Supplemental Information

Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters

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

[1]Department of Bioengineering and Therapeutic Sciences and the California Institute for Quantitative Biosciences, University of California, San Francisco, San Francisco, CA 94158, USA

[2]Department of Microbial Physiology and [3]Groningen Bioinformatics Centre, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, Nijenborgh 7, 9747AG Groningen, The Netherlands

[4]Department of Chemistry and Biochemistry, University of California, Santa Cruz, Santa Cruz, CA 95064, USA

[5]Department of Chemistry, Indiana University, Bloomington, IN 47405, USA

[6]US Department of Energy, Joint Genome Institute, Walnut Creek, CA 94598, USA

[7]The Broad Institute, Cambridge, MA 02142, USA

[8]Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA 02115, USA

[9]Manchester Institute of Biotechnology, Faculty of Life Sciences, University of Manchester, Manchester, United Kingdom

[10]Department of Pharmaceutical Chemistry, University of California, San Francisco, San Francisco, CA 94158, USA

#Present address: Max Planck Institute for Marine Microbiology, Celsiusstraße 1, 28359 Bremen, Germany

*Denotes equal contribution

Correspondence: fischbach@fischbachgroup.org

SUPPLEMENTAL TEXT

1. Systematic identification of gene clusters from bacterial genomes

A total of 1154 complete bacterial genomes were analyzed. Draft genomes were not included in the analysis, because biosynthetic gene clusters are often highly fragmented in their assemblies. Gaps in draft assemblies occur predominantly at genes encoding large biosynthetic enzymes (Klassen and Currie, 2012).

In our design of ClusterFinder, we chose not to train separate HMMs for specific gene cluster classes (e.g., polyketides or terpenes), since these narrower HMMs would be less effective at identifying hybrid and novel classes of biosynthetic gene clusters (BGCs).

To filter and analyze predicted gene clusters, we merged ClusterFinder's results with those of a second gene cluster identification algorithm, antiSMASH, which identifies and annotates gene clusters based on a complementary strategy: a hierarchical logic of conserved protein domains that are characteristic of one of ~20 gene cluster classes(Medema et al., 2011). Our probability threshold of 0.4 was chosen to keep the false-positive ratio based on classification of Pfam domains below 5% (an estimate based on a comparison of ClusterFinder results and manual annotations of 10 genomes). The true-positive ratio at this threshold is 55% (at the level of protein domains).

2. Prolific producers harbor exceptionally large complements of gene clusters

We addressed the question of how the genome size of a prokaryote relates to its biosynthetic capacity. Similar to a result from an earlier report (Donadio et al., 2007), we find that prokaryotes have an average of 2.4 gene clusters per Mb (SE = 0.03 and 0.10, simple least squares linear regression and generalized least squares linear regression corrected for phylogeny, respectively) (**Figure 1c**). Strikingly, however, certain strains are clear outliers in that they have more than the average number of gene clusters per Mb (defined as having residuals >8, 5.0% of the total). The scarcity of low-end outliers suggests that nearly all bacterial species harbor at least a minimal complement of biosynthetic gene clusters.

Likewise, we find that while the average species devotes 3.7% ± 3.1% of its genome to BGCs, a largely overlapping group of outlier species devote >7.5% of their genomes to natural product biosynthesis (defined as >1 SD above the mean, 6.7% of the total). This is comparable to the mean fraction of a bacterial genome devoted to transcription (7.2%) and translation (8.5%) (**Figure 1d**). One outlier, *Streptomyces bingchenggensis*, devotes a remarkable 22% of its genome to secondary metabolites; in aggregate, this strain's gene clusters (2.65 Mb) are larger than the entire genome of every sequenced strain of *Streptococcus*. The aggregate gene clusters of a less extreme strain, *Streptomyces griseus* (1.77 Mb), still dwarf most *Helicobacter* genomes.

Many of these outliers are strains of *Streptomyces*, *Myxococcus*, *Sorangium*, and *Burkholderia*. Our results suggest that it is probably no coincidence that these genera have long been known for their prolific production of natural products, since they harbor an exceptionally large complement of gene clusters. Importantly, other outliers are from genera that, to our knowledge, have not yet been mined for natural products: *Gloeobacter*, *Methylobacterium*, *Shewanella*, and *Teredinibacter*. In general, there is a vast discrepancy in phylogenetic distribution between experimentally characterized gene clusters in our training set and our set of predicted gene clusters (**SI Figure 1a**). Further highlighting the opportunity to identify new molecules by studying underexplored taxa, species from the genera *Legionella* and *Coxiella* stand out as intracellular pathogens that have retained multiple BGCs in spite of their reductive genome evolution (**SI Figure 1b**), indicating a strong selective pressure to retain the small molecule products of these gene clusters.

We next mined metadata on BGC-harboring organisms from the NCBI BioProject/BioSample databases (Barrett et al., 2012) to identify correlations between the numbers of gene clusters in a genome and the ecology or lifestyle of a microbe. We find that organisms that display a large degree of multicellularity, occur in terrestrial habitats, form endospores and/or have an aerobic lifestyle have more gene clusters on average than organisms that do not exhibit these features (**SI Table II**). Nonetheless, the biosynthetic potential from species without these features should not be underestimated: even though anaerobes have on average six times fewer gene clusters, these taxa have not been well explored and therefore hold great promise for further study(Letzel et al., 2013).

In a more general sense, we observe that the length of a bacterial genome correlates best with the size of coding regions for transcription-associated genes as well as primary and secondary metabolism, while the size of the coding regions for other functional categories remains constant (**Figure 1d** and **SI Figure 1c**). Thus, it would appear that bacterial genomes expand largely to increase their gene complements for transcription, primary metabolism, and secondary metabolism.

3. The relationship between phylogeny and gene clusters varies tremendously across the bacterial tree of life

The phylogenetic distribution of BGCs is a key factor in understanding their biological roles. If related species harbor similar BGCs, then their small molecule products could underlie phenotypes common to the taxon. Alternatively, if related species harbor different gene clusters, then these elements could play an important role in ecological specialization. Evidence for the latter has come from recent reports showing that genomes of *Mycobacterium* and *Bacillus* are 92-98% similar at the nucleotide level, yet differ markedly in their complement of gene clusters (Rückert et al., 2011; Tobias et al., 2013). However, it is not clear whether this phenomenon is general or specific to these taxa.

To answer this question, we used a quadratic entropy index to illustrate how the diversity of gene clusters can be decomposed among the nodes of the phylogenetic tree (Pavoine et al., 2010). This methodology allowed us to determine gene cluster diversity at internal nodes at different depths in the phylogeny (**SI Figure 2a & 3**). Surprisingly, we find that the degree to which gene clusters are shared within a taxon differs markedly among bacterial taxa. For example, while three strains of *Escherichia coli* and *Bacillus cereus* share 32% (6 out of 19) and 26% (9 out of 35) of their pan-gene-cluster complement, respectively, three strains of *Corynebacterium glutamicum* that span a comparable phylogenetic distance share 70% (9 out of 13) of their gene clusters (**SI Figure 2c**).

Even BGC repertoires of closely related strains from the latter (sub)phyla can display notable differences: for instance, *Bacillus subtilis* ATCC 6633(Zeigler, 2011) shares the bacillibactin, bacillaene, surfactin, subtilosin and bacilysin gene clusters with the common laboratory strain *B. subtilis* 168. However, *B. subtilis* ATCC 6633 harbors a mycosubtilin gene cluster in place of the plipastatin gene cluster found in *B. subtilis* 168 -- two nonribosomal peptide gene clusters of similar size that produce small molecule products in distinct families (**SI Figure 2b**). In addition, *B. subtilis* ATCC 6633 harbors the gene clusters for subtilin and rhizocticin, whereas *B. subtilis* 168 encodes the enzymatic routes to synthesize the cannibalistic SDP and SKF peptides(Liu et al., 2010).

In general, we find that the diversity of BGCs does not appear to be strongly skewed towards the root or the leaves of the phylogenetic tree (**SI Figure 1d**), indicating an ongoing process of gene cluster diversification. We observe many nodes of high diversity in the tree closer to the leaves, pointing to evolution independent of phylogeny, perhaps indicative of ecologically driven diversification.

4. Saccharides are the largest class of gene clusters

We began our analysis by grouping the 9,421 high-confidence gene clusters into classes based on the presence of characteristic protein domains and asking how many of each class we recovered

(**Figure 1b**). The prevalence of certain biosynthetic classes in the entire dataset could be compared with their prevalence in experimentally characterized gene clusters using our training set. This set of 732 experimentally characterized gene clusters is nearly exhaustive and was compiled in an unbiased manner, so it is a reasonable proxy for measuring how well each gene cluster class has been studied.

The predominance of saccharide gene clusters illuminates families of molecules that are not typically thought of as natural products. For example, based on Pfam domain content, 23% of the saccharide gene clusters are predicted to encode lipopolysaccharides and 3% capsular polysaccharides. These cell wall-mounted molecules play important roles in host-microbe and microbe-microbe interactions, and small changes in their structure can lead to large changes in their function (Rehm, 2010). Other saccharides have antibacterial activity. A recently discovered saccharide BGC (with an average ClusterFinder probability of 0.93) has been found to encode saccharomicin, a member of a novel family of heptadecaglycoside antibiotics with potent activity against Gram-positive pathogens (Strobel et al., 2012).

Besides saccharide gene clusters, several other gene cluster types are notable as well. Gene clusters encoding ribosomally synthesized and posttranslationally modified peptide natural products (recently termed RiPPs (Arnison et al., 2013)) are found in much larger numbers than polyketides and terpenes. RiPPs are difficult to detect because of their immense architectural diversity (Arnison et al., 2013); as a result, they are the most likely class to be underestimated by our approach. Consequently, gene clusters for RiPPs may be among the most widely distributed categories in bacterial genomes. Finally, we also detected and manually curated around 1,500 gene clusters (subdivided into low and middle confidence categories, see Methods) that have all the hallmarks of being BGCs, but do not clearly fall into any known class of BGCs. These provide a promising set of candidate BGCs that may lead to the discovery of novel chemical scaffolds, for which there is great need in current drug development approaches (Fischbach and Walsh, 2009).

5. A global map of biosynthesis based on a gene cluster distance metric

In order to draw a global network that shows the mutual evolutionary relationships between all the BGCs in our dataset, we used the distance metric of Lin et al. (2006). The distance metric has two components: the first is based on the Jaccard coefficient and measures the similarity between the gene families included in each gene cluster, and the second represents the copy number variation of gene families between the two clusters. We validated that the distance metric works in this setting by using it to measure the distances among every pair of gene clusters in our training set; we confirmed that the gene clusters for a natural product family (e.g., glycopeptides and lipopeptides) are collectively more similar to each other according to the metric than to other related clusters (e.g., other nonribosomal peptides) (**SI Figure 4a**). In addition, we created a high-resolution variant of the distance metric in which Pfam domain sequence similarity was also taken into account (see **Methods**). Since this version of the algorithm is more computationally intensive, we only applied it to the network of known BGCs.

We constructed and manually inspected the networks that result from making our threshold more or less stringent. The network structure shown in **Figure 2** is robust to small variations in the clustering threshold (+/- 0.1). Larger variations yielded networks that were almost fully connected or highly dissociated, neither of which provide biological insight into the large-scale relationships among gene cluster classes. While the network in **Figure 2** may appear densely connected, it contains just 0.6% of all possible edges (388,411 out of 63,286,875).

In the network displayed in **Figure 2**, oligosaccharides, nonribosomal peptides and polyketides/lipids feature prominently. The network reveals two key findings. First, one connected component harbors most of the gene clusters (72%), and is largely composed of two linked subgraphs: one dominated by oligosaccharide BGCs and the other a mixture of nonribosomal peptide (NRP) BGCs

and polyketide/lipid BGCs, indicating that BGC from these classes share a significant number of gene families with one another. Second, there are many prominent subgraphs in which no gene clusters have been characterized; some of these BGCs may encode entirely novel chemical scaffolds. From these unexplored subgraphs, many of which include 'low-confidence' BGCs, three common themes emerge, each pointing to a putative large class of chemically novel secondary metabolites: (*i*) There are dozens of gene cluster families ranging from 3-20 kb that harbor a 3-ketoacyl-ACP synthase (KAS) III enzyme and a diverse and varying set of auxiliary tailoring enzymes including desaturases, adenylation domains, and aminotransferases. These occur in well-studied organisms such as *Burkholderia* as well as unexplored genera such as *Anaeromyxobacter* and *Ochrobactrum*. (*ii*) There is an abundance of uncharacterized terpenoid, lipid, and glycolipid gene clusters in poorly studied genera such as *Zymomonas*, *Acetobacter*, *Nitrobacter*, and the archaeon *Sulfolobus*, which are unlike any known BGCs from these classes. (*iii*) There is a diverse set of gene clusters that are rich in redox enzymes without containing bond-forming enzymes for known compound classes, exemplified by a gene cluster consisting of four flavin-dependent halogenases and a TonB-dependent receptor from *Caulobacter.*

Unexpectedly, most gene clusters (84%) belong entirely to a single class. Hybrids therefore comprise a much larger proportion of known gene clusters than predicted gene clusters, suggesting that they may have been oversampled by experimental efforts to date. This may be partially explained by the fact that hybrids occur much more frequently (~50%) within the Actinobacteria, from which many known gene clusters originate. The distribution of known gene clusters in the network (black dots) is non-uniform, suggesting that efforts to experimentally characterize gene clusters have been biased toward specific BGC classes.

