HARNESSING SYNTHETIC BIOLOGY FOR THE BIOPROSPECTING AND ENGINEERING OF AROMATIC POLYKETIDE SYNTHASES

A thesis submitted to The University of Manchester for the degree of Doctor of Philosophy in the Faculty of Science and Engineering

(2017)

MATTHEW J. CUMMINGS

SCHOOL OF CHEMISTRY
List of contents

List of contents .............................................................................................................................. 3
List of figures ................................................................................................................................... 8
List of supplementary figures ...................................................................................................... 10
List of tables .................................................................................................................................. 11
List of supplementary tables ....................................................................................................... 11
List of boxes ................................................................................................................................ 11
List of abbreviations .................................................................................................................... 12
Abstract ....................................................................................................................................... 14
Declaration .................................................................................................................................. 15
Copyright statement .................................................................................................................... 16
Acknowledgements ..................................................................................................................... 17
Thesis preface ............................................................................................................................. 18
1 Introduction.......................................................................................................................... 19
   1.1 Specialised metabolites .............................................................................................. 19
       1.1.1 Antimicrobial discovery ....................................................................................... 19
   1.2 Biosynthetic genes cluster .......................................................................................... 24
       1.2.1 Operons and evolution ........................................................................................ 24
       1.2.2 BGC prediction .................................................................................................... 28
   1.3 Synthetic biology for specialised metabolites ............................................................. 32
       1.3.1 Rebuilding biosynthetic gene clusters ................................................................. 32
       1.3.2 Semantics in synthetic biology ............................................................................ 34
   1.4 Polyketides .................................................................................................................. 35
       1.4.1 Type I polyketide synthases ................................................................................ 37
       1.4.2 Type III polyketide synthases .............................................................................. 38
       1.4.3 Type II polyketide synthases ............................................................................... 39
   1.5 Aims of the thesis ........................................................................................................ 49
   1.6 References .................................................................................................................. 52
2 Steps towards the synthetic biology of polyketide biosynthesis ......................................... 66
   2.1 Preface ........................................................................................................................ 67
   2.2 Abstract ....................................................................................................................... 68
   2.3 Steps towards the synthetic biology of polyketide biosynthesis ..................................... 69
       2.3.1 Polyketides: magnificently modular ..................................................................... 69
       2.3.2 Modularity of scaffold biosynthesis .................................................................... 72
       2.3.3 Modularity of tailoring reactions ......................................................................... 75
   2.4 Synthetic biology: the future of combinatorial biosynthesis ......................................... 76
   2.5 Acknowledgements ....................................................................................................... 78
5.4 Results ...................................................................................................................... 125
5.4.1 Module 1: Identification of candidate ketosynthase: chain length factor dimer pairs for heterologous expression in E. coli ................................................................. 125
5.4.2 Evaluating solubility and dimer formation of the AntD and AntE in E. coli .... 125
5.4.3 Testing functionality of the PKS (Module 2) ...................................................... 127
5.4.4 Module 3: Exploring end-compound production of the anthraquinone biosynthetic gene cluster ....................................................................................................................... 128
5.4.5 Evaluation of a plug-and-play scaffold .............................................................. 131
5.5 Discussion ................................................................................................................. 134
5.6 Acknowledgement ................................................................................................. 137
5.7 References ................................................................................................................ 138
5.8 Supplementary results .............................................................................................. 141
5.8.1 Bioinformatics analysis of AntE; a partner for AntD ........................................ 141
5.8.2 Elucidating major metabolites produced by the anthraquinone biosynthetic gene cluster ....................................................................................................................... 141
5.8.3 SEK4 and SEK4b .............................................................................................. 142
5.8.4 Mutactin, SEK34 and their dehydrated counterparts ........................................ 143
5.8.5 Identification of trihydroxyanthrone (12) end product ........................................ 144
5.8.6 Dianthrone elucidation ...................................................................................... 144
5.8.7 Modified anthraquinone identification ............................................................... 145
5.9 Supplementary Methods ........................................................................................... 145
5.9.1 Bacterial strains and culture conditions ............................................................ 145
5.9.2 DNA isolation, Plasmid construction and Refactoring ....................................... 146
5.9.3 Protein purification and peptide identification ................................................... 148
5.9.4 Metabolite extraction and analysis .................................................................... 149
5.9.5 Octaketide shunt metabolite identification using HPLC-ESI-MS ...................... 149
5.9.6 Shunt metabolite identification via HPLC-UV-Vis-ESI-MS ................................ 150
5.9.7 Actinorhodin KR and CYC complementation and pACYCAnΔ86 analysis via UV-Vis ................................................................. 150
5.9.8 HPLC High resolution mass spectrum analysis ................................................. 150
5.9.9 Mass spectrum data analysis ............................................................................ 150
5.9.10 Code availability ............................................................................................ 151
5.9.11 Characterisation of AQ256 ............................................................................ 151
5.9.12 Phylogenetic analysis of amino acid sequences ............................................... 151
5.10 Supplementary Figures ............................................................................................. 152
5.11 Supplementary Tables .............................................................................................. 187
5.12 Supplementary References ....................................................................................... 192
6 Engineering Escherichia coli soluble type II polyketide synthases for biosynthesis of decaketides and above in vivo. ................................................................................................. 193
6.1 Preface ...................................................................................................................... 193
6.2 Abstract .................................................................................................................... 194
List of figures

Figure 1.1: Specialised metabolites with antimicrobial activity ................................................................. 21
Figure 1.2: First, Second and third generation tetracyclines .................................................................... 22
Figure 1.3: Schematic of a biosynthetic gene cluster and its corresponding biosynthetic pathway .......... 24
Figure 1.4: Two proposed models for the regulation of protein synthesis by Jacob and Monod ......... 25
Figure 1.5: Modular chemical diversity of aminocoumarin specialised metabolites ............................... 28
Figure 1.6: Specialised metabolites identified using non-rule based bioinformatics tools .................... 30
Figure 1.7: A guide to refactoring ............................................................................................................. 34
Figure 1.8: Polyketide synthases ................................................................................................................ 36
Figure 1.9: Polyketide biosynthesis ........................................................................................................... 37
Figure 1.10: Selection of type III polyketide core structures ..................................................................... 39
Figure 1.11: Genetic organisation of type II mPKS components ............................................................. 40
Figure 1.12: Conserved active site glutamine residues within CLF and KSQ ........................................... 41
Figure 1.13: Functionalisation and acylation of acyl carrier proteins ..................................................... 42
Figure 1.14: Early actinorhodin shunt metabolites .................................................................................. 44
Figure 1.15: Aromatic polyketide chemotypes, directed through cyclisation and aromatisation ........... 45
Figure 1.16: Biosynthesis of aromatic polyketide in "E. coli" by PKS4 ...................................................... 48
Figure 1.17: A specialised metabolite discovery pipeline .......................................................................... 51
Figure 2.1: Pictorial illustration of 6-dEB synthase, a modular type I PKS, and successful attempts at engineering this megasynthase ................................................................. 71
Figure 2.2: Schematic of the four steps of polyketide biosynthesis encoded by a prototypical polyketide biosynthetic gene cluster ......................................................................................... 73
Figure 3.1: Schematic overview of the MIBiG standard ......................................................................... 90
Figure 3.2: An example MIBiG entry ........................................................................................................ 92
Figure 3.3: The position of MIBiG in specialized metabolite research ................................................. 93
Figure 4.1: Schematic overview of BGC prioritisation parameters used in OOPS ................................ 103
Figure 4.2: OOPS graphical user interface .............................................................................................. 106
Figure 4.3: The relationship between gene number and CDS length ...................................................... 108
Figure 5.1: Biosynthesis of archetypal aromatic polyketides .................................................................... 122
Figure 5.2: Schematic representation of modularity of aromatic polyketide biosynthesis .................. 124
Figure 5.3: Identification and expression of the P. luminescens KS AntD and CLF AntE in "E. coli" .... 126
Figure 5.4: Expression of AntDEFBG in "E. coli" ......................................................................................... 128
Figure 5.5: Anthraquinone identification and characterization ................................................................. 130
Figure 5.6: Complementation of the Anthraquinone BGC with Actinorhodin components .................. 133
Figure 6.1: Chemistry of nascent polyketide chains ................................................................................ 196
Figure 6.2: Chemotype/ chromophores accessible by type II polyketide synthases and cognate tailoring enzymes ......................................................................................................................... 196
Figure 6.3: Type II PKS starter unit biosynthetic strategies .................................................................. 200
Figure 6.4: combinatorial biosynthesis using alternative initiation modules ........................................ 201
Figure 6.5: KS/CLF dimeric cavity ............................................................................................................ 202
Figure 6.6: Models of act and ant KS/ CLF ............................................................................................ 209
Figure 6.7: Purification strategy for AntDE heterodimer complex ........................................................ 210
Figure 6.8: Designing flexible linkers using the actinorhodin KS/ CLF crystal structure and AntDE homology model ..................................................................................................................... 212
Figure 6.9: Evaluation of His6-AntED fusion protein in "E. coli" BL21(DE3) ............................................ 213

8
Figure 6.10: Purification of His6\(^\text{\textregistered}\)AntED from *E. coli* soluble cell lysate .............................................................. 213
Figure 6.11: Δplu4190plu4191 AQ BGC complementation with episomally encoded his\(^{\text{\textregistered}}\)-antED-fus and his\(^{\text{\textregistered}}\)-antE and antD........................................................................................................... 215
Figure 7.1: Specialised metabolites from underexplored Organisms........................................................................................................... 232
Figure 7.2: Xentrivalpeptides, pericidins and streptochlorin – potential combinatorial therapy agents .... 233
Figure 7.3: Specialised metabolites identified by advanced culture techniques .............................................. 235
Figure 7.4: Non-Actinobacterial derived aromatic polyketides ............................................................................. 236
Figure 7.5: Translational dynamics of characterised KS sequences ................................................................. 240
Figure 7.6 Translational dynamics of characterised CLF sequences ............................................................ 241
Figure 7.7: Frequency of biosynthetic gene clusters identified by antiSMASH 3.0........................................... 245
Figure 7.8: Taxonomic distribution of type II PKS clusters defined by AntiSMASH 3.0 from a dataset of 2552 references genomes. ................................................................. 247
Figure 7.9: The *pab* type II PKS BGCs from *Paenibacilli*................................................................................. 247
Figure 7.10: Phylogenetic distribution of all curated non-Actinobacterial type II PKS clusters........ 253
Figure 7.11: KS/CLF phylogeny .................................................................................................................. 254
Figure 7.12: Phylogenetic reconstruction of 14 of 19 *Delftia* spp. analysed by antiSMASH. ................. 257
Figure 7.13: Genomic localisation of type II PKS BGC in *Delftia* spp.......................................................... 258
Figure 7.14: Evaluating solubility of DacA and DacB in *E. coli* BL21(DE3).................................................. 260
Figure 7.15: Purification of DacB from *E. coli* BL21(DE3)........................................................................... 261
Figure 7.16 Genomic localisation of type II PKS BGC in *Streptococcus* spp........................................ 263
Figure 7.17: Evaluating the solubility of SspA and SspB in *E. coli* BL21(DE3).......................... 265
Figure 7.18: Co-purification of SspA and SspB from *E. coli* BL21(DE3)...................................................... 266
Figure 7.19: Multilocus phylogenetic tree of *P. luteoviolacea* strains .................................................. 267
Figure 7.20: Genomic localisation of type II PKS BGC in *Pseudoalteromonas luteoviolacea*......... 268
Figure 7.21: Evaluation of PluA and B solubility in *E. coli* BL21(DE3)...................................................... 270
Figure 7.22: Co-purification of PluA and PluB from *E. coli* .................................................................. 271
Figure 7.23: *B. endophyticus* DSM 13796 and KCTC 13922 type II PKS BGC schematic.............. 273
Figure 7.24: Phylogeny of *Bacillus endophyticus* spp............................................................... 273
Figure 7.25: Evaluation of BenAB solubility in *E. coli* BL21(DE3)...................................................... 276
Figure 7.26: Purification of BenB from *E. coli* BL21(DE3) soluble cell lysate................................. 276
Figure 7.27: Schematic of putative type II PKS BGC from *Desulfobacterales* spp........................... 279
Figure 7.28: Amino acid sequence alignment of concatenated CLF and FAS parts from *Desulfobacterium* sp. 4572_20............................................................................................................. 280
Figure 7.29: Evaluation of DauAB solubility in *E. coli* BL21(DE3)...................................................... 281
Figure 7.30: Purification of DauB from *E. coli* BL21(DE3) soluble cell lysate ........................................ 282
Figure 7.31: Type II PKS BGCs from *Ktedonobacter racemifer* DSM 44963........................................ 284
Figure 7.32: Evaluation of KraAB solubility in *E. coli* BL21(DE3)...................................................... 287
Figure 7.33: Assessment of KraB solubility in *E. coli* BL21(DE3) ...................................................... 288
Figure 7.34: Purification of KraB from *E. coli* BL21(DE3) cell soluble lysate ..................................... 288
Figure 7.35: Evaluation RemAB solubility in *E. coli* BL21(DE3).......................................................... 290
Figure 7.36: Substitution of arginine and lysine codons in low and high G+C content Actinobacterial ketosynthases ........................................................................................................ 292
Figure 7.37: U-, J- and S- shaped aromatic polyketides .............................................................................. 298
Figure 7.38: mPKS from two aurachin producers ..................................................................................... 299
Figure 8.1: A refreshed perspective pipeline for specialised metabolite discovery .................................. 333
List of supplementary figures

Supplementary Figure 5.1: Multiple sequence alignment of FabF, AntD, AntE, ActIORFI and ActIORFII .................................................. 153
Supplementary Figure 5.2: Homology model of the AntDE dimer ............................................................. 154
Supplementary Figure 5.3: Cyclisation pathways of octaketide shunt metabolites .................................................. 155
Supplementary Figure 5.4: SEK4 and SEK4b MS² spectra ........................................................................ 156
Supplementary Figure 5.5: plausible oxidative coupling of 1,3,8 trihydroxyanthrone and dianthrone fragmentation pattern ............................................................. 157
Supplementary Figure 5.6: Comparative EIC's for metabolites of interest ............................................................. 158
Supplementary Figure 5.7: Graphical interpretation of AQ256 two dimensional NMR spectra ............................................................. 160
Supplementary Figure 5.8: ¹H NMR spectrum for AQ256 ........................................................................ 161
Supplementary Figure 5.9: ¹³C NMR spectrum for AQ256 ........................................................................ 162
Supplementary Figure 5.10: HSQC NMR spectrum of AQ256 ........................................................................ 163
Supplementary Figure 5.11: COSY NMR spectrum of AQ-256 ........................................................................ 164
Supplementary Figure 5.12: HMBC NMR spectrum of AQ256 ........................................................................ 165
Supplementary Figure 5.13: Evaluation of chrysophanol, emodin, aloesaponarin II and AQ256 MS² spectra (ES⁻) ........................................................................ 166
Supplementary Figure 5.14: UV-Vis spectra for chrysophanol, emodin and AQ256 ............................................................. 167
Supplementary Figure 5.15: Typical full HPLC chromatogram of E. coli BL21(DE3) pACYCAnthraquinone cultures expressing AntA-I at 434 nm ........................................................................ 168
Supplementary Figure 5.16: Full mass spectrum and MS² for two putative dianthrones ............................................................. 169
Supplementary Figure 5.17: HPLC-UV-Vis-MS analysis of dianthrones 1 and 2 ........................................................................ 170
Supplementary Figure 5.18: Extracted ion chromatogram of ketoreductase complemented anthraquinone biosynthetic pathways ........................................................................ 171
Supplementary Figure 5.19: Extracted ion chromatogram (EIC) of ARO/CYC complemented anthraquinone biosynthetic pathways ........................................................................ 172
Supplementary Figure 5.20: ES⁻ extracted ion chromatograms of polyketide shunt metabolites ........................................................................ 173
Supplementary Figure 5.21: ES⁻ extracted ion chromatograms of polyketide shunt metabolites ........................................................................ 174
Supplementary Figure 5.22: HPLC-UV-Vis-ESI-MS analysis of E. coli BL21(DE3) expressing AntB-I, ΔAntA ........................................................................ 175
Supplementary Figure 5.23: HPLC-UV-Vis-MS analysis of E. coli BL21(DE3) expressing AntA-G, I, ΔAntH ........................................................................ 176
Supplementary Figure 5.24: HPLC-UV-Vis-MS analysis of E. coli BL21(DE3) expressing AntA, B, D-I, ΔAntC ........................................................................ 177
Supplementary Figure 5.25: HPLC-UV-Vis-MS analysis of E. coli BL21(DE3) expressing AntA-I and AQ256 standards ........................................................................ 178
Supplementary Figure 6.1: PrDOS prediction of His⁵AntED-Fus disorder ........................................................................ 179
Supplementary Figure 6.2: PONDR predictions of His⁵AntED-Fus disorder ........................................................................ 180

10
List of tables

Table 1.1: Cyclisation patterns for different aromatic polyketide chemotypes .................................................. 45
Table 6.1: Prediction of KS/CLF tunnel volume for varying chain lengths .......................................................... 204
Table 6.2: Buffers used during purification of KS/CLF pairs from Streptomyces spp. ........................................ 217
Table 7.1: Non-Actinobacterial organisms comprising one or more predicted type II PKS BGCs .................... 252
Table 7.2: Polyketide chain length predictions for KS/CLF pairs from underexplored phyla ........................... 255
Table 7.3: Pairwise similarities of the bend KS/CLF and archetypal U and S shaped polyketide synthases ....... 275

List of supplementary tables

Supplementary Table 5.1: Chain length factor gatekeeper residues for biosynthesis of different length nascent poly-β-ketide chains. ................................................................................................................... 187
Supplementary Table 5.2: Plasmids used in this study ....................................................................................... 188
Supplementary Table 5.3: AQ256 carbon and proton assignment and coupling constants ............................ 189
Supplementary Table 5.4: Theoretical masses for all shunt metabolites .......................................................... 190
Supplementary Table 5.5: Primers used for plasmid construction ................................................................. 191
Supplementary Table 6.1: PCR Primers ............................................................................................................... 225
Supplementary Table 7.1: Organisation of the pab BGC from Paenibacillus borealis ............................... 311
Supplementary Table 7.2: Organisation of the dau BGC ................................................................................. 312
Supplementary Table 7.3: Organisation of the dac BGC from Azospirillum brasilense ............................ 314
Supplementary Table 7.4: Organisation of the dac BGC from Acidithiobacillus ferrivorans ....................... 315
Supplementary Table 7.5: amino acid sequence similarity of all dau BGCs .................................................. 315
Supplementary Table 7.6: Organisation of the ssp BGC from Streptococcus sp. GMDXs .......................... 316
Supplementary Table 7.7: Organisation of the ssp homologous BGC from Lactobacillus oris PB013-T2-3 ... 317
Supplementary Table 7.8: Organisation of the plu BGCs from Pseudoalteromonas luteoviolacea strains 318
Supplementary Table 7.9: Organisation of the bend BGC from Bacillus endophyticus ............................... 319
Supplementary Table 7.10: Organisation of the dau BGC from Candidatus Desulfofervidus auxilli ................ 320
Supplementary Table 7.11: Organisation of the kra1 BGC from Ktedonobacter racemifer ......................... 321
Supplementary Table 7.12: Organisation of the kra2 BGC from Ktedonobacter racemifer ......................... 322
Supplementary Table 7.13: Organisation of the kra BGC from Ktedonobacter racemifer .......................... 323

List of boxes

Box 1: The Genomic Standards Consortium and its MIxS framework ............................................................ 89
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>6-dEB</td>
<td>6-Deoxyerythronolide B</td>
</tr>
<tr>
<td>6-MSA</td>
<td>6-methylsalicylic acid</td>
</tr>
<tr>
<td>AGOS</td>
<td>Artificial gene operon assembly system</td>
</tr>
<tr>
<td>ANME</td>
<td>Anaerobic oxidation of methane Archeal partner organism</td>
</tr>
<tr>
<td>antiSMASH</td>
<td>Antibiotics and secondary metabolite analysis shell</td>
</tr>
<tr>
<td>ARO/CYC</td>
<td>Aromatase/cyclase</td>
</tr>
<tr>
<td>ARTS</td>
<td>Antibiotic resistance target seeker</td>
</tr>
<tr>
<td>AT</td>
<td>Acetyltransferase</td>
</tr>
<tr>
<td>BGC</td>
<td>Biosynthetic gene cluster</td>
</tr>
<tr>
<td>BIQ</td>
<td>Benzoisochromanequinone</td>
</tr>
<tr>
<td>BLAST</td>
<td>Basic local alignment search tool</td>
</tr>
<tr>
<td>bp</td>
<td>Base pair</td>
</tr>
<tr>
<td>CDA</td>
<td>Calcium dependant antibiotic</td>
</tr>
<tr>
<td>CDD</td>
<td>Conserved domains database</td>
</tr>
<tr>
<td>CDS</td>
<td>Protein coding sequence</td>
</tr>
<tr>
<td>CLF</td>
<td>Chain length factor</td>
</tr>
<tr>
<td>CLUSEAN</td>
<td>Cluster sequence analyzer</td>
</tr>
<tr>
<td>CoA</td>
<td>Coenzyme A</td>
</tr>
<tr>
<td>COG</td>
<td>Cluster of orthologous groups</td>
</tr>
<tr>
<td>CON</td>
<td>Acyl carrier protein condensation domain</td>
</tr>
<tr>
<td>COSY</td>
<td>Homonuclear correlation spectroscopy</td>
</tr>
<tr>
<td>CV</td>
<td>Column volume</td>
</tr>
<tr>
<td>CYC</td>
<td>Cyclase</td>
</tr>
<tr>
<td>DDBJ</td>
<td>DNA databank of Japan</td>
</tr>
<tr>
<td>DH</td>
<td>Dehydratase</td>
</tr>
<tr>
<td>DMAC</td>
<td>3,8-dihydroxy-1-methylanthraquinone-2-carboxylic acid</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic acid</td>
</tr>
<tr>
<td>DNPA</td>
<td>4-dihydro-9-hydroxy-1-methyl-10-oxo-3H-naphto[2,3-c]pyran-3-acetic acid</td>
</tr>
<tr>
<td>DoBiscuit</td>
<td>Database of biosynthesis clusters curated and integrated</td>
</tr>
<tr>
<td>DS</td>
<td>Diversity score</td>
</tr>
<tr>
<td>DTT</td>
<td>Dithiothreitol</td>
</tr>
<tr>
<td>EBI-ENA</td>
<td>European nucleotide archive</td>
</tr>
<tr>
<td>EDTA</td>
<td>Ethylenediaminetetraacetic acid</td>
</tr>
<tr>
<td>EF</td>
<td>Elution fraction</td>
</tr>
<tr>
<td>ER</td>
<td>Enoyl reductase</td>
</tr>
<tr>
<td>ESi</td>
<td>Electrospray ionization</td>
</tr>
<tr>
<td>FAS</td>
<td>Fatty acid synthase</td>
</tr>
<tr>
<td>FT</td>
<td>Flow through</td>
</tr>
<tr>
<td>gDNA</td>
<td>Genomic DNA</td>
</tr>
<tr>
<td>GSC</td>
<td>Genomics standard consortium</td>
</tr>
<tr>
<td>HEPES</td>
<td>Hydroxyethyl piperazineethanesulfonic acid</td>
</tr>
<tr>
<td>HESI</td>
<td>Heated electrospray ionization</td>
</tr>
<tr>
<td>HI</td>
<td>Hybrid isoprenoid</td>
</tr>
<tr>
<td>HMBC</td>
<td>Heteronuclear multiple-bond correlation spectroscopy</td>
</tr>
<tr>
<td>HMM</td>
<td>Hidden Markov model</td>
</tr>
<tr>
<td>HPLC</td>
<td>High-performance liquid chromatography</td>
</tr>
<tr>
<td>HSQC</td>
<td>Heteronuclear single quantum coherence spectroscopy</td>
</tr>
<tr>
<td>IDT</td>
<td>Integrated DNA technologies</td>
</tr>
<tr>
<td>iGEM</td>
<td>International genetically engineered machine</td>
</tr>
<tr>
<td>IMAC</td>
<td>Immobilized metal ion affinity chromatography</td>
</tr>
<tr>
<td>IMG</td>
<td>Integrated Microbial Genome</td>
</tr>
<tr>
<td>INSDC</td>
<td>International nucleotide sequence database collection</td>
</tr>
<tr>
<td>IPTG</td>
<td>Isopropyl β-D-1-thiogalactopyranoside</td>
</tr>
<tr>
<td>iTOL</td>
<td>Interactive tree of life</td>
</tr>
<tr>
<td>IUP</td>
<td>Intrinsically disordered structure</td>
</tr>
<tr>
<td>Kb</td>
<td>Kilobase</td>
</tr>
<tr>
<td>KCB</td>
<td>Known cluster BLAST</td>
</tr>
<tr>
<td>kDa</td>
<td>Kilodalton</td>
</tr>
<tr>
<td>KR</td>
<td>Ketoreductase</td>
</tr>
<tr>
<td>KS</td>
<td>Ketosynthase</td>
</tr>
<tr>
<td>Term</td>
<td>Definition</td>
</tr>
<tr>
<td>--------</td>
<td>--------------------------------------------------------------------------</td>
</tr>
<tr>
<td>KS&lt;sup&gt;G&lt;/sup&gt;</td>
<td>Decarboxylating ketosynthase</td>
</tr>
<tr>
<td>LB</td>
<td>Lysogeny broth</td>
</tr>
<tr>
<td>LC</td>
<td>Liquid chromatography</td>
</tr>
<tr>
<td>MAFFT</td>
<td>Multiple alignment fast fourier transform</td>
</tr>
<tr>
<td>Mb</td>
<td>Megabase</td>
</tr>
<tr>
<td>MCAT</td>
<td>Malonyl-CoA:acyl carrier protein transacylase</td>
</tr>
<tr>
<td>MCS</td>
<td>Multiple cloning site</td>
</tr>
<tr>
<td>MDa</td>
<td>Megadalton</td>
</tr>
<tr>
<td>MIBiG</td>
<td>Minimum information about a biosynthetic gene cluster</td>
</tr>
<tr>
<td>MIGS</td>
<td>Minimum information about a genome sequence</td>
</tr>
<tr>
<td>MIMARKS</td>
<td>Minimum information about a marker gene sequence</td>
</tr>
<tr>
<td>Mlxs</td>
<td>Minimum information about any sequence</td>
</tr>
<tr>
<td>mPKS</td>
<td>Minimal polyketide synthase</td>
</tr>
<tr>
<td>mRNA</td>
<td>Messenger ribonucleic acid</td>
</tr>
<tr>
<td>MS</td>
<td>Mass spectrometry</td>
</tr>
<tr>
<td>MS&lt;sup&gt;2&lt;/sup&gt;</td>
<td>Tandem mass spectrometry</td>
</tr>
<tr>
<td>MUSCLE</td>
<td>Multiple sequence comparison by log-expectation</td>
</tr>
<tr>
<td>NapDoS</td>
<td>Natural product domain seeker</td>
</tr>
<tr>
<td>NCBI</td>
<td>National center for biotechnology information</td>
</tr>
<tr>
<td>NMR</td>
<td>Nuclear magnetic resonance</td>
</tr>
<tr>
<td>NPC</td>
<td>No plasmid control</td>
</tr>
<tr>
<td>NRPS</td>
<td>Non-ribosomal peptide synthetase</td>
</tr>
<tr>
<td>OKS</td>
<td>Octaketide synthase</td>
</tr>
<tr>
<td>ORF</td>
<td>Open reading frame</td>
</tr>
<tr>
<td>PC</td>
<td>Positive control</td>
</tr>
<tr>
<td>PCR</td>
<td>Polymerase chain reaction</td>
</tr>
<tr>
<td>PHASTER</td>
<td>Phage search tool enhanced release</td>
</tr>
<tr>
<td>PKS</td>
<td>Polyketide synthase</td>
</tr>
<tr>
<td>PPTase</td>
<td>Phosphopantetheine arm</td>
</tr>
<tr>
<td>PPTase</td>
<td>Phosphopantetheinyl transferase</td>
</tr>
<tr>
<td>PRISM</td>
<td>Prediction informatics for secondary metabolomes</td>
</tr>
<tr>
<td>PS</td>
<td>Preferred similarity</td>
</tr>
<tr>
<td>RBS</td>
<td>Ribosome-binding site</td>
</tr>
<tr>
<td>RE</td>
<td>Restriction endonuclease</td>
</tr>
<tr>
<td>RiPP</td>
<td>Ribosomally synthesised and posttranslationally modified peptide</td>
</tr>
<tr>
<td>SAM</td>
<td>S-adenosyl methionine</td>
</tr>
<tr>
<td>SBOL</td>
<td>Synthetic biology open language</td>
</tr>
<tr>
<td>SBSPKS</td>
<td>Structure based sequence analysis of polyketide synthases</td>
</tr>
<tr>
<td>SDS-PAGE</td>
<td>Sodium dodecyl sulfate polyacrylamide gel electrophoresis</td>
</tr>
<tr>
<td>SEC</td>
<td>Size exclusion chromatography</td>
</tr>
<tr>
<td>SMBP</td>
<td>Secondary Metabolite Bioinformatics Portal</td>
</tr>
<tr>
<td>smCOG</td>
<td>Specialised metabolite cluster of orthologous groups</td>
</tr>
<tr>
<td>SMURF</td>
<td>Secondary metabolite unknown regions finder</td>
</tr>
<tr>
<td>SNP</td>
<td>Single nucleotide polymorphism</td>
</tr>
<tr>
<td>Sol</td>
<td>Soluble protein</td>
</tr>
<tr>
<td>SRB</td>
<td>Sulfate reducing bacteria</td>
</tr>
<tr>
<td>SU</td>
<td>Starter unit</td>
</tr>
<tr>
<td>Ta</td>
<td>Annealing temperature</td>
</tr>
<tr>
<td>TCEP</td>
<td>Tris(2-carboxyethyl)phosphine</td>
</tr>
<tr>
<td>THN</td>
<td>1,3,6,8-tetrahydroxynaphthalene</td>
</tr>
<tr>
<td>UI</td>
<td>User interface</td>
</tr>
</tbody>
</table>
Abstract

Antimicrobial resistant microorganisms are predicted to pose an existential threat to humanity inside of the next 3 decades. Characterisation of novel acting antimicrobial small molecules from microorganisms has historically counteracted this evolutionary arms race, however the bountiful source of pharmaceutically relevant bioactive specialised metabolites discovered in the Golden era of drug discovery has long since dried up. The clinicians’ arsenal of useful antimicrobials is diminishing, and a fresh perspective on specialised metabolite discovery is necessary. This call to action is being answered, in part, through advances in genome sequencing, bioinformatics predictions and the development of next generation synthetic biology tools aiming to translate the biological sciences into an engineering discipline. To expedite our route to new pharmaceutically relevant specialised metabolites using the synthetic biology toolbox several bottlenecks need to be addressed, and are tackled here in.

Biosynthetic gene clusters (BGCs) represent blueprints to pharmaceuticals, however to date the vast wealth of knowledge about biosynthetic gene clusters is inconsistently reported and sporadically disseminated throughout the literature and databases. To bring the reporting of BGCs in line with engineering principles we designed and built a community supported standard, the Minimum Information about a Biosynthetic Gene cluster (MIBiG), for reporting BGCs in a consistent manner, and centralised this information in an easy to operate and open access repository for rapid retrieval of information, an essential resource for the bioengineer.

Prioritisation represents the next bottleneck in specialised metabolite discovery. Bioinformatics tools have predicted a cache of thousands of BGCs within publicly available genome sequences, however high experimental attrition rates drastically slows characterisation of the corresponding specialised metabolite. We designed and built an Output Ordering and Prioritisation System (OOPS), to rank thousands of BGCs in parallel against molecular biology relevant parameters, pairing BGCs with appropriate heterologous expression hosts and facilitating a judicious choice of BGCs for characterisation to reduce experimental attrition.

To fully realise the potential of synthetic biology in specialised metabolite discovery a genetically amenable heterologous host, capable of completing rapid design-build-test-learn cycles, is necessary. This cannot be achieved for the pharmaceutically important type II polyketides, as their biosynthetic machinery is largely restricted to Actinobacteria. Using MIBiG datasets, antiSMASH and BLASTP we identify 5 sets of soluble type II polyketide synthases (PKS) in *Escherichia coli* for the first time. We construct and test the robustness of a plug-and-play scaffold for bioproduction of aromatic polyketides using one PKS in *E. coli*, yielding anthraquinones, dianthrones and benzoisochromanequinones intermediates. Through bioprospecting for biological ‘parts’ to expand the chemical diversity of our plug-and-play scaffold we describe a new lineage of type II PKSs predominantly from non-Actinobacteria.

The standards, softwares, and plug-and-play scaffold and biosynthetic ‘parts’ described here-in will act as an engine for rapid and automated bioproduction of existing, and novel, pharmaceutically relevant aromatic polyketides in *E. coli* using the synthetic biology toolbox.
Declaration

Portions of the work within this thesis have been submitted in support of an application for another degree or qualification at the University of Manchester or any other university or other institute of learning. In these instances the contributions made by each author are clearly stated within the preface of each chapter. A first software draft of the Output Ordering and Prioritisation System (OOPS) described in Chapter 2 was submitted in support of a Master of Science by Alejandro Pena De Alba. Furthermore, preliminary characterisation of the ketosynthase and chain length factor proteins from Streptococcus sp. GMDXs, described in Chapter 7, were reported in support of a Master of Research by Thomas McManus.
Copyright statement

i. The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the “Copyright”) and s/he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

ii. Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

iii. The ownership of certain Copyright, patents, designs, trade marks and other intellectual property (the “Intellectual Property”) and any reproductions of copyright works in the thesis, for example graphs and tables (“Reproductions”), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

iv. Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property University IP Policy (see http://documents.manchester.ac.uk/display.aspx?DocID=24420, in any relevant Thesis restriction declarations deposited in the University Library, The University Library’s regulations (see http://www.library.manchester.ac.uk/about/regulations/) and in The University's policy on Presentation of Theses.
Acknowledgements

First and foremost I would like to thank the BBSRC for supporting me with a DTP studentship and enabling me to learn and grow as a scientist. This educational process was greatly fostered by both Professor Eriko Takano and Professor Rainer Breitling, who not only provided me with exceptional opportunities present my work, interact with seminal researchers within the field of specialised metabolites and travel the globe, but also taught me how to think laterally and enabled me to explore any, and all, facets of work in a self directed manner.

I was thrown in at the deep end arriving to a bench space in Manchester without a centrifuge or set of pipettes. Accordingly, I must give thanks to Dr. Ashley Chessher, not only for taking the helm of the ship and helping build the lab, but also for his continued support during the early years of my PhD studies. From year 2 the Takano lab began to grow into a vibrant group of young scientists, all of whom I would like to give thanks to, including Dr. Victoria Jackson and Dr. Gajendar Komati Reddy. Special thanks must be given to Marc ‘Carrera-Bernaise’, for his dry humour and persistence in the lab, Adam Amara, as a Champion of synthetic biology, and Dr. William Finnigan, as a voice of sanity in an otherwise chaotic group of individuals. Furthermore, I would like to thank Eriko for providing me with an opportunity to supervise multiple students and have found great friends in Felix Räsch, Fabian Schäufele and Lucas Kruass. It was a pleasure to work with you, and I wish you well in your studies.

My departure from the South was not well received by all: most notably Palms and Pelicans lost a great bass player, however I would like to thank all of my friends and family for supporting me irrespectively, and for taking the time to visit me in the far flung and beautiful City of Manchester. I have had an exceptional time here and met some outstandingly driven and inspiring individuals. I can only hope some of your virtues have rubbed off on me.

Finally, and most importantly, I must give thanks to Julia, whose life I have categorically and holistically up-heaved. Yet despite this, she has been the most magical constant in my life, supporting me irrespective of the ludicrous situations I create for us both.

I would also like to thank you, the reader, for taking the time to read these acknowledgements, full of names you may or may not recognise, and hope you find the thesis here-in a worthwhile read.
Thesis preface

The thesis here-in is formatted in an ‘alternative’ or ‘journal’ style and comprises a collection of in press or published manuscripts. Two additional unpublished chapters are also included in an extended journal format. Accordingly, each chapter is supported by a discrete introduction comprising a detailed search of current literature supporting, and contextualising, the concepts described within, as well as materials and methods, and results and discussion sections. To avoid repetition as far as possible, this thesis is supported by a brief and broad introduction, and a published manuscript detailing the implication of synthetic biology in specialised metabolite / natural product discovery titled ‘Steps towards the synthetic biology of polyketide biosynthesis’¹, Chapter 2. Furthermore, each thesis chapter comprises its own discrete bibliography, and consistent Nature style citations. Citations are formatted using single line spacing for the sake of space. Where used, supplementary materials, results, figures and table are attached at the end of each chapter, opposed to the end of the thesis to improve readability and individually indexed. Furthermore, methods and materials consistently used throughout the thesis are cited between chapters to avoid unnecessary repetition.

The contribution made to each manuscript is explicitly detailed in a preface at the beginning of each chapter. The status of publication is also detailed, and where manuscripts are already published each is cited appropriately in efforts to prevent self-plagiarism.
1 Introduction

Here-in, I briefly describe the journey from the inception of antimicrobial discovery to modern day approaches to isolation of specialised metabolites. Additionally, I discuss the operon architecture of specialised metabolite genetics and visit current bioinformatics methods for specialised metabolite bioprospecting. Furthermore, I discuss the utility of synthetic biology tools in specialised metabolite discovery, and provide an overview of polyketide biosynthesis not introduced elsewhere within the thesis. The benefits and caveats of each point of interest are also discussed as perspective of the thesis, and the aims of the thesis are listed.

1.1 Specialised metabolites

The term specialised metabolite is not yet widely adopted in literature. Here-in, we use this term to describe ‘natural products’ or ‘secondary metabolites’, two widely used descriptors for an extensive array of organic compounds produced by Bacterial\textsuperscript{1}, Archaea\textsuperscript{2} and Eukaryotes\textsuperscript{3} with diverse metabolic function. To be more specific, ‘Secondary metabolite’ is a widely adopted umbrella term which constitutes any metabolites that are nonessential for the proliferation of an organism, when grown in isolation\textsuperscript{4}. Historically, this terminology was derived, in part, from the observation that these compounds were prolifically produced after switching from primary to secondary metabolism\textsuperscript{5}, during the transition to stationary phase of the Actinobacterial life-cycle\textsuperscript{6}. Natural products share much the same definition as secondary metabolites, and both terms are used synonymously within the literature\textsuperscript{7}. Whilst the natural products terminology remains sound, the notion that these small molecules play a secondary function to key metabolites, or are only produced during stationary phase\textsuperscript{6} is beginning to fall as their role in natura is described\textsuperscript{9}. Furthermore, secondary metabolites are widely observed outside of the Actinobacteria\textsuperscript{1} and have diverse and specialised functions\textsuperscript{6}, therefore the term ‘specialised metabolites’ is more appropriate.

Irrespective of semantics, specialised metabolites are of considerable interest because these molecules have historically been an exceptional pool of medicines\textsuperscript{10,11} with a myriad antibiotic\textsuperscript{12-14}, anticancer\textsuperscript{15}, antiviral\textsuperscript{16,17} antiparasitic\textsuperscript{18} activities, etc\textsuperscript{19,20}. Accordingly, concerted efforts have been made to explore all facets of their biology including, but not limited to, discovery\textsuperscript{21-24}, biosynthesis and function\textsuperscript{25-28}.

1.1.1 Antimicrobial discovery

Classically, antimicrobial medicines were systematically discovered using a platform developed by Selman Waksman in the 1940’s\textsuperscript{29}. This platform originally screened soil-borne Actinobacteria isolates for useful antimicrobial activity by measuring the inhibition of growth of susceptible target organisms on agar plates after application of Actinobacterial culture extracts. Such an approach led to the discovery of streptomycin (Figure 1.1), the first aminoglycoside, which showed activity against \textit{Mycobacterium tuberculosis} and collectively earned Waksman a Nobel Prize\textsuperscript{30}. The success of the Waksman platform led to wide adoption by the pharmaceutical industry and discovery of numerous classes of potent specialised metabolites e.g. the chloramphenicols\textsuperscript{31}, macrolides\textsuperscript{32}, tetracyclines\textsuperscript{33}, and glycopeptides\textsuperscript{34} (Figure 1.1). However,
roughly two decades after its implementation the Waksman platform was heading for redundancy resulting from continuous rediscovery of the same specialised metabolites\textsuperscript{30,35,36}. The interests of pharmaceutical companies moved towards synthetic chemistry\textsuperscript{32} to explore novel chemical space because the underlying processes were more readily transferable to high throughput screening pipelines than libraries of extracts, partially purified mixtures and pure compounds from natural origins\textsuperscript{10}. Investment into medicines treating acute infections were less economically viable prospects for pharmaceutical companies and many antimicrobial pipelines were deserted\textsuperscript{37} accordingly, leading to the dearth of newly discovered antimicrobials today.

Throughout the Golden era of antibiotic discovery, spanning from the 40’s to the 60’s\textsuperscript{38}, it is thought that tens of millions of Actinobacterial isolates were screened for antimicrobial activity\textsuperscript{23}. Such an enormous effort was hugely valuable, leading to development of broad libraries of extracts which have been successfully mined for other important bioactivities\textsuperscript{7} e.g. the clinically utilised doxorubicin derived anticancer agents\textsuperscript{39}. And, whilst resistance to the main classes of antibiotics was typically observed within 10 years of their discovery\textsuperscript{30}, at that time semi-synthetic derivatives remained active. Consequently, more interest was garnered in modification of specialised metabolite scaffolds to develop the next generation of therapeutics e.g. 2\textsuperscript{nd} and 3\textsuperscript{rd} generation tetracyclines\textsuperscript{33} (Figure 1.2). Semisynthetic drug derivatives often only differed from the parent compound by one or two small modifications (Figure 1.2) and therefore, whilst unpredictable, the observation of resistance was inevitable\textsuperscript{30,40,41}. The significant disproportions between both time and cost of drug development, and the first instances of resistance in previously susceptible organisms led to little enthusiasm to continually pursue semisynthetic and synthetic routes to antimicrobial discovery after the mid 90’s\textsuperscript{37}. However, more recently, a return to semi-synthetic modification of existing bioactive specialised metabolite core structures has been prompted by the dearth of potent clinically used antimicrobial agents, and the rise of pan-resistant organisms e.g. Multi drug resistant \textit{Neisseria gonorrhoeae}\textsuperscript{42}, which are collectively predicted to pose existential risk to humans in the next 3 decades\textsuperscript{43}. The pertinence and drawbacks of synthetic approaches to drug discovery are demonstrated by the recent development of peripherally modified vancomycin derivatives with a 200-fold improvement in potency towards vancomycin-resistant \textit{Enterococci} (VRE), however necessitates a 30 step synthesis strategy\textsuperscript{44} which cannot be feasibly scaled.
Figure 1.1: Specialised metabolites with antimicrobial activity

All specialised metabolites were discovered using the Waksman platform, or similar approach with the exception of the penicillins which were serendipitously discovered by Alexander Fleming. Extensive modifications to groups in blue resulted in increased antimicrobial potency of vancomycin.
Figure 1.2: First, Second and third generation tetracyclines

Chlortetracycline and oxytetracycline were isolated from Streptomyces aureomycin and S. rimosus respectively. Tetracycline was derived from chlortetracycline, synthetically in an attempt to improve the potency and solubility. Second and third generation tetracyclines were derived semi-synthetically. Additions to the core scaffold are shown in red.

The drought of newly discovered specialised metabolites of microbial origin after the decline of Waksman-like platforms lead to the assumption that all ‘low-hanging-fruit’ had been picked i.e. all specialised metabolites with antimicrobial activity identifiable using these methods had been isolated. This assumption is now refuted, and Baltz (2006) concludes that Actinobacteria producing new antibiotics will be identified at frequencies of $\leq 10^{-7}$ per random isolate, evidenced by Lilly who screen approximately 5 million Actinobacteria before isolating the novel lipopeptide antibiotic daptomycin from S. roseosporus. Whilst this prediction is promising - there are new antibiotics to be discovered – the exceptionally low discovery frequency necessitates the screening of Actinobacteria at orders of magnitude higher than the 10 million isolates analysed during the Golden era of discovery. Furthermore, antibiotic screens must be married with rapid and efficient de-replication strategies to prevent continual rediscovery of known compounds. Dereplication is not trivial. Efforts are being made to circumvent this...
discovery bottleneck. Twice as many academic papers were published under the de-replication topic in 2012 than 2005, and this value constituted a ~10-fold increase from 1993. However, de-replication remains a significant blockade in antimicrobial discovery. Excitingly, exotic culture methods are facilitating access to a wealth of previously uncultivable organisms (reviewed in Chapter 7) extending the utility of a Waksman-like approach, however the same caveats will inevitably apply.

It was not until DNA cloning and DNA sequencing became routine that the biological route to specialised metabolite synthesis began to be unpicked. Seminal studies by Malpartida and Hopwood (1984), Cortes et. al. (1990) and Donadio et. al. (1991) elucidated the biosynthetic machinery underpinning two classes of clinically harnessed polyketide specialised metabolites: the aromatic polyketides, also known as type II polyketides, and the macrolides produced by type I polyketide synthases. Interestingly, in each case all genes encoding the entire biosynthetic pathway were present upon a contiguous region of DNA and not sporadically organised across the genome. The notion that genes implicated in the biosynthesis of specialised metabolites are clustered was concurrently observed in *Penicillium chrysogenum*, during attempt to elucidate a biosynthetic route to penicillin production, demonstrating this important genetic organisation was not confined to Actinobacteria, but also observable in the Eukaryotic Kingdom.

At this time DNA sequencing technologies were in their infancy and improved substantially over the following decades. This improved capacity to sequence DNA led to construction and publication of the first entire prokaryotic genome which subsequently had a profound effect on the field of specialised metabolite discovery and confirmed the theory that biosynthetic genes cluster. In the pre-sequencing era just four specialised metabolites had been characterised from the well-studied Actinobacteria, *Streptomyces coelicolor* A3(2), including actinorhodin, undecylprodigiosin, the calcium dependant antibiotic and methylenomycin. Whilst the whiE genes responsible for biosynthesis of grey spore pigment were also known, sequencing of the *S. coelicolor* A3(2) genome resulted in identification of a further 15 clusters of biosynthetic genes, superseded today by a current total estimate of 94 clusters, suggesting many more specialised metabolites were produced by this organism. Furthermore, genome sequencing of the well-studied erythromycin A producer *Saccharopolyspora erythraea* NRRL23338, facilitated identification of 25 clusters of biosynthetic genes, only 3 of which had been linked to characterised specialised metabolites previously. These observations were not in isolation. The combined utility of next generation DNA sequencing platforms and advances in genome mining tools presented an emerging picture in which the vast majority of genes implicated in the biosynthesis of specialised metabolites were ‘switch-off’, or down regulated, when organisms were propagated in laboratory conditions.

Such cryptic, or silent, clusters of biosynthetic genes provided a bountiful pool of new and undiscovered specialised metabolites for exploration. Furthermore, the repeated observation that all genes necessary for the biosynthesis of complex specialised metabolites were clustered
on chromosomal elements or plasmids, irrespective of class, meant a blueprint for each evolutionarily honed molecule could be identified with ease.

1.2 Biosynthetic genes cluster

Genes involved in the regulation, resistance and export of specialised metabolites were also shown to associate with biosynthetic genes, forming discrete biosynthetic gene clusters (BGCs). Each BGC typically contained all the genetic information necessary to initiate transcription of biosynthetic machinery, in response to external elicitors or stimuli, through to maturation and export of each specialised metabolite\(^7\)\(^{1,21}\) (Figure 1.3).

![Figure 1.3 Schematic of a biosynthetic gene cluster and its corresponding biosynthetic pathway](image)

**Figure 1.3 Schematic of a biosynthetic gene cluster and its corresponding biosynthetic pathway**

- **(a)** Archetypal biosynthetic gene cluster (BGC) producing a specialised metabolite. Biosynthetic genes are shown as blue ORFs, genes with export function in green, a resistance gene in red and regulators in black. The two cluster regulators differentially regulate the transcription of different operons and monocistrons within the biosynthetic gene cluster, indicated by black arrows.
- **(b)** A schematic of functioning enzymatic machinery encoded by biosynthetic gene clusters. 1. Substrates e.g. amino acids or acyl-CoA. 2. Polymerisation of the universal substrates by the specialised metabolite synthase. 3. Action of a resistance protein to protect the native producer, this can be through a myriad of different mechanisms including modification of the specialised metabolite or modification of the host target e.g. the ribosome\(^30\). 4. Efflux of the specialised metabolite. 5. Environmental dissemination of the specialised metabolite. Steps 3, 4, and 5 are not functionally conserved between all specialised metabolite BGCs and can be omitted. In fact, in some instances resistance is mediated by efflux.

1.2.1 Operons and evolution

The observation of genes which co-localise, on chromosomal or episomal element, partaking in congruent biological processes was not new\(^7\)\(^2\). In fact, to the most part prokaryotic genes with
related functions are co-localised in the genome, co-regulated, and also co-transcribed in the form of operons. Jacob and Monod coined the bacterial operon in seminal work considering the interplay between metabolites and protein biosynthesis in lactose metabolism, ultimately leading to the characterisation of the lac operon and an operon definition, which, to date, remains relatively unchanged. Simplistically, the classical operon comprises a minimum of two genes which are co-localised in the chromosome immediately adjacent to one another upon the same strand. Both genes are co-regulated by the binding of a regulatory protein to an operator DNA sequence close to a single promoter and, upon gene expression, co-transcribed to form a poly-cistronic mRNA. The two proteins, in this instance, are translated from different regions of one poly-cistron and may form sub-units of a protein complex or have completely dissociative roles. For instance, the lac operon comprises three genes: lacYZA which fulfil the ‘operon’ requirements, and all of which have described functions in lactose metabolism. These genes are co-regulated by the lac repressor which inhibits unnecessary expression of the lac genes. As with many operons, the lac repressor sits directly upstream of lacYZA and forms a regulon.

Figure 1.4: Two proposed models for the regulation of protein synthesis by Jacob and Monod

Each model considers the ability of a regulatory gene (R) to repress or activate the biosynthesis of structural proteins (A and B) differentially in response to the presence of a metabolite elicitor through the use of an operator (O). The repressor was thought to be mRNA which interacted with a metabolite, shown by \( r \) or \( i \). The combined mRNA:metabolite was posited to modulate affinity for the operator and therefore differentially activate of the operon. In model I the operator is proposed to be attached to structural genes A and B, where as in model II the operator is attached to the cytoplasmic messenger (i.e. the mRNA), which was thought to form a single polycistron. In both models structural genes A and B are proposed to behave as a unit and constitute a 'genetic unit of co-ordinate expression', or operon. This figure is adapted from Jacob and Monod (1961).
The widespread adoption of operons in prokaryote genome architecture poses several interrelated questions: How are operons formed? Why are operons maintained? And, how are operons evolving? This topic is hotly debated but several prevailing and interlinking theories have arisen and will be touch upon here. The first posits horizontal transfer of operons to play a critical role in operon maintenance, the second posits operon formation as a consequence of complex co-regulation of functional gene products, and the third suggests operon evolution results from piece-wise construction from smaller mono- or bi-cistrons. In expansion: the compactness of operons and their dissemination horizontally between organisms led to the proposition that operons are ‘selfish’. Gene organisation in this manner more commonly results in gene maintenance in the pan-genome by transmission, possibly improving the fitness of the acquiring organisms through pleiotropic regulation, or evolution of regulation, opposed to providing a metabolic burden in the form of redundant, or infrequently used DNA. This hypothesis explains the survival of operons however doesn’t extensively comment on the evolution of functionally coordinated gene sequences. In support of the second: operons are thought to provide a fitness benefit to the host organism by providing just-in-time expression of a suite coordinating genes enabling an organism to rapidly respond to metabolic changes. Co-regulation of operonic genes has been proposed to reduce the stochastic difference of proteins forming a complex, localise protein biosynthesis and facilitate protein folding. Co-regulation of two mono-cistrons is argued to evolve more readily through simple modification of upstream promoters than through formation of a bi-cistron i.e. an operon, however this does not scale with complexity, and activation of gene expression is often differentially regulated by a multiplicity of elicitors. Thirdly, evolution of some operons is demonstrated to occur through a piece-wise assembly of scattered genes into bi- and multi-cistrons, and subsequent duplications, and/or horizontal gene transfer events, to form longer and complex operons, as is demonstrated for the proteobacterial histidine operon. Intermediates in operon evolution will of-course comprise functionally unrelated genes, however these are thought to be removed as a consequence of purifying selection during genomic streamlining.

Each theory/observation comprises significant interplay, and multiple divergent and convergent evolutionary pressures are likely to influence, and refine, the existence of operons. Whilst ‘alien’ operons, those containing genes with no known common function, are described, it is widely accepted that the genes within an operon commonly encode proteins with congruent purposes. In fact, a systematic analysis of operon composition from 113 bacterial genomes indicated genes more commonly identified in operons to function in linear pathways e.g. fatty acid biosynthesis and peptidoglycan synthesis, opposed to bioprocesses comprised of several independent steps which converge e.g. isoleucine, leucine and valine biosynthesis. Bratlie et. al., (2010) suggest this preference is underpinned by the relatively static stoichiometry of proteins in linear processes e.g. forming large multimeric complexes, by comparison to dynamic pathways comprising protein complexes with multiple functions. Furthermore, the same study also suggests operon forming genes evolve more rapidly than genes less frequently observed in operons. This argument may seem counterintuitive, however proteins derived from genes less commonly forming operons were shown to have more
interaction partners than operon derived genes, and therefore are presumed to be under more stringent selection pressures.

The prevailing theories of operon persistence, life, and death have many tenets fitting with biosynthetic gene clusters. 1) BGC evolution is known to derive from the acquisition and loss of operons, as well as duplication, reordering, acquisition, and loss of individual genes at an unprecedented rate when compared to primary metabolite genes. This natural recombineering is a principle factor leading to the extraordinary chemical complexity of specialised metabolites (Figure 1.5). 2) Operons within BGCs often comprise genes encoding subunits of large and complex biosynthetic enzymes which partake in long and sequential biosynthetic processes e.g. type I polyketide synthases (reviewed in chapter 2) and non-ribosomal peptide synthetases, where stoichiometry of proteins is important. 3) Biosynthesis of specialised metabolites is exceptionally well regulated, and induction of BGC gene expression is often multifactorial, exemplified by the vast number of cryptic or silent BGCs. 4) specialised metabolite biosynthetic pathways are discrete: all genes necessary for maturation of the specialised metabolite are typically encoded within the respective BGC and therefore one could speculate that the interaction partners would be limited to this pool of proteins. There are some obvious exceptions where interplay between fatty acid synthase components and polyketide biosynthetic components can be observed. However, to the most part, biosynthetic domains, or genes, missing from BGCs are paralogues of primary metabolic genes, and can be functionally complemented by their orthologue counterparts. Therefore, reconstitution of functional biosynthetic pathways represents less of a discovery challenge by comparison to singleton genes, which produce a protein product performing a specific chemical modification, as is the case in Higher Eukaryotes e.g. in Plantae. Medema et al., (2014) posit that BGCs are constructed in a bricks and mortar fashion, where operons, or subsets of genes tasked with biosynthesis of specific chemical moieties are linked by individual ‘mortar’ genes e.g. promiscuous glycosyl- or acyl- transferases. The dynamic evolution of ‘mortar’ proteins facilitates the biosynthesis of new specialised metabolites from existing bricks. The ‘bricks and mortar’ model is fantastically exemplified by aminocoumarins (Figure 1.5), and the capacity of promiscuous glycosyltransferases to extensively glycosylate vastly different specialised metabolite core structures. Not all specialised metabolite BGCs comprise such discrete building blocks, however, and the evolution of diversity is less obvious. Furthermore, replication of bricks and mortar, or ‘legoisation’ strategies to reconstruct BGCs de novo fails routinely and much remains to be learned about gatekeepers and quality control mechanisms during specialised metabolite biosynthesis (reviewed in Chapter 2).

‘Bricks and mortar’ or otherwise, the evolutionary forces driving co-association of biosynthetic genes into discrete clusters are virtuous for the bioinformatician and the experimental biologist alike. This phenomenon has expedited the elucidation of biosynthetic pathways underpinning thousands of bioactive specialised metabolites. By contrast, the less frequent clustering of biosynthetic genes in Higher Eukaryotes has hampered progress in these counterpart organisms e.g. in anthocyanin and hypericin biosynthesis. Estimates suggest more than
25,000 specialised metabolites with antibiotic activity have been isolated from plants and animals, compared with more than 23,000 discovered from microbial sources, however more than 90% of clinically used antimicrobials are microbial in origin\textsuperscript{23,107}. The unknown route of plant specialised metabolite biosynthesis has certainly stalled the utility of these metabolites in the clinic and concerted efforts have been made to rebuild biosynthetic routes to important plant specialised metabolites \textit{de novo} in microorganisms e.g. opioids\textsuperscript{108,109}.

Figure 1.5: Modular chemical diversity of aminocoumarin specialised metabolites

The chemical diversity of aminocoumarin specialised metabolites appears to be modular. The mixing and matching of operons responsible for the biosynthesis of each chemical moiety has produced highly modular metabolites\textsuperscript{99}. This approach is not restricted to one class of specialised metabolites and has been observed in enediyynes\textsuperscript{110,111}, non-proteinogenic amino acid containing specialised metabolites\textsuperscript{112} polyketide precursor biosynthesis\textsuperscript{113}, 6-methylsalicylic acid (6-MSA) hybrids\textsuperscript{114} and hybrid isoprenoids (HI)\textsuperscript{115}.

1.2.2 BGC prediction

The chemical space currently accessed by known specialised metabolites is vast (Figure 1.1), however the enzymology responsible for biosynthesis is often modest e.g. the type I polyketide synthases (PKSs) use variations of the same biosynthetic machinery\textsuperscript{116} to produce polyenes\textsuperscript{117}, structurally diverse macrolides\textsuperscript{118}, enediyynes\textsuperscript{119} and mono and bicyclic aromatic structures\textsuperscript{120}. The catalytic domains of biosynthetic enzymes are often exceptionally similar at the amino acid level\textsuperscript{121}, and within similar host species, at the nucleotide level, also. Conservation of nucleotide
sequences historically facilitated BGC discovery in organisms for which no genome sequence was available using a reverse genetic approach e.g. Southern blotting\textsuperscript{122}, or using degenerate primers to screen genetic libraries by polymerase chain reaction (PCR) methods\textsuperscript{123,124}. Owing to the decrease in genome sequence costs\textsuperscript{125} these processes are now carried out \textit{in silico}, however follow a similar premise i.e. characterised reference sequences specific for each class of biosynthetic machinery are used to identify protein homologues within a translated genome sequence, or contiguous string of DNA. The convenience of \textit{in silico} BGC identification stimulated development of a suite of computational tools for BGC discovery, many of which can be accessed via the secondary metabolite bioinformatics portal (SMBP)\textsuperscript{126}.

The first automated BGC discovery platform, Decipher, was developed in 2003\textsuperscript{127}. This software pipeline initially scanned 60 Actinobacterial genomes and identified approximately 700 BGCs. This large repository of predicted BGCs exceeded the number of identified specialised metabolites observed through traditional culture based identification by a factor of ten\textsuperscript{128}. Decipher not only identified BGCs but also predicted the chemical structure of each putative specialised metabolite, as well as the molecular weight and predicted UV-absorbance spectrum of each end compound. These values were compared with databases of known specialised metabolites to determine if BGCs encoded new chemical entities or known specialised metabolites, and were experimentally characterised accordingly. This genomic-driven discovery platform was before its time, and current BGC mining tools follow a similar, but refined, pipeline.

Comprehensive suites of bioinformatic tools are available today and enable prediction and annotation of particular classes of specialised metabolite biosynthetic machinery including non-ribosomal peptide synthetases (NRPSs) and PKSs e.g. CLUSEAN\textsuperscript{129}, 2metDB\textsuperscript{130}, SEARCHPKS\textsuperscript{131}, NRPS-PKS\textsuperscript{132}, SBSPKsv2\textsuperscript{133} and NaPDoS\textsuperscript{134}, ribosomally synthesis and post-translationally modified peptides (RiPPs) e.g. BAGEL3\textsuperscript{135} and type III PKSs\textsuperscript{136}, as well as more holistic programmes which identify suites of known and unknown classes of specialised metabolite BGCs in prokaryotes and fungi e.g. SMURF\textsuperscript{129}, PRISM\textsuperscript{3}\textsuperscript{137}, EvoMining\textsuperscript{138}, FunGeneClusterS\textsuperscript{139}, Cluster Finder\textsuperscript{24} and antiSMASH\textsuperscript{140-142}. Most tools use a rule based approach to BGC identification, paralleling the reverse genetics approach, however others use non-sequence based approaches. One common rule based method for BGC identification is exemplified by the antiSMASH pipeline.

The first version of the antibiotics and Secondary Metabolite Analysis Shell (antiSMASH)\textsuperscript{140} mined BGCs from a DNA input sequence using Hidden Markov Model (HMMs) searches to identify translated gene sequences with sequence commonalities to characterised biosynthetic enzymes and updated versions still utilise this as one BGC discovery method. In short, HMMs for core biosynthetic proteins, or domains of proteins, were constructed from multiple amino acid sequence alignments of characterised proteins for each specialised metabolite class. Each HMM was tested to determine a false positive threshold and essentially acts as a Southern blotting probe, \textit{in silico}. Predicted BGCs are defined by clustering of characteristic HMM hits within a contiguous stretch of DNA, with a maximum distance between HMM hits of 10 kb. To ensure enzymes with putative tailoring function are included within the predicted BGC
antiSMASH is inherently ‘greedy’ and includes an additional 5, 10 or 20 kb either side of the flanking most HMM. This approach has caveats as BGCs are known to localise at specific genomic loci\textsuperscript{143} and therefore using a ‘greedy’ approach, multiple BGCs can be amalgamated \textit{in silico} resulting in formation of superclusters. This phenomenon is often seen for ectoine\textsuperscript{144} and carotenoid\textsuperscript{145} BGCs, which comprise relatively few genes. BGC boundaries can be further refined bioinformatically using CASSIS\textsuperscript{146}, a tool which identifies co-regulated proteins in Eukaryotes, through commonalities in promoter motifs, and the Cluster Finder algorithm which identifies densities of specialised metabolite associated Pfam domains\textsuperscript{147} in protein coding sequences across the entire ‘greedily’ predicted BGC. Furthermore, BGC boundaries can also be defined by the expert and a ‘greedy’ and inclusive approach to BGC identification is favourable, therefore.

The Cluster Finder\textsuperscript{24} algorithm and EvoMining\textsuperscript{138} tool use non-rule based approaches to BGC discovery i.e. these methods do not rely on sequence information from characterised classes of biosynthetic gene clusters for discovery, and therefore enable identification of unknown, or unusual classes of BGC\textsuperscript{138}. So far non-rule based strategies have led to discovery of novel arseno-organic compounds in model \textit{Streptomyces} organisms\textsuperscript{138} and elucidated the extraordinary capacity of proteobacteria to produce aryl-polyenes, a previously underexplored class of specialised metabolite\textsuperscript{24} (Figure 1.6). These approaches to BGC identification may be more fruitful in Archaea, for instance, in which very few specialised metabolites have been identified, and new classes of compound may remain undiscovered.

Figure 1.6: Specialised metabolites identified using non-rule based bioinformatics tools

Irrespective of identification method, all softwares centre on the observation that biosynthetic genes cluster, without this phenomenon identification of the correct biosynthetic complement necessary for specialised metabolite synthesis would likely be confounded by the large number of biosynthetic paralogues and homologues within most genomes, as is the case in plants. In contrast, the clustering of biosynthetic genes has been demonstrated in plants more recently for the biosynthesis of dhurrin in \textit{Sorghum bicolor}\textsuperscript{148}, thalianol in \textit{Arabidopsis thaliana}\textsuperscript{149}, and noscapine in \textit{Papaver somniferum}\textsuperscript{150}, prompting the development of a plant specific antiSMASH platform, \textit{plantiSMASH}\textsuperscript{151}. \textit{PlantiSMASH} aims to systematically identify BGCs responsible for alkaloid, polyketide, saccharide, lignin and terpene biosynthesis, however with a lower
proportion of genes forming operons in plant genomes\textsuperscript{75}, and characterised BGCs spanning large regions of plant chromosomes, the utility of this tool remains to be seen. As well as identification of physically clustered genes, transcriptomics data, evolutionary gene co-occurrence and epigenetic co-regulation are necessary to support plant BGC predictions\textsuperscript{152}. PlantiSMASH is in its infancy, with only 773 jobs run, compared to ~322,000 processed by antiSMASH to-date.

A genes-to-molecules approach to specialised metabolite discovery is clearly an effective strategy to isolate and characterise bioactive molecules missed using Waksmann-like methods: bioinformatics searches for modular PKS genes in \textit{Streptomyces ambofaciens} led to identification of the cryptic stamobmycin BGC which spanned 150 Kb of the host genome and underpinned biosynthesis of the 51 membered macrolactam cytotoxic stambomycins A-D\textsuperscript{153}. A similar approach led to identification and activation of a cryptic BGC responsible for asperfuranone biosynthesis in \textit{Aspergillus nidulans}\textsuperscript{154}, which shows anti-proliferant activities against human non-small cell lung cancer cell lines\textsuperscript{155}. Furthermore, BGC predictions using antiSMASH have led to identification of new pentangular aromatic polyketides, hexarcins A-C from \textit{Streptosporangium} sp. CGMCC 4.7309\textsuperscript{156}, the non-ribosomal peptide thanapeptin from \textit{Pseudomonas} sp. SH-C52, with anti-oomycete activity against the late potato blight pathogen, \textit{Phytophthora infestans}\textsuperscript{157,158}, the lipopeptide antibiotic taromycin from \textit{Saccharomonospora} sp. CNQ490\textsuperscript{159} and the type IV lanthipeptide streptocollin produced by \textit{S. collinus} Tü 365\textsuperscript{160}, to name a few.

Identification of BGCs \textit{in silico} is now considered routine. The bottleneck to specialised metabolite discovery now lies firstly with a fervent need to prioritise the hundreds of thousands of predicted and uncharacterised BGCs in databases such as the Atlas of Biosynthetic Gene Clusters\textsuperscript{161}, and secondly, in advancement of molecular biology techniques necessary for rapid characterisation of BGC end products; the latter of which is being answered by the advent and development of synthetic biology toolbox\textsuperscript{162}. Despite development of numerous BGC mining tools, there are few dedicated softwares which sort the output of these tools into a coherent list of BGCs worthwhile characterising. Attempts to prioritise BGCs based on a variety of important discovery based parameters have been made by existing softwares e.g. the KnownClusterBlast (KCB) function\textsuperscript{163} of antiSMASH 3.0 and above\textsuperscript{142}, and the recently published Antibiotic Resistance Target Seeker pipeline (ARTS)\textsuperscript{164}, however both approaches fill different spaces within the prioritisation sphere and suffer caveats. KCB is tasked with de-replication of BGCs, however suffers from the need to visualise each BGC manually. Alternatively ARTS prioritises BGCs based upon the predicted cellular target of the end compound however, for best predictions, is currently restricted to Actinobacterial hosts\textsuperscript{164}. Neither considers the quagmire of molecular biology and synthetic biology techniques necessary to realise the gene-to-molecules discovery of specialised metabolites. The work here-in attempts to traverse this gap by prioritising predicted BGCs based upon key molecular biology and synthetic biology relevant parameters.
1.3 Synthetic biology for specialised metabolites

The definition of synthetic biology remains controversial, however themes conserved between multiple definitions include the utility of engineering principles in biology e.g. abstraction, standardisation, modularisation, and characterisation. All engineering processes comprise layers of complexity. Each layer can be sequentially deconstructed into multiple less complex layers, and following a reductionist approach one can reach the simplest constituents of each process. Should the individual constituent be sufficiently characterised, a new complex process with altered functions can be rebuilt. Furthermore, the workings of the new process should be easy to predict, given the properties of the starting constituents, enabling in silico prototyping. The standardisation and characterisation of biological parts e.g. promoters, ribosome binding sites (RBSs), insulator sequences, terminators, and various coding sequences, enzymes, etc., has been shown to facilitate construction of biological devices de novo with predictable and reproducible outputs, exemplified by construction of the 189 logic gates in the SynBioLGDB repository to-date, using an engineering approach to biology a.k.a synthetic biology. However, biological systems are inherently noisy and currently our understanding of the wider context dependencies of each constituent part is too poor to enable accurate and efficient prototyping in silico. Biophysical models for certain parts e.g. RBSs and terminators are available and somewhat reliable, however, modelling larger biosynthetic pathways constructed de novo using these parts remains to be realised, and still required empirical elucidation. For instance, a reductionist synthetic biology approach was used to completely rebuild the complex nitrogen fixation cluster from Klebsiella oxytoca using 89 characterised parts, including protein coding gene sequences, promoters, insulators, RBSs etc. (exemplified using the schematic in Figure 1.7). Complementation of this pathway in Klebsiella oxytoca only recovered 8.4% ± 2.4% of wild type nitrogen fixing activity. The construction process was refined using the design, build, test, learn cycle, and a further 122 gene clusters were built, each comprising over 100 characterised parts, and each with radically different operon architecture, gene orientation and order, and regulatory parts. Here, the best performing cluster recovered only 57% of wild type activity. Clearly huge efforts and time was necessary to multiplex the nitrogen fixation cluster and despite this the best synthetic pathway was suboptimal. Gene cluster redesign on such a scale is exceptionally complex, and accordingly unachievable for most laboratories, as well as cost ineffective for industrial processes, demonstrating the synthetic biology field is still within its infancy. This seminal study did succeed in decoupling and complementing a complex natural process using an orthogonal gene cluster constructed de novo however, and in doing so, set precedence within the field of synthetic biology.

1.3.1 Rebuilding biosynthetic gene clusters

The reconstruction of a biosynthetic pathway from first principles using characterised parts, as detailed above, is often termed refactoring, and is a principle borrowed from computer programming. In the computational sphere, this descriptor refers to the holistic overhauling of code underpinning a certain process without altering the output or function of this process. Typically, computer programs are refactored to improve their efficiency and stability, however in
biology refactoring is used to abolish or decouple the highly integrated, redundant, and often unknown native regulation within a system, modularise the system into its simplest parts and facilitate reconstruction of this system using characterised regulatory elements based upon a user defined specification\textsuperscript{176}. Refactoring is a powerful tool for awakening cryptic and silent BGCs which are stringently down regulated, or where regulatory elements have been lost abolishing expression\textsuperscript{159,179}. In addition to awakening BGCs, by using characterised parts, one is also able to tune gene expression\textsuperscript{180,181} and balance protein stoichiometry within a given pathway to optimise flux towards a compound of interest, theoretically concurrently improving yields and reducing the metabolic burden of biosynthetic intermediates to the heterologous host.

Refactoring strategies are popularising and have been successfully harnessed to awaken the lipopeptide taromycin from \textit{Saccharomonospora} sp. CNQ49\textsuperscript{159}, uncoupled the biosynthesis of the useful antiviral polyketide A-74528 from its cytotoxic shunt metabolite fredericamycin\textsuperscript{182}, translate a docosahexaenoic acid long-chain polyunsaturated fatty acid synthase from the \textit{Myxobacterium Aetherobacter fasciculatus} (SBSr002) to \textit{Pseudomonas putida} and \textit{Escherichia coli}\textsuperscript{183} and reconstruct the 56 Kb epothilone BGC from \textit{Sorangium cellulosum} for expression in the more genetically tractable \textit{Myxococcus xanthus} host\textsuperscript{184}. The latter two examples consider refactoring in its truest sense, i.e. completely optimising all DNA coding sequences to match codon usage and translational dynamics of the intended expression host. Whilst refactoring is a useful tool to awaken BGCs the corresponding specialised metabolites are typically produced at sub-preparative yields e.g. epothilone yielded 100 µg/L; suffering from complexity caveats. More holistic and informed routes to specialised metabolite biosynthesis are necessary to bridge this gap in yield i.e. through a better choice of enzymatic constituents (Chapter 2), as well as advances in chassis engineering improving orthogonality\textsuperscript{185}, precursor flux\textsuperscript{186} and resistance to end compounds. Refactoring is gathering momentum as a powerful tool for discovery of novel specialised metabolites and with associated advances in systems biology and cheminformatics aims to underpin a new golden age of antibiotic discovery\textsuperscript{38}.

\textbf{Step 1:} Identification of the BGC from sequence mining and subsequent identification of regulatory elements. Regulatory elements do not require refactoring as user-designed regulatory features will be added.

\textbf{Step 2:} BGC protein coding sequences (CDS) are deconstructed to their smallest constituent parts
Step 3: Codon usage of each CDS is altered accordingly to match the codon bias of the intended heterologous host, whilst considering translational dynamics.

Characterised orthogonal regulator parts (promoters, RBSs, insulators and terminators) are used to construct new BGC designed de novo. Synthetic BGCs are subsequently introduced into heterologous expression hosts, expressed and evaluated.

Figure 1.7: A guide to refactoring

1.3.2 Semantics in synthetic biology

Not only does synthetic biology call for the use of engineering principles when characterising and constructing biological devices, but also in standardisation of synthetic biology tools, methodologies, computational languages, semantics, and nomenclature. Endy (2005) details this succinctly, demonstrating the cost of poor standardisation to the biological engineer is vast both economically and in time. To elevate synthetic biology from the field of simple genetic engineering a holistic description of biological parts and the techniques used to manipulate these, in silico and in the wet lab, need to be comprehensively catalogued, routinely available and interoperable, opposed to being disseminated across the literature and specific databases, which are often poorly maintained, and described using synonymous terminologies. Canton et al., (2008) argue for the routine reuse of biological parts to be beneficial, each should be catalogued with a datasheet detailing up-to-date characterised attributes. Efforts have been made to standardise individual biological parts, largely driven by the iGEM (international Genetic Engineered Machines) competition and corresponding registry of standard biological parts. Implementation of a Synthetic Biology Open Language (SBOL) has gone some way to tackle syntax and semantics of synthetic biological parts, and SBOL version 1, 2 and SBOL visual are now integrated into more than 27 Synthetic biology relevant tools e.g. bio-Computer Aided Designers Cello and TinkerCell. In fact, using SBOL, synthetic biology ontologies have been constructed e.g. SyBiOnt, enabling mining and visualisation of all available data concerning a particular biological part from multiple highly disparate databases and generation of a corresponding knowledge base. Generation of synthetic biology ontologies and accompanying knowledge bases are already showing their utility in attempts to
automate identification of Pareto fronts when designing complex genetic circuits\textsuperscript{196}, rapidly linking semantics to genetics.

The implementation of ontologies, and use of standardised syntax and semantics, is not as stringently adopted by biological fields which are not considered ‘pure’ synthetic biology, despite being the source from which most ‘synthetic biological parts’ are derived. A fantastic example of this is demonstrated by the synonymous use of secondary metabolite, natural product, and specialised metabolite (1.1) throughout published literature. The synthetic biology toolbox offers great potential to build biosynthetic routes to new chemical entities, or more efficient/alternative routes known chemical entities, exemplified fantastically by the biosynthesis of hydrocodone in \textit{Saccharomyces cerevisiae} using mammalian, plant, yeast and bacterial derived enzymes\textsuperscript{108,109}, however, the robustness of a given pathway is dependent upon its least well characterised part. This is most likely going to be a biosynthetic constituent as there is no universal standard when reporting information about a BGC. Should a pathway malfunction as a result of a regulatory element, one could easily substitute this part for another characterised part, from a library\textsuperscript{197}, or the registry of standard biological parts\textsuperscript{189,190}. This is not currently the case for a biosynthetic enzyme, however, and one needs to scan the literature to identify functional homologues from alternative BGCs, or identify protein coding sequences in poorly maintained and out of date databases\textsuperscript{198}. After time-consuming identification of appropriate protein coding sequences, one must empirically characterise each candidate within an appropriate context. This lengthy process can fulfil a research project itself. To circumvent these issues, and facilitate rapid prototyping of BGCs designed de novo, Canton’s datasheet are needed not only for regulatory parts but also for specialised metabolite biosynthetic parts and resistance mechanisms. Implementation of a minimum standard of information when reporting a biosynthetic gene cluster is a stepping stone towards standardisation of BGC constituents, and will underpin a valuable knowledge base of vibrant catalytic enzymes which can be integrated into BioCAD software for the automated design of biosynthetic gene clusters. Using standardised syntax and semantics and reporting necessary information about a biosynthetic gene cluster we hope to move from genes-to-molecules more rapidly, and begin to build new specialised metabolites de novo.

1.4 Polyketides

Polyketides represent a pool of exceptional specialised metabolites with valuable bioactivities. Pharmaceutically relevant polyketides include frontline antibiotics e.g. erythromycin\textsuperscript{199}, the tetracyclines\textsuperscript{46}, rifamycin\textsuperscript{200} etc. as well as antiparasitics e.g. the avermectins\textsuperscript{201}, antivirals e.g. A-74528\textsuperscript{182}, anticancer agents e.g. doxorubicin and analogues thereof\textsuperscript{15}, antifungals e.g. amphotericin\textsuperscript{202} and nystatin\textsuperscript{203}, cholesterol lowering drugs e.g. lovastatin\textsuperscript{204}, insecticides, e.g. spinosyn\textsuperscript{205} and immunosuppressants e.g. rapamycin\textsuperscript{206}. These representative structures span an exceptional chemical space yet all core structures are derived from sequential decarboxylative Claisen condensations of a thioester malonate derivative and an acyl-thioester (Figure 1.9) using one of three core classes of polyketide synthase: type I, type II and type III PKSs (Figure 1.8). Polyketide biosynthesis shares many parallels with fatty acid synthases
(FASs)\(^{207}\), in fact type I PKSs resemble mammalian type I fatty acid megasynthases\(^{208}\), and type II PKSs resemble plant and bacterial FASs\(^{209,210}\). Both PKSs and FASs synthesise acyl chains through successive decarboxylative Claisen condensation (Figure 1.9), however whilst FASs are primarily restricted to acetyl- and malonyl- starter and extender units, polyketide synthases are able to use more vibrant acyl- substrates\(^{211}\). Furthermore, fatty acids are usually fully reduced whilst polyketides are not\(^{207}\). The degree and timing of β-keto reduction is another principle distinction between PKSs and FASs. During fatty acid biosynthesis the β-keto group of the growing fatty acid chain is sequentially reduced to form a fully saturated thioester after each condensation reaction\(^{207}\), however, during polyketide biosynthesis the β-keto group either remains unreduced until the poly-β-ketide chain is fully synthesised (typical for type II and III PKSs)\(^{210}\), or is selectively reduced to the corresponding secondary alcohol, olefin or alkane (as observed for type I PKSs)\(^{25}\).

