Insights into Modular Polyketide Synthase Loops Aided by Repetitive Sequences

Melissa Hirsch[1], Kaan Kumru[1], Ronak Desai[1], Brendan Fitzgerald[1], Takeshi Miyazawa[1], Katherine Ray[1], Nisha Saif[1], Samantha Spears[1], and Adrian Keatinge-Clay[1]

[1]The University of Texas at Austin

August 28, 2020

Abstract

The loops of modular polyketide synthases (PKSs) serve diverse functions but are largely uncharacterized. They frequently contain amino acid repeats resulting from genetic events such as slipped-strand mispairing. Determining the tolerance of loops to amino acid changes would aid in understanding and engineering these multidomain molecule factories. Here, tandem repeats in the DNA encoding 949 modules within 129 cis-acyltransferase PKSs were catalogued, and the locations of the corresponding amino acids within the module were identified. The most frequently inserted interdomain loop corresponds with the updated module boundary immediately downstream of the ketosynthase (KS), while the loops bordering the dehydratase (DH) were nearly intolerant to such insertions. An analysis of the loops bordering the acyl carrier protein (ACP) reveals they are relatively short (14±6 residues), that they resist large increases in length, and that ACP may rely on acyltransferase (AT) accessing a conformation like that observed through electron microscopy of the pikromycin PKS. From the 949 modules, no repetitive sequence loop insertions are located within ACP, and only 2 reside within KS, indicating the sensitivity of these domains to alteration.

INTRODUCTION

Modular polyketide synthases (PKSs) are among the most powerful molecular machines and synthesize some of the most important medicines (e.g., the antibiotic erythromycin and the anticancer agent epothilone), yet they are also some of the most difficult to study both *in vivo* and *in vitro* [1, 2]. Fortunately, the availability of genomic sequences provides an additional avenue for investigation*in silico* .

Megadalton PKS assembly lines are composed of several large polypeptides that house sets of domains termed modules (Figure 1)[1, 3]. The modules of *cis* -acyltransferase (*cis* -AT) assembly lines minimally contain an AT domain that selects the carbon building blocks (usually from malonyl- or methylmalonyl-CoA), a ketosynthase (KS) domain that fuses the building block with a growing polyketide chain, and an acyl carrier protein (ACP) domain that relays both building blocks and growing chains between the enzymatic domains. Modules may also contain processing enzymes such as a ketoreductase (KR) that reduces the β-keto group formed by KS, a dehydratase (DH) that dehydrates the resulting β-hydroxyacyl chain, and an enoylreductase (ER) that reduces the resulting α,β-unsaturated chain. Other enzymes often appear in the first and last modules to help load primer units and offload mature polyketide chains. Recently, a new module boundary was proposed that matches how sets of domains cooperate and evolutionarily co-migrate[4-6]. KS had been the most upstream domain but is now the most downstream, where it can play the role of a gatekeeper that ensures only properly processed polyketides are passed to the next module. The common module types in*cis* -AT assembly lines are α-modules, which do not contain processing domains, β-modules, which possess a KR, γ-modules, which contain a KR and DH, and δ-modules, which contain a KR, DH, and ER[1]. Modules are commonly split by C-terminal and N-terminal docking domains (^{C}DD and ^{N}DD) located between the ACP and KS[7]. The domains of PKS modules have each been characterized at atomic resolution, and structural biologists are endeavoring to determine their orientation relative to one another within an intact synthase[3, 8]. Less attention has been paid to the loops both between and within domains. While frequently not observed by x-ray crystallography, they play key roles like helping position and dock domains.

Amino acid repeats are present in most PKSs (Figure 2). A comparison of the ˜17,000-residue MlsA1 components of mycolactone PKSs harbored by various mycobacteria reveals sequence identities of 99%[9]. The differences include indels that encode variable numbers of amino acid repeats. For example, at the updated module boundary between the sixth and seventh modules, the sequence GSDPAV is repeated 2 times in *Mycobacterium ulcerans* Agy99 MlsA1, 3 times in *M. ulcerans* subsp. *shinshuense* MlsA1, and 4 times in*Mycobacterium liflandii* 128FXT MlsA1 (Figure 2a). These repetitive sequences are encoded by tandem repeats of the highly conserved 18-mer, 5'-GGTTCTGATCCCGCAGTG-3'. As another example, the loop at the boundary between the second and third modules of the nannocystin PKS from the myxobacterium *Nannocystis* sp. *MB1016* contains two repetitive sequences, one from a 46-mer repeated 3.5 times and the other from a 36-mer repeated 4.9 times (Figure 2b). This module also harbors a remarkable 152-residue insertion between the structural and catalytic subdomains of the KR (KR$_s$ and KR$_c$) resulting from a 48-mer repeated 9.7 times[3].

Each of these repeats is in a loop where it most likely does not impact PKS function, and their origin appears to be genetic, resulting from events such as slipped-strand mispairing[10, 11]. They provide an opportunity to identify the loops of PKSs tolerant to change, which could help elucidate the dynamics of assembly line domains, design better experiments, and engineer hybrid synthases that produce new molecules. Here, the tandem repeats in the genes encoding 949 modules within 129 *cis* -acyltransferase PKSs are catalogued, and the locations of the corresponding amino acids within the module are identified. The most frequently inserted interdomain loop corresponds with the updated module boundary immediately downstream of the ketosynthase (KS). An analysis of the loops bordering ACP reveals they are relatively short and that ACP relies on AT accessing a conformation such as that observed by electron microscopy of the pikromycin PKS[8]. The resistance of the ACP and KS domains to modifications indicates their sensitivity to alteration.

