EXPLORING FUNCTIONAL INFERENCE PATTERNS OF BIOSYNTHETIC GENE CLUSTERS VIA THEIR REGULATORY NETWORK

Master thesis Bioinformatics - Daan van Nassauw

Supervisors:

Hannah Augustijn
Zach Reitz
Marnix Medema

ABSTRACT

In the search for novel antibiotic compounds, the exploration of predicted Biosynthetic Gene Clusters (BGCs) with genome mining has been a widely used and successful strategy. The family Streptomycetaceae, known to produce most of the current clinically used antibiotics, has shown potential for even more novel secondary metabolites. However, the functional annotation of predicted BGCs with precise functions remains difficult, as validation through laboratory experiments is often necessary. However, due to the frequently cryptic nature of the BGCs, where genes remain inactive or produce no detectable products, prioritizing novel BGCs with antimicrobial properties is difficult. Addressing this challenge, this study focused on the possibility to extend our knowledge about the functions of predicted BGCs, by exploring the complex regulatory system that governs them. Regulatory protein families, particularly those recognized as pathway-specific regulators, are known to directly regulate BGCs. Therefore, we sought to assess the predictive capacities of associations between the regulator families and their BGCs.

This study investigated all 784 experimentally characterized Streptomycetaceae BGCs from the MIBiG database, on which regulatory protein homologs were detected with 1375 regulatory protein families from the Pfam database and 36 regulatory protein families from antiSMASH's smCOGs library. Unfortunately, no clear associations between regulatory protein families and functions or types of the curated BGCs were exhibited. Further exploration of protein families' locations (within or outside of a BGC) in 625 Streptomycetaceae genomes revealed that none of the regulatory families, including SARPs (Streptomyces antibiotic regulatory proteins), exclusively function as pathway-specific regulators for BGCs. Subsequently, a phylogenetic examination of SARP family members revealed clades characterized by shared BGC functions, types, or compound production. For instance, well-known antibiotics like beta-lactam and prodigiosin were associated with SARP homologs that claded together.

In conclusion, this study underscores the complexity of Streptomycetaceae BGC regulation. Even though the regulatory LuxR and SARP protein families showed to be present in BGC regions more than any other regulatory protein family, no unique associations between BGCs and regulatory families were made. Nevertheless, the observed connections among specific clades within the extensive SARP regulator family still indicate the potential to establish associations between BGC types/functions/compounds and subsets of larger regulatory protein families.

1. INTRODUCTION

Over the last decades, the extensive use of antimicrobials within the clinical and agricultural field has led to an acceleration of antimicrobial multi- or even pan drug-resistance among bacterial pathogens. The fast pace at which this happens, combined with the scary slow pace at which new medications are being developed emphasizes the necessity to discover new antimicrobial drugs [1–3]. Most of the current clinically used natural antimicrobial compounds find their origin in bacteria from the Actinomycetota phylum. The most well-known and extensively studied *Streptomyces* are the biggest contributors of all [4]. These members are part of Streptomycetaceae family, a prominent group of gram-positive bacteria that inhabit a wide variety of environments. Streptomycetes are especially well-adapted to thrive in the nutrient-rich and complex organic ground. Their filamentous growth allows them to explore the soil, where they play crucial roles in the degradation of organic matter and recycling of nutrients[5]. Additionally, their ability to produce a wide range of secondary metabolites serves as a competitive advantage and increases their survivability.

The biosynthesis of a secondary metabolite is often encoded within groups of co-localized (clustered) genes, so called biosynthetic gene clusters (BGCs). These BGCs are responsible for the biosynthesis of various bioactive compounds, including antibiotics, antifungals, and antitumor compounds. Numerous BGCs have been identified and characterized and their data is stored in ***MIBiG*** [6,7]. MIBiG (Minimum Information about a Biosynthetic Gene cluster) serves as a repository for curated and standardized information about known BGCs from different microbial sources. Currently, 2502 secondary metabolite clusters are stored in there, from which 784 are from in members in the Streptomycetaceae family. Besides the sequences and genetic locations, it also holds information about the predicted compound and properties of the natural product. All the compounds are classified into the currently available categories, based on their conserved enzymes or pathway types; Polyketides (PK), Ribosomally synthesized and post-translationally modified peptides (RiPPs), Saccharides, Alkaloids, Non-ribosomal peptides, Terpenes and Other (Appendix A:C). Even though most of the BGCs have a type classification, a larger part's functional annotation is still lacking. From the 2502 BGCs, 1410 (56,35%) has no known function linked to it (Appendix A:B), which is emphasizing the need to find methods that could aid in extracting new information using the limited information (types/functions) that is known of the annotated BGCs.

Currently, the validation of the eventual secondary metabolite functions needs to be performed by wet-lab techniques. Unfortunately, this is often not possible, due to the BGC being silent or cryptic[3]. The production of secondary metabolites, thus the expression of the BGCs, is regulated by proteins, which cascades are initiated by external stimuli. However, if the required conditions are unknown, it is difficult to induce the expression in laboratory conditions[8]. Approaches, such as the One-Strain-Many-Compounds (OSMAC), address the latter challenge by submitting the natural compound producing bacteria to different conditions (e.g., co-culture with different bacteria and environmental stimuli or overexpression of regulators). Nonetheless, such techniques are cost and labour intensive and require the metabolite to be already functionally predicted or prioritized.

The prediction of novel BGCs and their core genes depends almost entirely on genome-mining tools like antiSMASH (antibiotics and secondary metabolite analysis shell)[9]. Since 2011, this powerful bioinformatics tool is, with over 1.6 million processed jobs, the most extensively used online bacterial BGC predictor and can detect 81 different BGC types [10]. A combination of rule-based algorithms and profile hidden Markov models (pHMMs) is used to identify BGC regions. Among these is a small-molecule HMM database, also known as the smCOGs (secondary metabolite – Clusters of Orthologue Genes), which represents protein family pHMMs with specific roles in secondary metabolite biosynthesis. These families were curated and classified based on their conserved domains and aid antiSMASH in the annotation of predicted BGC core genes, such as transporters and regulators. Despite huge amounts of genomic data and comprehensive genome mining tools, many predicted metabolites in these predicted BGCs in Streptomycetaceae are hard to functionally annotate. Over the past few years, there has been an increase in the development of algorithms using machine and deep learning models to aid in the prediction and characterization of BGCs. These methods mainly use the sequences of protein families which are localised in BGCs and combine these with the BGCs' currently available and curated annotations. Even though the latter methods show great potential to assist in the search for novel BGCs, they all have a limitation in common; the availability of known and experimentally validated BGC types and classes. Nevertheless, these

recent studies do show the potential to use pattern-based approaches on protein families as a method to extend the knowledge on BGCs and their properties [11–13].

This shown potential opens new methods to, ultimately, extend our knowledge on known and novel BGCs by using the protein families within them. Especially regulatory protein families are expected to aid here. Within the complex regulatory network, BGCs are often directly regulated by a combination of global and pathway-specific regulators. Global regulators oversee the expression of multiple BGCs across the genome. They act as master regulators that coordinate the expression of various secondary metabolite biosynthetic pathways in response to cellular and environmental signals[14–16]. Pathway-specific regulators, on the other hand, are associated with a particular BGC and directly control its activation or repression[16–18]. Therefore, a deeper understanding of the biological activities is expected to be gained from associations of pathway-specific regulator families with the corresponding BGCs. For example, if a BGC is associated with a pathway-specific regulator that is known to control the biosynthesis of a specific class of antibiotics, it is reasonable to hypothesize that the silent or cryptic BGC, regulated by the same family, could also be involved in the production of a related antibiotic compound. Similarly, the presence of global regulators associated with multiple BGCs in the same gene cluster might suggest that these BGCs are co-ordinately regulated and may produce complementary or similar metabolites. Even though evidential literature to support this concept is currently lacking, large regulatory families like the Streptomyces antibiotic regulatory protein (SARP) family have been shown to be directly involved in the regulation of BGCs and their products[19,20]. Not only as pathway-specific regulators, but also from a more pleiotropic position[21,22].

Here, we report on the exploration of regulatory protein families in Streptomycetaceae BGCs and the consequent hypotheses that arise from it. We assessed the possibility to make associations between regulatory protein families and (curated) BGCs in the 625 members of the Streptomycetaceae family. BGCs and complete genomes were analysed to see whether there might be regulator families, that could solely or mainly be associated with the regulation of (specific types of) BGCs. Subsequently, the association concept was assessed by a case study of the earlier mentioned SARP family.

2. MATERIALS & METHODS

During this study, different genomic perspectives were highlighted and used to create possibilities to associate regulatory families with biosynthetic functions in the biosynthetic compound producing Streptomycetaceae family. In the section below, the approaches and tools are being described. Tools were run with default parameters if none are specified. If no tool is mentioned, the task is performed with an in-house written python script (indicated by an asterisk (*), followed by the script number). An overview of these can be found in Appendix D. A visual representation of the scripts (data flow), during this thesis, is shown in supplementary data Appendix I Content of all the scripts, raw data locations (also described in Appendix C) and (intermediate) output files can be found on https://git.wur.nl/daan.vannassauw/thesis_BGC_functional_inference.

2.1. DATA COLLECTION

Multiple data sets were obtained from various data sources to be used as starting points. Firstly, 2502 GenBank and JSON files were retrieved from the MIBiG database to extract the most recent, available, and curated BGCs. (v3.1) [downloaded April 12th, 2023 (GenBank) & April 24th, 2023 (JSON)]. Coding sequences and identifiers of BGCs, with their origin in members of the *Streptomycetaceae* family, were extracted from the GenBank files and transformed to fasta files, containing a single BGC each (*1). Pseudo genes or incomplete translations were neglected to improve credibility of future hypotheses. Information (names, types, and functions) of the BGCs and their compounds were extracted from the JSON files (*2). Furthermore, 625 representative Streptomycetaceae genomes were collected from NCBI RefSeq [Downloaded April 21st, 2023 – search term: *txid2062[Organism:exp] AND (latest[filter] AND "representative genome"[filter] AND all[filter] NOT anomalous[filter]*]. These were selected to include a representative of each Streptomycetaceae member, while excluding assemblies with anomalies. Like the BGC extraction, solely available coding sequences of non-pseudo proteins were extracted (*3). For the eventual extraction of regulatory protein family profile Hidden Markov Models (pHMMs), the complete PFAM HMM library was downloaded as the most extensive and updated source

of protein families[23] [Downloaded March 14th, 2023], together with the smCOG BGC associated HMM library from antiSMASH[24] [Downloaded April 26th, 2023].

2.2. SELECTION OF REGULATORY PROTEIN FAMILIES

Protein families in the HMM libraries of antiSMASH' smCOGs and the PFAM database are involved in a wide variety of functions. For this project, solely protein families with regulatory properties were desired from both sources. The regulatory HMM subsets were created and consecutively indexed using HMMER's (v3.3.2) *hmmfetch* on the PFAM and smCOG HMM libraries by feeding it a list of regulatory-related keywords, which were expected to be found in the protein family descriptions in the Pfam and AntiSMASH's smCOG libraries (Appendix B | *4,5)[25]. There is a significant difference in key terms for both sources (see Appendix B) is explainable by the different aims of both extractions; the smCOG library only contained 36 regulatory protein families (annotated by the tool with 'R')[26], where the total number of regulatory protein families in the PFAM database is unknown. Therefore, these were extracted with the aim to capture as many as possible. Detection of unwantedly captured smCOG HMMs was done manually by comparing it to antiSMASH' annotations[26]. These were removed to solely capture regulatory protein family homologs. From now on, the subsets with regulatory proteins will be referred to as Pfam-R and smCOG-R.

2.3. REGULATORY PROTEIN DOMAIN DETECTION APPROACH

The presence of regulatory protein families was detected using HMMER's *hmmscan,* which takes a BGC or genome fasta file and Pfam-R or smCOG-R as input and returns all captured homologs of protein family pHMMs (*6,7). Trustworthiness of the domain homologs was increased by setting cut-offs. Pfam-R homology hits followed the in-house trusted cut-off from the Pfam database (often bit scores > 22), where the smCOG-R homologs with E-values < 1E-16 were neglected. The latter was done as it is in line with antiSMASH' internal annotation process[27]. For both the BGCs and complete genomes, all the locations, protein family IDs and bit scores of the homology hits were extracted and summarized into one overview (*8).

2.4. REGULATORY PROTEIN FAMILIES IN CURATED STREPTOMYCETACEAE BGCs

The overview of Pfam-R and smCOG-R matches occurring in BGCs from MIBiG was combined with the compound and type information of BGCs to create one large dataset, which is ideal for pattern recognition (*9). Visualization of this annotated list of homologs followed a network-structured approach using CytoScape (v3.10.0) [28]. SmCOG-R and Pfam-R IDs were set as source nodes, while the BGC IDs functioned as target nodes. Pattern detection and prioritizing further interesting events was done manually. Visualization of the nine most frequently occurring functions was done of families that had at least ten homologs in the entire MIBiG BGC dataset. This was performed using Python's Matplotlib (v3.7.2)[29] (*17).

2.5. REGULATORY PROTEIN FAMILIES IN STREPTOMYCETACEAE GENOMES

BGC regions in the genomes were predicted by feeding the GenBank files to antiSMASH (v7.0) [10] and tabulated via Z. Reitz's existing workflow[30]. The locations of the Pfam-R and smCOG-R homologs were then compared to the predicted BGC region locations to determine whether the homolog was found in- or outside a predicted BGC region (*10). The same custom script also calculated the shortest distance from the homolog to the nearest predicted BGC edge. All information on the homolog locations, bit scores, predicted BGC locations, predicted products and distances to the nearest edges were combined into one single data frame (*11). Homologs occurring in draft genomes, indicated by a genome ID > 11 characters, were separated from complete genome homologs to improve the hit credibility. Simultaneously, homologs with close distances to the BGC edge (< 500 nucleotides) or with unknown values ("NA") were filtered out (*12). The fraction of in BGC-laying homologs was then calculated for the remaining homologs in complete genomes to identify possible BGC-associated protein families (*13).

2.6. SARP FAMILY OF REGULATORS - CASE STUDY

Matches of the SARP family in the complete, filtered genomes of the Streptomycetaceae family were located and its protein sequences were extracted (*14). Due to the large number of proteins, a dereplication approach was initiated. The collected sequences underwent a greedy incremental clustering with a 98% target coverage using MMseqs2 (v14.7.e284)[31]. It sorts the protein sequences from largest to smallest, aligns all sequences that

are covered for at least 98 percent & continues once there are none left. The representatives (from here on referred to as SARP-reps) of each cluster and their protein sequences were collected and visualized using CytoScape (v3.10.0)[28]. Subsequently, the SARP-reps' sequences were aligned against the original SARP profile HMM (SMCOG1041), which was caught from the smCOGs' HMM library using HMMER's, earlier mentioned, *hmmfetch*[25]. The alignment, executed by HMMER's *hmmalign,* simultaneously trimmed terminal tails of unaligned amino acids. After that, a transformation of the SARP-reps IDs took place (*15), while all insertions relative to the SARP HMM were removed from the alignment to compare the SARP homologs directly (*16). The resulted alignment of SARP-reps was used as input in IQ-TREE (v2.2.2.3)[32] for tree building by maximum likelihood (ML), which was performed with the ModelTest option and Ultrafast Bootstrap approximation to increase procedure speed[33]. Eventually, the tree was visualized and annotated with in/out BGC locations, BGC types and edge distances to the nearest BGC using the Interactive Tree of Life (iTOL) (v6.8)[34]. Pattern recognition and literature research was done manually.

3. RESULTS & DISCUSSION

3.1. EXTRACTION OF REGULATORY PROTEIN FAMILIES

Maximizing the amount of detected regulatory protein families started with the collection of them from the large PFAM and smCOG HMM libraries. To achieve this, regulatory related key terms that were expected to be present in regulatory family descriptions were created and used for the library subset creations, named Pfam-R and smCOG-R. The extraction of regulatory protein families led to library sizes of 1375 families in Pfam-R and 39 in smCOG-R. Since we were only interested in regulatory protein families, non-regulatory captured smCOGs ('SMCOG1132', 'SMCOG1210' and 'SMCOG1174') were manually removed from the list. These were unwantedly captured by HMMER's *hmmfetch* as a result. The Pfam-R could also contain non-regulatory or non-prokaryotic protein families; however, manual filtering 1375 families would take too much time and most of them were not expected to have a lot of hits anyways.

