Insights into Secondary Metabolism from a Global Analysis of Prokaryotic Biosynthetic Gene Clusters

Peter Cimermancic,1,10 Marnix H. Medema,2,3,10,12 Jan Claesen,1,10 Kenji Kurita,4 Laura C. Wieland Brown,5 Konstantinos Mavrommatis,6 Amrita Pati,9 Paul A. Godfrey,7 Michael Koehrsen,7 Jon Clardy,8 Bruce W. Birren,7 Eriko Takano,2,11 Andrej Sali,1,9 Roger G. Linington,4 and Michael A. Fischbach1,*

1Department of Bioengineering and Therapeutic Sciences and the California Institute for Quantitative Biosciences, University of California, San Francisco, San Francisco, CA 94158, USA
2Department of Microbial Physiology, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, Nijenborgh 7, 9747AG Groningen, The Netherlands
3Groningen Bioinformatics Centre, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, Nijenborgh 7, 9747AG Groningen, The Netherlands
4Department of Chemistry and Biochemistry, University of California, Santa Cruz, Santa Cruz, CA 95064, USA
5Department of Chemistry, Indiana University, Bloomington, IN 47405, USA
6US Department of Energy, Joint Genome Institute, Walnut Creek, CA 94598, USA
7The Broad Institute, Cambridge, MA 02142, USA
8Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA 02115, USA
9Department of Pharmaceutical Chemistry, University of California, San Francisco, San Francisco, CA 94158, USA
10Co-first author
11Present address: Manchester Institute of Biotechnology, Faculty of Life Sciences, University of Manchester, Manchester M1 7DN, UK
12Present address: Max Planck Institute for Marine Microbiology, Celsiusstraße 1, 28359 Bremen, Germany
*Correspondence: fischbach@fischbachgroup.org
http://dx.doi.org/10.1016/j.cell.2014.06.034

SUMMARY

Although biosynthetic gene clusters (BGCs) have been discovered for hundreds of bacterial metabolites, our knowledge of their diversity remains limited. Here, we used a novel algorithm to systematically identify BGCs in the extensive extant microbial sequencing data. Network analysis of the predicted BGCs revealed large gene cluster families, the vast majority uncharacterized. We experimentally characterized the most prominent family, consisting of two subfamilies of hundreds of BGCs distributed throughout the Proteobacteria; their products are aryl polyenes, lipids with an aryl head group conjugated to a polyene tail. We identified a distant relationship to a third subfamily of aryl polyene BGCs, and together the three subfamilies represent the largest known family of biosynthetic gene clusters, with more than 1,000 members. Although these clusters are widely divergent in sequence, their small molecule products are remarkably conserved, indicating for the first time the important roles these compounds play in Gram-negative cell biology.

INTRODUCTION

Microbial natural products are widely used in human and veterinary medicine, agriculture, and manufacturing, and are known to mediate a variety of microbe-host and microbe-microbe interactions. Connecting these natural products to the genes that encode them is revolutionizing their study, enabling genome sequence data to guide the discovery of new molecules (Bergmann et al., 2007; Challis, 2008; Franke et al., 2012; Freeman et al., 2012; Kersten et al., 2011; Laureti et al., 2011; Lautru et al., 2005; Letzel et al., 2013; Nguyen et al., 2008; Olilnyk et al., 2007; Schneiker et al., 2007; Walsh and Fischbach, 2010; Winter et al., 2011). The thousands of prokaryotic genomes in sequence databases provide an opportunity to generalize this approach through the identification of biosynthetic gene clusters (BGCs): sets of physically clustered genes that encode the biosynthetic enzymes for a natural product pathway.

