Peer reviewed version

License (if available):
Unspecified

Link to published version (if available):
[10.1038/nchembio.1342](#)

[Link to publication record in Explore Bristol Research](#)
PDF-document

A conserved motif flags Acyl Carrier Proteins for β-branching in polyketide synthesis

Anthony S Haines[2a], Xu Dong[1a], Zhongshu Song[1], Rohit Farmer[2], Christopher Williams[1], Joanne Hothersall[2], Eliza Płoskoń[1], Pakorn Wattana-amorn[1], Elton R. Stephens[2], Erika Yamada[2], Rachel Gurney[2], Yuiko Takebayashi[2], Joleen Masschelein[3], Russell J. Cox[1], Rob Lavigne[3], Christine L. Willis[1], Thomas J. Simpson[1], John Crosby[1], Peter J. Winn[2], Christopher M. Thomas[2,*], Matthew P. Crump[1,*]

From School of Chemistry, Cantock's Close, Clifton, Bristol, BS8 1TS, UK[1]; School of Biosciences, University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK[2]; Division of Gene Technology, KU Leuven, Kasteelpark Arenberg 21 - box 2462, 3001 Heverlee, Belgium[3]

Running head: **ACPs specific for beta-branching**

*Address correspondence to: Professor M. P. Crump Tel. +44 (0)117 3317163. Fax +44 (0)117 9251295. E-mail: Matt.Crump@bristol.ac.uk. Professor C. M. Thomas Tel +44 (0)121 4145903. Fax +44 (0)121 4145925. E-mail: c.m.thomas@bham.ac.uk

[a] These authors contributed equally to this paper.

Type I PKSs often utilise programmed β-branching, via enzymes of an "HMG-CoA synthase (HCS) cassette", to incorporate various side chains at the second carbon from the terminal carboxylic acid of growing polyketide backbones. We identified a strong sequence motif in Acyl Carrier Proteins (ACPs) where β-branching is known. Substituting ACPs confirmed a correlation of ACP type with β-branching specificity. While these ACPs often occur in tandem, NMR analysis of tandem β-branching ACPs indicated no ACP-ACP synergistic effects and revealed that the conserved sequence motif forms an internal core rather than an exposed patch. Modelling and mutagenesis identified ACP Helix III as a probable anchor point of the ACP-HCS complex whose position is determined by the core. Mutating the core affects ACP functionality while ACP-HCS interface substitutions modulate system specificity. Our method for predicting β-carbon branching expands the potential for engineering novel polyketides and lays a basis for determining specificity rules.

Polyketide synthases (PKSs)[1] are responsible for an extraordinarily large and diverse group of natural products that have important pharmaceutical applications such as antibiotic, antitumor, antifungal, anticholesterolemic and antiparasitic agents[2]. PKSs are classified on the basis of their protein architecture; bacterial type I PKSs are large multifunctional polypeptides with all core enzymatic functions for elongation and modification of the carbon backbone grouped as modules. Type I PKS biosynthetic pathways are normally constructed with one module for each condensation reaction, often with additional modules that can make non-elongating modifications, or iterative modifications incorporating multiple units. The minimal functions in an elongating module are the ketosynthase (KS) domain, which acquires either a starter unit or the oligoketide from the previous module, and an acyl carrier protein (ACP) domain that holds the extender unit (most commonly malonate or methylmalonate). The KS catalyses a Claisen condensation, creating a new carbon-carbon bond in the ACP bound intermediate. Canonically, type I modules also contain an acyl-transferase (AT) that loads the extender unit onto the ACP (known as *cis*-AT systems). However, an increasing number of known Type I PKSs, including pathways considered here, are *trans*-AT systems which lack integral AT domains in the extension modules but that encode one or more separate ATs to perform this function[3].

After the Claisen condensation, modifications of the acyl chain may take place before transfer to the next module, most commonly β-ketoreduction (KR) alone, KR and dehydration (DH) or KR, DH and enoyl reduction (ER), yielding hydroxyl, alkene and alkane moieties respectively. In addition, *trans*-AT systems can introduce α-methyl groups with a methyl-transferase (MT) domain as part of the Type I module. A more complex modification found in numerous known *trans*-AT and several *cis*- type I systems occurs at the β-keto group and is introduced by a *trans*-acting "HCS cassette", typically comprising an ACP, an hydroxymethylglutaryl-CoA (HMG-CoA) synthase (HCS), two proteins belonging to the crotonase superfamily and a decarboxylase. The HCS cassettes can introduce otherwise difficult modifications such as β-methyl, β-ethyl, β-methoxymethyl, cyclopropane and vinyl chloride moieties in different systems[4]. This cassette normally acts on a type I module characterised by the absence of KR, DH or ER functions, often with tandemly repeated ACPs. In the myxovirescin system there are two HCS enzymes, one of which acts specifically at one of two β–branch ACPs. Understanding what signals such specific modifications is

important for rational engineering of these synthases.

One of the first polyketide biosynthesis clusters in which the HCS cassette was discovered is responsible for mupirocin synthesis by *Pseudomonas fluorescens*[5] (Fig. 1). Mupirocin, a clinically important antibiotic used primarily against *Staphylococcus aureus*[6], is a mixture of pseudomonic acids (Fig 1), which are esters of monic acid (MA) and 9-hydroxynonanoic acid (9HN). A very similar pathway builds the marinolic acid component of thiomarinol[7]. The final module in MA backbone biosynthesis is module 3 of MmpA, which was deduced to associate with the *mup* HCS cassette (mAcpC, MupG, MupH, MupJ and MupK) that is responsible for addition of the C-15 methyl group. This association is thought to be via the tandem ACP-mupA3a and ACP-mupA3b (Fig. 1 and see legend for nomenclature). Previous mutational studies with ACP-mupA3a and ACP-mupA3b suggested that they function independently in parallel, increasing pathway flow rate[8]. However, some but not all recent biochemical studies in other systems suggest synergistic effects that could implicate interactions between the ACPs[9,10]. These polyketide synthases (PKSs) provide a good system to investigate HCS cassette specificity. Remarkably we have identified a highly conserved core that is characteristic of almost all ACPs associated with these modifications even when they have distinct HCS specificities. We propose that the core orients helix III within the ACP structure that in combination with the amino acid composition in and around helix III determines its ability to interact successfully with its cognate HMG-CoA synthase.

RESULTS

Type I PKS ACP sequences associated with HCS function. To search for sequence motifs specific for type I ACPs in modules where β-branching occurs, we collected the sequences of ACP domains from seven well studied *trans*-AT systems that have the HCS cassette (mupirocin[11], thiomarinol[7], curacin[12], jamaicamide[13], pederin[14], bacillaene[15,16] and kalimanticin[17]) and aligned them using ClustalW2. This showed β-branch-associated ACPs as a subgroup and revealed a tryptophan residue six amino acids downstream of the active site serine in all β-branch-associated ACPs (Supplementary Results, Supplementary Fig. 1) but absent from the other sequences.

To test the predictive power of the conserved motif, which includes the key residues of the phosphopantetheine binding site and the tryptophan residue (DSxxxxxW), we performed PHI-BLAST searches for each β-branch-associated ACP from the *trans*-AT PKSs listed above using this stretch of residues. Besides the 7 input systems described above, this identified 52 ACPs from 17 PKSs (discounting closely related biosynthetic systems likely to yield the same product; Supplementary Table 1). These comprised 21 ACPs in 13 modules of 8 PKSs with known products (bryostatin, difficidin, elansolid, etnangien, onnamide, myxovirescin, psymberin, thailandamide[3]), plus a further 31 ACPs in 17 modules of 9 unstudied PKSs. Of the 13 modules in PKSs with known products, 12 are assigned β-branching function in published pathways, while one (in BryX, of the bryostatin cluster) has no characterised role.

To further define the characteristics of this group of ACPs we created sequence logos[18] (Fig. 2a and 2b) for the β-branch-associated ACPs in well studied clusters (Supplementary Fig. 1) and for the non-branching ACPs from the same systems. The sequence logo for the β-branch-associated ACPs showed that in addition to the phosphopantetheine binding motif (GLDS) and the tryptophan discussed above, there are a number of other positions that are highly conserved, or where there is an obvious difference in the consensus. This implies a difference in function, especially given the high degree of diversity at most other positions.

The context of the putative β-branching ACPs from clusters with unknown products was examined to assess the likelihood of involvement in β-branching: all are in modules lacking ketoreductase domains (which if present would reduce the ketone group and thus prevent β-branching) and having no additional non-branching ACP domains. All of these systems appear to include HCS cassettes (Supplementary Table 1). Only two exceptions were found: the virginiamycin and leinamycin clusters incorporate β-branches but do not have ACPs identified in this screen, due to absence of the motif's tryptophan. Therefore this motif (DSxxxxxW) represents a very strong predictor for β-branching activity and thus elucidation of pathway chemistry.

Classification of ACPs using Hidden Markov Models. Because the DSxxxxxW motif is not infallible in identifying ACP type, we developed Hidden Markov Models (HMMs) as an alternative method. HMMs estimate, for a given sequence type, the probability of different amino acids occurring at each position along the polypeptide chain, including an allowance for them occurring through insertions and deletions. We generated two HMMs, one for β-branch-associated and one for non-β-branch-associated ACPs, using the 15 PKS clusters with known products listed in the previous section (Supplementary Fig. 1) and then determined

how each ACP is scored by the two models. Plotted as a scatter graph (Fig. 2c), these training sets form two discrete sets, divided by the y=x line where no distinction is made between β-branching and non-β-branching. Outliers are the predicted β-branching ACPs from the virginiamycin cluster which are slightly above the y=x line, while of two tandem ACPs in the leinamycin cluster, one falls just below the line and the other is within the non-branching cluster. Thus with this one exception, the classification is consistent with available information about ACP association with β-branching or non-branching modules (Supplementary Table 1).

