PNAS Nexus, 2024, 3, pgae384
https://doi.org/10.1093/pnasnexus/pgae384
Advance access publication 12 September 2024
Research Report

Natural diversifying evolution of nonribosomal peptide synthetases in a defensive symbiont reveals nonmodular functional constraints

Zhiyuan Li [a,b,c,1], Laura P. Ióca [d,1], Ruolin He [a] and Mohamed S. Donia [d,e,f,*]

[a]Center for Quantitative Biology, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing 100871, China
[b]Peking-Tsinghua Center for Life Sciences, Academy for Advanced Interdisciplinary Studies, Peking 8 University, Beijing 100871, China
[c]Center for the Physics of Biological Function, Princeton University, Princeton, NJ 08544, USA
[d]Department of Molecular Biology, Princeton University, Princeton, NJ 08544, USA
[e]Department of Chemical and Biological Engineering, Princeton University, Princeton, NJ 08544, USA
[f]Department of Ecology and Evolutionary Biology, Princeton University, Princeton, NJ 08544, USA
*To whom correspondence should be addressed: Email: donia@princeton.edu
[1]Z.L. and L.P.I. contributed equally to this work.
Edited By Li-Jun Ma

Abstract

The modular architecture of nonribosomal peptide synthetases (NRPSs) has inspired efforts to study their evolution and engineering. In this study, we analyze in detail a unique family of NRPSs from the defensive intracellular bacterial symbiont, *Candidatus* Endobryopsis kahalalidifaciens (*Ca.* E. kahalalidifaciens). We show that intensive and indiscriminate recombination events erase trivial sequence covariations induced by phylogenetic relatedness, revealing nonmodular functional constraints and clear recombination units. Moreover, we reveal unique substrate specificity determinants for multiple enzymatic domains, allowing us to accurately predict and experimentally discover the products of an orphan NRPS in *Ca.* E. kahalalidifaciens directly from environmental samples of its algal host. Finally, we expanded our analysis to 1,531 diverse NRPS pathways and revealed similar functional constraints to those observed in *Ca.* E. kahalalidifaciens' NRPSs. Our findings reveal the sequence bases of genetic exchange, functional constraints, and substrate specificity in *Ca.* E. kahalalidifaciens' NRPSs, and highlight them as a uniquely primed system for diversifying evolution.

Significance Statement

Nonribosomal peptide synthetases (NRPSs) represent an important class of biosynthetic pathways, responsible for the production of diverse small molecules with ecologically and biomedically relevant bioactivities. Here, we use a family of 20 NRPS pathways from the obligate intracellular bacterial symbiont, *Candidatus* Endobryopsis kahalalidifaciens, as a model system to study diversifying evolution in ecologically relevant NRPS pathways: they produce a diverse cocktail of defensive kahalalides that protect the algal host, *Bryopsis* sp., from predation. By resolving the evolutionary bases of diversification in these pathways, we were able to accurately predict and discover new kahalalide products from this symbiotic system. Finally, we show that similar evolutionary insights can be observed in environmental NRPSs at large, broadening the scope of our findings.

Introduction

Nonribosomal peptide synthetases (NRPSs) are multienzyme complexes that produce natural products (nonribosomal peptides, NRPs) with enormous chemical diversities and a wide variety of biological activities. Multimodular NRPSs build their products in an assembly line fashion, where each module incorporates a single amino acid—including nonproteinogenic ones—into the growing peptide chain (1, 2). Each basic catalytic module contains an adenylation domain (A) that selectively activates an amino acid, a thiolation domain (T) on which the growing chain is tethered via a thioester bond, and a condensation domain (C) that catalyzes the peptide bond formation. Amino acid substrates can also be modified by optional domains, e.g. an epimerization domain (E) that changes the configuration of the alpha carbon from L to D or an *N*-methyltransferase domain that methylates the amino acid on the backbone nitrogen, among others (1–3). The final product is typically released by a thioesterase domain (TE), giving a cyclic, linear, or branched product (1).

Between domains and modules, interdomain linkers sew catalytic sequences together. Such module and domain organization

Competing Interest: M.S.D. is a Scientific Co-Founder and CSO at Pragma Bio. The work described in this manuscript is unrelated to the work conducted at Pragma Bio.
Received: March 4, 2024. **Accepted:** July 19, 2024

tremendously potentiates the chemical diversity of NRPs and has been inspiring NRPS engineering for decades. NRPS engineering approaches focused on modifying the A domain specificity by swapping individual domains (4–8), multiple domains together (9–12), or even full modules (13, 14). Many of the engineering efforts using these strategies have resulted in lower yields or even no production at all, with the exception of very elegant designs that succeeded in creating functional NRPSs by modifying sequences outside of the standard module and domain framework (15–19). A full understanding of the principles limiting the rational design of NRPS pathways is still needed.

Duplication followed by divergence is the main evolutionary mode that has been proposed for generating new NRPSs, where the NRPS pathway is first copied to an exact replica then diversified by mutation, recombination, module deletion, or module duplication (20–22). Phylogenetic analyses, while valuable in reflecting potential evolutionary paths, also reflect trivial sequence covariations due to shared ancestry (23). When sequence covariation between two residues is observed, it is difficult to discern functional association from trivial co-occurrence in ancestral sequences by chance (24–27), which we term "phylogenetic relatedness" in this study. In principle, a natural evolutionary system where pathways are closely related in sequence but highly divergent in function would enable us to overcome this dilemma and directly link sequence covariations to functional constraints. Here, we study an NRPS system that is ideal for this purpose.

We have recently discovered an obligate bacterial symbiont, "*Candidatus* Endobryopsis kahalalidifaciens" (*Ca*. E. kahalalidifaciens) that lives intracellularly within the marine alga *Bryopsis* sp. and produces a library of defensive toxins for the benefit of its host: the kahalalides (28–30). Despite its highly reduced genome, *Ca*. E. kahalalidifaciens encodes 20 closely related NRPS pathways, each active one is responsible for the production of at least one distinct kahalalide product. These pathways appear to have evolved through duplication and divergence followed by pervasive diversifying recombination to give rise to new pathways. This genetic exchange occurred in extremely high frequency such that even sequence fragments in the same pathway do not follow the same phylogenetic tree. In this study, we report that the intensive recombination in *Ca*. E. kahalalidifaciens' NRPSs reduces sequence covariation induced by phylogenetic relatedness, leaving correlations induced by functional dependence clearly observable. We further capitalize on these findings to reveal rules of functional constraints and substrate specificity in this remarkable system, allowing us to accurately predict and experimentally characterize products of an orphan NRPS in *Ca*. E. kahalalidifaciens.

Results

Intensive recombination between NRPS pathways in *Ca*. E. kahalalidifaciens reduces phylogenetic relatedness

In our previous work, we uncovered several special features of the *Ca*. E. kahalalidifaciens' genome (28). First, its genome is significantly reduced (1.87 Mb without the NRPS pathways), half of the average size of its family, and missing complete pathways for amino acid biosynthesis and DNA repair. Second, its genome contains an unusually high number of transposons and transposases, enriched near the flanking regions of the NRPS pathways, with no significant signs for large-scale horizontal gene transfer. Third, a large proportion of the genome's coding capacity (~20%) and transcriptional activity (~26%) are dedicated to defensive kahalalides' biosynthesis by numerous NRPS pathways. Altogether, the *Ca*. E. kahalalidifaciens genome harbors 20 NRPS pathways with 120 modules and 434 domains that are highly homologous in sequence (Fig. 1A and B). We had labeled nine of these pathways as "active" because their chemical products were directly identified in the algal samples analyzed, whereas seven were labeled as "inactive" because they had disrupted domain structures by premature stop codons and transposases. The remaining four pathways were classed as "presumably active" since they had normal module and domain architectures, but their products had not yet been discovered. Among the four presumably active pathways, NRPS-3 and NRPS-19 are transcribed at an intermediate level in the environmental sample (Fig. 1B and Table S1). Throughout the rest of the manuscript, we will focus our analysis on the 13 active and presumably active pathways.

In a typical "duplication and divergence" scenario for NRPS evolution, domains in the same pathway share a similar evolutionary history. It is not clear whether the same degree of phylogenetic relatedness still holds for *Ca*. E. kahalalidifaciens' NRPSs, where intense recombination between duplicated pathways results in an unprecedented mode of diversifying evolution (28). To test whether domains from the same pathway in *Ca*. E. kahalalidifaciens' NRPSs continue to coevolve (or covary) despite constant diversification, we quantified the correlation coefficient of their sequence distance matrices using the following approach. Briefly, for domains or interdomains of type X (where X can be C domain, A domain, T domain, C–A interdomain, A–T interdomain, etc.), distances between its member x_i in the i-th domain and the member x_j in the j-th domain, $d(x_i, x_j)$, is calculated by the fraction of amino acids that are different after sequence alignment. For the type X, $d(x_i, x_j)$ for all pairs of i and j form its sequence distance matrix dX. The correlation coefficient between the distance matrix of type X and that of its neighboring type Y, $\rho_{X,Y} = \mathrm{Corr}(dX, dY)$, indicates the association degree between type X and type Y. The higher $\rho_{X,Y}$, the better the distance between a pair of X domains predicts the distance between their neighboring Y domains (Fig. 1C). Following this logic, if the pathway is under the typical duplication and divergence mode of evolution, across pathways and species, the similarity between two domains of type X can be attributed to them being in the same or closely related pathway or species. Such phylogenetic information is also shared by the neighboring domains of type Y, leading to strong covariation between X and Y even when there is no functional dependence between them (Fig. 1D). On the other hand, intensive recombination would likely act against the phylogeny-induced covariation, as fragments of sequences randomly shuffle so proximate regions may have distinct evolutionary histories. At the same time, sequence associations induced by functional dependence would resist random shuffling. If domain or interdomain X and Y function in concert for desired products, then combinations with unmatched pairs would lead to dead enzymes that eventually get eliminated through evolution (31, 32), leaving only domain pairs with strong correlation induced by functional dependence (Fig. 1E).

