopyrrole hybrids with eight bromines in total (see Fig. S4 in the supplemental material). The monomers corresponding to the likewise novel "tetrabromopseudilin" are also found in the extract, suggesting that these "*bis*-tetrabromopseudilins" are true compounds rather than artifacts arising from MS in-source chemistry. Full structural characterization of these low-proton-density compounds lies beyond the scope of this study but underlines the versatility of the *bmp* pathway and associated chemical diversity.

Identification of the indolmycin cluster shows resistance genes and potential quorum sensing (QS) control. Strains S4047-1, S4054, and CPMOR-1 are all producing the antibiotic indolmycin, as previously reported (36). Indolmycin was identified by GA/SVM as a discriminating feature for those three strains. In addition to indolmycin, the molecular family consisted of the N/C-demethyl- and N/C-didemethyl indolmycin analogues as well as indolmyceinic acid, a methylated analogue, and two hydroxylated analogues. Most of these analogues have not been reported from microbial sources, and their tentative structures were verified by their MS/MS fragmentation pattern (see Fig. S5 in the supplemental material).

Like violacein, indolmycin is derived from L-tryptophan, but even though the biosynthetic pathway has been described by feeding studies in *Streptomyces* (51–53) and recently characterized genetically (54), the biosynthetic cluster responsible has never been characterized. The pan-genome was probed for genes with presence/absence patterns matching the distribution of indolmycin and the related analogues, which led to the identification of 13 clustered genes, suggesting these to be the genetic basis for indolmycin biosynthesis (Fig. 3). The identified genes had predicted functions similar to those expected to be required for the synthesis of indolmycin such as an aromatic aminotransferase (*unk2*), aldoketomutase (*unk3*), *S*-adenosylmethionine (SAM) methyltransferase (*unk5*), and aminotransferase (*unk11*). We have compared our proposed indolmycin biosynthetic gene cluster to that characterized by Du et al. (54) and have identified homologues to the *Streptomyces griseus* ATCC 12648 genes involved in biosynthesis of indolmycin (see Fig. S3b in the supplemental material). Indolmycin has been identified as a competitive inhibitor of bacterial tryptophan-tRNA ligases (55, 56), and the putative cluster seems to incorporate a tryptophan-tRNA ligase (*unk1*), which

in *Streptomyces griseus* has been found to confer resistance to indolmycin (56). Interestingly, the cluster in *Pseudoalteromonas* is flanked by *luxI* and *luxR* homologues, something which is not observed in *S. griseus*, suggesting that the indolmycin pathway potentially could be under regulation by quorum sensing.

Thiomarinols add to the antibiotic cocktail. The strains 2ta16 and NCIMB1944 were identified as hot spots for biosynthetic diversity based on Fig. 2. This was supported by 313 chemical features (RT and *m/z* pairs) unique to these two strains. Based on the GA/SVM, they can be distinguished from the rest of the strains based on a feature with *m/z* 640 and an RT of 9.73 min ($C_{30}H_{44}N_2O_9S_2$), tentatively identified as thiomarinol A. Thiomarinols are hybrid NRPS-PKS compounds based on pseudomonic acid and pyrrothine. One of the gene clusters (hybrid NRPSPKS5) restricted to the pair 2ta16-NCIMB1944 was found to have high similarity to that of pseudomonic acid (*mup*) (57) and the recently characterized thiomarinol (*tml*) cluster (58), corroborating the finding of the compound class. Thiomarinols have previously reported antibacterial activities from *Pseudoalteromonas* sp. strain SANK 73390 (59, 60).

In the molecular network, it was possible to identify a whole series of thiomarinol and pseudomonic acid analogues (Fig. 4A and D), all restricted to NCIMB1944 and 2ta16. In addition to thiomarinols A to D, pseudomonic acid C amide and its hydroxyl analogue could be assigned based on the characteristic MS/MS fragmentation pattern (Fig. 4B and C). Besides the known analogues, two novel analogues with formulas $C_{25}H_{43}NO_8$ and $C_{34}H_{51}NO_{11}$ could be identified. Both shared the marinolic acid moiety based on the $C_6H_6O_2$ (*m/z* 110.0368) fragment and the loss of $C_{11}H_2OO_4$ (*m/z* 216.1362); however, they contained only a single nitrogen and no sulfur, indicating a completely new type of thiomarinol based on neither a holothine nor an ornithine "head" like the known analogues (Fig. 4C).