3.2. REGULATORY FAMILIES IN STREPTOMYCETACEAE BCGs

To find the regulatory protein families that are present in the current curated BGCs, 784 experimentally characterized BGC clusters with their origin in *Streptomycetaceae* were collected from MIBiG. Presence detection of these families was facilitated by the Pfam-R and smCOG-R pHMM subsets. Subsequently, the regulatory protein family homologs were assessed for possible links between families and BGC functions. Figure 1 shows the functional landscape of the BGCs in which the protein families occurred in.

As previously stated, more than 56% of both the currently identified and curated BGCs lacks definitive functional annotations. Ideally, BGCs with unknown functions could be annotated by the function of other BGCs, that share the same regulatory family. Almost all the regulatory families occur in one or more of the 784 BGCs that produce compounds annotated to have antibacterial, antifungal, cytotoxicity and inhibitory properties (often all of them).

The exception to occurring in multi-functional annotated BGCs was the Carboxypeptidase regulatory-like domain (CRL-D) (PF13620), which was found in BGCs with a single annotated function besides the non-annotated (NA) ones. Nine of the eleven BGCs harbouring this domain remain functionally uncharacterized, while the remaining two were annotated with an antibacterial function. Nevertheless, it is more likely that the CRL domain solely plays a role in the modification of EmrB/QacA drug resistance transporter proteins (smCOG1005), rather than fulfilling a regulatory position[35–42]. Therefore, following-up on the NA BGCs in which this CRL domain occurs to assess the potential association of the domain with the antibacterial function is unnecessary. At the same time, no consistent trends have been observed across all other protein families in relation to the functional annotations of available BGCs.

It was also noted that the Pfam-R library captured homologs in experimentally characterized BGCs, that were not captured by the smCOG-R library (see Appendix H). Among these were mainly helix-turn-helix domains that, after a literature search, have shown to have regulatory properties.

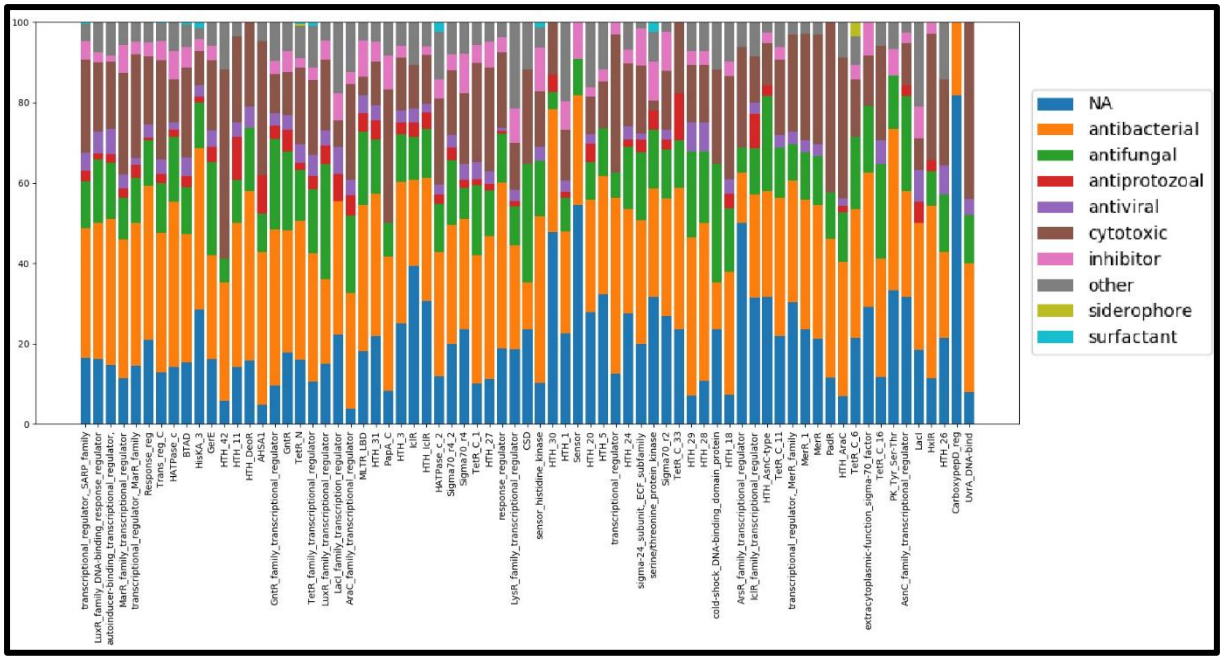


FIGURE 1 FRACTION OF BGC FUNCTIONS PER REGULATORY PROTEIN FAMILY (BOTH PFAM-R AND SMCOG-R) BASED ON THE 4444 DETECTED HOMOLOGS IN CURATED STREPTOMYCETACEAE BGCS. PROTEIN FAMILIES WITH TEN OR LESS DETECTED HOMOLOGS WERE EXCLUDED.

FRACTION OF REGULATORY FAMILIES WITHIN PREDICTED BGC AREAS

The identification of pathway-specific regulatory families that would exclusively be associated with the regulation of BGCs would lead to a significant increase in annotated BGCs. This detection method involved a comparison of occurrences of each regulatory protein family within the predicted BGC regions of antiSMASH against the total occurrences in Streptomycetaceae genomes. These values together yield the fraction of occurrences inside BGCs for each family.

Prior to this extraction, the detected homologs underwent filtering steps to improve the credibility of eventual associations and avoid missing broader patterns. Mainly, as the BGC regions, predicted by antiSMASH, consist of a core gene section and hard-coded extensions (between 5 and 20 kb) on both sizes. Therefore, only homologs that were found in complete genomes, without non-annotated distances to the nearest predicted BGC and distances larger than 500 bases were considered (see table 3).

TABLE 1 THE REMAINING NUMBER OF PROTEIN FAMILY MEMBERS PER TYPE OF STREPTOMYCETACEAE GENOME DATA TYPE BEFORE & AFTER APPLYING CONSECUTIVE FILTER CONDITIONS. THE NUMBER OF REMOVED HOMOLOGS ARE SHOWN BETWEEN BRACKETS

Filtering condition	All genomes	Complete genomes	Draft genomes
Unfiltered homologs	1,069,877	287,118	782,757
Homologs without an edge distance value (NA) excluded	723,698 (-346,179)	269,161 (-17,958)	454,538 (-328,219)
500 base pair distance to nearest BGC edge	710,161 (-13,539)	265,118 (-4,042)	445,042 (-9,496)
Remaining homologs	710,161 (-359,716; 33,63%)	265,118 (- 22,000; 7,66%)	445,042 (-337,715; 43,14%)

Starting with almost 1.07 million protein family homologs in the entire dataset turned into 265 thousand protein family homologs in the complete Streptomycetaceae genomes to continue with. The largest losses of the filtering steps are seen among the draft genomes (-43,1%; 337,715), where the complete genomes only lose a fraction of it (-7,2%; 22,000). The largest impact was delivered by the filtering on missing values, indicating the edge distances were not able to be calculated. Either no BGC regions were predicted in those genomes or, the data was too fragmented or lacked information, which is more likely to occur in draft genomes.

Figure 2A shows the total occurrences and the fraction of IN BGC locations for each protein family. Families containing high "IN BGC fraction" values often possess a relatively low overall count (<20), making them less reliable indicators of a family-BGC association. In contrary, protein families displaying a high number of matches still tend to fluctuate, but between a more constrained range (5% – 15%). The patterns demonstrate a 'baseline', which could be explained by the fact that 75% of the used Streptomyces genomes have a BGC content of 15% or less (see figure 2C). Therefore, IN BGC ratios around this baseline could be considered as random occurrences. Exceptions here are the families that have a significant number of occurrences and raise clearly above the 'baseline' (see figure 2B); the SARP family, LuxR family and the bacterial transcriptional activation domain (BTAD) family.

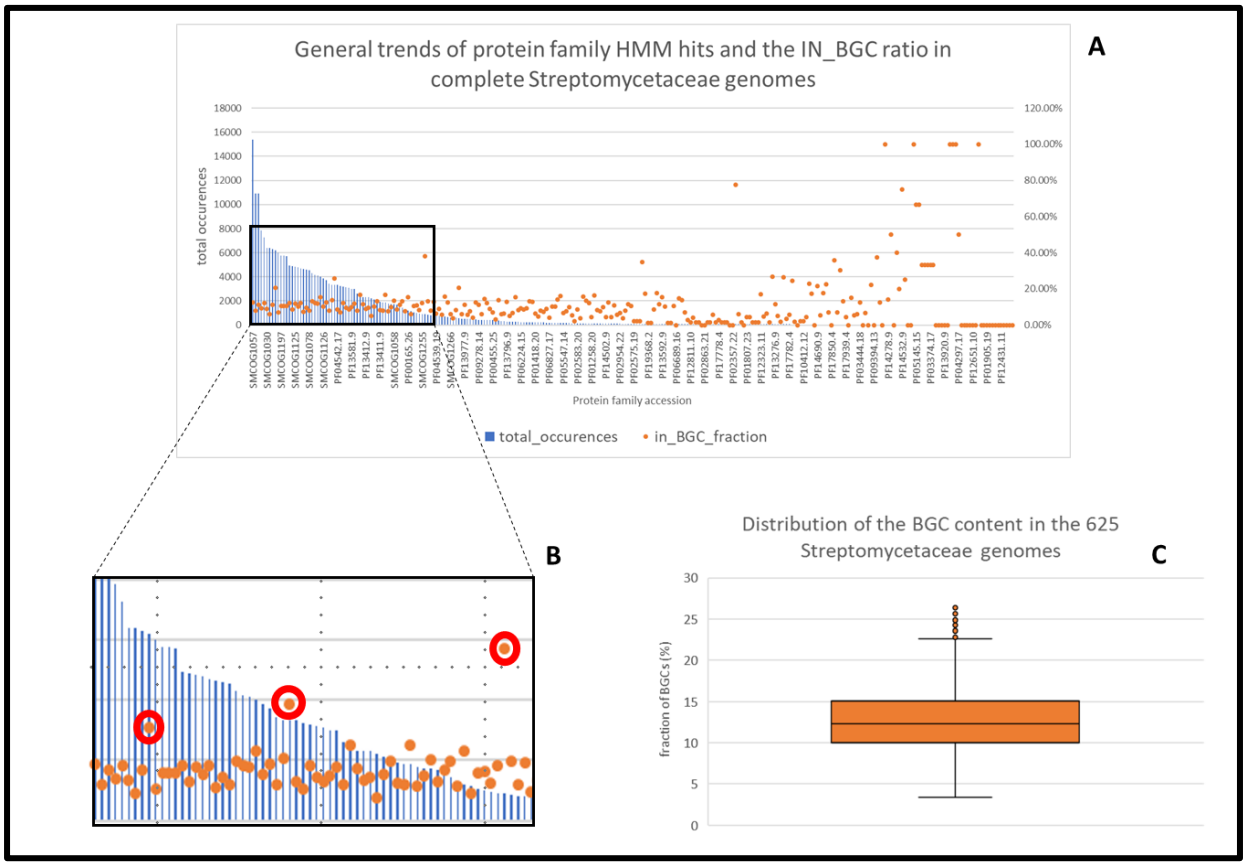


FIGURE 2 A) THE TOTAL AMOUNT OF OCCURRENCES PER PROTEIN FAMILY AND THE FRACTION THAT IS DETECTED INSIDE ANTISMASH'S PREDICTED BGC REGIONS. THE FAMILIES ARE SORTED FROM HIGHEST TO LOWEST NUMBER OF TOTAL HOMOLOGS IN STREPTOMYCETACEAE. **B)** REGULATORY PROTEIN FAMILIES WITH >900 TOTAL OCCURRENCES. RED CIRCLE 1 = SARP FAMILY (6191 HOMOLOGS - 20.7% IN BGC RATIO), RED CIRCLE 2 = LUXR FAMILY (3333 HOMOLOGS – 25.7% IN BGC RATIO), RED CIRCLE 3 = BTAD (906 HOMOLOGS – 38.0 % IN BGC RATIO), **C)** DISTRIBUTION OF BGC CONTENT FRACTIONS ACROSS ALL 625 STREPTOMYCETACEAE SPECIES (ALL VALUES CAN BE FOUND IN APPENDIX J).

In the dataset, the members of the LuxR regulatory family appeared 3333 times, with an IN BGC ratio of 25.7%. This observed elevated IN BGC ratio is in line with expectations, given the major role of the LuxR family in quorum sensing[43,44]. Furthermore, research onto the LuxR family in gram-positive bacteria has revealed an evolutionary history that contributed to a large diversity among its members. It led to LuxR regulators functioning within one-

or two-component signalling systems, with numbers up to 69,000 solo LuxR instances across 800 gram-positive bacterial genomes [45,46]. This indicates that finding high occurrences for this family is not entirely unexpected.

The BTAD is a domain located after the N-terminal DNA-binding domain of in Mycobacterium's EmbR regulator and in SARP family members[47–49]. This means that this domain is not entirely exclusive to the SARP family[50]. However, given the inclusion of solely Streptomycetaceae genomes, we expect that these 906 found homologs are found in members from the SARP family. Unfortunately, this has not been assessed in this study. Moreover, we still noticed a difference of 5285 detected homologs, that were detected with AntiSMASH's smCOG1041, but not with Pfam-R's BTAD. Differences between the HMM profile sizes (811AA for SARP vs. 146AA for BTAD) could aid as it makes the smCOG1041 more inclusive, capturing a broader range sequences. To better understand these differences, further analysis and comparison of the two HMMs and the detected homologs would be valuable by direct profile-profile comparison tools as HHsuite[51].

Furthermore, given the association of the SARP family with the regulation of BGCs and its usual localization inside those regions, raises questions why we observed such a low IN BGC ratio for this family. Multiple factors might contribute to this latter finding. Firstly, antiSMASH might not yet be able to identify a broader spectrum of novel BGC types, if SARP regulators exclusively regulate BGCs. However, given the extensive presence of SARP homologs in Streptomycetaceae, this scenario appears improbable. A second hypothesis was based on a study by *Krause et al*[52]., who performed a similar approach to detect SARPs using SMCOG1041 in Actinobacteria, revealing hits in Proteobacteria even without the BTAD or HTH motif [52]. This suggests the possibility that SMCOG1041's detection scope might be marginally wider than intended. Lastly, the most plausible explanation is based on the entire scientific perspective on SARP members. While their characterization as pathway specific BGC regulators is widely accepted, it is also known that some members, like AfsR, regulate multiple pathways from a more global perspective[22]. Therefore, it is not unlikely to think that there might be more SARP family members regulating from a more pleiotropic perspective[53].

VARIATION AMONG THE SARP FAMILY

To explore the diversity within the SARP family members and find the underlying reasons for the observed low IN BGC ratio, all 3093 homologs in the Streptomycetaceae genomes were collected. They were subjected to clustering and representatives for each cluster (SARP-reps) were subsequently aligned against the SARP HMM (smCOG1041). Alignment showed SARPs with minimal alignment lengths of 620 AA to the 811 AA long SARP HMM. A phylogenetic tree was reconstructed of the alignment and annotated with the IN_BGC Boolean, distances to the nearest BGC edge and compound annotations. The resultant tree contains 601 SARP-reps and 42 distinct BGC types, as shown in Figure 4.

Across the circular tree, elevated areas of SARP members found in predicted BGC regions are seen and indicated by the red circle ("SARP hit location"). These segments exhibit homologs that are evidently linked to BGCs. However, these segments frequently contain SARP-reps that are not localized in predicted BGC regions. When considering the distances to the nearest predicted BGC boundaries, some of these homologs display distances exceeding 150,000 nucleotides. Such distances could potentially signal the presence of novel, undetected BGCs. Moreover, during the clustering process of all SARPs, instances emerged where an entire cluster was situated within a BGC region, except for the cluster's representative member (Appendix E – red circles). Notably, these SARP-reps were primarily observed in the segments with BGC-associated SARP homologs (figure 4 – red labels).