Since the gene clusters for ribosomally synthesized and post-translationally modified peptides (RiPPs) do not share core domains (Arnison et al., 2013), their biosynthetic loci do not cluster in the network; rather, they constitute distinct clusters for different RiPP subclasses (e.g., lantipeptides, thiopeptides). This corroborates their mode of evolution: RiPP BGCs tend to be smaller and more diverse, and commonly incorporate tailoring genes from the other gene cluster classes.

Interestingly, the topology of the network offers important insights into BGC evolution. The BGC similarity graph is a small-world, scale-free network (Barabasi and Oltvai, 2004): the exponent of the degree distribution, the average shortest path, and the average clustering coefficient are 1.66 ± 0.07, 1.11, and 0.69, respectively (**SI Figure 4b-e**). In small-world networks, the path between two nodes selected at random is unusually short on average; this means that for most pairs of unrelated BGCs, there will be a third gene cluster that shares a substantial number of genes with each of them. The unusually gradual descent of a node degree distribution indicates that if a new node is added to the graph, an unusually large number of edges is likely to be added (Seyed-Allaei et al., 2006). Both of these characteristics are consistent with the view that the total set of BGCs is composed of a finite set of parts used in many different arrangements and contexts. Interestingly, highly linked nodes are unusually abundant (429 hubs with more than 200 links). Some of these nodes are small BGCs that are similar to common fragments of larger BGCs (here, a 'sub-cluster'), suggesting that such larger BGCs often evolve through the merger of smaller BGC modules.

6. Phylogenetic analysis of APE ketosynthase and adenylation enzymes

Structure-guided multiple sequence alignment and maximum likelihood phylogenetic reconstruction (**SI Figure 6b-c**) shed light on the evolutionary relationships between key enzymes from the APE gene clusters and other known enzymes from the same enzyme superfamilies. Although distantly related, the closest homologs of the two major clades of APE KS domains are FabF protein (Garwin et al., 1980) and enzymes putatively involved in ladderane biosynthesis (Rattray et al., 2009). Indeed, when we compared ladderane and APE BGCs at the whole-cluster level, there appeared to be a large overlap in gene content

between APE BGCs, the *Kuenenia stuttgartiensis* ladderane lipid BGC (Rattray et al., 2009; Strous et al., 2006) and a related polyunsaturated hydrocarbon BGC from *Desulfotalea psychrophila* (**SI Figure 6d**). This finding suggests that the APE superfamily may have evolved to include clusters that produce a wider range of chemically distinct metabolic products in different organisms.

The APE A domains comprise two separate, unrelated clades, suggesting convergent evolution of the enzyme that selects and activates the starter unit. The closest known homologs of these clades are a phenylacetate CoA-ligase and a 4-chlorobenzoyl-CoA ligase, respectively (Law and Boulanger, 2011; Reger et al., 2008). These results suggest that HMM-based approaches operating on Pfam domain frequencies, such as ClusterFinder's algorithm, can more sensitively predict noncanonical clusters than homology-based algorithms. They also support the notion that gene clusters harboring uncharacterized clades of well-known biosynthetic domains are a promising category to mine.

SUPPLEMENTAL FIGURE LEGENDS

SI Figure 1. **Global phylogenomic analysis of prokaryotic BGCs, Related to Figure 1. a,** The prokaryotic tree of life is mostly unexplored for BGCs. The phylogenetic tree of bacterial and archaeal classes (as stored in NCBI Taxonomy) shows the distribution of known (left) and predicted BGCs (right). A strong historical bias can be observed: some bacterial classes (such as Actinobacteria) have been heavily studied, whereas other classes with (on average) similarly large numbers of BGCs have been largely neglected. The two graphs are not scaled equally; the left bar plot shows the total number of known BGCs per class, whereas the bar plot on the right displays the average number of predicted BGCs per strain within a class. **b**, Examples of notable PKS and NRPS biosynthetic gene clusters detected in the genomes of the obligate intracellular pathogens *Legionella* and *Coxiella*. Letters above the PKS and NRPS genes signify domain structure, with adenylation domain substrates as predicted by NRPSPredictor2 (Röttig et al., 2011) in brackets. **c**, Cross-correlation matrix of COG protein functions in bacterial genomes. Although we focused on analyzing the association between the number of BGCs (or percentage of the genomes they occupy) and genome lengths (**Figure 1c**), we also investigated whether there are any other COG functions that correlate with genome length. Primary and secondary metabolism, as well as transcription regulation, are linked to genome length, suggesting that genomes become longer by incorporation of biosynthetic and regulatory genes. In contrast, COG functions such as translation, cell cycle regulation, RNA replication and repair, nucleotide metabolism and transport, post-translational modification, protein turnover, and chaperone functions do not seem to be linked to genome length. **d**, Histogram of cumulative quantitative entropy (QE) index with respect to the distance from the root of the phylogenetic tree. A decreasing trend in this histogram suggests decreasing diversification rates on a global evolutionary time-scale. However, a presence of nodes of high diversity closer to the leaves points to recent evolution of BGC repertoires. Each bar plots a sum of QE indices of all nodes within a given bar's limits with respect to the root of the phylogenetic tree. **e**, Examples of previously unknown saccharide gene clusters. The saccharide gene clusters are from unexplored or underexplored genera. Colors represent functions of the genes, as indicated in the figure legend. **f**, Type diversity of BGCs within the same taxonomic genera. The bar graph shows the percentage of gene clusters per class that is shared between two genomes randomly sampled from the same genus. While fatty acid biosynthesis gene clusters are often similar in species of the same genus, RiPP and saccharide BGC repertoires are often radically different between species within the same genus. **g**, Rarefaction analysis of numbers of BGC families (red) and Pfam families (green). BGC families (or "BGC clusters") were calculated from the BGC similarity network with a similarity threshold of 0.5 and MCL clustering with I = 2.0. For a given number of genomes, a random sample of organisms was selected 20 times (the thickness of the lines denote 68% confidence intervals based on these 20 bootstraps). **h**, Identification and classification of BGCs in 201 single-cell genomes from uncultivated organisms. Functional classification of the 947 BGCs identified in the set of 201 single-cell genomes from JGI (Rinke et al., 2013), using the same antiSMASH-based classification scheme used for the dataset of full genomes from JGI. Besides a significant number of saccharide-encoding gene clusters, the vast majority of putative BGCs falls outside known biosynthetic classes.

SI Figure 2. Diversity of BGCs is independent from the phylogeny, Related to Figure 1. a, The decomposition of BGC diversity among species of the phylum Actinobacteria. The diversity of each node in the phylogenetic tree is measured by the quadratic entropy index, and represented by the size of the circle (larger circle defines higher degree of diversity). Color bars at the leaf tips represent number of BGCs per species, with different colors denoting different BGC types (colors as in **Figure 1b**). Hybrid gene clusters (orange) are unusually prominent in Actinobacteria (~50%). For the entire phylogenetic tree, see **SI Figure 3**. **b**, The scatter plot shows no correlation between phylogenetic and BGC content distance for a given organism pair. **c,** The Venn diagrams show the number of BGCs shared among three different sets of closely related species. The phylogenetic tree sections to the right of the Venn diagrams are shown using the same scale.

SI Figure 3. Decomposition of BGC diversity among all sequenced prokaryotic genomes, Related to Figure 1. The diversity of each node in the phylogenetic tree is measured by the quadratic entropy index, and represented by the size of the circle (larger circle defines higher degree of diversity). Color bars at the leaf tips represent the number of BGCs per species, with different colors denoting different BGC types (colors as in **Figure 1b**). The outer ring shows the absence/presence of APE gene clusters in our initial set of 1154 genomes obtained from JGI-IMG. The discontinuous pattern of APE gene cluster conservation suggests frequent horizontal gene transfer and/or gene cluster loss. Pink indicates the presence of one APE gene cluster in a genome, red indicates the presence of two gene clusters in a genome. Several genomes from *Burkholderia* and *Ralstonia* have two different APE gene clusters located on two different chromosomes. The tree was generated using iTOL (Letunic and Bork, 2007).

SI Figure 4. BGC similarity networks, Related to Figure 2. a, Similarity network of known BGCs. The similarities between the BGCs were calculated by taking into account the architecture as well as the sequence similarity features of our distance metric (see Methods for details). This analysis shows that the gene cluster distance metric functions well in separating known families of BGCs, while maintaining links representing known genetic similarities between classes like aminoglycosides and saccharides. Cytoscape(Smoot et al., 2011) was used to visualize the network. **b**, Analysis of the global BGC similarity network. Network (or

graph) topology can be indicative of the relationships among its constituent nodes (here, BGCs). Tables **b** and **c** show different topology parameters for graphs with BGC similarity cutoffs of 0.6 and 0.8, respectively; #nodes indicates the number of nodes in the graph; #edges indicates the number of edges in the graph; gamma equals the exponent of the node degree frequency diagram (the steepness of the linear fit in **d**); L is the average shortest path between any two nodes; C is the average clustering coefficient, Lrand is the average shortest path between any two nodes in the randomized graphs; Crand is the average clustering coefficient in the randomized graphs; and K(k) is coefficient of the linear fit in **e**. The values of the parameters were calculated for all nodes in the graph, as well as for subgraphs of nodes corresponding to individual classes of BGCs. Parameters were calculated using the NetworkX library.

SI Figure 5. Full annotated APE superfamily clustered heat map including COG annotations, Related to Figure 3. Full version of the clustered heat map shown in Figure 3a. In this version, the COG annotations are shown at the bottom, and the accession number and source strain are shown on the right.

SI Figure 6. Phylogenetic analysis of APE gene clusters and key biosynthetic enzymes, Related to Figure 4. a, Pairwise sequence identities of the ketosynthase and adenylation domains in the four characterized gene clusters. The numbers in the graph represent the average percentage identity between the amino acid sequences of the pairs of most closely related adenylation / ketosynthase enzymes in the four gene clusters, as inferred from the structure-guided sequence sequence alignment. Three pairs of adenylation enzymes whose amino acid sequences are only 12% identical are shown as <20% identical, to account for the inexactness of sequence identity calculations for such distant relationships. **b**, Phylogenetic tree of APE ketosynthase domains with other ketosynthases. The maximum likelihood phylogenetic tree, based on a structure-guided multiple sequence alignment using PROMALS3D (Pei et al., 2008), shows that the ketosynthases from representative APE gene clusters belong to two evolutionary clades. One clade is most closely related to FabF proteins from *Escherichia coli* and *Bacillus subtilis*, while the other clade is most closely related to ketosynthases putatively involved in ladderane lipid biosynthesis in the anammox bacterium *Kuenenia stuttgartiensis*. The gene clusters from *Bacteroides* and *Flavobacterium* contain a duplicate of the ketosynthase from the latter clade, while the xanthomonadin gene cluster from *Xanthomonas campestris* contains no ketosynthase from the first clade. **c**, Phylogenetic tree of APE adenylation domains with other adenylation enzymes. The maximum likelihood phylogenetic tree, based on a structure-guided multiple sequence alignment using PROMALS3D (Pei et al., 2008), shows that the adenylation enzymes involved in APE biosynthesis cluster in two uncharacterized clades within the ANL superfamily that includes Acyl-CoA synthetases, NRPS adenylation domains, and Luciferase enzymes. Most closely related are two adenylation enzymes that are involved in the ligation of two different aryl group-containing compounds, suggesting that convergent evolution may have lead to the independent evolution of two mechanisms to attach an aryl group to the polyene that is synthesized by the same clades of ketosynthases. **d**, Comparison of APE gene clusters with related BGCs. Alignment of the two APE superfamily gene clusters from *Escherichia coli* CFT073 and *Flavobacterium johnsonii* ATCC 17061, the putative ladderane lipid biosynthesis gene cluster from *Kuenenia stuttgartiensis* and the polyunsaturated hydrocarbon biosynthesis gene cluster from *Desulfotalea psychrophila* LSv54. Colors signify homologous genes based on a MultiGeneBlast comparison with the blastp algorithm.