METHODS

Collecting and analyzing sequences

Sequences of the DNA encoding the abyssomicin (Aby, *Verrucosispora maris* AB-18-032, JF752342.1), ajudazol (Aju, *Chondromyces crocatus* , AM946600.1), akaeolide (Aka, *Streptomyces* sp.*NBRC 109706* , BBOM01000011.1), althiomycin (Alm, *Myxococcus xanthus* , FR831800.1), ambruticin (Amb, *Sorangium cellulosum* , DQ897667.1), amphotericin (Amp, *Streptomyces nodosus* , AF357202), anatoxin (Ana, *Oscillatoria sp. PCC 6506* , FJ477836.1), annimycin (Ann, *Streptomyces calvus* , KF683117.1), ansamitocin (Asm, *Actinosynnema pretiosum* subsp.*pretiosum* , KY4899977.1), apoptolidin (Apo, *Nocardiopsis*sp. *FU40* , JF819834.1), aurafuron (Auf, *Stigmatella aurantiaca* DW4/3-1, AM850130.1), aureothin (Aur, *Streptomyces thioluteus* , AJ575648.1), avermectin (Ave, *Streptomyces avermitilis* , AB032367.1), bafilomycin (Baf, *Streptomyces lohii* , GU390405.1), BE-14106 (Bec, *Streptomyces* sp. *DSM 21069* , FJ872523.1), bengamide (Ben, *Myxococcus virescens* , KP143770.1), borrelidin (Bor, *Streptomyces parvulus* , AJ580915.1), calcimycin (Cal, *Streptomyces chartreusis* NRRL 3882, HM452329.1), candicidin (Fsc, *Streptomyces* sp. *FR-008* , AY310323.2), chalcomycin (Chm, *Streptomyces bikiniensis* , AY509120.1), chaxamycin (Cxm, *Streptomyces leeuwenhoekii* , LN831790.1), chlorizidine (Clz, *Streptomyces* sp. *CNH-287* , KF585133.1), chlorothricin (Chl, *Streptomyces antibioticus* , DQ116941.2), chondramide (Cmd, *Chondromyces crocatus* , AM179409.1), chondrochloren (Cnd, *Chondromyces crocatus* , AM988861.1), coelimycin (Cpk, *Streptomyces coelicolor* A3(2), AL645882.2), concanamycin (Con, *Streptomyces neyagawaensis* , DQ149987.1), conglobatin (Cng, *Streptomyces conglobatus* , LN849060.1), cremimycin (Cmi, *Streptomyces* sp. *MJ635-86F5* , AB818354.1), crocacin (Cro, *Chondromyces crocatus* , FN547928.1), cryptophycin (Crp, *Nostoc* sp. *ATCC 53789* , ER159954.1), curacin (Cur,*Moorea producens* 3L, HQ696500.1), cyclizidine (Cyc,*Streptomyces* sp. *NCIB 11649* , KT327068.1), cylindrospermopsin (Cyr, *Cylindrospermopsis raciborskii* AWT205, EU140798.1), cystothiazole (Cta, *Cystobacter fuscus* , AY834753.1), divergolide (Div, *Streptomyces* sp. *HKI0576* , HF563079.1), DKxanthene (Dkx, *Stigmatella aurantiaca* , BN001209.1), E-837 (E837, *Streptomyces aculeolatus* , DQ292520.1), ebelactone (Ebe,*Streptomyces aburaviensis* , KC894072.1), ECO-02301 (Eco,*Streptomyces aizunensis* , AY899214.1), elaiophylin (Ela,*Streptomyces* sp. *ICBB 9297* , GP697151.1), epothilone (Epo,*Sorangium cellulosum* , GU063811.1), erythromycin (Ery,*Saccharopolyspora erythraea* , AM420283.1), FD-891 (Gfs,*Streptomyces graminofaciens* , AB469193.1), filipin (Pte,*Streptomyces avermitilis* MA-4680, BA000030.3), FK520 (Fkb,*Streptomyces hygroscopicus* subsp. *ascomyceticus* , AF235504.1), fostriecin (Fos, *Streptomyces pulveraceus* , HQ434551.1), geldanamycin (Gdm, *Streptomyces hygroscopicus* NRRL 3602, AY179507.1), gephyronic acid (Gph, *Cystobacter violaceum* Cb vi76, KF479198.1), guadinomine (Gdn, *Streptomyces* sp.*K01-0509* , JX545234.1), gulmirecin (Gul, *Pyxidicoccus fallax* , KM361622.1), halstoctacosanolide (Hls, *Streptomyces halstedii* , AB241068.1), hectochlorin (Hct, *Lyngbya majuscula* , AY974560.1), herbimycin (Hbm, *Streptomyces hygroscopicus* , AY947889.1), herboxidiene (Her, *Streptomyces chromofuscus* , JN671974.1), hitachimycin (Hit, *Streptomyces scabrisporus* , LC008143.1), hygrocin (Hgc, *Streptomyces* sp. *LZ35* , JX504844.1), incednine (Idn, *Streptomyces* sp. *ML694-90F3* , AB767280.1), indanomycin (Idm, *Streptomyces antibioticus* , FJ545274.1), jamaicamide (Jam, *Lyngbya majuscula* , AY522504.1), jerangolid (Jer, *Polyangium cellulosum* , DQ897668.1), kendomycin (Ken, *Streptomyces violaceoruber* , AM992894.1), kijanimicin (Kij,*Actinomadura kijaniata* , EU301739.1), lankamycin (Lkm,*Streptomyces rochei* , AB088224.2), lasalocid (Lsd,*Streptomyces lasaliensis* , AB449340.1), leupyrrin (Leu,*Sorangium cellulosum* , HM639990.1), lipomycin (Lip,*Streptomyces aureofaciens* , DQ176871.1), lobophorin (Lbp,*Streptomyces* sp. *SCSIO 01127* , KC013978.1), lobosamide (Lob, *Micromonospora* sp. *RL09-050-HVF-A* , KT209587.1), lorneic acid (Lor, *Streptomyces* sp. *NBRC 109706* , BBOM01000004.1), macbecin (Mbc, *Actinosynnema pretiosum* , EU827593.1), maklamicin (Mak, *Micromonospora* sp. *GMKU326* , LC021382.1), meilingmycin (Mei, *Streptomyces nanchangensis* , FJ952082.1), melithiazol (Mel, *Melittangium lichenicola* , AJ557546.1), meridamycin (Mer, *Streptomyces* sp. *NRRL 30748* , DQ351275.1), microcystin (Mcy, *Planktothrix agardhi*NIVA-CYA 126/8, AJ441056.1), microsclerodermin (Msc, *Jahnella* sp.*MSr9139* , KF657739.1), ML-449 (Mla, *Streptomyces* sp.*MP39-85, FJ372525.1*), monensin (Mon, *Streptomyces cinnamonensis* , AF440781.1), mycinamicin (Myc, *Micromonospora griseorubida* , AB089954.2), mycolactone (Mls, *Mycobacterium ulcerans* Agy99, BX649209.1), myxalamid (Mxa, *Stigmatella aurantiaca* , AF319998.1), myxothiazol (Mta, *Stigmatella aurantiaca* DW4/3-1, AF188287.1), nanchangmycin (Nan, *Streptomyces nanchangensis* , AF521085.1), nannocystin (Ncy, *Nannocystis* sp.*MB1016* , KT067736.1), naphthomycin (Nat, *Streptomyces* sp.*CS* , GQ452266.1), neoaureothin (Nor, *Streptomyces orinoci* , AM778535.1), niddamycin (Nid, *Streptomyces caelestis* , AF016585.1), nigericin (Nig, *Streptomyces violaceusniger* , DQ354110.1), nocardiopsin (Nsn, *Nocardiopsis* sp.*CMB-M0232* , KP339942.1), oligomycin (Olm, *Streptomyces avermitilis* , AB070940.1), nystatin (Nys, *Streptomyces norsei*ATCC 11455, AF263912.1), pellasoren (Pel, *Sorangium cellulosum* , HE616533.1), phenylnannolone (Phn, *Nannocystis pusilla* , KF739396.1), phoslactomycin (Plm, *Streptomyces* sp. *HK803* , AY354515.1), piericidin (Pie, *Streptomyces piomogenus* , HQ840721.1), pimaricin (Pim, *Streptomyces natalensis* , AJ278573.1), pikromycin (Pik, *Streptomyces venezuelae* , AF079138.1), pladienolide (Pld, *Streptomyces platensis* , AB435553.1), puwainaphycin (Puw, *Cylindrospermum alatosporum*CCALA 988, KM078884.1), pyoluteorin (Plt, *Pseudomonas protegens*Pf-5, AF081920.3), quartromicin (Qmn, *Amycolatopsis orientalis* , JF970188.1), rapamycin (Rap, *Streptomyces rapamycinicus* NRRL 5491, X86780.1), reveromycin (Rev, *Streptomyces* sp.*SN-593* , AB568601.1), rifamycin (Rif, *Amycolatopsis mediterranei* S699, AF040570.3), rubradirin (Rub, *Streptomyces achromogenes* subsp. *rubradiris* , AJ871581.1), salinilactam (Slm,*Salinispora tropica* CNB-440, CP000667.1), salinomycin (Sln,*Streptomyces albus* , JN033543.1), sanglifehrin (Sfa,*Streptomyces flaveolus* , FJ809786.1), soraphen (Sor,*Sorangium cellulosum* , U24241.2), spinosyn (Spn,*Saccharopolyspora spinosa* , AY007564.1), spirangien (Spi,*Sorangium cellulosum* , AM407731.1), stambomycin (Sta,*Streptomyces ambofaciens* ATCC 23877, AM238664.2), stigmatellin (Sti, *Stigmatella aurantiaca* Sg a15, AJ421825.1), streptazone (Stz, *Streptomyces* sp. *MSC090213JE08* , LC051217.1), streptolydigin (Slg, *Streptomyces lydicus* , FN433113.1), tautomycetin (Ttn, *Streptomyces* sp. *MSC090213JE08* , LC061217.1), tautomycin (Ttm, *Streptomyces spiroverticillatus* , EF990140.1), tetrocarcin (Tca, *Micromonospora chalcea* , EU443633.1), tetronasin (Tsn, *Streptomyces longisporoflavus* , FJ462704.1), tetronomycin (Tmn, *Streptomyces* sp. *NRRL 11266* , AB193609.1), thuggacin (Tga, *Sorangium cellulosum* , GQ981380.1), tiacumicin (Tia, *Dactylosporangium aurantiacum*subsp.

hamdenensis , HQ011923.1), tirandamycin (Tam,*Streptomyces* sp. *307-9* , GU385216.1), tubulysin (Tub,*Cystobacter* sp. *SBCb004* , GU002154.1), tylosin (Tyl,*Streptomyces fradiae* , U78289.1), versipelostatin (Vst,*Streptomyces versipellis* , LC006086.1), vicenistatin (Vin,*Streptomyces halstedii* , AB086653.1), and zwittermycin (Zma,*Bacillus cereus* , FJ430564.1) assembly lines were primarily obtained from MIBiG[12]. A FASTA file of the DNA encoding extension modules (the "GTNAH" motif near the end of the KS domain was used as the boundary) as well as the corresponding amino acid sequences was generated. Modules were named by the polypeptide containing its AT, its position within that polypeptide, and its type (e.g., the module EryA1_3b has its AT in the first polypeptide of the erythromycin PKS in the third position and is a β-module). Modules were divided into eight categories: α-modules without DD (n=62), α-modules with DD (n=9), β-modules without DD (n=168), β-modules with DD (n=111), γ-modules without DD (n=258), γ-modules with DD (n=183), δ-modules without DD (n=73), and δ-modules with DD (n=85). Those encoded on two polypeptides connected through a DD were treated as one sequence (5 C's represent the C-terminus of the first polypeptide and 5 N's represent the N-terminus of the second polypeptide). The biosynthetic model for each PKS helped determine which polypeptides contain the upstream and downstream portions of these split modules[1] (Supplementary Data Files 1-9).

The Tandem Repeats Finder server was employed to detect tandem repeats at the DNA level [advanced mode with alignment parameters (match, mismatch, indels): 2, 5, 7; minimum alignment score: 50; maximum period size: 50; maximum tandem repeat array size (bp, millions): 2][13]. These DNA repeats as well as the corresponding amino acids were made lowercase and sequentially highlighted yellow, green, and cyan (Supplementary Data File 1).

