Target-guided genome mining of natural product gene clusters and heterologous expression in *E. coli*

Introduction. 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, there are several limitations to accessing the encoded compounds: (1) the natural producers of these compounds are difficult or cannot at all be cultured in lab conditions; (2) the genes that encode the compounds are silent; (3) the structural complexity of many PKS and NRPS compounds hampers efforts of total chemical synthesis.

Heterologous expression of natural product genes in model hosts has been applied to access novel natural products including polyketides [4, 5]. *Escherichia coli* is an attractive host for many reasons: it is easy to culture with a well-developed genetic toolbox, the primary metabolism is well understood, and because it is not an endogenous producer of polyketides, potential interference of native proteins with heterologously expressed PKS pathways may be limited [6]. Despite these advantages, attempts at heterologous production of polyketides in *E. coli* have met with limited success [7,8], which have limited the utility of *E. coli* as a host for production of assembly-line polyketides, and have spurred efforts to improve general characteristics of this host to produce this class of compounds [8, 9].

Another significant challenge in the natural product discovery field is how to prioritize which biosynthetic gene cluster to express, without a-priori knowledge of the biological target of the produced compound. 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]. 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].

Here I, address both of these challenges by target-directed genome mining of natural product gene clusters and heterologous expression in *E. coli.*

Computational approach to selecting clusters for heterologous expression in *E. coli.* I have developed an automated method to identify and catalog clusters that harbor a potential antibacterial target protein, generating a non-redundant catalog of X PKSs, Y NRPSs, and Z PKS/NRPS hybrids. 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)

Figure 1. Pipeline

Figure S1. HMM models?

Figure 2. Phylogenetic tree of KS and C domains of PKS and NRPSs

Figure 3. Target and KS coevolution

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 199 894 protein records, of which 110174 unique NCBI nucleotide records/genomes.

Step 1: BLAST search for KS homologs.

Using 8 diverse KS query sequences, e-value < 1

199 894 *proteins;* 110174 *NCBI nucleotide records/genomes*

Step 2: Fetch all genbank ids that contain a KS

Genbank/gbdir (89449 gbids)

Genbank/assembly\_gb (21080 gbids)

STEP3. Run Antismash on all 89k genbank ids and 21k assembly ids

Some assembly gb files split into smaller files (26575 assembly files total)

STEP4. Extract gene sequences and KS sequences from all pks clusters found by antismash

cluster\_genes.21k.fasta

KS.21k.fasta

cluster\_genes.89k.fasta

KS.89k.fasta

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

STEP5. Make blast database from cluster\_genes.21k.fasta and cluster\_genes.89k.fasta

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.

**>**AdmT\_ACC 64

**>**SalI\_beta\_proteasome 18

**>**GriR\_DnaN 3

**>**EF-Tu 48

**>**PtmP3\_FabB-F 13

**>**BatG\_FabI 5

**>**GyrB-R 22

**>**mupM\_Ile-tRNA-syn 53

**>**borI\_Thr-tRNA-syn 7

**>**agnB2\_Leu-tRNA-syn 6

**>**rubR1\_TIF 8

**>**Ind0\_Trp-tRNA-syn 5

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

HG964936 **Mycobacterium asiaticum DSM 44297**

# LQPQ01000178 Mycobacterium riyadhense strain DSM 45176

# LZLF01000227

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?

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