Searching two protein databases (TrEMBL and RefSeq, on 09/03/2012) using the non-branching HMM model gave a list of 16,490 non-identical ACPs that were greater than 60 amino acids in length. As with the training sets, a scatter graph of these sequences gave two clusters, with the vast majority falling into the non-branching cluster (Fig. 2d). By examining the context of all newly-found ACPs in or near the branching cluster, we could assign most as either likely β-branch-associated or not branch-associated. This revealed that above a branching HMM score of about 45, the branching state of ACPs near the y=x line is not discriminated. One leinamycin cluster ACP is in this ambiguous region while for the other, β-branching is clearly not predicted. Only one ACP (in the sorangicin cluster, which has no HCS cassette) has a branching HMM score much greater than the standard HMM score yet appears non-β-branching. Interestingly, the KS of the following module was placed in a β-branch-accepting clade[19]. In combination this strongly suggests that the pathway previously specified a β-branch at this position. One further curiosity is the identification of two apparently β-branch-associated ACPs, present as individual genes - each adjacent to an HCS cassette, but not associated with a PKS cluster (CAM03837, EDY67341): these may bring β-branch chemistry to pathways other than type I PKS systems. Finally, we emphasise that the two ACP types in the myxovirescin system, associated with different substitutions are both clearly identified by the HMM model. Therefore, while the motif predicts β-branch association, it does not distinguish subtypes if they exist.

NMR solution structure of ACP-mupA3ab. To investigate structures of conserved sequences in β-branch-associated ACPs we cloned ACP-mupA3a and b with the natural linker between them but no natural sequences upstream or downstream. His$_6$ tags were added at both the N- and C-termini producing a soluble protein for NMR structural analysis (Supplementary Fig. 2a,b) whereas WT ACP-mupA3ab was stable for <24 hours (Supplementary Fig. 2c). A ^{1}H-^{15}N heteronuclear single quantum coherence (HSQC) spectrum of uniformly ^{13}C, ^{15}N labelled apo ACP-mupA3ab at 500 μM, (Supplementary Fig. 2d) showed 176 resolved cross-peaks out of a possible 183 non-his$_6$-tag residues indicating a single species in solution. Backbone and side-chain chemical shifts were assigned and covered >95% of the ^{1}H, ^{15}N and ^{13}C atoms. The Sfp-type[20] mupirocin cluster phosphopantetheinyl transferase MupN[21] was used to prepare holo ACP-mupA3ab. Both serines were 4'-phosphopantetheinylated (4'-PP) to completion (Supplementary Fig. 2e). The holo ACP-mupA3ab spectrum overlapped with the apo form[22] except for the modified serines, Ser38/142. The solution structure of apo ACP-mupA3ab was calculated based on 2002 distance restraints and 239 dihedral angle restraints (Supplementary Table 2). The paucity of restraints per residue may be due in part to the high percentage of flexible/loop regions in these structures and other ACPs (~30%) with long-range restraints limited to inter-helical contacts and the small hydrophobic core within each domain. Contributions from weaker nuclear Overhauser enhancements (NOE) may also be missing due to lowered sensitivity when working with a protonated 20 kDa protein at the slightly lowered temperature of 20 °C.

The structure of apo ACP-mupA3ab consists of two discrete folded ACP domains linked via a flexible unstructured linker (Fig. 3a). Each domain has an ACP fold with three main anti-parallel α-helices (I, II and IV) and a short α-helix (III). The 4'-PP attachment site of each domain (Ser38 and Ser142) is at the N terminus of helix II and both ACP-mupA3a and b have a long flexible loop linking helices I and II (Supplementary Fig. 3a,b). The flexible linker (Leu78-Ala105) between the two ACP domains is poorly defined due to a lack of distance restraints. Superposition of the ACP domains showed high similarity (backbone RMSD over the structured regions of 1.28 Å; See Supplementary Fig. 3c). No inter-domain or ACP-linker NOEs between the ACP domains and accordingly in the best twenty calculated structures, there were no consistent interactions observed between the two domains. Residual dipolar couplings (RDCs) were measured to explore whether the ACP domains were partially oriented with respect to each other but weak alignment and poor stability in several alignment systems meant no useful restraints could be derived. Finally the backbone ^{1}H-^{15}N chemical shifts of ACP-mupA3ab were compared to those from ^{15}N labelled, single expressed domain, ACP-mupA3a (Supplementary Fig. 2f) revealing that most were superimposable with the exception of small changes close to the truncation point at the C-terminal of ACP-mupA3a. The lack of observable chemical shift perturbations suggests there are no significant inter-ACP interactions as

suggested by related structural[10] and functional[8] studies.

These structures show that the hydrophobic cores of the ACPs are formed by the side chains of twelve residues (Fig. 3b). The most conserved residues are packed between helix I and II and behind the long loop joining helices I and II. The conserved Trp44/Trp148 of ACP-mupA3a and ACP-mupA3b pack between helices I and II (27 and 34 restraints per residue respectively) and replace smaller residues, such as threonine, valine or leucine, at the equivalent positions in non-β-branching structures. Therefore the majority of the key residues identified by our bioinformatic analysis (apart from conserved residues immediately around the active site serine, just after helix II and in or around helix III) pack around Trp44/Trp148 in this core.

Comparison to Type I and Type II ACP structures. Apart from two recently characterised HCS-associated ACPs from the curacin cluster[10] the closest published type I ACP counterparts are the fatty acid synthase (FAS) ACPs from rat FAS (2PNG) and yeast[23] (2UV8), as well as the PKS module 2 of the 6-deoxyerythronolide B synthase (DEBS)[24] (2JU2) and the norsolorinic acid synthase (NSAS)(2KR5)[25]. These ACP structures aligned with the well-ordered regions of both ACP-mupA3a and b (helix I, II and IV) with an RMSD ranging from 2.4 to 4.6 Å calculated using Profit. Superimposition on type II ACPs revealed similar RMSD values. A clear difference however between ACPs is in the orientation of helix III. This lies at ~60° to the axes of helices II and IV in ACP-mupA3a and ACP-mupA3b, rat FAS ACP and 6-DEBS ACP, almost perpendicular in NSAS ACP and more parallel in the *S. coelicolor* type II actinorhodin (act) ACP (PDB code: 2AF8)[26] (Fig. 4). Examination of these structures suggests that this orientation may be determined by the burial and distribution of key bulky hydrophobic side-chains around and within this helix. For ACP-mupA3a and ACP-mupA3b, helix III is anchored by I61/I165 with Y62/Y166 exposed at the surface. Helix III has been shown to be flexible and lacks packing interactions (Fig. 4) in several Type II ACPs and is an important hinge region allowing the structure to accommodate non-polar and, to a lesser extent, polar side chains[27]. Conversely ^{15}N relaxation data for Type I rat FAS ACP reveals the more stable helix III does not display significant flexibility and this protein shows no propensity to sequester fatty acid chains. Examination of the ensembles for both ACP-mupA3a and ACP-mupA3b reveals helix III is well defined by the structural restraints observed, consistent with stabilisation via packing of I61/I165, but this does not preclude the ACP accessing other stable sub-states if ligated.

Trp44/148 are important for ACP structure and function. Since either of the two ACPs, ACP-mupA3a or ACP-mupA3b, is sufficient for function in this module[8] we created chromosomal W>L mutations (L is the most common alternative at this position in non-branching ACPs) in either one or both ACPs: the single mutations significantly reduced PA-A production while the double mutant caused further reduction (Figure 5a). We expressed and attempted to purify both the wild type and W>L mutant ACP-mupA3a single domain and from bacteria with and without the MupN PPTase to convert it to holo-form. WT ACP-mupA3a was successfully isolated in both apo- and holo-form as established by NMR (Supplementary Fig. 2f) and ESMS (ESMS: observed/calculated mass: apo 13873.2/13872.8 Da, holo 14212.0/14212.8 Da), which matched the expected structures. By contrast, the W>L mutant protein was difficult to characterise for apo- and impossible for the holo form due to insolubility. Circular dichroism spectra of the soluble WT and mutant apo proteins suggested a change in alpha helical content and molecular dynamic simulations indicated that the backbone of helix III is more mobile in the mutant (Supplementary Fig. 4). The ^{1}H-^{15}N HSQC spectra obtained for the mutant were typical of partially folded protein confirming that the Trp44 is critical for core stability of ACP-mupA3a.

Replacement of ACP-mupA3ab with heterologous ACPs. To test whether ACPs provide the signal for β-branching, we replaced ACP-mupA3ab with heterologous β-branch-associated or non-β-branch-associated ACPs from the thiomarinol cluster[7]. The thiomarinol biosynthetic pathway builds a similar product to mupirocin but its DNA sequence is sufficiently different to avoid problems that might arise from homologous recombination. Bioassay, HPLC and LCMS analysis showed that replacement with the equivalent thiomarinol triplet of β-branch-associated ACPs tmlA3abc maintained and may actually increase mupirocin production (Fig. 5b). By contrast, insertion of the first two ACPs from TmpD module 3 (ACPs tmlD3ab), that are non-β-branch-associated[7], showed no PA-A or PA-B production. Interestingly this ACP-mupA3ab:tmlD3ab substituted strain grew more slowly than either wild-type ACP-mupA3ab deletion, double W>L mutations and ACP-tmlA3abc replacement strains. When we repeated the ACP-mupA3ab:tmlD3ab replacement in a ΔmupI strain[28] in which expression of the mupirocin biosynthesis pathway is greatly reduced, the resultant strain had the normal colony morphology. Thus growth rate

reduction may result from an abnormal product of the non-β-methylating module 3 of MmpA rather than a completely defective system. Thus, the *mup* HCS cassette can recognise β-branch-associated ACPs from another cluster but cannot work with ACPs from a non-β-branch-associated module, consistent with there being specific ACP recognition by enzymes of the HCS cassette. In addition, because each individual ACP varies from the consensus these replacements allow identification of alternative residues at conserved positions that are acceptable in the right context.