Type I polyketide synthases, and lessons learned from combinatorial biosynthesis of polyketide biosynthetic machinery, are the focus of published review article (Chapter 2), and introduction of diversity into nascent type II polyketide chains is discussed thoroughly in the introduction to Chapter 6, accordingly these will not be introduced extensively here. The biosynthesis of archetypal type II polyketides will be introduced along with Type III PKS biosynthesis.

**Figure 1.8: Polyketide synthases**

**Figure 1.9: Polyketide biosynthesis**

Archetypal two-step mechanism for carbon-carbon bond formation by fatty acid synthases (FAS) and polyketide synthases (PKSs). The FAS mechanism mirrors type II PKSs. Decarboxylation of the thioesterified malonate derivative is followed by Claisen condensation to produce a \(\beta\)-ketothioester.

### 1.4.1 Type I polyketide synthases

Type I polyketide synthases come in a variety of flavours including collinear acting enzymatic assembly lines e.g. the deoxyerythronolide B synthases (Figure 1.8a)\(^{213}\), and iterative type I PKSs (Figure 1.8b) identified in both fungi\(^{214}\) and prokaryotes\(^{110,215}\). In both cases, type I polyketide synthases are remarkable enzymatic complexes which often exceed 1 MDa in size\(^{213}\). These large ‘megasynthases’ are comprised of modules, and each module comprises a suite of enzymatic domains necessary for recruitment, installation, and selective reduction of acyl-CoA substrate into the growing polyketide chain. In collinear type I PKSs each module is responsible for introduction of a single acyl-group into the growing polyketide chain, whereas a single module is used repetitively in iterative type I PKSs\(^{210}\). Erythromycin biosynthesis is arguably the most well studied type I PKSs and comprises 3 megasynthases, DEBs 1-3\(^{199}\), encoding 29 enzymatic domains. DEBs 1-3 is modest in size, however, and collinear PKSs can be vast. One of the largest type I PKSs comprises 112 enzymatic domains within 25 modules, and produces the 51 membered macrolide stambomycin\(^{153}\).
1.4.2 Type III polyketide synthases

Unlike type I and II PKSs, type III PKSs have no parallel in fatty acid biosynthesis and this class of PKSs were originally thought to be restricted to plants. Indeed chalcone synthases, a type III PKS, is ubiquitous in higher plants\textsuperscript{216} however in the past two decades this enzymatic machinery has been characterised from numerous prokaryotes\textsuperscript{217}, and fungi\textsuperscript{127}: first discovered in \textit{Streptomyces griseus} (RppA)\textsuperscript{218} and \textit{Aspergillus oryzae} RIB40 (CysA)\textsuperscript{219}, respectively. Both \textit{rppA} and \textit{cysA} were successfully expressed and characterised in \textit{E. coli}\textsuperscript{218,220} and do not suffer for the insolubility issues described for type II PKSs\textsuperscript{221}. The core type III PKS biosynthetic machinery differs greatly from the collinear and iterative type I PKSs, and comprises a single discrete homodimeric ketosynthase protein which iteratively recruits and condensates acyl-CoA substrates to form a polyketide chain all catalysed by a single active site\textsuperscript{217} (Figure 1.10). Furthermore, the cyclisation of unreduced polyketide chains can also be directed by the type III PKS\textsuperscript{222} leading to formation of a variety of polyketide core scaffolds (Figure 1.10). Interestingly, the type III PKS is the only PKS not to employ an acyl carrier protein, instead acyl-CoA substrates are used directly\textsuperscript{217}. For compounds produced using modest biosynthetic machinery considerable chemical diversity exists in type III polyketides. This diversity is introduced through a vibrant choice of polyketide starter units\textsuperscript{223}, the number of decarboxylative Claisen condensation reactions and differential intramolecular aldol condensation reactions of the full length polyketide chain\textsuperscript{224}, sharing tenets with type II PKSs (Chapter 6).

The biosynthesis of type III polyketides in prokaryotes and fungi differs from plants in the choice of common starter units. Prokaryotic and fungal type III PKSs typically utilise aliphatic acyl-CoA starter units followed by successive decarboxylative Claisen condensation with up to 5 malonyl-CoA extender units\textsuperscript{225}, all of which are catalysed by the type III PKS, whereas plant type III PKSs typically harness aromatic acyl-CoA starting substrates e.g. 4-coumaroyl-CoA in naringenin biosynthesis\textsuperscript{226}, and can naturally catalyse up to 7 elongation steps\textsuperscript{227}. Exceptions exist, and recently \textit{Streptomyces clavuligerus} was shown to produce naringenin using a 4-coumaroyl-CoA starter unit\textsuperscript{228}. Interestingly, opening of the OKS type III PKS active site from \textit{Aloe arborescens} using rational mutagenesis enabled biosynthesis of much longer \textit{C}_{20} and \textit{C}_{24} polyketide chains\textsuperscript{229}, rivalling the typical length of type II polyketides. Mutagenesis of type III PKSs active site pocket is a well explored method to redirect type III polyketide biosynthesis, and was facilitated by the first crystal structure of any PKS from \textit{Medicago sativa} (Alfafa)\textsuperscript{230}.
Figure 1.10: Selection of type III polyketide core structures

The bioactivities of type III PKS derived specialised metabolites are broad including depigmentation properties of oxyresveratrol\textsuperscript{231}, antidepressant and anxiolytic properties of cannabidiol\textsuperscript{232}, the anticancer\textsuperscript{233}, antioxidant\textsuperscript{234} and anti-inflammatory\textsuperscript{235} properties of more than 6000 described flavonoids\textsuperscript{236} and the anti-parasitic activity and cytotoxicity of quinolone based type III polyketides, and semi-synthetic derivatives thereof\textsuperscript{237}. Accordingly, a continued effort is being made to characterise and engineer new type III PKSs to harvest the associated bioactive specialised metabolites\textsuperscript{224} and towards construction of optimised class specific bio-production chassis\textsuperscript{185}.

1.4.3 Type II polyketide synthases

The type II polyketides, or aromatic polyketides, are a vibrant source of valuable bioactive specialised metabolites synthesised by type II PKSs. These aromatic compounds comprise a diverse range of chemotypes (Figure 1.15) one of which, the anthracyclines possess some of the most effective anticancer activities discovered to-date, and accordingly are heavily used in the clinic\textsuperscript{238}. Another chemotype, the tetracyclines (Figure 1.2), have historically received great interest as antibiotics\textsuperscript{33}. Whilst this diminished in the 90’s, a renewed interest has seen tetracycline derivatives recently enter clinical trials\textsuperscript{239,240} and this resurgence has resulted in genesis of companies aiming to diversify tetracycline compounds (Tetraphase Pharmaceuticals)\textsuperscript{241} highlighting the usefulness of this pool of molecules in the clinic.

1.4.3.1 The minimal polyketide synthase (mPKS)

Type II polyketide synthases comprise three core components, a ketosynthase (KS), a chain length factor (CLF) and an acyl carrier protein (ACP), often referred to as the minimal polyketide synthase (mPKS) (Figure 1.8). Each mPKS components is expressed from an individual
gene\textsuperscript{242-244}, unlike type I PKSs, and the genetic organisation of the mPKS genes are exceptionally well conserved: genes are adjacent to one another and in the order described above, with few exceptions\textsuperscript{245,246} (Figure 1.11). The KS and CLF form a heterodimeric complex in its smallest quaternary form, which acts iteratively with a cognate ACP to catalyse formation of a poly-\(\beta\)-ketide chain\textsuperscript{244}. Whilst the KS and CLF typically share high sequence similarity e.g. 30\% identity and 96\% coverage for actinorhodin KS and CLF, each dimeric partner plays an individual role in biosynthesis. The KS catalyses C-C bond formation by Claisen type condensation which facilitates growth of the polyketide chain, however despite being a KS paralogue\textsuperscript{247}, the CLF plays a more passive role in biosynthesis\textsuperscript{248}. Initially the CLF was thought to be catalytically silent because the active site cysteine was consistently substituted for a glutamine residue, however, Leadley \textit{et al} (1999)\textsuperscript{249} implicated this conserved glutamine residue in decarboxylation of the ACP tethered malonyl- starter unit, priming the KS with an acetyl-group and initiating polyketide biosynthesis. Analogous KS\textsuperscript{5} condensation deficient ketosynthases have been described in modular PKSs, which consistently show a decarboxylating function, priming polyketide biosynthesis with an acetyl- starter unit\textsuperscript{250,251} (Figure 1.12). The exact enzymatic mechanism for decarboxylation remains to be elucidated, and several CLFs have been described without conserved glutamine residues including AurC\textsuperscript{252-254} and AntD\textsuperscript{255}, both of which are derived from non-actinobacterial hosts, as well as WhiE from \textit{S. coelicolor}\textsuperscript{256}.

![Figure 1.11: Genetic organisation of type II mPKS components](image)

\textbf{Figure 1.11: Genetic organisation of type II mPKS components}

Ketosynthases (KS), chain length factors (CLF) and acyl carrier proteins (ACP) are typically found in the genetic organisation shown by \textit{act}, actinorhodin, \textit{gra}, granaticin, and the \textit{frm}, frenolicin, BGCs. Exception to this rule include the \textit{med}, medermycin, BGC where the ACP is situated more than 15 Kb upstream of the KS/CLF genes, and the \textit{rem}, resistomycin, BGC where the ACP is separated from the KS/CLF genes by a PPTase encoding ORF.
Figure 1.12: Conserved active site glutamine residues within CLF and KS^Q

Active-site cystine residues of type I and type II PKS KSs, and corresponding glutamine residues from type II CLFs and type I KS^Q domains. Act KS, CLF and WhiE CLF are derived from S. coelicolor A3(2)^257, TcmL from S. glaucescens^257, OxyB from S. rimosus^45 and FrnM from S. roseofulvus^258. The erythromycin KS1 sequence is derived from Saccharopolyspora erythraea NRRL 233854,^259, and monensin, tylosin and niddamycin KS^Q from S. cinnamonensis^250, S. fradiae^360 and S. caelestis^251, respectively.

In addition to decarboxylation, the predominant function of the CLF is to control the length of the growing polyketide chain^244,248, the mechanism of which is introduced at depth in Chapter 6. In brief, an amphipathic cavity is formed at the dimeric interface between the KS and CLF into which the growing polyketide chain is extruded. The size of key amino acids, or ‘gatekeeper residues’ lining this cavity sterically limit the number of sequential iterative condensation reactions before the cavity is full, thereby capping the length of the polyketide chain^244,248. The polyketide chain length can vary from the modest octoketide of actinorhodin, C₁₆, to much larger C₃₀ pentangular aromatic polyketides^261. Attempts to engineer the CLF, derailing polyketide biosynthesis towards non-native polyketide chain lengths have been met with mixed success^247,248,262,263. The full complexities of the CLF are introduced in full in Chapter 6.

As well as the KS/CLF heterodimer each minimal PKS comprises an acyl carrier protein. This small highly soluble acidic protein shuttles starter and extender substrates to the KS/CLF heterodimer and tethers the growing polyketide chain during biosynthesis. The ACP cannot function in its translated, orapo, form, however, and must first be post-translationally modified through the addition of a phosphopantetheiny1 arm (PP): a reaction catalysed by a phosphopantetheinyl transferase (PPTase)^264 (Figure 1.13) and ubiquitous to fatty acid biosynthesis, non-ribosomal peptide biosynthesis and polyketide biosynthesis, excluding the ACP deficient type III PKSs^265. The PPTase functionalises apo-ACPs with a PP arm, derived from Coenzyme A, at a conserved serine residue situated on the solvent exposed elbow of ACP helix 2, forming holo-ACP, and it is the terminal sulfhydryl of the PP arm which is acylated^265. Whilst a PPTase is obligatory for polyketide biosynthesis it is not considered a member of the mPKS. Type II PKS BGCs rarely encode dedicated PPTases^264, and to the most part type II PKS ACPs are functionalised by endogenous FAS associated holo-acyl carrier protein synthases^92.
The phosphopanthethiene (PP) arm of Coenzyme A is transferred to a conserved serine residue of apo-ACP (Ser42 for the archetypal actinorhodin (act) ACP), by a PPTase forming holo-ACP. The holo-ACP is then acylated by a malonyl CoA:ACP transacylase, or by self-malonylation\textsuperscript{266}.

Type II PKS ACPs have a remarkable number of protein interaction partners, much like the primary metabolic FAS AcpP which must deliver growing fatty acid chains to FabHFBGAZ and I, as well as interact with a holo-ACP synthase, and FabD, a malonyl-CoA:acyl carrier protein transacylase (MCAT)\textsuperscript{267}. FabD catalyses the transacylation of malonyl-moieties from malonyl-CoA to holo-ACPs forming acyl-ACP (Figure 1.13) during fatty acid biosynthesis. The acyl-ACP serves as a substrate for fatty acid biosynthesis, facilitating elongation of the fatty acid chain, as the acyl-CoA analogue cannot be utilised directly\textsuperscript{267}. Like FASs, type II PKSs also require acyl-ACP substrates; however there is some debate over the involvement of an MCAT in acyl-transfer\textsuperscript{266}. Historically, it was thought that MCATs were essential for ACP acylation\textsuperscript{268}; however more recently completely synthetic ACPs have demonstrated self-malonylation properties\textsuperscript{266}, and polyketide shunt metabolites have been observed as a product of holo-ACP, KS/CLF and malonyl-CoA in vitro assays\textsuperscript{269}. The self-acylation capacity of type II PKS associated ACPs does not routinely extend to ACP counterparts involved in fatty acid synthesis: S. coelicolor FAS ACP cannot self-acylate\textsuperscript{270}, however type II FAS ACPs from Brassica napus (commonly known as
rapeseed) and *P. falciparum* buck this trend\(^{271}\). Ambiguity in ACP acylation patterns often necessitates the co-expression, or addition, of MCATs to *in vivo* and *in vitro* experiments to ensure the availability of acyl-ACPs\(^{272,273}\). This ambiguity also transcends the mPKS definition, and MCATs are often cited as mPKS components\(^{198}\). Interestingly, some type II PKS BGCs encode dedicated acyl-CoA:ACP transacylases e.g. *rem*\(^{246}\), auracin\(^{274}\) and gilvocarcin\(^{275}\) BGCs, where citation of MCATs as mPKS components may be warranted. The complexities of BGC associated MCATs and definition as mPKS component, or not, necessitates empirical testing\(^{276}\). It is fair to say the minimal complement of genes necessary for biosynthesis of the nascent polyketide chain includes the KS/CLF heterodimer, and all genes necessary for formation of acyl-ACP. The true description of these key enzymes differs between *in vivo* and *in vitro* experiments, and also between type II PKS BGCs.

### 1.4.3.2 Primary tailoring enzymes

The full length nascent polyketide chain synthesised by mPKS components is highly reactive and in the absence of any additional tailoring enzymes, and will readily form shunt metabolites through intramolecular aldol condensations\(^{210}\). Accordingly the polyketide chain is selectively modified by a suite of primary tailoring enzymes i.e. tailoring enzymes which modify an ACP linked polyketide, which catalyse regiospecific cyclisation, aromatisation and ketoreduction, forming the aromatic structures characteristic of type II polyketides\(^{210}\) (Figure 1.15).

The first octaketide shunt metabolites produced by the actinorhodin (*act*) mPKS are SEK4 and SEK4b, formed through C7-C12 and C10-C15 intramolecular aldol condensation reactions, respectively\(^{277}\) (Figure 1.14). Interestingly, SEK4 is more abundant than SEK4b, and the cyclisation pattern of SEK4 matches that of actinorhodin implicating the KS/CLF complex in first ring closure\(^{262}\). It has been proposed that during polyketide biosynthesis the nascent polyketide chain must buckle to fit into the KS/CLF cavity. Through buckling, the nascent polyketide chain is oriented correctly for C7-C12 cyclisation\(^{244}\). The *act* C9 ketoreductase, ActIII, has also been implicated in first ring cyclisation, where the nascent polyketide chain, delivered to ActIII tethered to the *act* ACP, exclusively produces the C7-C12 cyclised shunt metabolite mutactin\(^{278,279}\) (Figure 1.14). C7-C12 polyketide cyclisation is not always directed by the KS/CLF or C9 KR, and dedicated C7-C12 cyclases exist in unreduced systems e.g. R1128 (Zhul)\(^{280}\), mithramycin (MtmQ)\(^{281}\) and steffimycin (StfQ)\(^{282}\).

The first modification of the polyketide chain which is not thought to be catalysed by the KS/CLF heterodimer is often the regiospecific reduction of the C9 carbonyl to a hydroxyl group, and is catalysed by a cluster associated ketoreductase. C9 ketoreduction is not ubiquitous, however, and in unreduced systems the nascent polyketide chain is delivered directly to a suite of cyclases\(^{283,284}\).
Differential cyclisation of the polyketide chain is primarily the first non-congruent biosynthetic step during type II PKS biosynthesis. Regiospecific cyclisation commits polyketide biosynthesis towards one of several different chemotypes (Figure 1.15), which are further diversified to a bewildering array of chemically complex specialised metabolites\(^2\)\(^\text{10}\).

The linear cyclised polyketides (or U-shaped polyketides) constitute the most well studied aromatic polyketides\(^3\)\(^3\),\(^2\)\(^8\)\(^5\), and are typically substituted naphthalenes, naphacenes or anthracenes (Figure 1.15). Based upon the regiospecific cyclisation of the first and last rings, and polyketide chain length, linear aromatic polyketides can be classified widely into four chemotypes: tetracyclines, anthracyclines, tetracenomycins, benzoisochromanequinones (BIQs)\(^2\)\(^8\)\(^6\)\). The BIQs, anthracyclines and tetracyclines all follow congruent C7-C12 first ring cyclisation (Figure 1.15, Table 1.1) however, unlike the latter two, BIQs only comprise three rings resultant from a shorter C\(_{16}\) chain length, the last of which is lactonised. Closure of the third ring during BIQ biosynthesis is preceded by sterospecific ketoreduction of the C3 carbonyl group resulting in formation of either (S)- or (R)-DNPA, intermediates of actinorhodin and granaticin biosynthesis\(^2\)\(^4\)\(^5\),\(^2\)\(^8\)\(^7\), respectively. As well as first ring cyclisation, second and third ring cyclisation of the anthracyclines and tetracyclines is also congruent: these two chemotypes differ in cyclisation at the fourth ring. Intriguingly the mechanism underpinning formation of the fourth ring in oxytetracycline biosynthesis remains unknown. Whilst an acyl-CoA ligase has been implicated in fourth ring formation during biosynthesis of the tetracycline SF\(_{257}\)\(^5\), the acyl-CoA ligase homologue from the oxy BGC, OxyH, was shown to be expendable\(^4\)\(^5\). Differing to other U-shaped linear polyketides, first ring closure during tetracenomycin biosynthesis occurs between C9 and C14 and is catalysed by TcmN\(^2\)\(^8\)\(^0\). TcmN also facilitates closure of ring 2, C7-C16, however the mechanism of third ring closure, C5-C18 remains unknown. Tcml, is a dedicated C2-C19 fourth ring cyclases and together with TcmN results in formation of the carboxylated naphthacene TcmF\(^1\)\(^2\)\(^9\)\(^0\) (Figure 1.15).
Figure 1.15: Aromatic polyketide chemotypes, directed through cyclisation and aromatisation

Table 1.1: Cyclisation patterns for different aromatic polyketide chemotypes

<table>
<thead>
<tr>
<th>Polyketide chemotype</th>
<th>Cyclisation pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tetracyclines and aureolic acids</td>
<td>C7-C12, C5-C14, C3-C16, C1-C18</td>
</tr>
<tr>
<td>Anthracyclines</td>
<td>C7-C12, C5-C14, C3-C16, C2-C19</td>
</tr>
<tr>
<td>Tetracenomycin</td>
<td>C9-C14, C7-C16, C5-C18, C2-C19</td>
</tr>
<tr>
<td>Benzoisochromanquinones</td>
<td>C7-C12, C5-C14, O3-C15/O15-C3</td>
</tr>
<tr>
<td>Angucycline</td>
<td>C7-C12, C5-C14, C4-C17, C2-C19</td>
</tr>
</tbody>
</table>

Interestingly, polyketide chain length influences cyclisation pattern and polyketides ≥C_{24} in length predominantly form J-shaped aromatic chemotypes, opposed to U-shaped metabolites. This phenomenon is exemplified by pentangular polyphenols, however decaketide angucyclines also adopt an angular ring topology forming benz[a]anthracene cores (Figure 1.15). Remarkably the cyclisation patterns of anthracyclines and angucyclines differ only in the closure of the third ring (Table 1.1). Despite this redirection of aromatic polyketide biosynthesis through substitution of a single regiospecific cyclase e.g. third ring C3-C16 anthracycline and C4-C17 angucycline cyclase has shown mixed successes.
Substitution of an anthracycline second/third ring cyclase from the doxorubicin/daunorubicin BGC with JadI, a C4-C17 angucycline specific cyclase from the Jadomycin B BGC, failed to derail U-shaped polyketide biosynthesis to J-shaped polyketides. Conversely, substitution of SnoaM, a cyclase from the anthracycline nogalamycin BGC, with the C4-C17 cyclase PgaF, resulted in formation of UWM6, a angucyclic rabelomycin intermediate in Streptomyces lividans TK24. To access an alternative polyketide chemotype it may be fruitful to substitute the entire complement of enzymes required for cyclisation opposed to substitution of single regiospecific cyclases. There is a growing precedence for cyclases and other PKS tailoring enzymes to form complexes where substitution, or removal, of one part abolishes the function of all. This is thought to be the case during biosynthesis of the antifungal and antiviral angular dodecaketide pradimicin (pdm). Whilst the pdm mPKS and PdmD catalyse biosynthesis of a nascent polyketide chain and cyclisation of the first and second rings (C9-C14 and C7-C16, Figure 1.15), two additional cyclases (PdmK and PdmL) and a monooxygenase (PdmH) work synergistically to oxidise the second and close the third and fourth rings. In the absence of either PdmK, L or H, only a bicyclic intermediate shunt metabolite is formed. A complex comprising cyclases is also thought to form during the biosynthesis of the S-shaped polyketide resistomycin (rem). Heterologous expression of the rem mPKS and proposed first ring cyclase, remI, failed to produce polyketide in S. lividans TK23. Rather, to observe polyketide biosynthesis the mPKS and a combination of two additional cyclases, remI and remF or remI and remL, were necessary resulting in formation of RM80 and SEK15, and tetracenomycin shunt metabolites, respectively. Furthermore, the full complement of rem cyclases was necessary to form the S-shaped polyketide leading the conclusion that the mPKS components and RemI, L and K form a cage-like multienzyme complex into which the polyketide chain grows and is cyclised.

Substitution of regiospecific cyclases and selective C9 ketoreduction represent the first instances at which diversity can be introduced into fully formed nascent polyketide chains. The corresponding aglycone cores can be subsequently glycosylated, acylated, halogenated, oxidised and reduced etc. to form a myriad of specialised metabolites with ever increasing chemical complexities.

1.4.3.3 Unknowns in type II PKS biosynthesis
As evidenced in 1.4.3.2 the complexities of aromatic polyketide biosynthesis remain enigmatic even for the most well characterised systems. For instance, much is still to be learned about biosynthesis of the benzioisochromanequinone (BIQ) actinorhodin, despite cloning of the act BGC more than 3 decades ago. Recent advances have elucidated the role of ActIV, the second ring cyclases, as a bifunctional thioesterase leading to characterisation of a key bicyclic intermediate at the branch point before the first committed step to either actinorhodin or another BIQ, granaticin, as well as the molecular basis for the first regio- and stereospecific reduction of the nascent polyketide chain by ActIII, the act C9 ketoreductase, and the characterisation of completely new act shunt metabolites e.g. GTRI-02. Yet a number of questions still remain unanswered e.g. the enzymatic conversion of the penultimate monomeric act intermediate to the final dimeric BIQ structure is yet to be fully characterised.

46
Despite these, and other, knowledge gaps act and other PKS constituent parts have been routinely, and successfully, used in combinatorial engineering of aromatic polyketides\textsuperscript{256,300}.

### 1.4.3.4 Heterologous expression of type II PKSs in Escherichia coli

Heterologous expression experiments have facilitated the recombineering and exchange of polyketide synthase components in genetically amenable hosts\textsuperscript{213,221,292,301-308}, which is not always possible in native producers. Reconstruction of biosynthetic pathways in an alternative host enables expansion of the chemical diversity of specialised metabolites\textsuperscript{309,310}, derailment biosynthesis towards useful end products, opposed to by-products\textsuperscript{182} (discussed in Chapter 2), as well as characterisation of BGCs in a well described host background. Moreover, this strategy introduces a degree of orthogonality in gene expression i.e. if the heterologous expression host is phylogenetically distant to the native host redundant regulatory elements are less likely to functionally cross-talk\textsuperscript{311}. In attempts to satisfy the above, all polyketide synthases classes have been successfully reconstituted in the genetically amenable workhorse organism \textit{Escherichia coli} \textsuperscript{312}, with the exception of type II PKSs. Attempts to express KS and CLF genes in \textit{E. coli} has either resulted in 100% inclusion body formation of recombinant proteins\textsuperscript{221}, or no observable gene expression\textsuperscript{313}. Literature evidence is scant as a result of the poor success, but the cause of KS/CLF insolubility is thought to arise from disharmonious rates of protein folding and dimerisation\textsuperscript{221}. Strategies to solubilise KS/CLFs through addition of a short peptide linking each monomeric unit failed\textsuperscript{221}. Moreover, addition of dimerising peptides to the N- or C-terminus of each protein monomer failed, also. Interestingly, over expression of the \textit{E. coli} alternative sigma factor, $\sigma^{54}$, in a host harbouring the entire oxytetracycline BGC from \textit{Streptomyces rimosus} did produce detectable levels of polyketides\textsuperscript{313}, despite no perturbations to the nucleic acid sequence. Fortuitously each \textit{oxy} operon was shown to be preceded by a $\sigma^{54}$ promoter sequence, which are missing from the archetypal actinorhodin BGC, and consequently why identification of actinorhodin using the same approach might have failed. The KS and CLF (OxyA/B) could not be observed as recombinant proteins, however. The authors suggest gene expression using a T7 system, typically adopted for heterologous expression experiments, may be too harsh and gentler gene expression systems are favourable when working with fastidious proteins\textsuperscript{314}. The utility of this method remains to be seen, and the \textit{oxy} cluster is the only archetypal type II PKS to show functionality in \textit{E. coli} to date\textsuperscript{314}. The insolubility of type II PKS componentry is largely limited to the KS/CLF, and ketoreductases, cyclases, cyclase/aromatases etc. have been shown to be widely soluble in \textit{E. coli}\textsuperscript{272}. Complications at step 1 during type II PKS biosynthesis has inhibited the high throughput combinatorial biosynthesis strategies for metabolite diversification observed for other polyketide classes, routinely exemplified by erythromycin\textsuperscript{315} (Chapter 2). Type III PKSs can be engineered to produce unreduced $\geq \text{C}_{18}$ polyketide chains\textsuperscript{227} and may be an alternative to type II PKS machinery, however the absence of ACPs in these systems negates directionality of the complex regiospecific cyclisation reactions facilitated by type II PKS cyclases.

To circumvent the insolubility of type II PKS KS/CLF heterodimers a fungal non-reducing PKS has been engineered to produce aromatic polyketides in \textit{E. coli}\textsuperscript{316}. The PKS4 from \textit{Gibberella}
*Gibberella fujikuroi* was dissected to constituent MCAT, KS and ACP ‘parts’ and, in the presence type II PKS cyclases, produces C7-C12 cyclised PK8 and C9-C14 cyclised nonaSEK4 polyketides, and with the addition of ActIII, and cyclases from the griseusin BGC and oxy BGC produced the anthraquinone SEK26 (Figure 1.16). The use of non-reducing fungal PKSs has drawbacks, however: 1) Initiation modules enabling use of vibrant polyketide starter units cannot be routinely substituted, and 2) the chain length of the polyketide cannot be routinely altered. In both cases a new fungal megasynthase needs to be reengineered, whereas performing these alterations using type II PKSs are more trivial (see Chapter 6).

**Figure 1.16: Biosynthesis of aromatic polyketide in E. coli by PKS4**


The ability to recapitulate flexible biosynthetic pathways to aromatic / type II polyketides in *E. coli* is a largely unmet need. Through identification of *E. coli* soluble KS/CLF machinery one would be able to direct biosynthesis of aromatic polyketide to clinically important specialised metabolites e.g. tetracyclines and doxorubicin derivatives, engineer biosynthetic routes to clinically useful semi-synthetic analogues (Figure 1.2), reconstruct unknown biosynthetic routes to specialised metabolites important specialised metabolites *de novo* e.g. hypericin, and synthesis libraries of novel chemical entities in a high throughput manner, not currently achievable using Actinobacterial hosts. The generation of a plug-and-play scaffold in *E. coli* would represent an important step towards biosynthesis of aromatic polyketides outside of the Actinobacteria, enabling manipulation of these important biosynthetic pathways by almost any molecular biologist.
1.5 Aims of the thesis

Broadly, the aims of this thesis are to develop a suite of tools to expedite discovery of novel acting bioactive aromatic polyketides in a manner fitting with synthetic biology tenets by unblocking significant bottlenecks that exist in the field today, demonstrated in Figure 1.17.

1) The first bottleneck we aim to tackle is the issue of incongruence when reporting identification and characterisation of BGCs in manuscripts. We will design a community supported minimum standard of information which must be fulfilled when reporting a BGC. Furthermore, we aim to comb the literature to define a heritage dataset of previously characterised BGCs which we will recharacterize to conform with the minimum information about a biosynthetic gene cluster and use this data to populate a publicly available repository. This repository will comprise a comprehensive, highly accurate, and well curated dataset of hundreds of bioactive specialised metabolites and their corresponding BGCs which is easy to navigate and readily accessible (Chapter 3). We hope this standard will bring the science of specialised metabolites in line with the tenets of synthetic biology (Figure 1.17, Step 1).

2) Secondly, we aim to construct a prioritisation platform which enables one to rank the thousands of BGCs predicted *in silico* based upon key molecular biology relevant, and discovery pertinent, parameters. This prioritisation tool will reduce the bottlenecks in experimental characterisation of specialised metabolites by matching appropriate BGCs with appropriate heterologous expression hosts, thereby avoiding the frequent failures of molecular biological experiments resulting from unforeseen recombination events, or poor gene expression, etc. Furthermore, we will effectively incorporate a dereplication parameter to help prevent re-characterisation of known specialised metabolites. Collectively this software aims to expedite successful characterisation of specialised metabolite by providing insight into molecular biological characteristics of thousands of BGCs concurrently (Chapter 4) (Figure 1.17, Steps 2, 3 and 4).

3) Thirdly, using the tools developed in Chapters 3 and 4 we aim to migrate the biosynthesis of aromatic polyketides from Actinobacteria to the more genetically amenable host *Escherichia coli* (Chapters 5, 6 and 7). The biosynthesis of aromatic polyketides is largely restricted to the Actinobacteria resulting from the absolute insolubility of the KS/CLF heterodimer in biotechnologically virtuous hosts e.g. *E. coli* and the poor genetically tractability and complex life-cycles of Actinobacteria have been cited as two prominent inhibitors of effective high throughput approaches to combinatorial biosynthesis of this class of clinically important specialised metabolites. Using the well curated dataset generate from aim 1 we hope to identify, characterise (Chapter 5) and engineer (Chapter 6) a soluble KS/CLF heterodimer in *E. coli* for biosynthesis of aromatic polyketides. Furthermore, we aim to use this soluble machinery to build a ‘plug-and-play’ genetic platform enabling one to substitute and add biosynthetic ‘parts’ derived from all kingdoms of life to generate chemically diverse aromatic polyketides using *E. coli* (Figure 1.17, Step 7).
Additionally, using multiple bioinformatics approaches we aim to debunk the literature dogma which states type II polyketide synthases are predominantly identified in Actinobacteria and aim to evaluate the solubility of non-actinobacterial KS/CLF heterodimers in *E. coli* (Chapter 7). In doing we will catalogue a suite of biosynthetic ‘parts’ which could also be ‘plugged’ into our ‘plug-and-play’ scaffold to access a wealth of known clinically relevant aromatic polyketides and new chemical entities (Figure 1.17, Step 5).

In summary, this work seeks to expedite discovery and design of bioactive specialised metabolites by setting a community standard when reporting BGC metadata (Chapter 3), and by developing a tool for efficient prioritisation of BGCs of interest based upon key molecular biology relevant parameters (Chapter 4). Using these tools we aim to discover (Chapters 5 and 7), design, engineer (Chapter 6) and construct a plug-and-play platform for aromatic polyketide biosynthesis in *Escherichia coli* that is amenable to high throughput combinatorial biosynthesis. We hope this work will set a new precedence across the wide-reaching field of specialised metabolite biosynthesis, with a specific focus on aromatic polyketide biosynthesis.
Figure 1.17: A specialised metabolite discovery pipeline

Schematic of a synthetic biology led ‘genes-to-molecules’ approach to specialised metabolite biosynthesis. The input is all available information deposited across various media and outputs are new chemical entities or more efficient routes to known specialised metabolites. Step 1 is creation of a synthetic biology syntax and standard when reporting BGCs, as well as curation of information in one centralised repository. Step 2 represents the ability to rank/prioritise putative, predicted and characterised BGCs based upon relevant parameters, Steps 3 and 4 considers step 2 in high throughput (HTP) for the generation of refined and directed datasets of BGCs. Steps 5 through 7 consider all factors pertaining to the cyclical design-built-test-learn synthetic biology cycle. Red arrows represent current bottlenecks in the specialised metabolite discovery pipeline. Steps 5 and 7 are specific to biosynthesis of aromatic polyketides where the test cycle cannot be achieved in HTP due to heterologous host restrictions / lack of soluble parts in genetically amenable organisms.
1.6 References


2 Steps towards the synthetic biology of polyketide biosynthesis

Matthew Cummings¹, Rainer Breitling, and Eriko Takano*

Faculty of Life Sciences, Manchester Institute of Biotechnology, The University of Manchester, Manchester, UK.

*Corresponding author: Eriko Takano, Manchester Institute of Biotechnology, The University of Manchester, 131 Princess Street, Manchester, M1 7DN, UK.

Tel: +44 (0)161 306 4419.

E-mail: eriko.takano@manchester.ac.uk.

Keywords: Refactoring, plug-and-play biology, combinatorial biosynthesis, specialised metabolites, drug discovery

Running title: Synthetic biology of polyketides
2.1 Preface

‘Steps towards the synthetic biology of polyketide biosynthesis’ was published in FEMS Microbiology Letters. The draft manuscript was received on the 26th of November 2013, revised on the 16th of December 2013, accepted on the 17th of December 2013, and published online on the 7th of January 2014.

The first draft of the manuscript was authored by M.C., and figures were designed and drawn by M.C.. E.T. and R.B. provided useful comments and discussion on the manuscript and helped with the submission process. The published manuscript has been formatted in a style consistent throughout the thesis, opposed to FEMS Microbiology Letters formatting. Position of figures within the text is as consistent as possible with the published manuscript.

Citations as of submission date: 20
2.2 Abstract

Nature is providing a bountiful pool of valuable specialised metabolites, many of which possess therapeutic properties. However, the discovery of new bioactive specialised metabolites is slowing down, at a time when the rise of multidrug resistant pathogens and the realization of acute and long-term side effects of widely used drugs lead to an urgent need for new therapeutic agents. Approaches, such as synthetic biology, are promising to deliver a much-needed boost to specialised metabolite drug development through plug-and-play optimized hosts and refactoring novel or cryptic bacterial gene clusters. Here, we discuss this prospect focusing on one comprehensively studied class of clinically relevant bioactive molecules, the polyketides. Extensive efforts towards optimization and derivatisation of compounds via combinatorial biosynthesis and classical engineering have elucidated the modularity, flexibility and promiscuity of polyketide biosynthetic enzymes. Hence, a synthetic biology approach can build upon a solid basis of guidelines and principles, while providing a new perspective towards the discovery and generation of novel and new-to-nature compounds. We discuss the lessons learned from the classical engineering of polyketide synthases and indicate their importance when attempting to engineer biosynthetic pathways using synthetic biology approaches for the introduction of novelty and overexpression of products in a controllable manner.
2.3 Steps towards the synthetic biology of polyketide biosynthesis

2.3.1 Polyketides: magnificently modular

Polyketides represent an important class of compounds that are extremely diverse in structure and function. Natural screening strategies have brought more than 20 drugs to market, including the immunosuppressants FK506 and rapamycin\(^1\), hypcholesteroleemics, such as lovastatin\(^3\); anticancer agents, such as doxorubicin\(^4\); and a host of antimicrobials, including tetracycline and erythromycin\(^5\). Furthermore, it is predicted that more than 1% of polyketides described have potential drug activity\(^7\), and as a result of this polyketides have received tremendous attention in efforts to unearth new compounds with bioactive properties. Interest in the discovery of novel acting polyketides has been renewed with the recent surge in microbial genome sequences and the availability of accurate genome mining software to detect a previously unexpected abundance of uncharacterized specialised metabolite biosynthesis gene clusters (BGCs).

Advances in sequence analysis software such as antiSMASH are providing a facility to screen for BGCs in an automated computational fashion\(^8\)-\(^10\). They are particularly powerful in detecting polyketide BGCs, as these are defined by the presence of highly characteristic signature genes and motifs. Identification of putative BGCs using sequence-based analysis is also enabling the discovery of compounds that are cryptic, i.e. not expressed under laboratory conditions.

In addition to the clinical relevance and abundance of polyketides, there is one other reason behind the particular interest in polyketides as promising targets for synthetic biology: the highly modular architecture of both the BGCs and the constituent polyketide synthases (PKS) presents an ideal starting point from which to engineer chemical novelty.

The biosynthesis of polyketides is modular at many levels. First, the genes responsible for polyketide biosynthesis are typically clustered in the genome\(^11\), forming a BGC. Each BGC encodes the PKS responsible for the formation of the carbon backbone, together with the tailoring enzymes required for primary tailoring events, e.g. cyclization of the \(\beta\)-keto-acyl carbon chain, subsequent tailoring events to form the final polyketide structure as well as genes encoding the regulation of the BGC and resistance to the end product if applicable, e.g. in the case of antibiotic end products. Once transcribed and translated, the PKS enzymes themselves are also modular in nature. The best-characterized PKS, 6-deoxyerythronolide (6-dEB) synthase from *Saccharopolyspora erythraea*\(^12\), represents a good example of this. 6-dEB synthase consist of three megasynthases encoded by three ORFs, DEBS1–3 (Figure 2.1a). Each of these megasynthases comprises a series of modules responsible for the extension of the polyketide carbon backbone through addition and selective reduction of one acyl-CoA monomer to form an acyl intermediate. In addition to this, each module can be dissected further still into a series of domains. Each of these domains is unequivocally linked with one specific catalytic function required for chain extension and modification. Some domains are obligatory for recruitment of the acyl-CoA monomer and chain extension, e.g. acetyltransferase (AT), acyl carrier protein (ACP) and ketosynthase (KS), while others are accessory domains involved in the selective reduction of the \(\beta\)-keto-acyl intermediate to the corresponding alcohol, olefin or
methylene group catalysed by ketoreductase (KR), dehydratase (DH) and enoyl-reductase (ER) activity, respectively. Importantly, all of the modules encoded within DEBS1–3 are required for successful synthesis of 6-deoxyerythronolide B and act in succession, like a giant molecular assembly line. Because each domain is unequivocally linked with one specific catalytic function and polyketides are synthesized in a collinear fashion, addition, removal and/or substitution of these domains or modules will theoretically result in defined alterations of the end product (Figure 2.1b). Furthermore, the collinear architecture of these domains, and motifs within, can allow prediction of the structure of the polyketide and important elements of its stereochemistry from analysis of its coding sequence. With these rules in mind, theoretically, we have the potential to engineer rationally a desirable predefined polyketide end product if domains or modules can be stitched together like molecular lego bricks.
Figure 2.1. Pictorial illustration of 6-dEB synthase, a modular type I PKS, and successful attempts at engineering this megasynthase.

A) The native biosynthetic gene cluster and end product. B) Summary of engineered cluster variants and their products; alterations are indicated in red. Manipulation of the polyketide scaffold includes: 1) substitution of domains; 2) feeding with non-canonical substrates; 3) domain insertion; 4) inactivation of domains; and 5) domain deletions. The effects of modifications 1–5 to the 6-deoxyerythronolide B scaffold are also indicated in red, as are the positions at which engineered post-PKS tailoring modifications can occur. Abbreviations: AT; acetyltransferase domain, ACP; acyl carrier protein domain, KS ketosynthase domain, ER; enoyl reductase domain, DH; dehydratase domain, PKS; polyketide synthase.

With the increase in the number of characterized PKSs, it is becoming apparent that the collinear relationship between gene structure and chemical end product is not absolute; however, as a general rule, collinearity presents an ideal template for engineering the polyketide biosynthetic machinery. Consequently, manipulating polyketide assembly through domain alteration has been one major avenue that classical engineering has explored in order to derivatize known polyketides even before the current era of synthetic biology (Fig. 1B). However, not all PKSs show the ‘one domain–one reaction step’ modular organization that is seen in type I PKSs. Chain extension can also occur iteratively through a recursive approach, where domains that are part of a single polypeptide are used repeatedly. This is the case for type II and type III PKSs, as well as for many type I fungal PKSs. Although differences occur in enzymatic organization of PKSs, the underlying chemistry behind chain extension, through successive decarboxylative Claisen condensation of acyl-CoA monomers to form a β-keto-acyl intermediate and modification in cis or trans, remains the same for all. As such, all PKS are in principle amenable to engineering. Detailed reviews of the underlying biochemistry are available.

The immense diversity in the chemical structures of polyketides is the result of continuous evolutionary pressure for the development of chemical novelty facilitated by the modular nature of the PKS. On an evolutionary scale, diversity is introduced into polyketides through both simple mutations within domains as well as frequent horizontal co-transfer of genes between clusters. Evolutionary analysis reveals conserved synteny between gene clusters responsible for the biosynthesis of homologous products, as well as products of considerable structural difference and those in between. Transfer of gene units between BGCs, permitted by their inherent modularity and collinearity, generates a continuous interspecific flow of compounds with novel physicochemical properties, not only as polyketides, but also for the generation of hybrid products containing additional non-ribosomal peptide moieties. The recently described BGC encoding the biosynthesis of three zeamine-related antibiotics in the *Serratia plymuthica* RVH1 genome provides a good example of the plasticity of BGCs and their ability to co-transfer between organisms. This BGC comprises genes for five PKSs, three non-ribosomal peptide synthetases and one mixed fatty acid synthase (FAS)/PKS enzyme, which are required for the synthesis of the hybrid product backbones, as well as additional tailoring genes. Hybrid products, such as these, elucidate the tolerance of synthases to integrate non-canonical substrates from different biosynthetic systems into the growing carbon backbone successfully.
and are naturally occurring versions of domain alteration attempts paralleled in the laboratory-based engineering of PKSs.

Generation of novelty through exchange of domains between BGCs polished under evolutionary selection pressures, as above, invariably results in successful product assembly – as millions of failed “experiments” are rapidly discarded by natural selection. This process cannot be replicated easily in vitro, and simple domain substitutions between BGCs commonly result in the failure of product release and maturation37. Failure of product biosynthesis is regularly the result of the inflexibility of downstream enzymes to tolerate novel substrates. Without additional engineering, most domains incorporate only one substrate into the growing polyketide backbone and show little flexibility to introduce non-canonical substrates. Lessons learned from reprogramming PKSs using classical molecular biology approaches, detailed below, are supporting this general observation, but also, more interestingly, are revealing exceptions. This has provided an instruction manual that exemplifies the scope and limitations of plasticity of polyketide synthases to tolerate the integration of exogenous extenders into the growing β-keto-acyl chain.

Diversification of polyketides can occur at four steps throughout biosynthesis resulting from: 1) the choice of building blocks and chain length, 2) the extent of reduction and stereochemistry of β-keto-intermediates15, primary cyclisation, alkylation and branching, 3) rearrangements and specialised cyclisation and 4) post-polyketide tailoring: glycosylation, oxygenation etc. In the following discussion we focus on these events as two main phases of polyketide synthesis: core scaffold biosynthesis (steps 1–3) and subsequent or concurrent tailoring events (step 4) (Figure 2.2).

2.3.2 Modularity of scaffold biosynthesis

2.3.2.1 Initiation of biosynthesis

The composition of the polyketide backbone, or scaffold, structure is governed by the stringency of acetyltransferase (AT) domains to load a specific acyl-CoA substrate, but also through substrate stereochemistry and redox pattern38: each PKS assembles an individual product through the choice of acyl-CoA units, their level of reduction, and subsequent tailoring. Initiation of scaffold biosynthesis requires selection and recruitment of a starter unit onto a didomain, comprising an AT and an ACP, collectively termed the loading module. The resulting initial starter unit serves as the first substrate in the growth of the final β-keto-acyl chain. Generation of diversity through the promiscuity of AT domains to load multiple different starter units, termed poly-specificity, is more commonly observed than by poly-specificity of extender modules later in biosynthesis39,40. Introduction of diversity during initiation of biosynthesis also commonly occurs through the multiple different priming mechanisms used by the array of loading modules available34,41. Due to the mechanistic promiscuity of the starter domains, combinatorial biosynthesis attempts to manipulate PKS modules often start here. For example: the AT and ACP loading module of DEBS1 naturally recruits a propionate starter unit. Substitution with loading modules from tylosin and oleandomycin type I megasynthases from Streptomyces
fradiae and Streptomyces antibioticus, respectively, resulted in controlled integration of propionate or acetate as a starter unit42. Similarly, the replacement of the isobutryl-CoA–specific loading module initiating avermectin biosynthesis in Streptomyces avermitilis M1 by the unique phoslactomycin polyketide cyclohexanecarboxylic unit loading module from Streptomyces platensis resulted in production of the veterinary antiparasitic doremactin43. Alteration of loading modules for the initiation of biosynthesis is therefore one step showing promise for the generation of novel polyketides.

Figure 2.2. Schematic of the four steps of polyketide biosynthesis encoded by a prototypical polyketide biosynthetic gene cluster.

Each of these steps offers the potential for end product diversification by evolution or engineering as described in the text.

2.3.2.2 Chain extension

After initiation, continued assembly of the polyketide scaffold requires loading of extender units onto the AT and ACP and incorporation into the β-keto-acyl intermediate by the ketosynthase (KS). At this stage diversity can be introduced through the installation of non-canonical extender units resulting from the poly-specificity of loading domains, domain substitutions or by the iterative action of an otherwise modular PKS44. The collection of commonly used extender units nature provides is modest: canonical extender units mostly comprise malonyl- and methylmalonyl-CoA. Substitution for domains loading other, less commonly used, extender units will allow introduction of a broadened chemistry into the polyketide backbone. For example, reductive carboxylation of α,β-unsaturated acyl-CoA precursors via crotonyl-CoA reductase/carboxylase homologues facilitates inclusion of hexyl-, propyl-, chloroethyl- and
isobutylmalonyl-CoA into the polyketide scaffold\textsuperscript{45-47}. Alternatively, functionalization of extender units on stand-alone ACPs allows the incorporation of allyl-\textsuperscript{48}, amino-, hydroxyl-\textsuperscript{49} and methoxymalonyl-ACP\textsuperscript{50} extender units into the polyketide scaffold.

Predicting the poly-specificity of extender modules to introduce these rarer extender units is not straightforward. The mechanical processing and discrimination between acyl-CoA extender units by loading domains in type I modular PKSs is currently little understood. Investigations to elucidate why particular substrates are preferred or chosen are presenting a growing body of evidence suggesting that PKSs may be able to tolerate and incorporate exogenous natural and non-natural extender units into the $\beta$-keto-acyl chain. For example, analysis of the acyl-CoA substrate selectivity of PikAIV, a pikromycin synthase from \textit{Streptomyces venezuelae}, elucidated the poly-specificity of extender modules towards substrates not readily present in the producer. PikAIV successfully loaded malonyl-, propionyl-, ethyl- and native methylmalonyl-CoA to the ACP. In the case of malonyl- and propionyl-CoA, active site occupancy was low at 3\% and 19\%, respectively. More interestingly, the rare extender ethylmalonyl-CoA showed AT loading of 90\% and low levels of hydrolytic release indicating its potential for incorporation during assembly; the native substrate methylmalonyl-CoA showed 100\% AT saturation\textsuperscript{51}. In the case of PikAIV all acyl-CoA substrates were loaded; however, incorporation into the carbon chain depended upon the rate of subsequent hydrolytic release. These findings suggest that extender modules may show a greater tolerance to incorporate exogenous precursors lacking evolved selectivity, consistent with findings previously reported\textsuperscript{52}. Substituting extender domains for the addition of novel extender units showing limited hydrolytic release could, therefore, result in generation of novel polyketides; however, no concrete rules defining what properties extender modules require to do so have been elucidated, despite observed discrimination between sizes of extender units and incorporation.

\textbf{2.3.2.3 Product release}

Manipulating starter and extender modules of PKSs may permit the introduction of novel acyl-CoA substrates into the polyketide scaffold. However, for these to show activity they must reach maturation and be released from the PKS. For successful release, it is important to identify which catalytic domains act as decision gates, thereby permitting continuation of downstream biosynthesis of altered $\beta$-keto-acyl intermediates. Elucidating such points will significantly aid success when engineering BGCs. Yeh \textit{et al.} (2013)\textsuperscript{53} experimentally indicated that the phylogeny between non-reducing iterative PKS (nrPKS) modules is a good predictor of successful polyketide assembly and release from engineered BGCs. Increased phylogenetic proximity between gene units translated to improved domain–domain interactions, and as a result improved the release of the polyketide end product. In contrast, Xu \textit{et al.}, (2013)\textsuperscript{37} found that the best predictors of thioesterase (TE) acceptance, and therefore release, are the shape and size of the polyketide substrate, and consequently indicate the stringency of TE domains in carrying out discriminative decision gate functions. For example, if the native substrate of a TE was a nonaketide, but the engineered assembly line presented it with a heptaketide, the rate of release was almost zero\textsuperscript{54}. Substituting TE domains often resulted in abolished product
formation, despite the presence of an abundance of β-keto-acyl intermediates produced by the upstream domains, whereas judicious choices of TE substitution resulted in the successful production of an unnatural polyketide product, radilarin. Successful polyketide release from TE in the case of resorcylic acid lactones and dihydroxyphenylacetate acid lactones may be dependent upon substrate size. However, contrastingly, truncation of the DEBS1–3 megasynthase through relocation of the TE domains downstream of the modular DEBS1 resulted in assembly and successful release of a much shortened triketide lactone. These contrasting results indicated the complex nature of the TE and show the requirement for further work to build rules to predict TE domain tolerance for substrates. Currently, for successful incorporation of novel starter and extender units, and successful product release, analysis of domains must be done on a case-by-case basis.

### 2.3.3 Modularity of tailoring reactions

Introducing diversity within the polyketide scaffold provides the ability to diversify the backbone structure. Further tailoring of these structures generates an additional level of complexity, and pathway engineering over the past decade has generated new-to-nature products through novel glycosylation, acyl-transfer, hydroxylation, epoxidation, alkylation, transamination and desaturation reactions acting on naturally occurring products.

Tailoring enzymes can introduce chemical groups that often are more relevant to engineer for the alteration of specific activity of the polyketide than the backbone construct. 6-dEB is a precursor in the biosynthesis of the macroline antibiotic erythromycin. Biosynthesis of erythromycin requires the action of tailoring enzymes encoded by ORFs located within the BGC encoding DEBS1–3. Without the required glycosylation, hydroxylation and methylation reactions catalysed by tailoring enzymes, 6-dEB cannot become active as erythromycin. This is similarly the case for a group of type II aromatic polyketides with anticancer activities, the anthracyclines. The mechanisms of action of anthracyclines such as doxorubicin are mediated through DNA damage caused by inhibition of DNA topoisomerase II, DNA binding and subsequent alkylation, and intercalation within DNA of the target cells. While the basic aglycone structures comprise 7,8,9,10-tetrahydro-5,12-naphthacenequinones, the observed anticancer activities of anthracyclines are heavily dependent on the attached sugars. Furthermore, alteration of the attached sugars can modify not only activity, but also other parameters, such as toxicity. The clinical applications of doxorubicin are limited by dose-dependent cardiotoxic side effects. Epirubicin, an analogue of doxorubicin, with opposing configuration of a C-4 hydroxyl group on the deoxysugar, shows significantly less cardiotoxicity, while maintaining comparable antitumor properties. Therefore, exchanging the sugars attached to the aglycone scaffold can tune the overall properties of therapeutic polyketides. Derivatisation in this manner could be achieved through the addition of glycosyltransferases into BGCs, as opposed to introducing the sugar moieties semi-synthetically. First steps towards this have been undertaken and are revealing the substrate tolerance of individual tailoring enzymes.

In a parallel to the promiscuity of the scaffold biosynthesis genes, a case study of glycosyltransferases showcases the ability of tailoring enzymes to accept a broader range of
substrates and a tolerance to modifying foreign acceptor molecules. ElmGT from *Streptomyces olivaceus* involved in elloramycin biosynthesis can glycosylate 8-demethyltetracenomycin C with D-mycarose, D-olivose, L-olivose, L-rhodinose, L-rhamnose and a disaccharide comprising two D-olivose moieties, showing extensive tolerance to glycosylate scaffolds with multiple sugars. Other glycosylases show tolerance to introduce a more defined range of sugars to a wider number of acceptor scaffolds. For example, the L-olivosyl glycosyltransferase OleG2 from *Streptomyces antibioticus* involved in oleandomycin production has promiscuity for NDP-L-mycarose, NDP-L-rhamose and the foreign acceptor erythronolide B. The activities of ElmGT and OleG2 show a high tolerance to introduce non-canonical substrates onto aglycone scaffolds. This level of promiscuity for novel substrates and scaffolds is also consistent for EryCIII and EryBV from *S. erythraea* endogenous to the erythromycin BGC and the recombinant UrdGT2 from *Streptomyces fradiae* Tü2717 involved in urdamycin A production, for generation of novel C-glycosylated compounds.

Such flexibility of tailoring modifications may be a result of frequent co-transfer of tailoring genes between clusters, or as a result of multi-functional tailoring of a series of different polyketides within the host; however, in both cases, the ability to introduce moieties to a variety of scaffolds opens up promising prospects for further diversification by synthetic biology. Full characterization and cataloguing of tailoring enzymes in a standardized “biobrick” fashion would ultimately allow a user to pick and choose which modifications are desirable. Integration of tailoring enzymes into plug-and-play hosts as well as in standardized constructs would facilitate the rapid derivatisation of novel compounds. Tailoring in this fashion shows potential to speed up rational design of polyketides at the increased rate that will be necessary, for example, to overcome the rapid and unavoidable emergence of resistance against them in pathogenic bacteria.

### 2.4 Synthetic biology: the future of combinatorial biosynthesis

Combinatorial biosynthesis has successfully exploited the functional collinearity of PKS domains, their structural modularity across many levels, and their enzymatic flexibility and promiscuity for non-canonical substrates, to expand the accessible part of the polyketide universe. Progress, however, has been slower than expected. This is predominantly a result of our inability to predict the tolerance of enzymes to facilitate the downstream biosynthesis or incorporation of novel substrates. The inherent complexity of PKSs may exceed our capacity to define a set of pre-established rules, that, when followed ensure a judicious choice of modules or domains incorporated into a reengineered BGC for the successful generation of novel products. To overcome this limitation, high-throughput approaches, on a scale comparable to the working of evolutionary recombination, will be necessary. This is where synthetic biology’s recent advances of writing genetic code at an unprecedented scale and complexity will usefully complement the repertoire of classical genetic engineering methodologies.

The first and most obvious application of synthetic biology will be for the production of novel compounds discovered by genome mining and metagenomics. It has been generally observed that the majority of BGCs in newly sequenced genomes are cryptic or silent, and the corresponding products are not produced at detectable levels in normal culture conditions. This
is usually due to strict repressive control of gene expression by global regulation embedded within the coding sequence in the form of a complex combination of promoters, 5'-UTRs\textsuperscript{70,71}, feed-forward and feed-back loops\textsuperscript{4}, pause sites and small non-coding RNA\textsuperscript{72,73}. The intertwined nature of the control circuitry makes it difficult to circumvent native regulation and force expression. “Refactoring”, a synthetic biology methodology derived from software engineering, aims at decoupling this endogenous regulation through a comprehensive rewriting of BGCs. The resulting DNA sequence of a refactored BGC is as dissimilar as possible from the wild type DNA sequence, yet still encoding the same amino acid end products. Rewriting the BGC in this fashion will remove all internal redundancy and regulation, including those regulatory elements that are currently undiscovered\textsuperscript{74}. Once a gene cluster has been refactored, we are able to introduce new, controllable and desired regulation to allow biosynthesis and characterization of the end compound, e.g. using orthogonal T7-based promoter libraries\textsuperscript{75,76}.

Refactoring BGCs aims to by-pass the classical discovery limitations, decoupling desired product expression from the complex endogenous regulatory cascade. But the resulting engineered clusters can also be designed in such a way that they facilitate further reengineering through additions, deletions, substitutions, domain swapping or other modifications (Figure 2.1b). Of course, the same limitations for successful product assembly and release apply as they do in combinatorial biosynthesis. This also implies that the generation of novel products cannot reliably be achieved using random modularization of gene units from a multitude of different sources. Refactoring, however, is providing the methodology required to generate vast libraries of BGCs for the prospective biosynthesis of novel PKSs, and by its highly parallel approach may enable the elucidation of more informative assembly rules for the engineering of chemical novelty.

Refactoring approaches to synthetic biology do not need to be restricted to the nucleotide level. Ultimately, the aim would also be to optimize the encoded protein sequences to enhance their modularity and thus increase the engineering potential of PKSs. Advances towards this are already being made\textsuperscript{77}; however, as a general strategy, such protein-level refactoring is currently still unrealistic. Nonetheless, the deluge of new genome sequence data is providing increasingly detailed insights into the rules that govern domain compatibility during the natural evolution of polyketide diversity\textsuperscript{78-80}, and multiplexed genome engineering strategies\textsuperscript{81} can be used to systematically explore these rules in the context of specific biosynthetic pathways.

In conclusion, the engineering-inspired approach of synthetic biology raises the dissection, standardization, and decoupling of distinct catalytic units from highly integrated cellular processes to new levels of ambition. By learning from the pioneering efforts of combinatorial biosynthesis, as described above, these emergent technologies will soon yield the raw materials required to construct rationally designed biosynthetic machinery and regulatory circuits from first principles on a scale, and at a speed, far superseding our current capacity. Ultimately, the success of the next generation of polyketide bioprospecting for drug discovery will depend on an intimate interaction between protein chemistry, evolutionary genomics, and synthetic biology.
2.5 Acknowledgements

We thank Jason Micklefield and Ashley Chessher for their helpful comments on the manuscript. MC was funded by a BBSRC studentship (BB/J014478/1).
2.6 References


3 The Minimum Information about a Biosynthetic Gene cluster (MIBiG) specification


1. Microbial Genomics and Bioinformatics Research Group, Max Planck Institute for Marine Microbiology, Bremen, Germany. 2. Manchester Centre for Synthetic Biology of Fine and Specialty Chemicals (SYNBIOCHEM), Manchester Institute of Biotechnology, Faculty of Life Sciences, University of Manchester, Manchester, UK. 3. Laboratory of Genetically Encoded Small Molecules, Howard Hughes Medical Institute, The Rockefeller University, New York, NY, USA. 4. Norwegian Foundation Center for Biosustainability, Technical University of Denmark, Copenhagen, Denmark. 5. Netherlands Institute of Ecology (NIOO-KNAW), Department of Microbial Ecology, Wageningen, The Netherlands. 6. Department of Chemical and Biomolecular Engineering, Department of Chemistry and Biochemistry, University of California Los Angeles, Los Angeles, CA, USA. 7. Department of Bioengineering and Therapeutic Sciences and California Institute for Quantitative Biosciences, University of California, San Francisco, San Francisco, CA, USA. 8. DOE Joint Genome Institute, Walnut Creek, California, USA. 9. Evolution of Metabolic Diversity Laboratory, Unidad de Genómica Avanzada (Langebio), Cinvestav-IPN, Irapuato, Guanajuato, Mexico. 10. Helmholtz Institute for Pharmaceutical Research, Helmholtz Centre for Infection Research and Department of Pharmaceutical Biotechnology, Saarland University, Saarbrücken, Germany. 11. Institute for Molecular Biosciences, Goethe University, Frankfurt am Main, Germany. 12. Department of Chemistry and Biochemistry, California State University, Chico, CA, USA. 13. Microbiology and Biotechnology Division, Department of Food and Environmental Sciences, University of Helsinki, Helsinki, Finland. 14. Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California at San Diego, La Jolla, CA, USA. 15. Department of Microbial Biotechnology, John Innes Centre, Norwich Research Park, Norwich, UK. 16. Department of Pharmaceutical Biology and Biotechnology, Albert-Ludwigs-University of Freiburg, Freiburg.
Germany. 17. School of Biosciences, University of Birmingham, Birmingham, UK. 18. Institute of Microbiology, Eidgenössische Technische Hochschule (ETH) Zürich, Zürich, Switzerland. 19. Department of Chemistry, University of Pittsburgh, Pittsburgh, USA. 20. Leibniz Institute for Natural Product Research and Infection Biology (HKI), Jena, Germany. 21. Gordon and Betty Moore Foundation, Palo Alto, CA, USA. 22. Sustainable Studies Program, Roosevelt University Chicago, IL, USA. 23. Merck Stiftungsprofessor für Molekulares Biotechnologie, Goethe Universität Frankfurt, Fachbereich Biowissenschaften, Frankfurt, Germany. 24. Bioinformatics Research Center, Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea. 25. Laboratory of Gene Technology, KU Leuven, Heverlee, Belgium. 26. Scripps Institution of Oceanography, University of California at San Diego, La Jolla, CA, USA. 27. Department of Biology, William Paterson University, Wayne, NJ, USA. 28. Department of Biology, Memorial University of Newfoundland, St. John’s, Canada. 29. Center for Marine Biotechnology and Biomedicine, Scripps Institution of Oceanography, University of California San Diego, CA, USA. 30. Department of Metabolic Biology, John Innes Centre, Norwich Research Park, Norwich, UK. 31. Department of Chemistry, The Scripps Research Institute, Jupiter, FL, USA. 32. Institut für Chemie, Technische Universität Berlin, Berlin, Germany. 33. BIOMERIT Research Centre, School of Microbiology, University College Cork - National University of Ireland, Cork, Ireland. 34. Departamento de Bioquímica y Genómica Microbianas, IBCE, Montevideo, Uruguay. 35. Energy Biosciences Institute and Department of Chemical and Biomolecular Engineering, University of California, Berkeley, CA, USA. 36. State Key Laboratory of Bioorganic and Natural Products Chemistry, Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences, Shanghai, China. 37. Department of Pharmaceutical Sciences, University of Shizuoka, Shizuoka, Japan. 38. Department of Medical Microbiology and Immunology, University of Wisconsin-Madison, Madison, WI, USA. 39. Department of Molecular Biosciences and Institute for Cellular and Molecular Biology, The University of Texas, Austin, TX, USA. 40. Department of Biochemistry and Biomedical Sciences, The M. G. DeGroote Institute for Infectious Disease Research, McMaster University, Hamilton, ON, Canada. 41. Life Sciences Institute, Department of Medicinal Chemistry, Department of Chemistry, and Department of Pharmaceutical Sciences, University of Michigan, Ann Arbor, MI, USA. 42. CAS Key Laboratory of Tropical Marine Bio-resources and Ecology, Guangdong Key Laboratory of Marine Materia Medica, RNAM Center for Marine Biology, South China Sea Institute of Oceanology, Chinese Academy of Sciences, Guangzhou, China. 43. Université de Lorraine, Dynamique des Génomes et Adaptation Microbienne, UMR 1128, Vandœuvres-lès-Nancy, France. 44. INRA, Dynamique des Génomes et Adaptation Microbienne, UMR 1128, Vandœuvres-lès-Nancy, France. 45. Pharmaceutical Institute, Department of Pharmaceutical Biology, University of Tübingen, Tübingen, Germany. 46. German Center for Infection Research (DZIF), Partner Site Tübingen, Tübingen, Germany. 47. Infectious Disease Research, Merck Research Laboratories, Kenilworth, NJ, USA. 48. Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA, USA. 49. Buchmann Institute for Molecular Life Sciences (BMLS), Goethe Universität Frankfurt, Frankfurt, Germany. 50. Institut für Biologie/Bakteriengenetik, Humboldt Universität Berlin, Berlin, Germany. 51. UCD School of Biomolecular and Biomedical Science, University College Dublin, Dublin, Ireland. 52. UNT System College of Pharmacy, University of North Texas Health Science Center, Fort Worth, TX, USA. 53. Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA, USA. 54. Institut für Organische Chemie, Leibniz Universität Hannover, Hannover, Germany. 55. School of Chemistry, University of Bristol, Bristol, UK. 56. Centre of Microbial and Plant Genetics, Faculty of Bioscience Engineering, University of Leuven, Heverlee, Belgium. 57. Naicons Srl, Milano, Italy. 58. Department of Molecular Biology, Princeton University, Princeton, NJ, USA. 59. Department of Chemistry, University of Illinois at Urbana, Champaign, IL, USA. 60. Howard Hughes Medical Institute, USA. 61. Collaborative Mass Spectrometry Innovation Center, University of California San Diego, La Jolla, CA, USA. 62. Department of Biology, Maynooth University, Maynooth, Co. Kildare, Ireland. 63. Molecular Microbiology, Groningen Biomolecular Sciences and Biotechnology Institute and Zernike Institute for Advanced Materials, University of Groningen, Groningen, Netherlands. 64. Institute of Microbiology, Department of Pathobiology, University of Veterinary Medicine Vienna, Vienna, Austria. 65. Department of Bioengineering and Therapeutic Sciences and California Institute for Quantitative Biosciences, University of California San Francisco, San Francisco, CA, USA. 66. Center for Marine Biotechnology and Biomedicine, Scripps Institution of Oceanography, University of California San Diego, San Diego, CA, USA. 67. Friedrich Schiller University, Jena, Germany. 68. Department Crop Protection, Faculty of Bioscience Engineering, Ghent University, Gent, Belgium. 69. Department of Biological Sciences, University of Alberta, Edmonton, AB, Canada. 70. SynBERC – Synthetic Biology Engineering Research Center, University of California, Emeryville, CA, USA. 71. Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT, USA. 72. Department of Bacteriology, University of Wisconsin-Madison, Madison, WI, USA. 73. Institute of Molecular Biology, Slovak Academy of Sciences, Bratislava, Slovak Republic. 74. Department of Molecular Genetics, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, Groningen, The Netherlands. 75. Biotechnology Research Center, The University of Tokyo, Tokyo, Japan. 76. Department of Biological Sciences, Faculty of Science, King Abdulaziz University, Jeddah, Saudi Arabia. 77. Division of Bioscience and Bioinformatics, Myongji University, Yongin-si, Gyeonggi-Do, South Korea. 78. Institute of Integrative Biology of the Cell (I2BC), CEA, CNRS, Université Paris Sud, Orsay, France. 79. Laboratory of Food Microbiology, KU Leuven, Heverlee, Belgium. 80. Department of Microbiology and Immunology, University of Arkansas for Medical Sciences, Little Rock, AR, USA. 81. State Key Laboratory of Microbial Metabolism, and School of Life Sciences & Biotechnology, Shanghai Jiao Tong University, Shanghai, China. 82. Department of Pharmaceutical Sciences, Oregon State University, Corvallis, OR, USA. 83. Microbiology/Biotechnology, Interfaculty Institute of Sciences & Biotechnology, Shanghai Jiao Tong University, Shanghai, China. 84. Department of Pharmaceutical Microbiology, KU Leuven, Heverlee, Belgium. 85. Department of Microbiology and Immunology, University of Arkansas. 86. Infectious Disease Research, Merck Research Laboratories, Kenilworth, NJ, USA. 87. School of Chemistry, University of Manchester, Manchester, United Kingdom. 88. Department of Chemistry, University of Illinois at Urbana-Champaign, Urbana, IL, USA. 89. Institute for Bioengineering and Biosciences, Instituto Superior Técnico, Universidade de Lisboa, Lisboa, Portugal. 90. School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney, Australia. 91. Department of Chemical and Biological
Engineering, Chalmers University of Technology, Gothenburg, Sweden. 92. Curtin University, School of Biomedical Sciences, Perth, Australia. 93. Division of Chemistry, Graduate School of Science, Hokkaido University, Japan. 94. Tufts University School of Medicine, Department of Molecular Biology and Microbiology, Boston, MA, USA. 95. Department of Genetics and Biotechnology, Ivan Franko National University of Lviv, Lviv, Ukraine. 96. Institute of Microbiology ASCR, Prague, Czech Republic. 97. LIED, UMR8236 CNRS, Université Paris Diderot, Paris, France. 98. Novartis Institutes for BioMedical Research, Novartis Campus, Basel, Switzerland. 99. Institute of Fundamental Sciences, Massey University, Palmerston North, New Zealand. 100. Astbury Centre for Structural Molecular Biology, School of Molecular and Cellular Biology, Faculty of Biological Sciences, University of Leeds, Leeds, UK. 101. Molecular Therapeutics and Natural Products Library Initiative, The Scripps Research Institute, Jupiter, FL, USA. 102. Department of Biochemistry, Molecular Biology, and Biophysics and BioTechnology Institute, University of Minnesota – Twin Cities, Saint Paul, MN, USA. 103. INRA, UMR1290 BIOGER, Av. L. Brétignières, Grignon, France. 104. Department of Energy Great Lakes Bioenergy Research Center and Department of Energy Plant Research Laboratory, Michigan State University, East Lansing, MI, USA. 105. ChEM-H Institute, Stanford University, Stanford, USA. 106. Molecular Biotechnology, Institute of Biology, Leiden University, Leiden, The Netherlands. 107. Hofstra North Shore-Long Island Jewish School of Medicine, Hempstead, NY, USA. 108. Department of Biotechnology, Norwegian University of Science and Technology, Trondheim, Norway. 109. Jacobs University Bremen gGmbH, Bremen, Germany.

*Present address: Bioinformatics Group, Wageningen University, Wageningen, The Netherlands
Contributions to the manuscript are as follows: M.H.M., R.B., E.T. and F.O.G. initiated and coordinated the MIBiG standardization project. M.H.M. designed the first draft of the standard. M.H.M., R.K., P.Y. and F.O.G. coordinated integration within the MiXS framework. M.H.M. and R.K. constructed the submission system. M.H.M. and M.C. curated submitted community annotation entries. M.H.M., R.K., P.Y., M.C., R.B., E.T. and F.O.G. wrote the first draft of the paper. All authors contributed to the design of the standard, contributed to the community annotation of MIBiG entries and provided feedback on an early draft of the paper.

For the sake of space the author and affiliation lists are single spaced. Boxed comments within the published manuscript are written in italics to disambiguate this content from the main text. Supplemental data set 1 comprises the responses of a community survey and is not included within the thesis. The survey can be made available upon request, or found here (https://www.nature.com/nchembio/journal/v11/n9/full/nchembio.1890.html).

The manuscript was published in Nature Chemical Biology and was available online on the 18th of August 2015

Citations as of submission date: 96
3.2 Abstract

A wide variety of enzymatic pathways to produce specialized metabolites in bacteria, fungi and plants are known to be encoded in biosynthetic gene clusters. Information about these clusters, pathways and metabolites is currently dispersed throughout the literature, making it difficult to exploit. To facilitate consistent and systematic deposition and retrieval of data on biosynthetic gene clusters, we propose the Minimum Information about a Biosynthetic Gene cluster (MIBiG) data standard.
3.3 Introduction

Living organisms produce a range of specialised metabolites with exotic chemical structures and diverse metabolic origins. Many of these specialised metabolites find use as natural products in medicine, agriculture and manufacturing. Research on natural product biosynthesis is undergoing an extensive transformation, driven by technological developments in genomics, bioinformatics, analytical chemistry and synthetic biology. It has now become possible to computationally identify thousands of biosynthetic gene clusters (BGCs) in genome sequences, and to systematically explore and prioritize them for experimental characterization\textsuperscript{1,2}. A BGC can be defined as a physically clustered group of two or more genes in a particular genome that together encode a biosynthetic pathway for the production of a specialized metabolite (including its chemical variants). It is becoming possible to carry out initial experimental characterization of hundreds of such natural products, using high-throughput approaches powered by rapid developments in mass spectrometry\textsuperscript{3-5} and chemical structure elucidation\textsuperscript{6}. At the same time, single-cell sequencing and metagenomics are opening up access to new and uncharted branches of the tree of life\textsuperscript{7-9}, enabling scientists to tap into a previously undiscovered wealth of BGCs. Furthermore, synthetic biology allows the redesign of BGCs for effective heterologous expression in pre-engineered hosts, which will ultimately empower the construction of standardized high-throughput platforms for natural product discovery\textsuperscript{10,11}.

In this changing research environment, there is an increasing need to access all the experimental and contextual data on characterized BGCs for comparative analysis, function prediction, and for collecting building blocks for the design of novel biosynthetic pathways. For this purpose, it is paramount to have this information available in a standardized and systematic format, accessible in the same intuitive way as, for example, genome annotations or protein structures. Currently, the situation is far from ideal, with information on natural product biosynthetic pathways scattered across hundreds of scientific articles in a wide variety of journals; it requires in-depth reading of papers to confidently discern which of the molecular functions associated with a gene cluster or pathway have been experimentally verified and which have been predicted based solely on biosynthetic logic or bioinformatic algorithms. Although some valuable existing manually curated databases have data models in place to store some of this information\textsuperscript{12-14}, all are specialized towards certain subcategories of BGCs and include a limited number of parameters defined by the interests of a subset of the scientific community. To enable the future development of databases with universal value, a generally applicable community standard is required that specifies the exact annotation and metadata parameters agreed upon by a wide range of scientists, as well as the possible types of evidence that are associated with each variable in publications and/or patents. Such a standard will be of great value for the consistent storage of data, which will alleviate the tedious process of manually gathering information on BGCs. Moreover, a comprehensive data standard will allow future data infrastructures to enable the integration of multiple types of data, which will generate new insights that would otherwise not be attainable.

The Genomic Standards Consortium (GSC)\textsuperscript{15} (Box 1) previously developed the Minimum Information about any Sequence (MIxS) framework\textsuperscript{16}. This extensible ‘minimum information’
standardization framework includes the Minimum Information about a Genome Sequence (MIGS)\textsuperscript{17} and the Minimum Information about a MARKer gene Sequence (MIMARKS)\textsuperscript{16} standards. MIxS is a flexible framework that can be extended upon to serve a wide variety of purposes. The GSC facilitates the community effort of maintaining and extending MIxS, and stimulates compliance among the community.

Here, we introduce the "Minimal Information about a Biosynthetic Gene cluster" (MIBiG) specification as a coherent extension of the GSC’s MIxS standards framework. MIBiG provides a comprehensive and standardized specification of BGC annotations and gene cluster-associated metadata, which will allow their systematic deposition in databases. Through a community annotation of BGCs that have been experimentally characterized and described in the literature during previous decades, we have constructed an MIBiG-compliant seed dataset. Moreover, a large part of the research community has committed to continue submitting data on newly characterized gene clusters in the MIBiG format in the future. Together, the MIBiG standard and the resulting MIBiG-compliant data sets will allow data infrastructures to be developed that will facilitate key future developments in natural product research.

Box 1: The Genomic Standards Consortium and its MIxS framework

The Genomic Standards Consortium (GSC, \url{http://gensc.org/}) was founded in 2005 as an open-membership working body with the purpose to promote the standardization of genome descriptions as well as the exchange and integration of genomic data.

The GSC initiates and coordinates the design of and compliance to several minimum information standards (also called checklists). An overarching framework has been designed to connect and standardize these checklists themselves: the Minimum Information about any (x) Sequence (MIxS)\textsuperscript{16}. MIxS consists of three layers:

First, the MIxS standard includes a number of shared descriptors that are relevant to all types of nucleotide sequences, such as collection date, environmental origin, geographical location and sequencing method.

Second, for a wide range of different environmental origins, so-called ‘environmental packages’ are available that constitute checklists of measurements and observations that are specific to each environment: for example, for host-associated microbial DNA samples, the taxonomy of the host, the habitat of the host and several phenotypic characteristics of the host can be collected. In this manner, rich contextual information on the context of each microbial sample is stored.

Third, several checklists are available for specific sequence types, each having their own checklist-specific descriptors. Previous checklists include the Minimum Information about a Genome Sequence (MIGS)\textsuperscript{17}, Minimum Information about a Metagenome Sequence (MIMS)\textsuperscript{17} and the Minimum Information about a MARKer gene Sequence (MIMARKS)\textsuperscript{16}.

In spring 2013, the MIBiG project proposal was accepted by the board of the GSC to form a new standards project within the MIxS framework. Besides a number of general descriptors, it also includes pathway type-specific packages that function analogously for different classes of
biosynthetic pathways as the MiXs environmental packages do for different environmental origins.

3.4 Results and discussions

3.4.1 Design of the MIBiG standard

The MIBiG standard covers general parameters that are applicable to each and every gene cluster, as well as compound type-specific parameters that apply only to specific classes of pathways (Figure 3.1). Notably, the standard has been designed to be suitable for biosynthetic pathways from any taxonomic origin, including those from bacteria, archaea, fungi and plants.

The general parameters cover important data items that are universally applicable. First, they include identifiers of the publications associated with the characterization of the gene cluster, so that the full description of the experimental results that support the entire entry can be accessed easily.

Figure 3.1: Schematic overview of the MIBiG standard.

The MIBiG standard is composed of general and compound class-specific parameters. Wherever relevant, evidence coding is used to determine the experimental support for certain items in the checklist. Fields annotated with a star are absolutely mandatory; fields with two stars are conditionally mandatory.

