Bifurcation drives the evolution of assembly-line biosynthesis

Thomas J. Booth, Kenan A. J. Bozhüyük, Jonathon D. Liston, Sibyl F. D. Batey, Ernest Lacey & Barrie Wilkinson

Reprogramming biosynthetic assembly-lines is a topic of intense interest. This is unsurprising as the scaffolds of most antibiotics in current clinical use are produced by such pathways. The modular nature of assembly-lines provides a direct relationship between the sequence of enzymatic domains and the chemical structure of the product, but rational reprogramming efforts have been met with limited success. To gain greater insight into the design process, we wanted to examine how Nature creates assembly-lines and searched for biosynthetic pathways that might represent evolutionary transitions. By examining the biosynthesis of the anti-tubercular wollamides, we uncover how whole gene duplication and neofunctionalization can result in pathway bifurcation. We show that, in the case of the wollamide biosynthesis, neofunctionalization is initiated by intragenomic recombination. This pathway bifurcation leads to redundancy, providing the genetic robustness required to enable large structural changes during the evolution of antibiotic structures. Should the new product be non-functional, gene loss can restore the original genotype. However, if the new product confers an advantage, depreciation and eventual loss of the original gene creates a new linear pathway. This provides the blind watchmaker equivalent to the design, build, test cycle of synthetic biology.
Microbial natural products produced by modular biosynthetic assembly-lines, i.e. (type I) polyketide synthases (PKSs) and non-ribosomal peptide synthetases (NRPSs), are an important source of pharmaceutical and agrochemical agents. Examples include well known antibacterial molecules such as the polyketide insecticide spinosyn\(^2\) and NRPS derived penicillins\(^1\). Importantly, biosynthetic assembly-lines provide thousands of natural product scaffolds, including many of our essential clinical agents.

Essentially, NRPS and PKS modular megasynth(etas)es give rise to highly functionalised biopolymers from a broad variety of monomers, referred to as extender units. Hundreds of extender units have been reported, typically derived from malonate in the case of PKS\(^3\) or amino acids in the case of NRPSs\(^4,5\). They are likened to assembly-line processes due to their hierarchical and modular structures in which multiple, repeating modules of enzymatic domains catalyse the incorporation of an extender unit into the growing chain, along with any programmed additional enzymatic domains that can activate substrates with such varying physico-chemical properties before transferring the elongated chain to the next module. The archetypical minimal assembly-line module consists of three core domains. Firstly, a domain for the selection and activation of an extender unit, the acyltransferase (AT) domains for PKSs or adenylation (A) domains for NRPSs. The activated substrate is then covalently attached to a prosthetic phosphopantetheine group of a small acyl carrier protein (ACP; PKSs) or peptidyl carrier protein (PCP; NRPSs) domain. Finally, the ketosynthase (KS; PKSs) or condensation (C; NRPSs) domains then link the covalently bound substrates to the growing polyketide or peptide chain. Although the exact mechanisms and ancillary domains of PKSs and NRPSs differ, the fundamental principal is that modules condense covalently bound substrates in a linear fashion. The inherent logic of this mechanism means that there is a direct relationship between the sequence of domains in an assembly line and the chemical structure of the resulting molecule\(^6,7\). In principle this relationship enables the prediction of natural product chemical structures directly from DNA sequences. In turn, this logic has inspired efforts to rationally reprogramme assembly lines to produce tailor-made molecules.

Numerous examples of assembly-line engineering have been reported; however, many display productivities well below that of the parental (wild-type) system. Insights into structural flexibility, inter-domain communication and the role of proof-reading by catalytic domains pre-empted novel strategies to engineer assembly-line proteins\(^8\)–\(^12\). There is also an increasing body of evidence suggesting that we might further improve our ability to engineer these systems if we had a better understanding of their evolution\(^13\)–\(^18\).

With this latter point in mind, we have been searching for biosynthetic pathways that may represent transitional evolutionary states and provide exemplar systems to inform future work. Natural selection acts upon phenotypes, yet even a small change to the structure of a natural product can have profound effects on bioactivity and, thus, the fitness of the producing organism. Therefore, we hypothesised that strains encoding BGCs evolving new functionalities might be expected to produce multiple related products (congeners) (Fig. 1a). Here, we describe the BGC encoding of the wollamide-desotamide family of antibiotics, which represents an evolutionary snapshot of a modular NRPS assembly-line.