Critical residues at the ACP-HCS interface. To explore the ACP-HCS interface, models of the MupH:ACP-mupA3a and MupH:ACP-mupA3b complexes were created. By constraining the conserved serine of the ACP to the phosphopantetheine portion of the substrate in the MupH active site the HADDOCK webserver[29] produced just one cluster with four similar complexes for ACP-mupA3a and these were equivalent to two of the three possible clusters produced for ACP-mupA3b. All the solutions from Haddock had helix III of the ACP interacting with MupH, and this seems to be a consequence of the ACP-bound substrate being in the active site of MupH. The consensus docking solution (Fig. 6a, Supplementary Fig. 5) was further supported by evolutionary trace analysis[30] of sequence conservation and analysis of protein surface physicochemical properties using PIER[31] (Supplementary Fig. 6, Fig. 4). For ACPs-mupA3a/mupA3b residues Ser 38/142 (active site serine), Val39/143, Asp59/163, Tyr62/166, Thr63/167 (three residues on the surface of Helix III) were found at the interface of *all* complexes (Supplementary Tables 3 and 4 for a full list of interacting residues). For the consensus solution, Y62, conserved in 95% of beta-branching ACPs, sits in a pocket interacting with a methionine (257) conserved in all analysed MupH orthologues except TaF; interactions between aromatic residues and Met have recently been recognised as a potent stabiliser of protein structures[32]. Mutation of ACP-mupA3a Y62 to F or A in a ΔACP-mupA3b strain reduced antibiotic production 3- to 4-fold or up to 10-fold respectively (Figure 6b). Thus Y62 is critical for function and Phe, that is common at this position in non-branching ACPs, cannot replace it efficiently, indicating that the hydroxyl group is important for the interaction with HCS.

Seven of the eight HCS residues at the core of the interface with the cognate ACP and 4/5 at its periphery are highly conserved in the HCS proteins, despite only 142 overall conserved positions from 427 aligned residues (33%; Supplementary Fig. 7). The eighth core interface residue (residue 419 in MupH) is either L or M except in TaF where it is F, and is predicted to locate between the active site and the newly discovered pocket (Supplementary Table 4). Since the myxovirescin (Ta) cluster is known to have two HCS specificities associated with common additional HCS cassette functions[33] these differences may represent divergence in recognition despite conservation of the core. We chose the HCS homologues TmlH and BatC (with 68% and 64% identity to MupH respectively), from the thiomarinol[7] and the kalimantacin/batumin[17] clusters to test the effect of having M and L at this position (WT MupH and TmlH have M while BatC has L). Both *tmlH* and *batC* were found to restore PA-A production in a *mupH* deletion strain although BatC did so less well than TmlH indicating significant divergence of recognition and thus subtype of β-branching ACP. From this analysis we predicted that mutation of BatC from L to M would improve recognition of ACP-mupA3a/b and indeed it increased BatC ability to complement a ΔmupH mutant more than 3-fold (Fig. 6c,d), consistent with this residue being part of the recognition motif.

This analysis supports the model of ACP-HCS interaction and suggests that ACP:HCS recognition depends on both docking into both the active site of HCS and this newly identified pocket as well as the rest of the interface. It emphasises the importance of I-II loop and helix III regions as being key determinants of interaction specificity in the ACP.

DISCUSSION

This work aimed to identify features of ACPs associated with the action of "HCS cassette" enzymes in performing β-branching, and to establish whether structural interactions between these tandem ACPs might explain the apparent synergistic effects reported in other systems[9]. The structure of ACP-mupA3ab indicates that the ACP domains do not interact and should therefore function as two independent ACP units that facilitate parallel condensation plus β-methylation. We proposed this conclusion previously based on genetic studies since either ACP alone can catalyse the elongation of the polyketide chain while the number of ACPs determined the rate of polyketide production, presumably by overcoming a rate limiting step[8]. This is consistent with other reports relating polyketide yield to the concentration of ACP present[34-36]. However, Sherman and co-workers reported some limited synergy effects between the ACP domains in the ACP tri-domain of CurA, responsible for formation of the acetoacetyl substrate for the HCS cassette involved in curacin biosynthesis[9]. Analytical size exclusion chromatography was consistent with flexible structures

lacking domain-domain interactions and more recent functional data also contradicts these earlier findings[10]. Although ACP-ACP interactions may occur in other systems, our data suggest that, if deduced synergistic effects observed are real, they may result from increased efficiency related to the increased number of functional protein units. In line with queue theory[37] this effectively buffers the slow HCS processing steps against random surges in substrate flow. Not only would flow capacity be increased by multiple ACPs but the rate of interactions with the HCS cassette functions could also increase, leading to the non-linear increases observed.

Turning to the sequence motifs identified for the β-branching-associated ACPs that form a powerful predictive tool for analysing new systems, the striking result is that the majority of the highly conserved residues (including L10/114, L14/118, L18/122 and F31/135) pack with the bulky tryptophan in the core of the ACP (Fig. 3 and 4), so it was initially unclear whether this motif is functional in determining recognition. We considered the possibility that the absence of ketoreduction might be more important than the sequence motif since in many systems the only modules that lack a KR domain are HCS-associated. However, we show that heterologous ACPs from a β-branching module can substitute for ACP-mupA3ab function whereas ACPs from a non-β-branching module could not. Biochemical studies using purified proteins have also shown specificity[33]. It is therefore clear that it is the proteins that determine the specificity, not simply the presence of a β–keto group in the substrate.

The question therefore appears to be whether the core motif allows the HCS cassette to differentiate all β-branch-associated ACPs from other ACPs or whether it also relies on amino acid differences at the ACP:HCS interface? A key discovery was a pocket in HCS apparently acting as a second anchor point in the ACP-HCS complex (the first being the substrate arm and HCS active site), accommodating the conserved tyrosine on helix III. Mutation of this tyrosine to phenylalanine, the commonest residue at this position in non-β-branch-associated ACPs, reduced Pseudomonic acid production drastically (Fig. 6b), confirming the importance of this residue specifically in β-branch-associated ACPs. While this shows that the nature of the amino acid side chain at this point is critical, the distance and angle relative to the active site PPT arm as discussed in relation to Fig. 4, which need to be correct to allow docking, may help distinguish β-branch-associated ACPs from other ACPs and depend on the core. Indeed molecular dynamics simulations of WT and W>L ACP-mupA3a mutant showed much greater dynamic fluctuation in and around helix III of the mutant (Supplementary Fig. 4). Interestingly, helix III was also identified as important for the interaction of a similar ACP with the unique halogenase of the CurA module involved in curacin biosynthesis[10]. Our modelling and analysis of the ACP-HCS interface also revealed differences in key HCS residues that may further determine specificity, and by gain of function site-directed mutagenesis we could improve complementation of a *P. fluorescens* NCIMB10586ΔmupH mutant by BatC, the HCS from the *kal/bat* cluster.

Pulling together our observations and those of others[10], ACPs can be portrayed as having two arms that lie side by side: we can imagine the substrate on the active site and the residues on helix III as "pins" as in a two pin plug (Fig. 6a and graphical abstract). We propose that the conserved core of β-branching ACPs including the tryptophan presents these "pins" so that it docks correctly with HCS, but that specific interface amino acid interactions are required for productive docking of ACP/HCS pairs or differentiation of different ACPs by cognate HCS enzymes in a single system like myxovirescin biosynthesis[24]. While the exact details of this recognition require much further study and are beyond the scope of this paper, the work described here has established a means for reliably identifying β-branching-associated ACP specificity in polyketide and other synthetic pathways. Furthermore, it provides a basis for exploring the specific interaction of identified ACP/HCS pairs, which will allow the precise and rational incorporation of one or more of beta methylation, chlorination, cyclopropane and other HCS cassette associated modifications into re-engineered biosynthetic pathways.

Author contributions
X.D., purified ACP-mupA3a and ACP-mupA3ab and calculated the NMR structures; P.W-A., purified ACP mup-A3ab; E.P., purified wild type ACP mup-A3ab; C.W. calculated NMR structures; C.M.T. co-designed the study and wrote the paper; M.P.C, co-designed the study and wrote the paper; J.C. was critical for interpreting data on ACP structure and function; T.J.S. provided expertise in polyketide biosynthesis; A.S.H. discovered the conserved motif, carried out bioinformatic analysis and designed key elements of the experimental tests of our predictions; R.J.C. designed the LCMS analysis; Z.S. did the LCMS analysis; R.F., Y.T., and P.J.W. did bioinformatic analysis, the molecular modelling of the ACP mutants and of the ACP:HCS complex; R.L. and J.M. performed initial BatC complementation assays; E.Y. and R.G. isolated

and characterised the mutant ACP-mupA3a.

Acknowledgments

FIGURE LEGENDS

Figure 1. Biosynthetic pathway for PA-A and sequence comparisons of β-branching vs. non β-branching ACPs. Proposed biosynthetic pathway of monic acid and the formation of mupirocin H. ACP domains mupA3a and b are involved in the third module of MmpA. (this nomenclature[7] indicates a domain from the mupirocin ("mup") cluster, PKS "A" [MmpA] as the protein, "3" as the module number within the protein and then "a" and "b" to designate the tandem domains) HCS cassette enzymes are defined as follows; malonyl-ACP decarboxylase (KSQ), β-hydroxyl-β-methyl glutarate CoA synthase (HCS), enoyl-CoA hydratase (ECH). The red dot highlights the position of carbon 3 in pseudomonic acid as it progresses through the biosynthetic pathway.

Figure 2. Sequence logos showing sequence conservation in β-branch-associated ACPs and their classification using Hidden Markov Models. Sequence logos of: (a) non-HCS cassette-associated ACPs; and (b) β-branch-associated ACPs. Amino-acid conservation is shown by the relative height of the letters in individual stacks. The letter 'O' represents a gap. Key residues and buried hydrophobes in ACP-mupA3a/b are highlighted. (c) ACPs from 26 pks clusters with at least one known or predicted branching ACP: (□) Training set for HMM using known branching ACPs; (■) predicted branching except (▲) virginiamycin cluster branching ACPs and (×) leinamycin cluster β-branch-associated module ACPs; (◇) Training set for HMM using non-branching ACPs (◆) predicted non-branching ACPs. (d) 16,490 ACP-like sequences identified by screening the TrEMBL and RefSeq protein databases using the standard ACP HMM: (◆) branching ACP (See Supplementary Table 1); (■) unlisted variants in similar clusters; (×) known branching (not in Supplementary Table 1); (▼) predicted branching, identified in this screen; (+) ACPs which may add branches in a non-type I-pks pathway; (●) insufficient sequence context or conflicting information; (▲) predicted non-branching; (•) not examined. Sequences below the branching HMM cut-off were conferred a score of 10 for plotting.

