Title: Uncovering Lasonolide A biosynthesis using genome-resolved metagenomics

Running title (54 characters): Elucidation of Lasonolide A biosynthesis

Authors: Siddharth Uppal[a], Jackie L. Metz[b], René K.M. Xavier[b], Keshav Nepal[b], Dongbo Xu[b], Guojun Wang[b*], Jason C. Kwan[a*]

Affiliation: [a]Division of Pharmaceutical Sciences, School of Pharmacy, University of Wisconsin—Madison, Madison, Wisconsin, USA

[b]Harbor Branch Oceanographic Institute, Florida Atlantic University, Florida, USA

*Address correspondence to Guojun Wang, guojunwang@fau.edu; Jason C. Kwan, jason.kwan@wisc.edu

Abstract

Invertebrates, in particular sponges, have been a dominant source of new marine natural products. For example, lasonolide A (LSA) is a potential anti-cancer molecule isolated from the marine sponge *Forcepia* sp., with nanomolar growth inhibitory activity and a unique cytotoxicity profile against the National Cancer Institute 60 cell line screen. Here, we identified the putative biosynthetic pathway for LSA. Genomic binning of the *Forcepia* sponge metagenome revealed a gram-negative bacterium belonging to the phylum Verrucomicrobia as the candidate producer of LSA. Phylogenetic analysis showed this bacterium, herein named *Candidatus* Thermopylae lasonolidus, only has 88.78% 16S rRNA identity with the closest relative *Pedosphaera parvula* Ellin514, indicating it represents a new genus. The lasonolide A (*las*) biosynthetic gene cluster (BGC) was identified as a *trans*-AT polyketide synthase (PKS) pathway. When compared with its host genome, the *las* BGC exhibits a significantly different GC content and penta-nucleotide

frequency, suggesting a potential horizontal acquisition of the gene cluster. Furthermore, three copies of the putative *las* pathway were identified in the candidate producer genome. Differences between the three *las* repeats were observed including the presence of three insertions, two single-nucleotide polymorphisms and the absence of a stand-alone acyl carrier protein in one of the repeats. Even though the Verrucomicrobial producer shows signs of genome-reduction, its genome size is still fairly large (about 5Mbp) and when compared to its closest free-living relative contains most of the primary metabolic pathways, suggesting that it is in the early stages of reduction.

Importance

While sponges are valuable sources of bioactive natural products, a majority of these compounds are produced in small amounts by uncultured symbionts, hampering the study and clinical development of these unique compounds. Lasonolide A (LSA), isolated from marine sponge *Forcepia* sp., is a cytotoxic molecule active at nanomolar concentrations and causes premature chromosome condensation, blebbing, cell contraction and loss of cell adhesion, indicating a novel mechanism of action and making it a potential anti-cancer drug lead. However, its limited supply hampers progression to clinical trials. We investigated the microbiome of *Forcepia* sp. using culture-independent DNA sequencing to uncover how an uncultured bacterium produces LSA. This provides future opportunities for heterologous expression and cultivation efforts that may minimize LSA's supply problem.

Keywords: Lasonolide A, horizontal gene transfer, multiple repeats, Verrucomicrobia, *trans*-AT PKS, genome reduction

Introduction

Lasonolide A (LSA) is a cytotoxic polyketide derived from the marine sponge *Forcepia* sp. (**Fig. 1A and 1B**) (1). Out of its analogs (B–G) (**Fig. 1C**), LSA is the most potent (2) and exhibits IC_{50} values in the nanomolar range against certain cell lines in the National Cancer Institute 60 cell line screen (3). Furthermore, its unique mechanism of action - induction of premature chromosome condensation, loss of cell adhesion, activation of the RAF1 kinase in Ras pathway, along with cell blebbing and contraction (3–5) - makes it a promising candidate as a scaffold for future pharmaceutical development. However, a major challenge to its clinical development is the lack of availability. Scarcity and limited accessibility of the sponge prevent it from being a sustainable source of lasonolide A. Furthermore, the chemical synthesis of LSA is tedious and has poor yields, limiting its scalability (6–8).

It is well known that bacteria living in a symbiotic relationship with higher animals are valuable sources of novel bioactive secondary metabolites (9). In many instances, these molecules serve a protective function for the host but the identity of the microbial producer remains unknown (9–12). Furthermore, attempts to isolate these associated microbes are hampered by low cultivation success; it is estimated less than 1% of bacteria are currently culturable from the environment (13–15). These drawbacks have created the need to genetically engineer surrogate hosts for the sustainable and sufficient production of the desired natural products in the laboratory. The first step in engineering microbes for production of bioactive compounds is to identify the genes responsible for natural product synthesis, which can be elucidated through metagenomic analysis and cloning (16, 17). Based on its potent antitumor activity, it is likely that LSA also acts as a chemical defense within its host sponge. The structure of LSA very likely arises from an assembly-line type polyketide synthase (PKS), rather than the iterative PKSs that predominate in

fungi and other eukaryotes, and therefore the source is likely bacterial (18–20). Identifying the bacterium responsible for synthesizing LSA and elucidating its biosynthetic pathway will allow us to explore routes for LSA's heterologous expression, and potentially facilitate the synthesis of analogs.

Here, we describe a *trans*-AT PKS pathway (*las*) that is likely responsible for the biosynthesis of LSA. Furthermore, the entire *las* BGC has been captured on five overlapping fosmid clones and reassembled for the purpose of heterologous expression. We propose that the *las* BGC is present in a yet uncultivated bacterium belonging to a novel genus under the phylum Verrucomicrobia. Additionally, evidence suggests the *las* BGC is repeated thrice within the Verrucomicrobia symbiont with minor sequence variations between them. We also suggest that the *las* BGC has been horizontally acquired and has a codon adaptation index comparable to that of highly expressed genes. Finally, we show that the Verrucomicrobia symbiont is in very early stages of genome reduction and is likely to further reduce its size.

Results and discussion

Identification and capture of *las* BGC

In our initial studies, we constructed a high-capacity metagenomic DNA library consisting of ~600,000 *cfu* from *Forcepia* sp. sponges collected from the Gulf of Mexico (**Fig. 2A**) to search for potential *las* biosynthetic genes. The structure of LSA contains two tetrahydropyran rings and two β-methylations (21, 22) at C-13 and C-35 (**Fig. 2B**). These structural features have been identified in a variety of *trans*-AT PKS pathways but are rarely found in *cis*-AT PKS systems (23, 24) thus hinting that LSA is produced by a *trans*-AT PKS pathway (24). Therefore, clade-

guided degenerate primers targeted to conserved *trans*-AT PKS genes such as 3-hydroxy-3-methyglutaryl-CoA (HMG-CoA) synthase, free-standing ketosynthase (KS), acyl carrier protein (ACP), and two enoyl-CoA hydratases (ECH) were utilized for initial screening of the *Forcepia* fosmid library (**Table S1A**). From the metagenomic library, five fosmids were identified using these primers (fosmids 5-16, 6-71, 3-46, 1-80, and 4-77) resulting in the capture of approximately 48kb of the putative *las* BGC at its 3′ end (**Fig. S1A**). However, minimal progress was made toward capturing the remaining half of the BGC as primer walking failed to produce new hits in the region upstream of fosmid 5-16. Therefore, we sequenced the metagenome of *Forcepia* sp. and searched for *trans*-AT PKS BGCs. DNA was extracted from two different regions (referred to as Forcepia_v1 and Forcepia_v2) of the same sponge, and subjected to whole genome shotgun metagenomic sequencing. The reads were trimmed, assembled and then binned into metagenome assembled genomes (MAGs). The metagenomes were found to be abundant in Acidobacteria, Proteobacteria and Chloroflexota (**Fig. 2C and S1B**), with 56 and 55 MAGs recovered from the two metagenomes. Based on MiMAG (25) standards for completeness and contamination, 11 and 6 MAGs were high quality and 21 and 19 MAGs were medium quality, for Forcepia_v1 and Forcepia_v2, respectively (**Table S2**).

A tBLASTN (26) search of KS domains from publicly available *trans*-AT PKS pathways against our assembled metagenome was performed. In the case of Forcepia_v1, the top hits were all to a contig of length 98kbp labeled gnl|UoN|bin5_1_edit_8, thus strongly suggesting that this contig contains *trans*-AT PKS genes and may possess the potential LSA pathway. Contig gnl|UoN|bin5_1_edit_8 was manually inspected and corrected for sequence gaps (**Text S1**). With the exception of a 1.1kbp contig annotated as containing a *trans*-AT PKS pathway with a truncated condensation domain (in bin3674_131), analysis of the metagenome using

AntiSMASH (27) (**Fig. 2D**) did not reveal any other BGC with plausible size and genes for the synthesis of LSA. Contig gnl|UoN|bin5_1_edit_128 (3.6 kbp) which was connected to the 5′ end of gnl|UoN|bin5_1_edit_8 was found to contain a stand-alone ACP domain and about 47 amino acid residues which completed the terminal KS domain of gnl|UoN|bin5_1_edit_8 (see multiple repeats of the *las* BGC). Both the contigs were assembled together and annotation of genes and biosynthetic domains within this assembly re-affirmed that they are likely involved in LSA synthesis. We termed the gene cluster deemed relevant to LSA biosynthesis as *las* BGC_v1. Furthermore, the sequence of *las* BGC_v1 was also in alignment with fosmids identified from the metagenomic library. A screening strategy was then developed for isolating the previously missing 5′ end of the pathway from the metagenomic library (fosmids 5-41, 2-18 and 2-13) (**Fig. S1A and Table S1B**).

Inspection of the MAGs revealed that gnl|UoN|bin5_1_edit_8 binned with genome bin75_1. However, to our surprise, visual inspection of the assembly graph (**Fig. S1C**) in BANDAGE (29) indicated that gnl|UoN|bin5_1_edit_8 is present between contigs belonging to bin5_1 (phylum Verrucomicrobia). Furthermore, mapping the paired-end reads on the genomic bin (**Fig. S1D**) showed that multiple read pairs aligned across the contig junction. The terminal connections between contig gnl|UoN|bin5_1_edit_8 and several contigs in bin5_1 were verified via PCR (**Table S1C**) and Sanger sequencing of the amplicons using metagenomic DNA as the template. Based on this evidence gnl|UoN|bin5_1_edit_8 was manually placed with bin5_1, as well as additional contigs (**Text S1**).

In the case of Forcepia_v2, tBLASTN of KS domains hit to eight different contigs which could be assembled together (https://www.geneious.com) (**Fig. S1E**). Except for contig gnl|UoN|bin4_1_edit_10 the other seven contigs assembled into a single large contig of 102kbp

(*las* BGC_v2). Similar to *las* BGC_v1, inspection of the assembly graph (**Fig. S1F)** and mapping of paired-end reads (**Fig. S1G**) revealed that contigs forming *las* BGC_v2 have been binned incorrectly and should be part of the bin4_1 (phylum Verrucomicrobia). As a result, the contigs comprising *las* BGC_v2, as well as additional contigs (**Text S1**) were manually placed with bin4_1. No other contig containing a *trans*-AT PKS pathway was identified in the metagenome (**Fig. S1H**).

Alignment of *las* BGC from both Forcepia_v1 and Forcepia_v2 using clinker (28) revealed that these pathways are highly similar (**Fig. 2E**). The amino acid identity is 100% for most of the genes except for *lasJLO* where the amino acid identity is 98.37%, 99.84% and 99.83% respectively. The slightly lower identity of *lasJLO* is due to the insertion sequence present in *las* BGC_v2 but absent in *las* BGC_v1. These insertion variants were later identified to be present in some repeats of *las* BGC_v1 as well.