The second key group of general parameters describes the associated genomic locus (or loci) and its accession numbers and coordinates, as deposited in or submitted to one of the
databases of the International Nucleotide Sequence Database Collaboration (INSDC): the DNA Data Bank of Japan (DDBJ), the European Nucleotide Archive (EBI-ENA) or GenBank, all of which share unified accession numbers. The INSDC accession numbers are also used to link each MiBiG entry (which is given a separate MiBiG accession number) and its annotations to the corresponding nucleotide sequence(s) computationally; hence, a GenBank/ENA/DDBJ submission of the underlying nucleotide sequence is always required to file an MiBiG submission.

The third group of general parameters describes the chemical compounds produced from the encoded pathway, including their structures, molecular masses, biological activities and molecular targets. Additionally, these parameters allow documentation of miscellaneous chemical moieties that are connected to the core scaffold of the molecule (but synthesized independently) and the genes associated with their biosynthesis; this will facilitate the design of tools for the straightforward comparison of such 'sub-clusters', which are frequently present in different variants across multiple parent BGCs.

Finally, there is a group of general parameters describing experimental data on genes and operons in the gene cluster, including gene knockout phenotypes, experimentally verified gene functions and operons verified by techniques such as RNAseq.

Besides the general parameters, the MiBiG standard contains dedicated class-specific checklists for gene clusters encoding pathways to produce polyketides, nonribosomal peptides (NRPs), ribosomally synthesized and post-translationally modified peptides (RiPPs), terpenes, saccharides, and alkaloids. These include items like acyltransferase domain substrate specificities and starter units for polyketide BGCs, release/cyclization types and adenylation domain substrate specificities for NRP BGCs, precursor peptides and peptide modifications for RiPP BGCs, and glycosyltransferase specificities for saccharide BGCs. Where applicable, the standard was made compliant with earlier community agreements, such as the recently published classification of RiPPs\textsuperscript{18}. Hybrid BGCs that cover multiple biochemical classes can be described by simply entering information on each of the constituent compound types: the checklists have been designed in such a way that this does not lead to conflicts. Importantly, the modularity of the checklist system allows straightforward addition of further class-specific checklists when new types of molecules are discovered in the future.

The combination of general and compound-specific MiBiG parameters, together with the MIxS checklist, provides a complete description of the chemical, genomic and environmental dimensions that characterize a biosynthetic pathway (Figure 3.2). A minimal set of key parameters is mandatory, while other parameters are optional. For many parameters, a specific ontology has been designed in order to standardize the inputs and to make it easier to categorize and search the resulting data.
Figure 3.2: An example MIBiG entry

Entry describing the relatively simple hybrid NRPS-PKS biosynthetic gene cluster for isoflavipucine/dihydroisoflavipucine from *Aspergillus terreus*. Fields without information have been omitted and some JSON field abbreviations have been modified for clarity. The full entry is available from http://mibig.secondarymetabolites.org/repository/BGC0001122/BGC0001122.json.

Whenever possible, parameters are linked to a system of evidence attribution that specifies the kinds of experiments performed to arrive at the conclusions indicated by the chosen parameter values. Hence, each annotation entered during submission is assigned a specific evidence code: for example, when annotating the substrate specificity of a nonribosomal peptide synthetase (NRPS) adenylation domain, the submitter can choose between ‘activity assay’, ‘structure-based inference’ and ‘sequence-based prediction’ as evidence categories to support a given specificity.

During the design of the standard, great care was taken to make it compatible with unusual biosynthetic pathways, such as branched or module-skipping polyketide synthase (PKS) and NRPS assembly lines. Also, to ensure that the standard is compliant with the current state of the art in the various subfields of natural product research, we conducted an online community survey at an early stage of standard development (see Supplemental Data Set 1). Feedback was provided by 61 principal investigators from 16 different countries (most of whom also co-authored this paper), including at least ten leading experts for each major class of biosynthetic pathways covered.

3.4.2 MIBiG: designed to address key research needs

Adoption of the MIBiG standard will allow for the straightforward collation of all annotations and experimental data on each BGC, which would otherwise be dispersed across multiple scientific articles and resources. Moreover, there are at least three additional key ways in which MIBiG will facilitate new scientific and technological developments in the future: it will enable researchers to systematically connect genes to chemistry (and vice versa), to better understand...
specialised metabolite biosynthesis and the compounds produced in their ecological and environmental context, and to effectively use synthetic biology to engineer novel BGC configurations underpinned by an evidence-based parts registry (Figure 3.3).

First, the comprehensive dataset generated through MIBiG-compliant submissions will enable researchers to systematically connect genes and chemistry. Not only will it allow individual researchers to predict enzyme functions by comparing enzyme-coding genes in newly identified BGCs to a thoroughly documented dataset, it will also facilitate general advances in chemistry predictions. Substrate specificities of PKS acyltransferase domains, NRPS adenylation domains and their evidence codes will be registered automatically for all gene clusters. This enables automated updating of the training sets for key chemistry prediction algorithms, which can then be curated by the degree of evidence available, increasing the accuracy of predictions of core peptide and polyketide scaffolds. Also, because groups of genes associated with the biosynthesis of specific chemical moieties (such as sugars and non-proteinogenic amino acids) will be registered consistently, a continuously growing dataset of such sub-clusters will be available to use as a basis for chemical structure predictions.

**Figure 3.3: The position of MIBiG in specialized metabolite research**

The MIBiG data standard and submission system will lead to a continuously growing dataset (stored in the online MIBiG repository) that will be loaded into several databases and web services. The lower part of the figure shows the threefold potential of MIBiG for the study of BGCs, which will allow 1) systematically
connecting genes and chemistry, by identifying which genes are responsible for the biosynthesis of which chemical moieties; 2) understanding the natural genetic diversity of BGCs within their environmental/ecological context, by combining MiBiG- and MixS-derived metadata sets; 3) developing an evidence-based parts registry for engineering metabolic pathways through synthetic biology.

In addition, MiBiG has the potential to greatly enhance the understanding of specialised metabolite biosynthesis in its ecological and environmental context: the connection of MiBiG to the MixS standard should stimulate researchers to supply MiXs data on the genome and metagenome sequences that contain the BGCs. This will generate opportunities for a range of analyses, such as the biogeographical mapping of specialised metabolite biosynthesis thereby identifying locations and ecosystems harboring rich biosynthetic diversity. But even if the contextual data associated with the genome sequences cannot always be made MiXs-compliant (e.g., because the origin of a strain can no longer be traced), the MiBiG standard itself provides a comprehensive reference dataset for annotating large-scale MiXs-compliant metagenomic data from projects such as the Earth Microbiome Project, Tara Oceans and Ocean Sampling Day. This will enable scientists to obtain a better understanding of the distribution of BGCs in the environment. Altogether, the standard will play a significant role in guiding sampling efforts for future natural product discovery.

Finally, the data resulting from MiBiG-compliant submissions will provide an evidence-based parts registry for the engineering of biosynthetic pathways. Synthetic biologists need a toolbox containing genetic parts that have been experimentally characterized. The MiBiG standard, through its systematic annotation of gene function by evidence coding, knockout mutant phenotypes and substrate specificities, will streamline the identification of all available candidate genes and proteins available to perform a desired function, together with the pathway context in which they natively occur. In this manner, it will provide a comprehensive catalogue of parts that can be used for the modification of existing biosynthetic pathways or the de novo design of new pathways.

### 3.4.3 Community annotation effort

To accelerate the usefulness of new MiBiG-compliant data submissions, we initiated this project by annotating a significant portion of the experimental data on the hundreds of BGCs that have been characterized in recent decades. The resulting data will allow immediate contextualization of new submissions (see below) and comparative analysis of any newly characterized BGCs with a rich source of MiBiG-compliant data. Moreover, this annotation effort offered an ideal opportunity to evaluate the MiBiG standard in practice on a diverse range of BGCs. Hence, we carefully mined the literature to obtain a set of 1170 experimentally characterized gene clusters: 303 PKS, 189 NRPS, 147 hybrid NRPS-PKS, 169 RiPP, 78 terpene, 123 saccharide, 21 alkaloid and 140 other BGCs. Compared to the 288 BGCs currently deposited in ClusterMine and the 103 BGCs deposited in DoBISCUIT, this presents a significant advance in terms of comprehensiveness. We then annotated each of these 1170 BGCs with a minimal number of parameters (genomic locus, publications, chemical structure and
biosynthetic class/subclass). Subsequently, in a community initiative involving 81 academic research groups and several companies worldwide, we performed a fully MIBiG-compliant re-annotation of 405 of these BGCs according to the information available in earlier publications and laboratory archives. (All participants of this annotation effort are either listed as co-authors of this article or mentioned in the acknowledgements, depending on the size of their contribution.) An initial visualization of the full data set arising from this re-annotation is publicly available online at http://mibig.secondarymetabolites.org. Altogether, these submitted entries will function as a very useful seed dataset for the development of databases on specialised metabolism. Future data curation efforts will strive to achieve a fully MIBiG-compliant annotation of the remaining 765 BGCs that are currently annotated with a more restricted set of parameters.

3.4.4 Planned implementation in databases and web services

To allow straightforward and user-friendly access, the MIBiG standard will be implemented by multiple databases and web services for genome data and specialised metabolite research. For example, the MIBiG-curated dataset has already been integrated into a new module in antiSMASH\textsuperscript{26}, which compares any identified BGCs with the full MIBiG-compliant dataset of known BGCs. Moreover, a full-fledged database is currently under development that will be tightly integrated with antiSMASH and will build on the previously published ClusterMine\textsuperscript{360} framework\textsuperscript{12}. Additionally, MIBiG-compliant data will be integrated into the recently released Integrated Microbial Genomes Atlas of Biosynthetic Clusters (IMG-ABC) database from the Joint Genome Institute (https://img.jgi.doe.gov/ABC/)\textsuperscript{27}. Regular exchange of data will take place between the MIBiG repository and the IMG-ABC/antiSMASH/ClusterMine databases. Additional cross-links with the chemical databases ChemSpider\textsuperscript{28}, chEMBL\textsuperscript{29} and chEBI\textsuperscript{30} are being developed, so that researchers can easily find the full MIBiG annotation of the BGC responsible for the biosynthesis of given molecules. Finally, all community-curated data are freely available and downloadable in JSON format for integration into other software tools or databases, without any need to request permission, as long as the source is acknowledged.

For submission of new MIBiG-compliant data by scientists in the field, we prepared an interactive online submission form (available from http://mibig.secondarymetabolites.org), which was extensively tested through the community annotation effort. Data can also be submitted through the BioSynML plug-in\textsuperscript{26} (http://www.biosynml.de) that was recently built for use in the Geneious software. In this way, MIBiG-compliant data can easily be integrated with in-house BGC content management systems of individual laboratories or companies. Finally, it will be possible to submit updates to existing MIBiG entries based on peer-reviewed articles through dedicated web forms.

3.5 Future perspectives

The MIBiG coordinating team within the GSC is committed to ensuring the continued support and curation of the MIBiG standard, in cooperation with its partners. Compliance to the standard and interoperability with other standards and databases will also be guaranteed within GSC. In
order to stay relevant and viable, MIBiG is projected to be a ‘living’ standard: updates will be made as needed to remain technologically and scientifically current.

Coordination with relevant journals will be sought to make MIBiG submission of BGCs (evidenced by MIBiG accession codes) a standard item to check during manuscript review. To stimulate submission of MIBiG data during the process of publishing new biosynthetic gene clusters, unique MIBiG accession numbers are provided for each BCG, which can be used during article review (including data embargoed until after publication). The research community represented by this paper commits itself to submitting MIBiG-compliant data sets as well as updates to existing entries when publishing new experimental results on BGCs. We encourage the larger community to join in this endeavor.

3.6 Acknowledgements

M.H.M was supported by a Rubicon fellowship of the Netherlands Organization for Scientific Research (NWO; Rubicon 825.13.001). The work of R.K. was supported by the European Union’s Seventh Framework Programme (Joint Call OCEAN.2011-2: Marine microbial diversity – new insights into marine ecosystems functioning and its biotechnological potential) under the grant agreement no 287589 (Micro B3). M.C was supported by a BBSRC studentship (BB/J014478/1). The GSC is supported by funding from the Natural Environment Research Council, the NIEeS, the Gordon and Betty Moore Foundation, the National Science Foundation (NSF) and the U.S. Department of Energy. The Manchester Synthetic Biology Research Centre, SYNBIOCHEM, is supported by BBSRC/EPSRC grant BB/M017702/1. We thank Paul d’Agostino, Paul R. August, Rocky Chay, Caitlin D. Deane, Stephan Diethelm, Lorena Fernandez-Martinez, Abrahim El Gamal, Chantal Garcia De Gonzalo, Trudy H. Grossman, Chien-Jui Huang, Shinya Kodani, Ana Ligia Leandrini, Ian A. MacNeil, Mikhail Metelev, Evelyn M. Molly, Carlos Olano, Manuel Ortega, Lauren Ray, Kirk Reynolds, Avena Ross, Inês N. Silva, Robin Teufel, Gabrielle Thibodeaux, Jonathan Tietz and David Widdick for their contributions in the community annotation. We thank Richard Baltz, Mervyn Bibb, Christopher Boddy, Christoph Corre, Elke Dittmann, Hugo Gramajo, Natsuko Ichikawa, Haruo Ikeda, Paul Jensen, Chaitan Khosla, Rongfeng Li, Mohamed Marahiel, Debasisa Mohanty, Charles Moore, William Nierman, Dong-Chan Oh, Eric Schmidt, Yuemao Shen, David Stevens, Bettina Tudzynski and Steven Van Lanen for useful comments on an early draft version of the community standard. We are grateful to three anonymous referees for their constructive suggestions.

3.7 Competing Financial Interests Statement

There are no competing financial interests.
3.8 References


4 An Output Ordering and Prioritisation System (OOPS) for ranking biosynthetic gene clusters to enhance bioactive metabolite discovery

Matthew Cummings\textsuperscript{1†}, Alejandro Pena\textsuperscript{2†}, Francesco Del Carratore\textsuperscript{2†}, Eriko Takano\textsuperscript{1} and Rainer Breitling\textsuperscript{2*}

\textsuperscript{1}Manchester Institute of Biotechnology, School of Chemistry, The University of Manchester, UK

\textsuperscript{2}Manchester Institute of Biotechnology, School of Computing science, The University of Manchester, UK

†Joint first authors.

*Corresponding author: Rainer Breitling, Manchester Institute of Biotechnology, The University of Manchester, 131 Princess Street, Manchester, M1 7DN

Tel: +44 (0) 161 306 5117

E-mail: Rainer.breitling@manchester.ac.uk

Keywords: Prioritisation, biosynthetic gene clusters, specialised metabolites, antiSMASH, natural products

Submission to the Journal of Industrial Microbiology and Biotechnology – Synthetic Biology Thematic edition on the 15\textsuperscript{th} of September 2017
4.1 Preface

The project was proposed by E.T. and R.B.. Prioritisation parameters were decided by M.C., E.T. and R.B.. OOPS version 1 (1.0) was built by A.P, supervised by R.B and submitted as part of a M.Sc. thesis by A.P. Testing of OOPS 1.0 was undertaken by M.C. and the software was significantly refined, accordingly, by A.P.. Testing of OOPS version 2 (2.0) was undertaken by M.C and F.D.C. Actinobacterial BGCs were precomputed by A.P., M.C. and F.D.C.. The first draft of the manuscript was written by M.C., M.C., F.D.C., E.T. and R.B finalised the Abstract, Introduction and Methods and Materials sections of the manuscript. The Results and Discussion, and Conclusions were written by M.C.

The manuscript submitted to the Journal of Industrial Microbiology and Biotechnology consists of a shortened version of the manuscript here lacking the results, discussion and conclusion sections as well as Figure 4.1 and Figure 4.3. The manuscript reported here is in a plain format, consistent with the entire thesis, and not in the formatting require for the Journal of Industrial Microbiology and Biotechnology. Furthermore, Nature style citations are used for consistency, and are single line spaced for the sake of space.
4.2 Abstract

The exploitation of the rich biosynthetic resources revealed by recent microbial genome sequencing projects is one of the key drivers of modern drug discovery and synthetic biology. The rapid increase of publicly available microbial genome sequences has highlighted the presence of hundreds of thousands of biosynthetic gene clusters (BGCs) encoding the biosynthesis of specialised metabolites with potentially valuable bioactivities. Experimental characterization of new BGCs is still an extremely laborious process and laboratory research struggles to keep pace with the speed of in silico identification. Effective prioritization of the most promising candidates among the wealth of computationally predicted BGCs thus represents a need for the scientific community.

Here we propose an Output Ordering and Prioritisation System (OOPS) which helps prioritise in silico predicted BGCs using a wide variety of key biological, molecular biological and biochemical criteria in a flexible and user-friendly interface. OOPS facilitates a judicious choice of BGCs for experimental characterisation using G+C content, coding sequence length, gene number, cluster self-similarity and codon bias parameters, as well as enabling the user to rank BGC based upon BGC type, novelty and taxonomic distribution, each of which can be individually tuned. Effective prioritisation of BGCs using OOPS will help to reduce experimental attrition rates and improve the breadth of bioactive specialised metabolites characterised.
4.3 Introduction

Estimates indicate antimicrobial resistance to be one of the main existential crises to humanity in the next 33 years. Documented cases of pan-resistant bacterial infections are predicted to rise to 100 million by 2050\(^1\). Discovery of the next generation of antibiotic agents is imperative to circumvent such a catastrophe. One answer to our urgent requirement for novel bioactive compounds may lie in the wealth of publicly available genome sequences present in freely accessible online databases\(^2-4\). In recent years, this wealth of genome sequence data has revealed an unexpected diversity of biosynthetic gene clusters (BGCs) which are unequivocally linked to specialised metabolites\(^5-10\). Bacterial and fungal specialised metabolites have historically been the source of many clinically used medicines; however, the low-hanging-fruit has mostly been picked and classical routes to natural product-based drug discovery have been perceived as resulting in diminishing returns\(^11,12\), yielding little chemical novelty in the last two decades\(^13,14\). Instead a new retrospective approach to natural product identification is needed, linking BGCs to specialised metabolites\(^13\). The paradigm-shift in molecular biology techniques afforded by the advent of synthetic biology is facilitating this by moving the goal posts in our favour, enabling access to a new pool of silent, cryptic and poorly expressed BGC\(^15,16\).

As a result of the recent advances in computational biology, BGCs can be systematically predicted from DNA sequence data using freely available cluster mining tools (antiSMASH\(^17\), BAGEL\(^18\), CLUSEAN\(^19\), ClusterFinder\(^5\), PRISM\(^20\) etc.). The widespread use of these tools has provided an unprecedented view of the global distribution of specialised metabolites across all prokaryotic divisions\(^5\) and is beginning to unveil the true richness of prokaryotic specialised metabolism. A prevalence of BGCs across a broad range of Eukaryotic and Prokaryotic taxa is augmenting the natural products landscape, and blurring the historic heavy focus away from the pharmacologists favourite Phylum, Actinobacteria, and into new and lesser mined organisms e.g. \textit{Burkholderia} spp., \textit{Moorea} spp., \textit{Streptococcus} spp. etc., uncultivable organisms\(^21,22\), and metagenomes\(^22\). Furthermore, the next harvest of industrially and medically useful specialised metabolites may follow a more targeted pipeline using ecological niches to identify promising drug leads e.g. should an immunosuppressant be of interest perhaps BGCs derived from genome minimised intra- or extracellular obligate human pathogens are more fruitful candidates for exploration. These BGCs could be maintained under a more stringent and relevant selection pressure than those from the large genomes of soil-borne \textit{Actinobacteria}. Additionally, redundancy between biosynthetic gene clusters can be used to identify natural products with very similar and useful core structures but which are differentially tailored. Using this approach, the development of more potent antineoplastics, for example, via combinatorial biosynthesis or semi-synthesis\(^23,24\) could be superseded through a more informed search for analogues predicted from newly sequenced BGCs homologues, and achieved enzymatically.

Irrespective of the approach used to identify new BGCs of interest, shortlisting clusters for subsequent experimental characterization is a difficult and laborious task. BGC prediction software provides a useful, but usually unsorted, output per genome or input sequence. However, to our knowledge there is no easily used tool to combine the information from multiple
genomes for visualization, filtering and, most importantly, rigorous and consistent prioritisation based on key molecular biological and biochemical features.

Here we describe a freely available and easy-to-use system, OOPS, for the user-defined prioritisation of all classes of biosynthetic gene clusters using seven types of individually weighted parameters relevant for metabolite discovery (Figure 4.1) and the exploitation of BGCs using molecular biology and synthetic biology approaches.

Figure 4.1: Schematic overview of BGC prioritisation parameters used in OOPS.

Schematics of seven tuneable BGC prioritisation parameters. * The user can define a references species for a comparative analysis. Sorting BGCs according to class is not shown schematically.

### 4.4 Methods and materials

The OOPS software is a stand-alone Java-based application built with an embedded web server and accordingly is compatible with all operating systems. OOPS should be considered as
an extension of the antiSMASH pipeline\textsuperscript{17}: the input used for OOPS comprises BGCs predicted using any version of the prokaryotic or fungal antiSMASH pipelines.

### 4.4.1 Prioritization protocol

After optional filtering of the BGCs according to the cluster type, i.e., the predicted chemical class of the end compound, all seven prioritisation metrics are computed (or extracted from the antiSMASH output) for each BGC. The BGCs can be easily ranked according to each metric in ascending or descending order according to the user's choice. A final score associated to each BGC is then computed as a simple weighted sum of the all the ranks obtained in the previous step. The weighted scoring system provides the user with complete control over which parameters enter the final prioritization, and to which extent they dominate the results. This allows flexible adjustments of BGC prioritization according to the specific down-stream analysis envisaged (e.g., some users might search for novel chemistry, independent of the genetic structure of the BGC, while others may be looking only for clusters that are easy to manipulate using synthetic biology tools and therefore would want to prioritize BGC with few ORFs and relatively short total protein coding sequences).

### 4.4.2 Guanine and cytosine content

The G+C content for each cluster is computed from the antiSMASH output simply as the ratio G+C/A+T DNA base pairs over the length of the predicted BGC. If no reference species is selected by the user, BGCs are ranked by their total G+C content. Otherwise, the G+C content of the reference species is obtained from the Kazusa codon usage database (http://www.kazusa.or.jp/codon/) and the BGCs are ranked according to the absolute difference between the G+C content of each BGC and the G+C content of the reference species. This allows, e.g., prioritisation of BGCs that are maximally similar in G+C content to a prospective heterologous host species.

### 4.4.3 Codon bias

The codon bias parameter can only be used for the prioritization if a reference species is selected. The codon usage table is computed by OOPS for each BGC, while the table for the species of interest, typically the intended heterologous expression host is downloaded for the Kazusa website (http://www.kazusa.or.jp/codon/). The clusters are then ranked according the BGC codon bias score computed as follows:

\[
BGC\ codon\ bias\ score = \sum_{i=1}^{64} |x_{BGC}^{i} - x_{species}^{i}|
\]

where \(x_{BGC}^{i}\) and \(x_{species}^{i}\) represent the usage percentage of the \(i^{th}\) codon in the BGC and in the species respectively. Similar to the G+C content, this allows prioritisation of BGCs that have a codon bias similar to an intended host species, but it could also be employed more creatively, e.g., to prioritize clusters that match the coding patterns of a particular group of organisms with interesting known bioactivities.
4.4.4 Similarity to known cluster

OOPS retrieves from the antiSMASH output the percent similarity (calculated analogously to MultiGeneBlast\textsuperscript{25}) of the most similar BGC with a known end product. The score for this ranking parameter is calculated as the absolute difference between the similarity preference defined by the user (\%PS) and the similarity of the most similar cluster (\%known cluster).

\[ \text{known cluster score} = |\%PS - \%\text{known cluster}| \]

This enables users to prioritize BGC that hit the “sweet spot” of similarity to known clusters, which will differ depending on the specific application scenario: some users will search minor, but important variations of established known compounds; others prefer clusters that are completely different from anything studied before and will have a \%PS of zero.

4.4.5 Self-similarity

This parameter is used in order to prioritize the clusters according to how similar they are to themselves in terms of nucleotide sequences. This metric is computed using a modified version of the Smith-Waterman algorithm\textsuperscript{26} to find all suboptimal local alignments with a length equal to, or above, a minimum length specified by the user. This parameter allows elimination (or a preference) of BGCs with a large number of internal repeats, which might be challenging to engineer genetically (or might be chemically particularly interesting, depending on the use case).

4.4.6 Phylogenetic diversity

OOPS uses the cluster blast output provided for each BGC by antiSMASH\textsuperscript{17}, which contains all BGCs that are substantially similar to the BGC of interest, to compute a phylogenetic diversity score. The taxonomic identity of all host species harbouring significantly similar BGCs are obtained from the RESTful services provided by the EMBL-EBI databases\textsuperscript{27} and used to build a simplified phylogenetic tree, using only the taxonomic ranks provided by the database, and collapsing all intermediate nodes. The metric chosen to represent the phylogenetic diversity is the sum of nodes within the taxonomic tree above the species level. A high score indicates that similar clusters are very widespread across the tree of life, while low numbers indicate that similar clusters are found only in a small set of closely related species. Whether this parameter is used in ascending or descending order for the final prioritization will again strongly depend on the envisaged application.

All prioritization options are accessible via a unified intuitive user interface (Figure 4.2).
Figure 4.2: OOPS graphical user interface.

The weighing of each prioritization parameter can be adjusted using the slide bar, or ignored by checking the associated box. The sorting order (ascending or descending) is specified by clicking the blue arrow icon. The reference species is chosen by typing the species name in the ‘species’ field and is relevant for codon bias and G+C content parameters only. Multiple BGC type can be chosen using the “preferred cluster type” field by holding shift and selecting additional BGC classes.

4.4.7 Availability
OOPS is available at https://github.com/alexcpa/antSMASH-OOPS.

4.4.8 Precomputation of Actinobacterial BGCs
To illustrate the potential uses of our software, we provide a pre-computed dataset containing all actinobacterial genomes present within the antiSMASH database\textsuperscript{28}. Expression of heterologous Actinobacterial BGCs in various \textit{Streptomyces coelicolor} strains\textsuperscript{29,30} is a common first strategy when characterising new secondary metabolites; thus, \textit{S. coelicolor} A3(2) was chosen as our reference species for precomputation of codon bias and G+C content parameters. Additionally, to provide good resolution of BGC self-similarity (detection of relevant internal repeats), a threshold of 30 bp was used. Changing the reference species or self-similarity threshold when using the large precomputed dataset will result in lengthy prioritisation (> 4 days) and should be avoided. The pre-computed dataset here described is available for download at https://zenodo.com/OOPS1_RBreitling

4.4.9 Processing commercially sensitive data
All input data remains offline and is hosted locally by the user, so that commercially sensitive sequences can be analysed and prioritized using the OOPS pipeline.
4.5 Results and Discussion

4.5.1 Computational pipeline for OOPS

Thanks to the structured prioritization protocol and scoring system where all the parameters are considered differently, OOPS provides a remarkable tool for the prioritization where the user can easily define their own pipeline and rank the list of considered clusters according to their specific needs. The use and the biological relevance of the metrics used are discussed below.

4.5.2 Prioritisation Parameters

4.5.2.1 Cluster type

Conservative estimates predict *Streptomyces* spp. to produce over 150,000 specialised metabolites, less than 5% of which have been characterised. The chemical repertoire of these metabolites can predominantly be classified into different classes, hybrid classes, and subclasses thereof. For instance, polyketide synthases (PKSs) can be divided into type I, II, III and *trans* AT PKSs based upon the underlying enzymology of the biosynthetic machinery, but can also form hybrid clusters with non-ribosomal peptide synthases (NRPSs), saccharides and other PKS components, exemplified by the multi-hybrid DNA gyrase inhibitor simocyclinone.

OOPS allows the user to explore and prioritise the entire complement of BGCs defined by antiSMASH within one or more genomes by BGC class, should the user be an expert in one particular class, classes, or subclass, of biosynthetic machinery, using the ‘preferred cluster type’ parameter. The total specialised metabolome can also be prioritised, considering all BGCs irrespective of BGC class / type. We advise filtering according to cluster type to be considered as the first prioritisation step when using OOPS in order to save time and computational resources, although this is not obligatory. BGCs can be delimited by choosing one or more specialised metabolite types as defined by antiSMASH 4.0. Filtering is inclusive of hybrid clusters: the user is not required to select both cluster types should a polyketide-saccharide, or NRPS-PKS, hybrid be of interest. All BGC types / classes defined by antiSMASH 4.0, and 3.0, can be filtered against. Both prokaryotic and fungal genomes can be analysed concurrently, and putative BGCs predicted by the Clusterfinder algorithm e.g. cf_fattyacid, cf_saccharide, can also be used to delimit BGC type. OOPS does not currently support outputs from the sister software, plantiSMASH, or older version of antiSMASH (1.0 and 2.0).

Once the BGC cluster type is defined, and unwanted clusters are filtered, OOPS assigns each BGC 7 different values based on the remaining 7 parameters, the weighting of which, and order (increasing or decreasing) is dictated by the user.

4.5.2.2 Number of genes and CDS length

Gene synthesis remains costly and multiplexing, redesigning and rebuilding of refactored biosynthetic pathways is currently out of reach for most laboratories. Clusters comprising fewer genes may therefore be more attractive candidates when exploring the BGC reconstruction design space in a combinatorial manner, before expanding the rules learnt to
larger BGCs of the same class if necessary. To make a prudent choice of BGC prior to investing

time in wet work it may be necessary to rank BGC candidates against the number of genes

predicted within. Prioritisation of BGC by gene number can be performed by selecting a

weighting for the ‘Number of genes’ parameter in the OOPS user interface (UI). Simply, the

number of genes in each BGC associated Genbank file are counted and corresponding BGCs

are prioritised and pondered based upon this score. The prioritised BGC output is displayed in

the UI either in ascending or descending order depending on the user’s preference.

Whilst gene number is mostly a good proxy for BGC size (Figure 4.3) this is not universal, and
collinear acting cluster such as type I polyketide synthases36 and NRPSs have comparatively
few genes by comparison to the overall coding DNA sequence (CDS) length (Figure 4.3). The
cdaPSI gene, for example, is the largest CDS in the *Streptomyces coelicolor* genome37 at
22391 bp and encodes one of three NRPS megaenzymes responsible for biosynthesis of the
cyclic lipopeptide calcium dependant antibiotic (CDA)38. This CDS alone dwarfs the core
biosynthetic enzymes (minimal polyketide synthase) of the archetypal aromatic polyketide
actinorhodin with a combined CDS length of only 2889 bp.39

To disambiguate gene number from cluster size BGCs can alternatively be prioritised by OOPS
using overall CDS length as a metric. The user is once more able to define the importance of
prioritisation by CDS length across a sliding scale. Each BGC score is calculated by summing
the total CDS length of genes within the cluster defined by antiSMASH and multiplying this
number by the importance value selected.

**Figure 4.3 The relationship between gene number and CDS length**

a, The relationship between gene number and CDS length of 217 BGCs irrespective of type and including
hybrid gene clusters. Gene number loosely correlates with CDS Length. b, The relationship between gene
number and CDS length for NRPSs cluster only, cluster are taken from the dataset displayed in A. Hybrid
clusters are excluded. The correlation between these two metrics for NRPS clusters is poorer than for the
entire dataset.

**4.5.2.3 GC content and codon bias**

As mentioned previously, gene synthesis remains too expensive to refactor multiple complex
and large biosynthetic pathways for many laboratories. Obtaining genomic DNA from the host of
interest, cloning the desired BGC into an appropriate expression vector and expression of the
BGC complement in an appropriate heterologous host therefore remains a powerful tool in
natural product discovery\cite{40}. Success using this strategy relies on a prudent choice of
heterologous expression system as each host has a bespoke suite of transcriptional and
translational requirements which must be accommodated by the recombinant DNA. Two
important metrics underpinning such requirements are the G+C content and codon bias of the
recombinant DNA.

Organisms from the Genus *Streptomyces* have large genomes with a characteristically high
G+C content, ~70 %\cite{37}. The high G+C content of *Streptomyces* spp. protein coding sequences
stabilises mRNA secondary structure\cite{41-43}, alters mRNA half-life and processing, and alters
mRNA interactions with 16S and S1\cite{44} components of the host ribosome, necessary for efficient
translational initiation, these factors generally have a negative impact on protein abundance
when heterologously expressed in an organism with a globally lower G+C content e.g.
*Escherichia coli*. Despite the biotechnological virtues which are afforded by *E. coli* when utilised
as a heterologous host, a more fruitful alternative may be found in different genetically tractable
host with a genomic G+C content more similar to the recombinant DNA in this case.
Accordingly, pairing the BGC G+C content to that of the host organism may be judicious for
preliminary BGC heterologous expression experiments and the importance of this parameter
can be adjusted and prioritised for using OOPS in two fashions. The first enables a user to
prioritise BGC upon their G+C content in ascending or descending order. And, the second
enables a user to prioritise BGCs by similarity to the genomic G+C content of the intended
expression host. After selecting a reference organism BGCs with the most similar G+C content
are automatically ranked in descending order.

Whilst matching BGC G+C content to the average G+C content of the intended heterologous
host may be advantageous at the mRNA level this parameter does not directly consider factors
affecting translation, namely codon usage. The degenerate DNA code permits two species of
DNA with similar average G+C content to encode the same peptide with a significant difference
in codon usage. Therefore, balancing the average G+C content with the heterologous
expression host may circumvent issued related to mRNA structure and interactions but does not
necessarily immediately satisfy important factors facilitating efficient translation. To account for
this, BGCs can be prioritised upon their comparative codon usage frequencies to a reference
host of choice, also.

The full role of codon bias in protein translation is a hotly debated and wide-reaching topic\cite{45}.
The frequencies of rare codons and codon composition within mRNA sequences have been
evidenced to influence translational efficiencies\cite{46,47}, co-translational protein folding\cite{48,49}, alter
protein solubility and function\cite{50,51}, facilitate co-translational translocation\cite{52}, balance
stoichiometry of subunits within heterogenous protein complexes\cite{53} and regulate translating
ribosomes along a given mRNA\cite{54-56}. To reconstitute functional and efficient biosynthetic
pathways in heterologous hosts it is important to match the native translational dynamics as
best as possible. One pertinent example of this is the specific regulation of the TTA codon Leu-
tRNA gene, bldA, in Streptomyces spp. during stationary growth phase\textsuperscript{57}. Whilst most primary metabolic genes in Streptomyces spp. utilise the CTC Leu codon, genes associated with secondary/specialised metabolism are enriched in TTA preventing efficient translation in the absence of sufficient TTA acyl-tRNA\textsuperscript{54}. As a consequence, heterologous expression of genes comprising TTA codons during logarithmic growth phase will result in unfavourable yields of recombinant protein\textsuperscript{56}. Whilst the presence of a TTA codon is computed by antiSMASH 4.0\textsuperscript{17}, it is an opt-in function and not automatically identified. OOPS identifies the presence of TTA codons independently of antiSMASH and shows clusters comprising a TTA codon in red. Accordingly, genomes analysed with antiSMASH versions 1.0-3.0\textsuperscript{58,60} do not need to be recomputed to use this feature. For a more informed view of codon usage, the degree of similarity between each BGC and hosts preferred codon usage can be visualised as a heat map by clicking through the codon bias score.

G+C content and codon bias parameters can be ignored should the user intend on refactoring the BGC in questions, or alternatively used to test refactored clusters built \textit{de novo}.

\textbf{4.5.2.4 Self-similarity}

The recent explosion in synthetic biology relevant molecular biology tools has yielded a suite of DNA assembly techniques which rely on nucleotide homology in some form\textsuperscript{61-64}. Building biosynthetic pathways with genes comprising regions of extended nucleotide similarity can often interfere with DNA assembly, ultimately reducing construction efficiencies through unwanted or incorrect recombination, or mismatch construction events. Once constructed, additional homologous recombination events between DNA regions with extended sequence similarity can introduce deleterious mutations into the biosynthetic pathways, and this is further exacerbated in routinely used, \textit{recA} positive, heterologous host backgrounds e.g. \textit{E. coli} BL21. To improve construction efficiencies, construct maintenance, and reproducibility, sequences with extensive nucleotide sequence similarity are avoided where possible, although this is not a trivial task for the specialised metabolite experimentalist: extended regions of homology are hallmarks of evolving collinear type I PKSs and NRPSs\textsuperscript{65} and are often unavoidable without refactoring the gene sequence.

To circumvent homology associated issues, and to ensure an appropriate BGC is chosen, OOPS computes and ranks BGCs according to nucleotide self-similarity using a modified Smith-Waterman algorithm\textsuperscript{26} enabling the user to prioritise clusters without extended regions of self-similarity. The self-similarity cut-off threshold value, defined in base pairs, can also be customised allowing the user to tailor the prioritisation for different assembly techniques. Here, only BGCs with regions of self-similarity exceeding the threshold are penalised. Self-similarity is often a non-obvious and overlooked parameter however homology can be extensive, for instance two condensation domains within module 8 and module 11 of the CDA NRPS share 747 base pairs (bp) of extensive homology\textsuperscript{38} and adenylation domains in modules 5 and 8 share 898 bp of extended homology\textsuperscript{38}. All homology related data computed by OOPS is deposited, and can be view, in the ‘self-similarity’ directory enabling the user to map these regions directly onto the BGC.

110
4.5.2.5 Phylogenetic diversity and Similarity to known clusters

One criterion used by biologists to prioritise potentially interesting suites of hypothetical genes for characterisation is their broad dissemination across a variety of organisms. The role of such genes may be widely fundamental to propagation, and in light of the topic discussed here could comprise uncharacterised congruent biosynthetic pathways, as was the case for aryl polyene BGCs in the Proteobacteria. One could also argue that genes present only in a particular strain or species may be a rich source of novelty. Or, infer genes present exclusively in organisms isolated only from specific biogeographical or ecological niches to be important in environmental adaptation and specialisation, all of which are plausible assumptions. To explore each of these facets, OOPS enables users to prioritise BGCs based on both their novelty and phylogenetic distribution.

Firstly, BGCs can be prioritised based upon their similarity to characterised BGCs. This parameter uses the KnownClusterBlast score precomputed by antiSMASH to define the novelty of each BGC by comparison to curated entries within the MIBiG repository. Most characterised BGCs can be effectively filtered out by setting known cluster similarity to a high value (>80%) and ranking BGCs in an ascending order. Or, alternatively BGCs can be prioritised based upon a user defined threshold e.g. 40% ± 5, 10, 15% which in combination with additional prioritisation parameters e.g. preferred cluster type, enables identification of BGC homologues with small permutations in BGC gene complement. Such BGCs may encode interesting analogues of known bioactive compounds and therefore be a valuable source of new specialised metabolites. KnownClusterBlast was an important addition in antiSMASH 3.0, and subsequent versions, however the lack of a prioritisation strategy necessitates each BGC to be individually viewed, limiting its utility in large datasets.

Secondly, to observe the taxonomic distribution of organisms harbouring homologous BGCs the user can prioritise BGCs using the taxonomic diversity parameter. A simplistic taxonomic tree is drawn detailing the lineage of organisms containing BGC homologues which is precomputed by the ClusterBlast function embedded in antiSMASH. Each tree can comprise up to 100 entries. Phylogenetically diverse BGCs can be prioritized using a descending ranking, and BGCs only identified in phylogenetically proximal organisms can be prioritized in using an ascending ranking. Whilst the diversity score (DS) doesn’t provide any environmental, ecological or biogeographical context of BGC homologue containing hosts, the corresponding taxonomic tree, visualised by clicking through the DS, provides a good working document to piece together this information.

The ability to visualise and quantify the dissemination of genes between hosts is not easy to achieve. OOPS provides both an informative visual output, in the form of a basic taxonomic tree, and computes an arbitrary, but comparative, value to assess the phylogenetic of BGC homologues.
4.5.3 Exploring Actinobacterial BGCs

The ability of Actinobacteria to produce useful bioactive compounds appears to be a continuum, irrespective of the relatively poor yield of novel bioactive compounds in recent decades. Sequencing of diverse Actinobacteria continues to elucidate the remarkable coding capacity dedicated to specialised metabolism, outstripping organisms from most other phyla. Using this sequence data vast numbers of BGCs are annotated computationally, and potentially encode block buster pharmaceuticals, but ‘left on the shelf’ as experimental characterisation strategies cannot keep pace.

In an effort to expedite the journey from nucleotide sequences to characterised specialised metabolites we have precomputed and deposited prioritisation information for all predicted BGCs from the biosynthetically talented Actinobacteria indexed within the antiSMASH database (5037 BGCs). Precomputing data reduced prioritisation times from over 4 days to less than 10 seconds. The precomputed data is freely available to all users along with a walk through at https://zenodo.com/OOPS1_RBreitling.

4.6 Conclusion

The number of published and deposited biosynthetic gene clusters continues to grow, and the ability to retrieve this data is paramount for scientists within the specialised metabolite field. Accordingly, the framework in which this data is held, standardised, curated and viewed has received much attention and sparked the development of multiple supporting softwares. Despite this, to our knowledge, only one other published software enables prioritisation of BGCs: the Antibiotic Target Seeker (ARTS). The need to prioritise BGCs is widely accepted however the parameters of interest will duly change in accordance to the interests of each research group, therefore it is unlikely that one prioritisation software will be a one-stop-shop considering all needs. In light of this, the intended use of ARTS is towards the discovery of new antibiotics. Through identification of putative resistance mechanisms within BGCs, ARTS infers the target of the cognate specialised metabolite and enables the user to prioritise BGCs accordingly. Exploitation of self-resistance has been extensively used to enrich libraries of antibiotic producing organisms, and to good effect, predicting ARTS to be an important tool in our bioinformatics arsenal. However, to date, all references datasets used by ARTS are Actinobacterial limiting its usability. In opposition to identification of specialised metabolite with specific bioactivities e.g. antibiotics from Actinobacteria, OOPS considers all BGCs and enables the user to prioritise any antiSMASH computed BGC based upon molecular biology parameters. In this vein, OOPS aims to expedite specialised metabolite identification through prudent choice of BGC for experimental characterisation, where unfavourable translational, transcriptional and recombination characteristics can be weighted against, and BGCs can be fitted to an appropriate choice of heterologous host improving experimental success without extensive experimental recapitulation. Furthermore, the ability to prioritise BGCs based upon similarity to know clusters is an important ‘living’ filter to avoid specialised metabolite rediscovery and will be continually refined as more characterised BGCs are reported using the MIBiG standard. Lastly, whilst the phylogenetic diversity of organisms containing homologous BGCs is not a molecular
biology relevant parameter *per se*, ranking BGCs using this parameter may help elucidate new candidate heterologous expression host with a metabolism capable of supporting biosynthesis. Additionally, this parameter provides a targeted route to identification of specialised metabolites with ecologically inferred bioactivities, previously shown to be a fruitful discovery approach.

OOPS answers the need of wet lab biologists by linking the numerous, and undoubtedly valuable, BGCs predicted from the growing number of publicly available genomes with molecular biological caveats. In doing so, we hope to improve experimental success rate when characterising specialised metabolites. OOPS will continue to grow in complexity and richness in accordance with updates to the antiSMASH pipeline and the MIBiG standard remaining current without the need to extensive manual updates.
4.7 References


5 Assembling a modular production line for combinatorial biosynthesis of aromatic polyketides in *Escherichia coli*

Matthew Cummings, Rainer Breitling and Eriko Takano*

Manchester Synthetic Biology Research Centre SYNBioCHEM, Manchester Institute of Biotechnology, School of Chemistry, The University of Manchester, UK

*Corresponding author: Eriko Takano, Manchester Institute of Biotechnology, The University of Manchester, 131 Princess Street, Manchester, M1 7DN, UK.

**Tel:**+44 (0)161 306 4419.

**E-mail:** eriko.takano@manchester.ac.uk.

**Keywords:** Refactoring, plug-and-play biology, combinatorial biosynthesis, polyketide synthase, aromatic polyketides

5.1 Preface

All experiments were designed and undertaken by M.C.. The first draft of the manuscript was authored by M.C. and all authors finalised the manuscript.

The manuscript is formatted in line with all other chapters of the thesis and uses Nature style citation. In the interest of space the citations are single spaced. Supplementary material is frequently referenced within the manuscript and is attached to the end of this chapter and not the end of the thesis. An extended discussion is included and will differ from the shorter version present in the published manuscript. All numbered compounds are descrete to this chapter.
5.2 Abstract

Polyketides are a class of specialised metabolites synthesised by both Eukaryotes and Prokaryotes. These chemically and structurally diverse molecules are heavily used in the clinic and include frontline antimicrobial and anticancer drugs such as erythromycin and doxorubicin. To replenish the clinicians’ diminishing arsenal of bioactive molecules, a promising strategy aims at transferring polyketide biosynthetic pathways from their native producers into the biotechnologically superior host Escherichia coli. This approach has been successful for type I modular polyketide synthases; however, despite more than three decades of research, the large and important group of type II polyketide synthases has until now remained intractable in E. coli, hampering progress in this direction.

Here we report a new strategy to identify E. coli-compatible type II polyketide synthases. Using this approach, we successfully expressed a functional ketosynthase and chain length factor pair from Photorhabdus luminescens TT01 as soluble heterodimeric recombinant proteins in E. coli for the first time. We characterize the mechanism of the heterologous PKS, defining its minimal polyketide synthase components, and utilise this biosynthetic pathway to synthesise anthraquinones and dianthrones. Furthermore, we demonstrate the remarkable tolerance and promiscuity of the heterologous biosynthetic pathway in E. coli to act as a plug-and-play scaffold, showing it to function successfully when complemented with biosynthetic enzymes from phylogenetically distant species. By modularising the biosynthetic pathway to aromatic polyketides into four key stages, we show how chemical diversity can be introduced at each biosynthetic step.

This work enables combinatorial biosynthesis of aromatic polyketides using type II polyketide synthases in E. coli the first time, making full use of its many advantages in terms of easy and fast genetic manipulation and convenient biotechnological scale-up. Using the synthetic and systems biology toolbox in this novel plug-and-play biosynthetic platform will serve as an engine for the production of new and diversified bioactive polyketides in a rapid and versatile fashion.
5.3 Introduction

Natural products and their derivatives and mimics represent a lion’s share of clinically used therapeutic agents, accounting for 73% of antibacterial agents and 83% of anticancer agents approved by the FDA between 1981 and 2014\(^1\). Polyketides represent a central class of these natural products, with remarkably broad, potent, and diverse pharmacological properties. Although their native biological role is still debated, polyketides continue to have significant medical value as potent antitumoral agents, antibiotics, immunosuppressants, antiparasitics and cholesterol-lowering agents among others\(^2\).

The chemistry underpinning polyketide biosynthesis is widely conserved and carried out by biosynthetic machinery of three major classes\(^3\). In almost all cases, the polyketide biosynthesis machinery is highly modular at the genetic, enzymatic and chemical level\(^4\). This intrinsic modularity of polyketide synthases (PKSs) was a key motivation behind classical approaches to derivatization of natural products, and for the same reason PKSs have been favorite targets for the recent pathway engineering and natural product derivatization renaissance using synthetic biology\(^5\). The modularity observed throughout the type I PKS enzyme superfamily led to visions of pathway legoisation\(^6\) where discrete module or domain ‘blocks’ could be dropped in and out of model biosynthetic pathways to rationally generate ‘unnatural’ natural products\(^7\). However, reprogramming of type I PKS biosynthetic pathways using this approach was generally disappointing and often resulted in minimal returns\(^8\), because the highly integrated and complex protein–protein interactions which coevolved between the covalently linked PKS components are poorly understood\(^9\), do not typically allow crosstalk between pathways, are difficult to mimic, and often produce end compounds with diminished bioactivities\(^8\). Using a more dissociated and discrete biosynthetic machinery, based on type II PKSs (Figure 5.1), would circumvent this problem and remove the need to manipulate and redesign domains within highly integrated enzymes. However, despite major efforts over several decades, heterologous overexpression of the type II PKS machinery, specifically the minimal polyketide synthase (mPKS) comprising a ketosynthase (KS) and chain length factor (CLF) and acyl carrier protein (ACP) (Figure 5.1), in the biotechnologically favorable host species *Escherichia coli* has remained elusive.
Figure 5.1 Biosynthesis of archetypal aromatic polyketides.

The central dark grey pane (module 1) illustrates biosynthesis of the nascent poly-β-ketide chain by the KS:CLF heterodimer which remains tethered to the ACP. The red moiety represents the starter unit, which can be altered for increased product diversity. The left blue pane shows module 2: functionalization of the acyl carrier protein necessary to initiate polyketide biosynthesis via phosphopantetheinylation of the apo-ACP to holo-ACP catalyzed by a 4' phosphopantetheinyl transferase (PPTase), and acylation of the holo-ACP with the starter unit, e.g., via a malonyl-CoA:ACP transacylase (MCAT), Acyl-acyl carrier protein synthetase or via self-malonylation. The top right light grey pane (module 3) illustrates the typical primary tailoring enzymes, KR: ketoreductase, ARO: aromatase and CYC: cyclase that are responsible for biosynthesis of the aromatic carbon core scaffold. The bottom right yellow pane indicates the point at which the modified polyketide chain is released from the acyl carrier protein. The free metabolite can be further modified by secondary tailoring enzymes (module 4), e.g., by glycosylation, prenylation, or halogenation to introduce chemical diversity.

An engineered iterative type I fungal PKS from Gibberella fujikuroi has been expressed in E. coli together with exogenous bacterial derived tailoring enzymes to produce the nonaketide SEK2610,11. However, expanding the chemical space is difficult when using a less well understood iterative fungal PKS: to alter the polyketide chain length an entirely new PKS would be required with a fungal system10,12, whilst chain length can be altered easily using the bacterial type II machinery by including a small set of mutations13. Furthermore, without a starter unit loading domain (SAT) fungal PKS megasynthases cannot introduce important non-acetyl starter units into aromatic polyketides, whilst it is known that these can be easily introduced to dissociable bacterial type II PKS pathways14,15. The aromatic polyketide oxytetracycline has been detected in E. coli harbouring a recombinant oxytetracycline BGC, and overexpressing the alternative sigma factor σ54; however, the key enzymes, OxyA (KS) and OxyB (CLF), were not detectable among the soluble or insoluble proteins16, and oxytetracycline titers were low. Other attempts to achieve mPKS expression in E. coli have either resulted in unobservable expression or inactive inclusion body formation11,16,17. The cause of insolubility has not been experimentally
characterized, but inharmonious rates of translation, protein folding and heterodimerisation of the KS/CLF have been cited as the cause. The intractability of this class of bacterial PKS in E. coli has clearly hampered progress in the field, and necessitated the use of Streptomyces expression hosts.

Here we report a strategy for identification and expression of a soluble and heterodimeric bacterial type II PKS in E. coli for the first time, opening the door to successful biochemical diversification and biotechnological exploitation of polyketides in a versatile and tractable heterologous expression host. We deconstruct the biosynthetic route to aromatic polyketide biosynthesis into four key modules (Figure 5.1), and by satisfying the enzymatic requirement of each module we establish a functional and versatile biosynthetic route to multiple polyketide core scaffolds. We exemplify the value of this platform as a plug-and-play scaffold by demonstrating remarkable tolerance and promiscuity of the recombinant biosynthetic pathway to function successfully when complemented with sequence diverse and structurally diverse homologues from phylogenetically distant species. This work outlines the development of a novel E. coli based platform from which we can easily derivatise novel aromatic polyketides in a combinatorial fashion using a highly modular approach (Figure 5.2).
Figure 5.2: Schematic representation of modularity of aromatic polyketide biosynthesis

**a**, Expected octaketide shunt metabolites are designated by grey dotted arrows below each module. Each module is defined by a grey box and the boxes proceed in the order of biosynthesis. Circular arrows represent the ability to swap enzymes with homologues within each module. **b**, Examples of plausible biosynthetic pathway perturbations for introduction of chemical diversity: **b1**, the exchange of an octaketide producing polyketide synthase heterodimer with a decaketide producer, **b2**, Removal of the ketoreductase (KR), **b3**, alteration of polyketide starter unit by swapping enzymes in module 2, Aro/Cyc module exchange
and loss of the supplementary module. The supplementary module is an additional module which can be variable in function. c, Structures of AQ256 (1) and its dianthrone (13).

5.4 Results

5.4.1 Module 1: Identification of candidate ketosynthase: chain length factor dimer pairs for heterologous expression in E. coli

Identification of a catalytic 'module' tasked with aromatic polyketide biosynthesis in E. coli represented the first hurdle in this work. Both type I and type II PKSs have been the subject of in-depth evolutionary modelling and phylogenetic analysis in recent years\textsuperscript{19-22}. Phylogenetic analysis of large datasets of type II polyketide synthases indicates that canonical type II PKS ketosynthase (KS) and chain length factors (CLF) pairs arose from an ancient KS duplication event, most likely from a FabF-like fatty acid ketosynthase (FASII)\textsuperscript{19,22}. Therefore, the intrinsically soluble FabF protein from E. coli, sharing a common ancestor with canonical type II PKSs, was used to query candidate ketosynthase pairs for heterologous expression in E. coli. To do so, a dataset of 58 experimentally characterized type II KS sequences was acquired from the MiBIG repository\textsuperscript{23} and aligned with three FabF candidates from Streptomyces avermitilis, Bacillus subtilis and E. coli. Ketosynthase sequences were chosen, as these represent the catalytic half of the polyketide dimer and are more similar to FabF than the passive, and typically more sequence diverse, chain length factors. Phylogenetic reconstruction of the sequence alignment identified two ketosynthases, RemA from the resistomycin BGC and AntD from the anthraquinone BGC, to associate more closely with the FabF homologues than all other ketosynthase sequences acquired from the MiBIG dataset (Figure 5.3a); both KS and cognate CLF pairs were plausible first candidates for successful heterologous expression in active and soluble form in E. coli.

Biosynthesis of shunt metabolites by the resistomycin minimal PKS was previously shown to require co-expression of additional cyclases\textsuperscript{24}. This context dependency may limit the chemical diversity accessible through combinatorial biosynthesis of early biosynthetic shunt metabolites, which are shown to have valuable bioactivities\textsuperscript{25}. In contrast, the AntD-containing BGC, responsible for biosynthesis of anthraquinone pigments in the nematode symbiont Photorhabdus luminescens TT01, was a more attractive candidate: biosynthesis of its U-shaped polyketide core was proposed to be enzymatically congruent with biosynthesis of actinorhodin, the archetypal aromatic polyketide\textsuperscript{26}, and consequently was analysed first.

5.4.2 Evaluating solubility and dimer formation of the AntD and AntE in E. coli

Remarkably, the entire minimal PKS complement from P. luminescens, comprising the KS (AntD), CLF (AntE) and ACP (AntF), were successfully expressed as soluble recombinant proteins in E. coli BL21(DE3). Furthermore, AntD and E were observed as soluble proteins when overexpressed at 20°C and 30°C for >12 h, indicating transcription, translation and protein folding to be robust (Figure 5.3c)
Whilst AntD and E are both soluble recombinant proteins in *E. coli*, the role of AntE as a functional heterodimer partner for AntD was unknown. Sequence features of AntE defy conventions of characterized CLFs: alignment of CLF amino acid sequences in our dataset showed AntE exist outside of the canonical CLF clade (Figure 5.3b), and to also lack hallmark and gatekeeper residues which play important roles in polyketide biosynthesis (5.8.1, Supplementary Figure 5.1, Supplementary Figure 5.2, Supplementary Table 5.1). Most notably, the C-terminal third of AntE shows no sequence similarity to any CLF within our dataset; sequence divergence here might indicate that AntE is degenerate and no longer functional.

**Figure 5.3: Identification and expression of the *P. luminescens* KS AntD and CLF AntE in *E. coli.*

a and b, Phylogeny of KS and CLF sequences from a dataset of 58 characterized type II PKSs derived from the MiBIG repository, respectively. The clades representing canonical type II KS and CLF are denoted by red and yellow wedges, respectively. Both alignments include FabF sequences from *Streptomyces avermitilis, Escherichia coli* and *Bacillus subtilis*. Red and yellow dots denote His<sup>6</sup>-AntE and AntD, respectively. This coloring is conserved throughout the figure. Bootstrapping values below 50 are not shown. c, Denaturing PAGE showing soluble protein extracted from *E. coli* BL21(DE3) (lane 3) and *E. coli* BL21(DE3) pBbA2k-plumPKS, harboring the minimal PKS from *P. luminescens*, (lane 2) both induced at 30°C. Lanes 5 and 4 mirror those of 3 and 2; however, they show soluble protein expressed post incubation at 20°C. d, Denaturing PAGE gel of AntD and E purified by IMAC. Lane 1: protein flow through, lane 2: protein eluted at 50 mM imidazole, lane 3: 500 mM imidazole column wash. e, Western blot of
puriﬁed AntDE protein showing signal corresponding to a single AntE band. Lane 1: PAGE ladder, 2: Purified AntDE protein and 3: His-tagged mCherry (~29 kDa) fusion protein as positive control. All numbers correspond to standard protein ladders and are deﬁned in kDa. Theoretical size of His6-AntE and AntD is 42.43 kDa and 46.16 kDa, respectively.

To examine if a heterodimeric complex is formed by AntD and AntE, a His6 fusion of antE was co-expressed with antD in E. coli BL21(DE3) (5.9.2.2). Protein puriﬁed via immobilized metal ion afﬁnity chromatography (IMAC) and visualized by denaturing PAGE showed two distinct bands corresponding closely with the theoretical molecular weight of AntD and His6AntE and both of similar intensity (Figure 5.3d). Western blotting (Figure 5.3e) and LC/MS-MS analysis (5.9.3) of both bands conﬁrmed these to correspond to AntD and AntE, respectively. Co-purification of AntD with His6AntE agrees with the assumption that stable AntDE hetrodimers are formed in E. coli.

5.4.3 Testing functionality of the PKS (Module 2)

Expression of a soluble heterodimeric KS:CLF complex in E. coli is unprecedented and was the ﬁrst step towards development of an E. coli-based combinatorial biosynthetic platform for aromatic polyketide biosynthesis. We next sought to test the functionality of the AntDE heterodimer. To do so, the components of the P. luminescens anthraquinone mPKS were expressed in E. coli BL21(DE3) using the plasmid pBB1-a-plumPKS (Supplementary Table 5.2). Previous studies had shown that a P. luminescens TT01 strain lacking one anthraquinone-associated cyclase to accumulate mutactin and dehydromutactin, suggesting the mPKS to synthesize a 16-carbon octaketide primed with an acetyl starter unit26. Thus, the expected shunt metabolites formed by the minimal PKS were the acetyl-primed octaketides SEK4, SEK4b and their respective dehydrated forms (Figure 5.2, Supplementary Figure 5.3).

Expression of antDEF in E. coli BL21(DE3) did not produce any detectable masses corresponding to SEK4 / SEK4b (Figure 5.4, sample III). A lack of detectable octaketides suggested E. coli endogenous 4’-phosphopantetheinyl transferases, AcpS and EntD, could not efﬁciently functionalize AntF with a 4’ phosphopantetheine arm necessary for activity, and alternative auxiliary enzymes are required to fulﬁl this post-translation modiﬁcation (Module 2, Figure 5.2). Co-expression of antDEF with the anthraquinone-associated PPTase, antB, and CoA ligase, antG, conﬁrmed this assumption, resulting in detectable biosynthesis of molecules putatively identiﬁed as SEK4, SEK4b and AUR367 (Figure 5.4, sample V): MS2 fragmentation patterns of the putatively assigned octaketides were consistent with values reported in the literature27 (Supplementary Figure 5.4). Interestingly, in the absence of AntG, the putatively assigned CoA ligase, a decrease in SEK4 and SEK4b relative ion intensities was observed when compared with the antDEFBG expressing strain. This reduction in metabolite concentration indicates AntG to function in the transacylation of holo-ACP26, and to be a necessary component of Module 2 (Figure 5.1).
Figure 5.4: Expression of AntDEFBG in *E. coli*.

a and b, Three-dimensional and two-dimensional extracted ion chromatograms of all observable masses between 301.0661 and 301.0714 m/z: the theoretical masses of expected shunt metabolite [M+H]^+ adducts of AUR367 (5) and [M-H2O+H]^+ SEK4 (2) is 301.07067. c and d, three dimensional and two dimensional extracted ion chromatograms of all observable masses between 319.0764 and 319.086 m/z: The theoretical mass of expected shunt metabolites SEK4 (2) and SEK4b (3) is [M+H]^+ 319.08099, (15 ppm tolerance). Samples I through V are filtered supernatant from cultures of: I *E. coli* BL21(DE3), II *E. coli* BL21(DE3) pBbB1a-GFP pACYCDuet-1 (empty vector control), III *E. coli* BL21(DE3) pBbB1a-plumPKS pACYCDuet-1 (AntDEF), IV *E. coli* BL21(DE3) pBbB1a-plumPKS pACYC93 (AntDEFB) and V *E. coli* BL21(DE3) pBbB1a-plumPKS pACYC8893 (AntDEFBG). Extracted ion count was normalized to final cell density (OD₆₀₀). All bolded numbers correspond to Figure 5.2.

### 5.4.4 Module 3: Exploring end-compound production of the anthraquinone biosynthetic gene cluster

In addition to the mPKS (Module 1) and auxiliary CoA ligase and PPTase (Module 2), primary tailoring enzymes are common to all aromatic PKS pathways forming principal components of Module 3 (Figure 5.2). The anthraquinone biosynthetic gene cluster is no different and encodes 4 other enzymes with putative assigned functions, including a C9 ketoreductase, two cyclases and a hydrolase / peptidase. The full complement of biosynthetic genes was predicted to produce 1,3,8-trihydroxyanthrone (Figure 5.2 compound 12), and – in a similar fashion to aurachin biosynthesis\(^{29,30}\) – additional tailoring enzymes responsible for further modification of the polyketide core are thought to exist in extension clusters situated elsewhere on the *P. luminescens* genome. We sought to identify if Modules 1 through 3, AntA-I, were functional in *E. coli* and able to produce the rationalized anthrone (Figure 5.2 compound 12). To do so the entire anthraquinone cluster (accession No. BX470251.1, MIBiG No. BGC0000196, 9166 bp)
was introduced into the duel expression vector pACYCDuet-1 and expressed in *E. coli* BL21(DE3) (5.9.2). The exometabolome of the resulting strain was analyzed for the production of key expected octaketide shunt metabolites as well as plausible octaketide end products (5.8.3, 5.8.6).

Masses corresponding to 1,3,8-trihydroxyanthrone (Figure 5.2, compound 12) were only observed at trace levels using high resolution MS by both positive and negative electrospray ionization. Anthrone natural products have previously been shown to form their cognate anthraquinone or dianthrone either via spontaneous oxidation31 or enzymatically (Supplementary Figure 5.5). Should the trihydroxylated anthrone follow the same oxidation pathways in *E. coli*, or during the extraction process, masses corresponding to AQ256 and 1,3,8-trihydroxydianthrone (Figure 5.2, compound 1 and 13) would be expected. Both oxidized metabolites were identified from the exometabolome of *E. coli* BL21(DE3) antA-I. A targeted search led to the observation of AQ256 as the major product of the extended anthraquinone BGC (Figure 5.5, Supplementary Figure 5.6) which was fully characterized by NMR spectroscopy (Supplementary Figure 5.7, Supplementary Figure 5.12); high-resolution mass spectrometry, tandem MS, where fragmentation patterns for AQ256 follow those of other anthraquinones (Supplementary Figure 5.13) and UV/Vis absorbance in agreement with similar anthraquinones (λ<sub>max</sub>: 244, 265, 284 and 434 nm) (Supplementary Figure 5.14). Interestingly, whilst quinone formation is proposed to be catalyzed by the *plu0947* gene product in *P. luminescens* TT01<sup>26</sup>, the absence of *plu0947*, or other quinone forming monooxygenase (ActVA-ORF5/ActVB or ActVA-ORF6<sup>32</sup>) homologues within the *E. coli* BL21(DE3) genome indicates the quinone-forming oxygen is either introduced by an unknown alternative endogenous enzyme or through non-enzymatic oxidation as proposed for cladofulvin biosynthesis<sup>31</sup>. More extensively modified AQs isolated from *P. luminescens* TT01<sup>26</sup> were not identified as end compounds in the engineered *E. coli* strains, consistent with the absence of additional cognate tailoring enzymes (Supplementary Figure 5.15, 5.8.7).
Figure 5.5: Anthraquinone identification and characterization.

(a) Extracted ion chromatogram (EIC) for shunt metabolites described in Figure 5.2a, in addition to anthraquinones produced by *Photorhabdus luminescens* TT01, (AQ270a, b, AQ284a, b, AQ300 and AQ314), and predicted pathway end compounds. All EICs represent theoretical mass for [M-H] ± 5 ppm. Red, blue and black lines represent EICs of *E. coli* BL21(DE3) (host control), *E. coli* BL21(DE3) pACYCDuet-1 (plasmid control) and *E. coli* BL21(DE3) pACYCAnthraquinone normalized to final cell density (OD600).

(b) Schematic of full characterization of AQ256 by COSY (green), HSQC (grey) and HMBC (blue scale) 2D NMR. NMR characterization is detailed in full within supplementary data Supplementary Figure 5.7 – 5.12.

Masses corresponding to two dianthrones were also identified and characterized by high resolution MS. MS² spectra of both dianthrones show fragmentation to occur at the C10 – 10’ bond forming anthrone radicals (Supplementary Figure 5.16): a hallmark fragmentation pattern of a wide variety of glycosylated and aglycone dianthrones. Additionally UV-Vis absorbance of putative diathrones 1 and 2 showed similarities to emodin dianthrone with λ_max at 359, 263, 217 nm and λ_max 358, 275, respectively (Supplementary Figure 5.17). Full characterization by NMR was not possible, as neither compound was present in sufficient quantities.
5.4.5 Evaluation of a plug-and-play scaffold

5.4.5.1 Promiscuity in Module 3: Complementation of the C9 ketoreductase

A biosynthetic route to two pharmaceutically important octaketide scaffolds, anthraquinones and dianthrones was now established using type II PKSs in *E. coli*; however, to generate large libraries of privileged aromatic polyketide derivatives a plug-and-play scaffold in which functionally similar and divergent biosynthetic genes can be swapped in and out is necessary to generate diversity (Figure 5.2b).

During aromatic polyketide biosynthesis, the growing polyketide chain is tethered to an ACP, in this case AntF. To reach compound maturation the ACP tethered polyketide chain must be delivered to each biosynthetic enzyme within the biosynthetic pathway sequentially (Figure 5.1) through formation of specific protein–protein interactions. For a plug-and-play platform to work AntF must successfully form interactions with a multitude of non-cluster associated tailoring proteins in Modules 1 and 3. The promiscuity of AntF is therefore the major bottleneck and the key determinant in the success of the AntA-I cluster as a generic platform for aromatic polyketide derivatization in *E. coli*: the unusual charge character of the AntF recognition helices, and an uncommon DST pantetheinylation motif, cast doubt on its ability to do so.

To elucidate if uncommon ACP characteristics hamper AntF from functioning outside of the anthraquinone BGC the cluster associated C9 ketoreductase, AntA, was functionally replaced with ActIII (*sco*5086), a C9 ketoreductase from the phylogenetically distant actinorhodin BGC. Two vectors comprising the ant cluster but lacking *antA* were constructed; the first replaced *antA* with a fully refactored ActIII gene (*sco*5086) using *E. coli* BL21 codon preference (DNA optimization using GeneOptimiser, GeneArt. DNA synthesized by Gen9) and the second replaced *antA* with a modified wild type ActIII gene harboring two 5’ synonymous mutations.

Complementing AntA in the anthraquinone BGC with ActIII restored AQ256 biosynthesis in *E. coli* BL21(DE3) when expressed as either a refactored or wild type *sco*5086 gene sequence (Figure 5.6a). Removal of *antA* without subsequent addition of a KR abolished AQ256 biosynthesis (Figure 5.6a, ketoreductase AntA knockout mutant pACYCAntΔAntA, Supplementary Figure 5.18a and b), indicating AntF to successfully deliver biosynthetic intermediates to and from ActIII *in vivo*. In the ΔAntA host the expected shunt metabolites SEK4 and SEK4b accumulated at much higher intensities compared to cultures expressing the entire ant BGC normalized to final cell density (Supplementary Figure 5.18c and d), as expected for an interrupted biosynthetic pathway.
Figure 5.6: Complementation of the Anthraquinone BGC with Actinorhodin components

a, Typical exometabolome HPLC profiles of E. coli BL21(DE3), E. coli BL21(DE3) expressing empty vector pACYCDuet-1, C9 ketoreductase AntA knockout mutant pACYCAntΔAntA, wild type actII C9 KR complemented plasmid pACYCAntwtkR, refactored actIII C9 KR complemented plasmid pACYCAntrefKR and pACYCAntAntraquinone compared with fully characterized AQ256 standards and M9 growth media at 434 nm. The UV-vis spectrum for peaks designated 1 are as follows; AQ256 standard λ_{max} at 216, 264, 282, 434, 583 nm, E. coli BL21(DE3) pACYCAntAntraquinone λ_{max} 215, 263, 283, 434 nm, E. coli pACYCAntrefKR λ_{max} 215, 283, 435, E. coli BL21(DE3) pACYCAntwtkR λ_{max} 216, 282, 435, 585 in agreement with AQ256 (Supplementary Figure 5.14). Spectra were limited to 215-600 nm. b, Typical exometabolome HPLC profiles of E. coli BL21(DE3), E. coli BL21(DE3) expressing empty vector pACYCDuet-1, ARO/CYC knockout mutant pACYCAntΔAntH, wild type Sco5090 ARO/CYC complemented plasmid pACYCAntwCYC, refactored Sco5090 ARO/CYC complemented plasmid pACYCAntrefCYC and pACYCAntAntraquinone compared with standards as above. Compound reference numbers are as described in Figure 5.2; 1: AQ256, 11: Aloesaponarin II and 10: 3,8-dihydroxy-methylnthraquinone carboxylic acid (DMAC). The UV-vis spectrum for the three most abundant peaks are as follows: 1; λ_{max} at 216, 264, 282, 434, 583 nm in agreement with AQ256 (Supplementary Figure 5.14), 10; λ_{max} at 217, 408, 650 nm, 11; λ_{max} at 214, 277, 409, 582 nm with tR of 16.004 min (960 sec), 14.62 min (877.2 sec) and 16.742 min (1004.5 sec) respectively. c, Typical exometabolite HPLC profile of E. coli BL21(DE3), E. coli BL21(DE3) expressing empty vector pACYCDuet-1, an hydrolase/peptidase knock out mutant pACYCAntΔAntI, E. coli BL21(DE3) harboring pACYCAntAntraquinone, and antaquinone standards. All numbering and λ_{max} figures are as reported in b. d, by disrupting the function of final ring closure a bicyclic intermediate is produced. Modification of this bicyclic intermediate will enabled
biosynthesis of benzioisochromanquinone core structures e.g. actinorhodin, granaticin and medermycin, opposed to anthraquinone and dianthrone core structures.

5.4.5.2 Promiscuity in Module 3: Cyclase / aromatase complementation

In addition to functionally replacing AntA by ActIII, the structurally unique tridomain cyclase/aromatase, AntH, was successfully functionally replaced by the well-characterized didomain cyclase ActVII (sco5090) from the actinorhodin biosynthetic gene cluster. Two more constructs were built following the same strategy as the ketoreductase replacement; the first harbored a refactored ActVII gene (sco5090), and the second the wild type with a starkly different GC content, 73% by comparison to 36% average GC content of the antA-I cluster.

Functional replacement of AntH by ActVII (derived from either the refactored or wild type Sco5090 sequence) restored AQ256 biosynthesis, not observed in the ΔAntH host (Figure 5.6, Supplementary Figure 5.19a and b) evidencing the promiscuity of AntF to interface with structurally different enzymes in vivo. Interestingly, in both ActVII complemented strains the actinorhodin shunt metabolites aloesaponarin II and DMAC were also observed (Figure 5.6b, Supplementary Figure 5.19g-j). Identification of all three end products indicates that the maturing polyketide chain successfully undergoes congruent reduction, aromatization and cyclisation to form a bicyclic intermediate before differing in mechanism of chain release and final ring cyclisation. The observation of all three metabolites is evidence of disrupted function in final ring closure; however the complexities of the anthraquinone BGC cyclases are beyond the scope of this article. Polyketide cyclases comprise either one or two cyclases domain i.e. mono- or di-domain cyclases: AntH is the first reported tridomain cyclase and the involvement of each domain in ring closure is currently being evaluated. Deletion of plu4186, encoding AntI the hydrolase / peptidase, proposed to be involved in acyl-ACP release or final ring formation, confirmed the enzymatic congruence up to formation of a bicyclic intermediate. E. coli BL21(DE3) deficient in the AntI, but expressing antA-H, no longer produced AQ256, but rather produced aloesaponarin II and DMAC as end products of the truncated biosynthetic pathway exclusively, (Figure 5.6c, d) concluding pre-AntI to be the branch point between anthraquinone and dianthrone formation and BIQ biosynthesis.

5.5 Discussion

Being able to synthesize type II polyketides in E. coli opens up exciting avenues for the rapid and versatile diversification of novel bioactive compounds. The plug-and-play platform described here can be seamlessly integrated into a highly-automated design–build–test cycle which is currently unachievable in the slow-growing, poorly characterized and often genetically intractable native Actinobacterial hosts. The scope of this plug-and-play platform is yet to be fully realized however is fitting with two criteria necessary for a good drug discovery platform: the underlying enzymology comprises discrete and dissociated biosynthetic parts, which are well suited to combinatorial biosynthesis, arguably more so than their type I counterparts. And, aromatic polyketides, the output of these pathways, are privileged bioactive specialized metabolites which are heavily used in the clinic.
Furthermore, due to the pharmaceutical importance of aromatic polyketides a wealth of literature describes how to modify polyketide chain length\textsuperscript{13}, introduce novel starter units\textsuperscript{45}, and alter cyclisation patterns\textsuperscript{46} through rational point mutations and addition, substitution, or deletion of biosynthetic genes in type II BGCs; all of which can now be introduced / harnessed using our plug-and-play platform. As proof-of-principle we demonstrate a tolerance for substitution and deletion in module 3 (Figure 5.1, Figure 5.2). Firstly, by replacing the anthraquinone BGC-associated AntA (C9 ketoreductase) with ActIII (sco5086 ketoreductase from the actinorhodin BGC), as well as replacing AntH (cyclase/aromatase) by the well-characterized di-domain cyclase/aromatase ActVII (Sco5090, also from the actinorhodin BGC)\textsuperscript{39}, we restore AQ256 production in \textit{E. coli} evidencing the promiscuity of AntF to interface successfully with non-cluster associated enzyme homologues and structurally diverse biosynthetic homologues. And secondly, through gene deletion (plu4186 – AntI) we divert polyketide biosynthesis away from anthraquinone formation and towards DMAC (10) and aloesaponarin II (11), shunt metabolites of all benzoisochromanequinones (BIQs) e.g. actinorhodin, granaticin, and medermycin (Figure 5.6d). This enzymatic switch, which is now possible with our plug-and-play scaffold, enables access to three important polyketide core scaffolds by driving the flux towards biosynthesis of either AQ256, its derivatives and dianthrones as shown in this manuscript, or towards known, and new, benzoisochromanequinones, and their associated shunt metabolites.

The major challenge in establishing this platform was the identification of soluble mPKS systems. Instead of using a trial-and-error approach, we used evolutionary insights into the formation of type II PKSs\textsuperscript{19} to identify suitable KS/CLF pairs. Remarkably, the first KS/CLF pair tested was soluble using this approach. The properties affording solubility of the AntDE heterodimer are not known but could possibly be attributed to the phylogenetic proximity of \textit{E. coli} and the native AntDE endogenous host, \textit{Photorhabdus luminescens}. Phylogenetic proximity is plausibly a proxy for heterologous protein solubility. Firstly, the intracellular environment and protein maturation machinery might be less divergent between phylogenetically proximal host: the homeostatic interior of the cell is a stringent selection pressure influencing protein evolution\textsuperscript{47}. Additionally, recombinant protein solubility might be influenced by genomic DNA GC content of the endogenous host. Increased arginine-to-lysine ratios of proteins overexpressed in \textit{E. coli} correlated with insolubility\textsuperscript{48}, whilst proteins from notoriously GC-rich actinobacteria are expected to be enriched in arginine, as arginine codons are rich in guanine and cytosine compared to lysine codons. No one descriptor is likely to be the causative factor behind insubility of KS/CLF dimers in \textit{E. coli} and further Gram negative derived type II PKSs will help validate or disprove the influences of the above hypotheses. Identification of Module 1 (M1) represented a historic hurdle in this work however this is nonconsequential without a soluble ACP functionalized by components of Module 2 (M2).

Module 2 comprised enzymes necessary for post transcriptional functionalization and acylation AntF, the anthraquinone ACP e.g PPTase, MCATs, AasSs and CoA ligases (Figure 5.1). These pre-biosynthetic enzymes are integral for function, as without them no substrates can be loaded, tethered or delivered to M1 or M3 by the ACP. Despite this, many BGCs do not encode
dedicated PPTases or MCATs, as orthologues from primary metabolism can crosstalk to specialized metabolism. The promiscuity of E. coli BL21 endogenous primary metabolic PPTase (AcpS) and MCAT (FabD) does not extend to AntF, however. Characteristic residues and motifs forming key electrostatic and hydrophobic interactions known to orientate ACPs to the active site of MCAT and AcpS are poorly conserved in AntF. The uncharacteristic amino acid composition of AntF may prevent meaningful ACP:protein interactions thereby stopping functionalisation and acylation by E. coli endogenous enzymes. This phenomenon was previously reported for SMb20651, an ACP from Sinorhizobium meliloti, which shares the uncommon DST PP binding motif with AntF. Post transcriptional functionalization of apo- to holo-AntF, or indeed any ACPs or PCP, with a 4’ phosphopantetheine arm, has been pinpointed as one major hurdle to overcome when using E. coli as a heterologous host. To this end, the E. coli strain BAPI encoding a promiscuous PPTase sfp from Bacillus subtilis is often used to functionalise recombinant ACPs and PCPs, however Sfp cannot efficiently functionalise AntF (data not shown). Only the cluster associated PPTase, AntB, functionalised AntF in E. coli despite its poor solubility, suggesting the AntB:AntF interactions to be somewhat orthogonal. Similarly, the cluster associated CoA ligase was also a principle component of M2 with low levels of shunt metabolites observed in the ΔAntG hosts. Low octaketide titers in ΔAntG hosts may result from a comparatively slower ACP self-acylation demonstrated to occur with the actinorhodin ACP in the absence of an appropriate MCAT. Alternatively, the E. coli endogenous MCAT may cross-talk, thereby acylating holo-AntF, but with significantly reduced efficiency reducing polyketide biosynthesis. The presence of a CoA ligase within a type II BGC is atypical for a specialized metabolite primed with a non-acetyl- or malonyl- starter unit, and unexpected in a cluster utilizing malonyl groups as a starter and extender unit. The promiscuity of AntF to accept substrates larger than a malonyl group was not assessed here, however type II PKSs systems can be primed with a range of different starter groups and the additional of alternative CoA ligases, ACPs and ATs has been shown to introduce aminated, aromatic and long acyl groups to type II polyketides. By modifying the CoA ligase enzymes in M2 one may expand the chemical space accessible by each scaffold before addition of any alternative tailoring enzymes.

