|  |  |
| --- | --- |
| **POSSIBLE ACRONYMS FOR APPROACH** |  |

TargetSEARCH Target Search based on the Existence of Antibiotic self-Resistance genes in Clusters using Homology

MAGiC Mining for Antibiotic self-resistance biosynthetic Gene Clusters

CATCHY Curate AnTibiotiC Harboring polYketide

**FIGURES AND TABLES**

Table 1. Experimentally validated gene clusters harboring a self-resistance gene

Figure 1. Identifying PKS clusters harboring a putative self-resistance gene

Table 2. Number of PKS clusters harboring a putative self-resistance gene and more than one copy per nucleotide record.

Figure 2. Phylogenetic tree of KS domains proximal to a self-resistance gene (150 KSs)

Figure 3. Coevolution of KS and putative self-resistance gene. a) positive set, b) 5kb, c) 10kb, d) >50kb KS-self-resistance gene distance cutoff

Table 3. Top 20 ranked clusters harboring a putative self-resistance gene

FigureS1a. Phylogenetic tree of KS domains closest to a self-resistance gene 12 targets, 10kb, annotated

FigureS1b. Phylogenetic tree of KS domains closest to a self-resistance gene 92 targets, 10kb

FigureS1c. Phylogenetic tree of KS domains closest to a self-resistance gene 609 targets, 10kb

Figure S2a,b,c. Coevolution of KS and target 12 targets, 10kb, subplots, individual plots

Figure S2d,e,f. Coevolution of KS and target 92 targets, 10kb, subplots, individual plots

Figure S2g,h,i. Coevolution of KS and target 609 targets, 10kb, subplots, individual plots

Figure S3. Sequence similarity between bacillaene and PKS cluster from *Paenibacillus polymyxa.*

Figure S4a. Coevolution of KS and UDP-3-O-acyl-N-acetylglucosamine deacetylase (Lpxc) target. Left – within 10kb of a KS. Right: at least **50kb** away From a KS

Figure S4b. Coevolution of KS and Lon, DegP, FtsH, and ClpP protease targets within 10kb of a KS

Table S1. Scoring scheme

TableS2. Catalog of 150 clusters harboring a potential self-resistance gene (12 targets, 10kb)

TableS3. Catalog of 572 clusters harboring a potential self-resistance gene (92 targets, 10kb)

TableS4. Catalog of 3431 clusters harboring a potential self-resistance gene (609 targets, 10kb)

Table S5. Table of putative novel antibacterial targets

Supplementary file 1. Targets.12.fasta

Supplementary file 2. Targets.92.fasta

Supplementary file 3. Targets.609.fasta

**ABSTRACT**

With the increase in antibacterial resistant pathogens and the decline in discovery of novel antibiotics, the need for accelerating their discovery is urgent. More than two-thirds of the known antibiotics are natural products or natural product derivatives, with polyketides and non-ribosomal peptides being one of the most prolific classes. With the increasing throughput of DNA sequencing, hundreds of novel PKS and NRPS gene clusters have been identified in recent years. However, accessing novel and useful PKS activities remains complicated, as many host organisms are unculturable or do not express PKS genes in laboratory conditions. Furthermore, a significant challenge in the natural product discovery field is the identification of useful biosynthetic gene clusters, without the *a priori* knowledge of the biological target of the produced compound.

To identify PKS genes that are likely to produce an antibacterial compound, we developed an automated method to identify and catalog clusters that harbor potential self-resistance genes. These genes provide the cluster-harboring microorganism a defense against the antibiotic encoded by the cluster.

With this approach, we generated a catalog of 150 non-redundant PKS clusters. Manually curated list of known self-resistance genes was used to mine all NCBI nucleotide and genome databases. The algorithm takes into account (1) the distance of the potential self-resistance gene to a core enzyme in the biosynthetic gene cluster; (2) the presence of a duplicated housekeeping copy of the self-resistance gene; (3) the presence of close homologs in diverse species; and (4) evidence for coevolution of the self-resistance gene and core biosynthetic gene, and (5) self-resistance gene ubiquity, and scores the clusters based on these criteria. This method is generalizable since it can be applied to extract not only PKS-related gene clusters, but also other types of natural products. It can also be used to identify potentially bioactive clusters not only from bacterial origin, but also from fungal and plant native hosts. Most importantly, it can be used to prioritize gene clusters harboring putative novel targets and encoding compounds with new mechanisms of action.

**INTRODUCTION**

Resistance in pathogens occurs quickly after a drug is introduced in the clinic due to overuse and misuse of current antibiotics in the clinic. Thus, there is a constant need for the discovery and development of new antibacterials with novel mechanisms of action or novel targets of inhibition. There are even more challenges for antibiotic discovery for multi-drug resistant Gram-negative pathogens,, since they have a double cell wall, which prevents antibiotics molecules to reach thei intracellular targets.Despite extensive efforts, there have been no new classes of antibiotics approved against Gram-negative pathogens in over fifty years. (Smith et al, 2018). At least two thirds of drugs used in the clinic are natural products or derived from natural products, and polyketides constitute one of the well-represented classes [Ref]. They are synthesized by large enzymes called Type I polyketide synthases (PKSs) and nonribosomal peptide synthetases (NRPSs) respectively, many of which are multi-modular and act in assembly-line manner In microbial genomes, PKS and NRPS genes are often co-localized with other genes required for the production of a given compound and form a biosynthetic gene cluster.