The putative symbiont genome carrying the *las* BGC (Forcepia_v1 bin5_1 and Forcepia_v2 bin4_1) was identified to belong to phylum Verrucomicrobiota, order Pedosphaerales, and genus UBA2970 by GTDB-TK v1.5.0 (database r202) (30). Excluding the *las* genes, the ANI of Forcepia_v1 bin5_1 and Forcepia_v2 bin4_1 is 99.9%, suggesting little strain heterogeneity between the sites in the sponge, beyond a small amount perhaps attributable to sequencing errors. To our knowledge, this is the first time a *trans*-AT PKS BGC has been reported in an organism belonging to order Pedosphaerales. A phylogenetic tree of 51 different Verrucomicrobia genomes (**Fig. S1I**) placed the LSA producer in subdivision 3 (NCBI taxonomy). The closest relative of the symbiont with a publicly available genome is *Pedosphaera parvula* Ellin514 (GCA_000172555.1), with 88.78% identity to the 16S ribosomal RNA sequence. As per the 16S taxonomic cutoffs proposed by *Yarza* et al. (31), this represents a new genus within the family

AAA164-E04 (as classified by GTDB-Tk (30)). We named the bacterium "*Candidatus* Thermopylae lasonolidus": Thermopylae is a tribute to the 300 Spartan hoplites and other Greek soldiers that fought at the battle of Thermopylae. The Spartans fought to protect Greece from Persians and the LSA-producing bacterium with its three copies of the *las* BGC (see below) is proposed to be protecting the host sponge from predators. Lasonolidus suggests the bacterium is associated with lasonolide A and also rhymes with the Spartan king of the 300 hoplites, Leonidas. Despite being the putative producer of LSA, "*Ca* T. lasonolidus" is not highly abundant in the metagenome, having a relative abundance of just over 2.65% in Forcepia_v1 and 1.78% in Forcepia_v2 (**Fig. 2C and Fig. S1B**).

With the aid of metagenome sequence, additional fosmids covering the 5□-end were acquired (Fig. S1), which enabled us to capture the *las* BGC minimally on 5 fosmids (Fig. 3) and to subsequently reassemble the BGC into a plasmid for heterologous expression. Network analysis with BiG-SCAPE (32) revealed no shared families with MIBiG reference BGCs indicating the novelty of the *las* BGC.

Model for lasonolide biosynthesis by *las* BGC

The proposed biosynthetic scheme for the synthesis of LSA based on the *las* BGC is shown in **Fig. 3**. The complete *las* BGC consists of six *trans*-AT PKS proteins (*lasHJLMNO*), ten accessory genes (*lasCDEFIKPQRS*) and five genes with no or unknown role in LSA synthesis (*lasABGTU*). Phylogenetic analysis of 944 different KS domains (**Fig. S2**) was used to predict KS substrate specificity (33), which was found to be similar to the proposed biosynthetic model. The pathway is predicted to be collinear with the first KS domain of *lasH* clustering into the same clade as other starter KS domains in the KS phylogenetic tree. Moreover, the last *trans*-AT

178 PKS protein (*lasO*) contains a condensation domain, similar to those found in nonribosomal
179 peptide synthetase pathways, as its terminal domain, which is proposed to be responsible for
180 cyclizing and cleaving the final PKS product (24).

181 An acylhydrolase (AH) domain is often used in *trans*-AT PKS systems for proofreading by
182 cleaving the acyl units from stalled sites (34, 35). AHs are closely related to acyltransferase (AT)
183 domains, which are involved in the addition of malonyl-*S*-coenzyme A extender units on the
184 phosphopantetheine arms on ACP domains (24, 36). LasE (AH) and LasF (AT) were correctly
185 identified as AH and AT domains respectively, based on the presence of active site residues (**Fig.**
186 **S3A**) and phylogeny (35) (**Fig. S3B**). The accessory proteins LasCDRS include enzymes known
187 to be involved in β-branch formation at module 1 and 10 (21). The ACPs at module 1 and 10
188 contain a conserved tryptophan which is involved in interacting with β-branching enzymes (37,
189 38). LasR was identified to be responsible for dehydration (ECH1) while LasH ECHb and LasO
190 ECHb to be responsible for decarboxylation (ECH2) during β-formation (39) (**Fig. S3C**). Due to
191 their truncated size and lack of homology to the conserved sites needed for oxyanion hole
192 formation LasH ECHa and LasO ECHa are proposed to be inactive (40, 41) (**Fig. S3D and S3E**).
193 An endo-β-methyl ($\Delta^{\alpha,\beta}$-unsaturated β-methyl) is predicted to form on module 10. The presence
194 of a truncated ECH domain just upstream of ECH2 domain has been commonly observed with
195 the formation of exo-β-methylene (β,γ-unsaturated β-methylene), but to our knowledge this is
196 the first time such an architecture has been reported to form a endo-β-methyl (38). Based on the
197 collinearity of the pathway suggested above and the split module architecture (KS-DH MOX
198 ACP-KS) associated with different Baeyer-Villiger (BV) monooxygenases as seen in oocydin
199 and sesbanimide biosynthesis (42–44) we propose LasI to be involved in BV oxidation and LasK
200 in the addition of a hydroxyl group. Based on the recent reports that the most common

transformation by cytochrome P450 enzymes in PKS biosynthesis is C-H hydroxylation (45) we suggest LasP to be oxidizing C-31. Another accessory protein, the enoylreductase (ER) domain LasQ (46) is proposed to be acting in *trans* as observed in other pathways including lagriamide (47), patellazoles (48) and bacillaene (24, 49).

Due to the disruption of the catalytically active residues (CHH, **Fig. S3F**), we predict certain KS domains to be inactive (LasL KS1, LasL KS4, LasM KS5 and LasO KS7). We propose that the ACP domain of LasL directly takes the molecule from the first ACP of LasJ and thus we predict the KS domain in LasJ to be catalytically inactive despite the presence of catalytic residues, as observed in lagriamide, lankacidin, and etnangien pathways (24, 36). Likewise, the alignment of ketoreductase (KR) domains (**Fig. S3G**) allowed us to identify the ones lacking the KSY catalytic triad and thus spot the inactive KR domain in module 2 (LasL KR1). Additionally, it was found that the predicted stereoconfiguration of KR products (50, 51) in the *las* BGC, matched the configuration of the equivalent moieties within the LSA structure produced by total synthesis (8). The absence of a KR domain required in module 14 is proposed to be compensated by a *trans*-acting KR likely from the following module as proposed in the patellazole (48) pathway.

We were able to identify two pyran synthase (PS) domains (in module 7 and module 13) based on their phylogeny (**Fig. S4A**) and alignment (52, 53) (**Fig. S4B**). These PS domains are at the correct position in the *las* BGC to insert the pyran rings required to synthesize LSA. Even though module 13 lacks a DH domain required for pyran ring formation, we predict this role to be played by a *trans*-acting DH domain as commonly seen in *trans*-AT PKS pathways (24). Similarly, we were able to identify double bond-shifting DH domains in module 4 (LasM DH1) and 8 (LasN DH1) by the absence of both proline at the HxxxGxxxxP motif and of Glutamine/

Histamine at the DxxQ/H (**Fig. S4C**) (54) motif. Moreover, alignment of the DH domains allowed us to identify the presence of inactive DH domains in module 2 (LasH DH1) and 6 (LasM DH2) by the absence of catalytic histidine at the HxxxGxxxxP motif and catalytic aspartic acid at the DxxxQ/H motif (**Fig. S4D**). LasL DH3 has a serine in place of proline in its HxxxGxxxxP motif. Alignment with different DH domains with serine in the HxxxGxxxxP motif revealed a mixture of domains annotated as active and inactive (**Fig. S4E**). The majority of times, when the DH domain had the conserved histidine in the HxxxGxxxxP motif it was annotated as active. Based on this we propose LasL DH3 to be active. Specific primers were designed based on the *las* BGC sequence and used to identify additional fosmids so that the whole pathway could be assembled from five overlapping clones for future heterologous expression.

For the biosynthesis of other LSA analogs, we propose that all of them except for lasonolide D are modified post PKS (**Fig. 4**). The cytochrome P450 LasP is predicted to oxidize LSA at C-37 and C-36 leading to the synthesis of lasonolide B and C respectively. However, in the complete biosynthesis of lasonolide B it is unclear how the methyl group is transferred from C-38 in LSA to C-36 in lasonolide B. Recently it was shown that serine hydrolase activity of lipid droplet-associated hydrolase is responsible for cleaving the ester bond in LSA and yielding the active form of the molecule, i.e. lasonolide F (55). Due to its hydrophobicity LSA is able to easily diffuse into the plasma membrane and into lipid droplets, where it is converted into lasonolide F, which is more hydrophilic and therefore able to diffuse out of the lipid droplet and into the cytoplasm to exhibit its cytotoxic effect (55). Lasonolide C seems to undergo an esterification reaction with a long-chain fatty acid ($CH_3(CH_2)_{10}COOH$) to produce lasonolide G. We suggest that lasonolide E is also biosynthesized by a *trans*-esterification reaction, by reacting with an

ethanol molecule. As with the production of LSA, we suggest that for the biosynthesis of lasonolide D the molecule passes through the entire *las* BGC, however, the starter molecule in this case is an acetate instead of a malonate that gets loaded on the ACP of LasJ, with LasH and LasI being inactive.

Multiple repeats of the *las* BGC

The k-mer coverage of the *las* BGC (400.165× for *las* BGC_v1 and 159.02× for *las* BGC_v2) is roughly three times that of "*Ca.* T. lasonolidus" (135.16x in Forcepia_v1 and 48.24x in Forcepia_v2). The 3× coverage suggests three repeats of the putative *las* BGC. Visual inspection of the assembly graph as well as mapping of the paired-end reads onto "*Ca.* T. lasonolidus" allowed us to identify three connections on the 3′ end of *las* BGC but only two connections on the 5′ end of the pathway (contig 7 and 8) (**Fig. 5 and Table S3**). Another contig (contig 5) was observed to be connected to *las* BGC about 3 kbp (3.6 kbp for *las* BGC_v1 and 3.7 kbp for *las* BGC_v2) from the 5 end of *las* BGC. This suggests that the majority of *las* BGC (about 98kbp) is repeated thrice with a 3 kbp segment of the pathway (contig 6) is repeated twice (**Fig. 5**). The two repeats of contig 6 were further verified by more than twice paired end reads mapping to it as compared to contig 5 (56) as well as its 2× coverage when compared to "*Ca.* T. lasonolidus". All the connections between the *las* BGC and the bacterial genome were verified using PCR and Sanger sequencing of the amplicons. We believe that the three repeats of the *las* BGC might be involved in contributing to the increased expression of LSA through increased gene dosage (57).

On comparing the three repeats it was observed that the *las* BGC repeat connected to contig 5 lacks *lasC* (ACP domain, highlighted area in **Fig. 5**), which is predicted to play an important role

in *β*-branch formation. Furthermore, the same repeat which lacks *lasC* also shows the presence of an incomplete *lasD* (decarboxylating KS domain used in β-branching). Although this KS domain has the catalytic active site residues SHH, characteristic of decarboxylating KSs (38), it lacks about 47 amino acids that are present in the KS domain of the other two repeats connected to contig 6. On further investigation with GATK HaplotypeCaller (58, 59) we were able to detect three insertions and two single nucleotide polymorphisms (SNPs) between the three repeats of *las* BGC_v1 (**Fig. 6 and Table 1**). This was further supported by the allelic depth (AD) - informative reads supporting each allele - and phred-scaled likelihoods (PL) of the possible genotypes. The genotype quality (GQ) which represents the confidence in the PL values was 99 for all five variants, which is the maximum value GATK reports for GQ. Furthermore, alignment of *las* BGC_v1 with *las* BGC_v2 revealed that *las* BGC_v2 contains all the variants that were called by GATK, thus further supporting their presence. All the three insertions are multiples of three base pairs (60bp, 24bp and 54bp), and are thus not causing any frame-shift mutations. Moreover, all the three insertions lie between *trans*-AT PKS domains, suggesting they do not contribute to functional differences. Change in one base from G to A at 93,995 bp does not result in a change in the amino acid sequence as both codons (TAT and TAC) encode for tyrosine. Finally, the change in base from A to G at 95,154 bp lies just outside *lasS*, i.e. in the non-coding region. The above-mentioned differences in the three repeats of *las* BGC indicate that the repeats have been present long enough to allow divergence. However, the differences between the three repeats are not predicted to affect the function of the *las* BGC.