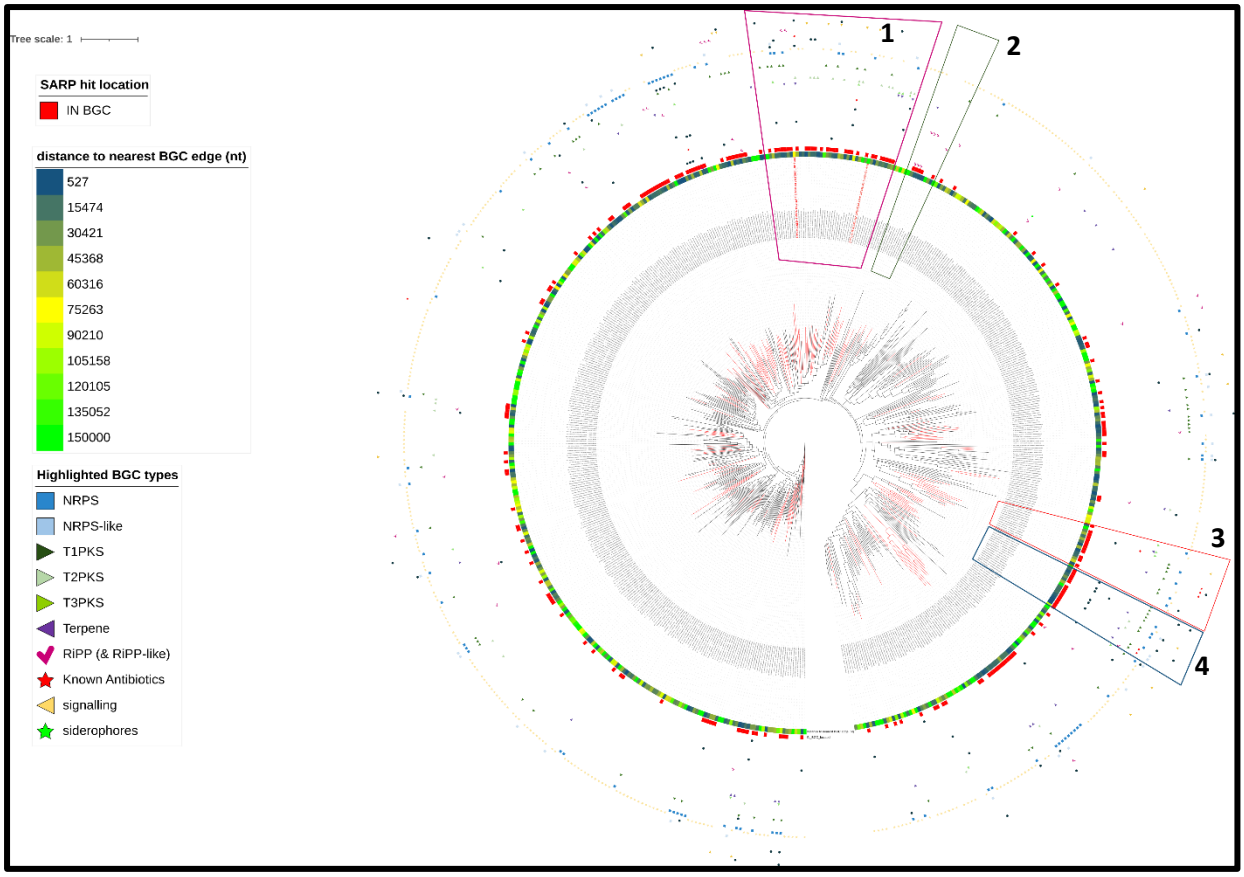


FIGURE 3 601 REPRESENTATIVE SARP HOMOLOGS WITH ANNOTATED LOCATION (IN OR OUT BGC), DISTANCE TO NEAREST BGC EDGE AND EVENTUAL BGC PROPERTIES. THE CIRCULAR TREE, BASED ON MAXIMUM LIKELIHOOD, WAS MIDPOINT-ROOTED AND BUILT WITH MODELTEST AND ULTRAFAST BOOTSTRAP APPROXIMATION. SUPPLEMENTARY VISUALIZATIONS OF COLOURED BOXES 1, 2, 3 AND 4 ARE AVAILABLE IN APPENDIX F.

Manual checking led to some noteworthy patterns (figure 4 - coloured boxes). Firstly, the purple area (#1) shows SARP homologs in a clade with visually elevated levels (30/44) of Type-II polyketide synthase (T2PKS) BGCs of which two are known antibiotics (2dos[54,55], phenazine[56,57]). Often, these T2PKS type BGCs are double annotated with either the T1PKS or T3PKS in the MIBiG database. In general, polyketide types cover a long list of clinically used antimicrobial compounds as tetracycline, anthracycline, amphotericin and avermectin[58–61]. Therefore, there might be potential to associate that clade of SARP-reps with T2PKS type BGCs, which could lead to the discovery of novel BGCs with compounds that have antimicrobial properties. Secondly, the dark-green area (#2) contains a clade with visually elevated levels (5/9) of SARPs in ribosomally synthesized and post-translationally modified peptide classes (RiPPs). This type embodies an enormous family of small molecule natural products with diverse functions, which caused the huge interest in assessing their potential in antibiotic compound discovery[62,63]. Lastly, the red box (#3) and blue box (#4) contain visually elevated (12/15 & 14/15) SARP-reps that are found in BGCs of the T1PKS type. Among the same segments, multiple known antibiotic compound producing BGCs are present. Multiple SARP-reps, that are found in BGCs for beta-lactam were clustered immediately next to each other. The same was shown for SARP-reps that are in prodigiosin producing BGCs. This indicates that there could be an association between those clades of SARP-reps and antibiotic compound synthesizing BGCs. Obviously, statistical substantiation is necessary to prove these predictions as there is phylogenetic non-independence among the Streptomycetaceae and their BGCs. This means that their characteristics are not statistically independent, due to their evolutionary history. If these are ignored, conclusions could be incorrect or overestimate the relationships between traits like a SARP homolog and the traits of the BGC they are found in. Phylogenetic regression models could aid in this situation, an example is the Caper package[64,65]. Unfortunately, this was outside of this project's scope due to the lack of time.

During this study, it was attempted to demonstrate associations between regulatory protein families and BGC properties (types, functions, etc.) in Streptomycetaceae with the idea to, ultimately, use these for functional predictions of yet unannotated BGCs. To the extent of our knowledge, such a concept has not been attempted and proofed before. An important necessity that allows making associations in the first place, is a high completeness of the dataset you are working with. Unfortunately, with over 56 percent of available, curated BGCs from MIBiG being unannotated, this has not come to our advance. Luckly, detection of regulatory protein families was not affected by the lack of annotative information, as you solely require the presence of complete genomes and the pHMMs of protein families. Both sources of the pHMM libraries (Pfam and antiSMASH) aided in the detection of protein families and followed the same principle, however, they displayed some fundamental differences along the way. The Pfam database is suggested to contain more specific pHMMs compared to the library of smCOGs. Not only does the Pfam database contain more pHMMs per regulatory protein family (example TetR: 41 vs. 4), but they also showed lower numbers of detected homologs per family in the complete Streptomycetaceae genomes (examples in Appendix G). Similarly, the Pfam database lacks a pHMM that covers the entire SARP as a protein family, but it does have pHMMs for SARP characterizing domains; BTAD, N-terminal winged HTH DNA-binding domain (Trans_reg_C) and TPR_12 (for larger SARPs)[52,66]. The latter was not included in this study. With the detected differences in specificity between the databases in mind, we would like to suggest an update of the smCOG library of AntiSMASH. Not only by adding novel regulatory domains, that capture regulators in experimentally validated BGCs (examples in Appendix H), but also as more specific pHMMs (e.g., of family subdivisions) might lead to more specific associations between regulator families and BGC annotations.

To gain more perspective in the locations of the regulatory homologs, the shortest distances from the homologs to the nearest predicted BGC edge was calculated and used as a label in the manual pattern finding in the SARP case. Initially, it was intended to predict potential novel BGC types, that could not be detected by antiSMASH yet (SARP homologs with huge differences to the nearest predicted BGC). Simultaneously, it could help manually assessing cases, where the regulator would be very close to the predicted BGC edge. Especially, knowing that the edges of the predicted BGC types are hardcoded by antiSMASH as extensions between 5 kb and 20 kb. Therefore, we would argue that an alternative approach i.e., calculating the distance to the nearest core genes of predicted BGCs would be more effective in the concept of making meaningful associations. Only not to establish associations between regulator families and complete BGCs, but to assess the possibility to predict associations between regulators and certain core genes.

No direct associations between a single regulatory protein family and one of the curated BGC types or known functions were displayed, which was not entirely unexpected. The BTAD, LuxR and SARP families of regulators showed a higher occupation in predicted BGC areas than any other regulatory protein family. This study displayed 1280 SARP homologs and 857 LuxR homologs within (predicted) BGC regions among the complete Streptomycetaceae genomes. SARPs have been described before to have multiple occurrences within a single BGC, which means that the 1280 homologs are probably not directly present in 1280 BGCs (detected and undetected)[67–69]. Estimation of the SARP distribution within this study has not been evaluated but could give additional insight into the complex regulatory network. Similarly, indications that subdivisions of the SARP family could be enriched in clades of certain BGC traits based on phylogenetic analysis were there, but they require statistical comparative analysis (phylogenetic regression models) to make the predictions meaningful. Nevertheless, phylogenetic analysis of large regulator families, involved in BGC regulation, is suggested to be helpful in prioritizing interesting cases of pathway-specific regulators in clades with specific BGC traits.

4. Conclusion

While the library of novel predicted BGCs in Streptomycetaceae keeps expanding, the need for functionally inferring them those does too. While the eventual functions of novel metabolites need to be validated experimentally, computational predictions could aid in prioritizing and save time and money that way. This study assessed the possibility to extend our knowledge on BGC functions by attempting to predict associations of enriched regulatory protein families in known with specific BGC types and functions. The initial linkage on protein regulator families to known BGCs and their types/functions did not show any cases with a full association

between a regulatory protein family and a BGC property (function or type). The locational assessment only showed the two large regulator families, LuxR and SARP, to have affinity with BGCs. It did become clear that, within the SARP family, there is potential to find associations between enriched sub-groups of the regulator family in specific BGCs with specific properties. Nevertheless, this requires further investigation and statistical substantiation. A follow-up on SARP homologs that are in the earlier mentioned boxes with large distances (>150k nucleotides) to the nearest BGC would be necessary to proof the concept of associating regulatory members to functionally infer unannotated BGCs. The functional diversity of the analysed SARP family does highlight the complex regulatory networks that the Streptomycetaceae uses to adapt to their environment and produce a wide range of bioactive compounds.

5. Recommendations

The following section presents key recommendations derived from the findings and analyses conducted in this study. Firstly, the assessment of SARP homologs that were found in clades with enriched BGC-associated SARPs and showing high distances to the nearest BGCs. Such cases could be an indication of novel BGC types in that area. One option would be the use of different BGC prediction tools, that use machine- or deep learning principles (e.g., GECCO or SanntiS) to evaluate the potential of those regions with, by antiSMASH, (yet) undetected BGCs[70,71]. Secondly, as there are large quantitative detection differences between the smCOG-R and Pfam-R libraries (examples in Appendix G), a direct comparison of Pfam-R and smCOG-R HMMs through tools like *HHsuite* might be worth exploring. This could aid in creating subdivisions among regulatory families and creation of the respective pHMMs. The main example being here the SARP smCOG and the SARP characterizing domains from the Pfam database. Lastly, it would be worth to propose an update of the smCOG library, as it showed not capturing all regulatory domains in experimentally validated BGCs (see Appendix H). Besides the addition of novel regulatory protein families, the update could focus on the creation of more specific pHMMs of already available families. This fine-tuning could potentially lead to more targeted connections in the future between regulator families and BGC annotations.

6. REFERENCES

1. Cooper, M. A. & Shlaes, D. Fix the antibiotics pipeline. *Nature 2011 472:7341* **472**, 32–32 (2011).
2. Chang, Q., Wang, W., Regev-Yochay, G., Lipsitch, M. & Hanage, W. P. Antibiotics in agriculture and the risk to human health: how worried should we be? *Evol Appl* **8**, 240–247 (2015).
3. Miethke, M. *et al.* Towards the sustainable discovery and development of new antibiotics. *Nature Reviews Chemistry 2021 5:10* **5**, 726–749 (2021).
4. Donald, L. *et al.* Streptomyces: Still the Biggest Producer of New Natural Secondary Metabolites, a Current Perspective. *Microbiol Res (Pavia)* **13**, 418–465 (2022).
5. Anandan, R. *et al.* An Introduction to Actinobacteria. *Actinobacteria - Basics and Biotechnological Applications* (2016) doi:10.5772/62329.
6. Medema, M. H. *et al.* Minimum Information about a Biosynthetic Gene cluster. *Nat Chem Biol* **11**, 625–631 (2015).
7. Terlouw, B. R. *et al.* MIBiG 3.0: a community-driven effort to annotate experimentally validated biosynthetic gene clusters. *Nucleic Acids Res* **51**, D603–D610 (2023).
8. Hoskisson, P. A. & Seipke, R. F. Cryptic or Silent? The Known Unknowns, Unknown Knowns, and Unknown Unknowns of Secondary Metabolism. *mBio* **11**, 1–5 (2020).
9. Medema, M. H. *et al.* antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Res* **39**, (2011).

10. Blin, K. *et al.* antiSMASH 7.0: new and improved predictions for detection, regulation, chemical structures and visualisation. *Nucleic Acids Res* **51**, W46–W50 (2023).

11. Walker, A. S. & Clardy, J. A Machine Learning Bioinformatics Method to Predict Biological Activity from Biosynthetic Gene Clusters. *J Chem Inf Model* **61**, 2560–2571 (2021).

12. Rios-Martinez, C., Bhattacharya, N., Amini, A. P., Crawford, L. & Yang, K. K. Deep self-supervised learning for biosynthetic gene cluster detection and product classification. *PLoS Comput Biol* **19**, e1011162 (2023).

13. Liu, M., Li, Y. & Li, H. Deep Learning to Predict the Biosynthetic Gene Clusters in Bacterial Genomes. *J Mol Biol* **434**, 167597 (2022).

14. Wei, J., He, L. & Niu, G. Regulation of antibiotic biosynthesis in actinomycetes: Perspectives and challenges. *Synth Syst Biotechnol* **3**, 229–235 (2018).

15. Novakova, R. *et al.* A New Family of Transcriptional Regulators Activating Biosynthetic Gene Clusters for Secondary Metabolites. *Int J Mol Sci* **23**, 2455 (2022).

16. Van Der Heul, H. U., Bilyk, B. L., McDowall, K. J., Seipke, R. F. & Van Wezel, G. P. Regulation of antibiotic production in Actinobacteria: New perspectives from the post-genomic era. *Nat Prod Rep* **35**, 575–604 (2018).

17. Kormanec, J., Novakova, R., Mingyar, E. & Feckova, L. Intriguing properties of the angucycline antibiotic auricin and complex regulation of its biosynthesis. *Appl Microbiol Biotechnol* **98**, 45–60 (2014).

18. Bibb, M. J. Regulation of secondary metabolism in streptomycetes. *Curr Opin Microbiol* **8**, 208–215 (2005).

19. Ramos, J. L. *et al.* The TetR family of transcriptional repressors. *Microbiol Mol Biol Rev* **69**, 326–356 (2005).

20. Chen, J. & Xie, J. Role and regulation of bacterial LuxR-like regulators. *J Cell Biochem* **112**, 2694–2702 (2011).

21. Tanaka, A., Takano, Y., Ohnishi, Y. & Horinouchi, S. AfsR Recruits RNA Polymerase to the afsS Promoter: A Model for Transcriptional Activation by SARPs. *J Mol Biol* **369**, 322–333 (2007).

22. Floriano, B. & Bibb, M. afsR is a pleiotropic but conditionally required regulatory gene for antibiotic production in Streptomyces coelicolor A3(2). *Mol Microbiol* **21**, 385–396 (1996).

23. Blum, M. *et al.* The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res* **49**, D344–D354 (2021).

24. antismash/antismash/detection/genefunctions/data/smcogs.hmm at master · antismash/antismash · GitHub. https://github.com/antismash/antismash/blob/master/antismash/detection/genefunctions/data/smcogs.hmm.