The wollamides are cyclic hexapeptides (Fig. 1b) that were disclosed in 2014\(^19\). The only reported producer of the wollamides is Streptomyces sp. MST-115088 and they exhibit potent antimycobacterial activity that has attracted the attention of synthetic medicinal chemists\(^20\)–\(^23\). Along with the wollamides, Streptomyces sp. MST-115088 produces a related group of hexapeptides, the desotamides\(^24\). The desotamides and wollamides share a common peptide scaffold, except for a single residue change from glycine in the desotamides to ω-ornithine in the wollamides (Fig. 1b). It is important to note that the previously identified desotamide producer, Streptomyces scopuliridis SCSIO ZJ46, is not reported to produce wollamides and the desotamide (dsa) BGC follows canonical NRPS logic\(^25\). The NRPS is encoded by three genes (dsaI, dsaH and dsaG) encoding two modules each. Importantly, module six, encoded by DsaG, incorporates glycine as the final amino acid. The ability of Streptomyces sp. MST-115088 to produce congeners with ω-ornithine and glycine in the same position is therefore difficult to rationalise as, under a canonical NRPS mechanism, A-domains are responsible for the selection and activation of specific amino acid substrates. While A-domains can activate structurally related amino acids giving rise to families of structurally similar congeners (for example the combinations of valine, leucine, or allo-leucine at positions 3 and 4 of the wollamides-desotamides), the ability of an A-domain to activate substrates with such varying physico-chemical properties as glycine and ornithine is without precedent. Therefore, to explain the production of wollamides and desotamides by a single strain we hypothesised three scenarios (Fig. 1a): dual specificity of module six for glycine and ornithine; duplicated genes encoding module six homologues, each specific to glycine or ornithine, respectively; or duplicated BGCs where, as above, the final modules of each BGC are specific to glycine or ornithine.

Herein, we describe our combined bioinformatics, in vivo engineering, and biochemical analysis demonstrating how the dsa BGC evolved from an ancestral wollamide producing BGC via the process of gene duplication and intragenomic recombination with a second NRPS encoding locus. This allows us to propose an update to the current model for the evolution of assembly-line BGCs through which duplication and the resulting bifurcation reduce selective pressure and drive the evolution of new functions.

**Results**

Desotamides and wollamides are produced by a bifurcated NRPS assembly-line. The genome of Streptomyces sp. MST-115088 was sequenced using Pacific Biosciences RS2\(^26\) single-molecule technology and assembled using the HGAP3\(^27\) pipeline to generate a single 7.9 Mb chromosomal assembly (Supplementary Table 1) (GenBank: CP074380). Analysis of the assembly using antiSMASH v4.0\(^8\) allowed us to rapidly identify the wollamide (wol) BGC (Supplementary Table 2), which was then compared to the desotamide (dsa) BGC previously reported from Streptomyces scopuliridis SCSIO ZJ46 (GenBank: KP769807) (Fig. 2a, b). Five additional desotamide producers (MST-70754, MST-71458, MST-71321, MST-94754 and MST-127221) were identified through metabolomic analysis of Microbial Screening Technologies’ unique collection of more than fifty thousand Australian actinomycete strains (Supplementary Information). The desotamide BGCs (Supplementary Tables 3–7) of these strains display functionally identical architectures (GenBank: MZ093610, MZ093611, MZ093612, MZ093613, MZ093614) to that reported from Streptomyces scopuliridis SCSIO ZJ46\(^25\) (>95% nucleotide identity) (Supplementary Fig. 1). The wol BGC displays a similar architecture to the dsa BGCs but contains two genes wolg1 and wolg2 that are duplicates of the dsaG gene that encodes the final two modules of the NRPS. Similarly, there are duplicated homologues of the dsaF gene (wolF1 and wolF2), which encodes an MbtH-like protein involved in A-domain functionality\(^28\)–\(^29\). The wol BGC also contains six additional genes wolRSTUV whose gene products are predicted to be involved in the biosynthesis of ω-ornithine consistent with the presence of ornithine in position 6 of the wollamides.
During desotamide biosynthesis, DsaG is responsible for the final two rounds of peptide chain elongation. We hypothesised, therefore, that WolG1 and WolG2 may encode two forks of a bifurcated biosynthetic pathway where the first four rounds of peptide elongation proceed via the colinear activity of WolI and WoH, but the final two elongation steps are catalysed either by WolG1 or WolG2, yielding desotamide or wollamide products, respectively (Fig. 2c). Consistent with this hypothesis, in silico analysis of A-domain specificities predicted the substrates for the final A-domains of WolG1 and WolG2 (henceforth WolG1A2 and WolG2A2) to be glycine and l-ornithine (Supplementary Table 8).

**Engineering wollamide production in a desotamide-only producing strain.** To confirm our biosynthetic hypothesis and to explore the evolutionary relationship between the wollamide and desotamide pathways, we sought to engineer the co-production of wollamides into a desotamide-only producing strain through the heterologous expression of wolG2.

As NRPS genes are typically large (for reference, wolG2 is 7.9 kb) and difficult to clone through conventional methods, we generated pBO1 (Supplementary Fig. 2), an integrative Streptomyces expression vector capable of propagating in both Escherichia coli and Saccharomyces cerevisiae. This allows larger genes to be assembled efficiently in yeast via transformation.

---

**Fig. 1 Intermediate chemotypes in the evolution of natural product biosynthetic gene clusters (BGCs).** A cartoon demonstrating the three main models of gene cluster evolution depicting the transition of a fictitious BGC from chemotype A to chemotype B (coloured circles). Evolutionary processes are represented by dashed lines. The structural diversity of the wollamides and desotamides. The d-ornithine and glycine residues, which define the wollamides and desotamides, respectively, are highlighted. The table on the right-hand side shows the variable positions of the various wollamide and desotamide congeners (NFK N-formyl kynunerine, Wol wollamide, Dsa desotamide).
associated homologous recombination downstream of a constitutive promotor and subsequently transferred from E. coli to Streptomyces spp. through conjugal transfer (Supplementary Information). To assemble a wolG2 expression plasmid, pBO1 along with target fragments amplified from genomic DNA were transformed into S. cerevisiae CEN.PK 2-1C. The resulting plasmid pBO1-wolG2 was transformed into the desotamide producer Streptomyces sp. MST-70754 via conjugation. Streptomyces sp. MST-70754 and its progeny carrying pBO1-wolG2 were grown in triplicate under desotamide producing conditions.