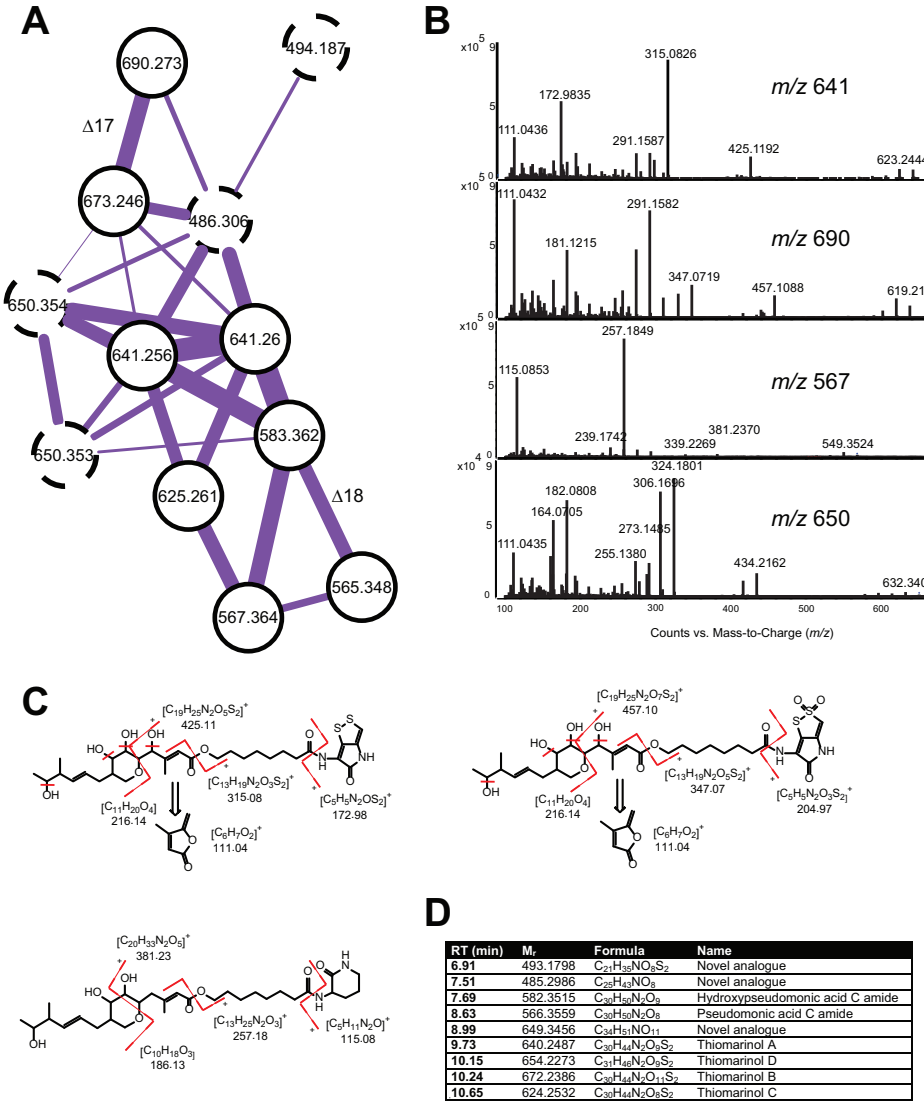
DISCUSSION

Advances in genomics and metabolomics have significantly increased our ability to generate high-quality data on microbial secondary metabolism at a very high speed. This, in turn, has enabled a completely new approach to drug discovery combining the two "-omics" approaches.

Using a combination of comparative metabolomics and genomics, we find a high potential and remarkable diversity in terms of secondary metabolite production for strains closely related to *P. luteoviolacea*. Overall, 8.6% of the genes are allocated to secondary metabolism, and on average, 10 NRPS/PKS-related OBUs are predicted. This is very high considering the relatively small size of the genomes (~6 Mb) and is comparable to that of recognized prolific species such as *Salinispora arenicola* (10.9% of 5.8 Mb) (13, 18, 61) and *Streptomyces coelicolor* (8% of 8.7 Mb) (62). Our data suggest an open pan-genome which is characteristic for species that are adapted to several types of environments (63), i.e., being both planktonic and associated with marine macroalgal surfaces. The pan-genome is a dynamic descriptor that will change with the number of strains and the specific subset. Nonetheless, our findings correlate with comparative genomic studies of other bacterial species (11, 12, 14, 63).

We found ~5-fold-higher genetic diversity in secondary metabolism compared to the full pan-genome, which supports the idea that production of secondary metabolites is a functionally adaptive trait (64, 65). More than half of the 41 predicted pathways are restricted to one or two strains, while only 10 pathways were shared between all. This is similar to findings in *Salinispora* (18), where 78% of the pan-genome is associated with one or two strains. Violacein (66, 67), indolmycin (68, 69), and pentabromopseudilin (49) are all examples of cosmopolitan antibiotics found in unrelated species; thus, we hypothesize that *P. luteoviolacea* acquired and retained biosynthetic genes linked to, e.g., antibiotic production as part of adapting to a specific niche that it commonly occupies.

Diversity is further supported at the chemical level. Using unbiased global metabolite profiling, we identify >7,000 putative chemical features among the 13 analyzed strains. As the number of chemical features depends on the filtering threshold, this



RT (min)	M_r	Formula	Name
6.91	493.1798	$C_{21}H_{35}NO_8S_2$	Novel analogue
7.51	485.2986	$C_{25}H_{43}NO_8$	Novel analogue
7.69	582.3515	$C_{30}H_{50}N_2O_9$	Hydroxypseudomonic acid C amide
8.63	566.3559	$C_{30}H_{50}N_2O_8$	Pseudomonic acid C amide
8.99	649.3456	$C_{34}H_{51}NO_{11}$	Novel analogue
9.73	640.2487	$C_{35}H_{44}N_2O_9S_2$	Thiomarinol A
10.15	654.2273	$C_{31}H_{46}N_2O_9S_2$	Thiomarinol D
10.24	672.2386	$C_{35}H_{44}N_2O_{11}S_2$	Thiomarinol B
10.65	624.2532	$C_{35}H_{44}N_2O_8S_2$	Thiomarinol C

FIG 4 (A) Molecular network of the thiomarinol/pseudomonic acid molecular family. Dashed nodes indicate novel analogues. Mass differences are highlighted for ion adducts only. (B) MS/MS spectra representing the four different analogue types. Parent mass *m/z* 641 is thiomarinol A representing the holothin head type; *m/z* 690 is $[M + NH_4]^+$ of *m/z* 673, thiomarinol B, representing the sulfone head type; *m/z* 567 is pseudomonic acid C amide, representing the nonsulfonated analogues; *m/z* 650 is a novel analogue with a nonsulfonated head. (C) Structures and suggested fragmentation of thiomarinols A and B and pseudomonic acid C amide. (D) Table of detected analogues in strains NCIMB1944 and 2ta16.

should not be seen as an absolute number of compounds that can be isolated and fully characterized. However, it provides an unbiased estimate of diversity, which in this case does not seem to change with the chosen threshold. Surprisingly, only 2% of the features were shared between all the strains. To the best of our knowledge, there is only one other similar study on chemical diversity in limited taxonomical spaces approaching the species level. Krug et al. (19, 70) analyzed 98 isolates of *Myxococcus xanthus* in a semitargeted approach and found 11 out of 51 identified compounds to be shared between all strains and a similar fraction present in only one or two strains. We found that almost half of all features and one-third of the 500 most intense features could be assigned to one or two strains (thus taking into account the almost clonal strains), which underlines a great potential for unique chemistry within a group of closely related strains. The detected chemical diversity is higher than what was found on the genetic level, which is to be expected, as the method at this initial screening level does

not allow for detecting differential regulation of complete pathways or individual analogues.

The remarkable chemical diversity can be found even within the same sample. Strains S4047, S4054, and S4060 were all collected from seaweed from the same geographical location (37). Strains S4047 and S4054 share 99% of their gene families (clonal) and 70% of their chemical features, but strain S4060 shares only 24% of gene families and 30% of features with the other two. It is also reflected in the biosynthetic pathways, where nine pathways were found in S4060 but not in S4047 and S4054. This is a fascinating ecological conundrum as the accessory metabolites and genes usually are considered to answer the immediate, more localized needs for the strains. Nonetheless, this is not the first report of such an occurrence. Vos and Velicer (71) found 21 genotypes of *M. xanthus* using multilocus sequence typing among 78 strains collected from soil on a centimeter scale. Likewise, significant differences have been found in the chemical profiles of cooccurring strains of *M. xanthus* (19) and *Salinibacter ruber* (72). In contrast, NCIMB1944 and 2ta16, which originate from the Mediterranean Sea (France) and the Florida Keys (United States), respectively, share 99% of their gene families and 70% of their features. That demonstrates that genomic content can be relatively conserved across biogeographical locations, suggesting a high selective pressure to conserve those genes despite an overall low degree of chemoconsistency.

