Cell

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

--Manuscript Draft--

Manuscript Number:	
Full Title:	Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters
Article Type:	Research Article
Keywords:	biosynthetic gene cluster; secondary metabolite; natural product; aryl polyene
Corresponding Author:	Michael Fischbach San Francisco, CA UNITED STATES
First Author:	Peter Cimermancic
Order of Authors:	Peter Cimermancic
	Marnix Medema
	Jan Claesen
	Kenji Kurita
	Laura Wieland Brown
	Konstantinos Mavrommatis
	Amrita Pati
	Paul Godfrey
	Michael Koehrsen
	Jon Clardy
	Bruce Birren
	Eriko Takano
	Andrej Sali
	Roger Linington
	Michael Fischbach
Abstract:	Although biosynthetic gene clusters (BGCs) have been discovered for hundreds of bacterial metabolites, our knowledge of their number and phylogenetic distribution remains limited. Here, we report the systematic identification of BGCs from across the prokaryotic tree of life, using a novel algorithm that was designed to detect BGCs of both known and unknown types. Network analysis of the resulting set of predicted BGCs indicated the existence of large and taxonomically widespread gene cluster families without characterized members. We experimentally characterized the most prominent family, consisting of two subfamilies of hundreds of BGCs distributed throughout the Proteobacteria. Unexpectedly, we found that they encode aryl polyenes, revealing a distant relationship to a third subfamily of aryl polyene BGCs; together, these three subfamilies constitute the largest known BGC family (>1000 clusters). Although these clusters are widely divergent in sequence, their small molecule products are highly conserved, suggesting they play an important role in Gram-negative cell biology.
Opposed Reviewers:	
Suggested Reviewers:	Joern Piel jpiel@ethz.ch
	Yi Tang

	yitang@ucla.edu
	Chaitan Khosla khosla@stanford.edu

Cover Letter (required for ALL submissions)

University of California
San Francisco
Schools of Pharmacy and Medicine
Bioengineering and Therapeutic Sciences

Michael Fischbach, PhD
Assistant Professor

UCSF, Mission Bay
Byers Hall, Room 308C
1700 4th Street
San Francisco, CA 94158

Tel: (415) 514-9435
Fax: (415) 514-9736
fischbach@fischbachgroup.org
http://www.fischbachgroup.org

Robert Kruger
Deputy Editor, Cell

February 26, 2014

Dear Robert:

Please find enclosed a manuscript, “Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters,” for consideration as a Research Article in Cell.

While biosynthetic gene clusters have been identified for hundreds of bacterial metabolites, genome mining has largely been an ad hoc pursuit that is motivated by interest in specific molecules or organisms rather than by a global view of biosynthesis. Here, we take the first step toward addressing that challenge, by systematically identifying and analyzing the biosynthetic gene clusters from across the prokaryotic tree of life. Our analysis yields two major findings:

1) The global landscape of biosynthetic gene clusters looks very different from the subset that has been well studied.

About 70% of the 10,000 gene clusters identified with our newly introduced algorithm fall outside the known classes detectable using the best existing computational tools. Surprisingly, we find that saccharides are the largest class of metabolites (46% of the ~10,000 putative gene clusters) and are vastly understudied relative to other classes (they comprise ~13% of gene clusters for known molecules). Saccharides are known to play an outsize role in microbe-host interactions (e.g., LPS, capsular polysaccharide, and *Bacteroides* polysaccharide A), and our results point to a great number and diversity of unmined saccharide gene clusters from almost every organism in the database.

By analyzing our predicted gene clusters in a similarity network together with known gene clusters, we discovered a wide range of gene cluster families that have not yet been characterized. These provide exciting opportunities to understand which classes of unknown metabolites play important roles in nature, and to identify novel classes of natural products in the future.

2) The largest uncharacterized family consists of >1000 aryl polyene gene clusters widely distributed among Gram-negative bacteria.

In order to determine which widely distributed classes of metabolites may have remained unrecognized by microbiologists, we zoomed in on the largest family of unknown gene clusters in our results set (two related subfamilies). From each subfamily, we selected a gene cluster (one from *E. coli* CFT073, the other from *V. fischeri* ES114), expressed it heterologously in *E. coli*, and solved the

chemical structure of its product; we also knocked out the *V. fischeri* cluster in its native host. In spite of the limited degree of homology between the clusters, their small molecule products are remarkably similar. Notably, the compounds also have a similar structure to two previously described pigments, xanthomonadin and flexirubin, yellow pigments whose biosynthetic genes are members of a separate, distantly related subfamily in our data set. Together, these three families constitute the most widely distributed gene cluster superfamily in the sequence database (>1000 clusters). Remarkably, the aryl polyenes are even more widely distributed than the well-known carotenoids (~870 clusters), and occur in many well-known genera such as *Escherichia*, *Vibrio*, *Pseudomonas*, *Burkholderia*, *Shewanella*, *Geobacter* and *Bacteroides*.

Overall, our data reveal a first glimpse into the thus far unexplored ‘dark matter’ of microbial secondary metabolism, which comprises numerous novel families of gene clusters and metabolites throughout a diverse range of organisms. Thus, our study paves the way towards a new understanding of the roles of microbial secondary metabolites in nature and towards their effective mining in the hunt for novel natural products.

We look forward to hearing from you after you have had a chance to consider the manuscript.

Sincerely,

Michael A. Fischbach

SUMMARY

Although biosynthetic gene clusters (BGCs) have been discovered for hundreds of bacterial metabolites, our knowledge of their number and phylogenetic distribution remains limited. Here, we report the systematic identification of BGCs from across the prokaryotic tree of life, using a novel algorithm that was designed to detect BGCs of both known and unknown types. Network analysis of the resulting set of predicted BGCs indicated the existence of large and taxonomically widespread gene cluster families without characterized members. We experimentally characterized the most prominent family, consisting of two subfamilies of hundreds of BGCs distributed throughout the Proteobacteria. Unexpectedly, we found that they encode aryl polyenes, revealing a distant relationship to a third subfamily of aryl polyene BGCs; together, these three subfamilies constitute the largest known BGC family (>1000 clusters). Although these clusters are widely divergent in sequence, their small molecule products are highly conserved, suggesting they play an important role in Gram-negative cell biology.

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

Contact: fischbach@fischbachgroup.org

*Denotes equal contribution

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

SUMMARY

Although biosynthetic gene clusters (BGCs) have been discovered for hundreds of bacterial metabolites, our knowledge of their number and phylogenetic distribution remains limited. Here, we report the systematic identification of BGCs from across the prokaryotic tree of life, using a novel algorithm that was designed to detect BGCs of both known and unknown types. Network analysis of the resulting set of predicted BGCs indicated the existence of large and taxonomically widespread gene cluster families without characterized members. We experimentally characterized the most prominent family, consisting of two subfamilies of hundreds of BGCs distributed throughout the Proteobacteria. Unexpectedly, we found that they encode aryl polyenes, revealing a distant relationship to a third subfamily of aryl polyene BGCs; together, these three subfamilies constitute the largest known BGC family (>1000 clusters). Although these clusters are widely divergent in sequence, their small molecule products are highly conserved, suggesting they play an important role in Gram-negative cell biology.

HIGHLIGHTS

- The ClusterFinder algorithm detects BGCs of both known and unknown classes
- There exist large and widely distributed BGC families with no characterized members
- We show that the most prominent family encodes the biosynthesis of aryl polyenes
- The aryl polyene clusters constitute the largest known family of BGCs

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., 2012; Nguyen et al., 2008; Oliynyk 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.

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 (**SI Table I**). 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**, **Methods**). 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 relatively little training set bias and is capable of identifying new classes of gene clusters effectively (See **Methods** 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 BCGs (with an estimated false-positive rate of 5%) in 1,154 genomes of organisms throughout the prokaryotic tree of life (**Figure 1c-d, SI Text 1**), which we subjected to an extensive phylogenomic analysis (**SI Text 2-3, SI Figures 1-6, SI Tables I-II**). 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 (**SI Text 4**). 93% 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 (**SI Figure 7**). 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, **SI Figure 8**). 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 multi-domain 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, SI Text 5, SI Figures 9-10**) 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. 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 thioesterase domains, as well as an outer membrane lipoprotein carrier protein

and an MMPL family transporter (**Figure 3a, SI Figures 11-12**). These clusters are found in a wide variety of Gammaproteobacteria (*Acinetobacter, Aggregatibacter*, *Escherichia*, *Klebsiella*, *Pantoea, Pseudoalteromonas, 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 (**SI Figure 24**). 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 (**Figure 3b, SI Figures 21, 23-25**).

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** 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 (**Figure 3b**, **SI Figures 22-25**). 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**). 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 **SI Figure 13**). 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 subfamily 3 and subfamilies 1 and 2. This suggests that the APE clusters

share a common ancestor and therefore comprise a single family of >1000 gene clusters (**Figure 3a**). This finding is further supported by a maximum-likelihood phylogenetic analysis of the ketosynthase and adenylation enzyme superfamilies based on structure-guided multiple sequence alignments (**SI Text 6, SI Figures 14-16**): the APE KS and A enzymes cluster together in separate uncharacterized clades that are only distantly related to all other known family members. 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 **SI Table III**).

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 bacteria" (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 (**Figure 4, SI Figure 17**). 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.

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 ~6,000 **(SI Figure 18)**, less than half of which are identified in our current set of genomes (~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 (**SI Figure 19**). Even among cultivated organisms, there are still many underexplored taxa (Letzel et al., 2012) (**SI Text 2**). 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 1154 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-BCG 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 > 1e-10). ClusterFinder source code is available from the GitHub repository (https://github.com/petercim/ClusterFinder).

ClusterFinder validation

The algorithm was validated in three ways. First, its output was compared to 10 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 (**SI Table IV**), 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 (**SI Table V**), 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 multi-domain 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 (08/2013), with a 20% sequence identity cut-off and 2000 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 Lin modified 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 MEGA5 (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 **SI Tables VI-VIII**. 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 tri-parental mating and integrants were identified by selection for kanamycin resistance. Second homologous recombination events were enriched by non-selective 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**p 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 APE_{EC} and APE_{VF}, 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 hour. 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 (**SI Figures 24-25**).

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: $C_{21}H_{22}O_3$ for APE_{EC} ($[M-H]^-$ adduct at 321.1496 *m/z* (Δppm = -0.310)) and $C_{22}H_{24}O_3$ for APE_{VF} ($[M-H]^-$ adduct at 335.1652 *m/z* (Δppm = 0.0)). Analysis of the ^{1}H-NMR, COSY, HSQC, HMBC, ROESY and TOCSY spectra of APE_{VF} in D_6 DMSO and APE_{EC} in D_6 acetone enabled the determination of their solution structure (**Figure 3b**). This procedure is described in detail in the supplementary methods section and shown in **SI Figures 21-23**.

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

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 Edward Ruby (University of Wisconsin) for providing us with *V. fischeri* ES114, Didier Mazel (Institut Pasteur) for plasmid pSW8197, and Mervyn Bibb (John Innes Centre) for plasmids pIJ773, pIJ790 and pIJ10257. This work was supported by an HHMI Predoctoral Fellowship (PC), a Boehringer Ingelheim Fonds travel grant (MHM), Grant 10463 from the GenBiotics programme of the Dutch Technology Foundation STW to ET (MHM), an NWO-Vidi fellowship (RB), NIH grant TW006634 (RGL), the James and Eleanor Delfino Charitable Trust (KK), a Medical Research Program Grant from the W.M. Keck Foundation (MAF), a Fellowship for Science and Engineering from the David and Lucile Packard Foundation (MAF), DARPA award HR0011-12-C-0067 (MAF), and NIH grants OD007290, AI101018, AI101722 and GM081879 (MAF). This project has been funded in part with Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Contract No.: HHSN272200900018C. M.A.F. is on the scientific advisory board of Warp Drive Bio.

REFERENCES

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.

Barbe, V., Vallenet, D., Fonknechten, N., Kreimeyer, A., Oztas, S., Labarre, L., Cruveiller, S., Robert, C., Duprat, S., Wincker, P.*, et al.* (2004). Unique features revealed by the genome sequence of Acinetobacter sp. ADP1, a versatile and naturally transformation competent bacterium. Nucleic Acids Res *32*, 5766-5779.

Bergmann, S., Schumann, J., Scherlach, K., Lange, C., Brakhage, A.A., and Hertweck, C. (2007). Genomics-driven discovery of PKS-NRPS hybrid metabolites from Aspergillus nidulans. Nat Chem Biol *3*, 213-217.

Blin, K., Medema, M.H., Kazempour, D., Fischbach, M.A., Breitling, R., Takano, E., and Weber, T. (2013). antiSMASH 2.0--a versatile platform for genome mining of secondary metabolite producers. Nucleic Acids Res *41*, W204-212.

Challis, G.L. (2008). Genome mining for novel natural product discovery. J Med Chem *51*, 2618-2628.

Dong, T., and Schellhorn, H.E. (2009). Global effect of RpoS on gene expression in pathogenic Escherichia coli O157:H7 strain EDL933. BMC Genomics *10*, 349.

Fautz, E., and Reichenbach, H. (1979). Biosynthesis of flexirubin: Incorporation of precursors by the bacterium Flexibacter elegans. Phytochemistry *18*, 957-959.

Fischbach, M.A., Walsh, C.T., and Clardy, J. (2008). The evolution of gene collectives: How natural selection drives chemical innovation. Proceedings of the National Academy of Sciences of the United States of America *105*, 4601-4608.

Flatt, P.M., and Mahmud, T. (2007). Biosynthesis of aminocyclitol-aminoglycoside antibiotics and related compounds. Nat Prod Rep *24*, 358-392.

Franke, J., Ishida, K., and Hertweck, C. (2012). Genomics-driven discovery of burkholderic acid, a noncanonical, cryptic polyketide from human pathogenic Burkholderia species. Angew Chem Int Ed Engl *51*, 11611-11615.

Frasch, H.J., Medema, M.H., Takano, E., and Breitling, R. (2013). Design-based re-engineering of biosynthetic gene clusters: plug-and-play in practice. Curr Opin Biotechnol.

Freeman, M.F., Gurgui, C., Helf, M.J., Morinaka, B.I., Uria, A.R., Oldham, N.J., Sahl, H.G., Matsunaga, S., and Piel, J. (2012). Metagenome mining reveals polytheonamides as posttranslationally modified ribosomal peptides. Science *338*, 387-390.

Fuchs, S.W., Bozhuyuk, K.A., Kresovic, D., Grundmann, F., Dill, V., Brachmann, A.O., Waterfield, N.R., and Bode, H.B. (2013). Formation of 1,3-cyclohexanediones and resorcinols catalyzed by a widely occurring ketosynthase. Angew Chem Int Ed Engl *52*, 4108-4112.

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.

Goel, A.K., Rajagopal, L., Nagesh, N., and Sonti, R.V. (2002). Genetic locus encoding functions involved in biosynthesis and outer membrane localization of xanthomonadin in Xanthomonas oryzae pv. oryzae. J Bacteriol *184*, 3539-3548.

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. Advances in applied microbiology *54*, 107-128.

Hassan, K.A., Johnson, A., Shaffer, B.T., Ren, Q., Kidarsa, T.A., Elbourne, L.D., Hartney, S., Duboy, R., Goebel, N.C., Zabriskie, T.M.*, et al.* (2010). Inactivation of the GacA response regulator in Pseudomonas fluorescens Pf-5 has far-reaching transcriptomic consequences. Environ Microbiol *12*, 899-915.

Jenkins, C.L., and Starr, M.P. (1982). The pigment of Xanthomonas populi is a nonbrominated aryl-heptaene belonging to xanthomonadin pigment group 11. Current Microbiology *7*, 195-198.

Kadioglu, A., Weiser, J.N., Paton, J.C., and Andrew, P.W. (2008). The role of Streptococcus pneumoniae virulence factors in host respiratory colonization and disease. Nat Rev Microbiol *6*, 288-301.

Kersten, R.D., Yang, Y.L., Xu, Y., Cimermancic, P., Nam, S.J., Fenical, W., Fischbach, M.A., Moore, B.S., and Dorrestein, P.C. (2011). A mass spectrometry-guided genome mining approach for natural product peptidogenomics. Nat Chem Biol *7*, 794-802.

Kersten, R.D., Ziemert, N., Gonzalez, D.J., Duggan, B.M., Nizet, V., Dorrestein, P.C., and Moore, B.S. (2013). Glycogenomics as a mass spectrometry-guided genome-mining method for microbial glycosylated molecules. Proc Natl Acad Sci U S A.

Khaldi, N., Seifuddin, F.T., Turner, G., Haft, D., Nierman, W.C., Wolfe, K.H., and Fedorova, N.D. (2010). SMURF: Genomic mapping of fungal secondary metabolite clusters. Fungal Genet Biol *47*, 736-741.

Laureti, L., Song, L., Huang, S., Corre, C., Leblond, P., Challis, G.L., and Aigle, B. (2011). Identification of a bioactive 51-membered macrolide complex by activation of a silent polyketide synthase in Streptomyces ambofaciens. Proc Natl Acad Sci U S A *108*, 6258-6263.

Lautru, S., Deeth, R.J., Bailey, L.M., and Challis, G.L. (2005). Discovery of a new peptide natural product by Streptomyces coelicolor genome mining. Nat Chem Biol *1*, 265-269.

Letzel, A.C., Pidot, S.J., and Hertweck, C. (2012). 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, M.H., Ung, P.M., Zajkowski, J., Garneau-Tsodikova, S., and Sherman, D.H. (2009). Automated genome mining for natural products. BMC Bioinformatics *10*, 185.

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

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. Nucleic Acids Res *40*, D115-122.

Mazmanian, S.K., Liu, C.H., Tzianabos, A.O., and Kasper, D.L. (2005). An immunomodulatory molecule of symbiotic bacteria directs maturation of the host immune system. Cell *122*, 107-118.

Mazmanian, S.K., Round, J.L., and Kasper, D.L. (2008). A microbial symbiosis factor prevents intestinal inflammatory disease. Nature *453*, 620-625.

McBride, M.J., Xie, G., Martens, E.C., Lapidus, A., Henrissat, B., Rhodes, R.G., Goltsman, E., Wang, W., Xu, J., Hunnicutt, D.W.*, et al.* (2009). Novel features of the polysaccharide-digesting gliding bacterium Flavobacterium johnsoniae as revealed by genome sequence analysis. Appl Environ Microbiol *75*, 6864-6875.

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. Nucleic Acids Res *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.

Nguyen, T., Ishida, K., Jenke-Kodama, H., Dittmann, E., Gurgui, C., Hochmuth, T., Taudien, S., Platzer, M., Hertweck, C., and Piel, J. (2008). Exploiting the mosaic structure of trans-acyltransferase polyketide synthases for natural product discovery and pathway dissection. Nat Biotechnol *26*, 225-233.

Oliynyk, M., Samborskyy, M., Lester, J.B., Mironenko, T., Scott, N., Dickens, S., Haydock, S.F., and Leadlay, P.F. (2007). Complete genome sequence of the erythromycin-producing bacterium Saccharopolyspora erythraea NRRL23338. Nat Biotechnol *25*, 447-453.

Park, B.S., Song, D.H., Kim, H.M., Choi, B.S., Lee, H., and Lee, J.O. (2009). The structural basis of lipopolysaccharide recognition by the TLR4-MD-2 complex. Nature *458*, 1191-1195.

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.

Quan, J., and Tian, J. (2009). Circular polymerase extension cloning of complex gene libraries and pathways. PLoS One *4*, e6441.

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.

Reichenbach, H., Kohl, W., Bottger-Vetter, A., and Achenbach, H. (1980). Flexirubin-type pigments in Flavobacterium. Archives of Microbiology *126*, 291-293.

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.

Schneiker, S., Perlova, O., Kaiser, O., Gerth, K., Alici, A., Altmeyer, M.O., Bartels, D., Bekel, T., Beyer, S., Bode, E.*, et al.* (2007). Complete genome sequence of the myxobacterium Sorangium cellulosum. Nat Biotechnol *25*, 1281-1289.

Segata, N., Bornigen, D., Morgan, X.C., and Huttenhower, C. (2013). PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. Nature communications *4*, 2304.

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.

Starcevic, A., Zucko, J., Simunkovic, J., Long, P.F., Cullum, J., and Hranueli, D. (2008). ClustScan: an integrated program package for the semi-automatic annotation of modular biosynthetic gene clusters and in silico prediction of novel chemical structures. Nucleic Acids Res *36*, 6882-6892.

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.

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.

Walsh, C.T., and Fischbach, M.A. (2010). Natural products version 2.0: connecting genes to molecules. J Am Chem Soc *132*, 2469-2493.

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.

Weber, T., Rausch, C., Lopez, P., Hoof, I., Gaykova, V., Huson, D.H., and Wohlleben, W. (2009). CLUSEAN: a computer-based framework for the automated analysis of bacterial secondary metabolite biosynthetic gene clusters. J Biotechnol *140*, 13-17.

Weitnauer, G., Muhlenweg, A., Trefzer, A., Hoffmeister, D., Sussmuth, R.D., Jung, G., Welzel, K., Vente, A., Girreser, U., and Bechthold, A. (2001). Biosynthesis of the orthosomycin antibiotic avilamycin A: deductions from the molecular analysis of the avi biosynthetic gene cluster of Streptomyces viridochromogenes Tu57 and production of new antibiotics. Chem Biol *8*, 569-581.

Winter, J.M., Behnken, S., and Hertweck, C. (2011). Genomics-inspired discovery of natural products. Curr Opin Chem Biol *15*, 22-31.

FIGURE LEGENDS

Figure 1. ClusterFinder flowchart and distribution of BGC classes and counts. a, Flowchart of the four-step BGC prediction pipeline: (*i*) annotation of a genome sequence and compression to a string of Pfam domains, (*ii*) calculation of posterior probabilities of a BGC hidden state, (*iii*) clustering of genes that contain Pfam domain(s) with posterior probabilities of BGC hidden state above the threshold, and (*iv*) annotation of the predicted BGCs using an expanded version of the antiSMASH algorithm. **b,** Distribution of BGC classes for known (inset) and predicted BGCs. “Other” gene clusters include gene clusters from other known classes as well as a manually curated set of 1,024 putative gene clusters that fall outside known biosynthetic classes. Unexpectedly, 40% of all predicted BGCs encode saccharides, more than twice the size of the next largest class. **c,** Number of predicted BGCs by genome size. Most bacterial species follow a linear trend (the equation in the bottom-right corner); outliers (defined as having residuals >8) are colored red. **d,** The proportions of bacterial genomes devoted to secondary metabolite biosynthesis (left panel; 6.7% of species that devote >7.5% of their genome to biosynthesis are marked red), transcription (middle panel), and translation (right panel).

Figure 2. A systematic analysis of bacterial BGCs. Similarity network of known and putative BGCs, with the BGC similarity metric threshold at 0.5. The topology of the network is robust to changes in the distance threshold, as described in the Methods. One connected component harbors most of the gene clusters (72%), and is largely composed of two linked subgraphs: one dominated by oligosaccharides and the other a mixture of nonribosomal peptides (NRPs) and polyketides/lipids, indicating that BGCs from these classes share a significant number of gene families with one another. A selection of node clusters within the network has been highlighted to show how gene cluster families form cliques within the network. The highlighted groups

include widely distributed gene cluster families for O-antigens, capsular polysaccharides, carotenoids, and NRPS-independent siderophores, along with one of the lantibiotic BGC families and an unknown family of BGCs with type III polyketide synthases. The aryl polyene family that we characterized further in this study is shown in the middle of the network.

Figure 3. APE gene clusters comprise the largest known BGC family. a, Heat map and dendrogram of all 1,021 detected APE family gene clusters, based on Clusters of Orthologous Groups generated by OrthoMCL (Li et al., 2003) using our adapted version of the Lin distance metric (Lin et al., 2006) that includes sequence similarity. Light grey indicates the presence of one gene from a COG, whereas darker grey tones indicate the presence of two or three genes from a COG. The two BGC subfamilies that functioned as the starting point of our analysis (subfamilies 1 and 2) are shown in green and red, respectively, while the smaller BGC subfamily that includes the xanthomonadin and flexirubin gene clusters (subfamily 3) is shown in blue. The positions of the two experimentally targeted gene clusters (*Ec* for *Escherichia coli* CFT073 and *Vf* for *Vibrio fischeri* ES114) as well as the *Xanthomonas campestris* ATCC 33913 xanthomonadin (*Xc*) and *Flavobacterium johnsonii* ATCC 17061 flexirubin (*Fj*) gene clusters are indicated below the heat map. **b,** Chemical structures obtained for the APE compounds from *E. coli* and *V. fischeri*, and the previously determined chemical structures of xanthomonadin and flexirubin. Note the difference in polyene acyl chain length as well as the distinct tailoring patterns on the aryl head groups. **c**, Bacterial pellets from strains harboring APE gene clusters showing the pigmentation conferred by aryl polyenes. **d,** Genetic architecture of the four characterized aryl polyene gene clusters. The inset in the *Flavobacterium johnsonii* flexirubin gene cluster is a sub-cluster putatively involved in the biosynthesis of dialkylresorcinol (Fuchs et al., 2013), which is acylated to an APE to form flexirubin.

Figure 4. APE gene clusters are widely but discontinuously distributed among Gram-negative bacteria. Presence/absence pattern of APE gene clusters across all complete genomes from selected bacterial genera, mapped onto the PhyloPhLan high-resolution phylogenetic tree (Segata et al., 2013). For each genus, the pie chart represents the percentage of sequenced genomes in which APE gene clusters are present (green) or absent (red). BGCs from the APE family occur throughout all subphyla of the Proteobacteria, as well as in a range of genera from the CFB group. The discontinuous presence/absence pattern suggests that gene cluster gain and/or loss has frequently occurred during evolution. A presence/absence mapping on all the genomes from our initial JGI dataset is provided in **SI Figure 17**.

Figure
Click here to download Figure: Figure1_v3.pdf

Figure 1

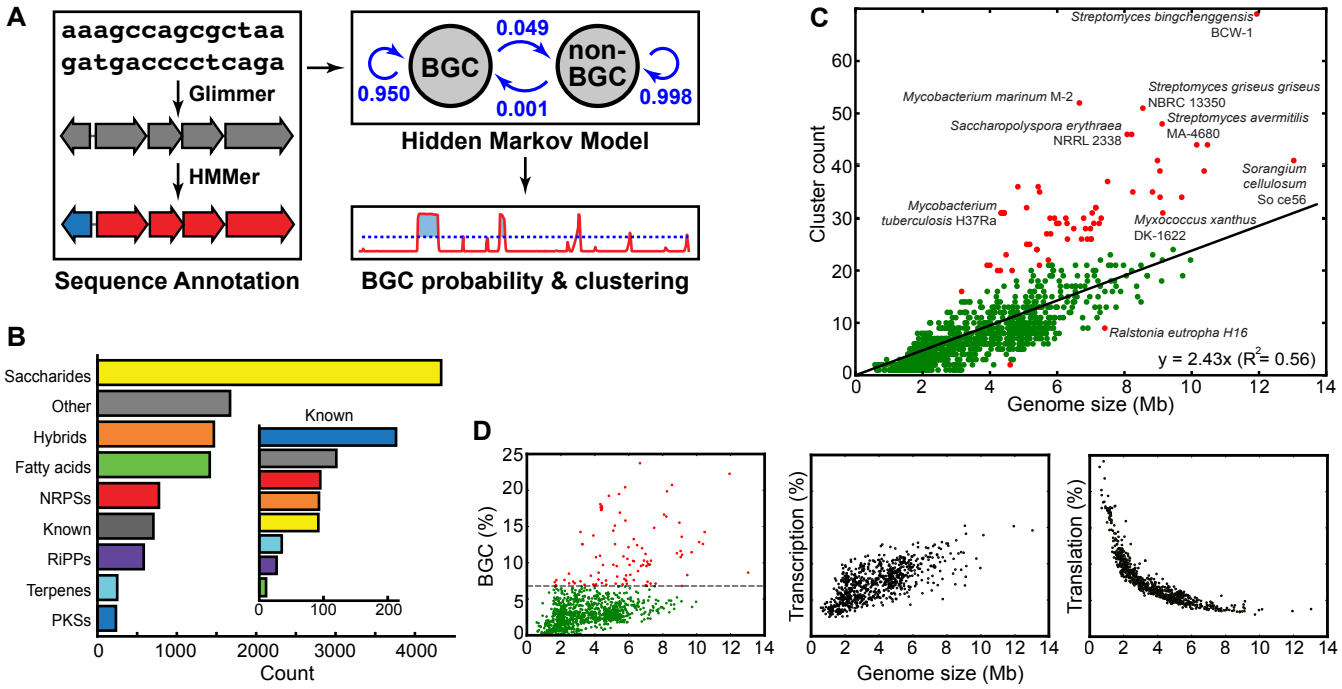


Figure
Click here to download Figure: Figure2_v10.pdf

Figure 2

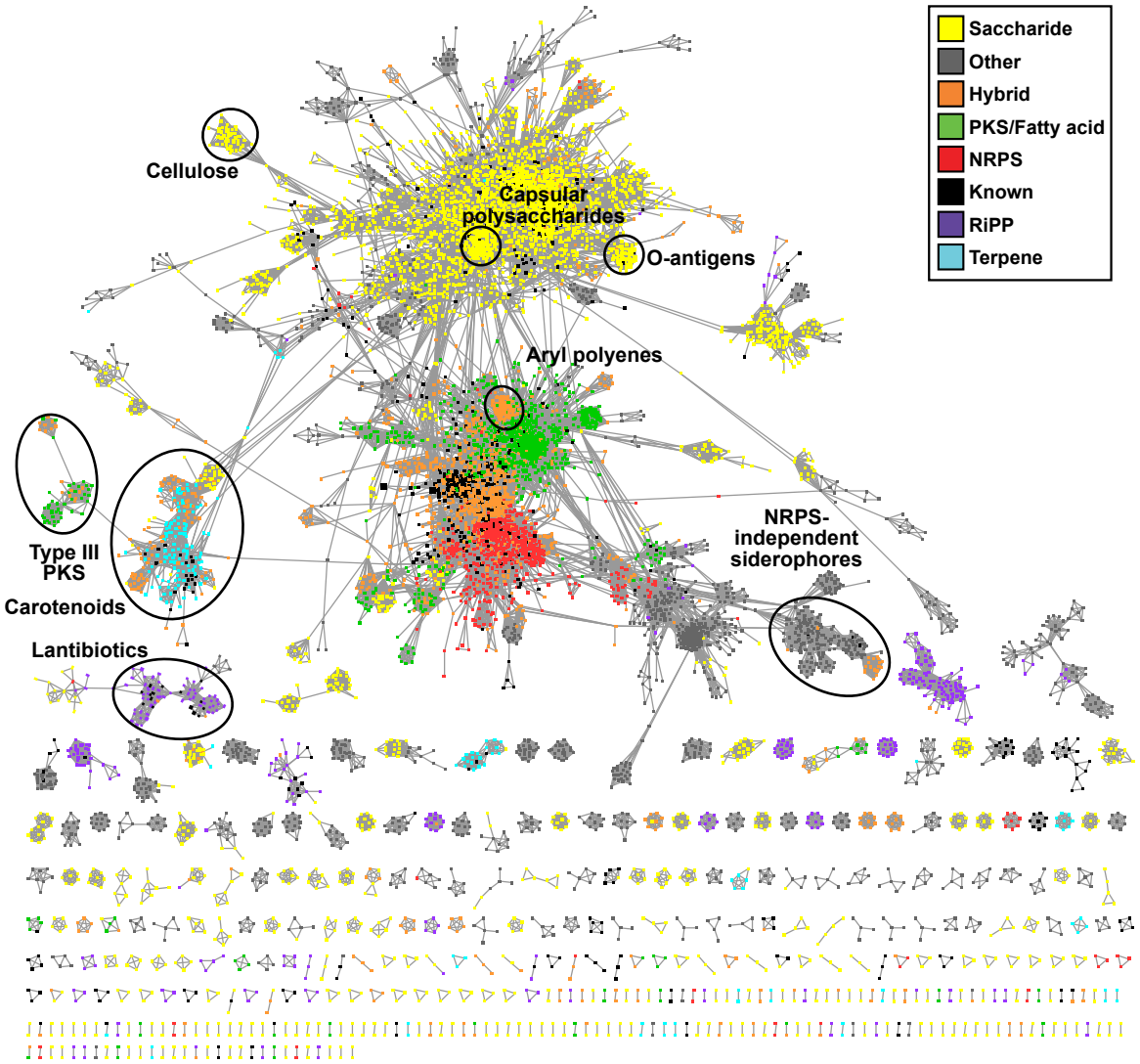
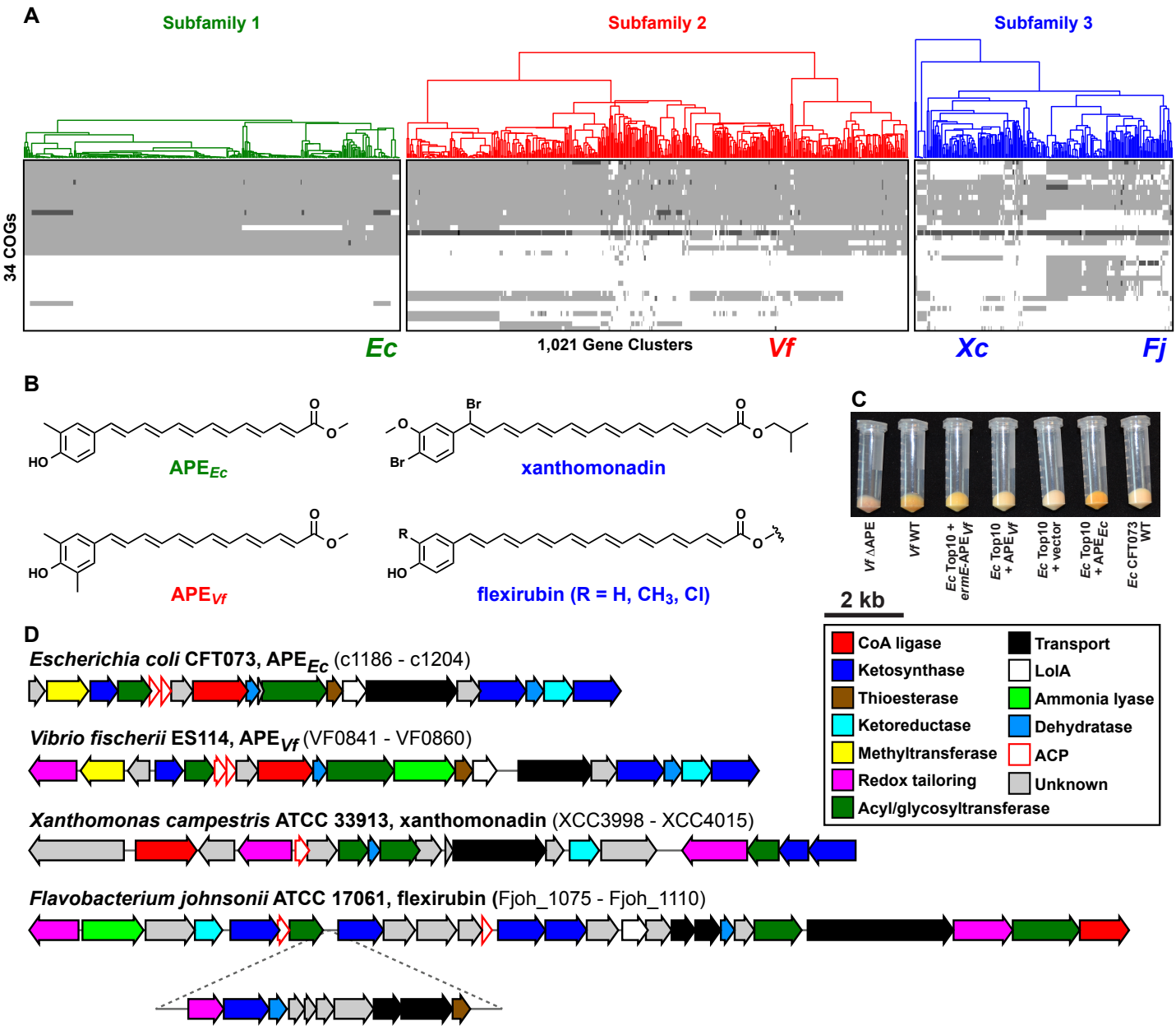


Figure
Click here to download Figure: Figure3_v15.pdf

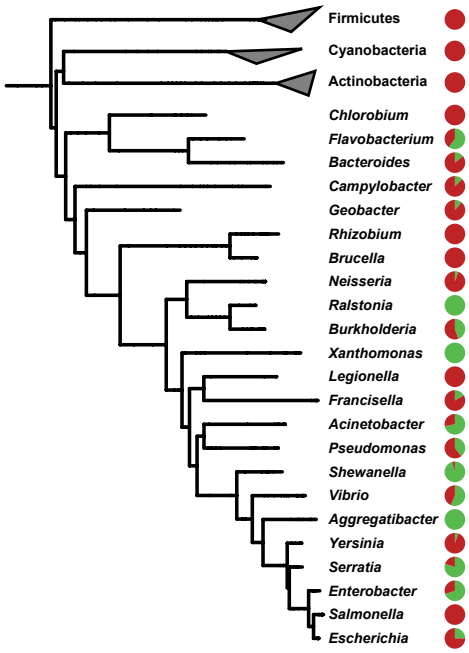
Figure 3



Figure

Click here to download Figure: Figure4_v3.pdf

Figure 4



Supplemental text:

Insights into secondary metabolism from a global analysis of 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

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).

SI Table I (in separate SI_Tables.XLSX file). 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.