Evidence for horizontal gene transfer

During the binning process by Autometa (60), Barnes-Hut Stochastic Neighbor Embedding (BH-tSNE) was used to reduce 5-mer frequencies to two dimensions. Generally, contigs belonging to the same genome would have similar 5-mer frequency and would be expected to cluster close to each other (61, 62). Visualization of the dimension-reduced data (**Fig. 7A–B and S5A–B)**, revealed that the *las* BGC contigs significantly differ in their 5-mer frequency from "*Ca.* T. lasonolidus", suggesting that the *las* BGC could have been recently horizontally acquired. Furthermore, the GC% of the *las* BGC is significantly different ($p < 0.05$, ANOVA followed by Tukey HSD) from annotated, hypothetical and pseudogenes (**Fig. 7C and S5C**) providing further evidence for horizontal transfer of the *las* BGC.

Codon adaptation index (CAI) compares the synonymous codon usage of a gene and that of a reference set along with measuring the synonymous codon usage bias (63). The CAI for the *las* BGC was significantly different ($p < 0.05$, ANOVA followed by Tukey HSD) from the annotated, hypothetical and pseudogenes, but matched that of highly expressed genes (i.e. ribosomal proteins) (**Fig. S5D-E**). Thus, despite its horizontal acquisition, the BGC's codon usage has been adapted to be efficiently translated even though the 5-mer composition is still different when compared to the rest of the "*Ca.* T. lasonolidus" genome.

The genome of the putative lasonolide producing symbiont

"*Ca.* T. lasonolidus", with multiple *las* BGC repeats, represents an important addition to the growing collection of symbiotic Verrucomicrobia ("*Candidatus* Didemnitutus mandela" and "*Candidatus* Synoicihabitans palmerolidicus") being identified with repeated *trans*-AT PKS

BGCs (57, 64, 65). Recently, two simultaneous studies have also identified a *trans*-AT PKS BGC for pateamine in a bacterium ("*Candidatus* Patea custodiens") belonging to phylum Kiritimatiellaeota (66, 67), a recently proposed phylum which was previously classified within Verrucomicrobia (68). This highlights the importance of this understudied phylum as an important producer of natural products. "*Ca.* T. lasonolidus" is little over 5 Mbp long and has GC percentage of about 53%. It is estimated to be 99% complete, is 1.35% contaminated (69), has tRNAs for all amino acids and complete 5S, 16S and 23S rRNA genes. Based on MIMAG standards (25) the bin is classified as a high-quality MAG. Detailed statistics of the putative LSA producer are provided in **Table 2**.

Eukaryotic-like proteins (ELPs) are known to be present in genomes of sponge symbionts and have been found to play an important role in regulating their interaction with the host sponge (70–73). It is hypothesized that interaction with ELPs allow the symbiotic bacteria to evade phagocytosis by the sponge, thus allowing discrimination between food and symbiont bacteria (72, 74). A number of ELPs were identified in "*Ca.* T. lasonolidus" (**Table 3, and Table S4A**), thus suggesting a symbiotic relationship of the bacterium with the *Forcepia* sp.

Bacterial microcompartments (BMCs) are organelles that enclose enzymes within a selectively permeable proteinaceous shell (75), and they are rare among bacteria. Members of the phyla Planctomycetes and Verrucomicrobia have a unique BMC gene cluster called the Planctomycetes-Verrucomicrobia bacterial microcompartment (PV BMC) which is responsible for production of microcompartment shell proteins BMC-P and BMC-H, as well as degradation of L-rhamnose, L-fucose and fucoidans (71, 76, 77). Genes encoding the PV BMC cluster were identified in "*Ca.* T. lasonolidus" (**Table S4B**), and the respective gene clusters in Forcepia_v1 "*Ca.* T. lasonolidus" and Forcepia_v2 "*Ca.* T. lasonolidus" were compared using clinker (28)

and were found to be 100% identical to each other. One interesting finding was that the identified PV BMC clusters had a DNA-methyltransferase and PVUII endonuclease gene between the first and the second BMC-H genes. This is different from the usual arrangement of the PV BMC gene cluster where both the BMC-H genes lie next to each other and the cluster lacks DNA-methyltransferase and PVUII endonuclease genes (**Fig. 8**). The presence of PV BMC genes in the "*Ca*. T. lasonolidus" genome suggests that it possesses bacterial microcompartments and that they might be involved in L-fucose and L-rhamnose degradation. Despite repeated attempts, we only found rhamnulokinase and fumarylacetoacetate hydrolase family proteins in the "*Ca*. T. lasonolidus" genomes and failed to identify other complementary enzymes involved in the degradation of L-fucose and L-rhamnose. However, other enzymes involved in carbohydrate metabolism including glycoside hydrolases, carbohydrate binding module, polysaccharide lyase, carbohydrate esterases and glycoside transferase were detected (**Table S4C**) indicating that "*Ca*. T. lasonolidus" is capable of polysaccharide degradation, something that is observed in a number of marine Verrucomicrobia (78–80).

A characteristic of obligate host-symbiont relationships is the loss of symbiont genes which are required for independent survival. The early stages of genome reduction are characterized by reduced coding density, and a high number of pseudogenes (81–83). We compared "*Ca* T. lasonolidus" with its closest free-living relative - *Pedosphaera parvula* Ellin514 (GCA_000172555.1). The draft genome of *P. paruva* Ellin514 is 7.41Mbp long, about 2.2 Mbp longer than "*Ca*. T. lasonolidus". Furthermore, in *P. paruva* Ellin514 only 0.5% of total ORFs were found to be pseudogenes (57, 84, 85) as opposed to about 16% in "*Ca*. T. lasonolidus" (**Fig. 9A-B**). Another indication of ongoing genome reduction comes from the fact that a much smaller percentage of genes with annotated function were identified in "*Ca.* T. lasonolidus" as compared

to *P. paruva* Ellin514 (**Fig. 9C**), perhaps indicating sequence degradation and divergence from functionally-annotated genes. Moreover, when compared with *P. paruva* Ellin514, "*Ca.* T. lasonolidus" lacks genes involved in DNA repair, DNA replication, chemotaxis and nucleotide metabolism (**Fig. 9D**), a trend which is commonly observed in symbionts undergoing genome reduction (81). However, "*Ca.* T. lasonolidus" contains most of the primary metabolic pathways (**Fig. 9E**) when compared to *P. paruva* Ellin514, and has a fairly large genome to be classified as reduced. Based on the above evidence we suggest that "*Ca.* T. lasonolidus" is in early stages of genome reduction. This hypothesis is also supported by its low coding density of ~72% (without pseudogenes), relative to the average coding density of 85-90% for free-living bacteria (81) which suggests a recent transitional event, such as host restriction (81).

Due to its potency and unique mechanism of action, LSA is considered a potential anti-cancer drug lead; however, its limited supply has hampered its transition to clinical trials. The evidence provided here suggests that LSA is synthesized by a yet uncultured Verrucomicrobial symbiont, which harbors three copies of the putative *las* BGC. The detailed analysis of the biosynthetic scheme, genome characteristics of the putative producer as well as the capture of the *las* BGC on a plasmid will aid future cultivation and heterologous expression efforts.

Acknowledgments

The authors acknowledge Amy Wright and Shirley Pomponi for providing sponge specimens, and Amy Wright, Shirley Pomponi and Peter McCarthy for valuable discussions during the project. The authors also thank Samantha C. Waterworth for fruitful discussions, John Barkei for discussions on heterologous expression strategies, and Chase Clark for providing DH sequences for sequence comparisons.

Funding

Support was provided by NCI (R21 CA209189) and a Start-up Fund from Harbor Branch Oceanographic Institute Foundation. The sample used in the study was collected with funds from a grant from the State of Florida Board of Education awarded to Florida Atlantic University for the Center of Excellence in Biomedical and Marine Biotechnology. This material is based upon work supported by the National Science Foundation under Grant No. DBI 1845890.

Data availability

The data associated with this study is deposited under BioProject RJNA833117. The WGS reads have been deposited in the sequence read archive (SRA) with accessions SRR18966768 (Forcepia_v1) and SRR18966767 (Forcepia_v2).

Methods

For full details see **Text S1.**

Sponge collection

Forcepia sp. (class, *Demospongiae*; order, *Poecilosclerida*; family, *Coelosphaeridae*) was collected in August of 2005 using the HBOI Johnson Sea Link submersible. Samples were collected at a depth of 70m from the Gulf of Mexico (26.256573N, 83.702772W) on the Pulley Ridge (https://shinyapps.fau.edu/app/bmr). The sponge samples were immediately frozen at −80°C. The sample ID was 12-VIII-05-1-006 200508121006 2005-08-12 JSL I-4837 (HBOI) *Forcepia* sp. 131921.

DNA purification and sequencing

The sponge hologenome was extracted using a modified cetyl trimethylammonium bromide (CTAB) DNA extraction method (87) and then size-fractionated by low melting point gel electrophoresis. DNA fragments greater than 40 kb were recovered from the gel and used for fosmid library preparation (**Text S1**) as well as metagenomic sequencing. Two rounds of sequencing were performed for different DNA extracts from the *Forcepia* sp. sponge. For the first round (referred to as Forcepia_v1) Illumina TruSeq DNA libraries were prepared and sequenced by RTL Genomics using an Illumina MiSeq sequencer giving us 108 million paired-end reads with length of 151bp. For the second round of sequencing (referred to as Forcepia_v2) Illumina Nextera libraries were prepared and sequenced using a NovaSeq 6000 sequencer giving us 303 million paired-end reads with length of 150 bp. Fosmids were sequenced by RTL Genomics and Genewiz.

Identification and annotation of *las* BGC

Identification of the *las* BGC was done using tBLASTN (26), where KS domains from different *trans*-AT PKS pathways were used as a query against the metagenomic assembly (assembled using MetaSpades (88), see **Text S1**). Genes for each bin were called and annotated using Prokka v1 (89, 90). MetaSpades contig headers were replaced by their respective Prokka headers to maintain consistency with the annotation file submitted to NCBI. Genes on contigs making up the *las* BGC were not called correctly by Prokka (89, 90) and were thus annotated manually in Artemis (91) with the help of AntiSMASH (27), CDD (92) and SMART (93, 94).

Functional analysis of the "*Ca*. T. lasonolidus" genome

Genes called using Prokka v1 were used for all functional analysis (89, 90). PV-BMC clusters were identified in "*Ca*. T. lasonolidus" using Interproscan v5.52-86.0 (95) and CDD (92). Initial identification of ELPs was done using Diamond BLASTP against the diamond-formatted nr database (using -k 1 --max-hsps 1 options) (96) and Interproscan v5.52-86.0 (95). This was followed by verification of non-pseudogenes using CDD (92). Enzymes involved in carbohydrate metabolism were detected using dbCAN2 (97) where genes annotated by ≥2 tools (out of HMMER, Diamond and Hotpep) were kept. COG categories were identified using the eggNOG mapper online server (98, 99).

The genome of *P. paruva* Ellin514 was downloaded from Genbank (GCA_000172555.1) and genes were called and annotated using Prokka v1 (89, 90). Primary metabolic pathways were identified for non-pseudogenes with kofamscan using the --mapper flag (100) and annotated against the KEGG database (101–103). The matrix with presence/absence of different enzymes was constructed in RStudio (104). Completeness of metabolic pathways was identified using KEGG-Decoder (86).

References

1. Horton PA, Koehn FE, Longley RE, McConnell OJ. 1994. Lasonolide A, a new cytotoxic macrolide from the marine sponge *Forcepia* sp. J Am Chem Soc 116:6015–6016.

2. Wright AE, Chen Y, Winder PL, Pitts TP, Pomponi SA, Longley RE. 2004. Lasonolides C-G, five new lasonolide compounds from the sponge *Forcepia* sp. J Nat Prod 67:1351–1355.

3. Isbrucker RA, Guzmán EA, Pitts TP, Wright AE. 2009. Early effects of lasonolide A on

pancreatic cancer cells. J Pharmacol Exp Ther 331:733–739.

4. Jossé R, Zhang Y-W, Giroux V, Ghosh AK, Luo J, Pommier Y. 2015. Activation of RAF1 (c-RAF) by the marine alkaloid lasonolide A induces rapid premature chromosome condensation. Mar Drugs 13:3625–3639.

5. Zhang Y-W, Ghosh AK, Pommier Y. 2012. Lasonolide A, a potent and reversible inducer of chromosome condensation. Cell Cycle 11:4424–4435.

