**Figures and Tables:**

Table 1. Positive set of clusters harboring a self-resistant target gene

Figure 1. Pipeline (12 known targets)

Table 2. Number of pks clusters harboring a putative self-resistant gene (out of the 12) and more than one copy per nucleotide record.

Figure 2. Phylogenetic tree of KS domains closest to a self-resistant gene (152 KSs), tree colored by target (Print phylogenetic distance, if > 30%, homologs?)

Figure 3. Coevolution of KS and target (12 targets, 5kb, 10kb, 50kb)

Table 3. 10 (or 20) top ranked clusters harboring a putative known antibacterial target or putative novel antibacterial target

Supplementary table 1. Catalog of 152 clusters harboring a potential self-resistant gene target (Columns: cluster, cluster type, target, **distance**, **phylogeny?, housekeeping copy, coevolution score, ARTS score?**)

Supplementary Figure 1. Phylogenetic trees of KS domains closest to a self-resistant gene

a) 501 KSs by mining 92 potential targets, 5kb cutoff

b) X KSs by mining 92 potential targets, 10kb cutoff

c) Y KS by mining E. coli essential genes, 10kb cutoff

Supplementary figure 2. Coevolution of KS and target

1. 92 targets, 5kb, 10kb
2. 600 targets, 5kb, 10kb

Supplementary table 2. Identifying clusters harboring a potential self-resistant target gene by mining MIBIG (target of molecule known, target copy in clusrter not tested if it is the resistant)

Supplementary table 3. Catalog of 501 clusters harboring a potential self-resistant gene target (92 targets)

Supplementary table 4. Catalog of X clusters harboring a potential self-resistant gene target (600 E. coli essential genes)

Supplementary file 1. Targets.12.fasta

Supplementary file 2. Targets.92.fasta

Supplementary file 3. Targets.600.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-products-derived, as polyketides and non-ribosomal peptides being one of the most prolific classes. They are synthesized by large multimodular enzymes called polyketide synthases (PKSs) and nonribosomal peptide synthetases (NRPSs), respectively [1]. In microbial genomes, PKS and NRPS genes are often co-localized with all other genes required for the production of a given compound in a biosynthetic gene cluster. With the increasing ease of DNA sequencing, hundreds of novel PKS and NRPS gene clusters have been identified in recent years [2, 3]. However, accessing novel and useful PKS activities remains complicated, as many host organisms are uncultivable or do not express PKS genes in lab conditions. Furthermore, a significant challenge in the natural product discovery field is how to identify useful biosynthetic gene clusters, without a-priori knowledge of the biological target of the produced compound.

To identify useful PKS genes, we developed an automated method to identify and catalog clusters that harbor potential self-resistance genes. These genes provide the cluster-harboring microorganisms a defense against the antibiotics encoded by the clusters. For example, the genes encoding the fatty acid synthase inhibitor thiotetronic acid are clustered with a resistant copy of the fatty acid synthase gene (fabB/F) [11].

With this approach, we generated a non-redundant catalog of 152 PKS clusters. Manually curated lists of known antibacterial target genes were used to mine all NCBI nucleotide and genome databases. 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 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 putative novel targets and encoding compounds with new mechanisms of action.

**Introduction**

* Polyketides, antibacterial value
* Advent in dna sequencing, number of modular clusters (2013 PKS paper), and typeII clusters (2016 PNAS paper)
* Mining self-resistant genes to discover clusters with potential antibacterial

properties – examples, narrow focus on fatty acid inhibitors, actinomycetes, etc.

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.

* ARTS – great tool for identifying novel targets based on BCG proximity, duplications, and HGT. However, focused on Actinobactera species

We have developed an automated method to identify and catalog clusters that harbor a potential antibacterial target protein, generating a non-redundant catalog of 152 PKS clusters. Manually curated list of known antibacterial target genes was used to mine the NCBI database. The algorithm takes into account the distance of the potential target to a core enzyme in the biosynthetic gene cluster and the presence of a duplicated housekeeping copy of the target gene. This 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. 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.