Figure 3. Solution NMR structures of ACP-mupA3a/b. (**a**) ACP-mupA3ab structure coloured as ACP mupA3a (maroon), ACP mupA3b (blue) and domain-domain linker (green). The four α-helices in ACP mupA3a and b are as follows: I, Val1-Leu18/Gln106-Leu122; II, Ser38-Gln53/Ser142-Phe156; and IV, Phe67-Gln77/Leu171-Gln183) and one short α-helix (helix III, Asp59-Thr63/Ala164-Trp168) (**b**) Top view comparison of the hydrophobic core packing in mupA3a (maroon) and mupA3b (blue) ACPs. For clarity only eight of the twelve residues comprising the core (Leu10/114, Leu14/118 and Leu18/122 in helix I [defined experimentally by 28 ± 5 restraints per residue]; Phe31/135 in the loop between I and II (37 and 17 restraints); Trp44/148, Ile45/Met149, Ile48/Val152 and Tyr52/Phe156 in helix II [30 ± 7 restraints per residue]; Ile61/Ile165 in helix III [42 and 30 restraints per residue respectively]; and Phe67/Leu171, Phe70/Leu174 and Val74/178 in helix IV [26 ± 10 restraints per residue]) are shown. Leu14/118, Leu18/122, Phe31/135 and Trp44/148 are conserved residues common to all HCS cassette associated ACPs.

Figure 4. Helix III packing in type I and II polyketide ACP structures. Cartoon image of helices II-IV of the (**a**) fungal type I NSAS ACP (yellow), (**b**) ACP-mupA3a (maroon), (**c**) ACP mupA3b (blue), (**d**) module 2 DEBS type I ACP from *S. erythrea* (orange), (**e**) type I rat FAS ACP (green) and (**f**) act ACP (purple). The orientation of helix III is highlighted with red lines. For clarity, helix I is omitted as its relative orientation is unchanged but it was included in the superposition (helices I, II and IV). Helix III is shown beneath each panel with key residues highlighted. Phe1768 and Leu1770 are buried in NSAS ACP. ACP-mupA3a and ACP-mupA3b have the single, partially exposed aromatic residue Tyr62 and Tyr166 respectively but lack

the equivalent residue to Phe1768 and helix III lies at 60° to the axes of helices II and IV. Likewise in helix III of 6-DEBS ACP, there are two packed hydrophobic residues, Leu76 and Val77 (equivalent to Leu1770 and Phe1771 in NSAS ACP) whilst Phe78 is again solvent exposed. Similarly in rat FAS ACP a single hydrophobe (Val61) packs within the hydrophobic core and again helix III lies at approximately 60º to helices II and IV. Finally, there are no amino acids with bulky hydrophobic side chains in helix III of actinorhodin (act) ACP which runs almost anti-parallel to the adjacent helices (which is maintained in the ligated states), perpendicularly in the NSAS ACP, but tilted at ~60° to the protein axis in 6-DEBS and ACP-mupA3a/b.

Figure 5. Analysis of Pseudomonic Acid A production by *P. fluorescens* NCIMB10586 wild type and ACP substitution strains. (**a**) Comparison of PA-A production by single and double W>L mutants of ACPs mupA3a and mupA3b. Assigned with purified standard compounds the yield of PA-A was measured at the retention time of 20.4 min and a [M-H]⁻ of *m/z* 499. The main peak is PA-A and the two leading minor peaks are due to two rearranged forms of PA-A, in which the epoxy group in PA-A cyclises with the C7 hydroxy group to give 5- and 6-membered rings respectively (4 replicas each). (**b**) Comparison of mupA3ab substitution mutants by tmlA3abc ACPs (β-branching) and tmlD3ab ACPs (non-β-branching) (2 replicas each).

Figure 6. Model for ACP-mupA3a-MupH complex and complementation of *P. fluorescens* NCIMB10586Δ*mupH* by MupH homologues. (**a**) The interface of the predicted complex of MupH (green) and ACP3/4 (cyan) with the ligand highlighted as spheres, the key interface residues on ACP (Ser 38, Val 39, Asp 59, Tyr 62 and Thr 63) and the interacting pairs on MupH (Arg 34, Met 257 and Arg 263) highlighted as green and yellow sticks respectively. (**b**) HPLC comparison of supernatants from strains demonstrating that ACP mupA3a Y62F (F commonly occurs at this position in non-β-branching ACPs) in a strain that lacks ACPmupA3b causes almost as big a decrease in Pseudomonic Acid production as ACP mupA3a Y62A. 2-4 replicas for each production experiment. (**c**) Plate bioassays showing restoration of Pseudomonic Acid production (clear zone around producer patch) to a Δ*mupH* mutant by expression of *batC* (WT and L>M mutant) of *P. fluorescens* BCCM_ID9359 in vector pJH10, with and without IPTG to induce transcription of the cloned gene from the *tac* promoter. Scale bars are 10mm. There were 12 replicas for each of the strains in the bioassay. (**d**) HPLC analysis of supernatants from IPTG-induced strains demonstrating the increased complementation of the mutation by the BatC L>M mutant compared to WT. WT + pJH10 2 replicas, Δ*mupH* 2 replicas, Δ*mupH* + BatC 7 replicas, Δ*mupH* + BatC L>M 18 replicas.

ONLINE METHODS

Bacterial strains and growth conditions. *Escherichia coli* strains used were: DH5α; BL21(DE3); S17-1; TOP10 (Invitrogen). Other bacteria were: *Pseudomonas fluorescens* NCIMB10586, the mupirocin producer, (WT and selected mutants); *P. fluorescens* BCCM_ID9359, the kalimantacin producer; *Pseudoalteromonas* spp SANK73390, the thiomarinol producer[38]. Growth medium was L Broth[39] or L Agar supplemented where necessary with antibiotics (ampicillin, 100 μg/ml; kanamycin, 50 μg/ml) unless stated. Polymerase chain reaction (PCR), gene cloning and expression were carried out by standard techniques[40] using oligonucleotide primers listed in Supplementary Table 5 where appropriate. Plasmid vectors for gene cloning and expression were as follows: pGEMT-Easy, a T-tailed vector used for cloning PCR products on the basis of the 3' A extensions; Zero Blunt TOPO vector (Invitrogen); pET28a (Novagen), the standard T7 promoter based vector that adds His_6 to the N-terminus of the gene; pAKE604, used for suicide mutagenesis[28]; pJH10, broad-host-range *tacp*-based expression plasmid[11].

Expression, mutagenesis and replacement of ACP-mupA3a/b. For expression the coding region was PCR amplified from *P. fluorescens* NCIMB10586 genomic DNA, ligated into Zero Blunt TOPO vector (Invitrogen) and transformed into *E. coli* TOP10 (Invitrogen). The correct insert was confirmed by sequencing before sub-cloning into pET28a. The generation of plasmid pJS563, pET28a-WT ACP-mupA3a, was described previously[7]. The W>L mutant was amplified from the mutant segment constructed for site-directed mutagenesis of the chromosomal genes in the cluster (see below). Seed cultures of BL21(DE3) with and without pJHN11 (pJH10 plus *mupN*[21]) were inoculated from glycerol stocks and, after collecting bacteria by centrifugation, used to inoculate 5x200 ml L Broth. Growth at 37 °C and 250 rpm to OD_{595}= 0.8-1.0, was followed by induction with 0.25 mM Isopropyl β-D-1-thiogalactopyranoside (IPTG) over two days at 18 °C and 250 rpm. Bacteria were harvested by centrifugation and lysed by sonication. The soluble fraction was collected by centrifugation and purified using a 5 ml HiTrap (GE Healthcare) column. For suicide mutagenesis a pAKE604 derivative was constructed containing both ACP-mupA3a and ACP mupA3b with their active sites flanked by arms of approximately 500 bp. The W>L mutations were introduced with mutagenic primers. Mutated segments were checked by sequencing. Recombination of mutations into the *P. fluorescens* NCIMB10586 chromosome was performed as described previously[28]. PCR amplification of the mutagenic segment followed by *Afl*III restriction digestion identified derivatives with mutations in one or both ACPs. Y62>F and Y62>A mutations were generated in cloned ACP-mupA3a with mutagenic primers and transferred into an ACP-mupA3b-deleted strain as above. To replace *mup* ACPs by those of the thiomarinol cluster[7] the latter were amplified by PCR and inserted between the arms used to make the in-frame *ACP-mupA3ab* deletion[8] before recombination into the chromosome of *P. fluorescens* as described before[28].

Expression and mutagenesis of *batC*. For complementation by *batC* the coding region was amplified with primers listed in Supplementary Table 5 and inserted into pJH10 under the control of *tacp*. The L>M change was introduced by two-step PCR using the internal primer batCLtoM in the first step. Bioassay images were processed as a group to improve contrast (IrfanView - "auto adjust colors").