Fig. 2 The genetic basis for bifurcated biosynthesis of desotamide and wollamide by Streptomyces sp. MST-110588. a The architecture of the desotamide producing dsa BGC from Streptomyces scopuliridis SCSIO ZJ46. b The architecture of the wollamide and desotamide producing wol BGC from Streptomyces sp. MST110588. c The proposed biosynthetic pathway of the wollamides and desotamides. The first three condensations are catalysed by WolI and WolH. The final two condensations are catalysed by WolG1 to produce desotamides or WolG2 to produce wollamides.
The presence of both wollamide and wollamide producer (desotamide: [M+H] fragmentation of the compounds produced by the native desotamide congeners was con-...showing the potential for the evolution of assembly-lines 1,42. More specifically, given the reduced rate of horizontal gene transfer between distantly related taxa and the high rate of heterogeneity between recombinant sequences, it has been speculated that intragenomic recombination within ancestral strains can provide opportunities for assembly-line diversification18,43. To assess the possible role of intragenomic recombination in the evolution of the wol BGC we generated a nucleotide sequence alignment of all thirty-four NRPS-associated adenylation domains present in the Strepto-...tionship between the conserved motif A2 (N-terminal) and motifs A5 and A6 (C-terminal) of the A-domain, thus comprising the flavodoxin subdomain18,39,41 and a large portion of the N-terminal subdomain (Fig. 4d). Crucially, this would allow for the substitution of the amino acid binding pocket and catalytic P-loop while maintaining the C-A-domain interface (Supplementary Fig. 13). This pattern is seen frequently in our dataset (Supplementary Fig. 12) and may suggest an advantage for maintaining the structural relationships between the P-loop and substrate binding pocket.

Importantly, the adenylation domain encoded by orf6595 was predicted in silico to select for glycine (Supplementary Table 8) meaning that the predicted recombination event could theoretically convert a wollamide producing assembly line into a desotamide producer.

Biochemical validation of A-domain substrate specificities. To gain insight into the function of the minor parent BGC, the sequences of Orf6595 and proteins encoded by the surrounding conditions and mehanic culture extracts were then analysed by LC/MS in comparison to the wollamide producer Streptomyces sp. MST-110588 (Fig. 3). The presence of both wollamide and desotamide congeners was confirmed for the engineered strain by comparison of retention time, isotopic masses and MS/MS fragmentation of the compounds produced by the native wollamide producer (desotamide: [M+H] = 697.4047, [M+Na] = 719.382; wollamide A [M+H] = 754.4655, [M+Na] = 776.4419) (Fig. 3) (Supplementary Figs. S3–S5).

This engineering strategy relied on the assumption that the docking domains (DDs) of DsaH and WolG2 that mediate NRPS interactions would still function as a pair35,36. Production of wollamide A by Streptomyces sp. MST-70754/pBo1-wolG2 shows that this interaction is still possible; however, the yields are relatively low (approximately 50-fold lower than Streptomyces sp. MST-110588, Fig. 3c, Supplementary Fig. 3). Based on this observation, we hypothesised that some depreciation of the DD interaction may have occurred.

To investigate this, we compared the N-terminal DD primary sequences of WolG1 and WolG2 (Supplementary Fig. 6), as well as calculated homology models of the DDs in complex with the WolH DD (Supplementary Fig. 7). Alignment of the DDs revealed a glutamate to alanine mutation at position (E16A). Homology modelling with the previously characterised PaxB docking domain (6TRP_1)37 revealed that the E16A amino acid change in WolG2 leads to the loss of a salt bridge formed with R3504 from WolH. Although other factors may play a role, the importance of this salt bridge in DD interactions suggests that this mutation is likely to account for the observed differences in yield.

Evolution of adenylation domain specificity via intragenomic recombination. The unusual architecture of the wol BGC led us to consider potential mechanisms for its evolution. The high similarity of wolG1 to wolG2 (80.9%) and wolG1/wolG2 to dsaG (74.3% and 71.2% respectively) is indicative of an ancestral gene duplication event38. Additionally, wolG1, wolG2 and dsaG share common patterns of nucleotide skew, a phenomenon indicative of recent divergence. However, despite this high similarity, there is a notable drop in nucleotide identity within the region coding for the final adenylation domains which is also manifest in the gene products (54.6% nucleotide and 29.2% protein sequence similarity) (Supplementary Fig. 8). Given the overall similarity of these genes, we deemed it unlikely that such high sequence variation could emerge through the accumulation of point mutations alone. Furthermore, phylogenetic reconstructions of A-domains hinted at independent evolutionary histories when compared to the rest of the assembly line (Supplementary Figs. S9–S11). Similar patterns have been observed in other bacterial NRPS clusters18,39–41.