SI Figure 7. Evaluation of the ClusterFinder algorithm and APE structural characterization, Related to Figure 3. A, The performance of the ClusterFinder algorithm was evaluated by calculating the ROC and AUC using 10 manually annotated genomes (SI Table VII) that were not used in the training of the algorithm. We obtained an AUC of 0.84, which is significantly better than the AUC of a random prediction (AUC of 0.5). The predictions were assessed on protein domain basis; for example, at each probability threshold, a given protein domain was assigned to the true-positive class if the probability of being in a BGC was higher than the threshold, and if it was manually annotated as being part of a BGC. **B,** We assessed the true-positive rate on a set of 74 BGCs from the literature (SI Table VIII). Only 7 BGCs (9.5%) did not pass our probability threshold of 0.4. **C**, Structure of APE_{Ec} with COSY (dashed lines) and HMBC (solid lines) correlations. **D**, Structure of APE_{Vf} with COSY (dashed lines) and HMBC (solid lines) correlations. **E**, HPLC traces for crude APE extracts. **a)** Overlay of traces for *V. fischeri* ES114 wild type (blue) and the *V. fischeri* ES114 Δ*ape* deletion strain (red). **b)** Overlay of traces for *E. coli* Top10 expressing the CFT073 cluster (blue) and the *E. coli* Top10 control strain containing the empty vector (red). HPLC conditions: gradient of acetonitrile in 0.02% formic acid water: 0% to 30% organic phase in 2 min, 30% to 90% organic phase from 2 min to 22 min, followed by a hold at 90% for 3 minutes and a 3 min wash at 100% organic phase. Detection was at λ = 441 nm. The peak purified and subjected to structural analysis is denoted with an asterisk. **F**, Second RP-HPLC purification for APE_{Ec} **(a)** and APE_{VF} **(b)**. **G**, UV spectrum for APE_{Ec} **(a)** and APE_{VF} **(b)**.

SI Data File 1. Overview of all 1,021 identified APE superfamily gene clusters (separate PDF file, 41 pages). Graphical overview of all 1,021 BGCs from the APE superfamily. Colors indicate annotation of gene function based on calculated COGs (see Methods).

SI Data File 2. NMR spectra for APE_{Vf} and APE_{Ec}.

I. *Vibrio fischeri* in D_6 DMSO
- *i.* ^1H NMR
- *ii.* Expanded ^1H NMR
- *iii.* COSY
- *iv.* Expanded COSY
- *v.* HSQC
- *vi.* HMBC
- *vii.* ROESY
- *viii.* Proton in D_6 Acetone
- *ix.* Expanded Proton in D_6 Acetone

II. *Escherichia coli* in D_6 Acetone
- *i.* ^1H NMR
- *ii.* Expanded ^1H NMR
- *iii.* COSY
- *iv.* Expanded COSY
- *v.* HSQC
- *vi.* HMBC
- *vii.* ROESY
- *viii.* TOCSY

SUPPLEMENTAL TABLE LEGENDS

Tables are provided in a separate SI_Tables.XLSX file, due to size limitations.

SI Table a. Training set composed of 732 experimentally identified BGCs. Columns contain further detailed information: the compound encoded by the BGC, GenBank accession number, description, compound type classification, PubMed IDs of relevant literature, PubChem IDs of the encoded compound, and SMILES string of chemical structure of the encoded compound.

SI Table b. Overview of the four environmental metadata features that show the most significant differences between genomes, depending on how many BGCs are encoded in these genomes. P-values are calculated with the Kruskall-Wallis test.

SI Table c. Set of 870 identified carotenoid gene clusters. The BGCs were identified using an MultiGeneBlast architecture search with CrtI, CrtE, and CrtB from *Rhodobacter capsulatus* SB 1003 as queries (same settings as used for the APE MultiGeneBlast search). Hits with at least two of the three genes present were classified as putative carotenoid gene clusters.

SI Table d. Benchmark of the ClusterFinder method on the *Pseudomonas fluorescens* Pf-5, *Streptomyces griseus* IFO13350 and *Salinispora tropica* CNB-440 genomes, compared to antiSMASH (Medema et al., 2011) and the manual genome annotations by Paulsen et al. (Paulsen et al., 2005) and Nett et al. (2009).

SI Table e. List of Pfam domains characteristic for saccharide gene clusters that were used for classification of this BGC type. Both Pfam accession numbers and descriptions are given. Data obtained from http://pfam.sanger.ac.uk.

SI Table f. Primers used in this study.

SI Table g . Plasmids used in this study.

SI Table h . Strains used in this study.

SI Table i. List of 100 randomly selected genomes. The table lists a hundred randomly selected genomes, whose protein domain information was used to train the emission frequencies of the hidden Markov model in ClusterFinder algorithm.

SI Table j. Overview of BGC class-specific domains used to classify BGCs. The first column contains PFAM accession numbers or 'ND' codes (domains from antiSMASH (Medema et al., 2011)). The second column gives the annotation of the domain. The third and final column displays the biosynthetic type associated with the domain or the class of associated tailoring reactions.

SI Table k. Predicted high-confidence BGCs from all genomes.

SI Table l. A list of BGCs from 10 manually annotated genomes, used to evaluate the performance of ClusterFinder algorithm.

SI Table m. A list of 74 BGCs from the literature, used to evaluate the performance of ClusterFinder algorithm.

SI Table n. ^{1}H and ^{13}C NMR data.

EXTENDED EXPERIMENTAL PROCEDURES

Genome information

For all available full genome sequences, gene and Pfam domain annotations were obtained from the JGI-IMG database (Markowitz et al., 2012), version 3.2 (08/17/2010). In the JGI-IMG database, coding regions in prokaryotic genomes are predicted with Glimmer (Salzberg et al., 1998), while domains are annotated with HMMER3 (Eddy, 2008, 2010; Krogh et al., 1994) using Pfam-A HMM profiles (Punta et al., 2012).

Training set generation

We first searched for all biosynthetic gene clusters in the NBCI Nucleotide database, (http://www.ncbi.nlm.nih.gov/nuccore/) using the search terms "biosynthetic gene cluster", "secondary metabolite", "natural product synthesis", and "biosynthesis". The results set was then manually curated and supplemented by gene clusters identified through a manual search through the scientific literature between 1990 and 2011. These also included known gene clusters from whole genome sequences. Next, by comparing the gene cluster entries with the descriptions of the gene clusters in the scientific literature, we manually checked that the biosynthetic gene clusters were full-length, and not deposited to the NCBI Nucleotide database as partial sequences or sequences with large flanking regions not belonging to the biosynthetic gene clusters. This procedure resulted in a set of 732 biosynthetic gene clusters (**SI Table I**). Finally, we filtered out 55 redundant gene clusters 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 (the similarities were calculated using a distance metric that adopts sequence similarity of Pfam domains in addition to Pfam domain architecture as described below in "Biosynthetic gene cluster prediction method: ClusterFinder").

Biosynthetic gene cluster prediction method: ClusterFinder

A two-state Hidden Markov Model (HMM) was designed, with one hidden state corresponding to biosynthetic gene clusters (BGC state) and a second hidden state to the rest of the genome (non-BCG state). A vector of observations fed to the HMM is a sequence of Pfam domains in the order in which they appear in the annotated genome. For each of the Pfam domains from the observation vector, the probability of being part of a biosynthetic gene cluster is computed as a posterior probability of the BCG hidden state using the Backward-Forward algorithm (Press et al., 1992). Emission probabilities of Pfam domain types for the BCG state of the HMM were trained by computing Pfam domain frequencies in our set of 677 known biosynthetic gene clusters, using balance training as follows: first, we binned BGCs into 6 classes (NRPS, PKS, terpene, oligosaccharide, ribosomal peptide, and other), based on antiSMASH predictions of biosynthetic classes. Frequencies of all Pfam domains observed in the training set were then calculated for each class separately, and then joined as an average frequency across all 6 classes. At the end, all frequencies were normalized to add up to 1.

To obtain Pfam domain frequencies for the non-BGC state, we first randomly selected one hundred genomes (**SI Table IX**), and aligned all their Pfam domain sequences to all Pfam domain sequences from the BGC training set using the blastp algorithm (Camacho et al., 2009). Only hits with an E-value larger than 1e-10 were used to calculate emission probabilities for the non-BGC state. Frequencies of Pfam domains that appear in BGC state but not in the non-BGC state (or *vice versa*) were set to 1% of the frequency of a single observation. The transition probabilities were inferred from manual annotation of biosynthetic gene clusters in the *Streptomyces avermitilis* genome. Around 30% of genes cannot be assigned to any current Pfam family. Consequently, emission probabilities of such cases were set to 1.0 for both states.

After obtaining the biosynthetic gene cluster probabilities for all domains from an input string of Pfam domains, ClusterFinder identifies gene clusters as sets of genes that are at most one gene apart and contain at least one domain with probability of more than 0.2. Finally, ClusterFinder filters out any biosynthetic gene clusters that do not meet any one of the following criteria: (i) having an average BGC probability of >0.4 (as chosen from the second evaluation set), (ii) being longer than the average length of two bacterial genes (2000 bp), and (iii) containing at least one of the class-specific domains (**SI Table X**). A summary of the ClusterFinder output on all analyzed genomes is given in **SI Table VIII**. ClusterFinder was implemented in Python, and is integrated in antiSMASH (Medema et al., 2011) as well as in the JGI-IMG platform (Markowitz et al., 2012). ClusterFinder source code is available from the GitHub repository (https://github.com/petercim/ClusterFinder).

ClusterFinder validation

The performance of the biosynthetic gene cluster prediction approach was tested in two ways. First, using ten manually annotated bacterial genomes, we plotted an ROC curve based on classification of Pfam domains (**SI Figure 7a-b and SI Table XII**), for which we determined an AUC of 0.84. Second, we searched for recently experimentally characterized biosynthetic gene clusters in the literature, and used them to assess the true-positives rate. We found a total of 74 biosynthetic gene clusters not used in our training sets (**SI Table XIII**). 91% of these gene clusters were predicted as biosynthetic gene clusters with a median probability (median across all Pfam domains of a given gene cluster) of >0.4. Two out of the six biosynthetic gene clusters with a median ClusterFinder probability lower than 0.4 were found to contain flanking regions not belonging to the actual gene cluster, while the actual gene cluster was detected in the center. Thus, we could conclude that 70 out of the 74 (95%) of the gene clusters had been detected successfully. The remaining four gene clusters from the test set encode two small terpenoid biosynthesis gene clusters, a putative phenolic lipid biosynthesis gene cluster and another putative BGC that did not contain enough Pfam domain similarity with our training set.

When we compared ClusterFinder with antiSMASH (Medema et al., 2011), antiSMASH proved to be more conservative than ClusterFinder. In spite of the increased power of ClusterFinder to find unknown gene cluster types, the algorithm has a low rate of clear false positives (4.6%). Another observation from the comparison of the two algorithms was that ClusterFinder algorithm is more accurate at predicting BGC borders (with 14.4±13.3 and 23.1±12.1 incorrectly predicted border genes per BGC for ClusterFinder and antiSMASH, respectively), which aids in calculating a BGC similarity network, since incorrectly predicted flanking regions would result in noisier BGC similarity values.

Annotation of biosynthetic gene clusters

Lipopolysaccharide gene clusters were specifically identified by detection of at least one of the following domains: PF01755 (Glycosyltransferase family 25, LPS biosynthesis protein), PF02706 (Chain length determinant protein), PF06176 (Lipopolysaccharide core biosynthesis protein WaaY), PF06293 (Lipopolysaccharide kinase Kdo/WaaP family), PF04390 (Lipopolysaccharide-assembly), PF06835 (Lipopolysaccharide-assembly, LptC-related), PF07507 (WavE lipopolysaccharide synthesis), PF10601 (LITAF-like zinc ribbon domain) and PF04932 (O-Antigen ligase). Capsular polysaccharide gene clusters were specifically identified by detection of at least one of the following domains: PF05704 (Capsular polysaccharide synthesis protein), PF10364 (Putative capsular polysaccharide synthesis protein), PF05159 (Capsule polysaccharide biosynthesis protein), PF09587 (Bacterial capsule synthesis protein PGA_cap). The percentage of saccharide gene clusters not closely related to known saccharide gene clusters was determined by counting the number of BGC in clusters in the BGC network (MCL clustering on >0.5 Lin distance network and I parameter set to 4.0) that do not contain any known gene clusters (see "Gene cluster distance metric and evolutionary network of BGCs" below).

Finally, gene clusters that could not be classified using the expanded antiSMASH-based annotation scheme were clustered into BGC families using MCL (BGC similarity network with a similarity threshold of 0.5 and MCL clustering with I = 2.0). These BGC families were then manually divided into low and high confidence BGC families based on the presence of biosynthetic characteristics in the blastp/HMMer search results against the Pfam, nr and SwissProt databases.

Phylogenetic distribution of BGCs

The phylogenetic distribution of BGCs across the microbial tree of life was plotted using iTOL 2 (Letunic and Bork, 2011). The phylogenetic tree used was based on 16S rRNA marker sequences from the corresponding genomes, and was obtained from JGI-IMG (Markowitz et al., 2012). Estimates of within-taxon variation across the tree were calculated using the quadratic entropy index, which allowed us to determine gene cluster diversity at different parts and depths in the phylogeny (Pavoine et al., 2010). Taxonomic classifications of organisms in genera, families, orders, classes and phyla were taken from NCBI Taxonomy (Federhen, 2012).