Besides core biosynthetic enzymes, many BGCs also harbor enzymes to synthesize specialized monomers for a pathway. For example, the erythromycin gene cluster encodes a set of enzymes for biosynthesis of two deoxysugars, d-desosamine and l-mycarose, that are appended to the polyketide aglycone (Olilnyk et al., 2007; Staunton and Weissman, 2001), while BGCs for glycopeptide antibiotics contain enzymes to synthesize the nonproteinogenic amino acids l-hydroxytyrosine, 4-hydroxyphenylglycine, and 3,5-dihydroxyphenylglycine that their core nonribosomal peptide synthetases use in the assembly of their peptidic scaffolds (Kahne et al., 2005; Pelzer et al., 1999). In many cases, transporters, regulatory elements, and genes that mediate host resistance are also contained within the BGC (Walsh and Fischbach, 2010). Although some BGCs are so well understood that the biosynthesis of their small molecule product has been reconstituted in heterologous hosts (Pfeifer et al., 2001) or in vitro using purified enzymes (Lowry...
et al., 2013; Sattely et al., 2008), little is known about the vast majority of BGCs, even those that have been connected to a small molecule product.

Here, we report the results of a systematic effort to identify and categorize BGCs in 1,154 sequenced genomes spanning the prokaryotic tree of life. We envisioned that the resulting “global map” of biosynthesis would enable BGCs to be systematically selected for characterization by searching for, e.g., biosynthetic novelty, presence in undermined taxa, or patterns of phylogenetic distribution that indicate functional importance. Surprisingly, the map revealed large and very widely distributed BGC families of unknown function. We experimentally characterized the most prominent of these families, leading to the unexpected finding that gene clusters responsible for producing aryl polyene carboxylic acids constitute the largest BGC family in the sequence databases.

RESULTS AND DISCUSSION

The ClusterFinder Algorithm Detects BGCs of Both Known and Unknown Classes

Several algorithms have been developed for the automated prediction of BGCs in microbial genomes (Khaldi et al., 2010; Li et al., 2009; Medema et al., 2011; Starcevic et al., 2008; Weber et al., 2009), but each of these tools is limited to the detection of one or more well-characterized gene cluster classes. As a more general solution to the gene cluster identification problem, we developed a hidden Markov model-based probabilistic algorithm, ClusterFinder, that aims to identify gene clusters of both known and unknown classes. ClusterFinder is based on a training set of 732 BGCs with known small molecule products that we compiled and manually curated (Table S1). To scan a genome for BGCs, it converts a nucleotide sequence into a string of contiguous Pfam domains and assigns each domain a probability of being part of a gene cluster, based on the frequencies at which these domains occur in the BGC and non-BGC training sets, and the identities of neighboring domains (Figure 1A, Experimental Procedures). Since ClusterFinder is based solely on Pfam domain frequencies, and nature uses distinct assemblages of the same enzyme superfamilies to construct unrelated natural product classes, ClusterFinder exhibits little training set bias and is capable of identifying new classes of gene clusters effectively (See Experimental Procedures for a detailed description of how we validated ClusterFinder).
A Global Phylogenomic Analysis of BGCs Provides a Quantitative Perspective on Bacterial Secondary Metabolite Biosynthesis

Our method predicted a total of 33,351 putative BGCs (with an estimated false-positive rate of 5%) in 1,154 genomes of organisms throughout the prokaryotic tree of life (Figures 1C and 1D, see also Extended Results Section Systematic Identification of Gene Clusters from Bacterial Genomes); we subjected these putative BGCs to an extensive phylogenomic analysis (Extended Results Sections Prolific Producers Harbor Exceptionally Large Complements of Gene Clusters and The Relationship between Phylogeny and Gene Clusters Varies Tremendously across the Bacterial Tree of Life, Figures S1A–S1D, S2, S3, Tables S1 and S2 available online). We divided the predicted BGCs into two categories—high confidence (10,724; used in all subsequent analyses) and low confidence (22,627)—based on assignment to one of ~20 well-validated BGC classes or on manual inspection for clusters that could not be assigned to any known class. Within the high-confidence set, 7,377 of the predicted gene clusters (69%) were not detected by antiSMASH (Blin et al., 2013; Medema et al., 2011); the difference is due primarily to the fact that antiSMASH does not detect certain BGC classes (including many oligosaccharides), highlighting the need for a tool that identifies BGCs independent of class (Figure 1B).