6. Yang L, Lin Z, Shao S, Zhao Q, Hong R. 2018. An enantioconvergent and concise synthesis of lasonolide A. Angew Chem Int Ed Engl 57:16200–16204.

7. Yang L, Lin Z, Shao S, Zhao Q, Hong R. 2019. Corrigendum: An enantioconvergent and concise synthesis of lasonolide A. Angew Chem Int Ed Engl 58:4431.

8. Trost BM, Stivala CE, Fandrick DR, Hull KL, Huang A, Poock C, Kalkofen R. 2016. Total synthesis of (-)-lasonolide A. J Am Chem Soc 138:11690–11701.

9. Piel J. 2009. Metabolites from symbiotic bacteria. Nat Prod Rep 26:338–362.

10. Lopanik NB. 2014. Chemical defensive symbioses in the marine environment. Funct Ecol 28:328–340.

11. Flórez LV, Biedermann PHW, Engl T, Kaltenpoth M. 2015. Defensive symbioses of animals with prokaryotic and eukaryotic microorganisms. Nat Prod Rep 32:904–936.

12. Oliver KM, Smith AH, Russell JA. 2014. Defensive symbiosis in the real world - advancing ecological studies of heritable, protective bacteria in aphids and beyond. Funct Ecol 28:341–355.

458 13. Bodor A, Bounedjoum N, Vincze GE, Erdeiné Kis Á, Laczi K, Bende G, Szilágyi Á,
459 Kovács T, Perei K, Rákhely G. 2020. Challenges of unculturable bacteria: Environmental
460 perspectives. Rev Environ Sci Technol 19:1–22.

461 14. Hofer U. 2018. The majority is uncultured. Nat Rev Microbiol 16:716–717.

462 15. Vartoukian SR, Palmer RM, Wade WG. 2010. Strategies for culture of “unculturable”
463 bacteria. FEMS Microbiol Lett 309:1–7.

464 16. Stevens DC, Hari TPA, Boddy CN. 2013. The role of transcription in heterologous
465 expression of polyketides in bacterial hosts. Nat Prod Rep 30:1391–1411.

466 17. Trindade M, van Zyl LJ, Navarro-Fernández J, Abd Elrazak A. 2015. Targeted
467 metagenomics as a tool to tap into marine natural product diversity for the discovery and
468 production of drug candidates. Front Microbiol 6:890.

469 18. Nivina A, Yuet KP, Hsu J, Khosla C. 2019. Evolution and diversity of assembly-line
470 polyketide synthases. Chem Rev 119:12524–12547.

471 19. Cox RJ. 2007. Polyketides, proteins and genes in fungi: Programmed nano-machines begin
472 to reveal their secrets. Org Biomol Chem 5:2010–2026.

473 20. Jenke-Kodama H, Sandmann A, Müller R, Dittmann E. 2005. Evolutionary implications of
474 bacterial polyketide synthases. Mol Biol Evol 22:2027–2039.

475 21. Calderone CT, Kowtoniuk WE, Kelleher NL, Walsh CT, Dorrestein PC. 2006.
476 Convergence of isoprene and polyketide biosynthetic machinery: Isoprenyl-S-carrier
477 proteins in the *pksX* pathway of *Bacillus subtilis*. Proc Natl Acad Sci U S A 103:8977–

8982.

22. Calderone CT. 2008. Isoprenoid-like alkylations in polyketide biosynthesis. Nat Prod Rep 25:845–853.

23. Gu L, Wang B, Kulkarni A, Geders TW, Grindberg RV, Gerwick L, Håkansson K, Wipf P, Smith JL, Gerwick WH, Sherman DH. 2009. Metamorphic enzyme assembly in polyketide diversification. Nature 459:731–735.

24. Helfrich EJN, Piel J. 2016. Biosynthesis of polyketides by *trans*-AT polyketide synthases. Nat Prod Rep 33:231–316.

25. Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, Schulz F, Jarett J, Rivers AR, Eloe-Fadrosh EA, Tringe SG, Ivanova NN, Copeland A, Clum A, Becraft ED, Malmstrom RR, Birren B, Podar M, Bork P, Weinstock GM, Garrity GM, Dodsworth JA, Yooseph S, Sutton G, Glöckner FO, Gilbert JA, Nelson WC, Hallam SJ, Jungbluth SP, Ettema TJG, Tighe S, Konstantinidis KT, Liu W-T, Baker BJ, Rattei T, Eisen JA, Hedlund B, McMahon KD, Fierer N, Knight R, Finn R, Cochrane G, Karsch-Mizrachi I, Tyson GW, Rinke C, Lapidus A, Meyer F, Yilmaz P, Parks DH, Murat Eren A, Schriml L, Banfield JF, Hugenholtz P, Woyke T. 2017. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nat Biotechnol 35:725–731.

26. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: Architecture and applications. BMC Bioinform 10:1–9.

27. Blin K, Shaw S, Steinke K, Villebro R, Ziemert N, Lee SY, Medema MH, Weber T. 2019.

499 antiSMASH 5.0: Updates to the secondary metabolite genome mining pipeline. Nucleic
500 Acids Res 47:W81–W87.

501 28. Gilchrist CLM, Chooi Y-H. 2021. Clinker & clustermap.js: Automatic generation of gene
502 cluster comparison figures. Bioinformatics 37:2473–2475.

503 29. Wick RR, Schultz MB, Zobel J, Holt KE. 2015. Bandage: Interactive visualization of *de*
504 *novo* genome assemblies. Bioinformatics 31:3350–3352.

505 30. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. 2020. GTDB-Tk: A toolkit to classify
506 genomes with the Genome Taxonomy Database. Bioinformatics 36:1925–1927.

507 31. Yarza P, Yilmaz P, Pruesse E, Glöckner FO, Ludwig W, Schleifer K-H, Whitman WB,
508 Euzéby J, Amann R, Rosselló-Móra R. 2014. Uniting the classification of cultured and
509 uncultured bacteria and archaea using 16S rRNA gene sequences. Nat Rev Microbiol
510 12:635–645.

511 32. Navarro-Muñoz JC, Selem-Mojica N, Mullowney MW, Kautsar SA, Tryon JH, Parkinson
512 EI, De Los Santos ELC, Yeong M, Cruz-Morales P, Abubucker S, Roeters A, Lokhorst W,
513 Fernandez-Guerra A, Cappelini LTD, Goering AW, Thomson RJ, Metcalf WW, Kelleher
514 NL, Barona-Gomez F, Medema MH. 2020. A computational framework to explore large-
515 scale biosynthetic diversity. Nat Chem Biol 16:60–68.

516 33. Nguyen T, Ishida K, Jenke-Kodama H, Dittmann E, Gurgui C, Hochmuth T, Taudien S,
517 Platzer M, Hertweck C, Piel J. 2008. Exploiting the mosaic structure of *trans*-
518 acyltransferase polyketide synthases for natural product discovery and pathway dissection.
519 Nat Biotechnol 26:225–233.

520 34. Jensen K, Niederkrüger H, Zimmermann K, Vagstad AL, Moldenhauer J, Brendel N, Frank
521 S, Pöplau P, Kohlhaas C, Townsend CA, Oldiges M, Hertweck C, Piel J. 2012. Polyketide
522 proofreading by an acyltransferase-like enzyme. Chem Biol 19:329–339.

523 35. Jenner M, Afonso JP, Kohlhaas C, Karbaum P, Frank S, Piel J, Oldham NJ. 2016. Acyl
524 hydrolases from *trans*-AT polyketide synthases target acetyl units on acyl carrier proteins.
525 Chem Commun 52:5262–5265.

526 36. Piel J. 2010. Biosynthesis of polyketides by *trans*-AT polyketide synthases. Nat Prod Rep
527 27:996–1047.

528 37. Haines AS, Dong X, Song Z, Farmer R, Williams C, Hothersall J, Płoskoń E, Wattana-
529 Amorn P, Stephens ER, Yamada E, Gurney R, Takebayashi Y, Masschelein J, Cox RJ,
530 Lavigne R, Willis CL, Simpson TJ, Crosby J, Winn PJ, Thomas CM, Crump MP. 2013. A
531 conserved motif flags acyl carrier proteins for β-branching in polyketide synthesis. Nat
532 Chem Biol 9:685–692.

533 38. Walker PD, Weir ANM, Willis CL, Crump MP. 2021. Polyketide β-branching: Diversity,
534 mechanism and selectivity. Nat Prod Rep 38:723–756.

535 39. Slocum ST, Lowell AN, Tripathi A, Shende VV, Smith JL, Sherman DH. 2018.
536 Chemoenzymatic dissection of polyketide β-branching in the bryostatin pathway, p. 207–
537 236. *In* Moore, BS (ed.), Marine Enzymes and Specialized Metabolism - Part A. Academic
538 Press.

539 40. Gu L, Jia J, Liu H, Håkansson K, Gerwick WH, Sherman DH. 2006. Metabolic coupling of
540 dehydration and decarboxylation in the curacin A pathway: Functional identification of a

mechanistically diverse enzyme pair. J Am Chem Soc 128:9014–9015.

41. Matilla MA, Stöckmann H, Leeper FJ, Salmond GPC. 2012. Bacterial biosynthetic gene clusters encoding the anti-cancer haterumalide class of molecules: Biogenesis of the broad spectrum antifungal and anti-oomycete compound, oocydin A. J Biol Chem 287:39125–39138.

42. Kačar D, Cañedo LM, Rodríguez P, González EG, Galán B, Schleissner C, Leopold-Messer S, Piel J, Cuevas C, de la Calle F, García JL. 2021. Identification of *trans*-AT polyketide clusters in two marine bacteria reveals cryptic similarities between distinct symbiosis factors. Environ Microbiol 23:2509–2521.

43. Meoded RA, Ueoka R, Helfrich EJN, Jensen K, Magnus N, Piechulla B, Piel J. 2018. A polyketide synthase component for oxygen insertion into polyketide backbones. Angew Chem Int Ed Engl 57:11644–11648.

44. Hemmerling F, Meoded RA, Fraley AE, Minas HA, Dieterich CL, Rust M, Ueoka R, Jensen K, Helfrich EJN, Bergande C, Biedermann M, Magnus N, Piechulla B, Piel J. 2022. Modular halogenation, α-hydroxylation, and acylation by a remarkably versatile polyketide synthase. Angew Chem Int Ed Engl 61:e202116614.

45. Greule A, Stok JE, De Voss JJ, Cryle MJ. 2018. Unrivalled diversity: The many roles and reactions of bacterial cytochromes P450 in secondary metabolism. Nat Prod Rep 35:757–791.

46. Bumpus SB, Magarvey NA, Kelleher NL, Walsh CT, Calderone CT. 2008. Polyunsaturated fatty-acid-like *trans*-enoyl reductases utilized in polyketide biosynthesis. J Am Chem Soc

130:11614–11616.

47. Flórez LV, Scherlach K, Miller IJ, Rodrigues A, Kwan JC, Hertweck C, Kaltenpoth M. 2018. An antifungal polyketide associated with horizontally acquired genes supports symbiont-mediated defense in *Lagria villosa* beetles. Nat Commun 9:2478.

48. Kwan JC, Donia MS, Han AW, Hirose E, Haygood MG, Schmidt EW. 2012. Genome streamlining and chemical defense in a coral reef symbiosis. Proc Natl Acad Sci USA 109:20655–20660.

49. Chen X-H, Vater J, Piel J, Franke P, Scholz R, Schneider K, Koumoutsi A, Hitzeroth G, Grammel N, Strittmatter AW, Gottschalk G, Süssmuth RD, Borriss R. 2006. Structural and functional characterization of three polyketide synthase gene clusters in *Bacillus amyloliquefaciens* FZB 42. J Bacteriol 188:4024–4036.

50. Keatinge-Clay AT. 2007. A tylosin ketoreductase reveals how chirality is determined in polyketides. Chem Biol 14:898–908.

51. Caffrey P. 2003. Conserved amino acid residues correlating with ketoreductase stereospecificity in modular polyketide synthases. ChemBioChem 4:654–657.

52. Pöplau P, Frank S, Morinaka BI, Piel J. 2013. An enzymatic domain for the formation of cyclic ethers in complex polyketides. Angew Chem Int Ed Engl 52:13215–13218.

