Journal of Industrial Microbiology & Biotechnology (2019) 46:1225–1235
https://doi.org/10.1007/s10295-019-02189-z

NATURAL PRODUCTS - ORIGINAL PAPER

Structural insights into dehydratase substrate selection for the borrelidin and fluvirucin polyketide synthases

Jesus F. Barajas[1,2] · Ryan P. McAndrew[3,4] · Mitchell G. Thompson[2,3] · Tyler W. H. Backman[2,3,5,6] · Bo Pang[2,3,6] · Tristan de Rond[2,3] · Jose H. Pereira[3,4] · Veronica T. Benites[1,2,3] · Héctor García Martín[1,2,3] · Edward E. K. Baidoo[1,2,3] · Nathan J. Hillson[1,2,3] · Paul D. Adams[2,3,4,5] · Jay D. Keasling[2,3,5,6,7,8,9]

Received: 23 March 2019 / Accepted: 16 May 2019 / Published online: 21 May 2019

Abstract

Engineered polyketide synthases (PKSs) are promising synthetic biology platforms for the production of chemicals with diverse applications. The dehydratase (DH) domain within modular type I PKSs generates an α,β-unsaturated bond in nascent polyketide intermediates through a dehydration reaction. Several crystal structures of DH domains have been solved, providing important structural insights into substrate selection and dehydration. Here, we present two DH domain structures from two chemically diverse PKSs. The first DH domain, isolated from the third module in the borrelidin PKS, is specific towards a *trans*-cyclopentane-carboxylate-containing polyketide substrate. The second DH domain, isolated from the first module in the fluvirucin B_1 PKS, accepts an amide-containing polyketide intermediate. Sequence-structure analysis of these domains, in addition to previously published DH structures, display many significant similarities and key differences pertaining to substrate selection. The two major differences between BorA DH M3, FluA DH M1 and other DH domains are found in regions of unmodeled residues or residues containing high B-factors. These two regions are located between α3–β11 and β7–α2. From the catalytic Asp located in α3 to a conserved Pro in β11, the residues between them form part of the bottom of the substrate-binding cavity responsible for binding to acyl-ACP intermediates.

Keywords Polyketide · Dehydratase · Borrelidin · Fluvirucin

Abbreviations

PK	Polyketide
PKS	Polyketide synthase
Flu	Fluvirucin
Bor	Borrelidin
DH	Dehydratase
ACP	Acyl carrier protein
PPant	Phosphopantetheine
MSA	Multiple sequence alignment
SNAC	*N*-Acetyl-cysteamine thioester

Jesus F. Barajas and Ryan P. McAndrew equally contributed.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10295-019-02189-z) contains supplementary material, which is available to authorized users.

✉ Jay D. Keasling
jdkeasling@lbl.gov

[1] Department of Energy Agile BioFoundry, Emeryville, CA 94608, USA

[2] Biological Systems and Engineering Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

[3] Joint BioEnergy Institute, 5885 Hollis St. 4th Floor, Emeryville, CA 94608, USA

[4] Molecular Biophysics and Integrated Bioimaging Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

[5] Department of Bioengineering, University of California, Berkeley, CA 94720, USA

[6] QB3 Institute, University of California, Berkeley, Emeryville, CA 94608, USA

[7] Department of Chemical and Biomolecular Engineering, University of California, Berkeley, Berkeley, CA 94720, USA

[8] Novo Nordisk Foundation Center for Biosustainability, Technical University Denmark, 2970 Horsholm, Denmark

[9] Synthetic Biochemistry Center, Institute for Synthetic Biology, Shenzhen Institutes for Advanced Technologies, Shenzhen, China

Introduction

Polyketide natural products are one of the largest classes of secondary metabolites, possessing vast structural and chemical diversity. The collinear biosynthetic logic of type I modular polyketide synthases (PKSs) make them a promising synthetic biology platform for the production of existing and novel compounds. The dehydratase (DH) domain within type 1 modular PKSs is responsible for the dehydration of specific C-3 hydroxyacyl-acyl carrier protein (ACP) intermediates, resulting in a corresponding enoyl-ACP [42]. These unsaturated ACP-tethered intermediates can be either reduced by an associated enoyl reductase (ER) domain or result in the production of alkenes in even-to-odd positions in the final polyketide structure [9]. The dehydratase domain from type 1 modular PKSs contains a canonical double hot-dog fold motif with an invariant His/Asp catalytic dyad. These structural features are well conserved and well depicted in DH domain crystal structures from erythromycin [25], curacin [3, 15], rifamycin [17] and the gephyronic acid PKSs [9]. Despite having several crystal structures of DH domains, the structural basis for substrate selectivity and specificity is not completely understood. There is a need to understand substrate selection to enable effective design of synthetic PKSs with different DH substrates. Two main factors important for substrate selection by DH domains include the stereochemistry of the C-3 hydroxyacyl group and the chemical structure of the acyl-ACP intermediate past the C-3 hydroxyacyl position. Dehydration proceeds via a *syn*-coplanar elimination of water and is, therefore, sensitive to the stereochemical configuration of the C-3 hydroxyacyl-ACP substrate [37], which is determined enzymatically by the ketoreductase (KR) domain within the module. Thus, DH domains are tied to KR domains that precede them biochemically [5].

The least understood factor in DH substrate selection is the length and chemical structure of the moiety past the C-3 hydroxyl group of the acyl-ACP intermediate. The structure and length of the acyl-ACP intermediate are determined by the upstream PKS module enzymatic architecture and the diverse starter-units incorporated in the nascent polyketide intermediate. How DH domains accommodate these diverse chemical structures while acting on the conserved C-3 hydroxyacyl position of the substrate remains unclear. To further understand substrate selection by DH domains, we conducted a sequence-structure analysis between DH domains from two chemically diverse modular type I PKSs. We chose to investigate two PKSs with distinct starter units and diverse enzymatic architecture: the macrolactone-producing borrelidin (**3**) PKS, and the macrolactam-producing fluvirucin B_1 (**6**) PKS (Fig. 1, Fig. S1). The borrelidin PKS from *Streptomyces parvulus* Tü4055 utilizes a rare *trans*-cyclopentane-1,2-dicarboxylate starter unit (Fig. 1a, Fig. S1A) [31]. To date, there are very few PKSs identified that are able to incorporate and extend a carboxylate-containing polyketide intermediate. The fluvirucin B_1 PKS utilizes amide-containing dipeptidyl intermediate **4** (Fig. 1b, Fig. S1B), likely derived from L-aspartic acid and L-alanine [7, 27, 29]. Here, we present the crystal structures of the borrelidin DH domain from module 3 (BorA DH M3) and the fluvirucin B_1 DH domain from module 1 (FluA DH M1) at 1.80 Å and 2.01 Å, respectively. The BorA DH M3 is specific towards *trans*-cyclopentane-carboxylate-containing polyketide substrate **1**, while the FluA DH M1 accepts amide-containing dipeptide polyketide intermediate **4** (Fig. 1). The DH monomers from both BorA DH M3 and FluA DH M1 possess the traditional double-hotdog fold with an invariant His/Asp catalytic dyad. A close inspection of both BorA DH