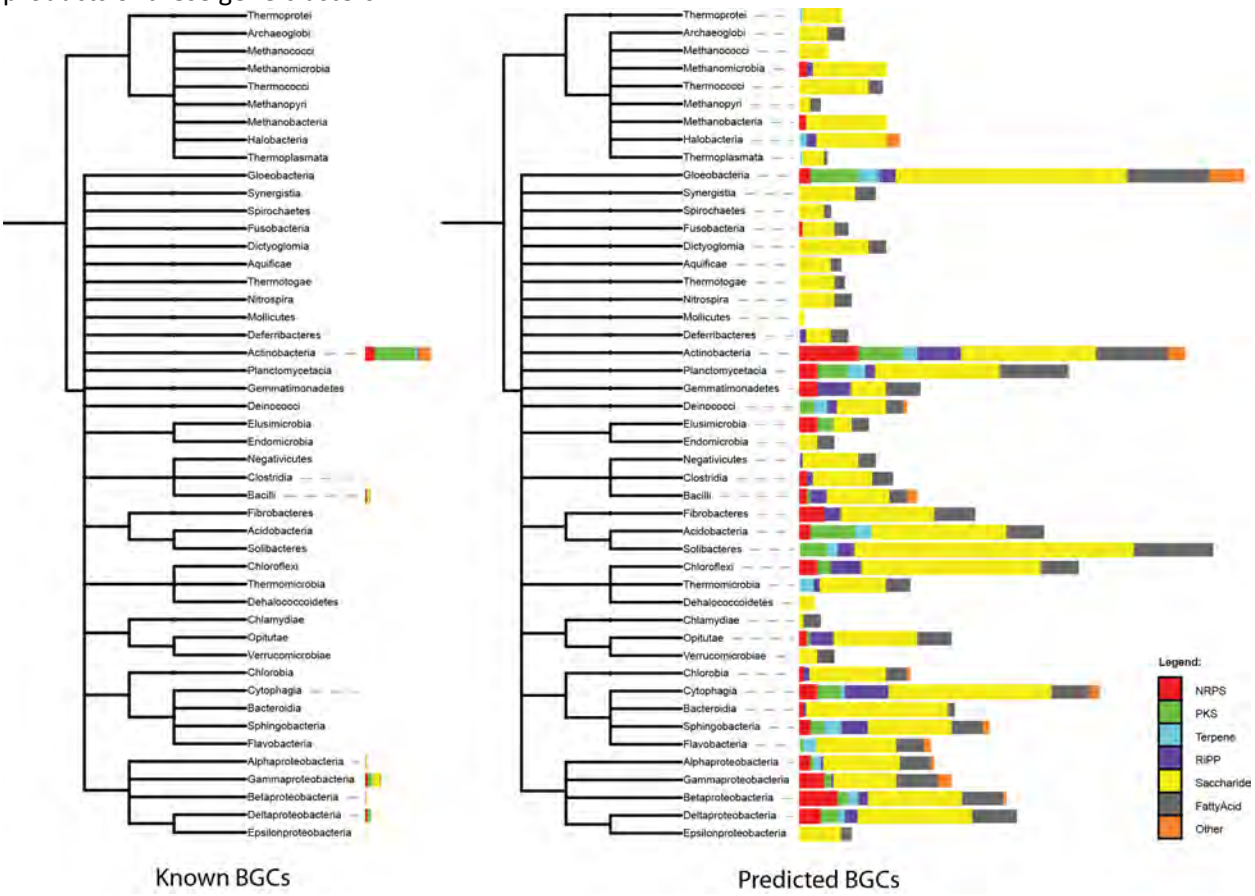
2. Prolific producers harbor exceptionally large complements of gene clusters

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 1**). 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

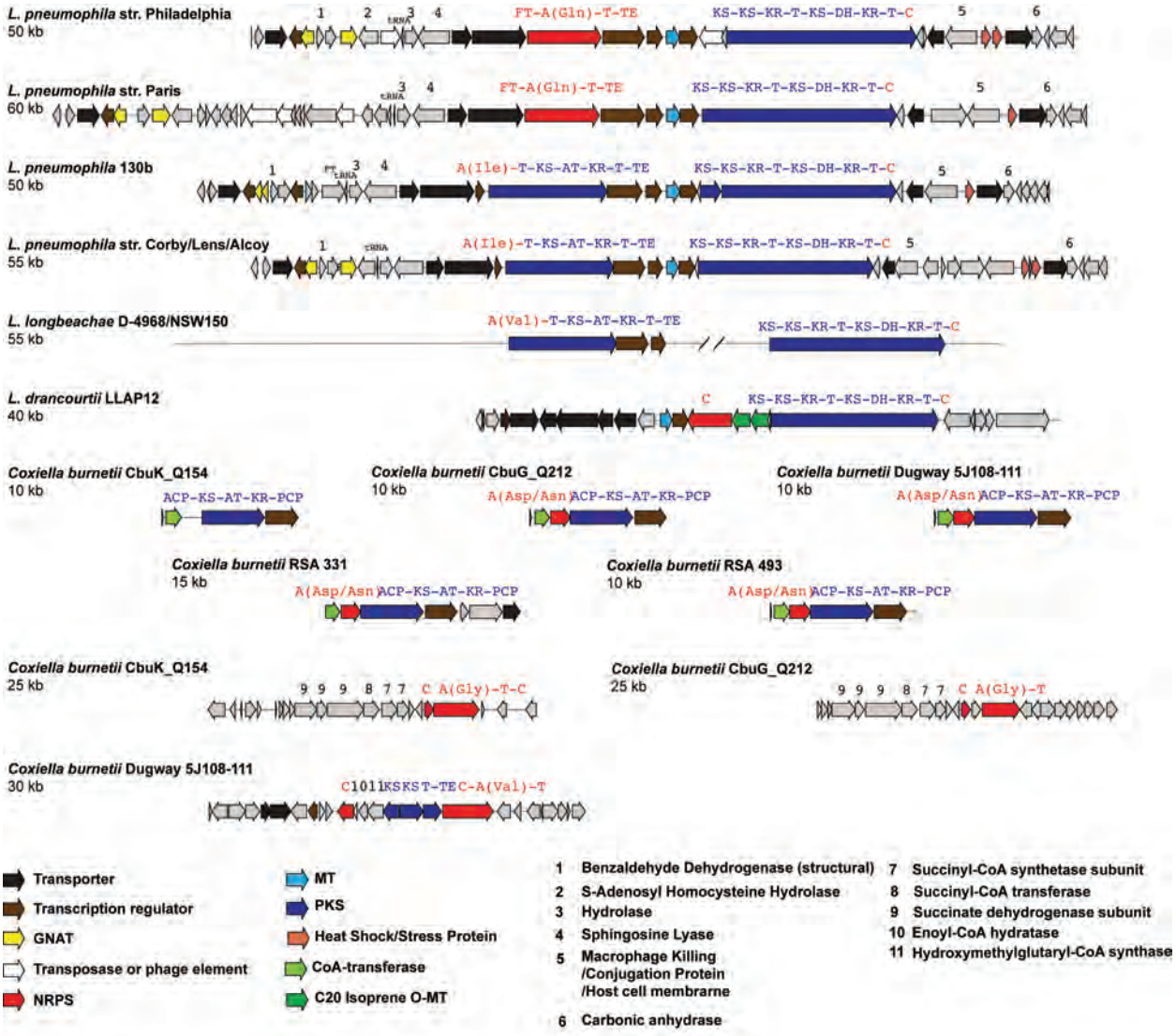
genome evolution (**SI Figure 2**), indicating a strong selective pressure to retain the small molecule products of these gene clusters.



SI Figure 1. 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.

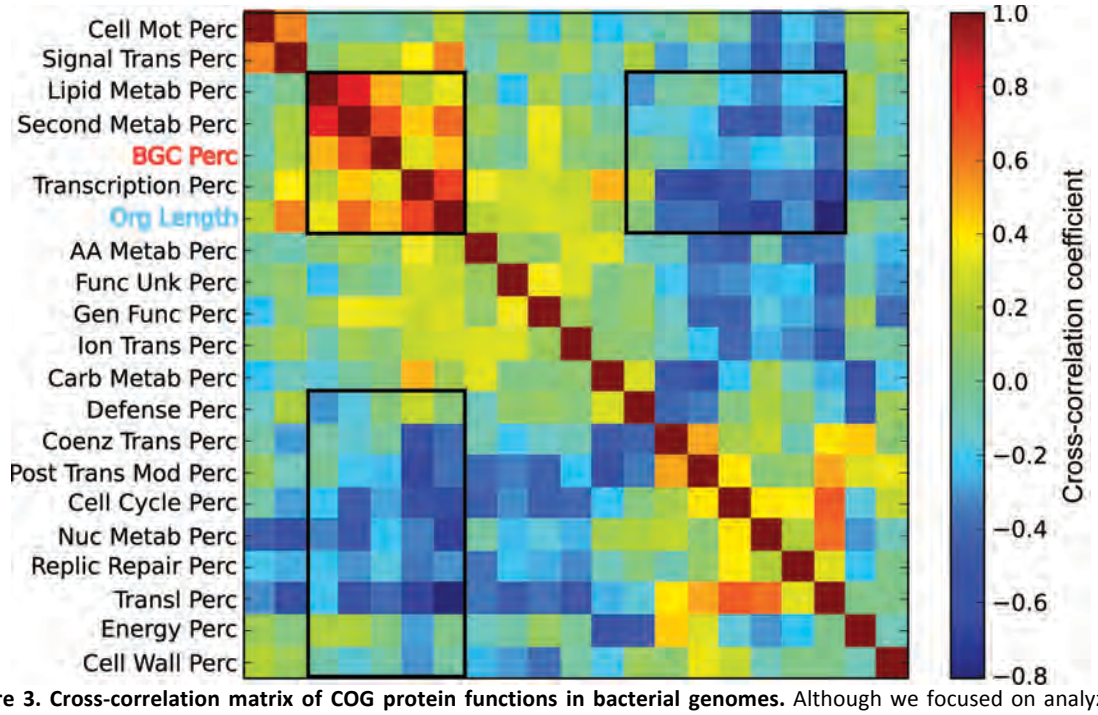
SI Table II (in separate SI_Tables.XLSX file). Overview of the four environmental metadata features that show the most significant differences between genomes, depending on how many BGCs are encoded in these genomes. P-values are calculated with the Kruskall-Wallis test.

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).



SI Figure 2. 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.

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 3**). Thus, it would appear that bacterial genomes expand largely to increase their gene complements for transcription, primary metabolism, and secondary metabolism.

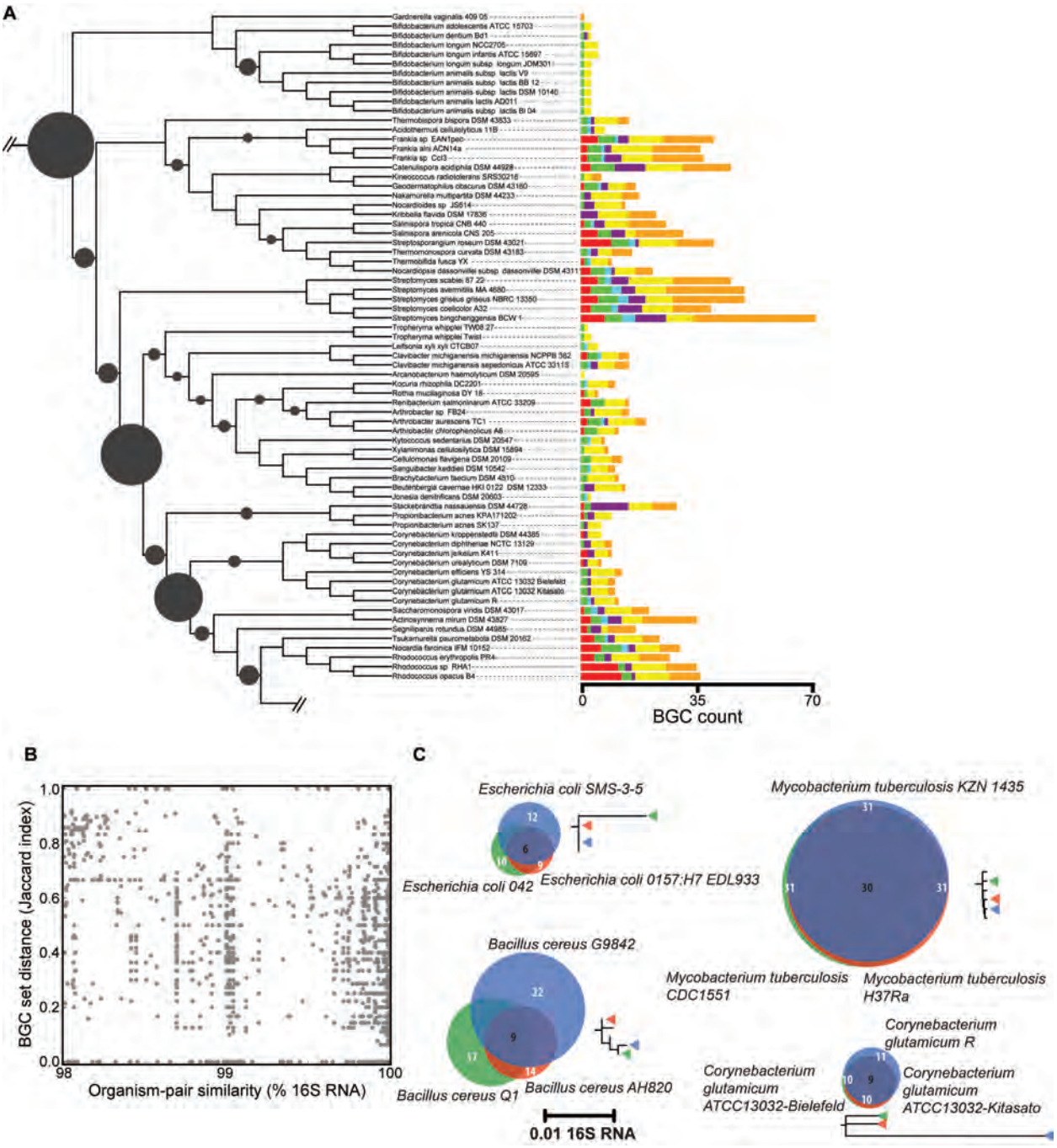


SI Figure 3. Cross-correlation matrix of COG protein functions in bacterial genomes. Although we focused on analyzing the association between the number of BGCs (or percentage of the genomes they occupy) and genome lengths (**Figure 1c**), we also investigated whether there are any other COG functions 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.

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(Ruckert et al., 2011; Tobias et al., 2013). However, it is not clear whether this phenomenon is general or specific to these taxa.

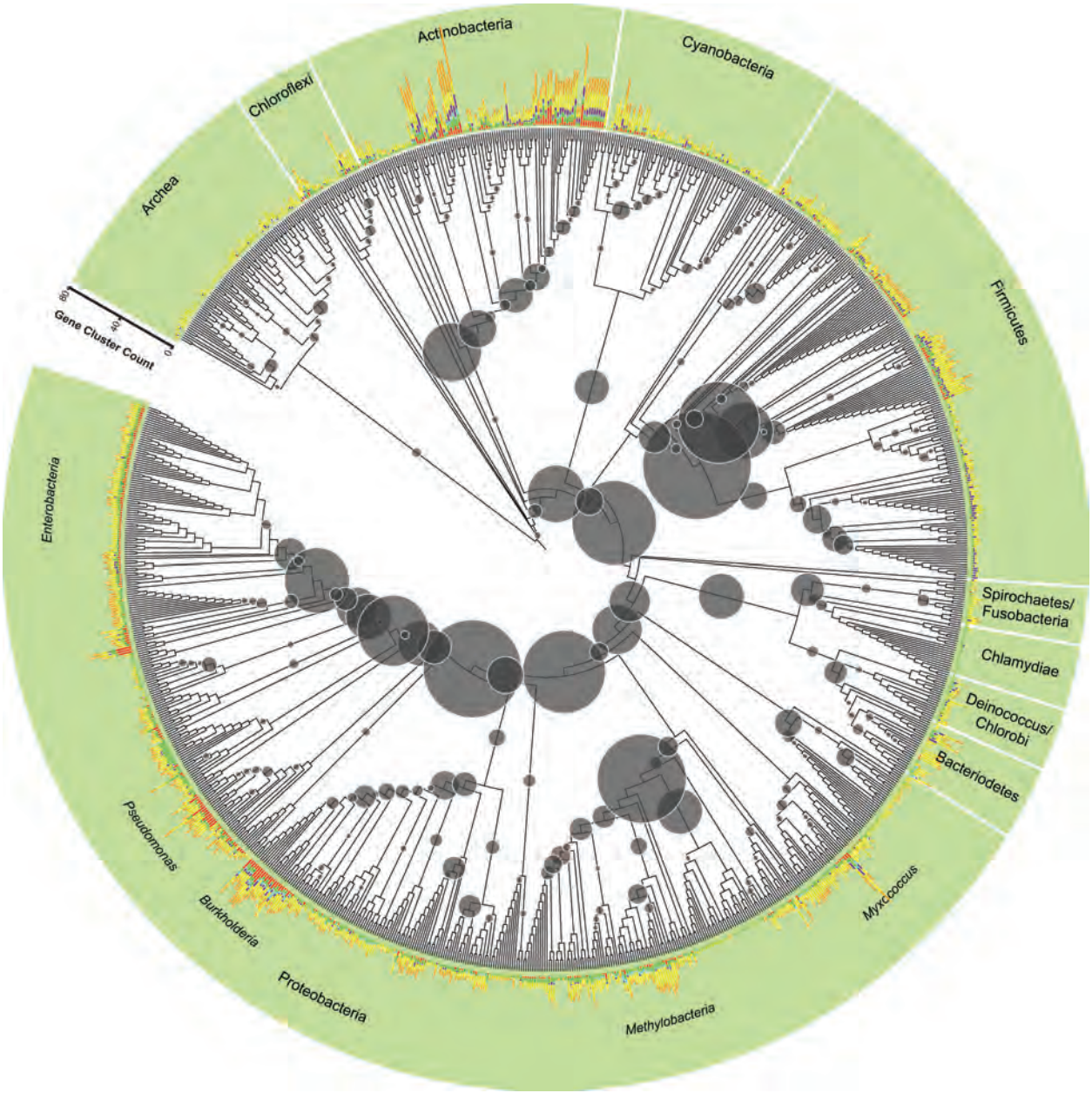
To answer this question, we used a quadratic entropy index to illustrate how the diversity of gene clusters can be decomposed among the nodes of the phylogenetic tree(Pavoine et al., 2010). This methodology allowed us to determine gene cluster diversity at internal nodes at different depths in the phylogeny (**SI Figure 4a & 5**). 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 4c**).



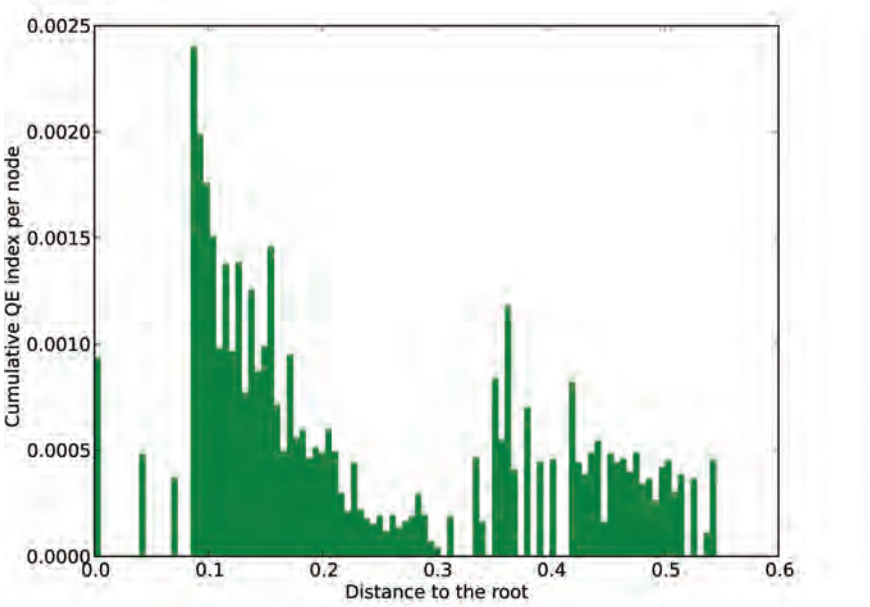
SI Figure 4. Diversity of BGCs is independent from the phylogeny. a, The decomposition of BGC diversity among species of the phylum Actinobacteria. The diversity of each node in the phylogenetic tree is measured by the quadratic entropy index, and represented by the size of the circle (larger circle 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 5**. **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.

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 4b**). 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).



SI Figure 5. Decomposition of BGC diversity among all sequenced prokaryotic genomes. 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**).



SI Figure 6. Histogram of cumulative 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.

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 6**), 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.

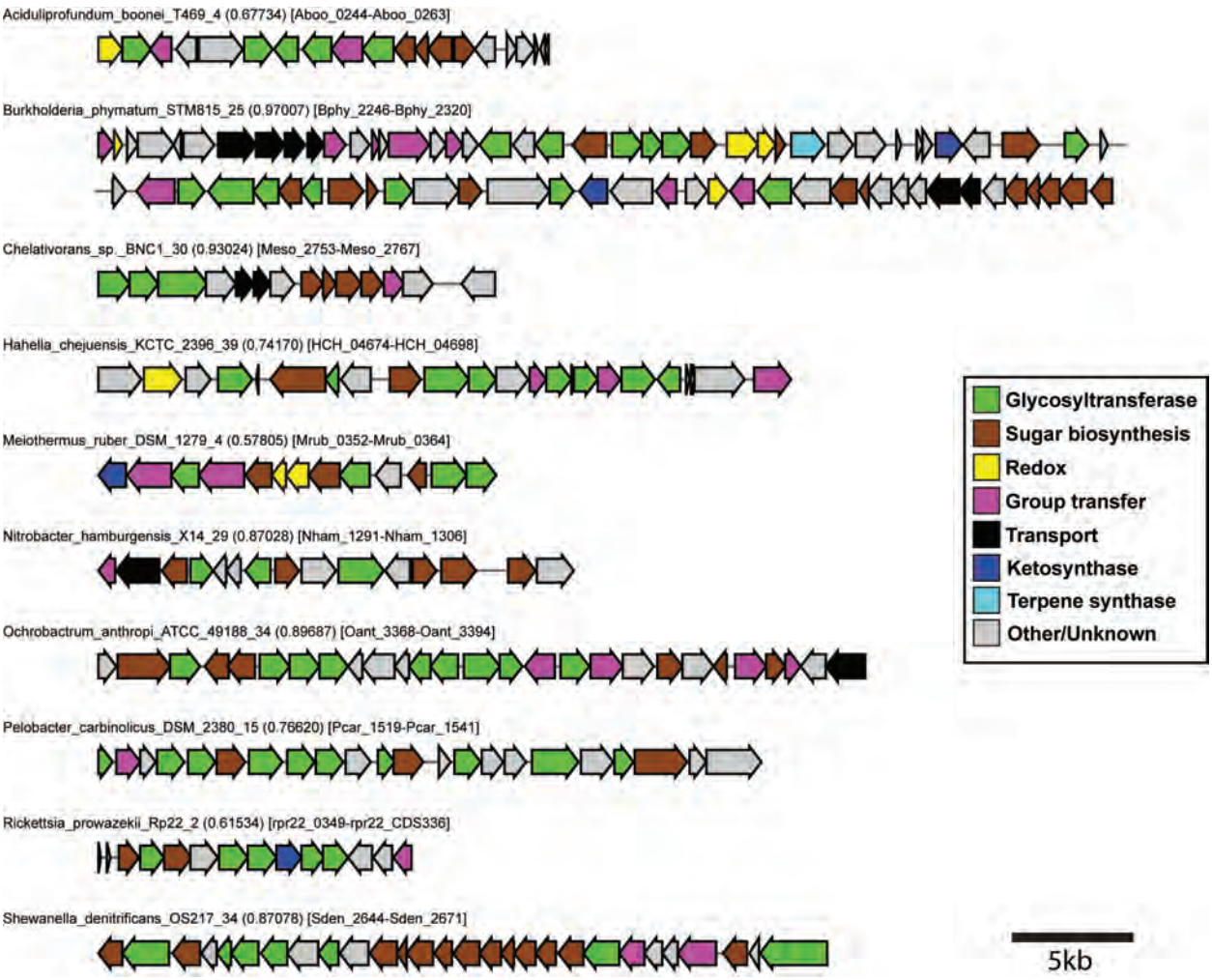
4. Saccharides are the largest class of gene clusters

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.

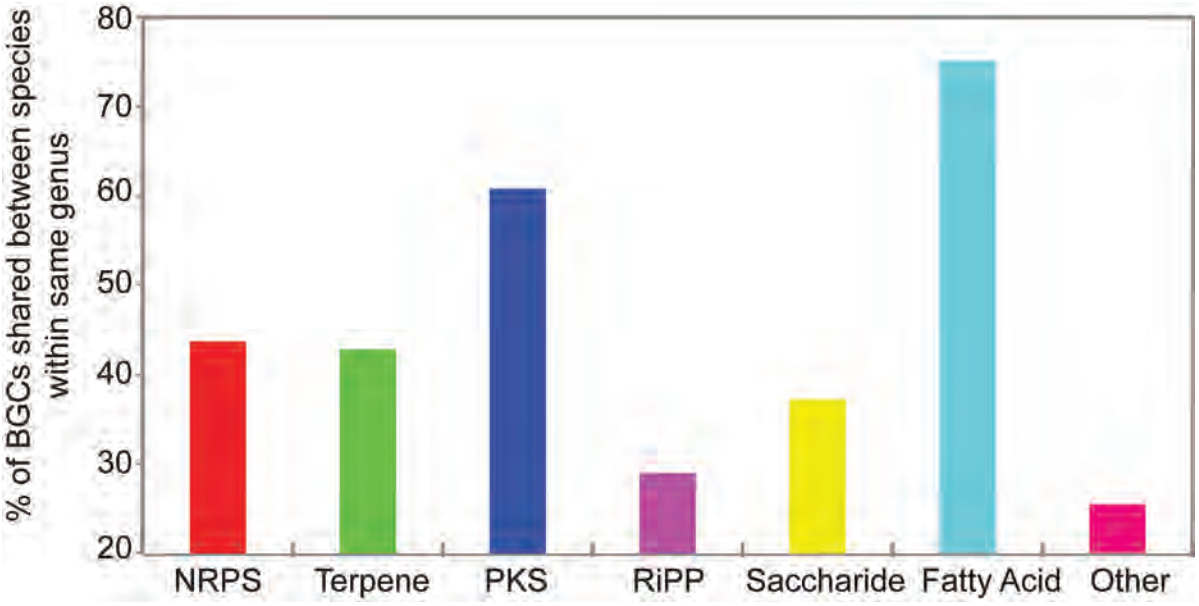
The predominance of saccharide gene clusters illuminates families of molecules that are not typically thought of as natural products. For example, based on Pfam domain content, 23% of the saccharide gene clusters are predicted to encode lipopolysaccharides and 3% capsular polysaccharides. These cell 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).



SI Figure 7. 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.



SI Figure 8. Type diversity of BGCs within the same 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.

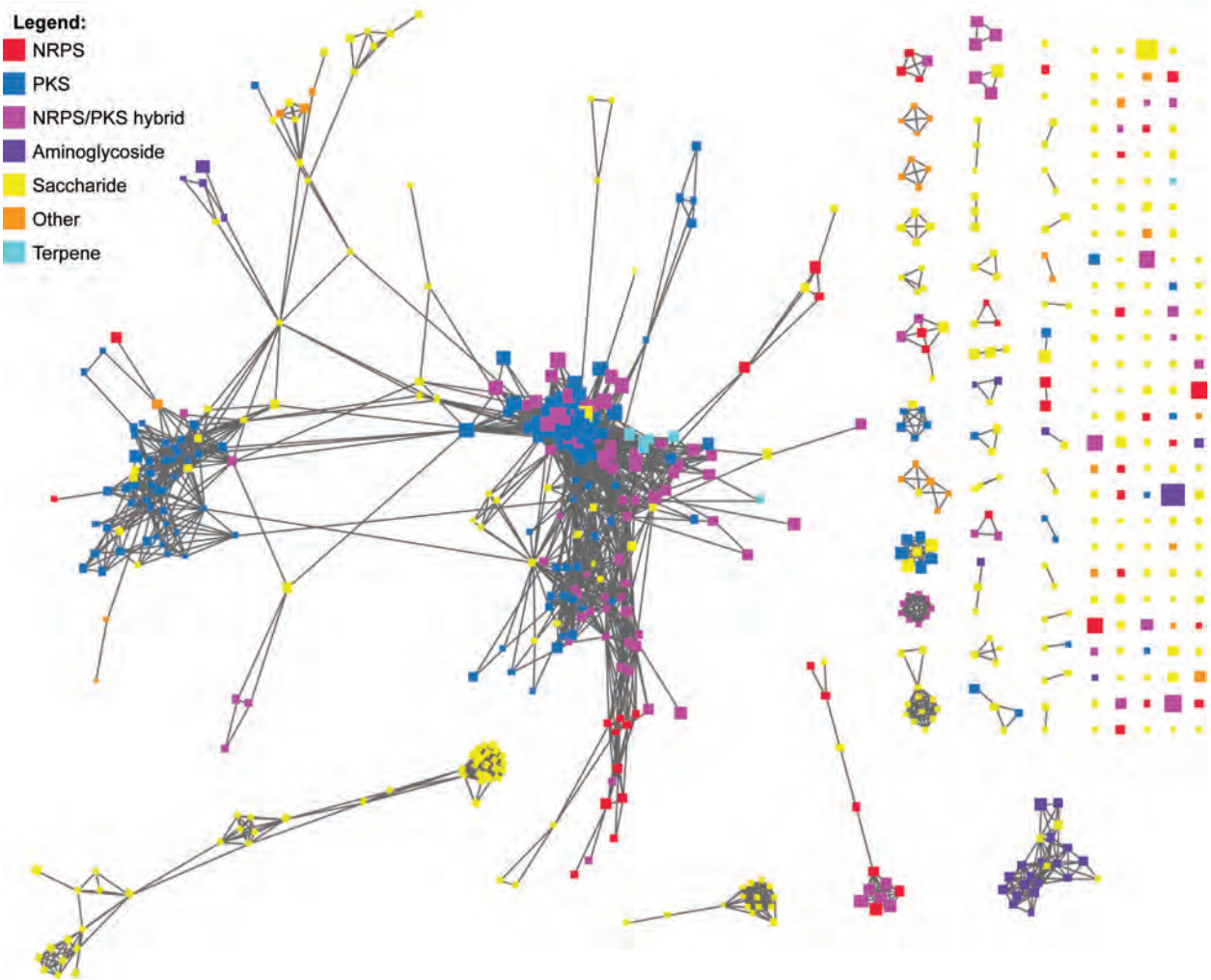
5. A global map of biosynthesis based on a gene cluster distance metric

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.(Lin et al., 2006b). 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 9**). 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.*



SI Figure 9. 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.

A

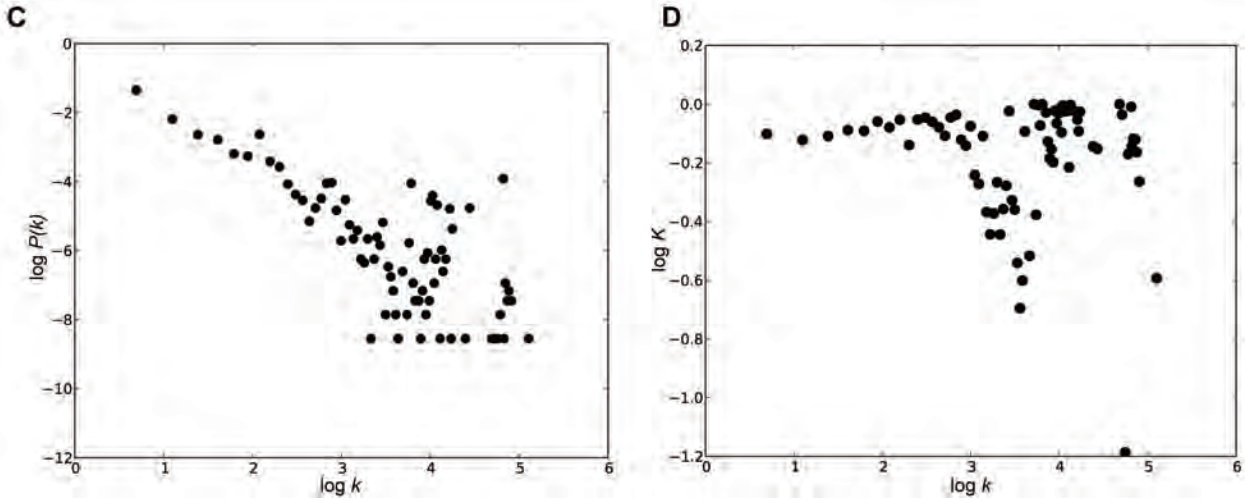
Statistics for the graph with >0.6 threshold

	#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

B

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



SI Figure 10. Analysis of the global BGC similarity network. Network (or graph) topology can be indicative of the relationships among its constituent nodes (here, BGCs). Tables **a** and **b** 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 **c**); 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 **d**. 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.

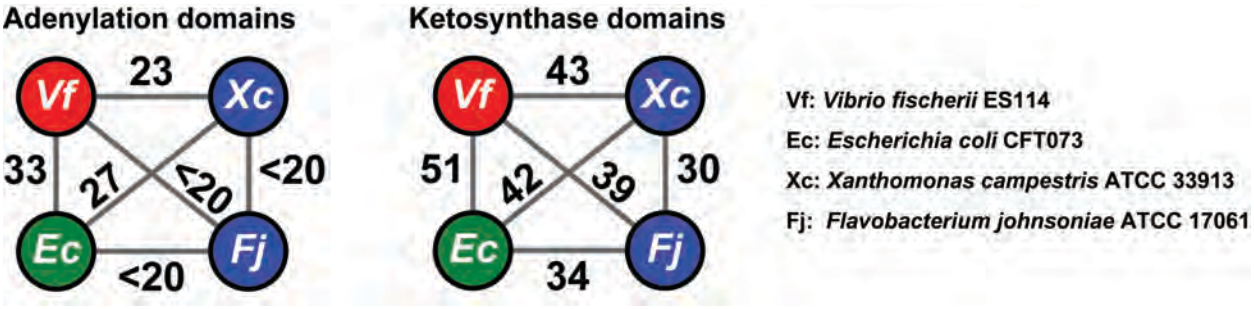
Unexpectedly, most gene clusters (84%) belong entirely to a single class. Hybrids therefore comprise a much larger proportion of known gene clusters than predicted gene clusters, suggesting that they may have been oversampled by experimental efforts to date. This may be partially explained by the fact 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 (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 10**)(Barabasi and Oltvai, 2004). 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.

SI Figure 11. 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).

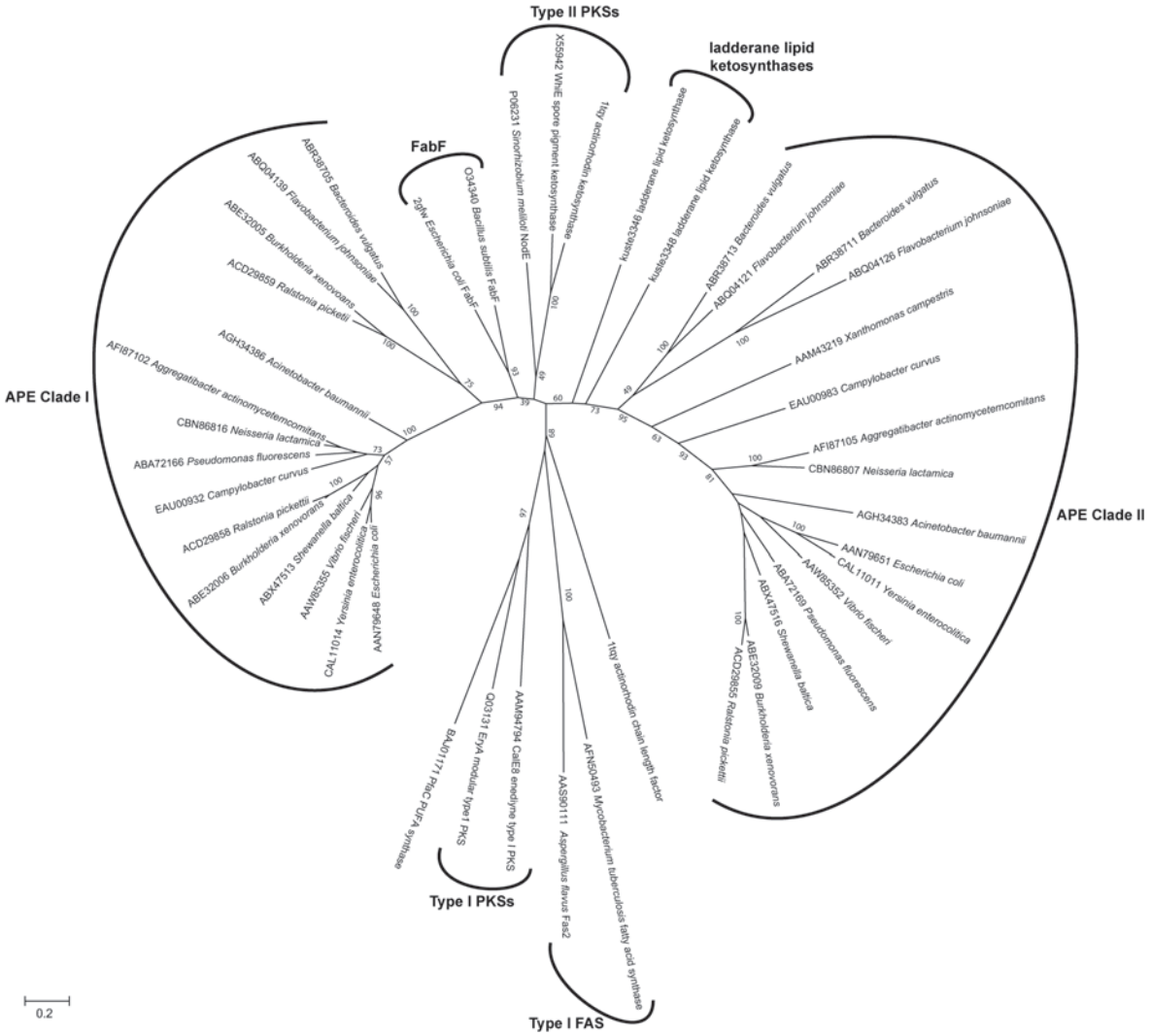
SI Figure 12. Full annotated APE superfamily clustered heat map including COG annotations (separate PDF file). Full version of the clustered heat map shown in Figure 3a. In this version, the COG annotations are shown at the bottom, and the accession number and source strain are shown on the right.