While this was sufficient for repetitive sequences located between domains, curation was necessary within domains to catalog repetitive sequences that alter the composition and/or length of known loops. Thus, repetitive sequences located in regions that correspond to structured elements were made uppercase. Likewise, when it was unclear whether a repetitive sequence was located in a (e.g., in the poorly conserved regions of KR$_s$ from γ/δ-modules), it was made uppercase. To help determine whether a repetitive sequence is in a known loop, multiple sequence alignments were generated with the program SEAVIEW (using the Clustal Omega algorithm) and ESPript with the aid of known structures [AT region (EryAT4, PDB 2QO3), KS (EryKS3, PDB 2QO3), KR region of β-modules (SpnKR4, PDB 4IMP), KR of γ/δ-modules (SpnKR3, PDB 3SLK), DH of γ/δ-modules, ER of δ-modules (SpnER3, PDB 3SLK), and ACP (MycACP8 from MlsB, PDB 6H0Q)] (Supplementary Figures 1-11) [14-16]. All lowercase repetitive sequences, along with their period size and repeat number, were tabulated (Tables I-II).

Inverted repeats were detected using the EMBOSS palindrome server (minimum length of palindrome: 10; maximum length of palindrome: 100; gap between repeated regions: 50; mismatches allowed: 0)[17]. Repeats were highlighted in magenta or red. Those separated by more than 10 bases were italicized to indicate a lower likelihood of being biologically significant (Supplementary Data File 1).

Constructing a model assembly line

A model of an assembly line composed of α-δ modules was constructed using the structures from the ESPript analysis. Domains were relatively positioned with PyMOL and polyalanine loops were built with Coot[18, 19]. The relative orientations of the domains were acquired from various crystal structures[20-22]; however, a 7-Å resolution electron microscopy reconstruction of a β-module with traditional boundaries from the pikromycin PKS shows a conformation of AT that is much closer to ACP and may be more accurate [8]. The coordinates of the model assembly line shown here can be freely downloaded from http://keatinge-clay.cm.utexas.edu/research/.

RESULTS

The DNA and amino acid sequences were obtained for 949 modules from 129 PKSs. Eight multiple protein sequence alignments (α-δ modules ± DD) were generated with ESPript to show the locations of known secondary structures [1, 3, 12-15] (Supplementary Figures 1-8). Tandem repeats were detected at the DNA level, and the amino acids encoded by them were examined on the protein level.

Interdomain insertions

Repetitive sequences in the loops between domains occur with the greatest frequency at the updated module boundary, where KS connects with the flanking subdomain (FSD, also referred to as the KS-AT adapter) (Figure 3, Table I)[20, 23]. This loop tolerates significant increases in length, as in NcyB_1c where it is 137 residues. Repetitive sequences occur least frequently in the loops upstream and downstream of DH and in the loop connecting ER and KR$_c$. Unless DH η1 is absent, the loop connecting the end of the AT region (the LPTYxFx$_5$W motif) to DH is strictly 5-7 residues. The composition and length of the loop connecting DH to KR$_s$are not significantly altered through the two observed insertions (SpiE_2c and TcaA4_1c). The ER-KR$_c$ loop is usually 2-3 residues long, as observed in the crystal structure of KR$_s$-ER-KR$_c$ from the third module of the spinosyn PKS (PDB 3SLK; although it is significantly longer in 3% of δ-modules, including StiF_1d where it is 16 residues)[21].

To determine the length of flexible interdomain loops, their interfaces with structured elements were identified. Absences of optional features such as DH η1, DH α4 (the terminal helix observed in a cremimycin DH, PDB 6K97), and ACP α1 were considered [24-26]. Which residues are unstructured in the loop between KR and ACP in β-modules is apparent; however, which residues are unstructured in the loop between these domains in γ- and δ-modules is less clear. Since 2 consecutive arginines are conserved 4 residues downstream of what was structurally-observed in KR$_s$-ER-KR$_c$ from the third module of the spinosyn PKS (PDB 3SLK), this region was considered to be structured [21].

The loops upstream and downstream of ACP are of particular interest in determining how much translational and rotational freedom an ACP domain possesses to access each of its cognate enzymes. Thus, the lengths of the loops upstream and downstream of ACP were analyzed in β-, γ-, and δ-modules not containing DDs. In β-modules, their lengths are 11 ± 4 and 14 ± 4 residues (median ± standard deviation), respectively; in γ-modules, 18 ± 3 and 13 ± 8 residues; and in δ-modules, 8 ± 1 and 14 ± 2 residues. Insertions, some of which likely resulted from slipped-strand mispairing, infrequently do increase the length of these loops; the longest observed upstream and downstream ACP linkers are, respectively, 26 and 33 residues (SlgA2_1b and BafA4_1b) in β-modules, 25 and 50 residues (Pik1_3c and PldA2_2c) in γ-modules, and 30 and 25 residues (FosA_2d and RapC_3d) in δ-modules. For comparison, the average length of linkers in the related *trans* -AT assembly lines, in which ACPs access separately-encoded *trans* -ATs, is 50 residues[2]. Instead of a flexible loop between ACP and KS, FscC_5c contains a fused Class 1b docking domain[7].

Intradomain insertions

Loops are present within domains as well, usually on the surface between two secondary structure elements. From the point of view of a protein engineer seeking to insert protease cut sites, purification tags, or domains, known loops tolerant of significant changes to their composition and/or length were catalogued; thus, repetitive elements that alter secondary structures or undefined structures were not considered (Figure 4, Table II).

AT region [including FSD, n=949): Most of the observed insertions (19/34) were before the first β-strand or after the last β-strand of AT, in the FSD [especially α1-α2 (8) and α3-β3 (10), but also β3-β4 (5) and the loop following η4 (4)]. Within the AT domain, the loops that tolerated the most insertions are α10-β7 (3) and α15-β12 (3).

DH [n=599]: The longest loop, β9-β10, is also the most frequently inserted (16 instances, including a GT dipeptide repeated 20 times in the cyclizidine PKS). The next most frequently inserted loop is η5-β14 (6), adjacent to the active site.

KR region of β-modules [including the dimerization element (DE), n=279]: None of the 3-helix DE, present in 77% of the β-modules, contain insertions [27]. Although KR$_s$ and KR$_c$ are similarly sized, KR$_s$ contains significantly more insertions (19 vs. 2), with β3-α5 (5) and η1-β6 (5) being the most frequently inserted loops.

KR from γ- and δ-modules [n=599]: Even with the region upstream of α2 in KR$_s$ not being analyzed due to low sequence conservation, KR$_s$ contains more insertions than KR$_c$ (18 vs. 4). The most frequently inserted loop of KR$_c$ is α6-β6 (2).

ER [n=158]: Each of the 4 observed insertions are located in the N-terminal portion of the substrate-binding subdomain.

ACP [n=949]: No insertions were observed within this 100-residue, helical domain. This analysis includes α1 (often referred to as "helix 0"), which is rarely absent[26].

DDs [n=388]: Most of the docking domain motifs,^{C}DD and ^{N}DD, could be grouped into Class 1a (n=226), Class 1b (n=63), or Class 2 (n=70) (Supplementary Figures 9-11)[7]. No insertions were observed in the Class 2 docking motifs. The majority of the insertions in Class 1a and Class 1b^{C}DDs were immediately upstream of the terminal helix (8/12 and 4/6). The only ^{N}DDs that possess insertions at their upstream end belong to Class 1a (3).

KS [n=949]: Two insertions were observed, both in β13-β14, the most downstream loop.

Inverted Repeats

Inverted repeats are another form of repetitive DNA sequence[28]. Perfect inverted repeats of at least 10 bases separated by 50 residues or less were catalogued[17]. Few repeats longer than 15 bases were observed. The locations of the encoded amino acids are distributed throughout the module, in loops and secondary structures, and usually did not significantly change the composition or length of the protein sequence (Supplementary Data File 1).

DISCUSSION

Assembly line PKSs, which represent some of the largest known proteins, contain modules and domains connected by loops that are experimentally challenging to study. To better understand the dynamics of PKSs and facilitate their engineering, we sought to better characterize these loops. As the insertion of repetitive sequences naturally occurs throughout the DNA encoding PKS assembly lines, cataloguing where insertions persist provides a measure for the tolerance of the encoded loops to modification.

Out of all of the loops in assembly line PKSs, the most tolerant to insertion is at the updated module boundary, between KS and FSD[4-6]. Even 137 unstructured residues at this location seems not to impair the function of the nannocystin module NcyB_1c. Connecting modules at this junction may preserve how its component domains structurally and functionally work together, as supported by recent*in vivo* and *in vitro* PKS engineering studies [29, 30]. Although this junction is upstream of the AT domain, AT swapping has been naturally observed and accomplished through engineering (by including FSD)[4, 31]. More studies are needed to elucidate the connection between the AT region and the downstream domains of the module.

The AT-DH loop of γ/δ-modules is more sensitive to modification than the AT-KR loop of β-modules. Out of 599 examined modules, there are no insertions in the AT-DH loop, nor are there deviations from its length of 5-7 residues when n1 is present. One explanation is that this loop is structured at the interface between KS and DH. This would mean that this connection is not hinge-like as in the highly related metazoan fatty acid synthase[32]. Other connections between enzymatic domains, in particular DH-KR$_s$ and ER-KR$_c$, are also resistant to modification. Perhaps these loops help position the catalytic sites of the module relative to ACP.