Many studies have highlighted the role recombination, e.g. the exchange of alleles via a double recombination event, plays in the evolution of assembly-lines1,42. More specifically, given the reduced rate of horizontal gene transfer between distantly related taxa and the high rate of heterogeneity between recombinant sequences, it has been speculated that intragenomic recombination within ancestral strains can provide opportunities for assembly-line diversification18,43. To assess the possible role of intragenomic recombination in the evolution of the wol BGC we generated a nucleotide sequence alignment of all thirty-four NRPS-associated adenylation domains present in the Strepto-...
genes were searched against the MIBiG database\textsuperscript{45} using clustal\textsuperscript{46}. The search identified the BGC as homologous to the rimosamide (\textit{rmo}) BGC from \textit{Streptomyces rimosus} ATCC 10970\textsuperscript{47} (Supplementary Table 10). More specifically, \textit{Orf6595} is a homologue of \textit{RmoG} which encodes a single NRPS module (C-A-T) that is known to incorporate glycine into the rimosamide peptide chain. This was consistent with bioinformatic predictions of the \textit{Orf6595} adenylation-domain active site providing strong evidence that recombination between the adenylation domains of \textit{orf6595} and \textit{wolG2} could confer specificity to glycine (Supplementary Fig. 14).

To verify this prediction and examine the substrate specificity of all of the A-domains of interest, pET28a hexa-histidine tagged \textit{WolG1A2}, \textit{WolG2A2} and \textit{Orf6595A} constructs were cloned for expression based upon the A-domain boundaries as described in Crüsemann et al.\textsuperscript{39}. These were expressed in \textit{E. coli} Rosetta (DE3)pLysS and purified using Ni-Affinity chromatography. Initially, the resulting protein was insoluble; however, co-expression with the MtB-like protein \textit{WolF2} (expressed from \textit{PCDFDuet-1}) enabled the purification of soluble protein in each case. The ability of these isolated A-domains to activate each of the twenty proteinogenic amino acids and L-ornithine was then measured using a hydroxylamine trapping assay\textsuperscript{48} (Supplementary Table 11). \textit{WolG2A2} adenylates L-ornithine, in line with our hypothesis (Fig. 5a); however, it was also capable of activating other substrates albeit with lower efficiency. Most noticeably, \textit{WolG2A2} accepted L-aspartate (58% activity relative to L-ornithine) and L-asparagine (44% activity relative to L-ornithine) as substrates, but wollamide analogues in which aspartate or asparagine were substituted for ornithine were not identified in culture extracts of \textit{Streptomyces sp. MST-110588} despite targeted LCMS analysis. In contrast, both \textit{WolG1A2} and \textit{Orf6595A} activate glycine in a highly specific manner (Fig. 5a). These data corroborate our hypothesis that historic recombination with \textit{orf6595} could alter the substrate specificity of the module six adenylation domain of \textit{WolG2} from L-ornithine to glycine. To test this hypothesis further, we produced a hybrid A-domain encoding gene sequence representing the hypothetical ancestral recombinant, based on \textit{wolG2A2} and \textit{orf6595} sequences (Fig. 5b). The resulting gene product was purified and assayed as above and found to be highly selective for glycine (Fig. 5a).

Taken together with combined genomic, in silico and in vivo data above, these biochemical data show how the contemporary desotamide BGCs have evolved from an ancestral wollamide-like BGC through the process of gene duplication, intragenomic recombination and gene loss.

**Discussion**

How NRPS and PKS assembly-line biosynthetic pathways evolve is a question of perennial interest in natural products research. An understanding of this process offers new avenues for developing rational approaches for bioengineering and the targeted
production of new molecules. Evolutionary analyses of publicly available BGC sequences are common in the literature, but it is unprecedented to find an example of a BGC that represents a snapshot of assembly-line evolution. In this work, we studied Streptomyces sp. MST110588, a co-producer of desotamide and wollamide hexapeptide antibiotics. Despite their close structural relationship, it was not obvious how such a mixture of congeners might be assembled based on canonical NRPS function, pointing to an unusual biosynthetic pathway.

The architecture of the wollamide BGC is highly similar to the previously described desotamide BGC; however, a striking difference is that it contains duplicates (wolG1 and wolG2) of dsaG which encodes the final NRPS protein of the desotamide assembly line. This suggested a bifurcated assembly-line was responsible for the observed mixture of desotamide and wollamide congeners (Fig. 2c). In this scenario, Woll and WolH are responsible for condensation of the first four amino acids prior to extension with either WolG1, producing desotamides, or WolG2, producing wollamides. This hypothesis was supported by bioinformatic (Fig. 4) and biochemical analysis (Fig. 5a) of the A-domains WolG1A2 and WolG2A2, which encode the selection and activation of glycine and L-ornithine respectively. Further in vivo evidence came from strain engineering in which expression of WolG2 in a desotamide-only producing strain led to the biosynthesis of additional wollamide congeners (Fig. 3).