In this study, SVM was applied in conjunction with GA to compile a list of 50 chemical features of interest for further structural characterization. Based on SVM, the reduced set of features are the ones that maximize the difference between samples, which in this study is exploited to select features unique to each strain or a subset of strains. GA works as a wrapper to select features to be evaluated in the SVM classifier (73). The intrinsic nature of the GA makes it highly suitable for discovery purposes as it favors diversity in how the subset of features is selected (47). To the best of our knowledge, there are only a few examples of the use of SVM in untargeted secondary metabolite profiling (74, 75). The list of discriminating features highlights key metabolites, both in the core and in the accessory metabolome. Of the 50 discriminating features, only 15 could be tentatively assigned to known compound classes. In this specific case, the list even reflects the four antibiotic classes identified in this species, underlining the utility of GA/SVM to prioritize not only strains but also compounds before the rate-limiting step of structural identification. The combination with molecular networking further strengthens this approach as it makes it possible to identify structural analogues that likely have similar biological activities.

This is the one of the first examples (20, 21, 29) of direct coupling of genomic and metabolomic data at a global level and at this early stage of the discovery process. By solely using the patterns of presence/absence across the pan-genome in conjunction with synteny, we could identify gene clusters without relying on the functions. This allowed for the identification of the pentabromopseudilin and indolmycin gene clusters. Combined with presence/absence of molecular features, this is an extremely powerful tool for translation back and forth between the genome and metabolome. Thus, it is possible to identify specific compounds using genomic queries or to specifically identify a gene cluster based on chemistry. Of course, in order to fully confirm the link between a compound and its genes, knockout mutants need to be analyzed or entire pathways recombinantly expressed, but here, single candidates for clusters could be directly and rapidly identified.

The combination of metabolomics and genomic data identifies obvious hot spots for chemical diversity among the 13 strains, which permits intelligent strain selection for more detailed chemical analyses. By randomly picking a single strain, in the worst case, only 38% of the 500 most intense chemical features (and thus most relevant from a drug discovery perspective) are covered (NCIMB2035). However, when maximizing strain orthogonality by selecting the two strains (NCIMB1944 and CPMOR-1) with the highest number of unique genes, pathways, and chemical features, 82% of the diversity can be covered. This is extremely important as the isolation and full structural characterization of these compounds still represent the greatest bottleneck in the discovery

process. This study shows that investigation of multiple closely related strains is a valuable strategy for detection of new compounds and is imperative for uncovering the full biosynthetic potential of a species.

MATERIALS AND METHODS

Strains, cultivation, and sample preparation for chemical analyses. The 13 strains included in the study were collected or donated to us as previously described (36, 37). We did attempt to build a larger collection; however, *P. luteoviolacea* autolyzes very easily, and in most laboratories, it has not been possible to store and revive strains. The strains were cultured in biological duplicates in marine broth (MB; Difco catalog no. 2216) at 25°C (200 rpm) for 48 h before extraction. See details in Text S1 in the supplemental material.

LC-MS and LC-MS/MS data acquisition. LC-MS and MS/MS analyses were performed on an Agilent 6550 iFunnel quadrupole-time of flight (Q-TOF) LC-MS (Agilent Technologies, Santa Clara, CA) coupled to an Agilent 1290 Infinity ultrahigh-performance liquid chromatography (UHPLC) system. Separation was performed using a Poroshell 120 phenyl-hexyl column (Agilent; 250 mm by 2.1 mm; 2.7 μm) with a water-acetonitrile (ACN) gradient. MS data were recorded in both positive and negative electrospray (ESI) mode in the *m/z* 100- to 1,700-Da mass range. Data for molecular networking were collected using a data-dependent LC-MS/MS as reported previously (76) with optimized collision energies and scan speed. See Text S1 in the supplemental material for the full experimental setup, procedures, and method parameters.

Feature extraction and multivariate analysis. Extraction of chemical features was performed using MassHunter (Agilent Technologies; v.B06.00) and the Molecular Features Extraction (MFE) algorithm and recursive analysis workflow. Feature lists were imported to Genespring-Mass Profiler Professional (MPP) (Agilent Technologies; v.12.6) and filtered with features resulting from the medium removed. The feature lists from ESI^+ and ESI^- data were merged in a table as generic data and reimported into MPP. The data were then normalized and aligned, resulting in a single list of chemical features for each sample. The list of discriminating features was generated in MPP using a genetic algorithm with a population size of 25, 10 generations, and a mutation rate of 1. The GA was evaluated using the SVM with a linear kernel type with an imposed cost of 100 and a ratio of 1. The feature list was validated via the leave-one-out method. Further details and settings can be found in the supplemental material. All 50 discriminating features (see Table S3 in the supplemental material) were manually verified to be present in the original data sets. Molecular formulas were predicted from the accurate mass of the molecular ion or related adducts (77) as well as the isotope pattern and matched against AntiMarin (v.08.13) and Metlin (78) databases to tentatively assign known compounds.

Molecular networking. For molecular networking, raw LC-MS/MS data were converted to .mgf using MSConvert from the ProteoWizard project (79) and analyzed with the algorithm described in the work of Watrous et al. (30). A new, public interface at http://gnps.ucsd.edu has been made public at the time of writing, and the data have been deposited (MSV000078988) in the corresponding database, http://massive.ucsd.edu. Likewise, the annotated MS/MS spectra for all the identified compounds have been uploaded and added to the GNPS spectral library. The network corresponding to a cosine value of more than 0.7 was visualized using Cytoscape 2.8.3 (80).

DNA extraction, genome sequencing, and assembly. Cultures were grown in MB for 1 to 2 days, and genomic DNA was isolated using either the JGI phenol-chloroform extraction protocol or the Qiagen 100/G kit. Library preparation and 150-base-paired end sequencing were done at the Beijing Genomics Institute (BGI) on the Illumina HiSeq 2000 system. At least 100-fold coverage was achieved for all genome sequences generated in this study. Raw sequence data for strain 2ta16 were downloaded from http://www.jcvi.org and assembled as described here. Genomes were assembled using CLC Genomics Workbench (v.2.1 for NCIMB2035, 2.04 for remaining whole-genome sequences) with default settings.