Since there is only one starting C domain (Cs) and one TE domain per NRPS pathway, we started by comparing the phylogenetic trees and computing distance matrix correlations (as described above) of these beginning and end domains from the 13 active and presumably active pathways in *Ca*. E. kahalalidifaciens at the amino acid sequence level (Fig. 1F). Indeed, the two phylogenetic trees looked dissimilar, and the distance matrix correlation coefficient between the two domains was only −0.05 (Fig. S1), supporting the hypothesis that intensive recombination caused large

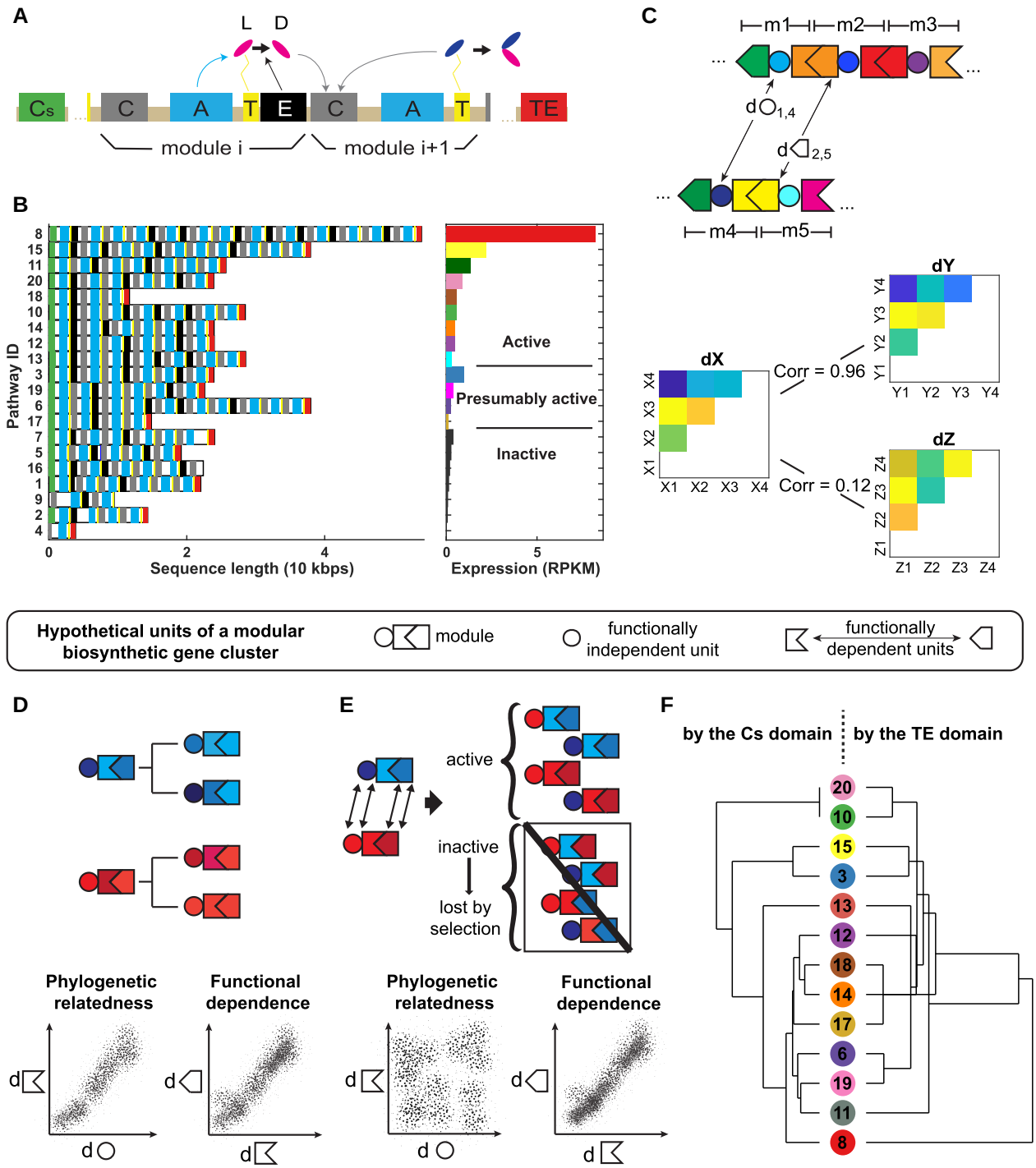

Fig. 1. A computational approach for quantifying modular biosynthetic gene cluster (BGC) evolution. A) Schematic representation of the modular structure of NRPSs. The same color code for domains is used for subsequent figures. B) Left, domain architecture of the 20 NRPS pathways encoded by *Ca.* E. kahalalidifaciens, following the same color code as in A. Right, a bar graph showing the expression level of the 20 NRPS pathways, in reads per kilobase pairs per millions of sequenced reads from the *Bryopsis* algae metatranscriptomic sequencing data. The same color code for the 20 pathways is used for subsequent figures. C) Schematic of the distance matrix correlation approach used in this study. In the upper panel, a hypothetical modular BGC is depicted, with modules made up of three types of domains represented by a concave polygon, a circle, and a convex polygon. Numbers are used to identify modules (m_1, m_2, etc.) and their domains. Modules 1–3 and 4–5 can be encoded on the same polypeptide, from the same pathways but encoded on two different polypeptides or from two different pathways. The fraction of amino acids that are different in a global alignment is used to calculate the distance between each two domains of the same type to produce comprehensive distance matrices. Hypothetical distance matrices of various types of domains (X, Y, and Z) are shown in the lower panel. The degree of association between two module types is indicated by the correlation coefficient between their two respective matrices. D) A diagram depicting modular BGC evolution under the "duplication and divergence" mode. Color shades distinguish subtypes within each type of domain. The distance matrices of domains that are functionally dependent (concave and convex polygons) and domains that are phylogenetically related but functionally independent (circle and polygons) are both strongly correlated. E) A diagram depicting modular BGC evolution under the "intensive recombination" mode, as in the case of *Ca.* E. kahalalidifaciens. A strong correlation can only be observed between the functionally dependent concave and convex polygons. On the other hand, any correlation due to the phylogenetic relatedness of functionally independent domains, e.g. the circle and the polygons, is eliminated by frequent sequence shuffling. F) Dendrograms based on the hierarchical clustering of the starting C domain sequences (left) and the ending TE domain sequences (right) from the 13 NRPS pathways (shown in colored circles in the middle). The two dendograms show different topologies.

disturbances in the evolutionary history of *Ca.* E. kahalalidifaciens' NRPSs and erased covariations that are typically induced by phylogenetic relatedness. Next, we computed sequence covariation between domains within the same module using the same approach, and again observed little evidence of covariation between all domain pairs, including C and A domains, A and T domains, and C and T domains ((Corr(dC, dA)) = 0.1, Corr(dA, dT) = 0.23, and Corr(dC, dT) = 0.12, Fig. 2A–C).

To be comprehensive, we also computed sequence covariation between domains in different modules. Surprisingly, we discovered that the amino acid sequence of T domain strongly covaries with its recipient C domain in the next module (Fig. 2D), with a correlation coefficient of 0.95. In addition, the C–A interdomain sequence also has a strong covariation with the preceding C and the preceding T domains, with correlation coefficients of 0.98 and 0.95, respectively (Fig. 2E and F). These strikingly high signals of sequence covariation across neighboring modules among a general lack of intramodule covariation suggest a case of functional dependence that constrains sequence divergence.

A nonmodular chirality unit in *Ca.* E. kahalalidifaciens

With strong intermodule covariation signals revealed by our analysis, we sought to explore their potential link to NRPS function. First, we decided to focus on chirality determinants. It has been known that the C domain has multiple chirality-dependent subtypes (33, 34). Likewise, the chirality of T domains has also been suggested (35), but following studies showed controversial results (36). Not much has been discussed about the chirality-related subtypes of C–A interdomains. In *Ca.* E. kahalalidifaciens' NRPSs, the T domain and C–A interdomain exhibit the same chirality dependence as that of the C domain. To further investigate this relationship, we performed hierarchical clustering for 90 T domains belonging to the 13 active and presumably active *Ca.* E. kahalalidifaciens' NRPS pathways. In the resulting phylogenetic tree, these domains did not group by their pathway, but, instead, they grouped exactly by whether they are followed by E, C, or TE domains (Fig. 2G). Furthermore, when we ordered the recipient C domains and the following C–A interdomains by their preceding T domains (for the C domains and C–A interdomains at the first module, they are ordered by the last T domain of the same pathway, the one right preceding the TE domain), this very grouping structure reappeared (Fig. 2G). Therefore, in *Ca.* E. kahalalidifaciens, T domains can be categorized into three subtypes: two of which follow the chirality groupings of C domains: L subtype, which has no E domain within its module; D subtype, which is followed by an E domain; and a third Ender-subtype, which directly precedes the TE domain. Furthermore, the C–A interdomain can also be categorized into three groups: L and D subtypes, such as the T domain, and a third starter subtype for C–A interdomain, which follows the Cs domain at the head of a pathway.

Next, we analyzed the multisequence alignments of T domains and C–A interdomains in search for positions that are highly predictive of their subgroups, as measured by the mutual information between residues and subgroups. Our analysis revealed a two-segment structure for both the T domain and the C–A interdomain (Figs. 2H and S2). In T domains, the highly conserved T1 motif "DDFFxLGGDS(LI)" appears in the middle of all subtypes (Fig. 2J), which has been known to be highly conserved and crucial for phosphopantetheinyl binding (37, 38). The T1 motif separates the T domain into two halves with distinctive properties. The first segment, extending from the start of the domain (Tα1 motif, Fig. 2I) to the end of the T1 motif (Fig. 2J), is relatively conserved and contains little information about subtypes. The second segment, which extends from the T1 motif to the end of the domain, diverges by subtypes, with residues containing high content of information on subtypes.

Across the C domain, high-information residues start from the known C1 motif, extend beyond the known C7 motif, and experience a sudden stop at a position between the known C7 motif and the known A1 motif (Fig. 2H). In this position located in the C–A interdomain, we noticed a new conserved motif "ELLETFNxTE(VA)YP" for all subtypes (Fig. 2K). We named it the "CA motif," located 101.8 (±7.9) aa after the end of the C7 motif, and 28.0 (±0.1) aa before the start of the A1 motif. Similar to the T1 motif, the CA motif separates the C–A interdomain into two halves, where the first half diverges by chirality dependence, and the second half is relatively conserved across all subtypes. By the known core T1 motif and the newly discovered CA motif, we can define a nonmodular and nondomain "chirality unit" for *Ca.* E. kahalalidifaciens, which starts from the middle of the T domain, contains the whole C domain, and ends around 28 aa before the A domain. This chirality unit has two subtypes: if this unit contains E domain, it is of D subtype; otherwise, it is of L subtype. Interestingly, all the three components of the proposed chirality unit share spatial proximity and appear to closely interact in previously solved crystal structures of multidomain NRPSs (9, 39), further supporting their potential functional dependence during enzymatic catalysis (Fig. S3).

An A-domain-based substrate specificity unit in *Ca.* E. kahalalidifaciens

While the T, E, and C domains may jointly dictate the substrate's chirality, the A domain has been known for decades to determine substrate specificity (40–42). Along the same lines, no other domain or interdomain in *Ca.* E. kahalalidifaciens covaries significantly with A domains within the same module (Fig. 2A and B). Because of the complex evolutionary history of NRPSs in *Ca.* E. kahalalidifaciens, a detailed analysis of the sequence–function association of A domains may reveal a more resolved view of substrate specificity determinants in this system. We quantified the mutual information between residues and substrate specificities in the nine active pathways whose products have been previously validated in the same sample (28). Detailed multisequence alignment revealed a three-segment structure for A domains in *Ca.* E. kahalalidifaciens (Fig. 3A), separated by two known core motifs A3 and A6 (Fig. 3B and C) (38): segments before the A3 motif and those after the A6 motif are moderately variable, with little information about the substrate specificity of A. In contrast, the segment between the A3 motif and the A6 motif (we termed it "A_{core}") is highly variable and harbors high content of mutual information about substrate specificity. Of note, the high-information region ends about seven residues earlier than the conventional A6 motif. A recent analysis suggested multiple conserved sites right before the known A6 motif (37). Therefore, we used this extended A6 motif (Fig. 3C) for further analyses. Not surprisingly, the A_{core} sequence overlaps with the pocket region of A, which has been long-known to be responsible for the substrate selectivity of A domains (40, 42, 43).