Strikingly, 40% of all predicted BGCs encode saccharides, more than twice the size of the next largest class. Notably, only 13% of previously reported BGCs encode the biosynthesis of saccharides (Extended Results Section Saccharides Are the Largest Class of Gene Clusters). Ninety-three percent of species harbor saccharide gene clusters, and in 33% of species, more than half of the predicted gene clusters encode saccharides. Cell-associated saccharides such as lipopolysaccharides (Park et al., 2009), capsular polysaccharides (Kadioglu et al., 2008), and polysaccharide A (Mazmanian et al., 2005; Mazmanian et al., 2008) are known to play key roles in microbe-host and microbe-microbe interactions, while diffusible saccharides have a range of biological activities, most notably antibacterial (Flatt and Mahmud, 2007; Weitnauer et al., 2001). The functions of many of the putative saccharide BGCs are still a mystery: 32%, including BGCs from entirely unexplored genera, are not closely related to any known gene cluster (Figure S1E). Saccharide BGC repertoires are also surprisingly diverse: only 37% occur in the genomes of two species chosen at random from the same genus (compared to 43% for polyketides, 60% for terpenoids, and 74% for fatty acids, Figure S1F). The abundance of novel oligosaccharide BGC families raises the possibility that more clinically relevant saccharides such as the antidiabetic drug acarbose and the antibiotics gentamicin and avilamycin will be discovered (Kersten et al., 2013). Another BGC class of unexpectedly large size is the ribosomally synthesized and posttranslationally modified peptides (RiPPs; Arnison et al., 2013). Notably, RiPP BGCs are as prevalent in our data set as those encoding nonribosomal peptides (Figure 1B).

A BGC Distance Network Reveals Unexplored Regions of the Biosynthetic Universe

We next sought to study the relationships among BGCs systematically, with the ultimate goal of creating a global BGC map that could be searched systematically to identify clusters of biosynthetic or taxonomic interest. We adapted a measure of the evolutionary distance between multidomain proteins (Lin et al., 2006) to calculate an all-by-all distance matrix for the 10,724 BGCs in our high-confidence set along with the 732 members of our training set. Using MCL clustering to identify groups of related nodes, we define 905 BGC families with distinct core genetic components. The resulting BGC distance network (Figure 2, see Extended Results Section A Global Map of Biosynthesis Based on a Gene Cluster Distance Metric, and Figure S4) revealed an unexpected finding: the presence of large cliques that represent very widely distributed BGC families without any experimentally characterized members.

While most known families of secondary metabolites are unique to a small set of organisms, a few are taxonomically widespread. These include the O-antigens, capsular polysaccharides, carotenoids, and NRPS-independent siderophores, which can all be clearly distinguished as prominent cliques within our distance network. From a fundamental microbiological perspective, these are among the most important families of molecules produced by microbes and, as such, they have been very intensively studied (Challis, 2005; Rehm, 2010; Samuel and Reeves, 2003; Walter and Strack, 2011). Although we had anticipated finding small gene cluster families of unknown function, we were surprised to discover families harboring hundreds of uncharacterized clusters, distributed widely throughout entire bacterial phyla.

We selected the most prominent of these families for experimental characterization: a set of 811 BGCs, distributed between two subfamilies (hereafter, subfamily 1 and 2), that were not detected by any of the existing BGC identification tools (e.g., antiSMASH), likely because the ketosynthase and adenylation domains they harbor are from uncharacterized, evolutionarily distant clades. BGCs in this family are ~20 kb in size and harbor a core set of genes that include adenylation, ketosynthase, acyl/glycosyltransferase, ketoreductase, dehydratase, thiolation, and thiosterase domains, as well as an outer membrane lipoprotein carrier protein and an MMPL family transporter (Figures 3A, S5, and Data S1). These clusters are found in a wide variety of Grammproteobacteria (Acinetobacter, Aggregatibacter, Escherichia, Klebsiella, Pantoaea, Pseudalteromonas, Pseudomonas, Serratia, Shewanella, Vibrio, and Yersinia), as well as a broader set of Beta-(Burkholderia, Neisseria) and Epsilonproteobacteria (Campylobacter) (Figure 3A).