Altering aromatic polyketide starter units is not the easiest way to diversify polyketide core scaffold, and alteration of primary tailoring enzymes (module 3) is a more common first strategy. The uncharacteristic motifs of AntF once more questioned its ability to interface with primary tailoring enzymes derived from alternative BGCs. Fortunately, this was non-problematic and replacing the cluster associated C9 ketoreductase with ActIII from the actinorhodin pathway restored AQ256 biosynthesis despite its structural similarities to emodin, a competitive inhibitor for ActII. A similar story is presented for substitution of structurally different but functionally similar aromatase/cyclase enzymes indicating major perturbations in enzyme structure are not limiting. Other than introduction of alternate starter units and chain length, alternate cyclisation patterns of nascent polyketide chains is the first step affording large variation in polyketide scaffolds. By choosing appropriate cyclases and aromatase/cyclases and modifying M1 to produce different chain lengths, U-shaped (linear), J-shaped (angular) and S-shaped aromatic.
polyketides could be accessed using this platform foreseeably enabling biosynthesis of BIQs, tetracyclines, aureolic acids, anthracyclines, tetracenomycins, angucyclines etc.

Many specialized metabolites are derived from similar precursors and follow congruent biosynthetic pathways before more specific biosynthetic enzymes modify each to a bewildering array of chemically and structurally diverse end compounds. The platform described here is no different, and through the cutting or pasting the AntI gene sequence into the AntA-H plug-and-play scaffold we are able divert polyketide biosynthesis towards to different classes of compounds. Identification of AntI as a biosynthetic branch point enables us to direct polyketide biosynthesis towards production of anthraquinones and dianthrone, structures which are associated with antiviral, antidepressant, antineoplastic, antiangiogenic, gastro-protectant and - purgative properties e.g. sennosides and hypericin, when plugged into the pathway, and bioproduction of two benzoisochromanequinone shunt metabolites (10, 11) of classical actinobacterial derived antibiotics and antitumoral agents, when removed.

In conclusion, we for the first time have been able to successfully express a soluble and functional type II PKS in E. coli to produce core scaffolds of pharmaceutically important polyketides. Modifications are not limited to the core polyketide scaffold however and toolboxes of glycosyltransferases, prenyltransferases, halogenases, sulfotransferases etc. previously described and fully characterized from biosynthetic pathways across all kingdoms can be integrated forming an extended biosynthetic pathway. Through careful curation of literature and community efforts to standardize biosynthetic enzymes we can rapidly identify, synthesize and introduce these secondary tailoring enzymes into module 4 of our biosynthetic platform in a semi-automated manner and generate large libraries of chemical diverse aromatic polyketides derived from pharmacologically privileged scaffolds. By utilizing the synthetic biology toolbox to do this in a highly-automated manner we can generate large libraries of chemical diverse aromatic polyketides derived from pharmacologically privileged scaffolds (Figure 5.2) holding the potential to unlock new vibrant bioactivities.

5.6 Acknowledgement
MC was funded by BBSRC DTP studentship BB/J014478/1. This is a contribution from the Manchester Centre for Synthetic Biology of Fine and Speciality Chemicals (SYNBIOCHEM) and acknowledges the Biotechnology and Biological Sciences Research Council (BBSRC) and Engineering and Physical Sciences Research Council (EPSRC) for financial support (Grant No. BB/M017702/1).
5.7 References


5.8 Supplementary results

5.8.1 Bioinformatics analysis of AntE; a partner for AntD

AntE does not comply with two structure–function based properties of chain length factors defined in the literature. 1) Gatekeeper residues in the CLF, which define the volume of the solvent excluded cavity at the dimer interface into which the polyketide chain grows, do not map onto AntE and the highly conserved glutamine (Q161) residue important in decarboxylation of the malonyl-ACP starter unit substrate initiating biosynthesis is substituted for an aspartic acid residue (Supplementary Figure 5.1). Together, these three points question AntE’s involvement as a component of the anthraquinones minimal polyketide synthase.

To gain a deeper insight into the secondary and tertiary structure of AntE a homology model was constructed using an actinorhodin (act) CLF backbone (Supplementary Figure 5.1). The homology model maintains the significant dissimilarity to ActI ORFII at the C-terminus, as expected, and while this does not comprise the region proposed to form the amphipathic tunnel, into which the polyketide grows, it does encompass a disordered region at the heterodimer interface. The region of dissimilarity may be important for ACP:CLF protein-protein interactions which, in canonical systems, is responsible for delivering acylated ACPS to the glutamine active residue catalysing decarboxylation of the polyketide starter unit. In agreement with this, residues shown to be important in ACP:FabF protein–protein interactions do not map to AntE (Supplementary Fig. 2). The anthraquinone cluster also harbors its own putative CoA ligase, this is a common feature of biosynthetic pathways which utilise a non-acetate starter unit suggesting the starter-unit-primed ACP may not require decarboxylation, or may be decarboxylated via an alternative as yet unknown mechanism. If substrate decarboxylation is not be catalyzed by AntE the significantly different C-terminus may reduce competitive and unfavourable interactions between AntE and AntF promoting end compound biosynthesis.

Interestingly, protein secondary structure predictions (Jpred4) show secondary structural motifs of AntE to have a similar organization to that observed in FabF, despite significant amino acid sequence dissimilarity at the C-terminus. The second beta sheet, β13 which form part of the major structural αβαβα thiolase fold common to fatty acid synthases and polyketide synthases, is not predicted in AntE homology models; rather an open solvent exposed structure is proposed with low confidence. In contrast, secondary structure predictions using Jpred4 indicate this structure to be maintained; the AntE AA sequence is predicted to form beta strands at positions corresponding to β8, β13 (QIIIQR) and β14 indicating the αβαβα fold topology of FabF, act CLF and other KSs and CLFs to be consistent in AntE (Supplementary Fig. 2). Unlike the homology models, protein secondary structure predictions were not modelled on existing crystal structures.

5.8.2 Elucidating major metabolites produced by the anthraquinone biosynthetic gene cluster

The first steps of the anthraquinones biosynthetic pathway are predicted to be congruent to
actinorhodin biosynthesis, the archetypal octaketide from *Streptomyces coelicolor*. Identification of expected octaketide shunt metabolites after each biosynthetic step enables interrogation of the AntA-I biosynthetic logic either confirming or refuting the biosynthetic congruence shown in Figure 2. Additionally, identification of shunt metabolites at each biosynthetic step provides a snapshot of any metabolic bottlenecks within the biosynthetic pathway which can guide pathway engineering to optimise end compound biosynthesis. Several octaketide shunt metabolites are isomeric and therefore discrimination by parent mass alone is not reliable; here retention time, MS² and HPLC-UV-Vis-MS aided characterisation when compared with existing literature. Masses corresponding to each modified octaketide shunt metabolite (2-11) were identified exclusively in the AntA-I expressing host, however at extremely low ion intensities preventing reliable identified using a targeted LC/MS approach; only protonated or deprotonated adducts were observed for each mass and isotopic masses were not visible (Figure 5.5, Supplementary Figure 5.6). The reduced ion intensity of shunt metabolites may indicate either the flux through the anthraquinone biosynthetic pathway is sufficient to avoid accumulation of intermediates or accumulated metabolites are being further derivatize to unknown end products which are not identified by our targeted MS analysis. To improve confidence when assigning MS peaks to expected shunt metabolites the corresponding extracted ion chromatogram (EIC) of AntA-I expressing hosts was compared to counterparts with individual ketoreductase, cyclase and cyclase/aromatase deletions (Supplementary Figure 5.20, Supplementary Figure 5.21); here, the absence of sequential biosynthetic enzymes results in accumulation of shunt metabolites after each biosynthetic step. The high accumulation of shunt metabolites in hosts deficient in one biosynthetic enzyme enabled characterization of intermediates at each step by HPLC-UV-Vis-MS (Supplementary Figure 5.22-Supplementary Figure 5.25). Intermediates were then used for comparative analysis of shunt metabolites from the AntA-I producing host.

5.8.3 SEK4 and SEK4b
Should a bottleneck arise at the C9 ketoreduction, directly after the biosynthesis of the nascent polyketide chain, accumulation of two cyclised octaketides, SEK4 (2) and SEK4b (3), would be expected (Figure 5.2). In the AntA-I expressing *E. coli* BL21(DE3) cultures the unreduced isomeric octaketides SEK4 and SEK4b were observed at tᵣ 527 sec and tᵣ 569.4 sec by positive ionization and tᵣ 516, 527, 569.4 and 619.8 sec by negative ionization (tᵣ = 527 sec, ES⁺ [M+H]⁺ 319.08084, theoretical [M+H]⁺ 319.08113 and ES⁻ [M-H]⁻ 317.06589, ES⁺ [M+H]⁺ 319.08005, ES⁻ [M-H]⁻ 317.06579) (Supplementary Figure 5.6). Theoretical exact, protonated and deprotonated masses for each shunt metabolite can be found in Supplementary Table 5.4.

Extracted ion chromatograms from Δketoreductase (KR) hosts also show that masses corresponding to SEK4 and 4b elute at 516 and 527 Sec, in good agreement with the AntA-I expressing strain and elution patterns of SEK4 and SEK4b in the literature (Supplementary Figure 5.20 and Supplementary Figure 5.21). Protonated and deprotonated adducts of each
parent mass were observed along with a comprehensive list of additional adducts in the $\Delta$KR strain. Interestingly, the dehydrated octaketides AUR367 and B26 were not observed reliably: the mass corresponding to a dehydrated octaketide eluting at 515.4 / 516 sec is more likely to be SEK4 where the predominant ion observed is [M-H$_2$O+H]$^+$ and not [M+H]$^+$, consistent with octaketide metabolite adduct profile from *E. coli* BL21 expressing antDEBF (Figure 5.2). Additional adducts at this retention time correspond to the hydrated octaketide rather than adducts of AUR367 or B26. This is a previously described characteristic of SEK4 and, together with the adducts described above, enables assignment of the first elution as SEK4, the second as SEK4b and the third as AUR367. Furthermore, UV-Vis spectra for SEK4 and SEK4b from $\Delta$AntA show $\lambda_{max}$ of 231 and 279 nm in agreement with previously published data\textsuperscript{4,5} (Supplementary Figure 5.22). Chromatographic peaks corresponding to SEK4 and SEK4b were not detected in the antA-I expressing host, at 279 nm, consistent with the both metabolites being present at extremely low concentrations.

### 5.8.4 Mutactin, SEK34 and their dehydrated counterparts

The next enzymatic step in the proposed biosynthesis of AQ256 is stereospecific reduction of C9 carbonyl group to a hydroxyl group catalyzed by a ketoreductase (Figure 5.2). Intramolecular aldol condensation of the ACP-linked polyketide chain results in formation of the first six-membered carbocyclic ring which is subsequently aromatised by a bifunctional aromatase/cyclase (ARO/CYC), or in the case of the anthraquinone cluster a tridomain aromatase/cyclase. Inactivation of the ARO/CYC during biosynthesis of actinorhodin or other aromatic octaketides results in the accumulation of mutactin (6) and its dehydrated form dehydromutactin (7). The aromatic nascent chain undergoes a second intramolecular aldol condensation between C5 and C14 catalyzed by an additional cyclase forming a common bicyclic intermediate\textsuperscript{8}. Inactivation of this cyclase results in accumulation of SEK34 (8) and SEK34b (19). Importantly, it is from the common bicyclic intermediate that the anthraquinone biosynthetic pathway diverges from benzoisochromanequinones.

Identification of mutactin and SEK34, and their dehydrated counterparts, poses a challenge in the AntA-I expressing cultures using LCMS/MS due to their isomeric nature. Comparative analysis of the exometabolome from $\Delta$Aro/Cyc (AntH) and $\Delta$Cyc (AntC) enabled identification and disambiguation of trace amounts of mutactin, dehydromutactin, SEK34 and SEK34b, however. Accumulation of mutactin is expected in the $\Delta$AntH strain and a mass pertaining to mutactin adducts are present at $t_R$ 582.6 sec ([M–H]$^-)$ 301.0712, [M+H]$^+$ 303.0861, theoretical masses [M–H]$^-)$ 301.0718, [M+H]$^+$ 303.0863) (Supplementary Figure 5.20, Supplementary Figure 5.21) which is in agreement with masses for the octaketide shunt product observed in the AntA-I expressing host ($t_R$ = 582.6, ES– [M–H]$^-)$ 301.0713, theoretical mass [M–H]$^-)$ 301.0717) (Supplementary Figure 5.6). Similarly, masses corresponding to SEK34 accumulated at $t_R$ 574.2 sec in $\Delta$AntC, (ES– [M–H]$^-)$ 301.0711, ES$^+$ [M+H]$^+$ 303.0856) matching an additional retention time corresponding to the ambiguous modified octaketide in the AntA-I expressing host exometabolome ($t_R$ = 574.2 ES– [M–H]$^-)$ 301.0713, ES$^+$ [M+H]$^+$ 303.0858), enabling annotation of masses eluting at 582.6 sec as mutactin and those eluting at 574.2 sec as SEK34. Mutactin
and SEK34 from ΔAntH and ΔAntC cultures were further characterized by HPLC-UV-Vis-MS where absorbance maxima and masses were in good agreement with previous publications\(^4,5\) (Supplementary Figure 5.23, Supplementary Figure 5.24) further validating the function of AntA and AntH. Discrimination of dehydromutactin and SEK34b was not possible using the HPLC conditions used here (Supplementary Figure 5.20, Supplementary Figure 5.21).

By sequentially knocking out one gene at each biosynthetic step of anthraquinone biosynthesis, with the exception of AntI, we show the biosynthetic pathway to be congruent with actinorhodin. Metabolites were normalised to final cell optical density (OD\(_{600}\)) and exometabolome was analysed in an unbiased manner using both positive and negative electrospray ionisation. Additionally, using this approach we identify accumulation of SEK34 in the antA–I expressing host indicating a biosynthetic bottleneck exists at AntC, the second ring cyclase, which can be further optimised to improve flux through the biosynthetic pathway.

### 5.8.5 Identification of trihydroxyanthrone (12) end product

The rationalised end compound of the anthraquinones cluster is proposed to be a trihydroxylated anthrone, as no monoxygenase is present within the anthraquinones BGC from \(P.\ luminescens\). Non-cluster associated monoxygenases are proposed to be involved in modification of this anthrone to the anthraquinones observed in the native producer\(^3\).

Masses corresponding to 1,3,8-trihydroxyanthrone (12) were observed in trace amounts exclusively in the supernatant of \(E.\ coli\) BL21(DE3) cultures expressing AntA-I at one main retention time (t\(_R\) 537 sec). The observed mass at t\(_R\) 537 sec, [M+H]** 243.06481, is within 1.6 ppm of theoretical mass [M+H]** 243.06519 of (12); however, only the protonated adduct was identified. Additional analysis using negative ionization corroborated the presence of C\(_{14}\)H\(_{10}\)O\(_4\) at t\(_R\) 537 sec, albeit at extremely low ion intensity rendering the annotation of 1,3,8-trihydroxyanthrone tentative. As mentioned in the main text anthrones have been shown to form cognate anthraquinones and dianthrones via various oxidative mechanisms which may occur during cultivation or during sample preparation (Supplementary Figure 5.5). Therefore the corresponding anthraquinones and dianthrones were targeted in subsequent analysis.

### 5.8.6 Dianthrone elucidation

Masses corresponding to 1,3,8 trihydroxydianthrone (13) were present exclusively in AntA–I expressing cultures at two retention times presumably corresponding to \(trans\) and \(meso\) forms (dianthrone 1: t\(_R\) 838.2 sec, [M+H]** 483.1074, [M-H]** 481.0918 and dianthrone 2: t\(_R\) 864.6 sec [M+H]** 483.1071, [M-H]** 481.0919, all masses within 2.5 ppm of the theoretical m/z [M+H]** 483.1074 and [M-H]** 481.0928)(Supplementary Figure 5.6). Furthermore, MS\(^2\) spectra of both dianthrone 1 and 2 indicate the putatively assigned 1,3,8 trihydroxydianthrones to fragment to the respective anthrone radical C\(_{14}\)H\(_{10}\)O\(_4\).\(^{●}\) (Supplementary Figure 5.5, Supplementary Figure 5.16): fragmentation of the C10 – 10’ bond is a hallmark of MS\(^2\) spectra from a wide variety of glycosylated and aglycone dianthrones\(^9\). UV-Vis absorbance of putative diathrones 1 and 2 showed similarities to emodin dianthrone\(^10\) and the dianthrone sennosides\(^11\) with λ\(_{\text{max}}\) at 359, 263, 217 nm and λ\(_{\text{max}}\) 358, 275, respectively (Supplementary Figure 5.17) collectively indicating
that both masses correspond to authentic dianthrones. Full characterization by NMR was not possible as neither compound was present in sufficient quantities.

5.8.7 Modified anthraquinone identification

AQ256 appeared to be the main anthraquinone detected from *E. coli* BL21(DE3) expressing antA-I however masses corresponding to the methylated AQ270a and b identified in *P. luminescens* were also detected with significantly reduced ion intensities (Supplementary Figure 5.6). Whilst addition or substitution of auxochromic substituents is well documented to cause small batho- or hypsochromic shifts in $\lambda_{\text{max}}$ of anthraquinones, no other intense peaks were identified at 434 nm suggesting the masses in question do not correspond to modified AQs (Supplementary Figure 5.15). Should the concentration of methylated anthraquinones drop below the detection limit of diode array detector, and the masses observed are authentic AQs, the low level of methylation is more likely performed by an enzyme endogenous to *E. coli* which can accept AQ256 as a poor substrate than a cluster-associated enzyme, as methoxy-substituted AQs are not expected intermediates in AQ256 biosynthesis.

The more extensively modified anthraquinones, AQ300 and AQ314 observed in *P. luminescens* TT01<sup>2</sup>, were not detected here. A mass corresponding to the actinorhodin shunt metabolites aloesaponarin II (11) was observed but was also present with diminished ion intensity in the exometabolome of *E. coli* BL21(DE3) and *E. coli* BL21(DE3) pACYCDuet-1 at the same retention time and was therefore discarded. Interestingly, masses corresponding to DMAC (10), the carboxylated precursor of aloesaponarin II, were observed at two retention times ($t_{\text{R}} = 702$ sec, ES$^+$ [M-H$^-$] 297.0397 and $t_{\text{R}} = 720$ sec, ES$^+$ [M-H$^-$] 297.0398), and fragmentation of the mass eluting at 702 sec is consistent with decarboxylation of the protonated parent ion to that of aloesaponarin II. Similarly, MS$^2$ of the mass at the second retention time shows decarboxylation of the parent mass at decreased ion intensity as well as a loss of water. Identification of DMAC but not aloesaponarin II has been previously reported during the biosynthesis of bhimamycins form *Streptomyces* sp. AK671<sup>12</sup>.

5.9 Supplementary Methods

5.9.1 Bacterial strains and culture conditions

*Escherichia coli* DH5α was used for routine cloning and plasmid preparation and maintenance. *E. coli* BL21(DE3) was used for expression of all recombinant proteins described here-in with the exception of AntDE purification where *E. coli* NiCo21(DE3) was used. For protein purification *E. coli* NiCo21(DE3) or BL21(DE3) was cultured in lysogeny broth (LB) at 37°C, 180 rpm, supplemented with appropriate antibiotics and induced with isopropyl β-D-1-thiogalactopyranoside (IPTG) at OD<sub>600</sub> 0.5-0.6 before reducing incubation temperature to 16°C for a further 16 h. For compound isolation *E. coli* BL21(DE3) was cultured in a minimal media as 2 mM MgSO<sub>4</sub>, 100 µM CaCl<sub>2</sub>, 238.8 µM Na<sub>2</sub>HPO<sub>4</sub>, 110.2 µM KH<sub>2</sub>PO<sub>4</sub>, 93.47 µM NH<sub>4</sub>Cl, 42.78 µM NaCl, 0.4% glucose in distilled H<sub>2</sub>O with appropriate antibiotics. Cultures were incubated at 37°C, 180 rpm, to OD<sub>600</sub> 0.35-0.4 and induced with 50 µM IPTG, for AQ256 and octaketide shunt metabolites before reducing the incubation temperature to 20°C for 3 to 5 days.
Photorhabdus luminescens TT01 was kindly provided by Dr. Ralf Heermann of Ludwig-Maximilian University, Munich. P. luminescens TT01 cultures were inoculated in casein-peptone soymeal-peptone broth (CASO) and routinely grown at 28°C. Working antibiotic concentrations were as follows: carbenicillin: 50 µg ml⁻¹, kanamycin: 25 µg ml⁻¹ and chloramphenicol: 25 µg ml⁻¹.

5.9.2 DNA isolation, Plasmid construction and Refactoring

5.9.2.1 DNA isolation

Genomic DNA (gDNA) from P. luminescens TT01 and Streptomyces coelicolor M145 was isolated from cultures grown to an OD₆₀₀ of 1 and OD₄₅₀ of 0.8 respectively, using a standard phenol chloroform DNA purification protocol and validated used routine PCR amplification. All primers used in this study are detailed in Supplementary Table 5.5 and Table 5.5. All vectors used in this study are listed in Supplementary Table 5.2. CloneAmp HiFi PCR Premix (TaKaRa) was used for all routine PCR amplification, for PCR products over 10 Kb PrimeSTAR Max DNA polymerase (TaKaRa) was used. All PCR products and restriction endonuclease digests were purified using the MinElute PCR purification kit (Qiagen) as per manufacturer’s instructions. All ligations were performed using the Rapid DNA ligation kit (Roche) as described by the manufacturer. All restriction endonucleases used in this study were obtained from New England Biolabs (NEB) and digests were performed for 1 h at 37°C unless stated otherwise.

5.9.2.2 Plasmid construction and refactoring

To construct the first KS/CLF expression vector pBbA2K-RFP was digested with EcoRI-HF and XhoI; the larger DNA fragment comprising the vector backbone was PCR purified. Primers EcoRI_Plu_for and Plu4189_XhoI_R were used to amplify a 2779 bp fragment encoding plu4191, plu4190 and plu4189 from P. luminescens TT01 genomic DNA by polymerase chain reaction (PCR), which was purified as above and ligated into the empty pBbA2k vector using T4 DNA ligase. The same procedure was followed to construct the BBR1 ori, T7 promoter and ampicillin resistant backbone mPKS expression vector pBbB1a-plumPKS. Aromatic polyketide KS and CLF genes are almost exclusively translationally coupled, and this is assumed to be the case for plu4191 and plu4190 due to the start:stop codon overlap. Transcriptional coupling has been proposed to colocalize proteins transcribed from the same polycistronic mRNA and may aid dimer formation, as such the native operon architecture of plu4191, 90 and 89, the KS, CLF and ACP, was retained in the pBbA2k expression vector. To construct the his⁶-AntE and AntD protein purification vector pETDuet419091 plu4191 was PCR amplified from P. luminescens TT01 gDNA using primers Plu4191_for_BglII and Plu4191_rev_KpnI, purified as above and digested with BglII and KpnI-HF. The digested plu4191 fragment was ligated into pETDuet-1 also digested with BglII and KpnI-HF to form pETDuet4191. Plu4190 was amplified from P. luminescens TT01 gDNA using plu4190_for_EcoRI, which removed the start ATG, and Plu4190_rev_PstI. The fragment was purified as above and digested using EcoRI-HF and PstI-HF before ligating into pETDuetplu4191 linearized with EcoRI-HF and PstI-HF to form the his⁶-plu4190 fusion vector pETDuetplu419091. To express recombinant AntB and AntG, the PPTase
and CoA ligase from the anthraquinone BGC, \textit{plu}4193 and \textit{plu}4188 sequences were amplified from \textit{P. luminescens} TT01 gDNA using primers Plu4193\_for\_Ncol\_untagged, Plu4193\_rev\_HindIII, Plu4188\_for\_Ndel and Plu4188\_rev\_XhoI respectively before purification, digestion with Ncol and HindIII, and Ndel and XhoI correspondingly and ligation into pACYCDuet-1 using appropriate restriction endonucleases to form both pACYCPlu4188 firstly and pACYCPlu418893 subsequently. Neither \textit{plu}4193 nor \textit{plu}4188 were tagged. The p15A ori enabled coexpression of AntB and G with the mPKS of pBbB1a-plumPKS.

The entire complement of genes responsible for anthraquinone biosynthesis were cloned into pACYCDuet-1 this enables further plasmids with compatible origins of replication to be easily introduced when derivatising the end compound in a combinatorial fashion. To construct this vector a fragment comprising \textit{plu}4192, 93 and 94 was first cloned into pACYCDuet-1 multiple cloning site (MCS) 2 after PCR amplification from \textit{P. luminescens} TT01 gDNA using primers Plu4194\_for\_NdeI and Plu4192\_rev\_XhoI. Both the fragment and pACYCDuet-1 vector were digested with NdeI and XhoI before ligation, as above, to form pACYCDuetPlu4192-94. The 6 remaining genes, \textit{plu}4186-91 were cloned into pACYCDuet-1 multiple cloning site (MCS) 1 in the same manner however used primers Plu\_for\_EcoRI and AnthraquinoneBGC\_Rev\_PstI for the PCR amplification forming the 13.6 Kb pACYCAntaraquinone vector.

To generate the ketoreductase (KR) complementation vector the anthraquinone KR, \textit{plu}4194, was swapped with sco5086, the C9 KR from the actinorhodin BGC, here pACYCAntaraquinone was linearized with primers KR\_Swap\_IF and Plu4194\_rev\_InFusion which both read outwards of \textit{Plu}4194 removing most of the \textit{plu}4194 CDS from the linear vector. The resulting 13.6 Kb fragment was purified by ethanol precipitation. A 767 bp DNA fragment encoding the wild type sco5086 sequence was PCR amplified from \textit{S. coelicolor} M145 using primers Sco5086\_for\_IF Sco5086\_rev\_IF introducing two synonymous mutations in the 5’ of the sequence. These primers added 15 bps of sequence homologous to each end of the linearised pACYCAntra\Delta\textit{plu}4194 to the \textit{sco}5086 containing sequence and enabled plasmid construction by In Fusion (Clonetech) as per manufacturer’s instruction forming pACYCAnt\textit{twt}KR. The anthraquinone KR, \textit{plu}4194, and downstream CDS PPTase, \textit{plu}4193, are transcriptionally coupled with a start stop codon overlap. The PPTase ribosome binding site, therefore, is within the 3’ end of \textit{plu}4194, fortuitously the N-terminal amino acid sequence of both ActIII and AntA are identical. The primer \textit{plu}4194\_rev\_InFusion binds a DNA sequence within the region encoding the identical sequence N-terminal amino acid sequence and maintains the putative PPTase Shine-Dalgarno. The same procedure was follow for the introduction of the refactored actinorhodin KR however Ref\_Sco5086\_IF\_for and Ref\_SC5086\_IF\_rev were used as primers to amplify a codon optimized \textit{sco}5086 sequence from pG9m-2-ActKRRef to form pACYCAn\textit{tref}KR. Additionally, pACYCAntara\DeltaKR was constructed as a KR negative control by linearising pACYCAntaraquinone via PCR as above using primers Plu4194\_del\_IF\_for and KR\_Swap\_IF\_For followed by DNA assembly using NEBuilder HiFi DNA Assembly Master Mix (NEB) following manufacturer’s instructions. The same process was undertaken to construct \textit{sco}5090 complementation vectors. Once more pACYCAntaraquinone was linearized with
primers plu4187_replace_fw and plu4188_rev both removing plu4187 to form the 12,169 bp linearized vector and purified by ethanol precipitation. A 989 bp DNA fragment encoding the wild type sco5086 sequence was PCR amplified from S. coelicolor M145 using primers Sco5090_for_IF Sco5090_rev_IF and purified using the MinElute Qiagen PCR purification kit. The vector pACYCAnntCYC was subsequently constructed via In Fusion DNA assembly from purified sco5090 DNA fragment and linearized pACYCAnthaΔplu4187. To introduce the refactored actinorhodin CYC/ARO, the same method was followed however Ref_Sco5090_IF_for and Ref_Sco5090_IF_rev were used as primers to amplify a codon optimized sco5090 sequence from pG9m-2-ActARO/CYCRef forming pACYCAnntrefCYC/ARO.
The sco5090 knock out plasmid pACYCAntΔAntH was generated by linearization of pACYCAnthraquinone using primers Plu4187_delta_for and Plu4187_delta_rev, purification of the linear DNA fragment by ethanol precipitation and DNA assembly using In Fusion as described above. Plasmids pACYCAnntΔAntC and pACYCAnntΔAntI were also generated by linearization of pACYCAnthraquinone using primers Plu4192Δ_for and Plu4192Δ_rev and plu4186Δ_for and plu4186Δ_rev respectively, followed by In Fusion DNA assembly as above. All codon optimized genes were designed using Gen Optimiser and manufactured by Gen9.

5.9.3 Protein purification and peptide identification
AntDE were purified from E. coli BL21 NiCo21(DE3), NEB, using immobilized metal affinity chromatography in 300 mM NaCl, 50 mM tris-HCl pH7.4, 50 mM imidazole and eluted with 90% purity. AntDE were further purified via anion exchange chromatography using a 6 ml resource Q (GE Healthcare Life Sciences) with a linear gradient from 95% to 5% 50 mM Tris-HCl pH7.4 against 50 mM Tris-HCl 1M NaCl at 3 ml min⁻¹. Samples containing AntDE were subsequently separated by size exclusion chromatography using Superdex 200 Increase 100/300 GL columns (GE Healthcare) eluted with 1.5 column volumes of 200 mM NaCl, 50 mM tris-HCl pH7.4 to isolate the complex in its dimeric form. Backbone vectors containing RFP or GFP were used as protein expression induction controls and to monitor protein extraction efficiency throughout. For SDS-PAGE visualization of protein bacterial cultures were suspended in 5% (vol/vol) bugbuster (Novagen) and incubated for 20 min at room temperature before centrifugation at 14000 rpm, 4°C, for 20 min where the supernatant was isolated from the pellet. A total of 15 µg of sample was boiled in 1x SDS loading dye (GBioscience) and 10 mM DTT for 8 min. Protein concentration was evaluated using Bradford Assay.

Protein bands of interest were isolated from polyacrylamide gels and coomassie stain was removed through alternating dehydration and hydration steps in 50% acetonitrile and 50 mM ammonium bicarbonate before digestion with mass spectrometry grade trypsin (Promega) at 37°C for 20 h. Extracts containing tryptic peptides were centrifuged at 13000 rpm for 10 min to remove particulate matter prior to separation and analysis using a C18 column (LC Packings, Acclaim Pep Map 100) and Bruker Esquire 3000plus ion trap mass spectrometer. Analysis was carried out in positive ion mode with an injection volume 20 µl and flow rate of 200 nL min⁻¹ over a gradient of water to 90% acetonitrile both acidified with 0.1% formic acid. Peptide fragments were identified using the Mascot MS/MS ion search software (Matrix Sciences). Mascot MS/MS
search results from band 1 (Figure 5.3, red circle) identified 5 peptide fragments consistent with AntD (FVLGESAGIIPSLK, LSSGFSGIHSVIMR, SEDYDFDSSAAUSVAK, SGAIGQVYSDGNNKEFVLK, GAHIYELAYASVNNAYHTDLPADGMAMAR). Similarly, results from band 2 (Figure 5.3, yellow circle) showed 6 peptide fragments consistent with AntE (LDVKLDPNR, INEFNITGIQQR, QPGDFSEGAAFLVLEER, IDEFSVYGIHAVEMALK, VVVATGVGAIHPDGDVTAIK, KIDEFSVYGIHAVEMALK, RVVVGTVGAIHPDGDVTAIK). Peptide fragments are identified as [2M+H]+ adducts.

### 5.9.4 Metabolite extraction and analysis

Both exo- and endo-metabolomes were analysed throughout. For analysis by HPLC and mass spectrometry exometabolomes were analysed rather than endometabolomes. For exometabolome preparation, cells were pelleted at 4000 x $g$, 4°C, for 20 mins before decanting supernatant and storing on ice. Supernatant was prepared by further centrifuged at 12000 x $g$, 4°C, for 20 min to remove cell debris before filtering through a 0.45 µm pore size filter (Merk Millipore) prior to analytics. For endometabolomes unbiased metabolite extraction of the cell pellet was done as previously described\(^\text{13}\). In brief, cell pellets were flash frozen in liquid nitrogen before being suspended in 80% MeOH (-48°C). Each sample was subsequently flash frozen in liquid nitrogen and allowed to thaw on ice three times. Samples were centrifuged at 14,000 x $g$ for 20 min, -9°C, before decanting the supernatant. The extraction process was repeated three times and the supernatant from each sample pooled and used to analyse the exometabolome. Metabolites monitored in each experimental analysis include SEK4 \(^\text{2}\), SEK4b \(^\text{3}\), B26 \(^\text{4}\), AUR367 \(^\text{5}\), mutactin \(^\text{6}\), dehydromutactin \(^\text{7}\), SEK34 \(^\text{8}\) SEK34b \(^\text{9}\), 1,3,8-trihydroxyanthrone \(^\text{12}\), 1,3,8-trihydroxydianthrone \(^\text{13}\) aloesaponarin II \(^\text{11}\), 3,8-dihydroxymethylanthaquinone carboxylic acid (DMAC) \(^\text{10}\), AQ256 \(^\text{1}\), AQ270a and b, AQ284a and b, AQ300 and AQ314 (Structures not shown). HPLC and mass spectrometers used throughout this study are detailed in individual figure legends. To normalise EICs to final culture optical density extracted ion intensities were divided by the average optical density (OD\(_600\)) of three technical replicates for each sample.

### 5.9.5 Octaketide shunt metabolite identification using HPLC-ESI-MS

The exometabolome of *E. coli* expressing antDEF, antDEFB and antDEFBG were analysed by HPLC-ESI-MS (Waters Acquity Ultra Performance LC, Thermo Scientific LTQ Orbitrap XL). HPLC conditions were as follows: 1 min: isocratic gradient of 5% solvent B, 7min: linear gradient 70% solvent B, 7.5: min linear gradient 95% solvent B, 8.5: min isocratic gradient 95% solvent B, 9 min: linear gradient to 5% B, 10 min: isocratic gradient of 5% B. Solvents A and B were HPLC grade water and HPLC grade acetonitrile both acidified with 0.1% formic acid using a C18 2.6 µm 2.1 x 100 mm LC column (Phenomenex) heated to 30°C with a flow rate of 0.3 ml min\(^{-1}\). Injection volume of 3 µl was analysed by electrospray injection mass spectrometry in positive ionisation (ES\(^+\)) mode with an ESI-HESI source over a mass scan range of 80-1200 m/z.
5.9.6 Shunt metabolite identification via HPLC-UV-Vis-ESI-MS

Shunt metabolites (2,3,6-9) (5.8.2), dianthones (13) (5.8.5, 5.8.6) and modified anthraquinones (structures not shown) (5.8.7) were identified using HPLC-UV-Vis-MS (Agilent 1100 series LC/MSD trap). HPLC conditions were as follows: 5 min isocratic gradient of 5% solvent B, 30 min linear gradient from 5 to 95% solvent B, 10 min isocratic gradient of 95% solvent B, 5 min linear gradient from 95 – 5% solvent B, 10 min isocratic gradient at 5% B using a C18 2.6 µm 2.1 x 100 mm accucore LC column (ThermoFisher Scientific) heated to 45°C with a flow rate of 0.25 ml min⁻¹. Solvents A and B are as in supplementary methods 2.4.1. Injection volume of 20 µl was analysed via electrospray injection mass spectrometry in negative mode.

5.9.7 Actinorhodin KR and CYC complementation and pACYCAntΔ86 analysis via UV-Vis

Biosynthesis of anthraquinone and benzoisochromanequinone production in actinorhodin KR and CYC complementation experiments was monitored at 434 nm using a Shimadzu prominence UFLC RX SPD-20A UV-Vis detector. Metabolites were separated using a 15 minute gradient as follows: 5 min isocratic gradient of 5% B, 15 min linear gradient to 95% B, 5 min isocratic gradient at 95% B, 3 min linear gradient to 5% B and 7 min isocratic gradient at 5% B. HPLC solvents and column are as in supplementary methods 5.9.4.

5.9.8 HPLC High resolution mass spectrum analysis

All experimental samples described here were principally analysed using HPLC high resolution mass spectrometry using the Dionex ultimate 3000 rapid separation HPLC coupled with QExactive plus mass spectrometer (Thermo Scientific). HPLC are as follows: 5 min isocratic gradient of 5% solvent B, 15 min linear gradient from 5 to 95% solvent B, 5 min isocratic gradient of 95% solvent B, 3 min linear gradient from 95 – 5% solvent B, 2 min isocratic gradient at 5% B. Column, column conditions and solvents were as described in Supplementary methods 5.9.4, however, with a flow rate of 0.3 ml min⁻¹. The QExactive plus mass spectrometer was operated in both positive and negative ionisation mode using an ESI-HESI source. All mass spectra were recorded using a full mass spectrum scan with data dependent MS² (Top5). Full-scan spectra were obtained over a scan range of 80-800 m/z with a resolution of 70000. A resolution of 17500 was used for routine MS² spectra with a default charge state of 1 and collision induced dissociation energy at 35 eV. Fragmentation patterns of AQ256, aloesaponarin II, emodin and chrysophanol (Supplementary Figure 5.13) were analysed using HPLC-tSIM-MS² with exact masses detailed in an inclusion list. A resolution of 35000 was used when recording MS² spectra. Default charge state and collision dissociation energy were as above.

5.9.9 Mass spectrum data analysis

Mass spectra were recorded in .raw format from all instruments before conversion to .mzML using Proteowizard 3.0.9393 with binary encoding precision of 64-bit, write index, zlib compression and TPP compatibility selected. Peak picking filters with MS level 1–2 were used as standard. Mass spectra were subsequently analysed using XCMS LC/MS and GC/MS data analysis package¹⁴ using R.
5.9.10 Code availability
Scripts used throughout are available upon request.

5.9.11 Characterisation of AQ256
For purification of AQ256 for NMR spectroscopy samples were evaporated to dryness under vacuum resulting in a brown oil which was further suspended in 50% methanol before purification via semi-preparative HPLC as: 5 % B for 0-10 min, 5 - 95% B linear gradient 10 - 55 min, 95% B 55 - 65 min, 95 - 5% B 65 - 75 min, 5% B 75 – 85 min with a and flow rate of 5 ml/min; where solvents A and B were water and acetonitrile acidified with 0.1% formic acid respectively. Yellow fractions were pooled and evaporated to dryness before resuspension in 1/10th volume 80% MeOH. Samples crystallised at 4°C over 48 h, excess solvent was removed and samples were desiccated for 48 h before suspension in 600 μl deuterated methanol, 1% TMS for ¹H and COSY NMR spectroscopy using 400 mHz Bruker NMR spectrometer and ¹³C, HSQC and HMBC NMR spectroscopy using 800 mHz Bruker nmr spectrometer. UV-Vis spectra of chrysophanol and emodin standards, and AQ256 were recorded using a Cary 60 UV-Vis spectrophotometer (Agilent Technologies) and are as described in Supplementary Figure 5.14.

5.9.12 Phylogenetic analysis of amino acid sequences
Multiple sequence alignments were performed with the Multiple Alignment Fast Fourier Transform (MAFFT) G-INS-1 progressive method¹⁵. Maximum likelihood phylogenetic trees were generated, bootstrapped with 500 iterations using MEGA6¹⁶. Sequence alignment with protein secondary structures in postscript were visualized using ESPript¹⁷.
5.10 Supplementary Figures
Supplementary Figure 5.1: Multiple sequence alignment of FabF, AntD, AntE, ActI ORF I and ActI ORF II

Multiple sequence alignment of FabF, AntD, AntE, ActI ORF I short (act KS), derived from its crystal structure\textsuperscript{18}, and ActI ORF II (act CLF) fatty acid synthesis and polyketide synthesis components. The FabF protein secondary structure overlaid is derived from the wild type \textit{E. coli} FabF crystal structure: 2GFW\textsuperscript{19}. The Blue arrow shows the catalytic cysteine of FabF, ActI ORF I and AntD, the glutamine in ActI ORF II intrinsic to starter unit decarboxylation and the corresponding aspartic acid in AntE. Black arrows at R207 and L209 show residues important in AcpP:FabF interaction in \textit{E. coli}\textsuperscript{1} and do not map onto AntE. The red dotted arrow indicates the QIIIQR motif predicted to form β-strand 13 by JPred4\textsuperscript{2}, the red bar indicates the region of nonaligned residues in AntE which form β-strand 13 in FabF, AntD and both ActI ORF I and ActI ORF II.
Supplementary Figure 5.2: Homology model of the AntDE dimer

Homology model of the AntE (grey and red) and AntD (blue) dimer built using SWISS-MODEL\textsuperscript{20} using 1TQY chains B and A respectively as target model template. The C-terminal dissimilar third of AntE is coloured red. Homology models were built individually and visualised and dimerised using PyMOL. The red predicted structure of AntE is considerably smaller, more open and disordered than the counterpart structure of the blue ketosynthase, AntD.
Supplementary Figure 5.3: Cyclisation pathways of octaketide shunt metabolites

Schematic showing C7-12 and C10-15 cyclisation of a nascent C16 poly-β-ketone forming octaketide shunt metabolites SEK4 and SEK4b, followed by dehydration to AUR367.
Supplementary Figure 5.4: SEK4 and SEK4b MS² spectra

a, Tandem mass spectrum showing possible fragmentation pattern of SEK4b from *E. coli* BL21(DE3) pBbB1a-plumPKS, pACYCPlu418893. Observed mass of each fragment is within a 2.8 ppm tolerance of expected masses. Adducts are as follows: [M-H-C\(_6\)H\(_6\)O\(_3\)]\(^-\) 191.03 and [M-H-C\(_{10}\)H\(_8\)O\(_4\)]\(^-\) 125.02. b, MS² mass spectrum showing possible fragmentation pattern of SEK4 as above, adducts are as follows: [M-H-CO\(_2\)] 273.08, [M-H-CH\(_2\)O\(_3\)]\(^-\) 255.07 and [M-H-C\(_3\)H\(_2\)O\(_3\)]\(^-\)
Mass spectra presented here are representative of three biological samples and recorded as described in supplementary methods 5.9.5.
Supplementary Figure 5.5: plausible oxidative coupling of 1,3,8 trihydroxyanthrone and dianthrone fragmentation pattern

Emodin anthrone, a C6 methyl substituted analogue of 12, follows two oxidative pathways to form the anthraquinone emodin or alternatively to an emodin dianthrone. The first pathway leads to emodin directly through oxidation of the C9 methylene and the second to emodin dianthrone via H atom abstraction followed by coupling of two anthrone radicals. Congruent pathways to AQ256 and its dianthrone are represented for 1,3,8 trihydroxyanthrone (12) in a, b, schematic of MS² fragmentation pattern observed for both anthraquinone 1 and 2. Compounds show hallmark fragmentation at the C10-C’10 bond. Anthraquinone carbon numbering follow standard convention for 9,10 anthraquinone
Supplementary Figure 5.6: Comparative EIC’s for metabolites of interest

A comparison of EICs for metabolites of interest extracted the exometabolome of *E. coli* BL21(DE3), *E. coli* BL21(DE3) pACYCDuet-1, *E. coli* BL21(DE3) pACYCAnthraquinone analysed in both positive and negative ionisation mode [M-H]⁻ (a and b) and [M+H]⁺ (c and d). EICs show all masses within ± 5 ppm of each metabolites theoretical mass. HPLC-ESI-MS conditions are as described in supplementary methods 5.9.8. Red, blue and black lines represent EICs of *E. coli* BL21 (host control), *E. coli* BL21 pACYCDuet-1 (plasmid control) and *E. coli* BL21 pACYCAnthraquinone all normalized by final cell density (OD₆₀₀). Figures b and d show a zoomed perspective of a and c, respectively, enabling identification of minor shunt metabolites. For the purpose of clarity the EIC displaying masses corresponding to AQ256 are greyed out in b. Each EIC is representative of three biological replicates. Collectively EICs show accumulation of AQ256 the predominant metabolite synthesised from the anthraquinone biosynthetic pathway identified using this targeted approach. Additionally, SEK34b also accumulates to high ion intensities.
Supplementary Figure 5.7: Graphical interpretation of AQ256 two dimensional NMR spectra

a, Significant correlations shown by COSY, green, and HSQC, grey, 2D NMR spectra. b, Assignment of AQ256 quaternary carbons via HMBC. Values for chemical shifts and coupling constants are described in Supplementary Table 5.3.
Supplementary Figure 5.8: $^1$H NMR spectrum for AQ256

$^1$H NMR spectrum for AQ-256 in deuterated methanol a, Complete $^1$H spectrum with TMS and methanol peaks annotated. b, Zoomed $^1$H spectrum showing aromatic proton signals.
Supplementary Figure 5.9: $^{13}$C NMR spectrum for AQ256

Complete $^{13}$C spectrum with TMS standard: solvent peaks are annotated.
Supplementary Figure 5.10: HSQC NMR spectrum of AQ256

a, Full HSQC NMR spectrum of AQ-256 in deuterated methanol. b, Zoomed HSQC showing relationship between aromatic protons and the corresponding carbon atoms.
Supplementary Figure 5.11: COSY NMR spectrum of AQ256

A two dimensional COSY NMR spectra showing proton–proton coupling in AQ256.
Supplementary Figure 5.12: HMBC NMR spectrum of AQ256

a, full HMBC NMR spectrum of AQ256 in deuterated methanol with TMS standard. b, Zoomed HMBC showing relationship between aromatic protons and the corresponding carbon atoms.
Supplementary Figure 5.13: Evaluation of chrysophanol, emodin, aloesaponarin II and AQ256 MS² spectra (ES)
Supplementary Figure 5.14: UV-Vis spectra for chrysophanol, emodin and AQ256

UV-Vis spectra for AQ256 (100% methanol), and two other anthraquinones, emodin and chrysophanol (80% methanol), recorded between 250 and 700 nm using a Cary 60 UV-Vis spectrophotometer (Agilent Technologies). All samples show common $\lambda_{\text{max}}$ at ~430 nm.
Supplementary Figure 5.15: Typical full HPLC chromatogram of *E. coli* BL21(DE3) pACYCAnthraquinone cultures expressing AntA-I at 434 nm

a, Full exometabolome HPLC profiles of M9 growth media, *E. coli* BL21(DE3), *E. coli* BL21(DE3) expressing empty vector pACYCDuet-1 and *E. coli* BL21(DE3) pACYCAnthraquinone expressing AntA-I at 434 nm. Chromatograms are representative of 3 biological samples described and analysed in Supplementary Figure 5.6. HPLC conditions are as described in supplementary methods 5.9.6. b, Zoomed chromatogram from a. A single distinct peak is present in the *E. coli* pACYCAnthraquinone sample. c, UV-vis spectrum from AQ256 (1) showing λmax 216, 283, 434, 585, 686 nm in agreement with λmax values observed form authentic AQ256 standards.
Supplementary Figure 5.16: Full mass spectrum and MS^2 for two putative dianthrones

Mass spectrum of dianthrones from antA-I expressing BL21(DE3) at 13.95 min and 14.39 min respectively a and b. Deprotonated parent masses for each dianthrone are written in red and are 2.1 ppm and 2.2 ppm from theoretical mass (Supplementary Table 5.4). The two retention times plausibly correspond to trans and meso dianthrone isomers. c and d represent MS^2 spectrum for both dianthrones.
Supplementary Figure 5.17: HPLC-UV-Vis-MS analysis of dianthrones 1 and 2

- **a**, UV-Vis spectrum for putative dianthrone 1, $\lambda_{\text{max}}$ at 359, 263, 217 nm.
- **b**, ES-MS spectrum corresponding to dianthrone 1 [M-H]$^-$ 418.1.
- **c**, UV-Vis spectrum for putative dianthrone 2, $\lambda_{\text{max}}$ at 358, 275, 261, 218 nm.
- **d**, ES-MS spectrum corresponding to dianthrone 1 [M-H]$^-$ 480.9.

Theoretical mass is [M-H]$^-$ 481.0929. Spectra are recorded from the *E. coli* BL21(DE3) antA-I exometabolome. Each figure is a consensus of 3 biological samples.
A comparison of extracted ion chromatograms from the exometabolome of *E. coli* BL21(DE3) expressing the anthraquinone pathway complemented with ActIII, a ketoreductase from the actinorhodin biosynthetic pathway. All samples are numbered as media blank, 1; *E. coli* BL21(DE3) wild type, 2; *E. coli* BL21(DE3) pACYCDuet-1, 3; *E. coli* BL21(DE3) pACYCAntΔAntA, 4; *E. coli* BL21(DE3) pACYCAntrrefKR (refactored sequence), 5; *E. coli* BL21(DE3) pACYCAnwtKR (modified wild type sequence), 6 and *E. coli* BL21(DE3) pACYCAanthraquinone, 7. Ion intensities were normalised to final cell density (OD$_{600}$). a and b show 2D and 3D extracted ion chromatograms (EIC) of all observable masses between 255.0286 and 255.0312 m/z: AQ256 theoretical mass of [M-H]$^-$ 255.02989. c and d show EIC of masses between 317.0651 and 317.0683 m/z: SEK4 and SEK4b theoretical mass of [M-H]$^-$ 317.0667. All EICs use a ±5 ppm cut off for identification of metabolite of interest and are representative of 3 biological samples and HPLC conditions are as described in supplementary methods 5.9.8. AQ256 biosynthesis is observed in *E. coli* BL21(DE3) expressing *antA-I*, and restored in *E. coli* BL21(DE3) expressing *AntB-I* complemented with the actinorhodin ketoreductase but not in the ΔAntA host. Instead, the exometabolome of *E. coli* BL21(DE3) pACYCAntΔAntA is enriched in SEK4 and SEK4b indicating a metabolic bottleneck to occur before keto reduction of C9, as expected. Bolded numbers represent metabolites detailed in Figure 5.2.
Supplementary Figure 5.19: Extracted ion chromatogram (EIC) of ARO/CYC complemented anthraquinone biosynthetic pathways

Comparative extracted ion chromatograms of metabolites present in the exometabolome of *E. coli* BL21(DE3) expressing the anthraquinone pathway complemented with ActVII, an aromatase/cyclase from the actinorhodin biosynthetic pathway. Samples are numbered as media blank, 1; *E. coli* BL21(DE3) wild type, 2; *E. coli* BL21(DE3) pACYCDuet-1, 3; *E. coli* BL21(DE3) pACYCAntΔAntH, 4; *E. coli* BL21(DE3) pACYCAntrefCYC (refactored sequence), 5; *E. coli* BL21(DE3) pACYCAntwtCYC (modified wild type sequence), 6; *E. coli* BL21(DE3) pACYCAntthraquinone, 7 and characterised AQ256 as sample 8 and are representative of biological triplicates. EICs use a ±5 ppm cut off for identification of metabolite of interest. Exact masses for metabolites of interest are detailed in Supplementary Table 5.4. HPLC conditions are as described in Supplementary Figure 5.18. a and b represent 3D and 2D EICs of masses ±5 ppm from AQ256 theoretical mass of [M-H] - 255.02989 respectively. AQ256 (1) biosynthesis is restored when complemented with the actinorhodin cyclase/dehydratase ActVII. c and d show EICs of masses ± 5 ppm from the theoretical deprotonated mutactin adduct (6), the main expected shunt metabolite from an antA-G, I expressing host (ΔAntH). e and f show EICs for masses ± 5 ppm of deprotonated dehydromutactin (7) adduct, the dehydrated mutactin counterpart. G and h show EIC for masses ± 5 ppm from the deprotonated adduct of DMAC (10) and i and j show EIC for masses ± 5 ppm from the deprotonated adduct of aloesaponarin II (11). Accumulation of aloesaponarin II (11) and its acid precursor DMAC (10) in exometabolomes of *E. coli* BL21(DE3) expressing the anthraquinone biosynthetic pathway complemented with ActVII is consistent with hydrolysis of the polyketide chain from the acyl carrier protein after cyclisation of the second ring. Bolded numbers are consistent with Figure 5.2.
Supplementary Figure 5.20: ES⁻ extracted ion chromatograms of polyketide shunt metabolites

Typical extracted ion chromatograms (EIC) for expected octaketide shunt metabolites (2,3 and 6-9) in analysed using negative ionisation mode (ES⁻). EIC’s a, c and e compare
exometabolomes from the background host *E. coli* BL21(DE3), host expressing and empty plasmid and host expressing antA-I, showing extracted ion chromatograms from *E. coli* BL21(DE3) in red, *E. coli* BL21(DE3) pACYCDuet-1 in blue and *E. coli* BL21(DE3) pACYCAnthraquinone in black respectively. EIC’s b, d and f additionally show chromatograms for the KR, ARO/CYC and Cyc biosynthetic pathway knock outs as *E. coli* BL21(DE3) pACYCAntΔAntA, green, *E. coli* BL21(DE3) pACYCAntΔAntH, orange, and *E. coli* BL21(DE3) pACYCAntΔAntC in sky blue. Ion intensities were normalised by final cell density (OD$_{600}$). Each extracted ion chromatogram was limited to the theoretical deprotonated mass ± 5 ppm for metabolites of interest. Masses are as follows: for a and b unreduced octaketide SEK4 (2) and SEK4b (3), [M-H]$^-$ 317.0651 - 317.0683 m/z; for c and d mutactin (6) and SEK34 (8), [M-H]$^-$ 301.0703 - 301.0733 m/z and for e and f dehydromutactin (7), SEK34b (9) [M-H]$^-$ 283.0598 - 283.0626 m/z. EIC’s are representative of three biological replicates and were analysed using conditions described in Supplementary Figure 5.18. Bolded numbers correspond to metabolites detailed Figure 5.2.
Supplementary Figure 5.21: ES⁺ extracted ion chromatograms of polyketide shunt metabolites

Typical extracted ion chromatograms (EIC) for expected octaketide shunt metabolites (2, 3 and 6-9) analysed in positive ionisation mode, (ES⁺). EIC’s follow conventions stated in Supplementary Figure 5.20. EIC’s show all masses within a window constrained to the theoretical protonated mass of each metabolite of interest ±5ppm. Masses are as follows: a and
b unreduced octaketide shunt metabolites SEK4 (2) and SEK4b (3) \([\text{M+H}]^+\) 319.0796 - 319.0828 \(m/z\); for c and d mutactin (6) and SEK34 (8) \([\text{M+H}]^+\) 303.0848 - 303.0878 \(m/z\) and for e and f dehydromutactin (7) and SEK34b (9) \([\text{M+H}]^+\) 285.0743 - 285.0772 \(m/z\). SEK34 is observed as a dehydrated \([\text{M-H}_2\text{O}+\text{H}]^+\) adduct. EICs are representative of samples analysed in triplicate. HPLC-ESI-MS conditions are as described in Supplementary Figure 5.18.
Supplementary Figure 5.22: HPLC-UV-Vis-ESI-MS analysis of E. coli BL21(DE3) expressing AntB-I, ∆AntA

HPLC-UV-Vis-ESI-MS analysis of E. coli BL21(DE3) expressing antB-I, ∆AntA, identifying SEK4 and SEK4b as predominant octaketide shunt metabolites in the ΔKR anthraquinone biosynthetic pathway. a, Typical chromatogram of culture supernatant from E. coli BL21(DE3) pACYCAntΔA monitored at 279 nm showing two peaks corresponding to SEK4 and SEK4b which are not present in E. coli BL21(DE3) or E. coli BL21(DE3) pACYCDuet-1. HPLC conditions were as follows: 5 min isocratic gradient of 5% solvent B, 30 min linear gradient from 5 to 95% solvent B, 10 min isocratic gradient of 95% solvent B, 5 min linear gradient from 95 – 5% solvent B, 10 min isocratic gradient at 5% B: Solvents A and B were HPLC grade water and HPLC grade acetonitrile both acidified with 0.1% formic acid using a C18 2.6 µM x 100 mm Accucore LC column (ThermoFisher Scientific) heated to 45°C. b, UV-Vis spectrum for SEK4 (2) with λ<sub>max</sub> at 231 and 297 nm consistent with previously reported literature<sup>4</sup>. c, ES<sup>-</sup> mass spectrum corresponding to SEK4 (2) observed [M-H]<sup>-</sup> 317.3, theoretical [M-H]<sup>-</sup> 317.1. d, UV-Vis spectrum corresponding to SEK4b (3) with λ<sub>max</sub> of 231 and 297, also consistent with literature<sup>4</sup>. e, as for c; ES<sup>-</sup> mass spectrum corresponding to SEK4b (3) observed deprotonated mass [M-H]<sup>-</sup> 317.3, theoretical mass is as for isomeric SEK4. Data presented represents three biological replicates and was acquired from E. coli BL21(DE3) pACYCAntΔAntA samples analysed in Supplementary Figure 5.6, Supplementary Figure 5.20 and Supplementary Figure 5.21.
Supplementary Figure 5.23: HPLC-UV-Vis-MS analysis of E. coli BL21(DE3) expressing AntA-G, I, ΔAntH

HPLC-UV-Vis-MS analysis of *E. coli* BL21(DE3) expressing *antA-G, I, ΔAntH*, identifying mutactin as predominant octaketide shunt metabolites in the ΔAntH anthraquinone biosynthetic
pathway. All chromatographic conditions and methods are as described in Supplementary Figure 5.22 and Supplementary Figure 5.18. 

a. Typical chromatogram of culture supernatant from *E. coli* BL21(DE3) pACYCAntΔH at 269 nm showing a peak corresponding mutactin which is not present in *E. coli* BL21(DE3) or *E. coli* BL21(DE3) pACYCDuet-1. 

b. UV-Vis spectrum for mutactin (6) with λ_max at 223 and 269 nm consistent with previously reported literature⁵. 


Data presented here is representative of three biological replicates and was acquired from *E. coli* pACYCAntΔAntH samples analysed in Supplementary Figure 5.6, Supplementary Figure 5.20 and Supplementary Figure 5.21.
Supplementary Figure 5.24: HPLC-UV-Vis-MS analysis of E. coli BL21(DE3) expressing AntA, B, D-I, ∆AntC

HPLC-UV-Vis-MS analysis of E. coli BL21(DE3) expressing antA, B, D-I, ∆AntC, identifying SEK34 as predominant octaketide shunt metabolites in the ∆AntC anthraquinone biosynthetic pathway. Chromatographic conditions and methods are as described in Supplementary Figure 5.22 and supplementary methods 4.2. a, Typical chromatogram of culture supernatant from E. coli BL21(DE3) pACYCAnt∆C at 258 nm showing a peak corresponding SEK34 which is not present in E. coli BL21(DE3) or E. coli BL21(DE3) pACYCDuet-1. b, UV-Vis spectrum for SEK34 (8) with λ<sub>max</sub> at 258 and 288 nm consistent with previously reported literature<sup>5</sup>. c, ES−
mass spectrum corresponding to SEK34 (8) observed [M-H]⁻ 301.1, theoretical [M-H]⁻ 301.1, presence of an abundant mass at 257.3 m/z is also previously reported for SEK34. Data presented is in strong agreement with two additional biological replicates and was acquired from *E. coli* BL21(DE3) pACYCAntΔAntC samples analysed in Supplementary Figure 5.6, Supplementary Figure 5.20 and Supplementary Figure 5.21.
Supplementary Figure 5.25: HPLC-UV-Vis-MS analysis of E. coli BL21(DE3) expressing AntA-I and AQ256 standards

HPLC-UV-Vis-MS analysis of E. coli BL21(DE3) expressing antA-I identifying AQ256 as the predominant end compound of the anthraquinone biosynthetic gene cluster. Data is representative of the same three biological replicates. All chromatographic conditions and methods are as described in Supplementary Figure 5.6, Supplementary Figure 5.20 and Supplementary Figure 5.21. a, Typical chromatogram of culture supernatant from E. coli BL21(DE3) pACYCAnthraquinone monitored at 434 nm showing a peak corresponding AQ256 which is not present in E. coli BL21(DE3) or E. coli BL21(DE3) pACYCDuet-1. Additionally, no other major peaks are present at this wavelength indicating this to be the only anthraquinone produced. b, Typical chromatogram of AQ256 analytical standard purified from E. coli BL21(DE3) pACYCAnthraquinone and characterised by $^1$H, $^{13}$C, COSY, HSQC and HMBC NMR spectroscopy. c, UV-Vis spectrum for AQ256 (1) with $\lambda_{max}$ at 244, 265, 284 and 434 nm consistent with the counterpart UV-Vis spectrum of the AQ256 standard in d. e and f, ES$^+$ mass
spectrum corresponding to AQ256 derived from *E. coli* BL21(DE3) pACYCAnthraquinone and AQ256 analytical standard respectively. Observed mass are [M-H]^{-} 254.9 and [M-H]^{-} 254.8, theoretical mass [M-H]^{-} 255. UV-Vis and mass spectrum are in good agreement between experimental samples and analytical standards.
### 5.11 Supplementary Tables

**Supplementary Table 5.1: Chain length factor gatekeeper residues for biosynthesis of different length nascent poly-β-ketide chains.**

<table>
<thead>
<tr>
<th>Host organism Gram stain</th>
<th>Cluster</th>
<th>Chain length factor gatekeeper residues</th>
<th>Chain length</th>
<th>Conserved Q residue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gram +</td>
<td>Actinorhodin</td>
<td>T F W F G</td>
<td>C16</td>
<td>Q</td>
</tr>
<tr>
<td>Gram +</td>
<td>Oxytetracycline</td>
<td>G L W F G</td>
<td>C18</td>
<td>Q</td>
</tr>
<tr>
<td>Gram +</td>
<td>WhiE spore pigment</td>
<td>G L I G S</td>
<td>C24</td>
<td>D</td>
</tr>
<tr>
<td>Gram +</td>
<td>Griseorhodin</td>
<td>G M L T G</td>
<td>C26</td>
<td>Q</td>
</tr>
<tr>
<td>Gram -</td>
<td>Anthraquinone</td>
<td>I V - W -</td>
<td>C16 (putative)</td>
<td>D</td>
</tr>
</tbody>
</table>

A table displaying the gatekeeper residues from a series of CLFs with bulky R-groups which sterically reduce the size of the amphipathic tunnel at the KS/CLF dimer interface. Residue order represents their proximity to the cavity entrance. Red residues define the bottom of the cavity while blue AAs are smaller residues from homologues producing longer polyketides. Gatekeeper residues do not map to the anthraquinone sequence, prediction of chain length using this method suggest the nascent poly-β-ketide to be C\(_{20}\). Figure adapted from Tang, Tsai and Khosla (2003).
## Supplementary Table 5.2: Plasmids used in this study

<table>
<thead>
<tr>
<th>Plasmid name</th>
<th>Insert Description</th>
<th>Backbone description</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>pACYCDuet-1</td>
<td>-</td>
<td>p15A ori, cat (Cm&lt;sup&gt;R&lt;/sup&gt;), T7 promoters</td>
<td>Novagen</td>
</tr>
<tr>
<td>pETDuet-1</td>
<td>-</td>
<td>pBR322 ori, bla (Ap), T7 promoters</td>
<td>Novagen</td>
</tr>
<tr>
<td>pG9m-2-ActKRRef</td>
<td>Refactored Sco5086, codon optimised for expression in E. coli</td>
<td>pUC19 ori, Amp&lt;sup&gt;R&lt;/sup&gt;</td>
<td>This study*</td>
</tr>
<tr>
<td>pG9m-2-ActARO/CYCRef</td>
<td>Refactored Sco5090, codon optimised for expression in E. coli</td>
<td>pUC19 ori, Amp&lt;sup&gt;R&lt;/sup&gt;</td>
<td>This study*</td>
</tr>
<tr>
<td>pBbA2k-RFP</td>
<td>RFP</td>
<td>p15A ori, kan/Neo (Kn&lt;sup&gt;R&lt;/sup&gt;/Neo&lt;sup&gt;R&lt;/sup&gt;), pTet</td>
<td>22</td>
</tr>
<tr>
<td>pBbB1a-GFP</td>
<td>GFP</td>
<td>BBR1 ori, bla (Ap), T7 promoter</td>
<td>22</td>
</tr>
<tr>
<td>pBbA2k-plumPKS</td>
<td>plu&lt;sub&gt;4191&lt;/sub&gt;, plu&lt;sub&gt;4190&lt;/sub&gt;, plu&lt;sub&gt;4189&lt;/sub&gt;</td>
<td>p15A ori, kan/Neo (Kn&lt;sup&gt;R&lt;/sup&gt;/Neo&lt;sup&gt;R&lt;/sup&gt;), pTet</td>
<td>This study</td>
</tr>
<tr>
<td>pBbB1a-plumPKS</td>
<td>plu&lt;sub&gt;4191&lt;/sub&gt;, plu&lt;sub&gt;4190&lt;/sub&gt;, plu&lt;sub&gt;4189&lt;/sub&gt;</td>
<td>BBR1 ori, bla (Ap), T7 promoter</td>
<td>This study</td>
</tr>
<tr>
<td>pETDuetPlu419091</td>
<td>plu&lt;sub&gt;4191&lt;/sub&gt;, His6:&lt;sub&gt;plu&lt;/sub&gt;</td>
<td>pBR322 ori, bla (Ap), T7 promoters</td>
<td>This study</td>
</tr>
<tr>
<td>pACYCPlu4188</td>
<td>plu&lt;sub&gt;4188&lt;/sub&gt;</td>
<td>This study</td>
<td></td>
</tr>
<tr>
<td>pACYCPlu418893</td>
<td>plu&lt;sub&gt;4188&lt;/sub&gt;, plu&lt;sub&gt;4193&lt;/sub&gt;</td>
<td>This study</td>
<td></td>
</tr>
<tr>
<td>pACYCDuetPlu4192-94</td>
<td>plu&lt;sub&gt;4194&lt;/sub&gt;, 93, 92</td>
<td>This study</td>
<td></td>
</tr>
<tr>
<td>pACYCAntraquinone</td>
<td>plu&lt;sub&gt;4194&lt;/sub&gt;, 93, 92, and plu&lt;sub&gt;4091&lt;/sub&gt;, 90, 89, 88, 87, 86</td>
<td>This study</td>
<td></td>
</tr>
<tr>
<td>pACYCAntrefKR</td>
<td>plu&lt;sub&gt;4193&lt;/sub&gt;, 92, and plu&lt;sub&gt;4091&lt;/sub&gt;, 90, 89, 88, 87, 86, (∆94), and refactored Sco5086, plu&lt;sub&gt;4193&lt;/sub&gt;, 92, and plu&lt;sub&gt;4091&lt;/sub&gt;, 90, 89, 88, 87, 86, (∆94), and wild type Sco5086</td>
<td>This study</td>
<td></td>
</tr>
<tr>
<td>pACYCAntwtKR</td>
<td>plu&lt;sub&gt;4194&lt;/sub&gt;, 93, 92, and plu&lt;sub&gt;4091&lt;/sub&gt;, 90, 89, 88, 87, 86, (∆94)</td>
<td>This study</td>
<td></td>
</tr>
<tr>
<td>pACYCAntΔAntA</td>
<td>plu&lt;sub&gt;4193&lt;/sub&gt;, 92, and plu&lt;sub&gt;4091&lt;/sub&gt;, 90, 89, 88, 87, 86, (∆94)</td>
<td>p15A ori, cat (Cm&lt;sup&gt;R&lt;/sup&gt;), T7 promoters</td>
<td>This study</td>
</tr>
<tr>
<td>pACYCAntrefCYC</td>
<td>plu&lt;sub&gt;4194&lt;/sub&gt;, 93, 92, and plu&lt;sub&gt;4091&lt;/sub&gt;, 90, 89, 88, 86, (∆87) refactored Sco5090 plu&lt;sub&gt;4194&lt;/sub&gt;, 93, 92, and plu&lt;sub&gt;4091&lt;/sub&gt;, 90, 89, 88, 86, (∆87) wild type Sco5090</td>
<td>This study</td>
<td></td>
</tr>
<tr>
<td>pACYCAntwtCYC</td>
<td>plu&lt;sub&gt;4194&lt;/sub&gt;, 93, 92, and plu&lt;sub&gt;4091&lt;/sub&gt;, 90, 89, 88, 86, (∆87)</td>
<td>This study</td>
<td></td>
</tr>
<tr>
<td>pACYCAntΔAntH</td>
<td>plu&lt;sub&gt;4194&lt;/sub&gt;, 93, 92, and plu&lt;sub&gt;4091&lt;/sub&gt;, 90, 89, 88, 86, (∆87)</td>
<td>This study</td>
<td></td>
</tr>
<tr>
<td>pACYCAntΔAntC</td>
<td>plu&lt;sub&gt;4194&lt;/sub&gt;, 93 and plu&lt;sub&gt;4091&lt;/sub&gt;, 90, 89, 88, 87 86, (∆92)</td>
<td>This study</td>
<td></td>
</tr>
<tr>
<td>pACYCAntΔAntI</td>
<td>plu&lt;sub&gt;4194&lt;/sub&gt;, 93 and plu&lt;sub&gt;4091&lt;/sub&gt;, 90, 89, 88, 87, (∆86)</td>
<td>This study</td>
<td></td>
</tr>
</tbody>
</table>

* Sequence optimised using GeneArt GeneOptimiser (ThermoFischer Scientific) and synthesised by Gen9.
Supplementary Table 5.3: AQ256 carbon and proton assignment and coupling constants

<table>
<thead>
<tr>
<th>Position</th>
<th>δ\text{c}, (ppm)</th>
<th>δ\text{H}, (ppm)</th>
<th>(mult., J in Hz)</th>
<th>HSQC</th>
<th>HMBC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>165.35</td>
<td>H-D</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>107.64</td>
<td>6.60, d (2.45)</td>
<td>C-2</td>
<td>C-1, 4</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>166.16</td>
<td>H-D</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>108.84</td>
<td>7.23, d (2.45)</td>
<td>C-4</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>7.77, dd (7.34, 0.98)</td>
<td>C-7, 10, 12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>119.13</td>
<td>0.98</td>
<td>C-5</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>136.12</td>
<td>7.71, t (8.13, 8.13)</td>
<td>C-6</td>
<td>C-11, 12</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>124.09</td>
<td>7.3, dd (8.31, 0.98)</td>
<td>C-7</td>
<td></td>
<td>12</td>
</tr>
<tr>
<td>8</td>
<td>162.00</td>
<td>H-D</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>190.86</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>181.67</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>133.56</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>115.37</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>109.16</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>135.50</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Supplementary Table 5.4: Theoretical masses for all shunt metabolites

<table>
<thead>
<tr>
<th>Name</th>
<th>Molecular formula</th>
<th>Exact mass</th>
<th>[M+H]^+</th>
<th>[M-H]^−</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEK4</td>
<td>C_{16}H_{14}O_{7}</td>
<td>318.0740</td>
<td>319.0812</td>
<td>317.0667</td>
</tr>
<tr>
<td>SEK4b</td>
<td>C_{16}H_{14}O_{7}</td>
<td>318.0740</td>
<td>319.0812</td>
<td>317.0667</td>
</tr>
<tr>
<td>B26</td>
<td>C_{16}H_{12}O_{6}</td>
<td>300.0634</td>
<td>301.0707</td>
<td>299.0561</td>
</tr>
<tr>
<td>AUR367</td>
<td>C_{16}H_{12}O_{6}</td>
<td>300.0634</td>
<td>301.0707</td>
<td>299.0561</td>
</tr>
<tr>
<td>Mutactin</td>
<td>C_{16}H_{14}O_{6}</td>
<td>302.0790</td>
<td>303.0863</td>
<td>301.0718</td>
</tr>
<tr>
<td>Dehydromutactin</td>
<td>C_{16}H_{12}O_{6}</td>
<td>284.0685</td>
<td>285.0758</td>
<td>283.0612</td>
</tr>
<tr>
<td>SEK34</td>
<td>C_{16}H_{14}O_{6}</td>
<td>302.0790</td>
<td>303.0863</td>
<td>301.0718</td>
</tr>
<tr>
<td>SEK34b</td>
<td>C_{16}H_{12}O_{6}</td>
<td>284.0685</td>
<td>285.0758</td>
<td>283.0612</td>
</tr>
<tr>
<td>Aloesaponarin II</td>
<td>C_{15}H_{10}O_{4}</td>
<td>254.0579</td>
<td>255.0652</td>
<td>253.0506</td>
</tr>
<tr>
<td>DMAC</td>
<td>C_{16}H_{10}O_{6}</td>
<td>298.0477</td>
<td>299.0550</td>
<td>297.0405</td>
</tr>
<tr>
<td>(s)-DNPA</td>
<td>C_{16}H_{14}O_{5}</td>
<td>286.0841</td>
<td>287.0914</td>
<td>285.0768</td>
</tr>
<tr>
<td>AQ256</td>
<td>C_{14}H_{6}O_{3}</td>
<td>256.0372</td>
<td>257.0445</td>
<td>255.0299</td>
</tr>
<tr>
<td>AQ270</td>
<td>C_{15}H_{10}O_{5}</td>
<td>270.0528</td>
<td>271.0601</td>
<td>269.0455</td>
</tr>
<tr>
<td>AQ270b</td>
<td>C_{15}H_{10}O_{5}</td>
<td>270.0528</td>
<td>271.0601</td>
<td>269.0455</td>
</tr>
<tr>
<td>AQ284</td>
<td>C_{16}H_{12}O_{5}</td>
<td>284.0685</td>
<td>285.0758</td>
<td>283.0612</td>
</tr>
<tr>
<td>AQ284b</td>
<td>C_{16}H_{12}O_{5}</td>
<td>284.0685</td>
<td>285.0758</td>
<td>283.0612</td>
</tr>
<tr>
<td>AQ300</td>
<td>C_{16}H_{12}O_{6}</td>
<td>300.0634</td>
<td>301.0707</td>
<td>299.0561</td>
</tr>
<tr>
<td>AQ314</td>
<td>C_{17}H_{14}O_{6}</td>
<td>314.0790</td>
<td>315.0863</td>
<td>313.0718</td>
</tr>
<tr>
<td>1,3,8-trihydroxyanthracen-9(10H)-one</td>
<td>C_{14}H_{10}O_{4}</td>
<td>242.0579</td>
<td>243.0652</td>
<td>241.0506</td>
</tr>
<tr>
<td>1,3,8-trihydroxy-Dianthrone</td>
<td>C_{28}H_{18}O_{6}</td>
<td>482.10017</td>
<td>483.1074</td>
<td>481.0929</td>
</tr>
</tbody>
</table>

Isomers are highlighted in corresponding colours. All masses are reported as atomic mass units.
**Supplementary Table 5.5: Primers used for plasmid construction**

<table>
<thead>
<tr>
<th>Primer name</th>
<th>Primer sequence 5’-3’</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plu4189_rev_XhoI</td>
<td>atcagtctgagatcttttat gttaacttgatagtt</td>
</tr>
<tr>
<td>Plu4191_for_BgIII</td>
<td>cggctagcatctgaagataaaatgaaacagaaatgtctt</td>
</tr>
<tr>
<td>Plu4191_rev_KpnI</td>
<td>atgcatgtaccttagcctgtctact</td>
</tr>
<tr>
<td>plu4190_for_EcoRI</td>
<td>atgcatactctgtaaaaaggtgctc</td>
</tr>
<tr>
<td>Plu4190_rev_PstI</td>
<td>tgcatactgcagnaacatgtggaaacagaggtct</td>
</tr>
<tr>
<td>Plu4193_for_NoI_unTAGGED</td>
<td>gcgtgagccttagggtgcagcatattcttttatcatctga</td>
</tr>
<tr>
<td>Plu4193_rev_HindIII</td>
<td>gcgtgagccttagggtgcagcatattcttttatcatctga</td>
</tr>
<tr>
<td>Plu4188_rev_Ndel</td>
<td>ggttcctagtagaaactaaatcagatgtgatctcag</td>
</tr>
<tr>
<td>Plu4188_rev_XhoI</td>
<td>ttaacctcggatttgaactatctaatgtgataatttaggttttagga</td>
</tr>
<tr>
<td>Plu4194_for_Ndel</td>
<td>gcgtgagccttagggtgcagcatattcttttatcatctga</td>
</tr>
<tr>
<td>Plu4192_for_XhoI</td>
<td>tataacctcggatttgaactatctaatgtgataatttaggttttagga</td>
</tr>
<tr>
<td>AQBGC_Rev_PstI</td>
<td>gcgtgagccttagggtgcagcatattcttttatcatctga</td>
</tr>
<tr>
<td>plu4186A_for</td>
<td>taaatcctttatacttaaactaggtgtcataacactcttttgc</td>
</tr>
<tr>
<td>plu4186A_rev</td>
<td>tataataacttttaataacttgctatgtgatagttgc</td>
</tr>
<tr>
<td>KR_swap_IF</td>
<td>atgtatatctcccttcataactaatataattaagatgtggttttagga</td>
</tr>
<tr>
<td>Plu4194_del_IF_for</td>
<td>tgtaacgctcaagcaacttaa</td>
</tr>
<tr>
<td>Sco5086_for_IF</td>
<td>gaaggagatatacagactgcgaccaggactcgaagtcgcactg</td>
</tr>
<tr>
<td>Sco5086_rev_IF</td>
<td>gcttgagccgttagccagccgagctcgaagctcgaagtcgcactg</td>
</tr>
<tr>
<td>Ref_Sco5086_IF_for</td>
<td>gaaggagatatacagactgcgaccaggactcgaagttgtaaaaggtggttttagga</td>
</tr>
<tr>
<td>Ref_Sco5086_IF_rev</td>
<td>gcttgagccgttagccagccgagctcgaagctcgaagttgtaaaaggtggttttagga</td>
</tr>
<tr>
<td>plu4187_replace_fw</td>
<td>taacttaaaagggaaaaaataagagtaaatttaggttttagga</td>
</tr>
<tr>
<td>plu4188_rev</td>
<td>tattgatcctactaatgtaaaatagtttttagga</td>
</tr>
<tr>
<td>Sco5090_for_IF</td>
<td>taacttggagctaatgtaaaatagagctcgaagctcgaagttgtaaaaggtggttttagga</td>
</tr>
<tr>
<td>Sco5090_rev_IF</td>
<td>tattgtccttttagaatgattgtgagctcgaagctcgaagttgtaaaaggtggttttagga</td>
</tr>
<tr>
<td>Ref_Sco5090_IF_for</td>
<td>taacttggagctaatgtaaaatagagctcgaagctcgaagttgtaaaaggtggttttagga</td>
</tr>
<tr>
<td>Ref_Sco5090_IF_rev</td>
<td>tattgtccttttagaatgattgtgagctcgaagctcgaagttgtaaaaggtggttttagga</td>
</tr>
<tr>
<td>plu4187_delta_for</td>
<td>acttaggcataaaactctaatgaaatagttgtctaatgtaaaataggtggttttagga</td>
</tr>
<tr>
<td>Plu4187_delta_rev</td>
<td>gtttactgctcaagttaagtttagtactctatctaatgtaaaataggtggttttagga</td>
</tr>
<tr>
<td>Plu4186A_for</td>
<td>taaatcctttatacttaaactaggtgtcataacactcttttgc</td>
</tr>
<tr>
<td>Plu4186A_rev</td>
<td>tataataacttttaataacttgaaaaataggtggttttagga</td>
</tr>
<tr>
<td>Plu4192A_for</td>
<td>attttagctgatactcaggtcgttagtcggaaagaaaac</td>
</tr>
<tr>
<td>Plu4192A_rev</td>
<td>tatcagactaaatattaactcatcttctttgtcttttataac</td>
</tr>
</tbody>
</table>

**Primer nomenclature** is typically: gene/region amplified_direction of amplification_restriction endonuclease site. All primers with additional 5’ restriction endonuclease sequences are preceded with an additional random 6 bp sequence to facilitate PCR product digestion. **Restriction endonuclease recognition sequences** are **bolded**. Primers were designed and verified with IDT oligoanalyser.
5.12 Supplementary References

6 Engineering *Escherichia coli* soluble type II polyketide synthases for biosynthesis of decaketides and above *in vivo*.

