

An updated catalog of diverse type II polyketide synthase biosynthetic gene clusters captured from large-scale nucleotide databases

1.1 Author names

Christina M. McBride (<https://orcid.org/0000-0001-9487-1525>)¹, Eric L. Miller² (<https://orcid.org/0000-0002-7157-6213>), Louise K. Charkoudian¹

1.2 Affiliation(s)

¹ Department of Chemistry, Haverford College, Haverford, PA, USA.

² Department of Biology, Haverford College, Haverford, PA, USA.

1.3 Corresponding author and email address

Eric L. Miller (emiller3@haverford.edu) and Louise K. Charkoudian (lcharkou@haverford.edu)

1.4 Keywords

Polyketide Synthase; Type II Polyketide; Biosynthetic Gene Cluster; Ketosynthase-Chain Length Factor

2. Abstract

Nature serves as a rich source of molecules with immense chemical diversity. Aptly named, these 'natural products' boast a wide variety of environmental, medicinal, and industrial applications. Type II polyketides, in particular, confer substantial medicinal benefits, including antibacterial, antifungal, anticancer, and anti-inflammatory properties. These molecules are produced by enzyme assemblies known as type II polyketide synthases (PKSs), which use domains such as the ketosynthase, chain-length factor, and acyl carrier protein to produce polyketides with varying lengths, cyclization patterns, and oxidation states. In this work, we use a novel bioinformatic workflow to identify biosynthetic gene clusters (BGCs) that encode for the core type II PKS enzymes. This method does not rely on annotation and thus was able to unearth previously 'hidden' type II PKS BGCs. This work led us to identify over 6,000 putative type II PKS BGCs spanning a diverse set of microbial phyla, nearly double those found in most recent studies. Notably, many of these newly identified BGCs were found in non-actinomycete phyla, which are relatively underexplored as sources of type II polyketides. Results from this work lay an important foundation for future bioprospecting and engineering efforts that will enable sustainable access to diverse and structurally complex molecules with medicinally relevant properties.

3. Impact statement

Nature—plants, bacteria, fungi, and even animals—has the unique ability to produce molecules with a diverse range of structures and functions. Since many of these molecules have been successfully repurposed by humans in medical, environmental, and industry settings, developing strategies to identify the organisms and biosynthetic systems that manufacture natural products could help to find novel molecules of interest. Type II polyketides have unique antimicrobial and anticancer properties, so identifying new sources of these molecules may offer renewed access to promising pharmaceuticals. By conducting a global phylogenetic analysis of the chain-length factor (CLF) protein involved in type II polyketide biosynthesis, our work provides access to previously unexplored gene clusters that may harbor unique biodiversity. Our curated CLF set nearly doubled the number of sequences available to study, increasing both the volume and diversity of potential polyketide synthases we can access *in vitro*.

4. Data summary

All sequence data were obtained from the precompiled National Center for Biotechnology Information (NCBI) Basic Local Alignment Search Tool (BLAST) databases. **The authors confirm all supporting data, code and protocols have been provided within the article or through supplementary data files.**

5. Introduction

Microbes have evolved the ability to manufacture a diverse profile of secondary metabolites that confer the host with a selective advantage over other organisms in the nearby environment (1,2). Humans have reaped the benefits of these molecules by repurposing secondary metabolites as key medicinally, environmentally, and industrially relevant agents. Type II polyketides, in particular, are

celebrated for their potent medicinal applications, including as antibiotics (e.g. tetracycline) and anticancer agents (e.g. doxorubicin) (3). These molecules are produced by type II polyketide synthases (PKSs), which are microbial protein assembly lines that iteratively assemble type II polyketides using discrete, monofunctional domains (4). The minimal type II PKS—the simplest set of enzymes needed to facilitate biosynthesis—consists of three core domains. The ketosynthase (KS, or KS α) and chain length factor (CLF, or KS β) exist as a heterodimer (namely, the KS-CLF) and drive polyketide elongation, while an acyl carrier protein (ACP) tethers and transports the acetate-based building blocks and growing polyketide intermediates to and from each enzymatic domain (5,6). Once working in concert with additional tailoring enzymes, these type II PKSs are capable of rapidly producing molecules of immense complexity (5).

Microbial enzyme assemblies like these are encoded by biosynthetic gene clusters (BGCs), which are physically grouped genes that program the production of a metabolite (7). Previous studies have established bioinformatic methods to study the evolution of type II PKS BGCs, which has enabled the inferred history to inform future bioprospecting efforts (8–10). Potential key duplication events in KS-CLF evolution were identified through the initial study, noting that type II KS-CLF BGCs may have diverged from the fatty acid synthase (FAS) FabF KS ancestor prior to the formation of the well-studied actinobacterial phylum (8,10). Most notably, this work highlighted the presence of type II PKS BGCs across a wide variety of non-Actinomycete microbes (8). Prior to this finding, research in the field was biased toward well-studied actinobacterial organisms (8), so shining light on these non-Actinomycete systems opens the way for exploration of new, previously untouched biosynthetic territory. Several non-Actinomycete type II PKS systems have been characterized, including BGCs from *Ktenobacter racemifer*, *Photorhabdus luminescens* TT01, *Streptococcus* sp. GMD2S, and *Pseudoalteromonas luteoviolacea*, revealing the possibility of innovation in this biosynthetic space (11,12). While a 2015 report ultimately identified 544 total putative type II PKS BGCs from all available nucleotide databases at the time of analysis (8), a 2022 inquiry reflects the rapid growth of deposited BGCs, reporting 3,421 total putative type II PKS BGCs found within the annotated bacterial reference sequence genomes alone (9). Analysis of the phylogeny of characterized clusters from this set highlighted how product class can overlay on the CLF evolutionary cladistics (9), further elucidating the previously noted connection to backbone chain length (8). Ancestral non-oxidative PKSs that span several bacterial phyla deviate from this trend, which suggests an ancient ancestor of these genes evolved before the divergence of these multiple phyla (9).