Genome analysis. Contigs were analyzed and plots were created using the CMG-biotools package as described in the work of Vesth et al. (39). Briefly, genes were predicted using Prodigal 2.00. Gene families were constructed by genome-wide and pairwise BLAST comparisons. Genes were considered part of the same gene family with a sequence identity of >50% over at least 50% of the length of the longest gene. A pan-genomic dendrogram based on occurrences of gene families was used to sort input order by clustering prior to generating the pan- and core-genome plots (14).

Putative biosynthetic pathways were predicted from sequences (FASTA) with antiSMASH 2.0 (8, 9), with KS and C domains of PKS and NRPS predicted with NaPDoS (10) using default settings. Pathways were assessed as being similar OBUs when MultiGeneBlast (81) analyses revealed that 80% of the genes in the pathway were present with homologues that show at least 60% amino acid identity. For assessment and assembly of pathways split between different contigs, the sequences of homologues on the same contig were used as the scaffold. MultiGeneBlast (81) was used for recursive OBU analysis across all 13 strains, thus providing pseudoscaffolds for larger pathways, which in turn give higher confidence in the assignments. Partial pathways with the same pattern of conservation were combined in order to avoid overestimation of diversity. Predicted genes involved in the putative indolmycin biosynthetic pathway are labeled *unk* for "unknown."

Mapping of genes shared by groups of strains. All predicted sets of protein sequences for the 13 strains were compared using the blastp function from the BLAST+ suite (82). These 169 whole-genome BLAST tables were analyzed to identify bidirectional best hits in all pairwise comparisons. Using custom Python scripts, this output was analyzed to identify, for all proteins, the strains in which orthologs were found. This allowed identification of unique genes, genes shared by clades and subclades of species, and

genes shared by all 13 strains of *Pseudoalteromonas*. The script also generates a binary 13-digit barcode of the presence/absence of gene orthologs across the 13 strains for all proteins in the pan-genome.

Nucleotide sequence accession numbers. The whole-genome shotgun projects have been deposited at GenBank under the accession numbers AUXS00000000, AUXT00000000, AUXU00000000, AUXV00000000, AUXX00000000, AUXY00000000, AUXZ00000000, AUYA00000000, AUYB00000000, and AUYC00000000. The versions described in this paper are versions AUXS01000000, AUXT01000000, AUXU01000000, AUXV01000000, AUXX01000000, AUXY01000000, AUXZ01000000, AUYA01000000, AUYB01000000, and AUYC01000000.

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at http://dx.doi.org/10.1128/mSystems.00028-15.

Text S1, DOCX file, 0.1 MB.
Figure S1, DOCX file, 0.2 MB.
Figure S2, DOCX file, 0.4 MB.
Figure S3, DOCX file, 0.1 MB.
Figure S4, DOCX file, 0.1 MB.
Figure S5, DOCX file, 0.4 MB.
Table S1, DOCX file, 0.1 MB.
Table S2, DOCX file, 0.1 MB.
Table S3, DOCX file, 0.1 MB.
Table S4, DOCX file, 0.1 MB.

ACKNOWLEDGMENTS

We acknowledge Farooq Azam and Krystal Rypien of the Scripps Institution of Oceanography, UCSD, for supplying strain 2ta16; Antonio Sanchez-Amat of the University of Murcia for supplying strains CPMOR-1 and CPMOR-2; and Tillman Harder of the University of New South Wales for supplying strains H33 and H33S. We thank Mingxun Wang for assistance with uploads to GNPS/MassIVE and Inge Kjærbølling for help with Fig. S2 in the supplemental material.

This study was supported by the Danish Research Council for Technology and Production Science with Sapere Aude (grant no. 116262). Instrumentation and software used in this study were supported by Agilent Technologies Thought Leader Donation. The work of J.L.N. and M.R.A. was partially supported by the Villum Foundation (grant VKR023437). P.C.D. and D.D.N. were supported by the National Institutes of Health (NIH) (grant GM097509), and L.M.S. was supported by the National Institutes of Health IRACDA K12 GM068524 grant award.

M.M., N.G.V., and L.G. designed the research; M.M., N.G.V., and J.M. carried out the experiments; M.M., A.K., N.G.V., D.D.N., L.M.S., N.Z., and M.R.A. analyzed the data; J.L.N., M.R.A., and P.C.D. provided methods and algorithms; M.M., A.K., M.R.A., and L.G. wrote the paper.

FUNDING INFORMATION

This work, including the efforts of Maria Månsson, Nikolaj G Vynne, Jette Melchiorsen, and Lone Gram, was funded by Danish Research Council for Technology and Production (116262). This work, including the efforts of Jane L Nybo and Mikael Rørdam Andersen, was funded by Villum Fonden (Villum Foundation) (VKR023437). This work, including the efforts of Don D Nguyen and Pieter C. Dorrestein, was funded by HHS | National Institutes of Health (NIH) (GM097509). This work, including the efforts of Laura M Sanchez, was funded by HHS | National Institutes of Health (NIH) (GM068524).

This study was supported by the Danish Research Council for Technology and Production Science with Sapere Aude (grant #116262). Instrumentation and software used in this study was supported by Agilent Technologies Thought Leader Donation. The work of J.L.N. and M.R.A. was partially supported by the Villum Foundation (grant VKR023437). P.C.D. and D.D.N. was supported by National Institute of Health (NIH)(grant GM097509) and L.M.S. by National Institutes of Health IRACDA K12 GM068524 grant award.