The Unexplored BGC Family Encodes the Biosynthesis of Aryl Polyene Carboxylic Acids

We set out to identify the small molecule product of two clusters in the family, one each from subfamilies 1 and 2. We used circular polymerase extension cloning (CPEC) (Quan and Tian, 2009) to amplify and assemble the 18 gene, 15.5 kb cluster from E. coli CFT073 (c1186-c1204), and we transferred a plasmid harboring the cluster into E. coli Top10. The transformants exhibited a strong yellow pigmentation that was absent in the empty vector control strain and not observed in the native host strain (Figure 3C), but the pigment did not appear to diffuse into liquid or solid culture medium. We liberated the pigment from an organic extract of the cell mass by mild
base hydrolysis and purified it by HPLC. Comparative HPLC analysis of extracts from the cluster+ and cluster− strains revealed the presence of a compound unique to the cluster+ strain with an absorption maximum of 425 nm, consistent with a yellow chromophore (Figure S7E). Purification of milligram quantities of the compound for structural characterization required the development of an isolation procedure that rigorously excluded exposure to light. A combination of 1D- and 2D-NMR experiments and high-resolution MS on the purified compound revealed that it was an aryl polyene (APE) carboxylic acid consisting of a 4-hydroxy-3-methylphenyl head group conjugated to a hexaenoic acid (Figures 3B, S7C, S7E, and S7F and Data S2).

To study the 20 gene, 18.9 kb cluster from Vibrio fischeri ES114 (VF0841-VF0860), we first deleted the cluster from its native producer. The yellow pigmentation that is observed in wild-type V. fischeri under normal laboratory growth conditions was absent in the V. fischeri knockout strain (Figure 3C). We then proceeded to amplify, assemble, and introduce the V. fischeri cluster into E. coli Top10, but the native cluster failed to confer yellow pigmentation on its heterologous host. We then constructed a modified variant of the cluster in which the ermE∗
A promoter was inserted upstream of the operon starting with VF0844. Introduction of this construct into E. coli resulted in a yellow-pigmented strain that produced a new compound with an absorption maximum at 425 nm (Figure 3C). Purification of the V. fischeri compound and analysis by a combination of 1D- and 2D-NMR experiments and high-resolution MS revealed a structure with a similar scaffold to the E. coli APE but a 4-hydroxy-3,5-dimethylphenyl head group (Figures 3B and 3D).
To our surprise, the Sequence Databases The Aryl Polyene BGCs Are the Largest Family in the ylic acids. the cluster representatives from this family encode APE carbox-S7D–S7F and Data S2). Taken together, these data suggest that the cluster representatives from this family encode APE carboxylic acids.

The Aryl Polyene BGCs Are the Largest Family in the Sequence Databases
To our surprise, the E. coli and V. fischeri APEs are similar in structure to flexirubin (Fuchs et al., 2013; McBride et al., 2009), a pigment that was previously isolated from the CFB group bacterium Flexibacter elegans, and xanthomonadin (Goel et al., 2002), the compound that gives Xanthomonas spp. their characteristic yellow color. The biosynthetic genes for flexirubin and xanthomonadin are known (Fuchs et al., 2013; Goel et al., 2002; McBride et al., 2009); both are part of a smaller, distinct subfamily in the ClusterFinder results set (subfamily 3 in Figure 3A), but little else is known about the genes in either cluster. Intriguingly, although the clusters in subfamily 3 share similar Pfam domain content to those in subfamilies 1 and 2, the percent identities of their constituent proteins are very low (<20% for some amino acid sequences, see Figure S6A). When we turned to a more sensitive approach in which we used MultiGeneBlast (Medema et al., 2013) to look for sequence similarity at the level of the entire gene cluster, we observed distant but recognizable homology between multiple gene pairs from BGCs from subfamily 3 and subfamilies 1 and 2, indicating that the APE clusters might share a common ancestor. Indeed, when we performed a maximum-likelihood phylogenetic analysis of the ketosynthase and adenylation enzyme superfamilies based on structure-guided multiple sequence alignments (Extended Results Section Phylogenetic Analysis of APE Ketosynthase and Adenylation Enzymes, Figures S6B–S6D), we found that the APE KS and A enzymes cluster together in separate uncharacterized clades that are only distantly related to all other known members of these enzyme superfamilies. Based on this evidence, we conclude that the three subfamilies together comprise a single BGC family of >1,000 gene clusters (Figure 3A). Notably, the APE family is, to our knowledge, the largest family of gene clusters in the database, even exceeding the size of the well-known carotenoids (870 clusters, as detected using the same methods, see Table S3).