[TODO: finish introduction section]

* Traditional discovery- bioactivity based; complicated, labor-intensive and yields fewer results, [ref]
* Sequence-guided discovery. Advent in dna sequencing, number of modular clusters (2013 PKS paper), and typeII clusters (2016 PNAS paper). Most clusters are orphan clusters – chemical structure not bioinformatically linked to a biosynthetic gen cluster (BGC).
* Computational prediction of clusters – trivial now (Automated genome mining tools like antiSMASH) but accessing novel and useful PKS activities remains the bottleneck.
* One way to prioritize: look for self-resistance genes. Many antibacterial compounds are produced by bacteria and, in order to avoid self-toxicity, these antibiotic-producing microorganisms have developed several self-resistance mechanisms. One resistance mechanism is target modification, where a resistant copy of the target gene is co-localized and co-expressed with the biosynthetic genes [10]. There are several examples in literature mining for specific targets of interest and specific class of microorganisms. Mining self-resistance genes to discover clusters with potential antibacterial properties – examples, narrow focus on fatty acid inhibitors, actinomycetes, etc. (Bradly Moore paper)
* Below, a self-resistance gene which is a modified target of the encoded compound, will be referred to as “self-resistance gene”
* The antibiotic resistant target seeker (ARTS) is a great web tool to detect known-resistance mechanism in a biosynthetic gene cluster, based on BCG proximity, duplications, and HGT. However, it has its limitations: 1) limited to Actinobacterial genomes and thus it cannot provide info for HGT or presence of housekeeping self-resistance genes for non-Actiobacterial species; (2) limited to known-resistant mechanisms; (3) one criteria: BGC proximity; it relies on Antismash to predict the boundaries of the BGC and antismash over-predicts boundaries to include all possible biosynthetic genes. As a result, the putative self-resistance gene is >20kb away from the core biosynthetic genes. [TODO: focus on how it is complimentary and different from my approach].

We have developed an automated method to identify and catalog clusters that harbor a potential self-resistance gene. To validate the algorithm's performance in identifying such clusters, we used a manually curated list of clusters known to harbor a self-resistance gene. We used this approach to detect new clusters harboring these genes, suggesting their possible function in providing self-resistance. We also applied this method to search for clusters harboring other genes that could be involved in self-resistance and present the highly ranked top 20 hits. We propose that by identifying clusters with possible self-resistance genes, our method allows to search the vast numbers of orphan PKS clusters for those likely to produce antibacterial compounds. It also allows to form hypotheses about the potential targets of those compounds, and as such could be a useful tool in biosynthetic cluster prospecting.

**Materials and Methods**

**Curation of a reference set of experimentally validated self-resistance genes**

In order to identify bacterial type I PKS biosynthetic gene clusters (BGCs) harboring a self-resistance copy of the target gene, we curated a list of known antibacterial self-resistance genes co-localized with the type I PKS BGCs within the annotated gene clusters in the Minimum Information About a Biosynthetic Gene Cluster (MIBiG) database (ref). All 1816 cluster files from MIBiG database as of August 6th, 2018 were parsed to find putative targets for each MIBiG entry (“Molecular Target” filed on the MIBiG website). UniProt database was searched for a set of proteins matching this description. We made a BLAST database for each cluster with all of the proteins in that clusters. Then for each protein in the set of UniProt protein sequences, we used BLASP to check if there is a homolog in the cluster. We filtered the BLASTP hits using identity threshold of 50% and “Molecular Target” field as “Antibacterial” or “Cytotoxic”. For any given hit from the BLAST search, we searched the literature for experimental evidence of self-resistance of that gene in that particular cluster. Of these results, combined with extensive literature search, we identified 16 clusters harboring one of 12 experimentally verified self-resistance protein families, spanning polyketides, non-ribosomal peptides and their hybrids, terpenes, ribosomally synthesized and post-translationally modified peptides (RiPPs) and aminocumarins. Table 1 lists the number of clusters harboring each of these 12 self-resistance gene families.

**Identifying Type I PKS clusters harboring a putative self-resistance gene in proximity to a KS domain**

*Step 1. Performing BLAST search for KS homologs.*

All NCBI nucleotide and genome databases were searched for KS homologs using tblastn. 8 diverse KS from modular Type I PKS (erythromycin), cis-AT PKS/NRPS (curacin, epothilone, guadinomine, rapamycin) and trans-AT PKS/NRPS (leinamycin, disorazol, chivosazol) clusters were used as query sequences against the major NCBI nucleotide and genome databases (nt, refseq\_genomic, other\_genomic, env\_nt, patnt, htgs, tsa\_nt, sts, gss, est\_others updated in April 2018 and wgs updated in January 2016), as well as the Bacterial Assemblies database updated in October 2018.

An initial relaxed blast search (e value < 1) identified 199,894 non-redundant protein sequences (99% identity cutoff) contained in 110,174 unique NCBI nucleotide records/genomes.