Isolation and LCMS analysis of metabolites in ACP-mupA3ab mutants. L-medium agar plates were inoculated with test strains and incubated at 30 °C for 30 hrs. A single colony from each mutant and wild type strain was picked and inoculated into 30 mL of L-broth with carbenicillin (50 μg/ml culture). The flasks were incubated at 25 °C and 200 rpm overnight. 5 ml of seed culture was used to inoculate each of three 500 mL flasks containing 100 mL of mupirocin production medium and then fermented (22 °C, 220 rpm, 48 hrs). After the cultures were combined, the cells were removed by centrifugation. The supernatant was then extracted by ethyl acetate (1:1), acidified to pH 5.0 followed by a further extra ethyl acetate extraction. The extracts were combined and ethyl acetate was evaporated *in vacuo*. The residue was dissolved in 3.0 mL MeOH and analytical samples were prepared by further 10-fold dilution with MeOH and analysed by LCMS using a Waters HPLC system. Detection was achieved by U.V. between 200 and 400 nm using a Waters 2998 diode array detector, and by simultaneous electrospray (ES) mass spectrometry using a Waters Quattro Micro[TM] spectrometer detecting between 150 and 600 *m/z* units. Chromatography (flow rate 1 mL·min^{-1}) was achieved using Phenomenex LUNA column (5 μ, C_{18}, 100 Å, 4.6 × 250 mm). Solvents were: A, HPLC grade H_2O containing 0.05% formic acid; B, HPLC grade CH_3CN containing 0.045% formic acid. Gradients were as follows: 0 min, 5% B; 22 min, 60% B; 24 min, 95% B; 26 min, 95% B; 27 min, 5% B; 30 min, 5% B.

Both positive ion (PI) and negative ion (NI) mode were employed for the characterization of the target compounds together with U.V. absorption pattern. The target compounds were measured by selected ion monitoring (SIM). Assigned with purified standard compounds, the yield of PA-A was measured at the retention time of 20.4 min and a [M-H]$^-$ of m/z 499, mupiric acid at retention time of 16.5 min and a [M-H]$^-$ of m/z 185. As negative ion signal for mupirocin H was very weak, the compound was measured at a retention time of 14.7 min and a (M+Na)$^+$ of m/z 295.

Production of ^{15}N/^{13}C ACP-mupA3ab. The pET28a (Novagen) mupA3ab plasmid was transformed into *E. coli* BL21(DE3), and a single colony used to inoculate 5 ml L-broth. After growth for 4 hrs at 37 °C and 250 rpm the culture was diluted into 2 L L-broth or M9 media with ^{15}NH$_4$Cl and ^{13}C-glucose. After growing at 37 °C and 250 rpm to OD$_{595}$ = 0.8-1.0, IPTG was added to 0.5 mM and growth continued overnight at 18 °C and 250 rpm. Cells were harvested by centrifugation and lysed by sonication. The inclusion bodies were separated from the cell debris by centrifugation, washed with 50 mM Tris buffer (100 mM NaCl, pH 8.5) and dissolved in 6 M guanidine HCl solution. The solution was then incubated with 5 mM DTT for 1 hr at room temperature before drop-wise addition into vigorously stirred refolding buffer (1 M arginine, 50 mM Tris-HCl, 2 mM EDTA, 5 mM DTT, pH 9.5) and left at 4 °C overnight with gentle stirring to refold followed by further purification using gel filtration in 50 mM Tris buffer (150 mM NaCl, pH 8.5). For unlabelled ACP-mupA3ab, ESMS confirmed production of the intact protein (observed 23598.0 Da, calculated 23601.5 Da).

Cloning, Expression and purification of his$_6$-MupN.   Cloning of his$_6$-MupN (holo-ACP synthase). The *mupN* gene from *P. fluorescens* NCIMB 10586, previously cloned into vector pJH10[21], was sub-cloned into pET28a (Novagen) vector to create pET28mupN using the EcoRI and SacI sites and sequenced (Lark) to confirm the correct sequence. *E. coli* BL21(DE3) (pET28a-mupN) was grown in 3-5 ml L-Broth at 37 °C for 4 hrs before inoculation into 2 L of the same medium. At exponential phase the cells were induced with 1mM IPTG and incubated at 37 °C and 250 rpm for 4 hrs before harvesting and lysis by sonication. The soluble fraction was separated by centrifugation and His$_6$-MupN purified using his-bind resin equilibrated with 50 mM potassium phosphate buffer (100 mM NaCl, glycerol 10%, pH 8.0) the column was then washed with 50 mM imidazole and the protein eluted with increasing concentrations of imidazole (150 mM, 250 mM and 0.5 M) before buffer-exchange into 50 mM phosphate buffer (glycerol 10%, pH 8.0).

Conversion of ^{15}N labelled apo to holo ACP-mupA3ab ACP. To convert (^{15}N) apoACP-mupA3ab (60 µM) to its holo form, MupN (5 µM), coenzyme A (2 mM), DTT (5 mM) and MgCl$_2$ (10 mM) were added and the reaction was incubated at room temperature overnight with shaking. Complete conversion was confirmed by ESMS (observed 24279.0 Da, calculated 24281.5 Da).

NMR experiments. The solvent conditions for NMR were optimized using micro-drop screening[41]. The protein was found to be stable for up to a month at concentrations below 500 µM in sodium phosphate buffer (pH 7.0) and a temperature of 20 °C. The refolded and purified isotopically labelled ACP mupA3ab was concentrated and buffer-exchanged (Millipore, Amicon Ultra-15, 10000 MW) into 20 mM sodium phosphate buffer (5 mM DTT, 8 % D$_2$O, pH 7.0). NMR experiments were acquired at 20 °C on a 600 MHz Varian VNMRS^{TM} spectrometer equipped with a cryogenically cooled probe (funded by the Wellcome Trust (WT082352)). Standard triple resonance experiments were used to assign the backbone and side-chains. One ^{15}N-edited and two ^{13}C-edited NOESY datasets (centred on the aliphatic and aromatic region) were acquired with 100ms mixing times. All the NMR data were processed using NMRPipe spectral processing and analysis system, and the assignment and NOE data collection were carried out using CCPN Analysis Version 2.1.3[42].

Real value evolutionary trace analysis. The real value evolutionary trace analysis for ACP mupA3a/mupA3b and MupH was performed using the Evolutionary Trace Viewer (http://mammoth.bcm.tmc.edu/traceview/index.html)[30]. The multiple sequence alignment needed for real value evolutionary trace analysis was carried out using ClustalW with 95 and 75 sequences for ACP mupA3ab and MupH respectively, followed by minor manual improvement in Seaview.

Residual Dipolar Couplings. For the measurement of RDCs, the refolded ACP mupA3ab was dissolved in 20 mM sodium phosphate (5mM DTT, 2mM EDTA, 5% D_2O and pH7.0) and incubated with pre-dried 5% polyacrylamide gel in a microfuge tube at 4 °C for 48 hrs (ACP mupA3ab was not stable in PEG/hexanol mixtures). The resultant polyacrylamide gel was squeezed into a bottomless NMR tube (ID 4.2 mm) using a syringe-like device and capped with a PTFE plug-in. The ^{1}H-^{15}N HSQC and IPAP HSQC spectra[43] were acquired using a Varian *INOVA*™ 600 spectrometer. A total of 38 RDCs could be determined although these were small (-2 to 3 Hz) with the majority of couplings being < 2Hz, indicating only weak interactions between the protein and gel matrix The outliers in the set of RDCs were selected using Module[44] based on the calculated ACP mA3ab structure. The alignment tensor, which was used in the structure calculation, was also defined using Module. Consequently when the RDC data was fitted to the calculated ACP mupA3ab structure, the alignment tensors of different helices in the same ACP domain were not collinear indicating a failure to converge to a single solution for the alignment tensor of either domain. As the addition of RDC restraints into the structure calculation did not improve the quality of the models, they were subsequently not included.

Structure calculation and refinement. The φ and ψ dihedral angle restraints were produced using Torsion Angle Likelihood Obtained from Shift and sequence similarity (TALOS)[45]. The structure calculation was performed using Ambiguous Restraints for Iterative Assignment (ARIA) version 2.2[46]. One hundred structures were progressed to the final round of calculations and the final 20 NMR structures were further refined in explicit water. The RMSD values were calculated using Molmol, and the structure qualities were predicted using WHATIF[47] and PROCHECK[48]. The final structures showed no distance violations greater than 0.5 Å and no angle violations greater than 5°. The Z-scores of the final ensemble of the calculated structures are given in Supplementary Table 1.[49] The model quality scores for ACP-mupA3b domain are poorer than those of ACP-mupA3a due to the longer unstructured helix I-II loop. The 2nd generation packing quality of only the secondary structure of ACP mupA3b domain improves to -1.380 with ~95% residues staying in the favoured region of the Ramachandran plot. The heavy atom RMSD is 1.78 and 2.21 Å for ACP-mupA3a and ACP-mupA3b respectively and each folded ACP domain showed single conformer solutions in the final 20 structures (total NOEs 2002, 12 per residue, 90 long range NOEs). The RMSDs are slightly higher than those reported for single ACP domains that are typically calculated with ~2000 NOE restraints (20/residue), approximately one third of which are long range NOE restraints.

Compilation of ACPs with HCS motifs. PHI-BLAST was accessed at the NCBI website with pattern DSxxxxxW and a cut-off probability of 0.0001 for each branching-module ACP of the bacillaene, kalimanticin, mupirocin, pederin and thiomarinol clusters, plus one representative from the six near-identical branching-module ACPs in the jamaicamide and curacin clusters (jamE1b).The resulting lists were merged, then filtered, discarding repeat sequences and short accessions incomplete for the multifunctional protein. Proteins were then scanned for domains using the PKS/NRPS Analysis server[50], and later confirmed with SMART[51]. The cluster was searched for HCS cassette functions using blastx against the mupirocin cluster HCS cassette proteins, with visual inspection using ACT. In the case of missing functionality the search was widened to other available contigs in the genome. ECH proteins & domains are distinguished as 1 or 2 (subscript), dependent on homology to MupJ and K respectively consistent with published results[52]. ECH domains with no subscript are dissimilar to both.

