**Title.** Self-resistant target mining of polyketide natural product gene

**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.

**Figures and Tables:**

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

Figure 1. Pipeline (12 known targets)

Supplementary table 1. Catalog of 152 clusters harboring a potential self-resistant gene target

Columns: cluster, cluster type, target, **distance**, **phylogeny**, **homologs?, housekeeping copy, coevolution score, ARTS score?**

Table 2. Number of pks clusters harboring a putative self-resistant gene 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 and high score

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

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. Mining of Mibig and identifying clusters harboring a potential self-resistant target gene (target of molecule known, target copy in clusrter not tested if it is the resistant)

Supplementary table 2. Catalog of 501 clusters harboring a potential self-resistant gene target (92 targets)

Columns: cluster, cluster type, target, **distance**, **phylogeny**, **homologs?, housekeeping copy, coevolution score, ARTS**

**Sorted by highest ranked ones**

Supplementary table 3. Catalog of X clusters harboring a potential self-resistant gene target (600 E. coli essential genes)

Columns: cluster, cluster type, target, **distance**, **phylogeny**, **homologs?, housekeeping copy, coevolution score, ARTS**

**Sorted by highest ranked ones**

Supplementary file 1. Targets.12.fasta

Supplementary file 2. Targets.92.fasta

Supplementary file 3. Targets.600.fasta

**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

0.88% abundance rate of all types

0.44% pks pks/nrps

**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 (0.44%). The higher rate might indicate that the pks class of natural product is rich in exmaples of self-resistant target genes colocalized with the biosynthetic genes.

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.

**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**

**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 – 54 targets – manually curate all genes associated with the mentioned targets

b) Mining Mibig for putative targets

Highlight a few interesting examples with hig score and interesting architecture

c) targets for gram negative bacteria

targeting the cell wall in gram positive and negative bacteria:

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

**Materials and Methods**

**STEP I. Curation of t1pks gene clusters**

Step1. Blast search for KS homologs. Blast 8 diverse KS against 11 NCBI nucleotide databases. Downloading traditional databases (10) and the NCBI Assembly database. An initial relaxed blast search (e value < 1) identified non-redundant 199 894 protein records, of which 110174 unique NCBI nucleotide records/genomes (<99% similar).

Step 2: Fetch all genbank ids that contain a core KS gene. Download all genbank files from the traditional databases. The genbank files from the Assembly database are already stored locally.

Genbank/gbdir (89449 gbids)

Genbank/assembly\_gb (21080 gbids)

STEP3. Run Antismash on all 100174 genbank files (89k genbank ids and 21k assembly ids). Some assembly gb files were split into smaller files, harboring only a potential pks cluster (26575 assembly files total). Antismash 4 with the minimal functionalities was run and total run time was 95h.

STEP4. Extract gene sequences and KS sequences from all pks-labeled clusters found by antismash. There are 29987 clusters annotated as t1pks or its derivatives in antismash.

78 clusters don’t have predicted KS domains (details\_data empty)

STEP5. Make blast database from all KS cluster genes.

STEP6. Blast 12 targets genes against antismash database from 110k genbank ids

4404 hits.

STEP7. Filter blast hits by evalue <1e-8 and identity > 0.3 (and FabB/F identity > 0.6)

806 hits.

STEP8. Require both KS+ target <10kb apart. Extract KS and target sequences.

252 hits. 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 colocalized with the biosynthetic genes.

**STEP II. Phylogeny**

Step 9. Remove redundant KS sequences (>90% identical)

Step 10. Build a MSA with mafft

Step 11. Build a phylogenetic tree with FastTree

Step 12. Visualize tree

Visualize tree by target?

Upload to mibig?(Talk to Marnix)

Visualize Coevolution plot by target

Make tanglegram?

Look at a few off-diagonal points and figure out if they are true negatives

Coevolution with >50kb cutoff

COEVOLUTION ANALYSES

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. Other reasons?

4. Plot pairwise identities. Points on diagonal

Coevolution of positive set

Add all putative targets (target co-localized with cluster, compound known to inhibit target homolog)

Take all antismash KS-target paris, >20kb away from each other?

Evolution of Type II PKS Accessory Enzymes. The origin and diversification of type II PKS enzymes outside of the KS and CLF

are not well understood. To study cluster-wide evolution, we first

identified classes of accessory genes frequently clustered within

30 kb of the KS-CLF gene pair (Fig. 2, Table 1, and SI Appendix,

Fig. S4). We developed a method to detect gene swap events,

building upon existing approaches (19, 20) to quantify gene pair

coevolution by comparing protein similarity scores between pairs

of homologs. Correlated similarity scores suggest that gene types

coevolved (Fig. 3 A and B). Our results confirmed that the core

KS and CLF coevolved with little to no gene swaps: when two

KSs from different genomes have high similarity, the neighboring

CLFs also have high similarity, and when the KSs have low similarity, the neighboring CLFs also have low similarity (Fig. 3C and

SI Appendix, Fig. S7A). We applied this framework to detect

coevolution of tailoring genes with the core KS, and extended it

to detect discrete homologous gene swap events as off-diagonal

groups (Fig. 3B).

Step1. Blast search for KS homologs. All ncbi nucleotide and genome databases were searched for KS homologs using tblastn. 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 (not updated), refseq\_genomic, other\_genomic, env\_nt, patnt, htgs, tsa\_nt, sts, gss, est\_others).

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 database 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)).

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4702866/

An initial relaxed blast search (e value < 1) identified 162984 protein records, of which 94516 unique NCBI nucleotide records/genomes.

A comprehensive set of clusters with experimentally verified and putative targets.

1. Putative positive gene cluster data set

2. Experimentally verified gene clusters data set – 9 unique targets, 12 clusters