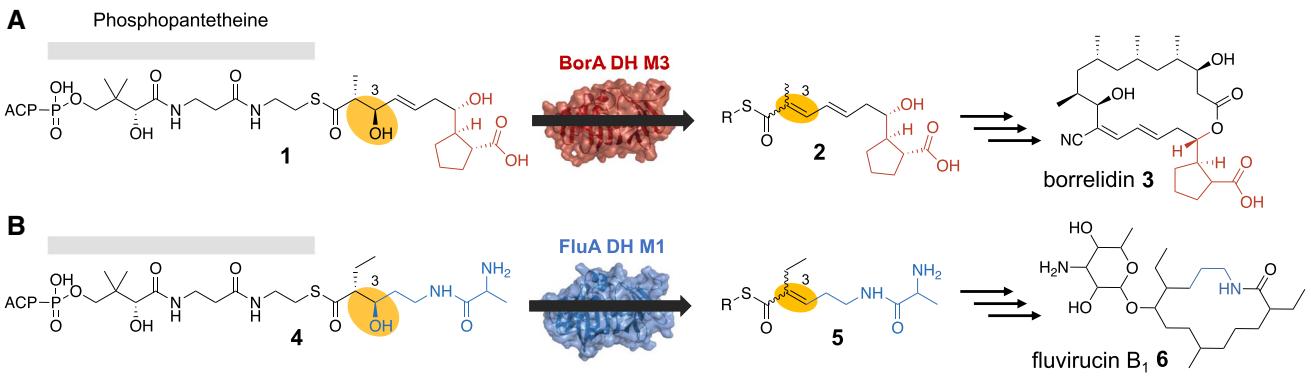

Fig. 1 Corresponding substrates (**1**, **4**) and product intermediates (**2**, **5**) for the BorA DH M3 and FluA DH M1. The final polyketide products are illustrated on the right (**3**, **6**). **a** The borrelidin PKS utilizes a *trans*-cyclopentane-dicarboxylate starter unit (highlighted in red) and **b** the fluvirucin B_1 PKS accepts an amide-containing polyketide starter unit (highlighted in blue). The dehydration reaction is highlighted in yellow. The full biosynthetic polyketide synthase pathways for borrelidin and fluvirucin B_1 are depicted in Fig. S1 (colour figure online)

M3 and FluA DH M1 reveal key structural differences in flexible regions, as observed by high-B-factor analysis, located in the substrate-binding region. In silico docking with their native substrates further supports the importance of these flexible regions. In addition, we aligned DH domain sequences within the ClusterCAD database [13] to identify residues and structural regions that may play a role in substrate binding.

Results

Structure analysis

The BorA DH M3 and FluA DH M1 possess the conventional double-hotdog fold homologous to previously determined type 1 DH domains from the erythromycin, curacin, rifamycin and gephyronic acid PKSs (Fig. 2a, b) [3,

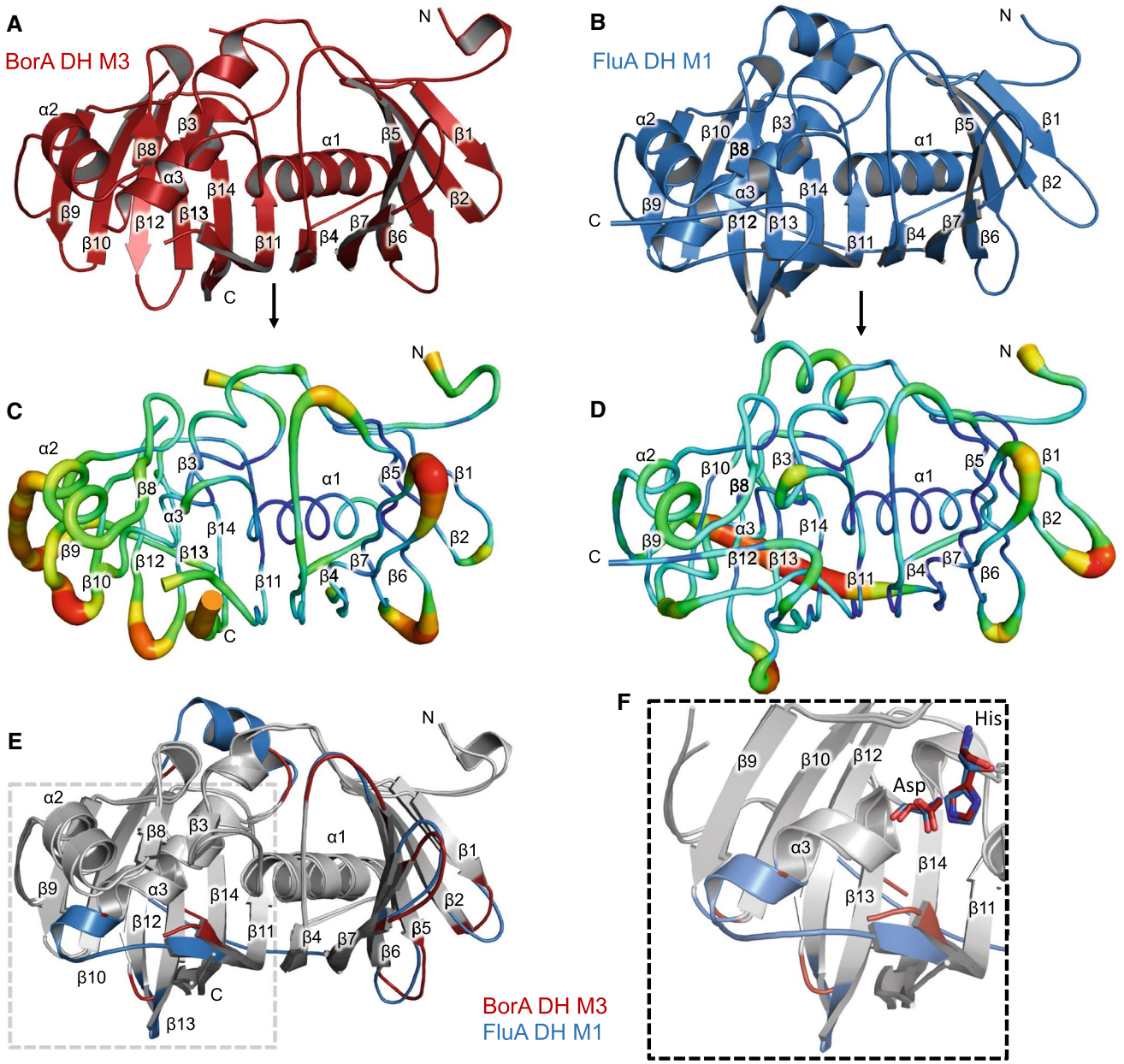

Fig. 2 Structure of the BorA DH M3 and FluA DH M1. Overall structure of the BorA DH M3 colored in red **a** and the FluA DH M1 colored in blue **b**. B-factor putty representation of BorA DH M3 **c** and FluA DH M1 **d**, where thin, blue tubes correspond to low B-factors and thicker, darker red tubes correspond to higher B-factors. **e** Structural alignment of the BorA DH M3 and FluA DH M1. Colored regions correspond to regions of higher α-carbon backbone deviation. **f** Close-up of **e** displaying the His/Asp catalytic dyad and the disordered region between α3 and β11, which form part of the substrate-binding cavity (colour figure online)

9, 17, 25]. Both BorA DH M3 and FluA DH M1 maintain the canonical His/Asp catalytic dyad (Fig. 2f) and share a sequence identity of 45.0%. Structural alignment between the BorA DH M3 and FluA DH M1 display minimal structural deviation within the double-hotdog motif (Fig. 2e). Cα superposition results in RMSD values of 1.29 Å within 209 residues. As reference with other type I DH domains, both BorA DH M3 and FluA DH M1 share a sequence identity of 45.5% with the erythromycin module 4 DH domain. Cα superposition between the erythromycin module 4 DH domain with either BorA DH M3 and FluA DH M1 results in RMSD values of 1.70 and 1.69 Å over a span of 236 residues. Compared to the dual-functioning dehydratase/isomerase DH domain from the gephyronic acid pathway, both BorA DH M3 and FluA DH M1 share 25.6% sequence identity. Cα superposition between the GphF DH1 and either BorA DH M3 or FluA DH M1 results in higher RMSD values of 2.32 and 2.42 Å over a span of 226 residues. The increase in structural differences between BorA DH M3, FluA DH M1 and GphF DH1 are not surprising, given that GphF DH1 has greater separation within the His/Asp catalytic dyad and other conserved residues surrounding the catalytic region [9].