With the ongoing growth of available genetic data (13), routinely updating bioinformatic searches like these serve as a roadmap to understand and explore type II PKS diversity. However, relying on annotated genetic data alone to identify these proteins of interest limits our access to unique environmental and uncultured samples. Here, we establish an updated bioinformatic pipeline to identify putative genes encoding for the expression of KS and CLF proteins without the need for genome annotation. Using this pipeline, we identified 6,352 KS-CLF pairs across 5,552 nucleotide records flagged from our curated NCBI nucleotide databases, nearly doubling the available repository of type II PKS BGCs to explore. Further interpreting the evolutionary relationships between these clusters may offer us the information needed to understand these BGCs and their encoded synthases and flag regions ripe for bioprospecting.

6. Methods

6.1. Nucleotide Data

We built our nucleotide databases from the following precompiled NCBI BLAST databases (14) and locally compiled bacterial / archaeal NCBI reference sequence genomic databases:

- env_nt (metagenome/environmental nucleotide sequences; 15 June 2022),
- nt (nucleotide sequences from GenBank/EMBL/DDBJ; 15 June 2022)
- patnt (patent nucleotide sequences; 15 June 2022)
- tsa_nt (non-project based Transcriptome Shotgun Assembly entries; 15 June 2022)
- ref_prok_rep_genomes (Refseq representative prokaryotic genomes; 29 July 2022)

Therefore, our nucleotide databases consisted of complete genomes, contigs from draft genomes, contigs from environmental sequencing, transcriptome sequences, and miscellaneous short nucleotide sequences.

6.2. BLAST Search for KS and CLF Homologous Sequences

These nucleotide databases were searched using translated nucleotide BLAST (tblastn) from the NCBI BLAST+ suite v.2.11.0 for sequence homology to a set of known KS and CLF protein query sequences (15). We selected proteins from the resistomycin, fredericamycin, and WhiE spore pigment clusters as our query sequences, as these proteins have established sequence diversity among described type II PKS BGCs (8). Significant alignments with an e-value ≤ 1 were processed; this threshold detected distantly related sequences, with spurious alignments removed in later stages of our analysis.

6.3. Identifying Open Reading Frames for each BLAST Hit

We designed a custom Python script to search between 500 nucleotides before and after a BLAST hit for the longest open reading frame (ORF) that contains the BLAST hit. We considered a set of six possible start codons (ATG, CTG, GTG, TTG, ATT, ATC). An ORF was defined as the first of these start codons after a stop codon (or the edge of a given DNA sequence) and continuing until an in-frame stop codon (or the edge of a given DNA sequence). By considering alternate start codons that are not ATG, our ORFs were perhaps longer on the 5' end of the gene compared to the true start codon; however, our method makes the fewest assumptions when calling ORFs. We required that ORFs were at least 300 nucleotides long and cover at least 50% of the BLAST hit with an identical translated sequence, although the vast majority of hits matched 100% of the BLAST hit. All custom Python scripts can be found at https://github.com/EricLMiller/KS_CLF_Search.

6.4. Protein Family Classification via a Profile Hidden Markov Model

Profile hidden Markov models (pHMMs) run on HMMER v. 3.3.2 were used to differentiate which KS and CLF ORFs belonged to type II PKS BGCs (16). Predicted proteins resulting from translated ORFs were compared to standard KS and CLF sequence models established by Hillenmeyer et al. to best distinguish a given type II PKS protein from other protein families (e.g. fatty acid synthases or type I PKSs) or proteins within the same cluster (i.e. a type II KS vs. CLF) (8). Sequence similarity score thresholds were selected based on the KS (score = 352) and CLF (score = 66) scores for the aurachin gene cluster, which is only distantly related to type II PKS systems and thus can be used as a cutoff to

retain KS and CLF homologous hits (8,17). Translated ORFs with a sequence similarity score above or equal to the aurachin score thresholds were retained.

6.5. Establishing Close Proximity of the KS and CLF Genes

Using the information from the ORF-Finder on sequences that scored high in our pHMM, we wrote a custom Python script to find KS / CLF ORF pairs from the same nucleotide accession number. This script also removed pairs with over 2 kb distance between the genes to ensure that the KS and CLF do function as a heterodimer and can be coexpressed, as noted by Hillenmeyer et al (2015) (8). KS / CLF matches that overlapped were included as passing this filter. Regions with more than 2 KS / CLF ORFs each within 2 kb of each other were counted as all possible, discrete KS / CLF pairs. All matches fitting these criteria with gene lengths less than 6 kb were retained.