*Step 2. Running antiSMASH.*

antiSMASH 4 [REF.] with minimal functionalities was run on all 79,053 non-redundant NCBI records over 1000bp in length, including 26,573 assembly records. To speed up computation time, assembly records were split into smaller ones by extracting sequence 150kb upstream and downstream of a KS domain and running antiSMASH in parallel.

*Step 3: Generating a database of proteins found in Type I PKS clusters.*

antiSMASH identified 29,987 clusters as Type I PKS, which constituted our cluster database. We extracted all proteins from these clusters, resulting in a database of 908,532 protein sequences. We also extracted all KS domains from these clusters and constructed a KS database.

*Step 4. Performing BLAST search for putative self-resistance genes in PKS clusters*

We searched the protein database for 12 experimentally verified self-resistance proteins (Table 1) using blastp with a relaxed e-value <1, which resulted in 4,404 hits. Filtering blast hits by e-value <1e-8 and identity > 0.3 (and FabB/F identity > 0.6) reduced the set to 806 hits.

*Step 5. Identifying putative self-resistance genes in proximity to KS domains.*

In most experimentally characterized clusters harboring a self-resistance gene (Table 1), it was within a 10kb-distance of a core KS domain, with the exception of FabB/F in platencimycin BGC (13kb away) and threonyl-tRNA synthetase in borrelidin BGC (14.5kb). We set an initial threshold for the maximum allowed distance between a KS domain and a putative self-resistance gene at 10kb. This filter reduced the set to 252 hits in 250 clusters. For additional analyses, we used thresholds of 5kb and 20kb.

**Phylogenetic analyses**

To remove redundant gene clusters, we selected a standard redundancy threshold of 90% KS protein sequence identity, which further reduced the number to 150 hits in 149 unique NCBI nucleotide records. Multiple sequence alignment of KS domains located in proximity to a putative self-resistance gene was performed using MAFFT and phylogenetic tree was generated using FastTreePMP. Tree was visualized using the APE package in R. Tree was rooted on *E.coli* FabB/F.

**Housekeeping copy of the putative self-resistance gene**

To test whether there is an additional, housekeeping copy of a putative self-resistance gene within an NCBI record containing hits, we used the 12 self-resistance protein sequences as queries in blastp search of all NCBI records harboring clusters with putative self-resistance genes. We used e-vaule cutoff of of 1e-8 and identity threshold of 0.3 (and FabB/F identity threshold of 0.6 to exclude false positive hits of KS domains). The number of copies per NCBI record was counted and the status of the record – complete genome vs. incomplete was noted.

**Coevolution**

For clusters harboring the same type of a putative self-resistance gene, we defined a coevolution score between the protein sequences of self-resistance proteins and proximal KS domains. For each pair of such clusters, we plotted the amino acid sequence identity of their putative self-resistance gene (IPSRG) vs. the amino acid sequence identity of the proximal KS domains (IKS), using either a 5kb, 10kb, or 20kb distance threshold. We fist plotted pairwise identities from our positive set, including clusters with predicted self-resistance gene (clarexpoxin, eponemycin, cinnabaramide: a homolog of the target gene is co-localized with cluster but it was not experimentally validated to be the self-resistance gene). We also calculated a coevolution score for each biosynthetic gene cluster:

where N is the number of clusters pairs that harbor the same type of putative self-resistance gene and n is each such pair.

**Scoring scheme**

We devised a scoring scheme to rank the identified Type I PKS clusters harboring a putative self-resistance gene, based on several parameters: (1) distance of self-resistance gene to the closest KS, (2) presence of homologs in diverse species; (3) presence of a housekeeping copy; (4) Coevolution score; (5) self-resistance gene ubiquity – the measure of how many times a putative self-resistance gene has been identified in the data set: high number of occurrences would increase the false discovery rate and vise-versa.

To select a threshold for defining homologous gene clusters, we manually examined pairwise amino acid sequence identities of KS domains from clusters known to be homologous. Niddamycin and spiramycin produce identical compounds and their Loading Module KS pairwise sequence identity is X%. Tylactone and chalcomycin produce closely related compounds and their Loading Module KS is Y% identical. Further characterization of the thiolactomycin synthase (TLMS) clade shows that the most distant TLMS homologs have X% identity between the KS domains of their Loading Modules [?], whereas the pairwise KS identity with the Loading Modules of clusters outside of that clade and the most distant TLMS homolog is Y%. We chose a threshold for homologous clusters of 80% amino acid identity.

Detailed description of the scores for each parameter is given in Table S1.

**RESULTS**

**Curation of a reference set of 12 experimentally verified self-resistance genes in 16 biosynthetic gene clusters (Table 1)**

In order to identify bacterial type I PKS biosynthetic gene clusters (BGCs) harboring a self-resistance copy of the target gene, we curated a list of known antibacterial self-resistance genes co-localized with the type I PKS BGCs by (a) automated mining of annotated gene clusters in the MIBiG database, and (b) manual literature search.. We identified 16 clusters harboring one of 12 experimentally verified self-resistance proteins, spanning polyketides, non-ribosomal peptides and their hybrids, terpenes, ribosomally synthesized and post-translationally modified peptides (RiPPs) and aminocumarins (Table 1). lists the number of clusters harboring each of these 12 self-resistance gene families.