The lack of homology even between the xanthomonadin and flexirubin biosynthetic genes (both in subfamily 3) is so profound that these pigments have never been connected in the literature: indeed, both previously discovered APEs have been proposed as chemosystematic markers of a genus (Flexibacter and Xanthomonas) because of their “limited distribution among bacte-ria” (Fautz and Reichenbach, 1979; Jenkins and Starr, 1982; Reichenbach et al., 1980; Starr et al., 1977; Wang et al., 2013). Our results, however, show that APE family BGCs are widely distributed throughout the Gram-negative bacterial tree of life (Figures 4 and S3). Notably, their pattern of phylogenetic distribution is markedly discontinuous: clusters are present in some strains but not others of most genera (36.4% of the complete genomes in a typical genus harbor the cluster, but note the high standard deviation of 37.9%). The most parsimonious explanations for this distribution pattern are frequent gene cluster loss from the descendants of a cluster-harboring ancestor, or frequent horizontal transfer among the descendants of a cluster-negative ancestor. Two lines of evidence support the possibility of frequent horizontal transfer: The family 1 cluster from E. coli O157:H7 is located on an O-island (Dong and Schellhorn, 2009), and the family 2 cluster from Acinetobacter sp. ADP1 resides on an element that has been identified as horizontally transferred (Barbe et al., 2004). Their broad distribution, and the fact that such widely divergent gene clusters have small molecule products that are so similar in structure, suggests the possibility that aryl polyenes play an important role in Gram-negative cell biology.

Aryl Polyenes Might Function as Protective Agents against Oxidative Stress
Xanthomonadin has been proposed to play a role in protection from photodamage by visible light (Poplawsky et al., 2000;...
Rajagopal et al., 1997), an effect that is thought to be due to its ability to quench the reactive oxygen species (ROS) that are generated when the photosensitizer used in these studies, toluidine blue, is exposed to visible light (Poplawska et al., 2000). Additionally, xanthomonaadin has been shown to protect cellular lipids from peroxidation in vitro (Rajagopal et al., 1997), and xanthomonaadin mutants show reduced epiphytic survival under conditions of natural light exposure (Poplawska et al., 2000).

Similarly, we hypothesize that other APEs play a role in protecting bacterial cells from exogenous oxidative stress. Membrane-bound APEs could reduce the concentration of free radicals that would otherwise cause damage to other cellular lipids, proteins, or nucleic acids. Notably, many bacteria that harbor APE BGCs are either commensals or pathogens of a eukaryotic host; consequently, they are likely to encounter oxidative stress from immune cells during colonization or infection.

A role for APEs in protecting Gram-negative bacteria against oxidative stress would make them analogous to the chemically similar but biosynthetically distinct Gram-positive carotenoids, whose antioxidant activity is well established. An important example is staphyloxanthin, a membrane-bound carotenoid virulence factor that is responsible for the characteristic yellow pigmentation of *S. aureus* and proposed to protect *S. aureus* from immune-mediated oxidative stress. An *S. aureus* mutant defective in the first committed step of staphyloxanthin biosynthesis exhibits higher susceptibility to various reactive oxygen species and in a neutrophil-killing assay (Clauditz et al., 2006; Liu et al., 2005). This mutant was also attenuated in murine models for subcutaneous abscess (Liu et al., 2005) and systemic infection (Liu et al., 2008). Experiments to test whether APE-deficient mutants of Gram-negative bacteria harbor colonization or pathogenesis defects will be an important step in testing this model and gaining insight into why APE gene clusters are so widely distributed throughout the Gram-negative parts of the bacterial tree of life.

