Resistance-guided mining of bacterial genotoxins defines a family of DNA glycosylases

Noah P. Bradley[1,4], Katherine L. Wahl[1,4], Jacob L. Steenwyk[1], Antonis Rokas[1,2], Brandt F. Eichman[1,3,*]

[1] Department of Biological Sciences, Vanderbilt University, Nashville, TN 37232

[2] Department of Biomedical Informatics, Vanderbilt University School of Medicine, Nashville, TN 37232

[3] Department of Biochemistry, Vanderbilt University School of Medicine, Nashville, TN 37232

[4] These authors contributed equally

Corresponding Author: Brandt F. Eichman; Box 351634 Station B, Nashville, TN 37235; (615) 936-5233; brandt.eichman@vanderbilt.edu

Keywords: AlkZ / natural product / biosynthetic gene cluster / DNA repair / DNA glycosylase / secondary metabolism / phylogenetic tree / self-resistance / genotoxin

Abbreviations: AZB, azinomycin B; AZL, AlkZ-like; HED, hedamycin; BER, base excision repair; BGC, biosynthetic gene cluster; GO, gene ontology; ICL, interstrand crosslink; NM, nitrogen mustard; TXNA, trioxacarcin A; YQL, YcaQ-like

Abstract

Unique DNA repair enzymes that provide self-resistance against therapeutically important, genotoxic natural products have been discovered in bacterial biosynthetic gene clusters (BGCs). Among these, the DNA glycosylase AlkZ is essential for azinomycin B production and belongs to the HTH_42 superfamily of uncharacterized proteins. Despite their widespread existence in antibiotic producers and pathogens, the roles of these proteins in production of other natural products are unknown. Here, we determine the evolutionary relationship and genomic distribution of all HTH_42 proteins from *Streptomyces* and use a resistance-based genome mining approach to identify homologs associated with known and uncharacterized BGCs. We find that AlkZ-like (AZL) proteins constitute one distinct HTH_42 subfamily and are highly enriched in BGCs and variable in sequence, suggesting each has evolved to protect against a specific secondary metabolite. As a validation of the approach, we show that the AZL protein, HedH4, associated with biosynthesis of the alkylating agent hedamycin excises hedamycin-DNA adducts with exquisite specificity and provides resistance to the natural product in cells. We also identify a second, phylogenetically and functionally distinct subfamily whose proteins are never associated with BGCs, are highly conserved with respect to sequence and genomic neighborhood, and repair DNA lesions not associated with a particular natural product. This work delineates two related families of DNA repair enzymes—one specific for complex alkyl-DNA lesions and involved in self-resistance to antimicrobials, and the other likely involved in protection against an array of genotoxins—and provides a framework for targeted discovery of new genotoxic compounds with therapeutic potential.

Significance Statement

Bacteria are rich sources of secondary metabolites that include DNA-damaging genotoxins with antitumor/antibiotic properties. Although *Streptomyces* produce a diverse number of therapeutic genotoxins, efforts toward targeted discovery of biosynthetic gene clusters (BGCs) producing DNA-damaging agents is lacking. Moreover, work on toxin-resistance genes has lagged behind our understanding of those involved in natural product synthesis. Here, we identified over 70 uncharacterized BGCs producing potentially novel genotoxins through resistance-based genome mining using the azinomycin B-resistance DNA glycosylase AlkZ. We validate our analysis by characterizing the enzymatic activity and cellular resistance of one AlkZ ortholog in the BGC of hedamycin, a potent DNA alkylating agent. Moreover, we uncover a second, phylogenetically distinct family of proteins related to *E. coli* YcaQ, a DNA glycosylase capable of

unhooking interstrand DNA crosslinks, that differ from the AlkZ-like family in sequence, genomic location, proximity to BGCs, and substrate specificity. This work defines two families of DNA glycosylase for specialized repair of complex genotoxic natural products and generalized repair of a broad range of alkyl-DNA adducts, and provides a framework for targeted discovery of new compounds with therapeutic potential.

Introduction

Bacteria are exceptionally rich sources of secondary metabolites, which are important for their survival and often have therapeutic value. *Streptomyces* produce 35% of all known microbial natural products and nearly 70% of all commercially useful antibiotics, with several being FDA-approved antitumor agents used as first-line cancer treatments (1-4). Secondary metabolites are often toxins used in ecological interactions with other organisms and can target any number of critical cellular functions (5). Natural products that damage DNA (genotoxins) form covalent or non-covalent DNA adducts that can inhibit replication and transcription, thus undermining genomic integrity through mutagenesis or cell death (6, 7). Consequently, genotoxins are particularly useful antineoplastic agents, as exemplified by several clinically relevant drugs including doxorubicin, bleomycin, mitomycin C, and duocarmycin analogs (8).

Streptomyces produce a wide variety of DNA alkylating and oxidizing agents that have antimicrobial and antitumor properties. Spirocyclopropylcyclohexadienones (duocarmycin A and SA, yatakemycin, CC-1065) (9, 10), pluramycins (pluramycin A, hedamycin, altromycin) (11-13), anthracycline glycosides (trioxacarcin A, LL-D49194α1) (14-16), and the leinamycin family (17) contain a single reactive group that covalently modifies purine nucleobases to form a broad spectrum of bulky alkyl-DNA monoadducts. *Streptomyces* also produce bifunctional alkylating agents that react with nucleobases on both DNA strands to create interstrand crosslinks (ICLs). Mitomycin C (MMC) from *S. lavendulae* crosslinks guanines at their N2 positions, and azinomycin A and B (AZA, AZB) from *S. sahachiroi* and *S. griseofuscus* crosslink purines at their N7 nitrogens (18). In addition to alkylating agents, several families of natural products, including bleomycins and enediynes, exert their toxicity by oxidative cleavage of DNA and RNA (19).

The production of secondary metabolites in *Streptomyces* is genetically organized into biosynthetic gene clusters (BGCs), which contain the genes necessary for their biosynthesis, export, regulation, and resistance. Resistance mechanisms protect antibiotic producers from toxicity of their own natural products, and include toxin sequestration, efflux, modification, destruction, and target repair/protection (20, 21). In the case of genotoxins, several DNA repair enzymes have been identified as target repair resistance mechanisms, including direct reversal of streptozotocin alkylation by AlkB and AGT (alkylguanine alkyltransferase) homologs (22), base excision of yatakemycin-adenine adducts by the DNA glycosylase YtkR2 (23, 24), nucleotide excision of DNA adducts of several intercalating agents including daunorubicin (25), and putative replication-coupled repair of distamycin-DNA adducts (26).

The AZB BGC in *Streptomyces sahachiroi* encodes a DNA glycosylase, AlkZ, that unhooks AZB-ICLs and that provides cellular resistance against AZB toxicity (27, 28). ICL unhooking by AlkZ involves hydrolysis of the N-glycosidic bonds of the crosslinked deoxyguanosine residues, producing abasic (AP) sites that can be repaired by the base excision repair pathway (29). AlkZ belongs to the relatively uncharacterized HTH_42 superfamily of proteins found in antibiotic-producing and pathogenic bacteria (28). The crystal structure of AlkZ revealed a unique C-shaped architecture formed by three tandem winged helix-turn-helix motifs, with two catalytically essential glutamine residues within a QΦQ motif (Φ is an aliphatic residue) located at the center of the concave surface (30). We recently characterized a second HTH_42 protein from *Escherichia coli*, YcaQ, as a DNA glycosylase that excises several types of *N7*-alkylguanine ICLs and monoadducts using a catalytic QΦD motif and that functions as a secondary pathway to nucleotide excision repair for bacterial resistance to the nitrogen mustard, mechlorethamine (31).

The targeted discovery of natural products has been employed to search for novel scaffolds in plants, fungi, and bacteria, and can be useful for identifying specific classes of compounds (32-34). Genome mining can be used to search for unidentified BGCs through analysis of core/accessory biosynthetic genes (PKS, NRPS, tailoring enzymes), comparative/phylogeny-based mining, regulatory genes, and more recently, resistance genes (35). Some of these resistance-based mining approaches focus on the experimental screening of antibiotic resistance, while others rely on bioinformatic tools to identify resistance genes within clusters based on homology to known resistance genes (36-39). However, many of these resistance-based methods have not been applied in bacteria for targeted discovery.

Here, we characterized the genomic differences of the HTH_42 proteins found in 435 species of *Streptomyces* to develop additional insight into this new family of DNA repair proteins, and applied this information in resistance-guided genome mining to characterize unknown BGCs or identify new genotoxins. We found that these proteins fall into two distinct subfamilies that are delineated by amino acid sequence, genomic context, and copy number. Proteins similar to *S. sahachiroi* AlkZ (AlkZ-like, AZL) are highly variable in sequence and enriched in BGCs, many producing known genotoxic alkylating agents. We show that the AZL protein within the BGC of the known DNA alkylating agent, hedamycin (HED), is a resistance DNA glycosylase specific for HED-guanine lesions, consistent with AZL-mediated DNA repair activity as a general self-resistance mechanism to genotoxins in antibiotic producers. Moreover, we found AZL proteins in BCGs that are either uncharacterized or that produce natural products not previously known to be genotoxic, validating resistance genome mining as an approach to discover new genotoxins. In contrast, *E. coli* YcaQ like (YQL) proteins are highly conserved in sequence and genetic neighborhood and are not associated with BGCs. We show that like *E. coli* YcaQ, two YQL enzymes from Actinobacteria have weaker substrate specificity than AZL proteins, suggesting a broader role of this subfamily of HTH_42 proteins outside of antibiotic self-resistance in bacteria.

Results

YQL and AZL proteins in Streptomyces are evolutionarily distinct

E. coli YcaQ and *S. sahachiroi* AlkZ are the only characterized members of the HTH_42 superfamily and are unique in their ability to unhook ICLs and to provide cellular resistance to crosslinking agents. Both enzymes fully unhook ICLs derived from AZB (Fig. 1A). While AlkZ is specific for AZB-ICLs and is essential to the AZB-producing organism, YcaQ unhooks a broader range of ICLs, including those derived from the simple bifunctional alkylating agent mechlorethamine (Fig. 1B), and displays robust excision activity for *N*7-methylguanine (7mG) monoadducts (28, 30, 31). YcaQ and AlkZ belong to one of five classes of HTH_42 proteins characterized by domain organization, which accounts for >95% of all HTH_42 proteins (Fig. S1A). Approximately two-thirds of the known HTH_42 proteins in prokaryotes are found in Actinobacteria, with ~25% of those sequences from *Streptomycetales* (Fig. S1B,C). The remainder are found in several different orders of Bacteria, and a very small number (12) in Archaea.

To better understand the evolutionary and phylogenetic breadth of this superfamily in *Streptomyces*, we collected and analyzed all HTH_42 protein sequences from available genomes using a combination of BLAST searches against *Streptomyces* genomes in GenBank and HHMR protein domain searches of the BLAST hits against the Pfam database (Table S1). Alignment of the 897 sequences showed that YQL and AZL proteins fall into distinct clades that represent 49% and 43% of the total number of sequences, respectively (Fig. 1C). The clades are defined in part by unique catalytic motifs QΦD (YQL) and (Q/H)ΦQ (AZL), where Φ is an aliphatic residue (30, 31). YQL proteins show a high degree (>75%) of amino acid sequence conservation, whereas the AZL subfamily is more diverse, with only ~40% amino acid similarity on average. The differences in conservation are consistent with mutation rates as approximated by tip-to-root branch lengths (0.23 for YQL and 0.59 for AZL). In addition, we found that 8% of sequences do not fall into either YQL or AZL clades and contain a unique catalytic consensus sequence, HΦ(S/T)(D/E) (Fig. 1C,D). Because these sequences exhibit greater sequence similarity overall to AZL than YQL, we refer to this third homolog AZL2. Interestingly, AZL2 is more similar to YQL in its copy number and genomic location (see below), and thus is somewhat of a hybrid between AZL and YQL. We verified that proteins within the AZL2 clade contain bona fide DNA glycosylase activity, as the *S. caeruleatus* AZL2 protein excised 7mG from DNA in a similar manner to *S. sahachiroi* AlkZ (Fig. S1E).

Another striking difference between the YQL and AZL families is that AZL genes are often found in multiple copies and in different combinations in many species of *Streptomyces*. The copy number differences between the different clades are significant, with the majority (90-95%) of YQL and AZL2 homologs found as a single copy and AZL mainly found in multiple (2-5) copies (Fig. 1E). The coincidence of YQL and AZL also varies. Although the most common combination is the presence of a copy of each YQL and AZL, many other combinations are observed (Fig. 1F). The number of species that contain both genes decreases as the copy number increases. For species containing either YQL or AZL (not both), the majority contain a single YQL copy, with just a few species having only AZL present. These results show that both YQL and AZL proteins are broadly distributed across *Streptomyces* and are distinct with respect to sequence, diversity, and copy number.