53. Wagner DT, Zhang Z, Meoded RA, Cepeda AJ, Piel J, Keatinge-Clay AT. 2018. Structural and functional studies of a pyran synthase domain from a *trans*-acyltransferase assembly line. ACS Chem Biol 13:975–983.

582 54. Gay DC, Spear PJ, Keatinge-Clay AT. 2014. A double-hotdog with a new trick: Structure
583 and mechanism of the *trans*-acyltransferase polyketide synthase enoyl-isomerase. ACS
584 Chem Biol 9:2374–2381.

585 55. Dubey R, Stivala CE, Nguyen HQ, Goo Y-H, Paul A, Carette JE, Trost BM, Rohatgi R.
586 2020. Lipid droplets can promote drug accumulation and activation. Nat Chem Biol
587 16:206–213.

588 56. Albertsen M, Hugenholtz P, Skarshewski A, Nielsen KL, Tyson GW, Nielsen PH. 2013.
589 Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of
590 multiple metagenomes. Nat Biotechnol 31:533–538.

591 57. Lopera J, Miller IJ, McPhail KL, Kwan JC. 2017. Increased biosynthetic gene dosage in a
592 genome-reduced defensive bacterial symbiont. mSystems 2:e00096–17.

593 58. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A,
594 Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella KV, Altshuler D, Gabriel S,
595 DePristo MA. 2013. From FastQ data to high-confidence variant calls: The genome analysis
596 toolkit best practices pipeline. Curr Protoc Bioinform 43:11.10.1–11.10.30.

597 59. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del
598 Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY,
599 Cibulskis K, Gabriel SB, Altshuler D, Daly MJ. 2011. A framework for variation discovery
600 and genotyping using next-generation DNA sequencing data. Nat Genet 43:491–498.

601 60. Miller IJ, Rees ER, Ross J, Miller I, Baxa J, Lopera J, Kerby RL, Rey FE, Kwan JC. 2019.
602 Autometa: Automated extraction of microbial genomes from individual shotgun

603 metagenomes. Nucleic Acids Res 47:e57.

604 61. Dick GJ, Andersson AF, Baker BJ, Simmons SL, Thomas BC, Yelton AP, Banfield JF.
605 2009. Community-wide analysis of microbial genome sequence signatures. Genome Biol
606 10:R85.

607 62. Laczny CC, Pinel N, Vlassis N, Wilmes P. 2014. Alignment-free visualization of
608 metagenomic data by nonlinear dimension reduction. Sci Rep 4:4516.

609 63. Sharp PM, Li W-H. 1987. The codon adaptation index-a measure of directional
610 synonymous codon usage bias, and its potential applications. Nucleic Acids Res 15:1281–
611 1295.

612 64. Avalon NE, Murray AE, Daligault HE, Lo C-C, Davenport KW, Dichosa AEK, Chain PSG,
613 Baker BJ. 2021. Bioinformatic and mechanistic analysis of the palmerolide PKS-NRPS
614 biosynthetic pathway from the microbiome of an Antarctic ascidian. Front Chem 9:802574.

615 65. Murray AE, Lo C-C, Daligault HE, Avalon NE, Read RW, Davenport KW, Higham ML,
616 Kunde Y, Dichosa AEK, Baker BJ, Chain PSG. 2021. Discovery of an Antarctic ascidian-
617 associated uncultivated Verrucomicrobia with antimelanoma palmerolide biosynthetic
618 potential. mSphere 6:e00759–21.

619 66. Michael R, N. HEJ, F. FM, Pakjira N, M. FC, Christian R, Tomas K, J. PM, L. WV, Jörn K,
620 Shinichi S, Jörn P. 2020. A multiproducer microbiome generates chemical diversity in the
621 marine sponge *Mycale hentscheli*. Proc Natl Acad Sci U S A 117:9508–9518.

622 67. Storey MA, Andreassend SK, Bracegirdle J, Brown A, Keyzers RA, Ackerley DF,

Northcote PT, Owen JG. 2020. Metagenomic exploration of the marine sponge *Mycale hentscheli* uncovers multiple polyketide-producing bacterial symbionts. mBio 11:e02997–19.

68. Spring S, Bunk B, Spröer C, Schumann P, Rohde M, Tindall BJ, Klenk H-P. 2016. Characterization of the first cultured representative of Verrucomicrobia subdivision 5 indicates the proposal of a novel phylum. ISME J 10:2801–2816.

69. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res 25:1043–1055.

70. Burgsdorf I, Handley KM, Bar-Shalom R, Erwin PM, Steindler L. 2019. Life at home and on the roam: Genomic adaptions reflect the dual lifestyle of an intracellular, facultative symbiont. mSystems 4:e00057–19.

71. Sizikov S, Burgsdorf I, Handley KM, Lahyani M, Haber M, Steindler L. 2020. Characterization of sponge-associated Verrucomicrobia: Microcompartment-based sugar utilization and enhanced toxin-antitoxin modules as features of host-associated Opitutales. Environ Microbiol 22:4669–4688.

72. Frank AC. 2019. Molecular host mimicry and manipulation in bacterial symbionts. FEMS Microbiol Lett 366:fnz038.

73. Díez-Vives C, Moitinho-Silva L, Nielsen S, Reynolds D, Thomas T. 2017. Expression of eukaryotic-like protein in the microbiome of sponges. Mol Ecol 26:1432–1451.

643 74. Nguyen MTHD, Liu M, Thomas T. 2014. Ankyrin-repeat proteins from sponge symbionts
644 modulate amoebal phagocytosis. Mol Ecol 23:1635–1645.

645 75. Kerfeld CA, Aussignargues C, Zarzycki J, Cai F, Sutter M. 2018. Bacterial
646 microcompartments. Nat Rev Microbiol 16:277–290.

647 76. Erbilgin O, McDonald KL, Kerfeld CA. 2014. Characterization of a planctomycetal
648 organelle: A novel bacterial microcompartment for the aerobic degradation of plant
649 saccharides. Appl Environ Microbiol 80:2193–2205.

650 77. Sichert A, Corzett CH, Schechter MS, Unfried F, Markert S, Becher D, Fernandez-Guerra
651 A, Liebeke M, Schweder T, Polz MF, Hehemann J-H. 2020. Verrucomicrobia use hundreds
652 of enzymes to digest the algal polysaccharide fucoidan. Nat Microbiol 5:1026–1039.

653 78. Cardman Z, Arnosti C, Durbin A, Ziervogel K, Cox C, Steen AD, Teske A. 2014.
654 Verrucomicrobia are candidates for polysaccharide-degrading bacterioplankton in an arctic
655 fjord of Svalbard. Appl Environ Microbiol 80:3749–3756.

656 79. Martinez-Garcia M, Brazel DM, Swan BK, Arnosti C, Chain PSG, Reitenga KG, Xie G,
657 Poulton NJ, Lluesma Gomez M, Masland DED, Thompson B, Bellows WK, Ziervogel K,
658 Lo C-C, Ahmed S, Gleasner CD, Detter CJ, Stepanauskas R. 2012. Capturing single cell
659 genomes of active polysaccharide degraders: An unexpected contribution of
660 Verrucomicrobia. PLoS One 7:e35314.

661 80. Herlemann DPR, Lundin D, Labrenz M, Jürgens K, Zheng Z, Aspeborg H, Andersson AF.
662 2013. Metagenomic *de novo* assembly of an aquatic representative of the verrucomicrobial
663 class *Spartobacteria*. mBio 4:e00569–12.

664 81. McCutcheon JP, Moran NA. 2011. Extreme genome reduction in symbiotic bacteria. Nat
665 Rev Microbiol 10:13–26.

666 82. Lo W-S, Huang Y-Y, Kuo C-H. 2016. Winding paths to simplicity: Genome evolution in
667 facultative insect symbionts. FEMS Microbiol Rev 40:855–874.

668 83. Dietel A-K, Merker H, Kaltenpoth M, Kost C. 2019. Selective advantages favour high
669 genomic AT-contents in intracellular elements. PLoS Genet 15:e1007778.

670 84. Lerat E, Ochman H. 2005. Recognizing the pseudogenes in bacterial genomes. Nucleic
671 Acids Res 33:3125–3132.

672 85. Waterworth SC, Flórez LV, Rees ER, Hertweck C, Kaltenpoth M, Kwan JC. 2020.
673 Horizontal gene transfer to a defensive symbiont with a reduced genome in a multipartite
674 beetle microbiome. mBio 11:e02430–19.

675 86. Graham ED, Heidelberg JF, Tully BJ. 2018. Potential for primary productivity in a
676 globally-distributed bacterial phototroph. ISME J 12:1861–1866.

677 87. Piel J, Hui D, Wen G, Butzke D, Platzer M, Fusetani N, Matsunaga S. 2004. Antitumor
678 polyketide biosynthesis by an uncultivated bacterial symbiont of the marine sponge
679 *Theonella swinhoei*. Proc Natl Acad Sci U S A 101:16222–16227.

680 88. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. 2017. metaSPAdes: A new versatile
681 metagenomic assembler. Genome Res 27:824–834.

682 89. Seemann T. 2014. Prokka: Rapid prokaryotic genome annotation. Bioinformatics 30:2068–
683 2069.

684 90. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal:
685 Prokaryotic gene recognition and translation initiation site identification. BMC Bioinform
686 11:119.

687 91. Carver T, Harris SR, Berriman M, Parkhill J, McQuillan JA. 2012. Artemis: An integrated
688 platform for visualization and analysis of high-throughput sequence-based experimental
689 data. Bioinformatics 28:464–469.

690 92. Lu S, Wang J, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI,
691 Marchler GH, Song JS, Thanki N, Yamashita RA, Yang M, Zhang D, Zheng C, Lanczycki
692 CJ, Marchler-Bauer A. 2020. CDD/SPARCLE: The conserved domain database in 2020.
693 Nucleic Acids Res 48:D265–D268.

694 93. Letunic I, Khedkar S, Bork P. 2021. SMART: Recent updates, new developments and status
695 in 2020. Nucleic Acids Res 49:D458–D460.

696 94. Letunic I, Bork P. 2018. 20 years of the SMART protein domain annotation resource.
697 Nucleic Acids Res 46:D493–D496.

698 95. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J,
699 Mitchell A, Nuka G, Pesseat S, Quinn AF, Sangrador-Vegas A, Scheremetjew M, Yong S-
700 Y, Lopez R, Hunter S. 2014. InterProScan 5: Genome-scale protein function classification.
701 Bioinformatics 30:1236–1240.

702 96. Buchfink B, Reuter K, Drost H-G. 2021. Sensitive protein alignments at tree-of-life scale
703 using DIAMOND. Nat Methods 18:366–368.

704 97. Zhang H, Yohe T, Huang L, Entwistle S, Wu P, Yang Z, Busk PK, Xu Y, Yin Y. 2018.
705 dbCAN2: A meta server for automated carbohydrate-active enzyme annotation. Nucleic
706 Acids Res 46:W95–W101.

707 98. Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. 2021. eggNOG-
708 mapper v2: Functional annotation, orthology assignments, and domain prediction at the
709 metagenomic scale. Mol Biol Evol 38:5825–5829.

710 99. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, Mende
711 DR, Letunic I, Rattei T, Jensen LJ, von Mering C, Bork P. 2019. eggNOG 5.0: A
712 hierarchical, functionally and phylogenetically annotated orthology resource based on 5090
713 organisms and 2502 viruses. Nucleic Acids Res 47:D309–D314.

714 100. Aramaki T, Blanc-Mathieu R, Endo H, Ohkubo K, Kanehisa M, Goto S, Ogata H. 2020.
715 KofamKOALA: KEGG ortholog assignment based on profile HMM and adaptive score
716 threshold. Bioinformatics 37:2251–2252.

717 101. Kanehisa M, Sato Y. 2020. KEGG Mapper for inferring cellular functions from protein
718 sequences. Protein Sci 29:28–35.

719 102. Kanehisa M, Sato Y, Kawashima M. 2021. KEGG mapping tools for uncovering hidden
720 features in biological data. Protein Sci 31:47–52.

721 103. Kanehisa M, Goto S. 2000. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic
722 Acids Res 28:27–30.