Most of the structural deviations between BorA DH M3, FluA DH M1 and other DH domains are evident in multiple loop regions connecting the α-helices and β-sheets of the core double-hotdog motif (Fig. 2e). Small loops connecting β1–β2, β5–β6, β6–β7 and β12–β13 displayed small structural deviations. However, the largest deviations were located in two loop regions connecting β7–α2 and α3–β11 (Fig. 2F, Fig. S2). This was further supported by either high B-factors within these residues or missing electron density for residues in the loop region (Fig. 2c, d). In the case of BorA DH M3, we were unable to model 11 residues between α3 and β11, suggesting a highly flexible region within the DH domain (Fig. S2A). A lack of modeled residues between α3 and β11 is also observed in the erythromycin M4, phthiocerol dimycocerosate, curacin F and the rifamycin M10 DH domains (Fig. S2C, E, G, I). Unlike the BorA DH M3, all residues between α3 and β11 were modeled in the FluA DH M1 (Fig. S2B). A closer inspection of the residues between α3 and β11 FluA DH M1 displayed partially higher B-factors. These higher B-factors between α3 and β11 FluA DH M1 are also observed in the curacin H, J, K and gephyronic acid DH domains (Fig. S2D, F, H, J). Overall, B-factor analysis and lack of electron density between α3–β11 and β7–α2 in type I DH structures suggest a highly flexible region within the domain.

Substrate-binding cavity

Structural and biochemical data suggest the DH substrate selection is based on DH/ACP protein–protein interactions, stereospecificity of the 3-hydroxyl group and substrate specificity (Fig. S3) [14, 16, 25, 30]. Within each DH monomer, the substrate-binding cavity contains (1) a hydrophobic tunnel/phosphopantetheine (PPant)-binding region, (2) a catalytic region and (3) an acyl intermediate-binding region (Fig. 3). The PPant-binding region is located at the surface of the DH domain between the β11 and β4, in close proximity to β14. An invariant arginine residue on the C-terminal end of β14 has been examined and proposed to be important for either ACP docking and/or PPant recognition [6, 25, 30]. An alignment of BorA DH M3 and FluA DH M1 with other type 1 DH domains displays the conserved arginine residue on the surface of the DH domain (Fig. S4). Measurement between the α-carbon of the conserved Arg and the α-carbon of the catalytic His is on average 17.8 Å (Fig. S4A–B). This highly conserved distance can accommodate the smaller PPant-tethered 3-hydroxyacyl intermediate to the catalytic region. On average, the distance between the phosphate group of the PPant and the C-3 hydroxyl position is 16.4 Å. This is evident in the co-crystal structure of PpsC DH complexed with *trans*-dodec-2-enoyl-CoA (Fig. S4C–E) [14]. These results suggest that the Arg and the residues outlining the substrate tunnel towards the catalytic region in BorA DH M3 and FluA DH M1 are highly conserved amongst type 1 DH domains and may interact extensively with the PPant moiety of the ACP-tethered substrate.

The catalytic region of the DH domain contains the catalytic His and Asp dyad responsible for the dehydration reaction via a *syn*-coplanar elimination of water. In addition to the catalytic His and Asp, both BorA DH M3 and FluA DH M1 contain conserved Leu and Tyr residues that outline the catalytic region, the latter orienting a water molecule (Fig. S5). The catalytic His is located at the beginning of β3 and is on average less than 3.8 Å away from the Asp, located within α3. Recent findings in the fostriecin DH domain suggest DH domains utilize a single-base mechanism, where the active site His residue acts as the base to deprotonate C-2, subsequently protonating the C-3 hydroxyl group to promote C–O bond-cleavage and elimination of water [42]. The carboxylate group of the Asp likely binds and orients the hydroxyl group of the substrate in the favored conformation [42].

The acyl intermediate-binding region can be defined as the residues outlining the bottom of the substrate-binding cavity, outside of the catalytic region. These residues bind to the chemical moiety past the C-3 hydroxyacyl group of acyl-ACP intermediate. The polyketide intermediate past the C-3 hydroxyacyl group can vary in chemical structure, length, and atom heterogeneity, and is dependent on both the upstream PKS module enzymatic architecture and the diverse starter-units incorporated in the nascent polyketide intermediate. The BorA DH M3 can select for a cyclopentane-carboxylate-containing intermediate (Fig. 1a), while the FluA DH M1 can select for an amide-containing dipeptidyl

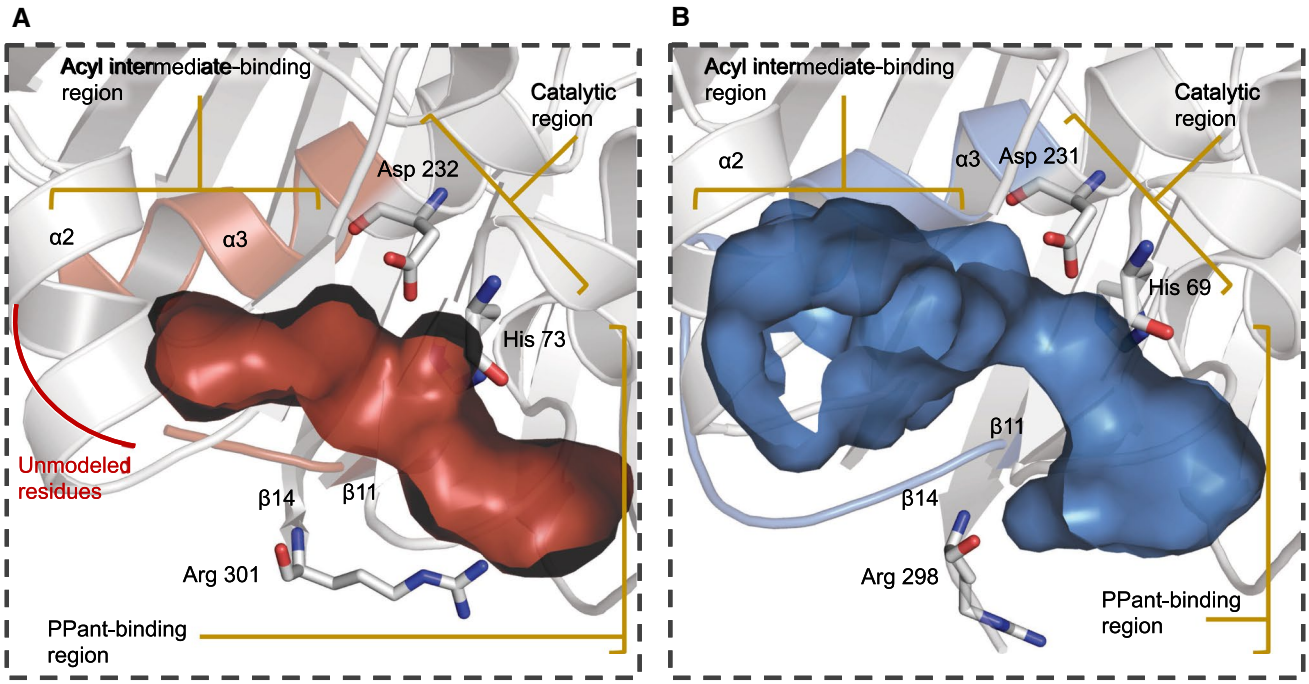

Fig. 3 Substrate-binding cavity of the BorA DH M3 and FluA DH M1. **a** Close-up of the BorA DH M3 substrate-binding cavity, key residues and depiction of the three regions. The three binding regions include the PPant-binding, catalytic and acyl intermediate-binding region. Due to poor electron density, 11 amino acids between α3 and β11 are unmodeled, therefore, given an incomplete model of the substrate-binding cavity. **b** Close-up of the FluA DH M1 substrate-binding cavity. The variable loop regions are highlighted in red and blue residues of the cartoon model (colour figure online)