AZL proteins are prevalent in biosynthetic gene clusters

Given the distinct phylogeny of YQL and AZL proteins, we next examined their proximity to BGCs and characterized the identities of clusters containing a putative homolog. To perform this analysis, we identified all BGCs in the genomes of known *Streptomyces* species containing an HTH_42 protein, determined the most similar known cluster via BLAST, and extracted the distance in base pairs between the YQL/AZL gene and the nearest 3' or 5' end of each BGC (Fig. 2A, Table S2). Strikingly, none of the 442 YQL genes localize to within 20 kb of the most proximal gene cluster in that organism (Fig. 2B). In contrast, AZL genes are primarily found inside or in close genomic proximity to clusters, with an average distance of roughly 2.3 kb from the nearest BGC (compared to 25 kb for YQL). Despite their sequence similarity to AZLs, the AZL2 proteins are more like YQL in that they are also not observed within 20 kb of a BGC (Table S2).

We found that AZL proteins are particularly enriched in uncharacterized *Streptomyces* BGCs, with 68 homologs localizing within a variety of different types of clusters (Fig. 2C,D; Table S3). Almost half (n=32; 47%) localize to clusters resembling those producing known DNA damaging agents, including AZB (n=5), LL-D49194α1 (LLD, n=6), HED (n=4), ficellomycin/vazabitide A (n=5), and C-1027/leinamycin (n=2) (12, 16-18, 40, 41). In addition, several other clusters are related to potential DNA damaging agents on the basis of a reactive epoxide functional group in the natural product, including angucycline-like molecules (n=4) herboxidiene and asukamycin. The remaining 10 uncharacterized BGCs are related to clusters that produce macrolides/terpenes, tambromycin-like compounds, and various RiPPs/depsipeptides (Fig. 2C,D).

Bacterial genes of similar function or in a particular pathway are frequently clustered into neighborhoods or operons within the genome, and thus we investigated the nearest neighbors of *Streptomyces* YQL and AZL genes. We collected gene ontology (GO) terms describing the biological functions of the five nearest neighbors on either side of 40 YQL genes, 40 AZL genes inside BGCs, and 40 AZL genes outside BGCs, which collectively represent ~15% of the total of all homologs. Biological processes were grouped into three categories—metabolism, signaling/cell function, and genetic information processing. Several key differences were found between the neighborhoods of AZL genes inside versus outside clusters (Fig. 2E; Fig. S2). AZL genes within BGCs were more often found near terpenoid/polyketide/non-ribosomal protein synthesis and resistance/defense genes. The defense genes fell into several types: ABC transporters/permeases, α/β-fold hydrolases (VOC resistance proteins), DinB DNA-damage inducible hydrolases, and other AZL proteins. For those AZL genes found outside of BGCs, there are an abundance of neighbors involved in cell wall biosynthesis, cell cycle control, and signal transduction. In contrast, there were no significant differences between AZL neighbors involved in processing genetic information inside versus outside clusters (Fig. 2E). In contrast to the variation in the function of AZL gene neighbors, the functions of YQL neighbors (outside clusters) are nearly invariant, and are composed of a variety of different gene types with no apparent functional connection between them (Fig. 2F). The functions of many of these neighbors have not been elucidated in *Streptomyces*, but some are homologous to N-acetyltransferase, a two-component transcription factor/histidine kinase, and a DNA helicase (ComF) involved in transformation competence. Thus, both the sequences and the genomic neighborhoods of YQL proteins are relatively conserved and always found outside of BGCs, in contrast to the more variable copy number, sequence, and neighborhood of AZL genes prevalent within BGCs.

Characterized BGCs containing AZL proteins

With the discovery that a significant proportion of AZL proteins reside within BGCs, we took a closer look at the nine characterized BGCs identified to contain an AlkZ homolog in the MIBiG database (Table S3). Four of these produce known DNA-alkylating agents (Fig. 3A), which contain reactive epoxide moieties like AZB that are scaffolded on diverse natural product backbones (Fig. 3A). Whereas AZB is a bifunctional alkylating agent, HED, trioxacarcin A (TXNA), and LL-D49194α1 (LLD) are monofunctional alkylating agents that react with N7 of guanine in specific nucleotide sequences via their epoxide rings, and also intercalate the DNA helix via their planar ring systems (12, 42). TXNA and LLD clusters each contain two AlkZ paralogs (TxnU2/U4 and LldU1/U5), whereas the HED cluster contains one (HedH4) that resides between the two polyketide synthase genes.

The remaining five AZL-containing clusters in MIBiG produce compounds that are not known to alkylate DNA, but that share some structural characteristics with the alkylating agents described above (Fig. 3B). Aclacinomycin contains an anthracycline core surrounded by sugars that allow it to intercalate into DNA and act as a topoisomerase I poison, potentially generating downstream DNA damage (43). Asukamycin contains a modified PKS scaffold and an electrophilic epoxide ring and has been shown to act as both a farsenyltransferase inhibitor and a molecular glue between the UBR7 E3 ubiquitin ligase and the TP53 tumor suppressor, leading to cell death (44, 45). Armeniaspirol contains a unique chlorinated pyrrole and inhibits the AAA+ proteases ClpXP and ClpYQ leading to cell division arrest in Gram positive bacteria (46). The other two BGCs produce compounds of known structure but unknown function—tambromycin and JBIR-34/35 are similar NRPS compounds containing densely substituted chlorinated indole and methyloxazoline moieties (47). The presence of AZL proteins in these clusters suggests that these compounds may be genotoxins or otherwise react with DNA, and/or that these particular AZL homologs may have a function outside of DNA repair.

The AZL protein within the HED BGC is a DNA glycosylase specific for HED-DNA lesions and provides cellular resistance to HED toxicity

The *alkZ* gene embedded within the AZB BGC provides exquisite resistance to the potent cytotoxicity of this natural product (27, 28). To determine if AlkZ homologs other than in the AZB BGC provide self-resistance to their cognate natural products, we characterized the DNA glycosylase and cellular resistance activities of HedH4 for HED-DNA adducts. HED is a potent antibiotic/antitumor agent that induces a strong DNA damage response (48). The bisepoxide side chain alkylates the N7 position of guanines in 5'-(C/T)G sequences (Fig. 4A), the highly oxidized aromatic polyketide intercalates the DNA helix, and two C-glycosidic linked aminosugars interact with the minor groove (12). We generated site-specifically labeled HED-guanosine adducts in DNA by reacting purified compound with an oligonucleotide containing a HED target sequence, d(T<u>G</u>TA). The HED-DNA adduct was stable relative to other *N*7-alkylguanine lesions as judged by thermal depurination (Fig. S3B) (31, 49). We first assessed the ability of purified HedH4 to hydrolyze HED-DNA using a gel-based glycosylase assay that monitors alkaline cleavage of the AP site product (30, 31). Reaction of HedH4 with HED-DNA followed by hydroxide work-up resulted in β- and δ-elimination products consistent with production of an AP site from DNA glycosylase-mediated excision of the N-glycosidic bond of the HED-guanosine nucleotide (Fig. 4A,B). We verified the identity of the excision product as HED-guanine by HPLC/MS (Fig. 4C). To verify that the HED-guanine product was not generated by a contaminating enzyme and to examine the conservation of the catalytic QΦQ motif, we purified alanine mutants of the two glutamine residues and tested their activity under single-turnover conditions (Fig. 4D, Fig. S3A,C). The calculated rate constant (k_{cat}) for wild-type HedH4 was at least 7.8 ± 0.5 min^{-1} (the reaction was complete at the earliest time point). Relative to wild-type, the Q41A mutant was at least 225-fold slower (k_{cat} = 0.04 ± 0.01 min^{-1}) and the Q43A mutant at least 10-fold slower (k_{cat} = 0.8 ± 0.2 min^{-1}), indicating that both Gln residues in the HedH4 QΦQ play a role in HED-guanine excision.

We probed specificity of HedH4 for HED-DNA adducts, first by asking whether the HED-guanosine lesion was a substrate for other bacterial alkylpurine DNA glycosylases with varying specificities. *E. coli* AlkA and YcaQ and *Bacillus cereus* AlkC and AlkD excise a relatively broad range of alkyl-DNA adducts (31, 50-55). *S. sahachiroi* AlkZ, *S. bottropensis* TxnU2 and TxnU4, and *S. vinaceusdrappus* LldU1 and LldU5, like HedH4, are found in BGCs that produce bulky *N*7-alkyl- and intercalating DNA adducts (Fig. 3A), and each is specific for their cognate toxin (31, 56). Compared to HedH4, which excises 100% of the HED-guanine from DNA, none of the ten alkylpurine DNA glycosylases tested showed any appreciable activity for HED-DNA after 1 hour (Fig. 4E). Thus, the HED-DNA adduct is hydrolyzed only by the glycosylase found in the HED BGC. We next examined the ability of HedH4 to excise *N*7-alkylpurine lesions that act as substrates for other YQL and AZL enzymes. Interestingly, HedH4 showed no significant activity for the simple methyl adduct 7mG, which is removed by most alkylpurine DNA glycosylases including *E. coli* YcaQ and *S. sahachiroi* AlkZ (Fig. 4F). HedH4 was also unable to hydrolyze TXNA-guanosine, a substrate for TxnU4 from the TXNA BGC (Fig. 4F) (56). We also tested the ability of HedH4 to unhook ICLs derived from AZB (Fig. 1A) and an 8-atom nitrogen mustard, NM_8 (Fig. 4G), which are substrates for *S. sahachiroi* AlkZ and *E. coli* YcaQ, respectively. Compared to AlkZ and YcaQ, HedH4 showed little to no activity for either ICL. Thus, HedH4 is highly specific for DNA adducts derived from its cognate natural product.

We next tested if the *hedH4* gene provides heterologous resistance to HED cytotoxicity in cells. *E. coli* transformed with either vector containing *hedH4* constitutively expressed at low levels or vector alone were grown in the presence of increasing amounts of HED (Fig. S3D-G).

HedH4 provided a modest protection against HED, as cells expressing HedH4 grew to a higher density at all HED concentrations (Fig. 4I, Fig. S3F-G) and had a higher IC_{50} than cells treated with vector alone (HedH4, 5.9 μM ± 0.7; vector, 3.9 μM ± 0.4). The sensitivity differences between HedH4 and vector control were more pronounced from a colony dilution assay performed under log-phase growth conditions (Fig. 4J). Cells expressing empty vector displayed an IC_{50} value of 11.1 μM ± 1.5, while cells expressing HedH4 displayed a 4-fold reduction in sensitivity to HED (48.1 μM ± 13.8). These results indicate that HedH4 is a DNA glycosylase specific for HED-DNA adducts and provides resistance to cells exposed to the antibiotic.

YQL proteins from Actinobacteria hydrolyze simple N7-alkylguanosine lesions and interstrand crosslinks

We previously characterized *E. coli* YcaQ to have robust activity toward 7mG and NM-ICLs (Figs. 1B and 4G), a substrate preference distinct from AZB- and HED-specific *S. sahachiroi* AlkZ and HedH4 (Fig. 4F,H) (31). We therefore were interested in determining if other proteins of the YQL subfamily were functional YcaQ orthologs. We purified YQL proteins from the Actinobacteria *Thermomonospora curvata* (Tcu) and *Thermobifida fusca* (Tfu) and tested their ability to hydrolyze 7mG and unhook NM_8-ICLs (Fig. 5). Both proteins showed significant activity for both substrates, providing evidence that the YQL subfamily in general has comparable specificity for simple N7-alkylguanine lesions, distinguishing it biochemically from the AZL subfamily.