6.1 Preface
This chapter constitutes currently unpublished work. To avoid unnecessary repetition, methods and materials developed and optimised for AQ256 production, isolation and characterisation in Chapter 5, ‘Assembling a modular production line for combinatorial biosynthesis of aromatic polyketides in *Escherichia coli*’ are cited opposed to being rewritten.

Furthermore, this results chapter is written in an extended journal style format comprising an introduction, methods and materials (with exceptions), results and discussion, and short perspectives. A Nature citation style is adopted for consistency across the thesis. All of the work here-in was undertaken by M.C.. As with all chapters, a self-contained bibliography is attached at the end.
6.2 Abstract

Crystal structures provide an exemplary view of a protein, and are capable of catching snapshots of vast megasynthase proteins such as non-ribosomal peptide synthases (NRPSs) in a suite of conformations. The crystal structures of specialised metabolite biosynthetic enzymes are of particular interest, as structural elucidation is fundamental when attempting rational reengineering for redirecting a catalytic process. Despite description over a decade ago, the actinorhodin (act) ketosynthase and chain length factor (KS/CLF) crystal structure remains the only representative of a type II polyketide synthase available publicly, and has been used as a cornerstone in type II PKS structural biology.\textsuperscript{1,2}

The act KS/CLF presents a good model for Actinobacterial KS/CLFs. However, in this work we show the amino acid sequence dissimilarities between the act KS/CLF and AntDE mires prediction of polyketide chain length using homology based approaches, impeding redirecting AntDE for production of decaketides and beyond. We therefore aim to solve the crystal structure of the first visibly soluble type II polyketide synthase (PKS) complex in \textit{E. coli} with two underlying motivations in mind. Firstly, reliable identification and modification of key gatekeeper amino acids residues encoded within the KS/CLF would enable one to rationally re-programme the biosynthetic machinery to produce an unreduced polyketide chain of choice length. Knowing which amino acid substitutions will confer a change in polyketide chain length from C\textsubscript{16} to a C\textsubscript{20} or C\textsubscript{24} etc., will facilitate biosynthesis of a wider range of pharmaceutically relevant aromatic polyketides in \textit{E. coli} than currently accessible. Secondly, the remarkable solubility of the anthraquinone KS/CLF complex is unique in the literature. Whilst the source of solubility is unknown, elucidation of the physiochemical make-up of the AntDE solvent accessible residues may provide insight when attempting to rationally engineer native type II PKS machinery from Actinobacteria for overexpression in \textit{E. coli}. Whilst we do not solve the AntDE crystal structure we attempt to evaluate the size of the KS/CLF dimeric cavity and take important steps towards purification of the ant KS/CLF as a heterodimeric complex. Furthermore, we demonstrate the remarkable solubility of an ant KS/CLF pseudo-dimer fusion protein.
6.3 Introduction

6.3.1 Modesty in type II nascent polyketide chain biosynthesis

Polyketide synthases (PKSs) were classically, and are routinely, categorised into three classes: type I, II and III. The dynamic evolution of specialised metabolism and experimental characterisation of vibrant hybrid PKS biosynthetic gene clusters (BGCs) shows this taxonomy to be an oversimplification\(^3\), however, the general assignment of PKS machinery remains a relic from fatty acid biosynthetic machinery and is consistent with literature\(^4\). Efforts are being made to modernise and redefine natural product classification to house exotic, hybrid, under-exploited and hitherto unknown classes and subclasses of natural product biosynthetic machinery\(^5\). This is reflected in the 8 different polyketide subclasses currently available on the MIBiG community annotation submission form\(^5\) (http://mibig.secondarymetabolites.org/form2.html). In view of the expanded chemistry introduced into specialised metabolites\(^6\) by the extended family of polyketide synthases, and hybrids thereof\(^7\)-\(^9\), the nascent polyketide chain derived from prokaryotic type II PKSs remain relatively conservative\(^10\) (Figure 6.1). Here, the diversity in the nascent polyketide is only two-fold and dictated firstly by the choice of starter unit incorporated, selected from a modest but growing pool of substrates\(^11\), and secondly, through variation in the number of subsequent Clasien condensation reactions, using malonyl-Coenzyme A (CoA) as an extender substrate exclusively. Remarkably, with the addition of tailoring enzymes, this modest polyketide chain can form a wide variety of different chemotypes core polyketide scaffolds including benzoisochromanequinones\(^12\), tetracyclines\(^13\), anthracyclines\(^14\), tetracenomycins\(^15\), aureolic acids\(^16\), pentangular polyphenols\(^17\), angucyclines\(^18\), pluramycin-type\(^19\), discoid polyphenols\(^20\) etc. (Figure 6.2). Both mechanisms for introduction of diversity into the nascent polyketide chain are discussed here.

\[ \text{R1: Me, Pr etc.} \]

\[ \text{R1: nitrobenzyol-, 4,5-dichloropyrroyl-, guanidinoacetyl- derived etc.} \]
\[ \text{R2: Me, Et, ChloroEt, Methoxy, Hydroxy, Amino etc.} \]
Figure 6.1: Chemistry of nascent polyketide chains

Schematic of nascent poly-β-ketide chains from type II (a) and type I (b) polyketide synthases after 8 condensation events. Variation is introduced into the type II polyketide by choice of starter (SU) unit only (a, R₁) e.g. actinorhodin comprises an acetyl- group and frenolicin comprises a butyryl- group. Contrastingly, variation can be introduced using a wider variety of starter units in type I PKSs (b, R₁) than the typical alkyl- group utilised by type II PKSs, including modified amino acids and halogenated substrates e.g. aureothin utilising a Nitrobenzyl- SU⁷, pyoluteorin utilising a 4,5-dichloropyrroyl- SU⁸ and guadinomine B utilising a guanidinoacetyl- SU⁹. Additional diversity can be introduced through the choice of the extender unit, which unlike type II PKSs is not constrained to malonyl-CoA. Here stereospecifically α-branched acyl- substrates can be incorporated (b, R₂). Moreover, each extender unit β-ketide (C₁₃) can be selectively reduced to the corresponding alcohol (C₁₁), olefin (C₉) or alkane (C₇) after each condensation event individually²¹. The mechanisms to introduce diversity during biosynthesis of type I derived nascent polyketides is much greater than the counterpart from type II PKSs derived from the same number of condensation reactions.

Figure 6.2: Chemotype/chromophores accessible by type II polyketide synthases and cognate tailoring enzymes.

Different type II chemotype core scaffolds including representatives of benzoisochromanequinones (frenolicin²²), tetracyclines (oxytetracycline²³), anthracyclines (doxorubicin²⁴), tetracenomycins (tetracenomycin C¹⁵), aureolic acids (mithramycin²⁵, shown as an aglycone. R₁: a disaccharide
comprising two olivose sugars. R₂: a trisaccharide comprising two olivose sugars and a terminal mycarose sugar), pentangular polyphenols (fredericamycin\textsuperscript{26}), angucyclines (Urdamycin\textsuperscript{27}, shown as an aglycone. R₁: Olivose sugar. R₂: trisaccharide comprising two rhodinose sugars and one olivose sugar), pluramycin-type (Hedamycin\textsuperscript{28}) and discoid polyphenols (resistomycin\textsuperscript{20}). Choice starter units discussed below are shown in red.

6.3.2 Diversity in type II polyketide chains: Priming biosynthesis

Priming the biosynthesis of type II polyketides with non-acetate starter units is a powerful mechanism to introduce diversity into the polyketide core. Understanding the biosynthetic logic leading to incorporation of alternate start units is critical when attempting to reengineer these systems effectively, however, these routes vary and can be orthogonal\textsuperscript{29}. Daunorubicin and doxorubicin, two heavily clinically utilised anthracyclines, are primed with a propionyl- starter unit\textsuperscript{30}. Installation of the propionyl- starter unit requires an additional priming ketosynthase III (KSIII - DpsC\textsuperscript{31}) similar to FabH, which catalyses the first Claisen condensation of propionyl-CoA with malonyl-acyl carrier protein (ACP) to form a $\beta$-ketidepentanoyl-ACP diketide, which is further elongated by the cognate ketosynthase/chain length factor (KS/CLF) PKS\textsuperscript{31} (Figure 6.3a). The same mechanism is observed during initiating of frenolicin biosynthesis which utilises a butyryl- starter unit\textsuperscript{22}. Here, the condensation of butyryl-CoA with the first malonyl group is catalysed by FrenJ (KSIII), however FrenJ is orthogonal for a second BGC encoded ACP (FrenI) and turn over with malonyl-FrenN, the ACP of the minimal PKS, is negligible\textsuperscript{22,29}. The co-occurrence of a KSIII and second ACP is common in specialised metabolites utilising nonacetate starter units. Whilst doxorubicin and frenolicin stringently utilise a single nonacetate starter unit, starter unit incorporation during R1128 biosynthesis is of particular interest because the underlying enzymology is promiscuous\textsuperscript{32,33}. R1128 biosynthesis can be initiated with either an acetyl-, propionyl-, isobutyryl- or isobutyryl-CoA primed diketide, with a preference for propionyl-CoA \textit{in vitro}\textsuperscript{32}. Once more the initial decarboxylative condensation between the acyl-CoA and malonyl-ACP is catalysed by an additional KSIII component, ZhuH, and is selective for the second cluster encoded ACP, ZhuG. The C3 $\beta$-ketide of each diketide starter unit is further reduced to a methylene prior to acyl-transfer and elongation by the R1128 KS/CLF complex. The reductive enzymes involved in this further modification are not known but are thought to be FabG, FabA and FabI functional homologues which are shared between primary and specialised metabolism\textsuperscript{32} (Figure 6.3b).

Complexities of starter unit (SU) biosynthesis do not stop here however: the anticancer agent hedamycin exploits two multidomain type I PKS proteins, HedT and HedU, to synthesise a hexadienyl- starter unit\textsuperscript{34} (Figure 6.3c) which is further elongated with 9 malonyl- extender units by the cognate KS/CLF and is ultimately modified into the bisepoxide moiety which confers its function\textsuperscript{35}. A hexadienyl- starter unit is also used during fredericamycin biosynthesis; however, the route to biosynthesis is markedly different to hedamycin\textsuperscript{19}. Here two routes are envisaged. Firstly, a single priming KS, FdmS, catalyses decarboxylative condensation of acetyl-CoA and two malonyl-ACP groups followed by successive reduction and dehydration to yield hexadienyl-
ACP (Figure 6.3d). Secondly, FdmS catalyses a single condensation of a butyryl-CoA or crotonyl-CoA with malonyl-ACP followed by ketoreduction and dehydration, and oxidation should butyryl-CoA be used, prior to further elongation by the cognate KS/CLF (Figure 6.3e)36. Both routes are thought to be plausible and employed preferentially with respect to the cellular abundance of each acyl-CoA36. More recently a bifunctional acetyltransferase/decarboxylase previously only associated with modular type I PKSs has been implicated in SU biosynthesis of lomaiviticin37. The BGC encoding the propionyl-primed lomaivitin does not encode an additional priming KSIII unlike doxorubicin, rather it encodes lom62 an acetyltransferase/decarboxylase lnmK homologue37. LmnK catalyses the transfer of methylmalonyl-CoA to an ACP and subsequent decarboxylation during biosynthesis of the type I polyketide leinamycin and a functionally similar reaction was observed during lomaivitacin biosynthesis37. Waldman and Balskus showed a dedicated priming ACP (Lom63) to be preferentially acylated by methylmalonyl-CoA, and not propionyl-CoA, forming methylmalonyl-Lom63 before subsequent decarboxylation to propionyl-Lom63, also catalysed by Lom62, presumably prior to elongation of the nascent polyketide chain by the cognate KS/CLF37 (Figure 6.3f). The co-opting of ACP priming mechanisms between type I and type II PKS is rare, and only two Lom62/LmnK homologues were found to associate with additional putative type II clusters (as of 2013). Engineering Lom62 homologues represents a novel way to introduce starter unit diversity by feeding more complex acyl-CoA substrates, circumventing the need for KSIII components37.

Most minimal PKS components are interchangeable between type II polyketide biosynthetic pathways, however, this is not necessarily the case for more complex priming PKS components, collectively termed the initiation module. Combinatorial biosynthesis using initiation module ACPs and KSIIIs has been demonstrated, exemplified in Figure 6.4, however the accompanying reductive enzymatic machinery, moonlighting from primary metabolism and necessary to form a comprehensive initiation module, introduces additional complexity when reconstructing biosynthetic pathways in alternative host backgrounds and components can be cluster orthogonal29. Furthermore, some initiation module components are poorly soluble when expressed in model heterologous hosts, as is the case for HedT and U in both E. coli and S. coelicolor19. Therefore, the first and easiest step to introduce a large number of structural perturbations is to alter a handful of gatekeeper residues which control the chain length of the nascent poly-β-ketide.
Figure 6.3: Type II PKS starter unit biosynthetic strategies

Variation in the biosynthesis of type II starter units from 6 different type II polyketides. a, biosynthesis of propionyl-CoA (doxorubicin) and butyryl-CoA (frenolicin), $R_1$: CH$_3$ and CH$_2$CH$_3$ respectively. During doxorubicin biosynthesis the minimal PKS associated ACP is used, rather than an orthogonal priming ACP. b, biosynthesis of R1128 starter units harnessing an unknown ketoreductase (KR), dehydratase (DH) and enoyl reductase (ER) thought to moonlight from fatty acid synthesis. $R_1$: CH$_3$, CH$_2$CH$_3$, CH(CH$_3$)$_2$ or (CH$_2$)$_2$CH$_3$. c, d, and e, three alternative routes to hexadienyl-ACP biosynthesis. f, KSIII independent route to propionyl-ACP biosynthesis during lomaivitacin biosynthesis. Cofactors, additional reactants and products and specific enzyme nomenclature are omitted where appropriate to improve clarity.
Figure 6.4: combinatorial biosynthesis using alternative initiation modules

A schematic of polyketide structures which have been produced through swapping initiation modules. a, AD284a, AD284b, AD284c novel polyketide synthesised by the act mPKS and fredericamycin initiation module comprising FdmS (KSIII), FdmC (KR), FdmH (ACP) and FdmW (PPTase)36. b, YT46, YT46b, YT84, YT84b polyketide synthesised by the act mPKS, C9 KR and the R1128 initiation module comprising ZhuH and ZhuG and accompanying MCAT. R1: H or CH₃38. c, canonical C₁₆ polyketide product of the act mPKS (SEK4 and 4b), and d, the canonical product of the act mPKS and C9 ketoreductase, ActIII. Starter unit derived moieties are shown in red.
6.3.3 Controlling chain length

6.3.3.1 Type II polyketide chain length: the general dogma

The general, and accepted, dogma for polyketide chain length states chain length of type II PKS derived aromatic polyketides is dictated by the size of a cavity situated at the KS/CLF heterodimer interface (Figure 6.5). The growing polyketide chain is extruded into the cavity until the cavity reaches capacity, upon which the polyketide chain cannot grow further and is sequentially modified to the mature polyketide by additional tailoring enzymes. The chain length factor is the primary PKS constituent controlling the length of the polyketide chain, although the KS and in some cases additional tailoring enzymes also contribute (expanded below).

Figure 6.5: KS/CLF dimeric cavity

The size of the dimeric cavity at the KS/CLF interface determines the length of the nascent polyketide chain produced. The CLF and KS are represented schematically as yellow and red heterodimeric complexes. The unreduced growing nascent polyketide chain is tethered to an acyl carrier protein (ACP, blue).

The first insights into the functional role of chain length factors in type II PKS biosynthesis were derived from heterologous and combinatorial expression of the decaketide producing tetracenomycin (tcm), and octaketide producing granaticin (gra) and actinorhodin (act) KS and CLF parts in Streptomyces coelicolor. Combinatorial swapping of the C16 benzoisochromanequinone forming polyketide synthases were successful: PKS components comprising the act KS and gra CLF, or vise versa, and act ACP produced aloesaponarin II and 3,8-dihydroxy-1-methylantraquinone-2-carboxylic acid (DMAC), known actinorhodin shunt metabolites. Complementation of the tcm KS (C20) with the act CLF (C16) also resulted in aloesaponarin II and DMAC biosynthesis, despite the native tcm cluster exclusively producing a
decaketide. These salient experiments showed the KSβ monomer (CLF) as the determiner of polyketide chain length and coined the chain length factor, however, at this time it was unknown whether the chain length was controlled by the number of condensation reactions, or by a measurement function.

An act KS/CLF homology model aided the deduction that the CLF governs chain length by measurement, opposed to the number of condensation reactions. The first homology model (PDB: 1QXG) was constructed using a combination of crystal structures comprising E. coli FabF and FabB, and a FabF homologue from Synechocystis sp. PCC 6803 (PDB: 1KAS, 1F91 and 1E5M, respectively) for the ketosynthase, and E. coli and Synechocystis sp. PCC 6803 FabF sequences for the CLF. This model showed an 18 Å long channel spanning from the catalytically active cysteine, C169, of the KS across the KS/CLF dimer interface which was proposed to be the tunnel into which the growing polyketide chain is extruded. Multiple sequence alignments with other chain length factor sequences showed the residues lining the channel to be well conserved however residues capping the tunnel were substituted for less bulky amino acids in CLFs responsible for biosynthesis of longer C20 and C24 polyketides. This observation led to the idea of a suite of ‘gatekeeper’ residues within the CLF which are encoded either in the open or closed position and control polyketide chain length by selectively occluding the size of the tunnel into which the polyketide grows, thereby sterically inhibiting polyketide elongation (Figure 6.5). In short, PKSs with a larger tunnel produce longer unreduced polyketide chains and vise versa. The gatekeeper residue hypothesis was experimentally validated through introduction of large-to-small AA mutations in the C16 act KS/CLF resulting in production of C20 decaketides, and small-to-large mutation in the C20 tcm KS/CLF resulting in production of shorter C16 octoketides. Remarkably, in the tcm CLF polyketide chain length could be altered through a single G116T point mutation, illustrating the power of this approach to engineer polyketide length.

The complement of gatekeeper residues identified from the KS/CLF homology model were expanded from F109, T112 and F116 to also include G193, W194 in the act CLF and L143 in the act KS after elucidation of the actinorhodin KS/CLF crystal structure (PDB: 1TQY). These additional residues are differentially functionally substituted in C20, C24 and longer (≥C26) PKSs, following the gatekeeper residue dogma. Opening further gatekeeper residues in the act KS/CLF, in attempts to produce and C24 polyketides abolished biosynthesis all together, suggesting these gatekeeper residues to also play structural roles in heterodimer formation, and that more refined mutations are necessary to access this chemical space. Calculating the size of the tunnel at the KS/CLF interface using these gatekeeper residues is now considered a benchmark to determine polyketide length in enzymatic machinery that have not been experimentally characterised and has recently been used to trace the evolution of type II PKS clusters. Furthermore, gatekeeper residues have facilitated a facile calculation of the tunnel which scales with polyketide length as expected (Table 6.1).
Table 6.1: Prediction of KS/CLF tunnel volume for varying chain lengths

<table>
<thead>
<tr>
<th></th>
<th>C_{16-18}</th>
<th>C_{20}</th>
<th>C_{24}</th>
<th>C_{26}</th>
<th>C_{28-30}</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>706</td>
<td>887</td>
<td>1053</td>
<td>1049</td>
<td>1134</td>
</tr>
</tbody>
</table>

Cavity volume is denoted as $\lambda^3$ and calculated from homology models of longer chain KS/CLF using the act KS/CLF (PDB: 1TQY). Polyketide length is denoted as $C_n$. Table 6.1 is adapted from Hillenmeyer et al., (2015).^2

6.3.3.2 Non-CLF determinants of polyketide chain length

The chain length factor is not the only PKS component which determines the chain length of a growing polyketide itself, however. In the case of the $C_{30}$ polyketide fredericamycin, co-expression of the minimal PKS components, MCAT and cognate PPTase in $S. coelicolor$ CH999 resulted in the formation of the truncated $C_{22}$ undecaketides TW94b and TW94c, and $C_{24}$ dodecaketide TW94d, and not the full length $C_{30}$ product. Co-expression of the fredericamycin initiation module components, described in 6.3.2, didn’t restore $C_{30}$ polyketide biosynthesis but rather resulted in the formation of the hexanoyl- primed dodecaketide $C_{24}$ AD210a. And, irrespective of priming the fredericamycin KS/CLF with a shorter acetyl- or longer hexanoyl- SU, the overall polyketide chain length did not exceed $C_{24}$. To achieve the $C_{30}$ polyketide backbone comprising fredericamycin an additional unknown auxiliary enzyme is thought to be necessary. In vitro reconstruction of the fredericamycin minimal PKS and additional MCAT lead to the formation of polyketides with unknown structure, and phosphopantetheine ejection assays only identified tri-, tetra-, and pentaketides, and cyclised variations thereof, and not full-length, $C_{22}$, or $C_{24}$ polyketide products, consistent with the need for additional stabilisation of the KS/CLF complex.

Oddly, expression of the A-74528 minimal PKS in $S. coelicolor$ CH999, also known to produce fredericamycin, yielded only tri- and tetraketides. Co-expression of sanl, the cluster associated C9-C14 cyclase/aromatase (ARO/CYC), was necessary for formation of TW95a and b, known cyclised $C_{24}$ shunt metabolites, however once more $C_{30}$ polyketides could not be detected. The lack of full-length polyketides from the A-74528 mPKS has not been experimentally elucidated, however it is clear that cluster associated ARO/CYC plays some role in stabilising the KS/CLF complex.

The observed KS/CLF:long chain acyl-ACP stabilising effect of an ARO/CYC has previously been reported during biosynthesis of spore pigments, and is thought to be a necessary PKS component to reach longer chain lengths in these systems. It is possible, however unlikely, that PKS components from whiE spore pigment BGC present within $S. coelicolor$ CH999 provided a stabilising effect during fredericamycin biosynthesis but not during that of A-74528, explaining the markedly different metabolic profiles of the congruent PKSs. Reconstitution of both mPKS and ARO/CYC components in $S. coelicolor$ ΔwhiE could substantiate this hypothesis. The complexities achieving a $C_{30}$ chain length are made further apparent by the restoration of A-
biosynthesis upon expression of sanJ, a monooxygenase, to the biosynthetic complement producing the C24 TW95a and b shunt metabolite. These findings suggest incorporation of a hexadienyl-starter unit, regiospecific cyclisation of 6 carbocycles, concomitant formation of 6 stereocentres and chain extension from a C24 to a C30 is facilitated by co-expression of sanJ, which may be necessary to form a multi-protein complex. The disparities between these two closely related biosynthetic pathways are indicative of our understanding of the complex protein-protein interactions necessary to reach maturation of longer polyketides using dissociable type II polyketide biosynthetic machinery.

Cyclases can be considered minimal polyketide synthase components during biosynthesis shorter chain polyketides also, e.g. resistomycin. Using a reductionist approach to characterise the rem BGC, polyketide biosynthesis was only achieved after co-expression of the KS/CLF, ACP, PPTase, first ring cyclase and one of two additional cyclases, remF or remL. The inclusion of two additional cyclases to the mPKS suggests these dissociated PKS components form a stabilised complex necessary for polyketide biosynthesis which is non-functional in the absence of one part.

### 6.3.3.3 High through-put (HTP) attempts to alter chain length

Attempts to alter polyketide chain length by mutagenesis of CLFs and high throughput screen were historically scuppered because no simple assay was available for polyketide identification, however high throughput mass spectrometry approaches can now be used. Structural and biochemical based inferences derived from CLF phylogenies have been used to generate a small number of frn CLF mutants which fill a large design space. Generation mutants within this design space elucidated residues important in CLF function however failed to alter chain length specificity. Publications on HTP CLF engineering are few because KS/CLF components are readily interchangeable between type II polyketide biosynthetic pathways meaning there is no real need to alter the polyketide chain produced by a specific PKS. Rather, another KS/CLF homologue known to produce a longer or short polyketide chain can be readily substituted into a biosynthetic pathway to achieve the desired chain length. This luxury is not afforded when using E. coli as a heterologous expression host as only one visibly soluble KS/CLF has been characterised (Chapter 5) and therefore engineering this unique PKS is a necessary strategy to access a more vibrant and useful chemical space using this heterologous host. Here we take important first steps towards altering the volume of the AntDE dimeric cavity and demonstrate an interesting strategy for purification of AntDE in its heterodimeric form.

### 6.4 Aims and objective

This work seeks to measure the volume and length of the cavity at the AntDE heterodimer interface into which the nascent polyketide chain grows and evaluate the ability to open this cavity through mutations of gatekeeper residues. Furthermore, to build a more informed picture of the AntDE complex we attempt to solve the crystal structure of this heterodimeric complex.
6.5 Materials and Methods

6.5.1 Building homology models

The AntDE dimeric homology model was built using SWISS-PROT\textsuperscript{48}. AntD and E (Accession CAE16563.1 and CAE16562.1) amino acid sequences were modelled upon act KS and CLF templates (PDB: ITQY, chains A and B respectively). Protein models were further manipulated using PyMOL v1.3. Ray-tracing was used to construct all figures in PyMOL. His\textsuperscript{6}AntED-Fus disorder was predicted by PONDR\textsuperscript{49}, a purely sequence based predictor of protein structures, PrDOS\textsuperscript{50} and Metadisorder\textsuperscript{51}. Graphical representation of Metadisorder output is not formattable for thesis submission however is in agreement with PONDR and PrDOS.

6.5.2 Plasmid construction

The his\textsuperscript{6}antED-fus expression vector pET419091\_30AA was built from two parts. The vector backbone was generated by polymerase chain reaction (PCR) linearization of pETDuetPlu419091 using primers which anneal to the very 3' end of plu4190 and the 5' end of plu4191 respectively. The poly-glycine serine linker insert was also generated by PCR using two long primers (69 and 71 bp) sharing a 20 bp complementary overlap, Ta ~ 60°C. Each long primer also shared complementarity with the respective end of linearized pETDuetPlu419091. Both PCR amplicons were assembled using In-Fusion HD cloning (Takara) as per manufacturer’s instructions. To swap the position of the His\textsuperscript{6} tag from the AntE N-terminus to the AntD C-terminus the antE-30aa-antD fusion was amplified from pET419091\_30AA using primers with 5' NcoI and XhoI restriction endonuclease sites. The resulting amplicon was cloned into pET28a using NcoI and XhoI resulting in formation of pET28a_AntDE30_Chis. Additional vectors comprising 20 aa and 40 aa failed during cloning. For construction of the ΔantDE complementation vector pACYCAntΔAntDE, pACYCAnthraquinone was linearized using plu419091_deletion_for and plu419091_deletion_rev. The corresponding PCR amplicon lacked plu4190 and plu4191. Each primer contained a 16 bp complementary 5' sequences facilitating circularised of the PCR amplicon by In-Fusion HD cloning (Takara), as above. All vectors were fully sequenced prior to use. All primers and vectors are detailed in Supplementary Table 6.1 and Supplementary Table 6.2).

6.5.3 Bacterial strains, culture techniques

\textit{Escherichia coli} BL21(DE3), was used for all routine protein expression studies. \textit{E. coli} DH5\textalpha was used for all routine cloning and maintenance and preparation of plasmids. For protein expression all bacterial cultures were grown in lysogeny broth supplemented with appropriate antibiotics as selection markers: chloramphenicol 25 µg ml\textsuperscript{-1}, carbenicillin, 50 µg ml\textsuperscript{-1}. AntDE heterodimeric complexes his\textsuperscript{6}-antE and antD were expressed from pETDuetPlu419091 (Chapter 5, Supplementary Table 5.2).

Recombinant protein was expressed as follows: LB broth was inoculated BL21(DE3), containing appropriate expression vector, and grown at 37°C, 180 rpm to an optical density (OD\textsubscript{600}) of 0.5, prior to induction of gene expression using Isopropyl β-D-1-thiogalactopyranoside (IPTG) (50
or 200 µM). Cultures were allowed to cool before induction and were incubated for a further 16 h at 16°C, 180 rpm, before extraction and purification of recombinant protein.

### 6.5.4 Protein purification and visualisation

Purification methods of recombinant AntDE proteins, or fusions there-of, were congruent. Bacterial cells were pelleted by centrifugation, 4,000 x g, for 20 min, 4°C, before removing supernatant. Cell pellets were suspended in buffer A (50 mM Tris-HCl, 300 mM NaCl, pH 7.4, 4°C), pH adjustment was undertaken at 4°C, before sonication on ice. Sonicated total cell lysate was centrifuged at 12,000 x g to remove cell debris and insoluble protein. Supernatant was collected and centrifuged once more as above. Supernatant was collected, designated as soluble cell lysate and stored on ice. Cell lysate was applied to an immobilised metal ion affinity chromatography (IMAC) column pre-equilibrated with buffer A, before washing with 5 column volumes (CV) of buffer A. Protein was eluted from the IMAC column in 1 CV of buffer A containing sequentially increasing concentrations of imidazole (20, 50, 200, 400 and 500 µM), the IMAC column was finally washed with 5 CV and stored in 20% ethanol. For anion exchange chromatography, proteins were desalted by buffer exchange into buffer B (50 mM Tris-HCl, pH7.4) by repeated concentration and dilution using centrifugal filters. Protein eluted over a 90 min linear gradient from 0 to 100% buffer C (400 mM NaCl, 50 mM Tris-HCl, pH 7.4) at 4°C using a RESOURCE Q 6 ml anion exchange column (GE Healthcare) and ÄKTAexpress protein purification system (GE healthcare), flow rate 3 mL/min. The RESOURCE Q column was pre-equilibrated with 5 column volumes of buffer B before loading protein and washed with 5 additional CV of buffer B after loading to remove unbound protein. The column was cleaned with 5 CV of buffer D (1 M NaCl, 50 mM Tris-HCl, pH 7.4) to ensure all protein eluted before storage in 20% ethanol. AntDE was further purified by size exclusion chromatography (SEC). In brief, AntDE containing fractions were loaded onto a Superdex 200 increase 10/300 GL (GE Healthcare) SEC column pre-equilibrated in buffer A and eluted in 1.5 CV, flow rate 4 mL/min. All protein purification was carried out at 4°C. Purified protein was visualised by SDS-PAGE and fractions containing AntDE were carried forward for additional purification steps.

### 6.5.5 Metabolite production extraction and analysis

All methods concerning AQ256 biosynthesis, metabolite extraction protocols and compound identification are detailed in Chapter 5, 5.9.4.

### 6.5.6 Crystallography trails

Purified AntDE proteins were concentrated to 2, 5 and 7 mg/ml prior to crystallography attempts using Vivaspin 20 30 kDa molecular weight cut off centrifugal filters (Sartorius) and maintained on ice. Protein concentration was calculated using Bradford reagent. A sitting drop method was used to stimulate protein crystal growth. JCSG, Morpheus, Pact and CS1 crystallisation screens were used to stimulate protein crystal growth.
6.6 Results

6.6.1 Prediction of AntE gatekeeper residues

Simple substitution of gatekeeper residues to alter polyketide chain length, as described for the act and tcm CLF, is not achievable using AntE. Multiple sequence alignments show the atypical residues lining the tunnel to be poorly conserved and the gatekeeper residues not to map onto the AntE amino acid sequence (Chapter 5, Supplementary Table 5.1). A more informed picture constructed from homology models confirms this (Figure 6.6). The first gatekeeper on alpha helix 3 is a tryptophan, W109, in line with the large phenylalanine residue F109 in the act CLF, therefore superficially it is possible that CLF W109A and V116A mutations would expand the tunnel sufficiently to generate decaketides. However, residues 193 and 194, also important in determining the cavity size, are missing on the multisequence alignment. Furthermore, residues corresponding to G193 and W194, determined using protein secondary structure as valine and lysine respectively, form a porous structure in the AntE homology model negating efficient calculation of the tunnel volume (Figure 6.6). Nevertheless, calculation of the tunnel length by measuring the distance from the catalytic KS C169 to the tunnel capping gatekeeper, AA116, as done for the act CLF42, showed the AntDE tunnel length to be 20 Å, 3 Å longer than the tunnel in the act KC/CLF (Figure 6.6), and 2 Å longer than previously reported values for the act KS/CLF42. Should the model be accurate, the additional 3 Å may accommodate the carboxylic acid group of the proposed malonyl- starter unit, however historically only acetyl- primed shunt metabolites have been identified from the AntDE complex52, (Chapter 5). Whilst this finding is of merit, the uncertainty in the AntDE homology model and noncanonical position of gatekeeper residues within AntE confound simple mutagenesis experiments to alter the PKS chain length. Obtaining the crystal structure of AntDE would provide a more holistic view of the vestigial tunnel and facilitate rational engineering to access different polyketide chain lengths. Towards this aim we optimised the extraction processes of dimeric, or pseudodimeric, AntDE from E. coli BL21(DE3) for crystallography trials.
Figure 6.6: Models of act and ant KS/CLF

a, Homology model of AntDE based upon act KS/CLF (PDB: 1TQY). AntD and E are green and light blue respectively. Gatekeeper residues are denoted as dark blue surfaces. Catalytic cysteine 169 is highlighted as magenta sticks. The tunnel length is shown as a white line (20 Å). b, model of the act KS/CLF (ActI ORFI and ORFII) denoted in yellow and red respectively. Gatekeeper residues and tunnel length (17 Å) are as indicated in a. The catalytic cysteine 169 is shown as green sticks. All models were constructed and viewed in PyMOL V1.3.

6.6.2 Purification and stability of dimeric of AntDE

AntD and His⁶AntE co-purify from E. coli BL21(DE3) soluble cell lysate as heterodimeric recombinant proteins by IMAC (Figure 6.7a), however, the stability of this complex ex vivo is in question and impeded crystallisation attempts. Purified AntDE routinely precipitated when maintained in sonication buffer A at 4°C for more than 24 h and further purification of AntDE complexes is challenging. Co-purification of His⁶AntE-AntD can be achieved via anion exchange chromatography (Figure 6.7b), however, complexes dissociate during size exclusion chromatography and are predominantly purified as monomeric proteins (Figure 6.7c), and not as heterodimeric complexes. Pooling and concentrating AntD and His⁶AntE monomer containing fractions to > 10 mg/ml does not cause precipitation; however concentrated protein precipitates quickly under Morpheus HT-96 crystallography assay conditions, at 4°C, preventing crystallisation efforts. Co-purification of AntDE complexes from soluble cell lysate indicates AntDE to complex in vivo, however, the purification methodology here appears to destabilise the heterodimeric interactions, and the KS and CLF appear to be loosely associated by comparison to other KS/CLF components e.g. the hedamycin¹⁹.
Figure 6.7: Purification strategy for AntDE heterodimer complex

a, IMAC co-purification of His$_6$AntED from the soluble cell lysate of *E. coli* BL21(DE3) expressing his$_6$-antE and antD from pETPlu419091. Unbound protein is demarked as flow through or FT. His$_6$-AntE and AntD co-elute after addition of 50 mM imidazole to sonication buffer A (elution fraction (EF) 50 mM), no further protein elutes upon addition of 500 mM imidazole (EF 500 mM). b and c show further purification steps of His$_6$-AntED. b, Anion exchange purification of elution fraction 50 mM (a) after buffer exchange. Dimeric His$_6$AntED co-elutes from the RESOURCE Q column over an increasing salt concentration gradient (fraction 3). At low salt concentrations, monomeric His$_6$-AntE elutes (lanes 1 and 2) as confirmed by trypsin digestion and LCMS/MS. c, separation of fraction 3 (b) by size exclusion chromatography (SEC). The similar molecular weight of His$_6$-AntE (42.34 kDa) and AntD (46.16 kDa) reduces the resolution achievable by separation on a Superdex 200 Increase 10/300 GL SEC column, however SDS-PAGE analysis of sequentially eluting fractions show the higher molecular weight AntD to elute from the column prior to His$_6$-AntE. Black arrows show His$_6$-AntE. Dashed black arrows show AntD. Protein size is calculated against a reference molecular weight ladder (L, kDa).

6.6.3 Fusing ketosynthase and chain length factors, AntD and AntE
To circumvent AntDE heterodimer dissociation during purification both proteins were fused together using a flexible serine and poly-glycine linker, forming a 90.61 kDa His$_6$AntED pseudo heterodimer. The AntDE homology model (Figure 6.6), was used to design flexible amino acid linkers of a sensible length.

The actinorhodin KS and CLF crystal structure (PDB: 1TQY) shows the respective C and N termini of each monomer to be located close together, and distant from the active sites of the proteins, and the proposed ACP binding regions$^{53}$ (Figure 6.8a and b) therefore modification between these termini i.e. introduction of a linker, does not appear to have any obvious sterically inhibitory properties. The same protein topology is assumed for AntDE structural homologues and inferred through similarities in secondary structure to act KS and CLF (Chapter 5, Supplementary Figure 5.1). Irrespective of structural similarity, amino acid sequence dissimilarity between AntE and act CLF at the C-terminus prevented reliable homology modelling of this region and resulted in a truncated AntE homology model, lacking 39 C-terminal AA, causing ambiguity in the location of the AntE C-terminus. Therefore, to assess the length of linker necessary to create the AntDE fusion protein the distance between AntD C- and AntE N-homology model terminal regions were measured. The distance between these termini was calculated to be 32.5 Å, and specifically spanned a gap between the AntD C-terminal Ala428 carboxylic acid and AntE Arg4 amino group, as the preceding Met1, Arg2 and Lys3 were omitted from the homology model (Figure 6.8c). Ten amino acids would be the minimum requirement necessary to bridge this distance using a low amino acid contour length value of 3.4 Å$^{54}$, however, this minimal linker does not account for the three-dimensional structure of the protein and therefore is likely to sterically inhibit correct heterodimer formation (Figure 6.8d). Accordingly, three linkers were designed comprising 20, 30 and 40 AA to compensate for the uncertainty in the homology model.
Protein or peptide linkers are typically categorised into three groups: rigid linkers, cleavable linkers and flexible linkers\textsuperscript{55}. Flexible linkers are characteristically disordered and do not perturb protein folding or interactions between fused protein domains, satisfying the criteria necessary to isolate correctly folded pseudo-heterodimeric AntDE complexes. The flexible linkers designed here comprise poly-glycine residues punctuated with serine residues (Gly\textsubscript{2}Ser\textsubscript{n}), adapted from the commonly used (Gly\textsubscript{4}Ser\textsubscript{n}) linker\textsuperscript{56}. Whilst Poly-glycine linkers may be the most flexible, introduction of polar serine or threonine residues reduces linker-protein interactions which may influence activity\textsuperscript{57}. Serine is overrepresented in the AntDE linkers, by comparison to described synthetic linkers\textsuperscript{58-61}, in an effort to further reduce linker-protein interactions.

Remarkably, His\textsuperscript{6}-AntED fused by a 30 AA linker (His\textsuperscript{6}-AntED-fus), was a largely soluble recombinant protein in \textit{E. coli} BL21(DE3) (Figure 6.9). Interestingly, the fusion pseudo-dimer co-purified with a His\textsuperscript{6}AntE fragment (Figure 6.10a). This smaller truncated protein was identified by peptide digestion and LCMS. Additionally, the early elution characteristics of the truncated His\textsuperscript{6}AntE fusion protein fragment from the RESOURCE Q anion exchange column (Figure 6.10b) match those of ‘monomeic’ His\textsuperscript{6}AntE (Figure 6.7b lanes 1 and 2). Pseudo-heterodimer truncation was initially thought to result from translational stalling and resultant premature dissociation of the ribosome during translation of the glycine and serine-rich linker\textsuperscript{62} and therefore a new vector was constructed with a C-terminal His\textsuperscript{6} tag, to avoid co-purification. Purification of C-terminal hexa-histidine tagged recombinant protein also resulted in co-purification of a truncated protein with an observed molecular weight consistent for the AntDHis\textsuperscript{6} counterpart (Figure 6.10c), demonstrating the fragmentation to arise from proteolysis of the pseudo-heterodimer opposed to translational inefficiencies. The exact cleavage mechanism remains unknown however and necessitates additional purification steps to achieve protein purity necessary for crystallography.
Figure 6.8: Designing flexible linkers using the actinorhodin KS:CLF crystal structure and AntDE homology model.

a, Actinorhodin ketosynthase (red) and chain length factor (yellow) heterodimer crystal structure (PDB: 1TQY Chains A and B respectively). N-terminal amino acid (AA) are indicated by blue spheres and C-terminal carbons by white spheres. All termini are orientated towards the top if the heterodimer and away from the active site (green residue: cysteine 169, denoted by a yellow sulphur atom) and two residues proposed to be important in ACP interaction (green residues: Arg211 and Thr213)\textsuperscript{53}. b, 90° rotation of a showing a top down view of the actinorhodin KS:CLF heterodimer. c, AntDE heterodimer homology model. Individual monomers were constructed using 1TQY as a template and docked together in PyMOL. The distance between N-Ante and AntD-C is detailed by a yellow dashed line. d, 90° rotation of c showing
AntE C-terminalΔ39 and AntD as surfaces. The AntE N-terminal region is illustrated in red and the buried AntD C-terminal amino acid in white. All structures were visualised using PyMOL.

Figure 6.9: Evaluation of His6-AntED fusion protein in *E. coli* BL21(DE3)

SDS-PAGE analysis of total cell lysate from *E. coli* BL21(DE3) expressing his6-antED-fus from pET419091_30AA (linker) and *E. coli* BL21(DE3) pETDuet-1 (empty vector control). Induction with IPTG is stated in black as 0, 50, 200, 250 or 500 µM. His6-AntED is both a both soluble and insoluble recombinant protein (white circle). A putative soluble truncated protein (39.56 kDa, accurately calculated by migration distance) is denoted in blue.

Figure 6.10: Purification of His6AntED from *E. coli* soluble cell lysate

a. Immobilised metal affinity chromatography (IMAC) purification of His6AntED-fus from *E. coli* BL21(DE3) soluble cell lysate. The His6AntED-fus recombinant protein (90.61 kDa) predominantly elutes from the IMAC column after addition of 50 mM imidazole to sonication buffer A (Elution fractions - EF: 50 mM),
detailed by a black arrow. His\textsuperscript{6}AntED-fus is also visible in column flow-through (FT), wash, and fractions eluted with sonication buffer comprising 20 mM, 200 mM and 500 mM imidazole respectively. Co-purification of truncated His\textsuperscript{6}AntE fragments is predominantly visible in EF 50 mM and 200 mM, (red arrow). Anion exchange chromatography purification of EF 50 mM (a). Elution fractions comprise increasing concentration of NaCl from left to right (black triangle). His\textsuperscript{6}AntED-fus (black arrow) elutes from fraction 4 to 8 with increasing NaCl concentration. His\textsuperscript{6}-AntE consistently elutes (red arrow). b, Purification of AntED-His\textsuperscript{6} (C-terminally hexa-histidine tagged) AntED fusion protein by anion exchange chromatography. Protein was purified from \textit{E. coli} BL21(DE3) soluble cell lysate by IMAC as in a, before further purification. A large protein is purified with a size consistent to AntED-His\textsuperscript{6} (Black arrow), along with a smaller protein putatively described as a AntD-His\textsuperscript{6} truncation (red arrow). All protein samples were separated and visualised by SDS-PAGE (10%).

6.6.4 Complementing Δ\textit{plu4190plu4191} AQ BGC with His\textsuperscript{6}-\textit{AntE} and \textit{AntD}, and His\textsuperscript{6}-AntED-fus

Whilst His\textsuperscript{6}AntED-fus, His\textsuperscript{6}AntE/D and AntED-His\textsuperscript{6} were soluble recombinant proteins in \textit{E. coli}, it was unknown if the constraints imposed by linking the two type II PKS components together would inhibit polyketide biosynthesis from this complex. Accordingly, an anthraquinone biosynthetic pathway deficient in AntD and AntE was complemented with the fusion protein His\textsuperscript{6}AntED-fus and the dissociable complex His\textsuperscript{6}AntE and AntD independently; corresponding genes were expressed from additional episomal elements. Biosynthesis of AQ256 (1) was abolished in hosts harbouring the Δ\textit{AntDE} anthraquinone biosynthetic complement, as expected, however, AQ256 (1) biosynthesis was successfully recovered when complemented with episomal elements expressing his\textsuperscript{6}antED-fus or his\textsuperscript{6}antE and antD (Figure 6.11). AQ256 was detectable in both complemented \textit{E. coli} BL21(DE3) strains by HR-ESI-LCMS/MS using negative ionisation mode with the retention times consistent with the anthraquinone BGC positive control and with previous experimental findings (Chapter 5, Supplementary Figure 5.18b) (AQ256 showed \textit{t}_{R} 785 sec, [M-H]\textsuperscript{-} 255.0298 for all samples in Figure 6.11, compared with \textit{t}_{R} 787 sec, [M-H]\textsuperscript{-} 255.0294 for AQ256 standard using the same analytical conditions). Furthermore, MS\textsuperscript{2} spectra of the peaks at \textit{t}_{R} 785 are consistent with those previously reported (data not shown) (Chapter 5, Supplementary Figure 5.13).
Figure 6.11: ∆plu4190plu4191 AQ BGC complementation with episomally encoded his⁶-
antED-fus and his⁶-
antE and antD.

Comparison of extracted ion chromatograms (EIC) from E. coli BL21(DE3) expressing the ∆AntDE anthraquinone biosynthetic pathway complemented with his⁶-
antED-fus (AntED fusion - red) and his⁶-
antE and antD (His⁶-AntE, AntD - orange). a and b, 3D and 2D EICs of all observable masses ± 1 ppm from AQ256 (1) theoretical mass, [M-H]- 255.02990. AQ256 (1) is detectable in the positive control, E. coli BL21(DE3) pACYCAntanthraquinone (AQ256 - yellow), at tR 785, in agreement with EICs from AntED fusion and His⁶-AntE, AntD. No masses corresponding to 1 are observable in E. coli BL21(DE3) (Host background control, HBC), E. coli BL21(DE3) expressing pACYCDuet1 and pETDuet1 (empty vector control – EVC) or E. coli BL21(DE3) expressing the ∆AntDE anthraquinone BGC (pACYCAnt∆AntDE).

Gene expression was induced by the addition of 200 µM IPTG at OD₆₀₀ 0.4, 16 h, 4°C for all cultures. Each trace is representative of three biological samples. All samples were normalised by final optical density.

The addition of an N-terminal histidine tag to the AntE chain length factor does not prevent biosynthesis of AQ256 as complementation of ∆AntDE BGC with his⁶-
antE and antD restores biosynthesis of 1. It is unclear whether the His⁶-AntED fusion protein is functional in vivo however, as the truncated dimeric parts visualised in Figure 6.10 (red arrows) may be the active KS/CLF constituents in the complemented biosynthetic pathway, opposed to the fused complex. The long 30 AA peptide linker between His⁶-AntED-fus is possibly overcompensating for uncertainty in the AntIDE homology model and the disordered peptide sequence is being cleaved in vivo. The introduction of sequentially shorter sequences between AntE and D could prevent nonspecific cleavage of the (Gly₂Ser)n linker whilst maintaining pseudo heterodimer solubility enabling assessment of functionality.

6.7 Discussion

Successfully modifying the size of the vestigial amphipathic tunnel formed between the KS and CLF dimeric PKS components is a powerful tool allowing access a wide variety of alternative polyketide chemotypes (Figure 6.2), which, arguably, comprise more vibrant chemical space than diversity introduced through combinatorial biosynthesis of initiation modules for alternative polyketide starter units. Modifying the tunnel into which the growing polyketide is extruded is not a trivial task for the anthraquinone cluster KS and CLF, AntDE, as the canonical gatekeeper residues defining the tunnel volume are ambiguous in AntE when compared to the model act CLF, and other Actinobacterial derived CLF sequences. The amino acid sequence dissimilarity of AntE has been previously reported⁵², yet we reason the secondary structure of this CLF to maintain the conserved αβαβα fold of act KS and CLF components⁵², irrespectively. Despite structural similarities, using Actinobacterial conventions to predict the position of gatekeeper residues within AntE resulted in identification of a longer 20 Å and porous tunnel (Figure 6.6) inconsistent with expectations for a C₁₆ polyketide. It is likely that a similar philosophy is applicable to modify chain length in AntDE, however the internal localisation of the tunnel may be different, and residues defining its volume will vary, as a consequence. A more informed view of the structure of the dimeric complex is therefore necessary to elucidate the volume and
localisation of the tunnel at the dimer interface. To highlight regions of AntE which are under stringent selection pressure, and therefore plausibly where the amphipathic tunnel lies, we compared amino acid substitutions between AntE homologues from different *Photorhabdus* spp.. Here, AntE from *P. luminescens* TT01 and the most dissimilar AntE homologue from *P. temperata* subsp. *khani* NC19, differ by 48 amino acids, 41 of which are solvent exposed. Of the remaining 7 amino acids only 1 residue, Q108K, was proximal to the gatekeeper residues and 1 additional substitution was located at the dimer interface. The entire dimer interface is broadly less tolerant to alteration and therefore cannot be used to discriminate the tunnel region. Interestingly, the Q108K mutation was present exclusively in *P. temperata* spp., whereas all *P. luminescens* spp. retained Q108. Unsupported claims in the literature state *P. temperata* spp. to produce a red pigment63, opposed to the orange or yellow pigments of *P. luminescens* spp., which could be indicative of a bathochromic shift in absorbance characteristic of some anthracyclines64. Other than the lack of supporting information, this seems unlikely for a number of reasons, however. Firstly, simple modifications of AQ256 hydroxylation pattern, or dimerisation can lead to a number of reported red anthraquinones65. Secondly, AQ256 is a pH indicator, altering from yellow to pink/red when acidified, or in a basic solution respectively (data not shown), without knowing the pH of the media no reliable conclusions can be drawn. And thirdly, lysine is not a small amino acid, therefore would likely also occlude the amphipathic tunnel should this residue be involved in tunnel lining. By obtaining a crystal structure of the AntDE complex we hope to circumvent the pitfall of our current bioinformatic analysis, elucidate the governance of the chain length, disambiguate the positions of the important gatekeeper residues and make a library of rational AntDE mutations to access a series of different polyketide chain lengths. The stability of the AntDE complex hampered abilities to do so.

AntD and AntE readily copurify, as expected for a heterodimeric complex (Figure 6.7a), however His6-AntE is also shown to elute from the anion exchange column without its heterodimeric counterpart in buffer A, with low salt. The absence of significant monomeric AntD counterpart indicates the His6-AntE to strike a more even equilibrium between monomer, or possible homodimer, and heterodimer than expected in vivo, or for the heterodimer to simply dissociate during IMAC purification. Supporting the latter is the presence a band similar in size to the theoretical size of AntD in the unbound flow through (Figure 6.7a). Purification of alternative KS/CLF pairs from *Streptomyces* spp. have used sodium and potassium phosphate buffers between pH7 and 834,42,66,67 (Table 6.2) which provide a better buffering capacity at physiological pH than Tris, and are likely used because these buffers are less temperature sensitive. Actinobacterial KS/CLF buffers also contain EDTA, reducing agents and glycerol. Addition of glycerol as a stabilising agent to buffer A, up to 15% (v/v), had no influence on AntDE heterodimer stabilisation, however. Furthermore, glycerol was not necessary to obtain a crystal structure of the act KS/CLF highlighting its more probable role as a cryoprotectant in protein storage buffers. The lack of reducing agent and EDTA may lead to protein oxidation during AntDE purification but there are no solvent exposed cysteine residues, excluding AntD
C169, so this was deemed nonconsequential. Identification of a better protein buffer may aid heterodimer purification attempts however, and optimization of buffer components using fluorescence-based thermal-shift assays would constitute a next step towards AntDE crystallisation. Protein stability is a prerequisite to crystallisation and identification of buffer components which improve stability of proteins in homogenously folded form have historically been key to facilitate structural elucidation.

Table 6.2: Buffers used during purification of KS/CLF pairs from *Streptomyces* spp.

<table>
<thead>
<tr>
<th>KS/CLF pair</th>
<th>Storage / lysis buffers</th>
</tr>
</thead>
<tbody>
<tr>
<td>act</td>
<td>100 mM NaH₂PO₄, pH 7.0, 150 mM NaCl, 2 mM EDTA, 2 mM DTT and 10 mM HEPES, pH 7.0, 150 mM NaCl and 1 mM DTT</td>
</tr>
<tr>
<td>fred</td>
<td>50 mM Tris-HCl, 1 mM TCEP, 10% glycerol, pH 8.0 and 100 mM NaH₂PO₄, pH 7.4, 2 mM TCEP, 2 mM EDTA, and 15% glycerol</td>
</tr>
<tr>
<td>hed</td>
<td>100 mM NaH₂PO₄, 2 mM DTT, 2 mM EDTA, 20% Glycerol (pH 7.4)</td>
</tr>
<tr>
<td>gil</td>
<td>50 mM KH₂PO₄, 300 mM NaCl, 10 mM imidazole pH 7.6</td>
</tr>
</tbody>
</table>


Alternatively, AntE homodimer formation could explain the elution characteristics of the recombinant AntDE. There are no published KS/CLF interaction constants, and little is known about the overall topology of the minimal PKS components however yeast two-hybrid systems have elucidated the ability of single act KS/CLF subunits to interact, albeit the act KS not the CLF. Yeast two-hybrid systems have limitations however: the act CLF was not reported to interact with any PKS components when used as prey, thought to be a consequence of incorrect folding. In all other reported examples of oligomeric KS/CLF complexes a 1:1 stoichiometry is maintained, therefore AntE homodimer formation is a plausible interaction but not well supported.

In light of this, the observation of the AntDE fusion as soluble recombinant protein partly circumvents the issue of heterodimer dissociation. Remarkably, the scant anecdotal evidence within the literature suggests expressing KS and CLF components as fusion proteins to be ineffective. It is plausible that the candidate was of Actinobacterial origin, however, which are historically insoluble. The cause of insolubility of Actinobacterial KS/CLFs therefore is not likely to be a result of inefficiencies in heterodimer formation but rather, disharmonious translational and folding dynamics of monomeric components. Unfortunately, the exact KS/CLF candidates are not detailed and nor is the composition of the fusion linker so no direct comparison can be made. The addition of a flexible linker may only be tolerated by KS/CLF pairs with inherent solubility in *E. coli*, as opposed to being a strategy to improve solubility.
E. coli BL21(DE3) is often the protein production strain of choice\textsuperscript{75}, firstly because of virtuous growth rates and ease of manipulation but more specifically due to its Lon protease and OmpT deficiency\textsuperscript{76}. Deletion of proteases for E. coli invariably improves the yield of recombinant protein productions however in the case of recombinant His\textsuperscript{6}AntDE-fus additional proteases appear to be cleaving the pseudo-heterodimer at the site of the flexible linker. The flexible properties of our (Gly\textsubscript{2}Ser\textsubscript{n} linker are afforded by its disorder i.e. its poor propensity to form a secondary structure, however this intrinsic disorder may target the fusion protein for proteolysis. For structurally characterised proteins it is known that sites of proteolysis overlap with protein regions with no recordable electron density\textsuperscript{77}. Proteolysis of recombinant proteins has also long been used as technique to assess their intrinsically unstructured properties (IUP)\textsuperscript{78}. Such an approach, in part, facilitated structural assignment of neuroligin 3, a human neuronal adhesion protein as partially unstructured and demonstrated the susceptibility of IUPs to degradation by E. coli proteases\textsuperscript{79}. Prediction of disordered regions within His\textsuperscript{6}AntED-Fus (Supplementary Figure 6.1 and Supplementary Figure 6.2) confirmed the poly-glycine and serine linker and the first 8 amino acids of AntD to form no strong secondary structure and this expanse of 38 amino acids surpasses the degree of unfolding thought necessary for proteolysis by most known proteases\textsuperscript{80}. Altering the length of the flexible linker will likely reduce the degree of proteolysis, however reducing the linker length significantly may sterically inhibit pseudo-dimer formation. Systematic reduction of linker length coupled to its effect protein solubility is the best solution to find the optimum linker length bespoke to the AntDE complex. Generation of fusion proteins using glycine rich linkers is to improve folding and protein stability is not a new idea, however most linker sequences are more modest in size and typically limited to a length of 15 amino acids\textsuperscript{55,81}, with very few examples of linkers more than 30 AA in length\textsuperscript{82}. Cleaved protein fragments can be removed during purification of His\textsuperscript{6}AntED-Fus, by size exclusion chromatography, however this is not an elegant solution, and copurification of proteases, observed for neuroligin 3\textsuperscript{79}, can reduce His\textsuperscript{6}AntED-Fus homogeneity impeding crystal growth. Motif specific proteolysis could not be identified for the His\textsuperscript{6}-AntED-Fus recombinant protein, and with the degradation of intrinsically disordered recombinant proteins a widely conserved characteristic of both prokaryotes and eukaryotes\textsuperscript{77}, a more prudent choice of expression host is unlikely to prevent pseudo-dimer cleavage.

The AntDE complex was remarkably soluble as a fusion protein, however the effects of KS and CLF fusion on polyketide biosynthesis cannot be quantified here because of in vivo His\textsuperscript{6}AntED-Fus cleavage. Structural elucidation of a functional protein complex will provide a better model to reengineer protein function than a soluble but non-functional counterpart and elucidation of differences in enzyme kinetics between fused and dissociated KS/CLF pairs needs to be quantified in vitro. Efforts were made to this end, however purification of holo-AntF ACP, functionalised with a 4’phosphopantetheine arm, necessary for polyketide biosynthesis could not be achieved, and while the AntF ACP was remarkably soluble, AntB, the cluster associated PPTase was poorly soluble in E. coli hampering phosphopantetheinylation in vitro. Furthermore, E. coli ACPS and EntD PPTases and the promiscuous SFP PPTase from Bacillus subtilis were
not able to functionalise AntF. Preliminary data indicated this to be a result of the noncanonical DST phosphopantetheinylation site, however requires further exploration.

The exact nature, size and shape of the vestigial tunnel into which the growing polyketide chain is extruded remains elusive for AntDE. Bioinformatically we predict the tunnel size to be akin to decaketide producing KS/CLFs however uncertainty in the homology model, introduced by poor amino acid sequence similarities between AntC and act CLF, brings this measurement into question. Here we take steps towards purification of an AntDE pseudo-heterodimer to homogeneity, laying the ground work for further crystallisation trials. Furthermore, we elucidate the robust solubility of a type II PKSs despite the addition of numerous peptide tags at the N- and C-terminus of each monomer.

6.8 Perspectives

Atomic dissection of the AntDE gatekeeper residues facilitated by a crystal structure is necessary to afford biosynthesis of polyketide chains with alternate lengths. Building a library of AntDE mutants specifically tasked with the biosynthesis of unreduced polyketides with a certain chosen length is an invaluable tool when producing therapeutically valuable aromatic polyketides in E. coli however additional biosynthetic components may be necessary to facilitate high and specific yields of polyketide with certain chain lengths. The mutation of F109A and F116A in act CLF resulted in >95% yield of decaketides in vitro. The same complex, with the addition of downstream primary tailoring enzymes from an octoketide producing cluster, produced just 66% deca- and 34% octoketides in vivo, indicating the tailoring enzymes to play a role in chain length, presumably preferentially modifying octaketide polyketide chains prior to the additional rounds of polyketide elongation. Co-expression of chain length specific primary tailoring enzymes may be one strategy to improve chain length specificity.
220

6.9 References


6.10 Supplementary Figures

Supplementary Figure 6.1: PrDOS prediction of His⁶AntED-Fus disorder

Residue number corresponds to each amino acid (AA) within the 849 AA protein. The red line corresponds to the order or disorder threshold with a false positive rate of 5%. The N-terminal His⁶ tag sequence (MGSSHHHHHHSQDPN) has a high disorder probability, as does the C-terminal 3 amino acids. The flexible poly-glycine serine linker (387 – 417) and the first 9 AA of AntD show are predicted to be disordered.

Supplementary Figure 6.2: PONDR predictions of His⁶AntED-Fus disorder

PONDR disorder predictors VSL2, VL3 and VL-XT all define the poly-glycine serine linker region and first 8 residues of AntD to be disordered, in agreement with the PrDOS prediction. One more the N-terminal His⁶ tag is also predicted to be disordered.
6.11 Supplementary Tables

**Supplementary Table 6.1. PCR Primers**

<table>
<thead>
<tr>
<th>Primer name</th>
<th>Sequence 5′-3′</th>
</tr>
</thead>
<tbody>
<tr>
<td>plu4190 NcoI for</td>
<td>ATGTCAACCATGGGCGGCTAAAGAGATTGTTCGT</td>
</tr>
<tr>
<td>Plu4191 XhoI rev</td>
<td>ATGCATCTCGAGCGCATTTGTCTACTGCCC</td>
</tr>
<tr>
<td>spacer_insert_rev</td>
<td>AATAGCTGAAAAACTCAACGCACAC</td>
</tr>
<tr>
<td>Spacer_insert_for</td>
<td>ATGGCAGATCTAATGATAATAAAATAACAGAAATGAATCTC</td>
</tr>
<tr>
<td>30 aa spacer for</td>
<td>AGTTTTTCAGCTATTGGGCAGCGGTGCTGCTGCTGTTCTGGCAGCGGA</td>
</tr>
<tr>
<td></td>
<td>GCAAGCGGTGTCAGCGAGCGGCGGT</td>
</tr>
<tr>
<td>30 aa spacer rev</td>
<td>CATTAGATCTGCGCGCCGCGCGCGCGCCGAGCGCCAC</td>
</tr>
<tr>
<td></td>
<td>GCCCGAACCACCGCGCTGCACCACC</td>
</tr>
<tr>
<td>Plu419091 deletion for</td>
<td>GAGATATACCATGGCAGATTAATCACGAGATAAAAATAAAA</td>
</tr>
<tr>
<td>Plu419091 deletion rev</td>
<td>CG</td>
</tr>
<tr>
<td></td>
<td>GCCCATGGTATATCTCCTTATTAAGTTAAACAAATTATTTT</td>
</tr>
</tbody>
</table>
## Supplementary Table 6.2 Vector list

<table>
<thead>
<tr>
<th>Plasmid name</th>
<th>Insert Description</th>
<th>Backbone description</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>pACYCDuet-1</td>
<td>-</td>
<td>p15A ori, cat (Cm&lt;sup&gt;R&lt;/sup&gt;), T7 promoters</td>
<td>Novagen</td>
</tr>
<tr>
<td>pETDuet-1</td>
<td>-</td>
<td>pBR322 ori, bla (Ap), T7 promoters</td>
<td>Novagen</td>
</tr>
<tr>
<td>pETDuetPlu419091</td>
<td>plu4191, His6:plu4190</td>
<td>pBR322 ori, bla (Ap), T7 promoters</td>
<td>(Chapter 5)</td>
</tr>
<tr>
<td>pET419091_30AA</td>
<td>His6:plu4190:30AAlinker:plu4191</td>
<td>pBR322 ori, bla (Ap), T7 promoters</td>
<td>This study</td>
</tr>
<tr>
<td>pET28aAntDE30_Chi</td>
<td>plu4190:30AAlinker:plu4191:His6</td>
<td>pBB322 ori, KnR, T7 promoters</td>
<td>This study</td>
</tr>
<tr>
<td>pET28a</td>
<td>-</td>
<td>pBB322 ori, KnR, T7 promoters</td>
<td>EMD Millipore</td>
</tr>
<tr>
<td>pACYCAnthraquinone</td>
<td>plu4194, 93, 92, and plu4091, 90, 89, 88, 87, 86</td>
<td>p15A ori, cat (Cm&lt;sup&gt;R&lt;/sup&gt;), T7 promoters</td>
<td>(Chapter 5)</td>
</tr>
<tr>
<td>pACYCAntDeltaAntDE</td>
<td>plu4194, 93, 92, and p89, 88, 87, 86</td>
<td>p15A ori, cat (Cm&lt;sup&gt;R&lt;/sup&gt;), T7 promoters</td>
<td>This study</td>
</tr>
</tbody>
</table>
7 Type II polyketide synthase biosynthetic gene clusters from underexplored Phyla

7.1 Preface
This chapter constitutes unpublished work, and accordingly the results detailed here are reported in more depth than expected for a journal publication. Despite this, Chapter 7 is written in the format of an extended journal article i.e., it comprises its own introduction, methods and materials, results, and discussion sections, as well as an independent bibliography in keeping with the rest of the thesis. To avoid repetition where possible methods and materials from previous chapters are cited opposed to rewritten. Compounds numbered within are discrete to this chapter.

All work undertaken here was designed and carried out by M.C.. The expression of *Streptococci*-derived KS/CLF was carried out as part of a Masters of Research project by both Thomas McManus and M.C.. All corresponding experimental data was collected by both M.C and T.M and repeated by M.C.. Furthermore, preliminary studies into solubility of *Delftia acidovorans* KS/CLF recombinant proteins in *Escherichia coli* were undertaken by Lucas Krauss along with M.C.. All preliminary data was reproduced in repeat experiments and built upon by M.C before description here. Additionally thanks must be given to Francesco Del Carratore who helped process the 2552 complete genomes using a standalone version of antiSMASH 3.0.5.
7.2 Abstract
As the cost of DNA sequencing tumbles and bioinformatics tools for genome mining become more sophisticated, we are uncovering a glut of biosynthetic gene clusters (BGCs) unequivocally linked to specialised metabolites. Specialized metabolites have historically been a well-harvested pool of bioactive molecules with uses as human medicines. Accordingly, significant efforts have been made to explore and identify potential new and abundant sources of these compounds. Using an objective view on BGC evolution we posit that the type II polyketide synthases (PKSs), historically constrained to the actinobacterial Phylum, are distributed widely throughout all Prokaryotic taxa, and further posit that this pool of non-Actinobacterial biosynthetic machinery is ripe for exploration using *Escherichia coli* as a heterologous host organism. By harnessing a two-pronged bioinformatics approach we identify 12 unique and unreported type II polyketide synthase BGCs across 17 non-Actinobacterial Genera. The principle polyketide synthase (PKS) components appear to be discrete from canonical Actinobacterial ketosynthases (KS) and chain length factors (CLFs) forming a new type II PKS sub-division.

Actinobacterial KS and CLFs are reported to form 100% inclusion bodies as recombinant proteins in *E. coli* heterologous hosts however, contrary to this, we identify 4 KS/CLF candidates from the discrete non-Actinobacterial clade which form soluble recombinant heterodimers in *E. coli*. This work builds upon the recently described *E. coli* based plug-and-play scaffold for type II polyketide biosynthesis. By ‘plugging’ characterised alternative KS/CLF modules into our plug-and-play scaffold we hope to produce $C_{20}$ and $\geq C_{24}$ polyketides in *E. coli*, diverting biosynthetis towards diverse type II polyketide chemotypes. The results described here promise new routes to clinically useful aromatic polyketide e.g. tetracycline and doxorubicin etc. using *E. coli* as a heterologous host.
7.3 Introduction

Biosynthetic gene clusters (BGCs) are collections of gene sequences encoding biosynthetic enzymes, regulatory proteins, and resistance, and export mechanisms which functionally coordinate to produce a specialised metabolite. Historically referred to as secondary metabolites, as a result of their non-essential functions in monoculture, specialised metabolites are now being ascribed essential functions for proliferation of prokaryotes in natura. BGCs, unequivocally linked to specialised metabolites, are globally distributed throughout the Prokaryotes and the typical prokaryote comprises between 2.4 and 6.4 BGCs per Mb of DNA, dedicating 3.7% ± 3.1% of its genome to BGCs. In exceptional cases vast proportions of the host genome are dedicated to specialised metabolism. *Streptomyces bingchenggensis* harbours more than 65 BGCs encoded by more than 21% (2.56 Mb) of the genome (11.94 Mb). This enormous coding capacity greatly exceeds the entire genome of the free living lactic acid bacterium *Weissella koreensis* KACC 1550 (1.4 Mb). However, this record was trumped more recently by the Actinobacteria *Actinoalloteichus cyanogriseus* and *Smaragdicoccus niigatensis* which dedicate 39% and 36% of their more modest genomes (6.03 Mb and 5.32 Mb, respectively) to BGCs. Many *Streptomyces*, *Myxococcus* and *Burkholderia* spp. commonly devote the same proportion of their genomes to transcription and translation, as to specialised metabolism, highlighting the importance of these metabolites in proliferation.

Soil-borne Actinobacteria have a remarkable capacity to produce specialised metabolites, many of which have been harvested over the last 6 decades and marketed as clinically used therapeutics. In fact, this Phylum has yielded over 5000 compounds with antimicrobial activity alone, 90% of which are derived from the *Streptomyces* Genus. Mathematic models predict the total number of antibiotic compounds produced by cultivatable Actinobacteria to surpass 100,000 however the low-hanging-fruit is considered to have mostly been picked. To reduce rediscovery rates, the biosynthetic capacity of less obvious and unculturable taxa are being identified and exploited; an effort which is facilitated by reducing costs, and technical improvements, in genome sequencing technologies.

The last few decades have elucidated the extensive biosynthetic capacity of *Proteobacteria*, and more specifically isolates from the *Burkholderia* species. Global analysis of all available genomes (2014) confirmed this observation by identifying four-fold more non-ribosomal peptide synthetase (NRPS) BGCs in *Burkholderia* spp. than *Streptomyces* spp. from a sample dataset. The specific bioactivities of *Burkholderia* spp. non-ribosomal peptide specialised metabolites are broad: BTH-II0204-207 A–D, more commonly known as betulinan/terferol analogues, isolated from *B. pseudomallei* K96243 are potent phosphodiesterase-4 inhibitors and have therapeutic applications in a range of neurodegenerative diseases. The bicyclic depsipeptides burkholdacs and thailandepsins produced by *B. thailandensis* E264 inhibit class I human histone deacetylases (HDACs) in the nM range, with an anticancer profile similar to that of the FDA approved anticancer drug FK228, and the cyclic glycopeptide occidiofungin isolated from *B. contaminans* MS14 shows potent and broad fungicidal activity against common fungal
pathogens associated with invasive pulmonary aspergillosis and dermatophytosis\textsuperscript{16}.

\textit{Burkholderia} spp. are also abundant reservoirs of polyketide synthase (PKS) containing BGCs: \textit{Burkholderia ambifaria}, a potential biocontrol agent\textsuperscript{17} isolated from the cystic fibrosis lung, produces enacyloxin Ila (1) with potent intra-species activity against the pan-resistant Gram negative pathogens \textit{Burkholderia multivorans} and \textit{Burkholderia dolosa} as well as \textit{Acinetobacter baumannii}\textsuperscript{18}. New compounds with valuable bioactivities continue to be discovered in \textit{Burkholderia} spp.. The recently described polyketide gladiolin (2), isolated from \textit{B. gladioli}, was recently shown to arrest RNA polymerase activity in isoniazid resistant \textit{Mycobacterium tuberculosis} (TB)\textsuperscript{19}. Isoniazid is ordinarily used to treat both latent and active TB infections and, identifying alternative treatments is essential to tackle persistent and resistant infections\textsuperscript{20}.

Specialised metabolites from Firmicutes are also playing a role in human health, however to-date this has been in homeostatic maintenance of the microbiome, opposed to the prescribed or prophylactic use of Actinobacterial and Proteobacterial specialised metabolites. Lactocillin (3), produced by \textit{Lactobacillus gasseri}, is a modulator of healthy vaginal flora with antimicrobial activity against the opportunistic vaginal pathogens \textit{Gardnerella vaginalis}, \textit{Enterococcus faecalis}, \textit{Corynebacterium aurimucosum} and \textit{Staphylococcus aureus}, but not common vaginal commensals\textsuperscript{21}. In opposition to the benefits provided by commensal organisms, specialised metabolites from \textit{Streptococcus} spp. may provide the competitive advantage needed for over-colonisation of the oral cavity, thereby exacerbating tooth decay. Analysis BGCs from 169 \textit{S. mutans} genomes isolated from dental caries identified 1 NRPS, 7 PKSs, 136 NRPS/PKS hybrids, 211 ribosomally synthesised and post-translationally modified peptides (RiPPs) and 615 bacteriocin BGCs\textsuperscript{22}. One of which, producing Mutanobactin A (4), causes persistence of a yeast-like morphological state in \textit{Candida albicans}\textsuperscript{23}, although most remain uncharacterised.

Obligate and opportunistic human pathogens historically represented a large number of sequenced genomes within publicly available databases and as recently as 2015 43\% of all sequence bacterial genomes covered just 10 human pathogens\textsuperscript{5}. However, efforts are been made to systematically sequence genomes across the entire breadth of the Prokaryotic Kingdom\textsuperscript{24} and initiatives aligning with this philosophy are popularising. The Genomic Encyclopaedia of Bacteria and Archaea (GEBA)\textsuperscript{25,26}, devised in 2009, aims to sequence a dataset of 11,000 genomes comprising a type strains of every archaeal and bacterial species characterised, named, published, and validated by the International Code of Nomenclature of Bacteria (Bacteriological code). This initiative recently realised the sequencing of 1,003 reference genomes, expanding the overall phylogenetic diversity of type strains by 25\%, and resultanttly identified 9 new phenazine biosynthetic pathways\textsuperscript{5}. The importance of casting the sequencing net widely is increasingly apparent, and is being adopted routinely. Kerfeld et. al. (2012)\textsuperscript{27} modelled their sequencing strategy of the Cyanobacterial Phylum on the GEBA approach and identified 54 genomes from phenotypically and phylogenetically diverse Cyanobacteria to sequence, forming the CyanoGEBA dataset\textsuperscript{27}. Mining a collated dataset of 72 previously sequenced Cyanobacterial genomes and the CyanoGEBA dataset for BGCs
elucidated the relative ubiquity of NRPS and PKS BGCs within this Phylum. *Fischerella* spp., *Cylindrosporum* spp. and *Nostoc* spp., isolates from three different morphological Cyanobacterial subdivisions, all encoded between 10 and 20 predicted NRPS, PKS and NRPS/PKS hybrid BGCs, rivalling frequencies observed in Actinobacteria\textsuperscript{27}, and surpassing the authors expectations. Whilst underestimated, the biosynthetic capacity of Cyanobacteria has been long known\textsuperscript{28}. Organisms of the *Moorea* Genus are prolific producers of specialised metabolites with over 190 new compounds being isolated and characterised since 1997\textsuperscript{29,30}. This library of compounds encompasses more than 40% of the characterised Cyanobacterial specialised metabolome\textsuperscript{31} and includes bioactive molecules with potent anticancer (curacin A, 5)\textsuperscript{32} and antifungal (hectochlorin, 6)\textsuperscript{33} activities.

![Chemical structures](image)
Figure 7.1: Specialised metabolites from underexplored Organisms

Enacylotoxin IIa (1), gladiolin (2) and lactobacillin (3) show antimicrobial activities. Mutanobactin A (4) and hectochlorin (6) modify fungal lifestyle and curacin A (5) has potent anticancer activity. Bioactivities are detailed in the supporting text.

The ability of macroorganisms, as well as microorganisms, to produce specialised metabolites is not exceptional in the literature. Whilst specialised metabolites are routinely isolated from higher organisms, the underlying biosynthetic pathway producing such compounds is rarely elucidated. One could argue that this results from comparatively fewer genome sequences of higher organisms being available, or attribute it to poorer clustering of biosynthetic genes within these organisms. However, in many instances the cognate BGC is not encoded within the genome of the macroorganisms, but rather within the genome of an unknown microbial symbiont. Mutualistic symbiotic bacteria tasked with biosynthesis of often remarkable specialised metabolites have been a recent panacea in the field of specialised metabolites.

Entomopathogenic nematode gut commensals from the *Photorhabdus* and *Xenorhabdus* Genera are biosynthetic talented microorganisms of this sort. The complement of specialised metabolite produced by these organisms is posited to provide a chemical arsenal which is beneficial for the host nematode whilst infecting insect larvae, and prevents insect larvae putrefaction after death by suppressing growth of bacterial and fungal competitors. GameX peptides, szentiamide, xenobactin and xenoamicin A isolated from *Xenorhabdus* spp., the obligate mutualists of *Steinernema* spp. nematodes, all show good anti-*Plasmodium* activity and linear peptides bicornutin A, rhabdopeptides 1-8, dithiolopyrrolone xenorhabdins, NRP/PK hybrid xenocoumacin I and indole-derived nematophin all poses antimicrobial activities. The potency of nematophin towards multidrug resistant methicillin resistant *S. aureus* is reported to rival vancomycin. Interestingly, many peptide specialised metabolites are produced as mixtures by *Xenorhabdus* spp.: biosynthesis of a notably broad range of NRP analogues is exemplified by the 17 xentrivalpeptides (A-Q) (Figure 7.2) derived in part from overly promiscuous adenylation and condensation domains within NRPSs modules of one BGC. The production of a mixture of specialised metabolites from the same biosynthetic machinery could be a combinatorial prophylaxis mechanism providing defence against a broad range of pathogens or competitors, or alternatively facilitate biosynthesis of specific mixtures of metabolites specific for different stages of the host or symbiont lifecycle.