 Out of all 1816 deposited in MIBiG, 176 have been annotated to have antibacterial activity. 16 of these 176 harbor a modified target gene, which was experimentally validated to be the self-resistance gene. Thus, in total, at least 10% of the clusters encode antibiotics, of which 10% harbor a modified target as the self-resistance gene. 9 out these 16 clusters are type I PKSs or type I PKS hybrids, which accounts for 1.22% of all type I PKS clusters deposited in MIBiG. Of all clusters deposited into MIBiG, only 0.88% have a characterized self-resistance gene that belongs to the same family as the target gene. These numbers suggest that (1) most natural product producers do not produce antibiotics. Currently, we have limited understanding to what all these clusters are encoding [REFS]; (2) we have characterized only a fraction of the self-resistance genes, but there are many more to be discovered; (3) the self-resistance mechanism through target modification is rare in nature and most producers protect themselves from the antibacterial compounds by either drug efflux or compound modification; (4) it is possible that the self-resistance gene is not co-localized with the biosynthetic genes. This phenomenon has been observed for the streptolydigin biosynthetic gene cluster: it inhibits the beta subunit of RNA polymerase (RNA PolB) and there is a self-resistance *RNA polB* gene outside of the cluster in the natural producer *Streptomyces lydicus,* located Xkb away from a KS domain (Marina Sánchez-Hidalgo et al 2010).

**Expanding the repertoire of type I PKS gene clusters harboring one of the 12 known self-resistance genes**

We performed a BLAST search of KS homologs in all NCBI databases using 8 diverse KS domain sequences as queries. We ran antiSMASH on all positive NCBI records and made a database of all proteins contained in ~30,000 clusters annotated as type I PKS. Then we performed a BLAST search of all 12 experimentally validated self-resistance protein sequences against our protein database to identify which clusters harbor a putative self-resistance gene. Lastly, we extracted all clusters in which the putative self-resistance gene is less than 10kB away from a KS domain.

**Validation of the [ACRONYM] approach**

There are 16 clusters harboring an experimentally validated self-resistance gene, including 9 type I PKS and type I PKS/NRPS hybrids. All of them were detected by our approach, except for the andrimid biosynthetic cluster, which was not annotated as a type I PKS in MIBIG, although it contains type I PKS KS domains. 7 out of 8 clusters were identified using a KS-self-resistance gene tandem distance cutoff of 10kb, whereas and 8 out of 8 we identified using a KS-target tandem distance cutoff of 20kb. Increasing the cutoff to 20kb allowed the discovery of the borrelidin biosynthetic cluster, which inhibits a Threonyl-tRNA synthase and the producer harbors a self-resistance copy of the threonyl- tRNA synthase gene located 12kb form a KS domain.

Out of the ~30, 000 clusters annotated as type I PKSs by antiSMASH, 252 clusters had a putative self-resistance gene located within 10kb of a KS domain (Table S2). Using a redundancy cutoff of 90% resulted in 150 non-redundant clusters. Increasing the distance cutoff to 20kb increased the number of redundant clusters to 587, and the number of non-redundant clusters to 306. These are discovery rates of 0.84% and 1.95%, respectively, which are close to the fraction of characterized type I PKS clusters from MIBiG which harbor a self-resistance homolog of the target gene (1.22%). As discussed above, one possible explanation is that the self-resistance mechanism through target modification is rare in nature.

Clusters encoding the acetyl-CoA carboxylase beta subunit (ACCB) and Isoluecyl-tRNA synthetases (Ile-tRNA synthetase) are the most abundant ones (Table 2). The fraction of clusters encoding a self-resistance *accB* gene in our positive set is 0.05%, whereas our approach detected a putative self-resistance *accB* gene in 0.16% of type I PKS clusters. It is probable that some of the identified clusters are false positives. One explanation is that ACCB is a biosynthetic enzyme providing the extender unit for the polyketide chain. Similarly, Ile-tRNA synthetase could be a biosynthetic enzyme involved in the biosynthesis of amino acids incorporated in the polypeptide moiety of a natural product. However, if the cluster harboring Ile-tRNA synthetase does not contain any NRPS modules, it is more likely that Ile-tRNA synthetase is a self-resistance protein. Similarly, for any putative self-resistance gene, if the corresponding protein is involved in the primary or secondary metabolism, it is difficult to assess whether its function within a cluster is to provide self-resistance or merely to encode a biosynthetic enzyme. Genes encoding non-metabolic enzymes, such as Gyrase B, are less likely to be biosynthetic.