Gene cluster distance metric and evolutionary network of BGCs

To estimate the evolutionary distance between gene clusters, we used a distance metric from Lin et al. (2006) that is a linear combination of two different indices: the Jaccard index 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). Additionally, sequence similarity information was incorporated in the distance metric, by replacing the term (a) in the exponent of the domain duplication index with term (b).

$$\text{(a)}\ \frac{\left|N_i^P - N_i^Q\right|}{S} \qquad \text{(b)}\ \frac{\left|N_i^P - N_i^Q\right| - Munkres\left(D\left(N_i^P, N_i^Q\right)\right)}{S}$$

Here, *Munkres* represents the Munkres (also known as Hungarian) algorithm (Munkres, 1957) for finding of the maximum bipartite matching in a bipartite graph of distances between domains *D* of the type *i* from the two sets to be compared. Due to the large number of domain sequences, the domain distance was defined as the degree of sequence identity. The sequence identities between domains were inferred from multiple sequence alignments constructed using MUSCLE (Edgar, 2004) for all the sequences of each Pfam domain. Default parameters were used (i.e., at most 5 iterations), except for domain types with more than 8,000 sequences, for which the number of iterations was set to 3. The distance between all domain pairs of the same type was defined as 1 – sequence identity. The final network was obtained by using a cluster-cluster distance cut-off of 0.5. Visualization was performed using Cytoscape (Smoot et al., 2011).

Delineation of the APE gene cluster superfamily

To expand the APE gene cluster family from the gene clusters present in our JGI dataset to all gene clusters present in the entire GenBank database (including unfragmented gene clusters from draft genomes), we used MultiGeneBlast (Medema et al., 2013) in architecture search mode with a combination of all the amino acid sequences encoded by the genes from the *E. coli*, *V. fischeri*, *X. campestris* and *F. johnsonii* gene clusters as query. To remove redundancy from this set of query sequences, we used CD-HIT (Li and Godzik, 2006) with a cut-off of 45% sequence identity. The

MultiGeneBlast architecture search was run with default settings, except that 20% was used as a minimal sequence identity cut-off for BLAST hits and 2000 BLAST hits were mapped per gene. The output was manually studied to generate a list of gene clusters that had all the characteristic hallmarks for APE gene clusters (>5 key genes shared with the known APE gene clusters, in which at least the key ketosynthase and adenylation enzyme should be present) and were also complete (no fragmentation due to being part of incomplete genome assemblies). Gene cluster borders of all the 1021 resulting gene clusters were estimated manually based on putative operon structures and predicted protein functions. Clusters of Orthologous Groups (COGs) were obtained using OrthoMCL (Li et al., 2003) (MCL I = 1.5, sequence identity cut-off 20%) and used for clustering of the 1021 gene clusters with the Lin distance metric (Lin et al., 2006).

Phylogenetic analysis of ketosynthase and adenylation domains

Structure-guided multiple sequence alignments involving available protein structures from the PDB database (Rose et al., 2013) were performed using PROMALS3D (Pei et al., 2008) using default settings. The phylogenetic trees were inferred with MEGA5 (Tamura et al., 2011) by using the Maximum Likelihood method based on the JTT matrix-based model. The trees with the highest log likelihood (-16196.4949 for A domains and -19511.3462 for KS domains) were used for the supplementary figures. Initial tree(s) for the heuristic search were obtained by applying the Neighbor-Joining method to a matrix of pairwise distances estimated using a JTT model. The trees are drawn to scale, with branch lengths measured in the number of substitutions per site. All positions containing gaps and missing data were eliminated. There were a total of 186 positions in the final A domain dataset and 263 in the final KS domain dataset.

General conditions for bacterial growth and DNA manipulations

Escherichia coli strains were grown at 37°C in LB medium (Sambrook and Russell, 2001), *Vibrio fischeri* at 30°C in LBS (Graf et al., 1994). Antibiotics were added at the following concentrations: for *E. coli* ampicillin (Ap; 100 μg/ml), kanamycin (Km; 50 μg/ml) and for *V. fischeri*: Km (25 μg/ml). Chemicals were purchased from Sigma-Aldrich. PCR reactions were performed with Phusion High-Fidelity DNA polymerase (New England Biolabs) according to the manufacturer's recommendations. Oligonucleotide primers (**SI Table VI**) were obtained from Elim Biopharmaceuticals. *E. coli* genomic DNA isolation, polymerase chain reaction, plasmid transformation and other general cloning methods were performed according to standard procedures (Sambrook and Russell, 2001). Plasmids and bacterial strains used in this study are summarized in **SI Tables VII and VIII**.

Construction of *V. fischeri* ES114 APE-cluster deletion mutant

The counterselectable suicide plasmid pSW8197 was used to make a construct for generating a stable and markerless deletion of the *V. fischeri* E114 APE-cluster (VF0841-VF0860). Primers JC_ES114_clusUP_FWD/_REV and JC_ES114_clusDOWN_FWD/_REV were used to amplify flanking regions of ~1000 bp up- and downstream of the cluster. The resulting flanking DNA sequences were assembled together into the JC_pSW8197_FWD/_REV amplified pSW8197 vector backbone via circular polymerase extension cloning (CPEC) (Quan and Tian, 2011). The resulting construct (pJC120) was introduced into π3813 competent cells via electroporation (yielding JC085) and it was verified by PCR and sequencing.

The generation of a *V. fischeri* deletion mutant was performed as described previously (Le Roux et al., 2007). Briefly, the deletion construct was introduced into *V. fischeri* ES114 by tri-parental mating using JC085 and helper strain DH5α λ*pir* pEVS104. Integrants were selected on LBS agar plates containing Km and subsequently grown non-selectively to allow for the second homologous

recombination to occur. A screening on plates containing 0.2% arabinose selected for colonies that had lost the integrated plasmid backbone (through induction of the *ccdB* toxin gene), resulting either in a successful deletion or a reversion to the wild-type state. The *V. fischeri* Δ*ape* deletion mutant (JC086) was identified by colony PCR, using primer pair JC_ES114_control_FWD/_REV, which only anneals on genomic DNA outside the flanking regions used in the deletion construct. The resulting PCR product was confirmed by sequencing.

Construction of the heterologous expression constructs for the *E. coli* CFT073 and *V. fischeri* ES114 APE-clusters

The *E. coli* CFT073 APE-cluster (c1186-c1204) was amplified in three parts from genomic DNA, using primer pairs JC_CFT073_pt1_FWD/_REV through JC_CFT073_pt3_FWD/_REV and assembled via CPEC(Quan and Tian, 2011) into SuperCos I amplified by JC_Super_CFT073_FWD/_REV. This yielded the 21.2 kb CFT073 APE-cluster heterologous expression construct pJC121. Oligonucleotide primers for this expression construct were designed with the aid of DeviceEditor(Chen et al., 2012). pJC121 was introduced into chemically competent *E. coli* Top10 cells with selection for Km (50 μg/ml), yielding the heterologous expression strain JC087.

The construct for expression of the *V. fischeri* APE was generated in a similar fashion, amplifying the region VF0841-VF0860 in three parts from genomic DNA, using primer pairs JC_ES114_pt1_FWD/_REV through JC_ES114_pt3_FWD/_REV. The cluster was assembled via the Gibson method(Gibson et al., 2009) into SuperCos I (amplified by JC_Super_ES114_FWD/_REV). The resulting 24.6 kb construct, pJC122, was introduced into chemically competent *E. coli* Top10 cells, yielding strain JC088. Since the cluster was not stably expressed in this form, we tested the introduction of a selection of promoters upstream of the operon starting in VF0844. Among the five different promoters that were introduced (p*cat*, T7, pBAD, *ermE** and the *E. coli* K-12 S20 ribosomal subunit *rpsT* promoter), we found that *ermE**p worked best for expression of the *V. fischeri* APE cluster without the need for induction (e.g., in the case for the pBAD or T7 promoters). Introduction of the *ermE** promoter was achieved by first generating a targeting cassette that consists of an Apra^R-cassette (amplified from pIJ773 using primers JC_VF0844_drop_FWD and JC_773_drop_REV) and *ermE**p (amplified from pIJ10257 using primers JC_permE_drop_FWD/_REV). The two resulting PCR products were gel-purified, digested with *Nco*I and ligated together. The resulting Apra^R-*ermE**p cassette was used to target the upstream region of VF0844 in JC089, as described previously (Gust et al., 2004). Plasmid DNA was isolated from apramycin resistant colonies (40 μg/ml) and the *ermE**p-targeted construct pJC123 was introduced into *E. coli* Top10, yielding JC090. Correct insertion of the promoter was verified by sequencing across the insertion site.

Phenotypic verification of mutant strains

V. fischeri and *E. coli* strains were grown for 3.5 days in the dark prior to harvesting. Cell pellets were collected by centrifugation (5180 x g, 30 min), washed with water and repeatedly extracted with acetone:MeOH (2:1 vol/vol) in a Waring blender (adapted from Ref. (Starr et al., 1977)). Extracts were combined and after concentration in a rotary evaporator, the aryl-polyenes were extracted in ethyl ether. The compounds were dried down and re-dissolved in CH_2Cl_2:MeOH (2:1 vol/vol), to be released from the cell material by mild base hydrolysis (by adding ½ volume of 0.5 M NaOH, for 30 min at 25°C). The reaction was neutralized with HCl, passed over Na_2SO_4 and dried in a rotary evaporator. The presence or absence of aryl polyenes in the extracts of different strains was detected by thin layer chromatography (TLC; developed in $CHCl_3$) and high-performance liquid chromatography (HPLC; monitoring absorbance at 441nm).

Special Considerations for Structure Elucidation

Difficulties in the isolation of related aryl-polyenes from *Lysobacter enzymogenes* are well known (Wang et al., 2013). To date structure elucidation efforts for this class of compounds have relied primarily on infrared spectroscopy (IR), ultraviolet spectroscopy (UV), mass spectrometry (MS), and some chemical manipulations, but due to the light sensitivity and limited material, no NMR spectra have previously been reported (Andrewes et al., 1973). By developing isolation conditions that rigorously exclude exposure to light, we have now isolated sufficient material to complete the first solution NMR characterization of a molecule of this type, and have confirmed all elements of the structure elucidation through careful and exhaustive examination of 1D and 2D NMR spectra.

Structure assignment for the *E. coli* CFT073 aryl polyene (APE_{Ec})

APE_{EC}, a red amorphous powder, was determined to have a molecular formula of $C_{21}H_{22}O_3$ based on the observation of the $[M-H]^-$ adduct at 321.1496 *m/z* (Δppm = -0.310) and analysis of one and two-dimensional NMR experiments (**SI Figure 7c** and **SI Table XIV**). Based on ^1H NMR and HSQC assignment of 15 aromatic and vinylic protons, one aromatic methyl singlet, one methoxy singlet, and one potential broad singlet phenolic proton at 8.43 ppm. From the TOCSY spectrum it was clear that the molecule contained two independent spin systems. One spin system was defined an phenyl ring with a 1,2,4 substitution pattern based on classical H18-H19 ortho-coupling constants ($^3J_{HH}$ = 7.2 Hz), meta-coupling between H15 and H19 ($^4J_{HH}$ = 2.1 Hz), and HMBC correlations from H19 and H20 to C17, H19 and H20 to C15, the aromatic methyl singlet to C15 and C16, and the phenolic proton to C17 and C16. The second spin system was defined as a long conjugated polyene terminating at a methyl ester and the 1,2,4-phenyl ring. The terminus of the polyene chain at the phenyl ring was identified based on HMBC signals from H15 and H19 to C13 as well as ROESY signals between H15 and H13, and H19 and H13.

The methyl ester was identified via an HMBC correlation from the singlet methoxy proton signal at 3.7 ppm to the quaternary carbon C1 at 167.7 ppm. Protons H2 (doublet, ^1H 5.93 ppm $^3J_{HH}$ = 15.2 Hz; ^{13}C 120.5 ppm) and H3 (doublet of doublets, ^1H 7.33 ppm $^3J_{HH}$ = 15.1, 11.4; ^{13}C 145.1 ppm) displayed strong COSY correlations to one another, and both possessed HMBC correlations to the ester carbonyl at C1. These chemical shifts and coupling constants are indicative of the presence of an alpha-beta unsaturated ester. The assignment of the polyene chain continued through H4 based on HMBC and COSY correlations. Of the remaining C_8H_7 one quaternary carbon is contained in the phenyl ring connecting the aromatic functionality to the polyene, leaving the remaining constituents (C_7H_7; all between ^1H 6.85 – 6.40 ppm and ^{13}C 126 – 138 ppm) as a contiguous all-*trans* polyene chain connecting the aromatic head group with the methyl ester tail. The all-*trans* configuration is suggested by the absence of the '*cis* peak' centered around 340 nm in the UV spectrum that is a diagnostic marker for alkene chains that possess at least one region of non-linear (angulated) region of lesser symmetry, caused by the presence of *cis*-olefin(s) (Baraldi et al., 2008).

Structure assignment for the *V. fischeri* ES114 aryl polyene (APE_{VF})

APE_{VF}, a red amorphous powder, was determined to have a molecular formula of $C_{22}H_{24}O_3$ based on the observation of the $[M-H]^-$ adduct at 335.1652 *m/z* (Δppm = 0.0) and analysis of one and two-dimensional NMR experiments (**SI Figure 7d** and **SI Table XIV**). Comparison of the NMR spectra in acetone-D6 to that of APE_{EC} in acetone-D6 indicated that the polyene segments of the two molecules were very similar based on related chemical shifts. To alleviate solubility issues, one and two-dimensional experiments were repeated in DMSO-D6. The alpha-beta unsaturated methyl ester motif was assigned based on both COSY correlations between H2 (doublet, ^1H 5.97 ppm $^3J_{HH}$ = 15.2 Hz) and H3 (doublet of doublets, ^1H 7.31 ppm $^3J_{HH}$ = 15.2, 11.5) and HMBC signals from both H2 and H3 to C1 at 166.7 ppm, as well as HMBC correlation from the methoxy proton singlet at 3.66 ppm to ester carbonyl C1. As with the previous

structure assignment, H4 was assigned based on COSY correlation to H3 and HMBC correlations to C2 and C3. While signal overlap complicated interpretation of the COSY spectrum, H5 could be assigned based on HMBC correlations to C4 and C3 as well as an HMBC correlation to C5 from H3 (assigned in conjunction with HSQC data).