The loops upstream and downstream of ACP (15 ± 5 and 13 ± 6 residues, for β-δ modules) may need to be short to prevent docking with non-cognate enzymes, such as AT of the downstream module; however, these tethers limit the degrees of freedom available to ACP. Surprisingly, δ-modules, which contain the most enzymatic domains, possess the shortest linkers (8 ± 1 and 14 ± 2 residues). As ACP cannot stretch to access its AT in the assembly line as presented (Figure 1), AT may need to access another conformation. One possibility is that ATs adopt the position observed in the electron microscopy reconstructions of a construct from the pikromycin PKS[8]. Another possibility is that a hinge between AT and the processing enzymes (immediately after the LPTYxFx$_5$W motif) facilitates these docking interactions[32]; however, this requires large, asymmetric motions throughout the assembly line that would hamper polyketide production compared to a more rigid, symmetric assembly line. In the yeast fatty acid synthase, active sites are fixed around reaction chambers in which the only mobile domain is ACP[33].

Those studying and engineering a PKS may wish to make changes to its polypeptides. Sequence alignments do not always sufficiently indicate the tolerance of loops to modification. As the repetitive sequences catalogued here are not likely to be functionally relevant, they serve as beacons illuminating where changes can be made both between and within domains. The reported repetitive sequence insertions could help in positioning 3C protease cut sites on each end of an ACP to liberate it from an assembly line for analysis by mass spectrometry, installing purification tags upstream of Class 1a ^{N}DDs or elsewhere, and adding domains such as fluorescent proteins or polyketide processing enzymes at desired locations.

As essential as loops are to the proper functioning of assembly line PKSs, they have been relatively ignored compared to PKS domains. However, they are informative as to how far domains can move from one another and provide restraints for structural biologists attempting to elucidate assembly line architecture. Some loops may be more structured than previously thought, such as the residues at the AT/DH interface and those downstream of KRs from γ/δ-modules. With the abundant sequence information that is now available the tolerance of loops to modification can be evaluated through bioinformatics. The resistance of domains such as ACP and KS to these modifications also reveals the importance of their surfaces for domain-domain and docking interactions. Complementing what is known about PKS domains by studying the loops that connect them and are present within them has advanced our understanding of PKS assembly lines as well as our ability to engineer them.

FIGURES

Figure 1. Loops connect domains into modules and modules into PKSs. a) A model assembly line colored by domain types. Acyltransferases (ATs, orange) add extender units to acyl carrier proteins (ACPs, salmon), which collect the growing acyl chain from ketosynthases (KSs, red) through carbon-carbon bond formation. Ketoreductases (KRs, blue), dehydratases (DHs, yellow), and enoylreductases (ERs, green) are often present to modify the α- and β-positions of the extended polyketide chain. A stereodiagram shows how these domains may be relatively oriented within an assembly line. b) The model assembly line colored by type. While α-modules do not contain processing enzymes, β-modules contain a KR, γ-modules contain both a KR and DH, and δ-modules contain a KR, DH, and an ER. The updated definition of the module places KS at the most downstream position. Docking domains (DDs) are present between the ACP and KS domains in 41% of the 949 modules studied here.

Figure 2. Examples of repetitive sequences. a) The MlsA1 polypeptides from the mycolactone PKSs of mycobacteria are 99% identical yet show variation in the number of GSDPAV repeats in the loops between modules. The DNA sequence encoding these repeats is the highly conserved 18-mer, 5'-GGTTCTGATCCCGCAGTG-3'. ml,*Mycobacterium liflandii* ; ms, *Mycobacterium ulcerans* subsp.*shinshuense* ; mu, *Mycobacterium ulcerans* Agy99. b) Two repetitive sequences are present in the loop between the second and third modules of the nannocystin PKS (encoded by a 46-mer repeated 3.5 times and a 36-mer repeated 4.9 times). Between the KR structural and catalytic subdomains (KR$_s$ and KR$_c$) of the same module a 48-mer repeated 9.7 times encodes a 152-residue insertion. Only the inserted regions of the alignment of module NcyB_1c with the reference γ-module (Cmod) are shown. Genetic events such as slipped-strand mispairing may give rise to these seemingly innocuous repetitive sequences.

Figure 3. Repetitive sequences between domains and modules. a) A heat map shows how tolerant or intolerant interdomain loops are to the insertion of repetitive sequences. The most tolerant loops are between modules, while the least tolerant are at the AT-DH, DH-KR$_s$, and ER-KR$_c$ junctions in γ- and δ-modules. A bar graph compares the frequencies of loop insertion. b) A stereodiagram of the model assembly line shows the location of interdomain loops, colored red.

Figure 4. Repetitive sequences in the loops of domains. The locations of loops that have been significantly altered compositionally or lengthwise are indicated in red. Frequencies of insertion are reported in the

accompanying bar graphs.

SUPPLEMENTARY INFORMATION

Supplementary Data 1. DNA and amino acid sequences for 949 modules grouped by module type. Tandem repeats and corresponding amino acids are sequentially highlighted yellow, green, and cyan. Loop insertions appear in lowercase. Inverted repeats and corresponding amino acids are highlighted in purple and red (italicized when repeats separated by more than 10 bases).

Supplementary Data 2-9 . Unannotated amino acid sequences of the 949 modules for each module type (α-DD, α+DD, β-DD, β+DD, γ-DD, γ+DD, δ-DD, and δ+DD, respectively).

Supplementary Figures 1-8. ESPript multiple sequence alignments annotated with known secondary structures for α-DD, α+DD, β-DD, β+DD, γ-DD, γ+DD, δ-DD, and δ+DD modules, respectively. Lowercase residues indicate repetitive sequence insertions.

Supplementary Figures 9-11. Multiple sequence alignments of Class 1a, Class 1b, and Class 2 DDs. Lowercase residues indicate repetitive sequence insertions.

ACKNOWLEDGMENTS

This research was supported by NIH (GM106112) and the Robert A. Welch Foundation (F-1712). We thank Shaochen You for helping collect and analyze sequences.

CONFLICT OF INTEREST

The authors have no conflict of interest to declare.

REFERENCES

1. Keatinge-Clay, A.T., *The Uncommon Enzymology of Cis-Acyltransferase Assembly Lines.* Chem Rev, 2017. **117** (8): p. 5334-5366.

2. Helfrich, E.J. and J. Piel, *Biosynthesis of polyketides by trans-AT polyketide synthases.* Nat Prod Rep, 2016. **33** (2): p. 231-316.

3. Keatinge-Clay, A.T., *The structures of type I polyketide synthases.* Nat Prod Rep, 2012. **29** (10): p. 1050-73.

4. Zhang, L., et al., *Characterization of Giant Modular PKSs Provides Insight into Genetic Mechanism for Structural Diversification of Aminopolyol Polyketides.* Angew Chem Int Ed Engl, 2017.**56** (7): p. 1740-1745.

5. Keatinge-Clay, A.T., *Polyketide Synthase Modules Redefined.*Angew Chem Int Ed Engl, 2017. **56** (17): p. 4658-4660.

6. Vander Wood, D.A. and A.T. Keatinge-Clay, *The modules of trans-acyltransferase assembly lines redefined with a central acyl carrier protein.* Proteins, 2018. **86** (6): p. 664-675.

7. Whicher, J.R., et al., *Cyanobacterial polyketide synthase docking domains: a tool for engineering natural product biosynthesis.*Chem Biol, 2013. **20** (11): p. 1340-51.

8. Dutta, S., et al., *Structure of a modular polyketide synthase.*Nature, 2014. **510** (7506): p. 512-7.

9. Pidot, S.J., et al., *Deciphering the genetic basis for polyketide variation among mycobacteria producing mycolactones.* BMC Genomics, 2008. **9** : p. 462.

10. Taylor, J.S. and F. Breden, *Slipped-strand mispairing at noncontiguous repeats in Poecilia reticulata: a model for minisatellite birth.* Genetics, 2000. **155** (3): p. 1313-20.

11. Zhou, K., A. Aertsen, and C.W. Michiels, *The role of variable DNA tandem repeats in bacterial adaptation.* FEMS Microbiol Rev, 2014.**38** (1): p. 119-41.

12. Kautsar, S.A., et al., *MIBiG 2.0: a repository for biosynthetic gene clusters of known function.* Nucleic Acids Res, 2020.**48** (D1): p. D454-D458.

13. Benson, G., *Tandem repeats finder: a program to analyze DNA sequences.* Nucleic Acids Res, 1999. **27** (2): p. 573-80.

14. Galtier, N., M. Gouy, and C. Gautier, *SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny.*Comput Appl Biosci, 1996. **12** (6): p. 543-8.

15. Sievers, F., et al., *Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega.* Mol Syst Biol, 2011. **7** : p. 539.

16. Robert, X. and P. Gouet, *Deciphering key features in protein structures with the new ENDscript server.* Nucleic Acids Res, 2014.**42** (Web Server issue): p. W320-4.

17. Rice, P., I. Longden, and A. Bleasby, *EMBOSS: the European Molecular Biology Open Software Suite.* Trends Genet, 2000.**16** (6): p. 276-7.

18. *The PyMOL Molecular Graphics System, Version 1.2r3pre, Schrödinger, LLC.*

19. Emsley, P., et al., *Features and development of Coot.* Acta Crystallogr D Biol Crystallogr, 2010. **66** (Pt 4): p. 486-501.