**Housekeeping copy of the putative self-resistance gene**

We reasoned that self-resistance genes would often be a cluster-localized duplicate of “housekeeping” gene elsewhere in the genome, and that often this genomic copy would not encode self-resistance. In the positive set of clusters, only 5 out of 9 NCBI records containing BCGs were of complete or nearly complete genomes; we examined only those records for the co-occurrence of “housekeeping” copies and we found ones in all 5 (Table 1). With our approach, we detected a “housekeeping” copy in four of them. Since we have set the sequence identity cutoff at 30%, our approach did not detect a second copy of the Ile-tRNA synthetase in the genome of *Pseudomonas fluorescens*, which is only 23% identical to the self-resistance Ile-tRNA synthase gene in the mupirocin biosynthetic cluster. Out of 150 identified clusters, only 24 are derived from complete genome sequences and a housekeeping copy is present in 12 of those NCBI records. The remaining records could not be exhaustively searched for the housekeeping copy of the self-resistance gene. In total, we identified at least two copies of the putative self-resistance gene in 27 NCBI records out of 150. The presence of a second housekeeping copy of the putative self-resistance gene is a useful feature by which to identify potential antibiotic producers. However, it is not required for the producer to keep both copies of the target, since harboring just the self-resistance one is sufficient for its survival [Ref?]. It is possibly evolutionary more advantageous for the producer to lose the sensitive copy of the target. Indeed, the rifamycin producer *Amycolatopsis mediterranei* harbors only a self-resistance copy of the RNA polymerase beta gene [ref?]. Furthermore, it is not clear whether there are extra copies of that gene per genome, since there is no reference copy number per gene per genome. We could apply the online tool ARTS to obtain a reference copy number, but ARTS output is limited to species belonging to the Actinobacteria phylum. Moreover, isolates from the same species might have different copy numbers of the same gene [Ref?]. Thus, although informative, the presence of a housekeeping copy of the putative self-resistance gene is not a required characteristic of the antibiotic producer organisms.

**Phylogeny**

We constructed a phylogenetic tree of all KS domains proximal to a putative self-resistance gene from the identified 150 clusters (Figure 2 and S1a). Two thirds of the clusters are from Actinobacteria, and the rest from Proteobacteria, Firmicutes, Cyanobacteria. A few clusters are from metagenomic sequences. We hypothesize that if a cluster was producing an antibiotic and thus conferring selective advantage, that cluster would likely have a close homolog, also harboring a homologous self-resistance gene. Indeed, there are several clades, in which the KS domains cluster by the putative self-resistance gene. Looking more closely at the architecture of the PKS clusters, we identified sets of close homologs: clusters which harbor the same number of modules and same number and set of domains. Interestingly, although there are 15 clusters harboring a gyrase B gene, the respective proximal KS domains have low sequence similarity. Possible explanations are: (1) close homologs of these 15 clusters exist, but there is a sparse coverage of the sequence space thus far; (2) phylogenetically distant type I PKS clusters target gyrase B ; or (3) there are no type I PKS clusters harboring a self-resistance gyrase B gene and all clusters we identified are false positives.

**Coevolution**

We previously used co-evolution analyses to detect coevolution and gene swap events in domains of type II PKS clusters (REF). We showed that the core KS and chain length factor (CLF) domain is a well conserved pair in type II PKS clusters and that it has coevolved: when two KSs from different NCBI records have high sequence identity, their corresponding CLFs also have high sequence identity, and when two KSs have low sequence identity, their corresponding CLFs also have low sequence identity.

In this study, we extended this approach to test if a KS domain coevolved with a putative self-resistance gene by comparing pairwise protein identity scores between pairs of homologs. We hypothesized that there will be a high correlation of KS sequence identities and the corresponding putative self-resistance protein identities, provided that the putative self-resistance gene is indeed providing protection to the bacterial producer from the biosynthesized antibiotic. However, there are several caveats of using coevolution of KS domains and self-resistance gene as a metric for presence of a self-resistance gene in a PKS cluster: the KS might have been horizontally gene transferred, which could result both in false positive and false negative results. With that caveat in mind, we first evaluated the co-evolution approach on the positive data set. For each family of self-resistance genes, we combined type I PKS clusters where the function of that gene was characterized (Table 1), and those where the gene was hypothesized to provide self-resistance, yet not experimentally validated (salinosporamide beta subunit inhibitors: cinnabaramide, clarexpoxcin, and eponemycin, Table S1). We then calculated a co-evolution score for each pair of clusters harboring self-resistance genes from the same family. As expected, sequence identities of KS domains and self-resistance proteins were highly correlated (Figure 3a). To further test this approach, we selected a KS-self-resistance gene distance cutoff of 5kb, 10kb, and >50kb (to capture random pairs) and constructed a coevolution plot for each data set (Figure 3). Pairwise identities are colored by target. There is a strong correlation of pairwise KS and self-resistance genes identities on the 5kb plot (R2 = 0.73, p-value < 0.001), which suggests that the KS and targets from these pairs coevolved. For the FabB/F target, there is even a stronger KS-FabB/F correlation, and most of these clusters harboring a FabB/F gene are close homologs of thiolactomycin and thioteramide synthases, which further increases the confidence of using coevolution plots for discovering clusters harboring self-resistance genes. There is a slight shift of the on-diagonal group: higher pairwise KS identity and lower pairwise self-resistance protein identity is generally observed. This phenomenon is possibly due to the fact that pairwise identities of KS pairs are calculated based on the KS domains only, not the entire protein sequences. KS domains are the most conserved PKS domains and do not contain more variable protein regions such as linkers, whereas the self-resistance protein sequence could be less constrained.

Increasing the distance cutoff to 10kb and then to 20kb decreased the correlation coefficient from 0.73 to 0.69 and 0.64, respectively with p-value < 0.001, suggesting that using a more stringent KS-self-resistance gene distance cutoff could reduce the false-discovery rate. Most importantly, increasing the cutoff to more than 50kb generated a plot with no significant correlation between the pairwise identities for all families of self-resistance genes (R2 = 0.19, p-value < 0.001, Figure 3d).