We then compared the ability of each of the following to predict the observed substrate specificity in the isolated kahalalides: sequence similarity of the A_{core} (Fig. 3D), sequence similarity of the full A domain (Fig. S4), and the pocket region of the A domain using standard methods (Table S1 and S2). Interestingly,

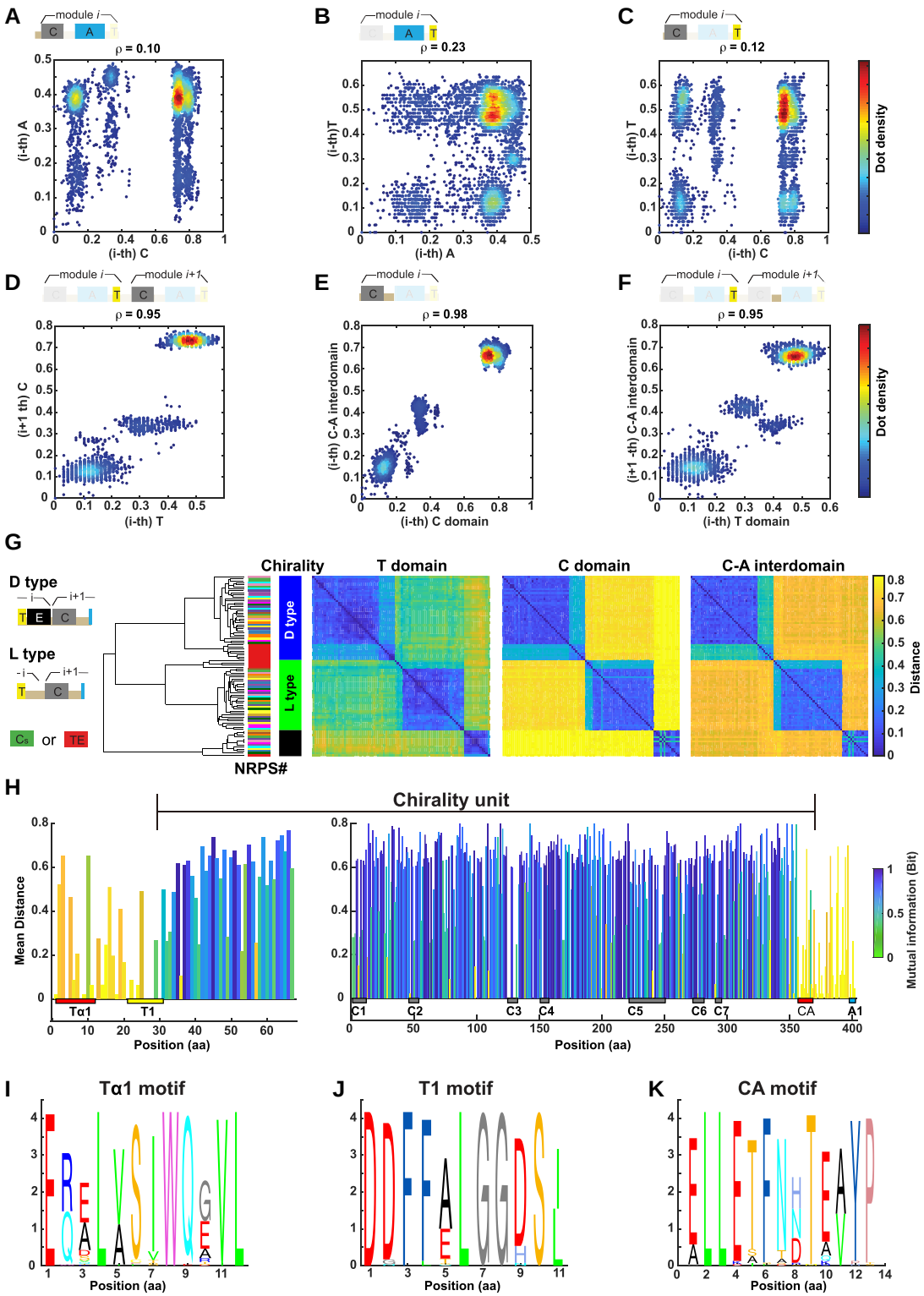

Fig. 2. Sequence covariations and the chirality unit in *Ca.* E. kahalalidifaciens' NRPSs. A–F) Correlation between the distance matrices of different types of domains within modules (A–C), between modules (D), and between domains and interdomains (E, F) of *Ca.* E. kahalalidifaciens' NRPSs. In A–F, axes indicate sequence distance, i.e. the fraction of amino acids that are different between the two sequences after alignment. Correlation coefficients are given on the top of each plot. G) The chirality dependence of the T domain, the C domain, and the C–A interdomain in *Ca.* E. kahalalidifaciens' NRPSs. From left to right: dendrogram based on the hierarchical clustering of T domain amino acid sequences and the associated distance matrix, followed by the sequence distance matrices of the C domains and the C–A interdomains, organized to reflect the same order as in the T domain matrix. Color codes indicating pathway identities and chirality subtypes are shown to the right of the dendrogram, while domain organization subtypes are shown to the left of the dendrogram. H) Bar graphs representing the average sequence distance (bar height) and the mutual information between residues and subtypes (bar color) calculated from the multisequence alignment of the T domain, the C domain and the C–A interdomain. Positions of known conserved motifs (Tα1, T1, and C1–C7) and the newly identified CA motif are marked at the bottom. The proposed chirality unit starts from the T1 motif and ends with the CA motif. I–K) Sequence logos of the Tα1, T1, and the CA motif are shown in I–K, respectively.

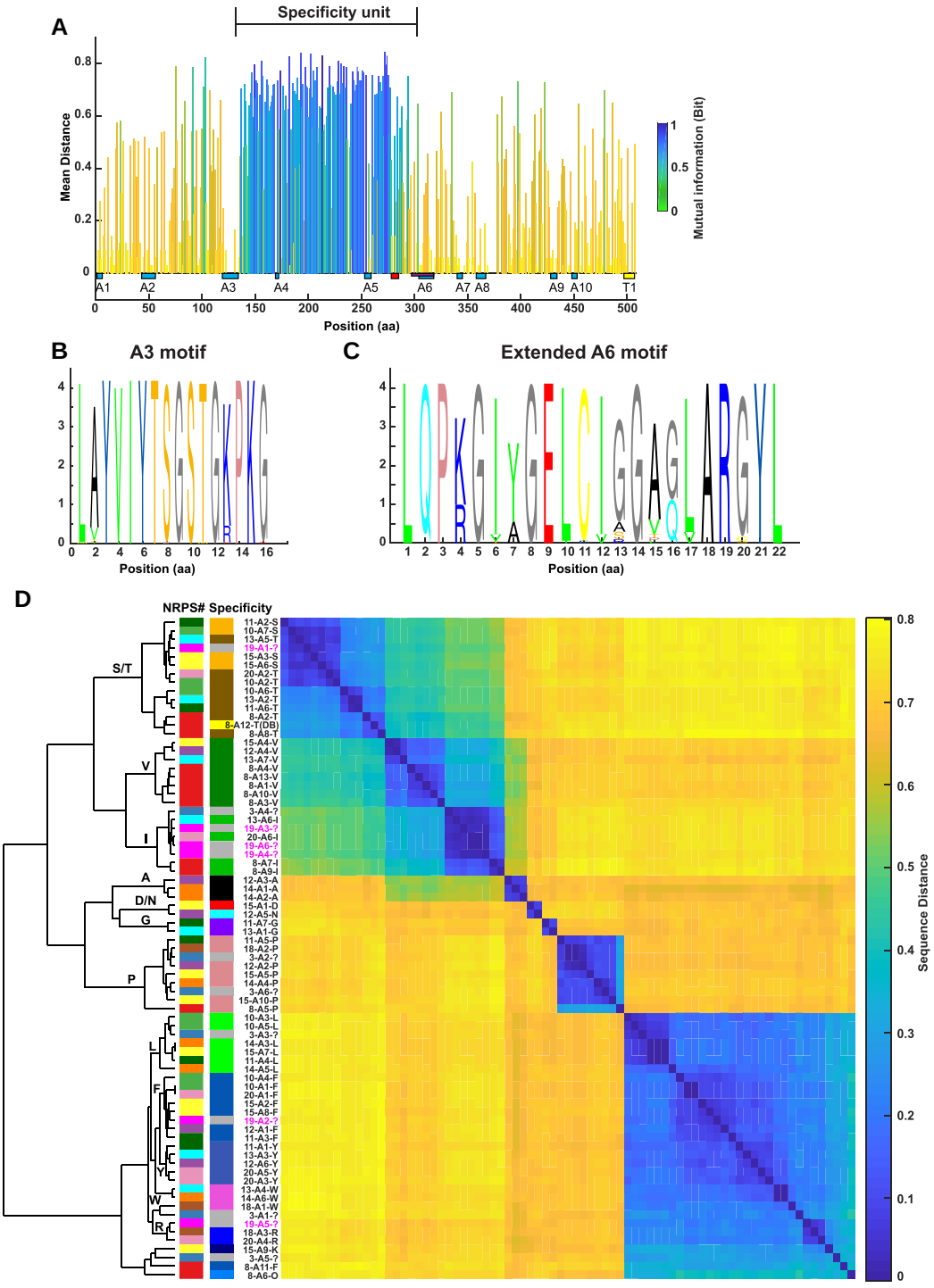

Fig. 3. A domain substrate specificity in *Ca.* E. kahalalidifaciens NRPSs. A) A bar graph representing the average sequence distance (bar height) and the mutual information between residues and substrate specificity (bar color) calculated from the multisequence alignment of the A domains in *Ca.* E. kahalalidifaciens NRPSs. Positions of known conserved motifs (the A1–A10 motifs of the A domain and the T1 motif at the beginning of the T domain) are marked at the bottom. The extended A6 motif is also marked on top of the canonical A6 motif. The proposed specificity unit starts at the end of the A3 motif and ends at the beginning of the A6 motif, also termed as the A_{core}. B, C) Sequence logos of the A3 and the extended A6 motif, respectively. D) Dendrogram based on the hierarchical clustering of A_{core} amino acid sequences and the associated distance matrix (A_{core} domains from the nine active NRPS pathways and two presumably active NRPS-3 and NRPS-19 are shown). NRPS pathway of origin (following the same color code as in Fig. 1) and amino acid substrate specificity for each sequence are shown in color to the right of the dendrogram, clearly highlighting the homogeneity of substrates but not pathway of origin within clades. Text annotation indicating "NRPS pathway number—A domain module number—observed amino acid specificity" is provided next to the heat map.

hierarchical clustering based on the distance matrix of A_{core} amino acid sequences produced a sharper clustering than that of the entire A domain (Fig. S4). More importantly, it revealed cases where few mutations between otherwise closely related A_{core} domains appear to flip substrate specificity: for example, NRPS14-A5 and NRPS15-A2 select Leu and Phe, respectively, yet they share 87% identity in the core regions and could be switched into one another with only 22 mutations. Of note, predictions from

commonly used tools that rely mostly on the A domain pocket, such as NRPSpredictor2 (in antiSMASH) (44, 45), SeMPI 2.0 (46), or the classical Stachelhaus code based on A domain crystal structure (40–42), were mostly inaccurate in the nine active pathways, which is likely due to the fact that they were trained on pathways from species that are distant from *Ca.* E. kahalalidifaciens (Table S1). Taken together, we denote the A_{core} piece as the "specificity unit" of *Ca.* E. kahalalidifaciens NRPSs.