20. Tang, Y., et al., *The 2.7-Angstrom crystal structure of a 194-kDa homodimeric fragment of the 6-deoxyerythronolide B synthase.*Proc Natl Acad Sci U S A, 2006. **103** (30): p. 11124-9.

21. Zheng, J., et al., *Divergence of multimodular polyketide synthases revealed by a didomain structure.* Nat Chem Biol, 2012.**8** (7): p. 615-21.

22. Herbst, D.A., et al., *Mycocerosic acid synthase exemplifies the architecture of reducing polyketide synthases.* Nature, 2016.**531** (7595): p. 533-7.

23. Gay, D.C., et al., *A close look at a ketosynthase from a trans-acyltransferase modular polyketide synthase.* Structure, 2014.**22** (3): p. 444-51.

24. Kawasaki, D., et al., *Functional and Structural Analyses of the Split-Dehydratase Domain in the Biosynthesis of Macrolactam Polyketide Cremimycin.* Biochemistry, 2019. **58** (48): p. 4799-4803.

25. Keatinge-Clay, A., *Crystal structure of the erythromycin polyketide synthase dehydratase.* J Mol Biol, 2008. **384** (4): p. 941-53.

26. Moretto, L., et al., *Modular type I polyketide synthase acyl carrier protein domains share a common N-terminally extended fold.* Sci Rep, 2019. **9** (1): p. 2325.

27. Zheng, J., et al., *The missing linker: a dimerization motif located within polyketide synthase modules.* ACS Chem Biol, 2013.**8** (6): p. 1263-70.

28. Bikard, D., et al., *Folded DNA in action: hairpin formation and biological functions in prokaryotes.* Microbiol Mol Biol Rev, 2010.**74** (4): p. 570-88.

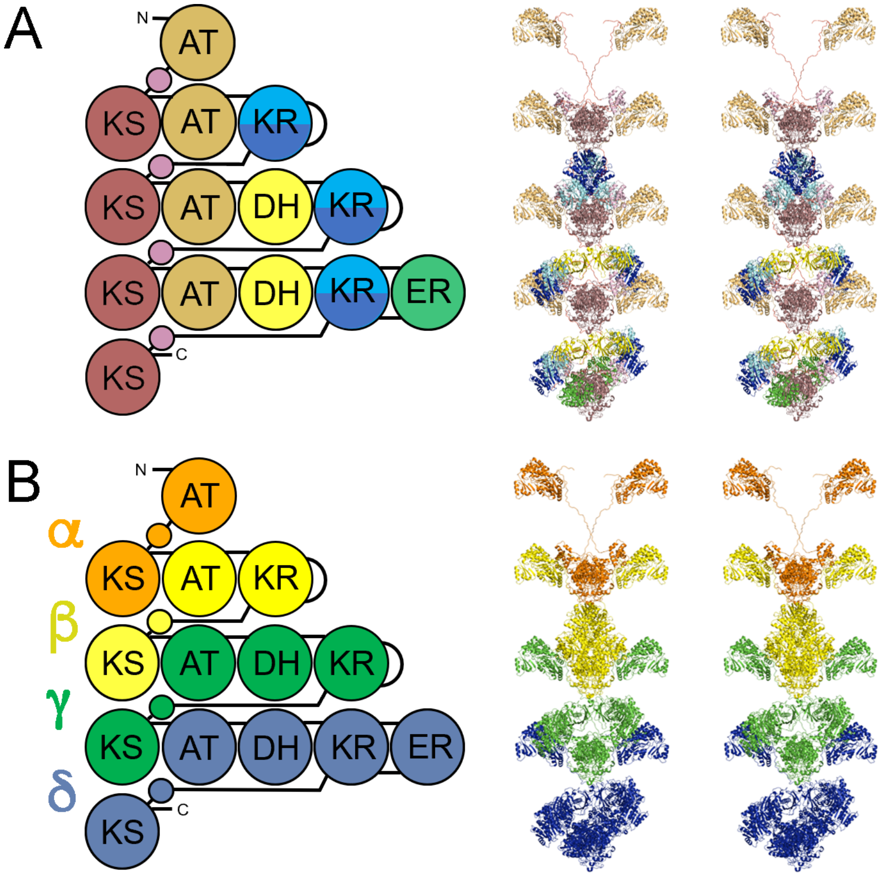
29. Peng, H., et al., *Emulating evolutionary processes to morph aureothin-type modular polyketide synthases and associated oxygenases.*Nat Commun, 2019. **10** (1): p. 3918.

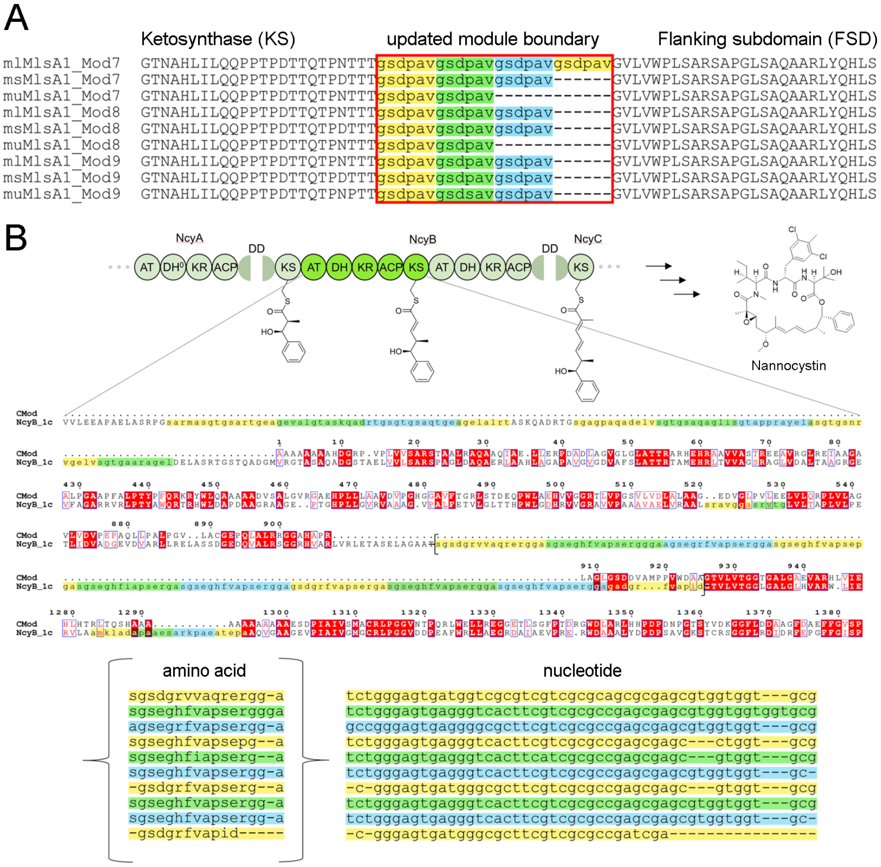
30. Miyazawa, T., et al., *An in vitro platform for engineering and harnessing modular polyketide synthases.* Nat Commun, 2020.**11** (1): p. 80.

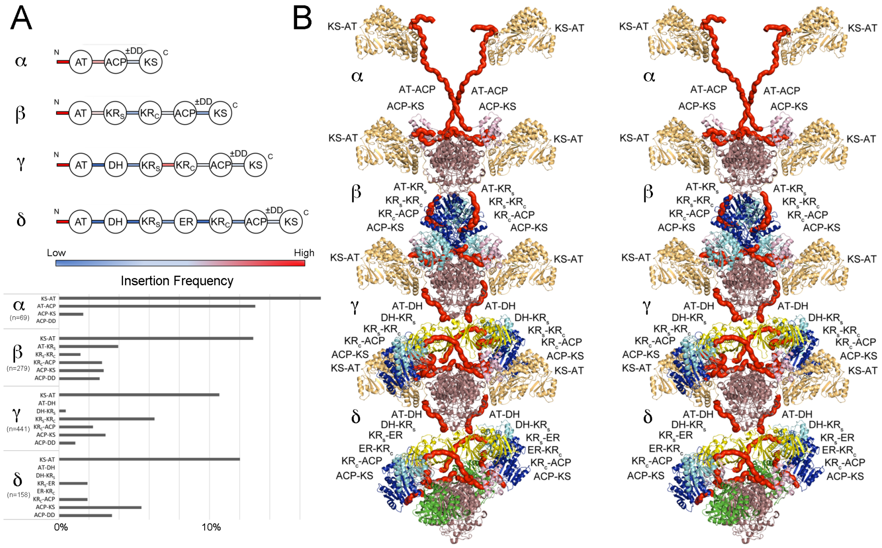
31. Yuzawa, S., et al., *Comprehensive in Vitro Analysis of Acyltransferase Domain Exchanges in Modular Polyketide Synthases and Its Application for Short-Chain Ketone Production.* ACS Synth Biol, 2017.**6** (1): p. 139-147.