Hidden Markov Models. The hidden Markov models were built using HMMER3 (http://hmmer.janelia.org/). From the 15 clusters described previously, 38 and 178 sequences were respectively used to build the HMMs for ACPs identified as β-branching and non-β-branching. The performance of these models was determined with a test set of ACPs from other clusters which we had identified as likely to include a beta-branching ACP. The model for standard ACPs was further used to fetch the ACP sequences from the RefSeq microbial and Uniprot TrEMBL databases. Extracted sequences were further processed to filter out all sequences shorter than 60 amino acids, duplicates and sequences without serine necessary for phosphopantetheinylation, ensuring that all sequences were true ACPs. The filtered set resulted in 16490 sequences which were then further extended by 7 residues on both the ends to ensure they would cover the full length of the models. The final set of 16490 extended sequences were then scored using both the HMM models and plotted (Fig. 2d). Low-scoring sequences for which no score was returned by HMMER were conferred a score of 10 so they could be plotted. ACPs in the relevant section of the graph

were assessed for likely branching state. The sequence context of novel ACPs was evaluated to determine whether the ACP was likely to be associated with β-branching, by examining the PKS module for compatibility with branching, and for presence of an HCS cassette. Where available, published analysis of a pathway was studied.

Homology Modelling of MupH. The homology modelling of MupH (NCBI accession no. AAM12922) was carried out using Modeller version 9v8[53]. Ten structural homologues of MupH were selected with sequence identity ranging from 27% to 32% using a BLAST search of the PDB. An automated alignment of the homologous sequences using ClustalW was subsequently manually refined. The secondary structures of the templates, determined by the DSSP[54,55], guided manual refinement of the final alignment using Seaview[56]. The HMG-CoA synthase structure from *Enterococcus faecalis* (PDB ID 1X9E) having 87% query coverage with 32% sequence identity was selected as the template. Modeller produced 5 structures which were further tested for stereochemical quality using PROCHECK[48] program (http://nihserver.mbi.ucla.edu/SAVES/). The model with the best PROCHECK score was selected for further analysis.

Modelling the polyketide intermediate in the MupH active site

To dock the polyketide intermediate in the MupH active site the modelled MupH structure was superimposed on the 1X9E crystal structure and the X-ray coordinates for the phosphopantetheine moiety from the bound ligand in 1X9E were copied. The rest of the polyketide intermediate was built manually inside the MupH active site using PyMol (www.pymol.org). To remove the steric clashes energy minimisation was carried out using Chimera. The AMBER99SB[57] force field and Antechamber program were used to assign the force-field parameters to the protein and ligand respectively.

The mupA3a/mupA3b and MupH interaction

The docking of each of ACP mupA3a and ACP mupA3b with the MupH ligand complex was carried out using the HADDOCK webserver[29]. One restraint was applied to keep a distance of 2.0 Å between the phosphorous of phosphopantetheine bound in the active site of MupH and the OG of the serine (S38/142) residue in the ACP mupA3a/mupA3b, and another restraint was applied to keep a distance of 9.13 Å between the sulphur of the thio-ester linkage of posphopantetheine-polyketide substrate and the Cα of the catalytic cysteine (Cys 115) of MupH to anchor the ligand within the active. The residues at the interface of the ACP and MupH docked complex (Supplementary Table 3) were determined using a PyMol script (http://www.pymolwiki.org/index.php/InterfaceResidues). The interacting pairs of residues between ACP mupA3a/b and MupH were determined using the CONTAC module of WHATIF[58] (Supplementary Table 4).

Molecular dynamics simulation

To predict the effect of the W>L mutation on the structural dynamics of ACP mupA3a we carried out molecular dynamics simulations of the wild type and mutant ACP in explicit water. The mutation from W to L at the 44th position in the ACP mupA3ab NMR structure was made using PyMol. The two constructs each underwent 20 independent simulations for 10 ns each. The GROMACS 4[59] molecular dynamics engine with AMBER99SB-ILDN[60] force field and TIP3P water model were used to carry out the simulation with a time step of 2fs. Each simulation had a water box extending 5 Å beyond the surface of the protein and the systems were each neutralized with 2 Na$^+$ counter ions, minimized for 1000 steps of conjugate gradient and steepest descent energy minimisation and were equilibrated for 100 ps. The final PME simulation run was performed for 10 ns at 300K and 1 atmosphere with other parameters set as the GROMACS4 defaults.

Accessing NMR data

Coordinates for the final 20 water-refined models of ACP-mupA3ab have been deposited in the RCSB PDB with access code 2L22. NMR chemical shift assignments have been deposited in the BioMagResBank with code 17111.

References.

1.   Staunton, J. & Weissman, K.J. Polyketide biosynthesis: a millennium review. *Nat. Prod. Rep.* **18**, 380-416 (2001).
2.   Butler, M.S. Natural products to drugs: natural product-derived compounds in clinical trials. *Nat. Prod. Rep.* **25**, 475-516 (2008).
3.   Piel, J. Biosynthesis of polyketides by *trans*-AT polyketide synthases. *Nat. Prod. Rep.* **27**, 996-1047 (2010).
4.   Calderone, C.T. Isoprenoid-like alkylations in polyketide biosynthesis. *Nat. Prod. Rep.* **25**, 845-853 (2008).
5.   Fuller, A.T*., et al.* Pseudomonic acid: an antibiotic produced by *Pseudomonas fluorescens*. *Nature* **234**, 416-417 (1971).
6.   Thomas, C.M., Hothersall, J., Willis, C.L. & Simpson, T.J. Resistance to and synthesis of the antibiotic mupirocin. *Nat. Rev. Microbiol.* **8**, 281-289 (2010).
7.   Fukuda, D*., et al.* A natural plasmid uniquely encodes two biosynthetic pathways creating a potent anti-MRSA antibiotic. *PLoS One* **6**, 1-9 (2011).
8.   Rahman, A.S., Hothersall, J., Crosby, J., Simpson, T.J. & Thomas, C.M. Tandemly duplicated acyl carrier proteins, which increase polyketide antibiotic production, can apparently function either in parallel or in series. *J. Biol. Chem.* **280**, 6399-6408 (2005).
9.   Gu, L.C*., et al.* Tandem acyl carrier proteins in the curacin biosynthetic pathway promote consecutive multienzyme reactions with a synergistic effect. *Angew. Chem. Int. Ed.* **50**, 2795-2798 (2011).
10.   Busche, A*., et al.* Characterization of molecular interactions between ACP and halogenase domains in the curacin A polyketide synthase. *ACS Chem. Biol.* **7**, 378-386 (2012).
11.   El-Sayed, A.K*., et al.* Characterization of the mupirocin biosynthesis gene cluster from *Pseudomonas fluorescens* NCIMB 10586. *Chem. Biol.* **10**, 419-430 (2003).
12.   Chang, Z*., et al.* 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 (2004).
13.   Edwards, D.J*., et al.* Structure and biosynthesis of the jamaicamides, new mixed polyketide-peptide neurotoxins from the marine cyanobacterium *Lyngbya majuscula*. *Chem. Biol.* **11**, 817-833 (2004).
14.   Piel, J., Wen, G.P., Platzer, M. & Hui, D.Q. Unprecedented diversity of catalytic domains in the first four modules of the putative pederin polyketide synthase. *ChemBioChem* **5**, 93-98 (2004).
15.   Calderone, C.T., Kowtoniuk, W.E., Kelleher, N.L., Walsh, C.T. & Dorrestein, P.C. Convergence of isoprene and polyketide biosynthetic machinery: Isoprenyl-S-carrier proteins in the pksX pathway of *Bacillus subtilis*. *Proc. Natl. Acad. Sci. USA* **103**, 8977-8982 (2006).
16.   Chen, X.H*., et al.* Structural and functional characterization of three polyketide synthase gene clusters in Bacillus amyloliquefaciens FZB 42. *J. Bacteriol.* **188**, 4024-4036 (2006).
17.   Mattheus, W*., et al.* Isolation and purification of a new kalimantacin/batumin-related polyketide antibiotic and elucidation of Its biosynthesis gene cluster. *Chem. Biol.* **17**, 149-159 (2010).
18.   Schneider, T.D. & Stephens, R.M. Sequence logos - A new way to display consensus sequences. *Nucleic Acids Res.* **18**, 6097-6100 (1990).
19.   Irschik, H*., et al.* Analysis of the sorangicin gene cluster reinforces the utility of a combined phylogenetic/retrobiosynthetic analysis for deciphering natural product assembly by trans-AT PKS. *ChemBioChem* **11**, 1840-1849 (2010).
20.   Nakano, M.M., Corbell, N., Besson, J. & Zuber, P. Isolation and characterization of sfp: a gene that functions in the production of the lipopeptide biosurfactant, surfactin, in *Bacillus*

subtilis. *Mol. Gen. Genet.* **232**, 313-321 (1992).

21. Shields, J.A.*, et al.* Phosphopantetheinylation and specificity of acyl carrier proteins in the mupirocin biosynthetic cluster. *ChemBioChem* **11**, 248-255 (2010).

22. Evans, S.E.*, et al.* An ACP structural switch: conformational differences between the apo and holo forms of the actinorhodin polyketide synthase acyl carrier protein. *ChemBioChem* **9**, 2424-2432 (2008).

23. Leibundgut, M., Jenni, S., Frick, C. & Ban, N. Structural basis for substrate delivery by acyl carrier protein in the yeast fatty acid synthase. *Science* **316**, 288-290 (2007).

24. Alekseyev, V.Y., Liu, C.W., Cane, D.E., Puglisi, J.D. & Khosla, C. Solution structure and proposed domain domain recognition interface of an acyl carrier protein domain from a modular polyketide synthase. *Protein Sci.* **16**, 2093-2107. (2007).

25. Wattana-Amorn, P.*, et al.* Solution structure of an acyl carrier protein domain from a fungal Type I polyketide synthase. *Biochemistry* **49**, 2186-2193 (2010).

26. Crump, M.P.*, et al.* Solution structure of the actinorhodin polyketide synthase acyl carrier protein from *Streptomyces coelicolor* A3(2). *Biochemistry* **36**, 6000-6008 (1997).

27. Evans, S.E.*, et al.* Probing the interactions of early polyketide intermediates with the actinorhodin ACP from *S. coelicolor* A3(2). *J. Mol. Biol.* **389**, 511-528. (2009).