Specificity of tailoring modifications predicted by starting C and terminal TE domains

An important modification in the lipopeptide kahalalides is the conjugation of fatty acyl moieties to the first amino acid in the peptide chain by the starter condensation domain (Cs), a process termed "lipoinitiation." Lipoinitiation has been investigated in other systems before, suggesting that Cs is not only responsible for the acylation of the first amino acid, but also for selecting it (47–52). In the reduced genome of *Ca.* E. kahalalidifaciens, fatty acid synthesis pathways are largely intact and highly active, frequently positioned in close proximity to NRPS pathways and likely contributing to the diversity of the NRPS products (Fig. S5A and B). Since the kahalalides harbor five different fatty acid chains, we next asked whether we can recover additional rules for the Cs domain specificity from sequence analyses. To answer this question, we calculated the distance matrix of all Cs domains, and used it to perform hierarchical clustering of these domains based on amino acid sequence identity. Satisfyingly, *Ca.* E. kahalalidifaciens' Cs domains grouped into four different clades, consistent with the size of the fatty acid chain incorporated (Fig. 4A): a group that selects butanoic acid (Bu) and 2-methylbutanoic acid (2MeBu; NRPS-20 and NRPS-10, respectively), a group that selects 5-methylhexanoic acid (5-MeHex; NRPS-13, NRPS-8, and NRPS-11), a group that selects 9-methyl-3-hydroxy-decanoic acid (9-Me-3Decol; NRPS-12, NRPS-14, and NRPS-15), and NRPS-18 that selects 7-methyl-3-hydroxy-octanoic acid (7-Me-3Octol). The high degree of sequence identity between Cs domains in *Ca.* E. kahalalidifaciens aids in identifying key regions that may determine fatty acid substrate specificity. We observed that sites with high mutual information regarding the fatty acid extend beyond the C7 motif (Fig. 4B and C), agreeing with the definition of a starter-unit extending from the first C1 motif to the CA motif. Structurally, these high-information sites can be located within or near the Cs domain's "Latch" (Fig. S5C), a region recently reported to change conformation during the catalytic reaction cycle (52).

Next, we turned into the last domain involved in kahalalide biosynthesis, the TE domain, and wondered if TE domain sequence can predict cyclization types observed. In a similar analysis to the one performed with the Cs domains, we computed the distance matrix of all 13 TE domains from the active and presumably active subset of *Ca.* E. kahalalidifaciens' NRPSs, and performed hierarchical clustering based on amino acid sequence identity. In the kahalalides, macrocyclization occurs via an ester bond formation between the carboxylic group of the last amino acid incorporated in the peptide chain and a hydroxyl group at the beginning of the chain. Interestingly, TE domains fell into two main clades. In the resulting molecules, these two clades differ in the source of the macrolactone hydroxyl group: a hydroxyl group from a hydroxylated fatty acid chain in one group (e.g. 9Me3Decol, 7Me3Octol in KY, KE, KQ, and KD of NRPS-12, -14, -15, and -18), and a hydroxyl group from a hydroxylated amino acid (e.g. Ser or Thr in KA, KB, KO, and KC of NRPS-10, -11, -13, and -20; Fig. 4D).

Aside from the NRPS-8 TE, which is over 80% dissimilar from other TEs, eight TEs from the other active pathways are extremely similar, with only 5–20% residue differences (Fig. 4D), allowing us to perform detailed sequence-to-function analyses. We uncovered two high-information sections concerning the cyclization type. One region is between the last T1 motif and the TE1 motif (residues 132–144, Fig. 4E and F). More high-information sites can be found after the TE1 motif (residues 233–235, Fig. 4E and F). This is the first helix of the "Lid" region (Fig. S5D), which has been shown previously to confer substrate selectivity (53–55). It has been reported that the substrate enters the TE domain through a channel between the Lid and the Core regions, consistent with our data suggesting that the first helix of the Lid area may exhibit selectivity for the cyclization type.

Sequence analysis of *Ca.* E. kahalalidifaciens NRPS domains allows accurate prediction and targeted discovery of novel kahalalides

To test whether the results we obtained from the detailed sequence analysis of *Ca.* E. kahalalidifaciens' NRPSs can lead to the accurate prediction of their products, we decided to focus on previously uncharacterized yet presumably active NRPSs encoded in the *Ca.* E. kahalalidifaciens genome. An extensive analysis of the nine orphan NRPSs, together with prior transcriptomics analysis (28), suggested that NRPS-19 is a promising candidate for discovery: it has an intact domain architecture with an intermediate expression level yet no kahalalide products have ever been linked to it (Fig. 1B). Guided by our sequence analyses described above, we based our predictions on the similarity of NRPS-19 domains to domains from the active NRPS pathways (i.e. previously linked to specific kahalalide products; Figs. 3D and 4C, F).

The A_{core} sequences of NRPS-19 cluster nicely with A domains with different substrate specificity from the nine active pathways (Fig. 3D). NRPS19-A1 locates within the S cluster, with 96% identity to NRPS10-A7 (selecting S) and 95% identity to NRPS15-A3 (S). NRPS19-A2 is most similar to domains selecting F (94% identity to NRPS15-A8 and 92% identity to NRPS15-A2). NRPS19-A3, NRPS19-A4, and NRPS19-A6 locate in the I/V cluster, most similar to NRPS13-A6 (97% identity), NRPS20-A6 (98% identity), and NRPS20-A6 (99% identity), respectively, both coding for I. NRPS19-A5 clusters with two other domains selecting R (NRPS18-A3, 95% identity and NRPS20-A4, 92% identity). Taken together, we speculated the major product of NRPS-19 to be: S-F-I-I-R-I. Notably, NRPSpredictor2 and other prediction tools fail to predict the specificity of all six domains of NRPS-19 (Table S1). Other than the A domains, the Cs domain of NRPS-19 falls into the 5MeHex cluster (Fig. 4A and C), and the TE domain is in the hydroxylated amino acid clade (Fig. 4D and F), which is in line with the first amino acid being a serine. Taken together, the product of NRPS-19 was predicted to be: 5MeHex-S-F-I-I-R-I, with an amino acid macrocyclization to the initial serine's hydroxyl. Using the E domain and chirality unit information, we further predicted that the amino acids F and R will be in the D configuration, while the rest of the amino acids will be in the L configuration. Following this prediction, NRPS-19's putative product was expected to have an exact mass of 841.54 Da (Fig. 5A).

To test whether our prediction is accurate, and whether the NRPS-19 product is indeed produced, we generated a crude extract of the *Bryopsis* algae and searched for the predicted mass ($m/z = 842.5498$, $[M+H]^+$) using High Performance Liquid Chromatography coupled with High Resolution tandem Mass Spectrometry (HPLC–HR–MS/MS) analysis. Satisfyingly, an ion

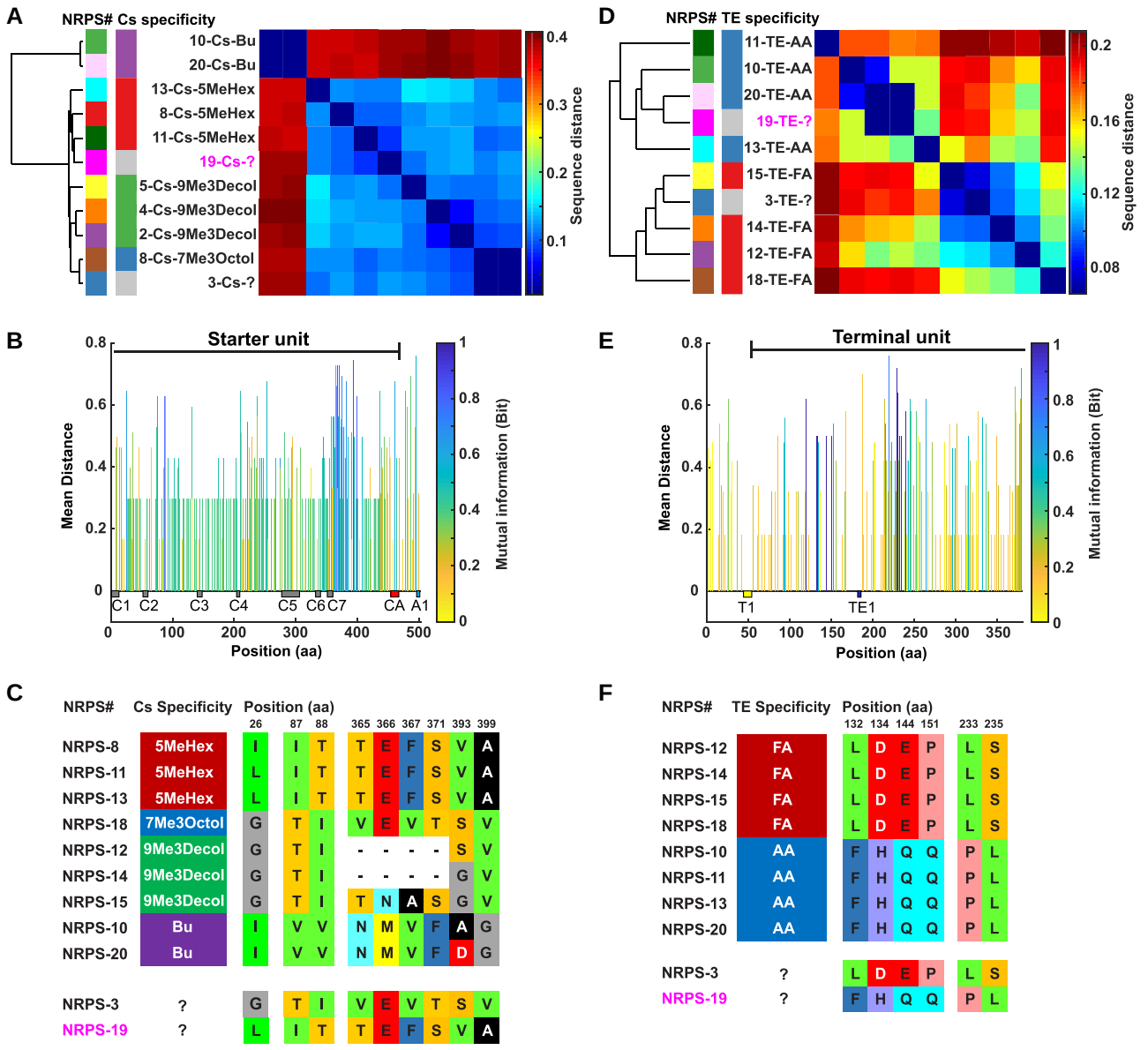

Fig. 4. Sequence–function specificity of Cs and TE domains in *Ca.* E. kahalalidifaciens' NRPSs. A) Dendrogram based on the hierarchical clustering of Cs domain amino acid sequences and the associated distance matrix (Cs domains from the nine active NRPS pathways and two presumably active NRPS-3 and NRPS-19 are shown). NRPS pathway of origin (following the same color code as in Fig. 1) and fatty acid substrate specificity for each sequence are shown in color to the right of the dendrogram, clearly highlighting the homogeneity of substrates within clades. Text annotation indicating "NRPS pathway number—Cs—observed fatty acid specificity" is provided next to the heat map. B) A bar graph representing the average sequence distance (bar height) and the mutual information between residues and fatty acid substrate specificity (bar color) calculated from the multisequence alignment of the Cs domains in *Ca.* E. kahalalidifaciens' NRPSs, starting from the first C1 motif until the first A1 motif. Positions of known conserved motifs are marked at the bottom. C) High mutual information sites for the pathways analyzed in A, B (numbers indicate amino acid positions in the multisequence alignment shown in B). D) Dendrogram based on the hierarchical clustering of TE domain amino acid sequences and the associated distance matrix (TE domains from eight active NRPS pathways, excluding NRPS-8, and two presumably active NRPS-3 and NRPS-19 are shown). NRPS pathway of origin (following the same color code as in Fig. 1) and cyclization type for each sequence are shown in color to the right of the dendrogram, clearly highlighting the homogeneity of cyclization types within clades. Text annotation indicating "NRPS pathway number—TE—cyclization type" is provided next to the heat map. E) A bar graph representing the average sequence distance (bar height) and the mutual information between residues and cyclization type (bar color) calculated from the multisequence alignment of the TE domains in *Ca.* E. kahalalidifaciens' NRPSs, starting from the end of the last Tα1 motif until the end of the pathway. Positions of known conserved motifs are marked at the bottom. F) High mutual information sites for the pathways analyzed in D, E (numbers indicate amino acid positions in the multisequence alignment shown in E).