For each cluster, we calculated a coevolution score by averaging all distances to the diagonal of that cluster with all of its pairs. The lower the coevolution score, the more partners this cluster has as on-diagonal group of pairwise data points, which would translate in higher confidence that the putative self-resistance gene coevolved with the KS and is thus the self-resistance gene in that cluster.

An interesting feature of the coevolution plots is the stripes pattern for some of the targets, like ACCB: there are diverse pairwise KS identities for the same value of self-resistance protein identities. This pattern indicates that the self-resistance gene does not coevolve with the KS gene. As a result, other factors should be taken into account when prioritizing a cluster based on a coevolution score alone.

**Ranking clusters**

In order to select the most interesting and useful clusters which could encode a compound with antibiotic activity, we developed a scoring scheme based on (1) distance, (2) presence of a “housekeeping” copy, (3) coevolution score, (4) presence of close homologs, and (5) ubiquity of the target, which attributed a score between 0 and 70 to each cluster and ranked them based on this score. The scoring scheme is described in Table S1. As expected, most clusters from the positive set ranked high (Table S2) with scores between 40 and 70, except for kalimantacin/batumin, which received a score of 15. We listed the top–ranking orphan clusters in Table 3.

**Type I PKS catalog application**

Our approach for searching type I PKS clusters harboring a putative self-resistance gene identified a total of 150 non-redundant PKSs and PKS/NRPS hybrids. Of them, we detected all characterized type I PKS clusters that harbor a self-resistance variant of the target and mapped them on the phylogenetic tree of KS domains proximal to those genes (Figure 2). This phylogenetic map provides the opportunity to either study structural analogs of known clusters, or explore orphan clusters for which the molecular structures of the encoded compounds are not known. Approximately 9% of these gene clusters encode a known natural product (TODO: map all known gene clusters on tree, not just the ones with a self-resistance gene); the rest are orphan PKS clusters with no known compound. Several clades lack any characterized cluster and could represent PKSs encoding new classes of polyketide antibiotics. For example, the Ile-tRNA inhibibor mupirocin belongs to a clade different that the one harboring most of the identified Ile-tRNA-synthetase-harboring clusters, which could encode a novel class of Ile-tRNA inhibitor. Similarly, a group of BGCs encode beta proteasome, even though their proximal KS domains are not similar in sequence to that of salinosporamide synthase. They could produce a new class of beta proteasomal inhibitors. Furthermore, we have identified several homologs of thiolactomycin and thiotetramide BGCs (Figure S1), which could be producing their analogs. Moreover, there are several clusters which harbor putative self-resistance proteins such as Gyrase B and Elongation factor Tu and DNA polymerase sliding clamp, which were not previously reported as self-resistance genes in type I PKS clusters. Moreover, ranking of the discovered clusters should aid the prioritization of useful and biologically active natural products and should guide the refactoring of these biosynthetic gene clusters in tractable heterologous hosts. Clusters from non-actinobacterial origin, such as from Proteobacteria and Firmicutes, are attractive for expression in tractable non-actinobacterial heterologous hosts, such as *E. coli* or *B. subtilis,* respectively.

Most importantly, our approach provides hypotheses about the potential function and mechanism of action of the encoded compounds of all identified BGCs and could provide a useful resource for antibiotic discovery.

**Extending the approach to other antibacterial drug targets**

Since the traditional molecular targets for antibiotic discovery involved in DNA/RNA synthesis, protein synthesis, cell wall biosynthesis, and fatty acid biosynthesis have been over-mined without success, there is a need to characterize and discover underexplored and novel targets (Culp 2017). We hypothesized that a type I PKS cluster harboring a self-resistance variant of any of these underexplored targets could produce a compound that inhibits that target. We performed an extensive literature search to identify such kind of antibacterial targets by curating biosynthetic pathways of interest and the enzymes involved at each metabolic step, as well as antibacterial targets specific to multidrug resistant Gram-negative pathogens (REF). Furthermore, we searched MIBiG database for clusters harboring a potential self-resistance gene (Materials and Methods). We identified several BGCs with a known molecular target, which is co-localized with the biosynthetic genes, but is not experimentally validated to be the self-resistance gene (Table SX). Combining both approaches, we curated a list of 92 targets from pathways of interest for antibacterial discovery. We applied our approach by mining NCBI records for these 92 genes and identified 225 clusters with a 5kb distance threshold between the KS and the putative self-resistance gene, and 572 clusters with a 10kb threshold. Out of the 92 targets, 55 were identified using a 5kb distance cutoff, and 73 using the 10kb distance cutoff. To be able to prioritize which of these clusters might encode compounds with antibacterial activity, we ranked them using the above-mentioned scoring approach (Table S1) and identified 134 clusters with a score of 50 or higher (Table S3). These high-ranking clusters harbor putative self-resistance genes involved primarily in fatty acid biosynthesis, folate biosynthesis, cell wall biosynthesis, protein synthesis, protein degradation, DNA/RNA synthesis, and quinol/quinolone metabolism. Below we discuss a few interesting antibacterial targets and the clusters that harbor them.