intermediate (Fig. 1b). Both of these substrates require interactions with residues of distinct chemical properties on the acyl-binding region of the DH domain. A thorough inspection of the polyketide acyl-binding region in BorA DH M3 and FluA DH M1 identified residues located between α3–β11 and α2–β8 (Fig. 3). Surprisingly, the same flexible loop region between α3 and β11, where residues are either missing (BorA DH M3) or have higher B-factors (FluA DH M1), form part of the acyl-binding region. Amino acid preference and multiple sequence alignment analysis suggest that most DH domains have a variant loop region between α3 and β11 (Fig. 5b, Fig. S6). Starting from the catalytic Asp in the middle of α3 to a conserved Pro at the beginning of β11, the loop region can vary in length and amino acid composition. These structural insights suggest a putative binding region responsible for selecting the diverse chemical moieties of different polyketide intermediates within DH domains.

Docking simulations of PPant-tethered substrate intermediates

In our efforts to further validate the importance of the acyl intermediate-binding region for substrate binding and identify key substrate-binding residues, we initially aimed to co-crystalize BorA DH M3 and FluA DH M1 with their corresponding starter units. These starter units included the *trans*-cyclopentane-dicarboxylic acid and β-alanine, in both the free acid and the *N*-acetyl-cysteamine thioester (SNAC)-coupled form. However, no electron density for the corresponding substrates was observed. The lack of substrate binding may be due to high specificity of both BorA DH M3 and FluA DH M1 for their native PPant-tethered polyketide intermediates **1**, **4** (Fig. 1). Using an inactive form of BorA DH M3 and FluA DH M1 with the corresponding native PPant-substrates would greatly increase our efforts in obtaining a substrate–DH domain complex structure. This is supported by the inactive H/F959 PpsC DH domain complexed with a PPant-containing *trans*-dodec-2-enoyl-CoA substrate [14]. Given the complexity to synthesize the native PPant-tethered borrelidin [20] and fluvirucin intermediates for co-crystallization experiments, we decided to conduct in silico docking as an alternative approach. Our initial docking analysis was focused on the crystal structure of FluA DH M1 and its corresponding native substrate **4**. The lack of 11 unmodeled residues in the BorA DH M3 acyl intermediate-binding region would yield inconclusive in silico docking results (Fig. 3a). Therefore, we opted not to conduct substrate docking analysis with the BorA DH M3 model.

In silico docking of the alanine-protected dipeptide PPant intermediate **4** with the FluA DH M1 revealed key structural regions and residues that may play a role in substrate selection. The fluvirucin PPant-tethered intermediate **4** revealed close interactions with all three regions of the substrate-binding cavity of FluA DH M1 (Fig. 4). On the surface of the DH domain, the conserved Arg 298 was in close proximity for potential electrostatic interactions with the phosphate group of **4**. Several hydrophobic residues (Leu 248, Phe 251, Ala 105) in the PPant-binding region of the DH were in close contact with the PPant moiety of **4**. Within the catalytic region, the C-3 hydroxyl group of **4** is in proximity and oriented toward the catalytic Asp 231. The orientation of the C-3 hydroxyl towards the Asp 231 positioned the C-2 of **4** next to the catalytic His 69, supporting the mechanism of deprotonation of C-2 by the single catalytic base His 69 [42].

Several residues in the acyl intermediate-binding region were in the vicinity of the nitrogen-containing dipeptide moiety of **4** (Fig. 4). The variable loop region between α3 and β11 (light blue) contained polar residues His 235 and Glu 244 within 4.0 Å from the terminal amino group of **4**. In addition, residues between α2 and β8 (dark blue), on the distal side of the variable loop region between α3 and β1, interacted with the dipeptide moiety of **4**. Residues between Ser 185 and the conserved Tyr 189 were in close contact with **4**. Surprisingly, the PPant-tethered fluvirucin intermediate **4** does not accommodate the entire catalytic region of the acyl intermediate-binding region. On average, there is a 9.5-Å distance gap between the terminal amino moiety of **4** and the bottom of the substrate-binding cavity. The bottom of the cavity contains several variant polar residues (Glu 241 and Gln 242). The gap identified between these polar residues with **4**, the high B-factors associated within the invariant loop and the lack of modeled residues for similar type I DH domains suggest this structural region may adopt a distinct conformational change upon binding to the native substrate. This is further supported by the small structural conformational changes observed between the variable loop regions α3–β11 of the *apo* FluA DH M1 crystal structure and the energy minimized docked PPant intermediate **4**-FluA DH M1 complex structure (Fig. S7). Cα superposition results in RMSD values of 2.10 Å within the 18 residues of the α3–β11 loop region.