Undoubtedly, important lessons can be learned from xentrivalpeptide NRPSs when engineering less promiscuous megasynthases. Combinatorial prophylactic strategies are not unheard of between symbiotic micro- and macroorganisms: *Streptomyces* spp. interwoven in Beewolf digger wasp (*Philanthus* spp.) cocoons produce a cocktail of at least 9 bioactive compounds, including 8 pericidin derivatives and streptochlorin (Figure 7.2). Combinatorial therapy of all 9 compounds inhibited growth of ten relevant entomopathogenic and putrefactive bacteria and fungi, improving maturation rates Beewolf Wasp larvae. A nine-month hibernation period in soil burrows likely exposes Beewolf Wasp larvae to a broad range of pathogens.
Development of combinatorial prophylaxis using a range of antimicrobial compounds is an important innovation thought to provide long term and effective defence\(^47\).

Figure 7.2: Xentrivalpeptides, pericidins and streptochlorin – potential combinatorial therapy agents

The xentrivalpeptides (A-Q) comprise different substituents at R groups 1-4, resulting from promiscuity of NRPS adenylation and condensation domains. R\(_1\): acetyl-, propionyl-, butyryl-, isovaleryl-, hexanoyl-, 5-methylhexyl- or benzyl- groups. R\(_2\) and R\(_3\) are derived from valine, isoleucine or α-amino butyric acid. R\(_4\): benzyl- or isopropyl- group\(^46\). Structures of piericidin analogues (24-32) are fully elucidated elsewhere\(^47\). In short, analogues comprise hydroxyl-, methyl- or β-D-glucopyranose at R\(_1\), R\(_2\): methyl-, ethyl- or dimethylethyl- substituents as well as epoxidation at C11 and 12 and demethylation at C9.

Insects (Arthropoda) and nematodes represent some of the most diverse multicellular organisms on the planet\(^48\). It is no surprise that over a long evolutionary history insect and nematode associated bacterial strains have coevolved and specialised to fulfil useful mutualistic functions in complex and specific ecological niches. This is fantastically exemplified by maintenance of healthy elaborate Basidiomycetous fungal ‘gardens’ tended over by Ants of the Myrmicininae subfamily\(^49\). Fungal ‘gardens’ provide a foodstuff for the tending Ants, and to prevent predation of the fungal crop by other opportunistic invasive parasitic fungal species e.g. Escovopsis spp., Ants utilise the specialised metabolite producing capacity of symbionts e.g. Pseudonocardia spp. living within specialised cuticular crypts\(^50\) on their exoskeletal plates. Isolation of Pseudonocardia spp. from Apterostigma dentigerum (Ant) led to purification of the cyclic depsipeptide dentigerumycin with specific anti-Escovopsis activity, and limited activity towards the crop species\(^51\). Further bioassays confirmed dentigerumycin was active against other pathogenic fungi\(^51\), and undisclosed human cancer cell lines\(^52\). Ant symbionts are considered an untapped resource of useful specialised metabolites: Burkholderia spp.\(^53\), Amycolatopsis spp.\(^54\), and Streptomyces spp.\(^55,56\) have all been isolated from farming Ants, and
new antifungal compounds continue to be isolated from the bacterial-fungal-ant tripartite relationship52.

As well as terrestrial environments, complex marine symbiotic relationships are also fruitful sources of specialised metabolites. Marine sponges are ancient metazoans and accordingly have garnered much interest as models of ancient animal-bacterial symbiosis57. Sponges are well studied Eukaryotic scaffolds shown to support exceptionally complex ecosystems of specialised metabolite producers, a vast number of which are extremely distant from any known cultivatable organism58. Therefore, elucidation of the underlying BGCs involved in natural product biosynthesis necessitates the development of novel culturing techniques or culture independent techniques. Recently the cognate BGC for the clinically approved anticancer compound ecteinascidein 743 (trade name Yondelis) was identified from the genome of the Ecteinascidia turbinata tunicate symbiont Candidatus Endoeecteinascidia frumentensis59. Attempts to culture the intracellular Gammaproteobacteria Ca. E. frumentensis repeatedly failed, and construction of the ~0.631 Mb genome was achieved from Tunicate metagenomic DNA59. Metagenomic DNA is an excellent and abundant resource for identification of exotic BGCs from ‘microbial dark matter’60: the number of publicly available metagenomes grew from 2 to more than 20,000 between 2006 and 201524.

Advances in in situ culturing techniques are also an important recent success story for discovery of novel specialised metabolites. The iChip (isolation chip technology) is one pertinent example61. This devise is a multi-welled diffusion chamber which houses single uncultivatable bacteria sealed by a semi-permeable membrane, and is incubated in the environment from which the uncultivatable bacteria is isolated e.g. soil. iChips are touted to enable access to almost half of the prokaryotic biological matter in soil: a significant improvement on the predicted 1% accessible using standard culture techniques62. This approach is powerful and has facilitated identification of an effective antibiotic with novel lipid II binding activity, teixobactin (33) from Eleftheria terrae62, and isolation of closthioamide (34), the first specialised metabolite from a strict anaerobe63, with potent activity against clinical Neisseria gonorrhoeae isolates64 and other antimicrobial resistant microorganisms65. The versatility of the iChip is intriguing. One could imagine using this devise in a range of terrestrial and marine environment however the versatility of such a devise is yet to be seen.
Figure 7.3: Specialised metabolites identified by advanced culture techniques

We are beginning to map the relative ubiquity of known classes of BGCs across all prokaryotic taxa. This map shows a promising future for the natural product / specialised metabolite chemist, and will most likely improve in resolution as we develop more powerful bioinformatics tools for BGC identification and harness the synthetic biology toolbox to automate specialised metabolite characterisation.

Curiously, some specialised metabolite classes on our map are continuously underrepresented. Systematic analysis of genomes within the NCBI database\(^4,12,21,22,27,66-68\) highlight most major classes of BGC to be ubiquitous distributed across all recognised prokaryotic Phyla. Obvious exceptions exist e.g. Cyanobactins are thought to be conserved to Cyanobacterial spp.\(^69\) etc., however non-obvious exceptions also exist. Type II PKSs, responsible for the biosynthesis of aromatic polyketides, have historically been exclusively identified in Gram positive Actinobacteria, with few exceptions\(^2,70,71\). The underrepresentation of this class of natural products is confusing as evolutionary studies posit evolution of the core biosynthetic machinery, the ketosynthase (KS) and chain length factor (CLF), and acquisition of two cyclases, responsible for cyclisation and aromatisation of the linear polyketide, to occur prior to branching of major bacterial Phyla\(^72\).
In spite of this, only three examples of non-actinobacterial type II polyketide synthases have been documented to-date\textsuperscript{2,70,71}. These include the aurachin (\textit{aur}) BGC from \textit{Stigmatella aurantiaca} Sg a15, the anthraquinone (\textit{ant}) BGC from \textit{Photorhabdus} spp. and clostrubin (\textit{clr}) BGC from \textit{Clostridium} spp. (Figure 7.4). Whereas, more than 60 characterised Actinobacterial derived type II PKS BGCs are present in the MiBiG repository\textsuperscript{1}, and more than 75 clusters have been characterised in the literature\textsuperscript{72}. Interestingly, the distribution of the non-Actinobacterial clusters is not limited to a single taxon and there are no obvious commonalities in lifestyles of the producing organisms. Furthermore, the gene complement of the BGC and the structure of the specialised metabolites are not heavily conserved, suggesting the evolutionary relationship between non-Actinobacterial BGCs to be distant, if present at all.

Figure 7.4: Non-Actinobacterial derived aromatic polyketides

Aurachin A is the most extensively modified aurachin produced by \textit{Stigmatella aurantiaca} Sg a15, also known to produce aurachins B, C and D\textsuperscript{73} (not shown) (MiBiG ID: BGC0001343). It is disputed whether the \textit{aur} BGC comprises a canonical KS/CLF\textsuperscript{72}. AQ256 is the core anthraquinone (AQ) aromatic polyketide produced by \textit{Photorhabdus luminescens} and \textit{P. temperate}. AQ256 is further methylated by non-BGC associated enzymes to form 6 different AQS (MiBiG ID: BGC0000196). Clostrubin is the first aromatic polyketide to be identified from the Firmicutes: BGCs have been identified in \textit{Clostridium beijerinckii} Ng 34 and \textit{C. puniceum} BL70/20 (\textit{clr} cluster).

Little is known about the distribution of non-Actinobacterial type II PKSs as a result of their rarity. In the case of clostrubin, the \textit{clr} BGC is only identifiable in 2 of 425 \textit{Clostridium} spp. within the Integrated Microbial Genomes database (July 2017): \textit{Clostridium beijerinckii} Ng 34 and \textit{C. puniceum} BL70/20, despite the dataset comprising 15 additional \textit{C. beijerinckii} strains. Furthermore, systematic analysis of BGC from large datasets of taxonomically diverse
organisms\textsuperscript{4,21,66} often fail to discriminate between polyketide synthase classes when reporting data, confounding our understanding of intra-PKS class diversity. And, dedicated databases for the identification, curation and collection of type II PKS BGCs only considered sequences of Actinobacterial origin\textsuperscript{74}.

We postulate that non-actinobacterial type II PKSs are distributed widely throughout prokaryotic Phyla. This is evidenced by the sporadic observation of aromatic polyketides in unrelated non-Actinobacterial hosts and the calculation that type II PKS BGC evolution predates major bacterial speciation events\textsuperscript{72}. Based upon the successful expression of non-actinobacterial AntDE heterodimeric complexes in \textit{Escherichia coli} (Chapter 5) we posit this underexplored biosynthetic sphere is ripe for exploitation using \textit{E. coli} as a heterologous host system.

To substantiate these claims and elucidate the ubiquity and biosynthetic capacity of this biosynthetic class we bioprospect using a two-pronged approach. The first tests brute force bioinformatics for type II PKS BGC discovery, where the specialised metabolome of 2552 complete prokaryotic genomes were analysed using the industry standard BGC mining tool antiSMASH 3.0.5\textsuperscript{75}. The second approach uses protein-protein BLASTP to identify anthraquinone KS/CLF homologues, and antiSMASH 4.0\textsuperscript{76} to confirm the inclusion of these sequences in type II BGCs. Using both approaches we evidence the existence of a new divergent clade of KS/CLF proteins predominantly from non-Actinobacteria. Furthermore, we elucidate the applicability of \textit{Escherichia coli} as an appropriate heterologous expression host for recombinant KS/CLFs derived from these new clades.

Using these newly described type II BGCs we hope to fully characterise additional biosynthetic ‘parts’ to supplement the plug-and-play pathway described in Chapter 5, thereby expanding the aromatic polyketide chemical space accessible in \textit{E. coli} from anthraquinones, benzoisochromanequinones and dianthrones to deca-, dodeca- and larger aromatic polyketides. This effort further circumvents the need to use difficult genetically intractable native Actinobacteria, or \textit{Streptomyces} spp. heterologous expression hosts when exploring the type II PKS class of natural products. In doing so we successfully migrate exploration of type II PKSs from the Actinobacterial specialist to any molecular biology laboratory.

\subsection{7.4 Methods}

\subsubsection{7.4.1 Acquiring a dataset of genome sequences and BGC mining}

All available prokaryotic genomes were downloaded from the NCBI Reference Sequence Database (RefSeq) assembly summary for bacteria (August, 2016) for BGC mining. The RefSeq summary comprises 7809 complete genomes, 48197 contigs, 45782 scaffolds and 1443 chromosomes. Analysing all sequences was predicted to be computationally heavy and expensive. To reduce computation burden 1 in 3 ‘complete genomes’ accession numbers were chosen randomly resulting in a dataset of 2556 nucleotide sequences. This dataset was further refined by removing any entries defined as ‘Contigs’ or ‘Chromosome’. Four entries were
removed (Contig titles: *Escherichia coli* NO114, *Escherichia coli* and *Clostridium sporogenes*. Chromosome titles: *Leptospira borgpetersenii* str. 4E) resulting in a dataset comprising 2552 nucleotide sequences corresponding to prokaryotic genomes. These sequences were analysed by antiSMASH 3.0.5 locally. Probabilistic prediction of BGCs using the cluster finder algorithm was not appropriate to identify type II PKSs and therefore not employed to further reduce computation burden, despite this all 2552 sequences took more than one month to analyse. RefSeq identifiers, sequence information and BGC complement are available upon request. To identify AntD and E homologues, corresponding amino acid sequences were obtained from the *Photorhabdus luminescens* TT01 genome sequence (*plu*4191 and *plu*4190) and analysed using the BLASTP suite against the non-redundant protein sequence (nr) database using all parameters as standard. Nucleic acid accession numbers for DNA sequences encoding the gene sequences for all 100 homologues were submitted to antiSMASH 4.0 to identify BGCs. Where possible genomes of the corresponding organism were also analysed.

To improve resolution of type II PKS distribution amongst Genera, Species and Strains of organisms identified to comprise at least one putative type II PKS cluster, complete, partial, and draft genomes sequences of those taxonomic groups were all considered for analysis by antiSMASH 4.0, where necessary. A sample set of genomes were also analysed using the PKMiner tool (http://pks.kaist.ac.kr/pkminer/index.cgi?menu=3).

### 7.4.2 Phylogenetic and bioinformatic analysis of putative type II BGCs

#### 7.4.2.1 Phylogenetic tree construction

Phylogenetic trees were constructed using a multi-locus sequence approach when 16s rRNA gene sequences showed similarities exceeding 98% between organisms of interest. Ribosomal RNA gene sequences were obtained from Silva, the NCBI database, or from the Integrated Microbial Genomes databases (IMG). Gene sequences for *recA* and *gyrB* were used as additional molecular chronographs where required, and were acquired as above. Each gene sequence was aligned individually using MUSCLE before being clipped and concatenated manually to generate one hybrid sequence representative of all three genes, per organism. Concatenated sequences were aligned as above and used to construct maximum likelihood phylogenetic trees using MEGA7. Phylogenetic trees were bootstrapped with 1000 resampling iterations as standard. iTOL and MEGA7 were both used for tree visualisation.

#### 7.4.2.2 Elucidation of BGC genetic components

For each candidate biosynthetic gene cluster (BGC), gene function was defined by protein-protein BLAST (BLASTP) of all predicted protein coding sequences against the non-redundant protein sequence (nr) database (as of June 2017). Putative function was defined by homology to characterised and putatively assigned proteins, as well as through identification of protein domains using Conserved Domains database search (CDD), NCBI. All parameters were as standard. To identify putative insertion sequences and transposable elements, BGC sequences,
as defined by antiSMASH 3.0.5 or 4.0, were analysed using the ISFinder\textsuperscript{81} BLAST function and PHAge Search Tool Enhanced Release (PHASTER)\textsuperscript{82}.

Identification of false positive type II PKS BGCs was achieved by manual curation. BGCs were rejected if each cluster did not comply with a set of type II polyketide synthase specific rules. Rule 1) The BGC must include a discrete ketosynthase (KS) and chain length factor (CLF) gene sequence adjacent to one another, within an operon. Rule 2) The BGC must include one or more acyl carrier proteins (ACP). Rule 3) The BGC must include two or more ‘primary tailoring enzymes’, where ‘primary tailoring enzymes’ are defined as any enzymes which act upon the polyketide chain whilst it remains attached to an ACP, e.g. ketoreductase (C9 KR), cyclase (CYC), or aromatase (ARO/CYC).

Gatekeeper residue predictions for putative CLF and KS sequences were identified by alignment of predicted KS/CLFs against 60 characterised KS/CLF sequences, where residues corresponding to characterised act CLF gatekeepers F109, T112, F116, W194 and G195 (PDB: 1TQY chain B number) and act KS gatekeepers F140 and L143 were tabulated\textsuperscript{83}. Predicted gatekeeper residues were compared against corresponding residues from characterised CLFs shown to produce U-shaped C16, C20, C24 and C26, and S-shaped C20, polyketides to aid prediction of chain length. Sequences were aligned using MUSCLE in MEGA7 with standard parameters (gap opening penalty: -2.9, Gap extend: 0, hydrophobicity multiplier: 1.2, Cluster method: UPGMB).

Characterisation of polyketide chemotype was defined by predicted polyketide cyclisation patterns and polyketide chain length. Polyketide chain length was predicted as above. Polyketide cyclisation patterns were predicted using candidate CYC and CYC/ARO homology to 128 CYC and CYC/ARO from characterised type II PKS BGCs with known chemotypes. CYC and CYC/ARO amino acid sequences were acquired from the MIBiG repository or from the NCBI nucleotide and protein databases. The compiled repository of CYC and CYC/ARO sequences is available upon reasonable request.

7.4.3 Bacterial strains and culture methods

\textit{Escherichia coli} DH5\textalpha was used for routine cloning experiments and plasmid maintenance. All recombinant protein expression trials were undertaken in \textit{E. coli} BL21(DE3). Gene expression experiments were undertaken as reported in Chapter 5, (5.9.3). In short, the solubility of recombinant KS and CLF components were evaluated in \textit{E. coli} BL21(DE3) harbouring an appropriate expression vector pETKS/CLF cultured in lysogeny broth (LB) at 37°C, 180 rpm, supplemented with kanamycin 25 µg/ml or carbenicillin 50 µg/ml, to an OD\textsubscript{600} of 0.5 before induction with appropriate concentrations of isopropyl β-D-1-thiogalactopyranoside (IPTG). Post induction, incubation temperatures were reduced to 16°C, 20°C, 25°C and 30°C as specified and cultures were incubated for 4 h, 5 h, 9 h or overnight (typically 16 h) as specified within the text. Benzyl alcohol (10 mM) was added to select cultures 30 minutes prior to induction with
IPTG to improve the cytosolic concentration of molecular chaperones, negating the need to co-express chaperone genes from additional episomal elements\(^8^4\).

### 7.4.4 Refactoring and gene synthesis

#### 7.4.4.1 Rare Codons Cluster
Clusters of rare codons can be important in efficient protein biogenesis, and these translational dynamics should be considered when refactored gene sequences. To identify if any such slow codon clusters are present within characterised ketosynthase and chain length factor gene sequences a set of 34 KS and 33 CLF gene sequences were analysed for slow codons using %MinMax Rare Codon Calculator\(^8^5\). Ketosynthase and CLF sequences were aligned separately by MUSCLE in MEGA7 with standard parameters, described previously, and matched with the codon usage score at each amino acid position, calculated by the %MinMax Rare Codon Calculator. No regions of slow codons were conserved between KS or CLF sequences (Figure 7.5, Figure 7.6), therefore no specialised translational dynamics were considered when refactoring KS and CLF gene sequences.

![Figure 7.5: Translational dynamics of characterised KS sequences](image)

Graphical illustration of average codon usage for 34 characterised ketosynthases, in their native producers, calculated using the %MinMax codon calculator. In short, average codon usage within a sliding 17 amino acid is calculated for each AA along the protein sequence using codon frequency as a metric. This is compared to theoretical windows comprising the best and worst combinations of codons for all
positions, these are defined as 100% and -100%. Should half of the codons be the least frequently used and half the more frequent the score for the window would be 0 (percentage (%)). Average %MinMax calculation for all 34 characterised KS sequence is shown as a black line, and the range is defined by a light blue box. The green box shows the interquartile range. No regions of slow codons are defined using a resolution of 17 AA. Interestingly, a translational ramp is observed at the 5’ region of the mRNA.

Figure 7.6 Translational dynamics of characterised CLF sequences

Graphical illustration of average codon bias for 33 characterised CLF sequences, in their native producing organisms, calculated as in Figure 7.5. Black line shows the average codon %MinMax usage, the light blue area defines the range and the green area defined the interquartile range. At a 17 AA resolution no regions of slow codons are defined, the average codon usage is consistently above 60%.

7.4.4.2 Assigning start codons

Assigning the correct start codon for putatively assigned translated ORFs is not trivial as KS and CLF N-terminal sequences are rich in hydrophobic residues e.g. GTG valine and TTG leucine, and ATG, GTG and TTG are demonstrated KS and CLF start codons. Accordingly, care was taken when defining ORF starts.

All KS and CLF start codons were validated using a multiple sequence alignment of both characterised and query KS and CLF amino acid sequences. The number of residues preceding the highly conserved (V/I/I)V(I/L)TG sequence motif forming the beginning of β1, must be in line with expectations: Actinobacterial type II PKS components average ~5.4. Start codons were further validated at the nucleotide level manually, with Glimmer v3.0286, and the RBS
7.4.5 Construction of KS/CLF expression vectors

Uncharacterised KS/CLF pairs were chosen from *Ktedonobacter racemifer* DSM 44953 (KraA/B), *Delftia acidovorans* CCUG 247B (DacA/B), *Streptococcus* sp. GMD2s (SspA/B), *Candidatus Desulfofervidus auxilii* (DauA/B), *Bacillus endophyticus* DSM 13796 (BendA/B), *Pseudoalteromonas luteoviolacea* DSM 6061 (PluA/B), *Blautia wexlera* DSM 19850 (BweA/B) and *Gloeocapsa* sp. PCC 7428 (GloeA/B) and two characterised KS/CLF sequences were chosen from *Streptomyces resistomycificus* (RemA/B) and *S. antibioticus* ATCC 11891 (OvmP/K) for expression in *E. coli*. Nucleotide sequences were refactored to achieve a codon adaptation index similar to highly expressed *E. coli* housekeeping genes using a simulated annealing approach. Refactored nucleotide sequences were verified graphically using the %MinMax Rare Codon Calculator. Unfavourable intragenic alternative start sites were identified using the ribosome binding calculator and substituted manually, where necessary. To future proof the use of the KS/CLF gene sequences as biosynthetic parts NdeI, BamHI, NcoI, PstI, XbaI, EcoRI, HindIII, NotI and XhoI restriction endonuclease recognition (RE) sites were omitted from each gene sequence.

A synthetic duel KS/CLF expression cassette was designed in silico based upon the pETDuet-1 expression vectors (Novagen) for *kraA/B, dauA/B, bendA/B, pluA/B, gloeA/B, bweA/B* and *remA/B*. In brief, the cassette comprised a His\(^6\)-\(\Delta\)Methionine1_CLF fusion gene sequence, designed using the N-terminal hexahistidine tag (MGSSHHHHHHHHQDPSN) nucleotide sequence from pETDuet-1, upstream of the standard pETDuet multiple cloning site (MCS-2) intergenic region, comprising the T7 promoter for MCS2, the cognate Shine Dalgarno sequence and NdeI methionine start codon. In frame with NdeI M1 (MSC2) was the cognate KS gene sequence preceded by an N-terminal Strep-II tag (ASWSHPQFEKG) from pET51b (Novagen). The 5' and 3' ends of each cassette was flanked by an Ncol and XhoI RE site to facilitate cloning into the expression vector, pETM11-b. All nucleotide sequences were synthesised by GeneArt (ThermoFisher Scientific). Synthetic operons were cloned directly into pETM11-b by GeneArt, forming a series of pETKS/CLF vectors. Gene synthesis *gloeA/B* and *bweA/B* systematically failed. To construct *dacA/B, ovmP/K* and *sspA/B* co-expression vectors each synthetic gene sequence amplified from holding vectors pHold[KS/CLF] template DNA by polymerase chain reaction (PCR) using PCR primers containing RE sites within 5' overhangs. CloneAmp HiFi PCR premix was used for all PCR reactions (Takara) as per manufactures instruction. All primers annealing temperatures (\(T_a\)) were calculated using Integrated DNA Technologies (IDT) OligoAnalyser 3.1 (https://eu.idtdna.com/calc/analyzer, 2016/17), with \(T_a\) as close to 50°C as possible. Resultant CLF PCR products were flanked by NdeI and XhoI, and KS PCR products by NdeI and HindIII and cloned into MCS1 and MCS2 of pETDuet expression vectors forming pETDacB, pETDacAB, pETSspA, pETSspAB, pETOvmP and pETOvmPK. In all expression vectors the CLF CDSs were fused with an N-terminal hexahistidine tag, with the
exception of Ssp containing vectors, where the corresponding KS was His-tagged. For ΔKS
expression vector construction GeneArt cloned pETKS/CLF expression vectors were linearized
by PCR, removing the corresponding KR sequence from the amplicon. PCR primers were
developed with complementary 20 bp overhangs to facilitate reaggregation via Gibson DNA
Assembly (New England Biolabs) or In Fusion HD cloning (Takara) as per manufacturer’s
instructions. All vectors are available upon reasonable request and are detailed in
Supplementary Table 7.14.

7.4.6 Protein extraction and visualisation
Total cell lysate was extracted from E. coli BL21(DE3) cultures normalised to a total OD_{600} of 4.
Normalised cells were centrifuged at 4,000 x g and supernatant was discarded. To lyse cells
300 µl of BugBuster (Novagen) protein extraction reagent was added to cell pellets and
incubated on a rocker for 30 minutes before centrifugation at 12,000 x g for 20 min, 4°C.
Supernatant was removed from cell debris and designated as soluble cell lysate. Cell pellets
were resuspended in equal volumes of BugBuster (Novagen) and designated insoluble cell
lysate. For protein purification 400 ml cultures were typically used, and cultured as in 7.4.3.
Once more, culture supernatant was removed by centrifugation, 4,000 x g at 4°C for 20 min. E.
coli BL21(DE3) cell pellets were resuspended in buffer A (50 mM Tris-HCl, 300 mM NaCl, pH
7.4 5% glycerol (v/v)) supplemented with cOmplete Mini EDTA-free protease inhibitor cocktail
(Roche). All buffers were filtered sterilised using a 0.22 µm syringe filter (Merck). Cell
suspenion was sonicated on ice for 5 minutes and centrifuged at 12,000 x g for 25 min, 4°C.
Supernatant was removed and centrifuged a second time as above. Supernatant was once
more removed and applied to an immobilised metal ion affinity chromatography (IMAC) column,
Ni-NTA agarose (Qaigen), pre-equilibrated with buffer A. Flow through was collected and
reapplied to the IMAC column. The column was washed with 5 x column volumes (CV) of buffer
A before sequential 1 CV washes with Buffer A comprising increasing concentrations of
imidazole. Typically 20, 50, 200, 400 and 500 mM solutions were prepared, and are denoted on
each SDS-PAGE gel image accordingly. IMAC columns were re-equilibrated in buffer A before
washing in 20% ethanol. The pH of all buffers was calculated at 4°C and all buffer and protein
purification was carried out at 4°C.

To visualise protein samples by sodium dodecyl sulphate polyacrylamide gel electrophoresis
(SDS-PAGE), protein aliquots were added to fresh 2 x Laemmli SDS-PAGE loading dye (4%
sodium dodecyl sulfate (SDS) (w/v), 0.2% bromophenol blue (w/v), 20% glycerol (v/v) and 200
mM dithiothreitol), made up to 15 µl. Samples were boiled for 10 minutes prior to loading onto
10-12% SDS-PAGE gels (Biorad). Gels were run at 250V as standard in Towbin buffer (25 mM
Tris, 192 mM glycine, 0.1% SDS). PageRule prestained protein ladder (ThermoFisher Scientific)
was used as a molecular weight reference, unless stated otherwise. SDS-PAGE gels were
stained using InstantBlue protein stain (Expedeon) before washing with water and visualisation
using a Gel Doc EZ system (BioRad). Corresponding Western blots followed the above
procedure, however were not stained with InstantBlue (Expedeon). Instead SDS-PAGE gels
were transferred onto nitrocellulose membranes using Trans-Blot Turbo transfer packs (BioRad) as per manufacturer’s instructions. After transfer SDS-PAGE gels were stained with InstantBlue (Expedeon) to assess protein transfer quality. Nitrocellulose membranes were washed in deionised water for 5 minutes before transfer to iBind Western system (ThermoFisher Scientific). Western blots were carried out following manufacturer’s instructions. Primary H1029_.02ml monoclonal anti-polyhistidine antibodies produced in mouse were purchased from Sigma. Primary anti-strep-II monoclonal antibodies produced in mouse (71590-100VG) were purchased from abcam. Secondary antibodies used throughout were ab216772 goat pAb to mouse IgG, IRDye 800CW.

7.5 Results

7.5.1 A brute force approach to Identification of KS/CLF sequences from underexplored Phyla

Detection and classification of natural products is continuously being improved and this is evidenced by the expansion and refinement of 17 detectable BGC classes in antiSMASH 1.0\(^8\) to 42 classes in antiSMASH 4.0\(^7\), excluding cluster finder algorithm specific classes. Expectedly, analysis of 2552 reference genome sequences identified predicted BGCs spanning all 42 of classes, including hybrids thereof.

In all, 13,183 BGCs were predicted from the 2552 genomes analysed, comprising a vast sequence space totalling 7,050,113,326 bp, predicting 1.87 BGCs per Mb of genomic DNA. Genomes within the dataset provided good taxonomic coverage including, but not limited to, 333 representative sequences from the Actinobacterial Phylum, 568 from Firmicutes, 1086 from Proteobacteria. NRPS containing clusters were the most abundant single class of BGC, representing more than 17% of all BGCs identified, and more than 26% when hybrid clusters were considered (Figure 7.7). Furthermore, the pangenomic biosynthetic capacity for RiPPs was remarkable, totalling 2441 predicted BGCs (18.5% of all clusters) from 13 sub-classes, of which 1739 BGCs were predicted bacteriocins. With the addition of siderophores the total peptide based natural product reservoir totalled 5159 BGCs (39.1% of all clusters), omitting hybrid clusters. By comparison, only 1846 ketosynthase containing clusters (~14% of all clusters) were identified, comprising type I, II, III and trans-AT PKSs, other KSs, and arylpolyene predicted BGCs, without considering hybrid clusters (Figure 7.7).
Figure 7.7: Frequency of biosynthetic gene clusters identified by antiSMASH 3.0

AG: aminoglycoside. PGL: phosphoglycolipid, AHL: N-acyl homoserine lactone, PKS: polyketide synthase, NRPS: non-ribosomal peptide synthase. Head-to-tail: HTT cyclised peptides, Fused: phageanomycin-like peptides. Aminocoumarins are not displayed graphically: one cluster was detected. Botromycins were detected in *Streptomyces scabies* 87.22 however denoted as a bacteriocin-bottromycin BGC and therefore assigned as a hybrid cluster. Absolute values for polyketide synthase BGC were 750 type I PKSs, 398 type III PKSs, 90 other KSs, 72 type II PKSs and 68 trans-AT PKSs. Ribosomally synthesised and posttranslationally modified peptides (RiPPs) included fused phageanomycin-like peptides, glycocins, proteusins, linaridins, microviridins, cyanobactins, head-to-tails, microcins, thiopeptides, sactipeptides, lassoipeptides, lantipeptides and bacteriocins.

Of the KS containing clusters remarkably few were characterised as type II PKS BGCs, or hybrids thereof, with 72 and 107 BGCs predicted, correspondingly. The complete complement of type II PKS containing BGCs spanned 71 organisms, 60 of which were Actinobacteria. Actinobacterial BGCs were predicted in the genomes of 28 *Streptomyces* spp., as well as lesser known Actinobacterial Genera from different biogeographical and ecological origins, including *Catenulispora* spp. 89, *Modestobacter* spp. 90, *Luteipulveratus* spp. 91, and *Kutzneria* spp., highlighting the broad dissemination of type II PKS BGC within this Phylum. Interestingly, BGCs from 11 genomes were non-Actinobacteria in origin, including organisms belonging to *Alphaproteobacteria* (3), *Gammaproteobacteria* (6), *Deltaproteobacteria* (1) and *Bacilli* (1) Classes (Figure 7.8). This facile analysis promingly indicates ~10% of the predicted type II PKS BGCs to lie outside of the expected Actinobacterial host reservoir. The pool of predicted type II PKS BGCs was refined manually (7.4.2.2) identifying 7 incorrectly defined BGCs from Actinobacteria, including predicted BGCs from *Mycobacterium* spp. and *Nocardiosis* spp., and 9 false positive BGCs from non-Actinobacterial genomes. The refined set of BGCs comprised 96 Actinobacterial and only 2 non-Actinobacterial derived sequences (Figure 7.8). False positive identification rates were 7.29% and 81.81% for Actinobacterial and non-Actinobacterial BGCs, respectively.

The anthraquinone type II PKS BGC, common to *Photorhabdus luminescens* and *P. temperate*71, was identified as one of the two remaining non-Actinobacterial BGCs. The
aurachin and clostrubin clusters were not identified as representative genomes for aurachin producers *Stigmatella aurantiaca* Sg a15, *S. erecta* Pd e32 and *Rhodococcus erythropolis* were not present in the RefSeq dataset, nor were genome sequences for *Clostridium beijerinckii* Ng 34 or *C. puniceum* BL70/20.

### 7.5.1.1 New type II PKS BGCs from *Paenibacillus* spp.

Interestingly, a previously unreported type II PKS BGC from the spore-forming, nitrogen-fixing bacterium *Paenibacillus borealis* DSM 13188 was identified here (Figure 7.9a) which is not present in any of the 25 other *Paenibacillus* spp. analysed. Identical clusters were present in three additional *Paenibacillus* spp., identified by gene neighbourhood regions comprising top COG hits (IMG) and collectively defined as the *pab* BGC. The *pab* minimal PKS comprises a translationally coupled putative KS (*pabA*, ORF 3) and CLF (*pabB*, ORF 4), and an ACP (*pabC*, ORF 5) in an operon architecture mirroring the actinorhodin (*act*) BGC. The KS catalytic cysteine (C169) is conserved in PabA (KS), but replaced with an arginine in the PabB, consistent with its function as a CLF (Figure 7.9b).
Figure 7.8: Taxonomic distribution of type II PKS clusters defined by AntiSMASH 3.0 from a dataset of 2552 references genomes.

One putative and one characterised non-Actinobacterial type II PKS BGC were identified from 2552 complete genomes, denoted by black and blue arrows respectively. Red lineages represent BGCs identified from Actinobacteria, blue lineages represent BGCs from Proteobacteria. *Paenibacillus borealis* is of the Phylum Firmicutes and represented in black. All non-Actinobacterial organisms are within the greyed area. BGCs incorrectly defined as type II PKSs were identified manually (7.4.2.2) and are denoted by a black dot. The unlabelled lineage is the root of the tree, the furthest node is cellular organisms, the second is the Kingdom bacteria. The dendrogram was constructed using NCBI taxonomy IDs, constructed using phyloT (http://phylot.biobyte.de/) and visualised using interactive Tree Of Life (iTOL)\(^3\). Sixty-two of 71 genomes are displayed; the remaining 9 genomes did not have available NCBI taxonomy identifiers including *Streptomyces venezuelae* ATCC 15439, *S. vietnamensis* GIM4.0001, *S. ambofaciens* ATCC 23877 and DSM 40697, *Mycobacterium abscessus* subsp. boletii MM1513, *M. abscessus* DOJ-44274, *Enterobacter aerogenes* EA1509E, *E. aerogenes* CAV1320 and *Bartonella henselae* BM1374165. Bacterial names are not italicised for clarity purposes.

**Figure 7.9:** The pab type II PKS BGCs from *Paenibacilli*
a, A schematic of the predicted type II PKS BGC from various *Paenibacillus* spp. (light grey box) and flanking ORFs. Light blue ORFs represent the minimal PKS (mPKS). The dark blue ORF is an AcpS type 4' phosphopantetheinyl transferase. Gene coding sequences with predicted primary tailoring functions are red. Putative regulators are shown in black, and resistance associated ORFs in green. Grey ORFs are conserved gene sequences flanking the predicted cluster. All ORFs predicted to be within the BGC are numbered. Gene designations are detailed in Supplementary Table 7.1. b, phylogenetic reconstruction of KS and CLF components from a variety of type II PKS and FAS biosynthetic pathways. Construction is as detailed in 7.4.2.1. Bootstrap values above 75 are shown. Scale details the number of substitutions per residue. c, Amino acid sequence similarity of actinorhodin (*act*), resistomycin (*rem*), *ant* and *pab* cluster ACPs. Values are pairwise protein-protein BLASTP E-values.

An AcpS family 4' phosphopantetheinyl transferase (PPTase) (ORF 6) is present within the *pab* BGC which may be orthogonal for the unusual DST 4' phosphopantetheine attachment consensus sequence, also present in *ant* and *rem* ACPs, AntF (Chapter 5) and RemC. Whilst both AntF and RemC contain unusual DST motifs, each is posttranslationally functionalised with a phosphopantetheine arm by a PPTase from a different protein family, SFP and AcpS97, respectively. This divergence is reflected in ACP similarities suggesting cluster associated PPTases and ACPs to coevolve, and each may not be interchangeable. Interestingly, PabC and *act* ACP show the least similarity despite both being functionalised by AcpS family PPTases in the native host98.

Prediction of gatekeeper residues within PabB, and the lack of initiation module components, suggests the *pab* mPKS to produce an acetyl- or malonyl-primed C20 nascent polyketide chain. Furthermore, phylogenetic characterisation of the three *pab* cyclases (ORFs 2, 8 and 10) indicates the polyketide chemotype to be either discoid or linear. Carbon-9 ketoreductases (C9 KRs) are primary tailoring enzymes to most type II PKS BGCs, however no C9 KR is present within the *pab* BGC. This finding is consistent with characterisation of Pab2 as a TcmN-like first ring cyclase, shown to accept unreduced polyketide chain substrates in other biosynthetic pathways99. Other than cyclases, few obvious additional tailoring enzymes are presenting within the BGC. Intriguingly, a number of gene sequences encoding putative coproporphyrinogen oxidase III and corresponding B12-binding radical S-adenosyl-L-methionine (SAM) proteins are conserved between all *pab* clusters which may play roles in oxidative decarboxylation100 of polyketide intermediates. A single predicted major facilitator family transporter (ORF7) likely confers host resistance to the end compound by efflux and was further classified as a tetracycline MFS efflux component by the Antibiotic Target Seeker (ARTS) pipeline101 (Resistance model: RF0134, E-value 6.1e-60). Detoxification of tetracyclines through modification of the metabolite is rare; reducing the cytosolic concentration of tetracyclines by efflux is a considerably more common resistance mechanism102.

The identification of a predicted discoid or linear decaketide aromatic polyketide from a non-Actinobacterial organism is exciting; however, the type II PKS BGC identification frequency using the brute force approach here is disappointing.
7.5.2 A KS/CLF homology approach to Identification of KS/CLF sequences from underexplored Phyla

Identification of predicted type II PKS BGCs by KS/CLF protein pairwise similarity to AntE or AntD is a more refined method than the brute force approach detailed in 7.5.1, however suffers from obvious caveats. Fishing for new type II PKS BGCs using AntD and AntE amino acid sequences will likely only identify a BGC reservoir with similar KSs and CLFs to the anthraquinone cluster, missing KS/CLF pairs which show sequence diversity but functional homology. Such could be the case for the rem KS/CLF, where isolation of remA using degenerate PCR primers based upon canonical type II KS sequences failed continuously\(^9\). Furthermore, *Frankia* spp. have been shown to harbour type II PKSs with unique KS/CLF sequence characteristics which clade together strongly, and distantly from canonical KS/CLFs\(^1\) and could be missed using canonical KS/CLF homology based approach.

The flaws in this discovery method may be advantageous for downstream purposes, however. Through biasing the elucidation of non-Actinobacterial KS/CLF sequence towards AntDE homologues we may identify a more fruitful pool of predicted BGCs for discovery of novel type II polyketides using *E. coli* as a heterologous expression host. The AntDE heterodimer is a remarkably soluble recombinant protein complex when overexpressed in *E. coli*, and unknown sequence characteristics facilitating protein solubility may be better conserved in AntDE homologue containing BGCs. Increasing the number of soluble and functional KS/CLF heterodimers in *E. coli* will provide a pool of useful parts for the plug-and-play scaffold described in Chapter 5.

7.5.2.1 AntD top homologues

Of the top 100 AntD BLASTP hits 52 homologues were Actinobacterial in origin and included KSs from previously documented BGCs e.g. RemB from rem, and spore pigment KS from *Streptomyces collinus* Tü 365\(^1\). A non-exhaustive search of the Actinobacterial homologues shows each to be associated with a specialised metabolite BGC, rather than primary metabolism e.g. fatty acid biosynthesis. All ketosynthase homologues were remarkably similar, AntD and homologue 100 share 100% sequence coverage and 41% identity (E value: 6e-104). The top 15 homologues were hypothetical *Photorhabdus* spp. proteins within highly similar anthraquinone BGCs, (E value ≤ 8.00E-180), and thus were excluded from further analysis. BGC mining of contigs, chromosomes and genomes encoding the remaining 35 AntD homologues predicted 38 type II PKS BGCs across 17 different non-Actinobacterial Genera. After manual refinement (7.4.2.2) this number dropped to 26 BGCs across 17 non-Actinobacterial Genera, including five clusters from the candidate Phylum *Onmitrophica* (Table 7.1).

7.5.2.2 AntE top homologues

The sequence diversity between different AntE BLASTP homologues was greater than counterpart ketosynthase sequences: AntE and homologue 100 shared 65% quality coverage
and 31% identity (E value: 9e-30). The CLF plays a sequence-structure-function role in defining the polyketide chain length and the decarboxylation state of polyketide starter unit, so this is to be expected. Furthermore, the AntE C-terminal region is non-canonical71. Remarkably, of the top 100 homologues only 3 were classified with the actinobacterial Phylum: *Thermoactinomyces vulgaris*, *Dietzia timorensis* and *Streptomyces sclerotialus* and unlike AntD homologues, not all AntE homologues were BGC associated: an exhaustive search of the top 60 homologues identified a false positive similarity threshold E-value of 4e-34. Thirty-six AntD homologues had an E-value below the false positive threshold, and all were encoded within a type II PKS BGC with one exception, *Sorangium cellulosum* (accession: KYF95834.1). Additional type II PKS BGCs were identified manually with E-values between 1e-33 and 2e-31, however all AntE homologues with E-values above the latter threshold were false positives with no exceptions. Delimiting in this fashion, and refining BGCs manually, identified 42 non-Actinobacterial type II PKS BGCs across 16 non-Genera (Table 7.1): 10 homologues were identified from *Photorhabdus* spp. and were removed from the analysis.

### 7.5.2.3 Type II PKS BGC discovery cross over

The sum of manually refined non-actinobacterial BGC identified in 7.5.2.1 and 7.5.2.2 totalled 58 and spanned 18 different Bacterial Genera, however of these only 12 BGCs were unique and unreported (Table 7.1). In most cases AntD and E BLASTP homologues were pairings of cognate KS and CLF sequences from one BGC respectively, and this is reflected in the number of BGC duplicates listed in Table 7.1. Only four unique BGCs were identified using AntE or AntD homologues, and not both. These include predicted BGCs from *Bacillus endophyticus*, *Gloeocapsa* sp. PCC 7428, *Desulfobacterium* sp. 4572_20, and the homologous *Lactobacillus* spp. and *Streptococcus* spp. BGCs.

Whilst the total number of unclassified predicted BGCs is low the taxonomic diversity of host organisms is broad (Figure 7.10): predicted type II PKS BGCs were identified in 5 *Proteobacteria* Classes: Acidithiobacilli (1), Alphaproteobacteria (2), Betaproteobacteria (1), Gammaproteobacteria (1) and Deltaproteobacteria (3), 3 Firmicute classes: Bacilli (4), Clostridia (2) and Negativicutes (1), and members of the Cyanobacteria, Chloroflexi and Candidate Omnitrophica phyla. Importantly, for the latter three phyla the predicted type II PKS containing BGCs are ‘first in Phylum’ examples. Concerted efforts to sequence the genomes of more exotic species within these Phyla will determine if these divisions are reservoirs of type II PKS BGCs.

### 7.5.2.4 Phylogeny of principle PKS components

Manually curated candidate non-Actinobacterial KS identified by our analysis (Table 7.1) clade discretely away from canonical Actinobacterial KSs (Figure 7.11d). Intriguingly, RemA, ClrA, AntD and PabA also clade with non-Actinobacterial KSs, despite not identifying the latter from the AntD top 100 homologues, suggesting the type II PKS class of specialised metabolites can be further subdivided into canonical Actinobacterial and non-Actinobacterial PKSs. A similar observation is made for non-Actinobacterial CLF sequences, where two discrete non-
Actinobacterial CLF clades (clade b and c) are formed (Figure 7.11b,c), however division of each non-Actinobacterial CLF clade is poorly supported by bootstrapping. Interestingly, clade C exclusively comprises CLFs from Proteobacteria, whilst clade B comprises CLF sequences from all Phyla described in 7.5.2.3. The existence of two clades is likely to result from the vertical acquisition of a CLF comprising a noncanonical C-terminal sequence, observed in AntD, in one instance, and the acquisition vertically, or otherwise, of a CLF comprising a more canonical C-terminal sequence in the other. Discrete clading of non-Actinobacterial principle PKS components is remarkable and confirms the presence of type II PKS evolving outside of the Actinobacterial phylum. The existence of a divergent lineage of type II PKSs has been theorised previously\textsuperscript{105}, however we described this subdivision for the first time here.
Table 7.1: Non-Actinobacterial organisms comprising one or more predicted type II PKS BGCs

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>antiSMASH 3.0</td>
<td></td>
<td></td>
<td>antiSMASH 3.0</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>Delftia acidovorans</td>
<td>3</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>Acidithiobacillus ferrorans</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>Azospirillum brasilense</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>Blautia spp.</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>Lachnospiraceae bacterium 3-1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>Lactobacillus salivarius</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>Lactobacillus oris</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>Streptococcus spp.</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>Candidatus Desulfofervidus auxilli</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>Deltaproteobacteria bacterium RBG_16_54_11</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>Desulfbacterium sp. 4572_20</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>N/A</td>
<td>Clostridium beijerinckii</td>
<td>1</td>
<td>N/A</td>
<td>Reported</td>
</tr>
<tr>
<td>1</td>
<td>N/A</td>
<td>Clostridium puniceum</td>
<td>1</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>Bacillus endophyticus</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>Dendrosporobacter quercicus</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>Gloeocapsa sp. PCC 7428</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>Hoeflea sp. IMCC20628</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>Ktedonobacter racemifer*</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>11</td>
<td>5</td>
<td>Omnitrophica spp. Pseudoalteromonas luteoviolacea</td>
<td>11</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td></td>
<td>Photorhabdus spp.</td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>52</td>
<td></td>
<td>Actinobacteria</td>
<td>3</td>
<td></td>
<td>N/A</td>
</tr>
<tr>
<td>100</td>
<td>28</td>
<td>Total</td>
<td>49</td>
<td>34</td>
<td>12</td>
</tr>
</tbody>
</table>

*K. racemifer comprises three predicted type II PKS BGCs: two satisfy manual curation criteria. Underlined organisms contain characterised BGCs. Coloured field's show organisms comprising BGCs predicted to produce the same or extremely similar specialised metabolites.
Figure 7.10: Phylogenetic distribution of all curated non-Actinobacterial type II PKS clusters

The blue lineage and the unlabelled lineage are described in Figure 7.8. The green lineage represents Firmicutes, and includes *P. borealis* identified using the brute force approach (bolded). The orange lineage comprises *Omnitropa* spp. from the Planctomycetes-Verrucomicrobia-Chlamydiae (PVC) Superphylum, and more specifically the invalidated *Omnitrophica* candidate Phylum. Cyanobacteria are shown in black and Chloroflexi in purple. Organisms harbouring characterised type II PKS BGCs are denoted with a black arrow. Not all non-Actinobacterial organisms described in Table 7.1 are shown, however at least one representative of each species is. Bacterial names are not italicised for clarity purposes.
Figure 7.11: KS/CLF phylogeny

Phylogenetic relationship of characterised Actinobacterial and uncharacterised non-Actinobacterial KS/CLFs (identified in 7.5.1 and 7.5.2). Type III PKSs (red) and FabH (orange) are used as outgroups, supported by bootstrap values above 95%. Clades a and e represent 55 characterised canonical Actinobacterial CLF and KS amino acid sequences, respectively. Non-Actinobacterial KSs are shown to clade together (Clade d, bootstrap values of >99%), and away from canonical Actinobacterial KSs. Non-Actinobacterial CLF sequences form two discrete clades (b and c, unsupported by bootstrapping), which clade apart from canonical Actinobacterial CLFs (bootstrap value: 100%). Maximum likelihood tree was computed as described in 7.4.2.1. Bacterial names are not italicised for clarity purposes.
<table>
<thead>
<tr>
<th>Cluster / host organism</th>
<th>Gatekeeper residues</th>
<th>CL</th>
<th>AS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>112</td>
<td>116</td>
<td>194</td>
</tr>
<tr>
<td>Actinorhodin</td>
<td>T</td>
<td>F</td>
<td>W</td>
</tr>
<tr>
<td>Oxytetracycline</td>
<td>G</td>
<td>L</td>
<td>W</td>
</tr>
<tr>
<td>WhiE spore pigment</td>
<td>G</td>
<td>L</td>
<td>I</td>
</tr>
<tr>
<td>Griseorhodin</td>
<td>G</td>
<td>M</td>
<td>L</td>
</tr>
<tr>
<td>Resistomycin*</td>
<td>A</td>
<td>L</td>
<td>G</td>
</tr>
<tr>
<td>Gram - Anthraquinone</td>
<td>I</td>
<td>V</td>
<td>-</td>
</tr>
<tr>
<td>Gram + L. oris PB013-T2-3</td>
<td>S</td>
<td>F</td>
<td>T</td>
</tr>
<tr>
<td>Gram + Streptococcus sp. GDM6s</td>
<td>S</td>
<td>F</td>
<td>T</td>
</tr>
<tr>
<td>Gram - Gloeocapsa sp. PCC 7428</td>
<td>C</td>
<td>F</td>
<td>M</td>
</tr>
<tr>
<td>Gram - D. acidovorans</td>
<td>T</td>
<td>L</td>
<td>F</td>
</tr>
<tr>
<td>Gram - C. Desulfobulbus auxilii</td>
<td>A</td>
<td>L</td>
<td>A</td>
</tr>
<tr>
<td>Gram - Pseudoalteromonas luteoviolacea</td>
<td>T</td>
<td>L</td>
<td>L</td>
</tr>
<tr>
<td>Gram + Bacillus endophyticus</td>
<td>A</td>
<td>L</td>
<td>G</td>
</tr>
<tr>
<td>Gram + K. racemifer C3</td>
<td>A</td>
<td>L</td>
<td>A</td>
</tr>
<tr>
<td>Gram + K. racemifer C2</td>
<td>A</td>
<td>L</td>
<td>A</td>
</tr>
<tr>
<td>Gram + Paenibacillus borealis</td>
<td>A</td>
<td>L</td>
<td>A</td>
</tr>
<tr>
<td>Gram + Blautia wexlerae</td>
<td>V</td>
<td>G</td>
<td>M</td>
</tr>
</tbody>
</table>

CL: Polyketide chain length. AS: Active site cysteine substitution residue. * Resistomycin is and S-shaped discoid polyketide, all others are U-shaped. **threonine and leucine may not sterically inhibit the chain sufficiently to prevent polyketide extension in the D. acidovorans CCUG274b CLF. In the case of Actinorhodin a bulkier phenylalanine follows the threonine residue and occludes the amphipathic cavity. †Predicted chain lengths. Gatekeeper residues 112, 116, 194, 109, 195 are CLF derived and 143’ and 140’ from the KS. The residues are ordered from the closest to the ketosynthase active cysteine. Phenylalanine 116 is important in dimer interactions. The genus of organisms sharing the same initial is written in full for clarity.
7.5.3 Non-Actinobacterial type II PKS components as soluble recombinant proteins in *Escherichia coli*

The observation of a discrete lineage of non-Actinobacterial type II PKSs is unprecedented and represents an untapped resource of specialised metabolites. The recent success in expressing antDE from this divergent lineage in *E. coli* posits this biotechnologically virtuous host as a plausible heterologous background to drive discovery of non-Actinobacterial aromatic polyketides. To test this hypothesis the solubility of a selection of KS/CLF sequences described in Table 7.1 were evaluated in *E. coli* BL21(DE3). Type II PKS BGCs from *Delftia* spp., *Streptoccocus* spp., *Pseudoalteromonas* spp., *Bacillus* spp., Ca. *D. auxilii*, and *Ktedonobacter* spp. are explored along with the biogeographical and ecological niche each organisms inhabits. The solubility of each corresponding KS and CLF proteins are evaluated as heterodimers and monomers in *E. coli* BL21(DE3).

### 7.5.3.1 Type II PKSs from *Delftia* spp.

An exhaustive search of currently available *Delftia* spp. genome sequences in the NCBI database (as of April 2017) shows putative type II polyketide synthase gene clusters to be present in the genomes of one additional *Delftia* spp. to the three identified from AntE BLASTP searches and analysis of 2552 genomes. *Delftia* spp. are members of the *Proteobacteria* Phylum, *Betaproteobacteria* Class and *Burkholderiales* Order, common also to *Burholderia* spp., known to be abundant reservoirs of biosynthetic gene clusters. Interestingly, the type II PKS clusters in *Delftia* spp. are not localised to one strain type, one species, or species’ isolated from a common environment (Figure 7.12). Specifically, clusters were identified in *Delftia acidovorans* CCUG 15835 and CCUG 274B isolated from the human microbiome, *Delftia lacustris* LZ-C isolated from metal contaminated industrial waste waters and *Delftia* sp. JD2 isolated from chromium contaminated soils. The sporadic observation of the type II PKS cluster may be due to the culmination of poor taxonomic classification of these closely related organisms here, recent speciation, or simply via horizontal gene transfer. All clusters are largely syntenic, with a maximum of 513 single nucleotide polymorphisms (SNPs) between the most dissimilar clusters as determined by antiSMASH 4.0, predicting the specialised metabolites to be highly similar, if not identical. Genomic regions proximal to the putative type II PKS BGC are rich in mobile genetic elements: the genomic location of the BGC in *Delftia acidovorans* CCUG247B is flanked upstream by a DDE transposase and downstream by ORFs with putative phage related functions, suggesting this area of the chromosome to tolerate insertions (Figure 7.13). Interestingly, a DDE transposase separates the BGC from a thiamine biosynthetic gene cluster comprising ThiE, ThiG, ThiS, ThiO and ThiD gene sequences conserved in all *Delftia* spp. tested (Figure 7.13). The presence of the ISS/IS1182 family transposase exclusively in *Delftia* sp. containing the type II BGC is suggestive of BGC acquisition horizontally; however the ORF represents only a DDE transposase fragment. No inverted repeats, or other mobile genetic element hallmarks, flank this gene sequence. Exceptionally similar BGCs were also identified in *Azospirillium brasilense* and *Acidothiobacillus ferrivorans* which both encode a cluster associated PPTase (ORF 11) opposed to the putative prenyltransferase in *Delftia* spp. (Figure 256).
7.13 Supplementary Tables 7.2 – 7.5). PPTases are not routinely associated with type II BGCs, and in canonical Actinobacterial native producers fatty acid synthase (FAS) PPTases are capable of functionalising BGC associated ACPs[8]. Despite being distributed across several members of the biotechnologically relevant *Delftia* genus this type II PKS BGC remains unannotated, with its function and specialised metabolite unknown. For all additional bioinformatics analysis and wet work the type II PKS BGC from *D. acidovorans* CCUG274B is considered.

![Phylogenetic reconstruction of 14 of 19 Delftia spp. analysed by antiSMASH.](image)

**Figure 7.12: Phylogenetic reconstruction of 14 of 19 *Delftia* spp. analysed by antiSMASH.**

Phylogenetic assignment is based upon concatenated 16s rRNA and recA nucleotide sequences to build a maximum likelihood phylogenetic tree. Bootstrap values are representative of 1000 resampling events, only values exceeding 70 are displayed. Red circles indicate strains harbouring a putative type II PKS BGC. *Xenophilus azovorans* DSM 13620 was used as an outgroup.
Figure 7.13: Genomic localisation of type II PKS BGC in *Delftia* spp.

The putative type II PKS cluster (Light blue ORFs: mPKS, red ORFs: tailoring enzymes) is inserted between thiamine biosynthetic genes (orange ORFs), and a suite of genes with varied functions common to a range of *Delftia* sp. and *Burkholderiales* sp. (white ORFs). Efflux pumps (MacB – macrolide transport system), macrolide specific efflux systems and an outer membrane factor lipoprotein are defined in green. *Delftia acidovorans* CCUG 15835 harbours a Mu-like prophage (dark blue ORFs). The putative transposase present in *D. acidovorans* CCUG 15835 and 274B is shown in grey. Numbered gene designations and amino acid pairwise sequence BLAST scores between *D. acidovorans* spp., *A. ferrivorans* YL15 and *A. brasilense* are described in Supplementary Table 7.2 - Supplementary Table 7.5.

### 7.5.3.1.1 *Delftia acidovorans* CCUG274B

The boundaries of the putative biosynthetic gene cluster were assigned manually through comparison of common genes within the four clusters identified in *Delftia* sp. and the genomic localisation of flanking genes among other *Delftia* sp. (Figure 7.13). Delimiting in this fashion isolated a gene cluster harbouring 12 ORFs encoding a putative ketosynthase (3), chain length factor (4) and acyl carrier protein (5), a cyclase/dehydratase (1), two ketoacyl reductases (2, 10), four antibiotic biosynthesis monoxygenases (6-9), a flavodoxin reductase (12) and a 1,4-dihydroxy-2-naphthoate octaprenyltransferase (11) (Supplementary Table 7.2). Inclusion of a putative prenyltransferase suggests the end product may be a polyketide terpene hybrid, also known as hybrid isoprenoids (HI). Polyketide terpene hybrid compounds are not uncommon secondary metabolites e.g. naphterin, napyridomycin A, furaquinocin A and marinone\(^\text{108}\), however the aforementioned metabolites all utilise a 1,3,6,8-tetrahydroxynaphthalene (THN) core synthesised by a type III PKS and not type II biosynthetic machinery present in the *Delftia* sp. derived BGC. Alternatively, the biosynthetic complement of the BGC could plausibly
synthesize quinone containing vitamins, coenzymes, similar alkylquinolone specialised metabolites\textsuperscript{109}, or related quorum sensing molecules\textsuperscript{110}.

The \textit{Delftia} spp. KS, CLF and ACP (termed DacA, DacB and DacC here-in) order is consistent with the gene architecture of canonical type II polyketide biosynthetic gene clusters; however, the KS and CLF gene sequences are not translationally coupled. BLASTP shows the putative CLF from the \textit{D. acidovorans} CCUG274B cluster to share 26\% identity and 92\% coverage with the actinorhodin (\textit{act}) CLF, maintaining a canonical C-terminus not present in AntE, despite residing within clade C (Figure 7.11). The gatekeeper residue profile of DacB is similar to OxyB indicating the KS/CLF pocket to accommodate a C\textsubscript{20} polyketide. Gatekeeper residues for \textit{A. brasilense} and \textit{A. ferrivorans} CLFs are consistent with DacB (Table 7.2).

7.5.3.1.2 Evaluation of DacA and DacB expression in \textit{E. coli} BL21(DE3)

DacA and DacB amino acid sequences are more similar to those of AntD and E than Actinobacterial derived type II PKSs and both clade together outside of the respective canonical actinobacterial KS and CLF clades (Figure 7.11). With \textit{Delftia} spp. and \textit{Escherichia} spp. both classified into the Phylum Proteobacteria this KS/CLF pair were plausible candidates to test the general applicability of \textit{E. coli} as a heterologous host for discovery of aromatic polyketides from non-actinobacterial hosts.

Unlike AntD and E, DacA and B are not soluble when expressed in \textit{E. coli} BL21(DE3) (Figure 7.14a,c). Visualisation of soluble cytosolic protein via SDS-PAGE shows no recombinant protein in \textit{E. coli} harbouring pETDacAB, a dacA and \textit{his\textsuperscript{6}}-dacB expression vector, when induced with 50 µM or 200 µM IPTG. As DacA and B are proposed to form a heterodimeric quaternary structure, expression of DacB alone would lead to inclusion body formation. Indeed, inclusion bodies were detectable by SDS-PAGE and Western blotting from the insoluble protein fraction of \textit{E. coli} BL21(DE3) expressing \textit{his\textsuperscript{6}}dacB alone, but also when coexpressed with dacA, its heterodimeric partner (Figure 7.14b,d).

Interestingly, \textit{his\textsuperscript{6}}DacB could be purified from soluble protein lysate extracted from \textit{E. coli} BL21(DE3) pETDacAB, confirmed by peptide digestion and identification (Figure 7.13a), suggesting the monomeric protein to be present in the \textit{E. coli} proteome at a low level. DacA was not copurified however, as was observed for the AntDE complex. The absence of DacA indicates \textit{his\textsuperscript{6}}DacB to exist as soluble aggregates, a poorly soluble monomeric protein, or to form heterodimeric complexes with \textit{E. coli} endogenous αβαβα thiolases including those from fatty acid biosynthesis, as homodimeric complexes or alternatively as a highly dissociable DacA/B heterodimer where protein:protein interactions are destabilised during purification.

Primarily, as analysed here, DacA and B follow the general dogmatic view which states type II PKS machinery is not soluble when overexpressed as recombinant protein in \textit{E. coli} heterologous hosts. Detection of oxytetracycline in \textit{E. coli} BAP1 did not require visualisation of soluble OxyA or B via SDS-PAGE\textsuperscript{111}, and therefore DacAB may be functional in \textit{E. coli}.
Figure 7.14: Evaluating solubility of DacA and DacB in *E. coli* BL21(DE3)

**Figure 7.14:** Evaluating solubility of DacA and DacB in *E. coli* BL21(DE3)

**a.** Soluble protein from *E. coli* BL21(DE3) (NPC: no plasmid control) and *E. coli* BL21(DE3) harbouring pETDuet-1, pETDacB and pETDacAB, empty, single and duel KS/CLF expression vectors. *dacA/B* gene expression was either uninduced or induced with 50 or 200 µM IPTG, for each lane. *E. coli* BL21(DE3) host control and empty plasmid controls were either uninduced or induced with 200 µM IPTG. Protein expression was induced for 5 h at 16°C. His₆mCherry (~29 kDa) (lane 12) was used a positive control for Western blotting and is denoted by a red arrow. Lanes designations are consistent in **b**, showing extracted insoluble protein, **c**, Western blot of **a**, and **d** Western blot of **b**. DacA and His₆DacB (44.14 and 41.35 kDa, respectively) cannot be resolved from soluble cytosolic protein via SDS-PAGE (**a**), or via Western blotting (**c**) using anti-polyhistidine IgG antibodies, where His₆mCherry is visible. Equivalent SDS-PAGE of insoluble protein fractions (**c**) show additional protein bands which correspond well with the theoretical size of recombinant His₆DacB (black arrow) and DacA (dashed arrow). Western blotting of insoluble protein
shows a signal consistent with His\textsuperscript{6}DacB in *E. coli* BL21(DE3) expressing dacB or dacAB, denoting the KS/CLF pair from *D. acidovorans* CCUG 274B to be predominantly insoluble in *E. coli* BL21(DE3). Standardised molecular weight protein ladders (lane 5) show relative protein size in kDa.

![Image of protein ladder and Western blot](image)

**Figure 7.15: Purification of DacB from *E. coli* BL21(DE3)**

a, purification of soluble polyhistidine tagged proteins from *E. coli* BL21(DE3) pETDacAB induced with 200 µM IPTG via immobilised metal ion affinity chromatography (IMAC). Elution of DacB occurs after addition of 200 mM imidazole to sonication buffer A and is denoted by a black arrow. DacB cannot easily be resolved by SDS-PAGE in the flow through fraction (FT), and is not visible in protein fractions eluted with less than 200 mM imidazole (e.g. 20 mM, shown here). A band of comparable intensity to DacB and corresponding to 44.14 kDa (DacA) is not visible in the protein fraction eluted with 200 mM imidazole. Lanes labelled X mM correspond to protein eluting in sonication buffer supplemented with X mM imidazole. b, Western blot of protein purified as in a using anti-polyhistidine IgG antibodies. A signal for DacB is only present in the protein flow through fraction (FT) and the fraction eluting in sonication buffer containing 200 mM imidazole. His\textsuperscript{6}mCherry is used as a positive control, denoted by the red arrow, as in Figure 7.14. L: standardised protein molecular weight ladder. Integers correspond to protein size (kDa).

### 7.5.3.2 *Streptococcus* sp.

*Streptococcus* spp. are medically relevant organisms. Accordingly, efficient taxonomic characterisation and identification is essential to enable the correct and most effective course of treatment, but is also useful from an epidemiological viewpoint when tracking infections and understanding evolution of virulence factors between isolates. Characterisation of *Streptococcus* spp. is not a trivial task however, and polyphasic approaches are necessary to correctly identify taxonomic novelty\textsuperscript{112}. The mitis group of the *Streptococcus* genus highlight the difficulty in classification. The archetypal molecular chronograph 16s rRNA genes of different mitis group species are often highly similar and even overlap\textsuperscript{113} while DNA-DNA hybridization shows significant intra-species variability, where isolates of the same species show DNA-DNA relatedness below the typical boundary for speciation\textsuperscript{113}. Poor intra-species DNA-DNA
hybridisation can be explained, in part, by the genomic plasticity of *Streptococcus* spp. afforded by an ability to become naturally competent and acquire exogenous environmental DNA. Pathogenicity islands, virulence factors and antibiotic resistance markers are commonly acquired and replaced between genetically competent organisms and lysed non-competent counterparts, or transitory bacteria, facilitating the rapid evolution and exchange of DNA. Genomic plasticity is a powerful survival tool in competitive environments such as plaque, the oral cavity and nasopharynx from which organisms of the mitis group are commonly isolated. Competence, tolerance to insertions, genome plasticity and isolation from highly competitive environments are all hallmarks attributed to *Streptomyces* spp. which pin-point these organisms as good sources of novel bioactive molecules. These features are shared with *Streptococcus* spp. and their utility in augmenting cellular physiology and lifestyle is exemplified by the diversity of capsular polysaccharides and virulence factors identified in closely related *Streptococcus* spp.. Despite these similarities, *Streptococcus* spp. are only now being identified as potential reservoirs of natural products.

In light of this, the secondary metabolome of 136 *Streptococcus* spp. from the 2552 genomes analysed by antiSMASH 3.0.5 did not include a type II PKS biosynthetic gene cluster and no type II PKS entries in the antiSMASH database are derived from *Streptococci*. Rather, a putative type II PKS cluster was identified by pairwise similarity to AntE using BLASTP. Putative clusters were identified in *Streptococcus* sp. GMD2s, GMD4s, GMD5s and GMD6s which characterise taxonomically within the *Streptococcus* mitis group, and more specifically within the *Streptococcus oralis* cluster of the *oralis* clade. Type II PKS biosynthetic machinery is uncommon in *Streptococcus* spp. and type II PKS clusters could not be identified in the closely related *Streptococcus oralis* 1212 SMIT, 206 SPSE, 274 SPSE and 727 SORA strains. The phylogenetic proximity of these organisms was defined by a maximum evolution tree computed from a 313,115 bp core genome derived from 195 mitis group *Streptococcus* spp. elsewhere.

### 7.5.3.2.1 *Streptococcus* sp. GMDXs

The type II PKS clusters from all 4 *Streptococcus* spp. are identical at the DNA level: sequence alignments of all clusters, with boundaries defined by antiSMASH, were 100% similar. Accordingly, the shared cluster is referred to as GMDXs here-in. Bidirectional searches for genomic DNA comprising similar COG assignments, as in 7.5.3.1.1, highlighted the 4 identical clusters from *Streptococcus* sp. GMDXs to also be strikingly similar to a type II PKS cluster from *Lactobacillus oris* PB013-T2-3 and *Lactobacillus salivarius* 609_LSAL, also isolated from the human oral flora. The *L. oris* PB013-T2-3 and *L. salivarius* 609_LSAL BGCs lack one ORF, with putative carboxymuconolactone decarboxylase activity, however. Both *Lactobacillus* species were identified by BLASTP analysis (Table 7.1, Supplementary Table 7.7).

BGC boundaries were refined manually and predict the *Streptococcus* BGC to contain 12 ORFs (Figure 7.16, Supplementary Table 7.6 and 7.7). ORFs flanking the ketosynthase in the 3’ orientation differ between the *Streptococcus* sp. and *Lactobacillus* sp. clusters in both putative protein function and amino acid identity: the *Streptococcus* sp. GMDXs neighbouring gene sequence encodes a putative MutR transcriptional regulator whereas in *L. oris* PB013-T2-3 the
gene is hypothetical. Similar analysis at the BGC 5’ region shows the BGC to be inclusive of a DUF2992 containing protein and a possible cluster associated regulator. The putative regulators from *Streptococcus* spp. and *L. oris* PB013-T2-3 share limited similarity at the amino acid level (37% coverage and 39% identity) and therefore may differentially regulate gene expression. Genes neighbouring each putative regulator show no functional conservation between the two clusters, nor do these gene products show plausible secondary metabolic function and are likely to be involved in other cellular processes. The *Streptococcus* spp. GMDXs BGC is flanked by *Streptococcus* phage phiNJ3 DNA polymerase III, and a serine recombinase and the *L. oris* PB013-T2-3 cluster is situated between two resolvases and a site-specific recombinase adding weight to BGC transmission horizontally.

![Figure 7.16 Genomic localisation of type II PKS BGC in *Streptococcus* spp.](image)

Schematics of the *Streptococcus* spp. GMDXs and *L. oris* PB013-T2-3 type II PKS BGC. Genes flanking 3’ of the BGC are conserved across other *Streptococcus* spp.. Light blue ORFs show minimal PKS components, red ORFs are genes encoding putative tailoring functions, dark blue ORFs encode an ACPS type PPTase (5) and a benzyl-CoA ligase (6). Transposon elements are shown in grey. Phage related ORFs and resolvases are shown in purple. Regulators are black. For ease of alignment a gap in the *Streptococcus* sp. AS20 genome is shown as a dotted line. All gene designations for both *Streptococcus* sp. GMDXs and *Lactobacillus oris* PB013-T2-3 are described in Supplementary Table 7.6 and Supplementary Table 7.7.

Coding DNA sequences common to *Streptococcus* spp. and *L. oris* PB013-T2-3 include genes with putative KS (1), CLF (2), ACP (3), AcpS type PPTase (5) and benzoate CoA ligase (6) protein function, as well as two cyclases (9, 10), methyltransferase (4), cupin 2 with similarity to TcmJ (7), alpha/beta hydrolase (11) and DUF2992 containing protein (12). As with *dacA*, *dacB* and *dacC*, the *Streptococcus* sp. GMDXs KS, CLF and ACP, termed sspA, sspB and sspC here-in, conform to the dogmatic gene order present in most type II PKSs, however once more sspA and sspB are not translationally coupled. BLASTP shows SspA to share 31% identity and 87% coverage with *act* KS and 37% identity and 95% coverage with AntD, and SspB to share 28% and 28% coverage and 98% and 87% identity with *act* CLF and AntE, respectively,
consistent with SspA and B clading with the non-actinobacterial type II PKS components (Figure 7.11). Gatekeeper residues controlling polyketide chain length are in agreement with Actinorhodin, where phenylalanine 116 occludes the amphipathic cavity at the KS/CLF dimer interface, predicting the polyketide to be C16 in length (Table 7.2).

Interestingly, all 6 clusters from the Order Lactobacillales lack a ketoreductase, the first primary tailoring enzyme in aromatic polyketide maturation. Ketoreductases specifically reduce one carbonyl- within the poly-β-ketone chain and the absence of a ketoreductase from a BGC is uncommon. Exceptions exist in the case of tetracenomycin\textsuperscript{122}, resistomycin\textsuperscript{97} and spore pigment biosynthesis all of which are classified as non-reduced systems as a result. The absence of a ketoreductase and presence of a putative benzoate CoA are also hallmarks of quinoline alkaloid biosynthesis. The aurachin cluster responsible for biosynthesis of a family of isoprenoid quinolone alkaloids from the Myxobacterium \textit{Stigmatella aurantiaca} Sg a15 lacks a ketoreductase\textsuperscript{70} and encodes an anthranilate CoA ligase. AuaB and C, the aurachin KS and CLF, are responsible for the elongation of the unusual anthranilate starter unit by condensation of two successive malonyl-CoA extender units, a reaction enzymatically more similar to those catalysed by canonical type III PKSs.

### 7.5.3.2.2 Solubility of SspA and SspB in the \textit{E. coli} BL21(DE3)

The Gram positive \textit{Streptococcus} spp. encoding sspA and sspB are not phylogenetically proximal to the intended heterologous host \textit{E. coli}, unlike antDE from \textit{Photorhabdus luminescens} TT01, however despite this, SspA and SspB do form soluble dimeric recombinant complexes in \textit{E. coli} BL21(DE3) (Figure 7.17 and Figure 7.18). The higher molecular weight hexa-histidine-SspA fusion protein can be resolved from the \textit{E. coli} BL21(DE3) soluble proteome when expressing both his\textsuperscript{6}-sspA and sspB, but not \textit{E. coli} BL21(DE3) expressing his\textsuperscript{5}sspA alone (pETSspA) or \textit{E. coli} BL21(DE3) plasmid and background controls. SspB cannot be resolved by SDS-PAGE as is masked by the highly abundant \textit{E. coli} endogenous elongation factor Tu, which shares a similar molecular weight (Figure 7.17a, lanes 10 and 11). While His\textsuperscript{6}SspA is faintly visible in the soluble proteome of \textit{E. coli} pETSspAB induced with IPTG, both proteins are abundant in the insoluble protein proteome, indicating SspA and SspB to be primarily insoluble recombinant proteins when overexpressed in \textit{E. coli} at 30°C (Figure 7.17a, lanes 7, 8, 10 and 11).

Western blotting confirmed His\textsuperscript{6}SspA to be completely insoluble in the absence of its CLF counterpart, SspB. Identification of His\textsuperscript{6}SspA as a soluble protein is only possible upon co-expression of sspA and sspB (Figure 7.17c and d) consistent with SspA and SspB forming heterodimeric quaternary complex. Untagged SspB readily co-purifies with the his\textsuperscript{6}SspA fusion protein when purified by IMAC (Figure 7.17), in agreement with SspAB forming a heterodimeric complex when co-expressed in \textit{E. coli} BL21(DE3) and conforming to the dogmatic view that KS and CLF proteins form soluble heterodimeric complexes as their smallest quaternary form. No information on high order protein complexes can be determined here, however. A significant metabolic burden is applied to heterologous hosts harnessing the T7 expression system at 30°C for 16 hours and therefore additional temperature for gene expression were also tested.
including 20°C and 16°C, where formation of soluble SspAB complexes were also observed (data not shown).

Figure 7.17: Evaluating the solubility of SspA and SspB in E. coli BL21(DE3)

a. Soluble protein cell lysate of E. coli BL21(DE3) (NPC) and E. coli BL21(DE3) harbouring pETDuet-1, pETSspB and pETSSpAB expression vectors induced with 0, 50 or 200 µM IPTG. pETSspA encodes a hisSsspA fusion gene sequence (HisSspA: 46.59 kDa) downstream of a T7 promoter. pETSspAB encodes HisSspA and sspB (42.86 kDa), both under the control of individual T7 promoters. Gene expression was induced for 16 h at 30°C. HisSspA is denoted by a black arrow and visible only in E. coli BL21(DE3) pETSspAB induced with 50 and 200 µM IPTG. Sample designation and lane number are consistent in b, showing insoluble cell lysate. HisSspA is visible as an insoluble protein in E. coli pETSspB and pETSspAB induced with 50 or 200 µM IPTG. No proteins corresponding to the theoretical mass of HisSspA are visible in protein extracted from the E. coli plasmid and background controls. Western blots
corroborate these findings (c and d), showing recombinant His\textsuperscript{6}SspA to soluble only upon co-expression of sspB. No signal is detected from soluble protein extracts of \emph{E. coli} pETSspA un/induced. His\textsuperscript{6}mCherry is positive control (PC: lane 12) as in Figure 7.14.

*Figure 7.18: Co-purification of SspA and SspB from \emph{E. coli} BL21(DE3)*

\begin{itemize}
  \item[a] SDS-PAGE of polyhistidine containing proteins from \emph{E. coli} pETSspAB soluble cell lysate purified by IMAC. His\textsuperscript{6}SspA and SspB co-elute after addition of 200 mM and 400 mM imidazole to elution buffer A, denoted by solid and dotted black arrows respectively. PAGE migration distances are in agreement with Figure 7.17, and respective molecular masses.
  \item[b] Western blot of a using Mouse anti-polyhistidine IgG \textalpha\ primary antibodies. Controls are as described in Figure 7.15. Signals corresponding to His\textsuperscript{6}SspA are present in the soluble protein fraction, flow through (FT) and 200 mM and 400 mM imidazole containing eluent. EF: elusion fraction, Sol: total soluble cell lysate.
\end{itemize}

\subsection*{7.5.3.3 \textit{Pseudoalteromonas luteoviolacea} spp.}

As with the \textit{Streptococcus} BGC, a putative type II polyketide synthase biosynthetic gene clusters in a \textit{Pseudoalteromonas} spp. was not identified from the initial 2552 RefSeq genome dataset analysed by antiSMASH 3.0.5. Seven \textit{Pseudoalteromonas} spp. genomes were within the dataset analysed including both non-pigmented (\emph{P. atlantica} T6c and \emph{P. translucida} KMM 520) and pigmented \textit{Pseudoalteromonas} strains (\emph{P. luteoviolacea} B, \emph{P. issachenkoni} KCTC 12958, \emph{P. phenolica} KCTC 12086, \emph{P. rubra} and \textit{Pseudoalteromonas} sp. Bsw20308)\textsuperscript{123}. Instead, a putative BGC was identified in \textit{P. luteoviolacea} DSM 6061 by amino acid pairwise similarity to AntE (Figure 7.10). An exhaustive search of all available \textit{P. luteoviolacea} spp. type strain genomes (NCBI, April 2017) identified 5 further putative type II PKS BGCs from this species (Figure 7.19).

\emph{P. luteoviolacea} spp. are ubiquitous in marine environments\textsuperscript{123}, however there is no obvious commonality between the ecological niches from which the 6 strains harbouring the putative
type II BGC are isolated: *P. luteoviolacea* DSM6061, NCIMB1942, NCIMB2035 and CPMOR-2 were all isolated from surface water whereas *P. luteoviolacea* S4060-1 and 2607 were isolated from seaweed and rock surfaces respectively, and most were distinct from one another in biogeographical origin124. Despite this, all 6 type II PKS BGC containing *P. luteoviolacea* strains are derived from a common ancestor suggesting BGC acquisition to be vertical (Figure 7.19). The specialised metabolome of *P. luteoviolacea* spp. has previously been shown to be ~5 fold more genetically diverse than the associated pangenome, enabling organisms to respond to changing environmental pressures123. Counterintuitively, the BGC complement of *P. luteoviolacea* strains is not unequivocally linked to its biogeographical origin/ecological niche125, and vertical acquisition of BGCs is common, in line with findings here.