matching the same mass was identified (observed: $m/z = 842.5497$, $[M + H]^+$, error −0.1 ppm, Fig. 5B), and further analysis of its MS/MS fragmentation pattern confirmed the proposed product as 5-MeHex-S-F-I-I-R-I (Fig. S6 and Table S3). To unequivocally determine the structure of the identified molecule (e.g. MS/MS fragmentation alone does not differentiate L and I), and to confirm the predicted stereochemistry, we performed MS-guided isolation and purification of the target ion from a large-scale extraction of *Bryopsis* algae, yielding 1.1 mg of pure molecule (Fig. S7). We then performed full structural elucidation of the purified new molecule, which we termed kahalalide L-1 (**1**, Fig. 5C), using a combination of 1D and 2D Nuclear Magnetic Resonance (NMR) analyses (Figs. S8–S10 and Table S4) and advanced Marfey's analysis (Figs. S11–S14, see Supplementary Methods for a complete

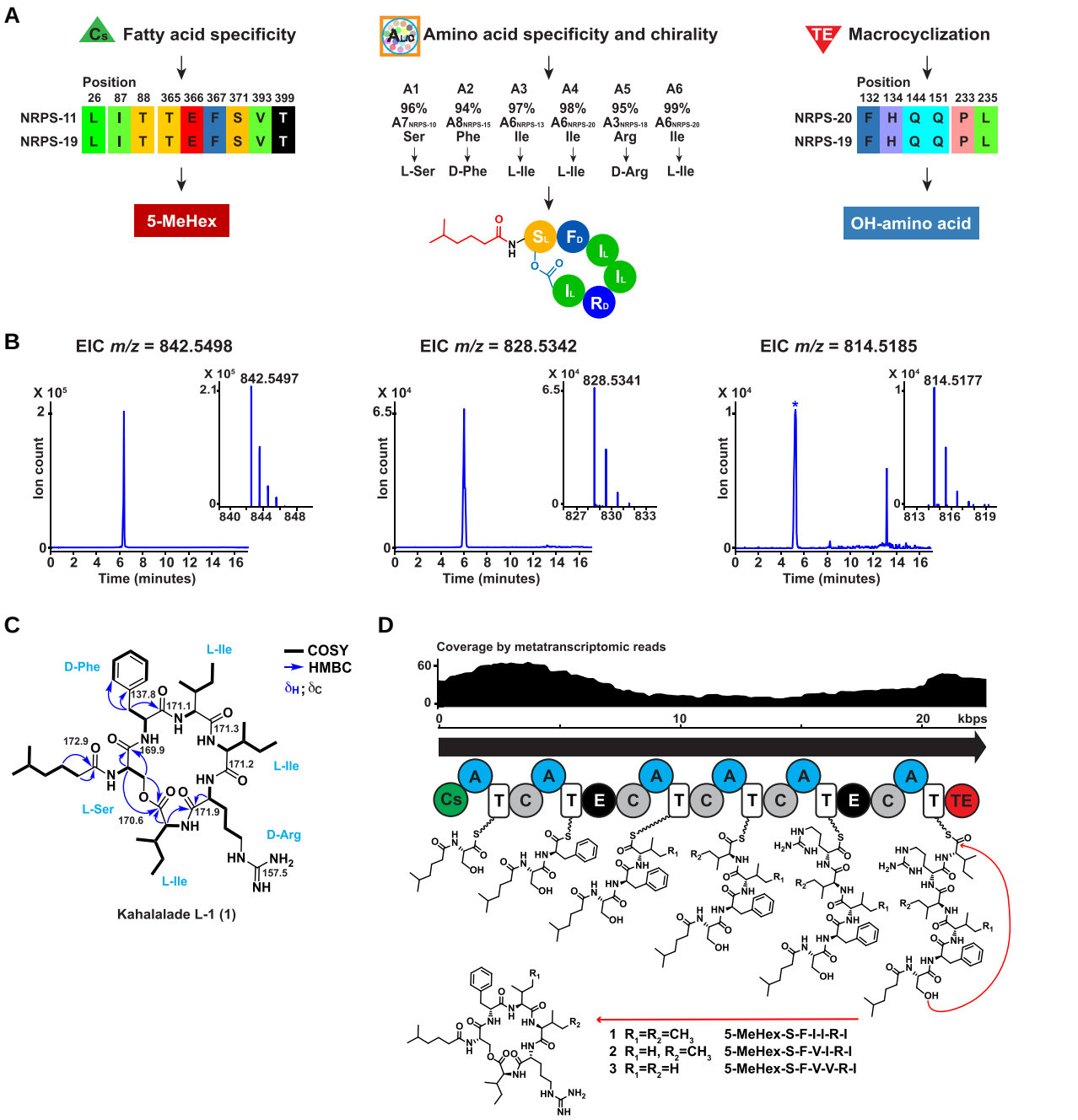

Fig. 5. Prediction, discovery, and structural elucidation of the products of *Ca.* E. kahalalidifaciens' NRPS-19. A) Overview of the prediction rules for *Ca.* E. kahalalidifaciens' NRPS products, applied to the orphan NRPS-19. B) Extracted ion chromatograms for the calculated mass to charge ratio (*m/z*) corresponding to the $[M+H]^+$ ions of the products of *Ca.* E. kahalalidifaciens' NRPS-19, as detected in the chemical extract of *Bryopsis* sp. Mass spectrum of the observed ion for each product, kahalalides L-1 (**1**), L-2 (**2**), and L-3 (**3**), respectively. C) The molecular structure of Kahalalide L-1 (**1**) and key COSY and HMBC correlations from the NMR analysis used to elucidate the structure. Results from the Marfey's analysis are indicated next to the amino acid residues. D) Domain organization of NRPS-19 from *Ca.* E. kahalalidifaciens is shown in the middle: Cs, starter condensation; A, adenylation; T, thiolation; C, condensation; E, epimerization; TE, thioesterase. Proposed biosynthetic scheme for kahalalides L-1-3 (**1–3**) is shown at the bottom, while transcriptional activity, based on prior metatranscriptomic analysis (28) is shown on top.

discussion of the structural elucidation procedures). In addition, two related variants, which we named kahalalides L-2 and L-3 (**2**–**3**), were detected and their structures could be proposed by comprehensive MS/MS analysis (Figs. 5B, D, S15 and S16, and Tables S5–S6). Notably, these congeners correspond to the substrate promiscuity predicted by our computational analysis: NRPS19-A3 and NRPS19-A4 are located in the I/V cluster, and therefore can select either amino acid (Fig. 3D). Unfortunately, the same approach was not successful when applied to NRPS-3: although we could confidently predict the substrate specificity of the Cs domain (7Me3Octol, Fig. 4A and C) and most A domains (NRPS-3-A1: W, NRPS-3-A2: P, NRPS-3-A3: L, NRPS-3-A4: I, NRPS-3-A6: P, Fig. 3D), as well as the macrocyclization type (a fatty acid-derived hydroxyl group, Fig. 4D and F), we were unable to confidently predict the substrate specificity of NRPS-3-A5, which falls in a promiscuous and undefined clade that includes A

domains responsible for the incorporation of F, K, as well as the nonproteinogenic amino acid O. Nevertheless, our ability to precisely predict natural products from the sequence of NRPS-19, and to experimentally verify these predictions by discovering novel products from a symbiotic system that has been studied for decades is remarkable. It is important to note that this level of predictability in such a complex system would not have been possible without the detailed sequence-level analysis performed on the *Ca.* E. kahalalidifaciens' NRPSs.

The unit-and-linker organization of *Ca.* E. kahalalidifaciens NRPSs reveals DNA recombination hotspots

The amino acid sequence covariation, multisequence alignment analysis, and the successful prediction of NRPS-19 products provide support to a unit-and-linker organization of NRPSs in *Ca.* E. kahalalidifaciens (Fig. 6A and B): the structural domains and modules are no longer unbreakable entities from an evolutionary standpoint. Instead, the sequences of domains and interdomains are divided by four highly conserved motifs to give four functional units with negligible sequence association: the starter unit, from the beginning of the pathway to the first CA motif; the specificity unit A_{core}, from the A3 motif to the extended A6 motif for each A domain; the chirality unit, from the T1 motif to its following CA motif; and the terminal unit from the last T1 motif to the tail of the pathway. In addition, two highly conserved linkers, one extends from the CA motif to its following A3 motif (termed CA–A3) and another extends from the extended A6 motif to its following T1 motif (termed A6–T1), sew these functional units together (Fig. 6B). In this assembly logic, beginning from the starter unit, *n* loops with the flow "→ CA–A3 linker → specificity unit → A6–T1 linker → chirality unit → CA–A3 linker," producing *n*-substrate product, then an exit from the A6–T1 linker to the terminal unit completes the synthesis.

Next, to test whether the units or linkers have sequence association, we quantified the residue covariation from the multisequence alignment of 88 A–T–C–A amino acid sequences from the 13 active or presumably active pathways using the average product corrected mutual information method (25). Strong covariations only occur inside each functional unit but not elsewhere else (Fig. 6C). We found no evidence of coevolution between the specificity unit and its following chirality unit, nor between any pairs of adjacent units and the linker. This finding demonstrates that units and linkers have little sequence covariation, implying that they might be used by *Ca.* E. kahalalidifaciens as the basic building blocks for sequence recombination and pathway diversification.