REFERENCES

1. **Peláez F**. 2006. The historical delivery of antibiotics from microbial natural products—can history repeat? Biochem Pharmacol **71:**981–990. http://dx.doi.org/10.1016/j.bcp.2005.10.010.
2. **Clardy J, Fischbach MA, Walsh CT**. 2006. New antibiotics from bacterial natural products. Nat Biotechnol **24:**1541–1550. http://dx.doi.org/10.1038/nbt1266.
3. **Müller R, Wink J**. 2014. Future potential for anti-infectives from bacteria—how to exploit biodiversity and genomic potential. Int J Med Microbiol **304:**3–13. http://dx.doi.org/10.1016/j.ijmm.2013.09.004.
4. **Aigle B, Lautru S, Spiteller D, Dickschat JS, Challis GL, Leblond P, Pernodet J-L**. 2014. Genome mining of Streptomyces ambofaciens. J Ind Microbiol Biotechnol **41:**251–263. http://dx.doi.org/10.1007/s10295-013-1379-y.
5. **Durkin AS, Eisen JA, Eisen J, Ronning CM, Barbazuk WB, Blanchard M, Field C, Halling C, Hinkle G, Iartchuk O, Kim HS, Mackenzie C, Madupu R, Miller N, Shvartsbeyn A, Sullivan SA**. 2006. Evolution of sensory complexity recorded in a myxobacterial genome. Proc Natl Acad Sci U S A **103:**15200–15205. http://dx.doi.org/10.1073/pnas.0607335103.
6. **Hanamoto A, Takahashi C, Shinose M, Takahashi Y, Horikawa H, Nakazawa H, Osonoe T, Kikuchi H, Shiba T, Sakaki Y, Hattori M**. 2001. Genome sequence of an industrial microorganism Streptomyces avermitilis: deducing the ability of producing secondary metabolites. Proc Natl Acad Sci U S A **98:**12215–12220. http://dx.doi.org/10.1073/pnas.211433198.
7. **Weber T**. 2014. In silico tools for the analysis of antibiotic biosynthetic pathways. Int J Med Microbiol **304:**230–235. http://dx.doi.org/10.1016/j.ijmm.2014.02.001.
8. **Medema MH, Blin K, Cimermancic P, de Jager V, Zakrzewski P, Fischbach MA, Weber T, Takano E, Breitling R**. 2011. antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Nucleic Acids Res **39:**W339–W346. http://dx.doi.org/10.1093/nar/gkr466.
9. **Blin K, Medema MH, Kazempour D, Fischbach MA, Breitling R, Takano E, Weber T**. 2013. antiSMASH 2.0—a versatile platform for genome mining of secondary metabolite producers. Nucleic Acids Res **41:**W204–W212. http://dx.doi.org/10.1093/nar/gkt449.
10. **Ziemert N, Podell S, Penn K, Badger JH, Allen E, Jensen PR**. 2012. The natural product domain seeker NaPDoS: a phylogeny based bioinformatic tool to classify secondary metabolite gene diversity. PLoS One **7:**e34064. http://dx.doi.org/10.1371/journal.pone.0034064.
11. **Mann RA, Smits TH, Bühlmann A, Blom J, Goesmann A, Frey JE, Plummer KM, Beer SV, Luck J, Duffy B, Rodoni B**. 2013. Comparative genomics of 12 strains of Erwinia amylovora identifies a pan-genome with a large conserved core. PLoS One **8:**e55644. http://dx.doi.org/10.1371/journal.pone.0055644.
12. **Park J, Zhang Y, Buboltz AM, Zhang X, Schuster SC, Ahuja U, Liu M, Miller JF, Sebaihia M, Bentley SD, Parkhill J, Harvill ET**. 2012. Comparative genomics of the classical Bordetella subspecies: the evolution and exchange of virulence-associated diversity amongst closely related pathogens. BMC Genomics **13:**545. http://dx.doi.org/10.1186/1471-2164-13-545.
13. **Penn K, Jensen PR**. 2012. Comparative genomics reveals evidence of marine adaptation in Salinispora species. BMC Genomics **13:**86. http://dx.doi.org/10.1186/1471-2164-13-86.
14. **Lukjancenko O, Wassenaar TM, Ussery DW**. 2010. Comparison of 61 sequenced *Escherichia coli* genomes. Microb Ecol **60:**708–720. http://dx.doi.org/10.1007/s00248-010-9717-3.
15. **Tagomori K, Iida T, Honda T**. 2002. Comparison of genome structures of vibrios, bacteria possessing two chromosomes. J Bacteriol **184:**431–4358. http://dx.doi.org/10.1128/JB.184.16.4351-4358.2002.
16. **Aylward FO, McDonald BR, Adams SM, Valenzuela A, Schmidt RA, Goodwin LA, Woyke T, Currie CR, Suen G, Poulsen M**. 2013. Comparison of 26 sphingomonad genomes reveals diverse environmental adaptations and biodegradative capabilities. Appl Environ Microbiol **79:**3724–3733. http://dx.doi.org/10.1128/AEM.00518-13.
17. **Penn K, Jenkins C, Nett M, Udwary DW, Gontang EA, McGlinchey RP, Foster B, Lapidus A, Podell S, Allen EE, Moore BS, Jensen PR**. 2009. Genomic islands link secondary metabolism to functional adaptation in marine Actinobacteria. ISME J **3:**1193–1203. http://dx.doi.org/10.1038/ismej.2009.58.
18. **Ziemert N, Lechner A, Wietz M, Millán-Aguiñaga N, Chavarria KL, Jensen PR**. 2014. Diversity and evolution of secondary metabolism in the marine actinomycete genus *Salinispora*. Proc Natl Acad Sci U S A **2014:**1–10. http://dx.doi.org/10.1073/pnas.1324161111.
19. **Krug D, Zurek G, Revermann O, Vos M, Velicer GJ, Müller R**. 2008. Discovering the hidden secondary metabolome of Myxococcus xanthus: a study of intraspecific diversity. Appl Environ Microbiol **74:**3058–3068. http://dx.doi.org/10.1128/AEM.02863-07.
20. **Doroghazi JR, Albright JC, Goering AW, Ju K-S, Haines RR, Tchalukov KA, Labeda DP, Kelleher NL, Metcalf WW**. 2014. A roadmap for natural product discovery based on large-scale genomics and metabolomics. Nat Chem Biol **10:**963–968. http://dx.doi.org/10.1038/nchembio.1659.
21. **Duncan KR, Crüsemann M, Lechner A, Sarkar A, Li J, Ziemert N, Wang M, Bandeira N, Moore BS, Dorrestein PC, Jensen PR**. 2015. Molecular networking and pattern-based genome mining improves discovery of biosynthetic gene clusters and their products from Salinispora species. Chem Biol **22:**460–471. http://dx.doi.org/10.1016/j.chembiol.2015.03.010.
22. **Lommen A**. 2009. MetAlign: interface-driven, versatile metabolomics tool for hyphenated full-scan mass spectrometry data preprocessing. Anal Chem **81:**3079–3086. http://dx.doi.org/10.1021/ac900036d.
23. **Katajamaa M, Miettinen J, Oresic M**. 2006. MZmine: toolbox for processing and visualization of mass spectrometry based molecular profile data. Bioinformatics **22:**634–636. http://dx.doi.org/10.1093/bioinformatics/btk039.
24. **Pluskal T, Castillo S, Villar-Briones A, Oresic M**. 2010. MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. BMC Bioinformatics **11:**395. http://dx.doi.org/10.1186/1471-2105-11-395.
25. **Kuhl C, Tautenhahn R, Böttcher C, Larson TR, Neumann S**. 2012. CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. Anal Chem **84:**283–289. http://dx.doi.org/10.1021/ac202450g.
26. **Tautenhahn R, Patti GJ, Rinehart D, Siuzdak G**. 2012. XCMS online: a web-based platform to process untargeted metabolomic data. Anal Chem **84:**5035–5039. http://dx.doi.org/10.1021/ac300698c.
27. **Katajamaa M, Oresic M**. 2007. Data processing for mass spectrometry-based metabolomics. J Chromatogr A **1158:**318–328. http://dx.doi.org/10.1016/j.chroma.2007.04.021.
28. **Lange E, Tautenhahn R, Neumann S, Gröpl C**. 2008. Critical assessment of alignment procedures for LC-MS proteomics and metabolomics measurements. BMC Bioinformatics **9:**375. http://dx.doi.org/10.1186/1471-2105-9-375.
29. **Nguyen DD, Wu C-H, Moree WJ, Lamsa A, Medema MH, Zhao X, Gavilan RG, Aparicio M, Atencio L, Jackson C, Ballesteros J, Sanchez J, Watrous JD, Phelan VV, van de Wiel C, Kersten RD, Mehnaz S, De Mot R, Shank EA, Charusanti P**. 2013. MS/MS networking guided analysis of molecule and gene cluster families. Proc Natl Acad Sci U S A **110:**E2611–E2620. http://dx.doi.org/10.1073/pnas.1303471110.
30. **Watrous J, Roach P, Alexandrov T, Heath BS, Yang JY, Kersten RD, van der Voort M, Pogliano K, Gross H, Raaijmakers JM, Moore BS, Laskin J, Bandeira N, Dorrestein PC**. 2012. Mass spectral molecular networking of living microbial colonies. Proc Natl Acad Sci U S A **109:**E1743–E1752. http://dx.doi.org/10.1073/pnas.1203689109.
31. **Ng J, Bandeira N, Liu W-T, Ghassemian M, Simmons TL, Gerwick WH, Linington R, Dorrestein PC, Pevzner PA**. 2009. Dereplication and de novo sequencing of nonribosomal peptides. Nat Methods **6:**596–599. http://dx.doi.org/10.1038/nmeth.1350.
32. **Liu W-T, Ng J, Meluzzi D, Bandeira N, Gutierrez M, Simmons TL, Schultz AW, Linington RG, Moore BS, Gerwick WH, Pevzner PA, Dorrestein PC**. 2009. Interpretation of tandem mass spectra obtained from cyclic nonribosomal peptides. Anal Chem **81:**4200–4209. http://dx.doi.org/10.1021/ac900114t.
33. **Yang JY, Sanchez LM, Rath CM, Liu X, Boudreau PD, Bruns N, Glukhov E, Wodtke A, De Felicio R, Fenner A, Wong WR, Linington RG, Zhang L, Debonsi HM, Gerwick WH, Dorrestein PC**. 2013. Molecular networking as a dereplication strategy. J Nat Prod **76:**1686–1699. http://dx.doi.org/10.1021/np400413s.
34. **Bowman JP**. 2007. Bioactive compound synthetic capacity and ecolog-