6.6. Phylogenetic Analysis of KS and CLF Genes

All multiple sequence alignments were produced using MAFFT Galaxy Version 7.475+galaxy0 (18). Alignment sites with more than 5% gaps across all sequences were removed to prevent undue influence of indel mutations. Maximum likelihood phylogenetic trees were built using FastTree2 Galaxy Version 2.1.10+galaxy1 with the default JTT+CAT model, and internal nodes with less than 70% bootstrap support were collapsed into polytomies (19). We visualized the tree and associated data using ggtree v. 3.0.4 and ggtreeExtra v. 1.2.3 in R Studio (20–22). We used *E. coli* FabF sequence WP_000044679.1 as an outgroup to root the tree.

6.7. Further Classification of KS/CLF Hits

We utilized the Natural Product Domain Seeker version 2 (NaPDoS2), a phylogeny-based classification webtool to predict ketosynthase and condensation domains, to verify and classify our CLF hits (23). A FASTA file containing the amino acid sequences of all relevant CLF hits was uploaded to the NaPDoS2 webpage (accessed 9/18/2022) and searched for KS domains using the default BLASTP e-value cutoff of 1×10^{-8} and minimum alignment length of 200 amino acids. We examined genomic regions 30kb before and after each KS-CLF pair; multiple pairs within this range in single genomes were combined together. We ran antiSMASH v6.1.1 (24) using a minimum ORF length of 300 bp, prodigal v2.6.3 (25) as a gene-finding tool, and ‘loose’ hmm-detection strictness. We searched the resulting antiSMASH .gbk files for gene_functions containing “t2ks”, “t2clf”, and “t2fas”. Additionally, the final CLF ORF set was also compared to sets provided by Hillenmeyer et al. (2015) and Chen et al (2022) (8,9).

7. Results

Using our pipeline (Figure 1), we identified 6,352 KS-CLF gene pairs across 5,552 unique nucleotide records. These pairs consisted of 6,209 KSs and 6,322 CLFs; this discrepancy in the number of proteins arises from one KS being located in proximity to multiple CLFs, or vice versa.

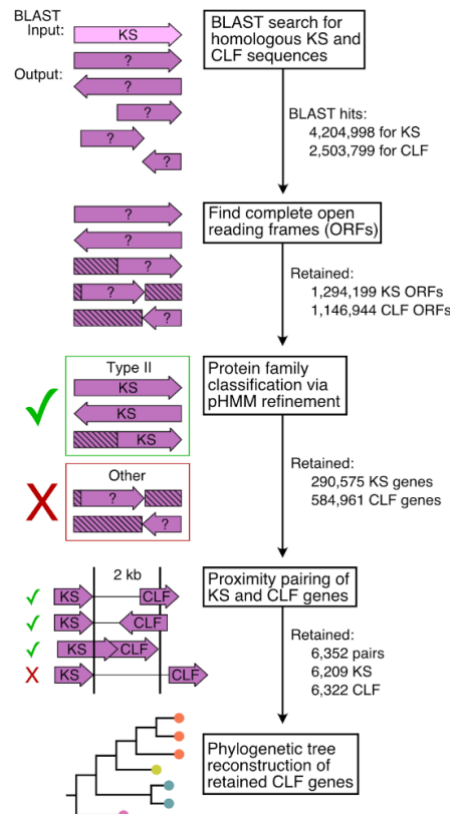


Figure 1. Bioinformatic pipeline to identify type II PKS KS-CLFs. Our workflow identified 6,352 putative genes encoding for the expression of KS-CLFs across 14 phyla.

We validated this pipeline by comparing the set of 167 characterized type II PKS CLFs reported by Chen et al (2022) with our 6,322 CLF hits (Figure S1). Our curated CLF set contained 95.2% of these characterized CLFs, with the following sequences missing in our analysis: five characterized CLFs that are no longer annotated on any genome available through NCBI; the dactylocycline A and thioangucycline CLFs reported by Chen, instead classifying these sequences as KSs; and the CLF for AQ-256, which was found using our BLAST-based search and retained based on our pHMM score but did not have a corresponding KS within 2,000 base pairs and was thus removed during the proximity pairing stage. Since we found nearly all of the previously reported characterized type II PKS CLFs, and all the CLFs we would expect to find given our workflow, we are confident that our pipeline can accurately identify a broad range of CLF diversity.

Our curated KS-CLFs were found across both actinobacterial and non-actinobacterial prokaryotic species, representing, in total, 14 phyla: Acidithiobacilla, Acidobacteria, Actinobacteria, Proteobacteria (alpha, beta, delta, and gamma), Armatimonadetes, Candidatus Aminicenantes, Candidatus Cryoseriata, Candidatus Omnitrophica, Chloroflexi, Cyanobacteria, Euryarchaeota, Firmicutes, Nitrospirae, and Planctomycetes (Figure 2). Also included in this dataset were KS-CLFs identified from environmental samples and patented sequences. The environmental KS-CLF hits originated from both uncultured bacteria and various soil, bioreactor, groundwater, gut, human gut, hot spring, hydrothermal vent, marine, sediment, symbiont, and wastewater metagenomes, illustrating the widespread prevalence of type II PKS enzymes. 3380 of these CLF hits were also

reported by Chen et al. in their 2022 study, meaning our remaining 2,942 CLF hits were newly identified by our pipeline. Notably, we unearthed a large, non-actinobacterial clade not identified by Chen et al (Figure 2, middle circle).