Nucleotide sequence–based analyses offer higher resolution on patterns of recombination than amino acid sequence–based analyses. Previously, we visualized *Ca.* E. kahalalidifaciens' recombination events as distinct stripes on genome-wide nucleotide identity dot plots (28). Here, we further analyzed these stripes to identify potential recombination "units" (Fig. 6D). Despite varying in length, composition of these stripes follows certain rules (Fig. 6E): first, between any two pairs of NRPSs, the starter and terminal units are always involved in stripes. Second, other than starter and terminal units, long stripes are more likely to start within around 80% of the C–A interdomains and end inside T domains at middle positions (Fig. 6F and G)—the exact two regions where the conserved CA motif and the T1 motifs are located. Sometimes, dips in stripes can be observed in the middle of A domains, corresponding to the highly variable specificity units. Consequently, most stripes start at the CA–A3 linker and end with the A6–T1 linker, with varying number of domains in between (Fig. 6H). As homologous recombination needs to be mediated by similar or identical DNA sequences, it is reasonable that these highly conserved motifs in linkers, which sew the functional units together, act as recombination hotspots.

Signatures of the chirality unit are broadly visible in non-*Ca.* E. kahalalidifaciens NRPS pathways

Our results suggest that *Ca.* E. kahalalidifaciens' NRPSs exhibit a special unit-and-linker organization that has not been previously observed in other NRPS pathways. It is not clear whether this organization is truly unique to *Ca.* E. kahalalidifaciens, stemming from its symbiotic lifestyle, or whether it reflects general principles of NRPSs that are masked by other factors. To answer this question, we explored a well-annotated BGC database containing 1,531 NRPS pathways from 991 species spanning across 3 domains of life using the same methods employed above for *Ca.* E. kahalalidifaciens (Fig. 7A–D) (56). In this database, domains exhibit extensive variation in sequence (33% mean amino acid sequence identity among 9,541 A domains, 27% among 7,946 C domains, and 34% among 8,518 T domains). However, despite such considerable sequence dissimilarity, we found that the chirality subtype in T domain and C–A interdomain, although weak, is also visually apparent and consistent. First, sequence distance matrices of T and C domains and C–A interdomains show clustering structure by their chirality subtypes (Fig. 7EG), where sequences within the same subtype are significantly more similar than sequences between different subtypes. Second, the grouping structure from the database is consistent with what was observed in *Ca.* E. kahalalidifaciens' NRPSs. For example, D-typed T domains from the database are closer to D-typed T domains from *Ca.* E. kahalalidifaciens than L-typed domains from *Ca.* E. kahalalidifaciens, and vice versa for subtype L. The same trend could be recovered for C domains and C–A interdomains. Third, using multisequence alignment, very similar sequence signatures that distinguish the chirality subtypes of T domain and C–A interdomain appear in both *Ca.* E. kahalalidifaciens' pathways and database pathways (Fig. S17). These results suggest that the chirality unit concept is likely a general feature for NRPSs.

Next, we decided to investigate further why the observed signal is so prominent in *Ca.* E. kahalalidifaciens' NRPSs but just discernible in the larger, more diverse database. The distinction can be attributed to the different modes of evolution: *Ca.* E. kahalalidifaciens is undergoing diversifying evolution, in which lineage information has been erased by extensive recombination so that even domains in the same pathway do not share the same history (Fig. 1E); meanwhile, pathways in the large database follow the model of "duplication and divergence," where pathways and modules duplicate then diversify through mutation and relatively infrequent recombination (20, 22, 31), leaving phylogenic-relatedness largely intact (Fig. 1D). Consequently, and as expected in a dataset containing diverse species where the interference from phylogenetic relatedness is strong, we found that distances between pairs of A domains correlate significantly with their phylogenetic relatedness: A domains from the same species tend to be more similar than those from different species, domains within the same pathways share even higher degree of similarity, and domains within the same protein coding sequences are the most similar (Fig. 7C). Such trend of increasing similarity with closer phylogeny produces sequence associations unrelated to functional dependence. In contrast, in the *Ca.* E. kahalalidifaciens system, there is no difference in distance distribution between all pairs of A domains, pairs from the same pathway, or pairs from the same protein coding sequence (Fig. 7D). We propose that this lack of

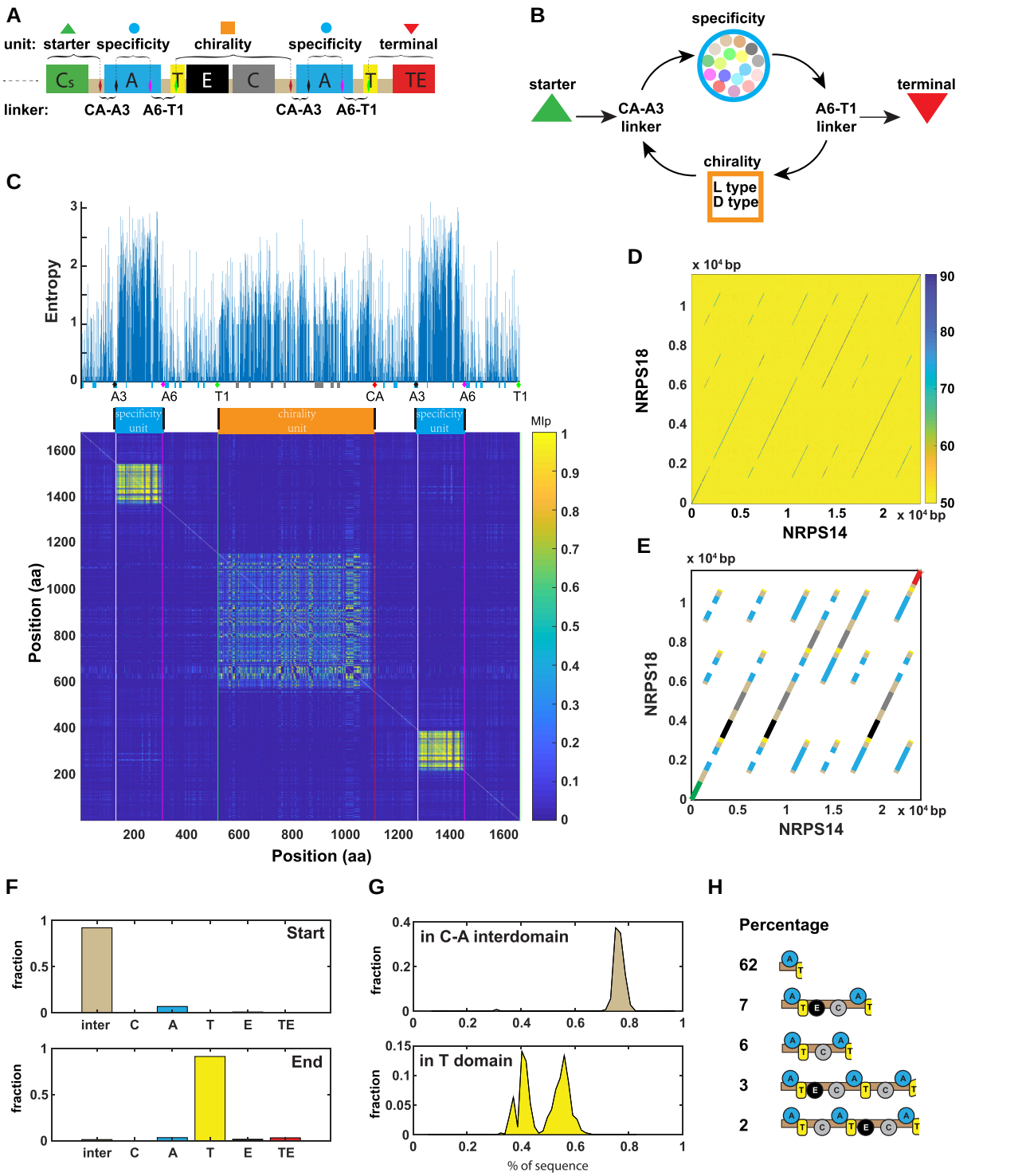

Fig. 6. A unit-and-linker view of NRPSs in *Ca.* E. kahalalidifaciens, marked by recombination hotspots. A) Cartoon illustration of the units and linkers of NRPSs in *Ca.* E. kahalalidifaciens. Colored diamonds represent conserved motifs (CA motif, A3 motif; extended A6 motif; T1 motif) that divide *Ca.* E. kahalalidifaciens' NRPSs into functional units (indicated by shapes above the pathway cartoon) and conserved linkers. B) The assembly logic of the units and linkers, with arrows indicating choices of the next unit or linker. C) The coevolution pattern between units and linkers quantified by normalized mutual information. Based on the multisequence alignment of A–T–(E)–C–A sequences (starting from the A1 motif, omitting the E1–E7 region, and ending before the T1 motif), the variability along the alignment was quantified by Shannon entropy (upper panel), with locations for core motifs marked by colored squares and the four conserved motifs in A marked by colored diamonds. The heat map in the lower panel shows the normalized mutual information (MIp) between pairs of positions in the alignment. Positions for the four conserved motifs are indicated by lines with corresponding colors, and positions match those in A. D) An example of a nucleotide sequence identity plot between NRPS-14 and NRPS-18. Percent identity between pairs of sequence fragments are represented by colors, ranging from yellow (dissimilar) to blue (highly similar), and following the color code on the right. E) Stripes formed by highly similar pairs in D, colored by the domains they harbor (domains follow the same color code as in A). F) The fraction of each domain type found at the start position (upper panel) or end position (lower panel) of stripes, marking potential start and end of recombination units. G) For stripes starting in C–A interdomain (higher panel), or ending in T domain (lower panel), the distribution of the position they start or end at in the corresponding domains is shown. Positions on the x axis are shown as a percentage of the total length of the domain. H) The five most frequent patterns of stripes among all pairs of NRPSs, as a percentage of all computed stripes.