Other core enzymes:

DMATS (alkaloid) Trichodiene synthase (terpene) GGPPS (terpene)

*“The discovery and development of several classes of safe and efficacious antibiotics has markedly reduced mortality from bacterial infections. However, the overuse and misuse of these same antibiotics—both in medicine and in agriculture—has driven the rapid evolution and dissemination of antibiotic resistance1. The ‘ESKAPE’ pathogens (Enterococcus faecium, Staphylococcus aureus, Klebsiella pneumoniae, Acinetobacter baumannii, Pseudomonas aeruginosa and Enterobacter species)2 present the most acute threat of developing untreatable multidrug-resistant (MDR) infections, with the Gram-negative members of this group (Escherichia coli, K. pneumoniae, P. aeruginosa and A. baumannii) posing a particular threat owing to their dual- membrane envelope that prevents many antibiotics from accessing their targets. Despite considerable effort, no new class of antibiotic has been approved for Gram-negative pathogens in over fifty years. “*

**Positive set of 12 known self-resistant targets in 18 biosynthetic gene clusters (Table 1)**

Manual literature search and automated mining of Mibig to identify all known self-resistant targets and their corresponding gene clusters

What is MiBig? 1800 clusters as of [date]

Pks, nrps, ripp, etc

12 targets curated

1% of pks pks/nrps clusters have a self-resistance gene colocalizrd with the biosynthetic genes (0.88% abundance rate of all types)

**Catalog of 152 non-redundant t1pks clusters harboring an antibacterial target gene**

This is a discovery rate of 252/29987= 0.84%, which is in a good agreement from the discovery rate from known clusters in Mibg (8/737 = 1%).

Target abundance – acc and ile-trna the most, some false positives

Others, not so many, perhaps they are no pks classes harboring such targets, but tharget specific to a certain class of natural product (lke RiPP)

**Second copy**

Most(?) positive examples have a self-resistant copy and a housekeeping copy. But there is at least one example, in which only the self-resistant copy is preserved.

Second copy is not required, since bacteria need only one resistant copy to survive from evolutionary standpoint

No reference number of copies per genomes

Not so many complete genomes, but many >4M bp

Could use ARTS for reference number, but it would be specific to Actinobacteria

There are examples in which copy number of genes for the same species but different isolates differs.

29 complete genomes, 14 of which have at least 2 copies of the target gene   
43 have at least 2 copies of the target gene

**Phylogeny**

Most are Actinobacteria (2/3rds)

Cluster by target

Many acc homologs, ile-trna synthetase homologs

No gyr homologs?

Place all positive examples on a tree – a roadmap - looking for analogs or thiolactomycin or orphan clusters?

Ile clade: Pos cluster in another clade, this one is an orphan clade. analyses. GyrB in that clade, but probably false positive

Middle of tree – at least two orphan clades

Use presence of homologs to access if cluster is real and if target is part of cluster

Use presence of homologs in distant species to suggest HGT

**Coevolution**

Co-evolution analyses used in TypeII catalog paper used to detect gene swap events. Here, we used the same approach of quantifying KS-target co-evolution by comparing pairwise protein identity scores between pairs of homologs. On-diagonal points suggest that those KS-target pairs co-evolved. We first build a co-evolution plot for pairs from the positive set (beta protesome inhibitors: salinosporamide, cinnabaramide, clarexpoxcin, eponemycin; and Isoleucyl tRNA synthetase inhibitors: mupirocin and thiomarinol)

Scoring scheme: abs(x-y)?