Discussion

Phylogenetic characterization of the HTH_42 superfamily proteins within *Streptomyces* reveals two distinct subfamilies, YQL and AZL (the latter of which contains the AZB2 clade). Most strikingly, AZL genes, which are most prevalent in environmental microbes such as those from the phylum Actinobacteria (Fig. S1B), are highly enriched in BGCs. We found AZL proteins in BGCs that produce a variety of verified and putative genotoxins, with approximately one-fifth of all AZL proteins located in BGCs predicted to produce a DNA alkylating agent. We show that the AZL protein, HedH4, within the HED cluster specifically excises HED-DNA adducts and improves viability of cells grown in the presence of the compound. In a separate study, we recently found that the two paralogs present in TXN and LLD clusters (TxnU2, TxnU4, LldU1, LldU5) are self-resistance glycosylases for these compounds (56). Thus, together with the previous example from the AZB BGC (28, 31), there is now mounting evidence that AZL family genes have evolved largely as DNA repair self-resistance proteins against a variety of natural products. Consistent with their role in resistance, the AZL genes found inside BGCs frequently localize around a variety of other resistance genes. Moreover, the relatively high copy number and low sequence conservation of AZL proteins are consistent with increased expression or possible horizontal gene transfer events that enable these enzymes to evolve specificity for particular natural product (57). We also found AZL homologs in BGCs that by homology were not expected to produce DNA alkylators or other genotoxins. The AZL proteins in these clusters could have regulatory or protective roles outside of DNA repair. Alternatively, these clusters could have additional uncharacterized enzymes such as cytochrome P450s, sulfate adenyltransferases, or epoxidases that could convert the natural products into DNA alkylators (58).

The fate of the AP sites generated by AZL enzymes is a key unanswered question regarding glycosylase-mediated self-resistance in antibiotic bacteria. While the DNA adducts of AZB and HED natural products would likely pose significant blocks to replication and transcription, their excision by AZL glycosylases also generate AP sites, which are highly toxic BER intermediates (59, 60). Although the modest protection we observed from HedH4 overexpression in HED-challenged *E. coli* could be a result of the weak-expression promoter used, it also suggests that either the AP sites generated are poor substrates for the AP endonucleases present in *E. coli* or that HED-DNA adducts are substrates for an alternative repair pathway. The intercalated HED-DNA adduct likely poses a unique challenge relative to other glycosylase substrates. It is likely that the HedH4-generated HED-guanine moiety remains intercalated at the AP site and requires a specialized AP endonuclease for repair. Indeed, we recently found that the excised guanine adduct of the related, intercalating natural product TXNA is a poor substrate for *E. coli* EndoIV (56). More pertinent to HED biosynthesis, the producing organism *S. griseoruber* contains two copies each of ExoIII- and EndoIV-like AP endonucleases that may have evolved to incise HED AP-sites, although none are located in the *hed* BGC. In addition, the bulky HED-DNA addicts lesions are likely substrates for nucleotide excision repair pathway, which is initiated by UvrA in bacteria and has been shown to play an important role in natural product self-resistance (25, 54, 61-63). Indeed, within the HED BGC there is a predicted UvrA-like drug resistance protein (HedH11) that contains a partial UvrA DNA-binding domain and a conserved ABC transporter domain that could initiate NER of HED-guanosine adducts or even HED-guanine/AP-site products generated by HedH4. There are also two additional putative UvrA homologs outside of the *hed* cluster. Additionally, there are three putative transporters within the cluster—HedH7 (ABC2 type), HedH6 (DrrA-like) and HedH1

(EmrB/QacA antiporter)—which could serve to physically bind to HED and direct it out of the cell through a transmembrane transport system.

In contrast to the genotoxin-specific AZL genes, YQL and AZL2 are always found outside clusters and thus are likely to provide a more general role in protecting the genome against environmental genotoxins, similar to that shown for *E. coli* YcaQ (31). YQL proteins and their gene neighborhoods are very highly conserved, suggesting they play a critical role as part of a unified pathway (64). Although that pathway is unknown, the presence of a two-component transcription factor/kinase and ComF DNA helicase within the YQL neighborhood in *Streptomyces* also hints at a signaling network for DNA uptake (65-67). Similarly, *E. coli* YcaQ is localized in a four-gene operon involved in cell wall biosynthesis and transformation competence (31). Continued exploration of the gene neighborhoods of YQL and AZL beyond *Streptomyces* will reveal a deeper understanding of the cellular roles played by these enzymes. This will be especially important for YQL, which are prevalent in human pathogens or commensal microbes (28).

A small subset of HTH_42 proteins contain additional domains often associated with nucleic acid transactions (Fig. S1A) (28). These multimodular HTH_42 proteins have been relatively understudied, although they do not appear to be associated with BGCs. Most contain an associated DEAD box helicase domain, including Lhr, a member of the helicase superfamily II (68). *Mycobacterium smegmatis* and *E. coli* Lhr have been characterized as ATP-dependent 3′→5′ ssDNA translocases with the ability to unwind RNA-DNA hybrids (69, 70). Studies in *Mycobacterium tuberculosis* have demonstrated a strong transcriptional activation of *lhr* in cells exposed to MMC (71), suggesting that Lhr may function as an RNA-DNA helicase in response to MMC-DNA crosslinks. While the structure of the C-terminal HTH_42 domain of *M. smegmatis* Lhr is similar to AlkZ, it lacks the catalytic QΦQ motif and adopts a tetrameric structure that occludes the putative DNA binding surface (70). Thus, the function of the Lhr HTH_42 domain and its interplay with the helicase core remains to be determined.

Resistance genome mining has emerged as a critical bioinformatically driven pipeline to discover novel natural products and gene clusters in several organisms (72, 73). A key benefit of resistance genome mining is the dramatically decreased candidate pool as a result of targeted identification of gene clusters containing a resistance gene. Generally, these methods require a basic understanding of the resistance mechanisms involved. We sought to use this approach for the first time to hunt for BGCs that produce alkylating genotoxins, using prior knowledge of the DNA repair functions of *S. sahachiroi* AlkZ within the AZB cluster (28, 30, 31). In this study, we examined 435 *Streptomyces* species for BGCs within which an AlkZ-related gene was located and found 62 uncharacterized clusters that are candidates for targeted elucidation of their products. Characterization of these orphan clusters could provide new analogs or types of DNA alkylating/damaging secondary metabolites, an important step in developing new antitumor or antibiotic treatments. This classification of YQL/AZL proteins in *Streptomyces* is an important first step in understanding their evolutionary connection to each other and to BGCs of different types, and demonstrates that targeted resistance genome mining is a viable approach to discover novel genotoxins and resistance mechanisms from uncharacterized BGCs.

Materials and Methods

Reagents. DNA oligonucleotides (Table S4) were purchased from Integrated DNA Technologies. *Escherichia coli* K-12 wild-type strain was purchased from the Keio *E. coli* knockout collection (Dharmacon, GE Healthcare). HED (*Streptomyces griseoruber* ATCC 23919) was obtained from the National Cancer Institute's Developmental Therapeutic Program (NCI DTP) Open Compound Repository (NSC 70929). Trioxacarcin A (TXNA) was isolated from *Streptomyces bottropensis* NRRL 12051 as described (56). AZB was prepared from organic extract of *Streptomyces sahachiroi* (ATCC 33158) as in (31). NM_8 compound was synthesized and purified by the Vanderbilt Molecular Design and Synthesis Center (31). AlkA, AlkC, AlkD, AlkZ, LldU1/5, and TxnU2/4, and YcaQ were purified as described (30, 31, 52, 56, 74, 75). Unless otherwise noted, all chemicals were purchased from Sigma-Aldrich, and all enzymes were purchased from New England Biolabs (NEB).

Taxonomy and phylogeny of Streptomyces HTH_42 proteins. To identify HTH_42 proteins in *Streptomyces*, the protein sequences for YcaQ (GenBank accession number QHB65847.1) and AlkZ (GenBank accession number ABY83174.1) were used for tBLASTn and BLASTp searches (BLAST+ v2.11.0) against all *Streptomyces* genomes (taxid:1883). Searches were run with the BLOSUM62 matrix, 1000 maximum target sequences, and 0.05 threshold using an e-value and identity cutoff of 10^{-4} and 25%, respectively. All hits were verified for the presence of the (H/Q)Φ(D/Q) catalytic motif, during which the (H/Q)Φ(S/T)(D/E) (AZL2) variant was identified. Truncated genes, poor sequence quality genes, and pseudogenes were eliminated. Additional sequences were obtained by searching the Pfam database v33.1 (76) for *Streptomyces* HTH_42 superfamily members (PF06224). Sequences from Pfam were sorted according to their domain classes (Fig. S1A), and only sequences from Class 1 with >75% coverage were included. Protein sequences were aligned using EMBL-EBI Clustal OmegaW or MAFFT v7 using default parameters (77, 78). The evolutionary history of YQL/AZL sequences were reconstructed using IQTREE2 with default settings (79), and the phylogenetic tree was assembled with the Interactive Tree of Life (v5) phylogeny display tool (80). Sequence logos were generated with WebLogo v2.8.2 (81). The copy number frequency and coincidence of YQL/AZL in the same genome was determined by manually counting the number and identity of homologs in each species. A list of all YQL/AZL/AZL2 proteins and *Streptomyces* genomes analyzed in this study can be found in Table S1.

Identification of AZL proteins in known biosynthetic gene clusters. To find AZL proteins in verified and/or published BGCs, we searched MIBiG v2.0 for the AZB BGC (BGC0000960) from *S. sahachiroi* (27, 82), followed by an iterative search using the *MIBiG Hits* function until no more hits were obtained. The homologs TxnU2 and TxnU4 were identified from the initial BLAST search within the deposited NCBI trioxacarcin BGC sequence (83). The homolog within the aclacinomycin BGC was also identified in the initial BLAST search as appearing in proximity to aclacinomycin biosynthesis genes. Closer inspection of the published sequence for the aclacinomycin BGC (GenBank accession number AB008466.1) revealed an AZL protein (Orf1) located immediately 3' of the cluster (84). A detailed list of the AZL proteins in known BGCs can be found in Table S3.

Identification of AZL proteins in uncharacterized biosynthetic gene clusters. To determine the physical distance in base pairs between the genomic coordinates of AZL proteins and those of BGCs present in the genome assemblies of *Streptomyces* (average number of scaffolds: 96.30; minimum: 1; maximum: 1,956), we first predicted the BGCs in each genome using antiSMASH v5.1.0 (38) with the *taxon* parameter set to *bacteria*. Using the BGC sequences identified from antiSMASH and AZL sequences, a custom python script using Biopython (85) determined the shortest base pair distance between the physical location of the YQL/AZL gene and the location of the nearest BGC on the same scaffold (less than 2 Mbps away). To be considered within a BGC, the homolog had to be observed within 5 genes or 2 kb of the nearest cluster. Known Cluster BLAST was performed within antiSMASH to determine the most similar BGC to the unknown clusters, and the result with the highest percentage of similar genes was recorded as the most similar cluster. A detailed list of the genome information, cluster IDs, and the closest 3' and/or 5' BGC can be found in Table S2.

Gene ontology analysis. To identify GO terms for nearest neighbors identified through BLAST, Pfam, and MIBiG searches, we randomly chose 40 homologs each of AZL inside BGCs, AZL outside BGCs, and YQL, which represent ~10% of the sequences for each. Amino acid sequences for the five genes on both sides of the YQL/AZL genes were downloaded from the NCBI database, for a total of 400 neighbors for each of the three classes. Cellular functions of any already annotated genes in the NCBI database were identified and recorded. The downloaded sequences were then run through the GhostKOALA (v2.2) and eggNOG (v5.0) GO annotation databases (86, 87). After known GO terms for all gene neighbors were identified, proteins were categorized by biological processes and molecular functions, and the values for these terms were used to create the GO term distributions. Proteins that had multiple GO terms associated with them were counted into each class of terms. A list of all proteins and their annotated GO terms can be found in Tables S5-S6.

Protein purification. Genes encoding *Streptomyces caeruleatus* AZL2, *Streptomyces griseoruber* HedH4, *Thermomonospora curvata* YQL, and *Thermobifida fusca* YQL were codon optimized and synthesized by GenScript and cloned into pBG102. The N-terminal His_6-SUMO fusion proteins were overexpressed in *Escherichia coli* Tuner (DE3) cells at 16°C for 18 hr in LB medium supplemented with 30 µg/mL kanamycin and 50 µM isopropyl β-D-1-thiogalactopyranoside (IPTG). Cells were lysed by sonication and cell debris removed by centrifugation at 45,000 × g at 4°C for 30 min. Clarified lysate was passed over Ni-NTA agarose equilibrated in buffer A (50 mM Tris•HCl pH 8.5, 500 mM NaCl, 25 mM imidazole, and 10% (vol/vol) glycerol) and protein eluted in 250 mM imidazole/buffer A. Protein fractions were pooled and supplemented with 0.1 mM EDTA and 1 mM tris(2-carboxyethyl)phosphine (TCEP) before incubation with 0.5 mg Rhinovirus 3C protease (PreScission) at 4°C overnight. Cleaved protein was diluted 10-fold in buffer B (50 mM Tris•HCl pH 8.5, 10% (vol/vol) glycerol, 0.1 mM TCEP, and 0.1 mM EDTA) and purified by heparin sepharose using a 0–1 M NaCl/buffer B linear gradient. Fractions were pooled and passed over Ni-NTA agarose in buffer A, concentrated and filtered, and buffer exchanged into buffer C (20 mM Tris•HCl pH 8.5, 100 mM NaCl, 5% (vol/vol) glycerol, 0.1 mM TCEP, and 0.1 mM EDTA). Protein was concentrated to 4 mg/mL, flash-frozen in liquid nitrogen, and stored at −80°C. Mutant protein expression vectors were generated using the Q5 Mutagenesis Kit (New England BioLabs) and proteins overexpressed and purified the same as wild-type.