The mechanism of cluster acquisition or loss by horizontal gene transfer is less obvious between *P. luteoviolacea* spp.. In contrast to the *ssp* and *dac* BGCs there are no obvious mobile genetic elements flanking the *P. luteoviolacea* BGC (*plu*). Rather, BGC flanking genes have putatively assigned biosynthetic functions e.g. GNAT acetyltransferases, SnoA-like aldol condensation catalysing enzymes and efflux pumps indicating this genomic region to be a hotspot for specialised metabolite associated genes. Consistent with this, *P. luteoviolacea* DSM6061 and CMPOR-2 both encode an additional uncharacterised NRPS BGC 10 kb upstream of the type II PKS BGC, indicating this region of the genome to be tolerance to BGC insertion and rearrangement.

*Pseudoalteromonas* spp. are well characterised organisms. Certain strains provide industrially relevant host chassis for bioproduction of high quality eukaryotic proteins126,127 and others are well known reservoirs of halogenated natural products128. Irrespectively, the type II PKS BGC distrusted widely in genome sequences of *P. luteoviolacea* spp. it is yet to be characterised.

**Figure 7.19: Multilocus phylogenetic tree of *P. luteoviolacea* strains**

Phylogeny was constructed using *recA*, *gyrB*, and 16s rRNA gene sequences. Bootstrapped values below 75 are not shown. Red dots indicate strains harbouring the putative type II PKS cluster. All 6 strains harbouring the putative type II PKS are shown to be derived from the same lineage and therefore from a
common ancestor, plu BGC acquisition is denoted by a black dot. The scale bar represents the number of substitutions per site.

7.5.3.3.1 Pseudoalteromonas luteoviolacea DSM 6061

The plu cluster boundaries were refined manually by identification of conserved flanking gene sequences present in Pseudoalteromonas spp. lacking the minimal type II PKS (Figure 7.20). Delimiting in this fashion identified gene sequences proposed to have chemotaxis related functions flanking the 3’ region of the cluster, as orientated in Figure 7.20, and hypothetical genes with homologues widely distributed among Pseudoalteromonas spp., in the 5’ direction. Oddly, this delimitation resulted in the inclusion of ORFs putatively encoding a cytochrome C556, cytochrome B and cytochrome C oxidase and points towards the end compound playing a role in the electron transport chain, as a toxin or otherwise70,73 (Supplementary Table 7.8). Production of electron transport chain inhibitors has been reported for Pseudomonads109. These quinolones bind the quinone reduction site (Q1) of the cytochrome bc1 complex inhibiting electron transfer109. The presence of additional cytochrome components in the putative type II BGC from P. luteoviolacea may be involved in self-resistance. Each plu cluster comprises CDSs putatively assigned with KS (12), CLF (13) and ACP (14) functions, as well as an SFP type PPTase (15), two ketoreductases (11 and 2), an aromatase (10), a putative oxidoreductase (6), three monooxygenases (3-5) and four hypothetical genes (9, 18-20) (Figure 7.20). The KS, CLF and ACP from all P. luteoviolacea spp. (termed pluA, pluB and pluC here-in) mirror the genetic organisation of canonical type II PKSs; however the pluA and pluB are not translationally coupled.

Gatekeeper residues defining the amphipathic cavity volume differ from the act CLF, and three of five amino acids gatekeepers are ‘open’ predicting the chain length to be C20 (Table 7.2). C20 linear polyketides form the backbone of a wide range of different aromatic polyketide chemotypes, including the tetracenomycins, angucyclines, anthracyclines and aureolic acids. Expression of pluA and pluB as soluble and functional proteins in E. coli would enable access to a larger aromatic polyketide chemical space than currently achievable.

Figure 7.20: Genomic localisation of type II PKS BGC in Pseudoalteromonas luteoviolacea.

A schematic of the plu BGC and corresponding genomic region of different Pseudoalteromonas spp.. Minimal PKS components are light blue ORFs, enzymes with tailoring function are encoded by red ORFs
and regulators are shown in black. No transposable elements are identified. Grey dotted line denotes an absence of sequence: in short, all numbered ORFs are missing from \( P. \ luteoviolacea \) S407-1, NCIMB 1944 and 2ta16. All gene designations are detailed in Supplementary Table 7.8.

### 7.5.3.3.2 Evaluation of PluA and PluB expression in \( E. \ coli \) BL21(DE3)

Remarkably, PluA and PluB are soluble recombinant proteins in \( E. \ coli \) BL21(DE3) (Figure 7.21). N-terminal Strep-II fusion tagged PluA (46.96 kDa) is visible in the \( E. \ coli \) BL21(DE3) soluble proteome after induction of pETPluAB, a His\(^6\)-pluB, StrepII-pluA co-expression vector, with 50 µM or 200 µM IPTG. His\(^6\)-PluB (43.31 kDa) is once more masked by \( E. \ coli \) endogenous elongation factor Tu however can be visualised by western blot using anti-polyhistidine IgG antibodies (Figure 7.21a and b). His\(^6\)-PluB is an abundant insoluble recombinant protein in the \( E. \ coli \) BL21(DE3) insoluble proteome whereas StrepII-PluA is comparatively unabundant. Expression of strepII-sspA from MCS 2 of the co-expression vector may be comparatively in efficient to MCS1 and explain this deficit.

Despite both proteins showing solubility in \( E. \ coli \) BL21, Strep-II-PluA and His\(^6\)-PluB copurify poorly (Figure 7.22). A dominant band corresponding to His\(^6\)-PluB elutes from gravity Ni-NTA IMAC column along with a much fainter band, confirmed to be Strep-II-PluA by Western blotting (Figure 7.22b). Repetition of the experiment using automated IMAC purification yielded identical findings (data not shown here). Disparity between band intensity of Strep-II-PluA and His\(^6\)-PluB is not in line with the expected 1:1 stoichiometry for heterodimeric complexes.

The poor stoichiometry of purified PluAB could be explained by weaker, less stable, or more transient KS/CLF interactions leading to heterodimer formation than corresponding interactions between C\(_{16}\) producing counterparts, AntDE and SspAB, which retain their quaternary form under identical purification conditions. The PluAB complex is predicted to accommodate a larger polyketide chain than AntDE, SspAB and act KS/CLF, with more gatekeeper residues in an open position, resulting in a larger amphipathic tunnel at the dimer interface. The larger amphipathic cavity of the putative C\(_{20}\) PKS may result in a KS/CLF protein-protein binding penalty, as fewer residues are forming interactions necessary to stabilise the heterodimeric complex than in C\(_{16}\) counterparts. Interactions between act KS Tyr118 and CLF Phe116 at the dimer interface are tighter than corresponding interactions between asymmetrical FabF homodimers\(^{83}\), and are proposed to aid correct heterodimer formation. PluA also a Tyr118Phe substitution and PluB has a Phe116Leu substitution as a consequence of ‘opening’ additional gatekeeper residues. Poorer interactions between PluA Phe118 and PluB Leu116 than corresponding Tyr118 and Phe116 could perturb stable KS/CLF complex formation. Gatekeeper residues are known to play important structural roles in KS/CLF stability mutation of all four act CLF gatekeepers to alanine resulted in abolition of polyketide biosynthesis\(^{129}\). Likewise, mutation of two gatekeeper residues in TcmL, the tetracenomycin CLF, also terminated polyketide biosynthesis resulting from extensive structural perturbations\(^{129}\).

The PluAB complex may require additional stabilising components to retain its quaternary form. The biosynthesis of resistomycin requires the rem mPKS and two additional cluster associated
cyclases to function\textsuperscript{130} and cyclases have previously been ascribed structural roles in spore pigment KS/CLF stabilisation\textsuperscript{130,131}. Furthermore, the unexpected stoichiometry of PluA and B may result from destabilisation of PluAB during purification, in the absence of additional stabilising PKS components, therefore coexpression with a BGC associated cyclase may improve recovery of dimeric complexes.

**Figure 7.21: Evaluation of PluA and B solubility in *E. coli* BL21(DE3)**

*a*, SDS-PAGE analysis of His\textsuperscript{6}-PluB and StrepII-PluA fusion protein solubility from *E. coli* BL21(DE3) pETPluAB total cell lysate. Lanes 1-5 show proteins isolated from *E. coli* BL21(DE3) soluble cell lysate and lanes 7 through 11 shows insoluble protein. *E. coli* BL21(DE3) and *E. coli* BL21(DE3) pETM11b-AntB, induced with 200 µM IPTG represent host background and vector controls in lanes 1, 2, 7 and 8. Soluble and insoluble protein extracted from *E. coli* BL21(DE3) pETPluAB uninduced and induced with 50 µ and 200 µM IPTG are visualised in lanes 3 through 5 and 9 through 11 respectively. StrepII-PluA (46.96 kDa) is visible in both soluble and insoluble protein fractions, *a*, lanes 4, 5, highlighted by a white dashed arrow, and lanes 10 and 11 identified by a black arrow dashed respectively. His\textsuperscript{6}-PluB (43.31 kDa) is visible in lanes 10 and 11 (solid black arrow).

*b*, Western blot counterpart of *a*, resolving StrepII-PluA from soluble and insoluble cell lysate protein using Anti-strep-II IgG primary antibodies.

*c*, Western blot of *a* resolving His\textsuperscript{6}-PluB from soluble and insoluble cell lysate protein. Lanes are as described in *a*. Anti-polyhistidine IgG
primary antibodies were used. His\textsuperscript{6}mCherry positive control is denoted by a red arrow (lane 12). PageRuler prestained protein ladder is used as a molecular weight reference (kDa).

![Image](image.png)

**Figure 7.22: Co-purification of PluA and PluB from E. coli**

a. Tandem purification of His\textsuperscript{6}PluB and StrepII-PluA does not show the 1:1 stoichiometry expected for dimeric complexes. Proteins were purified by IMAC. His\textsuperscript{6}PluB elutes in sonication buffer A with the addition of 20 mM imidazole (lane: 20 mM), and is not visible in column flow through (FT), when separated by SDS-PAGE (10%). b. Western blot confirmation of strepII-PluA copurification using anti-strep-II primary antibodies. StrepII-PluA coelutes with His\textsuperscript{6}PluB, and is not visible in the column flow through. Reference molecular weight ladder (L) is PageRuler prestained protein ladder (kDa).

### 7.5.3.4 Bacillus spp.

82 Gram-positive rod-forming *Bacillus* spp. were included in the 2552 RefSeq genome dataset analysed by antiSMASH 3.0.5 including 23 *B. subtilis* spp., 9 *B. cereus* spp., 9 *B. velezensis* spp., 7 *B. anthracis* spp. and 11 *B. amyloliquefaciens* spp. however using this approach no putative type II PKS biosynthetic gene clusters were detected. Instead, a putative type II PKS cluster was identified in *B. endophyticus* DSM 13796 and KCTC 13922 by amino acid pairwise similarity to AntE (Figure 7.23, Figure 7.24).

*Bacillus endophyticus* was first isolated from the inner tissues of *Gossypium* sp., commonly referred to as Cotton, and classified as an independent *Bacillus* lineage based on polymorphic DNA patterns and growth characteristics\textsuperscript{132}. Strains within this species possess plant growth promoting properties through antibiotic, antifungal\textsuperscript{133} and algicidal\textsuperscript{134} metabolite biosynthesis, repression of plant ethylene synthesis\textsuperscript{135} and promotion of plant pathogen defences. Reva and colleagues\textsuperscript{132} suggest *B. endophyticus* to persist predominantly in symbiosis with plants however strains are widely distributed and have been isolated from complex hypersaline environments void of higher *planta* as well as the gut of *Phlebotomus papatasi*\textsuperscript{134}. To determine if type II PKS BGCs are ubiquitous to *B. endophyticus* spp. all additional *B. endophyticus* spp. genomes were analysed using antiSMASH 4.0. Putative type II PKS clusters were identified in *B. endophyticus* DSM 13796 and KCTC 13922, but not *B. endophyticus* 2102 or Hbe603, the
latter two were isolated from hypersaline lakes and industrial co-cultures respectively\textsuperscript{134,136}. Literature states both DSM 13796 and KCTC 13922 to be the \textit{B. endophyticus} 2DT\textsuperscript{T} type strain isolated from \textit{Gossypium} sp., with different database identifiers, however the specialised metabolite BGC profiles between organisms suggests this to be false: \textit{B. endophyticus} DSM 13796 comprises an additional terpene and lantipeptide BGC. The putative type II PKS BGCs from both strains were highly similar at the DNA level however, and only differ in the loss of an RNA binding S1 protein gene sequence outside of the \textit{B. endophyticus} KCTC 13922 BGC.

### 7.5.3.4.1 \textit{Bacillus endophyticus} DSM 13796 and KCTC 13922

Genomic regions flanking the \textit{B. endophyticus} BGC (\textit{bend}), as defined by antiSMASH 4.0, comprise a plasmid replication / relaxation protein homologous to \textit{Bacillus} Poppyseed phage, however no additional transposable genetic elements were identified obscuring any obvious means of acquisition via HGT (Figure 7.23a). Defining the BGC gene complement as in Figure 7.13 and 7.5.3.1.1 was not possible as too few \textit{Bacillus endophyticus} spp. genome sequences are available, and flanking gene sequences are not conserved in other closely related \textit{Bacillus} spp.. Instead the cluster boundaries were defined manually, using predicted protein function (Figure 7.23a). The \textit{bend} cluster comprises two operons, the first encodes gene sequences for a TetR/AcrR transcriptional regulator (5), three ABC transporter components (2,3,4) and a pentachlorophenol-4-monoxygenase, or UbiH like monooxygenase (1). The second comprises a Helix-turn-helix domain containing protein (6), σ-70 family RNAP sigma factor (7), RemF, RemL and RemI cyclase homologues (8, 11 and 17), two methyltransferases (9 and 12), a tryptophan halogenase homologue (10), minimal PKS components, including a KS, CLF, SFP type PPTase and ACP (13-16) and a RemK homologue, with unknown function (Supplementary Table 7.9). This BGC was named the \textit{bend} cluster. The \textit{bend} cluster has a remarkably similar gene complement and operon architecture to the resistomycin biosynthetic gene cluster (Figure 7.23b) comprising a homologue of each cyclase tasked with formation of the discoid S-shaped polyketide core, suggesting this cluster to also produce a discoid polyketide. The \textit{rem} and \textit{bend} minimal PKS components share high AA similarity and clade together (Figure 7.11) consistent with the \textit{bend} and \textit{rem} BGC products being highly similar.
Figure 7.23: *B. endophyticus* DSM 13796 and KCTC 13922 type II PKS BGC schematic

**a**, Schematic of the *bend* BGC isolated from *B. endophyticus* DSM 13796 and KCTC 13922. The putative ABC efflux pump components are detailed as green ORFs, the mPKS as light blue ORFs, PPTase and MCAT components as royal blue ORFs, cyclases as red ORFs, methyltransferases as grey ORFs and regulators in black. White ORFs have variable, unknown or hypothetical putative functions. **b**, Comparison of the *bend* and *rem* clusters. **c**, Structures of two discoid polyketide products: resistomycin from *S. resistomycificus*105 (MiBiG ID: BGC0000264.1) and chlorxanthomycin from an undisclosed *Bacillus* sp.137. All gene designations are detailed in Supplementary Table 7.9.

Figure 7.24: Phylogeny of *Bacillus endophyticus* spp.
Phylogenetic tree is constructed using concatenated 16s rRNA and recA DNA sequences. The concatenated 16s rRNA and recA gene sequences of *B. endophyticus* KCTC 13922 and DSM 13796 differ by two base pairs, where nucleotides were called as N. Closely related *B. smithii* DSM 4216 and *B. magaterium* QM B1551 are included, as well as *Paenibacillus polymyxa* YC0573 as an outgroup. Strains denoted with a red dot contain the bend type II PKS BGC. Phylogeny is in agreement with NCBI dendrogram calculated by genomic BLAST (https://www.ncbi.nlm.nih.gov/genome/?term=bacillus+endophyticus). Bootstrapping values of below 90 are not shown. Tree scale corresponds to substitutions per base.

*B. endophyticus* species are poorly represented within DNA sequence databases and it is not remarkable that this type II PKS BGC is yet to be characterised in the literature. Interestingly, the halogenated pericyclic pyrene chlorxanthomycin, similar in structure to resistomycin (Figure 7.23c), has been identified from an undisclosed soil-borne *Bacillus* spp. which is phylogenetically proximal to *B. smithii*137. The striking similarities between the rem and bend BGC gene complement, the identification of discoid rem cyclase homologues and presence of a tryptophan halogenase determine the bend BGC to be a contender for the biosynthesis of chlorxanthomycin. Whilst chlorxanthomycin is yellow/orange and non-diffusible like resistomycin, *B. endophyticus* DSM 13796, isolated from cotton were shown to sporadically produce a red/pink non-diffusible pigment however pigment producing cell lines could not be isolated. Pigment biosynthesis could be induced when isolates were grown under stress e.g. with lysozyme or ampicillin132. It is likely that physiological and environmental stresses play a role in regulation of the bend cluster mediated a cluster associated sigma factor (7) (Figure 7.23). *B. endophyticus* strains have strong temporal adaptations to environmental change, supported by a repertoire of 17 alternative sigma factors, and nearly 300 regulatory associated genes138. Elucidation of the specialised metabolite encoded within the bend cluster using *E. coli* as a heterologous host first required evaluation of KS/CLF heterodimer solubility.

Interestingly the bend mPKS does not maintain the canonical Actinobacterial KS/CLF/ACP operon architecture but mirrors that of resistomycin, where an ORF putatively assigned as an SFP PPTase is present between the CLF and ACP gene sequences (Figure 7.23b). As with remA and remB, the bend KS and CLF, bendA and B, are not translationally coupled. However, unlike remB, the bendB Shine-Dalgarno sequence is not located within the coding sequence of the upstream KS gene. BLASTP shows BendA and B to be most similar to RemA and B, respectively, (Table 7.3) in agreement with the phylogeny of the bend KS and CLF (Figure 7.11). The BendB gatekeeper residues profile is identical to RemB indicating the bend mPKS to synthesise a C20 polyketide (Table 7.2). BendB shows poor similarity to AntE: this can be explained by the canonical CLF C-terminal sequence of BendB, absent in AntE.
Table 7.3: Pairwise similarities of the bend KS/CLF and archetypal U- and S-shaped polyketide synthases

<table>
<thead>
<tr>
<th>Protein Identity / coverage (%)</th>
<th>act KS*</th>
<th>AntD</th>
<th>RemA</th>
<th>act CLF</th>
<th>AntE</th>
<th>RemB</th>
</tr>
</thead>
<tbody>
<tr>
<td>BendA</td>
<td>41 / 99</td>
<td>42 / 99</td>
<td>53 / 98</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BendB</td>
<td>34 / 98</td>
<td>31 / 61</td>
<td>49 / 98</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*act KS amino acid sequence was acquired from chain A of PDB: 1TQY, rather than the translated sco5087 sequence (Accession: AL645882.2)

7.5.3.4.2 Evaluation of BendA and BendB expression in E. coli BL21(DE3)

The solubility of the proposed heterodimeric BendAB KS/CLF cannot be fully evaluated here. Induction of a his<sup>6</sup>bendB/bendA duel expression vector (pETBendAB) in E. coli BL21(DE3) resulted in visualization of a single insoluble recombinant protein corresponding to either bend KS/CLF by SDS-PAGE (Figure 7.25a). His<sup>6</sup>BendB and StrepII-BendA have similar theoretical molecular weights, 46.98 kDa and 47.412 kDa respectively, and resolving both bands via 1D SDS-PAGE is difficult. Accordingly, Western blots were used to improve resolution as each protein comprised a unique N-terminal epitope tag. Only His<sup>6</sup>BendB was detectable from E. coli insoluble cell lysate using this approach (Figure 7.25b): no signal in either soluble or insoluble protein lysate was detected for StrepII-BendA (data not shown). His<sup>6</sup>BendB showed low levels of solubility within the E. coli BL21(DE3) recombinant proteome when purified by IMAC (Figure 7.26a and c), however did not co-purify with StrepII-bendA (Figure 7.26c), unlike AntDE, SspAB and PluAB.
Figure 7.25: Evaluation of BenAB solubility in *E. coli* BL21(DE3)

Soluble and insoluble protein fractions from *E. coli* BL21(DE3) pETBendAB total cell lysate were separated visualised via SDS-PAGE. *E. coli* BL21(DE3) no vector (no plasmid control, NPC) and with an empty expression vector, pETM11b, were used as controls. **a**, A distinct single band of approximately 48 kDa is visible in the insoluble protein fraction of *E. coli* BL21(DE3) pETbendAB induced with 50 or 200 µM IPTG (black arrow), however not in uninduced cultures or controls. Theoretical mass of His<sup>6</sup>BendB and StreplII-BendA is 46.98 kDa and 47.412 kDa, respectively. A protein of similar size cannot be resolved in the soluble fraction by SDS-PAGE. **b**, Western blot of **a**, using anti-polyhistidine primary antibodies, confirms the recombinant protein to be His<sup>6</sup>BendB. His<sup>6</sup>mCherry (~ 29 kDa) and StreplIIPluA (46.96 kDa) were used as positive controls for Western blotting indicated by red and blue arrows respectively.

Figure 7.26: Purification of BenB from *E. coli* BL21(DE3) soluble cell lysate

**a**, IMAC purification of His<sup>6</sup>BendB from *E. coli* pETBendAB soluble cell lysate. Recombinant His<sup>6</sup>BendB cannot be resolved in any elution fractions (EF, 50-400 mM imidazole), the flow through (FT) or soluble fraction (SF) by SDS-PAGE. **b**, Western blot of **a**, using anti-strep-II primary antibodies. StreplIIPluA was used as a positive control (PC, blue arrow). **c**, Western blot of **a** using anti-polyhistidine primary antibodies.
showing His$_6$BendB to be present within the soluble fraction, flow through and elute from the IMAC column upon addition of 200 mM imidazole to elution buffer A.

The absence of StreplI-BendA could be explained by inefficient transcriptional initiation, degradation of the corresponding mRNA, or targeted proteolysis. Kudla et. al. (2009) implicate the mRNA secondary structure of nucleotides spanning -4 to +37 from the translational start to play an important role in translational initiation efficiency$^{139}$. This mRNA region comprises the 30S ribosome footprint$^{140}$, and is only accommodated into the ribosome entry channel if single stranded, requiring linearization before translation. 5’ mRNA secondary structure was accounted for during construct design, and to normalise translational efficiency refactored ketosynthase sequences were tagged with the same N-terminal strepII tag sequence acquired from pET52b (Novagen), a commonly used high expression level vector which spans the ribosome footprint. Wider analysis of the mRNA secondary structures from all duel expression vectors reported here-in shows no dissimilarities in translational initiation rate of N-terminal StreplI tag containing gene sequences, as calculated by the RBS calculator$^{87}$.

The precise mechanism underpinning the absence of recombinant StreplI-BendA in E. coli BL21(DE3) remains unknown. To circumvent factors associated with translational initiation we attempted to introduce a linker between his$_6$BendA and bendB as described in Chapter 6, however Gibson and In Fusion DNA assembly failed after multiple attempts.

7.5.3.5 Candidatus Desulfofervidus auxilii

Candidatus Desulfofervidus auxilii is a sulphate-reducing bacterium (SRB) which exists in a syntrophic relationship with an anaerobic methane-oxidising archaea (ANME), responsible for microbial anaerobic oxidation of methane in ocean sediments$^{141}$. ANME archaea are typically divided into three clades (ANME1-3)$^{142}$, each of which forms a syntrophic relationship with only one SRB partner$^{143}$. The role of SRB can be fulfilled by a considerably diverse range of Deltaaproteobacteria, however. ANME-1 and 2 are routinely associated with psychro- and mesophillic Desulfocarcinia spp. and Desulfococcus spp., whilst ANME-3 more commonly partner with Desulfobulbus spp.$^{144}$, all of which are of the Desulfobacterales Order$^{141,143}$. Very little is known about SRBs which live in consortia with ANMEs and purification to monoculture has not been achieved, suggesting these organisms live in obligate syntrophy$^{141}$. Interestingly, a putative type II polyketide synthase cluster was identified in C. D. auxilii, a thermophilic ANME-3 SRB partner, by pairwise similarity to AntE (Figure 7.27).

C. D. auxilii was not included within the 2552 genomes analysed by our brute force approach, however 4 organisms from 4 different genera of the Order Desulfovibionales were, none of which contained a putative type II PKS. Interestingly, aromatic polyketides are not common specialised metabolites within Desulfobacterales spp.: an exhaustive search of all 75 genomes of this order resulted in identification of only one additional partial cluster in Desulfobacterium sp. 4572_20 isolated from hydrothermal vents in the Guaymas Basin, Gulf of California (Figure 7.27), identified previously by AntD AA pairwise similarity (Table 7.1). Interestingly, this partial cluster comprises a canonical KS, and a truncated CLF comprising just 127 AA, which is
translationally overlapped by an ORF putatively described as a fatty acid synthase (FAS) type KS (Figure 7.27). The truncated CLF shares similarity to the N-terminal region of *Streptomyces sulphurous* CLF, whilst the overlapping FAS ORF shares similarity with the C-terminal regions of FabF from *Parageobacillus* genomosp. 1. A concatenation of the translated CLF and FAS ORFs from *Desulfobacterium* sp. 4572_20 shows excellent similarity to the N-terminal and C-terminal regions of the C. *D. auxilii* CLF (Figure 7.28), however 48 residues between these termini are missing. The large CLF deletion posits the *Desulfobacterium* sp. 4572_20 cluster to be a degenerated evolutionary relic of the C. *D. auxilii* BGC, and likely to be non-functional. These findings show a divide in the evolution of N- and C-terminal CLF domains where the C. *D. auxilii* CLF N-terminal region has specialised for polyketide biosynthesis whilst the C-terminal is more similar to FAS constituent parts.

As with the *bend* cluster, defining the C. *D. auxilii* BGC (termed the *dau* BGC) boundaries through identification of conserved flanking genes in organisms lacking the cluster was not possible, and the minimal cluster boundaries were defined manually. Neighbourhood COG searches did identify a conserved core BGC within the marine sub-sediment Chloroflexi *Dehalococcoidia* Bacterium SCGC AG-205-I10, AG-205-N23, AG-205-C05, AG-205-C16, and AG-205-G20145 and *Deltaproteobacteria* bacterium RGB_16_54_11 however (Figure 7.27), the latter of which was identified as an AntD and E homologue (Table 7.1).

The core *dau* BGC comprises a minimal PKS (7-9), (*dauA*: KS, *daub*: CLF and *dauC*: ACP), a cyclase (10), aromatase (11) and ketoreductase (12) but lacks a cluster specific PPTase or MCAT. The mPKS and primary tailoring enzymes are within an operon also encoding a flavin dependant amine oxidase, NADPH dependant FMN reductase, radical SAM YgiQ gene sequence, aldehyde:ferrodoxin oxidoreductase, ketopantoate *panE*, hypothetical gene and an aminocarboxymuconate-semialdehyde decarboxylase (1-6, 13) (Supplementary Table 7.10). Mostly, these gene products have plausible specialised metabolic functions and experimental analysis is necessary to further elucidate the cluster boundaries. Whilst no obvious specialised metabolite resistance mechanism is present within the *dau* BGC, the Antibiotic Resistance Target Seeker (ARTS)\textsuperscript{101} BGC prioritisation pipeline implicates the flavin dependent amine oxidase (1) as a putative tetracycline deuctrastase. This novel class of tetracycline inactivating enzymes is prevalent in environmental metagenomes opposed to tetracycline resistant clinical isolates, however the C. *D. auxilii* cluster associated flavin dependent amine oxidase shares limited sequence similarity (31% identity, 34% coverage) with any characterised tetracycline deuctrastase\textsuperscript{102}, and is not present in the highly similar *Dehalococcoidia* Bacterium BGC, questioning its applicability in resistance. The gatekeeper profile of DauA and B shows similarities to RemA/B indicating the mPKS to produce a C\textsubscript{20} polyketide, and the absence any initiation module components, typically comprising additional ACPs and a KSIII, a type I PKS module, or a dedicated CoA ligase/synthetase suggests the polyketide to be primed with a malonyl-CoA or acetyl-CoA start unit.

The complement of biosynthetic gene clusters within *Desulfobacterales* spp. is poor by comparison to the biosynthetically elastic *Actinobacteria* or *Burkholderia*: each genome
analysed contained an average of 3.6 BGCs, as computed by antiSMASH 4.0 without the clusterfinder algorithm. The C. D. auxilli type II PKS represents an exciting and rare opportunity to test the solubility of a KS/CLF from a thermophilic organism.

Open reading frames (ORFs) corresponding to minimal PKS components, KS (7), CLF (8) and ACP (9), are shown in light blue, and ORFs for primary tailoring enzymes are red. Additional ORF’s within the same operon as the mPKS are numbered but white. ORFs shared between Dehalococcoidia bacterium SCGC, C. D.auxilli and Desulfobacterium sp. 4572_20 are defined by a light grey box. A minimal ‘core’ BGC is highlighted in a dark grey box. The Desulfobacterium sp. 4572_20 deletion is represented by a black dotted line. The C. D. auxilli cluster is flanked by transposable elements detailed in purple. Ca. D. auxilli predicted ORF gene products are detailed in Supplementary Table 7.10.
Figure 7.28: Amino acid sequence alignment of concatenated CLF and FAS parts from *Desulfobacterium* sp. 4572_20

Concatenation of the two ‘CLF’ *Desulfobacterium* sp. 4572_20 ORFs shows high similarity to the C. *D.* auxilii CLF. Both putative CLF sequences are aligned against FabF (PDB: 2GFW). *E. coli* FabF secondary structure is displayed above the sequence alignment. FabF catalytic cysteine is denoted by a black star. Amino acids proximal to the FabF active site are shown to be conserved between phylogenetically diverse species146, but are not conserved within either CLF sequence.

### 7.5.3.5.1 Evaluating the solubility of DauAB in *E. coli* BL21(DE3)

Like *BendA* and *B*, the solubility of the putative *dau* KC/CLF cannot be evaluated here. His6*DauB* is observable as an insoluble recombinant protein in *E. coli* BL21(DE3) expressing *dubB/dauA*, however, its StrepII tagged heterodimeric counterpart, StrepII-DauA, is not observable (Figure 7.29a). Both proteins have similar theoretical molecular masses which prevents reliable identification via SDS-PAGE however Western blots using both anti-
polyhistidine and anti-StrepII primary antibodies corroborate the identification of His$^6$DauB (Figure 7.29b) and absence of StreplIDauA (data not shown). His$^6$DauB is poorly soluble within the *E. coli* recombinant proteomes, under expression conditions described here, (Figure 7.30a and c) and could only be detected after purification. His$^6$DauB did not co-purify with StreplIDauA, however (Figure 7.30b), unlike AntDE, SspAB and PluAB. This is to be expected should the corresponding *strepII-dauA* gene sequence not express. As with *bendAB*, attempts to create a *dauAB* fusion gene failed repeatedly.

**Figure 7.29: Evaluation of DauAB solubility in *E. coli* BL21(DE3)**

a, A Ca. *D. auxilii* mPKS component is visible as insoluble recombinant protein from *E. coli* BL21(DE3) pETDauAB induced with either 50 or 200 µM IPTG, but not from uninduced cultures (black arrow). No corresponding recombinant protein is visible in the insoluble fractions of empty vector controls, *E. coli* BL21(DE3) (pETM11b), or no plasmid controls (NPC: *E. coli* BL21). No recombinant proteins corresponding to His$^6$DauB nor StreplIDauA (45.25 and 47.1 kDa, respectively) can be visualised in the soluble protein from soluble protein cell lysate. Gene expression and induction conditions are described in 7.4.3. b, Western blot of a, using anti-polyhistidine primary antibodies. His$^6$DauB is observable in *E. coli* BL21(DE3) expressing *dauA* and *dauB* only. PC: His$^6$mCherry anti-polyhistidine antibody positive control also denoted by a red arrow).
Figure 7.30: Purification of DauB from E. coli BL21(DE3) soluble cell lysate

a, IMAC purification of His$_6^DauB$ (45.25 kDa) from E. coli BL21(DE3) pETDauAB, 200 mM IPTG, after 16 h incubation at 16°C. His$_6^DauB$ elutes from the IMAC column in buffer A containing 200 and 400 mM imidazole (black arrow), however cannot be resolved from the soluble fraction (SF) or flow through (FT).

b, Western blot of a, using anti-strep-II primary antibodies. StrepII-DauA (47.1 kDa) is not identified in any protein fraction. StrepII-PluA is used as a positive control (blue arrow).

c, Western blot of a, using anti-polyhistidine primary antibodies. His$_6^m$Cherry is used as a positive control (PC, red arrow). His$_6^DauB$ is visible in eluent containing 200 and 400 mM imidazole. Protein molecular weight ladders are shown on all gels and detailed by numerical ticks (kDa).

7.5.3.6 Ktedonobacter racemifer DSM44963 (SOSP1-21 – type strain)

*Ktedonobacter racemifer* is a member of a deep branching prokaryotic lineage tentatively assigned to the Chloroflexi phylum$^{147}$, first isolated from soil in Gerenzano, Northern Italy$^{148}$. These organisms show Actinobacterial cell morphologies and grow as mesophilic and thermophilic aerobic, non-motile, filamentous vegetative and aerial hyphae and form multiple grapelike spores per cell in a budding fashion$^{147}$. The genome of *K. racemifer* DSM 44963 was sequenced as part of the GEBA initiative and comprises 13.7 Mb with 11,453 putative proteins and G+C content of 53.8%$^{147}$. To-date this is the only complete genome from the genus *Ktedonobacter* publicly available. Metagenomic and environmental DNA profiling has highlighted the prominence of Ktedonobacteria in extreme environments including oligotrophic and copiotrophic Antarctic, volcanic and cave ecosystems, but also identified the ubiquity of Ktedonobacteria in mesophilic terrestrial at comparatively low relative abundance. One might assume Ktedonobacteria to prefer oligotrophic environments however the highest diversity of different *Ktedonobacterial* species was found in copiotrophic lower-layer forest soils and the lowest diversity in sands$^{149}$.

Unremarkably, *Ktedonobacter racemifer* DSM 44963 was not included in our RefSeq genome dataset. Rather, one putative type II PKS BGC (kra3) was identified by the AA pairwise similarity method. Further analysis of the *K. racemifer* DSM 44963 genome for additional BGCs elucidated two additional putative type II PKS BGCs (kra1 and kra2) however kra1 was
disregarded after manual curation because it lacked an ACP (Figure 7.31a) (Supplementary Table 7.11). Observation of more than one type II PKS BGC within a genome is relatively frequent in Actinobacteria, in such instances one BGC is typically responsible for the biosynthesis of a spore pigment\textsuperscript{131}, which may also be the case for Ktedonobacteria shown to produce ‘cream to pinkish orange’\textsuperscript{150} pigments when grown on solid media. To assess if type II PKSs are common place in other Ktedonobacteria a further 41 draft/scaffold genomes from Ktedonobacterales and Thermogemmatisporales Orders were analysed using antiSMASH 4.0, leading to identification \textit{kra1} and \textit{kra2} BGC homologues in the thermophile \textit{Thermosporothrix hazakensis}\textsuperscript{150} (Figure 7.31). Genomes of 6 \textit{K. racemifer} spp. were included in the analysis however each genome comprised hundreds of short contigs confounding identification of BGCs. Despite poor draft genome sequences, type II PKS BGC fragments were commonly identified, suggesting this lineage to be a reservoir of type II PKS BGCs and highlights the importance of genome quality when identifying specialised metabolite BGCs using commonly available genome mining tools.
Figure 7.31: Type II PKS BGCs from *Ktedonobacter racemifer* DSM 44963

**a**, Schematic of predicted type II PKS BGCs from *Ktedonobacter racemifer* and *Thermosporothrix hazakensis*. Clusters 1, 2 and 3 correspond to *kra1*, *kra2* and 3. All ORFs predicted to play a role in specialised metabolite biosynthesis, regulation, resistance and efflux are defined as putative cluster. ORF colour designations for cluster 3 and *rem* are as described in Figure 7.23. Colouring of ORFs in Cluster 1 and 2 are consistent with the above however all genes encoding putative tailoring enzymes are shown in red. Cluster 3 (*kra3*) is only identified in *K. racemifer*. **b**, comparison of *kra3* and *rem* BGC gene complements. Dotted lines show movement and/or loss or gain of BGC fragments. All gene designations are described in Supplementary Table 7.11 – 7.13.

The second type II BGC (*kra2*) comprises a minimal PKS (cluster 2, ORF3-5) which follows the dogmatic operon architecture present in the *act* BGC as well as two polyketide cyclases (6, 7), an antibiotic biosynthesis monooxygenase (8), O-methyltransferase (2) an acyl esterase (9) and 284
a MFS transporter (1) (Supplementary Table 7.12). No C9-ketoreductase is predicted. Intriguing, a C-methyltransferase, HrdB sigma factor and tetracycline resistance ribosomal protection protein (orange ORF) which may play a role in activation of aromatic polyketide biosynthesis, methylation and resistance, are separated from the putative cluster by two IS3 and one IS6 family transposases. Additionally, kra2 comprises an acyl esterase, enzymes with this function are rarely identified within type II PKSs BGCs. The acyl esterase may perform a thioesterase chain termination function, or alternatively police selection of non-malonyl- or acetyl- starter units priming biosynthesis, selectively hydrolysing more common substrates e.g. malonyl-CoA from the cluster associated ACP during initiation of biosynthesis. A similar mechanism polices benzyl-ACP and alkylacyl-ACP loading during enterocin\textsuperscript{151} and R1128\textsuperscript{152} biosynthesis, respectively.

Gatekeeper residues within KraA2 and KraB2 match RemA and B with the exception of a single AA substitution, A194G. Substitution of glycine for alanine should play no role in occlusion of the amphipathic tunnel hence the minimal PKS is predicted to produce a decaketide. Q169, implicated in decarboxylation of malonyl-ACP starter units during act biosynthesis is substituted for an arginine in Kra2B, consistent with RemB, and its function as a CLF, opposed to duplication of an active KS, where a cysteine would be expected. For the remainder of this thesis only the unique BGC, kra3 (Figure 7.31a, cluster 3), is considered and referred to as the kra cluster subsequently, for clarity.

Unlike kra2, the kra (Figure 7.31, cluster 3) mPKS operon architecture does not mirror the act BGC, but rather that of the rem and bend BGCs. The KS (KraA, 7) and CLF (KraB, 8) gene sequences are adjacent to one another, however are separated from the putative ACP gene sequence (KraC, 10) by an ORF encoding a putative ACPs type PPTase (9). Furthermore, the mPKS and PPTase is flanked either side by a methyltransferase (6) and RemI-like cyclase (11) which collectively form a 6 ORF operon common to kra, bend and rem BGCs (Figure 7.23, Figure 7.31b). The kra cluster also comprises additional gene sequences putatively assigned as RemL and RemF like cyclases (5 and 4), a Snoa-like cyclase (1), a methyltransferase (2), a putative 2-polyprenyl-6-methoxyphenol hydroxylase (3) and two ABC transporter components (12, 13) (Supplementary Table 7.13). The gatekeeper residue profile of KraAB shows similarities to RemAB (Table 7.2), predicting the kra BGC to produce a malonyl- or acetyl-primerd C\textsubscript{20} polyketide, which most likely forms an S-shaped discoid polyketide like resistomycin or an anthanthrene like structure, should the additional SnoaL-like cyclase facilitate closure of an additional ring\textsuperscript{153}. The kra cluster is flanked upstream by an additional ACP (1'), ACPs family PPTase (2'), acyl-CoA dehydrogenase (3') and saccharide associated genes (Supplementary Table 7.13), not shown on the schematic, however, the two BGCs are separated by 16s rRNA and 23s rRNA gene sequences, and two transposases questioning their role in biosynthesis.

7.5.3.6.1 Evaluating the solubility of kraAB in E. coli BL21(DE3)

Expression of kraA and kraB from the duel expression vector pETKraAB led to formation of soluble heterodimeric recombinant KraAB complexes in E. coli BL21(DE3). Recombinant proteins corresponding to the theoretical mass of His\textsuperscript{6}KraA or/and StreplI-KraB are visible in
both the soluble and insoluble fractions of *E. coli* BL21(DE3) pETKraAB induced with 50, 200 or 500 µM IPTG, but not total cell lysate from uninduced cultures, *E. coli* BL21(DE3) (NPC: no plasmid control) or *E. coli* BL21(DE3) pETM11b-AntB, backbone control (Figure 7.32a and b). His\textsuperscript{6}KraA and StrepII-KraB have similar theoretical masses (47 and 46.8 kDa, respectively) and cannot be resolved from one another via 1D SDS-PAGE. Western blotting using antibodies specific for each epitope tag (His\textsuperscript{6} or Strep-II) resolved both proteins from the soluble cell lysate of *E. coli* BL21(DE3) expressing both *kraA* and *kraB* (Figure 7.32c and d), however expression of *kraB* alone resulted in 100% of recombinant protein forming inclusion bodies (Figure 7.33). Interestingly, low expression of *kraA/B* caused by poor repression of chromosomal T7 RNAP and T7 promoter sites within the duel expression vector resulted in recombinant protein being predominantly soluble (Figure 7.32c, pETKraAB, 0 µM IPTG).

As with SspAB, KraB is only observed as a soluble recombinant protein upon co-expression of its cognate ketosynthase. This observation is consistent with the dogma that KS/CLFs form a heterodimeric complex which aids solvation by complementary shielding of the hydrophobic face covering more than 20% of each monomer\textsuperscript{83}. Despite this StrepII-KraA and His\textsuperscript{6}KraB do not co-purify, as observed for AntDE, SspAB and PluAB (Figure 7.34), suggesting the interactions between KraA and KraB to be transient or weak. Interestingly, His\textsuperscript{6}KraB eluted from nickel-NTA agarose upon addition of a relatively low concentration of imidazole (buffer A + 20 mM) indicating the hexahistidine tag to form poor interactions with the Ni-resin. Early elution may be explained by sequestration of the his\textsuperscript{6} tag by KraB. Furthermore, sequestration of the his\textsuperscript{6} tag may sterically inhibit efficient KS/CLF heterodimer formation in turn preventing co-purification of his\textsuperscript{6}-KraB and StrepII-KraA by IMAC. This observation is also consistent with the elution profile of PluAB (Figure 7.22). Cleavage of the N-terminal His\textsuperscript{6} tag from KraB followed by protein purification using *strep* Tactin sepharose resin would validate or refute this hypothesis.
Figure 7.32: Evaluation of KraAB solubility in *E. coli* BL21(DE3)

Separation and visualisation of soluble (a) and insoluble (b) cell lysate extracted from *E. coli* BL21(DE3) expressing *kraA* and *kraB* by SDS-PAGE. *E. coli* BL21(DE3) was used as host background control (NPC: no plasmid control), and *E. coli* BL21(DE3) pETM11b and pETM11b-AntB as vector controls. Recombinant protein is visible in both soluble and insoluble fractions of *E. coli* BL21(DE3) expressing *kraA* and *kraB*; however, the close molecular weight of His$^6$KraB and StreplI-KraA (47 and 46.8 kDa) mires identification. Both His$^6$KraB and StreplI-KraA are resolved independently by Western blots using anti-polyhistidine primary antibodies (c), and anti-strep-II primary antibodies (d) (black arrows). His$^6$mCherry and StreplII-PluA are used as positive controls, respectively.
Figure 7.33: Assessment of KraB solubility in E. coli BL21(DE3)

Western blot of soluble and insoluble protein cell lysate from E. coli BL21(DE3) expressing his^6 kraB. His^6 KraB is only detectable as insoluble protein. All controls are as described in Figure 7.33c. The black circle does not correspond to signal produced by the secondary antibody, but rather that of the pre-stained molecular weight reference ladder.

Figure 7.34: Purification of KraB from E. coli BL21(DE3) cell soluble lysate
SDS-PAGE (a) and corresponding Western blot (b) of His$^5$KraB purification from E. coli BL21(DE3) pETKraAB soluble cell lysate. a. His$^5$KraB is visible in fraction of protein eluting upon addition of 20 mM imidazole to buffer A. b. StrepII-KraA could not be detectable in the protein extracted from E. coli BL21(DE3) pETKraAB by Western blot. Anti-strep-II primary antibodies were used: StrepII-PluA was used as positive control. Protein molecular weight is characterised using a standard protein ladder (L). Elution fractions (EF) contain increasing concentrations of imidazole (20, 50 and 200 mM). Anti-Strep-II antibody positive control (PC) is strepII-PluA.

Alternatively, the substitution of gatekeeper residues within KraAB may play a role in destabilisation of the KS/CLF complex, as discussed for PluAB (7.5.3.3.2): both PluAB and KraAB have similar gatekeeper profiles to RemAB (Table 7.2). Furthermore, of our dataset KraB is the most phylogenetically proximal CLF to RemB and the kra BGC comprises RemI, L and F cyclases homologues shown to be necessary for biosynthesis of the C$_{20}$ polyketide chain during resistomycin biosynthesis$^{130}$. This hypothesis remains unsupported, and requires systematic co-expression of mPKS components and cluster associated cyclases.

7.5.4 Non-canonical Actinobacterial KS/CLF sequences

The resistomycin BGC (rem) from S. resistomycificus is curious. Phylogenetic analysis of non-Actinobacterial KS and CLF candidates identified here, and more than 50 previously characterised KS/CLFs from Actinobacteria shows formation of discrete Actinobacterial and non-Actinobacterial clades (Figure 7.11); however, this nomenclature is not strictly accurate. Both mPKS components of the rem BGC localise to non-Actinobacterial clades opposed to the respective Actinobacterial ones (Figure 7.11). Accordingly, the rem BGC is derived from the non-Actinobacterial type II PKS lineage opposed to the canonical Actinobacterial type II PKS. The rem cluster, therefore, presents an opportunity to identify if the causation of Actinobacterial KS/CLF insolubility arises in part from host specific evolutionary pressures$^{154}$, or from unknown Actinobacterial type II PKS specific structural and sequence based factors.

7.5.4.1 RemAB solubility in E. coli BL21

Co-expression of remA and B from the duel expression vector pETRemAB in E. coli BL21(DE3) predominantly led to formation of inclusion bodies (Figure 7.35a). His$^5$RemB and StrepII-RemB cannot be reliably resolved from the corresponding soluble protein fraction by SDS-PAGE, however, His$^5$RemB can be observed by Western blotting with anti-polyhistidine primary antibodies (Figure 7.35b). Curiously, a corresponding Western blot using anti-strep-II primary antibodies only resolved StrepII-RemA as insoluble recombinant protein (Figure 7.35c). Therefore, either RemB has some solubility without its cognate KS partner, or the difference in binding constant between both primary antibodies means the concentration of soluble recombinant StrepII-RemA falls below the detection limit for the corresponding primary antibody. The latter proposal is supported by the vibrant difference in signal detected from insoluble His$^5$RemB and StrepII-RemA Western blots despite both bands being visible by SDS-PAGE and protein concentrations normalised throughout the analysis. To determine if the solubility of RemB depends upon co-expression with remA an additional vector was constructed lacking remA. Expression of the ΔremA expression vector (pETRemB) resulted in RemB
forming 100% inclusion bodies: Western blotting using anti-polyhistidine primary antibodies could not resolve His\(^6\)RemB as a soluble recombinant protein (Figure 7.35d). This analysis was repeated three times using increasing concentrations of protein each time, however returned the same result.

We can infer that both RemA and RemB are soluble proteins within \textit{E. coli} (Figure 7.35). Furthermore, the solubility of one monomer depends on the co-expression of its cognate mPKS partner, suggesting the KS and CLF to form a heterodimer aiding protein solvation, consistent with the currently accepted dogma for type II PKS structures.

Figure 7.35: Evaluation RemAB solubility in \textit{E. coli} BL21(DE3)

\textbf{a}, Expression analysis of \textit{remA} and \textit{remB} coexpression in \textit{E. coli} BL21.. \textit{E. coli} BL21(DE3) (NPC: no plasmid control) and \textit{E. coli} BL21(DE3) pETM11b are host and empty vector controls. Gene expression was induced in \textit{E. coli} BL21(DE3) pETRemAB with 0, 50 or 200 µM IPTG at OD\(_{600}\) 0.5 prior to incubation at 16°C for 16 h. All controls were also induced accordingly. Recombinant His\(^6\)RemB and StrepII-RemA (44.9 and 45.93 kDa, respectively) can only be resolved from \textit{E. coli} BL21(DE3) expressing \textit{remA/B} insoluble cell lysate by SDS-PAGE. \textbf{b} and \textbf{c}, Western blots of SDS-PAGE \textbf{a} using anti-polyhistidine and anti-strep-II antibodies respectively. His\(^6\)mCherry (≈29 kDa) and StrepII-PluA (46.96 kDa) are used as
positive controls. His^6^RemB is visible as both soluble and insoluble recombinant protein from *E. coli* BL21(DE3) pETRemAB total cell lysate, exclusively. StrepII-RemA can only be resolved from *E. coli* BL21(DE3) insoluble cell lysate. d, Expression analysis of remB in *E. coli* BL21, controls are as described in a and b. In the absence of StrepII-RemA, His^6^RemB forms 100% inclusion bodies.

The G+C content of the host species may play a role in KS/CLF solubility. *Streptomyces* spp. have characteristically high G+C content genomes and resultantlly almost all codon 'wobble position bases', are either G or C. Bias in codon usage is a driver of protein evolution, and is correlated directly to CDS G+C content. For instance, coding sequences from characteristically low G+C content genomes are enriched in AT-rich codons, and *vise-versa*. The degeneracy of the genetic code does not exclusively account for this G+C bias, and 9 amino acids are predominantly encoded by A+T or G+C rich codons. Accordingly, low G+C CDS sequences are enriched in amino acids predominantly encoded by A+T rich codons: tyrosine, phenylalanine, asparagine, isoleucine and lysine amino acids, and in high G+C CDSs alanine, glycine, proline and arginine are overrepresented. Differential utility of amino acids between hosts will undoubtedly alter the physiochemical properties of proteins and protein folding and stability of homologous. Substitution of amino acids with similar biochemical characteristics but dissimilar average codon G+C contents are observable between *ant*, *act*, kinamycin (*kin*) and ovidomycin (*ovm*) KSs, which have CDS G+C contents of between 43% and 72% (Figure 7.36). The substitution of arginine for lysine and *vise-versa* has recently been demonstrated to play a role in recombinant protein solubility in *E. coli*, whereby proteins enriched in arginine, by comparison to lysine, are more likely to be insoluble. In support of this arginine residues have been found to increase the ‘stickiness’ of recombinant proteins in the *E. coli* cytosol. Addition of a contiguous string of more than 3 solvent exposed arginine residues to GB-1, an arginine poor soluble *Streptococcal* recombinant protein, resulted in protein macromolecule assembly, while untagged protein remained soluble in *E. coli*.

The effect of R->K substitution may play a role in KS/CLF solubility, however was not rigorously tested here. Warwicker et al., (2014) suggest recombinant proteins with a K:R threshold below 1.2 have a higher propensity to form inclusion bodies. The K:R ratio of the resistomycin KS/CLF is considerably below that threshold at 0.24 K/R and, expression of the ovm KS/CLF, which has the highest K:R ratio for any actinobacterial KS/CLF (~0.3), exceeding that of RemA/B, in *E. coli* resulted in both OvmP and K (KS and CLF, respectively) forming 100% inclusion bodies, even in the presence of elevated concentrations of cytosolic molecular chaperons (data not shown). The cause of insolubility of Actinobacterial KS/CLFs is likely multifactorial. In spite of the amino acid ‘drift’ of the rem BGC, presumably resultant from the higher GC content of the host genome, this KS/CLF remains soluble indicating the solubility of ‘non-Actinobacterial’ KS/CLFs results from an unknown sequence-structure-folding factor, and not host specific protein evolution.
Oviedomycin NGT AFWDLLSAG R TAT R SISSFFDASPFR R IAGEVDFAAABFGFR REIMRRTATQFA
Kinamycin NGT HFWSSLSEK R TAT R GIITFFDPTEK R KVWAIDFDPYEHGLSPQERVAAEAGQFA
Actinorhodin NGT QFWEWLTGK R TAT R ISFFDPSYPK R SQVAEAADFDPVAGFGFRELINNASQFA
Anthraquinone VGVNEFWNNIHNGSGVYKEKGERK R FGFGAIGQVYGSDDNNERFVLSEHERYLQFA

Figure 7.36: Substitution of arginine and lysine codons in low and high G+C content Actinobacterial ketosynthases

Amino acid alignment of a section of ovm, kin, act and ant KSs. Arginine residues appear to be functionally replaced with lysine residues in the ovm, kin, and ant KSs with respect to the act KS

7.6 Discussion

7.6.1 Refining type II PKS discovery in antiSMASH

Our understanding of the distribution of polyketide synthase containing BGCs is often mired by poor reporting of data i.e. only the specialised metabolite class is reported and the sub-class is omitted. This is the case for the recently described systematic analysis of the BGC complement of the human microbiome\(^2\)\(^1\), the cataloguing of 3,339 PKS and NRPS BGCs from 991 organisms\(^6\)\(^6\) and a paralleling global analysis of ‘secondary metabolite’ BGCs from prokaryotes (as of 2014)\(^4\). The latter provides sub-class details in the form of raw supplementary data, however. Interestingly, this raw supplementary data reports elucidation of 31 type II PKS BGCs from the 1,154 genomes analysed, 15 of which are derived from non-Actinobacterial origins\(^4\): 14 non-Actinobacterial hosts were Yersinia spp. and the other was Bartonella tribocorum.

Unremarkably, all predicted type II PKS BGCs from non-Actinobacterial origins were false positives following the rules described in 7.4.2.2. The exceptionally high false positive rates of type II PKS containing BGCs from non-Actinobacteria in this dataset was paralleled by our brute force discovery approach, where predicted type II PKS BGCs were also identified from Bartonella spp. (Figure 7.8), and combined, both datasets provide an important commentary on the accuracy of antiSMASH detection methods for type II PKSs\(^8\).

AntiSMASH identifies putative BGCs by searching for profile Hidden Markov Models (pHMMs) of characteristic core proteins within query translated gene sequences. These hits are listed and scored according to a manually defined similarity threshold for each protein\(^8\). pHMMs for a particular class of biosynthetic enzyme are built from seed datasets of characterised proteins and for the type II PKSs one must assume the KS and CLF pHMM seed sequences were almost exclusively from Actinobacterial as a result of their rarity outside of this taxon. For positive identification of a type II PKS BGC a KS or CLF pHMM hit scores must exceed 50, a manually curated threshold, and the KS or CLF pHMM score must exceed scores for enediyne, modular, iterative, hybrid, trans-AT KSs pHMMs, and also pHMM models built to detect false positives: bacterial type I FASs, type II FASs, HglD&E, fungal type I FASs and FabH pHMMs\(^8\). Such a detection method is problematic for three reasons. Firstly, the pHMM models may not

292
accommodate the sequence diversity of non-Actinobacterial KS and CLFs. Secondly, only detection of the KS or CLF pHMM is necessary to define a BGC as type II PKS containing, therefore BGCs comprising non-modular or monodomain ketosynthases including single FabF paralogous will be detected. Thirdly, BGCs lacking ketoreductases, cyclases and cyclase/aromatases, which are unlikely to form the aromatic structures representative of type II PKSs and are more likely to form saturated fatty acids as in bacterial fatty acid biosynthesis, will be identified. For instance, the predicted type II PKS BGC from *Bartonella* spp. (Figure 7.8) encodes the biosynthetic complement necessary for biosynthesis and transfer of 27-hydroxyoctacosanoic acid, or very long chain fatty acids (VLCFA) to lipid A, rather than a pericyclic / aromatic polyketide core structure. To improve the accuracy of type II PKS containing BGC prediction using antiSMASH we suggest the construction of additional pHMMs using seed data which include the characterised non-actinobacterial KS/CLF sequences described here. Furthermore, to improve confidence in BGC type we propose the construction of pHMMs for primary tailoring enzymes including cyclases, cyclase/aromatases and C9 ketoreductases which must be detected within a modest stretch of DNA flanking KS and CLF pHMM hits. Understandably antiSMASH authors wanted an inclusive, rather than exclusive, approach to genome mining, however, using more refined and specific selection criteria would allow assignment of a confidence score where BGCs detected using the published and currently implemented criteria would be weighted lower than candidates which satisfy more specific type II PKS criteria.

Similar specific identification criteria for multiple type II polyketide biosynthetic components are already implemented by two other BGC mining tools: PKMiner and PRediction Informatics for Secondary Metabolomes 3 (PRISM3). The predominant focus for PKMiner is as a repository for type II PKS BGCs and accordingly has not been widely adopted for BGC mining. Furthermore, PKMiner cannot be used as a standalone platform and suffers from a laborious input procedure. PRISM3 is a plausible alternative to antiSMASH, however also suffers from a laborious input procedure and should be considered as a polish step for improved structural predictions. PRISM requires both HMMs for KS and CLF hits for type II PKS BGC detection and a further suite of class specific HMMs for type II PKS associated cyclases, cyclase/aromatases, ketoreductases, C-glycosyltransferases, methyltransferases and oxygenases enables more reliable calling of type II PKS chemotypes, and the degree of subsequent tailoring of these core scaffolds. Unfortunately, we found this software to suffers from errors and could not be run in parallel here.

In accordance with previously published studies we detail important considerations to improve type II class specific BGC predictions using antiSMASH. Such improvements will be paramount as we begin to explore the non-Actinobacterial type II PKS lineage described here-in (Figure 7.11) and improve the accuracy of future systematic BGC analysis.

### 7.6.2 BGC discovery by brute force

Per mega base of DNA the BGC discovery frequency from our 2552 strong dataset of complete genomes sat between two previously reported values of 2.4 BGC per Mb, using the Cluster
Finder (CF) algorithm\(^4\), and 6.4, without the CF algorithm\(^5\). The latter calculation was derived from a dataset of 1,003 genomes, 396 of which were from first in Genus reference species and just 157 genomes were Actinobacteria highlighting the biosynthetic potential of phylogenetically diverse organisms outside of the Actinobacterial Phylum\(^5\).

Despite identifying over 13,000 predicted BGCs using the brute force method the frequency of unique and undiscovered type II PKS BGC from non-Actinobacterial hosts was disappointing and only a single predicted BGC conformed to the specific identification criteria described here (7.4.2.2). Using a brute force approach for specialised metabolite discovery is clearly applicable for identification of common specialised metabolite BGCs e.g. NRPS, terpene, bacteriocins, type I PKS and sideroregulated containing BGCs, where > 500 predicted BGCs were identified (Figure 7.7), however it is more prudent to use a targeted approach for identification of rarer and more specialised metabolite classes. The dataset described is undoubtedly useful however, and can be used to evaluate and hone the identification capacity of antiSMASH for rarer and less characterised specialise metabolite BGCs. For instance, the cyanobactin class of specialised metabolites are thought to be limited to Cyanobacteria\(^163\) however our analysis identified predicted cyanobactin BGCs in genomes of non-cyanobacteria including Streptomyces venezuelae ATCC 15439, Streptomyces reticuli TUE45, Shewanella piezotolerans WP3, 4 Dickeya spp., Rhodovulum sulfidophilum FACHB-1757 and 2 Yersinia enterocolitica subspps.. BGCs deposited in the antiSMASH database corroborate our findings, also listing cyanobactin BGCs from multiple non-cyanobacteria including\(^164\). One could posit either cyanobactin like BGCs are not phylogenetically delimited or the cyanobactin identification criteria utilised by antiSMASH 3.0 and antiSMASH 4.0 requires updating. It remains to be seen if these BGCs constitute false positives or true cyanobactins.

The rarity of type II PKSs drastically reduces identification frequencies when using fully sequenced genome datasets. This pool of DNA is relatively small by comparison to draft genomes, for instance 6,452 finished genomes are deposited within the IMG database (2017) compared with 48,726 draft genomes, and this figure doesn’t consider the sequence space within the thousands of publicly available metagenomes. Assessing the biosynthetic potential of the total pool of nucleic acids described constitutes a better approach to type II PKS BGC mining and this is validated by the identification of predicted type II PKS BGCs using a homology dependant approach in the uncultivatable Omnitrophica spp. (Table 7.1) persisting in a range of ecological niches e.g. hypolimnion permafrost lakes\(^165\) and the Bison hot spring, Yellowstone\(^166\).

### 7.6.3 BLASTP identification

Using a simple BLASTP approach to identify AntD and E homologues was a more bountiful way to identify non-Actinobacterial type II PKS BGCs. Here we describe 12 unique BGCs across 18 non-actinobacterial Genera, spanning 11 phylogenetically diverse Classes and comprising organisms with exceptionally diverse phenotypes e.g. photosynthetic cyanobacteria, Gloeocapsa sp. PCC 7426\(^167\), acidophilic psychrophiles, acidithiobacillus ferrivorans\(^168\), obligate symbiont thermophiles, Ca. Desulfotervidus auxillii\(^141\) and human gut commensals, 294
Observation of type II PKSs within genomes of phylogenetically, ecologically and biogeographically diverse organisms (Table 7.1) debunks the literature dogma predominantly restricting this class of specialised metabolite to the Actinobacterial Phylum, and more specifically the *Streptomyces* Genus\(^{74,103}\).

A targeted approach to mine type II PKS BGC from nucleotide databases using a BLASTp approach suffers from caveats, discussed in 7.5.2., i.e. AntD and E BGC homologues are not sequence diverse. Whilst this may be the case, components of the *pab* cluster, exclusively identified using a homology independent approach, also clade with all other non-Actinobacterial KS and CLF homologues (Figure 7.11). Despite not being a top homologue, considerable crossover exists between BLASTP results from AntE and PabB, including the *clr* BGC and predicted type II CLFs from *Gloeocapsa* sp. PCC 7428, *Ca. D. auxillii*, *K. racemifer*, *Blautia* spp. and the *Omnitrophica* implying the homology dependent approach provided good coverage of non-Actinobacterial type II PKSs. Multiple unique PabB homologues were also identified which may broaden the pool of aromatic polyketide producers further, including putative proteins from the strictly anaerobic SRB *Desulfosporosinus* spp.\(^{179}\), isolates from the rare Gemmatimonadetes Phylum\(^{171}\), three Chloroflexales, *Oscillochloris trichoides*\(^{172}\), *Chloroflexus aggregans*\(^{173}\) and *Chloroflexus islandicus*\(^{174}\), the Cyanobacterium *Chroogloeocystis siderophila*\(^{175}\), the Clostridia *Thermoanaerobacterium thermosaccharolyticum*\(^{176}\) and Bacilli including *Dehalobacterium formicoaceticum*\(^{177,178}\), however each new candidate genome is currently being mined for BGCs and therefore the biosynthetic potential for aromatic polyketides remains to be seen.

Concerted efforts have been made to explore the specialised metabolite producing capacity of Actinobacteria (Chapter 1), heavily biasing datasets of type II PKS containing organisms\(^{10}\). Concerted efforts to the same end have been made to explore the *Xenorhabdus* and *Photorhabdus* spp.\(^{36-42}\), however, despite this the *ant* BGC is the sole type II PKS described within these Genera\(^{71}\). Furthermore, the *clr* BGC is the only type II PKS to be characterised from *Clostridia* to-date\(^{2}\), despite being within an exceptionally well-studied Genus\(^{179}\) which has been broadly mined for cryptic BGCs\(^{180}\). Additionally, efforts to analyse all genomes of a bacterial Order or Class containing at least one AntD or AntE BLASTp hits, reported here, failed to identify additional unique BGCs e.g. of 75 Desulfobacterales, and 41 Chloroflexi genomes only fragments of homologous BGCs could be identified. Using these identification frequencies, one could argue that unique type II PKSs are common within *Streptomyces*, and the Actinobacteria, and comparatively rare in all other Phyla. However, the diversity of Actinobacteria screened for antimicrobial activity using Waksman-like platforms\(^{181}\) cannot be paralleled by any other organisms\(^{10,182,183}\), and it remains to be seen if the non-Actinobacterial hosts described here are bountiful reservoirs of unique aromatic polyketides, or if the observations reported here-in are truly sporadic.

Prediction of three type II PKS BGCs, two of which satisfied the type II PKS specific criteria described here, within *Ktedonobacter racemifer* (Figure 7.31) pinpoints this Genus as a good source of aromatic polyketides, however the lack of finished, complete or high-quality draft genomes of other *Ktedonobacter* spp. negated identification of further unique clusters, and only
core PKS components could be identified from exceptionally short contig sequences. *Ktedonobacter* spp. share remarkable phenotypic and physiological traits and ecological niches with *Streptomyces*\textsuperscript{147-149}, and the corresponding aromatic polyketides may play conserved roles e.g. as spore pigments\textsuperscript{131,147,184}. The genome sequence *K. racemifer* reveals an extraordinary capacity for acquisition of exogenous DNA by horizontal gene transfer and encodes 601 predicted transposases, 151 predicted integrases and 107 predicted resolvases\textsuperscript{147}. Coupled with its large genome size (13.66 Mb) *Ktedonobacter racemifer* spp. may be a hotbed for evolution of specialised metabolite BGCs and the capacity of this Genus to produce specialised metabolites has not gone unnoticed\textsuperscript{149}: two novel specialised metabolites were recently described from *Thermosporothrix hazakensis* SK20-1T\textsuperscript{185,186}, which also encodes a *kra*\textsubscript{2}-like type II BGC (Figure 7.31). In stark contrast to *K. racemifer*, the *dau* BGC was only 1 of 2 predicted BGCs identified within the genome of Ca. *D. auxilii* and therefore may play a more fundamental role in its biology, or provide a competitive advantage over other SRB when forming consortia with ANME. It is too early to speculate the roles these specialised metabolites *in natura*, however the broad dissemination of unique BGCs phylogenetically, ecologically and biogeographically presents a strong precedence that type II PKS are ubiquitous, albeit rare, specialised metabolites.

Generally, the poor quality of sequence data e.g. short contig length, and lack of depth within a given species or Genus is a heavily contributing factor to the dearth of non-Actinobacterial type II PKSs described. The GEBA approach\textsuperscript{25,26} to genome sequencing needs to be more widely adopted to populate publicly accessible databases with nucleotide sequences from a breadth of organisms. Such an approach will further broaden the pool of organisms harbouring type II PKS BGCs and help define the existence of the new and predominantly non-Actinobacterial lineage of type II PKS machinery described here (Figure 7.11).

### 7.6.4 Non-Actinobacterial type II PKSs as recombinant proteins in *Escherichia coli*

*Escherichia coli* is a virtuous host for the molecular biologist, and often the first port of call when isolating protein for crystallography or *in vitro* reactions etc.. Heterologous gene expression and purification of recombinant proteins from *E. coli* has played a fundamental role in our understanding of polyketide biosynthesis\textsuperscript{187-198}. However, disregarding the work presented in this thesis, the use of *E. coli* for the heterologous expression of KS and CLF genes has been completely confounded by persistent insolubility of the class of recombinant protein, the cause of which is unknown\textsuperscript{199}. Accordingly, both *in vitro* and *in vivo* experimentation necessitated the use of a *Streptomyces* heterologous expression host\textsuperscript{200}. In previous work we demonstrate the solubility and functionality of the *ant* KS/CLF in *E. coli* BL21(DE3), chosen because of its amino acid sequence similarities to *E. coli* FabF (Chapter 5 and 6). Here, we provide a better phylogenetic context for AntD and AntE and show these PKS components to clade separately from Actinobacterial KS/CLFs and into a second lineage of KS/CLFs predominantly comprised of non-Actinobacterial hosts. Characterised non-Actinobacterial KS and CLF sequences (*clr*\textsuperscript{2}, *ant*\textsuperscript{71} and *aua*\textsuperscript{70,92,201}), shown previously to be outliers of the canonical Actinobacterial KS/CLF
clades, also clade within the second lineage of type II PKSs alongside the only characterised
Actinobacterial member of this group: the rem KS/CLF, derived from S. resistomycificus\textsuperscript{130}
(Figure 7.11).

We expressed 7 KS/CLF candidates from this second discrete lineage using the same gene
expression strategy reported previously for ant KS/CLF (Chapter 5) and identified a further 4
soluble recombinant KS/CLF heterodimers in E. coli BL21(DE3) originating from Streptococcus
sp. GMDXs of the Mitis group\textsuperscript{112}, isolated from the Human oral microbiome\textsuperscript{119}, Ktedonobacter
racemifer type strain SOSP1-21\textsuperscript{T} isolated from soil\textsuperscript{148}, a variety of marine Pseudoalteromonas
luteoviolacea from discrete biogeographical and ecological niches and the previously
characterised KS/CLF from the rem BGC\textsuperscript{105}. Usefully, the latter three PKSs are either predicted
(Table 7.2), or characterised\textsuperscript{105}, to produce C\textsubscript{20} polyketide which can form a variety of aromatic
polyketide chemotypes including the clinically relevant tetracyclines\textsuperscript{202} and angucyclines e.g.
doxorubicin\textsuperscript{203} upon co-expression with chemotype specific cyclases\textsuperscript{130}. Combinatorial co-
expression of cyclases and cyclase/aromatases can direct cyclisation of nascent poly-β-ketide
chain towards different C\textsubscript{20} scaffold: Metsa¨-Ketela et al., (2003)\textsuperscript{204} show substitution of the third
ring cyclase during biosynthesis of the linear U-shaped anthracyclines aclacinomycin and
nogalamycin with, pgaF, from the angucycline gaudimycin BGC\textsuperscript{205} resulted in formation of a J-
shaped polyketide core\textsuperscript{204} (Figure 7.38). Furthermore, selective co-expression of remABCDIL
with or without remF resulted in formation of a discoid, S-shaped, polyketide or the U-shaped
TcmD1 and TcmD3: known linear tetracenomycin shunt metabolites\textsuperscript{130}. Combinatorial swapping
of kra, rem and plu KS/CLFs with their respective cyclases, and additional characterised
cyclases e.g. tcmN and pgaF, may feasibly enable biosynthesis and modification of S-, J- and
U-shaped decaketides in E. coli (Figure 7.37). Full characterisation of the mPKS components
and primary PKS tailoring enzymes will greatly expand the aromatic polyketide chemical space
currently accessible in E. coli.

\begin{align*}
\text{Resistomycin} & \quad \text{S-shaped} \\
\text{Tetracenomycin A2} & \quad \text{U-shaped} \\
\text{Oviedomycin} & \quad \text{J-shaped}
\end{align*}
Figure 7.37: U-, J- and S- shaped aromatic polyketides

The cyclase complement of the ssp BGC from *Streptococcus* sp. GMDXs, *L. oris* and *L. salivarius* is similar to the ant BGC and, like the ant KS/CLF, SspAB is predicted to synthesis a C_{16} octoketide. Therefore, the ssp cluster may be less interesting candidate for engineering specialised metabolites in *E. coli*. The ssp BGC encodes a putative benzoyl-CoA ligase, sspG, however, which may introduce this more exotic starter unit into the growing polyketide chain, as observed in enterocin and the wailupemycins^{193,206}. Manipulation of the polyketide starter units and chain length are the two main determiners of diversity during type II PKS biosynthesis (Chapter 6), further characterisation of the reservoir of biosynthetic enzymes within the second lineage of type II PKSs will facilitate manipulation of both facets when using *E. coli* as a heterologous host.

### 7.6.5 KS/CLF quaternary structures

The role of predicted CLFs in polyketide biosynthesis remains enigmatic and cannot immediately be assigned the passive function ascribed of the well characterised act CLF^{83}. For instance, aurachin biosynthesis is observed in both *S. aurantiaca* Sg a15 and *Rhodococcus erythropolis* (rau) using type II PKS-like machinery i.e. a KS/CLF heterodimer is proposed to catalyse the condensation of an anthranilate starter unit with two malonyl extender units^{70,92,201,207}, however the CLF from both clusters differ dramatically in size. The aua CLF comprises 407 AA resides, not dissimilar in size to most CLFs, however the rau CLF is severely truncated, comprising only 176 AA and questioning its function in biosynthesis^{207}. Moreover, early steps of aurachin biosynthesis in *R. erythropolis* may use a homodimer forming KS, like FabF, H and B FASs^{146}, and the lack of a CLF glutamine residue, involved in decarboxylation of starter units led to the proposition of *auaD* being inactive^{208}. Neither aua or rau BGCs encode polyketide cyclases indicating the aurachin polyketide intermediate to be spontaneously cyclised to form a quinolone core, which would not occur should the polyketide chain be fully reduced, as might be expected for a FAS system. Chain length factors and cyclases are known to prevent spontaneous intramolecular condensations during polyketide biosynthesis^{209}, implicating CLFs in rau and aua biosynthetic pathways, yet it remains to be seen if these unusual KS/CLF complexes form the 1:1 heterodimeric complexes of canonical Actinobacterial type II PKSs^{210}. Alternatively, the nascent polyketide chain could be synthesised in a manner analogous to type III PKSs, however using ACP activated substrates. In accordance with synthetic biology tenets it is important to characterise the quaternary structure of each type II PKS to enable successful, and non-redundant, substitution into the plug-and-play platform described in Chapter 5.
Minimal PKS components are ordered as ACP, KS and CLF and shown as light blue ORFs. The *R. erythropolis* CLF, rauE, is truncated with respect to auaD, the counterpart from *S. aurantiaca* Sg a15. The anthranilate-CoA ligase is shown in royal blue.

Whilst the CLFs from the *ssp*, *plu*, *kra* and *rem* clusters are canonical in length, no candidates comprise the decarboxylative glutamine residue CLF active site (Table 7.2). Additionally, the sequence divergence from canonical actinobacterial type II PKSs questions the quaternary nature of these biosynthetic complexes. Interestingly, the predicted octoketide forming *Streptococcus* sp. GMDXs KS/CLF was the only PKS machinery shown to readily copurify as a heterodimer (Figure 7.18) consistent with the *ant* KS/CLF (Chapters 5 and 6), all other soluble KS/CLFs either co-purify poorly (PluAB), or not at all (KraAB). Deletion of KSs genes from the co-expression vectors showed the solubility of each monomer to depend upon co-expression of its cognate partner *in vivo*, consistent with a heterodimeric complex being formed, however the cognate PKS partner is lost during purification suggesting the KS/CLF interactions to be transient or weak as discussed in 7.5.3.3.2. Additional ancillary KS/CLF stabilising enzymes may be required for efficient PKS complex purification and mPKS constituents will need to be evaluated empirically. The observation of KS/CLF stabilising enzymes is well documented in the literature and the complement of proteins necessary for polyketide biosynthesis, the mPKS, is fluid for different aromatic polyketides e.g. the resistomycin mPKS requires two additional cyclases, *remI* and *remL* or *remF* to function\textsuperscript{130}, the enterocin mPKS comprises a EncD, a C9 ketoreductase as well as the KS/CLF and hexanoyl-primed ACP\textsuperscript{206,211} and cyclases are necessary to stabilise the KS/CLF complex during biosynthesis of *S. coelicolor* A3(2) spore pigment\textsuperscript{131}, A-74528\textsuperscript{212} and fredericamycin\textsuperscript{213}. However, the enterocin\textsuperscript{193}, spore pigment\textsuperscript{188} and fredericamycin\textsuperscript{214} KS/CLF have been demonstrated to co-purify successfully. The dissociative nature of the non-Actinobacterial KS/CLFs may be a characteristic of this lineage. It remains to be seen if this is an issue for polyketide biosynthesis *in vivo*, however the dissociative nature of these complexes confounds *in vitro* experiments and crystallisation attempts. Introduction of a flexible linker between the two monomeric subunits, demonstrated for AntD and AntE (Chapter 6) may be a generally applicable strategy for co-purification. Chain length factors from the *bend*, *dau* and *dac* BGCs showed some solubility in the *E. coli* recombinant proteome, however the topology of these recombinant proteins is unknown, and in each case soluble protein microaggregates may be forming. The addition of a linker between the respective KS and CLF components is shown in royal blue.

**Figure 7.38:** mPKS from two aurachin producers
monomers may facilitate KS/CLF solubility however the general applicability of this strategy to improve solubility remains untested. Scant literature states linking Actinobacterial KS/CLFs does not aid solvation\textsuperscript{198} however the solubility of each monomeric component is unknown in \textit{E. coli}.

### 7.7 Conclusion

Using a suite of bioinformatics tools we debunk the literature dogma which states type II PKSs are predominantly identified within the Actinobacteria. We identify 12 unique BGCs from across the Prokaryotic Kingdom, from 5 different Phyla and show PKS components from each of these to form a type II PKS lineage which is discrete from the canonical Actinobacterial KSs and CLFs. Furthermore, we show KS and CLF sequences from previously described non-Actinobacteria to sit within this clade. Using previous observations that non-Actinobacterial KS/CLF heterodimers may have a propensity to form soluble recombinant proteins in \textit{E. coli} we test 7 KS/CLF pairs, 4 of which are soluble and will be an exceptionally valuable source for the dervitisation of aromatic polyketide in \textit{E. coli}.

Through bioprospecting we identify putative protein coding sequences within the 12 unique non-Actinobacterial BGCs with a variety of predicted tailoring functions e.g. prenyltransferases and halogenases and access to a variety of polyketide chain lengths, C\textsubscript{16}, C\textsubscript{20} and C\textsubscript{24}, which can be added to the bio-engineer’s toolbox. After characterisation, each enzyme can be plugged into the various modules described in Chapter 5 to generate a wide array of diverse aromatic structures in \textit{Escherichia coli}, and in the case of altering polyketide chain length, the KS/CLF can be substituted opposed to the more complex protein engineering of the \textit{ant} KS/CLF heterodimer (Chapter 6). Coupled with the synthetic biology toolbox, the utility of \textit{E. coli} permits the high throughput combinatorial biosynthesis of aromatic polyketides circumventing the poor genetic tractability and growth characteristic of \textit{Streptomyces} spp.\textsuperscript{215}, expediting our route to new designer bioactive aromatic polyketides.
7.8 References


67 Motley, J. L. et al. Opportunistic Sampling of Roadkill as an Entry Point to Accessing Natural Products Assembled by Bacteria Associated with Non anthropoidal Mammalian Microbiomes. *Journal of Natural Products* 80, 598-608, (2017).