723 104. Team R. 2020. RStudio: Integrated Development for R. RStudio, PBC, Boston, MA.

724

Figure Legends

Fig. 1. (A) Sponge *Forcepia* sp. as seen in the field. (B) The *Forcepia* sp. specimen used for DNA extraction (sample ID: 12-VIII-05-1-6). Photo credit: HBOI Marine Biomedical and Biotechnology Program. (C) The chemical structures of lasonolide A (LSA) and its analogs.

Fig. 2. (A) Collection site of *Forcepia* sp. sponge (dark red diamond, 26.256573N, 83.702772W). **(B)** Features in lasonolide A (LSA) characteristic of biosynthesis by a *trans*-AT PKS pathway. **(C)** Relative abundance of different phyla in the sequenced Forcepia_v1 metagenome. Each block shows the relative abundance of each metagenome-assembled genome (MAG), with colors representing the phylum they belong to. The *las* biosynthetic gene cluster (BGC)-carrying bin is marked with a star. **(D)** BGC distribution in *Forcepia*_v1 sp. metagenome. AntiSMASH (27) annotations of bacterial contigs greater than 500 bp are shown. Each bar indicates a MAG. Bars have been grouped by phylum. The star represents the MAG possessing *las* BGC. BGC annotations have been simplified into polyketide synthase (PKS), Type 1 PKS, Type 3 PKS, *trans*-AT PKS, nonribosomal peptide synthetase (NRPS), ribosomally synthesized, post-translationally modified peptide (RiPP), hgIE-KS, hgIE-KS-T1PKS, terpenes, RiPP-terpene and others. **(E)** Comparison of *las* BGC_v1 and *las* BGC_v2 using clinker (28). V1 refers to *las* BGC_v1 while V2 refers to *las* BGC_v2. Numbers in the boxes indicate amino acid identity as a fraction of 1.

Fig. 3. Proposed LSA biosynthetic scheme. Colored lines above the *las* BGC represent an alignment of individual fosmids to the pathway, the fosmids were subsequently assembled together in a plasmid. A cross indicates a domain predicted to be catalytically inactive. Open reading frames colored in gray represent genes with unknown or no role in LSA synthesis.

Numbers below domains indicate the module number and 'A' and 'B' denote the predicted stereoconfiguration of the KR product, as previously described (50, 51). Predicted substrate specificity of KS domains, obtained through phylogeny (**Fig. S3**) (33), are shown above each respective KS domain. Carbon 31 is highlighted in blue to represent the site where P450 LasP is predicted to act. Abbreviations: ACP, acyl carrier protein, also denoted by a filled black circle; AH, acylhydrolase; AT, acyltransferase; C, condensation; DH, dehydratase; ECH, enoyl-CoA reductase; ER, enoylreductase; HMGS, 3-hydroxy-3-methylglutaryl-CoA synthase; KR, ketoreductase; KS, ketosynthase; MOX, monooxygenase; PS, pyransynthase; P450, cytochrome P450; THP, Tetrahydropyran.

Fig. 4. Proposed biosynthesis of different analogs from lasonolide A. We could not identify an enzyme which would explain the migration of the methyl from C-38 in LSA to C-36 in lasonolide B.

Fig. 5. Model for three repeats of the *las* BGC. The 5′ end of *las* BGC is highlighted to demonstrate the location where one of the *las* BGC repeats lacks *lasC*. Contig(s) making up the 98 kbp segment of *las* BGC (one in *las* BGC_v1 and six in *las* BGC_v2) have been collectively referred to as contig 1. Contigs except for the *las* BGC are not shown to scale.

Fig. 6. Variants identified between the three repeats of *las* BGC_v1.

Fig. 7. (A) 2D visualization of Automata binning of Forcepia_v1. The "*Ca.* T. lasonolidus" genome is circled in black and the *las* BGC contigs are marked with an arrow. Axes represent dimension-reduced Barnes-Hut Stochastic Neighbor Embedding (BH-tSNE) values (BH-tSNE x and BH-tSNE y). **(B)** 3D visualization of contigs present in "*Ca.* T. lasonolidus". The *las* BGC is colored red. Axes represent BH-tSNE values (BH-tSNE x and BH-tSNE y) along with k-mer

769 coverage. **(C)** GC percentage of different sets of genes in Forcepia_v1 "*Ca.* T. lasonolidus". The
770 *las* BGC genes are colored in red.

771 **Fig. 8.** Comparison of PV BMC gene cluster in Forcepia_v1 "*Ca.* T. lasonolidus" with the PV
772 BMC cluster from other Verrucomicrobia. "*Ca.* T. lasonolidus" has DNA-methyltransferase and
773 PVUII endonuclease genes (in gray, labeled 1 and 2) between the first and the second BMC-H
774 genes. This kind of arrangement was not observed in other PV BMC clusters.

775 **Fig. 9. (A and B)** Comparison of the gene length in **(A)** Forcepia_v1 "*Ca.* T. lasonolidus" and
776 **(B)** Forcepia_v2 "*Ca.* T. lasonolidus", respectively with their closest homologs in the nr
777 database. Genes with length less than 80% of the closest homolog (below the lower black line)
778 are classified as putative pseudogenes (57, 84, 85). The graphs have been truncated for clarity, as
779 some genes are many kbp long. **(C)** Comparison of functional COG categories in Forcepia_v1
780 "*Ca.* T. lasonolidus", Forcepia_v2 "*Ca.* T. lasonolidus" and *P. paruva* Ellin514 for non-
781 pseudogenes. A gene is considered to have a functional annotation when it belongs to a COG
782 category, except for category S which represents unknown function. **(D)** Comparison of genes in
783 different metabolic pathways for Forcepia_v1 "*Ca.* T. lasonolidus", Forcepia_v2 "*Ca.* T.
784 lasonolidus" and *P. paruva* Ellin514, including only non-pseudogenes. Colored squares represent
785 presence of a gene while white squares represent absence of gene. *K00940 is involved in both
786 purine and pyrimidine metabolism. Genes absent in all three genomes have been removed. **(E)**
787 Comparison of completeness of different metabolic pathways in Forcepia_v1 "*Ca.* T.
788 lasonolidus", Forcepia_v2 "*Ca.* T. lasonolidus" and *P. paruva* Ellin514 (including only non-
789 pseudogenes) as determined by KEGG decoder (86). Pathways have been grouped into
790 categories wherever possible. Pathways absent in all three genomes have been removed. V1 and
791 V2 refer to Forcepia_v1 "*Ca.* T. lasonolidus" and Forcepia_v2 "*Ca.* T. lasonolidus" respectively.

792 **Table 1.** Description of the variants identified between the three repeats of *las* BGC_v1. Both
793 AD and PL values are represented in the manner "reference, variant". A lower PL value
794 represents a higher likelihood of the sample being that genotype.

ID	Location (bp)	Change	Length (bp)	Allelic depth	PL
Insertion 1	Between 15,246 and 15,247	+GGAGGATGGGGTGGAGGATGGGGTGGAGGATGGGGTGGAGGATGGGGT	60	7, 15	1129, 0
Insertion 2	Between 24,776 and 24,777	+GGGGTCGGATGGGGGGTCGGATGG	24	8, 57	3092, 0
Insertion 3	Between 67,976 and 67,977	+GCGGCGGTTGAGGCGGAGGCGGCGGTTGAGGCGGAGGCGGCGGTTGAGGCGGAG	54	23, 164	5987, 0
SNP 1	93,995	G → A	1	363, 457	2973, 0
SNP 2	95,154	A → G	1	175, 621	16101, 0

795 **Table 2.** Genome statistics for "*Ca.* T. lasonolidus". *Coding density is weighted by length
796 taking into account the 97.11% coding density of *las* BGC repeats.

Characteristic	**Forcepia_v1 "Ca. T. lasonolidus"**	**Forcepia_v2 "Ca. T. lasonolidus"**
Size	4.85 Mbp	4.93 Mbp
Size after adding the three *las* repeats	5.05 Mbp	5.13 Mbp
checkM completeness	99.24%	99.32%
checkM contamination	1.35%	1.35%
No. of contigs	144	92
Longest contig	204,102 bp	649,894 bp
N50	52,980 bp	96,223
Average GC percentage	53.81%	53.88%
Percentage of pseudogenes	16.31% of total ORFs	16.62% of total ORFs
Transposase genes	6	15
Coding density*	79.45%	79.41%
Coding density without pseudogenes*	72.58%	72.38%
Eukaryotic like proteins		

Ankyrin repeats	3	3
Tetratricopeptide repeat	43 (9 were Sel-1 repeats)	42 (9 were Sel-1 repeats)
Pyrrolo-quinoline quinone	21	21
Leucine-rich repeat	16	16
WD40	4	5

797

798

799 # Legend for supplementary material

800 **Fig S1. (A)** Alignment of fosmids to the *las* BGC. Fosmids are depicted as arrows above the *las*
801 BGC. Fosmids captured before WGS are colored orange (3-46, 5-16, 6-17, 4-77 and 1-80),
802 whereas fosmids captured after WGS are colored blue (5-41, 2-18, and 2-13). **(B)** Relative
803 abundance of different phyla in the sequenced Forcepia_v2 metagenome. Each block shows the
804 relative abundance of each metagenome-assembled genome (MAG), with different colors
805 representing the phylum they belong to. The *las* BGC-carrying bin is highlighted is marked with
806 a star. **(C)** Assembly graph of *las* BGC_v1 visualized in BANDAGE (Wick RR, Schultz MB,
807 Zobel J, Holt KE, Bioinformatics 31:3350–3352, 2015,
808 https://doi.org/10.1093/bioinformatics/btv383). **(D)** Mapping of paired-end reads to contigs
809 making up *las* BGC_v1. Contigs in green boxes represent the *las* BGC, red boxes represent the 5′
810 end of *las* BGC and blue boxes represent the 3′ end of *las* BGC. **(E)** Assembly of the seven
811 contigs making up *las* BGC_v2. **(F)** Assembly graph of *las* BGC_v2 visualized in BANDAGE
812 (Wick RR, Schultz MB, Zobel J, Holt KE, Bioinformatics 31:3350–3352, 2015,

https://doi.org/10.1093/bioinformatics/btv383). **(G)** Mapping of paired-end reads to *las* BGC_v2. Contigs in green boxes represent the *las* BGC, red boxes represent the 5□ end of *las* BGC and blue boxes represent the 3□ end of *las* BGC. Panels D, E, G, and H were edited for clarity by removing contigs which had either very few paired-end read connections, were mapping to themselves or were very small. **(H)** BGC distribution in Forcepia_v2 sp. Metagenome. AntiSMASH (Blin K, Shaw S, Steinke K, Villebro R, Ziemert N, Lee SY, Medema MH, Weber T, Nucleic Acid Res 47:W81–W87, 2019, https://doi.org/10.1093/nar/gkz310) annotations of bacterial contigs greater than 3000bp are shown. Each bar indicates a MAG, grouped by phylum. The star represents the MAG containing the *las* BGC. BGC annotations have been simplified into polyketide synthase (PKS), Type 1 PKS, Type 3 PKS, *trans*-AT PKS, nonribosomal peptide synthetase (NRPS), ribosomally synthesized, post-translationally modified peptide (RiPP), hgIE-KS, hgIE-KS-T1PKS, terpenes, and others. **(I)** Phylogenetic tree of 51 different Verrucomicrobia genomes. Bootstrap values were calculated using RaxML with 1000 bootstrap replicates.

Fig S2. Clades from a phylogenetic tree of 944 KS domains from *trans*-AT PKS and the erythromycin BGC as an outgroup, containing KS domains from the *las* BGC. Color within the individual clades corresponds to the chemical structure shown on its right. ‘i’ and ‘c’ in *lasD* KS1 indicate the incomplete and complete KS domain in *lasD* respectively.