SI Figure 13. Pairwise sequence identities of the ketosynthase and adenylation domains in the four characterized gene clusters. The numbers in the graph represent the percentage identity between the amino acid sequences of the 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.

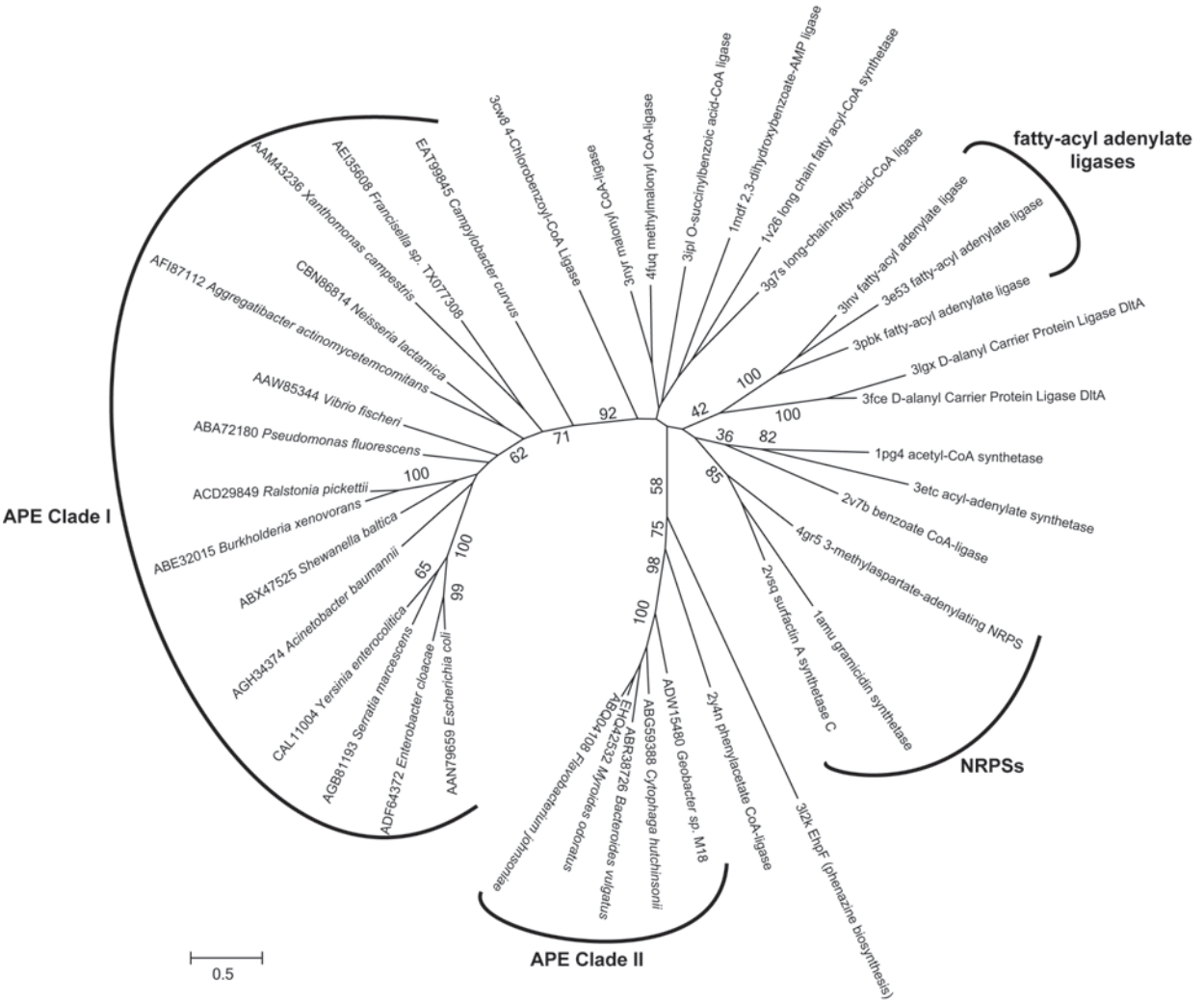
6. Phylogenetic analysis of APE ketosynthase and adenylation enzymes

Structure-guided multiple sequence alignment and maximum likelihood phylogenetic reconstruction (**SI Figure 14&15**) 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 16**). 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.



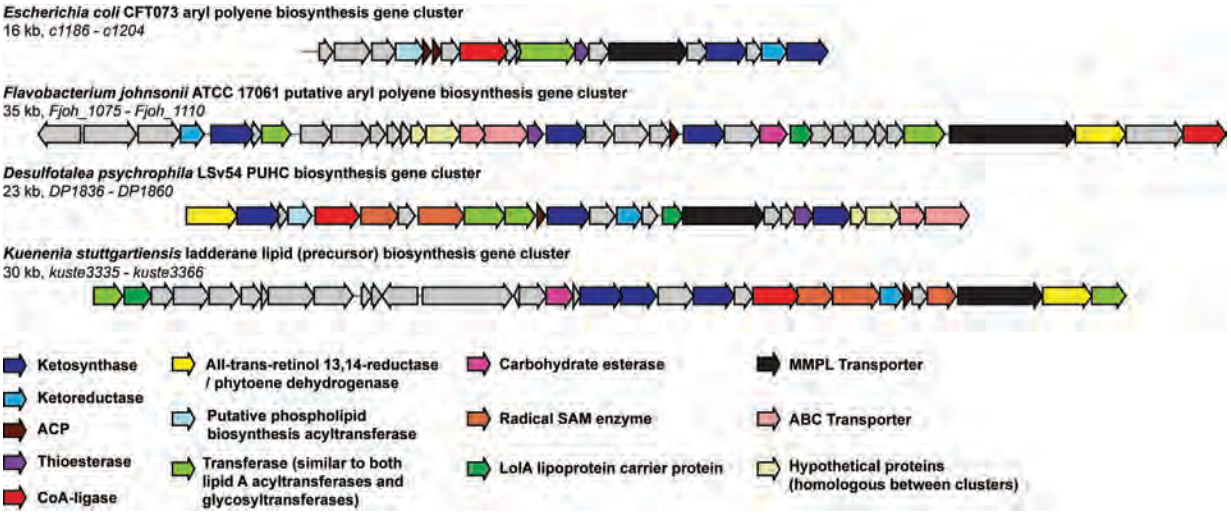
SI Figure 14. Phylogenetic tree of APE ketosynthase domains with other ketosynthases. The maximum likelihood phylogenetic tree, based on a structure-guided multiple sequence alignment using PROMALS3D(Pei et al., 2008), shows that the ketosynthases from representative APE gene clusters belong to two evolutionary 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.



SI Figure 15. 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.

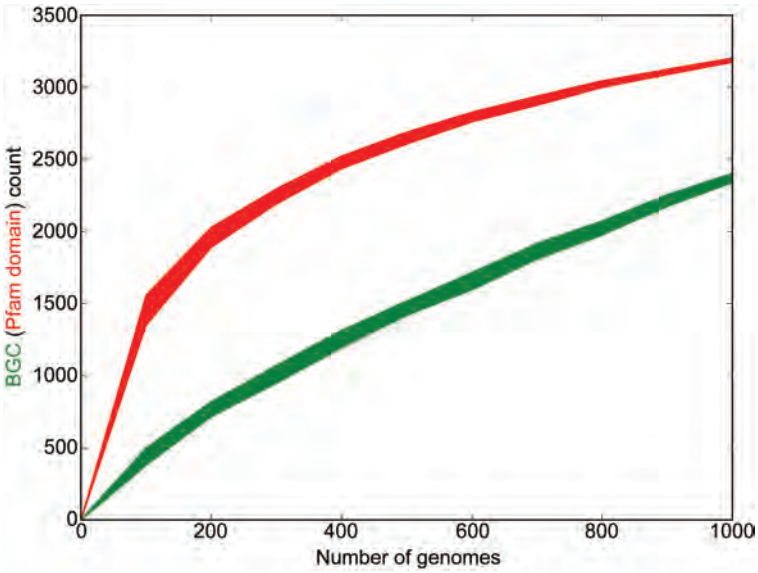
The APE A domains comprise two separate, unrelated clades, suggesting convergent evolution of the enzyme that selects and activates the starter unit. The closest known homologs of these clades are a phenylacetate CoA-ligase and a 4-chlorobenzoyl-CoA ligase, respectively(Law and Boulanger, 2011; Reger 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.



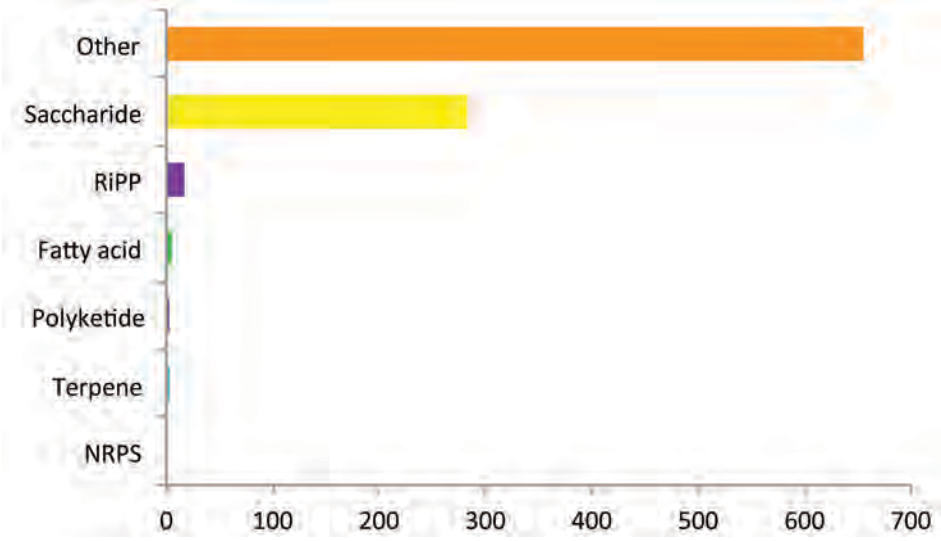
SI Figure 16. Comparison of APE gene clusters with related BGCs. Alignment of the two APE superfamily gene clusters from *Escherichia coli* CFT073 and *Flavobacterium johnsonii* ATCC 17061, the putative ladderane lipid biosynthesis gene cluster from *Kuenenia stuttgartiensis* and the polyunsaturated hydrocarbon biosynthesis gene cluster from *Desulfotalea psychrophila* LSv54. Colors signify homologous genes based on a MultiGeneBlast comparison with the blastp algorithm.

SI Table III (in separate SI_Tables.XLSX file). 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 Figure 17 (in separate PDF). 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 18. Rarefaction analysis of numbers of BGC families and Pfam families. 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).



SI Figure 19. 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 Table IV (in separate SI_Tables.XLSX file). 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.(Nett et al., 2009).

SI Table V (in separate SI_Tables.XLSX file). 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 VI (in separate SI_Tables.XLSX file). Primers used in this study.

SI Table VII (in separate SI_Tables.XLSX file). Plasmids used in this study.

SI Table VIII (in separate SI_Tables.XLSX file). Strains used in this study.

METHODS

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.

SI Table IX (in separate SI_Tables.XLSX file). 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.

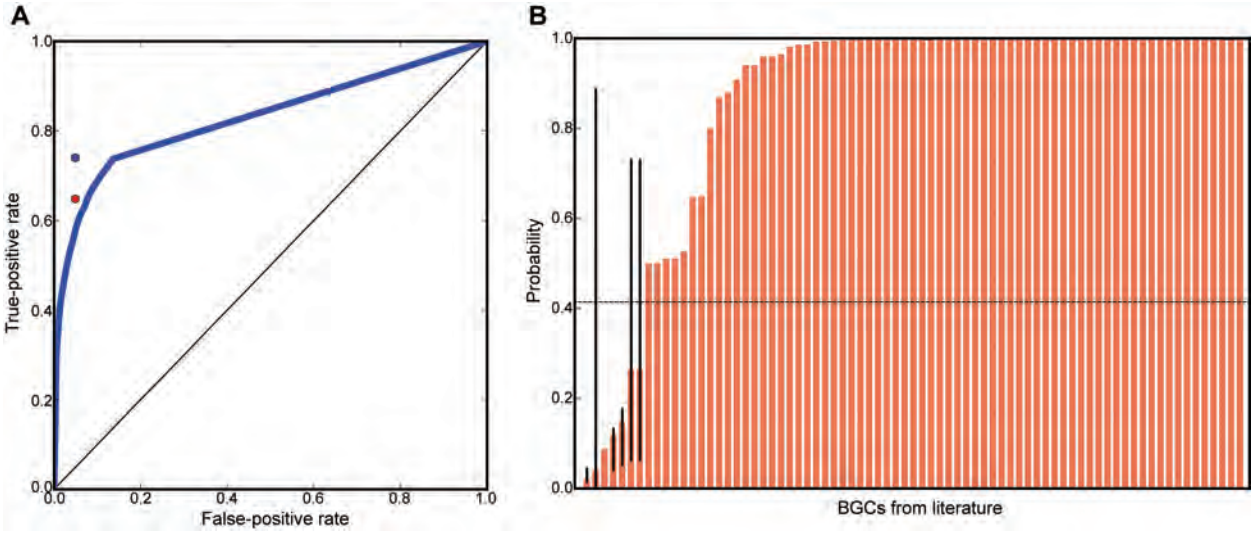
To obtain Pfam domain frequencies for the non-BGC state, we first randomly selected one hundred genomes (**SI Table IX**), and aligned all their Pfam domain sequences to all Pfam domain sequences from the BGC training set using the blastp algorithm(Camacho et al., 2009). Only hits with an E-value larger 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).

SI Table X (in separate SI_Tables.XLSX file). 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 XI (in separate SI_Tables.XLSX file). Predicted high-confidence BGCs from all genomes.



SI Figure 20. Evaluation of the ClusterFinder algorithm. 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.

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 20 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.

SI Table XII (in separate SI_Tables.XLSX file). A list of BGCs from 10 manually annotated genomes, used to evaluate the performance of ClusterFinder algorithm.

SI Table XIII (in separate SI_Tables.XLSX file). A list of 74 BGCs from the literature, used to evaluate the performance of ClusterFinder algorithm.

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

(a) (b)

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.(Lin et al., 2006b) 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).

$$\frac{N_i^P - N_i^Q}{S} \qquad \frac{N_i^P - N_i^Q - Munkres(D(N_i^P, N_i^Q))}{S}$$

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., 2006a).

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 introduced the *ermE** promoter from pIJ10257 upstream of the operon starting in VF0844. This was done 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

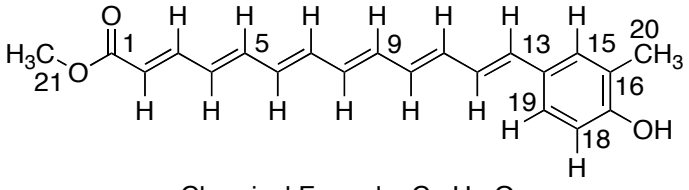
V. fischeri and *E. coli* strains were grown for 3.5 days in the dark prior to harvesting. Cell pellets were collected by centrifugation (5180 x g, 30 min), washed with water and repeatedly extracted with

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).

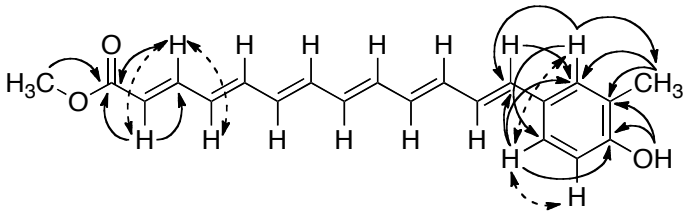
Special Considerations for Structure Elucidation

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})



Chemical Formula: $C_{21}H_{22}O_3$
Exact Mass: 322.1569



SI Figure 21. Structure of APE_{Ec} with COSY (dashed lines) and HMBC (solid lines) correlations.

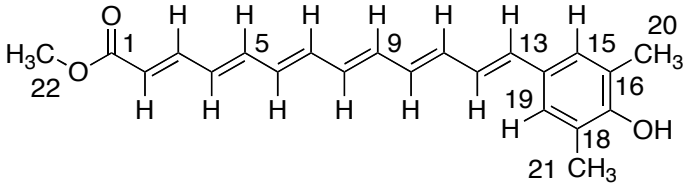
APE_{EC}, a red amorphous powder, was determined to have a molecular formula of $C_{21}H_{22}O_3$ based on the observation of the $[M-H]^-$ adduct at 321.1496 *m/z* (Δppm = -0.310) and analysis of one and two-dimensional NMR experiments (**SI Figure 21** and **SI Table XIV**). Based on ^1H NMR and HSQC 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.

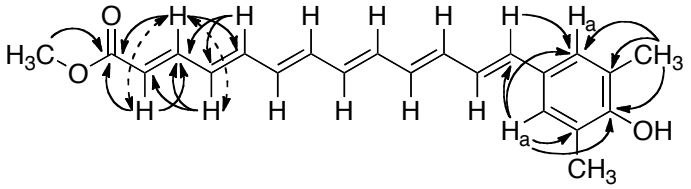
The methyl ester was identified via an HMBC correlation from the singlet methoxy proton signal at 3.7 ppm to the quaternary carbon C1 at 167.7 ppm. Protons H2 (doublet, ^{1}H 5.93 ppm $^{3}J_{HH}$ = 15.2 Hz; ^{13}C 120.5 ppm) and H3 (doublet of doublets, ^{1}H 7.33 ppm $^{3}J_{HH}$ = 15.1, 11.4; ^{13}C 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 ^{1}H 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).

SI Table XIV (in separate SI_Tables.XLSX file). ^{1}H and ^{13}C NMR data.

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



Chemical Formula: $C_{22}H_{24}O_3$
Exact Mass: 336.1725



SI Figure 22. Structure of APE$_{Vf}$ with COSY (dashed lines) and HMBC (solid lines) correlations.

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 22** 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, ^{1}H 5.97 ppm $^3J_{HH}$ = 15.2 Hz) and H3 (doublet of doublets, ^{1}H 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 26**).

SI Figure 23. NMR spectra for APE_{Vf} and APE_{Ec}. (in separate file)

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

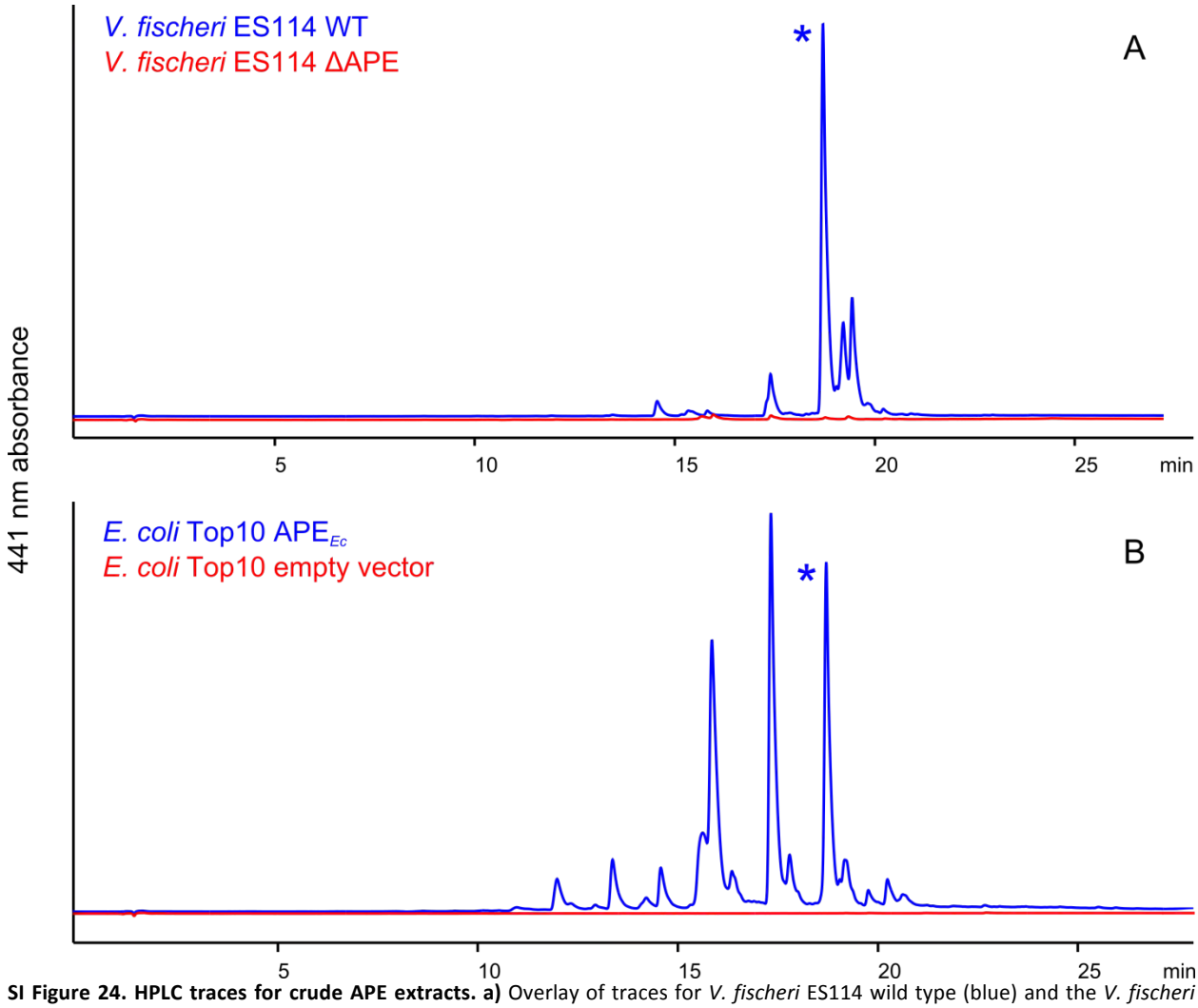
Growth and Purification

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 24**).

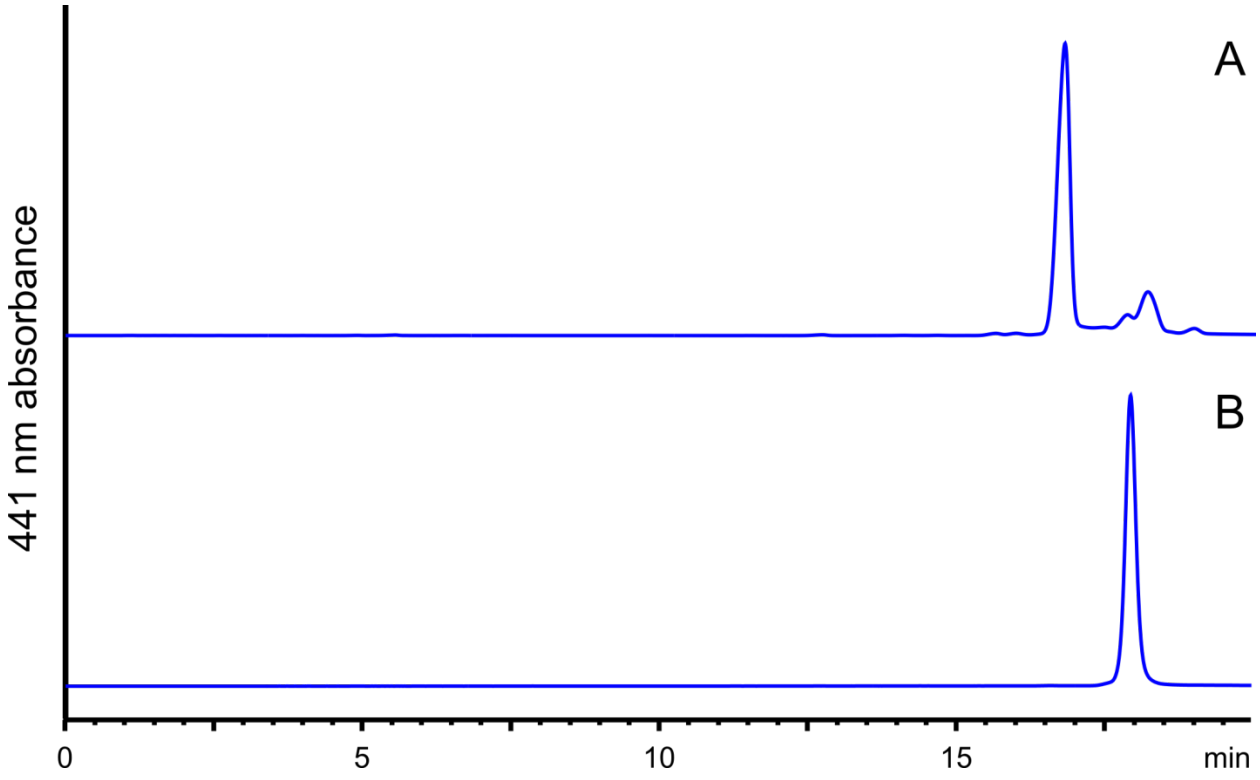


SI Figure 24. HPLC traces for crude APE extracts. a) Overlay of traces for *V. fischeri* ES114 wild type (blue) and the *V. fischeri* ES114 Δ*ape* deletion strain (red). **b)** Overlay of traces for *E. coli* Top10 expressing the CFT073 cluster (blue) and the *E. coli* Top10 control strain containing 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.

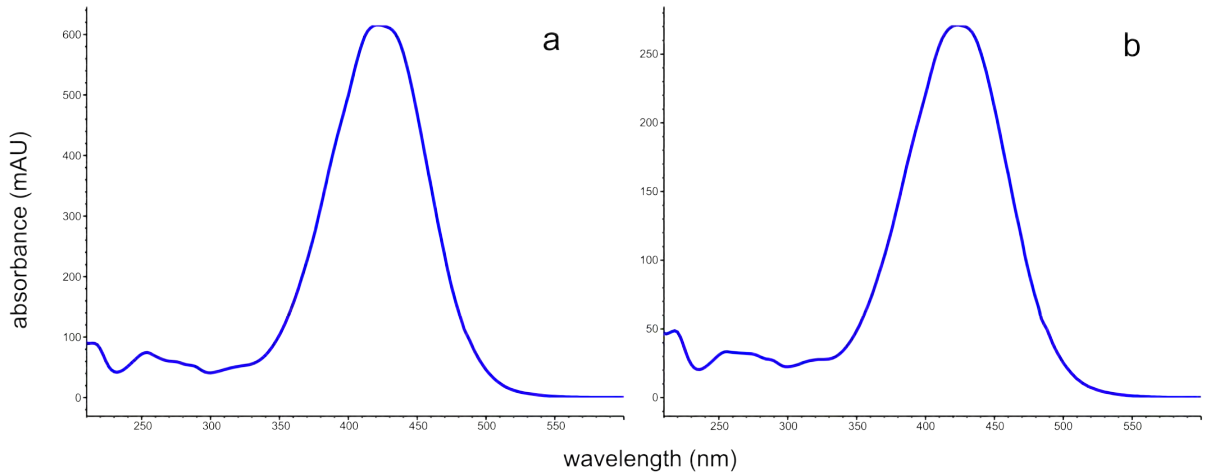
E.coli materials were purified on RP-HPLC using a two step purification protocol. Firstly, crude material was purified on a semi-prep RP column (Phenomenex Synergi Fusion-RP, 250 x 10 mm, 10 μm) using a gradient of acetonitrile MeCN:H_2O + 0.02% formic acid (32% MeCN for 26 minutes, 100% MeCN for 9 min, 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 25a**). 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 25b**). 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.



SI Figure 25. Second RP-HPLC purification for APE_{Ec} (a) and APE_{VF} (b).



SI Figure 26. UV spectrum for APE_{Ec} (a) and APE_{VF} (b).

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.

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 chemica Scandinavica *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. Natural product reports *30*, 108-160.

Barabasi, A.L., and Oltvai, Z.N. (2004). Network biology: understanding the cell's functional organization. Nature reviewsGenetics *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. Spectrochimica acta Part A, Molecular and biomolecular spectroscopy *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. Nucleic 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. Natural product reports *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). HMMER3.
Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research *32*, 1792-1797.
Federhen, S. (2012). The NCBI Taxonomy database. Nucleic 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. Proceedings of the National Academy of Sciences of the United States of America *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. Advances in applied microbiology *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.

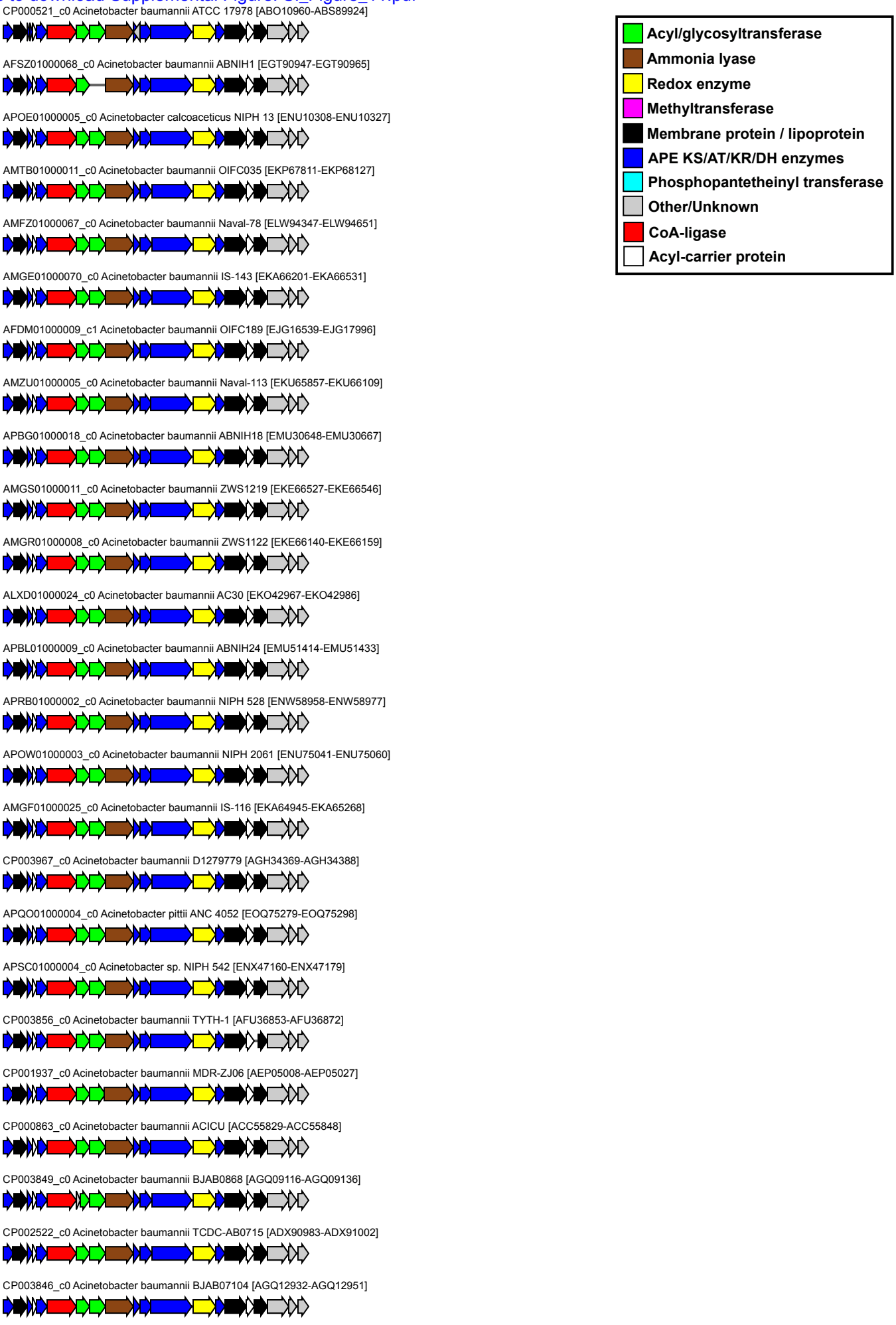
Letunic, I., and Bork, P. (2011). Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. Nucleic acids research *39*, W475-478.
Letzel, A.C., Pidot, S.J., and Hertweck, C. (2013). A genomic approach to the cryptic secondary metabolome of the anaerobic world. Natural product reports *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. (2006a). An initial strategy for comparing proteins at the domain architecture level. Bioinformatics *22*, 2081-2086.
Lin, K., Zhu, L., and Zhang, D.Y. (2006b). An initial strategy for comparing proteins at the domain architecture level. Bioinformatics (Oxford, England) *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. Proceedings of the National Academy of Sciences of the United States of America *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. Nucleic 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. Nucleic 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. Journal of the Society for Industrial & Applied Mathematics *5*, 32-38.
Nett, M., Ikeda, H., and Moore, B.S. (2009). Genomic basis for natural product biosynthetic diversity in the actinomycetes. Natural product reports *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. Nature biotechnology *23*, 873-878.
Pavoine, S., Baguette, M., and Bonsall, M.B. (2010). Decomposition of trait diversity among the nodes of a phylogenetic tree. Ecological Monographs *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. Nature reviewsMicrobiology *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. Nucleic acids research *39*, W362-W367.
Ruckert, 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. Journal of Biotechnology *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. Physical reviewE, Statistical, nonlinear, and soft matter physics *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 (Oxford, England) *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 Ruckert, C. (2012). Complete genome sequence of Saccharothrix espanaensis DSM 44229(T) and comparison to the other completely sequenced Pseudonocardiaceae. BMC genomics *13*, 465-2164-2113-2465.
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. Journal of Bacteriology *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.

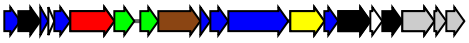
Wang, Y., Qian, G., Li, Y., Wang, Y., Wang, Y., Wright, S., Li, Y., Shen, Y., Liu, F., and Du, L. (2013). Biosynthetic mechanism for sunscreens of the biocontrol agent Lysobacter enzymogenes. PLoS One *8*, e66633.
Zechmeister, L., and Polgar, A. (1943). cis-trans isomerization and cis-peak effect in 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 (Reading, England) *157*, 2033-2041.