DNA glycosylase activity. DNA substrates containing a single *N*7-methyl-2′- deoxyguanosine lesion and a 5'-Cy5 fluorophore were prepared as described previously (88). AZB- and NM_8-ICL substrates were generated and purified as in (31). DNA substrates containing a single HED-guanosine or trioxacarcin A (TXNA)-guanosine adduct were prepared by annealing 5'-Cy5-labeled DNA containing the target sequence to the complementary unlabeled oligodeoxynucleotide (Table S4). HED and TXNA were dissolved in DMSO to a concentration of 5 mM. 100 µM DNA was incubated with 200 µM HED or TXNA in 10% methanol and 20% DMSO at 4°C on ice in the dark for 24 hr. Unreacted drug was removed using an Illustra G-25 spin column (GE Healthcare) equilibrated in TE buffer (10 mM Tris•HCl pH 8.0, 1 mM EDTA pH 8.0), and the DNA was stored at -80°C.

In each glycosylase reaction, 1 µM enzyme was incubated with 50 nM DNA in glycosylase buffer (50 mM HEPES pH 8.5, 100 mM KCl, 1 mM EDTA, and 10% (vol/vol) glycerol) at 25°C. At various time points, 4-µL aliquots were added to 1 µL of 1M NaOH and heated at 70°C for 2 min. Samples were denatured at 70°C for 5 min in 5 mM EDTA pH 8.0, 80% (wt/vol) formamide, and 1 mg/mL blue dextran prior to electrophoresis on a 20% (wt/vol) acrylamide/8 M urea sequencing gel at 40 watts for 1 hr in 0.5 × TBE buffer (45 mM Tris, 45 mM borate, and 1 mM EDTA pH 8.0). Gels were imaged on a Typhoon Trio variable mode imager (GE Healthcare) using 633-nm excitation/670-nm emission fluorescence for Cy5, and bands were quantified with ImageQuant (GE Healthcare). All excision assays were performed in triplicate.

HPLC-MS analysis of HED and HED-guanine. HPLC was performed on an Agilent Series 1100 system equipped with an analytical SymmetryShield RP-C18 column (3.5 µm, 4.6 mm × 7.5 mm, 100 Å pore size) and using a linear gradient from 90% buffer A (10 mM ammonium formate) / 10% buffer B (100% methanol) to 100% B over 40 min and a flow rate of 0.4 mL/min. HED was diluted to 50 µM in 10% methanol and stored on ice prior to HPLC injection. To analyze the product of HedH4 activity, HED-DNA was diluted to 10 µM in glycosylase buffer and reacted with 50 µM HedH4 for 1 hr at room temperature before injection. Mass spectrometry was performed with an LTQ Orbitrap XL Hybrid FT Mass Spectrometer (Thermo Fisher Scientific) in positive ion mode from 300-1000 m/z.

Cellular assays for HED resistance. The *hedH4* wild-type gene was sub-cloned from pBG102 into pSF-OXB1 using NcoI and XbaI restriction sites. The pSF-OXB1 vector contains a kanamycin resistance gene and allows for constitutive low-level expression from a modified

AraBAD promoter. pSF-OXB1 and HedH4/pSF-OXB1 were transformed into *E. coli* K-12 cells. Cloning of *hedH4* was confirmed by sequencing, restriction digest using NcoI-HF/XbaI (Fig. S4C), and colony PCR of K-12 transformants using the HedH4 NcoI and XbaI primers (Fig. S4D, Table S4). Cultures were grown at 37°C in LB media supplemented with 30 µg/mL Kan. Growth curves were generated by diluting overnight cultures to 0.01 OD_{600} in LB/Kan supplemented with 0 nM-100 µM HED in a 96-well flat-bottom plate. The plate was incubated at 30°C with shaking for 24 hr and cell density was measured at 600 nm every 20 min using a Bio-Tek Synergy 2 microplate reader. IC_{50} values were determined from a fit to the equation, Lag time = $Min_{lag} + (Max_{lag}-Min_{lag})/(1+(IC_{50}/[HED])^{h})$, where *h* is the Hill slope. Growths were performed in triplicate.

E. coli survival curves after HED treatment were performed using a colony dilution assay. A saturated overnight LB/Kan culture from a single colony was diluted to 0.01 OD_{600} in 1 mL fresh LB/Kan media and grown to 0.6 OD_{600} at 37°C. The cells were treated with various concentrations of HED for 1 hr at 37°C. Treated cells were transferred to fresh LB/Kan media, serially diluted by 10^{-6} in LB/Kan media, and 100 µL of diluted cells were plated on LB/Kan agar plates and grown at 37°C overnight. Colonies were counted the next morning and the CFU/mL culture was determined. The percent survival was calculated as CFU/mL(treated) / CFU/mL(untreated). Curves were plotted on a logarithmic scale and IC_{50} values determined by non-linear regression fits to the data. Growths were performed in triplicate.

Acknowledgements

The authors wish to thank Gong-Li Tang (Chinese Academy of Sciences) for providing trioxacarcin A. This work was supported by grants from the National Institutes of Health (R01GM131071) and the National Science Foundation (MCB-1928918) to B.F.E. Research in A.R.'s lab is supported by grants from the National Science Foundation (DEB-1442113 and DEB-2110404), the National Institutes of Health/National Institute of Allergy and Infectious Diseases (R56 AI146096), and the Burroughs Wellcome Fund. N.P.B. was supported by the Vanderbilt Training Program in Environmental Toxicology (NIH T32ES007028) and an NSF Graduate Research Fellowship (DGE-1445197). J.L.S. and A.R. are supported by the Howard Hughes Medical Institute through the James H. Gilliam Fellowships for Advanced Study program.

Author Contributions

B.F.E. and N.P.B. conceived of the study; N.P.B., K.L.W., and J.L.S. performed experiments; N.P.B., K.L.W., and B.F.E. designed experiments, analyzed data, and wrote the manuscript. All authors provided feedback on the interpretation of the results and the manuscript.

Declaration of interests

A.R. is a scientific consultant for LifeMine Therapeutics, Inc.

References

1. Demain AL, Sanchez S. 2009. Microbial drug discovery: 80 years of progress. J Antibiot (Tokyo) 62:5-16.
2. Procopio RE, Silva IR, Martins MK, Azevedo JL, Araujo JM. 2012. Antibiotics produced by Streptomyces. Braz J Infect Dis 16:466-71.
3. Jacob C, Weissman KJ. 2017. Unpackaging the Roles of Streptomyces Natural Products. Cell Chem Biol 24:1194-1195.
4. Law JW, Law LN, Letchumanan V, Tan LT, Wong SH, Chan KG, Ab Mutalib NS, Lee LH. 2020. Anticancer Drug Discovery from Microbial Sources: The Unique Mangrove Streptomycetes. Molecules 25.
5. Tyc O, Song C, Dickschat JS, Vos M, Garbeva P. 2017. The Ecological Role of Volatile and Soluble Secondary Metabolites Produced by Soil Bacteria. Trends Microbiol 25:280-292.
6. Gates KS. 2009. An Overview of Chemical Processes That Damage Cellular DNA: Spontaneous Hydrolysis, Alkylation, and Reactions with Radicals. Chemical Research in Toxicology 22:1747–1760.
7. Chumduri C, Gurumurthy RK, Zietlow R, Meyer TF. 2016. Subversion of host genome integrity by bacterial pathogens. Nat Rev Mol Cell Biol 17:659-673.
8. Huang M, Lu JJ, Ding J. 2021. Natural Products in Cancer Therapy: Past, Present and Future. Nat Prod Bioprospect 11:5-13.
9. Boger DL, Garbaccio RM. 1997. Catalysis of the CC-1065 and Duocarmycin DNA Alkylation Reaction: DNA Binding Induced Conformational Change in the Agent Results in Activation Bioorganic and Medicinal Chemistry 5:263-276.
10. Parrish JP, Kastrinsky DB, Wolkenberg SE, Igarashi Y, Boger DL. 2003. DNA Alkylation Properties of Yatakemycin. Journal of the American Chemical Society 125:10971-10976.
11. Nagai K, Yamaki H, Tanaka N, Umezwa H. 1967. Inhibition by pluramycin A of nucleic acid biosynthesis. J Biochem 62:321-7.

12. Hansen M, Yun S, Hurley L. 1995. Hedamycin intercalates the DNA helix and, through carbohydrate-mediated recognition in the minor groove, directs N-alkylation of guanine in the major groove in a sequence-specific manner Chemistry & Biology 2:229-240.
13. Hansen M, Hurley L. 1995. Altromycin B Threads the DNA Helix Interacting with Both the Major and the Minor Grooves To Position Itself for Site-Directed Alkylation of Guanine N7. J Am Chem Soc 117:2421-2429.
14. Tamaoki T, Shirahata K, Iida T, Tomita F. 1981. Trioxacarcins, novel antitumor antibiotics. II. Isolation, physico-chemical properties and mode of action. J Antibiot (Tokyo) 34:1525-30.
15. Tomita F, Tamaoki T, Morimoto M, Fujimoto K. 1981. Trioxacarcins, novel antitumor antibiotics. I. Producing organism, fermentation and biological activities. J Antibiot (Tokyo) 34:1519-24.
16. Maiese WM, Labeda DP, Korshalla J, Kuck N, Fantini AA, Wildey MJ, Thomas J, Greenstein M. 1990. LL-D49194 antibiotics, a novel family of antitumor agents: taxonomy, fermentation and biological properties. J Antibiot (Tokyo) 43:253-8.
17. Nooner T, Dutta S, Gates KS. 2004. Chemical Properties of the Leinamycin-Guanine Adduct in DNA. Chemical Research in Toxicology 17:942-949.
18. Terawaki A, Greenberg J. 1966. Effect of carzinophillin on bacterial deoxyribonucleic acid: formation of inter-strand cross-links in deoxyribonucleic acid and their disappearance during post-treatment incubation. Nature 209:481-4.
19. Galm U, Hager MH, Van Lanen SG, Ju J, Thorson JS, Shen B. 2005. Antitumor antibiotics: bleomycin, enediynes, and mitomycin. Chem Rev 105:739-58.
20. Cundliffe E, Demain AL. 2010. Avoidance of suicide in antibiotic-producing microbes. J Ind Microbiol Biotechnol 37:643-72.
21. Tenconi E, Rigali S. 2018. Self-resistance mechanisms to DNA-damaging antitumor antibiotics in actinobacteria. Curr Opin Microbiol 45:100-108.
22. Ng TL, Rohac R, Mitchell AJ, Boal AK, Balskus EP. 2019. An N-nitrosating metalloenzyme constructs the pharmacophore of streptozotocin. Nature 566:94-99.
23. Xu H, Huang W, He QL, Zhao ZX, Zhang F, Wang R, Kang J, Tang GL. 2012. Self-resistance to an antitumor antibiotic: a DNA glycosylase triggers the base-excision repair system in yatakemycin biosynthesis. Angew Chem Int Ed Engl 51:10532-6.
24. Mullins EA, Dorival J, Tang GL, Boger DL, Eichman BF. 2021. Structural evolution of a DNA repair self-resistance mechanism targeting genotoxic secondary metabolites. Nat Commun 12:6942.