The one aromatic singlet in the downfield region of the spectrum (H15, H19; ^{1}H 7.06 ppm; ^{13}C 126.7 ppm) integrated for two protons, suggesting a 1,2,4,6-tetra-substituted symmetric aromatic group. The aromatic methyl singlet (^{1}H 2.15 ppm; ^{13}C 16.3 ppm) integrating for six protons and the phenol signal at 8.47 ppm suggested para substitution of the polyene and phenolic OH moieties, with the methyl groups either ortho or meta to the phenolic OH. An HMBC correlation from singlet aromatic protons H15 and H19 to C13, coupled with through space ROESY correlations between H13 and H15/H19 proved that the substitution pattern of the phenol was 1,2,4,6 substituted. As with the previous structure assignment, completion of the structure elucidation was accomplished by consideration of the remaining double bond equivalents and the chemical shifts for the ^{1}H and ^{13}C resonances for the remaining atoms, which unequivocally determined that the aromatic head group and the methyl ester tail be connected via a linear polyene chain.

Consideration of the UV-profiles of the isolated peaks with previously reported data on alpha and beta carotenoids suggests and all-*trans* structure for both molecules (Jurkowitz et al., 1959; Tsukida et al., 1982; Zechmeister and Polgar, 1943). A *cis*-double bond within extended polyene chains breaks the linearity of the molecule, resulting in a shorter chain and new absorption axis. The result is what is known as a *cis*-peak in the UV spectra between 310 and 370 nm. In both the *V. fischeri* and *E. coli* UV-profiles there is little or no absorbance between 310-370 nm, indicating all-*trans* configurations for both structures (**SI Figure 7g**).

Growth and Purification

Cultures were grown in LB Broth Miller from Fischer (tryptone 10 g, yeast extract 5 g, sodium chloride 10 g) buffered with 50 mM TRIS at pH 7.5. After autoclaving the media and letting it cool to 60°C, kanamycin and ampicillin were added via sterile filtration at final concentrations of 50 μg/ml to maintain plasmids. Where necessary, 1.5% agar was added to prepare solid media. For large-scale preparation the following growth process was repeated eight times, 4 l per iteration, to produce a total of 32 l of culture. Bacteria were grown on solid media at 37°C overnight after streaking them on solid media. Colonies were used to inoculate 10 ml of media in a 50 ml culture tube. Cultures were grown in the dark at 37°C and shaken at 250 rpm. After 8 hours the small-scale culture was used to inoculate 100 ml of antibiotic-containing media in a 250 ml wide neck Erlenmeyer flask and grown under the same conditions overnight. Finally 50 ml of this medium-scale culture was used to inoculate 1 l that was subsequently grown for 3 days, spun down at 4000 rpm at 4°C for 20 minutes, transferred to a 50 ml falcon tube, and lyophilized. After the cells were spun down, all the subsequent steps were conducted in the dark with the use of red LED headlamp.

Four LBS agar plates were streaked with a lawn of *V. fischeri* ES114 and incubated at 30ºC over night. The following day, cells were scraped off the plates, suspended in a small volume of LBS and used to inoculate 80 l of LBS. These *V. fischeri* ES114 production cultures were grown for 60 hours (30ºC, 150 rpm) in a light protected environment and cell pellets were subsequently harvested by centrifugation (5180 x g, 4ºC, 20 min) and lyophilized.

The same process was used to extract both *E. coli* and *V. fischeri* separately. The dried cell pellets were split into two 1 l Erlenmeyer flasks containing 500 ml of 1:2 methanol/dichloromethane, shaken for 1 hour at 180 rpm, stirred vigorously with a magnetic stir bar for 1 hour, then vacuum filtered, and the solution concentrated to dryness under vacuum. The cell debris was re-extracted three times in this fashion and all extracts for each strain were combined into a 1 l round bottom flask. The

dried extract was suspended in 400 ml of 1:2 methanol/dichloromethane at room temperature. A saponification reaction was performed on each extract by stirring the solution rapidly with a magnetic stir bar and adding 200 ml of 0.5 M potassium hydroxide. The reaction was carried out for 1 hour at which time the mixture was neutralized with 2.0 M sulfuric acid to pH 7.0 and transferred to a 2 l separatory funnel. The organic layer was collected, washed three times with brine, once with deionized water, dried over sodium sulfate, transferred through a paper filter into a 500 ml round bottom flask, and concentrated to dryness under vacuum. The dried extracts were suspended in 10 ml of acetone and carried forward to purification (**SI Figure 7e**).

E.coli materials were purified on RP-HPLC using a two step purification protocol. Firstly, crude material was purified on a semi-prep RP column (Phenomenex Synergi Fusion-RP, 250 x 10 mm, 10 μm) using a gradient of acetonitrile MeCN:H_2O + 0.02% formic acid (32% MeCN for 26 minutes, 100% MeCN for 9 min, 20% MeCN for 2 minutes, and a 9 minute re-equilibration) at a flow rate of 4 ml min^{-1}. The peak eluting at 16 min displaying a strong UV absorbance at 441 nm was collected and re-purified using an analytical column (Phenomonex Kinetix 2.6 μm XB-C18 100 x 4.6 mm) using a gradient of MeCN:H_2O + 0.02% formic acid (50% MeCN for 2 min, 50%-65% MeCN over 20 min) at a flow rate of 2 ml min^{-1} (**SI Figure 7f**). APE_{EC}, the peak eluting at 16 min that displayed the correct UV spectra, was collected, dried under vacuum, and stored at -20°C in a 5 ml amber vial. Standard one and two-dimensional NMR experiments were performed on a Varian 600 MHz cryoprobe NMR in acetone-D6.

The *V. fischeri* extract was first purified by RP-HPLC analytical column (Phenomonex Kinetex 5μm XB-C18 250 x 4.6 mm) using a gradient of MeCN:H_2O + 0.02% formic acid (50%-60% MeCN 2 min, 60%-73.8% MeCN over 11 min, 73.8%-95% over 1 min, 95%-100% over 3 min, 100% for 1 min) at a flow rate of 2 ml min^{-1}. The peak at 9.5 min with absorbance at 441 nm was collected and re-purified on an analytical column (Phenomonex Synergi 10 μm Fusion-RP 250 x 4.6 mm) using a gradient of methanol (MeOH):H_2O + 0.02% formic acid (50% MeOH for 2 min, 50%-90% MeOH over 15 min, 100% MeOH for 2 min) at a flow rate of 2 ml min^{-1} (**SI Figure 7f**). APE_{VF}, the peak eluting at 18 min that displayed the correct UV spectra, was collected, dried under vacuum, and stored at -20°C in a 5 ml amber vial. Standard one and two-dimensional NMR experiments were performed on a Varian 600 MHz cryoprobe NMR in both acetone-D6 and DMSO-D6.

Mass Spectrometry

Compounds were analyzed on an Agilent uPLC-ESI-TOF mass spectrometer, comprising a 1260 binary pump in low dwell volume mode, an Agilent column oven heated to 45°C, and an Agilent 6230 Time-of-flight Mass Spectrometer with an electrospray ionization (ESI) source. 1 μl of sample, dissolved in 50% v/v methanol/water, was injected onto a 1.8 μm particle size, 50 x 2.3 mm I.D. ZORBAX RRHT column. Each sample was subjected to a MeCN:H_2O gradient from 10% to 90% MeCN over 4 min followed by 1.5 min at 90% MeCN at a flow rate of 0.8 ml min^{-1}. Formic acid, 200 μl/l, was added to both the water and the acetonitrile. Water, 1 ml min^{-1}, was added to the acetonitrile. The mass spectrometer was run with a detector mass range of 100 to 1700 *m/z*. The ESI source was operated with a desolvation temperature of 350° C and a drying gas flow rate of 11 l min^{-1}. The fragmentor voltage was held at 135 V. In positive ESI mode, the capillary voltage was ramped from 2500 V at 0 min to 2750 V at 1 min, and to 3000 V at 3 min. In negative ESI mode, the capillary voltage was held at 2750 V. Each sample was run in high resolution (4GHz) detector mode.

SUPPLEMENTAL REFERENCES

Andrewes, A.G., Hertzberg, S., Liaaen-Jensen, S., and Starr, M.P. (1973). *Xanthomonas* pigments. 2. The *Xanthomonas* "carotenoids"--non-carotenoid brominated aryl-polyene esters. Acta Chem Scand *27*, 2383-2395.

Arnison, P.G., Bibb, M.J., Bierbaum, G., Bowers, A.A., Bugni, T.S., Bulaj, G., Camarero, J.A., Campopiano, D.J., Challis, G.L., Clardy, J.*, et al.* (2013). Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature. Nat Prod Rep *30*, 108-160.

Barabasi, A.L., and Oltvai, Z.N. (2004). Network biology: understanding the cell's functional organization. Nat Rev Genet *5*, 101-113.

Baraldi, I., Benassi, E., and Spalletti, A. (2008). *cis* peak as probe to investigate the molecular structure. Application to the rotational isomerism of 2,5-diphenylethenyl(hetero)arenes. Spectrochim Acta A *71*, 543-549.

Barrett, T., Clark, K., Gevorgyan, R., Gorelenkov, V., Gribov, E., Karsch-Mizrachi, I., Kimelman, M., Pruitt, K.D., Resenchuk, S., Tatusova, T.*, et al.* (2012). BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. Nucl Acids Research *40*, D57-63.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC Bioinformatics *10*, 421.

Chen, J., Densmore, D., Ham, T.S., Keasling, J.D., and Hillson, N.J. (2012). DeviceEditor visual biological CAD canvas. J Biol Eng *6*, 1.

Donadio, S., Monciardini, P., and Sosio, M. (2007). Polyketide synthases and nonribosomal peptide synthetases: the emerging view from bacterial genomics. Nat Prod Rep *24*, 1073-1109.

Eddy, S.R. (2008). A probabilistic model of local sequence alignment that simplifies statistical significance estimation. PLoS Comput Biol *4*, e1000069.

Eddy, S.R. (2010). Hidden Markov model speed heuristic and iterative HMM search procedure. BMC Bioinformatics *11***,** 431.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucl Acids Research *32*, 1792-1797.

Federhen, S. (2012). The NCBI Taxonomy database. Nucl Acids Research *40*, D136-143.

Fischbach, M.A., and Walsh, C.T. (2009). Antibiotics for emerging pathogens. Science *325*, 1089-1093.

Fischbach, M.A., Walsh, C.T., and Clardy, J. (2008). The evolution of gene collectives: How natural selection drives chemical innovation. Proc Natl Acad Sci USA *105*, 4601-4608.

Garwin, J.L., Klages, A.L., and Cronan, J.E., Jr. (1980). Beta-ketoacyl-acyl carrier protein synthase II of *Escherichia coli*. Evidence for function in the thermal regulation of fatty acid synthesis. J Biol Chem *255*, 3263-3265.

Gibson, D.G., Young, L., Chuang, R.Y., Venter, J.C., Hutchison, C.A., 3rd, and Smith, H.O. (2009). Enzymatic assembly of DNA molecules up to several hundred kilobases. Nat Methods *6*, 343-345.

Graf, J., Dunlap, P.V., and Ruby, E.G. (1994). Effect of transposon-induced motility mutations on colonization of the host light organ by *Vibrio fischeri*. J Bacteriol *176*, 6986-6991.

Gust, B., Chandra, G., Jakimowicz, D., Yuqing, T., Bruton, C.J., and Chater, K.F. (2004). Lambda red-mediated genetic manipulation of antibiotic-producing *Streptomyces*. Adv App Microbiol *54*, 107-128.

Jurkowitz, L., Loeb, J.N., Brown, P.K., and Wald, G. (1959). Photochemical and stereochemical properties of carotenoids at low temperatures. Nature *184*, 614-624.

Klassen, J.L., and Currie, C.R. (2012). Gene fragmentation in bacterial draft genomes: extent, consequences and mitigation. BMC Genomics *13*, 14.

Krogh, A., Brown, M., Mian, I.S., Sjolander, K., and Haussler, D. (1994). Hidden Markov models in computational biology. Applications to protein modeling. J Mol Biol *235*, 1501-1531.

Law, A., and Boulanger, M.J. (2011). Defining a structural and kinetic rationale for paralogous copies of phenylacetate-CoA ligases from the cystic fibrosis pathogen *Burkholderia cenocepacia* J2315. J Biol Chem *286*, 15577-15585.