28. El-Sayed, A.K., Hothersall, J. & Thomas, C.M. Quorum-sensing-dependent regulation of biosynthesis of the polyketide antibiotic mupirocin in Pseudomonas fluorescens NCIMB 10586. *Microbiology* **147**, 2127-2139 (2001).

29. De Vries, S.J., van Dijk, M. & Bonvin, A.M.J.J. The HADDOCK web server for data-driven biomolecular docking. *Nat. Protoc.* **5**, 883-897 (2010).

30. Mihalek, I., Res, I. & Lichtarge, O. A family of evolution-entropy hybrid methods for ranking protein residues by importance. *J. Mol. Biol.* **336**, 1265-1282 (2004).

31. Kufareva, I., Budagyan, L., Raush, E., Totrov, M. & Abagyan, R. PIER: Protein interface recognition for structural proteomics. *Proteins: Struct. Funct. Bioinform.* **67**, 400-417 (2007).

32. Valley, C.C.*, et al.* The Methionine-aromatic Motif Plays a Unique Role in Stabilizing Protein Structure. *J. Biol. Chem.* **287**, 34979-34991 (2012).

33. Calderone, C.T., Iwig, D.F., Dorrestein, P.C., Kelleher, N.L. & Walsh, C.T. Incorporation of nonmethyl branches by isoprenoid-like logic: Multiple beta-alkylation events in the biosynthesis of myxovirescin A1. *Chem. Biol.* **14**, 835-846 (2007).

34. Khosla, C., Ebertkhosla, S. & Hopwood, D.A. Targeted gene replacements in a streptomyces polyketide synthase gene-cluster - Role for the acyl carrier protein. *Mol. Microbiol.* **6**, 3237-3249 (1992).

35. Decker, H., Summers, R.G. & Hutchinson, C.R. Overproduction of the acyl carrier protein component of a Type II polyketide synthase stimulates production of tetracenomycin biosynthetic intermediates in *Streptomyces Glaucescens*. *J. Antibiot.* **47**, 54-63 (1994).

36. Beltran-Alvarez, P., Cox, R.J., Crosby, J. & Simpson, T.J. Dissecting the component reactions catalyzed by the actinorhodin minimal polyketide synthase. *Biochemistry* **46**, 14672-14681 (2007).

37. Daley, D.J. Queuing output processes. *Advances in Applied Probability* **8**, 395-415 (1976).

38. Shiozawa, H.*, et al.* Thiomarinol, a new hybrid antimicrobial antibiotic produced by a marine bacterium fermentation, isolation, structure and antimicrobial activity. *J. Antibiot.* **46**, 1834-1842 (1993).

39. Bertani, G. Studies on lysogenesis .1. The mode of phage liberation by lysogenic *Escherchia coli*. *J. Bacteriol.* **62**, 293-300 (1951).

40. Sambrook, J., Fritsch, E.F. & Maniatis, T. *Molecular Cloning: A laboratory manual. Second edition*, (Cold Spring Harbour Press, New York, 1989).

41. Lepre, C.A. & Moore, J.M. Microdrop screening: a rapid method to optimize solvent conditions for NMR spectroscopy of proteins. *J. Biomol. NMR.* **12**, 493-499 (1998).

42. Vranken, W.F.*, et al.* The CCPN data model for NMR spectroscopy: Development of a

software pipeline. *Proteins-Structure Function and Bioinformatics* **59**, 687-696 (2005).

43. Ottiger, M., Delaglio, F. & Bax, A. Measurement of J and dipolar couplings from simplified two-dimensional NMR spectra. *J. Magn. Reson.* **131**, 373-378 (1998).

44. Dosset, P., Hus, J.C., Marion, D. & Blackledge, M. A novel interactive tool for rigid-body modeling of multi-domain macromolecules using residual dipolar couplings. *J. Biomol. NMR* **20**, 223-231 (2001).

45. Cornilescu, G., Delaglio, F. & Bax, A. Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *J. Biomol. NMR* **13**, 289-302 (1999).

46. Habeck, M., Rieping, W., Linge, J.P. & Nilges, M. NOE assignment with ARIA 2.0: the nuts and bolts. *Methods Mol. Biol.* **278**, 379-402 (2004).

47. Vriend, G. WHAT IF: A molecular modeling and drug design program. *J. Mol. Graph.* **8**, 52-56 (1990).

48. Laskowski, R.A., Macarthur, M.W., Moss, D.S. & Thornton, J.M. Procheck - a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* **26**, 283-291 (1993).

49. Spronk, C.A.E.M., Nabuurs, S.B., Krieger, E., Vriend, G. & Vuister, G.W. Validation of protein structures derived by NMR spectroscopy. *Prog. Nucl. Magn. Reson. Spectrosc.* **45**, 315-337 (2004).

50. Bachmann, B.O. & Ravel, J. Methods for in silico prediction of microbial polyketide and non-ribosomal peptide biosynthetic pathways from DNA sequence data. in *Complex Enzymes in Microbial Natural Product Biosynthesis, Part A: Overview Articles and Peptides*, Vol. 458 (ed. Hopwood, D.A.) 181-217 (Elsevier Inc., 2009).

51. Letunic, I., Doerks, T. & Bork, P. SMART 7: recent updates to the protein domain annotation resource. *Nucleic Acids Res.* **40**, D302-D305 (2012).

52. Gu, L.*, et al.* Metabolic coupling of dehydration and decarboxylation in the curacin A pathway: Functional identification of a mechanistically diverse enzyme pair. *J. Am. Chem. Soc.* **128**, 9014-9015 (2006).

53. Eswar, N.*, et al.* Comparative protein structure modeling using Modeller. in *Current protocols in bioinformatics* Vol. Unit5.6 (ed. Baxevanis A. D. et al.) (Wiley, 2006).

54. Joosten, R.P.*, et al.* A series of PDB related databases for everyday needs. *Nucleic Acids Res.* **39**, D411-D419 (2011).

55. Kabsch, W. & Sander, C. Dictionary of protein secondary structure - Pattern-recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577-2637 (1983).

56. Galtier, N., Gouy, M. & Gautier, C. SEAVIEW and PHYLO_WIN: Two graphic tools for sequence alignment and molecular phylogeny. *Comput. Appl. Biosci.* **12**, 543-548 (1996).

57. Hornak, V.*, et al.* Comparison of multiple amber force fields and development of improved protein backbone parameters. *Proteins: Struct. Funct. Bioinform.* **65**, 712-725 (2006).

58. Vriend, G. WHAT IF: a molecular modeling and drug design program. *J. Mol. Graph.* **8**, 52-56, 29 (1990).

59. Hess, B., Kutzner, C., van der Spoel, D. & Lindahl, E. GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.* **4**, 435-447 (2008).

60. Lindorff-Larsen, K.*, et al.* Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins: Struct. Funct. Bioinform.* **78**, 1950-1958 (2010).