Comparison of the protein sequences of the NRPSs encoded by wolG1/G2 indicated a high degree of sequence conservation and identical GC-skews, indicative of a gene duplication event. The regions coding for the final adenylation domains showed a marked drop in identity. These data were indicative of a recombination event rather than of divergence through the accumulation of mutations alone, and we subsequently analysed all 34 A-domain sequences present in the Streptomyces sp. MST110588 genome using the recombination detection program RDP4. This identified wolG2 as the major parent, and an NRPS gene orf6595 (encoded ~3 Mb away on the chromosome) as the minor parent, of a recombination event that formed wolG1. Using the blastclaster tool we identified the BGC containing orf6595 as a homologue of the rimosamide BGC. This was verified by subsequent bioinformatics and biochemical analysis. To validate the recombination event predicted by RDP4 we used the contemporary sequences of wolG2A2 and orf6595 to recapitulate the predicted ancestor and generated a synthetic A-domain. Subsequent biochemical analysis of the isolated protein showed it was selective for glycine activation as predicted.

Based on these combined data we can confidently trace the evolutionary history of the wollamide and desotamide BGCs (Fig. 6a). First, a gene duplication event in an ancestral wollamide (or wollamide-like) BGC resulted in a redundant copy of the bimodular NRPS encoding modules 5 and 6 of the assembly line. Subsequently, an intragenic recombination event between the DNA encoding the adenylation domains of the duplicated module six in the wolB homologue and orf6595 resulted in an intermediate NRPS selective for glycine (leading to WolG1) and L-ornithine (leading to WolG2). This progenitor is the common ancestor of the wollamide and desotamide producing wol BGC sequenced here. In one lineage, selection and mutation resulted in the wol BGC observed in the contemporary genome of Streptomyces sp. MST110588 capable of bifurcated biosynthesis. In a divergent lineage, the ancestral gene encoding the L-ornithine-specific adenylation domain along with the duplicated MbtH-like protein-encoding gene and associated genes encoding L-ornithine biosynthesis were lost through gene deletion, ultimately resulting in the contemporary dsa BGCs producing only desotamides. These observations are consistent with the presence of a redundant epimerase domain located in the second modules of WolG1 and all DsaG homologues. Moreover, in Streptomyces MST110588, production of wollamide congeners is approximately three times lower than that of desotamides. Our engineered strain of the desotamide producer, Streptomyces sp. MST-70754, expressing wolG2 also showed an even larger difference in titer. This observation suggested that the interactions between the C- and N-terminal docking domains of WolH/DsaH and WolG2 had depreciated. To assess this, we calculated homology models which revealed that a key E16A amino acid change in WolG2 leads to the loss of a salt bridge formed with R3504 from WolH, most likely causing decreased DD pair affinities, explaining the observed differences in peptide titers. The depreciation of this interaction is consistent with a lack of selective pressure for wollamide production and may indicate drift towards a desotamide-only BGC. Although DD deprecation is likely to account for the difference in yield between the wollamides and desotamides in WT MST-110588, other factors may account for further loss of titre in our engineered strains, such as differences in the MbtH domains or the reduced capacity to biosynthesise L-ornithine.

Gene duplication and divergence is a central mechanism behind the evolution of new gene functions. Gene duplications may occur as whole gene duplications (characterised by the duplication of an entire gene), or as partial or intragenic duplications (characterised by incomplete duplication of the
ancestral gene and can result in attenuated or lengthened duplicates, respectively\textsuperscript{38}. There is convincing evidence for the role of intragenic duplications during assembly-line evolution (Fig. 6b). Such processes are often cited as the origin of multi-modularity\textsuperscript{1,55} and there is significant evidence for this process, especially in PKS assembly-lines. In contrast, whole gene duplication has the potential to create bifurcated biosynthetic pathways, as evidenced here by the wof BGC, or parallel pathways, as evidenced by the recently published BE-18257A-C and pentaminomycins A-E BGC\textsuperscript{56}. Through either mechanism, whole gene duplication creates redundancy and has the potential to reduce selective pressure for maintenance on both alleles\textsuperscript{38,54}. In this fashion, a duplicate may evolve new functionality (neofunctionalisation) without the organism losing the original product. Thus, neofunctionalisation of one copy becomes possible without strong selective pressure against the loss of a phenotype and may follow a more gradual route for the emergence of a new activity. Subsequently, the original allele can be lost if there is no advantage and selection pressure for its maintenance leading to a new linear pathway (Fig. 6b). Moreover, this model provides a scenario where rounds of duplication and neofunctionalisation can occur but if no advantage is gained then the duplicated allele can be lost, restoring the original genotype. As genes and pathways are subject to multiple selective forces it is unclear how common this mechanism may be in nature. This is compounded by the homogenising effects of gene conversion/concerted evolution that can make it difficult to reliably distinguish duplication events\textsuperscript{1,55,57,58}. However, the wof BGC provides compelling evidence that this process has occurred at least once, during the evolution of desotamide biosynthesis. A growing number of hexapeptide BGCs are becoming available, for example, those of the related ulleungmycin and curacomyclin\textsuperscript{59,60}. Given the small yet diverse differences of the peptide backbones, the hexapeptides present an enticing model for further understanding NRPS evolution.