25. Lomovskaya N, Hong SK, Kim SU, Fonstein L, Furuya K, Hutchinson RC. 1996. The Streptomyces peucetius drrC gene encodes a UvrA-like protein involved in daunorubicin resistance and production. J Bacteriol 178:3238-45.
26. Ma L, Sun S, Yuan Z, Deng Z, Tang Y, Yu Y. 2020. Three putative DNA replication/repair elements encoding genes confer self-resistance to distamycin in Streptomyces netropsis. Acta Biochim Biophys Sin (Shanghai) 52:91-96.
27. Zhao Q, He Q, Ding W, Tang M, Kang Q, Yu Y, Deng W, Zhang Q, Fang J, Tang G, Liu W. 2008. Characterization of the azinomycin B biosynthetic gene cluster revealing a different iterative type I polyketide synthase for naphthoate biosynthesis. Chem Biol 15:693-705.
28. Wang S, Liu K, Xiao L, Yang L, Li H, Zhang F, Lei L, Li S, Feng X, Li A, He J. 2016. Characterization of a novel DNA glycosylase from S. sahachiroi involved in the reduction and repair of azinomycin B induced DNA damage. Nucleic Acids Res 44:187-97.
29. Mullins EA, Rodriguez AA, Bradley NP, Eichman BF. 2019. Emerging Roles of DNA Glycosylases and the Base Excision Repair Pathway. Trends Biochem Sci 44:765-781.
30. Mullins EA, Warren GM, Bradley NP, Eichman BF. 2017. Structure of a DNA glycosylase that unhooks interstrand cross-links. Proc Natl Acad Sci U S A 114:4400-4405.
31. Bradley NP, Washburn LA, Christov PP, Watanabe CMH, Eichman BF. 2020. Escherichia coli YcaQ is a DNA glycosylase that unhooks DNA interstrand crosslinks. Nucleic Acids Res 48:7005-7017.
32. Kersten RD, Weng JK. 2018. Gene-guided discovery and engineering of branched cyclic peptides in plants. Proc Natl Acad Sci U S A 115:E10961-E10969.
33. Kjaerbolling I, Vesth T, Andersen MR. 2019. Resistance Gene-Directed Genome Mining of 50 Aspergillus Species. mSystems 4.
34. Belknap KC, Park CJ, Barth BM, Andam CP. 2020. Genome mining of biosynthetic and chemotherapeutic gene clusters in Streptomyces bacteria. Sci Rep 10:2003.
35. Ziemert N, Alanjary M, Weber T. 2016. The evolution of genome mining in microbes - a review. Nat Prod Rep 33:988-1005.
36. Thaker MN, Wang W, Spanogiannopoulos P, Waglechner N, King AM, Medina R, Wright GD. 2013. Identifying producers of antibacterial compounds by screening for antibiotic resistance. Nat Biotechnol 31:922-7.
37. Skinnider MA, Dejong CA, Rees PN, Johnston CW, Li H, Webster AL, Wyatt MA, Magarvey NA. 2015. Genomes to natural products PRediction Informatics for Secondary Metabolomes (PRISM). Nucleic Acids Res 43:9645-62.

38. Blin K, Shaw S, Steinke K, Villebro R, Ziemert N, Lee SY, Medema MH, Weber T. 2019. antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. Nucleic Acids Res 47:W81-W87.

39. Mungan MD, Alanjary M, Blin K, Weber T, Medema MH, Ziemert N. 2020. ARTS 2.0: feature updates and expansion of the Antibiotic Resistant Target Seeker for comparative genome mining. Nucleic Acids Res 48:W546-W552.

40. Reusser F. 1977. Ficellomycin and feldamycin; inhibitors of bacterial semiconservative DNA replication. Biochemistry 16:3406-12.

41. Sugimoto Y, Otani T, Oie S, Wierzba K, Yamada Y. 1990. Mechanism of action of a new macromolecular antitumor antibiotic, C-1027. J Antibiot (Tokyo) 43:417-21.

42. Pfoh R, Laatsch H, Sheldrick GM. 2008. Crystal structure of trioxacarcin A covalently bound to DNA. Nucleic Acids Res 36:3508-14.

43. Nitiss JL, Pourquier P, Pommier Y. 1997. Aclacinomycin A Stabilizes Topoisomerase I Covalent Complexes. Cancer Research 57:4564-4569.

44. Hara M, Akasaka K, Akinaga S, Okabe M, Nakano H, Gomez R, Wood D, Uh M, Tamanoi F. 1993. Identification of Ras farnesyltransferase inhibitors by microbial screening. Proc Natl Acad Sci U S A 90:2281-2285.

45. Isobe Y, Okumura M, McGregor LM, Brittain SM, Jones MD, Liang X, White R, Forrester W, McKenna JM, Tallarico JA, Schirle M, Maimone TJ, Nomura DK. 2020. Manumycin polyketides act as molecular glues between UBR7 and P53. Nat Chem Biol 16:1189-1198.

46. Labana P, Dornan MH, Lafreniere M, Czarny TL, Brown ED, Pezacki JP, Boddy CN. 2021. Armeniaspirols inhibit the AAA+ proteases ClpXP and ClpYQ leading to cell division arrest in Gram-positive bacteria. Cell Chem Biol doi:10.1016/j.chembiol.2021.07.001.

47. Muliandi A, Katsuyama Y, Sone K, Izumikawa M, Moriya T, Hashimoto J, Kozone I, Takagi M, Shin-ya K, Ohnishi Y. 2014. Biosynthesis of the 4-methyloxazoline-containing nonribosomal peptides, JBIR-34 and -35, in Streptomyces sp. Sp080513GE-23. Chem Biol 21:923-34.

48. Tu LC, Melendy T, Beerman TA. 2004. DNA damage responses triggered by a highly cytotoxic monofunctional DNA alkylator, hedamycin, a pluramycin antitumor antibiotic. Molecular Cancer Therapeutics 3:577-585.

49. Hemminki K, Peltonen K, Vodicka P. 1989. Depurination from DNA of 7-methylguanine, 7-(2-aminoethyl)-guanine and ring-opened 7-methylguanines. Chemico-Biological Interactions 70:289-303.

50. O'Brien PJ, Ellenberger T. 2004. Dissecting the broad substrate specificity of human 3-methyladenine-DNA glycosylase. J Biol Chem 279:9750-7.

51. Alseth I, Rognes T, Lindbäck T, Solberg I, Robertsen K, Kristiansen KI, Mainieri D, Lillehagen L, Kolstø AB, Bjørås M. 2006. A new protein superfamily includes two novel 3-methyladenine DNA glycosylases from *Bacillus cereus*, AlkC and AlkD. Molecular Microbiology 59:1602-9.

52. Shi R, Mullins EA, Shen XX, Lay KT, Yuen PK, David SS, Rokas A, Eichman BF. 2018. Selective base excision repair of DNA damage by the non-base-flipping DNA glycosylase AlkC. EMBO Journal 37:63-74.

53. Parsons ZD, Bland JM, Mullins EA, Eichman BF. 2016. A Catalytic Role for C-H/π Interactions in Base Excision Repair by Bacillus cereus DNA Glycosylase AlkD. J Am Chem Soc 138:11485-8.

54. Mullins EA, Shi R, Eichman BF. 2017. Toxicity and repair of DNA adducts produced by the natural product yatakemycin. Nat Chem Biol 13:1002-1008.

55. Brooks SC, Adhikary S, Rubinson EH, Eichman BF. 2013. Recent advances in the structural mechanisms of DNA glycosylases. Biochim Biophys Acta 1834:247-71.

56. Chen X, Bradley NP, Lub W, Wahl KL, Zhang M, Yuan H, Hou XF, Eichman BF, Tang GL. 2022. Base excision repair system targeting DNA adducts of antibiotics trioxacarcin/LL-D49194 for self-resistance. Accepted for publication.

57. Hastings PJ, Lupski JR, Rosenberg SM, Ira G. 2009. Mechanisms of change in gene copy number. Nat Rev Genet 10:551-64.

58. Thibodeaux CJ, Chang WC, Liu HW. 2012. Enzymatic chemistry of cyclopropane, epoxide, and aziridine biosynthesis. Chem Rev 112:1681-709.

59. Posnick LM, Samson LD. 1999. Imbalanced base excision repair increases spontaneous mutation and alkylation sensitivity in Escherichia coli. J Bacteriol 181:6763-71.

60. Tomicic M, Franekic J. 1996. Effect of overexpression of E. coli 3-methyladenine-DNA glycosylase I (Tag) on survival and mutation induction in Salmonella typhimurium. Mutat Res 358:81-7.

61. Jin SG, Choi JH, Ahn B, O'Connor TR, Mar W, Lee CS. 2001. Excision repair of adozelesin-N3 adenine adduct by 3-methyladenine-DNA glycosylases and UvrABC nuclease. Mol Cells 11:41-7.

62. Kiakos K, Sato A, Asao T, McHugh PJ, Lee M, Hartley JA. 2007. DNA sequence selective adenine alkylation, mechanism of adduct repair, and in vivo antitumor activity of the novel

achiral seco-amino-cyclopropylbenz[e]indolone analogue of duocarmycin AS-I-145. Mol Cancer Ther 6:2708-18.

63. Burby PE, Simmons LA. 2019. A bacterial DNA repair pathway specific to a natural antibiotic. Mol Microbiol 111:338-353.
64. Rogozin IA, Makarova KS, Murvai J, Czabarka E, Wolf WS, Tatusov RL, Szekely LA, Koonin EV. 2002. Connected gene neighborhoods in prokaryotic genomes. Nucleic Acids Res 30:2212-2223.
65. Londono-Vallejo JA, Dubnau D. 1993. comF, a Bacillus subtilis late competence locus, encodes a protein similar to ATP-dependent RNA/DNA helicases. Mol Microbiol 9:119-31.
66. Turgay K, Hahn J, Burghoorn J, Dubnau D. 1998. Competence in Bacillus subtilis is controlled by regulated proteolysis of a transcription factor. EMBO J 17:6730-8.
67. Veening JW, Blokesch M. 2017. Interbacterial predation as a strategy for DNA acquisition in naturally competent bacteria. Nat Rev Microbiol 15:629.
68. Reuven NB, Koonin EV, Rudd KE, Deutscher MP. 1995. The gene for the longest known Escherichia coli protein is a member of helicase superfamily II. J Bacteriol 177:5393-400.
69. Ordonez H, Shuman S. 2013. Mycobacterium smegmatis Lhr Is a DNA-dependent ATPase and a 3'-to-5' DNA translocase and helicase that prefers to unwind 3'-tailed RNA:DNA hybrids. J Biol Chem 288:14125-14134.
70. Warren GM, Wang J, Patel DJ, Shuman S. 2021. Oligomeric quaternary structure of Escherichia coli and Mycobacterium smegmatis Lhr helicases is nucleated by a novel C-terminal domain composed of five winged-helix modules. Nucleic Acids Res 49:3876-3887.
71. Boshoff HI, Reed MB, Barry CE, 3rd, Mizrahi V. 2003. DnaE2 polymerase contributes to in vivo survival and the emergence of drug resistance in Mycobacterium tuberculosis. Cell 113:183-93.
72. Panter F, Krug D, Baumann S, Muller R. 2018. Self-resistance guided genome mining uncovers new topoisomerase inhibitors from myxobacteria. Chem Sci 9:4898-4908.
73. Yan Y, Liu Q, Zang X, Yuan S, Bat-Erdene U, Nguyen C, Gan J, Zhou J, Jacobsen SE, Tang Y. 2018. Resistance-gene-directed discovery of a natural-product herbicide with a new mode of action. Nature 559:415-418.
74. Rubinson EH, Metz AH, O'Quin J, Eichman BF. 2008. A new protein architecture for processing alkylation damaged DNA: the crystal structure of DNA glycosylase AlkD. J Mol Biol 381:13-23.
75. Taylor EL, O’Brien PJ. 2015. Kinetic mechanism for the flipping and excision of 1,N6-ethenoadenine by AlkA. Biochemistry 54:898-908.

76. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A, Sonnhammer ELL, Hirsh L, Paladin L, Piovesan D, Tosatto SCE, Finn RD. 2019. The Pfam protein families database in 2019. Nucleic Acids Res 47:D427-D432.
77. Madeira F, Park YM, Lee J, Buso N, Gur T, Madhusoodanan N, Basutkar P, Tivey ARN, Potter SC, Finn RD, Lopez R. 2019. The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Res 47:W636-W641.
78. Katoh K, Rozewicki J, Yamada KD. 2019. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Brief Bioinform 20:1160-1166.
79. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. 2020. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol Biol Evol 37:1530-1534.
80. Letunic I, Bork P. 2019. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res 47:W256-W259.
81. Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: a sequence logo generator. Genome Res 14:1188-90.
82. Kautsar SA, Blin K, Shaw S, Navarro-Munoz JC, Terlouw BR, van der Hooft JJJ, van Santen JA, Tracanna V, Suarez Duran HG, Pascal Andreu V, Selem-Mojica N, Alanjary M, Robinson SL, Lund G, Epstein SC, Sisto AC, Charkoudian LK, Collemare J, Linington RG, Weber T, Medema MH. 2020. MIBiG 2.0: a repository for biosynthetic gene clusters of known function. Nucleic Acids Res 48:D454-D458.
83. Zhang M, Hou XF, Qi LH, Yin Y, Li Q, Pan HX, Chen XY, Tang GL. 2015. Biosynthesis of trioxacarcin revealing a different starter unit and complex tailoring steps for type II polyketide synthase. Chem Sci 6:3440-3447.
84. Chung JY, Fujii I, Harada S, Sankawa U, Ebizuka Y. 2002. Expression, purification, and characterization of AknX anthrone oxygenase, which is involved in aklavinone biosynthesis in Streptomyces galilaeus. J Bacteriol 184:6115-22.
85. Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, de Hoon MJ. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25:1422-3.
86. Kanehisa M, Sato Y, Morishima K. 2016. BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. J Mol Biol 428:726-731.