Le Roux, F., Binesse, J., Saulnier, D., and Mazel, D. (2007). Construction of a *Vibrio splendidus* mutant lacking the metalloprotease gene *vsm* by use of a novel counterselectable suicide vector. Appl Environ Microbiol *73*, 777-784.

Letunic, I., and Bork, P. (2007). Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics *23*, 127-128.

Letunic, I., and Bork, P. (2011). Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. Nucl Acids Research *39*, W475-478.

Letzel, A.C., Pidot, S.J., and Hertweck, C. (2013). A genomic approach to the cryptic secondary metabolome of the anaerobic world. Nat Prod Rep *30*, 392-428.

Li, L., Stoeckert, C.J., Jr., and Roos, D.S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res *13*, 2178-2189.

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics *22*, 1658-1659.

Lin, K., Zhu, L., and Zhang, D.Y. (2006). An initial strategy for comparing proteins at the domain architecture level. Bioinformatics *22*, 2081-2086.

Liu, W.T., Yang, Y.L., Xu, Y., Lamsa, A., Haste, N.M., Yang, J.Y., Ng, J., Gonzalez, D., Ellermeier, C.D., Straight, P.D*., et al.* (2010). Imaging mass spectrometry of intraspecies metabolic exchange revealed the cannibalistic factors of *Bacillus subtilis*. Proc Natl Acad Sci USA *107*, 16286-16290.

Markowitz, V.M., Chen, I.M., Palaniappan, K., Chu, K., Szeto, E., Grechkin, Y., Ratner, A., Jacob, B., Huang, J., Williams, P*., et al.* (2012). IMG: the Integrated Microbial Genomes database and comparative analysis system. Nucl Acids Research *40*, D115-122.

Medema, M.H., Blin, K., Cimermancic, P., de Jager, V., Zakrzewski, P., Fischbach, M.A., Weber, T., Takano, E., and Breitling, R. (2011). antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Nucl Acids Research *39*, W339-346.

Medema, M.H., Takano, E., and Breitling, R. (2013). Detecting sequence homology at the gene cluster level with MultiGeneBlast. Mol Biol Evol *30*, 1218-1223.

Munkres, J. (1957). Algorithms for the assignment and transportation problems. J Soc Ind Appl Math *5*, 32-38.

Nett, M., Ikeda, H., and Moore, B.S. (2009). Genomic basis for natural product biosynthetic diversity in the actinomycetes. Nat Prod Rep *26*, 1362-1384.

Paulsen, I.T., Press, C.M., Ravel, J., Kobayashi, D.Y., Myers, G.S., Mavrodi, D.V., DeBoy, R.T., Seshadri, R., Ren, Q., and Madupu, R. (2005). Complete genome sequence of the plant commensal *Pseudomonas fluorescens* Pf-5. Nat Biotechnol *23*, 873-878.

Pavoine, S., Baguette, M., and Bonsall, M.B. (2010). Decomposition of trait diversity among the nodes of a phylogenetic tree. Ecol Monogr *80*, 485-507.

Pei, J., Tang, M., and Grishin, N.V. (2008). PROMALS3D web server for accurate multiple protein sequence and structure alignments. Nucleic Acids Res *36*, W30-34.

Press, W.H., Teukolsky, S.A., Vetterling, W.T., and Flannery, B.P. (1992). Numerical recipes in C: the art of scientific computing. 2. Cambridge: CUP.

Punta, M., Coggill, P.C., Eberhardt, R.Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J.*, et al.* (2012). The Pfam protein families database. Nucleic Acids Res *40*, D290-301.

Quan, J., and Tian, J. (2011). Circular polymerase extension cloning for high-throughput cloning of complex and combinatorial DNA libraries. Nat Protoc *6*, 242-251.

Rattray, J.E., Strous, M., Op den Camp, H.J., Schouten, S., Jetten, M.S., and Damste, J.S. (2009). A comparative genomics study of genetic products potentially encoding ladderane lipid biosynthesis. Biology Direct *4*, 8.

Reger, A.S., Wu, R., Dunaway-Mariano, D., and Gulick, A.M. (2008). Structural characterization of a 140 degrees domain movement in the two-step reaction catalyzed by 4-chlorobenzoate:CoA ligase. Biochemistry *47*, 8016-8025.

Rehm, B.H. (2010). Bacterial polymers: biosynthesis, modifications and applications. Nat Rev Microbiol *8*, 578-592.

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N.N., Anderson, I.J., Cheng, J.F., Darling, A., Malfatti, S., Swan, B.K., Gies, E.A.*, et al.* (2013). Insights into the phylogeny and coding potential of microbial dark matter. Nature *499*, 431-437.

Rose, P.W., Bi, C., Bluhm, W.F., Christie, C.H., Dimitropoulos, D., Dutta, S., Green, R.K., Goodsell, D.S., Prlic, A., Quesada, M.*, et al.* (2013). The RCSB Protein Data Bank: new resources for research and education. Nucleic Acids Res *41*, D475-482.

Röttig, M., Medema, M.H., Blin, K., Weber, T., Rausch, C., and Kohlbacher, O. (2011). NRPSpredictor2—a web server for predicting NRPS adenylation domain specificity. Nucl Acids Research *39*, W362-W367.

Rückert, C., Blom, J., Chen, X., Reva, O., and Borriss, R. (2011). Genome sequence of *B. amyloliquefaciens* type strain DSM7(T) reveals differences to plant-associated *B. amyloliquefaciens* FZB42. J Biotechnol *155*, 78-85.

Salzberg, S.L., Delcher, A.L., Kasif, S., and White, O. (1998). Microbial gene identification using interpolated Markov models. Nucleic Acids Res *26*, 544-548.

Sambrook, J., and Russell, D.W. (2001). Molecular cloning: A laboratory manual (Cold Spring Harbor Laboratory Press).

Seyed-Allaei, H., Bianconi, G., and Marsili, M. (2006). Scale-free networks with an exponent less than two. Phys Rev E *73*, 046113.

Smoot, M.E., Ono, K., Ruscheinski, J., Wang, P.L., and Ideker, T. (2011). Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics *27*, 431-432.

Starr, M.P., Jenkins, C.L., Bussey, L.B., and Andrewes, A.G. (1977). Chemotaxonomic significance of the xanthomonadins, novel brominated aryl-polyene pigments produced by bacteria of the genus Xanthomonas. Arch Microbiol *113*, 1-9.

Strobel, T., Al-Dilaimi, A., Blom, J., Gessner, A., Kalinowski, J., Luzhetska, M., Puhler, A., Szczepanowski, R., Bechthold, A., and Rückert, C. (2012). Complete genome sequence of *Saccharothrix espanaensis* DSM

44229(T) and comparison to the other completely sequenced Pseudonocardiaceae. BMC Genomics *13*, 465.

Strous, M., Pelletier, E., Mangenot, S., Rattei, T., Lehner, A., Taylor, M.W., Horn, M., Daims, H., Bartol-Mavel, D., Wincker, P*., et al.* (2006). Deciphering the evolution and metabolism of an anammox bacterium from a community genome. Nature *440*, 790-794.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol *28*, 2731-2739.

Tobias, N.J., Doig, K.D., Medema, M.H., Chen, H., Haring, V., Moore, R., Seemann, T., and Stinear, T.P. (2013). Complete genome sequence of the frog pathogen *Mycobacterium ulcerans* ecovar Liflandii. J Bacteriol *195*, 556-564.

Tsukida, K., Saiki, K., Takii, T., and Koyama, Y. (1982). Separation and determination of *cis*/*trans*-beta-carotenes by high-performance liquid chromatography. J Chromatography *245*, 359-364.

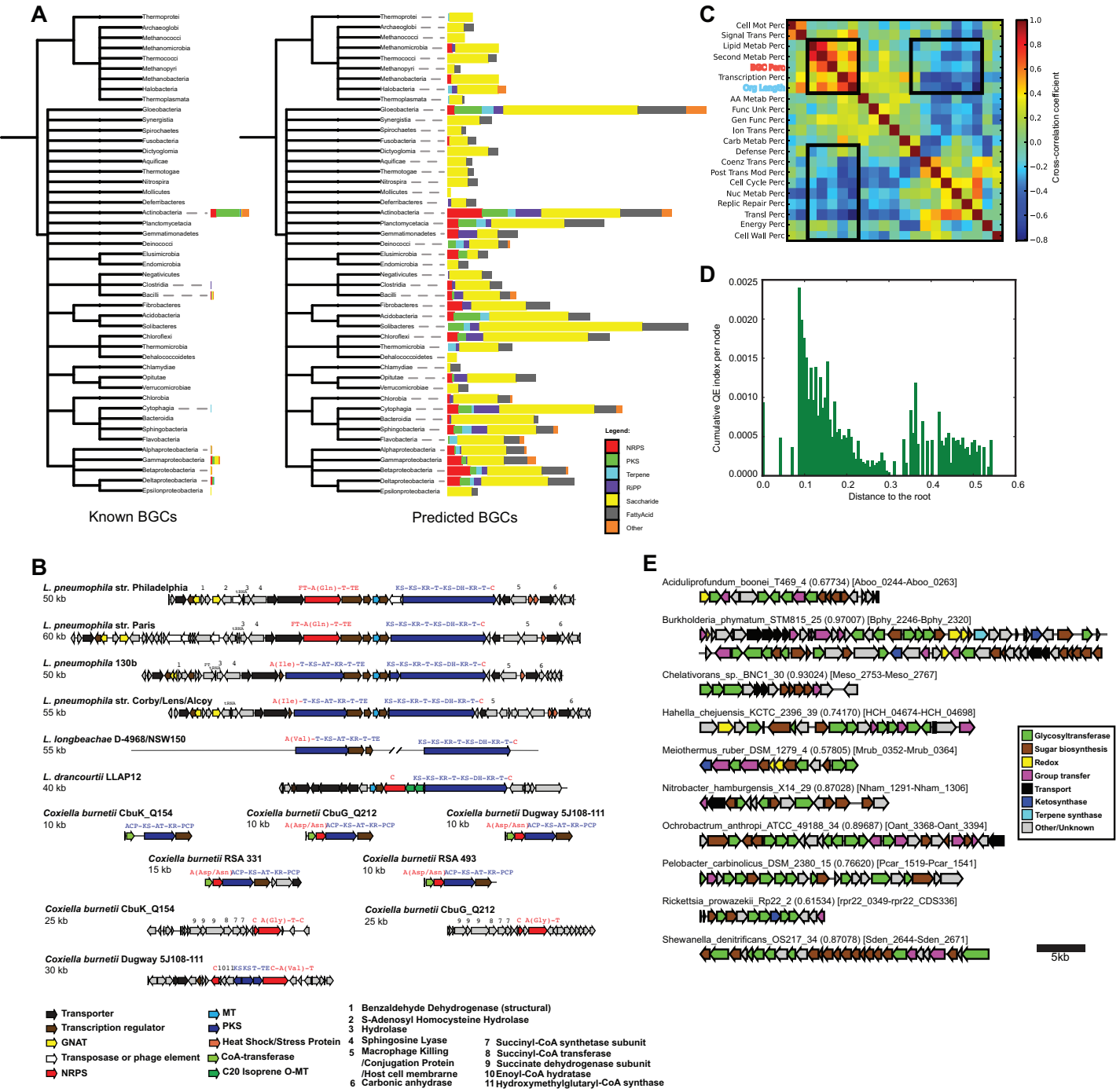
Wang, Y., Qian, G., Li, Y., Wang, Y., Wang, Y., Wright, S., Li, Y., Shen, Y., Liu, F., and Du, L. (2013). Biosynthetic mechanism for sunscreens of the biocontrol agent *Lysobacter enzymogenes*. PLoS One *8*, e66633.

Zechmeister, L., and Polgar, A. (1943). *cis-trans* isomerization and *cis*-peak effect in the alpha-carotene set and in some other stereoisomeric sets. J Am Chem Soc *66*, 137-144.

Zeigler, D.R. (2011). The genome sequence of *Bacillus subtilis* subsp. *spizizenii* W23: insights into speciation within the *B. subtilis* complex and into the history of *B. subtilis* genetics. Microbiology *157*, 2033-2041.