Thiolactomycin analogs on diagonal

**Ranking clusters**

Scoring of clusters based on the following criteria:

distance,

phylogeny (orphan clade),

homologs?,

housekeeping copy,

coevolution score,

ARTS

Target scored by randomly expected vs observed occurrence rate

**Mining of interesting antibacterial targets**

a) Reviews – 92 targets – manually curate all genes associated with the mentioned targets

b) Mining Mibig for putative targets

Highlight a few interesting examples with high score and interesting architecture

**Mining of novel antibacterial targets**

gvandova@maguro:/mnt/gnpn/gnpn/projects/orphanpks/TargetMining/Antismash\_gbids/bak\_2019$ cat out.609.filtered.5kb|cut -d"|" -f2|sort -u |wc

237 237 12365

3238 clusters harbor an E. coli essential genes. 237 out of the 609 essential E. coli genes were found within 5kb of a KS gene.

E. coli essential genes : targets for gram negative bacteria

targeting the cell wall in gram positive and negative bacteria:

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5039524/>

Metazoan and Fungal sequences

<https://www.ncbi.nlm.nih.gov/pubmed/17207587>

# A novel group of type I polyketide synthases (PKS) in animals and the complex phylogenomics of PKSs.

To our knowledge, no assembly-line PKS has been functionally characterized in eukaryotes. However, assembly-line PKSs are not thought to exist in metazoans. We were therefore surprised that the above analysis revealed the existence of an orphan clade that spanned a range of nematode species.

Parakeets PKS – pigment, but other functions in ther etazoans

<https://www.sciencedirect.com/science/article/pii/S0092867417309418#mmc2>

**NW\_017609824**|sp\_P0AAI3\_FTSH\_ECOLI|351430|351854|344264|349693|cluster-1|t1pks|331423-390511|1737 **NW\_017609824**\_sp\_P0AAI3\_FTSH\_ECOLI\_t1pks\_Larimichthys crocea (large yellow croaker)

**FQ310508**|sp\_P0AAI3\_FTSH\_ECOLI|13090530|13090953|13099159|13105634|cluster-1|t1pks|13059378-13117256|8206 **FQ310508**\_sp\_P0AAI3\_FTSH\_ECOLI\_t1pks\_Dicentrarchus labrax (European seabass)

**NW\_011337187**|sp\_P0AAI3\_FTSH\_ECOLI|493661|494021|484264|491318|cluster-1|t1pks|481956-532900|2343 **NW\_011337187**\_sp\_P0AAI3\_FTSH\_ECOLI\_t1pks\_Notothenia coriiceps (black rockcod)

**CP026263**|sp\_P0AAI3\_FTSH\_ECOLI|12324637|12325061|12332900|12338728|cluster-1|t1pks|12295860-12351358|7839 **CP026263**\_sp\_P0AAI3\_FTSH\_ECOLI\_t1pks\_Scophthalmus maximus (turbot)

**AE014134**|sp\_P0A8L1\_SYS\_ECOLI|3045730|3046137|3035645|3037150|cluster-1|t1pks|3025712-3087502|8580 **AE014134**\_sp\_P0A8L1\_SYS\_ECOLI\_t1pks\_Drosophila melanogaster (fruit fly)

<file:///Users/gvandova/TargetMining/Antismash_gbids/antismash_output/NW_017609824/index.html#cluster-1>

<file:///Users/gvandova/TargetMining/Antismash_gbids/antismash_output/NW_011337187/index.html#cluster-1>

<file:///Users/gvandova/TargetMining/Antismash_gbids/antismash_output/CP026263/index.html#cluster-1>

<file:///Users/gvandova/TargetMining/Antismash_gbids/antismash_output/FQ310508/index.html#cluster-1>

https://fungismash.secondarymetabolites.org/upload/fungi-f03b84e4-fee7-4975-89cd-c730ad8f2b70/index.html#r1c1

NW\_017609824|sp\_P0AAI3\_FTSH\_ECOLI|351430|351854|344264|349693|cluster-1|t1pks|331423-390511|1737 **Metazoa**

FQ310508|sp\_P0AAI3\_FTSH\_ECOLI|13090530|13090953|13099159|13105634|cluster-1|t1pks|13059378-13117256|8206 **Metazoa**

NW\_011337187|sp\_P0AAI3\_FTSH\_ECOLI|493661|494021|484264|491318|cluster-1|t1pks|481956-532900|2343 **Metazoa**