To assess the significance of our data, we plotted the pHMM score for each CLF ORF onto the phylogeny, observing lower pHMM scores (corresponding to a weaker match to the selected model) for non-actinobacterial PKSs. The pHMM scores for actinobacterial hits had a median of 801.8 (with lower and upper quartiles of 747.8 and 893.7, respectively), while the pHMM scores for non-actinobacterial hits were much lower with a median of 132.25 (with lower and upper quartiles of 87 and 194.85, respectively). Since the pHMM score of the aurachin CLF was 66, these clusters score better against the model than the most distant type II BGC included in our tree, suggesting that these lower scores do not necessarily indicate that they are not true type II CLFs. As a result, to further verify the reasonableness of our CLF set, we used NaPDoS2 to detect and classify KS and CLF domains from our dataset. Ultimately, the NaPDoS2 algorithm identified 5,497 type II CLF domains, 181 type II KS domains, 464 type II fatty acid synthase (FAS) domains, and 22 KS or condensation domains from other synthases (type II polyene, type II aryl polyene, and type I modular cis-acyl transferase). 158 of our CLF sequences were not classified.

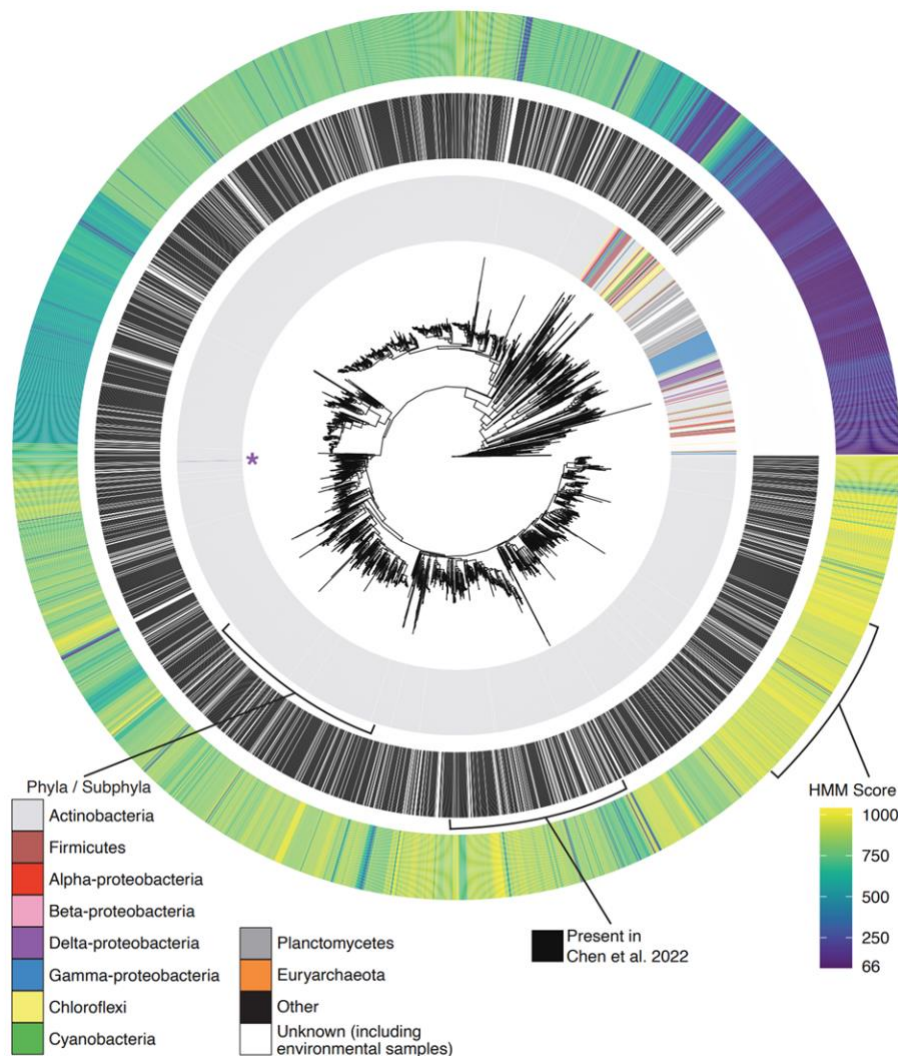


Figure 2. Phylogenetic tree representing 6,322 type II PKS CLF protein sequences. The inner ring represents the phyla classification for the sequence, the middle ring signifies which sequences were also identified by Chen et al. in their recent publication, and the outer ring notes the pHMM score for each CLF amino acid sequence. The purple star indicates the location of a single non-Actinomycete hit (the Deltaproteobacterium *Melittangeum boletus* DSM 14713) within the large Actinomycete clade. This phylogeny reveals a previously unexplored clade (indicated by the large white section of the middle ring) of primarily non-actinobacterial type II PKS CLFs.

We continued our analysis by focusing on the non-actinobacterial sequences identified by our pipeline (Figure 3). For accuracy of phyla determination, we excluded environmental samples and patented sequences from these analyses. Out of the 565 non-actinobacterial CLF hits, NaPDoS2 classified 163 hits as type II CLFs, 107 hits as type II KSs, 263 hits as type II FAS KS domains, and 12 KS or condensation domains from other synthases (type II polyene or type II aryl polyene). 20 of these non-actinobacterial CLF sequences were not classified. Out of those sequences classified as type II KSs or CLFs, only 145 had predicted product types. Interestingly, whereas the large actinobacterial

clade showed a connection between product type and phylogenetic similarity, we did not observe evidence of this association for the non-actinobacterial CLF sequences (Figure S2).