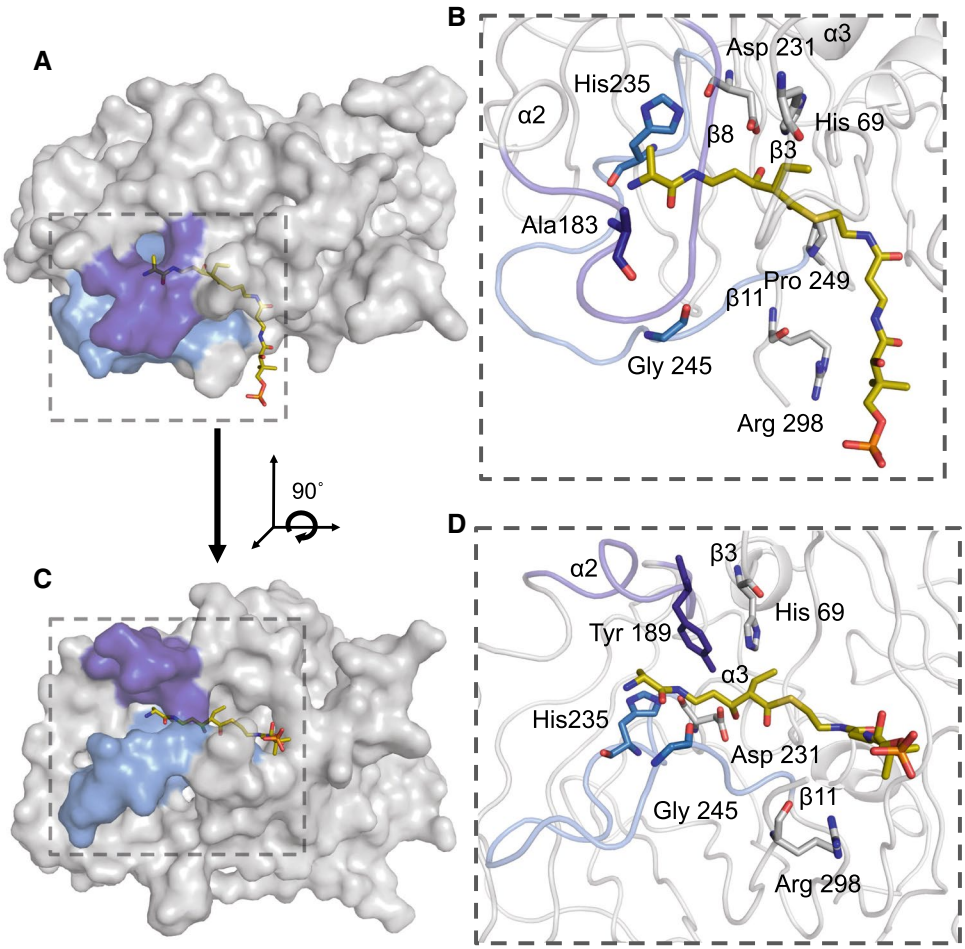

Fig. 4 In silico docking analysis of FluA DH M1 with the natural PPant-substrate intermediate **4**. Substrate **4** interacts with all three DH substrate-binding regions **a**. **b** A close inspection of the dipeptide moiety of **4** shows close interactions with residues located between α3 and β11 (light blue) and α2 and β8 (dark blue). **c** A 90˚ horizontal rotation of **a** and close-up of the substrate **4** and interacting residues (colour figure online)

Analysis of DH domain conservation

To quantitatively examine the level of sequence conservation amongst DH domains, the Shannon entropy of aligned DH sequences was calculated at each position within a refined multiple sequence alignment (MSA). 330 DH domain sequences derived from modular type I PKSs were exported from the ClusterCAD database [13]. The secondary structure of the FluA DH M1 was used to refine the DH boundaries of the MSA and visualize Shannon entropy (*H*) at each position (Fig. 5a). Overall, residues that were conserved and displayed low *H* values were both catalytic residues and residues important for structural integrity of the double-hotdog fold motif. The catalytic His/Asp dyad, located within β3 and α3, and related DH motifs (HxxxGxxxxP, GYxYGPxF, LPFxW) displayed low *H* values (Fig. 5a). Similarly, the known structural motifs that make up the DH boundaries (HPLL and LxLxR) prior to β1 and within β14 show low Shannon entropy values.

With the aim of identifying residues and/or structural regions that may play a role in substrate binding past the C-3 hydroxyacyl position, we looked for residues that may be easily substituted with others to accommodate the diversity of the acyl-ACP substrate intermediate. A closer inspection of DH domains showed two regions with high *H* values. These higher *H* values were present in the loop regions between β7–α2 and α3–β11 (Fig. 5b). As observed in the BorA DH M3, FluA DH M1 and other DH structures, these two regions with higher *H* values are located in the same region of unmodeled, high-B-factor residues in the PDB structures. Residues between α3 and β11, and within α2 form part of the acyl intermediate-binding region (Figs. 5, S6). These results further support our DH domain structural and in silico docking analysis regarding the importance of α3–β11 and α2 in substrate selection. These two variant regions between α3–β11 and β7–α2 may possess distinct amino acid properties across different DH domains and may be important for the structural architecture of the acyl intermediate-binding region. In particular, the residues between α3 and β11 and within α2 may play a direct role in substrate selectivity past the C-3 hydroxyacyl position.

Discussion

Substrate selection by DH domains is attributed to both the stereochemistry of the C-3 hydroxyacyl group and the chemical structure on the acyl-ACP intermediate past the C-3 hydroxyacyl position. The recent structural and biochemical insights of various DH domains have significantly advanced our understanding of stereoselection by DH domains [3, 15, 17, 25, 26]. However, DH substrate selection pertaining to the acyl-ACP intermediate past the C-3 hydroxyacyl position remains elusive. For example, the Rif DH M10, CurK DH,

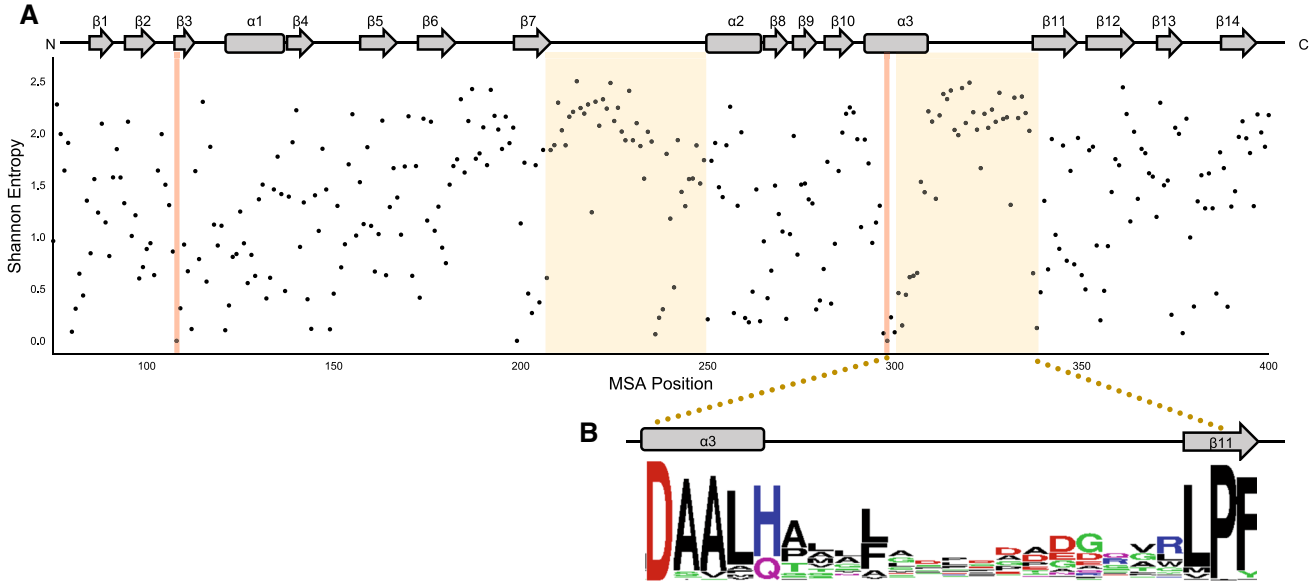

Fig. 5 **a** Prediction of conserved residues through Shannon entropy measurements on a multiple sequence alignment of modular type I DH domains taken from ClusterCAD. Low Shannon entropy measurements correspond to low levels of amino acid substitution. The secondary structure of the FluA DH M1 was used to give a relative position within the MSA. Regions highlighted in red display the His/Asp catalytic dyad and the regions highlighted in yellow correspond the two most highly variable regions amongst the known modular type I DH domain structures. **b** Close-up of the region between α3 and β11. The overall height of the stack indicates the sequence conservation at that position (colour figure online)

and most DH domains act on a C-3 hydroxyl-ACP intermediate. Yet, the Rif DH M10 from rifamycin is selective towards a C-18, naphthoquinone-containing intermediate [17, 38] while the CurK DH acts on a shorter thiazoline, cyclopropyl-containing intermediate [3, 8]. A key question in DH substrate selectivity is what residue(s) and/or structural features within the DH domain are responsible for selection of the acyl intermediate past the C-3 hydroxyacyl position. Further structural and biochemical insights on DH domains would provide a better understanding of substrate selection past the PPant and C-3 hydroxyacyl moiety.