CP026263|sp\_P0AAI3\_FTSH\_ECOLI|12324637|12325061|12332900|12338728|cluster-1|t1pks|12295860-12351358|7839 **Metazoa**

AE014134|sp\_P0A8L1\_SYS\_ECOLI|3045730|3046137|3035645|3037150|cluster-1|t1pks|3025712-3087502|8580 **Metazoa**

HK492615|beta\_lactamase|4652|5076|12706|13566|cluster-1|t1pks|1-15108|7630 **None**

HK492619|beta\_lactamase|5033|5333|13870|14730|cluster-1|t1pks|1-16272|8537 **None**

GV586860|sp\_P0A8L1\_SYS\_ECOLI|2793125|2793477|2795949|2797232|cluster-15|t1pks-butyrolactone-otherks|2746969-2811734|2472 **None**

HK492617|beta\_lactamase|4658|5094|14020|14880|cluster-1|t1pks|1-16422|8926 **None**

PFVX01000026|sp\_P0AGJ9\_SYY\_ECOLI|46189|46613|39712|40980|cluster-1|t3pks-t1pks|1-52732|5209 **None**

KU234530|mpaF\_IMDH|12042|12332|4345|6052|cluster-1|t1pks|1-23325|5990 **Fungi**

BAUL01000113|sp\_P0CE47\_EFTU1\_ECOLI|13762|14187|22558|24113|cluster-1|t1pks|1-36426|8371 **Fungi**

JNNS01000353|mpaF\_IMDH|106535|106710|98706|99887|cluster-1|t1pks|87760-122045|6648 **Fungi**

HG792019|mpaF\_IMDH|1307601|1307776|1299771|1300952|cluster-3|t1pks|1288742-1333376|6649 **Fungi**

HG529702|sp\_P21889\_SYD\_ECOLI|41809|42243|37119|39452|cluster-1|t1pks|37899-83970|2357 **Fungi**

NW\_020194399|fpaI\_methionine\_aminopeptidase|3172242|3172554|3177711|3178235|cluster-9|t1pks-nrps|3140930-3192111|5157 **Fungi**

NW\_019154045|sp\_P06986\_HIS8\_ECOLI|66270|66703|58414|59800|cluster-1|t1pks-nrps|1-92737|6470 **Fungi**

KV875099|sp\_P18335\_ARGD\_ECOLI|1965330|1965666|1974928|1975860|cluster-1|t1pks|1942840-1985323|9262 **Fungi**

**Targets which are rare**

**Materials and Methods**

**Curating of t1pks gene clusters harboring a putative self-resistant gene**

*Step 1. Blast search for KS homologs. Blast 8 diverse KS against 11 NCBI nucleotide databases.*

All ncbi nucleotide and genome databases were searched for KS homologs using tblastn. The tblastn algorithm searches a protein query against nucleotide databases by translating the nucleotide records into all 6 possible open reading frames. This allows search into unannotated databases, such as Assembly database. 8 diverse KS from modular type1 pks (erythromycin), cisat pks/nrps (curacin, epothilone, guadinomine, rapamycin) and transat pks/nrps hybrids (leinamycin, disorazol, chivosazol) polyketide classes were used as query sequences against the major ncbi nucleotide and genome databases (nt, wgs, refseq\_genomic, other\_genomic, env\_nt, patnt, htgs, tsa\_nt, sts, gss, est\_others (all updated April 25, 2018, except for wgs, Jan-27-2016). Since we were unable to update wgs, we downloaded all Bacterial Assemblies database (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4702866/>). As of Oct-26-2018, there are 172 642 Bacterial Assemblies.

An initial relaxed blast search (e value < 1) identified non-redundant 199,894 protein records, of which 110,174 unique NCBI nucleotide records/genomes (<99% similar).