Using Systematic Searches to Prioritize BGCs for Experimental Characterization

BGCs are commonly selected for characterization on the basis of chemical or enzymatic novelty. Following the example of the APE family, we anticipate that our global BGC map will enable gene clusters to be selected in a new way that is based on a criterion biologists have long used to prioritize genes: what are the most widely distributed gene clusters of unknown function? Various other prioritization criteria could be used to select BGCs of interest (Frasch et al., 2013). For example, one could select BGCs likely to encode new chemical scaffolds by searching for clusters that do not harbor conventional monomer-coupling enzymes.

Many gene cluster families still await characterization: even with conservative assumptions, we estimate the total number of bacterial BGC families (such as those encoding carotenoids or calcium-dependent lipopeptides) present in the biosphere to be \(\approx 6,000\) (Figure S1G), less than half of which are identified in our current set of genomes (\(\approx 2,400\)). Importantly, each of these 6,000 families will likely contain a range of molecules with distinct biological activities. As developments in single-cell genomics and metagenomics are opening up the exploration of a vast microbial dark matter, this number may grow even further: just in the 201 single-cell genomes of uncultivated organisms recently obtained by the JGI (Rinke et al., 2013), our method identified 947 candidate BGCs, of which 655 fall outside all known BGC classes (Figure S1H). Even among cultivated organisms, there are still many underexplored taxa (Lettzel et al., 2013) (Extended Results Section Prolific Producers Harbor Exceptionally Large Complements of Gene Clusters). For the foreseeable future, the number of gene clusters encoding molecules with distinct scaffolds will continue to rise as new genomes are sequenced, and computational approaches to systematically study their relationships will be of great value in prioritizing them for experimental characterization.

EXPERIMENTAL PROCEDURES

Genome Sequences

A set of 1,154 complete genome sequences was obtained from JGI-IMG (Markowitz et al., 2012), version 3.2 (08/17/2010).

ClusterFinder Algorithm and Training Data

The ClusterFinder prediction algorithm for BGC identification is a two-state Hidden Markov Model (HMM), with one hidden state corresponding to biosynthetic gene clusters (BGC state) and a second hidden state corresponding to the rest of the genome (non-BGC state). The training set for the BGC state was gathered using a comprehensive search of the scientific literature, which yielded 732 clusters. From these, 55 redundant BGCs were filtered out by selecting one random member from each biosynthetic gene cluster family, with a cluster family defined as a connected component in the >0.7 similarity network (see below). Thus, the final BGC state training set consisted of 677 experimentally characterized gene clusters. For the non-BGC state, non-BGC regions were collected from 100 randomly selected genomes, defined as those regions without significant sequence similarity to the BGC state training set sequences (Pfam domain similarities with E-value > 1 x 10^-10). ClusterFinder source code is available from the GitHub repository (https://github.com/petercm/ClusterFinder).

ClusterFinder Validation

The algorithm was validated in three ways. First, its output was compared to ten bacterial genomes manually annotated for BGCs (leading to an area under the ROC curve of 0.84). Second, its performance was assessed on 74 experimentally characterized BGCs outside the training set. Out of these, 70 (95%) were detected successfully. When tested alongside antiSMASH (Medema et al., 2011) on the genomes of *Pseudomonas fluorescens* Pf-5, *Streptomyces griseus* IFO13350 and *Salinispora tropica* CNB-440 (Table S4), antiSMASH detected 62 out of 65 (95%) manually annotated secondary metabolite gene clusters, while ClusterFinder detected 59 of these (91%). However, ClusterFinder identified 43 (66%) unannotated gene clusters that appeared likely to synthesize small molecule metabolites on manual inspection, whereas antiSMASH detected only five (8%). This highlights the strength of ClusterFinder in detecting gene clusters irrespective of whether they belong to known or a priori specified classes. Among the additional gene clusters detected by ClusterFinder are known gene clusters encoding the biosynthesis of, e.g., alginate and lipopolysaccharides, as well as an uncharacterized cluster that was previously predicted to encode a novel type of secondary metabolite (Hassan et al., 2010).