ical significance of marine bacterial genus *Pseudoalteromonas*. Mar Drugs **5:**220–241. http://dx.doi.org/10.3390/md504220.

35. **Holmström C, Kjelleberg S**. 1999. Marine Pseudoalteromonas species are associated with higher organisms and produce biologically active extracellular agents. FEMS Microbiol Ecol **30:**285–293.
36. **Vynne NG, Mansson M, Gram L**. 2012. Gene sequence based clustering assists in dereplication of Pseudoalteromonas luteoviolacea strains with identical inhibitory activity and antibiotic production. Mar Drugs **10:** 1729–1740. http://dx.doi.org/10.3390/md10081729.
37. **Gram L, Melchiorsen J, Bruhn JB**. 2010. Antibacterial activity of marine culturable bacteria collected from a global sampling of ocean surface waters and surface swabs of marine organisms. Mar Biotechnol (NY) **12:**439–451. http://dx.doi.org/10.1007/s10126-009-9233-y.
38. **Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R**. 2005. The microbial pan-genome. Curr Opin Genet Dev **15:**589–594. http://dx.doi.org/10.1016/j.gde.2005.09.006.
39. **Vesth T, Lagesen K, Acar Ö, Ussery D**. 2013. CMG-biotools, a free workbench for basic comparative microbial genomics. PLoS One **8:**e60120. http://dx.doi.org/10.1371/journal.pone.0060120.
40. **Thomas T, Evans FF, Schleheck D, Mai-Prochnow A, Burke C, Penesyan A, Dalisay DS, Stelzer-Braid S, Saunders N, Johnson J, Ferriera S, Kjelleberg S, Egan S**. 2008. Analysis of the Pseudoalteromonas tunicata genome reveals properties of a surface-associated life style in the marine environment. PLoS One **3:**e3252. http://dx.doi.org/10.1371/journal.pone.0003252.
41. **Médigue C, Krin E, Pascal G, Barbe V, Bernsel A, Bertin PN, Cheung F, Cruveiller S, D'Amico S, Duilio A, Fang G, Feller G, Ho C, Mangenot S, Marino G, Nilsson J, Parrilli E, Rocha EP, Rouy Z, Sekowska A, Tutino ML, Vallenet D, von Heijne G, Danchin A**. 2005. Coping with cold: the genome of the versatile marine Antarctica bacterium Pseudoalteromonas haloplanktis TAC125. Genome Res **15:** 1325–1335. http://dx.doi.org/10.1101/gr.4126905.
42. **Moffitt MC, Neilan BA**. 2003. Evolutionary affiliations within the superfamily of ketosynthases reflect complex pathway associations. J Mol Evol **56:**446–457. http://dx.doi.org/10.1007/s00239-002-2415-0.
43. **Jenke-Kodama H, Börner T, Dittmann E**. 2006. Natural biocombinatorics in the polyketide synthase genes of the actinobacterium *Streptomyces avermitilis*. PLoS Comput Biol **2:**e132. http://dx.doi.org/10.1371/journal.pcbi.0020132.
44. **Zucko J, Long PF, Hranueli D, Cullum J**. 2012. Horizontal gene transfer and gene conversion drive evolution of modular polyketide synthases. J Ind Microbiol Biotechnol **39:**1541–1547. http://dx.doi.org/10.1007/s10295-012-1149-2.
45. **Jenke-Kodama H, Dittmann E**. 2009. Evolution of metabolic diversity: insights from microbial polyketide synthases. Phytochemistry **70:** 1858–1866. http://dx.doi.org/10.1016/j.phytochem.2009.05.021.
46. **Lin X, Yang F, Zhou L, Yin P, Kong H, Xing W, Lu X, Jia L, Wang Q, Xu G**. 2012. A support vector machine-recursive feature elimination feature selection method based on artificial contrast variables and mutual information. J Chromatogr B Analyt Technol Biomed Life Sci **910:** 149–155. http://dx.doi.org/10.1016/j.jchromb.2012.05.020.
47. **Lin X, Wang Q, Yin P, Tang L, Tan Y, Li H, Yan K, Xu G**. 2011. A method for handling metabonomics data from liquid chromatography/mass spectrometry: combinational use of support vector machine recursive feature elimination, genetic algorithm and random forest for feature selection. Metabolomics **7:**549–558. http://dx.doi.org/10.1007/s11306-011-0274-7.
48. **Zhang X, Enomoto K**. 2011. Characterization of a gene cluster and its putative promoter region for violacein biosynthesis in *Pseudoalteromonas* sp. 520P1. Appl Microbiol Biotechnol **90:**1963–1971. http://dx.doi.org/10.1007/s00253-011-3203-9.
49. **Agarwal V, El Gamal AA, Yamanaka K, Poth D, Kersten RD, Schorn M, Allen EE, Moore BS**. 2014. Biosynthesis of polybrominated aromatic organic compounds by marine bacteria. Nat Chem Biol **10:**640–647. http://dx.doi.org/10.1038/nchembio.1564.
50. **Laatsch H, Renneberg B, Hanefeld U, Kellner M, Pudleiner H, Hamprecht G, Kraemer HP, Anke H**. 1995. Structure-activity-relationships of phenylpyrroles and benzoylpyrroles. Chem Pharm Bull **43:**537–546. http://dx.doi.org/10.1248/cpb.43.537.
51. **Hornemann U, Hurley LH, Speedie MK, Floss HG**. 1971. The biosynthesis of indolmycin. J Am Chem Soc **93:**3028–3035.
52. **Woodard RW, Mascaro L, Horhammer R, Eisenstein S, Floss HG**. 1980. Stereochemistry of indolmycin biosynthesis. Steric course of C- and N-methylation reactions. J Am Chem Soc **102:**6314–6318.