25. HMMER. http://hmmer.org/.

26. antismash/antismash/detection/genefunctions/data/cog_annotations.txt at master · antismash/antismash · GitHub. https://github.com/antismash/antismash/blob/master/antismash/detection/genefunctions/data/cog_annotations.txt.

27. antismash/antismash/detection/genefunctions/smcogs.py at master · antismash/antismash · GitHub. https://github.com/antismash/antismash/blob/master/antismash/detection/genefunctions/smcogs.py.

28. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498–2504 (2003).

29. Hunter, J. D. Matplotlib: A 2D graphics environment. *Comput Sci Eng* **9**, 90–95 (2007).

30. GitHub - zreitz/multismash. https://github.com/zreitz/multismash/.

31. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology 2017 35:11* **35**, 1026–1028 (2017).

32. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol Biol Evol* **37**, 1530–1534 (2020).

33. Hoang, D. T., Chernomor, O., Von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol Biol Evol* **35**, 518–522 (2018).

34. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res* **49**, W293–W296 (2021).

35. BGC0000705. https://mibig.secondarymetabolites.org/repository/BGC0000705/index.html#r1c1.

36. BGC0000704. https://mibig.secondarymetabolites.org/repository/BGC0000704/index.html#r1c1.

37. BGC0000703. https://mibig.secondarymetabolites.org/repository/BGC0000703/index.html#r1c1.

38. BGC0001003. https://mibig.secondarymetabolites.org/repository/BGC0001003/index.html#r1c1.

39. Yanai, K., Murakami, T. & Bibb, M. Amplification of the entire kanamycin biosynthetic gene cluster during empirical strain improvement of Streptomyces kanamyceticus. *Proc Natl Acad Sci U S A* **103**, 9661–9666 (2006).

40. Piepersberg, W., Aboshanab, K. M., Schmidt-Beißner, H. & Wehmeier, U. F. The Biochemistry and Genetics of Aminoglycoside Producers. *Aminoglycoside Antibiotics: From Chemical Biology to Drug Discovery* 15–118 (2007) doi:10.1002/9780470149676.CH2.

41. Kharel, M. K. *et al.* A gene cluster for biosynthesis of kanamycin from Streptomyces kanamyceticus: Comparison with gentamicin biosynthetic gene cluster. *Arch Biochem Biophys* **429**, 204–214 (2004).

42. Bihlmaier, C. *et al.* Biosynthetic gene cluster for the polyenoyltetramic acid α-lipomycin. *Antimicrob Agents Chemother* **50**, 2113–2121 (2006).

43. Fuqua, W. C., Winans, S. C. & Greenberg, E. P. Quorum sensing in bacteria: the LuxR-LuxI family of cell density-responsive transcriptional regulators. *J Bacteriol* **176**, 269–275 (1994).

44. Brotherton, C. A., Medema, M. H. & Greenberg, E. P. luxR Homolog-Linked Biosynthetic Gene Clusters in Proteobacteria. *mSystems* **3**, (2018).

45. Rajput, A. & Kumar, M. In silico analyses of conservational, functional and phylogenetic distribution of the LuxI and LuxR homologs in Gram-positive bacteria. *Scientific Reports 2017 7:1* **7**, 1–13 (2017).

46. Santos, C. L., Correia-Neves, M., Moradas-Ferreira, P. & Mendes, M. V. A Walk into the LuxR Regulators of Actinobacteria: Phylogenomic Distribution and Functional Diversity. *PLoS One* **7**, 46758 (2012).

47. Rehakova, A., Novakova, R., Feckova, L., Mingyar, E. & Kormanec, J. A gene determining a new member of the SARP family contributes to transcription of genes for the synthesis of the angucycline polyketide auricin in Streptomyces aureofaciens CCM 3239. *FEMS Microbiol Lett* **346**, 45–55 (2013).

48. Novakova, R., Rehakova, A., Kutas, P., Feckova, L. & Kormanec, J. The role of two SARP family transcriptional regulators in regulation of the auricin gene cluster in Streptomyces aureofaciens CCM 3239. *Microbiology (Reading)* **157**, 1629–1639 (2011).

49. Kurniawan, Y. N., Kitani, S., Maeda, A. & Nihira, T. Differential contributions of two SARP family regulatory genes to indigoidine biosynthesis in Streptomyces lavendulae FRI-5. *Appl Microbiol Biotechnol* **98**, 9713–9721 (2014).

50. Alderwick, L. J. *et al.* Molecular structure of EmbR, a response element of Ser/Thr kinase signaling in Mycobacterium tuberculosis. *Proc Natl Acad Sci U S A* **103**, 2558–2563 (2006).

51. Steinegger, M. *et al.* HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics* **20**, 1–15 (2019).

52. Krause, J., Handayani, I., Blin, K., Kulik, A. & Mast, Y. Disclosing the Potential of the SARP-Type Regulator PapR2 for the Activation of Antibiotic Gene Clusters in Streptomycetes. *Front Microbiol* **11**, (2020).

53. Huang, J. *et al.* Cross-regulation among disparate antibiotic biosynthetic pathways of Streptomyces coelicolor. *Mol Microbiol* **58**, 1276–1287 (2005).

54. Kudo, F. & Eguchi, T. Biosynthetic genes for aminoglycoside antibiotics. *J Antibiot (Tokyo)* **62**, 471–481 (2009).

55. Llewellyn, N. M. & Spencer, J. B. Biosynthesis of 2-deoxystreptamine-containing aminoglycoside antibiotics. *Nat Prod Rep* **23**, 864–874 (2006).

56. Wang, Y., Kern, S. E. & Newman, D. K. Endogenous phenazine antibiotics promote anaerobic survival of Pseudomonas aeruginosa via extracellular electron transfer. *J Bacteriol* **192**, 365–369 (2010).

57. Kudo, F. & Eguchi, T. Biosynthetic genes for aminoglycoside antibiotics. *The Journal of Antibiotics 2009 62:9* **62**, 471–481 (2009).

58. Barajas, J. F., Blake-Hedges, J. M., Bailey, C. B., Curran, S. & Keasling, J. D. Engineered polyketides: synergy between protein and host level engineering. *Synth Syst Biotechnol* **2**, 147–166 (2017).

59. Klaus, M. & Grininger, M. Engineering strategies for rational polyketide synthase design. *Nat Prod Rep* **35**, 1070–1081 (2018).

60. Katz, L. & Baltz, R. H. Natural product discovery: past, present, and future. *J Ind Microbiol Biotechnol* **43**, 155–176 (2016).

61. Fang, L., Guell, M., Church, G. M. & Pfeifer, B. A. Heterologous erythromycin production across strain and plasmid construction. *Biotechnol Prog* **34**, 271–276 (2018).

62. Arnison, P. G. *et al.* Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature. *Nat Prod Rep* **30**, 108–160 (2013).

63. Montalbán-López, M. *et al.* New developments in RiPP discovery, enzymology and engineering. *Nat Prod Rep* **38**, 130–239 (2021).

64. Orme, D. The caper package: comparative analysis of phylogenetics and evolution in R. (2023).

65. Ives, A. R. & Garland, T. Phylogenetic Logistic Regression for Binary Dependent Variables. *Syst Biol* **59**, 9–26 (2010).

66. Liu, G., Chater, K. F., Chandra, G., Niu, G. & Tan, H. Molecular Regulation of Antibiotic Biosynthesis in Streptomyces. *Microbiol Mol Biol Rev* **77**, 112 (2013).

67. Karray, F., Darbon, E., Nguyen, H. C., Gagnat, J. & Pernodet, J. L. Regulation of the biosynthesis of the macrolide antibiotic spiramycin in Streptomyces ambofaciens. *J Bacteriol* **192**, 5813–5821 (2010).

68. Bunet, R. *et al.* Characterization and manipulation of the pathway-specific late regulator AlpW reveals Streptomyces ambofaciens as a new producer of kinamycins. *J Bacteriol* **193**, 1142–1153 (2011).

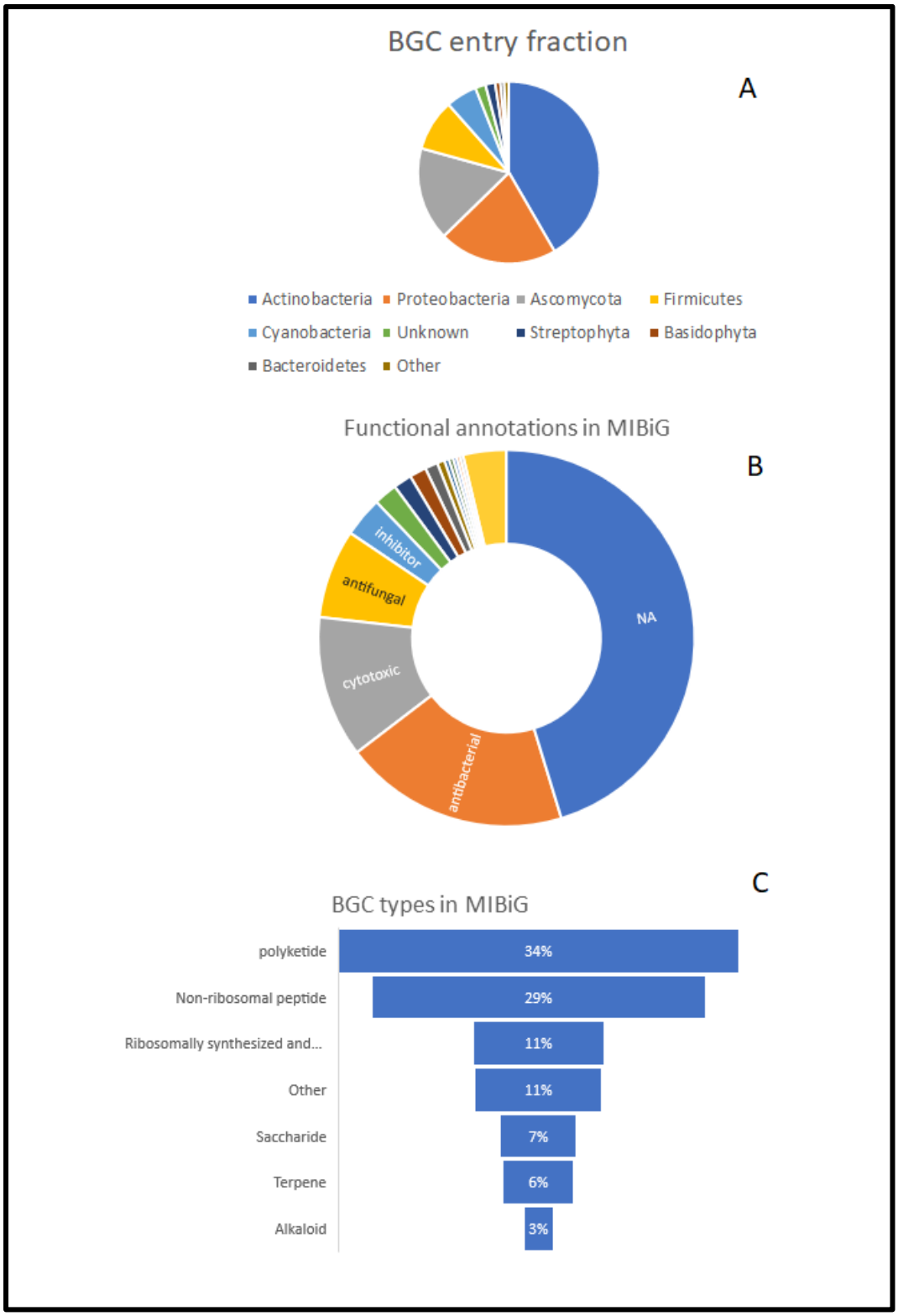
69. Bate, N., Butler, A. R., Gandecha, A. R. & Cundliffe, E. Multiple regulatory genes in the tylosin biosynthetic cluster of Streptomyces fradiae. *Chem Biol* **6**, 617–624 (1999).

70. Carroll, L. M. *et al.* Accurate de novo identification of biosynthetic gene clusters with GECCO. *bioRxiv* 2021.05.03.442509 (2021) doi:10.1101/2021.05.03.442509.

71. Sanchez, S. *et al.* Expansion of novel biosynthetic gene clusters from diverse environments using SanntiS. *bioRxiv* 2023.05.23.540769 (2023) doi:10.1101/2023.05.23.540769.

7. APPENDICES

APPENDIX A: CURRENT CURATED BGCS AND THEIR ANNOTATIONS IN THE MIBIG DATABASE. A) the fraction of present phyla, B) division of functions and C) fraction of BGC types



APPENDIX B: KEY TERMS USED TO CAPTURE REGULATORY PROTEIN FAMILIES FROM THE HMM LIBRARIES

Key terms	
PFAM	**antiSMASH's smCOG**
Regulator	Regulator
Transcription	Transcriptional
Repressor	Factor
Activator	Kinase
Histidine kinase	DNA-binding
DNA-binding	Repressor
DNA binding	ECF
Inducer	
Helix-turn-helix	
HTH	
Helix-loop-helix	
HLH	
Winged	
Sigma	
Serine/threonine	
Fork	
Ribbon-helix-helix	
RHH	
Sensor	
Inhibitor	
Quorum-sensing	
Leucine zipper	
sRNA regular	
Zinc-finger	
Zinc finger	
Co-activator	
Response	

APPENDIX C: LOCATIONS OF THE THESIS'S INITIAL DATASETS

Stored data name	Location on the server
MIBiG Genbank files	/lustre/BIF/nobackup/nassa006/MIBiG_regulatory_Pfams/mibig_gbk_3.1
MIBiG JSON files	/lustre/BIF/nobackup/nassa006/MIBiG_regulatory_Pfams/mibig_json_3.1
Extracted BGCs from MIBiG	/lustre/BIF/nobackup/nassa006/MIBiG_regulatory_Pfams/BGC_seq_files
Streptomycetaceae Genbank files	/lustre/BIF/nobackup/reitz001/seq_data/streptomycetaceae/genbanks/

APPENDIX D: CUSTOM SCRIPT NAMES

All can be found in: https://git.wur.nl/daan.vannassauw/thesis_BGC_functional_inference/-/tree/main/scripts

#	Script name
1	extract_MIBiG_clusters.py
2	parse_json_info.py
3	extract_streptomycetacaea_cds.py
4	filter_pfams.py
5	filter_smcogs.py
6	regs_from_mibig.py
7	regs_from_streptomycetaceae.py
8	Hmmer_parser.py
9	attribute_table_maker.py
10	location_analysis.py
11	regions_overview.py
12	homolog_split_and_filtering.py
13	location_stats.py
14	extract_prot_fam.py
15	fasta_ID_converter.py
16	alignment_trimming.py
17	functions_rep_view.py
18	Prediction_stats.py
19	
20	