Finally, the question arises as to whether our knowledge of the NRPS evolution can also be used to develop new engineering approaches. Recently, several efficient and/or highly productive engineering strategies have been published\textsuperscript{61–63}. Yet efficient engineering of these often-huge biosynthetic machinery to produce novel bioactive NRPs is an ongoing challenge. Understanding the evolution of these multifunctional enzymes might provide new insight for engineering and discovering new peptide-based therapeutic agents\textsuperscript{63}. A key feature of desotamide evolution that could be exploited for future NRPS engineering efforts is the intragenomic recombination between distinct NRPS encoding BGCs. The recombination event we predicted takes place within the boundaries of A domains, consistent with previous\textsuperscript{39,40,66} and more recent work\textsuperscript{18,41}. More specifically, the predicted recombination allows for the substitution of the substrate specificity conferring the active site and the ATP/phosphate-binding catalytic P-loop\textsuperscript{67} while functionally maintaining most of the A-T\textsuperscript{58} and all of the C-A inter-domain contacts\textsuperscript{69–71}. This pattern is frequently seen in our dataset, suggesting an evolutionary advantage for maintaining the structural relationships between the P-loop and substrate binding pocket. Our results are supported by two other independent studies, further highlighting its significance\textsuperscript{18,41}.

Calcott and co-workers analysis of recombination hotspots within C-A-T tri-domains (modules) from \textit{Pseudomonas, Bacillus} and \textit{Streptomyces} species identified the same recombination sites as predicted for the formation of wofG1A2\textsuperscript{41}. Subsequently, Baunach and co-workers systematic in silico dissection of many individual recombination events unveiled the striking commonality of A domain recombination events in nature\textsuperscript{18}. Specifically, these recombination events target variable portions of the A\textsubscript{core} domains to modulate A domain substrates while domain-domain interactions and the flexible A\textsubscript{sub} domain largely remained unaffected. These studies, taken together with the evolutionary and biochemical evidence presented here, must inevitably lead to a change in paradigm of future NRPS engineering experiments. Established engineering principles are being overhauled by the increasing evidence base. In particular, the idea that C and A domains have coevolved\textsuperscript{72,73}, resulting in strong acceptor site specificity of C domains\textsuperscript{74,75} is in question\textsuperscript{18,41,63}. This hypothesis is not congruent with the observation that A domain recombination is responsible for the diversification of many NRPs\textsuperscript{18,39,76}, not least the L-ornithine to glycine change demonstrated here. Such diversification would be impossible if C domain ‘gate-keeping’ was universal.

The ability to exchange A domains with greater accuracy should aid engineering efforts in the future, however many issues remain. As we have observed in nature, there is unlikely to be a single recombination site that will work in every case. Nevertheless, the growing body of genomic data will continue to reveal evolutionary snapshots akin to the wollamide BGC presented here. Over a decade of sequencing has provided us with only a threadbare sample of microbial genomic diversity and, as a result, we are still largely ignorant of the evolutionary processes that govern small molecule biosynthesis. Fortunately, this means there remains a wealth of information yet to be gleaned from Nature.

**Methods**

**Strains and culture conditions.** All strains and plasmids used in the study can be found in Supplementary Tables 12 and 13. \textit{Escherichia coli} strains were cultured on
were cultured on SF + 1 g/L calcium carbonate, in deionised water). Apramycin (100 µg/mL), kanamycin (50 µg/mL) and G418 (200 µg/mL) were used as selection markers.

Streptomyces spp. MST-70754, MST-71321, MST-71458, MST-94754 and MST-127221 were isolated by Microbial Screening Technologies (Smithfield, NSW, Australia) from soil samples collected across New South Wales between 1994 and 1998.

Genome sequencing and genomic analysis. High molecular weight genomic DNA was extracted according to a modified version the salting out procedure described by Kieser et al.34 with the modifications as described here: wet mycelium (0.5 mL) from a 3 h old SV2 culture was washed with 10% sucrose (10 mL) before resuspension in SET buffer (5 mL; 57 mM NaCl, 25 mM EDTA, 20 mM Tris HCl pH 8.0) to which lysozyme (200 µL; 50 mg/mL) and ribonuclease A (15 µL; 10 mg/mL) were added. The cells were incubated overnight at 37 °C. Fresh lysozyme (300 µL) was added after ca. 17 h followed by an additional 2 h incubation.

Genomic DNA of Streptomyces spp. MST-110588 was sequenced with Pacific Biosciences (PacBio) RSII SMRT technology (commissioned to the Earlham Institute, Norfolk, UK) and assembled via the HGAP2.0 pipeline. Genomic DNA of Streptomyces spp. MST-70754, MST-71321, MST-71458, MST-94754 and MST-127221 were sequenced with Illumina MiSeq using paired-end sequencing and the specific primers described above. The genome sequences of the desmotidamate BGCs were deposited at GenBank as follows: from Streptomyces spp. MST-70754 at accession code GenBank: MZ093610; Streptomyces spp. MST-71321 at accession code GenBank: MZ093611; Streptomyces spp. MST-71458 at accession code GenBank: MZ093612; Streptomyces spp. MST-94754 at accession code GenBank: MZ093613; and Streptomyces spp. MST-127221 at accession code GenBank: MZ093614.