53. **Speedie MK, Hornemann U, Floss HG**. 1975. Isolation and characterization of tryptophan and indolepyruvate C-methyltransferase. Enzymes involved in indolmycin biosynthesis in Streptomyces griseus. J Biol Chem **250:**7819–7825.
54. **Du Y-L, Alkhalaf LM, Ryan KS**. 2015. In vitro reconstitution of indolmycin biosynthesis reveals the molecular basis of oxazolinone assembly. Proc Natl Acad Sci U S A **112:**2717–2722. http://dx.doi.org/10.1073/pnas.1419964112.
55. **Vecchione JJ, Sello JK**. 2009. A novel tryptophanyl-tRNA synthetase gene confers high-level resistance to indolmycin. Antimicrob Agents Chemother **53:**3972–3980. http://dx.doi.org/10.1128/AAC.00723-09.
56. **Kitabatake M, Ali K, Demain A, Sakamoto K, Yokoyama S, Söll D**. 2002. Indolmycin resistance of Streptomyces coelicolor A3(2) by induced expression of one of its two tryptophanyl-tRNA synthetases. J Biol Chem **277:**23882–23887. http://dx.doi.org/10.1074/jbc.M202639200.
57. **El-Sayed AK, Hothersall J, Cooper SM, Stephens E, Simpson TJ, Thomas CM**. 2003. Characterization of the mupirocin biosynthesis gene cluster from Pseudomonas fluorescens NCIMB 10586. Chem Biol **10:** 419–430. http://dx.doi.org/10.1016/S1074-5521(03)00091-7.
58. **Fukuda D, Haines AS, Song Z, Murphy AC, Hothersall J, Stephens ER, Gurney R, Cox RJ, Crosby J, Willis CL, Simpson TJ, Thomas CM**. 2011. A natural plasmid uniquely encodes two biosynthetic pathways creating a potent anti-MRSA antibiotic. PLoS One **6:**e18031. http://dx.doi.org/10.1371/journal.pone.0018031.
59. **Shiozawa H, Kagasaki T, Torikata A, Tanaka N, Fujimoto K, Hata T, Furukawa Y, Takahashi S**. 1995. Thiomarinols B and C, new antimicrobial antibiotics produced by a marine bacterium. J Antibiot **48:**907–909. http://dx.doi.org/10.7164/antibiotics.48.907.
60. **Shiozawa H, Shimada A, Takahashi S**. 1997. Thiomarinols D, E, F and G, new hybrid antimicrobial antibiotics produced by a marine bacterium; isolation, structure and antimicrobial activity. J Antibiot **50:**449–452. http://dx.doi.org/10.7164/antibiotics.50.449.
61. **Udwary DW, Zeigler L, Asolkar RN, Singan V, Lapidus A, Fenical W, Jensen PR, Moore BS**. 2007. Genome sequencing reveals complex secondary metabolome in the marine actinomycete Salinispora tropica. Proc Natl Acad Sci U S A **104:**10376–10381. http://dx.doi.org/10.1073/pnas.0700962104.
62. **Thomson NR, James KD, Harris DE, Quail MA, Bentley SD, Harper D, Bateman A, Brown S, Collins M, Cronin A, Fraser A, Goble A, Hidalgo J, Hornsby T, Howarth S, Larke L, Murphy L, Oliver K, Rabbinowitsch E, Rutherford K, Rutter S, Seeger K, Saunders D, Sharp S, Squares R, Squares S, Taylor K, Warren T, Woodward J, Barrell BG, Parkhill J**. 2002. Complete genome sequence of the model actinomycete Streptomyces coelicolor A3(2). Nature **417:**141–147. http://dx.doi.org/10.1038/417141a.
63. **Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS, Deboy RT, Davidsen TM, Mora M, Scarselli M, Margarit y Ros I, Peterson JD, Hauser CR, Sundaram JP, Nelson WC, Madupu R, Brinkac LM, Dodson RJ, Rosovitz MJ, Sullivan SA, Daugherty SC, Haft DH, Selengut J, Gwinn ML, Zhou L, Zafar N, Khouri H, Radune D, Dimitrov G, Watkins K, O'Connor KJB, Smith S, Utterback TR, White O, Rubens CE, Grandi G, Madoff LC, Kasper DL, Telford JL, Wessels MR, Rappuoli R, Fraser CM**. 2005. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome." Proc Natl Acad Sci U S A **102:**13950–13955.
64. **Osbourn A**. 2010. Secondary metabolic gene clusters: evolutionary toolkits for chemical innovation. Trends Genet **26:**449–457. http://dx.doi.org/10.1016/j.tig.2010.07.001.
65. **Firn RD**. 2003. Bioprospecting—why is it so unrewarding? Biodivers Conserv **12:**207–216. http://dx.doi.org/10.1023/A:1021928209813.
66. **Tobie WC**. 1935. The pigment of Bacillus violaceus: I. The production, extraction, and purification of violacein. J Bacteriol **29:**223–227.
67. **Yada S, Wang Y, Zou Y, Nagasaki K, Hosokawa K, Osaka I, Arakawa R, Enomoto K**. 2008. Isolation and characterization of two groups of novel marine bacteria producing violacein. Mar Biotechnol (NY) **10:** 128–132. http://dx.doi.org/10.1007/s10126-007-9046-9.
68. **Von Wittenau MS, Els H**. 1963. Chemistry of indolmycin. J Am Chem Soc **85:**3425–3431. http://dx.doi.org/10.1021/ja00904a028.
69. **Månsson M, Phipps RK, Gram L, Munro MH, Larsen TO, Nielsen KF**. 2010. Explorative solid-phase extraction (E-SPE) for accelerated microbial natural product discovery, dereplication, and purification. J Nat Prod **73:**1126–1132. http://dx.doi.org/10.1021/np100151y.
70. **Krug D, Zurek G, Schneider B, Garcia R, Müller R**. 2008. Efficient