**Aminoacyl-tRNA synthetases as antibacterial targets**

There are several known natural polyketide inhibitors of tRNA-synthetases: borrelidin (Thr-tRNA synthetase inhibitor), granaticin (Leucyl-tRNA synthase inhibitor), mupirocin A, (Ile-tRNA synthetase inhibitor), reveromycin, (Ile-tRNA synthetase inhibitor). Two of these, mupirocin and borrelidin, harbor a self-resistance Ile- and Thr- tRNA synthetase gene, respectively. However, this class of antibacterials are usually (1) hard to penetrate the bacterial cell wall and enter the cell; (2) their targets are structurally very similar to eukaryotic tRNA-synthetase, (3) some of these compounds show cross-reactivity against targets other than tRNA synthetases*.*  Mupirocin is a naturally occurring Ile-tRNA synthetase inhibitor produced by *Pseudomonas fluorescens*and is the only aminoacyl-tRNA synthetase inhibitor used in clinic. We mined the NCBI database for all 20 aminoacyl-tRNA synthetases and identified several high-scoring PKS clusters in Firmicutes, Proteobacteria and Actinobacteria, harboring a putative tRNA synthetase (See Table S2).

A recently identified activity of bacterial Asp-tRNA and Glu-tRNA synthetase may provide yet unexplored antibacterial targets. Some bacteria do not have Asn-tRNA and Gln-tRNA synthetase and instead, use Asp-tRNA and Glu-tRNA synthetases to load the tRNAAsn and tRNAGln with Asp and Glu respectively. The mischarged tRNA molecules are then converted to Asn-tRNAAsn and Gln-tRNAGln and used in translation. These tRNA synthetases are structurally dissimilar from the canonical Asp-tRNA and Glu-tRNA synthetase (Salazar et al 2003). Since they are also very different from eukaryotic tRNA synthetases, they are attractive antibacterial targets.

One of the high-scoring clusters in *Paenibacillus polymyxa* (CP015423) harbors an Asp-tRNA synthetase gene and is architecturally similar to the known antibiotic bacillaene. Bacillaene is a bacteriostatic antibiotic, known to inhibit prokaryotic protein synthesis but not the eukaryotic one. Three out of the four PKS proteins show 46%, 49%, and 51% sequence identity to bacillaene PKS proteins (Figure S3). However, there are two major architectural differences: (1) the bacillaene BGC has 8 modules harboring a DH domain, whereas cluster CP015423 has only 5; and (2) The bacillaene BGC has two NRPS modules, whereas CP015423 only has one. These comparisons suggest that the orphan cluster might encode a structurally similar product to bacillaene and possess antibacterial properties. Further experiments are needed to determine whether CP015423 produces antibiotic and whether the target of bacillaene and CP015423 is indeed Asp-tRNA synthetase.

**Lipid A biosynthesis**

With the continuous increase in multidrug-resistant Gram-negative bacteria, there is imminent need for discovery of antibiotics with novel mechanisms of action. UDP-3-O-acyl-N-acetylglucosamine deacetylase (Lpxc) is an enzyme that catalyzes the first committed step in lipid A biosynthesis. Lipid A is a conserved component of lipopolysaccharide, which is a major component of the outer membrane of Gram-negative bacteria. LpxC is conserved in most Gram-negative bacteria (except in *Acinetobacter baumannii*) and it does not show homology to any mammalian protein and thus is considered an attractive target [Ref]. Although drug-discovery efforts for LpxC inhibitors have been successful [Ref?], no inhibitor has been approved to treat human infections to date (Alice Erwin, 2016)..The KS and LpxC coevolution plot shows a strong correlation between pairwise KS and LpxC protein sequence identities. We also made the negative control plot for a KS and LpxC pairs encoded at least 50kB away from each other. As expected, we observed no significant correlation (Figure S4a). The high pairwise KS identities and the high correlation of KS identities and the LpxC identities suggest that either LpxC is a biosynthetic enzyme in all of these PKS clusters, or that all of these clusters are homologs. KS domains from all clusters encoding LpxC form a single clade. All 15 non-redundant clusters are annotated as Type I PKS by antiSMASH and their scores range from 30 to 70. They indeed have similar architecture, consisting of one fully reducing PKS module, and one or two additional KR genes. No function could be appointed by antiSMASH to a ~680 amino acid sequence in the PKS module in all 15 clusters, which could encode a novel class of domain. The lowest pairwise KS domains amino acid sequence identity is between CP003350 and AP013062 and is 73%. All of them belong to the Proteobacteria phylum, mostly from the *Burkholderia* genus, but also from *Caballeronia*, *Frateuria, and Ralstonia* genera. The presence of 15 homologs of the same cluster, all of which harbor a LpxC gene, suggests that this cluster might be of of evolutionary importance and the target of the encoded molecule might be LpxC. Since there are no known homologs of these clusters, the set of 15 BGCs is a potential source of LpxC inhibitors.

