**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