*Step 2. Run Antismash on all 100174 genbank files (89k genbank ids and 21k assembly ids).*

Download all 89449 genbank files from the traditional databases. The genbank files from the Assembly database (21080 total) are already stored locally. Antismash 4 with the minimal functionalities was run on all 100k sequences. While running antismash on assembly ids, the server was running out of memory because (1) genbank files were very big and (2) Antismash is also parallel. To avoid breaking the server, we split the assembly genbank files into smaller ones by extracting sequence 150kb upstream and downstream of a KS, resulting in a total of 26575 assembly files, and then ran antismash with 5 parallel processes on all sequences. Total run time for all 110k sequences was 95h.

Antismash ran on 79053 genbank files total (52480 traditional ncbi gebank files and 26573 assembly files). The rest 21121 genbank files were below 1000b and antismash does not run on shorter than 1000bp sequences.

*Step 3: Extract KS and gene sequences from all PKS-labeled clusters found by antismash and make a database.*

Otut of the 79k gbids, there are 29987 clusters annotated as type 1 pks (this includes cisat pks, transat pks, pks/nrps hybrids, and hybrids of them). We extracted all KS sequences, as well as all genes from all t1pks-annotated clusters and we made two databases: there are 244 196 sequences from nucleotide records in NCBI Bacterial assemblies database and 664,336 sequences from nucleotide records in the traditional NCBI databases. 78 clusters don’t have predicted KS domains (details\_data empty).

*Step 4. Blast search for a self-resistant target gene in a PKS cluster*

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

*Step 5. Require both KS and target to be less than 10kb apart*

In most experimentally characterized clusters harboring a self-resistant target gene (Table 1), the target gene was within a 10kb-distance of a core KS gene, with the exception of FabB/F in platencimycin (13kb away) and threonyl-tRNA synthetase in borrelidin (14.5kb). We chose an initial maximum allowed distance between a KS and a putative self-resistant gene as 10kb. This filter reduced the set to 252 KS + target proteins, in 250 unique nucleotide records.

[X gbids had a target but its partner KS is not annotated KS by antismash, skipped]

[Move to Results and Discusion]. This is a discovery rate of 252/29987= 0.84%, which is in a good agreement from the discovery rate from known clusters in Mibg (0.44%). The higher rate might indicate that the pks class of natural product is rich in exmaples of self-resistant target genes co-localized with the biosynthetic genes.

**Phylogenetic analyses**

To remove redundant gene clusters, we selected a standard redundancy threshold of 90% KS sequence identity, which further reduced the number of pks gene clusters to 152. Multiple sequence alignment 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 and colored either by phyla or self-resistant gene.

**Housekeeping copy**

To count the number of copies per genbank record, the 12 self-resistant targets were used as queries in blastp search of all genbanks records harboring clusters with these 12 targets using e-vaule cutoff of of 1e-8 and identity threshold of 0.3 (and FabB/F identity of 0.6). The number of copies per nucleotide record were counted and the status of the genbank record – complete genome vs incomplete was noted.

*There are some empty files because there is no sequence in genbank file. Blast manually or ignore???*

[Move to Results and Discusion?]. 43 out of the 252 clusters harbor at least two copies of the self-resistant gene, one of which is colocalized with the biosynthetic gene cluster. Out of the 252 nucleotide records, there are 29 complete genomes, 14 of which harbor at least 2 copies of the self-resistant gene.

[Move to Results and Discusion?]. **Discovery of PKS clusters from the positive set:**

There are 15 clusters harboring a self-resistant gene. Of them, 8 are PKS or PKS/NRPS hybrids. 7 out of 8 were identified in the pipeline (andrimid was mis-annotated as arylpolyene-nrps, and thus it was not extracted from the antismash results). 6/7 were identified using a KS-target tandem distance cutoff of 10kb, and 7/7 we identified using a KS-target tandem distance cutoff of 20kb (increasing the cutoff to 20% included the discovery of borrelidin, which thr-tRNA self-resistant target is 12kb form a KS gene).