The most common resistance mechanism in the Gram-negative pathogens to LpxC inhibitors is a point mutation in FabZ. FabZ is a R-3-hydroxymyristoyl acyl carrier protein dehydrase (Clements et al. 2002; Zeng et al. 2013; Tomaras et al. 2014) involved in phospholipid synthesis. The mechanisms of resistance in FabZ mutants is presumably due to restoration of the balance between lipid A synthesis and phospholipid synthesis by reducing the flux through both pathways while maintaining the permeability barrier of the outer membrane. This observation suggests that FabZ could also be a potential antibacterial target.

**Proteases as antibacterial targets**

Bacterial proteases are one set of underexploited antibacterial targets (REF). Proteases are protein degrading enzymes that play key roles in bacterial physiology, biochemistry, and pathogenicity. They have been shown to be druggable, since in mammals, proteases are targets for an estimated 5–10% of all drugs being developed [ref]. However, currently there are no approved antimicrobial agents targeting bacterial proteases. There are four families of intracellular proteases in bacteria: Lon, HslUV, ClpXP and FtsH, and periplasmic/secreted proteolytic complex, DegP [ref]. Additionally, prokaryotic proteasomes are mainly constrained to Actinomycetes phylum (Mot 2007). Several of these complexes have been explored as antibacterial targets. ClpP has been the most well characterized one and is the only complex for which natural product bacterial inhibitors have been found so far. Athough cyclomarin, ecumicin and lassomycin are all active inhibitors of *M. tuberculosis* ClpP protease with high potency, further optimization of their pharmacological properties is required [ref]. The peptidyl boronate MG262 has been the only compound identified as a Lon inhibitor (Frase, H. & Lee 2007). However, it is 2000-fold more potent against the human 20s proteasome, thus exhibiting off-target effects. There are no known HsIUV, DegP and FtsH inhibitors.

In order to identify novel protease inhibitors, we applied out approach in mining NCBI records for ClpP, Lon, HsIUV, FtsH, and DegP and identified 60 clusters encoding at least one of them, including 19 clusters with scores of 50 or higher (Table S3). The coevolution plots for Lon, DegP, FtsH, ClpP and their corresponding proximal KS domains shows a high correlation (Figure S4b). Although most of the clusters are present in Proteobacteria, surprisingly, two clusters harboring a putative self-resistance FtsH gene, are from fish: *Larimichthys crocea* (large yellow croaker) and *Scophthalmus maximus* (turbot). Both clusters consist of two PKS modules. Putative type I PKS clusters have been identified in metazoans, but only a few were functionally characterized: a single module PKS, encoding the yellow pigment of feathers in parakeets (Cooke et al 2017), and a multi-modular PKS expressed during all larval stages in nematodes, possibly encoding a signaling molecule (Shou et al 2016). Although unlikely, it is possible that PKSs from fish might produce antibacterial compounds against bacterial proteases. These PKS clusters, together with the rest of the 60 highly ranked clusters, provide a great resource of potential inhibitors of bacterial proteases.

**Extending the pipeline to novel antibacterial drug targets**

In order to identify novel antibacterial targets and the type I PKS clusters which might encode self-resistance copies, we hypothesized that any gene family that is known to be essential in *E. coli* could be an antibacterial target if a gene from this family is in the proximity of a type I PKS cluster and if that clusters ranked highly in our scoring scheme. We mined the NCBI databases for type I PKS clusters harboring a putative self-resistance gene from a set of 609 gene families that are essential in *E.coli* (REF). We identified 3431 non-redundant clusters that harbor one or more such genes within 10kb of a core KS domain (Table S4). We selected all clusters which scored 50 or higher and compiled a table of all putative self-resistance genes in these clusters (Table S5), resulting in 114 unique putative self-resistance genes. Some of the identified self-resistance gene families, such as FabB/F and EF-Tu, are already considered as antibacterial targets, but even though the antibacterial target might not be novel, the identified clusters might encode novel compounds. However, to the best of our knowledge, most of the identified genes have not been evaluated as potential antibacterial targets. They could be divided in two categories: metabolic and non-metabolic. As mentioned above, it will be hard to distinguish whether a gene is the self-resistance gene or a gene involved in the biosynthesis of the compound if that gene is a metabolic enzyme.

* One way to filter – metabolic vs non-metabolic. Low false discovery rate
* Another – is there a human (or mitochondrial) homolog? If not, great target, since there won’t be cross-reactivity with the human homologs
* Interesting ones – involved in cell wall biosynthesis

**Conclusion**

[TODO: Finish section]. Despite recent advances in laboratory techniques (cite Harvey et al., a Sean Brady paper, maybe one of Nancy Kelleher’s efforts), experimental identification of novel antimicrobial compounds and of the BGC that encodes their synthesis lags behind the rate at which sequencing efforts identify BGCs. We used a purely bioinformatic technique to identify BGCs that may harbor self-resistance genes. Thus our technique provides guidance on the possible targets and modes of action of compounds produced by BCGs without experimental validation. Our method is generalizable since it can be applied to extract gene clusters from any class of natural products, which can be characterized by a core enzyme (DMATS (alkaloid) Trichodiene synthase (terpene) GGPPS (terpene)). It can also be used to identify potential clusters not only from bacterial origin, but also from fungal and plant native hosts. Most importantly, it can be used to prioritize gene clusters harboring novel targets and encoding compounds with new mechanisms of action.

Here we presented a catalog of antibiotics in reserve: when resistance emerges for one, go onto the next in the list.