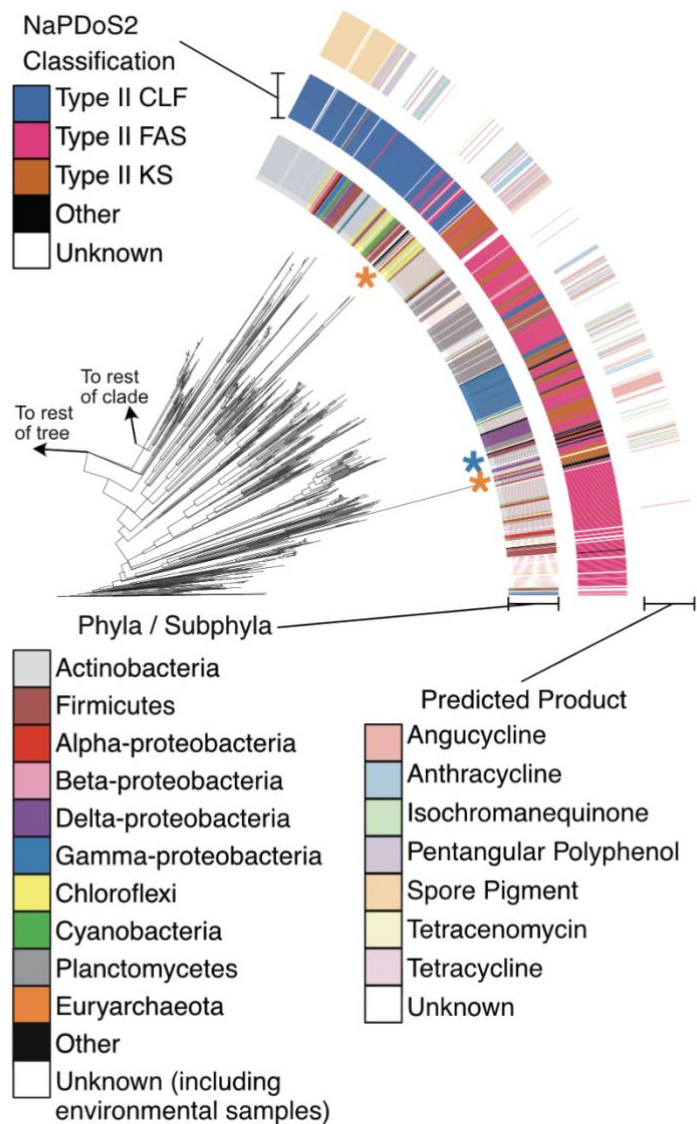


Figure 3. Subset of phylogenetic tree representing the primarily non-Actinomycete clade of type II PKS CLF proteins. The inner ring represents the taxonomic classification, the middle ring identifies how the sequences were classified by NaPDoS2, and the outer ring represents the predicted molecular product type per NaPDoS2. The orange stars represent the locations of our archaeal hits, while the blue star represents the aurachin gene cluster chosen as the cutoff for similarity to type II PKS BGCs. The existence of type II PKS KS and CLF hits within this paraphyletic group closer to the root (as determined by outgroup *E. coli* FabF sequence WP_000044679.1) confirms that this previously unexplored region could represent an evolutionary distinct subset of type II PKS BGCs, rather than false positive hits.

While NaPDoS2 classified some of the hits within this previously unexplored non-actinobacterial clade as type II CLFs, others were instead classified as type II FAS KSs, type II KSs, other condensation domains, or even unknown proteins. To further understand these classifications, we utilized antiSMASH to look for type II condensation domains in our nucleotide records of interest. As described in Table 1, we identified the presence of three different condensation domain pairings: 1) type II KS and type II CLF, 2) type II KS and type II FAS, and 3) two type II FASs. The most traditional type II PKS would be represented by those containing the type II KS and type II CLFs, which occurred in 5267 Actinomycete pairs and 87 non-Actinomycete pairs. However, the type II KS and type II FAS pairing (233 pairs in Actinomycetes, 94 pairs in non-Actinomycetes) and the type II FAS – FAS pairing (196 pairs in Actinomycetes, 319 pairs in non-Actinomycetes) have a significant presence in our curated CLF set. The existence of these different pairings of condensation domains may indicate that our pipeline captured CLF sequence diversity extending beyond the traditional understanding of the differences between type II PKS and type II FAS systems.

Table 1. antiSMASH-identified genes within antiSMASH-detected BGCs within 30kb of our KS/CLF pairs.

	Type II KS & CLF	Type II KS & FAS	Type II FAS & FAS
Actinomycetes			
KS-CLF Pairs (5348 total)	5267 (98.5%)	233 (4.4%)	196 (3.7%)
Unique GenBank Accessions ¹ (4862 total)	4799 (98.7%)	232 (4.8%)	196 (4.0%)
Non-Actinomycetes			
KS-CLF Pairs (546 total)	87 (15.9%)	94 (17.2%)	319 (58.4%)
Unique GenBank Accessions ¹ (501 total)	87 (17.4%)	91 (18.2%)	315 (62.9%)

¹8.6% of Actinomycete accessions and 3.0% of non-Actinomycete accession numbers have multiple regions >30 kb apart, each with KS-CLF pairs, indicating the presence of multiple type II PKS BGCs within the same nucleotide record.