Supplemental Figure

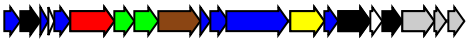
Click here to download Supplemental Figure: SI_Figure_11.pdf



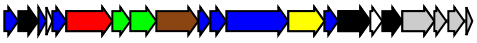
CP002177_c0 Acinetobacter calcoaceticus PHEA-2 [ADY84006-ADY84025]



GG704951_c0 Acinetobacter calcoaceticus RUH2202 genomic scaffold supercont1.3 [EEY76566-EEY76585]



APQA01000024_c0 Acinetobacter ursingii DSM 16037 = CIP 107286 [ENV75005-ENV75025]



APQR01000011_c0 Acinetobacter haemolyticus NIPH 261 [ENW19503-ENW19522]

APOS01000030_c0 Acinetobacter guillouiae CIP 63.46 [ENU58484-ENU58502]

ASQG01000031_c0 Acinetobacter guillouiae MSP4-18 [EPH35229-EPH35247]

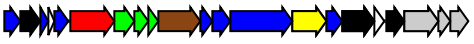
APPR01000012_c0 Acinetobacter brisouii ANC 4119 [ENV46937-ENV46955]

ATGI01000039_c0 Acinetobacter rudis CIP 110305 [EPF70006-EPF70024]

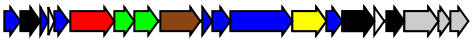
AQFM01000036_c0 Acinetobacter tandoii DSM 14970 = CIP 107469 [EOR08269-EOR08287]

APOH01000003_c0 Acinetobacter sp. ANC 3994 [ENU21359-ENU21377]

BAEB01000033_c0 Acinetobacter sp. NBRC 100985 [GAB02540-GAB02560]



APPO01000005_c0 Acinetobacter venetianus RAG-1 = CIP 110063 [ENV38510-ENV38529]



ASQH01000004_c0 Acinetobacter gyllenbergii MTCC 11365 [EPH34876-EPH34895]

APOI01000007_c0 Acinetobacter sp. NIPH 809 [ENU25111-ENU25130]

APRN01000042_c0 Acinetobacter sp. CIP 70.18 [ENX53257-ENX53276]

APRL01000015_c0 Acinetobacter sp. ANC 4105 [ENW90095-ENW90114]

APRW01000014_c0 Acinetobacter sp. NIPH 2168 [ENX20242-ENX20261]

APPC01000012_c0 Acinetobacter sp. NIPH 758 [ENU93619-ENU93638]

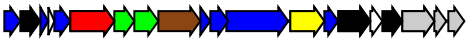
AMFX01000021_c0 Acinetobacter baumannii OIFC338 [ELX03099-ELX03405]

AMSX01000054_c0 Acinetobacter baumannii Naval-2 [EKP52632-EKP52882]

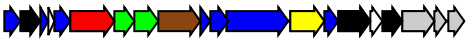
AFTB01000222_c0 Acinetobacter baumannii ABNIH3 [EGT92681-EGT92700]

APBM01000050_c0 Acinetobacter baumannii ABNIH25 [EMT85842-EMT85861]

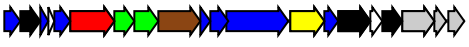
APBI01000013_c0 Acinetobacter baumannii ABNIH20 [EMU47165-EMU47184]



APBJ01000018_c0 Acinetobacter baumannii ABNIH22 [EMU41192-EMU41211]



APBK01000012_c0 Acinetobacter baumannii ABNIH23 [EMU45963-EMU45982]

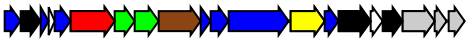


ALAM01000002_c0 Acinetobacter baumannii AC12 [EJN44840-EJN44859]

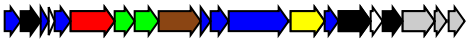
AKAT01000028_c0 Acinetobacter baumannii Ab44444 [EKB40539-EKB40558]

APOR01000015_c0 Acinetobacter baumannii NIPH 1362 [ENU53910-ENU53929]

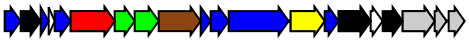
AMFV01000034_c0 Acinetobacter baumannii OIFC065 [EKP33754-EKP34073]



AMFL01000012_c0 Acinetobacter baumannii OIFC110 [EKL59996-EKL60320]



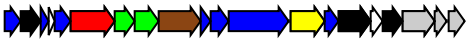
AMGG01000003_c0 Acinetobacter baumannii WC-692 [EKA74729-EKA75017]



APQH01000003_c0 Acinetobacter calcoaceticus ANC 3680 [ENV95698-ENV95717]

APQM01000009_c0 Acinetobacter pittii ANC 4050 [EOQ67307-EOQ67326]

AMSS01000050_c0 Acinetobacter baumannii WC-141 [EKU38416-EKU38584]



APQI01000002_c0 Acinetobacter calcoaceticus DSM 30006 = CIP 81.8 [ENW01205-ENW01224]

APQJ01000011_c0 Acinetobacter calcoaceticus ANC 3811 [EOQ61784-EOQ61803]

APPF01000031_c0 Acinetobacter sp. NIPH 817 [ENV01333-ENV01352]

CP003500_c0 Acinetobacter baumannii MDR-TJ [AFI96817-AFI96836]

CP002080_c0 Acinetobacter oleivorans DR1 [ADI92246-ADI92265]

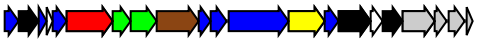
APOF01000017_c0 Acinetobacter baumannii NIPH 24 [ENU11702-ENU11723]

AKAQ01000022_c0 Acinetobacter baumannii Ab11111 [EKB35122-EKB35143]

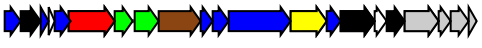
AFDO01000006_c1 Acinetobacter baumannii Naval-17 [EJG29922-EJG31639]

CP001921_c0 Acinetobacter baumannii 1656-2 [ADX02182-ADX02204]

APQB01000015_c0 Acinetobacter ursingii NIPH 706 [ENX50059-ENX50079]

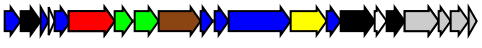


APPT01000018_c0 Acinetobacter baylyi DSM 14961 = CIP 107474 [ENV52628-ENV52648]



APQC01000006_c0 Acinetobacter ursingii ANC 3649 [ENV80295-ENV80317]

CR543861_c0 Acinetobacter sp. ADP1 complete genome. [CAG67490-CAG67510]

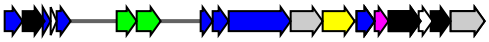


AKVH01000112_c0 Pseudomonas sp. Ag1 [EJF68113-EJF68131]

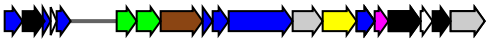
ATLM01000050_c0 Pseudomonas sp. CFT9 [EPJ82030-EPJ82049]

ATLQ01000028_c0 Pseudomonas sp. CF150 [EPL12227-EPL12246]

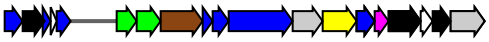
AOKQ01000159_c0 Pseudomonas syringae pv. actinidiae ICMP 18801 [EPN82856-EPN82873]



AOJW01000082_c0 Pseudomonas syringae pv. actinidiae ICMP 19072 [EPM79023-EPM79041]



AOJS01000116_c0 Pseudomonas syringae pv. actinidiae ICMP 19071 [EPM61183-EPM61201]



CM001763_c0 Pseudomonas syringae pv. syringae B64 chromosome [ELS40171-ELS40188]

CP000075_c0 Pseudomonas syringae pv. syringae B728a [AAY35492-AAY35509]

CM001986_c0 Pseudomonas syringae pv. syringae SM chromosome [EPF64562-EPF64579]

AKCK01000025_c0 Pseudomonas syringae pv. avellanae str. ISPaVe037 [EKG41147-EKG41164]

AMXL01000056_c0 Pseudomonas syringae BRIP34881 [ELQ00231-ELQ00248]

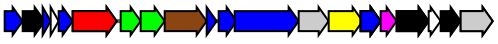
AMZW01000118_c0 Pseudomonas fluorescens BRIP34879 [ELQ12746-ELQ12764]

AEGG01000052_c0 Pseudomonas syringae pv. glycinea str. B076 [EFW78118-EFW78136]

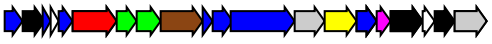
AEAF01000057_c0 Pseudomonas syringae pv. lachrymans str. M301315 [EGH82798-EGH82816]

AOKJ01000284_c0 Pseudomonas syringae pv. actinidiae ICMP 19094 [EPN49744-EPN49762]

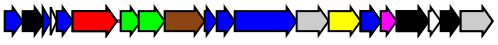
AKJM01000179_c0 Pseudomonas sp. GM48 [EJM48427-EJM48446]



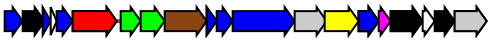
AKJJ01000034_c0 Pseudomonas sp. GM55 [EJM76488-EJM76507]



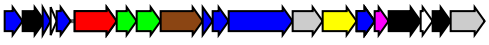
AKJI01000024_c0 Pseudomonas sp. GM60 [EJM86620-EJM86639]



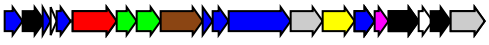
AKJK01000051_c0 Pseudomonas sp. GM50 [EJM66763-EJM66782]



AE016853_c0 Pseudomonas syringae pv. tomato str. DC3000 [AAO58518-AAO58537]



AMQP01000022_c0 Pseudomonas viridiflava UASWS0038 [EKN47841-EKN47860]



AOJV01000204_c0 Pseudomonas syringae pv. theae ICMP 3923 [EPM67672-EPM67691]

AEAM01000165_c0 Pseudomonas syringae pv. lachrymans str. M302278 [EGH96047-EGH96066]

CP003041_c0 Pseudomonas fluorescens A506 [AFJ54760-AFJ59733]

AKXH01000026_c1 Pseudomonas fluorescens BBc6R8 [EKA24007-EKA24025]

AEAZ01000004_c1 Pseudomonas fluorescens WH6 [EFQ65655-EFQ65675]

AHPP01000002_c0 Pseudomonas synxantha BG33R [EIK67178-EIK69712]

AHPN01000001_c0 Pseudomonas fluorescens SS101 [EIK58908-EIK64019]

ATLN01000036_c0 Pseudomonas sp. CFII68 [EPJ96370-EPJ96390]

AHPO01000001_c0 Pseudomonas fluorescens Q8r1-96 [EIK70011-EIK71017]

AKJQ01000029_c0 Pseudomonas sp. GM25 [EJM29402-EJM29422]

AKJR01000081_c0 Pseudomonas sp. GM24 [EJM39234-EJM39254]

CP000058_c0 Pseudomonas syringae pv. phaseolicola 1448A [AAZ32967-AAZ37878]

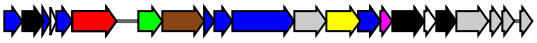
KB644101_c0 Pseudomonas savastanoi pv. savastanoi NCPPB 3335 genomic scaffold WGS_Scaffold_009 [EFI01481-EFI01499]

CP003150_c0 Pseudomonas fluorescens F113 [AEV60495-AEV60515]

CP000094_c0 Pseudomonas fluorescens Pf0-1 [ABA72165-ABA72185]

AGBM01000001_c0 Pseudomonas fluorescens Q2-87 [EJL01050-EJL05486]

CP000076_c0 Pseudomonas protegens Pf-5 [AAY95867-AAY95889]

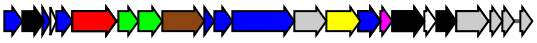


AMWJ01000053_c0 Pseudomonas putida CSV86 [EKX85913-EKX85934]

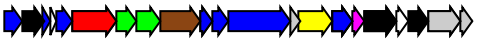
AHOT01000030_c1 Pseudomonas chlororaphis O6 [EIM12842-EIM13307]

AHHJ01000010_c1 Pseudomonas chlororaphis subsp. aureofaciens 30-84 [EJL00434-EJL00905]

CP003190_c0 Pseudomonas protegens CHA0 [AGL82264-AGL82286]



AJWX01000006_c0 Pseudomonas sp. M47T1 [EIK96902-EIK96922]



AKJP01000114_c0 Pseudomonas sp. GM30 [EJM32134-EJM32154]

AKJV01000005_c0 Pseudomonas sp. GM16 [EJM07516-EJM07536]

ATLR01000017_c0 Pseudomonas sp. CF149 [EPJ94091-EPJ94111]

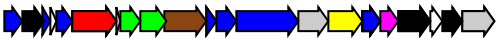
AOGS01000005_c0 Pseudomonas sp. Lz4W [EMI08472-EMI08492]

AEGH01000039_c0 Pseudomonas syringae pv. glycinea str. race 4 [EFW86828-EFW86846]

CP004045_c0 Pseudomonas poae RE*1-1-14 [AGE28653-AGE28671]

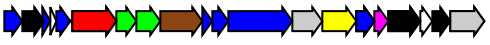
CP003880_c1 Pseudomonas sp. UW4 [AFY17566-AFY17585]

AKJG01000041_c1 Pseudomonas sp. GM74 [EJM94200-EJM94220]

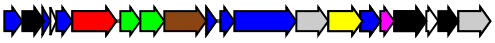


ABSM01000015_c0 Pseudomonas syringae pv. tomato T1 [EEB59391-EEB59410]

AOKE01000079_c0 Pseudomonas syringae pv. actinidiae ICMP 19098 [EPM49524-EPM49543]

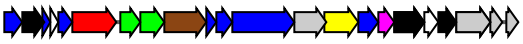


ATLO01000035_c0 Pseudomonas sp. CFII64 [EPJ82304-EPJ82323]

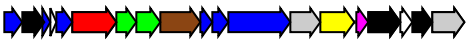


AKJN01000054_c1 Pseudomonas sp. GM41(2012) [EJM52909-EJM52928]

CM001561_c0 Pseudomonas fluorescens R124 chromosome [EJZ56134-EJZ56155]



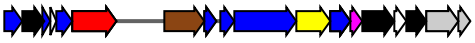
AKJF01000074_c1 Pseudomonas sp. GM78 [EJN29470-EJN29488]



APIO01000014_c0 Pseudomonas sp. G5(2012) strain G5 [EPA98294-EPA98312]

AKCJ01000039_c0 Pseudomonas syringae pv. avellanae str. ISPaVe013 [EKG43403-EKG43418]

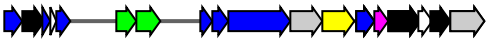
AM181176_c0 Pseudomonas fluorescens SBW25 complete genome. [CAY46696-CAY46714]



CT573326_c0 Pseudomonas entomophila str. L48 chromosome [CAK13303-CAK13324]

CP002585_c0 Pseudomonas brassicacearum subsp. brassicacearum NFM421 [AEA66583-AEA66603]

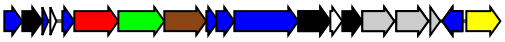
AOKN01000419_c0 Pseudomonas syringae pv. actinidiae ICMP 19097 [EPN73760-EPN73777]



AMZX01000012_c0 Pseudomonas syringae BRIP39023 [ELQ13515-ELQ13532]

AMXK01000020_c0 Pseudomonas syringae BRIP34876 [ELQ02356-ELQ02373]

CP004013_c0 Ralstonia solanacearum FQY_4 megaplasmid [AGH86472-AGH86490]



CP001645_c0 Ralstonia pickettii 12D chromosome 2 [ACS64942-ACS64961]

CP001069_c0 Ralstonia pickettii 12J chromosome 2 [ACD29844-ACD29863]

ABLD01000010_c0 Burkholderia graminis C4D1M [EDT09611-EDT09630]

CP001043_c1 Burkholderia phymatum STM815 chromosome 1 [ACC69878-ACC69897]

JH603161_c2 Burkholderia sp. Ch1-1 genomic scaffold BCh11scaffold_3 [EIF30102-EIF30121]

CP003516_c0 Burkholderia sp. KJ006 chromosome 3 [AFJ89793-AFJ89812]

CP001026_c0 Burkholderia ambifaria MC40-6 chromosome 2 [ACB66537-ACB66556]

CP000616_c0 Burkholderia vietnamiensis G4 chromosome 3 [ABO58878-ABO58897]

CP000441_c0 Burkholderia ambifaria AMMD chromosome 2 [ABI89151-ABI89170]

FP885907_c0 Ralstonia solanacearum CFBP2957 plasmid RCFBPv3_mp [CBJ53166-CBJ53186]

ACFB01000016_c0 Burkholderia multivorans CGD1 [EED97410-EED97429]

ACFD01000001_c0 Burkholderia multivorans CGD2M [EEE15577-EEE15596]

ACFC01000001_c0 Burkholderia multivorans CGD2 [EEE09654-EEE09673]

APMQ01000007_c0 Ralstonia pickettii OR214 [ENZ77178-ENZ77197]

AP013061_c0 Burkholderia sp. RPE64 plasmid p1 DNA [BAN27554-BAN27573]

AKKD01000442_c0 Burkholderia sp. BT03 [EJL42518-EJL42537]

AKAU01000036_c0 Burkholderia terrae BS001 [EIN02239-EIN02258]

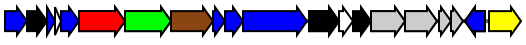
CP000270_c1 Burkholderia xenovorans LB400 chromosome 1 [ABE32001-ABE32020]

CP001052_c0 Burkholderia phytofirmans PsJN chromosome 1 [ACD17438-ACD17457]

CP002217_c0 Burkholderia sp. CCGE1003 chromosome 1 [ADN58596-ADN58615]

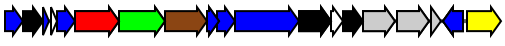
CP002519_c1 Burkholderia sp. CCGE1001 chromosome 1 [ADX56090-ADX56109]

CP003863_c1 Burkholderia phenoliruptrix BR3459a chromosome 1 [AFT86820-AFT86839]

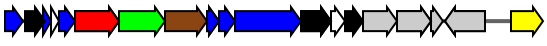


AL646053_c0 Ralstonia solanacearum GMI1000 megaplasmid complete sequence. [CAD17505-CAD17523]

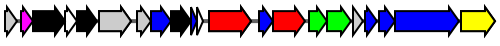
KE392095_c0 Ralstonia solanacearum SD54 scaffold34 [EPR99678-EPR99696]



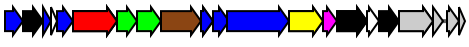
FP885896_c0 Ralstonia solanacearum CMR15 plasmid CMR15_mp [CBJ39980-CBJ39999]



CP003745_c1 Bibersteinia trehalosi USDA-ARS-USMARC-192 [AGH39460-AGH39480]



AHBD01000006_c0 Pseudomonas psychrotolerans L19 [EHK70429-EHK70449]



AOBS01000052_c0 Pseudomonas stutzeri NF13 [EMD99880-EMD99899]

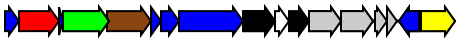
CP003071_c0 Pseudomonas stutzeri RCH2 [AGA88164-AGA88183]

CP003725_c0 Pseudomonas stutzeri DSM 10701 [AFN76630-AFN76649]

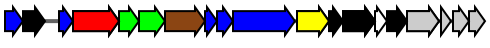
ADMS01000059_c0 Achromobacter piechaudii ATCC 43553 [EFF75994-EFF76016]

CP002287_c0 Achromobacter xylosoxidans A8 [ADP15664-ADP15682]

ALIX01000343_c0 Burkholderia multivorans CF2 [EJO54637-EJO54681]



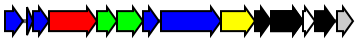
AMWD01000002_c0 Janthinobacterium sp. HH01 [ELX09985-ELX10003]



APPV01000006_c0 Acinetobacter soli NIPH 2899 [ENV61115-ENV61132]

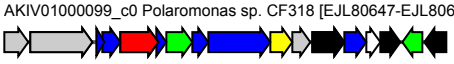
APPU01000004_c0 Acinetobacter soli CIP 110264 [ENV58668-ENV58685]

CP000285_c2 Chromohalobacter salexigens DSM 3043 [ABE60095-ABE60108]



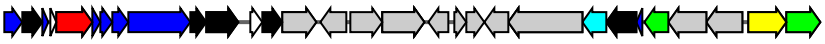
CABR01000151_c0 Mine drainage metagenome [CBI11562-CBI11582]

AKIV01000099_c0 Polaromonas sp. CF318 [EJL80647-EJL80663]



CP002056_c1 Methylotenera versatilis 301 [ADI30676-ADI30691]

AGSB01000067_c0 Holophaga foetida DSM 6591 [EHP08344-EHP08373]

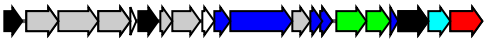


CM001368_c1 Desulfovibrio sp. FW1012B chromosome [EHJ48830-EHJ48850]

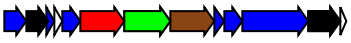
CP001649_c0 Desulfovibrio salexigens DSM 2638 [ACS78586-ACS78610]

FO203522_c0 Desulfovibrio hydrothermalis AM13 = DSM 14728 chromosome [CCO23417-CCO23437]

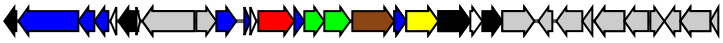
CP000142_c4 Pelobacter carbinolicus DSM 2380 [ABA89884-ABA89903]



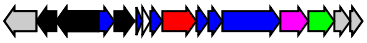
ABYL01000065_c0 Burkholderia sp. H160 [EEA00465-EEA00477]



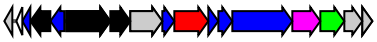
BAEO01000002_c0 Glaciecola arctica BSs20135 [GAC17065-GAC17096]



CBIZ010000029_c0 Clostridium sp. CAG:768 [CDE44553-CDE44595]



CAXB010000032_c0 Clostridium sp. CAG:715 [CCX80058-CCX80074]



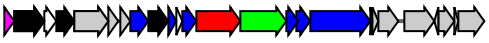
ATFC01000009_c0 Treponema vincentii F0403 [EPF46086-EPF46112]

AFAY01000010_c0 Neisseria bacilliformis ATCC BAA-1200 [EGF11745-EGF11764]

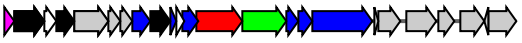
AJXU01000044_c0 Rhodanobacter fulvus Jip2 [EIL88874-EIL88890]

CP000767_c0 Campylobacter curvus 525.92 [EAT99837-EAU00983]

AMZQ01000020_c0 Campylobacter showae CSUNSWCD [EKU10174-EKU10198]



ACVQ01000006_c0 Campylobacter showae RM3277 [EET80551-EET80574]



CBJW010000010_c0 Fusobacterium sp. CAG:815 [CDE91565-CDE91577]

FN995097_c1 Neisseria lactamica 020-06 complete genome. [CBN86805-CBN86825]

ACDY02000011_c0 Neisseria cinerea ATCC 14685 [EEZ71120-EEZ71145]

ACEQ02000001_c0 Neisseria lactamica ATCC 23970 [EEZ76861-EEZ76886]

CP000792_c2 Campylobacter concisus 13826 [ABW74826-EAT98502]

ALJJ01000009_c0 Campylobacter sp. FOBRC14 [EJP75333-EJP75762]

ACFU01000031_c0 Campylobacter rectus RM3267 [EEF12950-EEF12982]

ABCS01000057_c0 Plesiocystis pacifica SIR-1 [EDM76834-EDM76850]

CBJP010000054_c0 Sutterella sp. CAG:521 [CDE75998-CDE76033]

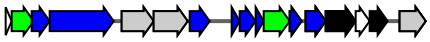
ATFE01000004_c0 Treponema medium ATCC 700293 [EPF29422-EPF29444]

CP002158_c1 Fibrobacter succinogenes subsp. succinogenes S85 [ADL24669-ADL26788]

CP001792_c4 Fibrobacter succinogenes subsp. succinogenes S85 [ACX76579-ACX76597]

ACRA01000003_c0 Fibrobacter succinogenes subsp. succinogenes S85 [EER86488-EER86506]

AJXS01000088_c0 Rhodanobacter sp. 115 [EIL98887-EIL98903]



AICM01000001_c0 Cellvibrio sp. BR [EIK45999-EIK46024]

CP001614_c3 Teredinibacter turnerae T7901 [ACR11254-ACS93584]

CP000947_c3 Haemophilus somnus 2336 [ACA31786-ACA31807]

CP000436_c0 Haemophilus somnus 129PT [ABI24381-ABI24399]

CP003402_c0 Francisella noatunensis subsp. orientalis str. Toba 04 [AFJ43169-AFJ43184]

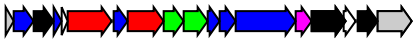
CP002738_c2 Methylomonas methanica MC09 [AEG02524-AEG02541]

CP003496_c0 Aggregatibacter actinomycetemcomitans D7S-1 [AFI87102-AFI87118]

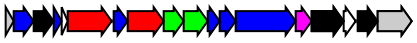
AEWU01000010_c0 Haemophilus parainfluenzae ATCC 33392 [EGC72772-EGC72789]

AEJN01000002_c0 Aggregatibacter actinomycetemcomitans serotype e str. SCC393 [EGY44278-EGY44297]

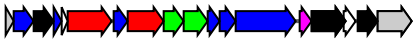
CP001733_c1 Aggregatibacter actinomycetemcomitans D11S-1 [ACX81935-ACX81952]



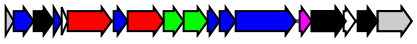
AEJP01000064_c0 Aggregatibacter actinomycetemcomitans serotype b str. SCC1398 [EGY44428-EGY44445]



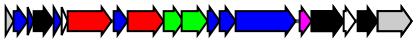
ADOB01000004_c0 Aggregatibacter actinomycetemcomitans D17P-2 [EGY76112-EGY76129]



AJMG01000058_c0 Aggregatibacter actinomycetemcomitans serotype c str. AAS4A [ELT51330-ELT51347]



AJMH01000048_c0 Aggregatibacter actinomycetemcomitans serotype b str. S23A [ELT57684-ELT57702]



AEJK01000042_c0 Aggregatibacter actinomycetemcomitans serotype a str. H5P1 [EGY41037-EGY41053]

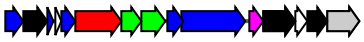
AEJO01000028_c0 Aggregatibacter actinomycetemcomitans serotype f str. D18P1 [EGY42136-EGY42152]

AEJQ01000049_c0 Aggregatibacter actinomycetemcomitans serotype b str. I23C [EGY50823-EGY50844]

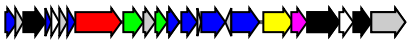
AMEN01000009_c0 Aggregatibacter actinomycetemcomitans Y4 [EKX98651-EKX98669]

AEJR01000092_c0 Aggregatibacter actinomycetemcomitans serotype c str. SCC2302 [EGY44811-EGY44829]

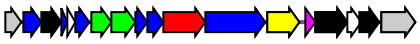
HE798385_c3 Achromobacter xylosoxidans NH44784-1996 complete genome. [CCH07903-CCH07917]



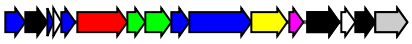
CP003555_c2 Advenella kashmirensis WT001 [AFK64538-AFK64560]



ADCY02000011_c0 Simonsiella muelleri ATCC 29453 [EFG31382-EJZ50251]



AEQP01000024_c0 Lautropia mirabilis ATCC 51599 [EFV93721-EFV93736]



AHGR01000004_c1 Aggregatibacter actinomycetemcomitans RhAA1 [EHK90689-EHK90704]

ADOA01000301_c0 Aggregatibacter actinomycetemcomitans D17P-3 [EGY69384-EGY69400]

AEJM01000020_c0 Aggregatibacter actinomycetemcomitans serotype e str. SC1083 [EGY33908-EGY33926]

AFUV01000022_c0 Haemophilus pittmaniae HK 85 [EGV04971-EGV05012]

CP003099_c1 Aggregatibacter actinomycetemcomitans ANH9381 [AEW76841-AEW76858]

AJMF01000067_c0 Aggregatibacter actinomycetemcomitans serotype b str. SCC4092 [ELT56141-ELT56158]

AALE02000025_c0 Yersinia frederiksenii ATCC 33641 [EEQ13475-EEQ13489]

AALF02000030_c0 Yersinia intermedia ATCC 29909 [EEQ17791-EEQ17804]

FO203355_c2 Enterobacter aerogenes EA1509E complete genome. [CCG32865-CCG32881]

AP012032_c0 Pantoea ananatis AJ13355 DNA [BAK13773-BAK13787]

CBJA010000174_c0 Sutterella sp. CAG:351 [CDE49044-CDE49067]

AGJF01000002_c0 Opitutaceae bacterium TAV5 [EHP33933-EHP33947]

FP929040_c2 Enterobacter cloacae subsp. cloacae NCTC 9394 draft genome. [CBK86164-CBK86177]

BADY01000095_c0 Pseudoalteromonas sp. BSi20495 [GAA80316-GAA80336]

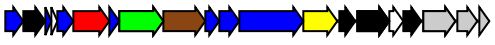
CP000447_c0 Shewanella frigidimarina NCIMB 400 [ABI70200-ABI70217]

CP000931_c1 Shewanella halifaxensis HAW-EB4 [ABZ74941-ABZ74958]

CP000472_c1 Shewanella piezotolerans WP3 [ACJ27126-ACJ27144]

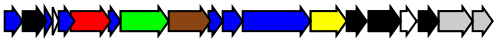
ABIC01000019_c0 Shewanella benthica KT99 [EDQ00533-EDQ00550]

CP000821_c0 Shewanella sediminis HAW-EB3 [ABV34921-ABV34940]

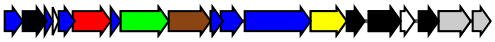


CP000469_c1 Shewanella sp. ANA-3 chromosome 1 [ABK46568-ABK46589]

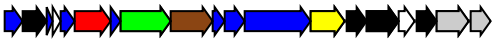
EP901500_c0 marine metagenome JCVI_SCAF_1096627205315 genomic scaffold [EDA30068-EDA30124]



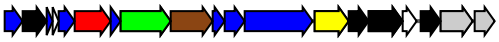
CP000446_c1 Shewanella sp. MR-4 [ABI37414-ABI37432]



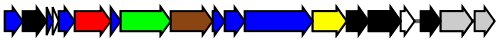
CP002457_c0 Shewanella putrefaciens 200 [ADV52781-ADV52799]



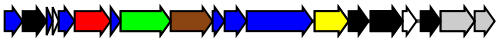
AGEX01000008_c0 Shewanella baltica OS625 [EHC05108-EHC05126]



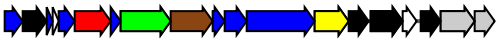
CP002811_c0 Shewanella baltica OS117 [AEH12226-AEH12244]



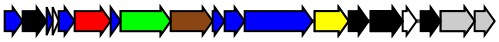
CM001435_c0 Shewanella baltica OS183 chromosome [EHQ13314-EHQ13332]



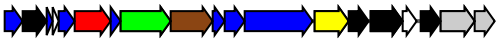
CP001252_c1 Shewanella baltica OS223 [ACK44869-ACK44887]



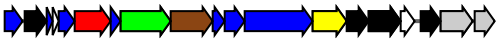
CP000891_c1 Shewanella baltica OS195 [ABX47512-ABX47530]



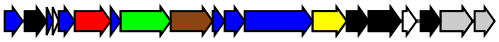
CP002383_c1 Shewanella baltica OS678 [ADT92538-ADT92556]



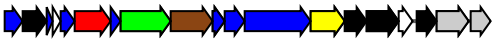
CP000563_c0 Shewanella baltica OS155 [ABN59859-ABN59877]



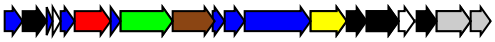
CP000753_c1 Shewanella baltica OS185 [ABS06494-ABS06512]



CP000503_c0 Shewanella sp. W3-18-1 [ABM23135-ABM23153]

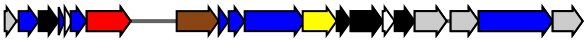


CP000681_c1 Shewanella putrefaciens CN-32 [ABP74162-ABP74180]



AMRQ01000009_c0 Thalassospira xiamenensis M-5 = DSM 17429 strain M-5 [EKF12238-EKF12255]

CP000774_c0 Parvibaculum lavamentivorans DS-1 [ABS63105-ABS63124]



CP002039_c1 Herbaspirillum seropedicae SmR1 [ADJ65179-ADJ65197]

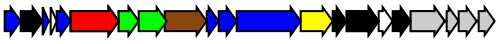
FP929140_c1 gamma proteobacterium HdN1 complete genome. [CBL46048-CBL46071]

CP000749_c0 Marinomonas sp. MWYL1 [ABR72772-ABR72788]

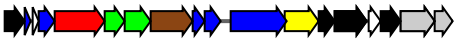
CP002771_c0 Marinomonas posidonica IVIA-Po-181 [AEF53605-AEF53619]

BAFK01000018_c0 Rheinheimera nanhaiensis E407-8 [GAB59852-GAB59872]

AKJW01000059_c0 Herbaspirillum sp. CF444 [EJL88278-EJL88298]

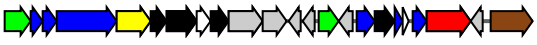


FJ965829_c0 Collimonas sp. MPS11E8. [ADU90641-ADU90658]



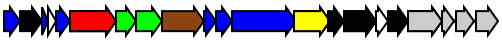
AKCV01000021_c0 Ralstonia sp. PBA [EIZ04091-EIZ04110]

CP001678_c0 Hirschia baltica ATCC 49814 [ACT58985-ACT59007]



CP002013_c2 Burkholderia sp. CCGE1002 chromosome 1 [ADG16323-ADG16339]

AKJA01000017_c0 Herbaspirillum sp. YR522 [EJN09274-EJN09294]



AM743169_c1 Stenotrophomonas maltophilia K279a complete genome, strain K279a. [CAQ47911-CAQ47927]

JH109153_c1 Methylobacter tundripaludum SV96 genomic scaffold Mettuscaffold_2 [EGW20157-EGW20176]

HE798556_c0 Stenotrophomonas maltophilia D457 complete genome. [CCH14596-CCH14613]

AMXM01000010_c0 Stenotrophomonas maltophilia EPM1 [EMF60029-EMF60046]

CP002986_c0 Stenotrophomonas maltophilia JV3 [AEM53266-AEM53283]

CP001111_c1 Stenotrophomonas maltophilia R551-3 [ACF53606-ACF53624]

ALOG01000019_c0 Stenotrophomonas maltophilia Ab55555 [EJP78728-EJP78746]

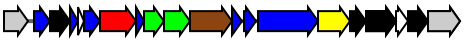
CP004025_c0 Myxococcus stipitatus DSM 14675 [AGC44608-AGC44626]

ASTJ01000036_c0 Halomonas anticariensis FP35 = DSM 16096 strain FP35 [EPC01160-EPC01178]

CP003746_c0 Simiduia agarivorans SA1 = DSM 21679 [AFU98575-AFU98597]

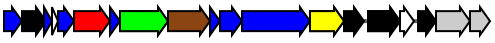
APME01000006_c0 Pseudoalteromonas agarivorans S816 [ENO00274-ENO00293]

ADOP01000007_c0 Pseudoalteromonas haloplanktis ANT/505 [EGI74423-EGI74442]

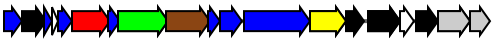


BADV01000015_c0 Pseudoalteromonas sp. BSi20429 [GAA66502-GAA66523]

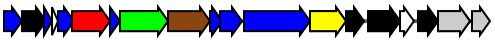
AE014299_c5 Shewanella oneidensis MR-1 [AAN57334-AAN57352]



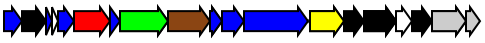
EQ082401_c0 marine metagenome JCVI_SCAF_1096627386215 genomic scaffold [ECV22695-ECV22714]



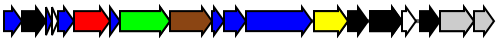
CP000444_c5 Shewanella sp. MR-7 [ABI44654-ABI44672]



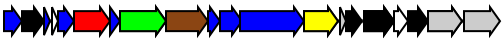
AFOZ01000044_c0 Shewanella sp. HN-41 [EGM68205-EGM68223]



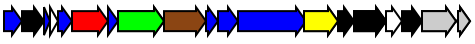
CP002767_c5 Shewanella baltica BA175 [AEG13123-AEG13141]



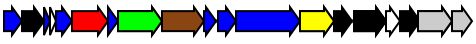
CP000961_c4 Shewanella woodyi ATCC 51908 [ACA88844-ACA88863]



CP000606_c3 Shewanella loihica PV-4 [ABO25375-ABO25393]



CP000507_c4 Shewanella amazonensis SB2B [ABM01580-ABM01598]



AMRI01000004_c0 Gallaecimonas xiamenensis 3-C-1 [EKE76651-EKE76671]

AP011177_c1 Shewanella violacea DSS12 DNA [BAJ04061-BAJ04078]

CP000851_c4 Shewanella pealeana ATCC 700345 [ABV89196-ABV89213]

AMRG01000009_c0 Idiomarina xiamenensis 10-D-4 [EKE83410-EKE83428]

CP001707_c3 Kangiella koreensis DSM 16069 [ACV27980-ACV27997]

CR954247_c0 Pseudoalteromonas haloplanktis str. TAC125 chromosome II [CAI89428-CAI89448]

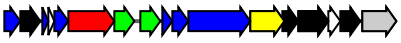
CP002745_c0 Collimonas fungivorans Ter331 [AEK61718-AEK61736]

AE017340_c0 Idiomarina loihiensis L2TR [AAV80968-AAV80987]

CP005964_c0 Idiomarina loihiensis GSL 199 [AGM34992-AGM35011]

AJVC01000004_c0 Herbaspirillum sp. GW103 [EIJ46038-EIJ46057]

BAER01000060_c0 Glaciecola polaris LMG 21857 [GAC33406-GAC33422]



CP003389_c5 Corallococcus coralloides DSM 2259 [AFE06615-AFE10737]

AQPZ01000017_c0 Ralstonia sp. GA3-3 [EON20095-EON20113]

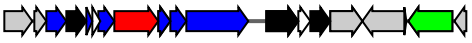
CP000353_c0 Cupriavidus metallidurans CH34 megaplasmid [ABF11241-ABF11258]

AKJX01000119_c0 Acidovorax sp. CF316 [EJE52268-EJE52285]

BAEN01000015_c0 Glaciecola lipolytica E3 [GAC13124-GAC13143]

AQOU01000042_c0 Pandoraea sp. SD6-2 [EON11484-EON11500]

HE577327_c0 Azospirillum brasilense Sp245 main chromosome complete genome. [CCC97284-CCC97302]



DS999312_c0 Francisella philomiragia subsp. philomiragia ATCC 25015 supercont1.5 genomic scaffold [EET20836-EET20852]

CP000937_c0 Francisella philomiragia subsp. philomiragia ATCC 25017 [ABZ87301-ABZ87318]

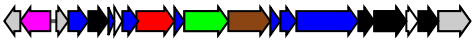
ACZR01000006_c0 Pasteurella dagmatis ATCC 43325 [EEX50718-EEX50735]

CP002872_c1 Francisella sp. TX077308 [AEI35601-AEI35619]

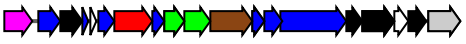
CP000934_c0 Cellvibrio japonicus Ueda107 [ACE82656-ACE86309]

AQGQ01000013_c0 Aeromonas molluscorum 848 [EOD56383-EOD56400]

CP004404_c0 Psychromonas sp. CNPT3 [AGH79953-AGH79972]



ABCQ01000003_c0 Moritella sp. PE36 [EDM68716-EDM68734]