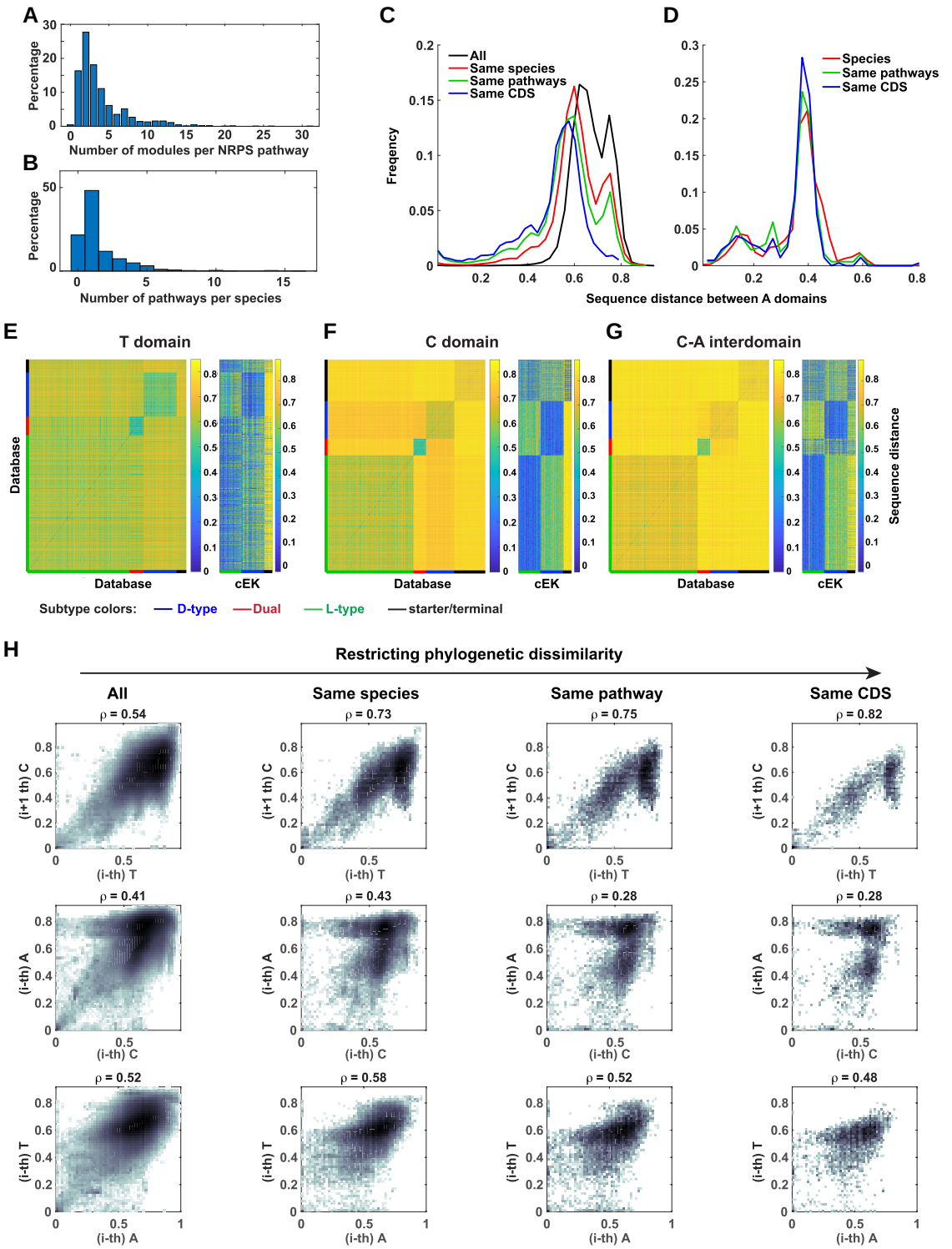

Fig. 7. Validation of the unit and linker model in a large database of diverse NRPSs. A) Distribution of the number of modules in each NRPS pathway in the analyzed database. Compared with NRPS-8 in *Ca.* E. kahalalidifaciens (13 modules), only 2.9% of pathways in the database have the same or more number of modules. B) Distribution of the number of NRPS pathways encoded by each species in the analyzed database. Compared to *Ca.* E. kahalalidifaciens (20 pathways), the species with the largest number of NRPS pathways in the Helsinki database is *Aspergillus oryzae* RIB40 (16 pathways). C) Distribution of the amino acid sequence distances between A domains in the analyzed database, based on comparing all pairs of A domains (black), only pairs from the same species (red), only pairs from the same pathways (green), and only pairs from the same protein coding sequence (blue). D) Distribution of the amino acid sequence distances between A domains in *Ca.* E. kahalalidifaciens, based on comparing all pairs of A domains in this species (red), only pairs from the same pathways (green), and only pairs from the same protein coding sequence (blue). E–G) Left heat maps: distance matrices of T domain (E), C domain (F), and C–A interdomain (G) from the analyzed database, sorted by their chirality subtypes. Note that domains and interdomains in the database cluster by chirality in the same manner as in *Ca.* E. kahalalidifaciens. Right heat maps: distance matrix between domain or interdomain sequences from the analyzed database and *Ca.* E. kahalalidifaciens' domains or interdomains in the corresponding chirality subtypes. H) The correlation between distance matrices of NRPS domains in the analyzed database, as in the analysis shown for *Ca.* E. kahalalidifaciens' NRPSs in Fig. 2A–E. Axes indicate sequence distance, i.e. the fraction of amino acids that are different between the two sequences after alignment. From left to right, the phylogeny gets more and more restricted from all pairs, to only pairs originating from the same species, pathways, and protein coding sequences. Only in the first row (the correlation between T domains and the next C domains), the correlation coefficients increase.

phylogenic relatedness in *Ca.* E. kahalalidifaciens' NRPSs sharpens the functional dependence signals in its sequence covariations.

If true, we hypothesized that it is possible to computationally recover very similar signals of functional dependence in the large database by progressively restricting the phylogenetic distance (Fig. 7H, see "Methods" section for detail). When the phylogeny is not restricted, i.e. all domain pairs are taken into calculation, the correlation between the distance matrices of T domains and the following C domains Corr(dT, dC + 1) is 0.54, not drastically different than that of C domains and the following A domains (Corr(dC, dA) = 0.41), or that of A domains and the following T domains (Corr(dA, dT) = 0.52), first column of Fig. 7H). However, if we set the phylogeny to be more and more restricted, from "only comparing pairs of domains from the same species" (second column, Fig. 7H) to "from the same pathways" (third column, Fig. 7H) then to "from the same gene" (fourth column, Fig. 7H, Corr(dT, dC + 1)) monotonically raises to 0.73, 0.75, and 0.82, respectively. Meanwhile, the correlation between other neighboring domains remains flat or even decreases (Corr(dC, dA|same gene) = 0.28, Corr(dA, dT|same gene) = 0.48). In summary, among the neighboring domain pairs, T domains and their recipient C domains increase in correlation as the phylogeny gets restricted, while C domains and the following A domains decrease in correlation with reduced phylogeny information, consistent with our initial observation in *Ca.* E. kahalalidifaciens (Fig. 2A–F). This analysis further supports the notion that closely related NRPS pathways in *Ca.* E. kahalalidifaciens uncover nonmodular functional dependencies that may otherwise be masked by phylogenetic relatedness.

Discussion

Candidatus Endobryopsis kahalalidifaciens is unique in several aspects, from its ecological role to the evolution mode and organization of its NRPSs. Ecologically, *Ca.* E. kahalalidifaciens is strictly intracellular, which limits genetic communication with other bacteria. On the other hand, its defensive role in the symbiotic relationship with its algal host inspires chemical innovation, forcing *Ca.* E. kahalalidifaciens to generate chemical diversity through intensive recombination events. Here, we uncover that under this mode of evolution, frequent recombination erased the phylogenetic relatedness of duplicated pathways, leaving functional constraints of NRPS assembly lines sharply observable. Under this organization, NRPS pathways in *Ca.* E. kahalalidifaciens are composed of four types of functional units sewed by conserved linkers. Within each type of functional unit, the high degree of sequence homology helps in unveiling the sequence–structure–substrate relationship. By applying metagenome mining in this diversely evolved system, we deorphanized an NRPS pathway in *Ca.* E. kahalalidifaciens and discovered new symbiont-derived marine natural products, named kahalalides L-1-3.

It is worth mentioning that in *Ca.* E. kahalalidifaciens, sequences from adjacent functional units appear to be nearly decoupled. Various couplings have been proposed in NRPS systems, including the influence of the C domains on the substrate selectivity of the following A domain, and linkages between adjacent A domains (57, 58). In a recent computational work performing coevolution analysis on sequences of 7,329 NRPS domains from the MiBiG database (59), overlapped coevolving sectors across C–A–T modules were revealed, limiting domain and subdomain swapping (37). Diversifying evolution may have pushed functional units in *Ca.* E. kahalalidifaciens to a "ready-to-recombine" state, where the specificity and chirality units decoupled to encourage the generation of new functional pathways. Because of this decoupling, NRPSs in *Ca.* E. kahalalidifaciens may be ideal candidates for recombination-based reengineering in a nonmodular fashion. Excitingly, a nonmodular "exchange-unit" (XU) design was recently used as the building block for combinatorial reengineering of another diverse system of NRPSs and achieved good yields (15, 16). Remarkably, XUs are fused at specific positions that connect C domain and A domain, about nine amino acids ahead of the *Ca.* E. kahalalidifaciens CA motif. More recently, the T1 and T2 motifs were also identified in this system, along with the correlation between the T2 motif and subsequent chirality determining domains, enabling NRPS engineering between pathways from different bacteria with unprecedented success (60). Likewise, swapping sequence pieces within the specificity-unit-like sequence instead of the whole A domain had shown better yields while engineering NRPSs (17–19). With rapid advances in sequence editing technologies, and vast growing databases of natural product biosynthetic pathways and their characterized products, more light is constantly being shed on sequence–function associations and specific modes of diversifying evolution in these complex pathways (19, 61).

We envision that the broad utilization of the coevolutionary methodology described here would stimulate both the targeted discovery of new molecules as well as the rational engineering of their biosynthetic machineries. Recombination eliminates sequence associations, a consequence that has been thoroughly investigated in linkage disequilibrium in human population genetics (62). In this specific case of a marine symbiont, diversifying evolution results in a phenomenon that we would like to refer to here as the "ocean wave model": pervasive recombination in *Ca.* E. kahalalidifaciens erases the sequence associations induced by phylogenetic relatedness, as if a tidal wave washes away the sand from the rocky beach, leaving the essential organization for NRPS functions clearly visible. Using computational techniques, we found that artificially constraining phylogenetic information reveals similar functional constraints even in large and diverse databases. With more databases of biosynthetic pathways and their products compiled, and more genomic and metagenomic sequencing data generated from diverse environments, similar coevolutionary approaches to the one described in this work can be systematically applied. Importantly, the logic of the ocean wave model is not limited to examining the design principles of NRPS assembly lines, but can be extended to discover functional constraints in other types of multienzyme biosynthetic and metabolic pathways.

Methods

Detailed methods are available in the Supplementary Material of this paper.

Acknowledgments

The authors thank members of the Donia Lab for useful discussions.

Supplementary Material

Supplementary material is available at *PNAS Nexus* online.

Funding

Funding for this project has been provided by the Gordon and Betty Moore Foundation (grant GBMF9199 to M.S.D., https://doi.org/10.37807/GBMF9199), and the National Natural Science Foundation of China (grant nos. 32071255 and T2321001) to Z.L.

Author Contributions

M.S.D., Z.L., and L.P.I. designed the study. Z.L., R.H., and L.P.I. performed the computational analysis. L.P.I. performed the experimental work. Z.L., R.H., L.P.I., and M.S.D. analyzed the data and wrote the manuscript.

Data Availability

All data reported in the manuscript are provided in the main text or in the supplementary material. Original code used in this manuscript is available at: https://github.com/donia-lab/Oceanwavemodel_NRPS_cEK.