87. Huerta-Cepas J, Szklarczyk D, Heller D, Hernandez-Plaza A, Forslund SK, Cook H, Mende DR, Letunic I, Rattei T, Jensen LJ, von Mering C, Bork P. 2019. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res 47:D309-D314.

88. Mullins EA, Rubinson EH, Pereira KN, Calcutt MW, Christov PP, Eichman BF. 2013. An HPLC-tandem mass spectrometry method for simultaneous detection of alkylated base excision repair products. Methods 64:59-66.

Figure Legends

Figure 1. Phylogenetic Organization of YQL/AZL Proteins in *Streptomyces*. (A) Azinomycin B reacts with opposite strands of DNA to form an ICL, which is unhooked by AlkZ. (B) Structure of a nitrogen mustard ICL derived from mechlorethamine and unhooked by *E. coli* YcaQ. (C) Phylogenetic tree of YcaQ-like (YQL, blue) and AlkZ-like (AZL, red/orange; AZL2, grey) *Streptomyces* proteins (n=897). The red and orange AZL clades distinguish HΦQ and QΦQ catalytic motifs. *E. coli* YcaQ and S. sahachiroi AlkZ proteins are labeled. (D) Sequence logos for the catalytic motifs in YQL, AZL, and AZL2 proteins. Catalytic residues are marked with asterisks. Colors correspond to side chain chemistry. (E) Copy number frequency per *Streptomyces* genome as a percentage of the total species analyzed (n=436 species, 897 sequences). One-way ANOVA significance (P) values of copy number variance are 0.0078 (YQL-AZL), 0.0033 (AZL-AZL2), and 0.3305 (YQL- AZL2), the latter of which is not significant. (F) YQL/AZL coincidence frequency. Blue shaded section represents species containing both subfamilies; tan shaded section represents species containing either YQL or AZL.

Figure 2. *Streptomyces* AZL proteins are found in diverse uncharacterized biosynthetic gene clusters. (A) Schematic depicting the workflow for identification of HTH_42 homologs in uncharacterized *Streptomyces* BGCs. Homologs were identified through the presence of the catalytic motif (red text in sequence alignment). The amino acid numbering is in relation to *S. sahachiroi* AlkZ. The corresponding *Streptomyces* genomes were input into antiSMASH, from which genomic distances between YQL/AZL and the nearest BGC, as well as homologous clusters were extracted. (B) Violin plot showing the distribution of distances of YQL (n=167) and AZL (n=154) genes to the nearest BGC (in kbps; ±100 kb). The dotted line at 0 kb represents the 5' (+) / 3' (-) termini of the nearest BGC. Thick and thin dashed lines within the plot represent the median and upper/lower quartiles, respectively. The Chi-square significance (P) value between YQL and AZL data is less than 0.0001. (C) Frequency of various types of BGCs in which AZL genes were found (n=68 clusters identified). The y-axis denotes the natural product/scaffold type to which that cluster is most homologous. Black bars represent known DNA alkylators or DNA interacting metabolites, and hashed bars represent potential DNA damaging metabolites. Lowercase letters to the right of the bars correspond to structures shown in panel D. (D) Representative compounds corresponding to BGC types in panel C. Potential reactive sites are colored red. LL-D4919α1 and hedamycin structures are shown in Fig. 3. (E,F) Nearest neighbor analysis of AZL (E) and YQL (F). (E) Nearest genes to AZL proteins found inside and outside clusters, shown as the ratio of GO terms inside and outside, and grouped by

function (blue, metabolic; green, cell signaling and function; orange, genome maintenance). (F) Representative example from *Streptomyces griseoviridis* of nearest neighbor analysis for YQL proteins. Genes are colored according to function as in panel E (grey, unknown/hypothetical gene). These genes are invariant for all YQL proteins, with the exception of the outermost genes, in which only one instance of variance was observed.

Figure 3. AZL proteins found in characterized *Streptomyces* biosynthetic gene clusters. (A,B) Gene diagrams for AZL-containing BGCs producing DNA alkylating agents (A) and compounds not known to alkylate DNA (B). Gene names are labeled below the cluster diagrams. The biosynthetic scaffold produced by specific genes in the cluster are shaded grey and labeled above the respective genes. NRPS, non-ribosomal peptide synthetase; PKS1/PKS2, type 1/2 polyketide synthase; (PKS), PKS-like. Chemical structures of the metabolites produced by each cluster are shown at the bottom of each panel.

Figure 4. HedH4 excises hedamycin-guanine adducts from DNA and provides cellular resistance to hedamycin toxicity. (A) HED modification of deoxyguanosine in DNA forms a HED-DNA adduct that is hydrolyzed by HedH4 to generate an abasic (AP) site in the DNA and free HED-guanine. The reactions within the dashed line are not catalyzed by HedH4. The AP nucleotide is susceptible to base-catalyzed nicking to form shorter DNA products containing either a 3′-phospho-α,β-unsaturated aldehyde (PUA, β-elimination) or a 3′-phosphate (δ-elimination). The asterisk (*) denotes the original 5′-end of the DNA. (B) Denaturing PAGE of 5'-Cy5-labeled HED-DNA substrate and β- and δ-elimination products after treatment with enzyme or buffer (mock) for 1 h, followed by NaOH to nick the AP site. The HED-DNA reaction only goes to ~50% completion under our reaction conditions, as shown by the two bands of equal intensity in the mock reaction. (C) HPLC-MS analysis of HED (blue) and the HED -guanine excision product from reaction of HedH4 and HED-DNA (red). Axis represents elution time (x-axis) versus relative abundance from total ion count (y-axis). Insets show mass spectra of each elution peak. (D) Wild-type and mutant HedH4 glycosylase activity for HED-DNA. Spontaneous depurination from a no-enzyme reaction (mock) is shown as a negative control. Data are mean ± SD (n=3). Curves were fit to a single exponential. Representative data are shown in Fig. S3C. (E) Denaturing PAGE of HED-DNA adducts after 1 h incubation with either buffer (mock) or bacterial alkylpurine-DNA glycosylases. (F) Denaturing PAGE of 1-hr reaction products of *E. coli* YcaQ and HedH4 with 7mG-DNA (left), and *S. bottropensis* TxnU4 and HedH4 with TXNA-DNA (right). (G) Structure of NM_8-ICL. (H) Denaturing PAGE of AZB-ICL unhooking by *S. sahachiroi*

AlkZ and HedH4 (left), and NM_8-ICL unhooking by *E. coli* YcaQ and HedH4 (right). Reactions were treated with buffer (mock) or enzyme for 1 hr, followed by alkaline hydrolysis. MA, monoadduct. (I) HED inhibition of *E. coli* K-12 transformed with *hedH4*/pSF-OXB1 (constitutively expressed) or empty vector pSF-OXB1. The lag time is defined as the time elapsed before cells start to grow exponentially. Data are mean ± SD (n=3). Growth curves are shown in Fig. S3F-G. Significance values were determined by unpaired t test of the mean lag time values (*: $0.05 \leq P \leq 0.01$; ***: $0.001 \leq P \leq 0.0001$). (J) Colony dilution assay for *E. coli* strains ±HedH4 exposed to increasing concentrations of HED for 1 hr. Surviving fraction (%) is relative to untreated cells. Values are mean ± SD (n = 3). Significance values were determined by unpaired t test of the mean sensitivity values (*: $0.05 \leq P \leq 0.01$; **: $0.01 \leq P \leq 0.001$).

Figure 5. *YQL proteins from Actinobacteria hydrolyze simple N7-alkylguanosine lesions and interstrand crosslinks*. (A,B) Denaturing PAGE of reaction products of YQL orthologs from *E. coli* (*Eco*), *Thermomonospora curvata* (*Tcu*), and *Thermobifida fusca* (*Tfu*) with 7mG-DNA (A) and NM_8-ICL (B) after 5 min and 1 hr. Lane 1 of each gel is a no-enzyme control.

Supplementary Tables

Table S1. List of HTH_42 proteins by organism. List of all *Streptomyces* YQL (YcaQ-like), AZL (AlkZ-like), or AZL2 (AlkZ-like 2) proteins in this study, along with the GenBank/RefSeq genome/assembly ID for each organism. Homologs are alphabetized by organism.

Table S2. Proximity of HTH_42 proteins to biosynthetic gene clusters. Results from BGC proximity analysis organized by AZL and YQL protein distances to the nearest antiSMASH-predicted cluster in the species' genome. Nearest cluster upstream (5'/(-)) and/or downstream (3'/(+)) is recorded with the cluster ID, and the most related cluster BLAST hit is denoted with the % gene similarity. *No BGC identified* denotes the cluster BLAST could not find a similar cluster which compares to the hit by homology.

Table S3. AZL proteins found in characterized and uncharacterized biosynthetic gene clusters. *%I/S to AlkZ* (column C) is the percent identity or similarity to *S. sahachiroi* AlkZ. *Cluster BLAST* (column E) is the most similar BGC as determined by cluster BLAST analysis (*% similarity* is the percentage of genes in uncharacterized BGC that have homology to genes in the known similar BGC).

Table S4. Cellular strains, plasmids, and oligodeoxynucleotides used in this study. All oligos were dissolved in TE buffer (10 mM Tris•HCl pH 8.0, 1 mM EDTA pH 8.0) to 200 µM, and the DNA was stored at -20°C (stored in the dark for the Cy5/FAM oligos). The underlined nucleotide in the 7mG_Top, HED_Top, TXNA_Top, AZB_Top/_Bottom, and NM_8_Top/_Bottom oligos is the site of the *N*7-alkylguanine lesion. PCR was performed with a primer concentration of 500 nM.

Table S5. AZL nearest neighbor GO term analysis. Nearest 5 open reading frames (ORFs) upstream (-) (3' → 5') and downstream (+) (5' → 3') of AlkZ-like proteins predicted to be within (A) or outside (B) BGCs. ORFs are listed with their GenBank/RefSeq ID and biological pathway and molecular GO terms as determined by NCBI, GhostKOALA, and eggNOG databases. Empty cells mean no GO terms could be assigned to these proteins through homology search.

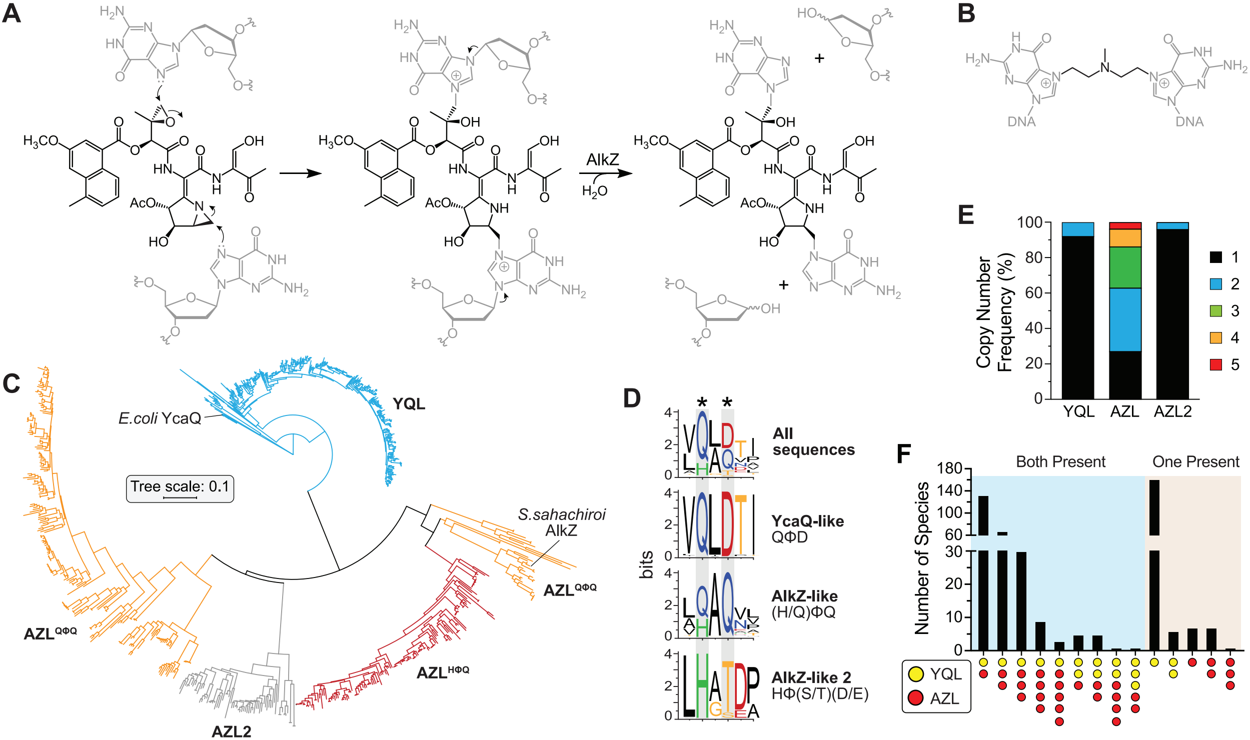
Table S6. YQL nearest neighbor GO term analysis. Nearest 5 open reading frames (ORFs) upstream (-) (3' → 5') and downstream (+) (5' → 3') of YcaQ-like proteins assigned to be outside BGCs. ORFs are listed with their GenBank/RefSeq ID and biological pathway and molecular GO terms as determined by NCBI, GhostKOALA, and eggNOG databases. Empty cells mean no GO terms could be assigned to these proteins through homology search.