112 Jensen, A., Scholz, C. F. P. & Kilian, M. Re-evaluation of the taxonomy of the Mitis group of the genus Streptococcus based on whole genome phylogenetic analyses, and proposed reclassification of Streptococcus dentisani as Streptococcus oralis subsp dentisani comb. nov., Streptococcus tigurinus as Streptococcus oralis subsp tigurinus.


Magli, A., Wendt, M. & Leisinger, T. Isolation and characterization of Dehalobacterium formicoacetitc gen nov sp nov, a strictly anaerobic bacterium utilizing...


### 7.9 Supplementary tables

Supplementary Table 7.1: Organisation of the pab BGC from *Paenibacillus borealis* DSM 13188

<table>
<thead>
<tr>
<th>ORF</th>
<th>Putative function</th>
<th>homologue protein / protein domains</th>
<th>homologue organism</th>
<th>Identity / coverage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>HTH domain containing protein</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>cyclase/dehydratase</td>
<td>SRPBCC (monodomain)</td>
<td><em>Gloeocapsa</em> sp. PCC 7428</td>
<td>47/96</td>
</tr>
<tr>
<td>3</td>
<td>beta-ketoacyl synthase</td>
<td>KAS_1 II</td>
<td><em>Omnitrophica</em> bacterium GWA2 41 15</td>
<td>30/98</td>
</tr>
<tr>
<td>4</td>
<td>beta-ketoacyl-[ACP] synthase II</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>acyl carrier protein</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>holo-[ACP] synthase</td>
<td>ACPS superfamily</td>
<td><em>Planomicrobium flavidum</em></td>
<td>34/99</td>
</tr>
<tr>
<td>7</td>
<td>Major facilitator superfamily transporter</td>
<td>MFS_1 superfamily</td>
<td><em>Methanohalobium evestigatum</em></td>
<td>31/97</td>
</tr>
<tr>
<td>8</td>
<td>cupin domain-containing protein</td>
<td>schB</td>
<td><em>Streptomyces halstedii</em></td>
<td>43/99</td>
</tr>
<tr>
<td>9</td>
<td>4-hydroxyphenyl-beta-ketoacyl-CoA hydrolase</td>
<td>metal-dependent hydrolase superfamily</td>
<td><em>Streptomyces</em> sp. CB03238</td>
<td>33/82</td>
</tr>
<tr>
<td>10</td>
<td>cyclase</td>
<td>cyclase superfamily</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>TIGR04076 family protein</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>coproporphyrinogen III oxidase</td>
<td>HemN superfamily</td>
<td><em>Myxococcus stipitatus</em> DSM 14675</td>
<td>32/98</td>
</tr>
<tr>
<td>13</td>
<td>carboxymuconolactone decarboxylase</td>
<td>CMD superfamily</td>
<td><em>Clostridium beijerinckii</em></td>
<td>36/74</td>
</tr>
<tr>
<td>14</td>
<td>coproporphyrinogen III oxidase</td>
<td>HemN superfamily</td>
<td><em>Myxococcus stipitatus</em></td>
<td>36/92</td>
</tr>
<tr>
<td>15</td>
<td>peptidase U32</td>
<td>peptidase U32</td>
<td><em>Desulfomicrobium baculatum</em></td>
<td>35/79</td>
</tr>
<tr>
<td>16</td>
<td>Multimeric flavodoxin WrbA</td>
<td>FMN red superfermaly, WrbA</td>
<td><em>Proteiniborusr ethanologenes</em></td>
<td>38/43</td>
</tr>
<tr>
<td>17</td>
<td>Pimeloyl-ACP methyl ester carboxylesterase</td>
<td></td>
<td><em>Nonomuraea pusilla</em></td>
<td>34/96</td>
</tr>
<tr>
<td>18</td>
<td>B12 binding domain containing radical SAM protein</td>
<td>Radical SAM superfamily</td>
<td><em>Geobacter</em> sp. DSM 9736</td>
<td>40/96</td>
</tr>
<tr>
<td>19</td>
<td>B12 binding domain containing radical SAM protein</td>
<td>B12-binding SF, Radical SAM superfamily</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>radical SAM protein</td>
<td>HemN superfamily</td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>HEAT repeat domain-containing protein</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

311
### Supplementary Table 7.2: Organisation of the dac BGC

<table>
<thead>
<tr>
<th>ORF</th>
<th>Gene designation</th>
<th>Putative function</th>
<th>homologue organism</th>
<th>Identity/coverage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMPREF9701_RS26195</td>
<td>HYPOTHETICAL / DNA POLYMERASE II SUBUNIT ALPHA</td>
<td>Pseudoalteromonas agarivorans</td>
<td>36/47</td>
<td></td>
</tr>
<tr>
<td>HMPREF9701_RS26200</td>
<td>HYPOTHETICAL PROTEIN</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HMPREF9701_RS26205</td>
<td>HYPOTHETICAL PROTEIN / SMALL TERMINASE SUBUNIT</td>
<td>Pseudomonas phage B3</td>
<td>35/79</td>
<td></td>
</tr>
<tr>
<td>HMPREF9701_RS26210</td>
<td>HYPOTHETICAL PROTEIN / LARGE TERMINASE SUBUNIT</td>
<td>Pseudomonas phage B3</td>
<td>99/69</td>
<td></td>
</tr>
<tr>
<td>HMPREF9701_RS26215</td>
<td>DUF935 / PHAGE GP37GP68 FAMILY PROTEIN</td>
<td>Medicago truncatula</td>
<td>57/89</td>
<td></td>
</tr>
<tr>
<td>HMPREF9701_RS26220</td>
<td>HEAD MORPHOGENESIS PROTEIN SPP1</td>
<td>Polaronas sp. CG9_12</td>
<td>65/98</td>
<td></td>
</tr>
<tr>
<td>HMPREF9701_RS26225</td>
<td>VIRION MORPHOGENESIS PROTEIN</td>
<td>Simonsiella muelleri</td>
<td>31/95</td>
<td></td>
</tr>
<tr>
<td>HMPREF9701_RS26230</td>
<td>ACYL-COA DEHYDROGENASE</td>
<td>Delftia tsuruhatensis</td>
<td>99/100</td>
<td></td>
</tr>
<tr>
<td>HMPREF9701_RS26235</td>
<td>AMINGLYCOSIDE PHOSPHOTRANSFERASE / ACYL-COA DEHYDROGENASE</td>
<td>Delftia acidovorans</td>
<td>99/100</td>
<td></td>
</tr>
<tr>
<td>HMPREF9701_RS26240</td>
<td>CRP/FNR TRANSCRIPTIONAL REGULATOR</td>
<td>Comamonas composti</td>
<td>95/99</td>
<td></td>
</tr>
<tr>
<td>HMPREF9701_RS26245</td>
<td>3-HYDROXYACYL-COA DEHYDROGENASE</td>
<td>Delftia sp. ZNC0008</td>
<td>99/100</td>
<td></td>
</tr>
<tr>
<td>HMPREF9701_RS26250</td>
<td>HYPOTHEtical PROTEIN / HYDROXIMETHYLPyRIMIDINE/PHOSPHOMETHYLPyRIMIDINE KINASE</td>
<td>Acidovorax sp. MR-S7</td>
<td>54/32</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>HMPREF9701_RS26255</td>
<td>HYPOTHETICAL PROTEIN / FLAVODIXIN REDUCTASE (FERREDOXIN-NADPH REDUCTASE) FAMILY 1</td>
<td>Phaeobacter galleaeciens</td>
<td>58/86</td>
</tr>
<tr>
<td>11</td>
<td>HMPREF9701_RS26260</td>
<td>DUF1772 / 1,4-DIHYDROXY-2-NAPHTHOATE OCTAPRENYLTRANSFERASE</td>
<td>marine Actinobacterium PHSC20C1</td>
<td>31/47</td>
</tr>
<tr>
<td>10</td>
<td>ORF#1</td>
<td>3-KETOACYL ACP REDUCTASE</td>
<td>Azospirillum brasilense</td>
<td>47/94</td>
</tr>
<tr>
<td>9</td>
<td>HMPREF9701_RS26265</td>
<td>HYPOTHETICAL / ANTIBIOTIC BIOSYNTHESIS MONOOXYGENASE - ABM SUPERFAMILY DOMAIN</td>
<td>Methylobacterium platani</td>
<td>38/44</td>
</tr>
<tr>
<td>8</td>
<td>HMPREF9701_RS26270</td>
<td>ANTIBIOTIC BIOSYNTHESIS MONOOXYGENASE</td>
<td>Desulfovibrio africanus</td>
<td>29/36</td>
</tr>
<tr>
<td></td>
<td>Accession</td>
<td>Gene</td>
<td>Function</td>
<td>Species/Moniker</td>
</tr>
<tr>
<td>---</td>
<td>-----------</td>
<td>------</td>
<td>--------------------------------------------------------------------------</td>
<td>----------------------------------</td>
</tr>
<tr>
<td>7</td>
<td>HMPREF9701_RS26275</td>
<td>putative enzyme involved in extracellular polysaccharide biosynthesis / antibiotic biosynthesis monooxygenase</td>
<td>Desulfomonile tiedjei DSM 6799</td>
<td>29/40</td>
</tr>
<tr>
<td>6</td>
<td>HMPREF9701_RS26280</td>
<td>antibiotic biosynthesis monooxygenase</td>
<td>Pseudomonas carica papayae</td>
<td>32/57</td>
</tr>
<tr>
<td>5</td>
<td>HMPREF9701_RS26285</td>
<td>acyl carrier protein</td>
<td>Labilithrix luteola</td>
<td>40/77</td>
</tr>
<tr>
<td>4</td>
<td>HMPREF9701_RS26290</td>
<td>beta-ketoacyl-ACP synthase II</td>
<td>Gloeocapsa sp. PCC 7428</td>
<td>33/98</td>
</tr>
<tr>
<td>3</td>
<td>HMPREF9701_RS26295</td>
<td>beta-ACP synthase</td>
<td>Streptomyces sp. NRRL WC-3723</td>
<td>47/98</td>
</tr>
<tr>
<td>2</td>
<td>HMPREF9701_RS26300</td>
<td>Ketoacyl reductase</td>
<td>Delftia lacustris</td>
<td>97/100</td>
</tr>
<tr>
<td>1</td>
<td>HMPREF9701_RS26305</td>
<td>cyclase/dehydratase</td>
<td>Crinalium epipsammum</td>
<td>35/97</td>
</tr>
<tr>
<td></td>
<td>HMPREF9701_RS26310</td>
<td>Transposase (possibly from ISPsy6)</td>
<td>Pseudomonas syringae</td>
<td>44/84</td>
</tr>
<tr>
<td></td>
<td>HMPREF9701_RS26315</td>
<td>thiamine phosphate synthase</td>
<td>Delftia lacustris</td>
<td>99/100</td>
</tr>
<tr>
<td></td>
<td>HMPREF9701_RS26320</td>
<td>thiazole synthase</td>
<td>Comamonas testosteronei</td>
<td>93/98</td>
</tr>
<tr>
<td></td>
<td>HMPREF9701_RS26325</td>
<td>thiamine biosynthesis protein ThiS</td>
<td>Delftia acidovorans CCUG 15835</td>
<td>100/100</td>
</tr>
<tr>
<td></td>
<td>HMPREF9701_RS26330</td>
<td>FAD dependent oxidoreductase / glycine oxidase (100%, 99%)</td>
<td>Delftia acidovorans SPH-1</td>
<td>100/100</td>
</tr>
<tr>
<td></td>
<td>HMPREF9701_RS26335</td>
<td>hydroxymethylpyrimidine/phosphomethylypyrimidine kinase</td>
<td>Delftia sp. Cs1-3</td>
<td>99/100</td>
</tr>
<tr>
<td></td>
<td>HMPREF9701_RS26340</td>
<td>Peptide binding protein</td>
<td>Delftia sp. Cs1-4</td>
<td>98/100</td>
</tr>
<tr>
<td></td>
<td>HMPREF9701_RS26345</td>
<td>ankyrin</td>
<td>Comamonas testosteronei 544</td>
<td>71/75</td>
</tr>
<tr>
<td></td>
<td>HMPREF9701_RS26350</td>
<td>TonB-dependent receptor</td>
<td>Delftia sp. 670</td>
<td>98/100</td>
</tr>
<tr>
<td></td>
<td>HMPREF9701_RS26355</td>
<td>peptidylprolyl isomerase</td>
<td>Comamonas compositi</td>
<td>79/100</td>
</tr>
<tr>
<td></td>
<td>HMPREF9701_RS26360</td>
<td>BolA transcriptional regulator</td>
<td>Comamonas compositi</td>
<td>92/85</td>
</tr>
<tr>
<td></td>
<td>HMPREF9701_RS26365</td>
<td>intracellular septation protein A</td>
<td>Delftia sp. RIT313</td>
<td>100/97</td>
</tr>
<tr>
<td></td>
<td>HMPREF9701_RS26370</td>
<td>peptide methionine (R)-s-oxide reductase</td>
<td>Comamonas compositi</td>
<td>100/99</td>
</tr>
<tr>
<td>ORF</td>
<td>Gene designation</td>
<td>Putative function</td>
<td>homologue protein</td>
<td>homologue organism</td>
</tr>
<tr>
<td>-------</td>
<td>------------------</td>
<td>-----------------------------------------</td>
<td>--------------------------</td>
<td>-----------------------------------------</td>
</tr>
<tr>
<td>1</td>
<td>AMK58_25500</td>
<td>cyclase/dehydratase</td>
<td>Di-SRPBCC superfamily</td>
<td>Crinalium epipsammum</td>
</tr>
<tr>
<td>2</td>
<td>AMK58_25505</td>
<td>Ketoacyl reductase</td>
<td>SRD superfamily</td>
<td>Pseudoalteromonas luteoviolacea</td>
</tr>
<tr>
<td>3</td>
<td>AMK58_25510</td>
<td>Ketosynthase</td>
<td>KAS_I-II</td>
<td>Acidithiobacillus ferrivorans</td>
</tr>
<tr>
<td>4</td>
<td>AMK58_25515</td>
<td>Chain length factor</td>
<td></td>
<td>Acidithiobacillus ferrivorans</td>
</tr>
<tr>
<td>5</td>
<td>AMK58_25520</td>
<td>Acyl carrier protein</td>
<td>AcpP</td>
<td>Schistosoma haematobium</td>
</tr>
<tr>
<td>6</td>
<td>AMK58_25525</td>
<td>antibiotic monooxygenase</td>
<td>ABM superfamily</td>
<td>Acidithiobacillus ferrivorans</td>
</tr>
<tr>
<td>7</td>
<td>AMK58_25530</td>
<td>antibiotic monooxygenase</td>
<td>ABM superfamily</td>
<td>Desulfobacca sp. RBG_16_60_12</td>
</tr>
<tr>
<td>8</td>
<td>AMK58_25535</td>
<td>antibiotic monooxygenase</td>
<td>ABM superfamily</td>
<td>Desulfovibrio sp. X2</td>
</tr>
<tr>
<td>9</td>
<td>AMK58_25540</td>
<td>antibiotic monooxygenase</td>
<td>ABM superfamily</td>
<td>Desulfobacca sp. RBG_16_60_12</td>
</tr>
<tr>
<td>10</td>
<td>AMK58_25545</td>
<td>NAD(P)-dependent oxidoreductase</td>
<td>SRD superfamily</td>
<td>Acidithiobacillus ferrivorans</td>
</tr>
<tr>
<td>11</td>
<td>AMK58_25550</td>
<td>4’-phosphopantetheinyl transferase</td>
<td>SFP PPTase</td>
<td>Solibacillus silvestris</td>
</tr>
</tbody>
</table>
# Supplementary Table 7.4: Organisation of the dac BGC from *Acidothiobacillus ferrivorans*

<table>
<thead>
<tr>
<th>ORF</th>
<th>Gene designation</th>
<th>Putative function</th>
<th>homologue protein</th>
<th>homologue organism</th>
<th>Identity / coverage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>BBC27_01920</td>
<td>cyclase/dehydratase</td>
<td>Di-SRPBCC superfamily (OctcD1)</td>
<td>Crinalium epipsammum</td>
<td>35/96</td>
</tr>
<tr>
<td>2</td>
<td>BBC27_01915</td>
<td>Ketoacyl reductase</td>
<td>SRD superfamily</td>
<td>Azospirillum brasilense</td>
<td>77/99</td>
</tr>
<tr>
<td>3</td>
<td>BBC27_01910</td>
<td>Ketosynthase</td>
<td>KAS_I_II</td>
<td>Azospirillum brasilense</td>
<td>67/98</td>
</tr>
<tr>
<td>4</td>
<td>BBC27_01905</td>
<td>Chain length factor</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>BBC27_01900</td>
<td>AcrL carrier protein</td>
<td>AcpP</td>
<td>Enterobacter lignolyticus</td>
<td>35/95</td>
</tr>
<tr>
<td>6</td>
<td>BBC27_01895</td>
<td>Antibiotic monoxygenase</td>
<td>ABM superfamily (HmoA)</td>
<td>Azospirillum brasilense</td>
<td>42/72</td>
</tr>
<tr>
<td>7</td>
<td>BBC27_01890</td>
<td>Antibiotic monoxygenase</td>
<td>ABM superfamily</td>
<td>Pseudoalteromonas luteoviolacea</td>
<td>39/94</td>
</tr>
<tr>
<td>8</td>
<td>BBC27_01885</td>
<td>Antibiotic monoxygenase</td>
<td>ABM superfamily</td>
<td>Azospirillum brasilense</td>
<td>50/97</td>
</tr>
<tr>
<td>9</td>
<td>BBC27_01880</td>
<td>Antibiotic monoxygenase</td>
<td>ABM superfamily</td>
<td>Azospirillum brasilense</td>
<td>52/92</td>
</tr>
<tr>
<td>10</td>
<td>BBC27_01875</td>
<td>NAD(P)-dependent oxidoreductase 4'</td>
<td>SRD superfamily</td>
<td>Azospirillum brasilense</td>
<td>61/97</td>
</tr>
<tr>
<td>11</td>
<td>BBC27_01870</td>
<td>4'-phosphopantetheiny transferase</td>
<td>SFP PPTase</td>
<td>Streptococcus mutans</td>
<td>31/63</td>
</tr>
<tr>
<td>12</td>
<td>BBC27_01865</td>
<td>Flavodoxin reductase (ferredoxin-NADPH reductase) family 1</td>
<td></td>
<td></td>
<td>58/94</td>
</tr>
</tbody>
</table>

# Supplementary Table 7.5: Amino acid sequence similarity of all dau BGCs

<table>
<thead>
<tr>
<th>Azospirillum brasilense</th>
<th>Delftia spp.</th>
<th>Acidothiobacillus ferrivorans</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>49/97</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>69/98</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>61/95</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>52/98</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>35/50</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>34/63</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>34/73</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td>36/92</td>
<td>8</td>
</tr>
<tr>
<td>9</td>
<td>41/92</td>
<td>9</td>
</tr>
<tr>
<td>10</td>
<td>47/95</td>
<td>10</td>
</tr>
<tr>
<td>11</td>
<td>24/8</td>
<td>11</td>
</tr>
</tbody>
</table>

Identity and coverage values are reported as percentages
Supplementary Table 7.6: Organisation of the ssp BGC from *Streptococcus* sp. GMDXs

<table>
<thead>
<tr>
<th>ORF</th>
<th>Gene designation</th>
<th>Putative function</th>
<th>homologue protein / protein domains</th>
<th>homologue organism</th>
<th>Identity / coverage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMD2S_RS08875</td>
<td></td>
<td>Regulatory protein</td>
<td>MutR family transcriptional regulator</td>
<td><em>Streptococcus equinus</em></td>
<td>64, 98</td>
</tr>
<tr>
<td>1</td>
<td>GMD2S_RS08880</td>
<td>3-keto-acyl ACP synthase</td>
<td>FabB and KAS-I-II</td>
<td><em>Lactobacillus oris</em></td>
<td>57, 99</td>
</tr>
<tr>
<td>2</td>
<td>GMD2S_RS08885</td>
<td>beta-keto-synthase</td>
<td>FabF</td>
<td><em>Lactobacillus oris</em></td>
<td>52, 97</td>
</tr>
<tr>
<td>3</td>
<td>GMD2S_RS08890</td>
<td>PP-binding protein</td>
<td>Acyl Carrier Protein</td>
<td><em>Streptomyces sp. Ncost-T10-10d</em></td>
<td>30, 69</td>
</tr>
<tr>
<td>4</td>
<td>GMD2S_RS08895</td>
<td>Methyltransferase</td>
<td>SAM-dependent methyltransferase</td>
<td><em>Thermoactinomyces vulgaris</em></td>
<td>63, 97</td>
</tr>
<tr>
<td>5</td>
<td>GMD2S_RS08900</td>
<td>ACPS PPTase</td>
<td>ACP synthase</td>
<td><em>Thermoactinomyces vulgaris</em></td>
<td>40, 65</td>
</tr>
<tr>
<td>6</td>
<td>GMD2S_RS08905</td>
<td>Benzoate CoA ligase</td>
<td>AFD_Class-I superfamly</td>
<td><em>Ethanoligenens harbinense</em></td>
<td>33, 97</td>
</tr>
<tr>
<td>7</td>
<td>GMD2S_RS08910</td>
<td>Cupin 2</td>
<td>TcmJ / Cupin 2</td>
<td><em>Streptomyces glaucescens</em></td>
<td>33, 80</td>
</tr>
<tr>
<td>8</td>
<td>GMD2S_RS08915</td>
<td>carboxymuconolactone decarboxylase</td>
<td>Yurz</td>
<td><em>Gloeocapsa sp. PCC7428</em></td>
<td>42, 87</td>
</tr>
<tr>
<td>9</td>
<td>GMD2S_RS08920</td>
<td>Cyclase</td>
<td>SRPBCC domain, polyketide cyclase/dehydratase</td>
<td><em>Gloeocapsa sp. PCC7428</em></td>
<td>48, 99</td>
</tr>
<tr>
<td>10</td>
<td>GMD2S_RS08925</td>
<td>Cyclase</td>
<td></td>
<td><em>Lactobacillus oris</em></td>
<td>37, 96</td>
</tr>
<tr>
<td>11</td>
<td>GMD2S_RS08930</td>
<td>alpha/beta hydrolase</td>
<td>EnCL possibly</td>
<td><em>Lactobacillus oris</em></td>
<td>48, 99</td>
</tr>
<tr>
<td>12</td>
<td>GMD2S_RS08935</td>
<td>DUF2992</td>
<td></td>
<td><em>Streptococcus mutants</em></td>
<td>48, 99</td>
</tr>
</tbody>
</table>
### Supplementary Table 7.7: Organisation of the ssp homologous BGC from *Lactobacillus oris* PB013-T2-3

<table>
<thead>
<tr>
<th>ORF</th>
<th>Gene designation</th>
<th>Putative function</th>
<th>homologue protein / protein domains</th>
<th>homologue organism</th>
<th>Identity / coverage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HMPREF9265_RS04280</td>
<td>Resolvase</td>
<td>SR_ResInv domain, serine recombinase and HTH superfamily</td>
<td><em>Lactobacillus fermentum</em></td>
<td>98, 99</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ORF57</td>
<td></td>
<td>MarR family transcriptional regulator</td>
<td>MarR-2 domain</td>
<td><em>Lactobacillus salivarius</em></td>
<td>99, 97</td>
</tr>
<tr>
<td>12</td>
<td>HMPREF9265_RS04285</td>
<td>DUF2992 - unknown</td>
<td>DUF2992 domain</td>
<td><em>Lactobacillus salivarius</em></td>
<td>25, 87</td>
</tr>
<tr>
<td>11</td>
<td>HMPREF9265_RS04290</td>
<td>alpha/beta hydrolase</td>
<td>AHydrolase_5 domain, MhpC multidomain</td>
<td><em>Streptococcus porci</em></td>
<td>25, 87</td>
</tr>
<tr>
<td>10</td>
<td>HMPREF9265_RS04295</td>
<td>aromatase/cyclase</td>
<td>SRPBC superfamily</td>
<td><em>Geodermatophilus obscurus</em></td>
<td>39, 82</td>
</tr>
<tr>
<td>9</td>
<td>HMPREF9265_RS04300</td>
<td>cyclase</td>
<td>Cyclase superfamily</td>
<td><em>Chroogloeocystis siderophila</em></td>
<td>46, 99</td>
</tr>
<tr>
<td>7</td>
<td>HMPREF9265_RS04305</td>
<td>cupin 2</td>
<td>Cupin_2</td>
<td>*Actinobacteria bacterium 13 1 20CM 3 71 11</td>
<td>36, 61</td>
</tr>
<tr>
<td>6</td>
<td>HMPREF9265_RS04310</td>
<td>Benzoate-CoA ligase</td>
<td>AFD_Class_I</td>
<td><em>Clostridium beijerinckii</em></td>
<td>36, 89</td>
</tr>
<tr>
<td>5</td>
<td>HMPREF9265_RS04315</td>
<td>ACPS PPTase</td>
<td>ACPS domain</td>
<td><em>Lactobacillus salivarius</em></td>
<td>99, 99</td>
</tr>
<tr>
<td>4</td>
<td>HMPREF9265_RS04320</td>
<td>SAM dependant methyltransferase</td>
<td>Methyltransf_11</td>
<td><em>Thermoactinomyces vulgaris</em></td>
<td>62, 97</td>
</tr>
<tr>
<td>3</td>
<td>HMPREF9265_RS04325</td>
<td>Acyl carrier protein</td>
<td>PP-binding domain</td>
<td><em>Gloeocapsa sp. PCC 7428</em></td>
<td>55, 67</td>
</tr>
<tr>
<td>2</td>
<td>HMPREF9265_RS04330</td>
<td>beta-ketoacyl synthase</td>
<td>FabB domain</td>
<td><em>Lactobacillus salivarius</em></td>
<td>99, 99</td>
</tr>
<tr>
<td>1</td>
<td>HMPREF9265_RS04335</td>
<td>3-ketoacyl-ACP synthase</td>
<td>KAS_I_II</td>
<td><em>Lactobacillus salivarius</em></td>
<td>99, 99</td>
</tr>
<tr>
<td></td>
<td>HMPREF9265_RS04340</td>
<td>Unknown</td>
<td>Hypothetical protein</td>
<td><em>Lactobacillus salivarius</em></td>
<td>99, 99</td>
</tr>
<tr>
<td></td>
<td>HMPREF9265_RS04345</td>
<td>integrase</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>HMPREF9265_RS04350</td>
<td>addiction toxin ReLE</td>
<td>GP49 domain, ParE_toxin superfamily</td>
<td><em>Lactobacillus gastricus</em></td>
<td>99, 99</td>
</tr>
</tbody>
</table>
### Supplementary Table 7.8: Organisation of the *plu* homologous BGCs from *Pseudoalteromonas luteoviolacea* strains

<table>
<thead>
<tr>
<th>ORF</th>
<th>Gene designation</th>
<th>Putative function</th>
<th>homologue protein / protein domains</th>
<th>homologue organism</th>
<th>Identity / coverage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>17</td>
<td>N475_19075</td>
<td>cytochrome c</td>
<td>Cytochrom_C_2</td>
<td><em>Vibrio coralliilyticus</em></td>
<td>47, 93</td>
</tr>
<tr>
<td>16</td>
<td>N475_19070</td>
<td>hydrogenase</td>
<td>Cytochrom_b_N</td>
<td><em>Vibrio coralliilyticus</em></td>
<td>53, 98</td>
</tr>
<tr>
<td>15</td>
<td>N475_19065</td>
<td>phosphopantetheine–protein transferase</td>
<td>SFP PPTase</td>
<td><em>Clostridium</em> sp. CAG:609</td>
<td>35, 57</td>
</tr>
<tr>
<td>14</td>
<td>N475_19060</td>
<td>ACP</td>
<td>PP-binding superfamily</td>
<td><em>Ruminococcus</em> sp. CAG:488</td>
<td>36, 68</td>
</tr>
<tr>
<td>13</td>
<td>N475_19055</td>
<td>beta-ketoacyl synthase</td>
<td></td>
<td><em>Streptococcus</em></td>
<td>31, 98</td>
</tr>
<tr>
<td>12</td>
<td>N475_19050</td>
<td>beta-ketoacyl-[acyl-carrier-protein] synthase II</td>
<td>KAS_I_II</td>
<td><em>Azospirillum brasilense</em></td>
<td>63, 98</td>
</tr>
<tr>
<td>11</td>
<td>N475_19045</td>
<td>ketoacyl reductase</td>
<td>SDR superfamily</td>
<td><em>Azospirillum brasilense</em></td>
<td>71, 99</td>
</tr>
<tr>
<td>10</td>
<td>N475_19040</td>
<td>cyclase/dehydrase</td>
<td>didomain SRPBCC superfamily, OtcD1</td>
<td></td>
<td>36, 96</td>
</tr>
<tr>
<td>9</td>
<td>N475_19035</td>
<td>oxidoreductase</td>
<td></td>
<td><em>Phaeobacter inhibens</em></td>
<td>48, 98</td>
</tr>
<tr>
<td>8</td>
<td>N475_19030</td>
<td>LuxR family transcriptional regulator</td>
<td>LuxR family</td>
<td><em>Chloroflexi bacterium</em> OL14</td>
<td>41, 83</td>
</tr>
<tr>
<td>7</td>
<td>N475_19025</td>
<td>MarR family transcriptional regulator</td>
<td>HTH_MARR superfamily</td>
<td><em>Photobacterium proteolyticum</em></td>
<td>51, 97</td>
</tr>
<tr>
<td>6</td>
<td>N475_19020</td>
<td>antibiotic biosynthesis monooxygenase</td>
<td></td>
<td><em>Acidithiobacillus ferrivorans</em></td>
<td>36, 55</td>
</tr>
<tr>
<td>5</td>
<td>N475_19015</td>
<td>antibiotic biosynthesis monooxygenase</td>
<td></td>
<td><em>Azospirillum brasilense</em></td>
<td>26, 96</td>
</tr>
<tr>
<td>4</td>
<td>N475_19010</td>
<td>antibiotic biosynthesis monooxygenase</td>
<td></td>
<td><em>Mycobacterium vulneris</em></td>
<td>29, 46</td>
</tr>
<tr>
<td>3</td>
<td>N475_19005</td>
<td>antibiotic biosynthesis monooxygenase</td>
<td></td>
<td><em>Labrenzia aggregata</em></td>
<td>41, 38</td>
</tr>
<tr>
<td>2</td>
<td>N475_19000</td>
<td>NAD(P)-dependent oxidoreductase</td>
<td>SDR superfamily</td>
<td><em>Acidithiobacillus ferrivorans</em></td>
<td>51, 99</td>
</tr>
<tr>
<td>1</td>
<td>N475_18995</td>
<td>cytochrome-c oxidase</td>
<td>Heme_Cu_oxidase_I superfamily</td>
<td><em>Bacillus azotoformans</em></td>
<td>30, 88</td>
</tr>
<tr>
<td></td>
<td>N475_18990</td>
<td>hypothetical protein</td>
<td></td>
<td><em>Pseudoalteromonas sp. R3</em></td>
<td>44, 86</td>
</tr>
<tr>
<td></td>
<td>N475_18985</td>
<td>hypothetical protein</td>
<td></td>
<td><em>Pseudoalteromonas phenolica</em></td>
<td>66, 98</td>
</tr>
</tbody>
</table>
Supplementary Table 7.9: Organisation of the \textit{bend} BGC from \textit{Bacillus endophyticus}

<table>
<thead>
<tr>
<th>ORF</th>
<th>Gene designation</th>
<th>Putative function</th>
<th>homologue organism</th>
<th>Identity / coverage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BM012_RS20615</td>
<td></td>
<td>hypothetical protein</td>
<td>\textit{Bacillus} sp. REN51N</td>
<td>100 / 93</td>
</tr>
<tr>
<td>BM012_RS20610</td>
<td></td>
<td>Plasmid replication / relaxation protein</td>
<td>\textit{Bacillus} phage poppyseed</td>
<td>40 / 96</td>
</tr>
<tr>
<td>BM012_RS20605</td>
<td></td>
<td>Cell division protein</td>
<td>\textit{Bacillus} sp. REN51N</td>
<td>94 / 100</td>
</tr>
<tr>
<td>BM012_RS20600</td>
<td></td>
<td>XRE family transcriptional regulator</td>
<td>\textit{paenibacillus} sp. soil522</td>
<td>55 / 89</td>
</tr>
<tr>
<td>1</td>
<td>BM012_RS20595</td>
<td>Pentachlorophenol-4-monoxygenase</td>
<td>\textit{Chlamydia abortus}</td>
<td>61 / 99</td>
</tr>
<tr>
<td>2</td>
<td>BM012_RS20590</td>
<td>ABC transporter</td>
<td>\textit{Actinomyces naeslundii}</td>
<td>36 / 90</td>
</tr>
<tr>
<td>3</td>
<td>BM012_RS20585</td>
<td>ABC-type multidrug transport system</td>
<td>\textit{Actinomyces oris}</td>
<td>26 / 84</td>
</tr>
<tr>
<td>4</td>
<td>BM012_RS20580</td>
<td>putative ABC transporter ATP-binding protein</td>
<td>\textit{Peptococcaceae bacterium} CEB3</td>
<td>45 / 99</td>
</tr>
<tr>
<td>5</td>
<td>BM012_RS20575</td>
<td>TetR/AcrR transcriptional regulator</td>
<td>\textit{paenibacillus algorifonticola}</td>
<td>30 / 64</td>
</tr>
<tr>
<td>6</td>
<td>BM012_RS20570</td>
<td>HTH domain containing protein</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>BM012_RS20565</td>
<td>Sigma-70 family RNAP sigma factor</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>BM012_RS20560</td>
<td>Cupin domain-containing protein</td>
<td>\textit{Streptomyces resistomycinicus} (crystal structure)</td>
<td>61 / 97</td>
</tr>
<tr>
<td>9</td>
<td>BM012_RS20555</td>
<td>tRNA (cmo5U34)-methyltransferase</td>
<td>\textit{Mucilanginibacter mallensis}</td>
<td>30 / 29</td>
</tr>
<tr>
<td>10</td>
<td>BM012_RS20550</td>
<td>tryptophan halogenase</td>
<td>\textit{Streptomyces sp. NBRC} 110030</td>
<td>39 / 71</td>
</tr>
<tr>
<td>11</td>
<td>BM012_RS20545</td>
<td>cyclase</td>
<td>\textit{Ktedonobacter racemifer}</td>
<td>55 / 100</td>
</tr>
<tr>
<td>12</td>
<td>BM012_RS20540</td>
<td>methyltransferase</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>BM012_RS20530</td>
<td>beta-ketoacyl synthase</td>
<td>\textit{Ktedonobacter racemifer}</td>
<td>51 / 98</td>
</tr>
<tr>
<td>15</td>
<td>BM012_RS20525</td>
<td>NAD(P)H-hydrate dehydratase</td>
<td>\textit{Anaerococcus sp. Marseille-P2143}</td>
<td>35 / 91</td>
</tr>
<tr>
<td>16</td>
<td>BM012_RS20520</td>
<td>hypothetical protein</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>BM012_RS20515</td>
<td>polyketide cyclase</td>
<td>\textit{Ktedonobacter racemifer}</td>
<td>47 / 94</td>
</tr>
<tr>
<td>18</td>
<td>BM012_RS20510</td>
<td>isoprenylcysteine carboxymethyltransferase / phospholipid methyltransferase</td>
<td>\textit{Sedimentibacteria} sp. SX930</td>
<td>28 / 71</td>
</tr>
<tr>
<td>19</td>
<td>BM012_RS20505</td>
<td>cation:proton antiporter</td>
<td>\textit{Paenibacillus} sp. FSL RS-0345</td>
<td>60 / 95</td>
</tr>
<tr>
<td></td>
<td>BM012_RS20500</td>
<td>hypothetical protein</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>BM012_RS20495</td>
<td>RNA binding protein S1</td>
<td>\textit{Bacillus} sp. REN51N</td>
<td>96 / 89</td>
</tr>
<tr>
<td></td>
<td>BM012_RS20490</td>
<td>Phage infection protein</td>
<td>\textit{Bacillus} sp. REN51N</td>
<td>80 / 100</td>
</tr>
</tbody>
</table>
Supplementary Table 7.10: Organisation of the *dau* BGC from *Candidatus Desulfofervidus auxilii*

<table>
<thead>
<tr>
<th>ORF</th>
<th>Gene designation</th>
<th>Putative function</th>
<th>homologue protein / protein domains</th>
<th>homologue organism</th>
<th>Identity / coverage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>HS1_001348</td>
<td>amine oxidase</td>
<td>amino_oxidase superfamily</td>
<td><em>Streptomyces chartreusis</em></td>
<td>25,61</td>
</tr>
<tr>
<td>2</td>
<td>HS1_001349</td>
<td>flavodoxin family protein</td>
<td>WrbA</td>
<td><em>Paenibacillus</em> sp. FSL R7-0337</td>
<td>36,90</td>
</tr>
<tr>
<td>3</td>
<td>HS1_001350</td>
<td>Fe-S oxidoreductase</td>
<td>radical SAM superfamily</td>
<td><em>Methanobacterium sp. Maddingley</em> MBC34</td>
<td>32,92</td>
</tr>
<tr>
<td>4</td>
<td>HS1_001351</td>
<td>aldehyde:ferredoxin oxidoreductase</td>
<td>AFOR_C superfamily</td>
<td><em>Caldarchaeum subterraneum</em></td>
<td>44,99</td>
</tr>
<tr>
<td>5</td>
<td>HS1_001352</td>
<td>2-dehydropantoate 2-reductase</td>
<td>Apba_C superfamily</td>
<td><em>Rubrobacter xylanophilus DSM 9941</em></td>
<td>32,99</td>
</tr>
<tr>
<td>6</td>
<td>HS1_001353</td>
<td>TIGR04076 family protein</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>HS1_001354</td>
<td>beta-ketoacyl-[acyl-carrier-protein] synthase II</td>
<td>KAS_I_II</td>
<td><em>Pseudoalteromonas luteoviolacea</em></td>
<td>50,99</td>
</tr>
<tr>
<td>8</td>
<td>HS1_001355</td>
<td>beta-ketoacyl-[acyl-carrier-protein] synthase II</td>
<td>KAS_I_II</td>
<td><em>Bacillus</em> sp. FIAT-18017</td>
<td>46,99</td>
</tr>
<tr>
<td>9</td>
<td>HS1_001356</td>
<td>ACP</td>
<td>PP-binding protein</td>
<td><em>Omnitrophica WOR_2 bacterium</em></td>
<td>45,96</td>
</tr>
<tr>
<td>10</td>
<td>HS1_001357</td>
<td>cyclase</td>
<td>Cyclase family</td>
<td><em>Desulfobacterium</em> sp. 4572_20</td>
<td>98,100</td>
</tr>
<tr>
<td>11</td>
<td>HS1_001358</td>
<td>Cyclase/aromtase</td>
<td>SRPBCC superfamily</td>
<td><em>Streptomyces vinaceus</em></td>
<td>44,96</td>
</tr>
<tr>
<td>12</td>
<td>HS1_001359</td>
<td>NAD(P)-dependent oxidoreductase</td>
<td></td>
<td><em>Bacillus</em> sp. NC2-31</td>
<td>35,96</td>
</tr>
<tr>
<td>13</td>
<td>HS1_001360</td>
<td>amidohydrolase</td>
<td>metall_dependant_hydrolase</td>
<td><em>Clostridium beijerinckii</em></td>
<td>34,100</td>
</tr>
<tr>
<td></td>
<td>HS1_001361</td>
<td>Glycosyl hydrolase family 76</td>
<td>DUF4135</td>
<td><em>Sarcina</em> sp. DSM 11001</td>
<td>35,89</td>
</tr>
</tbody>
</table>
Table 7.11: Organisation of the kra1 BGC from *Ktedonobacter racemifer*

<table>
<thead>
<tr>
<th>ORF</th>
<th>Gene designation</th>
<th>Putative function</th>
<th>homologue organism</th>
<th>Identity / coverage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>KraRS12430</td>
<td>group II intron reverse transcriptase</td>
<td><em>Cyanobacterium</em> sp. IPPAS B-1200</td>
<td>52, 99</td>
</tr>
<tr>
<td>2</td>
<td>KraRS12435</td>
<td>putative transposase</td>
<td><em>Alicyclobacillus</em> sp. USBA-503</td>
<td>66, 38</td>
</tr>
<tr>
<td>3</td>
<td>KraRS12440</td>
<td>Nuclear transport factor 2 / polyketide cyclase</td>
<td><em>Thermogemmatispora onikobensis</em> / <em>Multispecies pseudomonas</em></td>
<td>46, 80 / 27, 77</td>
</tr>
<tr>
<td>4</td>
<td>KraRS12445</td>
<td>alpha/beta hydrolase</td>
<td><em>Streptomyces</em> sp. PRh5</td>
<td>33, 83</td>
</tr>
<tr>
<td>5</td>
<td>KraRS12450</td>
<td>3-oxoacyl-[acyl-carrier-protein] synthase II</td>
<td><em>Dendrosporobacter querciclus</em></td>
<td>46, 93</td>
</tr>
<tr>
<td>6</td>
<td>KraRS12455</td>
<td>beta-ketoacyl-[ACP] synthase II</td>
<td><em>Planococcus maritimus</em></td>
<td>39, 97</td>
</tr>
<tr>
<td>7</td>
<td>KraRS65805</td>
<td>3-(3-hydroxyphenyl)propionate hydroxylase</td>
<td><em>Pseudomonas fulva</em></td>
<td>39, 94</td>
</tr>
<tr>
<td>8</td>
<td>KraRS12465</td>
<td>carbohydrate kinase, gluconokinase</td>
<td><em>Rokubacteria bacterium</em> CSP1-6</td>
<td>44, 97</td>
</tr>
<tr>
<td>9</td>
<td>KraRS12470</td>
<td>uracil-DNA glycosylase</td>
<td><em>Tolypothrix bouteillei</em></td>
<td>72, 99</td>
</tr>
<tr>
<td>10</td>
<td>KraRS12475</td>
<td>aldo/keto reductase</td>
<td><em>Nostoc</em> sp. 106C</td>
<td>63, 98</td>
</tr>
<tr>
<td>11</td>
<td>KraRS12480</td>
<td>cysteine desulfurase-like protein</td>
<td><em>Chloroflexi bacterium</em> RBG_16_57_11</td>
<td>62, 99</td>
</tr>
<tr>
<td>12</td>
<td>KraRS12485</td>
<td>hypothetical protein</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>KraRS12490</td>
<td>hypothetical protein</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>KraRS12495</td>
<td>No homology</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Supplementary Table 7.12: Organisation of the *kra2* BGC from *Ktedonobacter racemifer*

<table>
<thead>
<tr>
<th>ORF</th>
<th>Gene designation</th>
<th>Putative function</th>
<th>homologue protein / protein domains</th>
<th>homologue organism</th>
<th>Identity / coverage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Krac_RS42015</td>
<td>MFS transporter</td>
<td><em>Sulfobacillus acidophilus</em> DSM 10322</td>
<td>52, 97</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Krac_RS42020</td>
<td>hypothetical Protein</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Krac_RS42025</td>
<td>Transposase</td>
<td><em>Anaeromyxobacter</em> sp. PSR-1</td>
<td>55, 36</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Krac_RS42030</td>
<td>IS5 family transposase</td>
<td><em>Polromonas</em> sp. CG9 12</td>
<td>64, 96</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Krac_RS42035</td>
<td>hypothetical Protein</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Krac_RS42040</td>
<td>Tetracycline resistance ribosomal protection protein</td>
<td><em>Paenibacillus</em> sp. OSY-SE</td>
<td>63, 99</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Krac_RS42045</td>
<td>HrdB sigma factor protein</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Krac_RS42050</td>
<td>Ubiquinone/maenaquinone biosynthesis C-methyltransferase</td>
<td><em>Paenibacillus</em> sp. UNCA49MF</td>
<td>62, 97</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Krac_RS42055</td>
<td>IS6 family transposase</td>
<td><em>Rhodococcus</em> sp. YL-1</td>
<td>70, 89</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Krac_RS42060</td>
<td>IS3 family transposase</td>
<td><em>Moorella thermoacetica</em></td>
<td>56, 95</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Krac_RS42065</td>
<td>IS3 family transposase</td>
<td><em>Archangium violaceum</em></td>
<td>58, 64</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Krac_RS42075</td>
<td>acyl esterase</td>
<td><em>Legionella jordania</em></td>
<td>35, 96</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Krac_RS42080</td>
<td>antibiotic biosynthesis monoxygenase</td>
<td><em>Streptomyces coaingensis</em></td>
<td>43, 89</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Krac_RS42085</td>
<td>polyketide cyclase</td>
<td><em>Streptomyces</em> sp. NRRL B-3229</td>
<td>54, 95</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Krac_RS42090</td>
<td>cyclase</td>
<td><em>Streptomyces paucisporeus</em></td>
<td>43, 90</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Krac_RS42095</td>
<td>Acyl carrier protein / Ppant binding protein</td>
<td><em>Streptomyces</em> sp. AA0539</td>
<td>47, 63</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Krac_RS42100</td>
<td>Beta-ketoacyl-synthase II</td>
<td><em>Streptomyces</em> sp. CB02400</td>
<td>48, 98</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Krac_RS42105</td>
<td>Beta-ketoacyl-synthase</td>
<td><em>Streptomyces resistomycticus</em></td>
<td>54, 97</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Krac_RS42110</td>
<td>O-methyltransferase family 2</td>
<td><em>Leptolyngbya</em> sp. NIES-3755</td>
<td>33, 94</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Krac_RS42115</td>
<td>MFS transporter</td>
<td><em>Filbrilla</em> sp. ES10-3-2-2</td>
<td>29, 84</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Krac_RS42120</td>
<td>hypothetical Protein</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Krac_RS42125</td>
<td>pentapeptide repeat containing protein</td>
<td><em>Nostoc</em> sp. PCC 7107</td>
<td>54, 90</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Krac_RS42130</td>
<td>luciferase</td>
<td>*Actinobacterium 13_1_65_1240CM_4_</td>
<td>49, 98</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Krac_RS42135</td>
<td>hypothetical Protein</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Krac_RS42140</td>
<td>Transposase</td>
<td>*Candidatus Atribacteria bacterium RBG_16_35_8</td>
<td>37, 65</td>
<td></td>
</tr>
</tbody>
</table>
### Supplementary Table 7.13: Organisation of the kra BGC from *Ktedonobacter racemifer*

<table>
<thead>
<tr>
<th>ORF</th>
<th>Gene designation</th>
<th>Putative function</th>
<th>homologue protein / protein domains</th>
<th>homologue organism</th>
<th>Identity / coverage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Krac_RS46070</td>
<td>Hypothetical protein</td>
<td>RRXRR</td>
<td>Moorea bouillanii</td>
<td>47, 99</td>
<td></td>
</tr>
<tr>
<td>Krac_RS46075</td>
<td>serine hydrolase</td>
<td>beta-lactamase class C histidine transporter, GGDEF, HDc</td>
<td>Dehalococcoides maccartyi</td>
<td>39, 41</td>
<td></td>
</tr>
<tr>
<td>Krac_RS46080</td>
<td>diguanylate cyclase</td>
<td>FAD binding 3 superfamily</td>
<td>Acidobacteria bacterium 13_1_20CM_3_53_8</td>
<td>50, 89</td>
<td></td>
</tr>
<tr>
<td>Krac_RS46085</td>
<td>monoxygenase</td>
<td>methyl accepting chemotaxis protein</td>
<td>Thermogemmatispora onikobensis</td>
<td>49, 68</td>
<td></td>
</tr>
<tr>
<td>Krac_RS46090</td>
<td>HAMP domain-containing histidine kinase</td>
<td>cupin domain containing protein</td>
<td>Bacillus endophyticus</td>
<td>55, 99</td>
<td></td>
</tr>
<tr>
<td>Krac_RS46095</td>
<td>DNA-response regulator</td>
<td>OmpR</td>
<td>Candidatus Methanoperedens sp. BLZ1</td>
<td>34, 94</td>
<td></td>
</tr>
<tr>
<td>Krac_RS46100</td>
<td>alpha/beta hydrolase</td>
<td>SnoA-like PK cyclase</td>
<td>Burholderia sp. BT03</td>
<td>34, 86</td>
<td></td>
</tr>
<tr>
<td>Krac_RS46105</td>
<td>Hypothetical protein</td>
<td>UbI, dimerisation2 domain</td>
<td>Streptomyces resistomycificus</td>
<td>66, 93</td>
<td></td>
</tr>
<tr>
<td>Krac_RS46110</td>
<td>Hypothetical protein</td>
<td>UbI, dimerisation2 domain</td>
<td>Streptomyces resistomycificus</td>
<td>34, 94</td>
<td></td>
</tr>
<tr>
<td>Krac_RS46115</td>
<td>ester cyclase</td>
<td>UbI, dimerisation2 domain</td>
<td>Methanosarcina barkeri</td>
<td>35, 93</td>
<td></td>
</tr>
<tr>
<td>Krac_RS46120</td>
<td>dimethylmenaquinone methyltransferase</td>
<td>UbI, dimerisation2 domain</td>
<td>Methanosarcina barkeri</td>
<td>35, 93</td>
<td></td>
</tr>
<tr>
<td>Krac_RS46125</td>
<td>2-polyprenyl-6-methoxyphenol hydroxylase</td>
<td>UbI, dimerisation2 domain</td>
<td>Methanosarcina barkeri</td>
<td>35, 93</td>
<td></td>
</tr>
<tr>
<td>Krac_RS46130</td>
<td>RemF cyclase</td>
<td>UbI, dimerisation2 domain</td>
<td>Streptomyces resistomycificus</td>
<td>66, 93</td>
<td></td>
</tr>
<tr>
<td>Krac_RS46135</td>
<td>Cyclase</td>
<td>UbI, dimerisation2 domain</td>
<td>Streptomyces resistomycificus</td>
<td>66, 93</td>
<td></td>
</tr>
<tr>
<td>Krac_RS46140</td>
<td>O-methyltransferase</td>
<td>UbI, dimerisation2 domain</td>
<td>Streptomyces sp. NRRL B-3229</td>
<td>65, 98</td>
<td></td>
</tr>
<tr>
<td>Krac_RS46145</td>
<td>beta-ketoacyl synthase</td>
<td>UbI, dimerisation2 domain</td>
<td>Streptomyces sp. NRRL B-3229</td>
<td>65, 98</td>
<td></td>
</tr>
<tr>
<td>Krac_RS46150</td>
<td>beta-ketoacyl synthase</td>
<td>UbI, dimerisation2 domain</td>
<td>Streptomyces sp. NRRL B-3229</td>
<td>65, 98</td>
<td></td>
</tr>
<tr>
<td>Krac_RS46155</td>
<td>Holo-[ACP] synthase</td>
<td>UbI, dimerisation2 domain</td>
<td>Streptomyces sp. NRRL B-3229</td>
<td>65, 98</td>
<td></td>
</tr>
<tr>
<td>Krac_RS46160</td>
<td>acyl carrier protein</td>
<td>UbI, dimerisation2 domain</td>
<td>Streptomyces sp. NRRL B-3229</td>
<td>65, 98</td>
<td></td>
</tr>
<tr>
<td>Krac_RS46165</td>
<td>polyketide cyclase</td>
<td>UbI, dimerisation2 domain</td>
<td>Streptomyces sp. NRRL B-3229</td>
<td>65, 98</td>
<td></td>
</tr>
<tr>
<td>Krac_RS46170</td>
<td>ABC transporter ATP-binding protein</td>
<td>UbI, dimerisation2 domain</td>
<td>Streptomyces sp. NRRL B-3229</td>
<td>65, 98</td>
<td></td>
</tr>
<tr>
<td>Krac_RS63905</td>
<td>ABC transporter permease</td>
<td>UbI, dimerisation2 domain</td>
<td>Streptomyces sp. NRRL B-3229</td>
<td>65, 98</td>
<td></td>
</tr>
<tr>
<td>Krac_RS46180</td>
<td>23s rRNA</td>
<td>UbI, dimerisation2 domain</td>
<td>Streptomyces sp. NRRL B-3229</td>
<td>65, 98</td>
<td></td>
</tr>
<tr>
<td>Krac_RS46185</td>
<td>16s rRNA</td>
<td>UbI, dimerisation2 domain</td>
<td>Streptomyces sp. NRRL B-3229</td>
<td>65, 98</td>
<td></td>
</tr>
<tr>
<td>Krac_RS46190</td>
<td>Hypothetical protein</td>
<td>UbI, dimerisation2 domain</td>
<td>Streptomyces sp. NRRL B-3229</td>
<td>65, 98</td>
<td></td>
</tr>
<tr>
<td>Krac_RS46195</td>
<td>ferredoxin</td>
<td>UbI, dimerisation2 domain</td>
<td>Streptomyces sp. NRRL B-3229</td>
<td>65, 98</td>
<td></td>
</tr>
<tr>
<td>Krac_RS67810</td>
<td>cation transporter</td>
<td>UbI, dimerisation2 domain</td>
<td>Streptomyces sp. NRRL B-3229</td>
<td>65, 98</td>
<td></td>
</tr>
<tr>
<td>Krac_RS46205</td>
<td>S9 family peptidase</td>
<td>UbI, dimerisation2 domain</td>
<td>Streptomyces sp. NRRL B-3229</td>
<td>65, 98</td>
<td></td>
</tr>
</tbody>
</table>

323
<table>
<thead>
<tr>
<th>Gene</th>
<th>Description</th>
<th>Organism</th>
<th>Reference(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Krac_RS46210</td>
<td>transposase IS605 family protein</td>
<td><em>Scytonema</em> sp. NEW-4073</td>
<td>43, 84</td>
</tr>
<tr>
<td>Krac_RS46215</td>
<td>IS4/IS5 family transposase</td>
<td><em>Sphaerobacter thermophilus</em></td>
<td>57, 100</td>
</tr>
<tr>
<td>Krac_RS46220</td>
<td>no homology</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1'</td>
<td>Krac_RS46225</td>
<td>acyl carrier protein</td>
<td><em>Labrenzia alba</em></td>
</tr>
<tr>
<td>2'</td>
<td>Krac_RS46230</td>
<td>Holo-[ACP] synthase</td>
<td><em>Nevskia soli</em></td>
</tr>
<tr>
<td>3'</td>
<td>Krac_RS46235</td>
<td>acyl-CoA dehydrogenase</td>
<td><em>Nevskia soli</em></td>
</tr>
</tbody>
</table>
## Supplementary Table 7.14: Plasmid list

<table>
<thead>
<tr>
<th>Plasmid name</th>
<th>Insert Description</th>
<th>Backbone description</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>pETDuet-1</td>
<td>-</td>
<td></td>
<td>Novagen</td>
</tr>
<tr>
<td>pETM-11b</td>
<td>-</td>
<td>pBR322 ori, bla (Ap), T7 promoters</td>
<td>Novagen</td>
</tr>
<tr>
<td>pETM-11b AntB</td>
<td><em>plu4193</em> (AntB, PPTase)</td>
<td></td>
<td>This study</td>
</tr>
<tr>
<td>pHOLDKraAB</td>
<td>Refactored <em>kra</em> KS and CLF</td>
<td></td>
<td>This study</td>
</tr>
<tr>
<td>pHOLDDacAB</td>
<td>Refactored <em>dac</em>KS and CLF</td>
<td></td>
<td>This study</td>
</tr>
<tr>
<td>pHOLDSspAB</td>
<td>Refactored <em>ssp</em> KS and CLF</td>
<td></td>
<td>This study</td>
</tr>
<tr>
<td>pHOLDDauAB</td>
<td>Refactored <em>dau</em> KS and CLF</td>
<td></td>
<td>This study</td>
</tr>
<tr>
<td>pHOLDBendAB</td>
<td>Refactored <em>bend</em> KS and CLF</td>
<td></td>
<td>This study</td>
</tr>
<tr>
<td>pHOLDPluAB</td>
<td>Refactored <em>plu</em> KS and CLF</td>
<td></td>
<td>This study</td>
</tr>
<tr>
<td>pHOLDRemAB</td>
<td>Refactored <em>rem</em> KS and CLF</td>
<td></td>
<td>This study</td>
</tr>
<tr>
<td>pHOLDOvmPK</td>
<td>Refactored <em>ovm</em> KS and CLF</td>
<td></td>
<td>This study</td>
</tr>
<tr>
<td>pETDacAB</td>
<td>Refactored <em>dac</em>KS and His&lt;sup&gt;6&lt;/sup&gt;CLF</td>
<td></td>
<td>This study</td>
</tr>
<tr>
<td>pETDacB</td>
<td>Refactored <em>dac</em> His&lt;sup&gt;6&lt;/sup&gt;CLF</td>
<td></td>
<td>This study</td>
</tr>
<tr>
<td>pETSspAB</td>
<td>Refactored <em>ssp</em> His&lt;sup&gt;6&lt;/sup&gt;KS and CLF</td>
<td></td>
<td>This study</td>
</tr>
<tr>
<td>pETSspA</td>
<td>Refactored <em>ssp</em> His&lt;sup&gt;6&lt;/sup&gt;KS</td>
<td></td>
<td>This study</td>
</tr>
<tr>
<td>pETPluAB</td>
<td>Refactored <em>plu</em> KS and His&lt;sup&gt;6&lt;/sup&gt;CLF</td>
<td></td>
<td>This study</td>
</tr>
<tr>
<td>pETBendAB</td>
<td>Refactored <em>bend</em> KS and His&lt;sup&gt;6&lt;/sup&gt;CLF</td>
<td></td>
<td>This study</td>
</tr>
<tr>
<td>pETDauAB</td>
<td>Refactored <em>dau</em> KS and His&lt;sup&gt;6&lt;/sup&gt;CLF</td>
<td></td>
<td>This study</td>
</tr>
<tr>
<td>pETKraAB</td>
<td>Refactored <em>kra</em> KS and His&lt;sup&gt;6&lt;/sup&gt;CLF</td>
<td></td>
<td>This study</td>
</tr>
<tr>
<td>pETKraB</td>
<td>Refactored <em>kra</em> His&lt;sup&gt;6&lt;/sup&gt;CLF</td>
<td></td>
<td>This study</td>
</tr>
<tr>
<td>pETRemAB</td>
<td>Refactored <em>rem</em>KS and His&lt;sup&gt;6&lt;/sup&gt;CLF</td>
<td></td>
<td>This study</td>
</tr>
<tr>
<td>pETRemB</td>
<td>Refactored <em>rem</em> His&lt;sup&gt;6&lt;/sup&gt;CLF</td>
<td></td>
<td>This study</td>
</tr>
<tr>
<td>pETOvmPK</td>
<td>Refactored <em>ovm</em> KS and His&lt;sup&gt;6&lt;/sup&gt;CLF</td>
<td></td>
<td>This study</td>
</tr>
<tr>
<td>pETOvmK</td>
<td>Refactored <em>ovm</em> His&lt;sup&gt;6&lt;/sup&gt;CLF</td>
<td></td>
<td>This study</td>
</tr>
</tbody>
</table>
8 Final discussions, conclusions and future perspectives

We set out to expedite the synthetic biology led bioprospecting route from genes to specialised metabolites with a particular focus towards aromatic polyketides. In doing so we holistically considered the specialised metabolite bioprospecting pipeline and identified a suite of bottlenecks which can be circumvented or solved by adopting a synthetic biological philosophy i.e. abstraction, standardisation, modularisation and characterisation.

The synthetic biology toolbox has shown its worth in the field of specialised metabolites\(^1\,^2\), however to the most part these examples correspond to advances in individual molecular biological techniques e.g. refactoring BGCs\(^3\), design of regulatory parts\(^4\,^6\) and use CRISPR/CAS for appropriate chassis engineering\(^7\), and have not considered synthetic biology in its holistic form. In our pipeline from genes to specialised metabolites we consider all aspects which are fundamental to rapidly link genes to specialised metabolites, novel or otherwise, and identify 5 predominant bottlenecks (Figure 8.1). The first lies in the archaic reporting of specialised metabolite biosynthetic gene clusters and associated biosynthetic parts across all platforms e.g. journals, books, databases and repositories and the general ontologies used in the community (Figure 8.1, step 1).

8.1 Step 1: Setting standards in specialised metabolite discovery

Standardisation is a tenet which sits at the beginning of our genes to molecules pipeline and represents the first bottleneck tackled within this work.

The design and implementation of a community supported standard for reporting biosynthetic gene clusters makes headway towards the synthetic biology of specialised metabolite biosynthesis, to which there was no unified publically available parallel prior to the minimum information about ab biosynthetic gene cluster (MIBiG) (Chapter 3). Forming a centralised repository in which well curated information about BGCs is deposited is a useful resource for the bioengineer, the specialised metabolite chemist, the ecologist, the pharmacologist etc.. The MIBiG standard and its accompanying repository are tackling bottlenecks in rapid retrieval of BGC relevant information. By removing this bottleneck we aimed to accelerate prototyping of new BGCs \textit{de novo} by describing a suite of characterised vibrant catalytic enzymes which can foreseeably be introduced into BGC BioCAD software, as well as more generally providing a useful resource to expedite specialised metabolite discovery using phylogenomics and by dereplication i.e. through integration of up-to-date reference dataset in genome mining tools (Chapter 4). Acceptance of our standard by the community is integral for the maintenance of MIBiG, and was key a motivation behind the community involvement in the project. Continued support of the MIBiG standard can be qualified by the growth of the MIBiG repository, which now houses 1416 entries (Sept, 2016) exceeding 1183 in June 2015, as well as using citations as a metric. The growth of 233 entries over a 16 month period is impressive; however MIBiG is not yet capturing all characterised BGCs\(^8\). An additional call is needed for existing, and new, databases to conform to the MIBiG standard prior to deposition to facilitate interoperability. Despite serving a similar purpose, databases housing specific classes of specialised
metabolites e.g. the ProCarDB indexing bacterial carotenoids, makes no reference to MIBiG standard.

Most MIBiG entries in our seed data set were prokaryotic in origin however the MIBiG standard has inspired a comprehensive curation of more than 190 published fungal BGCs. This has contributed to the breadth of specialised metabolites within the MIBiG repository, improved the visibility of the MIBiG standard in the Eukaryotic specialised metabolite field and augmented synthetic biology approaches to fungal specialised metabolite biosynthesis. These contributions are helping realise the function of new fungal specialised metabolites in natura and systematically define the phylogenetic distribution of fungal BGC homologues. The MIBiG standard was constructed to support BGC metadata from all Kingdoms and the advent of plantiSMASH will undoubtedly help populate the MIBiG repository further.

Centralising BGC metadata is impacting bioprospecting, exemplified in Chapter 5 but also more widely in the literature. MIBiG acquired datasets have been used to demonstrate the biosynthetic capacity of entire Genera, or Species, linking genes to ecology, helped redefine the modularity of specialised metabolite biosynthetic enzymology e.g. collinear acting polyketide synthases, and also validate the accuracy of new bioinformatics tools e.g. the RiPP structural predictor RiPP-PRISM, the PKS genome mining tool SeMP, and a peptide natural product dereplicator pipeline. There is no doubt that the MIBiG standard is holistically beneficial for all fields associated with specialised metabolites, however, to realise the benefits of standardisation the utility of MIBiG needs to be ubiquitous. A community curation effort must be established to ensure the data contained within the repository is up-to-date. This can be achieved by an employee, or more realistically by the charitability of researchers utilising the MIBiG repository.

Currently all entries within the MIBiG repository described native BGCs. A powerful next step towards bioengineering BGCs using BioCAD would be an extension of the MIBiG standard to report and visualise synthetic BGCs designed and constructed using the synthetic biology toolbox e.g. the reconstituted hydrocodone BGC which comprised 23 sequentially functioning enzymes derived from mammals, yeast, plants and bacteria. The repository of synthetic BGCs will provide critical information about biosynthetic parts in a suite of contexts and ameliorate the promiscuity of and interplay of each part, drastically reducing the attrition rate of combinatorial biosynthetic experiments. One could imagine implementation of both the native and synthetic MIBiG repositories into BioCAD pipelines which identify plausible retrosynthetic routes to metabolites of interest e.g. RetroPath for the rapid and automated design of novel biosynthetic pathways.

The second two bottlenecks in our specialised metabolite discovery pipeline consider the need to prioritise the thousands of BGCs predicted by in silico genome mining tools for experimental characterisation (Figure 8.1, step 2) and how to achieve this in a high throughput manner (Figure 8.1, step 3 and 4). These bottlenecks are broadly applicable for all specialised
metabolites and through solving each we unlock and hope to accelerate the design-build-test-learn cycle for all classes of specialised metabolite.

8.2 Prioritisation of the masses

Standardisation of ontologies, syntax and semantics is a powerful tool for effectual databasing, indexing and cataloguing information. To efficiently make use of this well curated information one must appropriately filter, prioritise and rank standardised entries using appropriate descriptors. For biosynthetic gene clusters this is not a trivial task, however, and few dedicated software pipelines are able to do so in a high throughput manner, despite the explosion in specialised metabolite associated bioinformatics tools available\textsuperscript{22}. The only published dedicated prioritisation platform for BGCs is the antibiotic resistance target seeker (ARTS\textsuperscript{23}), which prioritises BGCs using cluster associated resistance mechanisms to infer activity. Whilst this approach has been previously been fruitful e.g. in identification of the thiolactomycin BGC\textsuperscript{24}, a type II fatty acid synthase inhibitor\textsuperscript{25} and the pekiskomycin glycopeptide BGC\textsuperscript{26}, it does not consider the complexities of corresponding molecular biological experiments necessary for characterisation of biosynthetic pathways. Attrition and failure of wet lab experiments is a phenomenal bottleneck in the route to specialised metabolite discovery and therefore considering factors which improve successful characterisation in the laboratory is an imperative next step. The Output Ordering and Prioritisation System (OOPS) aims to fill this gap, enabling users to prioritise large datasets of BGCs in parallel, consideration of factors underpinning molecular biological complexities when rebuilding BGCs, maintaining BGCs in heterologous hosts and expressing large recombinant genetic constructs in non-native organisms.

Whilst the virtues of OOPS as a stand-alone system are exemplified in Chapter 4, there is precedence for OOPS to function as a ‘bolt-on’ for software pipelines prioritising, ranking or sorting BGCs upon predicted specialised metabolite properties. The fervent need for new antimicrobials is exemplified by remarkable epidemiological statistics of carbapenem-resistant organisms in Southeast Asia: more than 60\% of \textit{Acinetobacter baumanii} infections reported in Thailand and Pakistan, and >10\% of \textit{Enterobacteriaceae} clinical isolates from Indian and Pakistani patients were resistant to meropenem and other carbapenems\textsuperscript{27}. Epidemiological studies such as these cements ARTS as an important tool when targeting BGCs for characterisation as it is often difficult to determine the bioactivities of specialised metabolites \textit{a priori}. Using a directional approach to BGC prioritisation may result in characterisation of fewer specialised metabolites with disappoint bioactivities however experimental attrition rates will remain high. To this end OOPS can sit between the output of ARTS and the molecular / synthetic biologist and facilitate an additional prioritisation step, refining the dataset of ‘interesting’ BGCs further.

In addition to molecular ‘target seeking’, the use of ecological niches as a proxy to infer specialised metabolite function is a progressive and targeted method to improve the frequency of bioactive specialised metabolite hits per organism, and is currently receiving significant attention\textsuperscript{28} e.g. the identification of antifungal agents from fungi supressing soils\textsuperscript{29}, and ant commensal bacteria\textsuperscript{30}. By ranking BGCs using the phylogenetic diversity parameter OOPS
enables one to build and visualise ecological relationships between certain BGCs. Currently this must be carried out manually, however it is foreseeable that biotic interactions, growth characteristics and requirements, biogeographical distribution and other isolation, organism and environmental metadata covered by the MiXs specification can be retrieved from online databases e.g. GloBI, JGI GOLD, BioCyc etc. and integrated into OOPS providing a high-resolution genes-to-organism-to-ecology map.

OOPS is currently limited to prioritisation based on 8 parameters, none of which detail the enzymology within BGC beyond the class level. Refining BGCs through identification of specific enzymatic Pfam domains, or specialised metabolite clusters of orthologous groups (smCOGs), would enable the user to prioritise BGCs encoding specialised metabolites with specific chemical characteristics e.g. halogenation, known to have profound effects on specialised metabolite bioactivity and potency. Furthermore, this approach could be extended to ranking BGCs based upon the presence of larger chemical moieties of interest. A systematic analysis of characterised and predicted BGCs from more than 4000 prokaryotic derived DNA sequences (Del Carratore et al. in press) is validating the presences of thousands of biosynthetic ‘bricks’ or modules often with prescribed functions linked to biosynthesis of specific moieties. One could imagine prioritising BGCs based upon the presence of absence of single or multiple ‘bricks’ or modules would facilitate identification of specialised metabolites with characteristic functions e.g. the warhead moiety of an enediyne type I polyketide with the specific DNA localisation properties of sugars. Construction of designer molecules de novo routinely fails (Chapter 2), therefore identification of evolutionarily honed molecules with desirable characteristics circumvents the engineering challenges of pathway recapitulation or alternatively may identify crucial ‘mortar’ genes necessary to ‘glue’ together the two chemical moieties of interest.

The current version of OOPS is a stepping stone in the path to a more comprehensive prioritisation tool needed to effectively sort the ‘tsunami’ of sequence data available today. Judicious prioritisation of BGCs is extremely important when progressing in silico predictions to characterised specialised metabolites with useful functions, and it is remarkable that there are not more tools which effectively perform this function.

The last two bottlenecks in our specialised metabolite discovery platforms are more specific to the field of aromatic polyketides.

8.3 Accessing aromatic polyketides in Escherichia coli

The polyketides represent an exceptional pool of specialised metabolites and as such biosynthetic routes have been greatly explored. As a first step in specialised metabolite discovery biosynthetic machinery is routinely expressed in an amenable heterologous host, however prior to the work demonstrated in thesis, recapitulation of type II PKS BGCs in the workhorse organism Escherichia coli has been largely unachievable. Complete insolubility of the type II PKS was cited as the predominant bottleneck (Figure 8.1, step 7) and accordingly less genetically tractable Actinobacterial hosts were needed, slowing the test and learn steps of the design, built, test and learn cycle.
Using the standardised and easily retrievable datasets from the MIBiG repository we identified a ketosynthase and chain length factor pair from *Photorhabdus luminescens* as soluble type II PKS components in *E. coli* immediately, and by doing so overcame a hurdle which has persisted in the literature for the last 30 years. We constructed a plug-and-play scaffold using the *ant* biosynthetic machinery which can easily be adapted to synthesis a range of pharmaceutically relevant specialised metabolites in *E. coli* through expedition of the design built test learn cycle (Figure 8.1, step 7).

Building and testing vast arrays of biosynthetic pathways with slight genetic perturbations in an automated fashion, driven by synthetic biology principles, has been repeatedly realised for *Escherichia coli* however this is not paralleled in Actinobacteria. For instance, using a multiplex iterative plasmid engineering strategy to alter Shine-Dalgarno sequences in the ribA-E operon, Li *et. al.* (2013) improved production of riboflavin by *E. coli* MG1655 2.67 fold in less than 1 week. Furthermore, the same engineering strategy was used to simultaneously manipulate the translation frequencies of 24 chromosomal gene sequences, improving yields of lycopene in *E. coli* five-fold in just three days, and generated as many as 15 billion genetic variants. The *ant* plug-and-play platform could be easily optimised using either of these strategies to vastly improve the aromatic polyketide yields, however the same could not easily be achieved for the *act* BGC in *S. coelicolor*. A recently described artificial gene operon assembly (AGOS) system shows promise for the deconstruction and reconstruction of BGCs in *Streptomyces* spp. however suffered from poor production efficiencies: AGOS refactored novobiocin BGCs yielded just ~40 µg/L by comparison to ~36 mg/L from wild type producers. Multiplexing BGC designs in Actinobacteria is additionally confounded by the lack of well characterised biosynthetic parts. The refactored novobiocin BGC utilised the same tetracycline inducible promoter upstream of each operon, reducing degrees of freedom during reconstruction.

In addition to the wealth of wet lab tools, our exceptional knowledge of the *E. coli* metabolism has facilitated development of accurate metabolic models. Using metabolic models one is able to rationally manipulate *E. coli* metabolism in silico facilitating rapid chassis engineering for multiple aims e.g. robustness – improving host resistance to a specialised metabolite end compound, or titres – ensuring substrates and cofactors are available and not diverted towards primary metabolism, etc.. Expression of genes underpinning these processes can be concurrently optimised using the multiplexing strategies described above thereby simultaneously engineering the chassis and the biosynthetic pathway. Metabolic models exist for some *Streptomyces* spp. however ambiguity in genomic annotation currently mires the utility of these tools. Whilst important steps are being made in this direction e.g. for enhanced rapamycin and FK506 production, there is often a significant lag between development and utility of new tools in *E. coli* and their translation to the Actinobacteria.

The work described here is predominantly qualitative. Therefore, to identify if the biotechnological virtues of *E. coli* outweigh the production capacity of some Actinobacterial a
direct comparison of economic costs, time and yield must be undertaken and will be considered as next steps in this work.

8.4 Bioprospecting for further soluble type II PKS components

Regardless of the high throughput capacity to manipulate biosynthetic pathways afforded by E. coli, a plug-and-play scaffold is of limited use if few compatible biosynthetic parts are available (Figure 8.1, step 5). We therefore demonstrated the robustness of the ant minimal PKS to function when complemented with biosynthetic parts from the act BGC, derived from an evolutionarily distant host, and different lineage of type II PKSs. Remarkably, ant ACP (AntF) successfully interface with biosynthetic enzymes outside of the BGC in which it has been coevolving. The ability to successfully substitute ACP between Actinobacterial type II PKS BGCs is well documented50-53, and we show this characteristic extends to AntF: it is unlikely that the observed promiscuity only extends to biosynthetic enzymes from the act cluster, and therefore probable that AntF will also be able to successfully deliver nascent polyketide chains to a gamut of characterised type II PKS cyclases from Actinobacteria, directing biosynthesis towards U- J- and S-shaped polyketide chemotypes.

Whilst octaketides are of pharmaceutical relevance e.g. antiparasitic agent frenolicin B54, the most clinically used type II PKSs are decaketides55-57. Accessing decaketides in E. coli could be achieved by rational engineering of the ant KS/CLF heterodimer (Chapter 6), or by bioprospecting for new biosynthetic parts which also show solubility as recombinant proteins in E. coli (Chapter 7). By answering either of these aims we remove the bottlenecks in BGC design (Figure 8.1, step 5) and facilitate access to a more diverse pool of aromatic polyketide core scaffolds to further derivatise. We took small steps towards engineering the ant KS/CLF, and this remains for focus of on-going work, however succeeded in identification of a suite of additional soluble recombinant KS/CLF heterodimers in E. coli. Specifically 4 additional KS/CLF pairs were identified from Streptococcus sp. GMDXs, Pseudoalteromonas luteoviolacea DSM6061, Ktedonobacter racemifer DSM44963 and Streptomyces resistomycificus. The latter of which is a characterised decaketide synthase, and all of which are currently being functionally evaluated in E. coli. In our quest to identify new biosynthetic parts for the ant plug-and-play scaffold we describe a new and predominantly non-Actinobacterial lineage of type II PKSs for the first time, which possess remarkably soluble characteristics when comparison to the fastidious canonical Actinobacterial KS/CLF complexes.

The interoperability of biosynthetic parts from the predominantly non-Actinobacterial lineage of type II PKSs remains to be seen, however this pool of enzymes comprises potential ‘constituent parts’ necessary to reconstruct BGCs for production of clinically useful tetracyclines and doxorubicin analogues. Should biosynthetic routes to these compounds be realised in E. coli we begin to introduce diversity to these privileged polyketide core scaffolds in a combinatorial and automated fashion competing and superseding the fully synthetic and semi-synthetic routes currently in operation58,59. Using a similar synthetic biology led strategy one can imagine developing routes to useful bioactive compounds for which no pathways has been characterised e.g. towards the antidepressant, antiviral, anti-carcinogen and photodynamic therapy agent
hypericin\textsuperscript{60-62}, a polycyclic naphthodianthrone polyketide constituent of the \textit{Hypericum} Genus. There is no doubt that the plug-and-play platform described here-in is within its infancy, yet we feel it shows great promise as an engine to develop the next generation of useful aromatic polyketides.
Figure 8.1

A refreshed perspective pipeline for specialised metabolite discovery

The minimum information about a gene cluster reporting standard tackles bottlenecks at Step 1. Bottlenecks in prioritisation of BGCs in a highly paralleled manner and refining BGC datasets are tackled by OOPS (Steps 2, 3 and 4). Identification of a suite of soluble KS/CLF biosynthetic parts in E. coli unlocks this biotechnologically virtuous organism as a heterologous host for type II polyketide biosynthesis, expediting the design-build-test-learn synthetic biology cycle, unblocking bottlenecks at steps 5 and 7.
In summary, herein we present and exemplify a suite of tools and platforms which act successively to expedite identification, construction and testing of biosynthetic routes to new and characterised specialised metabolites, with a specific focus towards the combinatorial biosynthesis of aromatic polyketides. We describe efforts to remove key sequential bottlenecks confounding the implementation of synthetic biology principles when moving from genes to molecules, and set a new precedence in the field of aromatic polyketide biosynthesis. The work here-in migrates the field of type II polyketides from the actinobacterial specialist to the domain of any molecular biologist.

The synergistic use of the MIBiG standard, OOPS prioritisation pipeline and plug-and-play scaffold for aromatic polyketide biosynthesis in *E. coli* underpin the potential for high throughput automation of combinatorial biosynthetic strategies for the bioproduction of next generation therapeutics, elevating this discipline from genetic engineering to synthetic biology. The next milestones for this pipeline will be to demonstrate scalability. In the vein of type II polyketides we aim to produce next generation tetracyclines with extended utility as antimicrobials, and new doxorubicin anticancer analogues with reduced cardiotoxicity.

**Final Thoughts**

Concerted global efforts are necessary to tackle the looming threat of pan-resistant organisms: it is naïve in the extreme to think there is a single solution to this threat. For a satisfactory outcome, a multifaceted approach must consider socio-economic factors, epidemiology of pan-resistant organisms, the prudent use of antibiotics in prophylaxis, education of the populace, FDA approval requirements and public funding contributions, etc. as well as effective routes to the discovery, de-replication and evaluation of specialised metabolites as new medicines. The work here-in should be considered as a small but meaningful contribution to this aim.

### 8.5 References