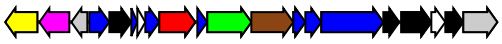
AFWH01000024_c0 Vibrio orientalis CIP 102891 = ATCC 33934 strain CIP 102891 [EGU50402-EGU50418]

ARYZ01000082_c0 Aeromonas salmonicida subsp. pectinolytica 34mel [EQC02816-EQC02832]

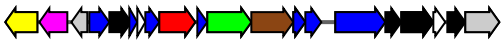
AMPD01000037_c0 Vibrio alginolyticus E0666 [EMD78370-EMD78385]

FM178379_c0 Aliivibrio salmonicida LFI1238 chromosome 1 complete genome. [CAQ79775-CAQ79793]

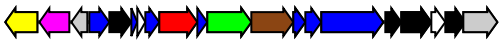
AHIH01000003_c0 Vibrio fischeri SR5 [EHN71078-EHN71097]



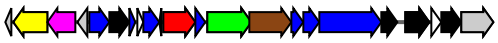
CP000020_c0 Vibrio fischeri ES114 chromosome I [AAW85336-AAW85355]



CP001139_c0 Vibrio fischeri MJ11 chromosome I [ACH65227-ACH67203]



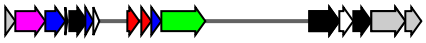
DS999345_c0 Vibrio sp. 16 scf_1108854221887 genomic scaffold [EED25983-EED26052]



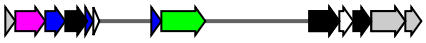
DF093600_c0 Photobacterium leiognathi subsp. mandapamensis svers.1.1. DNA, Scaffold08, whole genome shotgun sequence. [GAA05732-GAA05753]

AAOJ01000001_c0 Photobacterium angustum S14 [EAS66692-EAS66712]

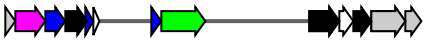
AERR01000015_c0 Escherichia coli O157:H7 str. 1125 [EGD68441-EGD68456]



AERQ01000011_c0 Escherichia coli O157:H7 str. EC1212 [EFW66147-EFW66159]



AERP01000070_c0 Escherichia coli O157:H7 str. 1044 [EGD61380-EGD61392]



EP887286_c0 marine metagenome JCVI_SCAF_1096627191101 genomic scaffold [EDA70219-EDA70240]

AGWR01000003_c0 Aeromonas hydrophila SSU [EKB29774-EKB29793]

CP000644_c0 Aeromonas salmonicida subsp. salmonicida A449 [ABO89039-ABO89060]

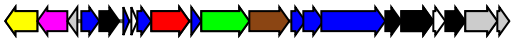
AGVO01000058_c0 Aeromonas salmonicida subsp. salmonicida 01-B526 [EHI51538-EHI51559]

ACZB01000079_c0 Vibrio alginolyticus 40B [EEZ81726-EEZ81746]

AEVT01000018_c0 Vibrio sinaloensis DSM 21326 [EGA71639-EGA71659]

ACZV01000004_c0 Vibrio orientalis CIP 102891 = ATCC 33934 strain CIP 102891 [EEX94525-EEX94546]

AFWG01000022_c0 Vibrio splendidus ATCC 33789 [EGU42738-EGU42758]

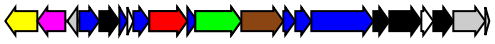


BA000031_c0 Vibrio parahaemolyticus RIMD 2210633 DNA, chromosome 1, complete sequence. [BAC59144-BAC59164]

ACFN01000131_c0 Vibrio parahaemolyticus AQ4037 [EFO46161-EFO46205]

CP006004_c0 Vibrio parahaemolyticus O1:Kuk str. FDA_R31 chromosome I [AGQ90428-AGQ92049]

CP001805_c2 Vibrio sp. Ex25 chromosome 1 [ACY52219-ACY52239]



ASXT01000002_c0 Vibrio fluvialis I21563 [EPP27825-EPP27844]

AAND01000007_c0 Vibrio sp. MED222 [EAQ54247-EAQ54266]

CP002469_c2 Vibrio vulnificus MO6-24/O chromosome I [ADV87164-ADV87183]

AE016795_c0 Vibrio vulnificus CMCP6 chromosome I [AAO08584-AAO08603]

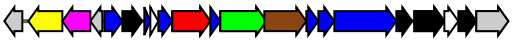
AAPS01000033_c0 Vibrio alginolyticus 12G01 [EAS74689-EAS74708]

ACKB01000116_c0 Vibrio parahaemolyticus K5030 [EFO52887-EFO52927]

AFBW01000012_c0 Vibrio parahaemolyticus 10329 [EGF44192-EGF44211]

AAWQ01000027_c0 Vibrio parahaemolyticus AQ3810 [EDM59489-EDM59528]

CP006008_c2 Vibrio parahaemolyticus O1:K33 str. CDC_K4557 chromosome I [AGQ98870-AGR00210]

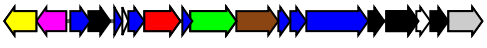


ACFM01000029_c0 Vibrio parahaemolyticus Peru-466 [EFO36448-EFO36487]

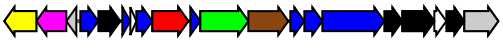
CP000462_c2 Aeromonas hydrophila subsp. hydrophila ATCC 7966 [ABK36231-ABK39644]

CP005966_c1 Aeromonas hydrophila ML09-119 [AGM45378-AGM45397]

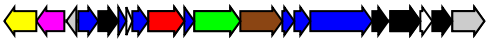
CP002284_c1 Vibrio anguillarum 775 chromosome I [AEH32794-AEH32812]



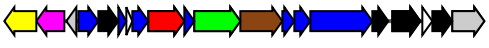
AAZW01000009_c0 Vibrionales bacterium SWAT-3 [EDK29441-EDK29460]



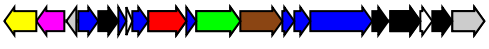
AHHQ01000018_c0 Vibrio harveyi CAIM 1792 [EMR37797-EMR37816]



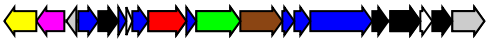
AJSQ01000019_c0 Vibrio sp. HENC-01 [EKM24821-EKM24840]



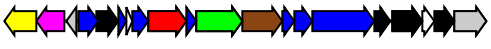
CP003972_c0 Vibrio parahaemolyticus BB22OP chromosome 1 [AGB09344-AGB09363]



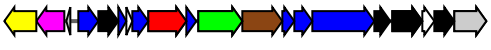
ACFO01000003_c0 Vibrio parahaemolyticus AN-5034 [EFO42486-EFO42516]



ACZP01000013_c0 Vibrio furnissii CIP 102972 [EEX41288-EEX41307]

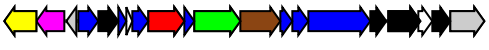


ASXS01000009_c1 Vibrio fluvialis PG41 [EPP22687-EPP22706]

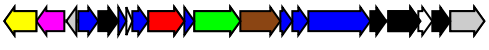


AEVS01000051_c0 Vibrio brasiliensis LMG 20546 [EGA66014-EGA66033]

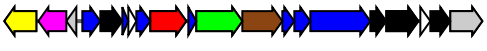
AFWI01000145_c0 Vibrio tubiashii ATCC 19109 [EGU55389-EGU55408]



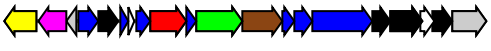
AHHF01000096_c0 Vibrio tubiashii NCIMB 1337 = ATCC 19106 strain NCIMB 1337 [EIF02011-EIF02030]



ANFM02000012_c0 Grimontia sp. AK16 [EOD80564-EOD80583]



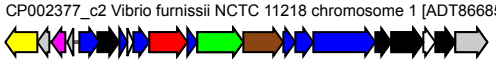
BA000037_c0 Vibrio vulnificus YJ016 DNA, chromosome I, complete sequence. [BAC93830-BAC93849]



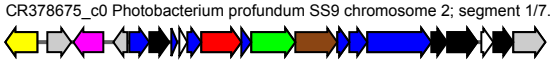
ABCH01000017_c0 Vibrio shilonii AK1 [EDL53458-EDL53476]

AAOU01000055_c0 Photobacterium sp. SKA34 [EAR53454-EAR53472]

CP002377_c2 Vibrio furnissii NCTC 11218 chromosome 1 [ADT86685-ADT86705]



CR378675_c0 Photobacterium profundum SS9 chromosome 2; segment 1/7. [CAG21977-CAG21997]



AAPH01000001_c0 Photobacterium profundum 3TCK [EAS45461-EAS45481]

FM954972_c1 Vibrio splendidus LGP32 chromosome 1. [CAV17752-CAV17774]

AAMR01000020_c0 Vibrio splendidus 12B01 [EAP93995-EAP94015]

CALB01000097_c0 Cronobacter turicensis 564 [CCJ90332-CCJ90349]

CP001790_c2 Pectobacterium wasabiae WPP163 [ACX90253-ACX90270]

CM001230_c0 Brenneria sp. EniD312 chromosome [EHD23765-EHD23782]

CP001836_c1 Dickeya dadantii Ech586 [ACZ78990-ACZ79007]

CP003415_c3 Pectobacterium sp. SCC3193 [AFI92791-AFI92808]

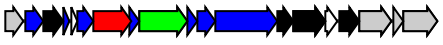
FP236843_c0 Erwinia billingiae strain Eb661 complete chromosome. [CAX59118-CAX59134]

AEDL01000006_c0 Pantoea sp. aB [EFM19674-EFM19690]

CP002206_c0 Pantoea vagans C9-1 [ADO09028-ADO09044]

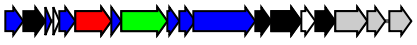
CP003938_c0 Enterobacteriaceae bacterium strain FGI 57 [AGB76883-AGB76899]

CP002521_c0 Acidovorax avenae subsp. avenae ATCC 19860 [ADX45298-ADX45316]

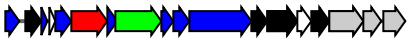


ANKX01000019_c0 Pantoea agglomerans 299R [ELP26058-ELP26073]

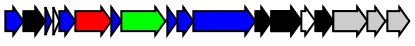
AJHJ01000027_c0 Serratia sp. M24T3 [EIC83206-EIC83223]



CP003403_c0 Rahnella aquatilis HX2 [AFE57102-AFE57119]



CP002505_c0 Rahnella sp. Y9602 [ADW72521-ADW72538]



CP002775_c0 Serratia sp. AS13 [AEG26624-AEG26640]

CP002773_c0 Serratia plymuthica AS9 [AEF43964-AEF43980]

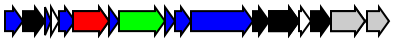
CP002774_c0 Serratia sp. AS12 [AEF48916-AEF48932]

ADBY01000013_c0 Serratia odorifera DSM 4582 [EFE98049-EFE98067]

AJTB01000101_c0 Serratia plymuthica PRI-2C [EIJ17890-EIJ17907]

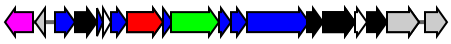
CP003244_c1 Rahnella aquatilis CIP 78.65 = ATCC 33071 [AEX50859-AEX50875]

CP003942_c0 Serratia marcescens FGI94 [AGB81188-AGB81204]



AKKR01000075_c0 Yersinia enterocolitica subsp. enterocolitica WA-314 [EKA26751-EKA26769]

AM286415_c0 Yersinia enterocolitica subsp. enterocolitica 8081 complete genome. [CAL10997-CAL11015]



CP000826_c0 Serratia proteamaculans 568 [ABV39993-ABV40010]

CP006252_c0 Serratia liquefaciens ATCC 27592 [AGQ29572-AGQ29589]

CP006566_c0 Serratia plymuthica S13 [AGP43101-AGP46803]

CP006250_c0 Serratia plymuthica 4Rx13 [AGO53741-AGO53758]

AOBV01000007_c0 Wohlfahrtiimonas chitiniclastica SH04 [ELV08035-ELV08053]

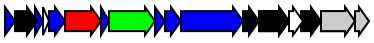
BX950851_c2 Erwinia carotovora subsp. atroseptica SCRI1043 [CAG77386-CAG77402]

AJFP01000001_c1 Pantoea sp. Sc1 [EIB99256-EIB99272]

AKIT01000025_c0 Pantoea sp. YR343 [EJN02598-EJN02614]

AKIU01000028_c0 Pantoea sp. GM01 [EJL90506-EJL90522]

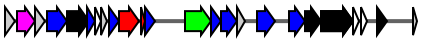
AJXP01000013_c0 Enterobacter cloacae subsp. cloacae GS1 [EIM35114-EIM35130]



CP004142_c2 Raoultella ornithinolytica B6 [AGJ88710-AGJ88727]

CP002272_c0 Enterobacter lignolyticus SCF1 [ADO46532-ADO46549]

CALA01000393_c1 Cronobacter dublinensis 582 [CCJ86615-CCJ86639]

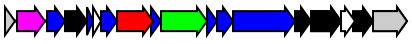


ANXF01000010_c0 Escherichia coli KTE215 [ELH82795-ELH82810]

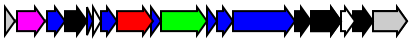
AGDM01000048_c0 Klebsiella oxytoca 10-5246 [EHT05026-EHT05043]

AKNF01000067_c0 Shigella flexneri 1235-66 strain 1236-66 [EIQ79426-EIQ79445]

CP001135_c0 Edwardsiella tarda EIB202 [ACY83668-ACY83685]



CP004141_c0 Edwardsiella tarda C07-087 [AGH72922-AGH72939]



FN543502_c1 Citrobacter rodentium ICC168 [CBG91021-CBG91039]

ADLG01000006_c0 Citrobacter freundii 4_7_47CFAA [EHL85574-EHL85592]

AKYD01000006_c1 Enterobacter radicincitans DSM 16656 [EJI92556-EJI93039]

AGCL01000032_c0 Yokenella regensburgei ATCC 43003 [EHM48770-EHM48788]

ATCK01000059_c0 Enterobacter cloacae str. Hanford [EPR32599-EPR32617]

ATHX01000010_c0 Enterobacter cloacae EC_38VIM1 [EPY95414-EPY95523]

CP002154_c0 Edwardsiella tarda FL6-60 [ADM40889-ADM40906]

CP002824_c0 Enterobacter aerogenes KCTC 2190 [AEG96044-AEG96060]

CP001918_c2 Enterobacter cloacae subsp. cloacae ATCC 13047 [ADF64367-ADF64383]

CP003678_c3 Enterobacter cloacae subsp. dissolvens SDM [AFM62009-AFM62025]

AFHR01000070_c0 Enterobacter hormaechei ATCC 49162 [EGK57308-EGK57324]

AKVS01000073_c0 Pectobacterium wasabiae CFBP 3304 [EJS92093-EJS92111]

CP000653_c2 Enterobacter sp. 638 [ABP62537-ABP62556]

CU928158_c2 Escherichia fergusonii ATCC 35469 chromosome [CAQ90926-CAQ90944]

CU928163_c1 Escherichia coli UMN026 chromosome [CAR15084-CAR15102]

CP001164_c1 Escherichia coli O157:H7 str. EC4115 [ACI34636-ACI39646]

ABHP01000001_c0 Escherichia coli O157:H7 str. EC4113 [EDU56375-EDU56600]

ABHO01000001_c0 Escherichia coli O157:H7 str. EC4196 [EDU35216-EDU35661]

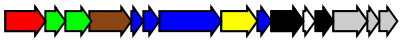
CP001619_c2 Dyadobacter fermentans DSM 18053 [ACT96977-ACT96991]

CM001403_c0 Mucilaginibacter paludis DSM 18603 chromosome [EHQ30840-EHQ30858]

AMDQ01000104_c0 Acinetobacter baumannii OIFC180 [EKL42876-EKL42946]

APBF01000038_c0 Acinetobacter baumannii ABNIH17 [EMU34284-EMU34298]

ACYS02000039_c0 Acinetobacter baumannii 6014059 [EGJ68701-EGJ68715]

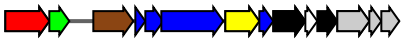


APBE01000048_c0 Acinetobacter baumannii ABNIH16 [EMU20424-EMU20438]

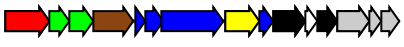
APBD01000028_c0 Acinetobacter baumannii ABNIH15 [EMU20326-EMU20340]

APBC01000030_c0 Acinetobacter baumannii ABNIH14 [EMU14064-EMU14078]

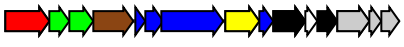
AOGD01000034_c0 Acinetobacter baumannii ABNIH26 [EMT90858-EMT90871]



AFTA01000105_c0 Acinetobacter baumannii ABNIH2 [EGT89678-EGT89692]



APBB01000030_c0 Acinetobacter baumannii ABNIH13 [EMU10532-EMU10546]

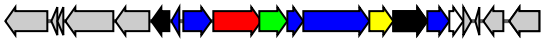


AKJB01000073_c0 Pseudomonas sp. GM102 [EJL98438-EJL98454]

CP002727_c0 Pseudomonas fulva 12-X [AEF20417-AEF20450]

CP002620_c1 Pseudomonas mendocina NK-01 [AEB56627-AEB56660]

CP000378_c1 Burkholderia cenocepacia AU 1054 chromosome 1 [ABF77040-ABF77059]



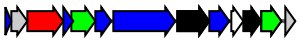
CP000458_c1 Burkholderia cenocepacia HI2424 chromosome 1 [ABK09502-ABK09522]

CP003093_c0 Pseudoxanthomonas spadix BD-a59 [AER54729-AER54745]

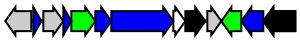
FP565176_c0 Xanthomonas albilineans GPE PC73 complete genome. [CBA14671-CBA14688]

CP003154_c2 Thiocystis violascens DSM 198 [AFL74194-AFL74209]

CP001905_c0 Thioalkalivibrio sp. K90mix [ADC70985-ADC70997]



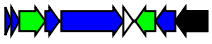
AJXV01000044_c0 Rhodanobacter sp. 116-2 [EIM00799-EIM00811]



CP003470_c3 Rhodanobacter sp. 2APBS1 [AGG90814-AGG90829]

AJXU01000065_c0 Rhodanobacter fulvus Jip2 [EIL88016-EIL88028]

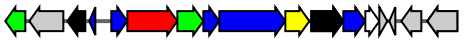
CP003350_c0 Frateuria aurantia DSM 6220 [AFC84575-AFC84583]



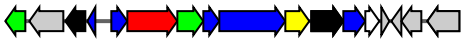
AE017282_c0 Methylococcus capsulatus str. Bath [AAU92800-AAU92829]

AM747720_c0 Burkholderia cenocepacia J2315 chromosome 1 [CAR51138-CAR51155]

CP001025_c2 Burkholderia ambifaria MC40-6 chromosome 1 [ACB65152-ACB65168]



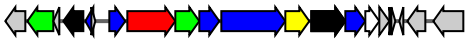
CP000440_c2 Burkholderia ambifaria AMMD chromosome 1 [ABI88364-ABI88380]



ALJA02000005_c1 Burkholderia cenocepacia K56-2Valvano [EPZ90960-EPZ91589]

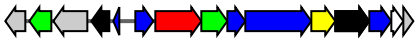
CP000614_c1 Burkholderia vietnamiensis G4 chromosome 1 [ABO55851-ABO55868]

CP003514_c2 Burkholderia sp. KJ006 chromosome 1 [AFJ86874-AFJ86894]



CP003774_c0 Burkholderia cepacia GG4 chromosome 1 [AFQ47121-AFQ47135]

CP000151_c1 Burkholderia sp. 383 chromosome 1 [ABB09679-ABB09693]



CAHW01000040_c0 Ralstonia solanacearum MolK2 [CAQ56716-CAQ56731]

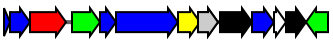
KE392123_c0 Ralstonia solanacearum SD54 scaffold6 [EPR98166-EPR98179]

CP004012_c0 Ralstonia solanacearum FQY_4 [AGH82969-AGH82982]

FR854088_c0 Ralstonia syzygii R24, genomic contig 00003-1628. [CCA88790-CCA88803]

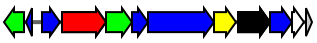
AL646052_c0 Ralstonia solanacearum GMI1000 chromosome complete sequence. [CAD13952-CAD13964]

CR555306_c2 Azoarcus aromaticum EbN1 complete genome. [CAI10128-CAI10140]

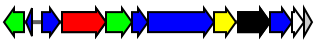


FR854068_c0 blood disease bacterium R229, genomic contig 00012-1626. [CCA81068-CCA81080]

ACTT02000004_c0 Ralstonia sp. 5_2_56FAA [EGY66057-EGY66068]



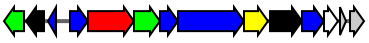
APMQ01000011_c0 Ralstonia pickettii OR214 [ENZ76235-ENZ76246]



CP000091_c0 Ralstonia eutropha JMP134 chromosome 2 [AAZ63282-AAZ63294]

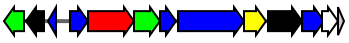
CP002819_c2 Ralstonia solanacearum Po82 [AEG70254-AEG70265]

FP885897_c2 Ralstonia solanacearum CFBP2957 chromosome complete genome. [CBJ44226-CBJ44239]

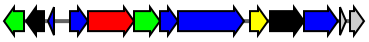


FP885895_c2 Ralstonia solanacearum CMR15 chromosome [CBJ39218-CBJ39230]

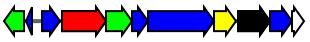
FP885906_c2 Ralstonia solanacearum str. PSI07 chromosome [CBJ52298-CBJ52310]



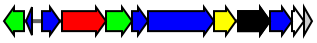
CU914168_c0 Ralstonia solanacearum strain IPO1609 Genome Draft. [CAQ60588-CAQ60600]



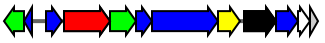
CP001068_c0 Ralstonia pickettii 12J chromosome 1 [ACD25460-ACD25470]



CP001644_c0 Ralstonia pickettii 12D chromosome 1 [ACS61618-ACS61629]

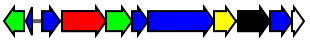


ASZV01000046_c0 Ralstonia sp. AU12-08 [EPX95721-EPX95732]

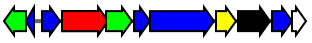


AAKL01000009_c0 Ralstonia solanacearum UW551 [EAP73845-EAP73855]

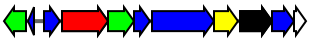
ACUF01000024_c0 Ralstonia sp. 5_7_47FAA [EFP66887-EFP66897]



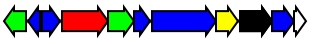
CU633750_c0 Cupriavidus taiwanensis str. LMG19424 chromosome 2 [CAQ72064-CAQ72074]



AQPZ01000012_c0 Ralstonia sp. GA3-3 [EON20573-EON20583]



CP002878_c0 Cupriavidus necator N-1 chromosome 2 [AEI80608-AEI80618]



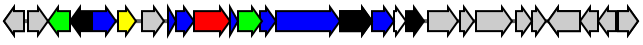
AM260480_c0 Ralstonia eutropha H16 chromosome 2. [CAJ96442-CAJ96454]

CP000529_c0 Polaromonas naphthalenivorans CJ2 [ABM36681-ABM36693]

CP001715_c1 Candidatus Accumulibacter phosphatis clade IIA str. UW-1 [ACV33718-ACV33733]

AP012304_c0 Azoarcus sp. KH32C DNA [BAL24083-BAL24097]

AGEZ01000056_c0 Acetobacteraceae bacterium AT-5844 [EHM01873-EHM01899]



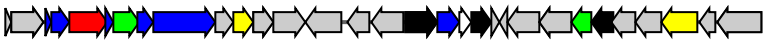
AJWL01000118_c0 Hydrogenophaga sp. PBC [EIK88477-EIK88496]

ACIS01000009_c0 Pseudogulbenkiania ferrooxidans 2002 [EEG07341-EEG07361]

AP012224_c0 Pseudogulbenkiania sp. NH8B DNA [BAK76073-BAK76093]

FP475956_c1 Thiomonas sp. str. 3As chromosome [CAZ89217-CAZ89235]

CP000089_c2 Dechloromonas aromatica RCB [AAZ48895-AAZ48925]



CP000116_c1 Thiobacillus denitrificans ATCC 25259 [AAZ98695-AAZ98719]

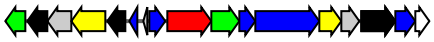
CP000245_c0 Ramlibacter tataouinensis TTB310 [AEG91316-AEG91333]

CP000555_c0 Methylibium petroleiphilum PM1 [ABM94278-ABM94295]

AJHK02000012_c0 Burkholderia sp. SJ98 [EKS70027-EKS70045]

CP003090_c0 Burkholderia sp. YI23 plasmid byi_1p [AET93953-AET93968]

AP013061_c1 Burkholderia sp. RPE64 plasmid p1 DNA [BAN27666-BAN27682]



BAFJ01000008_c0 Sulfuricella denitrificans skB26 [GAB72826-GAB72845]

CABP01000033_c0 Mine drainage metagenome [CBI03870-CBI03888]

CP001219_c0 Acidithiobacillus ferrooxidans ATCC 23270 [ACK77996-ACK80936]

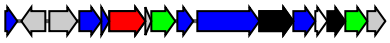
CP002985_c2 Acidithiobacillus ferrivorans SS3 [AEM48911-AEM48926]

CP001132_c0 Acidithiobacillus ferrooxidans ATCC 53993 [ACH83178-ACH83191]

CP001635_c1 Variovorax paradoxus S110 chromosome 1 [ACS19294-ACS19307]

CU207211_c2 Herminiimonas arsenicoxydans chromosome [CAL62975-CAL62996]

AFWT01000027_c0 Thiorhodococcus drewsii AZ1 [EGV29108-EGV29123]



AJGB01000028_c0 Kingella kingae PYKK081 [EIC13698-EIC13714]

ACRG01000003_c2 Neisseria mucosa C102 [EFV81388-EFV81398]

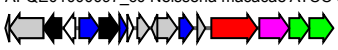
ACQV01000024_c0 Neisseria flavescens SK114 [EER55857-EER56044]

ACEO02000005_c0 Neisseria subflava NJ9703 [EFC52190-EFC52200]

AJMT01000060_c0 Neisseria sicca VK64 [EIG29576-EIG29602]

ACKO02000006_c0 Neisseria sicca ATCC 29256 [EET45001-EET45017]

AFQE01000097_c0 Neisseria macacae ATCC 33926 [EGQ76362-EGQ76378]



ACWS01000066_c0 Neisseria sp. GT4A_CT1 [EGY60422-EGY60436]

CM001488_c0 Desulfobacter postgatei 2ac9 chromosome [EIM63332-EIM63351]

DS267837_c0 Vibrio sp. Ex25 scf_1101759099835 genomic scaffold [EDN56412-EDN56423]

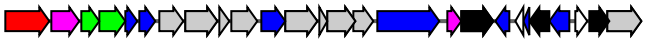
ADBD01000010_c0 Vibrio sp. RC586 [EEY99481-EEY99493]

CP003985_c0 Desulfocapsa sulfexigens DSM 10523 [AGF76704-AGF76740]

CP000267_c0 Rhodoferax ferrireducens T118 [ABD71670-ABD71707]

ACJW02000005_c1 Kingella oralis ATCC 51147 [EEP67023-EEP67070]

AMER01000225_c0 Neisseria sp. oral taxon 020 str. F0370 [EKY02451-EKY02477]



CP003875_c1 Actinobacillus suis H91-0380 [AFU20050-AFU20079]

AEPS01000002_c0 Aggregatibacter segnis ATCC 33393 [EFU68276-EFU68308]

CP000967_c0 Xanthomonas oryzae pv. oryzae PXO99A [ACD57125-ACD57137]

AE013598_c1 Xanthomonas oryzae pv. oryzae KACC 10331 [AAW77566-AAW77578]

AP008229_c1 Xanthomonas oryzae pv. oryzae MAFF 311018 DNA [BAE70818-BAE70829]

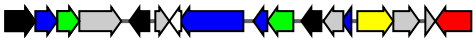
AEQX01000349_c0 Xanthomonas gardneri ATCC 19865 [EGD17711-EGD17722]

CP002446_c0 Pseudoxanthomonas suwonensis 11-1 [ADV28610-ADV28626]

CP004399_c0 Xanthomonas axonopodis Xac29-1 [AGH79504-AGH79521]

CP003057_c0 Xanthomonas oryzae pv. oryzicola BLS256 [AEQ94646-AEQ94673]

ACPY01000189_c0 Xanthomonas fuscans subsp. aurantifolii str. ICPB 10535 [EFF47835-EFF47851]



CP003778_c0 Xanthomonas citri subsp. citri Aw12879 [AGI06022-AGI06041]

AEQW01000005_c0 Xanthomonas perforans 91-118 [EGD16059-EGD16077]

AHIB01000024_c0 Xanthomonas axonopodis pv. malvacearum str. GSPB1386 [EKQ63958-EKQ63976]

CAGJ01000010_c0 Xanthomonas axonopodis pv. punicae str. LMG 859 [CCF67355-CCF67373]

CP002789_c0 Xanthomonas campestris pv. raphani 756C [AEL05176-AEL05194]

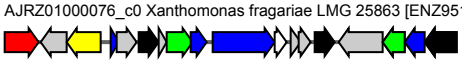
CP002914_c0 Xanthomonas axonopodis pv. citrumelo F1 [AEO44191-AEO44209]

AM920689_c1 Xanthomonas campestris pv. campestris complete genome, strain B100. [CAP53559-CAP53577]

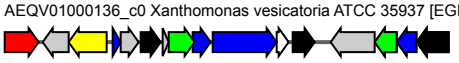
AE008922_c1 Xanthomonas campestris pv. campestris str. ATCC 33913 [AAM43219-AAM43236]

CP000050_c2 Xanthomonas campestris pv. campestris str. 8004 [AAY51126-AAY51143]

AJRZ01000076_c0 Xanthomonas fragariae LMG 25863 [ENZ95116-ENZ95133]



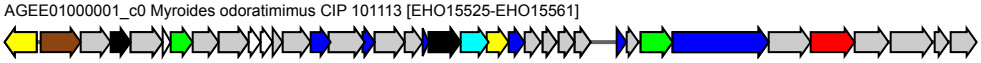
AEQV01000136_c0 Xanthomonas vesicatoria ATCC 35937 [EGD08308-EGD08323]



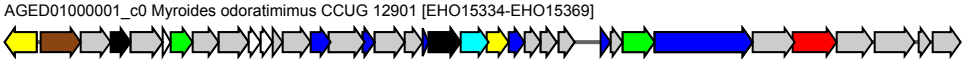
AE008923_c0 Xanthomonas axonopodis pv. citri str. 306 [AAM38921-AAM38940]

AM039952_c2 Xanthomonas campestris pv. vesicatoria complete genome. [CAJ25903-CAJ25923]

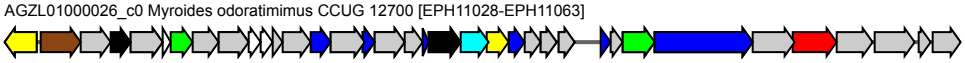
AGEE01000001_c0 Myroides odoratimimus CIP 101113 [EHO15525-EHO15561]



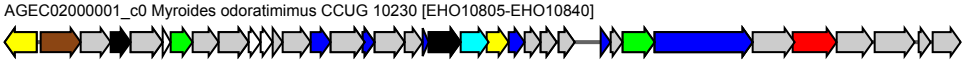
AGED01000001_c0 Myroides odoratimimus CCUG 12901 [EHO15334-EHO15369]



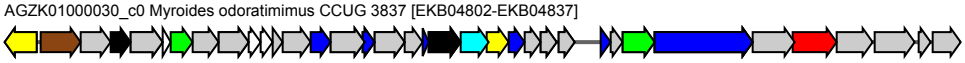
AGZL01000026_c0 Myroides odoratimimus CCUG 12700 [EPH11028-EPH11063]



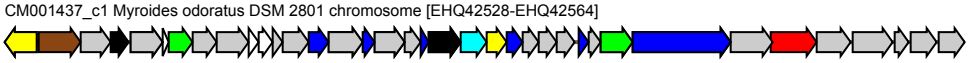
AGEC02000001_c0 Myroides odoratimimus CCUG 10230 [EHO10805-EHO10840]



AGZK01000030_c0 Myroides odoratimimus CCUG 3837 [EKB04802-EKB04837]



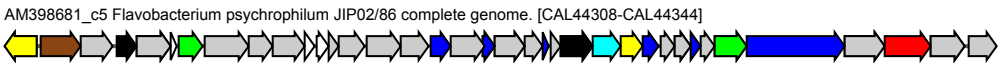
CM001437_c1 Myroides odoratus DSM 2801 chromosome [EHQ42528-EHQ42564]



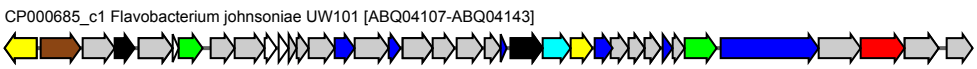
AGZJ01000038_c0 Myroides odoratimimus CIP 103059 [EKB07909-EKB07945]

FP476056_c0 Zobellia galactanivorans strain DsiJT chromosome [CAZ96183-CAZ96240]

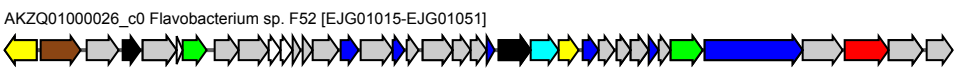
AM398681_c5 Flavobacterium psychrophilum JIP02/86 complete genome. [CAL44308-CAL44344]



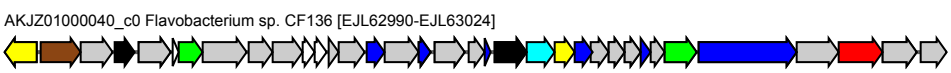
CP000685_c1 Flavobacterium johnsoniae UW101 [ABQ04107-ABQ04143]



AKZQ01000026_c0 Flavobacterium sp. F52 [EJG01015-EJG01051]



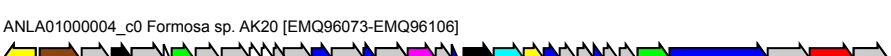
AKJZ01000040_c0 Flavobacterium sp. CF136 [EJL62990-EJL63024]



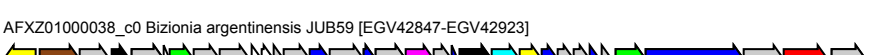
CP003222_c2 Flavobacterium columnare ATCC 49512 [AEW87146-AEW87177]

CP003280_c0 Aequorivita sublithincola DSM 14238 [AFL80411-AFL80443]

ANLA01000004_c0 Formosa sp. AK20 [EMQ96073-EMQ96106]

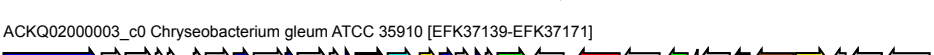


AFXZ01000038_c0 Bizionia argentinensis JUB59 [EGV42847-EGV42923]



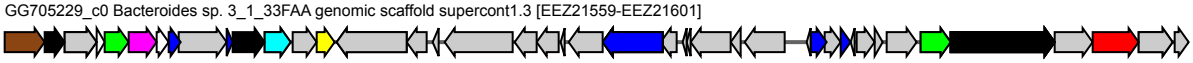
CP002825_c0 Lacinutrix sp. 5H-3-7-4 [AEH01912-AEH01944]

ACKQ02000003_c0 Chryseobacterium gleum ATCC 35910 [EFK37139-EFK37171]



AKJY01000045_c0 Chryseobacterium sp. CF314 [EJL71072-EJL71107]

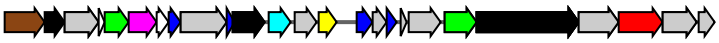
GG705229_c0 Bacteroides sp. 3_1_33FAA genomic scaffold supercont1.3 [EEZ21559-EEZ21601]



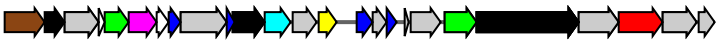
ASSP01000003_c0 Bacteroides massiliensis dnLKV3 [EOS16461-EOS16484]

ACDI02000014_c0 Bacteroides dorei 5_1_36/D4 [EEO45930-EEO45955]

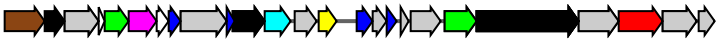
CP000139_c1 Bacteroides vulgatus ATCC 8482 [ABR38703-ABR38727]



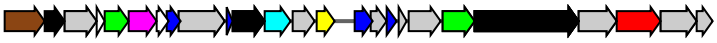
ADKO01000081_c0 Bacteroides vulgatus PC510 [EFG17203-EFG17227]



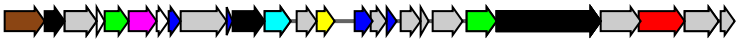
ACDR02000004_c0 Bacteroides sp. 4_3_47FAA [EET14269-EET14293]



ACRT01000010_c0 Bacteroides sp. 3_1_40A [EFV69017-EFV69041]



EQ973126_c0 Bacteroides sp. 9_1_42FAA supercont1.2 genomic scaffold [EEO59889-EEO59914]

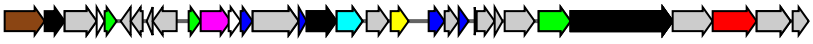


CBDO010000036_c0 Bacteroides dorei CAG:222 [CDB36799-CDB36824]

AGXH01000039_c0 Bacteroides dorei CL02T00C15 [EIY24411-EIY24436]

AGXJ01000027_c0 Bacteroides dorei CL02T12C06 [EIY35825-EIY35850]