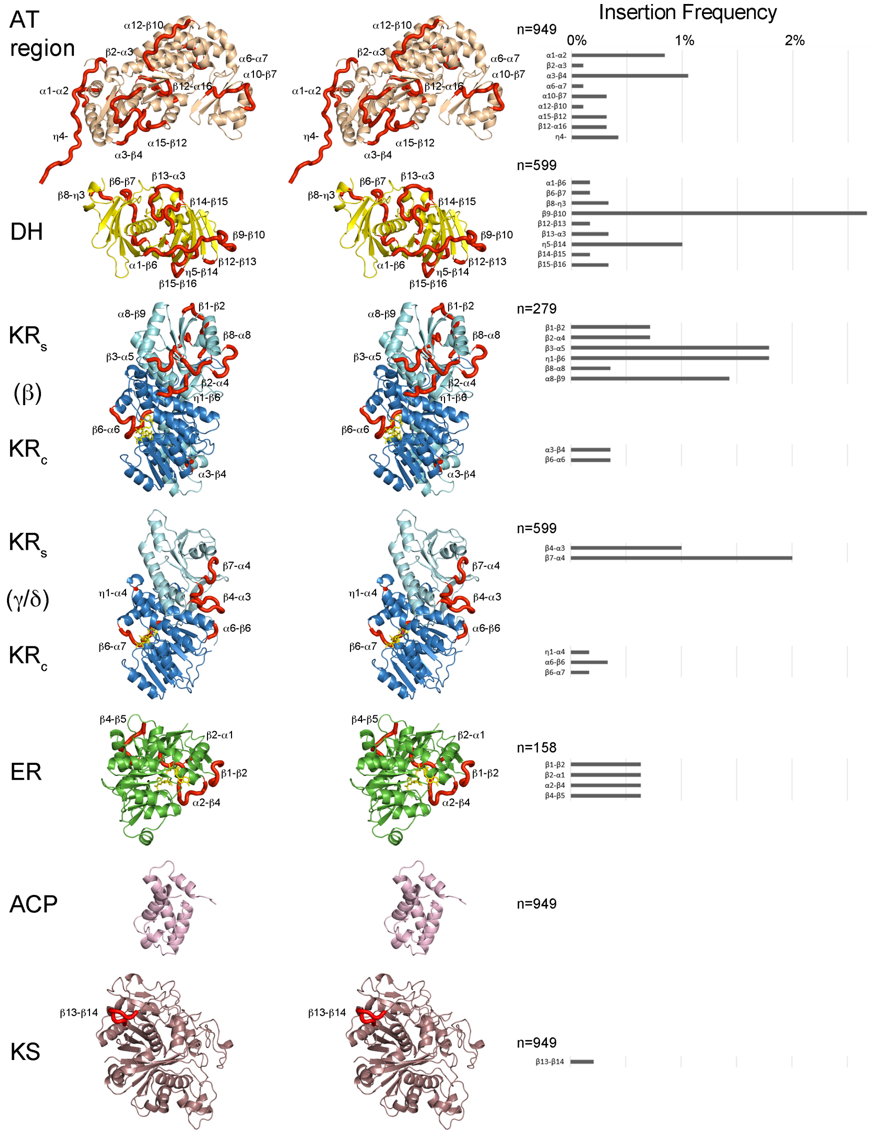
32. Brignole, E.J., S. Smith, and F.J. Asturias, *Conformational flexibility of metazoan fatty acid synthase enables catalysis.* Nat Struct Mol Biol, 2009. **16** (2): p. 190-7.

33. Leibundgut, M., et al., *Structural basis for substrate delivery by acyl carrier protein in the yeast fatty acid synthase.*Science, 2007. **316** (5822): p. 288-90.

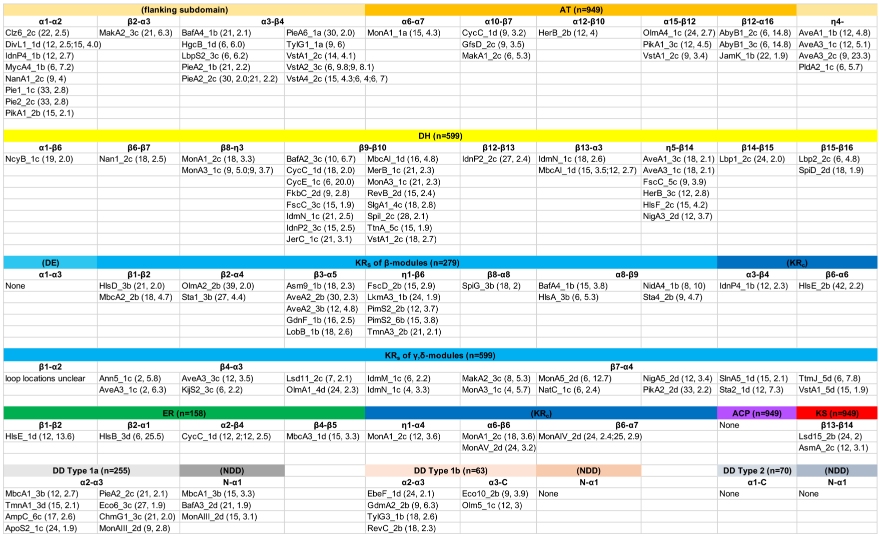



A
Ketosynthase (KS)
updated module boundary
Flanking subdomain (FSD)
mlMlsA1_Mod7 GTNAHLILQQPPTPDTTQTPNTTTgsdpavgsdpavgsdpavgsdpavGVLVWPLSARSAPGLSAQAARLYQHLS
msMlsA1_Mod7 GTNAHLILQQPPTPDTTQTPDTTTgsdpavgsdpavgsdpav------GVLVWPLSARSAPGLSAQAARLYQHLS
muMlsA1_Mod7 GTNAHLILQQPPTPDTTQTPNTTTgsdpavgsdpav------------GVLVWPLSARSAPGLSAQAARLYQHLS
mlMlsA1_Mod8 GTNAHLILQQPPTPDTTQTPNTTTgsdpavgsdpavgsdpav------GVLVWPLSARSAPGLSAQAARLYQHLS
msMlsA1_Mod8 GTNAHLILQQPPTPDTTQTPDTTTgsdpavgsdpavgsdpav------GVLVWPLSARSAPGLSAQAARLYQHLS
muMlsA1_Mod8 GTNAHLILQQPPTPDTTQTPNTTTgsdpavgsdpav------------GVLVWPLSARSAPGLSAQAARLYQHLS
mlMlsA1_Mod9 GTNAHLILQQPPTPDTTQTPNTTTgsdpavgsdpavgsdpav------GVLVWPLSARSAPGLSAQAARLYQHLS
msMlsA1_Mod9 GTNAHLILQQPPTPDTTQTPDTTTgsdpavgsdpavgsdpav------GVLVWPLSARSAPGLSAQAARLYQHLS
muMlsA1_Mod9 GTNAHLILQQPPTPDTTQTPNPTTgsdpavgsdsavgsdpav------GVLVWPLSARSAPGLSAQAARLYQHLS
B
NcyA
NcyB
NcyC
AT
DH⁰
KR
ACP
DD
KS
AT
DH
KR
ACP
KS
AT
DH
KR
ACP
DD
KS
Nannocystin
CMod
NcyB_1c
amino acid
nucleotide
sgsdgrvvaqrergg-a
sgseghfvapsergga
agsegrfvapsergg-a
sgseghfvapsepg--a
sgseghfiapserg--a
sgseghfvapsergg-a
-gsdgrfvapserg--a
sgseghfvapsergg-a
sgseghfvapsergg-a
-gsdgrfvapid-----
tctgggagtgatggtcgcgtcgtcgcgcagcgcgagcgtggtggt---gcg
tctgggagtgagggtcacttcgtcgcgccgagcgagcgtggtggtggtgcg
gccgggagtgaggggcgcttcgtcgcgccgagcgagcgtggtggt---gcg
tctgggagtgagggtcacttcgtcgcgccgagcgagc---ctggt---gcg
tctgggagtgagggtcacttcatcgcgccgagcgagc---gtggt---gcg
tctgggagtgagggtcacttcgtcgcgccgagcgagcgtggtggt---gc-
-c-gggagtgatgggcgcttcgtcgcgccgagcgagc---gtggt---gcg
tctgggagtgagggtcacttcgtcgcgccgagcgagcgtggtggt---gcg
tctgggagtgagggtcacttcgtcgcgccgagcgagcgtggtggt---gc-
-c-gggagtgatgggcgcttcgtcgcgccgatcga---------------


A
α
β
γ
δ
N
AT
ACP
±DD
KS
C
AT
KR$_S$
KR$_C$
ACP
±DD
KS
AT
DH
KR$_S$
KR$_C$
ACP
±DD
KS
AT
DH
KR$_S$
ER
KR$_C$
ACP
±DD
KS
Low
High
Insertion Frequency
α (n=69)
KS-AT
AT-ACP
ACP-KS
ACP-DD
β (n=279)
KS-AT
AT-KR$_S$
KR$_S$-KR$_C$
KR$_C$-ACP
ACP-KS
ACP-DD
γ (n=441)
KS-AT
AT-DH
DH-KR$_S$
KR$_S$-KR$_C$
KR$_C$-ACP
ACP-KS
ACP-DD
δ (n=158)
KS-AT
AT-DH
DH-KR$_S$
KR$_S$-ER
ER-KR$_C$
KR$_C$-ACP
ACP-KS
ACP-DD
0%
10%
B
KS-AT
α
AT-ACP
ACP-KS
β
AT-KR$_S$
KR$_S$-KR$_C$
KR$_C$-ACP
ACP-KS
γ
AT-DH
DH-KR$_S$
KR$_S$-KR$_C$
KR$_C$-ACP
ACP-KS
δ
AT-DH
DH-KR$_S$
KR$_S$-ER
ER-KR$_C$
KR$_C$-ACP
ACP-KS