To further improve our current understanding of DH selectivity towards different acyl-ACP intermediates past the C-3 hydroxyacyl position, we have performed structural analysis on two chemically diverse DH domains. The first DH domain, isolated from the third module in the borrelidin PKS, is specific towards a *trans*-cyclopentane-carboxylate-containing polyketide substrate. The second DH domain, isolated from the first module in fluvirucin B_1 PKS, accepts an amide-containing polyketide intermediate. Both DH domains maintain the canonical double-hotdog fold motif with the invariant His/Asp catalytic dyad. Surprisingly, a structural comparison of the BorA DH M3, FluA DH M1 and other type I DH domains show few significant deviations from each other. The two major differences between BorA DH M3, FluA DH M1 and other DH domains are found in regions of unmodeled residues or residues containing high B-factors. These two regions are located between α3–β11 and β7–α2. From the catalytic Asp located in α3 to a conserved Pro in β11, the residues between them form part of the bottom of the substrate-binding cavity responsible for binding to the acyl-ACP intermediates past the C-3 hydroxyacyl position. The distal region from α3 to β11 is also part of substrate-binding cavity and is located in α2–β8. The loop connecting β7–α2 is one of the most variable regions in modular type I DH domain structures. The variable β7–α2 loop region may provide flexibility to part of the substrate-binding cavity located in α2–β8 (Fig. 4). These structural insights suggest both α3–β11 and β7–α2 are important for binding of the acyl-ACP intermediate past the C-3 hydroxyacyl position. This was further supported by our in silico docking analysis of the FluA DH M1 and prediction of conserved residues through Shannon entropy measurements (Figs. 4, 5). Structural evidence for this can be observed in the PpsC DH structures in complex with *trans*-dodec-2-enoyl-CoA (PDB ID: 5NJI) or crotonyl-CoA ((PDB ID: 5I0K). Comparison of the α3–β11 and β7–α2 regions in both structures displays key differences. More residues were modeled in the α3–β11 of PpsC DH structure in the presence of the longer *trans*-dodec-2-enoyl-CoA substrate, suggesting higher stability of the α3–β11 region upon binding of a substrate that closely resembles the native acyl-ACP intermediate.

The identification of both the α3–β11 and β7–α2 regions involved in substrate selection may provide insights into rational engineering of DH domains. In our efforts to improve DH-mediated dehydration in adipic acid production using the engineered BorA2 PKS [19], we tested various DH domains and chimeric DH domains with α3–β11 loop swaps in vitro (Fig. S8a). Surprisingly, all chimeric DH domains containing the α3–β11 loop swaps were soluble (Fig. S8b). This suggests that a key secondary structure is maintained within the DH domain's α3–β11 loop swap region. However, a close inspection of adipic acid production using the chimeric DH domains was inconclusive (Fig. S8c). We speculate that the lack of conclusive results may be due to both the high background of adipic acid that is likely derived from *E. coli* during protein purification, and low efficiency of the BorA2 PKS to produce adipic acid. This assay may not be optimal for testing chimeric DH activity. Nonetheless, generating soluble and stable chimeric DH domains is a first step towards DH engineering. Further efforts in investigating the activity of chimeric DH domains can be simplified by testing standalone chimeric DH domains containing the α3–β11 loop swaps with more naturally relevant acyl-ACP, acyl-PPant or acyl-SNAC substrates [20].

Our structures of the BorA DH M3 and FluA DH M1 provide further structural evidence of substrate selection by DH domains within modular type I PKSs. We present both a sequence-structural analysis between DH domains and identify significant similarities and key differences pertaining to substrate selection. The work presented here in combination with existing DH domain studies should facilitate further engineering efforts of modified PKSs that can process non-natural substrates with improved catalytic properties.

Materials and methods

Cloning of *BorA DH M3* and *FluA DH M1*

Vector backbone template originated from JPUB_008800 plasmid (Table S1). The genes encoding for BorA DH M3 and FluA DH M1 were PCR amplified from the pDVA00936 and pTL-A01 plasmids, and cloned into a pET28a vector using Gibson assembly methodology [18]. The j5 DNA assembly design automation software [22] was utilized to generate Gibson assembly primers for the BorA DH M3, FluA DH M1 and pET28a vector backbone. Primer sequences are available in Table S1. Polymerase chain reaction amplification of the BorA DH M3, FluA DH M1 and pET28a backbone was conducted using Phusion Hot Start II DNA polymerase (Thermo Fisher) using the vendor's recommended protocol at an annealing temperature of 65 °C and 1 min extension time for the DHs and 3 min extension time for the pET28 backbone. The pET28a PCR

product was Dpn1 digested (Thermo Fisher) using the vendor-recommended protocol and subjected to PCR clean up (Zymo Research). The DH genes and pET28a PCR product were ligated using Gibson assembly master mix (New England BioLabs), following the vendor-recommended protocol. Both DH constructs were designed with a thrombin-cleavable N-terminal 6×His-tag. The strains and plasmid sequences utilized in this study are listed in Table S1. All strains and plasmid sequences may be accessed and requested through the Joint BioEnergy's public registry (https://public-registry.jbei.org/folders/412) [21].

Protein expression and purification

The recombinant BorA DH M3 and FluA DH M1 containing an N-terminal 6×His-tag were produced in BL21 (DE3) *E. coli* cells (Novagen). Cells containing the pET28-DH plasmids were grown to OD_{600} = 0.8 at 37 °C in TB medium containing 50 µg/mL kanamycin. The cell cultures were cooled to 18 °C and expression was induced using 1 mM IPTG. The cell cultures were incubated for an additional 16 h at 18 °C and harvested by centrifugation at 5525 r.c.f. for 10 min. The cell pellets were resuspended in 50 mM Hepes, pH 7.6, 10% glycerol, 10 mM imidazole, 300 mM NaCl. Resuspended cells were cooled on ice for 30 min and the cells were disrupted using sonication. The cell debris was cleared by centrifugation at 21,036 r.c.f. for 1 h. The supernatant was collected and batch bound to HisPur TM Cobalt Resin (Thermo Scientific) for 1 h at 4 °C. BorA DH M3 and FluA DH M1 were purified according to the manufacturer's instructions using an imidazole step-gradient. Fractions containing pure protein were determined by SDS-PAGE and fractions containing either BorA DH M3 or FluA DH M1. Removal of the N-terminal 6×His-tag was conducted by incubating the fractions containing the proteins at 14 °C for 30 h with thrombin from bovine plasma (Sigma-Aldrich) at a concentration of 2 U/mg of recombinant protein and 3.5 mM $CaCl_2$. After thrombin digestion and assessment of N-terminal 6×His-tag cleavage via SDS-PAGE, we dialyzed the DH domains against 50 mM Hepes, pH 7.6, 5% glycerol. Removal of thrombin and further purification of BorA DH M3 and FluA DH M1 were conducted by anion-exchange chromatography using HiTrap Q FF (GE Healthcare) according to the manufacturer's instructions. Purified BorA DH M3 and FluA DH M1 were dialyzed against crystallization buffer, which consisted of 25 mM Hepes, pH 7.6, and 1 mM dithiothreitol (Fig. S9). Both BorA DH M3 and FluA DH M1 were concentrated to 12 mg/mL for crystallographic studies.

Crystallization and structure determination

Crystallization screening was carried out on a Phoenix robot (Art Robbins Instruments, Sunnyvale, CA) using a sparse matrix screening method [23]. BorA DH M3 concentrated to 10 mg/mL was crystallized by sitting drop vapor diffusion in drops containing a 1:1 ratio of protein solution and 0.20 M $MgCl_2$, 0.10 M Tris, pH 8.5, and 29% (w/v) PEG 4000. For FluA DH M1, 10 mg/mL protein was crystallized by sitting drop vapor diffusion in drops containing a 1.5:1 ratio of protein solution and 0.40 M NaCl, 0.10 M Tris, pH 8.5, and 29% (w/v) PEG 3350. For both proteins, a final concentration of 15% glycerol was added before flash freezing in liquid nitrogen.