Genomic DNA sequences were annotated using prodigal77 as implemented by antiSMASH 3.0.41. Strain taxonomy was performed using multi-locus sequence typing was implemented in AutoMLST78. Adenylation-domain specificity predictions were carried out using NRPSsp79 or by NRPSpredictor280 and the specificity was determined by comparison with the crystal structure coordinates of PaxC_NDD as well as the crystal structure coordinates of PfarX_NDD–PfarX_BDD (retrieved from the RSCB Protein Data Bank file PDB-ID: 5TRP_1 [https://doi.org/10.2210/pdb5trp/pdb])37 for WolG2_NDD. Prior to homology modeling the crystal structure of 6TRP_1 was prepared (i.e. wrong protonation, chirality, and hybridisation) and a structural alignment was made. A series of 10 models per protein were constructed with MOE using a Boltzmann-weighted randomised procedure combined with specialised logic for their initiation and sequence insertion/deletion38. The models with the best packing quality function was selected for full energy minimisation. Using the AMBER14 forcefield parameters for proteins (Amber14: EHT), the calculated MOE packing scores for models of WolG1_NDD–WolH_CDD and WolG2_NDD–WolH_CDD were 2.4286 and 2.2322, respectively. The stereochemical qualities of the models were assessed using Ramachandran plots and by calculating the Root-Mean-Square-Deviation (RMSD) values of the superposed Ca-atoms of the model with its respective template structure. RMSDs of WolG1_NDD–WolH_CDD and WolG2_NDD–WolH_CDD with 6TRP_1 were 0.1724 and 0.1762, respectively.

Protein expression, purification and biochemical assays. Adenylation domains were assembled into pET28a using the methodology described in section 1.5 using the primers from Supplementary Table 14. Additionally, the MbtH-like protein-coding sequence wolF2 was cloned into pCDF-Duet-1 using the same procedure. E. coli (DE3) pLysS containing both expression plasmids were grown overnight in Luria-Bertani (LB) medium containing 50 µg/mL kanamycin and 30 µg/mL apramycin. Cells were harvested at high optical density (OD600 ~0.5) and induced with IPTG (1 mM). Following induction, cultures were harvested at 18 °C. Cultures were centrifuged at 8000 × g for 20 min and the supernatant was discarded. Pellets were resuspended in 20 mL P-Buffer (1 M K2HPO4 9.4%; 1 M KH2PO4, 0.6%, pH 8.0). Cells were homogenised using an AV8 Bioxpress homogeniser (30 s) and the supernatant was discarded. Samples were centrifuged at 14,000 × g for 20 min. Supernatants were then filtered through a 0.22 µm filter and frozen at –80 °C. Protein expression was confirmed by SDS-PAGE (12% RunBlue) and Western blotting (see Supplementary Information).

Data availability

Data supporting the findings of this work are available within the paper and its Supplementary Information file, or in publicly available databases. The DNA sequence data for the genome assembly and biosynthetic gene clusters (BGCs) generated in this study have been deposited in GenBank. The genome sequence of Streptomyces spp. MST-
110588 has been deposited under the accession code GenBank: CP074380. The desolate BGCs have been deposited under the accession codes as follows: Streptomyces sp. MST-70754, GenBank: MZ093610; Streptomyces sp. MST-71321, GenBank: MZ093611; Streptomyces sp. MST-71458, GenBank: MZ093612; Streptomyces sp. MST-94745, GenBank: MZ093613; and Streptomyces sp. MST-127221, GenBank: MZ093614. The crystal structure coordinates of the docking domain protein PaxC_NDD-PaxAt_CDD were retrieved from the RCSB Protein Data Bank file PDB-ID: 6TRP [https://doi.org/10.2210/pdb6trp/pdb]. Biochemical data for analysis of adenylation-domain substrate specificity are available as a Source Data file. A reporting summary for this Article is available as a Supplementary Information file. Source data are provided with this paper.