8. Discussion

Our bioinformatic pipeline identified 6,322 individual CLF hits, nearly double those found by Chen et al. in April 2022 and around eleven times larger than the set reported in 2015 (8,9), confirming the repeated need for updated bioinformatic efforts as the nucleotide databases grow. The variation of phyla represented in the tree exceeds that of previous searches (ours represented fourteen phyla, while Chen et al had eight and Hillenmeyer et al had ten), opening up routes for new type II polyketide biosynthetic diversity outside of the traditional actinobacterial space. Notably, we identified a primarily non-actinobacterial clade that has not been previously identified in the literature. Past bioinformatic searches have utilized a variety of criteria (Pfam domain annotation, pHMM searches, antiSMASH BGC classification) with varying stringency, so it is unsurprising that

these different methods produce different sequence sets. Our model may be less stringent than recent searches, leading to a broader selection of CLFs that could represent more distantly evolved, biosynthetically unique BGCs.

It is important to highlight the archaeal KS-CLF hits identified by our pipeline. Previously, modern archaea and their last common ancestor have been shown to contain FAS pathways, but these archaeal systems lack an ACP (26). More recently, a survey of modular enzymatic synthase BGCs—which produce similar small molecule products to type II PKSs—was performed across bacterial, archaeal, and eukaryotic genomes. Three archaeal non-ribosomal peptide synthetase (NRPS) pathways were identified in strains of *Methanobacteria* and *Methanomicrobia* but no type I PKS BGCs were found in archaea, suggesting that these small-molecule-producing clusters are elusive and rare (27). Our pipeline ultimately reported three CLFs from *Euryarchaeota*, all members of the *Candidatus Methanoperedenaceae* archaeon species. While NaPDoS2 classified one of these CLF hits as belonging to a type II FAS cluster, the other two hits were classified as type II PKS KS proteins, posing the question as to whether archaea may actually be capable of producing polyketide-like products.

Mirroring the results of recent phylogenetic analyses (9), the Deltaproteobacterium *Melittangeum boletus* DSM 14713 was the only non-actinobacterial CLF found within a large, otherwise actinobacterial clade. Our results are consistent with a single large horizontal gene transfer (HGT) event, as suggested by Chen et al. (2022).

As a result of their relative evolutionary separation, there are distinct differences between the various phyla represented in our tree. On average, the actinobacterial CLFs scored substantially higher in our pHMM model than non-actinobacterial CLFs. To some extent, this discrepancy indicates that the non-actinobacterial CLFs are less similar to known, characterized CLF proteins. However, since all of the CLF proteins used to build the model were from Actinomycete organisms, this lower score may actually reflect the current bias in the field toward actinobacterial type II PKSs.

To further verify our CLF hits, we ran all 6,322 CLF sequences through NaPDoS2, a phylogeny-powered webtool that classifies various KS and condensation domains (23). Approximately 95% of our actinobacterial CLF hits were classified as type II PKS CLF domains by NaPDoS2, verifying the relative accuracy of our pipeline. On the other hand, only 29% of our non-actinobacterial CLF hits were classified as type II PKS CLF domains via NaPDoS2, whereas 48% were classified as type II FAS, 17% were classified as type II KS, 2% were classified as some other KS domain, and the remaining 5% were not classified. Most of the non-actinobacterial hits classified as type II CLFs have a closer evolutionary relationship to actinobacterial CLFs, indicating that the characteristic sequence motifs of CLF proteins are dictated by those found across actinobacterial CLFs. Nonetheless, this presence of CLFs across both actinobacterial and non-actinobacterial phyla supports the inference that the ancient FAS *ks* gene that was duplicated to form the type II PKS KS and CLF genes originated prior to the most recent common ancestor of the actinobacteria and of *E. coli*, whose FAS is used as an outgroup for these genes (8,10).

While the CLF proteins with closer homology to the *E. coli* FAS KS were less likely to be classified as type II PKS CLFs by NaPDoS2, this does not mean that these hits are extraneous sequences accidentally retained by our pipeline as a previously unexplored clade. Many of these CLF hits were classified as either type II FAS or type II KS domains by NaPDoS2; since our workflow mandated the presence of a KS and CLF gene within 2000 bp of each other to retain them as a KS-CLF pair, each of these hits must have another KS-like domain within these bounds. Therefore, these seemingly extraneous hits could represent BGCs that evolved directly after the ancient FAS *ks* gene duplication and thus have two near-identical KS domains that lack the typical motifs seen across most actinobacterial CLFs. Despite their relative infrequency, there were CLF hits within this previously unexplored clade that were classified as type II CLF genes, indicating that our workflow is capturing distant CLF genes that may harbor unique sequence and biosynthetic diversity.

Our antiSMASH analysis of the regions surrounding our hits further support this inference. While 15.9% of our non-actinobacterial hits had the traditional type II KS – type II CLF pairing, 17.2% of these examined regions contained type II KS – type II FAS pairs and 58.4% of these regions contained type II FAS – type II FAS pairs. Differences between the percentage of Actinomycete and non-Actinomycete pairs classified as type II KS / type II CLF are hard to interpret due to the Actinomycetes sequenced at a much high phylogenetic density compared to the non-Actinomycetes; genomes are often not randomly selected to be sequenced, and so the resulting collection of genomic information in NCBI can be biased accordingly. The type II KS – type II FAS pairing represents a relatively unexplored region of BGC diversity. Hits of this type contain two condensation domains: one of which is homologous to known type II KSs, the other of which is most similar to known type II FAS KS domains. This pairing may indicate that hybrid type II PKS / type II FAS systems exist, and that our definition of what constitutes a “CLF” may need to be broadened, such that type II CLFs may actually look much more similar to type II FASs than we previously understood. We also note several non-canonical clusterings of three condensation domains, such as type II KS / type II FAS / type II CLF or type II KS / type II CLF / type II CLF triads. Further characterizing atypical BGC types may reveal biosynthetic diversity beyond our current understanding of type II PKS and type II FAS systems. While the type II FAS – type II FAS pairs may just represent regions where the FAS KS gene duplicated and thus still belong to a traditional type II FAS system, these clusters still represent an area of biosynthetic diversity that should be explored. Delineating the differences between these pairs of condensation enzymes may provide further insight into what constitutes a type II PKS vs. a type II FAS and how we might manipulate these systems to allow them to combinatorially interact.