References

1 Walsh CT. 2016. Insights into the chemical logic and enzymatic machinery of NRPS assembly lines. *Nat Prod Rep*. 33:127–135.
2 Walsh CT, O'Brien RV, Khosla C. 2013. Nonproteinogenic amino acid building blocks for nonribosomal peptide and hybrid polyketide scaffolds. *Angew Chem Int Ed Engl*. 52:7098–7124.
3 Süssmuth RD, Mainz A. 2017. Nonribosomal peptide synthesis—principles and prospects. *Angew Chem Int Ed Engl*. 56:3770–3821.
4 Krätzschmar J, Krause M, Marahiel MA. 1989. Gramicidin S biosynthesis operon containing the structural genes grsA and grsB has an open reading frame encoding a protein homologous to fatty acid thioesterases. *J Bacteriol*. 171:5422–5429.
5 Schneider A, Stachelhaus T, Marahiel MA. 1998. Targeted alteration of the substrate specificity of peptide synthetases by rational module swapping. *Mol Gen Genet*. 257:308–318.
6 Stachelhaus T, Schneider A, Marahiel MA. 1995. Rational design of peptide antibiotics by targeted replacement of bacterial and fungal domains. *Science*. 269:69–72.
7 Turgay K, Krause M, Marahiel MA. 1992. Four homologous domains in the primary structure of GrsB are related to domains in a superfamily of adenylate-forming enzymes. *Mol Microbiol*. 6:529–546.
8 Yan F, *et al.* 2018. Synthetic biology approaches and combinatorial biosynthesis towards heterologous lipopeptide production. *Chem Sci*. 9(38):7510–7519.
9 Tanovic A, Samel SA, Essen L-O, Marahiel MA. 2008. Crystal structure of the termination module of a nonribosomal peptide synthetase. *Science*. 321(80):659–663.
10 Calcott MJ, Owen JG, Ackerley DF. 2020. Efficient rational modification of non-ribosomal peptides by adenylation domain substitution. *Nat Commun*. 11:4554.
11 Duerfahrt T, Doekel S, Sonke T, Quaedflieg PJLM, Marahiel MA. 2003. Construction of hybrid peptide synthetases for the production of α-l-aspartyl-l-phenylalanine, a precursor for the high-intensity sweetener aspartame. *Eur J Biochem*. 270:4555–4563.
12 Nguyen T, *et al.* 2006. Combinatorial biosynthesis of novel antibiotics related to daptomycin. *Proc Natl Acad Sci U S A*. 103:17462–17467.
13 Bozhueyuek KAJ, Watzel J, Abbood N, Bode HB. 2021. Synthetic zippers as an enabling tool for engineering of non-ribosomal peptide synthetases. *Angew Chem Int Ed Engl*. 60:17531–17538.
14 Huang H-M, Stephan P, Kries H. 2021. Engineering DNA-templated nonribosomal peptide synthesis. *Cell Chem Biol*. 28:221–227.e7.
15 Bozhüyük KAJ, *et al.* 2018. De novo design and engineering of non-ribosomal peptide synthetases. *Nat Chem*. 10:275–281.
16 Bozhüyük KAJ, *et al.* 2019. Modification and de novo design of non-ribosomal peptide synthetases using specific assembly points within condensation domains. *Nat Chem*. 11:653–661.
17 Crüsemann M, Kohlhaas C, Piel J. 2013. Evolution-guided engineering of nonribosomal peptide synthetase adenylation domains. *Chem Sci*. 4:1041–1045.
18 Kries H, Niquille DL, Hilvert D. 2015. A subdomain swap strategy for reengineering nonribosomal peptides. *Chem Biol*. 22:640–648.
19 Thong WL, *et al.* 2021. Gene editing enables rapid engineering of complex antibiotic assembly lines. *Nat Commun*. 12:6872.
20 Fischbach MA, Walsh CT, Clardy J. 2008. The evolution of gene collectives: how natural selection drives chemical innovation. *Proc Natl Acad Sci U S A*. 105:4601–4608.
21 Booth TJ, *et al.* 2022. Bifurcation drives the evolution of assembly-line biosynthesis. *Nat Commun*. 13:3498.
22 Medema MH, Cimermancic P, Sali A, Takano E, Fischbach MA. 2014. A systematic computational analysis of biosynthetic gene cluster evolution: lessons for engineering biosynthesis. *PLoS Comput Biol*. 10:e1004016.
23 Felsenstein J. 1985. Phylogenies and the comparative method. *Am Nat*. 125:1–15.
24 Qin C, Colwell LJ. 2018. Power law tails in phylogenetic systems. *Proc Natl Acad Sci U S A*. 115:690–695.
25 Dunn SD, Wahl LM, Gloor GB. 2008. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics*. 24:333–340.
26 Dutheil JY. 2012. Detecting coevolving positions in a molecule: why and how to account for phylogeny. *Brief Bioinform*. 13:228–243.
27 Wang S-W, Bitbol A-F, Wingreen NS. 2019. Revealing evolutionary constraints on proteins through sequence analysis. *PLoS Comput Biol*. 15:e1007010.
28 Zan J, *et al.* 2019. A microbial factory for defensive kahalalides in a tripartite marine symbiosis. *Science*. 364:eaaw6732.
29 Hamann MT, Scheuer PJ. 1993. Kahalalide F: a bioactive depsipeptide from the sacoglossan mollusk *Elysia rufescens* and the green alga *Bryopsis* sp. *J Am Chem Soc*. 115:5825–5826.
30 Hamann MT, Otto CS, Scheuer PJ, Dunbar DC. 1996. Kahalalides: bioactive peptides from a marine mollusk *Elysia rufescens* and its algal diet *Bryopsis* sp.1. *J Org Chem*. 61:6594–6600.
31 Salzberg SL, White O, Peterson J, Eisen JA. 2001. Microbial genes in the human genome: lateral transfer or gene loss? *Science*. 292:1903–1906.
32 Lawrence JG, Hendrix RW, Casjens S. 2001. Where are the pseudogenes in bacterial genomes? *Trends Microbiol*. 9:535–540.
33 Clugston SL, Sieber SA, Marahiel MA, Walsh CT. 2003. Chirality of peptide bond-forming condensation domains in nonribosomal peptide synthetases: the C5 domain of tyrocidine synthetase is a DCL catalyst. *Biochemistry*. 42:12095–12104.
34 Rausch C, Hoof I, Weber T, Wohlleben W, Huson DH. 2007. Phylogenetic analysis of condensation domains in NRPS sheds light on their functional evolution. *BMC Evol Biol*. 7:78.
35 Linne U, Doekel S, Marahiel MA. 2001. Portability of epimerization domain and role of peptidyl carrier protein on epimerization activity in nonribosomal peptide synthetases. *Biochemistry*. 40:15824–15834.
36 Calcott MJ, Ackerley DF. 2015. Portability of the thiolation domain in recombinant pyoverdine non-ribosomal peptide synthetases. *BMC Microbiol*. 15:162.
37 He R, *et al.* 2023. Knowledge-guided data mining on the standardized architecture of NRPS: subtypes, novel motifs, and sequence entanglements. *PLoS Comput Biol*. 19(5):e1011100.
38 Marahiel MA, Stachelhaus T, Mootz HD. 1997. Modular peptide synthetases involved in nonribosomal peptide synthesis. *Chem Rev*. 97:2651–2674.

39 Tarry MJ, Haque AS, Bui KH, Schmeing TM. 2017. X-ray crystallography and electron microscopy of cross- and multi-module nonribosomal peptide synthetase proteins reveal a flexible architecture. *Structure*. 25:783–793.e4.

40 Challis GL, Ravel J, Townsend CA. 2000. Predictive, structure-based model of amino acid recognition by nonribosomal peptide synthetase adenylation domains. *Chem Biol*. 7:211–224.

41 Conti E, Stachelhaus T, Marahiel MA, Brick P. 1997. Structural basis for the activation of phenylalanine in the non-ribosomal biosynthesis of gramicidin S. *EMBO J*. 16:4174–4183.

42 Stachelhaus T, Mootz HD, Marahiel MA. 1999. The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. *Chem Biol*. 6:493–505.

43 Bachmann BO, Ravel J. 2009. Chapter 8. Methods for in silico prediction of microbial polyketide and nonribosomal peptide biosynthetic pathways from DNA sequence data. In: Hopwood D, editor. *Complex enzymes in microbial natural product biosynthesis, part A: overview articles and peptides*. London: Academic Press. p. 181–217.

44 Blin K, *et al*. 2021. antiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucleic Acids Res*. 49:W29–W35.

45 Röttig M, *et al*. 2011. NRPSpredictor2—a web server for predicting NRPS adenylation domain specificity. *Nucleic Acids Res*. 39: W362–W367.

46 Zierep PF, Ceci AT, Dobrusin I, Rockwell-Kollmann SC, Günther S. 2021. SeMPI 2.0—a web server for PKS and NRPS predictions combined with metabolite screening in natural product databases. *Metabolites*. 11:13.

47 Fuchs SW, Proschak A, Jaskolla TW, Karas M, Bode HB. 2011. Structure elucidation and biosynthesis of lysine-rich cyclic peptides in *Xenorhabdus nematophila*. *Org Biomol Chem*. 9:3130–3132.

48 Imker HJ, Krahn D, Clerc J, Kaiser M, Walsh CT. 2010. N-acylation during glidobactin biosynthesis by the tridomain nonribosomal peptide synthetase module GlbF. *Chem Biol*. 17:1077–1083.

49 Kraas FI, Helmetag V, Wittmann M, Strieker M, Marahiel MA. 2010. Functional dissection of surfactin synthetase initiation module reveals insights into the mechanism of lipoinitiation. *Chem Biol*. 17:872–880.

50 Kraas FI, Giessen TW, Marahiel MA. 2012. Exploring the mechanism of lipid transfer during biosynthesis of the acidic lipopeptide antibiotic CDA. *FEBS Lett*. 586:283–288.

51 Liu Q, Fan W, Zhao Y, Deng Z, Feng Y. 2020. Probing and engineering the fatty acyl substrate selectivity of starter condensation domains of nonribosomal peptide synthetases in lipopeptide biosynthesis. *Biotechnol J*. 15:e1900175.

52 Zhong L, *et al*. 2021. Engineering and elucidation of the lipoinitiation process in nonribosomal peptide biosynthesis. *Nat Commun*. 12:296.

53 Chakravarty B, Gu Z, Chirala SS, Wakil SJ, Quiocho FA. 2004. Human fatty acid synthase: structure and substrate selectivity of the thioesterase domain. *Proc Natl Acad Sci U S A*. 101:15567–15572.

54 Huguenin-Dezot N, *et al*. 2019. Trapping biosynthetic acyl-enzyme intermediates with encoded 2,3-diaminopropionic acid. *Nature*. 565:112–117.

55 Samel SA, Wagner B, Marahiel MA, Essen L-O. 2006. The thioesterase domain of the fengycin biosynthesis cluster: a structural base for the macrocyclization of a non-ribosomal lipopeptide. *J Mol Biol*. 359:876–889.

56 Wang H, Fewer DP, Holm L, Rouhiainen L, Sivonen K. 2014. Atlas of nonribosomal peptide and polyketide biosynthetic pathways reveals common occurrence of nonmodular enzymes. *Proc Natl Acad Sci U S A*. 111:9259–9264.

57 Farag S, *et al*. 2019. Inter-modular linkers play a crucial role in governing the biosynthesis of non-ribosomal peptides. *Bioinformatics*. 35:3584–3591.

58 Lott JS, Lee TV. 2017. Revealing the inter-module interactions of multimodular nonribosomal peptide synthetases. *Structure*. 25:693–695.

59 Terlouw BR, *et al*. 2023. MIBiG 3.0: a community-driven effort to annotate experimentally validated biosynthetic gene clusters. *Nucleic Acids Res*. 51(D1):D603–D610.

60 Bozhuyuk KAJ, *et al*. 2024. Evolution-inspired engineering of nonribosomal peptide synthetases. *Science*. 383(6689):eadg4320.

61 Burgard C, *et al*. 2017. Genomics-guided exploitation of lipopeptide diversity in myxobacteria. *ACS Chem Biol*. 12(3):779–786.

62 Slatkin M. 2008. Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. *Nat Rev Genet*. 9: 477–485.