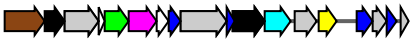
AGXI01000015_c0 Bacteroides dorei CL03T12C01 [EIY37286-EIY37317]



ABWZ01000034_c0 Bacteroides dorei DSM 17855 [EEB25678-EEB25707]

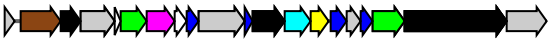
ASSN01000019_c0 Bacteroides vulgatus dnLKV7 [EOS02408-EOS02421]

CBKK010000183_c0 Bacteroides vulgatus CAG:6 [CDF19264-CDF19281]



AGXZ01000015_c0 Bacteroides vulgatus CL09T03C04 [EIY78925-EIY78945]

CBDX010000253_c0 Bacteroides ovatus CAG:22 [CDB58853-CDB58872]



ADLF01000019_c0 Bacteroides oleiciplenus YIT 12058 [EKU88648-EKU88671]

ABJL02000008_c3 Bacteroides intestinalis DSM 17393 [EDV05389-EDV05412]

AGXG01000001_c0 Bacteroides cellulosilyticus CL02T12C19 [EIY40174-EIY40197]

CBHY010000207_c0 Bacteroides intestinalis CAG:315 [CDD93436-CDD93487]

ADLV01000054_c0 Dysgonomonas gadei ATCC BAA-286 [EGJ99689-EGJ99712]

EQ973364_c0 Bacteroides sp. 2_2_4 supercont1.10 genomic scaffold [EEO57331-EEO57353]

ASSM01000015_c0 Bacteroides thetaiotaomicron dnLKV9 [EOR97803-EOR97827]

GG774955_c0 Bacteroides sp. 3_1_23 genomic scaffold supercont1.7 [EFI36881-EFI36904]

ACGA02000059_c0 Bacteroides sp. D2 [EFS30908-EFS30931]

AGXU01000010_c0 Bacteroides ovatus CL03T12C18 [EIY69285-EIY69308]

AGXT01000038_c0 Bacteroides ovatus CL02T12C04 [EIY55272-EIY55295]

AAXF02000041_c0 Bacteroides ovatus ATCC 8483 [EDO13172-EDO13197]

AGXW01000011_c0 Bacteroides finegoldii CL09T03C10 [EKJ90212-EKJ90235]

ACWH01000015_c0 Bacteroides ovatus 3_8_47FAA [EGN02391-EGN02414]

ADMO01000110_c0 Bacteroides ovatus SD CMC 3f [EFF51355-EFF51378]

CBCP010000022_c0 Bacteroides sp. CAG:754 [CDA82670-CDA82694]

AGXE01000002_c0 Bacteroides xylanisolvens CL03T12C04 [EIY88475-EIY88499]

ADCL01000038_c0 Bacteroides sp. 1_1_30 [EGN06482-EGN06506]

CP001619_c0 Dyadobacter fermentans DSM 18053 [ACT93207-ACT93228]

AENZ01000003_c0 Alistipes sp. HGB5 [EFR59028-EFR59112]

JH594606_c0 Gillisia limnaea DSM 15749 genomic scaffold Gilliscaffold_2 [EHQ03901-EHQ03918]

CP001650_c1 Zunongwangia profunda SM-A87 [ADF53082-ADF53099]

CU207366_c2 Gramella forsetii KT0803 complete circular genome. [CAL66624-CAL66642]

CBGS010000073_c0 Alistipes sp. CAG:29 [CDD24765-CDD24784]

CP003274_c1 Alistipes finegoldii DSM 17242 [AFL77872-AFL77892]

AMFJ01033692_c0 Groundwater metagenome [EKD31955-EKD31977]

CP001699_c1 Chitinophaga pinensis DSM 2588 [ACU59349-ACU59373]

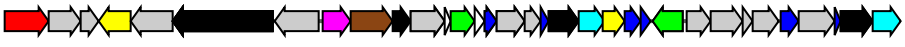
ATMR01000206_c0 Winogradskyella psychrotolerans RS-3 [EPR70008-EPR70032]

CP002046_c0 Croceibacter atlanticus HTCC2559 [EAP87851-EAP87873]

CP003156_c5 Owenweeksia hongkongensis DSM 17368 [AEV32214-AEV32239]

AMZN01000043_c0 Fulvivirga imtechensis AK7 [ELR71286-ELR71315]

CP002542_c0 Fluviicola taffensis DSM 16823 [AEA43416-AEA43448]

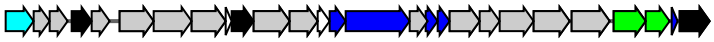


CP002305_c2 Leadbetterella byssophila DSM 17132 [ADQ17186-ADQ17223]

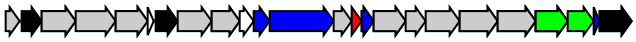
CP002455_c0 Weeksella virosa DSM 16922 [ADX68247-ADX68275]

CP002479_c2 Geobacter sp. M18 [ADW15459-ADW15485]

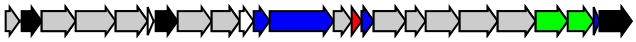
ACJN02000003_c0 Desulfonatronospira thiodismutans ASO3-1 [EFI33459-EFI33485]



CP003221_c6 Desulfovibrio africanus str. Walvis Bay [EGJ51938-EGJ51961]



AOSV01000005_c0 Desulfovibrio africanus PCS [EMG38296-EMG38319]



FO203503_c2 Desulfobacula toluolica Tol2 complete genome. [CCK81989-CCK82019]

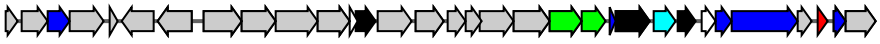
FR695877_c1 Uncultured Desulfobacterium sp. genomic DNA, clone selsg002P03.q1c. [CBX31085-CBX31103]

CP000482_c2 Pelobacter propionicus DSM 2379 [ABK98534-ABK98560]

CR522870_c0 Desulfotalea psychrophila LSv54 chromosome. [CAG36565-CAG36589]

AAEW02000007_c0 Desulfuromonas acetoxidans DSM 684 [EAT15869-EAT15898]

CP001940_c1 Desulfurivibrio alkaliphilus AHT2 [ADH85811-ADH85842]



CBGN010000064_c0 Alistipes sp. CAG:435 [CDD17963-CDD18014]

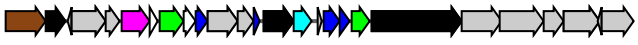
AGSA01000028_c1 Niabella soli DSM 19437 [EHP51820-EHP51839]

CP000383_c4 Cytophaga hutchinsonii ATCC 33406 [ABG59372-ABG59390]

ADCT01000073_c0 Prevotella sp. C561 [EGW47205-EGW47228]

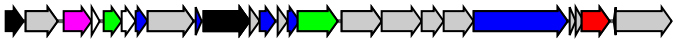
GG739927_c0 Prevotella buccae D17 genomic scaffold supercont1.2 [EFC76438-EFC76461]

ALJQ01000038_c0 Prevotella sp. MSX73 [EJP28514-EJP28771]



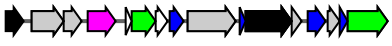
AEPD01000028_c0 Prevotella buccae ATCC 33574 [EFU30391-EFU30416]

AEQO01000205_c0 Prevotella salivae DSM 15606 [EFV03144-EFV03168]

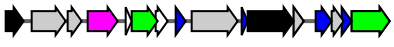


ADFQ01000003_c0 Prevotella amnii CRIS 21A-A [EFN92048-EFN92194]

ACZS01000058_c0 Prevotella sp. oral taxon 472 str. F0295 [EEX53244-EEX53259]



GG740073_c0 Prevotella sp. oral taxon 317 str. F0108 genomic scaffold supercont2.2 [EFC68344-EFC68359]



ADGI01000054_c0 Prevotella oulorum F0390 [EGV30225-EGV30244]

GL349572_c0 Prevotella oris C735 genomic scaffold supercont1.9 [EFI47756-EFI47778]

ACUZ02000036_c0 Prevotella oris F0302 [EFB31484-EFB31508]

CAWS010000052_c0 Bacteroides sp. CAG:598 [CCX61298-CCX61331]

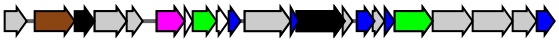
AFXP01000021_c0 Prevotella histicola F0411 [EHG15485-EHG15510]

CBAM010000097_c0 Bacteroides sp. CAG:702 [CCZ69614-CCZ69633]

CP002589_c1 Prevotella denticola F0289 [AEA19903-AEA22264]

AEXO01000081_c0 Prevotella denticola CRIS 18C-A [EGC85974-EGC86026]

AGWK01000009_c0 Prevotella micans F0438 [EHO74169-EHO74190]



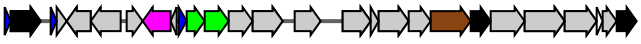
AEEI01000015_c0 Prevotella marshii DSM 16973 [EFM02695-EFM02714]

AMEP01000089_c0 Prevotella saccharolytica F0055 [EKY00113-EKY00130]

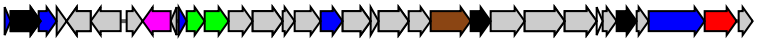
CBFR010000146_c0 Bacteroides sp. CAG:770 [CDC65793-CDC65809]

CBCT010000053_c0 Bacteroides sp. CAG:709 [CDA94209-CDA94262]

DS995286_c0 Campylobacterales bacterium GD 1 scf_1106149034639 genomic scaffold [EDZ62797-EDZ63838]



AFRZ01000001_c1 Sulfurimonas gotlandica GD1 [EHP29891-EHP29923]



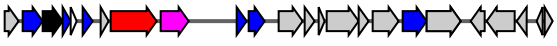
CP001999_c0 Arcobacter nitrofigilis DSM 7299 [ADG93952-ADG93982]

CP003333_c1 Sulfurospirillum barnesii SES-3 [AFL69512-AFL69538]

CP001816_c0 Sulfurospirillum deleyianum DSM 6946 [ACZ13108-ACZ13136]

CP001965_c0 Sideroxydans lithotrophicus ES-1 [ADE10591-ADE10618]

GL349412_c0 Neisseria sp. oral taxon 014 str. F0314 genomic scaffold supercont1.2 [EFI23540-EFI23563]



CP001607_c0 Aggregatibacter aphrophilus NJ8700 [ACS98029-ACS98075]

ACZJ01000014_c0 Aggregatibacter aphrophilus F0387 [EHB89425-EHB89468]

AEWB01000005_c0 Aggregatibacter aphrophilus ATCC 33389 [EGY32201-EGY32243]

AE005174_c1 Escherichia coli O157:H7 EDL933 [AAG58584-AAG58602]

CP001925_c1 Escherichia coli Xuzhou21 [AFJ31117-AFJ31135]

CM001142_c1 Escherichia fergusonii ECD227 chromosome [EGC96883-EGC96901]

GG749177_c0 Escherichia coli B185 genomic scaffold supercont1.16 [EFF04440-EFF04458]

CP001846_c1 Escherichia coli O55:H7 str. CB9615 [ADD58649-ADD58667]

GL884131_c0 Escherichia coli M718 genomic scaffold supercont1.31 [EGI19151-EGI19169]

AMTU01000377_c0 Escherichia coli FDA504 [EKH22461-EKH22480]

ANLT01000316_c0 Escherichia coli 96.0428 [EKW38841-EKW39034]

AKLK01000448_c0 Escherichia coli PA25 [EIN91717-EIN92008]

AOEM01000133_c0 Escherichia coli PA47 [ELV94531-ELV94551]

AIFI01000073_c0 Escherichia coli DEC3E [EHU72951-EHU72971]

AODY01000173_c0 Escherichia coli 99.0814 [ELV15843-ELV15863]

AMVF01000429_c0 Escherichia coli EC1869 [EKJ32978-EKJ32998]

AOEA01000182_c0 Escherichia coli 99.0816 [ELV33078-ELV33098]

AODZ01000226_c0 Escherichia coli 99.0815 [ELV24775-ELV24795]

AKMI01000666_c0 Escherichia coli EC4402 [EIP29379-EIP29418]

AOES01000138_c0 Escherichia coli PA35 [ELW25235-ELW25255]

AMUZ01000422_c0 Escherichia coli EC1856 [EKI94853-EKI94873]

ANME01000135_c0 Escherichia coli 99.0678 [EKW87206-EKW87226]

AIFL01000069_c0 Escherichia coli DEC4B [EHU91519-EHU91539]

AMUY01000419_c0 Escherichia coli EC1850 [EKI92137-EKI92334]

AMUU01000495_c0 Escherichia coli EC1846 [EKI66709-EKI66729]

AMTK01000459_c0 Escherichia coli 10.0833 [EKK52137-EKK52157]

AKMC01000510_c0 Escherichia coli EC4196 [EIO93935-EIO94029]

AMTM01000402_c0 Escherichia coli 88.0221 [EKK68650-EKK68670]

AKMH01000441_c0 Escherichia coli EC4013 [EIP25478-EIP25498]

AKMJ01000464_c0 Escherichia coli EC4439 [EIP36785-EIP36873]

AKMQ01000383_c0 Escherichia coli EC1863 [EIP75086-EIP75106]

AKLC01000435_c0 Escherichia coli PA3 [EIN51267-EIN51287]

AKMP01000475_c0 Escherichia coli EC1845 [EIP75752-EIP75772]

AMUX01000447_c0 Escherichia coli EC1849 [EKI84354-EKI84374]

AMUV01000465_c0 Escherichia coli EC1847 [EKI74599-EKI74619]

AMUW01000417_c0 Escherichia coli EC1848 [EKI78372-EKI78392]

AKLJ01000440_c0 Escherichia coli PA24 [EIN93240-EIN93326]

AMVA01000562_c0 Escherichia coli EC1862 [EKJ02535-EKJ02555]

AOEC01000197_c0 Escherichia coli 99.0848 [ELV37607-ELV37627]

AKLL01000431_c0 Escherichia coli PA28 [EIN97478-EIN97744]

AMVE01000451_c0 Escherichia coli EC1868 [EKJ22343-EKJ22539]

AMVD01000425_c0 Escherichia coli EC1866 [EKJ23251-EKJ23271]

AMVG01000513_c0 Escherichia coli EC1870 [EKJ38170-EKJ38190]

AMVB01000436_c0 Escherichia coli EC1864 [EKJ08044-EKJ08064]

AKMM01000491_c0 Escherichia coli EC4448 [EIP52735-EIP52755]

AOEB01000156_c0 Escherichia coli 99.0839 [ELV32645-ELV32754]

AOEJ01000118_c0 Escherichia coli PA19 [ELV79498-ELV79518]

AOEK01000124_c0 Escherichia coli PA13 [ELV78930-ELV78950]

ACTQ01000045_c0 Escherichia coli 4_1_47FAA [EHP64722-EHP64741]

AIFK01000044_c0 Escherichia coli DEC4A [EHU88061-EHU88080]

AKLG01000570_c0 Escherichia coli PA14 [EIN72050-EIN72068]

ANTH01000026_c0 Escherichia coli KTE193 [ELC94943-ELC94961]

ANXP01000111_c0 Escherichia coli KTE112 [ELI21775-ELI21793]

AIFN01000071_c0 Escherichia coli DEC4D [EHV02174-EHV02457]

ANLU01000500_c0 Escherichia coli 96.0427 [EKW41247-EKW41265]

AIFS01000041_c0 Escherichia coli DEC5C [EHV34059-EHV34294]

AIFT01000033_c0 Escherichia coli DEC5D [EHV35646-EHV35664]

AMUJ01000363_c0 Escherichia coli 5412 [EKI07764-EKI07782]

AIFJ01000081_c0 Escherichia coli DEC3F [EHU82351-EHU82579]

AKLT01000037_c0 Escherichia coli TW06591 [EIO47018-EIO47578]

AKME01000445_c0 Escherichia coli TW14301 [EIP07485-EIP07573]

ANMC01000224_c0 Escherichia coli 97.0007 [EKW74207-EKW74225]

ANLY01000341_c0 Escherichia coli 96.0107 [EKW59460-EKW59645]

ANLQ01000475_c0 Escherichia coli 95.0183 [EKW24042-EKW24060]

ANLM01000390_c0 Escherichia coli 90.2281 [EKV90606-EKV90815]

ANLO01000361_c0 Escherichia coli 93.0056 [EKW06760-EKW06778]

AKMG01000369_c0 Escherichia coli EC4422 [EIP21324-EIP21342]

AMVI01000484_c0 Escherichia coli FRIK523 [EKJ49501-EKJ49693]

AKKW01000483_c0 Escherichia coli FDA505 [EIN19421-EIN19439]

AMTX01000444_c0 Escherichia coli NE1487 [EKH38535-EKH38553]

ANLK01000545_c0 Escherichia coli 90.0091 [EKV87651-EKV87669]

ANLZ01000115_c0 Escherichia coli 97.0003 [EKW61593-EKW61611]

AOEU01000165_c0 Escherichia coli 95.0083 [ELW33112-ELW33130]

ANLX01000582_c0 Escherichia coli 96.0109 [EKY36142-EKY36160]

ANLS01000418_c0 Escherichia coli 95.0943 [EKW25115-EKW25133]

AMTZ01000436_c0 Escherichia coli FRIK2001 [EKH50469-EKH50487]

AMUD01000427_c0 Escherichia coli PA45 [EKH73876-EKH73894]

AMUH01000451_c0 Escherichia coli CB7326 [EKH98300-EKH98318]

ANLN01000370_c0 Escherichia coli 93.0055 [EKW06901-EKW07087]

AMTT01000428_c0 Escherichia coli FDA507 [EKH14926-EKH14944]

AIFP01000061_c0 Escherichia coli DEC4F [EHV18047-EHV18065]

AKLH01000486_c0 Escherichia coli PA15 [EIN72811-EIN73006]

AKLW01000064_c0 Escherichia coli TW11039 [EIO61268-EIO61286]

AMTV01000443_c0 Escherichia coli FRIK1999 [EKH28226-EKH28244]

AMTE01000409_c0 Escherichia coli 5.2239 [EKK23525-EKK23721]

AMTO01000091_c0 Escherichia coli 10.0821 [EKK83665-EKK83683]

AIFQ01000027_c0 Escherichia coli DEC5A [EHV21587-EHV21605]

AMUA01000413_c0 Escherichia coli PA4 [EKH56054-EKH56236]

AKLN01000537_c0 Escherichia coli PA32 [EIO10370-EIO10451]

ANMF01000102_c0 Escherichia coli 99.0713 [EKW88632-EKW88650]

ANMD01000470_c0 Escherichia coli 99.0672 [EKW78716-EKW78734]

AKLD01000499_c0 Escherichia coli PA5 [EIN54312-EIN54330]

AKLQ01000056_c0 Escherichia coli PA40 [EIO25888-EIO25906]

AKLO01000534_c0 Escherichia coli PA33 [EIO13467-EIO13485]

ANLL01000352_c0 Escherichia coli 90.0039 [EKV93920-EKV94044]

AMTD01000405_c0 Escherichia coli 3.4870 [EKK24007-EKK24025]

AMTF01000400_c0 Escherichia coli 6.0172 [EKK24857-EKK24875]

AOEF01000159_c0 Escherichia coli 99.1793 [ELV53166-ELV53184]

AMTI01000517_c0 Escherichia coli 8.0586 [EKK41099-EKK41117]

AIFR01000047_c0 Escherichia coli DEC5B [EHV26243-EHV26261]

AIFG01000061_c0 Escherichia coli DEC3C [EHU67415-EHU67433]

AFDZ01000022_c0 Escherichia coli STEC_MHI813 [EGX02124-EGX02142]

AKLY01000060_c0 Escherichia coli TW09109 [EIO71785-EIO71803]

AEUA01000075_c0 Escherichia coli O55:H7 str. 3256-97 strain 3256-97 TW 07815 [EFX23685-EFX23703]

AMUG01000382_c0 Escherichia coli 5905 [EKH89875-EKH89893]

AIFU01000021_c0 Escherichia coli DEC5E [EHV44114-EHV44132]

AEUB01000047_c0 Escherichia coli O55:H7 str. USDA 5905 [EFX28960-EFX28978]

AEUC01000076_c0 Escherichia coli O157:H7 str. LSU-61 [EFX33551-EFX33569]

AETY01000088_c0 Escherichia coli O157:H- str. 493-89 [EFX14274-EFX14292]

AETZ01000094_c0 Escherichia coli O157:H- str. H 2687 [EFX19035-EFX19053]

AOEG01000223_c0 Escherichia coli 99.1805 [ELV66257-ELV66275]

AOEV01000120_c0 Escherichia coli 99.0670 [ELW40001-ELW40019]

AOEH01000109_c0 Escherichia coli ATCC 700728 [ELV65824-ELV65842]

AOEQ01000140_c0 Escherichia coli 99.1781 [ELW11852-ELW11870]

AMUE01000337_c0 Escherichia coli TT12B [EKH81367-EKH81385]

AMUF01000435_c0 Escherichia coli MA6 [EKH86120-EKH86138]

AOED01000208_c0 Escherichia coli 99.1753 [ELV46457-ELV46475]

ANMB01000658_c0 Escherichia coli 97.0010 [EKY37370-EKY37388]

AMUO01000508_c0 Escherichia coli PA38 [EKI38866-EKI38884]

ANLV01000420_c0 Escherichia coli 96.0939 [EKW45677-EKW45695]

ANLJ01000470_c0 Escherichia coli 89.0511 [EKV73429-EKV73447]

AMVH01000440_c0 Escherichia coli NE098 [EKJ40279-EKJ40449]

ANLI01000399_c0 Escherichia coli 88.1042 [EKV72529-EKV72547]

AETX01000100_c0 Escherichia coli O157:H7 str. G5101 [EFX09352-EFX09370]

ABHU01000002_c0 Escherichia coli O157:H7 str. EC869 [EDU92648-EDU93115]

ABHT01000001_c0 Escherichia coli O157:H7 str. EC4501 [EDU87869-EDU88328]

AMUI01000467_c0 Escherichia coli EC96038 [EKI04633-EKI04651]

AOEN01000108_c0 Escherichia coli PA48 [ELV95166-ELV95184]

ANLW01000401_c0 Escherichia coli 96.0932 [EKW53122-EKW53140]

AOEI01000107_c0 Escherichia coli PA11 [ELV65552-ELV65570]

ANLR01000529_c0 Escherichia coli 95.1288 [EKW27119-EKW27137]

ANLP01000370_c0 Escherichia coli 94.0618 [EKW11081-EKW11099]

AIFF01000057_c0 Escherichia coli DEC3B [EHU56536-EHU56554]

AIFM01000044_c0 Escherichia coli DEC4C [EHV03138-EHV03368]

AKLM01000539_c0 Escherichia coli PA31 [EIO09781-EIO09852]

AOEE01000210_c0 Escherichia coli 99.1775 [ELV50180-ELV50320]

AKLZ01000980_c0 Escherichia coli TW09195 [EIO90759-EIO90777]

AMVK01000522_c0 Escherichia coli 0.1304 [EKJ58295-EKJ58313]

AODX01000546_c0 Escherichia coli 09BKT078844 [ELV17349-ELV17367]

AKLI01000061_c0 Escherichia coli PA22 [EIN85039-EIN85377]

AKKY01000541_c0 Escherichia coli FRIK1996 [EIN17967-EIN18054]

AKLS01000445_c0 Escherichia coli PA42 [EIO35276-EIO35473]

AKLV01000041_c0 Escherichia coli TW10246 [EIO55180-EIO55534]

AMTP01000115_c0 Escherichia coli FRIK920 [EKG99170-EKG99188]

AKMN01000051_c0 Escherichia coli EC1738 [EIP57435-EIP57787]

AKMA01000048_c0 Escherichia coli TW10119 [EIO80528-EIO80930]

AMTY01000582_c0 Escherichia coli NE037 [EKH44561-EKH44747]

AMUB01000434_c0 Escherichia coli PA23 [EKH65168-EKH65186]

AMTS01000362_c0 Escherichia coli FDA506 [EKH11144-EKH11162]

AMTQ01000060_c0 Escherichia coli PA7 [EKG96817-EKG96835]

AMUC01000499_c0 Escherichia coli PA49 [EKH67720-EKH67915]

AKKX01000475_c0 Escherichia coli FDA517 [EIN19317-EIN19335]

AIFH01000069_c0 Escherichia coli DEC3D [EHU71470-EHU71488]

AIFE01000054_c0 Escherichia coli DEC3A [EHU54512-EHU54734]

AKMF01000395_c0 Escherichia coli EC4421 [EIP11881-EIP12080]

AIFO01000056_c0 Escherichia coli DEC4E [EHV08817-EHV08835]

ANWB01000031_c0 Escherichia coli KTE84 [ELG33101-ELG33119]

AFAT01000067_c0 Escherichia coli PCN033 [EGP23207-EGP23225]

ANWM01000032_c0 Escherichia coli KTE146 [ELG85576-ELG85594]

AFDR01000026_c0 Escherichia coli STEC_C165-02 [EGW65097-EGW65115]

ANUT01000030_c0 Escherichia coli KTE83 [ELE77699-ELE77717]

ANUR01000039_c0 Escherichia coli KTE80 [ELE67774-ELE67792]

ANTU01000023_c0 Escherichia coli KTE228 [ELD57885-ELD57903]

ANTC01000042_c0 Escherichia coli KTE181 [ELC71139-ELC71157]

ANTM01000036_c0 Escherichia coli KTE208 [ELD17908-ELD17926]

ANTJ01000035_c0 Escherichia coli KTE204 [ELD02640-ELD02658]

AJWS01000228_c0 Escherichia coli 576-1 [EIL70105-EIL70123]

ANSN01000053_c0 Escherichia coli KTE2 [ELB96505-ELB96523]

ANSX01000024_c0 Escherichia coli KTE26 [ELC43979-ELC43997]

ASUM01000033_c0 Escherichia coli KTE200 [EOV20488-EOV20506]

ASUQ01000004_c0 Escherichia coli KTE225 [EOX15556-EOX15574]

ASUW01000061_c0 Escherichia coli KTE68 [EOV57526-EOV57544]

ASVM01000069_c0 Escherichia coli KTE127 [EOW32100-EOW32118]

ASVT01000031_c0 Escherichia coli KTE172 [EOW73122-EOW73140]

ASVS01000054_c0 Escherichia coli KTE170 [EOW64918-EOW64936]

ANVV01000002_c0 Escherichia coli KTE54 [ELG11654-ELG11672]

ANWP01000040_c0 Escherichia coli KTE158 [ELG96894-ELG96912]

ANWT01000039_c0 Escherichia coli KTE190 [ELH17706-ELH17724]

ANYT01000103_c0 Escherichia coli KTE82 [ELJ63228-ELJ63246]

ASTN01000049_c0 Escherichia coli KTE1 [EOW88845-EOW88863]

ASTT01000033_c0 Escherichia coli KTE231 [EOU48667-EOU48685]

ANXT01000092_c0 Escherichia coli KTE122 [ELI39888-ELI39906]

ANXM01000056_c0 Escherichia coli KTE105 [ELI04319-ELI04337]

ANXW01000087_c0 Escherichia coli KTE128 [ELI52476-ELI52494]

ANYP01000063_c0 Escherichia coli KTE177 [ELJ40772-ELJ40790]

AEIA01000017_c0 Escherichia fergusonii B253 [EGC05776-EGC05794]

FN554766_c1 Escherichia coli 042 complete genome. [CBG36561-CBG36579]

AKLR01000483_c0 Escherichia coli PA41 [EIO33479-EIO33497]

AKLB01000453_c0 Escherichia coli 93-001 [EIN35483-EIN35501]

AMTL01000386_c0 Escherichia coli 10.0869 [EKK63752-EKK63770]

AKLA01000514_c0 Escherichia coli FRIK1990 [EIN38647-EIN38665]

ANLH01000439_c0 Escherichia coli 88.1467 [EKV76198-EKV76216]

AKKZ01000664_c0 Escherichia coli FRIK1985 [EIN35242-EIN35326]

AKLF01000548_c0 Escherichia coli PA10 [EIN67931-EIN67949]

AMTW01000431_c0 Escherichia coli FRIK1997 [EKH34078-EKH34096]

AMTN01000119_c0 Escherichia coli 8.0416 [EKK73509-EKK73527]

CP003109_c1 Escherichia coli O55:H7 str. RM12579 [AEZ42548-AEZ42566]

AOET01000109_c0 Escherichia coli 3.4880 [ELW30166-ELW30184]

AOEP01000128_c0 Escherichia coli 7.1982 [ELW09370-ELW09388]

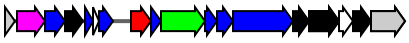
AIGM01000044_c0 Escherichia coli DEC9C [EHW44571-EHW44589]

ANYV01000050_c0 Escherichia coli KTE88 [ELJ67938-ELJ67956]

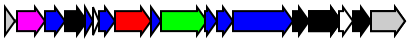
AIHM01000058_c0 Escherichia coli DEC14C [EHX86959-EHX86977]

ANYZ01000042_c0 Escherichia coli KTE97 [ELJ92437-ELJ92455]

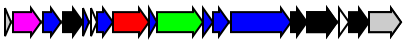
AFDV01000055_c0 Escherichia coli STEC_DG131-3 [EGW87039-EGW87056]



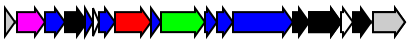
AIGT01000030_c0 Escherichia coli DEC10E [EHW88181-EHW88198]



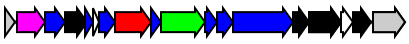
AMUN01000153_c0 Escherichia coli 3006 [EKI34462-EKI34479]



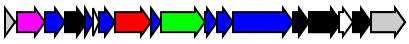
APZX01000401_c0 Escherichia coli P0304799.3 [ENE04697-ENE04721]



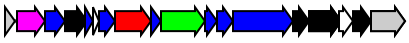
AMVJ01000117_c0 Escherichia coli 0.1288 [EKJ55730-EKJ55747]



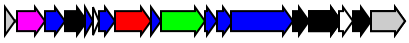
AIGI01000046_c0 Escherichia coli DEC8D [EHW20022-EHW20039]



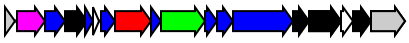
AFEX01000038_c0 Escherichia coli STEC_O31 [EJK94118-EJK94135]



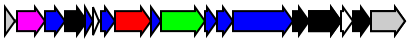
ANUO01000043_c0 Escherichia coli KTE75 [ELE51505-ELE51522]



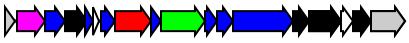
ADTK01000149_c0 Escherichia coli MS 84-1 [EFJ87565-EFJ87582]



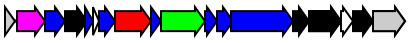
ADWQ01000033_c0 Escherichia coli MS 85-1 [EFU33359-EFU33376]



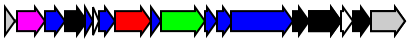
ACGN01000007_c0 Escherichia coli 83972 [EEJ49197-EEJ49214]



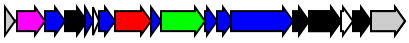
ANYX01000054_c0 Escherichia coli KTE94 [ELJ82656-ELJ82673]



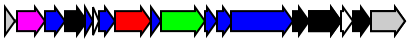
ANXN01000062_c0 Escherichia coli KTE106 [ELI09147-ELI09164]



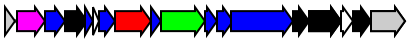
ANYL01000050_c0 Escherichia coli KTE167 [ELJ24293-ELJ24310]



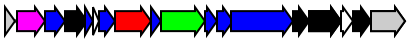
ANYI01000052_c0 Escherichia coli KTE160 [ELJ09673-ELJ09690]



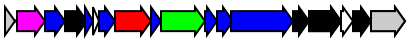
ANZA01000041_c0 Escherichia coli KTE99 [ELJ96339-ELJ96356]



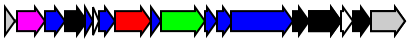
ANXL01000056_c0 Escherichia coli KTE104 [ELI04713-ELI04730]



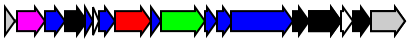
ANUM01000027_c0 Escherichia coli KTE67 [ELE38798-ELE38815]



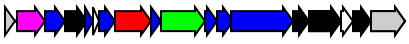
ALIN02000002_c0 Escherichia coli J96 [ELL43789-ELL43806]



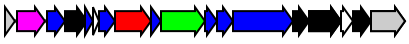
ANXU01000066_c0 Escherichia coli KTE124 [ELI40196-ELI40213]



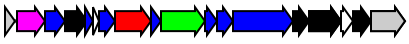
AEZJ02000017_c0 Escherichia coli 97.0246 [EIG92959-EIG93698]



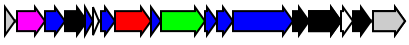
FN649414_c1 Escherichia coli ETEC H10407 [CBJ02668-CBJ02685]



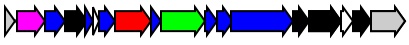
AQGH01000047_c0 Escherichia coli 2735000 [EMZ66112-EMZ66311]



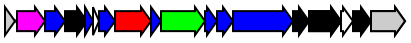
AQDF01000056_c0 Escherichia coli 2867750 [EMV56501-EMV56518]



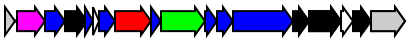
ANXF01000032_c0 Escherichia coli KTE215 [ELH75583-ELH75600]



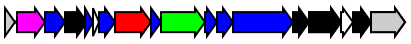
AQDY01000152_c0 Escherichia coli 2756500 [EMW59633-EMW59650]



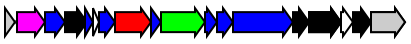
APYH01000074_c0 Escherichia coli P0299438.8 [ENC25601-ENC25618]



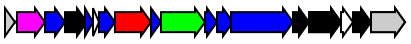
AIGK01000055_c0 Escherichia coli DEC9A [EHW33120-EHW33137]



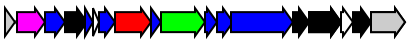
AIGO01000050_c0 Escherichia coli DEC9E [EHW54456-EHW54473]



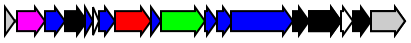
AJVR01000262_c0 Escherichia coli O103:H2 str. CVM9450 [EIL03366-EIL03383]



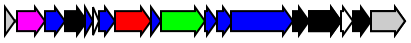
AEZT02000046_c0 Escherichia coli 93.0624 [EIH64283-EIH64530]



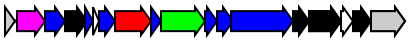
ANYM01000056_c0 Escherichia coli KTE168 [ELJ25051-ELJ25068]



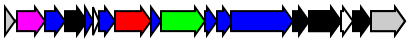
ANSU01000028_c0 Escherichia coli KTE16 [ELC25831-ELC25848]



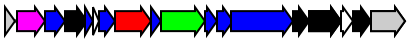
ANVI01000028_c0 Escherichia coli KTE8 [ELF47070-ELF47087]



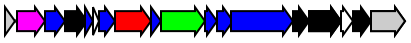
ANTT01000034_c0 Escherichia coli KTE224 [ELD50565-ELD50582]



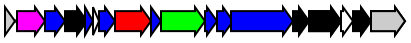
ANVF01000035_c0 Escherichia coli KTE169 [ELF35341-ELF35358]



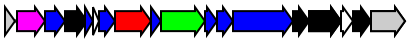
ANXC01000028_c0 Escherichia coli KTE207 [ELH59328-ELH59345]



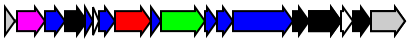
ANXZ01000053_c0 Escherichia coli KTE133 [ELI66031-ELI66048]



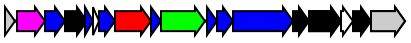
AQDO01000098_c0 Escherichia coli 2850750 [EMW06729-EMW06746]



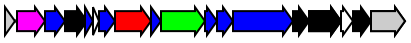
AQDB01000098_c0 Escherichia coli 2875000 [EMV37347-EMV37364]



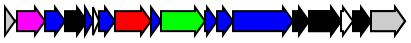
AQDM01000102_c0 Escherichia coli 2853500 [EMW00618-EMW00635]



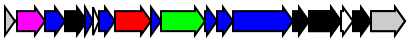
AQDP01000045_c0 Escherichia coli 2850400 [EMW17011-EMW17028]



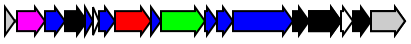
AQDH01000105_c0 Escherichia coli 2866550 [EMV69296-EMV69313]



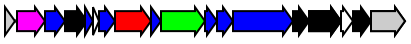
AQDG01000049_c0 Escherichia coli 2866750 [EMV74115-EMV74132]



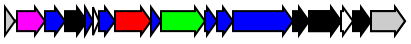
AQDI01000097_c0 Escherichia coli 2866450 [EMV70642-EMV70659]



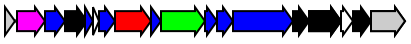
APXB01000285_c0 Escherichia coli 179550 [ENA61854-ENA61871]



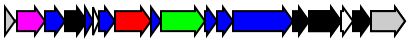
APXC01000247_c0 Escherichia coli 180200 [ENA65958-ENA65975]



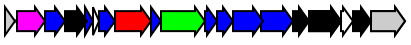
AQDC01000088_c0 Escherichia coli 2872800 [EMV45105-EMV45122]



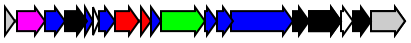
AQDN01000085_c0 Escherichia coli 2851500 [EMW00866-EMW00883]



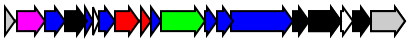
AIHL01000015_c0 Escherichia coli DEC14B [EHX89382-EHX89400]



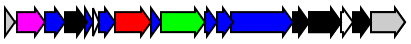
ANTD01000011_c0 Escherichia coli KTE187 [ELC74288-ELC74306]