APPENDIX E: SARP CLUSTERS

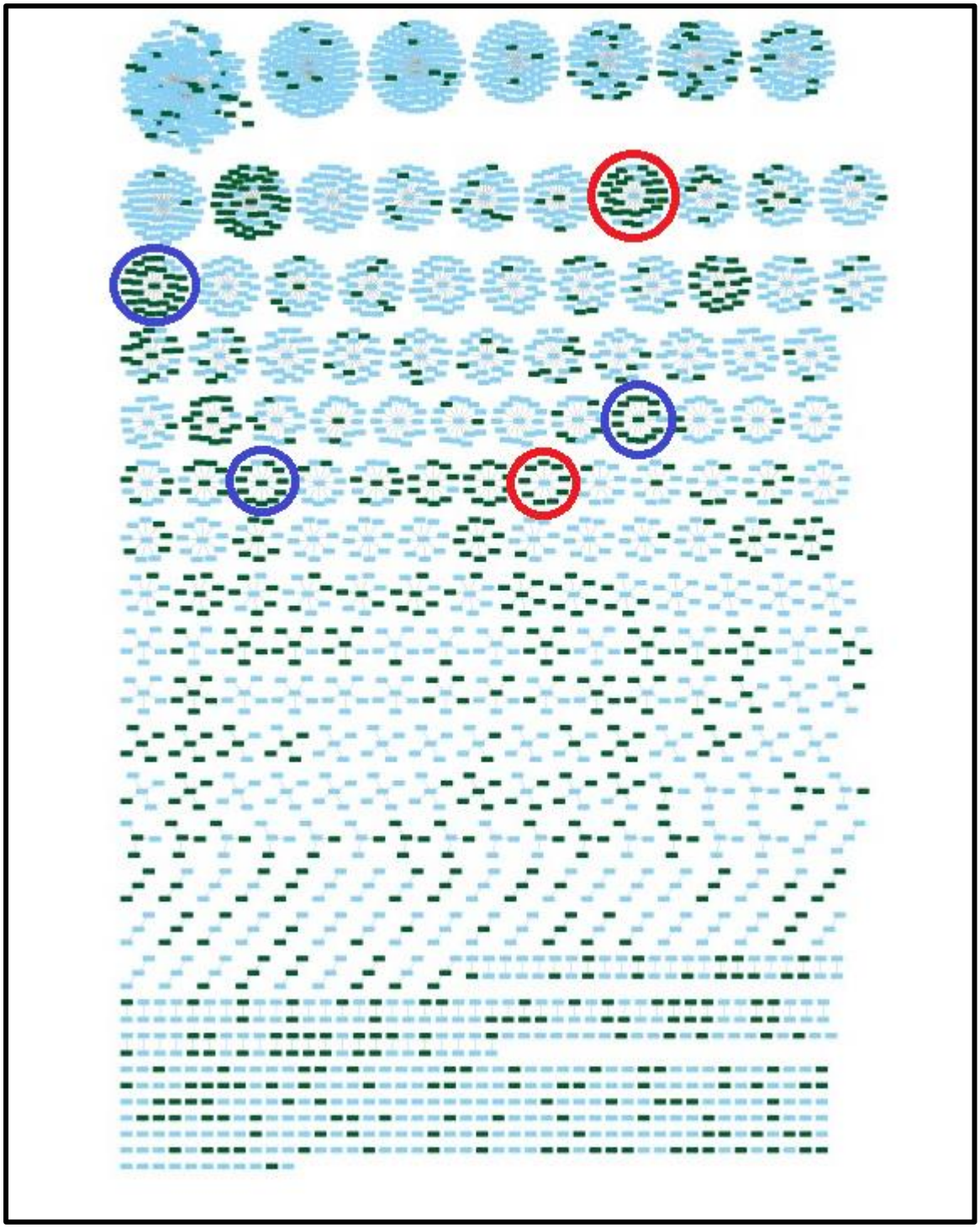
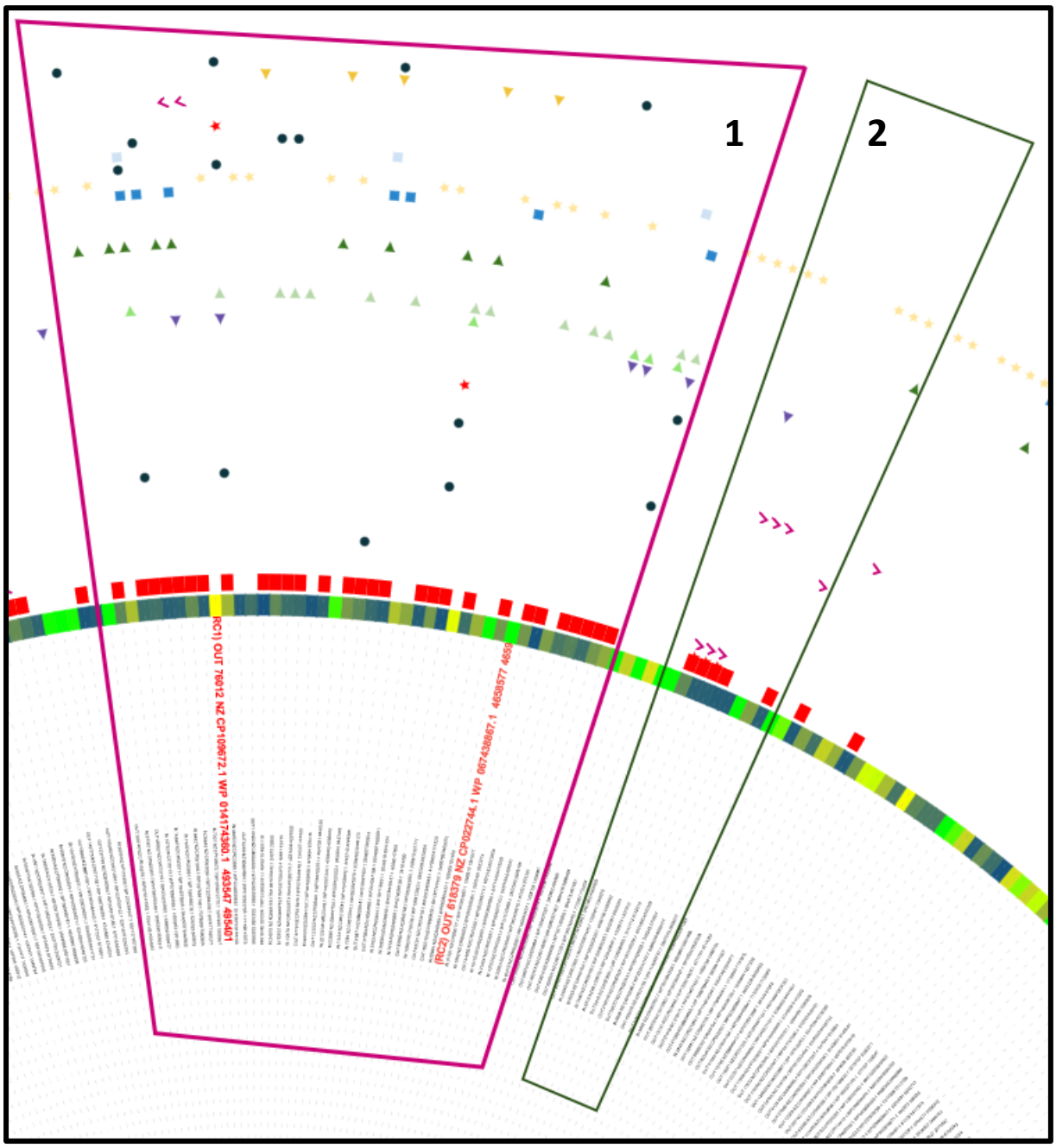
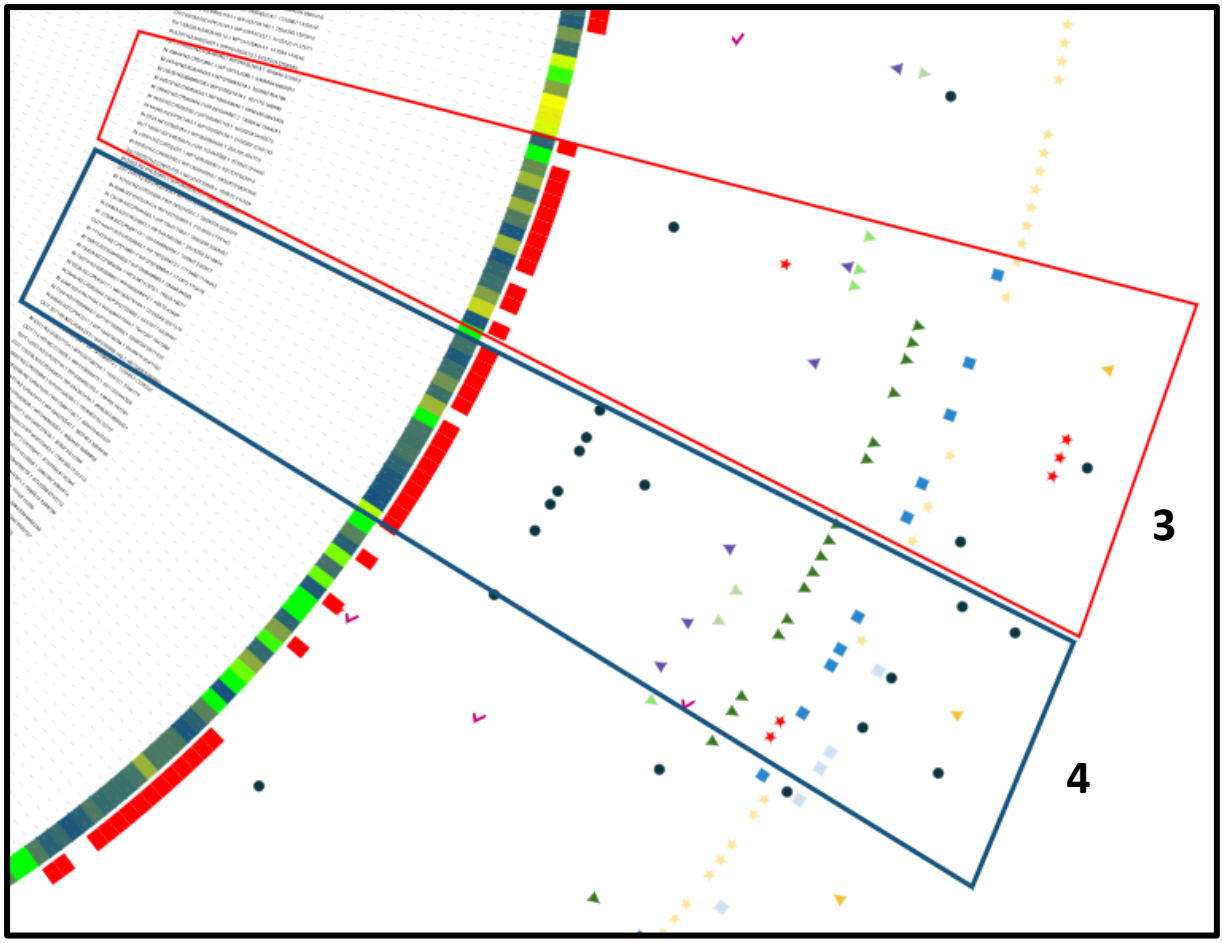


FIGURE 4 THE SARP HOMOLOGS BASED ON INCREMENTAL CLUSTERING WITH 98 TARGET COVERAGE. DARK GREEN SARPS ARE LOCATED IN PREDICTED BGC REGIONS. A LIST OF NAMES OF EVERY HOMOLOGUE CAN BE FOUND ON THE GIT. RED CIRCLE INDICATES CLUSTERS WHOSE REPRESENTATIVE HAS A DIFFERENT LOCATION THAN IT MEMBERS. BLUE CIRCLES INDICATE CLUSTERS WITH INDIVIDUAL OUT BGC-LAYING SARPS AS INTERESTING FOLLOW-UP CASES

APPENDIX F: ZOOMED-IN IMAGES OF THE SARP REPRESENTATIVE TREE





APPENDIX G: TOTAL AMOUNT OF DETECTED HOMOLOGS PER DATABASE (FEW EXAMPLES)

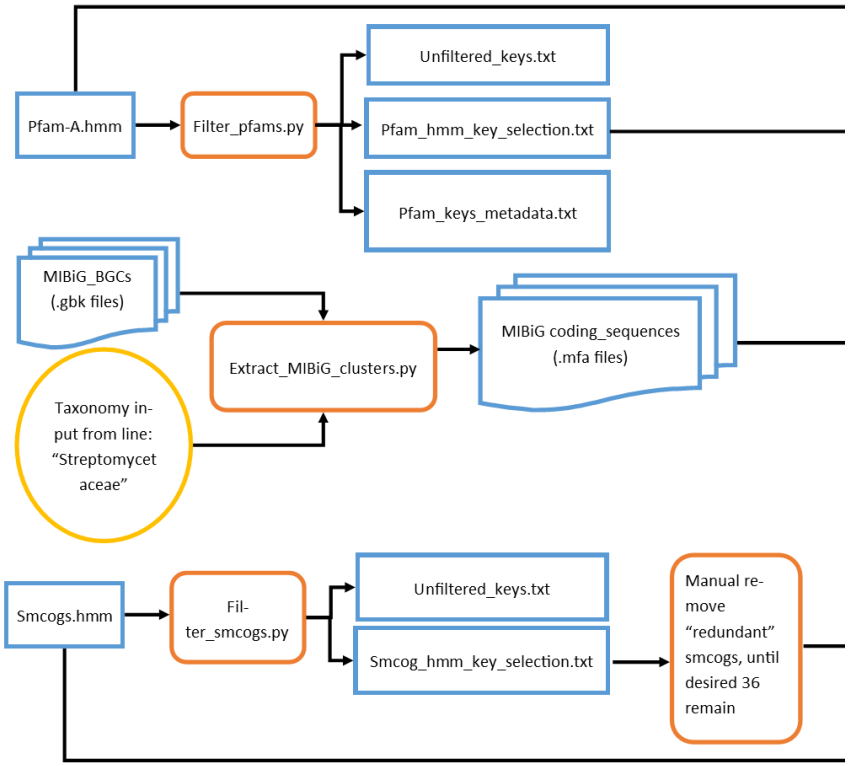
The full list of occurrences per accession can be found on the git (filename; Genomes_filter_and_count_comparison.txt). The values are the sum of all the HMMs dedicated to a single family. Thus, overlapping hits or HMMs were not analysed.

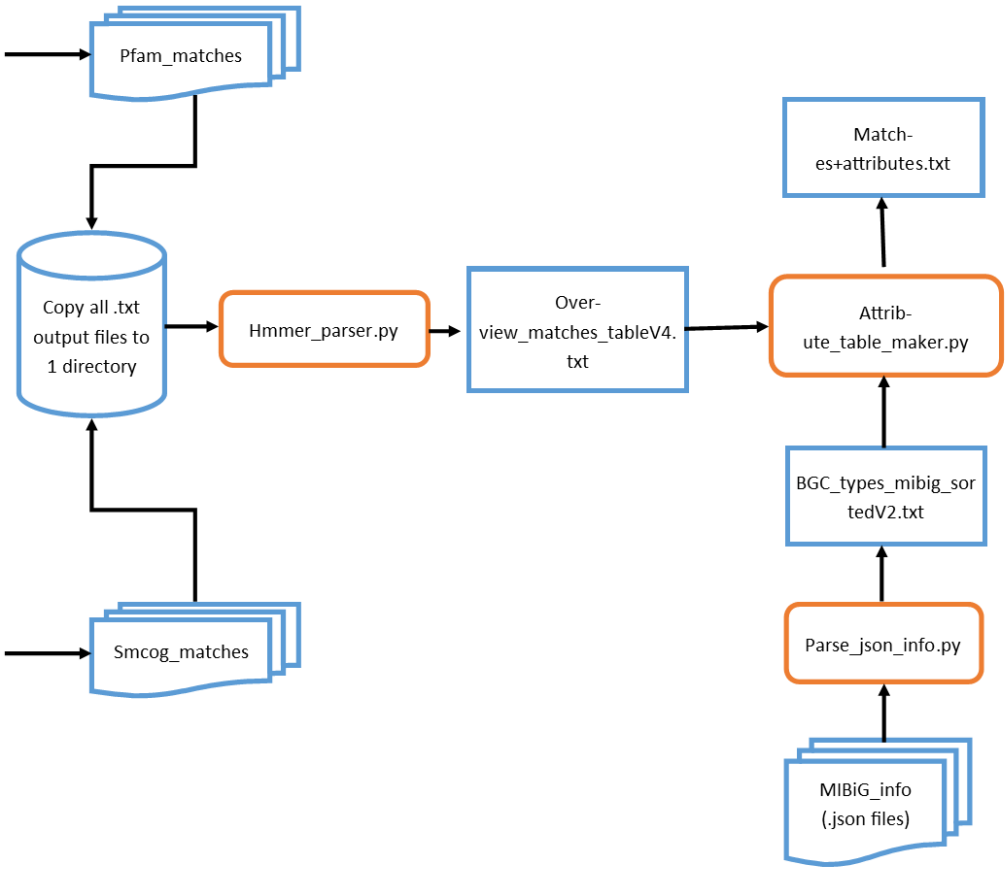
FAMILY NAME	**smCOG-R**	**Pfam-R**
LuxR	10929	3997
LysR	14473	2104
Sigma-70 factor	3845	4617
GntR	10646	3224
SARP (BTAD in Pfam-R)	6191	906

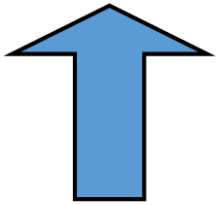
APPENDIX H: (REGULATORY) PROTEIN FAMILY OCCURRENCES IN BGCS, NOT CAPTURED BY ANTISMASH' SMCOGS