Type Classification of BGCs

ClusterFinder-detected biosynthetic gene clusters were classified by antiSMASH (Medema et al., 2011) to determine their subtypes (e.g., type I polyketide, nonribosomal peptide, terpenoid). The native antiSMASH types were supplemented by a list of profile HMMs for protein domains characteristic of saccharide gene clusters (Table S3), as well as by fatty acid gene clusters, which could be assigned based on the HMMs that antiSMASH uses in...
polyketide synthase annotation. Gene clusters lacking protein domains characteristic of gene cluster classes included in antiSMASH were binned in a separate class.

**BGC Distance Metric and Similarity Network**

BGC similarity networks were calculated using a modified version of the distance metric from Lin and coworkers (Lin et al., 2006) for multidomain proteins. The modified version consists of two different indices: the Jaccard index (which measures the similarity in Pfam domain sets from two BGCs) and the domain duplication index, with weights of 0.36, and 0.64, respectively. The Goodman-Kruskal γ index, which was included in the original similarity metric with a low weight of 0.01, was omitted, since the conservation of the order between two sets of domains does not appear to have an important effect on the structure of the small molecule product, except in the case of NRPS and PKS gene clusters (Fischbach et al., 2008). BGC families were calculated with a Lin similarity threshold of 0.5 and MCL clustering with I = 2.0. The similarity network was obtained using the same Lin similarity threshold and visualized using Cytoscape (Smoot et al., 2011).

**Bioinformatic Analysis of APE Gene Clusters**

Expansion of the APE BGC family was performed using manual parsing of MultiGeneBlast (Medema et al., 2013) architecture search results (with the E. coli, V. fischeri, X. campestris, and F. johnsonii APE gene clusters as query) against GenBank version 197 (08/2,013), with a 20% sequence identity cut-off and 2,000 blastp hits mapped per query sequence. APE Clusters of Orthologous Groups (COGs) were obtained using OrthoMCL (Li et al., 2003) (MCL I = 1.5, sequence identity cutoff 20%), and were used to construct a cladogram with hierarchical clustering using the modified Lin distance metric. Structure-guided multiple sequence alignments of APE A and KS domains were performed using PROMALS3D (Pei et al., 2008), and phylogenetic trees were inferred with MEGAS (Tamura et al., 2011) using the Maximum Likelihood method.

**Construction of the V. Fischeri ES114 APE-Cluster Deletion Mutant**

Oligonucleotide primers, plasmids and bacterial strains used and generated in this study are summarized in Tables S6, S7, and S8. A deletion construct was generated by fusing the ~1 kb up- and downstream regions of the V. fischeri cluster into a counterselectable suicide plasmid backbone using circular polymerase extension cloning (CPEC; Quan and Tian, 2011). This construct was introduced into V. fischeri ES114 by triparental mating and integrants were identified by selection for kanamycin resistance. Second homologous recombination events were enriched by nonselective growth, followed by induction of the counterselectable marker to identify cells that had lost the integrated plasmid backbone. Successful deletion mutants were separated from revertants and verified by colony PCR and sequencing.

**Heterologous Expression of APE Gene Clusters**

The E. coli CFT073 and V. fischeri ES114 APE clusters were amplified by PCR in three parts from genomic DNA and assembled into the SuperCos I vector backbone using either the CPEC (Quan and Tian, 2011) or Gibson (Gibson et al., 2009) method. The V. fischeri APE cluster was further modified by introducing an apramycin-resistant cassette containing the ermE’ promoter upstream of the operon starting with VF0844 using PCR targeting (Gust et al., 2004). Correct insertion of ermE’ was verified by sequencing. The heterologous expression constructs for the E. coli CFT073 and V. fischeri APE clusters were introduced into chemically competent E. coli Top10 yielding strains JC087 and JC090, respectively.