Fig S3. (A) Alignment of *las* AT and AH domains with AT and AH domains from different *trans*-AT PKS pathways. Active sites as well as sites distinguishing AT and AH domains (Jenner M, Afonso JP, Kohlhaas C, Karbaum P, Frank S, Piel J, Oldham NJ, Chem Commun 52:5262–5265, 2016, https://doi.org/10.1039/C6CC01453D) have been marked. **(B)** Phylogenetic tree of AT and AH domains. The different types of domain separate into different clades (Jenner M, Afonso JP, Kohlhaas C, Karbaum P, Frank S, Piel J, Oldham NJ, Chem Commun 52:5262–5265,

2016, https://doi.org/10.1039/C6CC01453D). **(C)** Phylogenetic tree of ECH1 and ECH2 domains. Both the domains separate into different clades (Slocum ST, Lowell AN, Tripathi A, Shende VV, Smith JL, Sherman DH, Methods Enzymol 604:207–236, 2018, https://doi.org/10.1016/bs.mie.2018.01.034). **(D-E)** Alignment of *las* **(D)** ECH1 and **(E)** ECH2 domains with respective ECH domains from other PKS pathways. Sequence that is required for the formation of the oxyanion hole which stabilizes the enolate anions is marked. LasH_a and LasO_a are proposed to be inactive as they are truncated and show poor homology to the rest of the ECH domains (Gu L, Jia J, Liu H, Håkansson K, Gerwick WH, Sherman DH, J Am Chem Soc 128:9014–9015, 2006, http://doi.org/10.1021/ja0626382, Matilla MA, Stöckmann H, Leeper FJ, Salmond GPC, J Biol Chem 287:39125–39138, 2012, https://doi.org/10.1074/jbc.M112.401026). **(F)** Alignment of *las* KS domains with active site (CHH) marked. 'i' and 'c' in the LasD KS indicate the incomplete and complete KS domain in different repeats of LasD respectively. LasD is a decarboxylating KS which are known to lack the active site cysteine (Walker PD, Weir ANM, Willis CL, Crump MP, Nat Prod Rep 38:723–756, 2021, https://doi.org/10.1039/D0NP00045K). **(G)** Alignment of *las* KR domains with two from the erythromycin BGC to allow comparison. Active site residues and conserved motifs are marked. The presence or absence of the second aspartate in the LDD motif is supposed to predict the stereochemistry of the hydroxyl group (Keatinge-Clay AT, Chem Biol 14:898–908, 2007, https://doi.org/10.1016/j.chembiol.2007.07.009, Caffrey P, ChemBioChem 4:654–657, 2003, https://doi.org/10.1002/cbic.200300581). Figures have been truncated for clarity and to show only the relevant sites. In phylogenetic trees, *las* BGC domains are highlighted in white.

Fig S4. **(A)** Phylogenetic tree of DH and PS domains, which separate into different clades (Wagner DT, Zhang Z, Meoded RA, Cepeda AJ, Piel J, Keatinge-Clay AT, ACS Chem Biol

13:975–983, 2018, http://doi.org/10.1021/acschembio.8b00049). *Las* BGC DH/PS domains are highlighted in white. **(B)** Alignment of PS domains identified in the *las* BGC with PS domains from other *trans*-AT PKS pathways. The DH domain from the erythromycin BGC is used for comparison. LasO DH2 and LasM DH4 are annotated as putative PS domains. Generally, PS domains have a Hx_4P motif instead of a Hx_8P and they lack the catalytic aspartate at the DxxxQ/H motif (Wagner DT, Zhang Z, Meoded RA, Cepeda AJ, Piel J, Keatinge-Clay AT, ACS Chem Biol 13:975–983, 2018, http://doi.org/10.1021/acschembio.8b00049, Pöplau P, Frank S, Morinaka BI, Piel J, Angew Chem Int Ed Engl 52:13215–13218, 2013, https://doi.org/10.1002/anie.201307406). This was found to be true only for LasM DH4 and not LasO DH2. However, identical variations from a traditional PS domain architecture are also seen in PS domains found in the mandelalide pathway (MndC DH3 and MndD DH3) (Lopera J, Miller IJ, McPhail KL, Kwan JC, mSystems 2:e00096–17, 2017, https://doi.org/10.1128/mSystems.00096-17). **(C)** Alignment of double bond-shifting DH domains identified in *las* BGC with similar domains found in other *trans*-AT PKS pathways. The DH domain from the erythromycin BGC is used for comparison. LasM DH1 and LasN DH1 are annotated as putative double bond-shifting DH domains. Generally, in DH shifting domains the conserved proline (P) in Hx_8P motif is often replaced by either valine (V) or leucine (L). In the case of LasM DH1, a methionine (M) instead of V or L appears in the place of P, which is in line with what is observed in difficidin biosynthesis as well (Chen X-H, Vater J, Piel J, Franke P, Scholz R, Schneider K, Koumoutsi A, Hitzeroth G, Grammel N, Strittmatter AW, Gottschalk G, Süssmuth RD, Borriss R, J Bacteriol 188:4024–4036, 2006, https://doi.org/10.1128/JB.00052-06). Furthermore, DH shifting domains are sometimes characterized by the replacement of the conserved aspartic acid (D) with asparagine (N) and substitution of glutamine (Q) or histidine

(H) with V or L in the DxxxQ/H motif. Even though LasN DH1 has an N in place of D in the DxxxQ/H motif, it substitutes Q/H with a serine (S). This is unusual and not found in any other double bond shifting DH. **(D)** Alignment of DH domains present in the *las* BGC with the DH domain from the erythromycin BGC. Putative PS and double bond-shifting DH domains have been excluded. LasH DH1 and LasM DH2 are annotated as inactive domains due to disrupted catalytic motifs Hx_8P and DxxxQ/H. Even though in LasL DH1 the catalytic aspartic acid is replaced by glutamic acid (DxxxQ/H motif), we propose it is active, as a similar mutation is observed in the palmerolide BGC (Avalon NE, Murray AE, Daligault HE, Lo C-C, Davenport KW, Dichosa AEK, Chain PSG, Baker BJ, Front Chem 9, 2021, https://doi.org/10.3389/fchem.2021.802574). **€** Alignment of LasL DH3 with other DH domains having a serine in place of proline in Hx_8P motif. The DH domain from the erythromycin BGC is used for comparison. Sequence headers in blue represent DH domains annotated as active while in red represent the ones annotated as inactive.

Fig S5. **(A)** 2D visualization of the initial Autometa binning of Forcepia_v2. The "*Ca.* T. lasonolidus" genome is circled in black and contigs making up the *las* BGC are marked with arrows. Axes represent dimension-reduced Barnes-Hut Stochastic Neighbor Embedding (BH-tSNE) values (BH-tSNE x and BH-tSNE y). **(B)** 3D visualization of contigs present in the "*Ca.* T. lasonolidus" genome. Contigs making up the *las* BGC are colored red. **(C)** GC percentage of different sets of genes in Forcepia_v2 "*Ca.* T. lasonolidus". *Las* BGC genes are colored in red. **(D)** and **(E)** Codon adaptation index (CAI) of different categories of genes present in Forcepia_v1 "*Ca.* T. lasonolidus" and Forcepia_v2 "*Ca.* T. lasonolidus", respectively. *Las* BGC genes are colored in red. P values for pairwise comparison between different categories of genes are shown in the matrix below their respective plots. Values with $p < 0.05$ are considered

905 significant. Other non-significant p values are colored red. Annotated and hypothetical genes
906 represent the genes annotated with a function and genes annotated as hypothetical respectively
907 by Prokka (Seemann T, Bioinformatics 30:2068–2069, 2014,
908 https://doi.org/10.1093/bioinformatics/btu153).

909 **Table S1.** List of oligonucleotide primers used for different purposes. **(A)** Primers used for
910 screening the *Forcepia* sp. fosmid library before WGS. (**B**) Primers used for screening the
911 *Forcepia* sp. fosmid library after WGS. **(C)** Primers used for confirming the presence of terminal
912 connections with the *las* BGC.

913 **Table S2.** Metadata and taxonomic classification of all the MAGs

914 **Table S3.** Contigs making up the three repeats of the *las* BGC in "*Ca*. T. lasonolidus" in
915 Forcepia_v1 and Forcepia_v2. Contig IDs represent the labels in **Fig. 5** of the main text.

916 **Table S4.** Gene annotation in Forcepia_v1 and Forcepia_v2 "*Ca*. T. lasonolidus". **(A)** Non-
917 pseudogenes annotated as Eukaryotic-like proteins. **(B)** Genes forming the PV BMC cluster. **(C)**
918 Non-pseudogenes annotated by dbCAN2

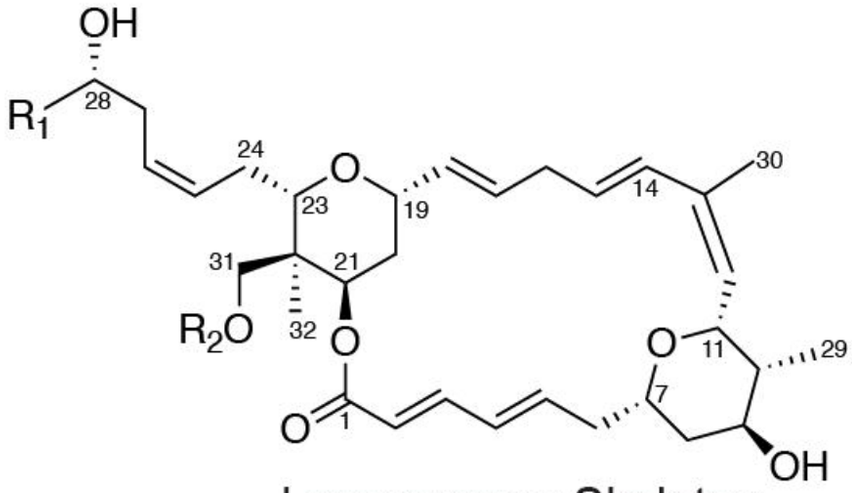
919 **Text S1.** Supplementary methods

920

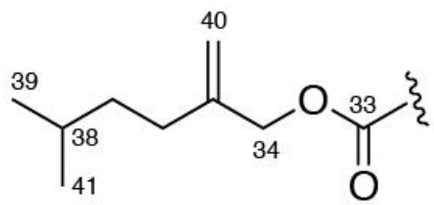
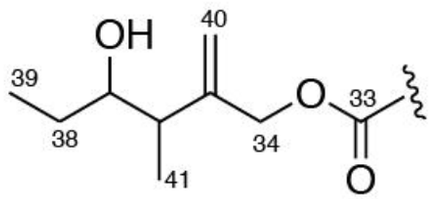
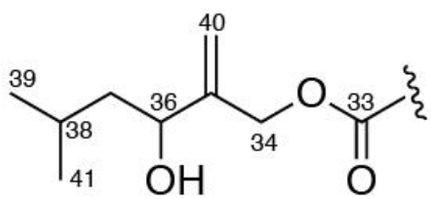
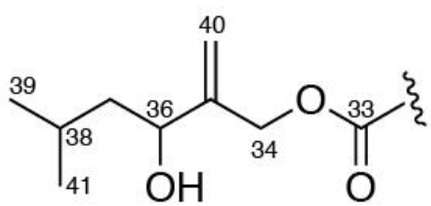
A