HMM_accession	HMM_match_protein_family_description	domain_description	Comments	literature	GO-terms
PF17755.4	UvrA DNA-binding domain	UvrA hypothesized to locate DNA damage as a part of a UvrABC complex		https://doi.org/10.1038%2Fnsmb.1973	
PF08220.15	HTH_DeoR	HTH DNA-binding TF (negative) regulator family (involved in sugar catabolism)		PMID: 10714997	GO:0006355, GO:000
PF03551.17	PadR; TF regulator family	winged helix-like DNA-binding TF family, possibly involved in nitrone-oligomycin A resistance & produ		https://doi.org/10.1186/s13036-018-0103-x , https://d	
PF01381.25	HTH_3	the full protein fold incorporates a helix-turn-helix motif, but the function of this member is unlikely to		https://doi.org/10.1002/prot.22698	
PF13693.9	HTH_35	winged HTH DNA-binding	has 1 hit --> BGC0001386_BAU98050.1_[23801:24392](+)		
PF13560.9	HTH_31	showed homology hits with lambda repressor-like & Cro/XRE domains (xenobiotic response element)		https://www.uniprot.org/uniprotkb/Q53895/entry	
PF13443.9	HTH_26	Cro/C1-type = Cro/XRE domains (xenobiotic response element)		https://www.ebi.ac.uk/interpro/entry/pfam/PF1344	
PF13413.9	HTH_25	'probably" binds to DNA	has 1 hit --> BGC0002350_QTT77481.1_[89142:89760](-)		
PF12844.10	HTH_19	This family contains many example antitoxins from bacterial toxin-antitoxin systems. In other domain databases, these are refered to as lambda repressor-li			
PF03444.18	HrcA_DNA-binding TF_repressor (also HTH domain)	This domain is always found with a pair of CBS domains. CBS doma	has 1 hit --> BGC0001200_ctg1_orf262	PMID:14722619	
PF06224.15	HTH_42	tend to include recently found DNA glycosylases, that play essenti	This family contains two copies of a v	PMID: 28396405 , PMID: 35311535 , PMID: 26400161	
PF02082.23	Rrf2; Iron-dependent Transcriptional regulator	HTH, Several proteins in this family form iron-sulfur clusters enabling iron dependent DNA transcriptior		PMID:23644595	
PF13589.9	HATPase_c_3 (histidine kinases)	ATPase domains of histidine kinase, DNA gyrase B and HSP90			
PF13581.9	HATPase_c_2 (histidine kinases)	ATPase domains of Sensor histidine kinases			
PF13551.9	HTH_29	This helix-turn-helix domain is often found in transferases and is l			
PF13518.9	HTH_28	This helix-turn-helix domain is often found in transposases and is	sharing hits		
PF13592.9	HTH_33	This helix-turn-helix domain is often found in transferases and is l			
PF08327.14	AHSA1	It is probably a general upregulator of Hsp90 function, particularly contributing to its efficiency in condi		PMID:12504007	
PF13556.9	HTH_30	often found at the C-terminus of PucR-like transcriptional regulat	activator for purine metabolic proces	https://www.uniprot.org/uniprotkb/O32138/publicat	
PF17765.4	MLTR_LBD	MmyB-like transcription regulator ligand binding domain, found in a family of actinobacterial transcript		PMID:22844465	
		> high chance of being a transcriptional regulator			
		> indication to be a TF, but lack of evidental literature			
		> low chance of being a transcriptional regulator			

APPENDIX I: DATA FLOW VISUALISATION WITH SCRIPT NAMES



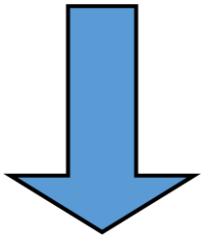

Pfam_matches
Copy all .txt output files to 1 directory
Hmmer_parser.py
Over-view_matches_tableV4.txt
Match-es+attributes.txt
Attrib-ute_table_maker.py
BGC_types_mibig_sortedV2.txt
Parse_json_info.py
MIBiG_info (.json files)
Smcog_matches

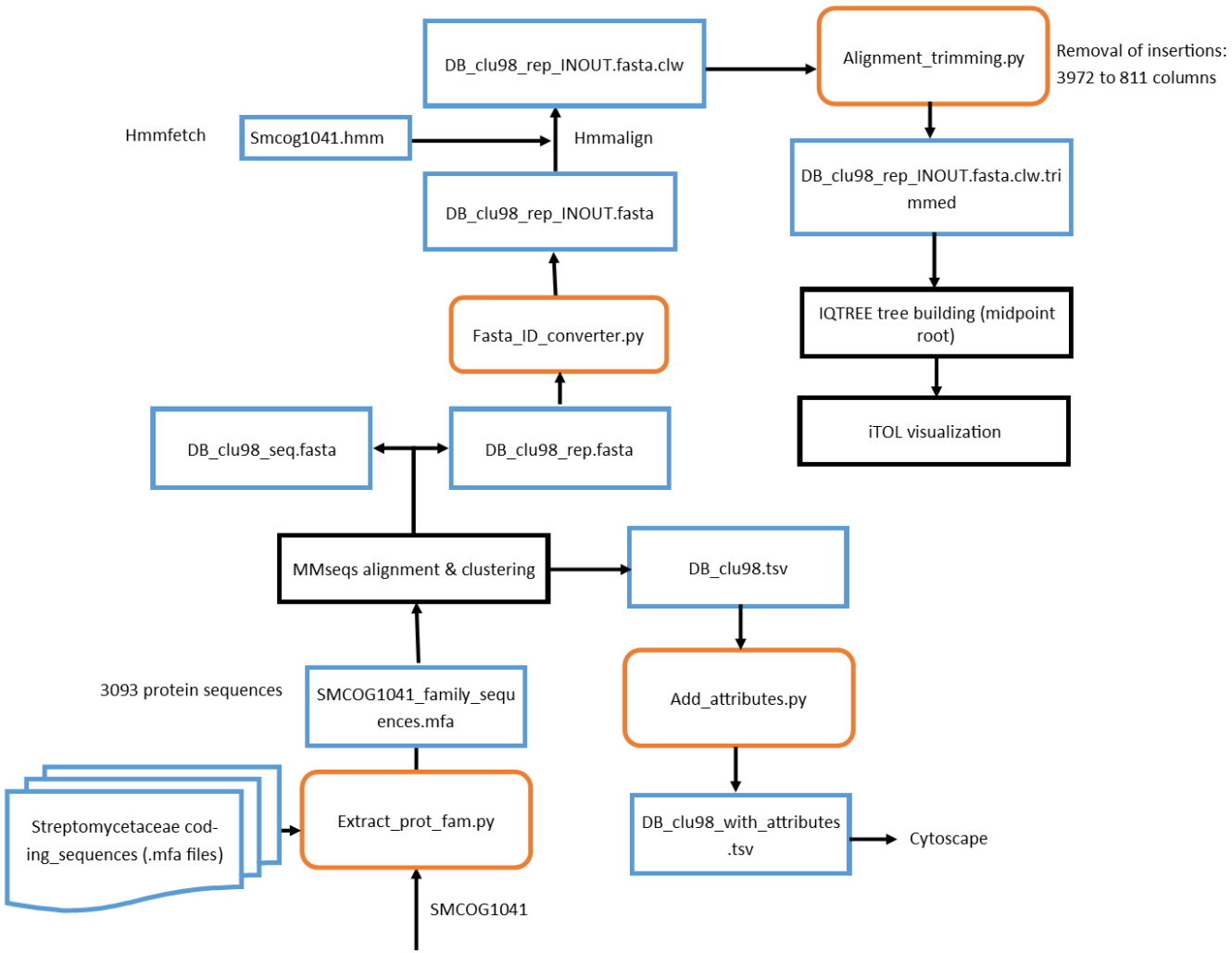


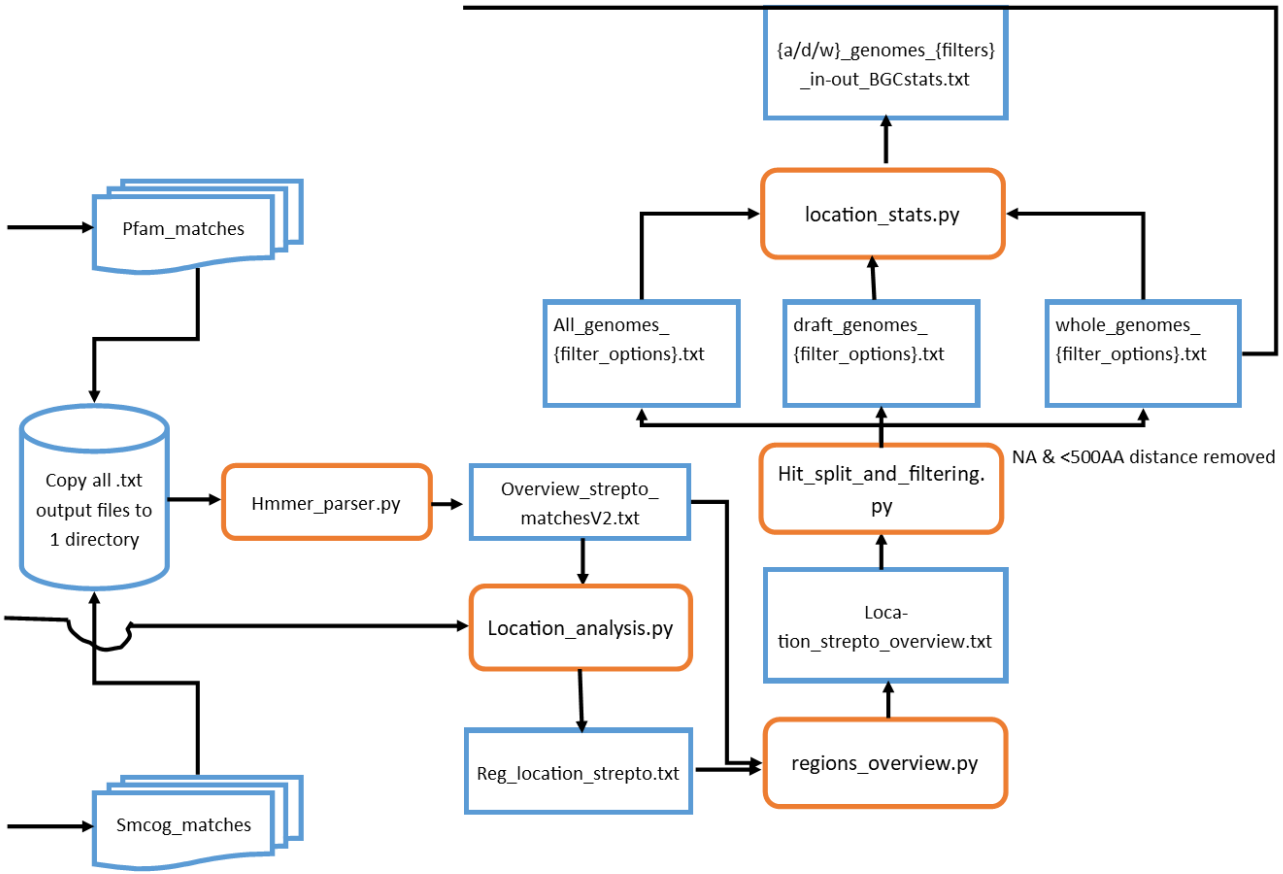
BGCs only

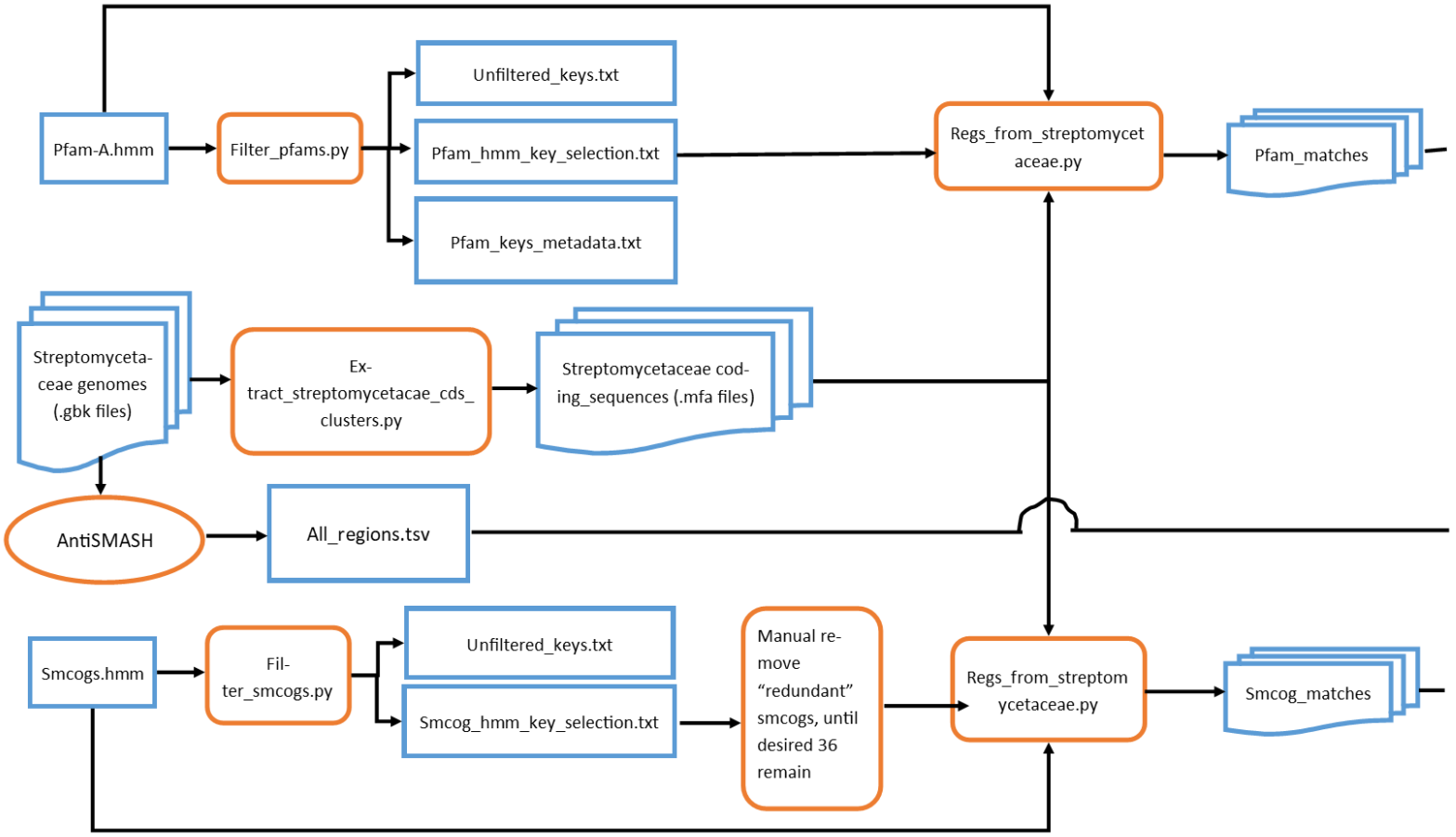
Streptomycetaceae

Whole-genomes




DB_clu98_rep_INOUT.fasta.clw
Alignment_trimming.py
Removal of insertions:
3972 to 811 columns
Hmmfetch
Smcog1041.hmm
Hmmalign
DB_clu98_rep_INOUT.fasta
DB_clu98_rep_INOUT.fasta.clw.trimmed
Fasta_ID_converter.py
IQTREE tree building (midpoint root)
DB_clu98_seq.fasta
DB_clu98_rep.fasta
iTOL visualization
MMseqs alignment & clustering
DB_clu98.tsv
3093 protein sequences
SMCOG1041_family_sequences.mfa
Add_attributes.py
Streptomycetaceae coding_sequences (.mfa files)
Extract_prot_fam.py
DB_clu98_with_attributes.tsv
Cytoscape
SMCOG1041


{a/d/w}_genomes_{filters}
_in-out_BGCstats.txt
location_stats.py
Pfam_matches
All_genomes_
{filter_options}.txt
draft_genomes_
{filter_options}.txt
whole_genomes_
{filter_options}.txt
Copy all .txt
output files to
1 directory
Hmmer_parser.py
Overview_strepto_
matchesV2.txt
Hit_split_and_filtering.
py
NA & <500AA distance removed
Location_analysis.py
Loca-
tion_strepto_overview.txt
Reg_location_strepto.txt
regions_overview.py
Smcog_matches