AT region
α12-β10
β2-α3
α6-α7
α10-β7
α1-α2
β12-α16
η4-
α3-β4
α15-β12
DH
β8-η3
β6-β7
β13-α3
β14-β15
β9-β10
α1-β6
η5-β14
β12-β13
β15-β16
KRs
(β)
KRc
α8-β9
β1-β2
β8-α8
β3-α5
β2-α4
η1-β6
β6-α6
α3-β4
KRs
(γ/δ)
KRc
β7-α4
η1-α4
β4-α3
α6-β6
β6-α7
ER
β4-β5
β2-α1
β1-β2
α2-β4
ACP
KS
β13-β14
Insertion Frequency
0%
1%
2%
n=949
α1-α2
β2-α3
α3-β4
α6-α7
α10-β7
α12-β10
α15-β12
β12-α16
η4-
n=599
α1-β6
β6-β7
β8-η3
β9-β10
β12-β13
β13-α3
η5-β14
β14-β15
β15-β16
n=279
β1-β2
β2-α4
β3-α5
η1-β6
β8-α8
α8-β9
α3-β4
β6-α6
n=599
β4-α3
β7-α4
η1-α4
α6-β6
β6-α7
n=158
β1-β2
β2-α1
α2-β4
β4-β5
n=949
n=949
β13-β14

α-modules (n=71)						
KS-AT			**AT-ACP**		**ACP-**	
AurC_1a (24, 2.2)	PieA6_1a (15, 2.9)	BafA1_3a (30, 4.9)	Clz6_1a (21, 2.3)	OlmA1_1a (18, 3.8)	**KS (n=62)**	**CDD (class 1b) (n=3)**
Divk_2a (24, 3.4)	RevA_1a (18, 2.4)	StzD_1a (18, 1.9)	HerB_1a (24, 2.1)	RubA_2a (15, 1.9)	RubA_2a (12, 2.7)	None
MerA_1a (6, 15.8)	SlnA1_1a (18, 2.8)	TsnA1_1a (16, 3.7)	IdmL_1a (36, 3.2)	SlnA1_1a (19, 2.0)		
MonA1_1a (18, 4.7)	Sta1_1a (21, 1.9)	TtnA_1a (19, 1.9)	MerA_1a (30, 1.9)	Sta1_1a (18, 2.3)	**CDD (class 1a) (n=6)**	**CDD (class 2) (n=0)**
			NorC_1a (12, 4.4)		None	None
β-modules (n=279)						
KS-AT		**AT-KR$_s$**	**KR$_s$-KR$_c$**	**KR$_c$-ACP**	**ACP-**	
AmbB_1b (12, 4.2)	JamK_1b (7, 4.3)	AveA1_1b (18, 2.9)	AmpI_4b (21, 2)	CalA4_1b (18, 2.3)	**KS (n=168)**	**CDD (class 1a) (n=72)**
AmpB_1b (9, 4.4)	JerB_1b (21, 1.9)	BafA2_2b (15, 3.3)	ConB_2b (18, 2.8)	E837.6_2b (27, 2.7)	BafA4_1b (15, 4.1)	Sta2_2b (22, 1.9)
AmpB_2b (12, 2.7)	Ken16_3b (9, 5)	Ela2_1b (21, 1.9)	NidA4_1b (12, 2.7)	EryA2_1b (20, 2)	HlsC_1b (6, 7)	Sta7_2b (24, 3.2)
AmpI_4b (15, 1.9)	NsnB_1b (12, 3.8)	FscD_5b (18, 2.3)	TgaC_1b (27, 2.1)	GulB_1b (24, 2.7)	NysI_3b (27, 2.1)	
AveA1_1b (6, 7.5)	OlmA2_2b (12, 5.4)	GfsA_2b (12, 3.1)		Lsd14_1b (18, 1.9)	PteA1_2b (18, 2.6)	**CDD (class 1b) (n=24)**
AveA2_2b (27, 3.8)	OlmA6_3b (24, 2.1)	IdmN_3b (27, 1.9)		PimS2_4b (22, 1.9)	TiaA2_1b (24, 2)	None
BafA2_1b (21, 2.1)	PikA2_1b (18, 3.2)	MycA1_2b (18, 2.4)		SfaG_1b (15, 2.9)		
ConB_1b (18, 2.4)	PldA2_3b (9, 2.8)	NidA1_2b (15, 3.5)		TgaA_4b (21, 3.9)		**CDD (class 2) (n=15)**
ConE_1b (18, 2.8)	PteA4_1b (21, 2.8)	SlmB_2b (15, 4.7)				E837.6_2b (15, 2.8)
DivK_3b (15, 10.5)	RevC_2b (6, 7.8)	SpiE_1b (30, 2.2)				
Eco1_4b (18, 2.9)	RevD_3b (12, 2.8)	TylG2_1b (18, 2.2)				
Eco5_2b (21, 4.6)	SfaI_1b (33, 2.5)					
Ela1_3b (12, 3.4)	SfaI_1b (27, 2.4)					
FosE_2b (21, 4.2)	SlgA1_3b (36, 2.4)					
FscD_3b (18, 2.2)	SlnA7_1b (24, 2.4)					
GfsA_2b (15, 1.9)	SlnA8_1b (15, 3.3)					
IdmP_1b (12, 2.1)	Sta7_1b (24, 2.4)					
IdnP4_1b (12, 2.4)	TgaC_1b (21, 4)					
γ-modules (n=441)						
KS-AT		**AT-DH**	**DH-KR$_s$**	**KR$_s$-KR$_c$**	**KR$_c$-ACP**	**ACP-**
AsmB_1c (18, 4.3;18, 2.6;15, 3.1)	NanA5_1c (21, 2.0)	None	SpiE_2c (21, 2.1)	AmbC_1c (14, 2.2)	DivK_1c (21, 2.1)	**KS (n=258)**
AveA3_1c (6, 7.5)	NatB_1c (12, 3.4)		TcaA4_1c (18, 2.1)	ApoS6_1c (47, 2.4)	Eco6_1c (12, 2.4)	AsmB_1c (24, 2.0)
BafA2_3c (35, 1.9)	NatB_3c (15, 1.9)			CpkB_1c (12, 3.1)	FscC_2c (18, 3.5)	FkbB_3c (24, 3.0)
BafA3_1c (15, 2.9)	NcyB_1c (46, 3.5;36, 4.9)			Eco3_1c (21, 2.1)	LbpS1_3c (17, 2.1)	MonA4_1c (19, 3.3)
ChlA1_2c (30, 2.7)	NigA3_1c (30, 2.8)			Ela1_2c (17, 3.1)	MakA1_3c (20, 2.2)	NatB_2c (15, 2.4)
DivK_1c (9, 6.1)	NysI_1c (18, 2.9)			FkbA_2c (26, 2.7)	NysC_1c (30, 2.3)	NcyB_1c (18, 4.0)
E837.4_1c (18, 2.4)	NysJ_3c (24, 2.7)			FkbB_1c (21, 3.7)	NysC_2c (24, 2.3)	RubA_1c (21, 2.6)
EbeD_1c (9, 3.6)	OlmA2_3c (24, 2.7;21, 2.2)			GfsD_2c (18, 2.3)	OlmA4_1c (27, 2.0)	SlnA2_1c (18, 2.5)
Eco2_1c (12, 2.7)	OlmA4_1c (18, 3.6;18, 4.1)			IdnP2_3c (12, 3.9)	RevD_2c (18, 2.3)	VstA1_2c (9, 3.7)
Eco6_2c (18, 2.1)	OlmA5_1c (6, 7)			JerC_1c (14, 2.2)	TgaB_2c (18, 3)	
Eco9_1c (9, 3.3)	PelE_1c (18, 2.6)			LbpS1_3c (27, 2.9)		**CDD (class 1a) (n=121)**
FosE_1c (21, 4.2)	PieA1_2c (15, 3.3)			LbpS1_4c (6, 10.5)		MakA1_6c (21, 3)
HerB_3c (18, 2.5)	PieA2_2c (15, 2.9)			Lsd11_2c (6, 8.7)		
HerC_1c (18, 2.0)	PimS1_2c (21, 2.7)			MakA2_2c (21, 2.3)		**CDD (class 1b) (n=24)**
HerC_2c (15, 2.0)	RevA_4c (12, 2.4)			MeiA2_4c (12, 2.6)		None
IdnP1_1c (6, 5.7)	RevD_2c (15, 3.3)			MlaF_1c (24, 2.0)		
IdnP2_3c (36, 2.8)	SlmN_2c (15, 6.2)			MonA5_1c (15, 2.1)		**CDD (class 2) (n=38)**
IdnP5_1c (27, 2.8)	SlmN_3c (15, 6.2)			NcyB_1c (48, 9.7)		MlaB_2c (6, 6.7)
LobA_1c (18, 3.5)	SlmN1_4c (15, 6.5)			NysC_6c (16, 4)		
MakA1_2c (18, 1.9)	Sta4_3c (18, 4.7)			NysJ_3c (18, 2.1)		
MakA1_3c (21, 2.0)	StzC_1c (24, 2.0)			OlmA4_1c (15, 2.4)		
MakA1_4c (33, 4.1)	TmnA5_3c (18, 2.2)			OlmA5_1c (24, 2.6)		
MonA3_1c (12, 3.6)	TtnA_2c (15, 3.4)			SlgA1_4c (12, 3.6)		
NanA4_1c (18, 3.9)				SpiI_2c (24, 3)		
				StzC_1c (9, 3.9)		
				StzD_3c (9, 2.8)		
				TcaA1_2c (21, 2.0)		
				TcaA4_1c (15, 5.1)		
δ-modules (n=158)						
KS-AT	**AT-DH**	**DH-KR$_s$**	**KR$_s$-ER**	**ER-KR$_c$**	**KR$_c$-ACP**	**ACP-**
AmpJ_2d (27, 2.1)	None	None	GfsB_1d (42, 1.8)	None	CngC_2d (21, 1.9;15, 1.9)	**KS (n=73)**
AsmA_1d (9, 3.4)			HlsE_1d (23, 6.0)		NanA2_1d (27, 2.0)	FscB_2d (22, 2.3)
CmiP4_3d (15, 2.9)			PikA2_2d (18, 3.0)		NsnB_3d (24, 2.3)	MbcAI_1d (27, 2.1)
DivL1_1d (15, 3.3)						Sta3_2d (21, 2.2)
FscB_2d (36, 1.9)						TmnA5_1d (15, 2.0)
HerB_4d (18, 7.4)						
HgcB_1d (24, 2.6)						**CDD (class 1a) (n=56)**
KijS4_1d (27, 2.0)						None
LkmAII_2d (21, 3.8)						
MeiA3_1d (18, 2.1)						**CDD (class 1b) (n=14)**
NigA3_2d (19, 1.9)						E837.3_1d (18, 3.8)
PldA1_4d (9, 2.8)						EbeF_1d (18, 3.6)
RevB_2d (24, 2.0)						
RevC_1d (12, 3.4;24, 2.2;15, 2.0)						**CDD (class 2) (n=15)**
SfaG_2d (18, 1.9)						Ken13_1d (6, 10.0)
TcaA3_1d (18, 2.9)						
TmnA4_1d (12, 6.7)						
TtmJ_5d (15, 2.6)						
TtnB_1d (21, 2.3)						