The X-ray data sets for BorA DH M3 and FluA DH M1 were collected at the Berkeley Center for Structural Biology on beamlines 8.2.1 and 8.2.2 of the Advanced Light Source at Lawrence Berkeley National Laboratory. Diffraction data were recorded using ADSC Q315R detectors (Area Detector Systems Corporation, San Diego, CA). Processing of image data was performed using the HKL2000 suite of programs [32]. For both proteins, phases were calculated by molecular replacement with the program Phaser [28], using the structure of Rif DH M10 (PDB id: 4LN9) [6] as a search model. Automated model building was conducted using AutoBuild [39–41] from the Phenix [1] suite of programs resulting in a model that was mostly complete. Manual building using Coot [12] was alternated with reciprocal space refinement using Phenix [2]. Waters were automatically placed using Phenix and manually added or deleted with Coot according to peak height (3.0σ in the *Fo-Fc* map) and distance from a potential hydrogen bonding partner (< 3.5 Å). TLS refinement [33] using ten groups, chosen using the TLSMD web server [33], was used in later rounds of refinement. All data collection, phasing, and refinement statistics are summarized in Table S2.

In silico substrate docking

We conducted in silico substrate docking between **4** and the FluA DH M1. The natural PPant-tethered substrate **4** was initially drawn in Chemdraw (PerkinElmer) and transferred in SMILES format to the PHENIX software suite, eLBOW ligand and constraints generator [1]. The eLBOW program generated a PDB file of the PPant-tethered substrate **4** and a constraints (.cif) file. Both files were used to dock **4** into the FluA DH M1. As a reference for substrate binding, we used the co-crystal structure of PpsC DH complexed with *trans*-dodec-2-enoyl-CoA (PDB ID: 5NJI) to position substrate **4** into the FluA DH M1. We utilized the program *Coot* to overlay the PpsC DH with chain A of the FluA DH M1 and model in substrate **4** [11]. The entire FluA DH M1-substrate

4 complex was then energy minimized using the program UCSF CHIMERA [34]. The lack of 11 unmodeled residues in the BorA DH M3 acyl intermediate-binding region made it difficult to conduct precise in silico docking on this target and obtain conclusive information. Therefore, we opted not to conduct substrate docking analysis with the BorA DH M3 model.

Protein structure and visualization

All of the protein structure analysis and figures were generated using UCSF CHIMERA [34] and PyMOL (The PyMOL Molecular Graphics System, PyMOL1.2edu1 2009, Schrödinger, LLC). Both PyMOL and the online I-TASSER server were used for PDB structural alignment analysis [36]. Multi-sequence alignment for Fig. S6 was generated using MUSCLE [10] and visualized using ESPript 3.0 [35].

Analysis of DH sequence conservation

The Shannon entropy of aligned DH sequences was calculated at each position within a refined multiple sequence alignment (MSA). Briefly, 330 sequences derived from PKS non-loading modules containing AT, DH, KR, and ACP domains with active KR and DH domains were exported from the ClusterCAD database [13]. Sequences contained the first residue after the AT domain to the last residue immediately preceding the downstream domain (usually a KR or ER). Sequences were then aligned using MAFFT [24]. The MSA was then refined using the ProDy python library so that all positions within the MSA maintained at least 10% occupancy [4]. ProDy was then used to calculate the Shannon entropy of each position within the resulting MSA. The boundaries of the DH domain within the MSA were further refined using PDB structural information.

Acknowledgements We thank Dr. Sangeeta Nath and Dr. Samuel Deutsch at the Joint Genome Institute for providing the template pDVA00936 plasmid, from which the pET28-BorA DH M3 was cloned. We also thank Dr. Nathan A. Schnarr for providing the pTL-A01 plasmid from which the pET28-FluA DH M1 was cloned. Finally, we thank Dr. Leonard Katz and Christopher Eiben for their suggestions and helpful discussions. This work was part of the DOE Joint BioEnergy Institute (https://www.jbei.org) supported by the U. S. Department of Energy, Office of Science, Office of Biological and Environmental Research, and was part of the Agile BioFoundry (https://agilebiofoundry.org) supported by the U. S. Department of Energy, Energy Efficiency and Renewable Energy, Bioenergy Technologies Office, through contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U. S. Department of Energy. The views and opinions of the authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, expressed or implied, or assumes any legal liability or responsibility or the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

Author contributions JFB, RPM, JHP, PDA, JDK designed the research. JFB, RPM, BP, TdR, VTB, EEKB performed wet lab experiments and MGT, TWHB performed bioinformatic analysis. JFB, RPM, HGM, NJH, PDA, and JDK wrote and edited the manuscript.

Compliance with ethical standards

Conflict of interest J. D. K has financial interests in Amyris, Lygos, Constructive Biology, Demetrix, Napigen, and Maple Bio.

References

1. Adams PD, Afonine PV, Bunkóczi G et al (2010) PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr D Biol Crystallogr 66:213–221. https://doi.org/10.1107/S0907444909052925
2. Afonine PV, Grosse-Kunstleve RW, Echols N, Headd JJ, Moriarty NW, Mustyakimov M, Terwilliger TC, Urzhumtsev A, Zwart PH, Adams PD (2012) Towards automated crystallographic structure refinement with phenix.refine. Acta Crystallogr D Biol Crystallogr 68:352–367. https://doi.org/10.1107/S0907444912001308
3. Akey DL, Razelun JR, Tehranisa J, Sherman DH, Gerwick WH, Smith JL (2010) Crystal structures of dehydratase domains from the curacin polyketide biosynthetic pathway. Structure 18:94–105. https://doi.org/10.1016/j.str.2009.10.018
4. Bakan A, Dutta A, Mao W, Liu Y, Chennubhotla C, Lezon TR, Bahar I (2014) Evol and ProDy for bridging protein sequence evolution and structural dynamics. Bioinformatics 30:2681–2683. https://doi.org/10.1093/bioinformatics/btu336
5. Barajas JF, Blake-Hedges JM, Bailey CB, Curran S, Keasling JD (2017) Engineered polyketides: synergy between protein and host level engineering. Synth Syst Biotechnol 2:147–166. https://doi.org/10.1016/j.synbio.2017.08.005
6. Barajas JF, Shakya G, Moreno G et al (2017) Polyketide mimetics yield structural and mechanistic insights into product template domain function in nonreducing polyketide synthases. Proc Natl Acad Sci USA 114:E4142–E4148. https://doi.org/10.1073/pnas.1609001114
7. Barajas JF, Zargar A, Pang B, Benites VT, Gin J, Baidoo EEK, Petzold CJ, Hillson NJ, Keasling JD (2018) Biochemical characterization of β-amino acid incorporation in fluvirucin b2 biosynthesis. ChemBioChem 19:1391–1395. https://doi.org/10.1002/cbic.201800169
8. Chang Z, Sitachitta N, Rossi JV, Roberts MA, Flatt PM, Jia J, Sherman DH, Gerwick WH (2004) Biosynthetic pathway and gene cluster analysis of curacin A, an antitubulin natural product from the tropical marine cyanobacterium *Lyngbya majuscula*. J Nat Prod 67:1356–1367. https://doi.org/10.1021/np0499261