Pfam-A.hmm
Filter_pfams.py
Unfiltered_keys.txt
Pfam_hmm_key_selection.txt
Pfam_keys_metadata.txt
Regs_from_streptomycetaceae.py
Pfam_matches
Streptomyceta-ceae genomes (.gbk files)
Ex-tract_streptomycetacae_cds_clusters.py
Streptomycetaceae cod-ing_sequences (.mfa files)
AntiSMASH
All_regions.tsv
Smcogs.hmm
Fil-ter_smcogs.py
Unfiltered_keys.txt
Smcog_hmm_key_selection.txt
Manual re-move "redundant" smcogs, until desired 36 remain
Regs_from_streptomycetaceae.py
Smcog_matches

APPENDIX J: FRACTION OF BGCS IN RELATION IN ALL STREPTOMYCETACEAE GENOMES (extracted with *18)

mean fraction: 12.852339167026784
median fraction: 12.252232381438137

genome	amount_of_BGCs	fraction_of_genome (%)
GCF_000717595.1	35	10.428897599280038
GCF_001008345.1	27	10.284363331050178
GCF_026339895.1	18	9.863593357909451
GCF_006334995.2	23	9.989899061464302
GCF_003751605.1	27	11.695062325544015
GCF_014649895.1	30	8.17498659473775
GCF_000444875.1	36	16.983340427914243
GCF_000009765.2	36	15.752209866451313
GCF_001514065.1	32	9.317470222169831
GCF_000242715.1	47	15.863426539020503
GCF_014648775.1	37	19.623643703507536
GCF_000158915.1	41	21.259099023302745
GCF_019933235.1	32	15.56169978199646
GCF_023155275.1	26	10.542432797355024
GCF_014650095.1	33	10.68187120016651
GCF_016741875.1	78	23.573752286815264
GCF_016745505.1	29	7.776394305373099
GCF_000725565.1	46	21.771716808750362
GCF_019890725.1	27	6.15916886482877
GCF_016598475.1	24	7.456380323158306
GCF_009811555.1	39	14.859561429487075
GCF_000966975.1	52	6.229302866888363
GCF_001418575.1	29	3.6239014056646943
GCF_001418415.1	39	7.808344806206508
GCF_014650995.1	30	8.847237494181329
GCF_016103465.1	97	11.869909017408682
GCF_020907985.1	43	13.96348992240515
GCF_023614315.1	20	12.455304771737291
GCF_009709555.1	22	5.990510962825775
GCF_900114215.1	31	14.313772384052218
GCF_014141525.1	35	5.816087157394179
GCF_004684805.1	76	13.723134522341345
GCF_014649415.1	37	14.787749127200694
GCF_003721215.1	83	14.00519452110743
GCF_014156695.1	38	7.566574296737903
GCF_008868685.1	40	12.151178560150528
GCF_001493375.1	32	10.81702401022406
GCF_003112535.1	40	14.566736051529263
GCF_014648815.1	24	8.908182526000797
GCF_020404845.1	32	10.132727725102614
GCF_003258295.1	37	11.908277059517312
GCF_004348415.1	47	7.409196375187922
GCF_000225525.1	34	10.194857653422762
GCF_003097515.1	29	16.4612226224176
GCF_003696235.1	30	8.459893413147013
GCF_014649015.1	32	10.659401176627501
GCF_000158955.1	31	12.504556085323662
GCF_001294335.1	49	15.115260477341428
GCF_016598615.1	36	11.426766074450084
GCF_023516615.1	36	8.407449440888563
GCF_019104725.1	20	9.846687629974934
GCF_002155905.1	79	10.024109082559624
GCF_000717055.1	29	9.088207077377172
GCF_012034175.1	32	16.31835619102516
GCF_021474425.1	45	11.028804645575969
GCF_000719135.1	21	8.557918228187903
GCF_001514205.1	39	11.967561956840553
GCF_027947595.1	25	9.986140586261213
GCF_008974245.1	26	10.950698550081988
GCF_003391135.1	45	25.718504396187285
GCF_001005085.2	46	11.627331747920332
GCF_023923245.1	32	15.181258387530244
GCF_005768555.2	46	7.537369132779865
GCF_000717745.1	27	11.057215747111385
GCF_009796285.1	29	9.943148873860016
GCF_014779715.1	29	20.09054959441399
GCF_014648695.1	34	10.269529058644062
GCF_000269985.1	40	15.069567421183756
GCF_002804165.1	30	15.539436679944346
GCF_014656115.1	43	24.846696059470013
GCF_000980885.2	38	14.721424188837679

GCF_005047355.1	26	6.428952091537023	
GCF_015160855.1	30	10.631857053577596	
GCF_001418475.1	45	6.1150594807796494	
GCF_006547165.1	49	12.89978345273361	
GCF_014216335.1	22	14.357449680526138	
GCF_000154965.1	27	8.510350680421332	
GCF_008704515.1	24	10.991649515914762	
GCF_014295035.1	29	17.49487841561234	
GCF_014649595.1	49	15.004352212537508	
GCF_021462825.1	27	9.683318104505462	
GCF_003344445.1	32	10.893436442866172	
GCF_016803985.1	21	13.056799569008612	
GCF_008806595.1	30	10.44873880389085	
GCF_018138715.1	53	24.02307204719209	
GCF_008704555.1	26	11.648246100209956	
GCF_900110255.1	23	8.91323888579517	
GCF_000331185.2	43	22.78126121500129	
GCF_008704935.1	24	9.759772224154556	
GCF_002217755.1	41	13.076248892302523	
GCF_000725125.1	27	10.76753156988939	
GCF_001514055.1	39	13.74844112937261	
GCF_002982015.1	37	8.19507008713831	
GCF_008634015.1	25	10.093002177538253	
GCF_000091305.1	33	11.967154397683643	
GCF_019599145.1	33	11.590218873567713	
GCF_020328095.1	46	16.25377787741709	
GCF_020312215.1	26	7.793825865829916	
GCF_009755605.1	47	20.96121778467533	
GCF_003261055.1	32	11.033391827075206	
GCF_014646335.1	37	19.510432878771105	
GCF_014649195.1	37	12.767348732258032	
GCF_900105695.1	45	19.486176629093464	
GCF_000787835.1	36	12.8786414922436	
GCF_009377185.1	56	6.975386197143714	
GCF_026340005.1	37	9.19153164243531	
GCF_015910445.1	22	9.446406573959209	
GCF_900111245.1	29	14.143382403665369	
GCF_016467295.1	33	18.74490588416051	
GCF_000830005.1	31	11.275782663265668	
GCF_028807635.1	29	16.41830407333676	
GCF_014650915.1	32	12.16463687027844	
GCF_014645895.1	20	10.720167721837017	
GCF_000787815.1	41	13.549611859705719	
GCF_014656275.1	37	12.408131886648817	
GCF_014649755.1	36	12.661771876072208	
GCF_017813245.1	24	7.909694885288471	
GCF_000725495.1	29	7.396566073932119	
GCF_014649735.1	27	11.83063838357128	
GCF_001751365.1	20	8.751628823897626	
GCF_005280215.1	30	16.01352342202929	
GCF_003112595.1	29	11.359914515285988	
GCF_000720835.1	35	13.865472775660995	
GCF_007280575.1	30	6.113084822050114	
GCF_003626535.1	29	13.55832035825673	
GCF_003627815.1	34	14.357508431982485	
GCF_016103505.1	126	13.325111341558957	
GCF_000717025.1	37	12.812323253296807	
GCF_006636205.1	33	19.199767866257403	
GCF_014650615.1	34	11.581716349020427	
GCF_000010605.1	37	19.497903621712748	
GCF_020639365.1	33	12.174340379168754	
GCF_018138705.1	51	24.47453985326938	
GCF_002154675.1	38	8.4223651835196	
GCF_025908395.1	48	22.636096049933553	
GCF_014653855.1	54	13.763764308237194	
GCF_002082175.1	32	15.140692100663905	
GCF_016917755.1	42	18.623043184731277	
GCF_008905045.1	47	10.848752839900774	
GCF_000716875.1	40	16.604485222320484	
GCF_014648995.1	34	11.29836806432297	
GCF_000981895.1	34	8.658944492322913	
GCF_017315755.1	29	5.21487720834863	
GCF_014650055.1	39	9.509105392770962	
GCF_000380165.1	23	13.263561241433946	
GCF_001514235.1	41	13.718397414312916	
GCF_000720455.1	33	17.219900900855414	
GCF_000716805.1	27	6.374980733824376	
GCF_002242805.1	44	12.75759853658518	

GCF_000745345.1	24	10.079988295768189	
GCF_026427415.1	31	11.29822040874911	
GCF_000718135.1	30	13.107103209302132	
GCF_001886595.1	30	12.232074670179658	
GCF_000744655.1	27	8.675065479192167	
GCF_001748305.1	44	17.036648052527102	
GCF_020312145.1	20	10.204778619356428	
GCF_002150735.1	37	8.759792311396874	
GCF_014648075.1	33	15.482711495581505	
GCF_018255875.1	27	14.318264571639483	
GCF_012033785.1	30	8.8638351534126	
GCF_021474405.1	42	9.362383012440924	
GCF_003323735.1	29	12.364539356740398	
GCF_014651055.1	25	8.355317322880348	
GCF_002891295.1	39	25.969372535640467	
GCF_014649515.1	36	9.797483827657434	
GCF_024436055.1	31	14.553774055379801	
GCF_001419795.1	74	8.597386563373247	
GCF_008704445.1	23	12.132276408051057	
GCF_014649395.1	50	16.566491745590874	
GCF_024508375.2	31	8.50558818972905	
GCF_003112575.1	32	14.189700450115117	
GCF_900142575.1	36	14.022299811351122	
GCF_003429565.1	20	5.647135476268679	
GCF_009600885.1	59	17.299137928488527	
GCF_002082605.1	26	15.701722428997108	
GCF_001513975.1	31	11.416197975119378	
GCF_001027185.1	30	11.43295309394305	
GCF_016860545.1	31	12.33178189028226	
GCF_014650355.1	48	15.453704661359719	
GCF_003122365.1	28	13.499508454731806	
GCF_001751255.1	58	9.561238446446337	
GCF_900103455.1	24	9.997987451548953	
GCF_014647875.1	37	7.316097662402937	
GCF_011766325.1	34	16.41411610231578	
GCF_006716135.1	32	15.363568602451966	
GCF_014650255.1	29	9.583978016266297	
GCF_907177275.1	37	14.222082318383134	
GCF_001513965.1	55	17.449574150535373	
GCF_013407765.1	17	9.336930755886117	
GCF_003675325.1	31	17.68987917824289	
GCF_000716675.1	39	15.840121431971083	
GCF_014651135.1	28	10.247275664268157	
GCF_000829715.2	28	12.158399451701047	
GCF_020564935.1	29	12.27377666449735	
GCF_011045015.1	39	8.991226556993775	
GCF_014645835.1	37	11.272650324343282	
GCF_015160875.1	19	7.947007743393875	
GCF_005795905.1	33	13.546242183724836	
GCF_000718095.1	44	14.366227937162613	
GCF_016804005.1	28	16.258292915771193	
GCF_025399795.1	34	9.019887225714198	
GCF_014651115.1	62	21.990925277036055	
GCF_002954775.1	32	11.076879442194878	
GCF_003323715.1	33	12.550713496560586	
GCF_002148965.1	46	8.59451378463296	
GCF_016918855.1	36	13.155612636997095	
GCF_014645815.1	58	15.583026803585739	
GCF_020819595.1	39	9.658212504842279	
GCF_003626575.1	46	22.382998410219408	
GCF_000935125.1	33	12.73890886562572	
GCF_014650155.1	41	12.742662610649738	
GCF_900188405.1	34	9.938073103140876	
GCF_006335015.1	27	17.494980539226145	
GCF_021216675.1	23	12.910175743988036	
GCF_009811575.1	34	13.850837499083502	
GCF_009377205.1	48	7.696598407558013	
GCF_017874715.1	68	23.623312243189215	
GCF_014649035.1	37	12.584793184871284	
GCF_020010925.1	33	12.373328935277902	
GCF_017676385.1	23	9.435551546247211	
GCF_020521255.1	21	8.160842970065158	
GCF_000429085.1	24	11.892584386243165	
GCF_028657195.1	51	15.433186346169256	
GCF_008369065.1	42	9.634496141500202	
GCF_000696185.1	32	12.950278630317172	
GCF_000716445.1	36	9.788047742851145	
GCF_001879105.1	10	3.3305538915350175	

GCF_001270025.1	27	6.848975649600892	
GCF_002891435.1	35	4.981317417306841	
GCF_002843305.1	34	12.175102557375826	
GCF_000725555.1	25	11.966155552749482	
GCF_008312835.1	39	13.286053519573013	
GCF_001905345.1	36	15.938621918823184	
GCF_023218175.1	26	12.025071791861103	
GCF_002335465.1	39	10.015815779885308	
GCF_900112355.1	24	12.971634701040994	
GCF_000787775.1	33	8.96242054182905	
GCF_014656295.1	31	11.510072792834976	
GCF_007829875.1	34	17.817194666292252	
GCF_000802245.2	29	13.904459183986006	
GCF_014701095.1	23	8.697565309143592	
GCF_019857225.1	25	7.557172002712539	
GCF_015767775.1	35	17.06594630896187	
GCF_026339705.1	18	6.435309259161276	
GCF_001514305.1	32	10.284285895546098	
GCF_013912435.1	18	5.509771024407793	
GCF_014656215.1	24	9.117032028298482	
GCF_019399205.1	40	8.234714647971195	
GCF_015689475.1	48	24.871206206158046	
GCF_000359525.1	23	14.325859160562022	
GCF_008634025.1	40	21.378726724772886	
GCF_020521275.1	27	13.257332291403188	
GCF_014534645.1	38	11.654249924933007	
GCF_003665435.1	24	12.368783759391592	
GCF_014205055.1	29	12.770776703242904	
GCF_005786655.1	33	9.099142534140206	
GCF_001704635.1	64	13.130522667007336	
GCF_001751245.1	25	18.097531727663874	
GCF_017676345.1	35	14.472964382299908	
GCF_001513955.1	38	10.833068889907127	
GCF_021556455.1	23	5.2215307587873845	
GCF_000725745.1	45	17.665849269430645	
GCF_014648875.1	45	17.05478574659076	
GCF_009604385.1	27	12.202979886504272	
GCF_014493765.1	22	7.180482709296249	
GCF_000993785.3	21	13.630568782309933	
GCF_018141485.1	32	20.56921494528963	
GCF_002082585.1	38	19.91415318889836	
GCF_014654785.1	42	21.836074044803794	
GCF_009739465.1	30	12.844732048275572	
GCF_014203855.1	51	26.342489552636923	
GCF_013394065.1	38	22.31968697956296	
GCF_014650895.1	30	10.632083170921137	
GCF_016741855.1	42	15.870751242731334	
GCF_000721185.1	28	14.620527999921459	
GCF_014649635.1	46	13.38880241365187	
GCF_004028635.1	35	12.879109824916682	
GCF_013364315.1	37	16.4690604043758	
GCF_003675955.1	54	25.646584247787622	
GCF_014650395.1	24	10.974957158212819	
GCF_029223525.1	40	17.666638583795592	
GCF_000715845.1	47	14.601856837344657	
GCF_024349285.1	13	9.40113325616543	
GCF_008704535.1	35	15.9006460771595	
GCF_000717245.1	27	11.600750185465454	
GCF_000716435.1	35	10.15953964340367	
GCF_000718455.1	31	12.509578071207192	
GCF_003205575.1	35	14.043713794819315	
GCF_020521295.1	37	13.590724869912082	
GCF_014650595.1	62	13.828110419511997	
GCF_009299385.1	23	12.164817527803603	
GCF_003112515.1	24	14.453335942371412	
GCF_004122735.1	46	14.62758722187783	
GCF_002082195.1	30	11.670351595758602	
GCF_016741935.1	90	22.921668920076936	
GCF_003074055.1	19	8.0388649225124	
GCF_002939475.1	30	22.15172850605386	
GCF_001418325.1	56	10.166127377254263	
GCF_024760485.1	15	5.467094293500809	
GCF_913919575.1	27	6.930758003433487	
GCF_007828955.1	26	11.71508039086008	
GCF_011045075.1	93	12.808461685977099	
GCF_900101585.1	25	11.04900404612839	
GCF_017676365.1	33	8.494027224610887	
GCF_014655715.1	35	11.805869103214599	