(flanking subdomain)				AT (n=949)					
α1-α2	β2-α3	α3-β4		α6-α7	α10-β7	α12-β10	α15-β12	β12-α16	η4-
Clz6_2c (22, 2.5)	MakA2_3c (21, 6.3)	BafA4_1b (21, 2.1)	PieA6_1a (30, 2.0)	MonA1_1a (15, 4.3)	CycC_1d (9, 3.2)	HerB_2b (12, 4)	OlmA4_1c (24, 2.7)	AbyB1_2c (6, 14.8)	AveA1_1b (12, 4.8)
DivL1_1d (12, 2.5;15, 4.0)		HgcB_1d (6, 6.0)	TylG1_1a (9, 6)		GfsD_2c (9, 3.5)		PikA1_3c (12, 4.5)	AbyB1_3c (6, 14.8)	AveA3_1c (12, 5.1)
IdnP4_1b (12, 2.7)		LbpS2_3c (6, 6.2)	VstA1_2c (14, 4.1)		MakA1_2c (6, 5.3)		VstA1_2c (9, 3.4)	JamK_1b (22, 1.9)	AveA3_2c (9, 23.3)
MycA4_1b (6, 7.2)		PieA2_1b (21, 2.2)	VstA2_3c (6, 9.8;9, 8.1)						PldA2_1c (6, 5.7)
NanA1_2c (9, 4)		PieA2_2c (30, 2.0;21, 2.2)	VstA4_2c (15, 4.3;6, 4;6, 7)						
Pie1_1c (33, 2.8)									
Pie2_2c (33, 2.8)									
PikA1_2b (15, 2.1)									
DH (n=599)									
α1-β6	β6-β7	β8-η3	β9-β10		β12-β13	β13-α3	η5-β14	β14-β15	β15-β16
NcyB_1c (19, 2.0)	Nan1_2c (18, 2.5)	MonA1_2c (18, 3.3)	BafA2_3c (10, 6.7)	MbcAI_1d (16, 4.8)	IdnP2_2c (27, 2.4)	IdmN_1c (18, 2.6)	AveA1_3c (18, 2.1)	Lbp1_2c (24, 2.0)	Lbp2_2c (6, 4.8)
		MonA3_1c (9, 5.0;9, 3.7)	CycC_1d (18, 2.0)	MerB_1c (21, 2.3)		MbcAI_1d (15, 3.5;12, 2.7)	AveA3_1c (18, 2.1)		SpiD_2d (18, 1.9)
			CycE_1c (6, 20.0)	MonA3_1c (21, 2.3)			FscC_5c (9, 3.9)		
			FkbC_2d (9, 2.8)	RevB_2d (15, 2.4)			HerB_3c (12, 2.8)		
			FscC_3c (15, 1.9)	SlgA1_4c (18, 2.8)			HlsF_2c (15, 4.2)		
			IdmN_1c (21, 2.5)	SpiI_2c (28, 2.1)			NigA3_2d (12, 3.7)		
			IdnP2_3c (15, 2.5)	TtnA_5c (15, 1.9)					
			JerC_1c (21, 3.1)	VstA1_2c (18, 2.7)					
(DE)	KR$_S$ of β-modules (n=279)							(KR$_c$)	
α1-α3	β1-β2	β2-α4	β3-α5	η1-β6	β8-α8	α8-β9		α3-β4	β6-α6
None	HlsD_3b (21, 2.0)	OlmA2_2b (39, 2.0)	Asm9_1b (18, 2.3)	FscD_2b (15, 2.9)	SpiG_3b (18, 2)	BafA4_1b (15, 3.8)	NidA4_1b (8, 10)	IdnP4_1b (12, 2.3)	HlsE_2b (42, 2.2)
	MbcA2_2b (18, 4.7)	Sta1_3b (27, 4.4)	AveA2_2b (30, 2.3)	LkmA3_1b (24, 1.9)		HlsA_3b (6, 5.3)	Sta4_2b (9, 4.7)		
			AveA2_3b (12, 4.8)	PimS2_2b (12, 3.7)					
			GdnF_1b (16, 2.5)	PimS2_6b (15, 3.8)					
			LobB_1b (18, 2.6)	TmnA3_2b (21, 2.1)					
KR$_s$ of γ,δ-modules (n=599)									
β1-α2		β4-α3		β7-α4					
loop locations unclear	Ann5_1c (2, 5.8)	AveA3_3c (12, 3.5)	Lsd11_2c (7, 2.1)	IdmM_1c (6, 2.2)	MakA2_3c (8, 5.3)	MonA5_2d (6, 12.7)	NigA5_2d (12, 3.4)	SlnA5_1d (15, 2.1)	TtmJ_5d (6, 7.8)
	AveA3_1c (2, 6.3)	KijS2_3c (6, 2.2)	OlmA1_4d (24, 2.3)	IdmN_1c (4, 3.3)	MonA3_1c (4, 5.7)	NatC_1c (6, 2.4)	PikA2_2d (33, 2.2)	Sta2_1d (12, 7.3)	VstA1_5d (15, 1.9)
ER (n=158)				(KR$_c$)				ACP (n=949)	KS (n=949)
β1-β2	β2-α1	α2-β4	β4-β5	η1-α4	α6-β6	β6-α7		None	β13-β14
HlsE_1d (12, 13.6)	HlsB_3d (6, 25.5)	CycC_1d (12, 2;12, 2.5)	MbcA3_1d (15, 3.3)	MonA1_2c (12, 3.6)	MonA1_2c (18, 3.6)	MonAIV_2d (24, 2.4;25, 2.9)			Lsd15_2b (24, 2)
					MonAV_2d (24, 3.2)				AsmA_2c (12, 3.1)
DD Type 1a (n=255)		(NDD)		DD Type 1b (n=63)		(NDD)		DD Type 2 (n=70)	(NDD)
α2-α3		N-α1		α2-α3	α3-C	N-α1		α1-C	N-α1
MbcA1_3b (12, 2.7)	PieA2_2c (21, 2.1)	MbcA1_3b (15, 3.3)		EbeF_1d (24, 2.1)	Eco10_2b (9, 3.9)	None		None	None
TmnA1_3d (15, 2.1)	Eco6_3c (27, 1.9)	BafA3_2d (21, 1.9)		GdmA2_2b (9, 6.3)	Olm5_1c (12, 3)				
AmpC_6c (17, 2.6)	ChmG1_3c (21, 2.0)	MonAIII_2d (15, 3.1)		TylG3_1b (18, 2.6)					
ApoS2_1c (24, 1.9)	MonAIII_2d (9, 2.8)			RevC_2b (18, 2.3)					