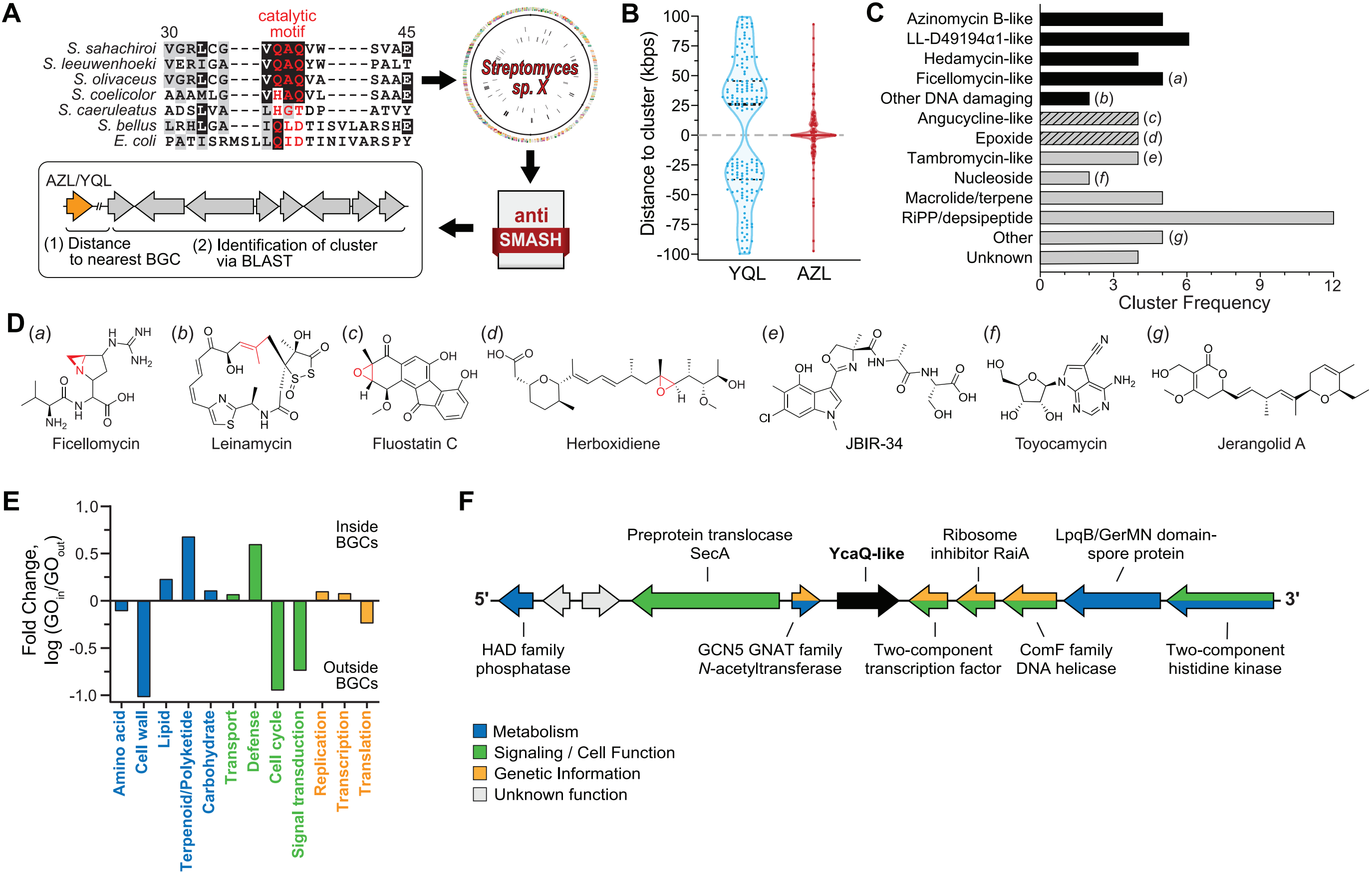
Supplementary Figures

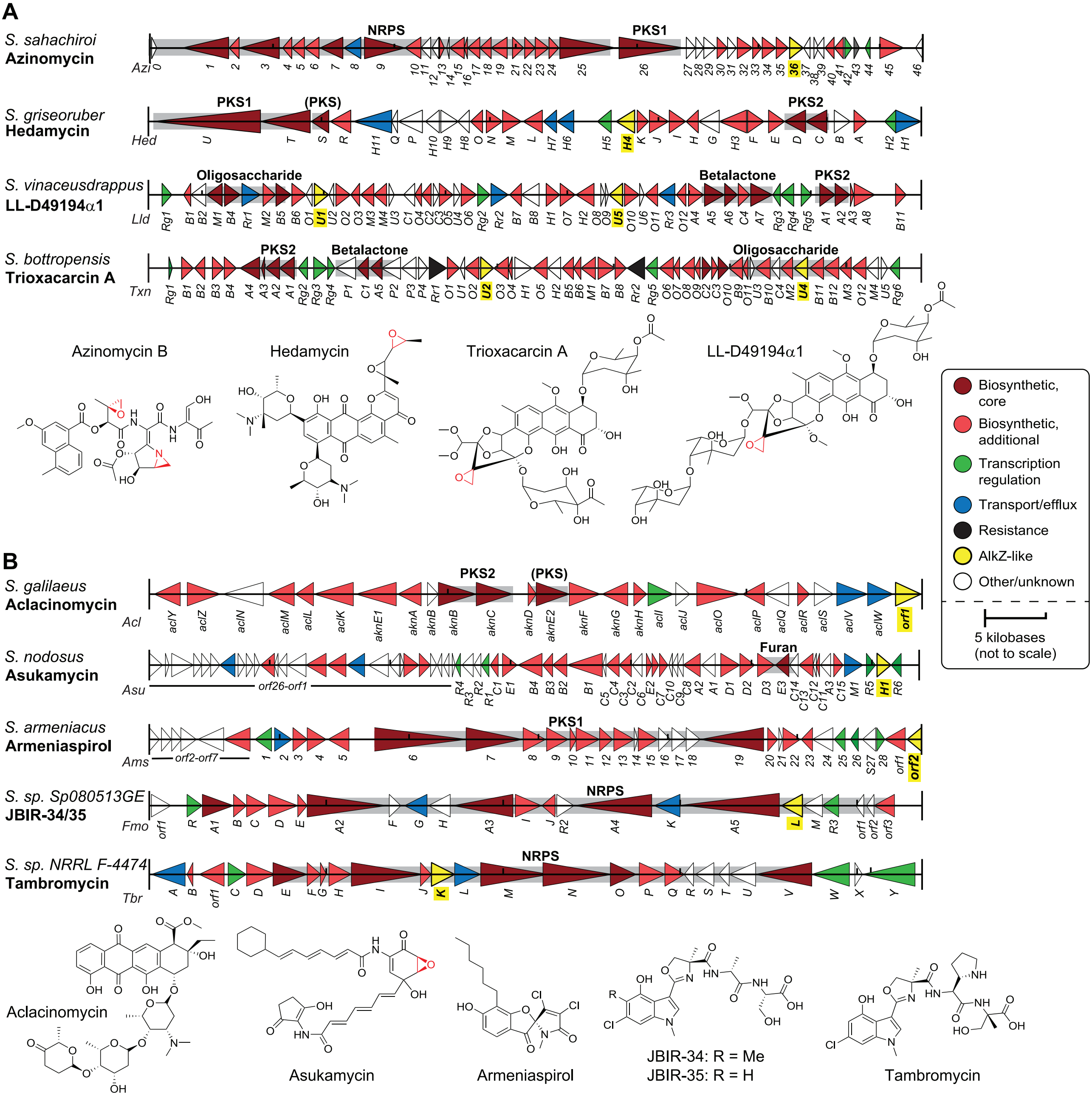
Figure S1. HTH_42 superfamily taxonomy, phylogeny, and copy number analysis. (**A**) Domain schematics for top 5 Pfam classes of HTH_42 superfamily proteins. Number of sequences for each organization is labeled to the right, along with the domain key. Dark vertical lines represent predicted unstructured regions. (**B**) Taxonomic distribution of HTH_42 proteins from prokaryotes (Class, 4,797 sequences). (**C**) Taxonomic distribution of HTH_42 proteins from Actinobacteria (Order, 3,033 sequences). (**D**) Sequence alignment of YcaQ-like (YQL, blue), AlkZ-like (AZL, red; AZL2, yellow) proteins in the region surrounding the catalytic motif (asterisks). *E. coli* YcaQ and *S. sahachiroi* AlkZ are shown at the top of each block as a reference. (**E**) Denaturing PAGE of 5′-Cy5 labeled d7mG-DNA substrate (S) and nicked AP-DNA product (P) after treatment with either buffer (mock), *E. coli* YcaQ, *S. sahachiroi* AlkZ, or *Streptomyces caeruleatus* AZL2. AP-DNA resulting from glycosylase activity was treated with 0.1 M NaOH to generate β,δ-elimination products, which are quantified below the gel.

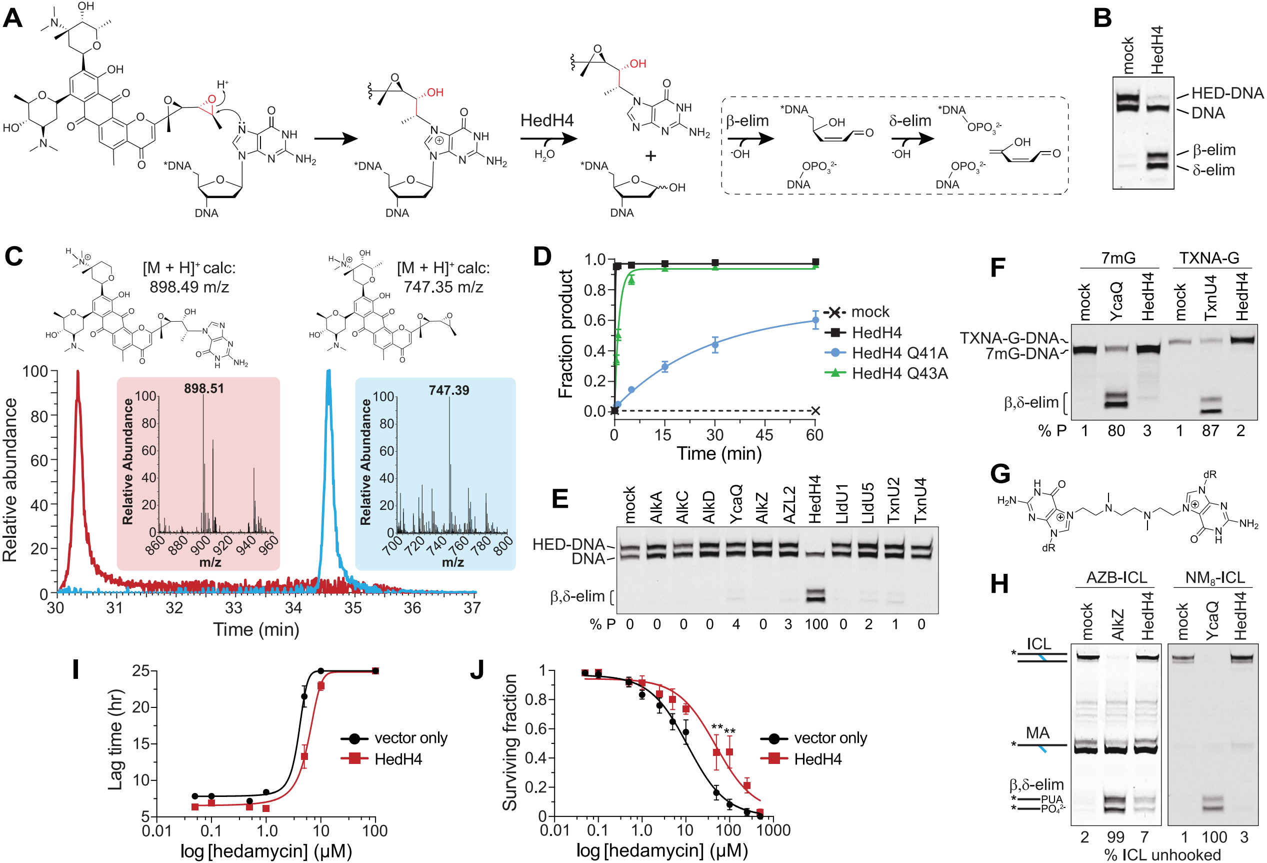
Figure S2. Nearest neighbor analysis of AZL proteins. Gene ontology (GO) analysis for AZL nearest neighbors (± 5 open reading frames) inside (**A**) and outside (**B**) BGCs. Venn diagrams depict the number of neighbors involved in metabolism (blue), signaling and cell function (green), and processing of genetic information (orange). The boxes represent subdivisions of each of the three functions, colored with respect to the key below. Uncharacterized/hypothetical proteins (40 inside, 90 outside) that could not be identified by homology are not included in these data. Full GO term analysis can be found in SI Tables S5-S7.