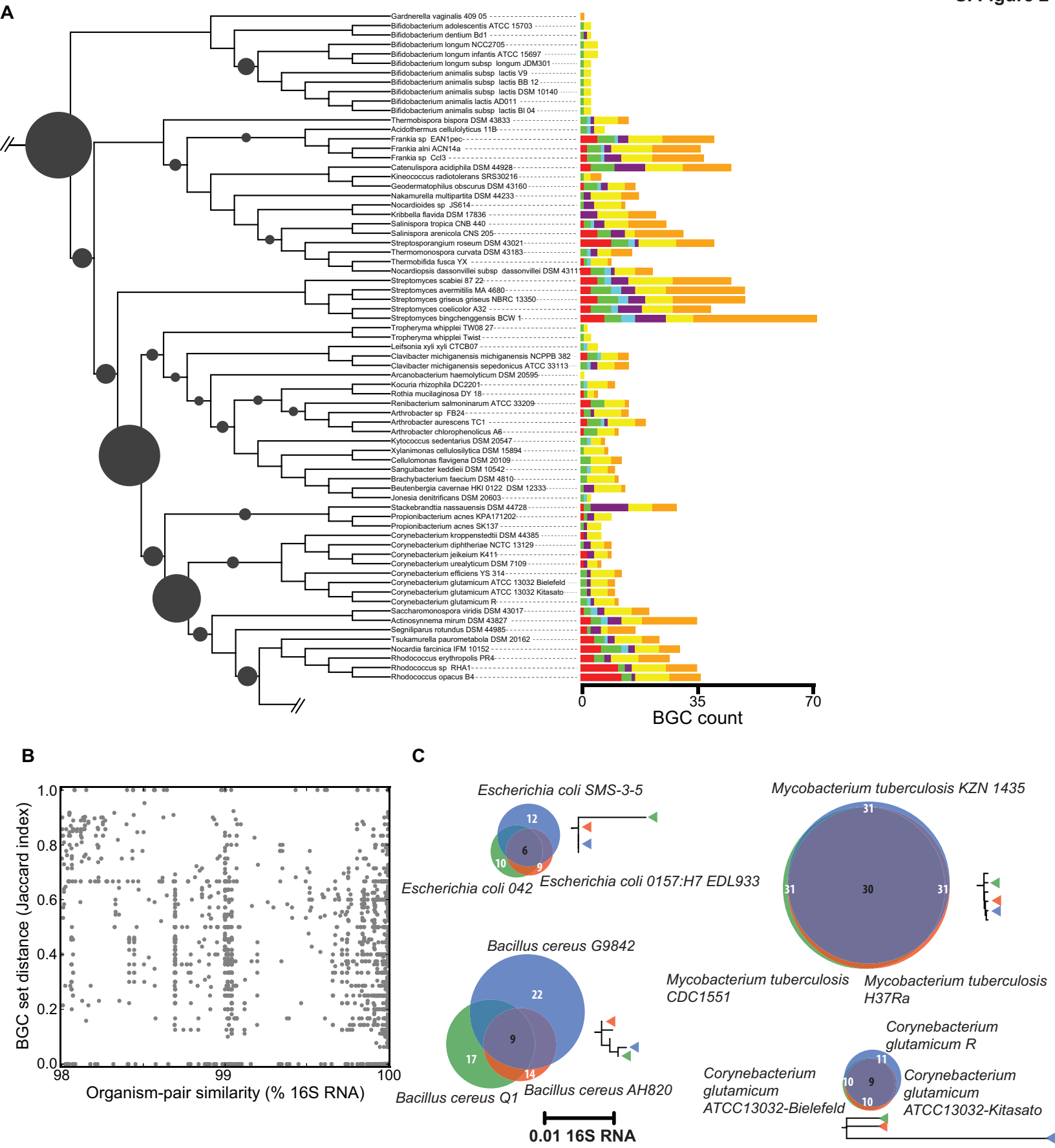
Supplemental Figure

SI Figure 1



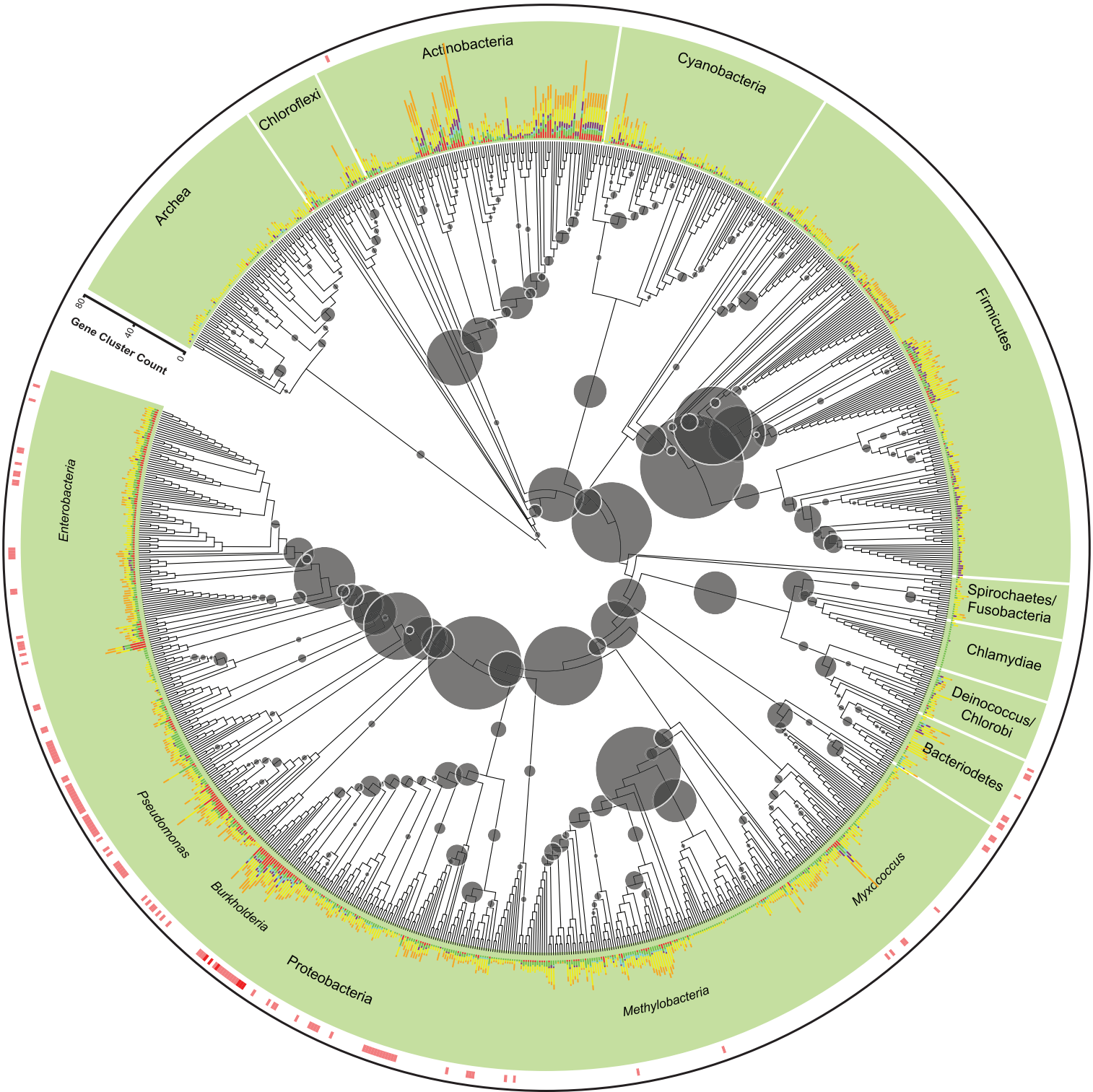
Supplemental Figure

SI Figure 2



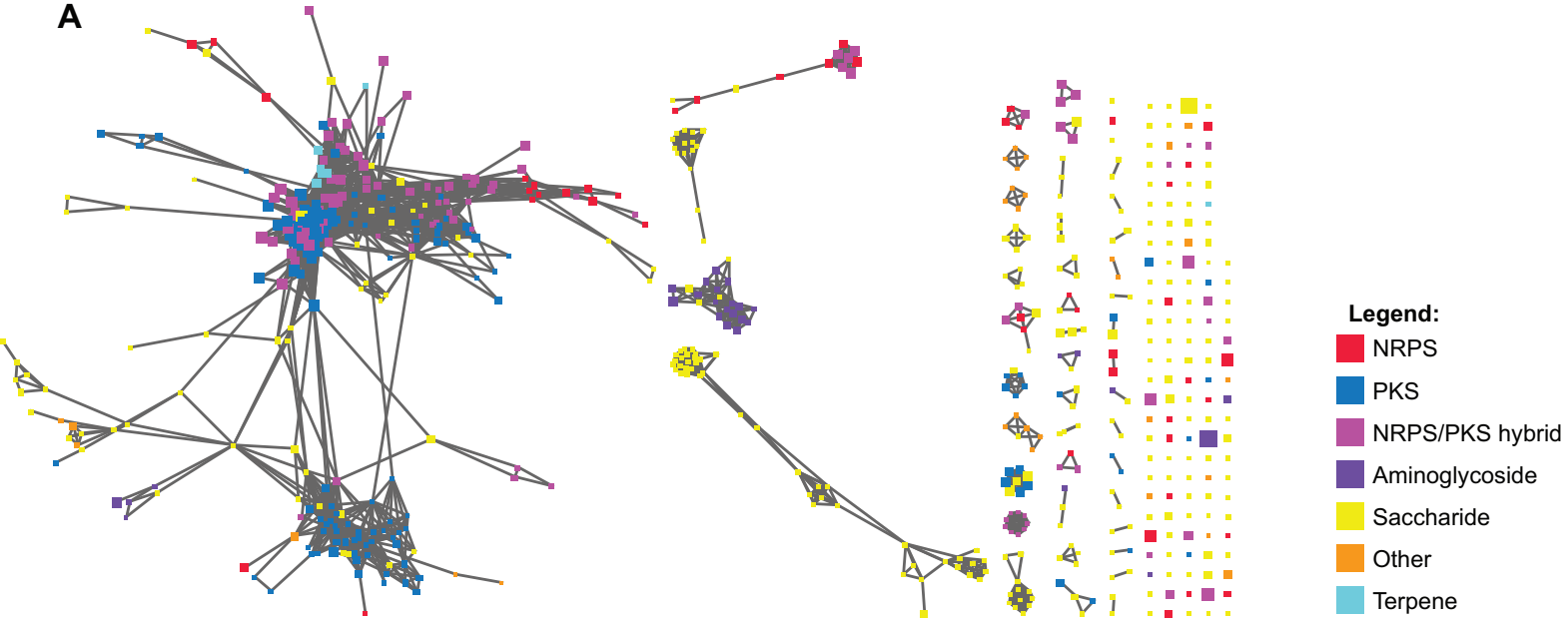
Supplemental Figure

SI Figure 3



SI Figure 4

A



B

Statistics for the graph with >0.6 threshold

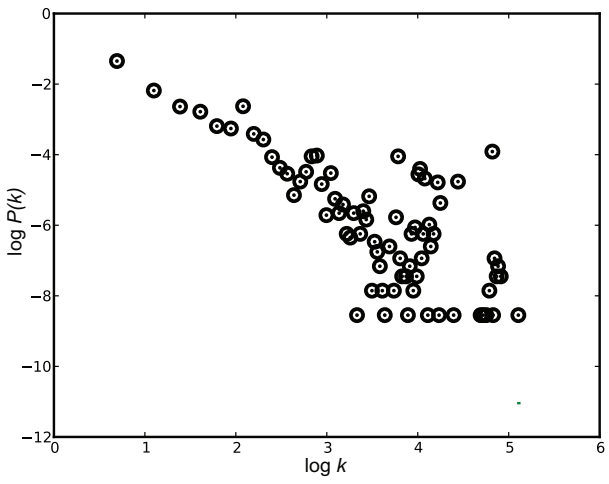
	#nodes	#edges	γ	L	C	L_{random}	C_{random}	K(k)	p-value $_{K(k)}$
ALL	7,391	136,483	1.66 ± 0.07	1.11	0.69	2.83	0.005	-0.012 ± 0.009	0.45
PKS	1,344	94,357	1.03 ± 0.07	1.11	0.78	1.90	0.100	-0.017 ± 0.010	0.39
Terpene	137	417	0.70 ± 0.49	1.12	0.54	2.90	0.050	0.071 ± 0.061	0.54
Saccharide	2,588	18,896	1.78 ± 0.21	1.1	0.66	3.21	0.006	-0.035 ± 0.032	0.39
RP	290	1,414	0.83 ± 0.29	1.11	0.72	2.73	0.038	-0.131 ± 0.041	0.24
Siderophore	200	1,213	0.47 ± 0.36	1.43	0.79	2.39	0.060	-0.147 ± 0.054	0.19
Hybrid	694	5,525	0.92 ± 0.19	1.13	0.73	2.66	0.023	-0.018 ± 0.016	0.49
NRPS	524	4,431	1.43 ± 0.21	1.16	0.7	2.53	0.030	0.016 ± 0.048	0.74

C

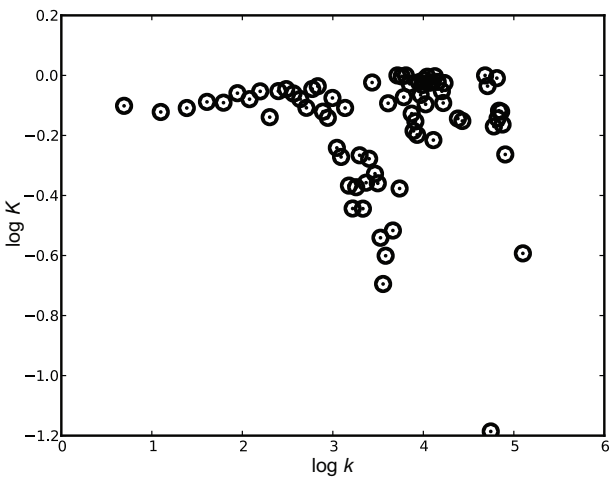
Statistics for the graph with >0.8 threshold