GCF_014650655.1	29	8.214353475853244	
GCF_019890635.1	30	10.858225654405393	
GCF_014621695.1	19	8.55568894671389	
GCF_023516595.1	46	12.729585495841485	
GCF_000718165.1	31	7.7710991887862315	
GCF_011045025.1	40	8.893187765217744	
GCF_000719265.1	43	21.343541864057023	
GCF_014650515.1	34	13.20026598853433	
GCF_003024195.1	42	13.0894202706625	
GCF_014651175.1	27	12.6419506016723	
GCF_024666385.1	23	11.793950036547923	
GCF_020783455.1	33	16.517472036146845	
GCF_000384175.1	27	11.719869343475366	
GCF_009735685.1	26	7.1954915505763575	
GCF_009498275.1	24	7.7988993652262915	
GCF_008704855.1	34	15.24736113770925	
GCF_014650175.1	27	14.618048696426994	
GCF_016921115.1	36	13.949828905706458	
GCF_008704715.1	35	15.064187218487954	
GCF_021028635.1	32	18.54167177598669	
GCF_005280195.1	30	12.274814724448717	
GCF_014649675.1	32	11.232210307952178	
GCF_022647665.1	25	9.328446432593335	
GCF_000718015.1	29	14.346024433130733	
GCF_003963535.1	34	13.284881518332114	
GCF_008932075.1	33	11.657989170147003	
GCF_014649135.1	32	11.654339907463246	
GCF_014649795.1	46	14.837705744405724	
GCF_014654935.1	38	14.754417006106952	
GCF_000381025.1	27	15.699486780487751	
GCF_000725475.1	43	14.16913667771505	
GCF_017942185.1	47	23.35374710971838	
GCF_000739045.1	28	11.700603218553592	
GCF_000988945.1	46	5.497746359673623	
GCF_001509475.1	43	11.080894580916354	
GCF_014203895.1	30	9.69893571518539	
GCF_014655295.1	36	18.702610158807246	
GCF_014655955.1	41	12.615319270935647	
GCF_020400655.1	36	12.10777639208193	
GCF_014649775.1	43	11.699753801398026	
GCF_020400605.1	41	14.226270835252889	
GCF_002154375.1	55	6.5735942664211695	
GCF_020881015.1	29	13.722071398394492	
GCF_001700515.1	43	10.071555553669848	
GCF_014489635.1	47	20.881267324168704	
GCF_004117935.1	25	9.250239864325446	
GCF_027626975.1	32	13.514367872450686	
GCF_000280865.2	32	16.827970352790857	
GCF_014649855.1	30	8.889556559046428	
GCF_014650695.1	46	15.077882478719554	
GCF_014203555.1	24	15.374638585352246	
GCF_014650875.1	39	11.98509421624931	
GCF_001013905.1	35	17.01468182605558	
GCF_000237305.1	39	19.800648818734953	
GCF_003865155.1	45	21.26847348248882	
GCF_003955715.1	31	9.181250659636191	
GCF_001705785.1	75	19.0743261463552	
GCF_014646055.1	72	19.438899181347256	
GCF_005869865.1	33	10.578261294498605	
GCF_000725785.1	35	10.572829253450765	
GCF_001514145.1	33	8.243214594590343	
GCF_900102095.1	24	10.834711037529404	
GCF_008704425.1	29	13.742043621820516	
GCF_014649375.1	36	9.50930737695454	
GCF_013364095.1	20	14.636904954755295	
GCF_009600895.1	48	13.763458702624828	
GCF_001485145.1	36	11.84679650651208	
GCF_024761905.1	19	13.992466359354797	
GCF_000744785.1	30	13.09795499413162	
GCF_001642695.1	36	10.12194433577926	
GCF_017114865.1	25	12.015881654462696	
GCF_014650115.1	49	14.584278825752955	
GCF_018101125.2	19	12.890664388242671	
GCF_015690355.1	24	7.306476766299928	
GCF_014650295.1	31	13.028789308280158	
GCF_004784475.1	22	10.836822539272463	
GCF_002154385.1	34	8.465622209774395	
GCF_014650335.1	41	19.50117915437602	

GCF_001642995.1	20	6.192286639939643	
GCF_008704395.1	28	15.67122361444851	
GCF_001514265.1	28	10.674229702709283	
GCF_014203645.1	48	17.167662652387644	
GCF_014648955.1	32	14.710712906193171	
GCF_021462265.1	35	10.340857973108305	
GCF_020037025.1	66	19.2917353851466	
GCF_001611795.1	28	10.980856008099195	
GCF_000497445.1	44	16.02228401891302	
GCF_014646275.1	35	16.52033856571017	
GCF_000262345.1	30	8.424130452195476	
GCF_014650755.1	36	10.837377073377738	
GCF_014650435.1	34	9.485560664493752	
GCF_014650215.1	41	13.492711797504498	
GCF_002154505.1	27	4.1424246166383725	
GCF_001514035.1	53	16.596553375084355	
GCF_003121295.1	35	19.208025157393664	
GCF_013618545.1	30	11.644885077398376	
GCF_014651075.1	25	8.576511486362083	
GCF_001866645.1	27	8.488489647154216	
GCF_001906585.1	25	9.047196043705153	
GCF_000968685.2	85	19.477937479059428	
GCF_011694815.1	21	12.13366863709027	
GCF_004794175.1	54	9.84255303397582	
GCF_001278075.1	26	12.252232381438137	
GCF_009811635.1	30	14.379230959680317	
GCF_900103985.1	22	10.357507305519546	
GCF_003355155.1	40	21.892187163849584	
GCF_000372745.1	48	15.024166000228744	
GCF_000718985.1	19	9.174229123422815	
GCF_014648635.1	47	17.21050850561009	
GCF_001660045.1	20	12.088485837499393	
GCF_016031615.1	35	18.835598457675065	
GCF_023887685.1	32	13.715192358474862	
GCF_003999195.1	18	4.68053758249621	
GCF_001189035.1	34	10.83236683984848	
GCF_016906185.1	32	14.871591255445923	
GCF_001700505.1	53	13.597700494240414	
GCF_016860525.1	51	18.00243422892203	
GCF_014257025.1	34	11.204893084790829	
GCF_023498005.1	29	13.375279591021453	
GCF_020532645.1	40	12.806193606053368	
GCF_003932715.1	34	20.84011467977996	
GCF_001445655.1	51	24.707879124024533	
GCF_025402955.1	31	14.300236762696649	
GCF_008704795.1	44	24.253973627431385	
GCF_014651015.1	35	12.918885770763133	
GCF_012034385.1	34	12.2416809879204	
GCF_014852565.2	35	11.475844166727162	
GCF_014646115.1	37	12.349654206638185	
GCF_001484625.1	22	7.973548362947726	
GCF_014650715.1	35	11.24945923894761	
GCF_910593825.1	29	14.5432134322761	
GCF_009569385.1	29	14.246020706469839	
GCF_004803895.1	32	11.668834083142276	
GCF_001267885.1	26	13.19224485137363	
GCF_900104815.1	28	13.74579125274481	
GCF_009377235.1	38	7.609909831998589	
GCF_008312845.1	44	18.755434825686926	
GCF_014649055.1	29	7.592995714550481	
GCF_014649875.1	28	11.7764525528802	
GCF_029223485.1	49	14.487852006291838	
GCF_014656195.1	48	18.884434333863563	
GCF_003967355.1	45	18.367517088420936	
GCF_900110735.1	22	12.164313767116017	
GCF_000717995.1	34	9.584893981768463	
GCF_900109465.1	31	11.951468487026487	
GCF_019880305.1	30	11.241496535599873	
GCF_019890615.1	27	15.537862600205138	
GCF_000719095.1	21	6.8745927170185555	
GCF_003994395.1	33	15.61096028812236	
GCF_028401405.1	53	15.864429416653875	
GCF_014650555.1	31	15.428572314789287	
GCF_014673495.1	34	18.78599997893968	
GCF_008704575.1	23	9.885583057876067	
GCF_001953875.1	35	10.959905291070813	
GCF_021261325.1	36	10.345083592615955	
GCF_000383595.1	29	10.028433205750154	

GCF_001514215.1	36	14.819843828896293
GCF_014646095.1	39	12.103263949330035
GCF_015244315.1	35	18.659408102690275
GCF_900100315.1	15	5.438446045767504
GCF_014647675.1	23	7.491274770547609
GCF_003270085.1	25	6.541382473343224
GCF_011044995.1	19	4.582096257498387
GCF_017349075.1	33	19.32212554518065
GCF_001418565.1	71	12.41069532963945
GCF_002794255.1	27	9.46337657655706
GCF_014141535.1	41	6.633366733936512
GCF_000719285.1	33	10.971996704452406
GCF_003054555.1	28	12.680639206290733
GCF_008705135.1	37	14.249271988018464
GCF_017639205.1	29	13.555229817056924
GCF_014649655.1	51	14.6490091730271
GCF_014649495.1	46	14.325204178484121
GCF_009811595.1	33	14.336476940036313
GCF_016755875.1	26	11.161029940798487
GCF_024172095.1	28	12.10630059048042
GCF_000718025.1	43	17.53725648511464
GCF_003865135.1	43	20.319207270060907
GCF_008386495.1	36	9.931909398513369
GCF_011044975.1	27	4.806868939085099
GCF_017353455.1	33	4.820707585999716
GCF_000497425.1	31	13.582466394618189
GCF_003344965.1	31	16.96272152368841
GCF_000478605.2	18	10.842712476179
GCF_001044425.1	28	11.030829094308402
GCF_001419745.1	60	13.578946995238752
GCF_001187435.1	38	25.395045298584268
GCF_006715785.1	33	10.729271742893012
GCF_019059395.1	51	24.5023620921529
GCF_028421465.1	28	13.629382820466434
GCF_900107965.1	30	12.591277383475674
GCF_000718305.1	38	12.024781808038213
GCF_015710995.1	31	12.639753147638189
GCF_004023625.1	34	8.590825950829597
GCF_000744815.1	18	4.994656405530591
GCF_022699385.1	29	12.181909784724969
GCF_012033735.1	39	12.467093959530903
GCF_000744225.1	34	19.591543697183095
GCF_007829885.1	21	11.497748607998913
GCF_014650955.1	43	12.83806238034625
GCF_014650795.1	30	10.321970493867308
GCF_001984445.1	25	10.752817571958586
GCF_000744705.1	39	16.208309896841108
GCF_014649935.1	20	7.68647557104698
GCF_017876625.1	41	25.379130098717102
GCF_000836635.1	23	11.937412230263602
GCF_014202475.1	36	10.958252589585372
GCF_002286695.1	66	15.861993756279144
GCF_014648975.1	39	12.103888642810068
GCF_008704495.1	38	21.585881522182433
GCF_000961885.1	68	16.871095608128194
GCF_013433285.1	40	11.776155754294367
GCF_014648935.1	37	15.861995775912849
GCF_014650575.1	31	11.201316931276619
GCF_005405925.1	44	18.353118644119366
GCF_026343715.1	26	9.74934937577577
GCF_003814885.1	26	11.791427364722102
GCF_024519315.1	67	16.5674157994618
GCF_009377175.1	33	6.690290511691896
GCF_014655595.1	36	18.789252600389446
GCF_002155915.1	28	7.541944298350427
GCF_003330865.1	25	15.079624216183563
GCF_000955965.1	41	10.122743813952216
GCF_001746455.1	34	15.047831108921214
GCF_014656095.1	66	17.138530800226732
GCF_001418495.1	28	5.279111643033543
GCF_000716545.1	23	9.97902477333699
GCF_024436035.1	27	12.191859710929075
GCF_014203595.1	47	17.484040246275296
GCF_900230195.1	33	15.12205419548572
GCF_018070025.1	38	14.7268269657855
GCF_014649695.1	23	8.88424991941227
GCF_016901035.1	41	8.756929135005755
GCF_004328625.1	22	12.417827207023697

GCF_000718625.1 45	15.48036096980435
GCF_016741775.1 39	15.935310651269555
GCF_009184865.1 29	9.173450976075728
GCF_000220705.2 23	9.596237880348692
GCF_016919245.1 35	15.288349480748087
GCF_000709915.1 19	8.158954076543006
GCF_014656055.1 25	10.625135657803773
GCF_001735805.1 33	15.972948818779109
GCF_014655855.1 30	11.36772319994811
GCF_007856155.1 31	20.937548026854287
GCF_020099395.1 33	24.718093735014957
GCF_026342395.1 35	8.972853397903743
GCF_003346515.1 56	13.721395876199905
GCF_018927715.1 42	10.391950551321628
GCF_014650235.1 28	10.086110263048077
GCF_003258605.2 19	13.619141430467804
GCF_014648835.1 39	20.0312596432959
GCF_014489615.1 27	15.497963142011445
GCF_000720485.1 38	12.737635531368591
GCF_009739905.1 26	12.223730757928525
GCF_000718635.1 36	18.04179556019829
GCF_001542625.1 50	18.69433296354353
GCF_001514125.1 33	15.263660258239762
GCF_028401765.1 34	10.046909508199974
GCF_014650815.1 49	15.609526885636827
GCF_024752535.1 26	13.656757431304122
GCF_026343615.1 32	11.766779977301793
GCF_013409565.1 25	14.84021906176828
GCF_004305975.1 54	11.527502253110251
GCF_003595235.1 23	7.3219756405811465
GCF_014651095.1 41	13.367912535346688
GCF_000716625.1 28	9.960522299529375
GCF_000725545.1 34	11.714794125225064
GCF_014650975.1 18	8.619756233846683
GCF_005981925.1 35	7.3516477617610265
GCF_002150845.1 57	12.633390021549094
GCF_014654675.1 28	7.063071148866635
GCF_009709575.1 25	15.696739329875891
GCF_014649715.1 25	8.853426275008358
GCF_002128305.1 29	18.71524616430568
GCF_017526105.1 33	16.14270918778527
GCF_014649335.1 25	10.565518227658583
GCF_014649535.1 25	10.586605785985633
GCF_000376565.1 33	7.574507370421597
GCF_014650775.1 33	14.896865336842918
GCF_022221585.1 26	9.127117132075627
GCF_006539505.1 34	9.775117832581152
GCF_022385335.1 27	8.564098979975089
GCF_009908195.1 33	5.679721149989549
GCF_014656135.1 23	10.80102307051528
GCF_002946835.1 24	9.460893477292236
GCF_020024005.1 34	9.183531149260906
GCF_000974985.2 26	7.236326244500632
GCF_002911015.1 40	11.169546048465568
GCF_001419765.1 57	7.723628952675772
GCF_014656035.1 22	10.646724888536454
GCF_007829815.1 50	22.48059034566365
GCF_001866665.1 61	15.890177698236066
GCF_026341945.1 26	11.844329061323984
GCF_014650135.1 44	13.103671358106036
GCF_014649995.1 26	8.304900507094574
GCF_014650035.1 38	13.490224655458533
GCF_014649915.1 25	13.243161149973934
GCF_900112845.1 19	8.045117245904128
GCF_002261115.1 37	17.114912340081336
GCF_011008945.1 79	12.607304553124198
GCF_018069625.1 42	20.45740315705337
GCF_014651035.1 36	12.792244586659052
GCF_002154615.1 28	4.9986247879982555
GCF_014647975.1 35	11.309594640726072
GCF_019219635.1 27	15.04568428945006
GCF_014701115.1 27	10.224245001242524
GCF_014650735.1 54	13.811205247643365
GCF_014648915.1 41	12.7483945266841
GCF_005048155.1 33	14.783608769185788
GCF_014203705.1 27	11.340539197116337
GCF_016107395.1 32	10.6655112979884
GCF_017352335.1 21	11.807799969584368

GCF_027270315.1	32	14.875831465077818
GCF_012273655.1	38	22.56655051060125
GCF_000787855.1	36	12.324305999977202
GCF_014649115.1	51	19.03276821910648
GCF_003626645.1	42	23.864782581543615
GCF_000349325.1	35	13.973304590333772