At least 4 clusters of the positive set harbor a second housekeeping copy and we identified 3 out of 4 (the Isoleucyl tRNA synthetase housekeeping copy of the mupirocin producer has lower than 30% sequence identity with the self-resistant copy nd thus was not identified by our pipeline (23%).

The other 3 clusters from the positive set are on incomplete genomes (the nucleotide record depositied in genbank harbors only the biosynthetic gene cluster and thus it is not clear if there is another housekeeping copy). We also found only one copy per nucleotide record.

**Coevolution**

We plotted KS1-KS2 pairwise amino acid identities vs. pairwise identities of the corresponding target within 5, 10, or 20kb of a core KS gene. We fist plotted pairwise identities from positive set, including clusters with putative self-resistance gene (clarexpoxin, eponemycin, cinnabaramide: target co-localized with cluster, compound known to inhibit target homolog). We also calculated a coevolution score for each biosynthetic gene cluster:

Score = sum of all distances to diagonal abs(x-y) divided by the number of pairs this cluster occurs

**[Results and Discussion Coevolution]**

The lower the coevolution score, the more partners this cluster has on the diagonal and thus the higher the chance that the target coevolved with the cluster and is thus the self-resistant gene. Clusters from positive dataset are exclusively on diagonal.

Out of the 92 (have at least 2 clusters harboring the same target)

Targets exclusively on diagonal (Good targets)

Dna pol II epsilon subunit

degP

dihydrohynaphotil coa synthase

amnodeoxychorismate synthase

FtsH

Alanine racemase

n-acetylglucosamine deacetylase

FabG

Ser trna ligase

Aminodeoxychorismate ynthase component 1

From E. coli

TorCAD

Octulosonate

ygaP

deoxy d xylyose

L19

Phenylacetic degradation protein

Hypothetical protein yfch

Coevolution of positive set : Add all putative targets (target co-localized with cluster, compound known to inhibit target homolog)

1. Make blast database from KS and target sequences

makeblastdb -in KS.12.10kb.fasta -dbtype prot -out KS.12.10kb

makeblastdb -in targets.12.10kb.fasta -dbtype prot -out targets.12.10kb

2. Pairwise blast of all KSs and pairwise blast of all targets

blastp -db blastdb/KS.12.10kb -query KS.12.10kb.fasta -outfmt "6 qseqid sseqid sstart send nident qlen slen evalue" -evalue 1 -out KS.12.10kb.out

blastp -db blastdb/targets.12.10kb -query targets.12.10kb.fasta -outfmt "6 qseqid sseqid sstart send nident qlen slen evalue" -evalue 1 -out targets.12.10kb.out

3. Parse blast hits and output a file with KS and corresponding target pairwise identities

AM889285-KV411304 0.40 0.39

!!! Some low KS-KS seq identity because antismash couldn't recognize the domains, just the active sites. Or there was a CLF domain miss annotated as KS domain. Other reasons?

4. Plot pairwise identities.

Whole Genome Shotgun (WGS) projects are genome assemblies of incomplete genomes or incomplete chromosomes of prokaryotes or eukaryotes that are generally being sequenced by a whole genome shotgun strategy. WGS projects may be annotated, but annotation is not required. The NCBI Assembly databas

) provides stable accessioning and data tracking for genome assembly data. The Assembly database stores the names and identifiers for the sequences in each genome assembly and records the organization of the component sequences into scaffolds and chromosomes.Likewise, the NCBI Reference Sequence (RefSeq) database ([2](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4702866/#B2); [www.ncbi.nlm.nih.gov/refseq/](http://www.ncbi.nlm.nih.gov/refseq/" \t "_blank)) issues identifiers only to single genomic sequences selected from INSDC for inclusion in its non-redundant sequence collection. INSDC Whole Genome Shotgun (WGS) project identifiers cannot adequately distinguish between different versions of an assembly because the project version does not always change when the number of sequences in an assembly is increased or decreased ([www.ncbi.nlm.nih.gov/genbank/wgs/](http://www.ncbi.nlm.nih.gov/genbank/wgs/" \t "_blank)).