**APE Compound Purification**

For large-scale isolation and purification of APEEC and APEVF, all steps were performed in a way that avoided exposure to light. Cells were harvested from 32 L of E. coli JC087 and 80 L of V. fischeri ES114 liquid cultures, respectively. Following lyophilization, the cell material was extracted four times with 1:2 methanol/dichloromethane and the extracts were concentrated, resuspended in 1:2 methanol/dichloromethane and subjected to mild saponification with 0.5 M potassium hydroxide for 1 hr. The mixture was neutralized and the organic layer was collected, washed, dried, and resuspended in acetone for further purification by a two-step RP-HPLC method. For both extracts, the peaks with absorbance at 441 nm were collected, dried under vacuum and stored at −20°C in an amber vial prior to structural analysis (Figures S7E and S7F).

**APE Structural Characterization**

Purified APE methyl esters were analyzed by a combination of high-resolution uPLC-ESI-TOF mass spectrometry and 1D and 2D-NMR experiments, enabling the determination of their molecular formula: C12H17O2 for APEEC ([M-H]− adduct at 321.1496 m/z (Δppm = −0.310)) and C20H23O2 for APEVF ([M-H]− adduct at 335.1652 m/z (Δppm = 0.00)). Analysis of the 1H-NMR, COSY, HSQC, HMBC, ROESY and TOCSY spectra of APEEC in D3 DMSO and APEVF in D3 acetone enabled the determination of their solution structure (Figure 3B). This procedure is described in detail in the Extended Experimental Procedures section and shown in Figure S7D and S7E and Data S2.

For further details regarding the materials and methods used in this work, see the Extended Experimental Procedures.

**SUPPLEMENTAL INFORMATION**

Supplemental information includes Extended Results, Extended Experimental Procedures, seven figures, fourteen tables, and three data files and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2014.06.034.

**AUTHOR CONTRIBUTIONS**

P.C., M.H.M., J. Claesen, K.K., R.G.L., and M.A.F. designed the research, analyzed the data, and wrote the paper, with substantial input from E.T., J. Clardy, and A.S. P.C. and M.H.M. performed the computational research. J. Claesen and K.K. performed the experimental research. K.M. and A.P. provided input data and data integration into the JGI-IMG database. L.C.W.B., P.A.G., M.K., B.W.B., and M.A.F. designed an earlier version of the gene cluster identification algorithm that served as a model for the current version.

**ACKNOWLEDGMENTS**

We thank members of the Fischbach lab for helpful discussions and Rainer Bretting and an anonymous reviewer for constructive feedback on the manuscript. We thank Edward Ruby (University of Wisconsin) for providing us with V. fischeri ES114, Didier Mazel (Institut Pasteur) for plasmid pSW1917, and Mervyn Bibb (John Innes Centre) for plasmids pJL773, pJL790 and pJL10257. This work was supported by an HHMI Predoctoral Fellowship (P.C.), a Boehringer Ingelheim Fonds travel grant (M.H.M.), Grant 10463 from the GenBiotics programme of the Dutch Technology Foundation STW to ET (M.H.M.), NIH grant T2006834 (R.G.L.), the James and Eleanor Delfino Charitable Trust (K.K.), a Medical Research Program Grant from the W.M. Keck Foundation (M.A.F.), a Fellowship for Science and Engineering from the David and Lucile Packard Foundation (M.A.F.), DARPA award HR0101-12-C-0067 (M.A.F.), and NIH grants OD007290, AI101018, AI101722, and GM081879 (M.A.F.). This project has been funded in part with federal funds from the National Institute of Allergy and Infectious Diseases, NIH, Department of Health and Human Services, under Contract No.: HHSN272200900018C. M.A.F. is on the scientific advisory board of Warp Drive Bio.

Received: March 4, 2014
Revised: June 5, 2014
Accepted: June 23, 2014
Published: July 17, 2014

**REFERENCES**