Received: 6 June 2021; Accepted: 18 May 2022; Published online: 17 June 2022

References
1. Nivina, A., Yue, K. P., Hsu, J. & Khosla, C. Evolution and diversity of assembly-line polyketide synthases. Chem. Rev. 119, 12524–12547 (2019).
2. Süssmuth, R. D. & Mainz, A. Nonribosomal peptide synthesis—principles and perspectives. Angew. Chem. Int. Ed. 56, 377–402 (2017).
3. Kirst, H. A. The spinosyn family of insecticides: realizing the potential of natural products research. J. Antibiotics 63, 101–110 (2011).
4. Fierro, F. et al. Transcriptional and bioinformatic analysis of the 56.8 kb DNA region amplified in tandem repeats containing the penicillin gene cluster in Penicillium chrysogenum. Fungal Genet. Biol. 43, 618–629 (2006).
5. Ray, L. & Moore, B. S. Recent advances in the biosynthesis of unusual polyketide substrate synthases. Nat. Prod. Rep. 33, 150–161 (2016).
6. Caboche, S., Leclère, V., Pupin, M., Kucherov, G. & Jacques, P. Diversity of monomers in nonribosomal peptides: towards the prediction of origin and biological activity. J. Bacteriol. 192, 5143–5150 (2010).
7. Walsh, C. T., O’Brien, R. V. & Khosla, C. Nonproteinogenic amino acid building blocks for nonribosomal peptide and hybrid polyketide scaffolds. Angew. Chem. Int. Ed. 52, 7098–7124 (2013).
8. Blin, K. et al. antiSMASH 5.0—updates to the secondary metabolite genome mining pipeline. Nucleic Acids Res. 47, W81–W87 (2019).
9. Kouprina, N. & Larionov, V. TAR cloning: insights into gene function, long-range haplotypes and genome structure and evolution. Nat. Rev. Genet. 7, 805–812 (2006).
10. Nijkamp, J. F. et al. De novo sequencing, assembly and analysis of the genome of the laboratory strain Streptomyces coccureus CEN.PK113-7D, a model for modern industrial microbiotechnology. Microb. Cell Factor. 11, 36 (2012).
11. Hahn, M. & Stachelhaus, T. Harnessing the potential of communication-mediated domains for the biocombinatorial synthesis of nonribosomal peptides. Proc. Natl Acad. Sci. USA 103, 275–280 (2006).
12. Hacker, C. et al. Structure-based redesign of docking domain interactions modulates the product spectrum of a rhodopeptide-synthesizing NRPS. Nat. Commun. 9, 1–11 (2018).
13. Watzel, J., Hacker, C., Darder-Ferné, E., Bode, H. B. & Wöhnert, J. A new docking domain type in the peptide-antimicrobial-xenobabudus peptide producing nonribosomal peptide synthetase from Xenobabudus bovienii. ACS Chem. Biol. 15, 982–989 (2020).
14. Innan, H. & Kondrashov, F. The evolution of gene duplications: classifying and distinguishing between models. Nat. Rev. Genet. 11, 97–108 (2010).
15. Cruesemann, M., Kohlhais, C. & Piel, J. Evolution-guided engineering of nonribosomal peptide synthetase adenylation domains. Chem. Sci. 4, 1041–1045 (2013).
16. Kries, H., Niquelle, D. L. & Hilvert, D. A subdomain swap strategy for reengineering nonribosomal peptides. Chem. Biol. 22, 640–648 (2015).
17. Baltz, R. H. Function of MbtH homologs in nonribosomal peptide cyclase and hydrolase in surugamide non-ribosomal peptide biosynthesis. Curr. Opin. Microbiol. 19, 2259–2257 (2016).
18. Bozhuyuk, K. A., Micklefield, J. & Wilkinson, B. Engineering enzymatic assembly lines to produce new antibiotics. Curr. Opin. Microbiol. 51, 88–96 (2019).
19. Jenke-Kodama, H. & Dittmann, E. Evolution of metabolic diversity: insights from microbial polyketide synthases. Phytochemistry 70, 1858–1866 (2009).
20. Zhang, L. et al. Characterization of giant modular PKSs provides insight into genetic mechanism for structural diversification of anaplyglol polyketides. Angew. Chem. Int. Ed. 56, 1740–1745 (2017).
21. Wlodek, A. et al. Diversity oriented biosynthesis via accelerated evolution of modular gene clusters. Nat. Commun. 8, 1206 (2017).
22. Baunach, M., Chowdhury, S., Stallforth, P. & Dittmann, E. The landscape of recombination events that create nonribosomal peptide diversity. Mol. Biol. Evol. 38, 2116–2120 (2021).
23. Khalil, Z. G., Salim, A. A., Lacey, E., Blumenthal, A. & Capon, R. J. Wollamides: anticymbobacterial cyclic heptapeptides from an Australian soil Streptomyces. Org. Lett. 16, 5120–5123 (2014).
24. Song, Y. et al. Cyclic heptapeptides from the deep South China sea-derived Streptomyces scopoletidis CSIO 7246 active against pathogenic Gram-Positive bacteria. J. Nat. Products 78, 944–948 (2015).
25. Li, Q. et al. Identification of the biosynthetic gene cluster for the anti-fungal desoxamides and production of a new analogue in a heterologous host. J. Nat. Products 78, 944–948 (2015).
26. Eil, J. et al. Real-time DNA sequencing from single polymerase molecules. Science 349, 377–380 (2015).
27. Chow, Y. et al. Investigation of Penicillin binding protein (PPB)-like peptide cyclase and hydrolase in surugaride non-ribosomal peptide biosynthesis. Cell Chem. Biol. 26, 737–744 (2019).
28. Fazal, A., Webb, M. E. & Seipke, R. F. The desomamide family of antibiotics. Antibiotics 9, 1–14 (2020).
29. Kadi, N. & Challis, G. L. Chapter 17 siderophore biosynthesis. Methods in Molecular Biology. Biochemistry 38, 703–787 (2019).
30. McCleary, R. A. et al. Elucidating the rimosamide-detoxin natural product families and their biosynthesis using metabolite/gene cluster correlations. ACS Chem. Biol. 11, 3452–3460 (2016).
31. Mechinaud, F. L. & Challis, G. L. Nitisinone as a model for natural product biosynthesis. Curr. Opin. Microbiol. 15, 435–441 (2012).
32. Khalil, Z. G. et al. Structure-activity relationships of wollamides cyclic heptapeptides with activity against drug-resistant and intracellular mycobacterium tuberculosis. Antimicrob. Agents Chemother. 63, e01773-18 (2019).
33. Prior, A. M. & Sun, D. in Methods in Molecular Biology, vol. 2103, 175–187 (Humana Press Inc., 2020).
Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-022-30950-z.

Correspondence and requests for materials should be addressed to Barrie Wilkinson.

Peer review information Nature Communications thanks Marnix Medema, Martin Baunach and the other anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.