Our bioinformatic pipeline, developed using the computational resources made available through a primarily undergraduate institution, can be easily understood and manipulated by researchers with limited computational experience. While initially designed to identify type II PKS KS-CLF pairs, our pipeline is not unique to type II PKS systems; rather, each step can be tailored to incorporate the parameters needed to produce a phylogeny of any evolutionarily conserved gene or protein, such as by changing the reference set used for comparison in the pHMM model or changing the conditions for gene length or proximity. Since our pipeline does not rely on previous annotation, users can systematically identify homologs of their sequence of interest from sequence data of varying completeness or accuracy.

9. Conclusions

Unveiling new routes to type II polyketide-like molecules could offer swift access to promising bioactivity. Using the bioinformatic pipeline described herein, we report an updated profile of CLF diversity – a critical and unique protein to type II PKSs – across fourteen distinct microbial phyla. Along with unveiling a previously undescribed, primarily non-actinobacterial clade, we share potential evidence of type II polyketide biosynthetic potential in archaea. We further provide additional support for the inference that the type II KS and CLF genes diverge from a common ancestor of the *E. coli* KS prior to the formation of the actinobacterial clade.

The curated CLF set itself serves as a “biosynthetic catalog”, offering insight into potential leads for type II PKS bioprospecting efforts. Exploring the previously undescribed non-actinobacterial clade may unveil new type II PKS enzymatic and molecular diversity, guiding future efforts in the biosynthesis of never-before-seen secondary metabolites. Understanding the existence of this clade further may allow us to dissect the differences between type II FAS and PKS KS-like domains, gaining insight into the divergent evolution of these systems. We share these results with the community with the hope that they will spark efforts to characterize these diverse BGCs and encoded enzymes. Uncovering the biosynthetic prowess of these systems represents a critical step in gaining sustainable access to important molecules.

10. Author statements

10.1 Author contributions

L.K.C., E.L.M., and C.M.M. designed the research; E.L.M. wrote original python scripts; C.M.M. and E.L.M. collected and curated the data; L.K.C., E.L.M., and C.M.M. analyzed the data; and L.K.C., E.L.M., and C.M.M. wrote the manuscript.

10.2 Conflicts of interest

The author(s) declare that there are no conflicts of interest.

10.3 Funding information

We acknowledge generous support from the National Science Foundation (CHE1652424 and CHE2201984 to L.K.C.), a 2021 Arnold and Mabel Beckman Foundation Scholarship (to C.M.M.), and a 2022 Goldwater Scholarship (to C.M.M.).

10.4 Acknowledgements

We thank Joe Cammisa and Dr. Sara Matheison (Haverford College) for their technical guidance. We also thank Dr. Marnix Medema (Wageningen University) for helpful discussions.

11. References

1. Katz L, Baltz RH. Natural product discovery: past, present, and future. *J Ind Microbiol Biotechnol*. 2016 Mar;43(2–3):155–76. doi:10.1007/s10295-015-1723-5