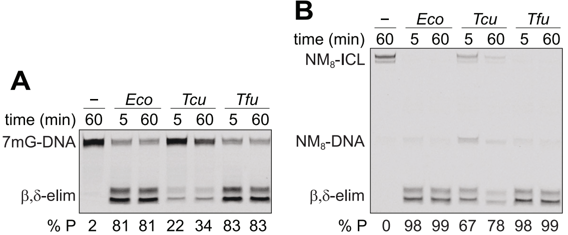
Figure S3. HedH4 biochemistry and cellular resistance. (**A**) Coomassie-stained SDS-PAGE of purified HedH4, *S. sahachiroi* AlkZ, and *E. coli* YcaQ proteins. MW, molecular weight standards. Calculated protein molecular weights are 40.8 kDa (HedH4), 41.2 kDa (AlkZ), and 47.7 kDa (AlkX). (**B**) Thermal and enzyme-catalyzed depurination of HED-DNA adducts. Denaturing PAGE of 5'-Cy5-labeled HED-DNA oligodeoxynucleotide substrate and β- and δ-elimination products formed from hydroxide treatment of the abasic site generated from hydrolysis of the HED-deoxyguanosine N-glycosidic bond. Formation of HED-DNA goes to ~50% completion under our reaction conditions. Lane 1, HED-DNA; lanes 2-3, HED-DNA heated to 95°C for 5 min followed by treatment with either water or NaOH; lanes 4-5, HED-DNA treated with either buffer (mock) or 1 μM HedH4 for 1 hr at 25°C, followed by NaOH. (**C**) Denaturing PAGE of hedamycin excision by HedH4 wild-type and catalytic mutants Q41A and Q43A. Mock, reaction with buffer alone. Quantification of this gel and the replicates are in Fig. 4D. (**D**,**E**) Verification of HedH4 cloning. (**D**) 1% agarose gel of analytical restriction digest of empty pSF-OXB1 and HedH4/pSF-OXB1 using NcoI-HF and XbaI restriction enzymes. Calculated molecular weights for pSF-OXB1 and HedH4 are 3.9 kb and 1.1 kb, respectively. (**E**) 1% agarose gel of colony PCR of HedH4 transformants in *E. coli* using the HedH4 NcoI and XbaI primers (Table S4). Wild-type *E. coli* K-12 served as the negative control, while the protein expression vector HedH4/pBG102 served as a positive control. (**F**,**G**) Growth curves for WT *E. coli* K-12 containing either pSF-OXB1 (**F**) or HedH4/pSF-OXB1 (**G**) grown in LB/Kan media supplemented with increasing concentrations of hedamycin. Values are mean ± SD (n=3).


A
H_2N
HN
N
O
H_3CO
AcO
HO
OH
HN
NH
NH_2
AlkZ
H_2O
+
B
DNA
DNA
C
E.coli YcaQ
YQL
Tree scale: 0.1
S.sahachiroi AlkZ
$AZL^{Q\Phi Q}$
$AZL^{Q\Phi Q}$
$AZL^{H\Phi Q}$
AZL2
D
bits
All sequences
YcaQ-like QΦD
AlkZ-like (H/Q)ΦQ
AlkZ-like 2 HΦ(S/T)(D/E)
E
Copy Number Frequency (%)
0
20
40
60
80
100
YQL
AZL
AZL2
1
2
3
4
5
F
Number of Species
Both Present
One Present
0
10
20
30
60
100
140
180
YQL
AZL


A
catalytic motif
30
45
S. sahachiroi VGRLCG---VQAQVW----SVAE
S. leeuwenhoeki VERIGA---VQAQYW----PALT
S. olivaceus VGRLCG---VQAQVA----SAAE
S. coelicolor AAAMLG---VHAQVL----SAAE
S. caeruleatus ADSLVA---LHGTDP----ATVY
S. bellus LRHLGA---IQLDTISVLARSHE
E. coli PATISRMSLLQIDTINIVARSPY
Streptomyces sp. X
anti SMASH
AZL/YQL
(1) Distance to nearest BGC
(2) Identification of cluster via BLAST
B
Distance to cluster (kbps)
100
75
50
25
0
-25
-50
-75
-100
YQL
AZL
C
Azinomycin B-like
LL-D49194α1-like
Hedamycin-like
Ficellomycin-like (a)
Other DNA damaging (b)
Angucycline-like (c)
Epoxide (d)
Tambromycin-like (e)
Nucleoside (f)
Macrolide/terpene
RiPP/depsipeptide
Other (g)
Unknown
0
3
6
9
12
Cluster Frequency
D
(a) Ficellomycin
(b) Leinamycin
(c) Fluostatin C
(d) Herboxidiene
(e) JBIR-34
(f) Toyocamycin
(g) Jerangolid A
E
Fold Change, log (GO_in/GO_out)
1.0
0.5
0
-0.5
-1.0
Inside BGCs
Outside BGCs
Amino acid
Cell wall
Lipid
Terpenoid/Polyketide
Carbohydrate
Transport
Defense
Cell cycle
Signal transduction
Replication
Transcription
Translation
F
5'
3'
Preprotein translocase SecA
YcaQ-like
Ribosome inhibitor RaiA
LpqB/GerMN domain-spore protein
HAD family phosphatase
GCN5 GNAT family N-acetyltransferase
Two-component transcription factor
ComF family DNA helicase
Two-component histidine kinase
🟦 Metabolism
🟩 Signaling / Cell Function
🟧 Genetic Information
⬜ Unknown function


A
S. sahachiroi
Azinomycin
NRPS
PKS1
Azi
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46
S. griseoruber
Hedamycin
PKS1
(PKS)
PKS2
Hed
U T S R H11 Q P H10 H9 H8 O N M L H7 H6 H5 H4 K J I H G H3 F E D C B A H2 H1
S. vinaceusdrappus
LL-D49194α1
Oligosaccharide
Betalactone
PKS2
Lld
Rg1 B1 B2 M1 B4 Rr1 M2 B5 B6 O1 U1 U2 O2 O3 M3 M4 U3 C1 O4 C2 C3 O5 U4 O6 Rg2 Rr2 B7 B8 H1 O7 H2 O8 O9 U5 O10 U6 O11 Rr3 O12 A4 A5 A6 C4 A7 Rg3 Rg4 Rg5 A1 A2 A3 A8 B11
S. bottropensis
Trioxacarcin A
PKS2
Betalactone
Oligosaccharide
Txn
Rg1 B1 B2 B3 B4 A4 A3 A2 A1 Rg2 Rg3 Rg4 P1 C1 A5 P2 P3 P4 Rr1 O1 U1 O2 U2 O3 O4 H1 O5 H2 B5 B6 M1 B7 B8 Rr2 Rg5 O6 O7 O8 O9 C2 C3 O10 B9 O11 U3 B10 C4 M2 U4 B11 B12 M3 O12 M4 U5 Rg6
Azinomycin B
Hedamycin
Trioxacarcin A
LL-D49194α1
Biosynthetic, core
Biosynthetic, additional
Transcription regulation
Transport/efflux
Resistance
AlkZ-like
Other/unknown
5 kilobases (not to scale)
B
S. galilaeus
Aclacinomycin
PKS2
(PKS)
Acl
aclY aclZ aclN aclM aclL aclK aknE1 aknA aknB aknB aknC aknD aknE2 aknF aknG aknH aclI aclJ aclO aclP aclQ aclR aclS aclV aclW orf1
S. nodosus
Asukamycin
Furan
Asu
orf26-orf1
R4 R3 R2 R1 C1 E1 B4 B3 B2 B1 C5 C4 C3 C2 C6 E2 C7 C10 C9 C8 A2 A1 D1 D2 D3 E3 C14 C13 C12 C11 A3 C15 M1 R5 H1 R6
S. armeniacus
Armeniaspirol
PKS1
Ams
orf2-orf7
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 S27 28 orf1 orf2
S. sp. Sp080513GE
JBIR-34/35
NRPS
Fmo
orf1 R A1 B C D E A2 F G H A3 I J R2 A4 K A5 L M R3 orf1 orf2 orf3
S. sp. NRRL F-4474
Tambromycin
NRPS
Tbr
A B orf1 C D E F G H I J K L M N O P Q R S T U V W X Y
Aclacinomycin
Asukamycin
Armeniaspirol
JBIR-34: R = Me
JBIR-35: R = H
Tambromycin


A
OH
N
O
OH
O
HO
N
O
O
O
O
H+
O
O
N
NH
NH2
N
N
*DNA
O
DNA
N
OH
⊕
NH
NH2
*DNA
DNA
HedH4
H2O
OH
NH
NH2
*DNA
+
OH
DNA
β-elim
-OH
*DNA
OH
O
OPO3 2-
DNA
δ-elim
-OH
*DNA
OPO3 2-
OH
O
OPO3 2-
DNA
B
mock
HedH4
HED-DNA
DNA
β-elim
δ-elim
C
[M + H]+ calc: 898.49 m/z
[M + H]+ calc: 747.35 m/z
Relative abundance
100
80
60
40
20
0
30
31
32
33
34
35
36
37
Time (min)
898.51
Relative Abundance
860
880
900
920
940
960
m/z
747.39
Relative Abundance
700
720
740
760
780
800
m/z
D
Fraction product
1.0
0.8
0.6
0.4
0.2
0.0
0
15
30
45
60
Time (min)
mock
HedH4
HedH4 Q41A
HedH4 Q43A
E
mock
AlkA
AlkC
AlkD
YcaQ
AlkZ
AZL2
HedH4
LldU1
LldU5
TxnU2
TxnU4
HED-DNA
DNA
β,δ-elim
% P
0
0
0
0
4
0
3
100
0
2
1
0
F
7mG
TXNA-G
mock
YcaQ
HedH4
mock
TxnU4
HedH4
TXNA-G-DNA
7mG-DNA
β,δ-elim
% P
1
80
3
1
87
2
G
H2N
N
H
O
N
⊕
N
dR
N
N
dR
N
⊕
N
NH2
O
H
H
AZB-ICL
NM8-ICL
mock
AlkZ
HedH4
mock
YcaQ
HedH4
ICL
*
MA
*
β,δ-elim
*PUA
*PO4 2-
2
99
7
1
100
3
% ICL unhooked
I
Lag time (hr)
25
20
15
10
5
0.01
0.1
1.0
10
100
log [hedamycin] (μM)
vector only
HedH4
J
Surviving fraction
1
0.8
0.6
0.4
0.2
0
0.01
0.1
1.0
10
100
1000
log [hedamycin] (μM)
**
**
vector only
HedH4


A
–
Eco
Tcu
Tfu
time (min) 60 5 60 5 60 5 60
7mG-DNA
β,δ-elim
% P 2 81 81 22 34 83 83
B
–
Eco
Tcu
Tfu
time (min) 60 5 60 5 60 5 60
NM8-ICL
NM8-DNA
β,δ-elim
% P 0 98 99 67 78 98 99