Competing financial interests
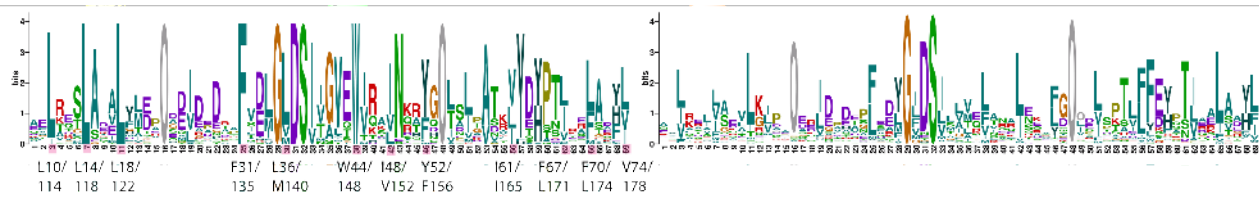
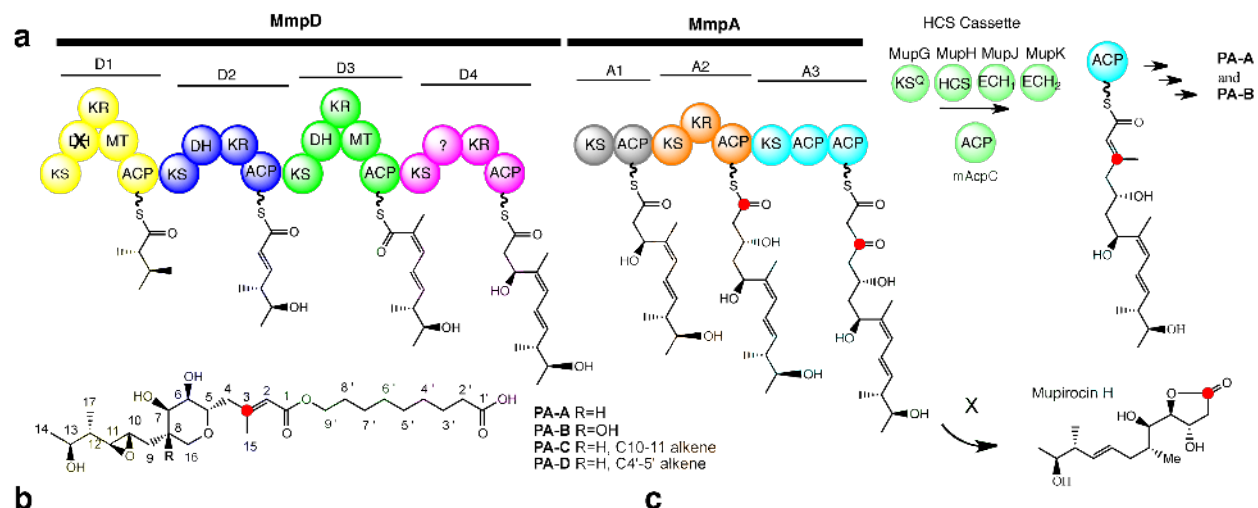The authors declare no competing financial interests.
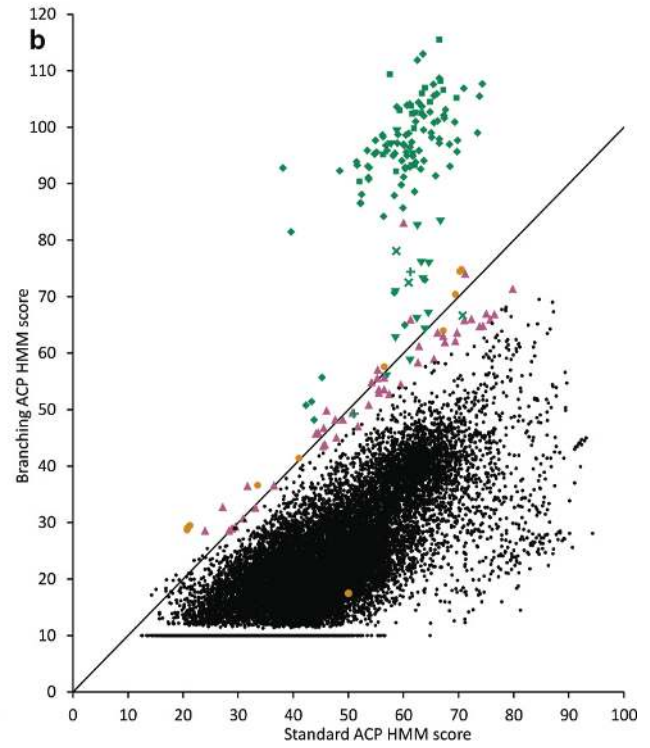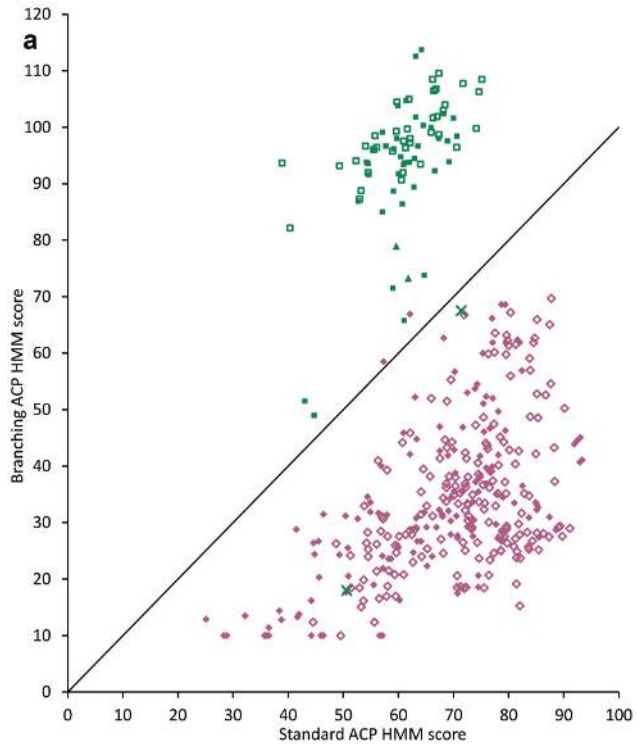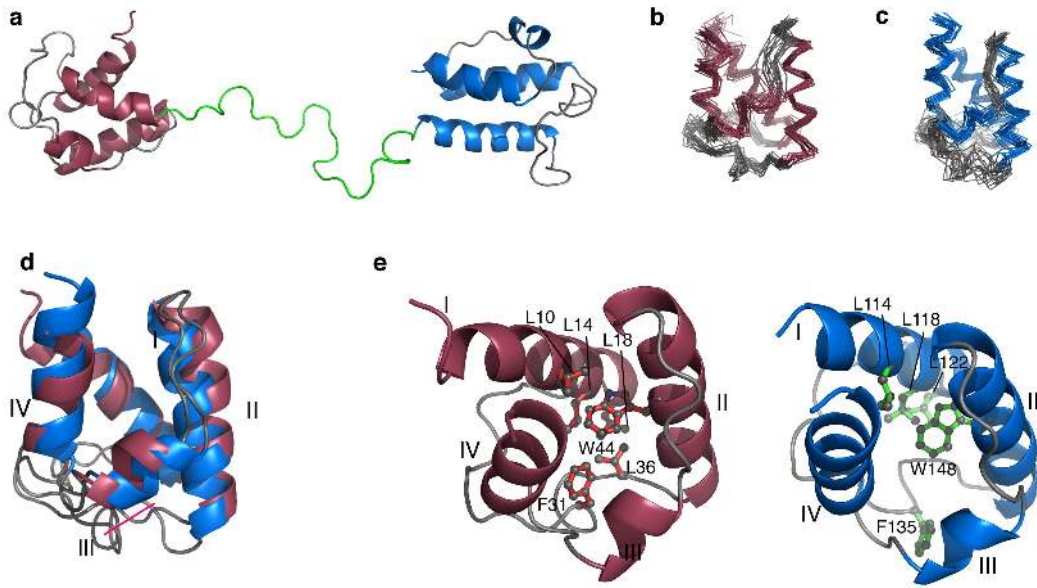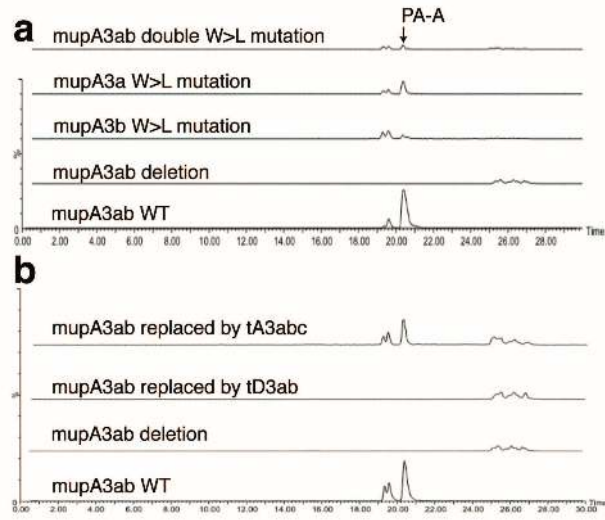
Figure 1 Crump
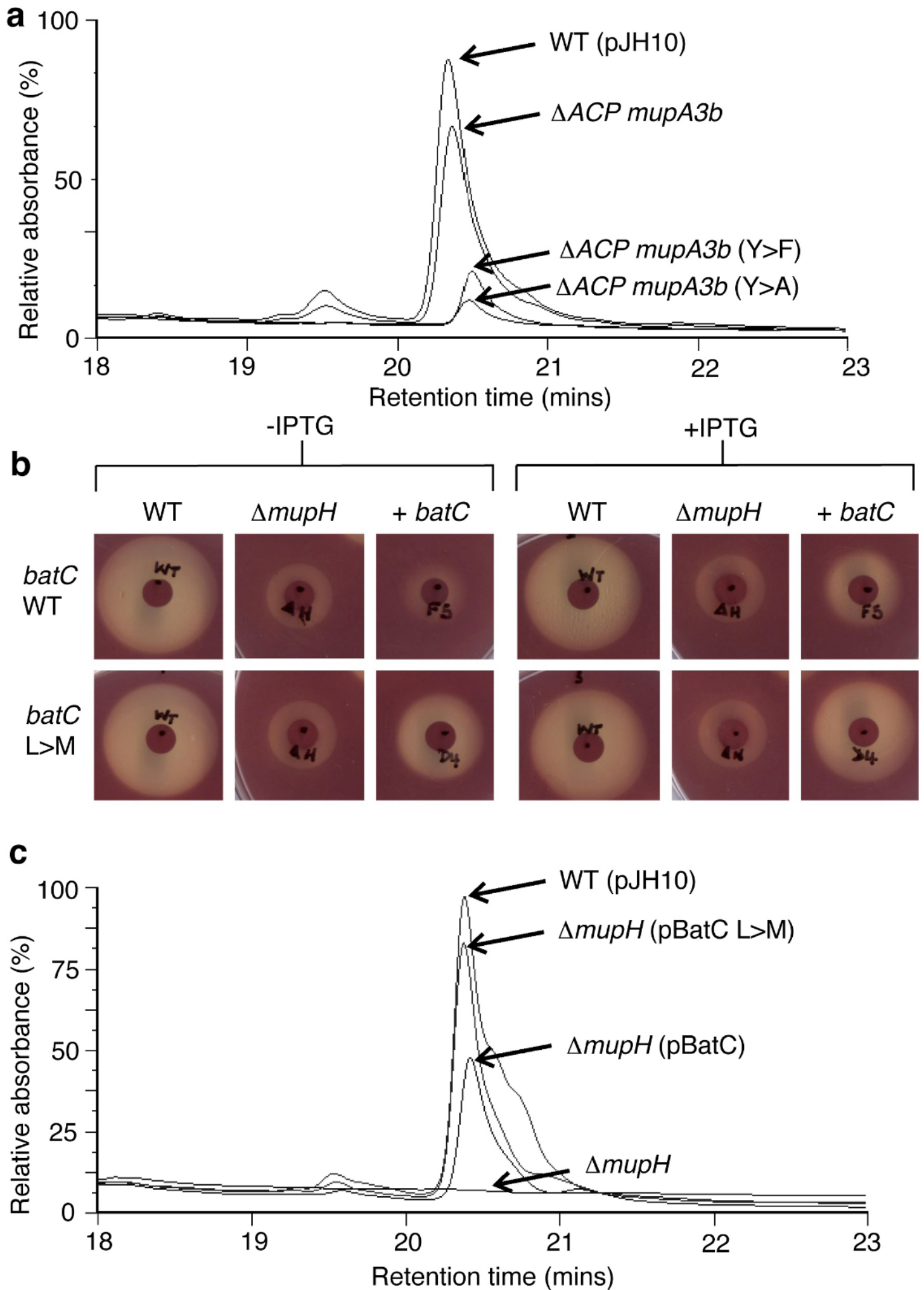
Figure 2. Crump

Figure 3 Crump

Figure 4 Crump

Figure 5 Crump

Figure 6 Crump