- 398 2. Dias DA, Urban S, Roessner U. A historical overview of natural products in drug discovery.
399 Metabolites. 2012 Apr 16;2(2):303–36. doi:10.3390/metabo2020303
- 400 3. Hertweck C. The Biosynthetic Logic of Polyketide Diversity. *Angewandte Chemie International*
401 *Edition*. 2009;48(26):4688–716. doi:10.1002/anie.200806121
- 402 4. Fischbach MA, Walsh CT. Assembly-line enzymology for polyketide and nonribosomal Peptide
403 antibiotics: logic, machinery, and mechanisms. *Chem Rev*. 2006 Aug;106(8):3468–96.
404 doi:10.1021/cr0503097
- 405 5. Wang J, Zhang R, Chen X, Sun X, Yan Y, Shen X, et al. Biosynthesis of aromatic polyketides in
406 microorganisms using type II polyketide synthases. *Microb Cell Fact*. 2020 Dec;19(1):110.
407 doi:10.1186/s12934-020-01367-4
- 408 6. Sulpizio A, Crawford CEW, Kowec RS, Charkoudian LK. Probing the structure and function of
409 acyl carrier proteins to unlock the strategic redesign of type II polyketide biosynthetic pathways.
410 *J Biol Chem*. 2021 Jun;296:100328. doi:10.1016/j.jbc.2021.100328
- 411 7. Walsh CT, Fischbach MA. Natural Products Version 2.0: Connecting Genes to Molecules. *J Am*
412 *Chem Soc*. 2010 Mar 3;132(8):2469–93. doi:10.1021/ja909118a
- 413 8. Hillenmeyer ME, Vandova GA, Berlew EE, Charkoudian LK. Evolution of chemical diversity by
414 coordinated gene swaps in type II polyketide gene clusters. *Proc Natl Acad Sci USA*. 2015 Nov
415 10;112(45):13952–7. doi:10.1073/pnas.1511688112
- 416 9. Chen S, Zhang C, Zhang L. Investigation of the Molecular Landscape of Bacterial Aromatic
417 Polyketides by Global Analysis of Type II Polyketide Synthases. *Angewandte Chemie*
418 *International Edition*. 2022;61(24):e202202286. doi:10.1002/anie.202202286
- 419 10. Ridley CP, Lee HY, Khosla C. Evolution of polyketide synthases in bacteria. *Proceedings of the*
420 *National Academy of Sciences*. 2008 Mar 25;105(12):4595–600. doi:10.1073/pnas.0710107105
- 421 11. Klein JG, Wu Y, Kokona B, Charkoudian LK. Widening the bottleneck: Heterologous expression,
422 purification, and characterization of the Ktedonobacter racemifer minimal type II polyketide
423 synthase in Escherichia coli. *Bioorg Med Chem*. 2020 Oct 15;28(20):115686.
424 doi:10.1016/j.bmc.2020.115686
- 425 12. Cummings M, Peters AD, Whitehead GFS, Menon BRK, Micklefield J, Webb SJ, et al. Assembling
426 a plug-and-play production line for combinatorial biosynthesis of aromatic polyketides in
427 Escherichia coli. *PLOS Biology*. 2019 Jul 18;17(7):e3000347. doi:10.1371/journal.pbio.3000347
- 428 13. NCBI. GenBank and WGS Statistics [Internet]. [cited 2022 Aug 24]. Available from:
429 <https://www.ncbi.nlm.nih.gov/genbank/statistics/>
- 430 14. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol*
431 *Biol*. 1990 Oct 5;215(3):403–10. doi:10.1016/S0022-2836(05)80360-2
- 432 15. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture
433 and applications. *BMC Bioinformatics*. 2009 Dec 15;10:421. doi:10.1186/1471-2105-10-421

- 434 16. Eddy SR. Accelerated Profile HMM Searches. *PLOS Computational Biology*. 2011 Oct
435 20;7(10):e1002195. doi:10.1371/journal.pcbi.1002195
- 436 17. Pistorius D, Li Y, Sandmann A, Müller R. Completing the puzzle of aurachin biosynthesis in
437 *Stigmatella aurantiaca* Sg a15. *Mol Biosyst*. 2011 Dec;7(12):3308–15. doi:10.1039/c1mb05328k
- 438 18. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements
439 in Performance and Usability. *Mol Biol Evol*. 2013 Apr;30(4):772–80.
440 doi:10.1093/molbev/mst010
- 441 19. Price MN, Dehal PS, Arkin AP. FastTree 2 – Approximately Maximum-Likelihood Trees for Large
442 Alignments. *PLOS ONE*. 2010 Mar 10;5(3):e9490. doi:10.1371/journal.pone.0009490
- 443 20. Yu G. Using ggtree to Visualize Data on Tree-Like Structures. *Current Protocols in Bioinformatics*.
444 2020;69(1):e96. doi:10.1002/cpbi.96
- 445 21. Xu S, Dai Z, Guo P, Fu X, Liu S, Zhou L, et al. ggtreeExtra: Compact Visualization of Richly
446 Annotated Phylogenetic Data. *Molecular Biology and Evolution*. 2021 Sep 1;38(9):4039–42.
447 doi:10.1093/molbev/msab166
- 448 22. RStudio Team. RStudio: Integrated Development for R.
- 449 23. Klau LJ, Podell S, Creamer KE, Demko AM, Singh HW, Allen EE, et al. The Natural Product Domain
450 Seeker version 2 (NaPDoS2) webtool relates ketosynthase phylogeny to biosynthetic function.
451 *Journal of Biological Chemistry* [Internet]. 2022 Sep 11 [cited 2022 Sep 18];0(0). Available from:
452 [https://www.jbc.org/article/S0021-9258\(22\)00923-1/abstract](https://www.jbc.org/article/S0021-9258(22)00923-1/abstract) doi:10.1016/j.jbc.2022.102480
- 453 24. Blin K, Shaw S, Kloosterman AM, Charlop-Powers Z, van Wezel GP, Medema MH, et al.
454 antiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucleic Acids*
455 *Research*. 2021 Jul 2;49(W1):W29–35. doi:10.1093/nar/gkab335
- 456 25. Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene
457 recognition and translation initiation site identification. *BMC Bioinformatics*. 2010 Mar
458 8;11(1):119. doi:10.1186/1471-2105-11-119
- 459 26. Lombard J, López-García P, Moreira D. An ACP-Independent Fatty Acid Synthesis Pathway in
460 Archaea: Implications for the Origin of Phospholipids. *Molecular Biology and Evolution*. 2012
461 Nov 1;29(11):3261–5. doi:10.1093/molbev/mss160
- 462 27. Wang H, Fewer DP, Holm L, Rouhiainen L, Sivonen K. Atlas of nonribosomal peptide and
463 polyketide biosynthetic pathways reveals common occurrence of nonmodular enzymes.
464 *Proceedings of the National Academy of Sciences*. 2014 Jun 24;111(25):9259–64.
465 doi:10.1073/pnas.1401734111