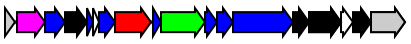
ANWK01000015_c0 Escherichia coli KTE141 [ELG79830-ELG79848]



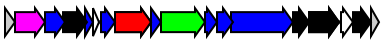
GG657385_c0 Shigella sp. D9 genomic scaffold supercont1.2 [EGJ05961-EGJ05978]



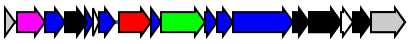
CP001671_c0 Escherichia coli ABU 83972 [ADN45644-ADN45661]



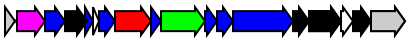
ADTJ01000303_c0 Escherichia coli MS 198-1 [EFJ73400-EFJ73417]



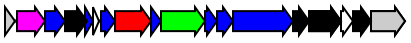
APYE01000082_c0 Escherichia coli P0299438.5 [ENC05540-ENC05557]



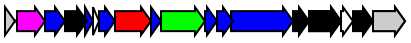
CAPM01000033_c0 Escherichia coli Nissle 1917 [CCQ04547-CCQ04564]



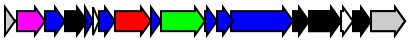
ADTO01000021_c0 Escherichia coli MS 45-1 [EFJ94481-EFJ94498]



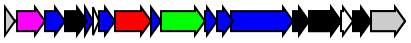
ASUH01000016_c0 Escherichia coli KTE185 [EOX23403-EOX23420]



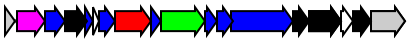
ANTG01000014_c0 Escherichia coli KTE191 [ELC91813-ELC91830]



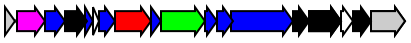
ANYN01000029_c0 Escherichia coli KTE174 [ELJ41811-ELJ41828]



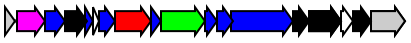
ASUR01000010_c0 Escherichia coli KTE226 [EOX11314-EOX11331]



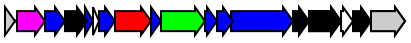
ANXI01000021_c0 Escherichia coli KTE223 [ELH91027-ELH91044]



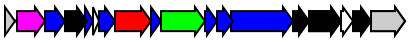
ANTS01000004_c0 Escherichia coli KTE220 [ELD52209-ELD52226]



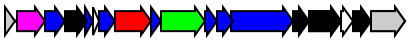
ANTI01000020_c0 Escherichia coli KTE201 [ELD00918-ELD00935]



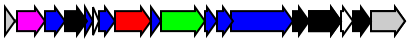
ANTF01000019_c0 Escherichia coli KTE189 [ELC85691-ELC85708]



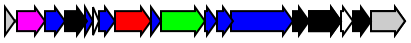
ASUJ01000011_c0 Escherichia coli KTE195 [EOV10005-EOV10022]



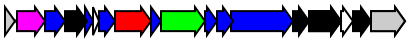
ASHD01000021_c0 Escherichia coli ATCC 25922 [EOR53367-EOR53384]



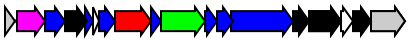
AJWO01000019_c0 Escherichia coli KD1 [EIL54074-EIL54091]



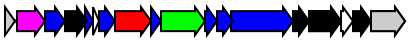
ANTE01000018_c0 Escherichia coli KTE188 [ELC83008-ELC83025]



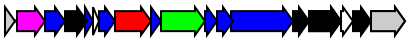
ANTV01000013_c0 Escherichia coli KTE230 [ELD63599-ELD63616]



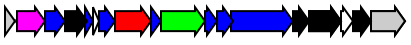
ANUE01000012_c0 Escherichia coli KTE53 [ELE07492-ELE07509]



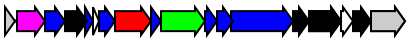
ANUJ01000011_c0 Escherichia coli KTE60 [ELE34007-ELE34024]



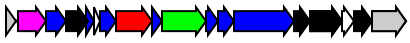
ANST01000015_c0 Escherichia coli KTE15 [ELC31676-ELC31693]



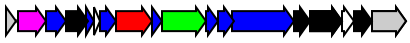
ANUN01000007_c0 Escherichia coli KTE72 [ELE51778-ELE51795]



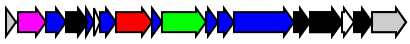
AEFA01000002_c0 Escherichia coli NC101 [EFM54990-EFM55007]



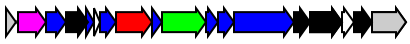
AEZS02000010_c1 Escherichia coli 3.2608 [EIH53734-EIH57150]



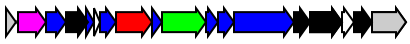
APYB01000082_c0 Escherichia coli P0299438.11 [ENB91608-ENB91625]



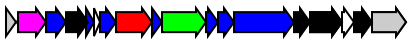
APYF01000084_c0 Escherichia coli P0299438.6 [ENC09154-ENC09185]



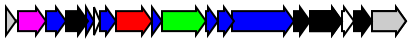
APYC01000074_c0 Escherichia coli P0299438.3 [ENB93740-ENB93777]



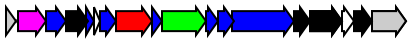
APYG01000117_c0 Escherichia coli P0299438.7 [ENC10092-ENC10109]



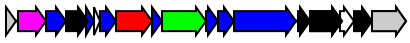
ASTV01000003_c0 Escherichia coli KTE19 [EOU64781-EOU64798]



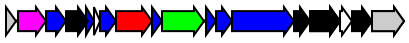
AJQW01000018_c0 Escherichia coli O32:H37 str. P4 [EIF16889-EIF16906]



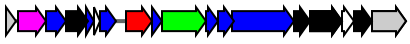
AIGL01000022_c0 Escherichia coli DEC9B [EHW46009-EHW46026]



ANSZ01000012_c0 Escherichia coli KTE39 [ELC57468-ELC57485]



ANTQ01000018_c0 Escherichia coli KTE214 [ELD40492-ELD40509]



CP003683_c2 Klebsiella oxytoca E718 [AFN29926-AFN34581]

BAFF01000007_c0 Escherichia hermannii NBRC 105704 [GAB52565-GAB52583]

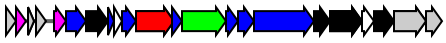
FN543093_c2 Cronobacter turicensis z3032 complete genome. [CBA34311-CBA34329]

CAKZ01000175_c0 Cronobacter dublinensis 1210 [CCJ83060-CCJ83078]

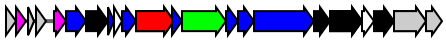
BA000007_c1 Escherichia coli O157:H7 str. Sakai DNA [BAB37747-BAB37765]

GG749227_c0 Escherichia coli FVEC1412 genomic scaffold supercont1.22 [EFE98902-EFE98919]

ANTY01000039_c0 Escherichia coli KTE235 [ELD75831-ELD75852]



ANUY01000044_c0 Escherichia coli KTE116 [ELE96671-ELE96692]



GG774916_c0 Escherichia coli FVEC1302 genomic scaffold supercont1.18 [EFI18157-EFI18175]

GG657369_c0 Citrobacter sp. 30_2 genomic scaffold supercont1.4 [EEH95332-EEH95350]

CP003218_c0 Klebsiella oxytoca KCTC 1686 [AEX02728-AEX02746]

CP000822_c0 Citrobacter koseri ATCC BAA-895 [ABV15951-ABV15969]

ASQK01000014_c0 Citrobacter sp. KTE151 [EOQ45039-EOQ45057]

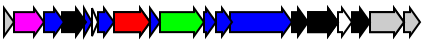
ASQL01000014_c0 Citrobacter sp. KTE32 [EOQ26323-EOQ26341]

CP001875_c0 Pantoea ananatis LMG 20103 [ADD75714-ADD75732]

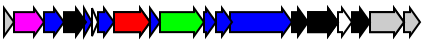
ALNS01000038_c0 Enterobacter sp. SST3 [EJO45130-EJO45148]

CP002886_c1 Enterobacter cloacae EcWSU1 [AEW75686-AEW75705]

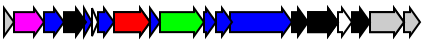
AEXG01000005_c0 Escherichia coli E101 [EHN93015-EHN93033]



ABKY02000001_c0 Escherichia coli O157:H7 str. TW14588 [EEC28832-EEC31285]

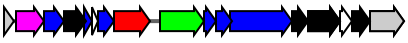


AEHW01000034_c0 Escherichia coli TW10509 [EGB69976-EGB69994]

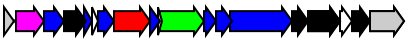


ASVP01000011_c0 Escherichia coli KTE134 [EOW57694-EOW57712]

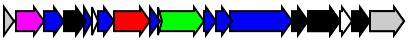
AJ586887_c0 Escherichia coli Genomic Island I, strain Nissle 1917. [CAE55664-CAE55680]



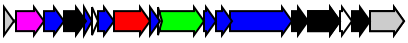
CP002211_c0 Escherichia coli str. 'clone D i2' [AER83663-AER83681]



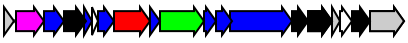
CP002212_c0 Escherichia coli str. 'clone D i14' [AER88582-AER88600]



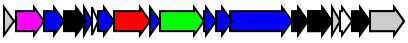
AE014075_c0 Escherichia coli CFT073 [AAN79648-AAN79666]



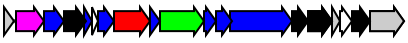
ANUU01000009_c0 Escherichia coli KTE86 [ELE82503-ELE82521]



ASVC01000019_c0 Escherichia coli KTE89 [EOV96058-EOV96076]



ANWX01000033_c0 Escherichia coli KTE183 [ELH32760-ELH32778]



AOER01000122_c0 Escherichia coli 99.1762 [ELW16247-ELW16268]

AKLE01000396_c0 Escherichia coli PA9 [EIN57705-EIN57726]

AMUT01000424_c0 Escherichia coli EC1737 [EKI62373-EKI62394]

AKML01000398_c0 Escherichia coli EC4437 [EIP50725-EIP50939]

AMTJ01000362_c0 Escherichia coli 8.2524 [EKK54776-EKK54797]

AKLX01000430_c0 Escherichia coli TW09098 [EIO68607-EIO68691]

AKMO01000044_c0 Escherichia coli EC1734 [EIP65253-EIP65521]

AOEO01000109_c0 Escherichia coli PA8 [ELW01424-ELW01445]

AMUS01000378_c0 Escherichia coli EC1736 [EKI58734-EKI58825]

AMUR01000380_c0 Escherichia coli EC1735 [EKI48152-EKI48247]

AKMD01000603_c0 Escherichia coli TW14313 [EIP06243-EIP06281]

AKLP01000753_c0 Escherichia coli PA39 [EIO31432-EIO31502]

AKMK01000399_c0 Escherichia coli EC4436 [EIP41931-EIP42027]

AKLU01000359_c0 Escherichia coli TW07945 [EIO54411-EIO54553]

AOEL01000094_c0 Escherichia coli PA2 [ELV87514-ELV87535]

AMTR01000359_c0 Escherichia coli PA34 [EKH01866-EKH01887]

CP001368_c1 Escherichia coli O157:H7 str. TW14359 [ACT74156-ACT74173]

ABHK02000001_c1 Escherichia coli O157:H7 str. EC4206 [EDZ75096-EDZ79074]

ABHM02000001_c1 Escherichia coli O157:H7 str. EC4042 [EDZ86063-EDZ89805]

ABHL02000001_c1 Escherichia coli O157:H7 str. EC4045 [EDZ81874-EDZ84564]

ABHQ01000003_c0 Escherichia coli O157:H7 str. EC4076 [EDU71196-EDU71662]

ABHW01000001_c0 Escherichia coli O157:H7 str. EC508 [EDU98320-EDU98579]

ABHR01000001_c0 Escherichia coli O157:H7 str. EC4401 [EDU77212-EDU77798]

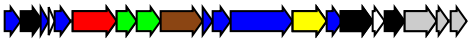
ABHS01000001_c0 Escherichia coli O157:H7 str. EC4486 [EDU82818-EDU83221]

APPJ01000010_c0 Acinetobacter guillouiae NIPH 991 [ENV17389-ENV17407]

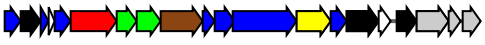
APPZ01000009_c0 Acinetobacter johnsonii ANC 3681 [ENV71593-ENV71612]

APON01000043_c0 Acinetobacter johnsonii CIP 64.6 [ENU38387-ENU38406]

APQD01000021_c0 Acinetobacter bouvetii DSM 14964 = CIP 107468 [ENV81659-ENV81678]



APRM01000007_c0 Acinetobacter sp. NIPH 298 [ENW95283-ENW95302]



ATGG01000007_c0 Acinetobacter gyllenbergii CIP 110306 [EPF92360-EPF92379]

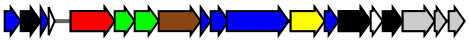
AQFL01000005_c0 Acinetobacter sp. CIP 110321 [EOR09621-EOR09640]

APPH01000004_c0 Acinetobacter sp. CIP 56.2 [ENV11019-ENV11038]

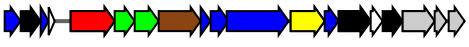
APSD01000015_c0 Acinetobacter sp. ANC 3880 [ENX62915-ENX62934]

AEOX01000007_c0 Acinetobacter baumannii AB210 [EGK48008-EGK48026]

APAW01000039_c0 Acinetobacter baumannii ABNIH5 [EMT92604-EMT92622]

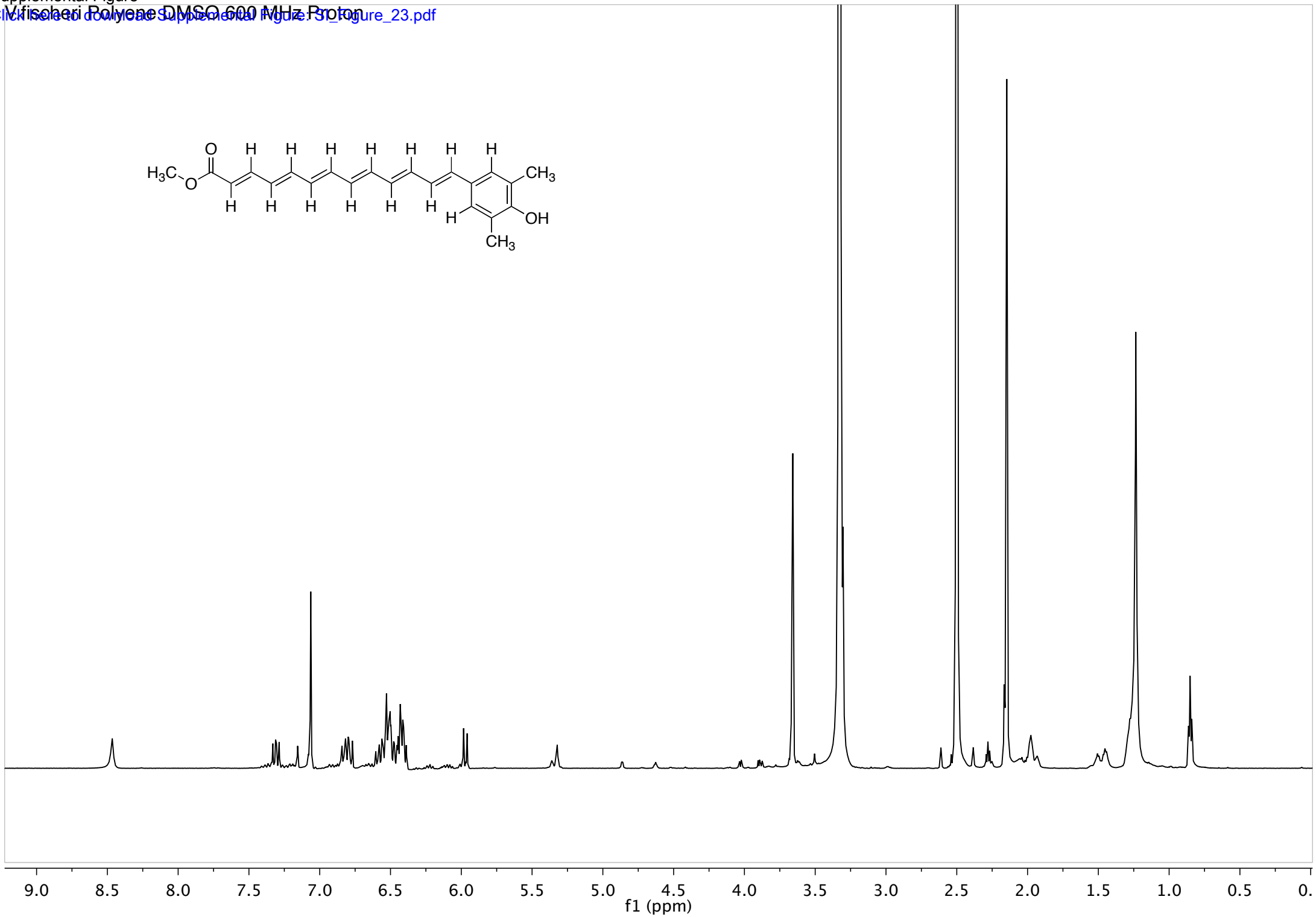


AFTC01000040_c0 Acinetobacter baumannii ABNIH4 [EGU02829-EGU02847]

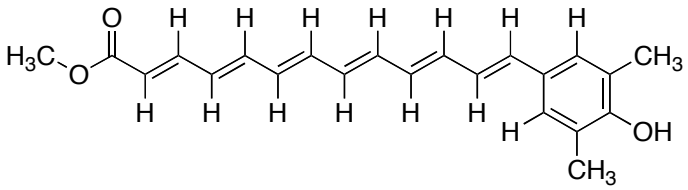


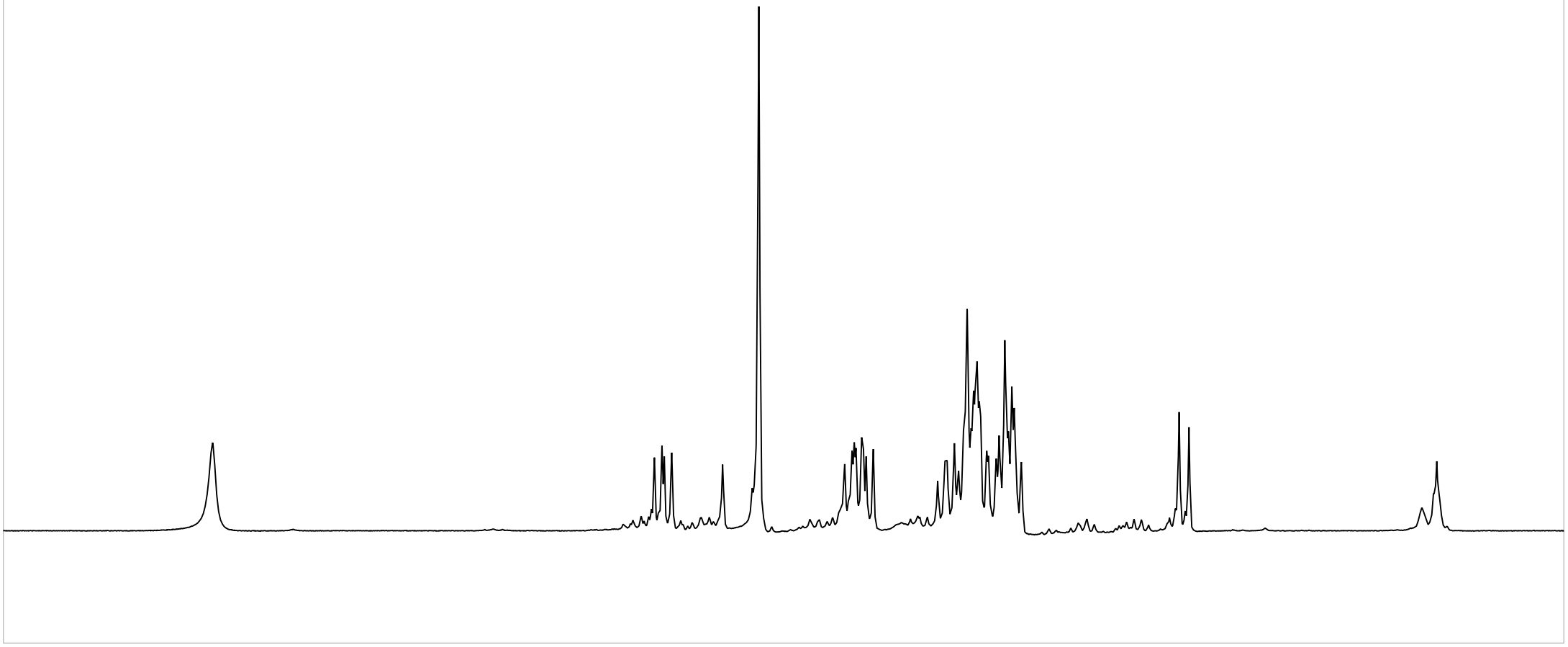
Supplemental Figure

Click here to download Supplemental Figure 1 S1_Figure_23.pdf



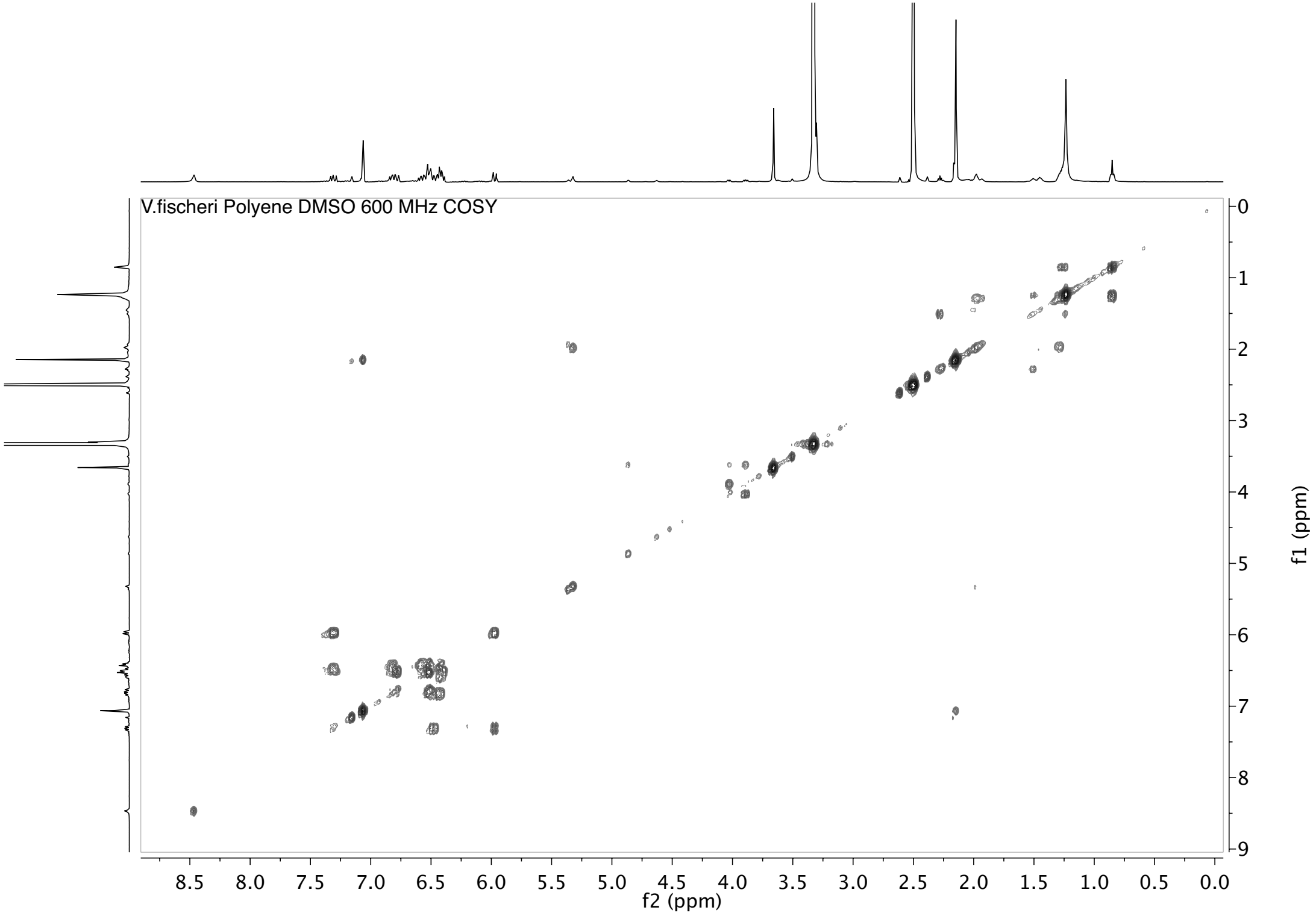
V.fischeri Polyene DMSO 600 MHz Proton

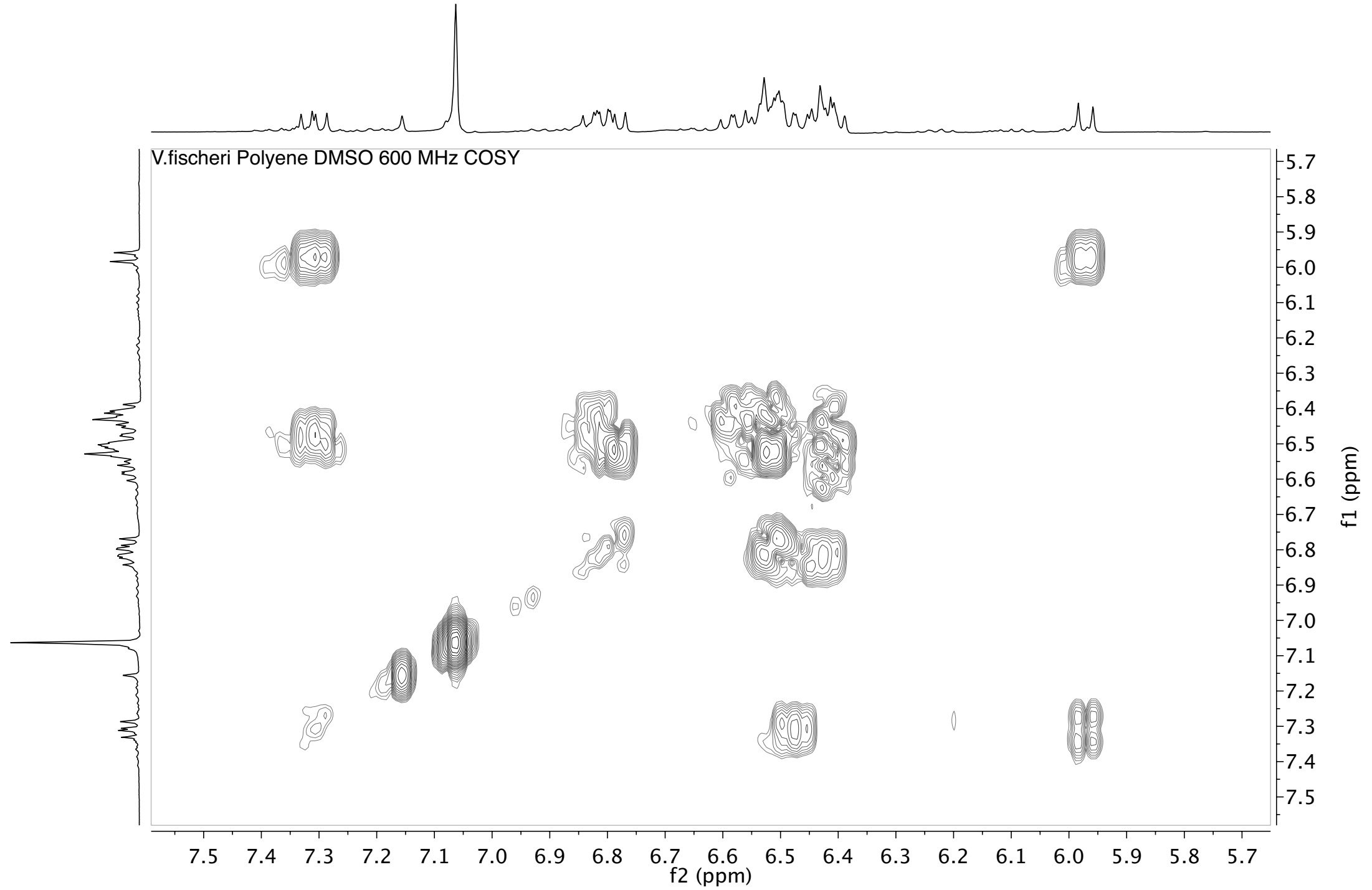


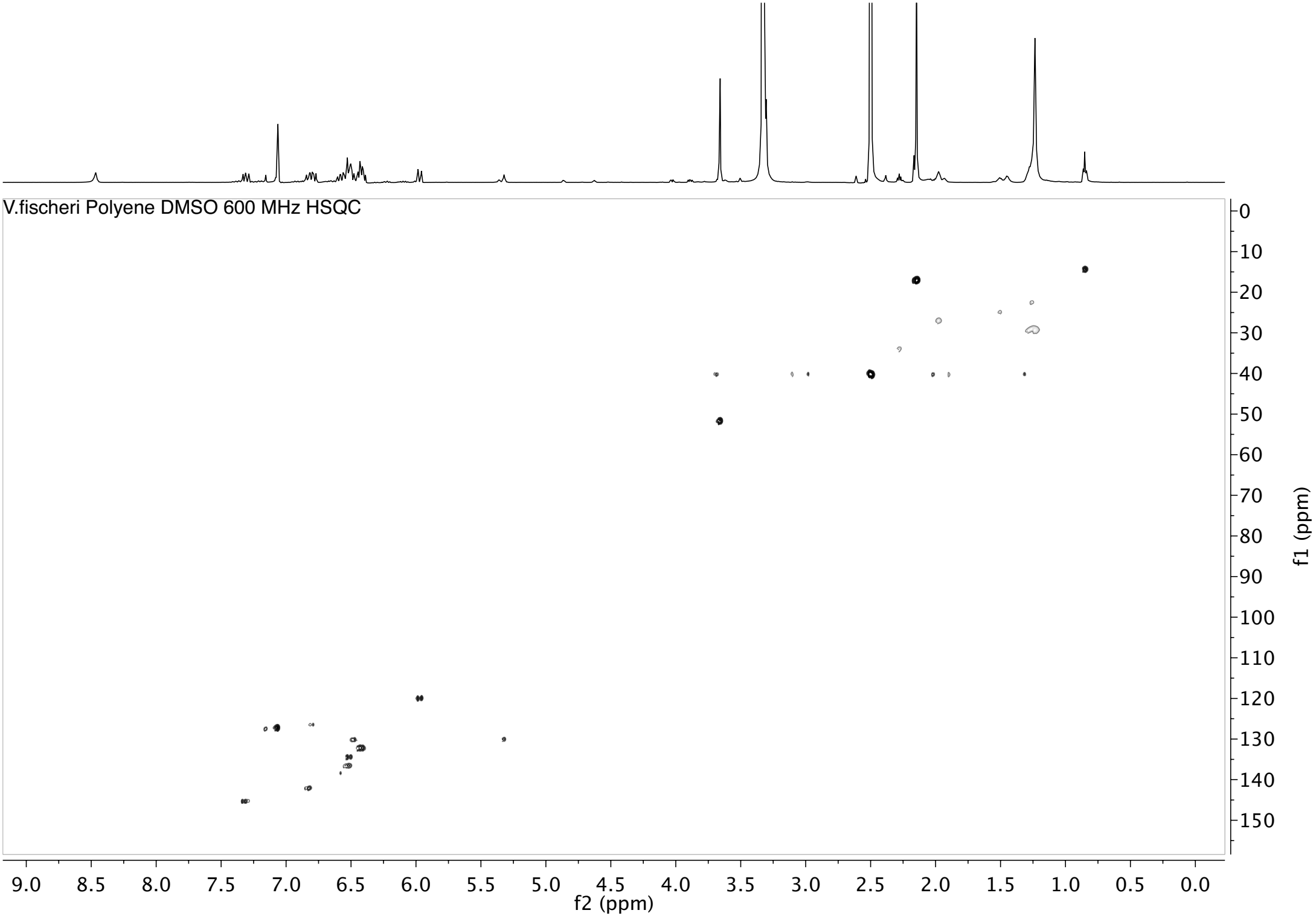


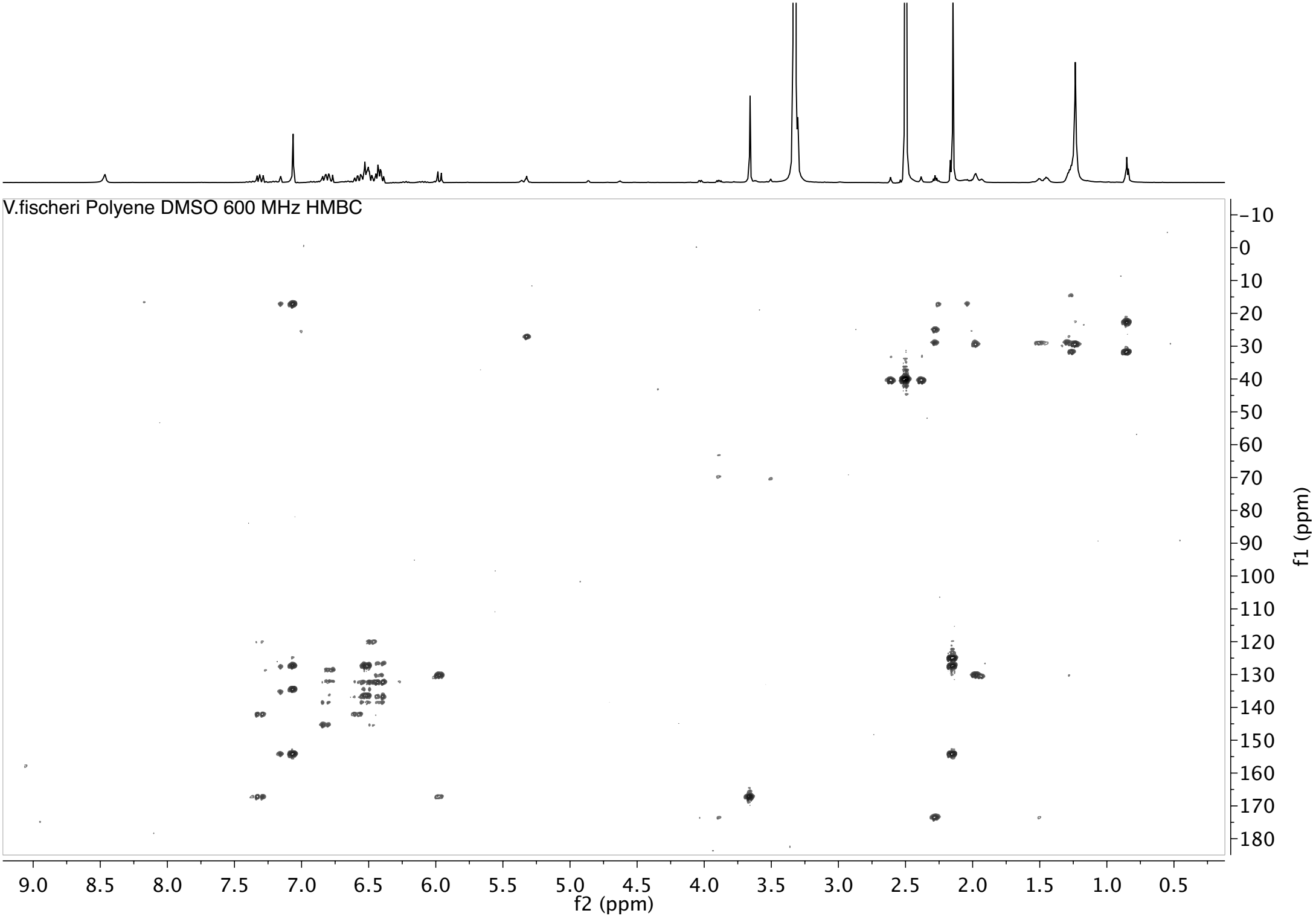
.0 8.9 8.8 8.7 8.6 8.5 8.4 8.3 8.2 8.1 8.0 7.9 7.8 7.7 7.6 7.5 7.4 7.3 7.2 7.1 7.0 6.9 6.8 6.7 6.6 6.5 6.4 6.3 6.2 6.1 6.0 5.9 5.8 5.7 5.6 5.5 5.4 5.3 5.2 5.1 5.