9. Dodge GJ, Ronnow D, Taylor RE, Smith JL (2018) Molecular basis for olefin rearrangement in the gephyronic acid polyketide synthase. ACS Chem Biol 13:2699–2707. https://doi.org/10.1021/acschembio.8b00645
10. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797. https://doi.org/10.1093/nar/gkh340
11. Emsley P, Lohkamp B, Scott WG, Cowtan K (2010) Features and development of Coot. Acta Crystallogr D Biol Crystallogr 66:486–501. https://doi.org/10.1107/S0907444910007493
12. Emsley P, Cowtan K (2004) Coot: model-building tools for molecular graphics. Acta Crystallogr D Biol Crystallogr 60:2126–2132. https://doi.org/10.1107/S0907444904019158
13. Eng CH, Backman TWH, Bailey CB, Magnan C, García Martín H, Katz L, Baldi P, Keasling JD (2018) ClusterCAD: a computational platform for type I modular polyketide synthase design. Nucleic Acids Res 46:D509–D515. https://doi.org/10.1093/nar/gkx893
14. Faille A, Gavalda S, Slama N, Lherbet C, Maveyraud L, Guillet V, Laval F, Quémard A, Mourey L, Pedelacq J-D (2017) Insights into substrate modification by dehydratases from type I polyketide synthases. J Mol Biol 429:1554–1569. https://doi.org/10.1016/j.jmb.2017.03.026
15. Fiers WD, Dodge GJ, Sherman DH, Smith JL, Aldrich CC (2016) Vinylogous dehydration by a polyketide dehydratase domain in curacin biosynthesis. J Am Chem Soc 138:16024–16036. https://doi.org/10.1021/jacs.6b09748
16. Finzel K, Nguyen C, Jackson DR, Gupta A, Tsai S-C, Burkart MD (2015) Probing the substrate specificity and protein–protein interactions of the *E. coli* fatty acid dehydratase, FabA. Chem Biol 22:1453–1460. https://doi.org/10.1016/j.chembiol.2015.09.009
17. Gay D, You Y-O, Keatinge-Clay A, Cane DE (2013) Structure and stereospecificity of the dehydratase domain from the terminal module of the rifamycin polyketide synthase. Biochemistry 52:8916–8928. https://doi.org/10.1021/bi400988t
18. Gibson DG, Young L, Chuang R-Y, Venter JC, Hutchison CA, Smith HO (2009) Enzymatic assembly of DNA molecules up to several hundred kilobases. Nat Methods 6:343–345. https://doi.org/10.1038/nmeth.1318
19. Hagen A, Poust S, de Rond T, Fortman JL, Katz L, Petzold CJ, Keasling JD (2016) Engineering a polyketide synthase for in vitro production of adipic acid. ACS Synth Biol 5:21–27. https://doi.org/10.1021/acssynbio.5b00153
20. Hahn F, Kandziora N, Friedrich S, Leadlay PF (2014) Synthesis of complex intermediates for the study of a dehydratase from borrelidin biosynthesis. Beilstein J Org Chem 10:634–640. https://doi.org/10.3762/bjoc.10.55
21. Ham TS, Dmytriv Z, Plahar H, Chen J, Hillson NJ, Keasling JD (2012) Design, implementation and practice of JBEI-ICE: an open source biological part registry platform and tools. Nucleic Acids Res 40:e141. https://doi.org/10.1093/nar/gks531
22. Hillson NJ, Rosengarten RD, Keasling JD (2012) j5 DNA assembly design automation software. ACS Synth Biol 1:14–21. https://doi.org/10.1021/sb2000116
23. Jancarik J, Kim SH (1991) Sparse matrix sampling: a screening method for crystallization of proteins. J Appl Crystallogr 24:409–411. https://doi.org/10.1107/S0021889891004430
24. Katoh K, Kuma K, Toh H, Miyata T (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res 33:511–518. https://doi.org/10.1093/nar/gki198
25. Keatinge-Clay A (2008) Crystal structure of the erythromycin polyketide synthase dehydratase. J Mol Biol 384:941–953. https://doi.org/10.1016/j.jmb.2008.09.084
26. Li Y, Dodge GJ, Fiers WD, Fecik RA, Smith JL, Aldrich CC (2015) Functional characterization of a dehydratase domain from the pikromycin polyketide synthase. J Am Chem Soc 137:7003–7006. https://doi.org/10.1021/jacs.5b02325
27. Lin T-Y, Borketey LS, Prasad G, Waters SA, Schnarr NA (2013) Sequence, cloning, and analysis of the fluvirucin B1 polyketide synthase from *Actinomadura vulgaris*. ACS Synth Biol 2:635–642. https://doi.org/10.1021/sb4000355
28. McCoy AJ, Grosse-Kunstleve RW, Adams PD, Winn MD, Storoni LC, Read RJ (2007) Phaser crystallographic software. J Appl Crystallogr 40:658–674. https://doi.org/10.1107/S0021889807021206
29. Miyanaga A, Hayakawa Y, Numakura M, Hashimoto J, Teruya K, Hirano T, Shin-Ya K, Kudo F, Eguchi T (2016) Identification of the fluvirucin B2 (Sch 38518) biosynthetic gene cluster from *Actinomadura fulva* subsp. *indica* ATCC 53714: substrate specificity of the β-amino acid selective adenylating enzyme FlvN. Biosci Biotechnol Biochem 80:935–941. https://doi.org/10.1080/09168451.2015.1132155
30. Nguyen C, Haushalter RW, Lee DJ et al (2014) Trapping the dynamic acyl carrier protein in fatty acid biosynthesis. Nature 505:427–431. https://doi.org/10.1038/nature12810
31. Olano C, Wilkinson B, Sánchez C et al (2004) Biosynthesis of the angiogenesis inhibitor borrelidin by Streptomyces parvulus Tü4055: cluster analysis and assignment of functions. Chem Biol 11:87–97. https://doi.org/10.1016/j.chembiol.2003.12.018
32. Otwinowski Z, Minor W (1997) Processing of X-ray diffraction data collected in oscillation mode. Methods Enzymol 276:307–326
33. Painter J, Merritt EA (2006) Optimal description of a protein structure in terms of multiple groups undergoing TLS motion. Acta Crystallogr D Biol Crystallogr 62:439–450. https://doi.org/10.1107/S0907444906005270
34. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE (2004) UCSF Chimera—a visualization system for exploratory research and analysis. J Comput Chem 25:1605–1612. https://doi.org/10.1002/jcc.20084
35. Robert X, Gouet P (2014) Deciphering key features in protein structures with the new ENDscript server. Nucleic Acids Res 42:W320–W324. https://doi.org/10.1093/nar/gku316
36. Roy A, Kucukural A, Zhang Y (2010) I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc 5:725–738. https://doi.org/10.1038/nprot.2010.5
37. Sedgwick B, Morris C, French SJ (1978) Stereochemical course of dehydration catalysed by the yeast fatty acid synthetase. J Chem Soc Chem Commun. https://doi.org/10.1039/c39780000193
38. Stratmann A, Toupet C, Schilling W, Traber R, Oberer L, Schupp T (1999) Intermediates of rifamycin polyketide synthase produced by an *Amycolatopsis mediterranei* mutant with inactivated rifF gene. Microbiology (Reading, England) 145(Pt 12):3365–3375. https://doi.org/10.1099/00221287-145-12-3365
39. Terwilliger TC (2003) Statistical density modification using local pattern matching. Acta Crystallogr D Biol Crystallogr 59:1688–1701
40. Terwilliger TC (2003) Automated side-chain model building and sequence assignment by template matching. Acta Crystallogr D Biol Crystallogr 59:45–49
41. Terwilliger TC (2003) Automated main-chain model building by template matching and iterative fragment extension. Acta Crystallogr D Biol Crystallogr 59:38–44
42. Xie X, Cane DE (2018) pH-Rate profiles establish that polyketide synthase dehydratase domains utilize a single-base mechanism. Org Biomol Chem 16:9165–9170. https://doi.org/10.1039/c8ob02637h