mining of myxobacterial metabolite profiles enabled by liquid chromatography-electrospray ionisation-time-of-flight mass spectrometry and compound-based principal component analysis. Anal Chim Acta **624:**97–106. http://dx.doi.org/10.1016/j.aca.2008.06.036.

71. **Vos M, Velicer GJ**. 2006. Genetic population structure of the soil bacterium Myxococcus xanthus at the centimeter scale. Appl Environ Microbiol **72:**3615–3625. http://dx.doi.org/10.1128/AEM.72.5.3615-3625.2006.
72. **Antón J, Lucio M, Peña A, Cifuentes A, Brito-Echeverría J, Moritz F, Tziotis D, López C, Urdiain M, Schmitt-Kopplin P, Rosselló-Móra R**. 2013. High metabolomic microdiversity within co-occurring isolates of the extremely halophilic bacterium Salinibacter ruber. PLoS One **8:**e64701. http://dx.doi.org/10.1371/journal.pone.0064701.
73. **Li S, Kang L, Zhao X-M**. 2014. A survey on evolutionary algorithm based hybrid intelligence in bioinformatics. Biomed Res Int **2014:**362738. http://dx.doi.org/10.1155/2014/362738.
74. **Boccard J, Kalousis A, Hilario M, Lantéri P, Hanafi M, Mazerolles G, Wolfender J, Carrupt P, Rudaz S**. 2010. Standard machine learning algorithms applied to UPLC-TOF/MS metabolic fingerprinting for the discovery of wound biomarkers in Arabidopsis thaliana. Chemometr Intell Lab Syst **104:**20–27. http://dx.doi.org/10.1016/j.chemolab.2010.03.003.
75. **Mahadevan S, Shah SL, Marrie TJ, Slupsky CM**. 2008. Analysis of metabolomic data using support vector machines. Anal Chem **80:**7562–7570. http://dx.doi.org/10.1021/ac800954c.
76. **Kildgaard S, Mansson M, Dosen I, Klitgaard A, Frisvad JC, Larsen TO, Nielsen KF**. 2014. Accurate dereplication of bioactive secondary metabolites from marine-derived fungi by UHPLC-DAD-QTOFMS and a MS/HRMS Library. Mar Drugs **12:**3681–3705. http://dx.doi.org/10.3390/md12063681.
77. **Nielsen KF, Månsson M, Rank C, Frisvad JC, Larsen TO**. 2011. Dereplication of microbial natural products by LC-DAD-TOFMS. J Nat Prod **74:**2338–2348. http://dx.doi.org/10.1021/np200254t.
78. **Smith CA, O'Maille G, Want EJ, Qin C, Trauger SA, Brandon TR, Custodio DE, Abagyan R, Siuzdak G**. 2005. METLIN: a metabolite mass spectral database. Ther Drug Monit **27:**747–751. http://dx.doi.org/10.1097/01.ftd.0000179845.53213.39.
79. **Chambers MC, Maclean B, Burke R, Amodei D, Ruderman DL, Neumann S, Gatto L, Fischer B, Pratt B, Egertson J, Hoff K, Kessner D, Tasman N, Shulman N, Frewen B, Baker TA, Brusniak M, Paulse C, Creasy D, Flashner L, Kani K, Moulding C, Seymour SL, Nuwaysir LM, Lefebvre B, Kuhlmann F, Roark J, Rainer P, Detlev S, Hemenway T, Huhmer A, Langridge J, Connolly B, Chadick T, Holly K, Eckels J, Deutsch EW, Moritz RL, Katz JE, Agus DB, MacCoss M, Tabb DL, Mallick P**. 2012. A cross-platform toolkit for mass spectrometry and proteomics. Nat Biotechnol **30:**918–920. http://dx.doi.org/10.1038/nbt.2377.
80. **Smoot ME, Ono K, Ruscheinski J, Wang P-L, Ideker T**. 2011. Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics **27:**431–432. http://dx.doi.org/10.1093/bioinformatics/btq675.
81. **Medema MH, Takano E, Breitling R**. 2013. Detecting sequence homology at the gene cluster level with MultiGeneBlast. Mol Biol Evol **30:**1218–1223. http://dx.doi.org/10.1093/molbev/mst025.
82. **Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL**. 2009. BLAST+: architecture and applications. BMC Bioinformatics **10:**421. http://dx.doi.org/10.1186/1471-2105-10-421.
83. **Sullivan MJ, Petty NK, Beatson SA**. 2011. Easyfig: a genome comparison visualizer. Bioinformatics **27:**1009–1010. http://dx.doi.org/10.1093/bioinformatics/btr039.