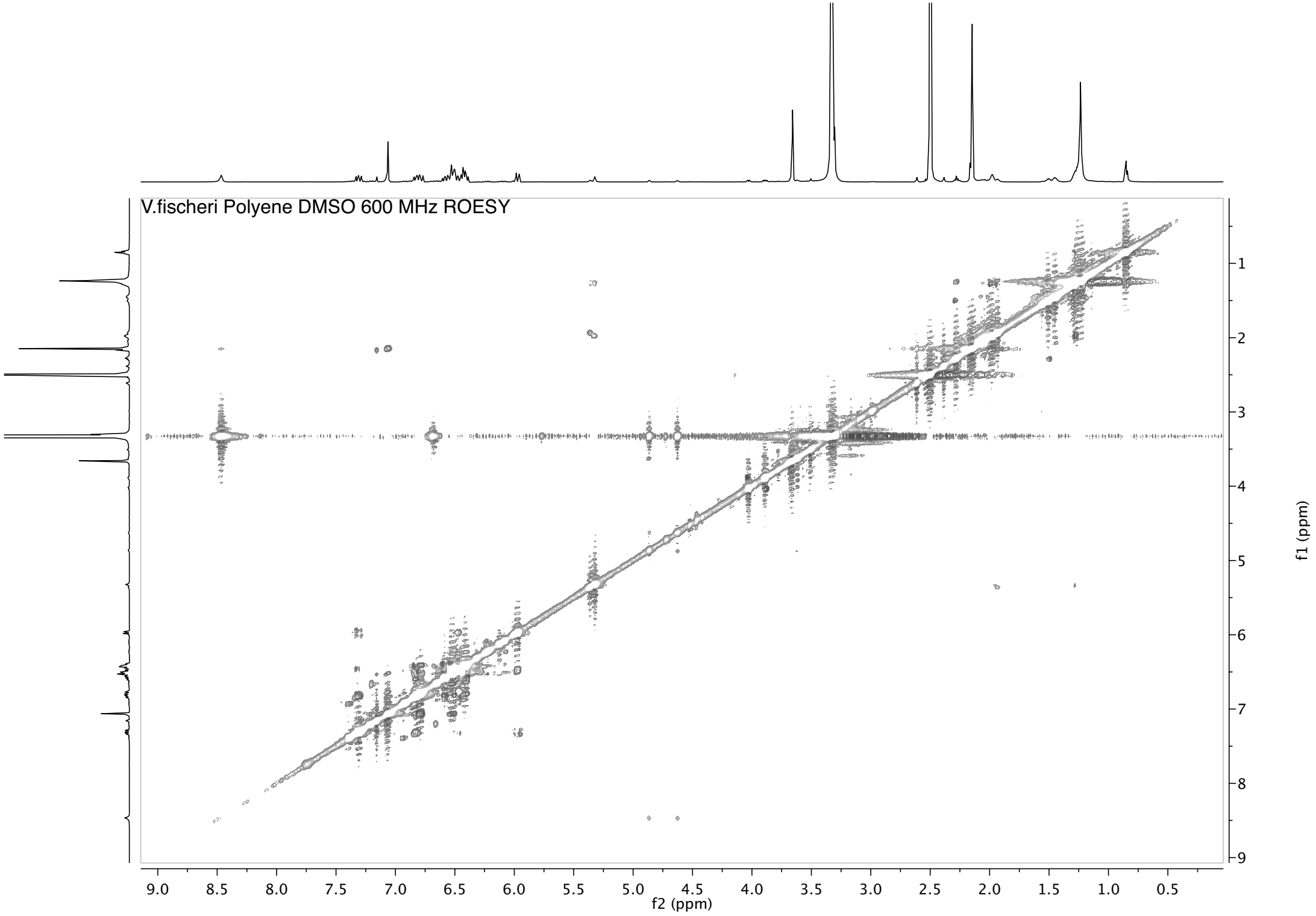
f1 (ppm)


V.fischeri Polyene DMSO 600 MHz COSY
f2 (ppm)
f1 (ppm)
8.5
8.0
7.5
7.0
6.5
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0
1
2
3
4
5
6
7
8
9

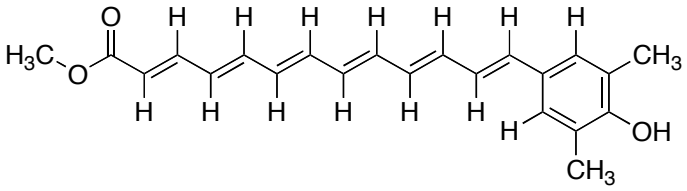

V.fischeri Polyene DMSO 600 MHz COSY
f2 (ppm)
f1 (ppm)
7.5 7.4 7.3 7.2 7.1 7.0 6.9 6.8 6.7 6.6 6.5 6.4 6.3 6.2 6.1 6.0 5.9 5.8 5.7
5.7 5.8 5.9 6.0 6.1 6.2 6.3 6.4 6.5 6.6 6.7 6.8 6.9 7.0 7.1 7.2 7.3 7.4 7.5


V.fischeri Polyene DMSO 600 MHz HSQC
f2 (ppm)
f1 (ppm)

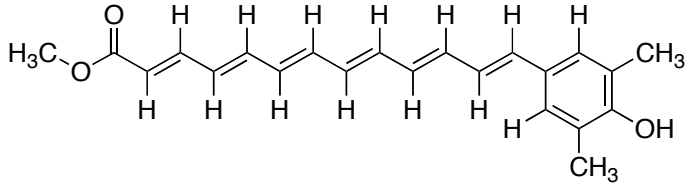

V.fischeri Polyene DMSO 600 MHz HMBC
f2 (ppm)
f1 (ppm)


V.fischeri Polyene DMSO 600 MHz ROESY
9.0
8.5
8.0
7.5
7.0
6.5
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
f2 (ppm)
1
2
3
4
5
6
7
8
9
f1 (ppm)

V.fischeri Polyene Acetone 600 MHz Proton

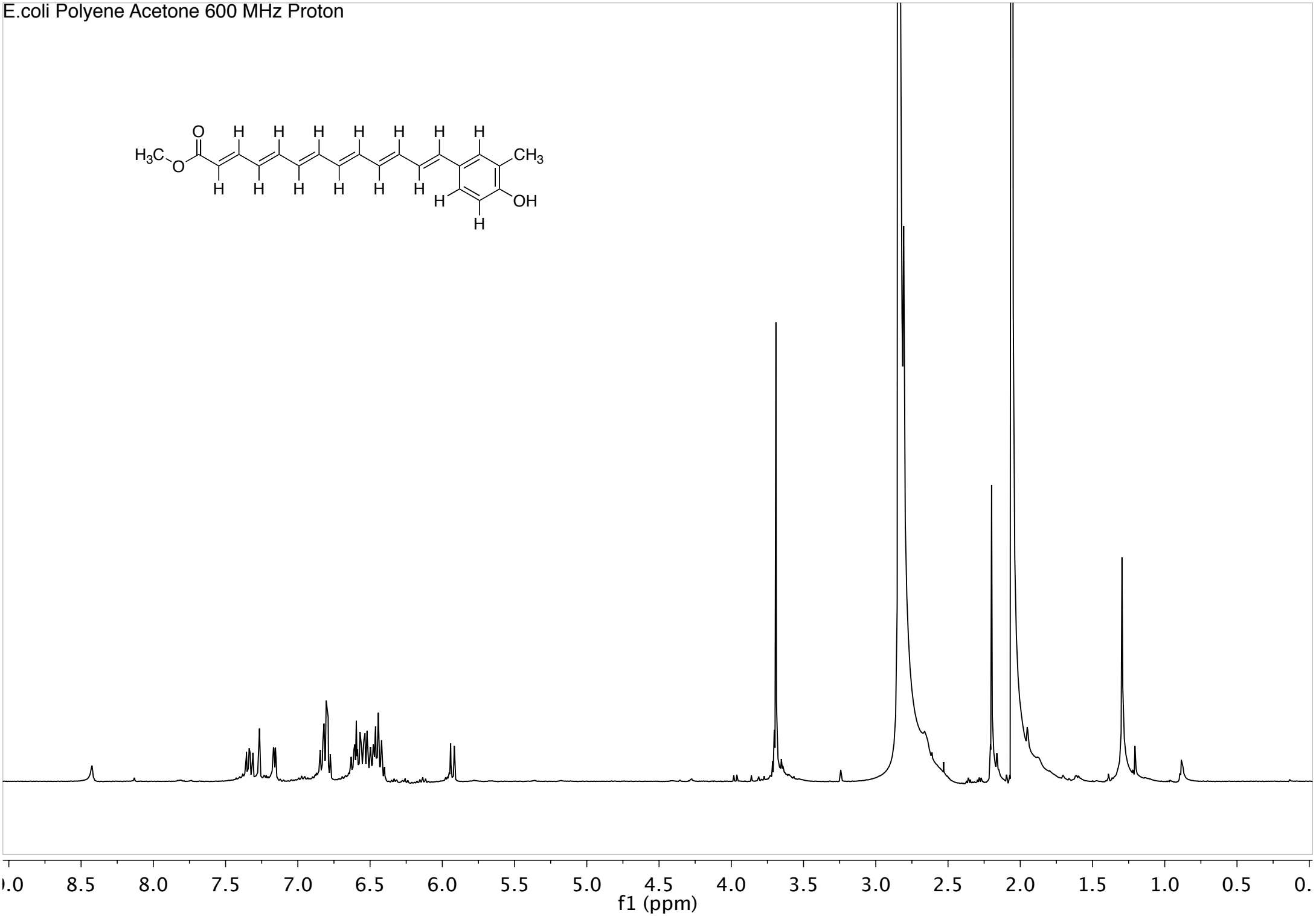


V.fischeri Polyene Acetone 600 MHz Proton

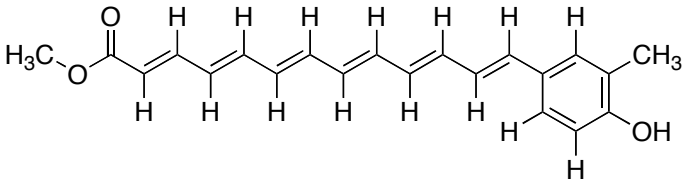


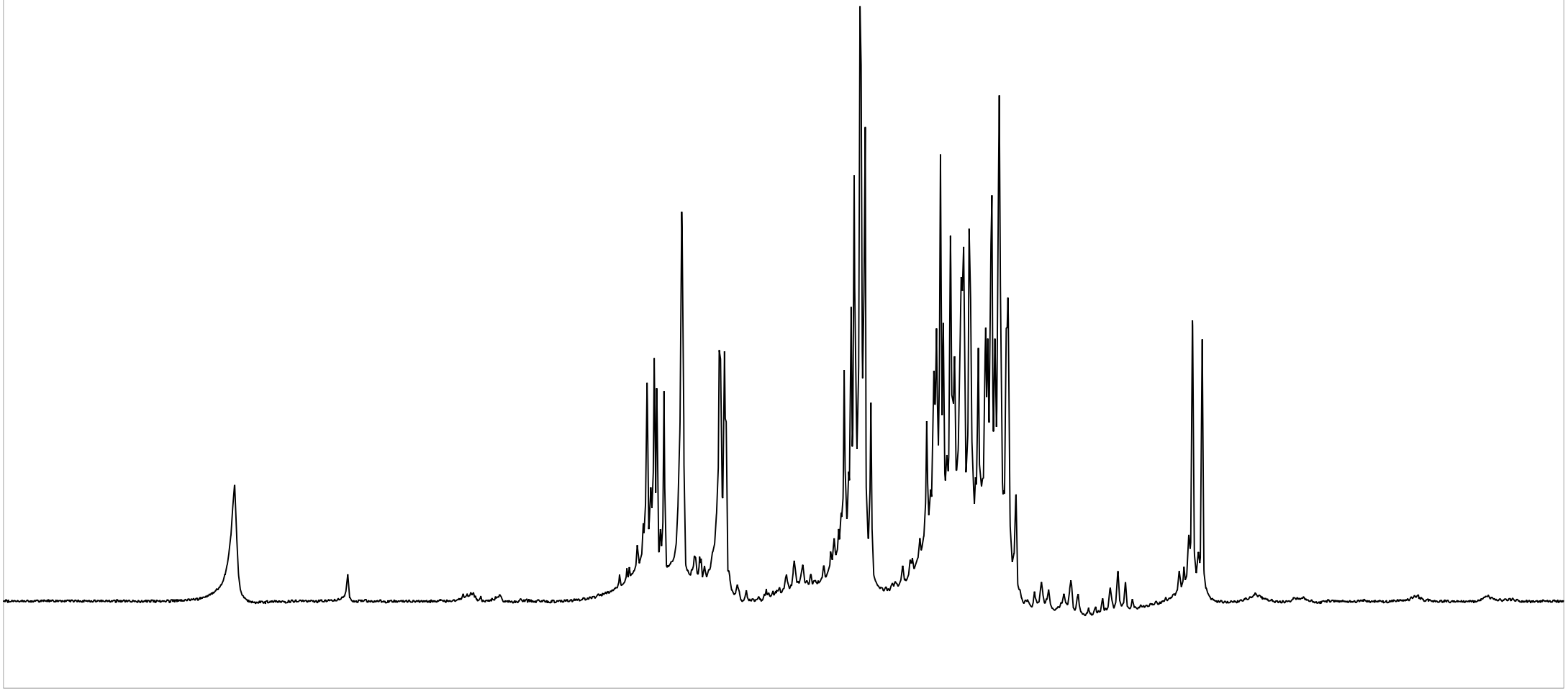
9.0 8.9 8.8 8.7 8.6 8.5 8.4 8.3 8.2 8.1 8.0 7.9 7.8 7.7 7.6 7.5 7.4 7.3 7.2 7.1 7.0 6.9 6.8 6.7 6.6 6.5 6.4 6.3 6.2 6.1 6.0 5.9 5.8 5.7 5.6 5.5 5.4 5.3 5.2 5.1 5.0

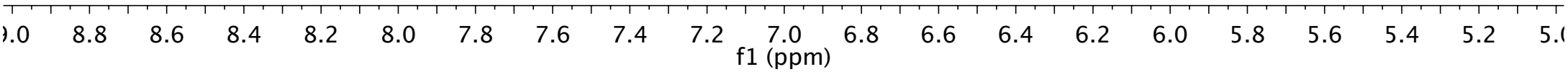
f1 (ppm)

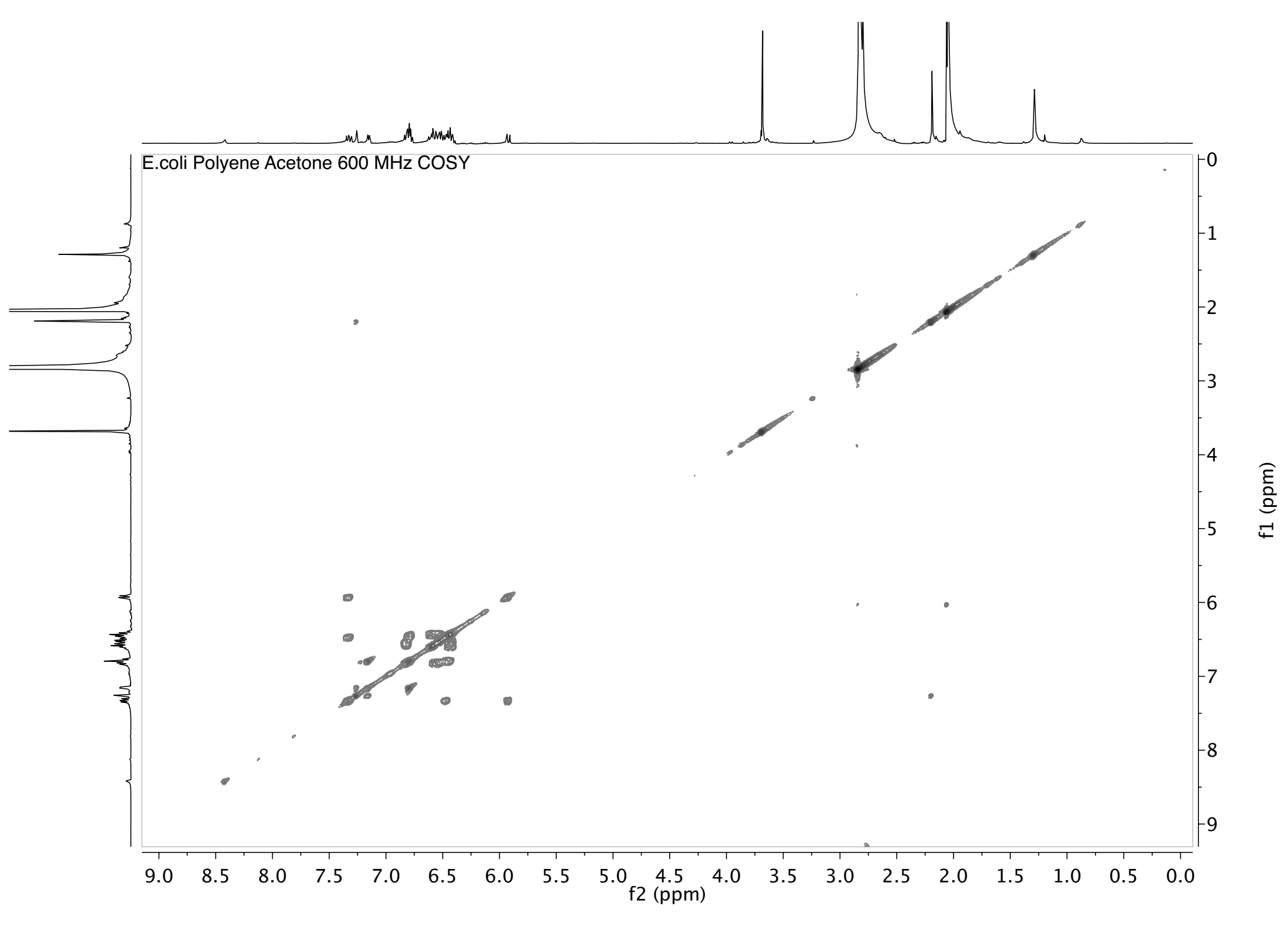

E.coli Polyene Acetone 600 MHz Proton
f1 (ppm)

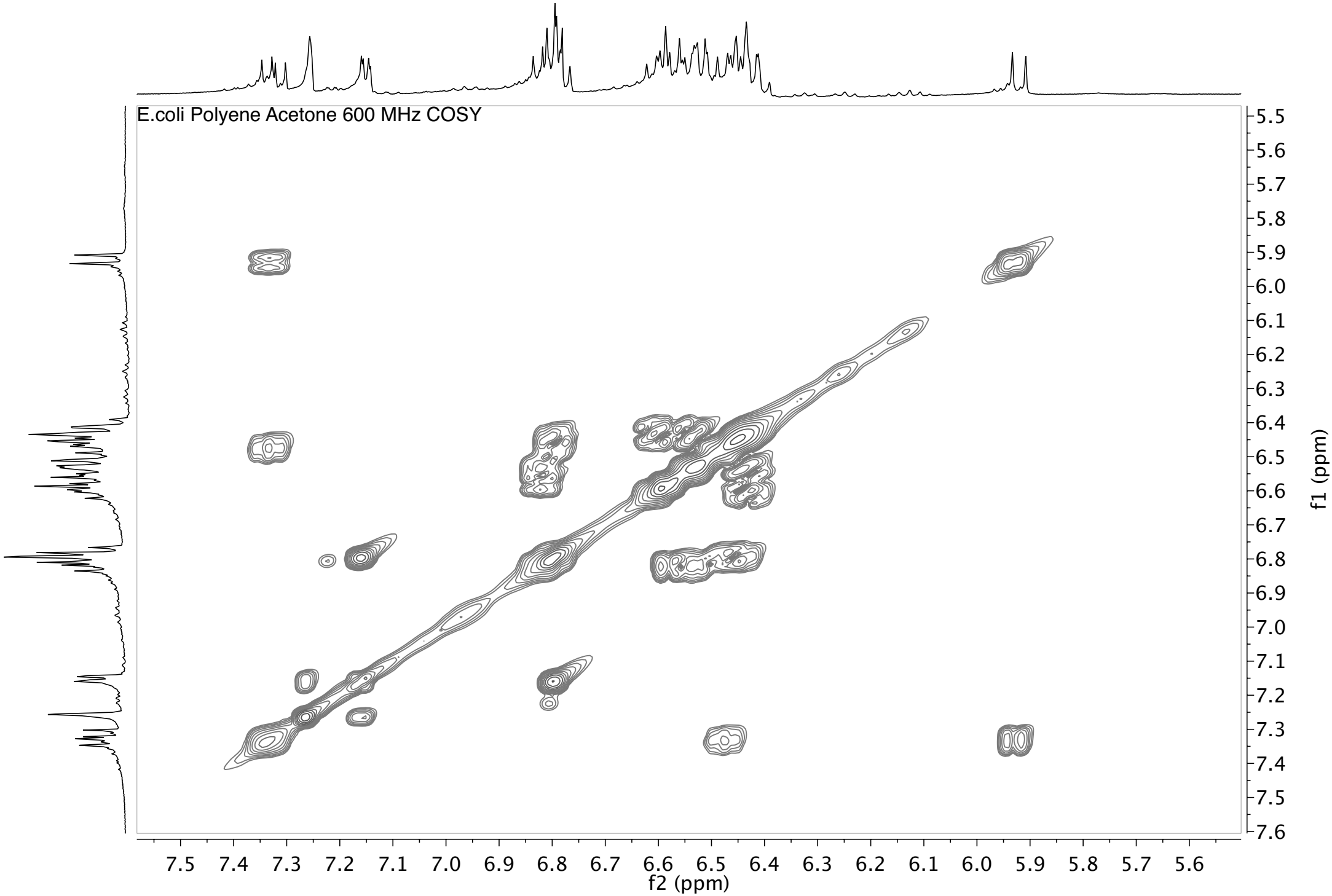
E.coli Polyene Acetone 600 MHz Proton

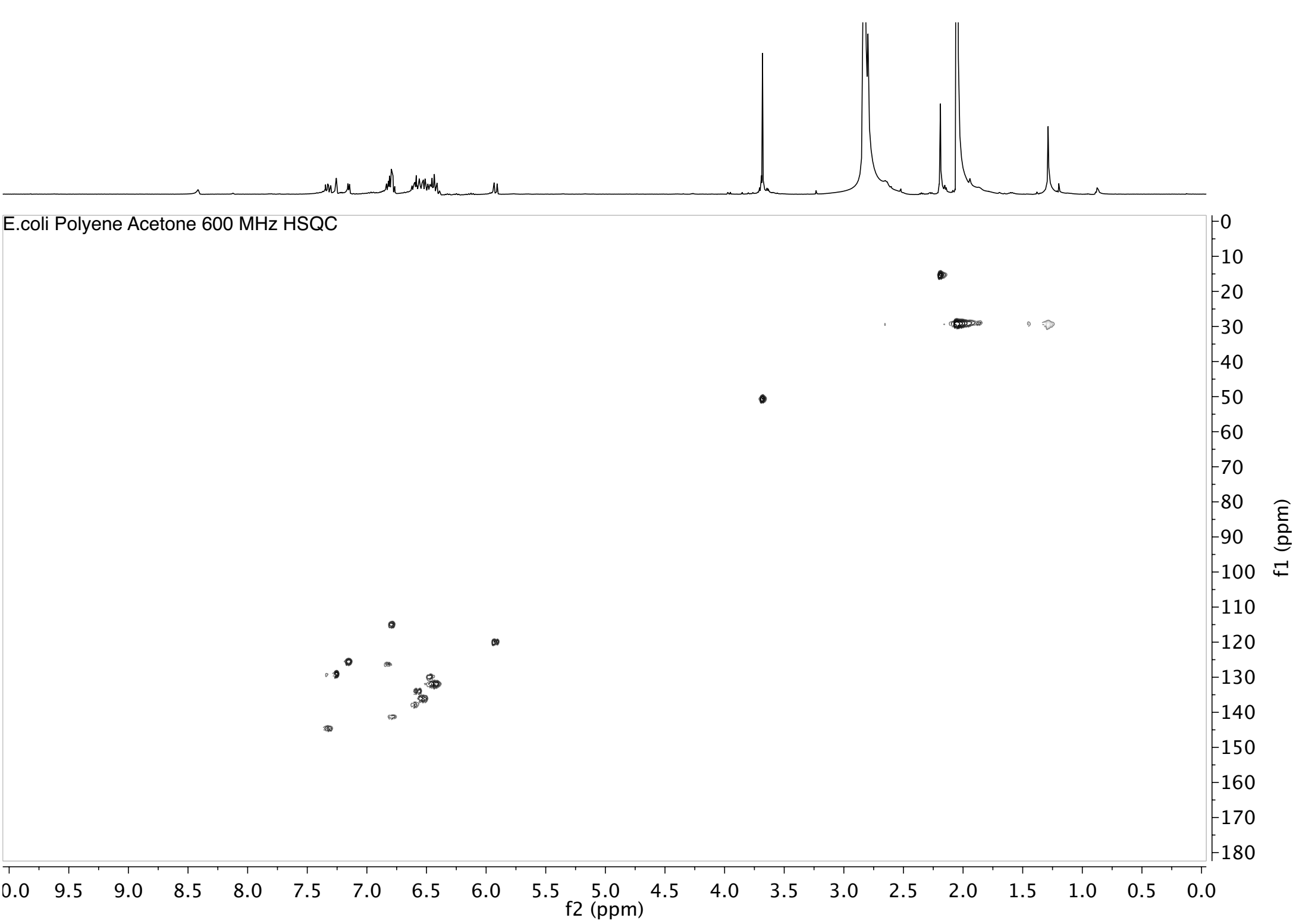


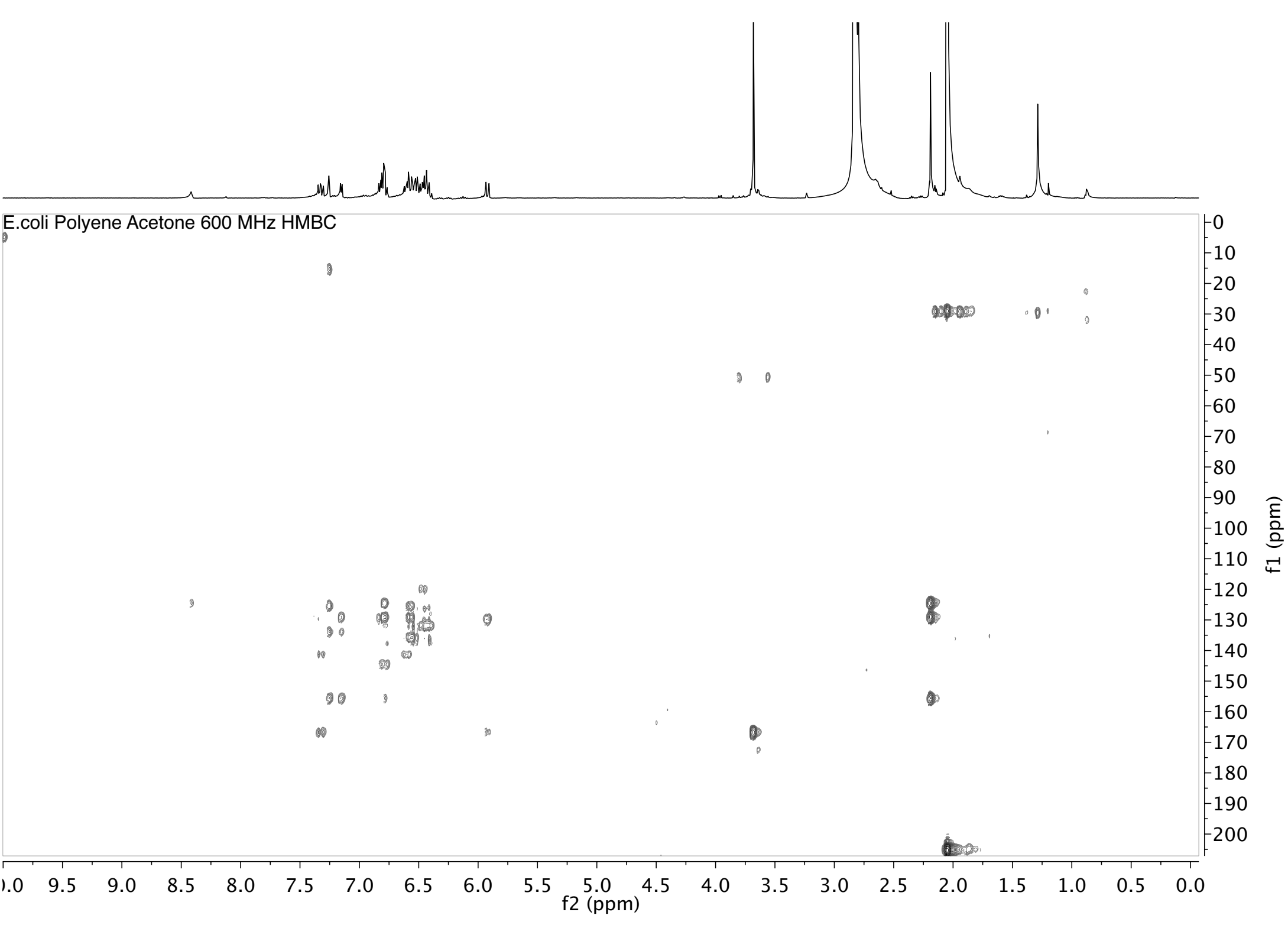


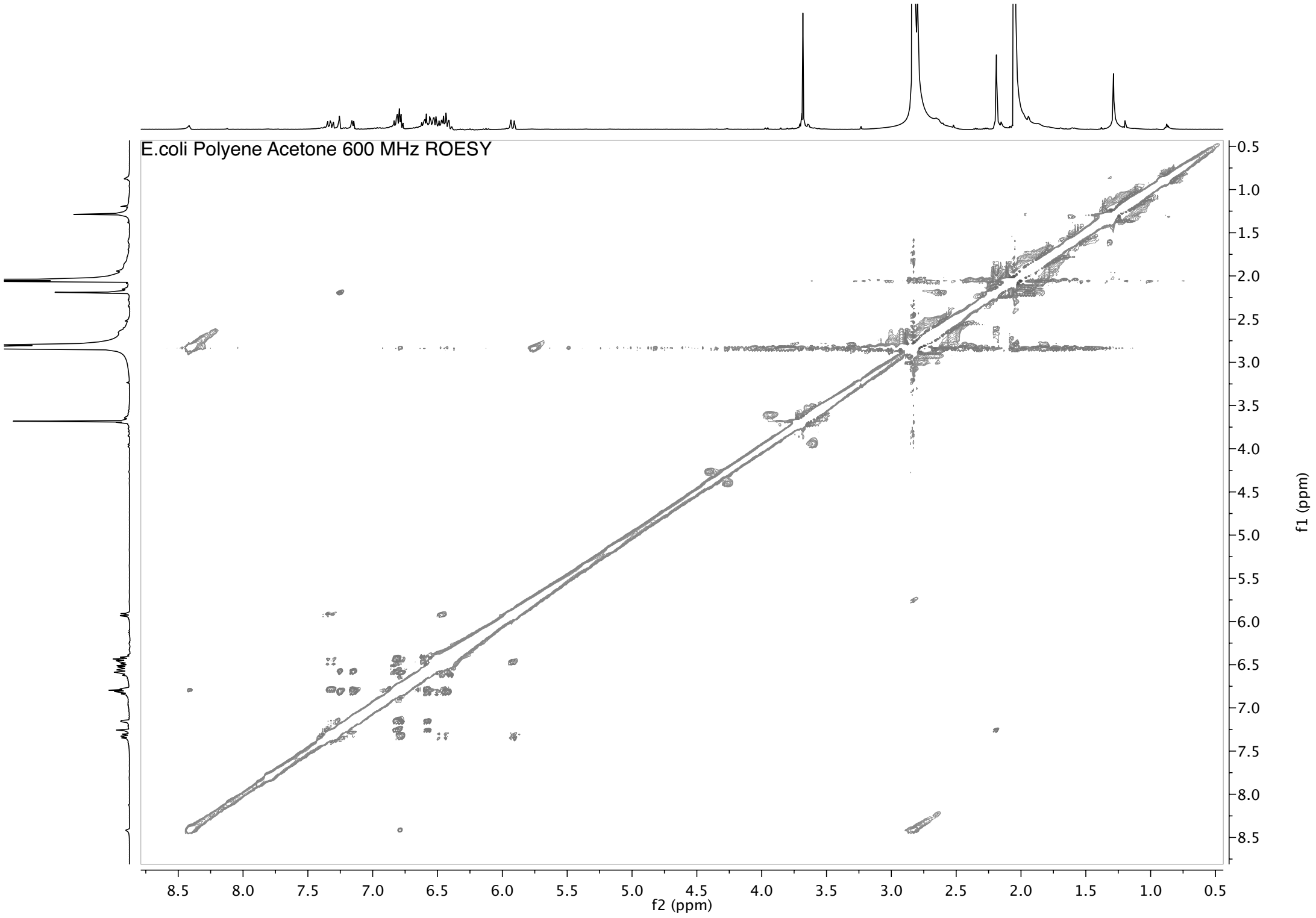


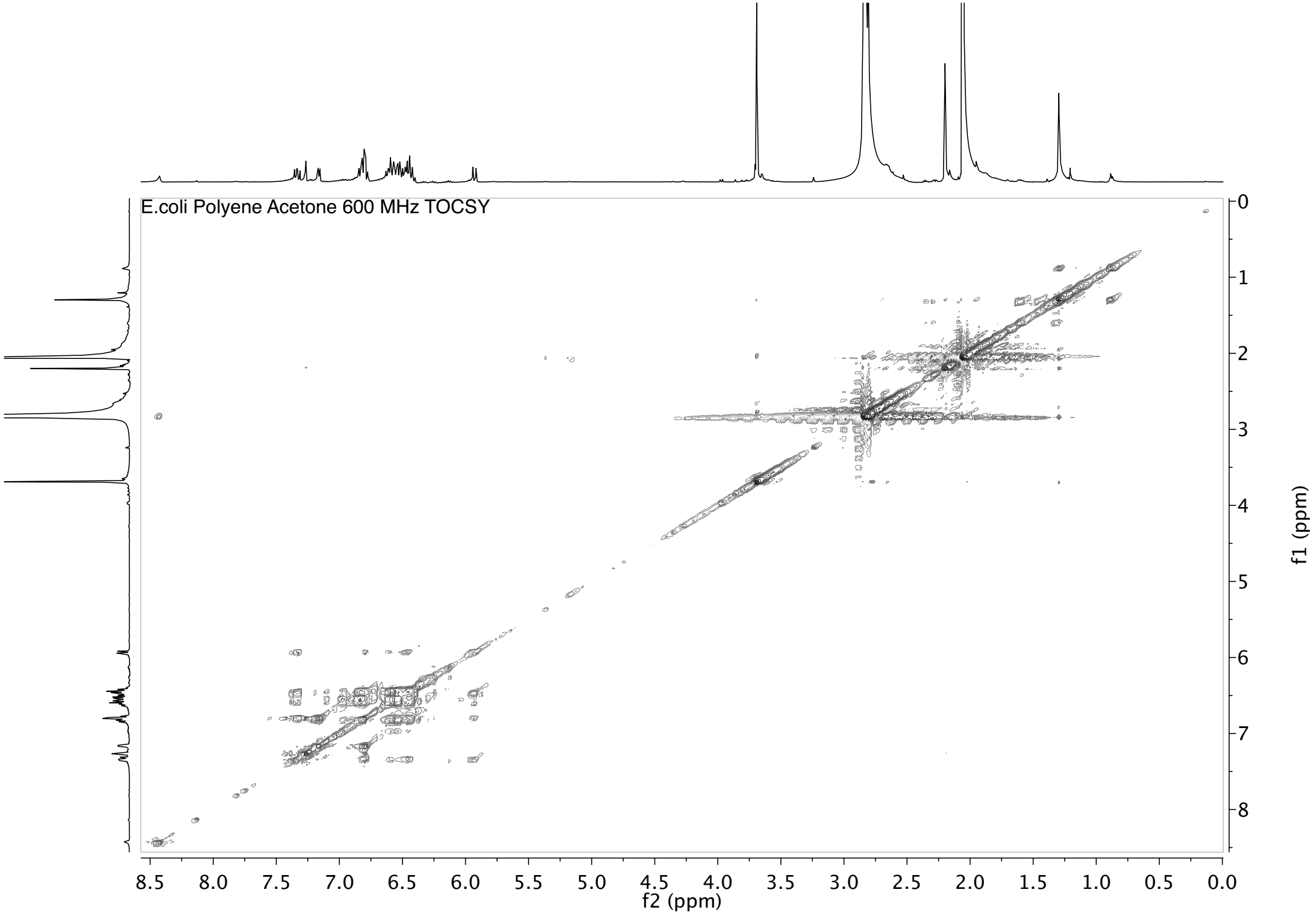

E.coli Polyene Acetone 600 MHz COSY
f2 (ppm)
f1 (ppm)
9.0
8.5
8.0
7.5
7.0
6.5
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0
1
2
3
4
5
6
7
8
9


E.coli Polyene Acetone 600 MHz COSY
7.5 7.4 7.3 7.2 7.1 7.0 6.9 6.8 6.7 6.6 6.5 6.4 6.3 6.2 6.1 6.0 5.9 5.8 5.7 5.6
f2 (ppm)
5.5 5.6 5.7 5.8 5.9 6.0 6.1 6.2 6.3 6.4 6.5 6.6 6.7 6.8 6.9 7.0 7.1 7.2 7.3 7.4 7.5 7.6
f1 (ppm)


E.coli Polyene Acetone 600 MHz HSQC
f1 (ppm)
f2 (ppm)
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
180
0.0
9.5
9.0
8.5
8.0
7.5
7.0
6.5
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0


E.coli Polyene Acetone 600 MHz HMBC
f2 (ppm)
f1 (ppm)


E.coli Polyene Acetone 600 MHz ROESY
f2 (ppm)
f1 (ppm)


E.coli Polyene Acetone 600 MHz TOCSY
8.5
8.0
7.5
7.0
6.5
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
f2 (ppm)
0
1
2
3
4
5
6
7
8
f1 (ppm)

Supplemental Figure

Click here to download Supplemental Movies and Spreadsheets: SI_Figure_12_v2.pdf

Supplemental Figure
Click here to download Supplemental Movies and Spreadsheets: SI_Figure_17.pdf

Table

Click here to download Supplemental Movies and Spreadsheets: SI_Tables.xlsx