	#nodes	#edges	γ	L	C	L_{random}	C_{random}	K(k)	p-value $_{K(k)}$
ALL	5,152	34,976	2.16 ±0.15	1.07	0.67	3.57	0.0026	-0.028 ± 0.025	0.49
PKS	1,151	18,836	1.52 ±0.14	1.14	0.79	2.36	0.029	-0.005 ± 0.300	0.86
Terpene	97	146	0.03 ±0.99	1.10	0.44	NaN	NaN	0.370 ± 0.460	0.73
Saccharide	1,776	8,199	1.50 ±0.30	1.06	0.64	3.62	0.0056	-0.057 ± 0.030	0.36
RP	221	537	0.73 ±0.71	1.07	0.67	2.29	0.03	-0.044 ± 0.066	0.52
Siderophore	159	609	0.55 ±0.60	1.10	0.67	2.7	0.04	0.120 ± 0.101	0.53
Hybrid	489	1,661	1.03 ±0.41	1.06	0.73	3.45	0.017	-0.130 ± 0.041	0.0034
NRPS	369	2,572	1.07 ±0.30	1.06	0.72	2.53	0.037	0.048 ± 0.240	0.27

D

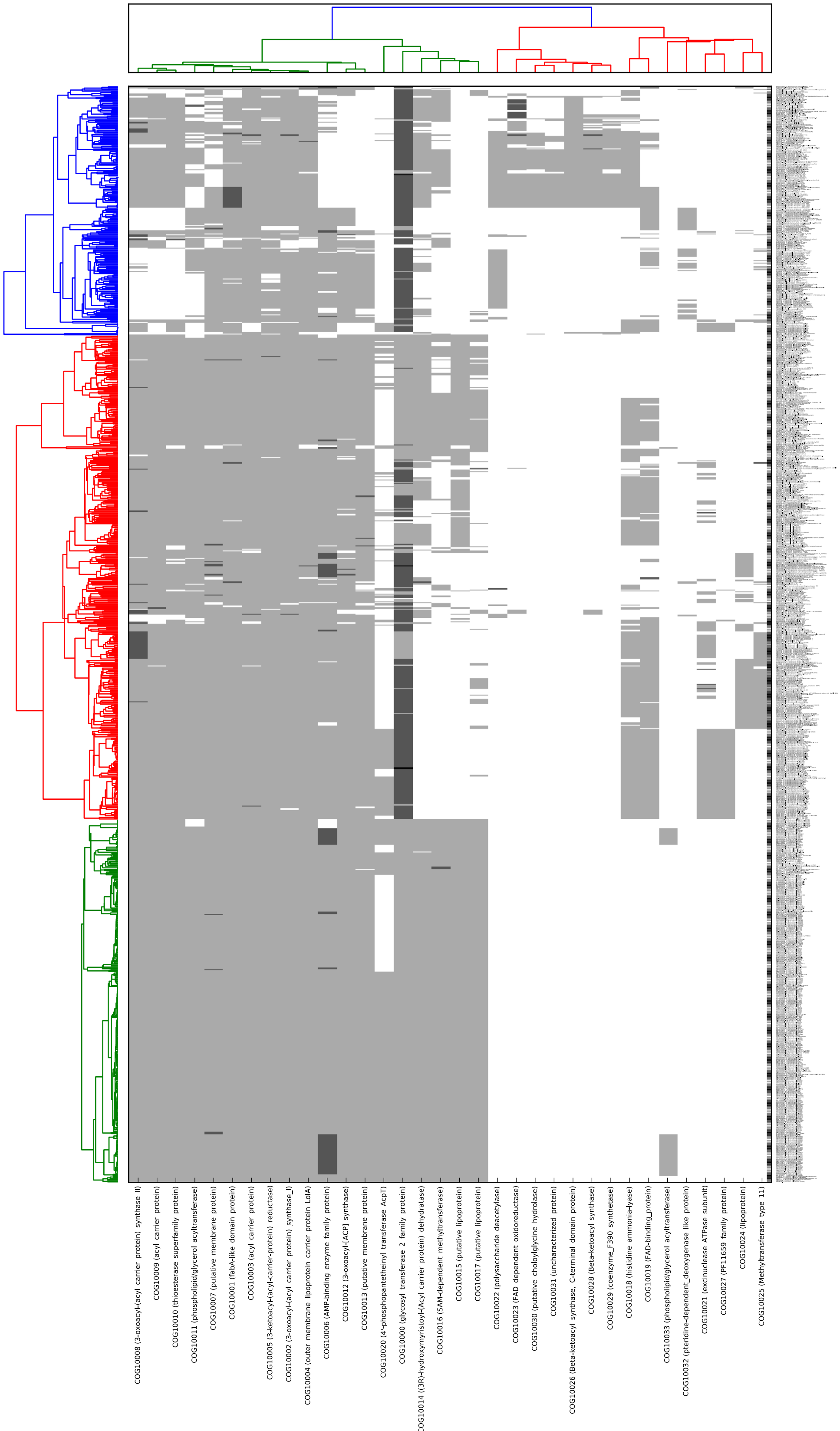


E



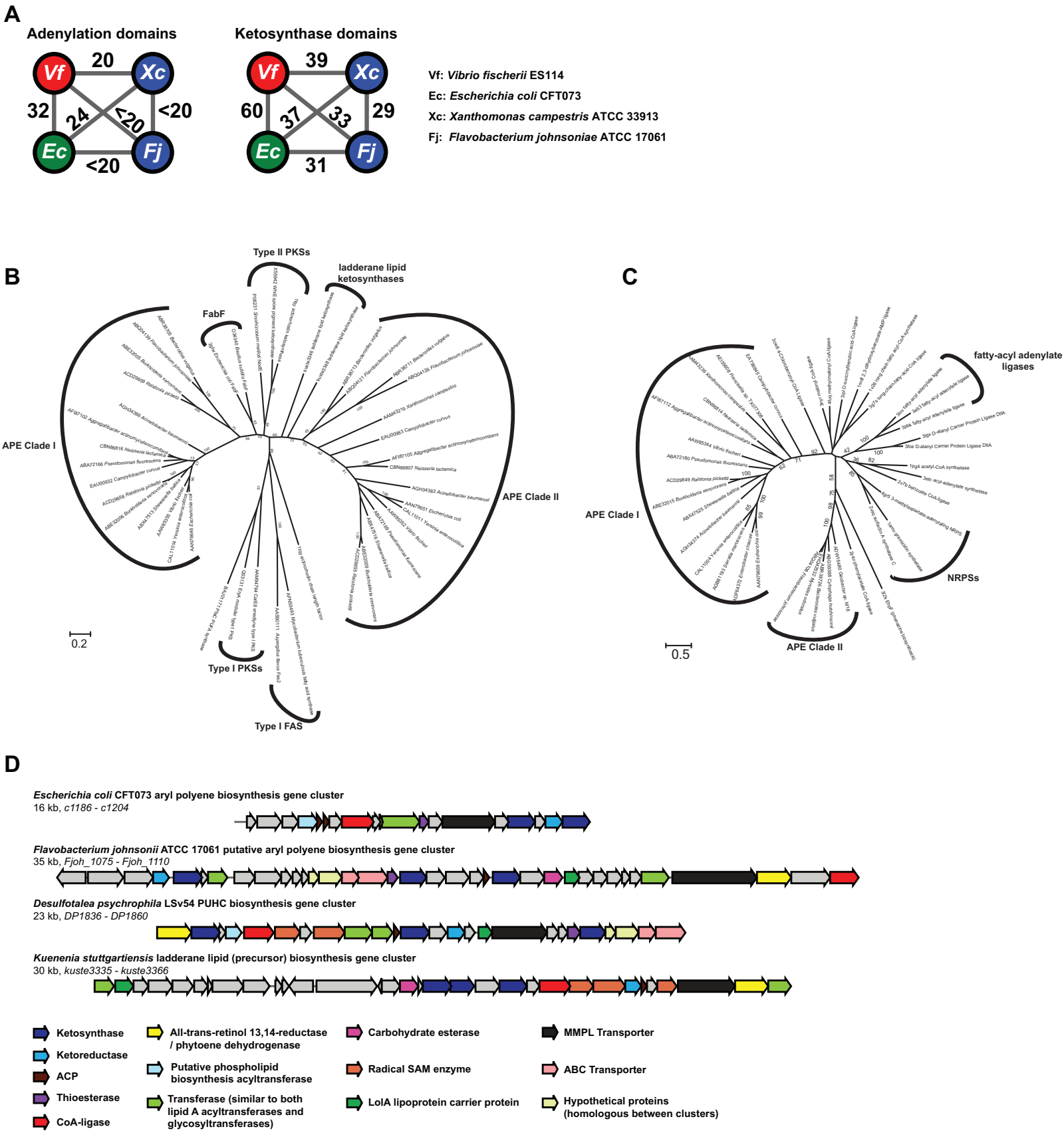
Supplemental Figure

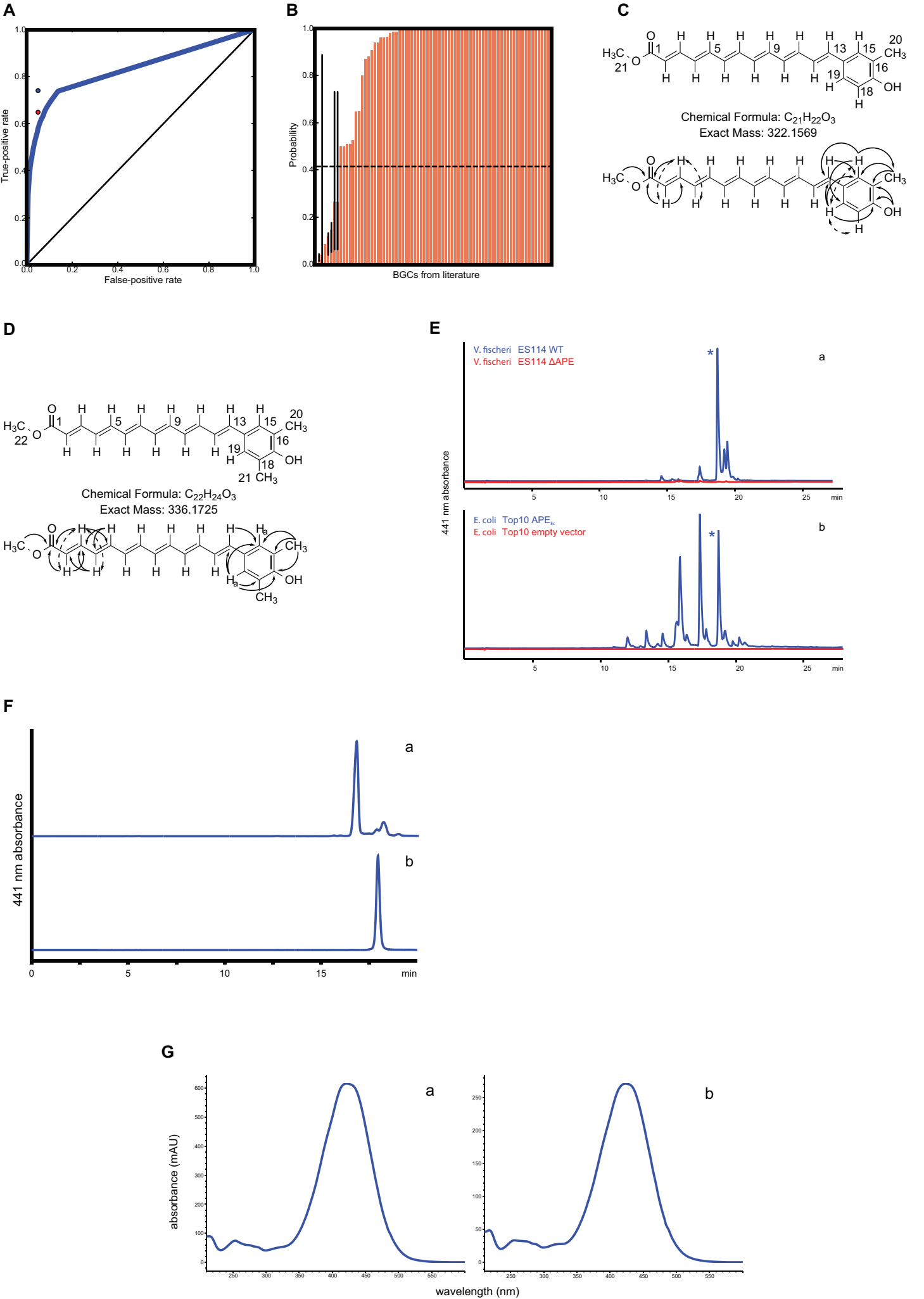
SI Figure 5



Supplemental Figure

SI Figure 6




A
True-positive rate
False-positive rate
B
Probability
BGCs from literature
C
Chemical Formula: C21H22O3
Exact Mass: 322.1569
D
Chemical Formula: C22H24O3
Exact Mass: 336.1725
E
V. fischeri ES114 WT
V. fischeri ES114 ΔAPE
E. coli Top10 APEEc
E. coli Top10 empty vector
441 nm absorbance
min
F
441 nm absorbance
min
G
absorbance (mAU)
wavelength (nm)