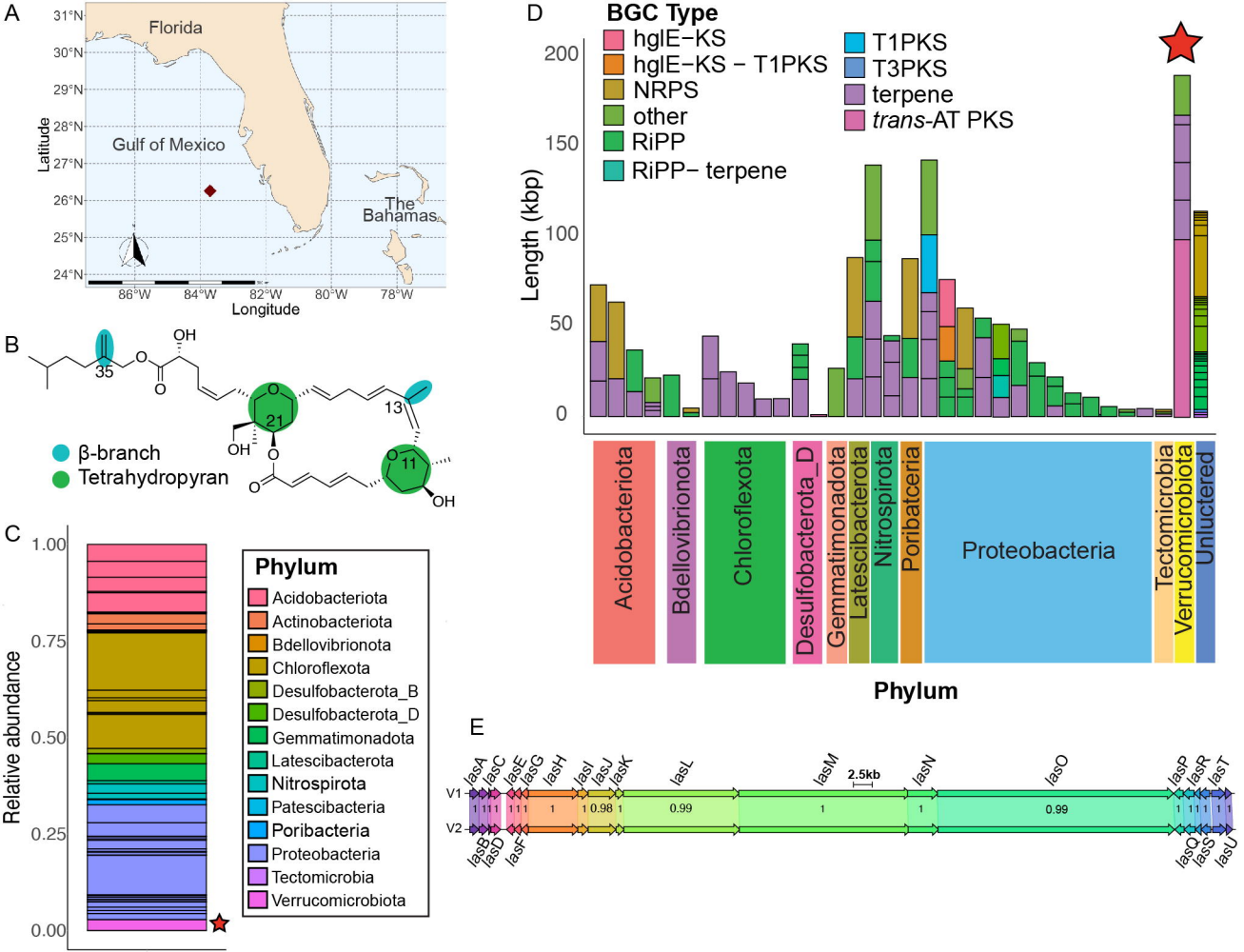
B

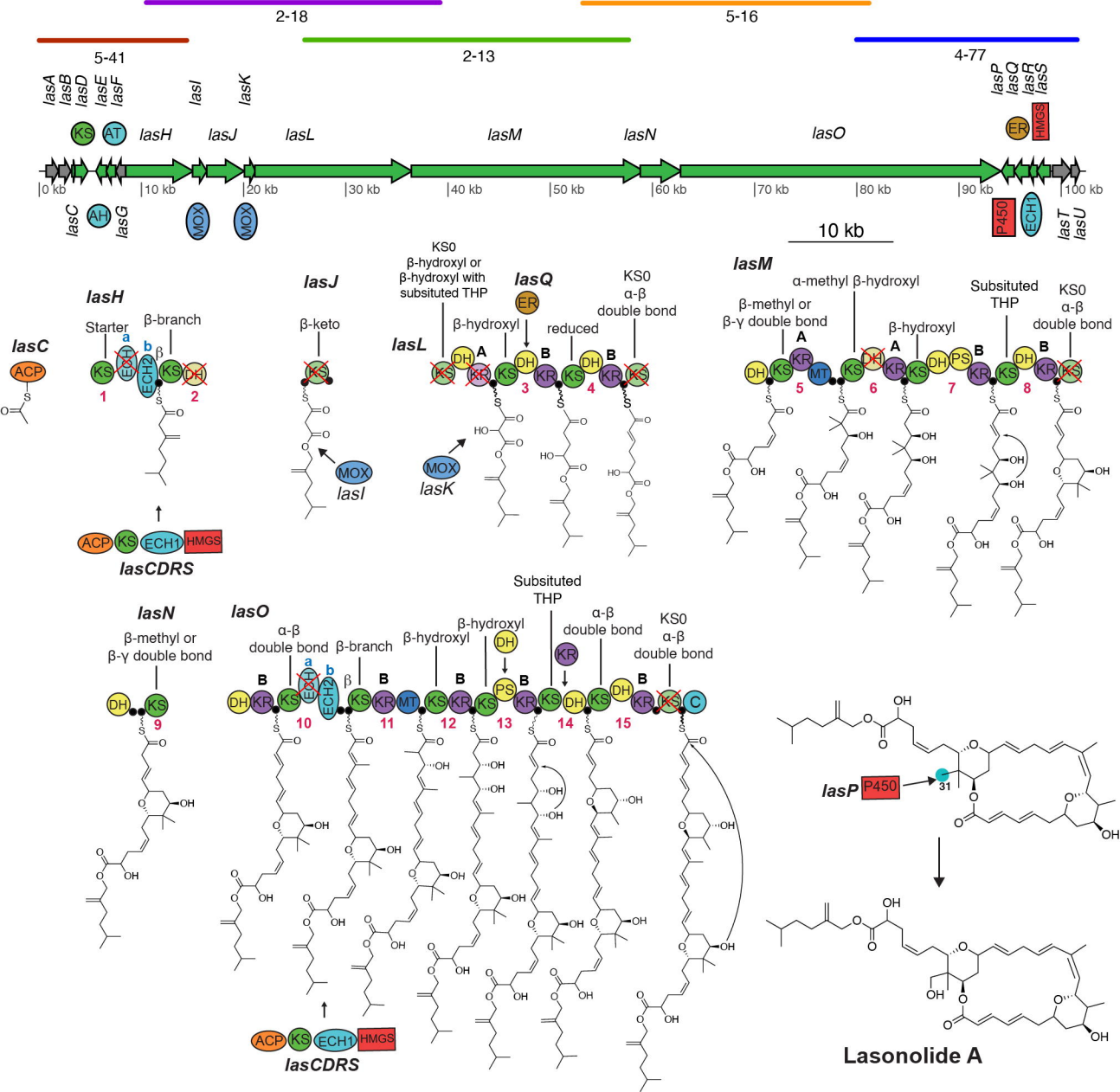
C

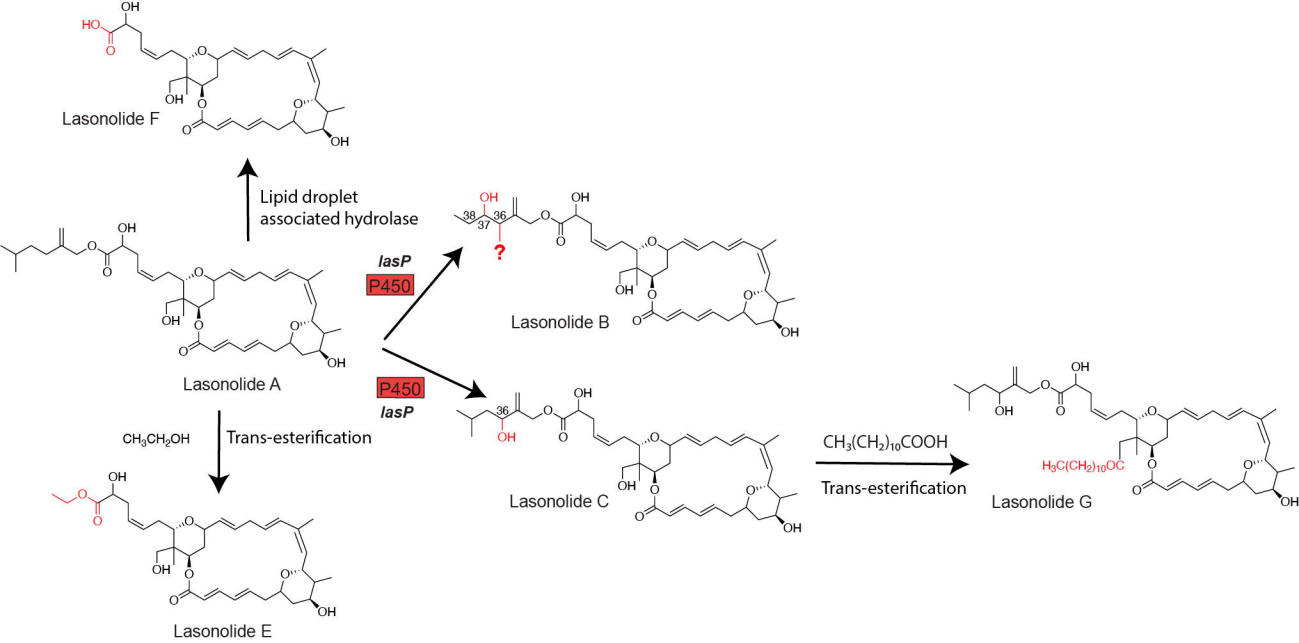


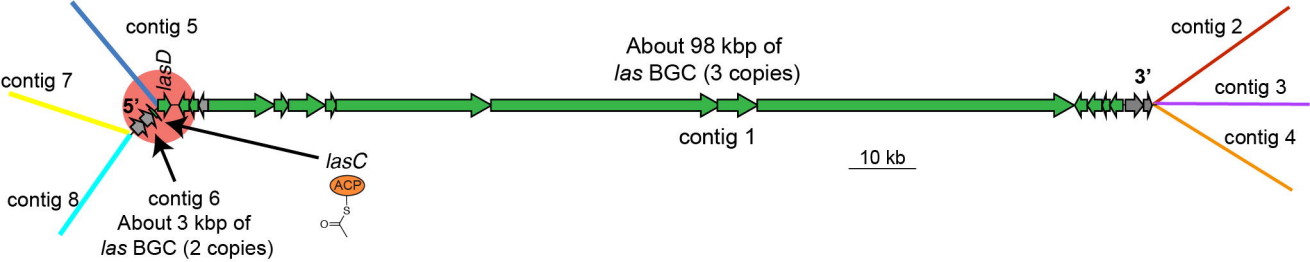
Lasonopyran Skeleton

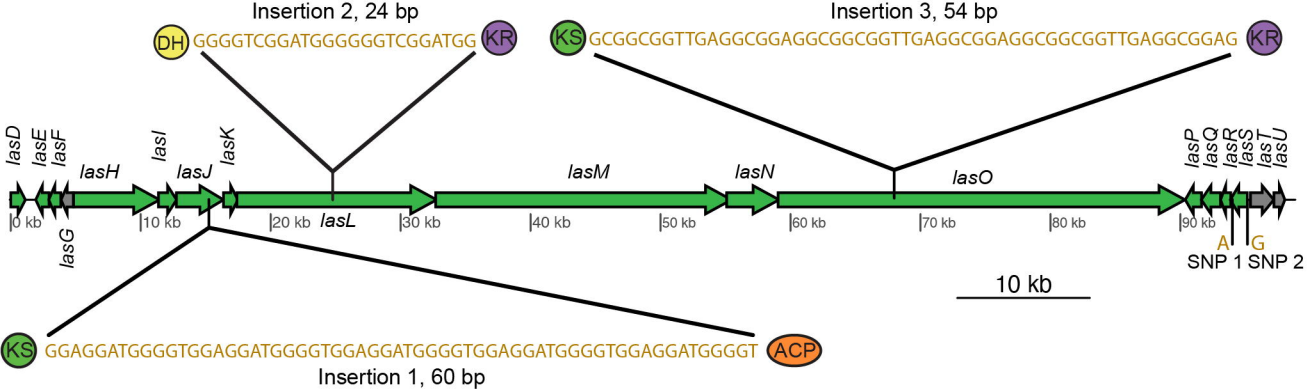
R_1	R_2	Analog Name
	H	Lasonolide A
	H	Lasonolide B
	H	Lasonolide C
H	H	Lasonolide D
CH_3CH_2OOC	H	Lasonolide E
HOOC	H	Lasonolide F
	$CH_3(CH_2)_{10}CO$ (53–42)	Lasonolide G

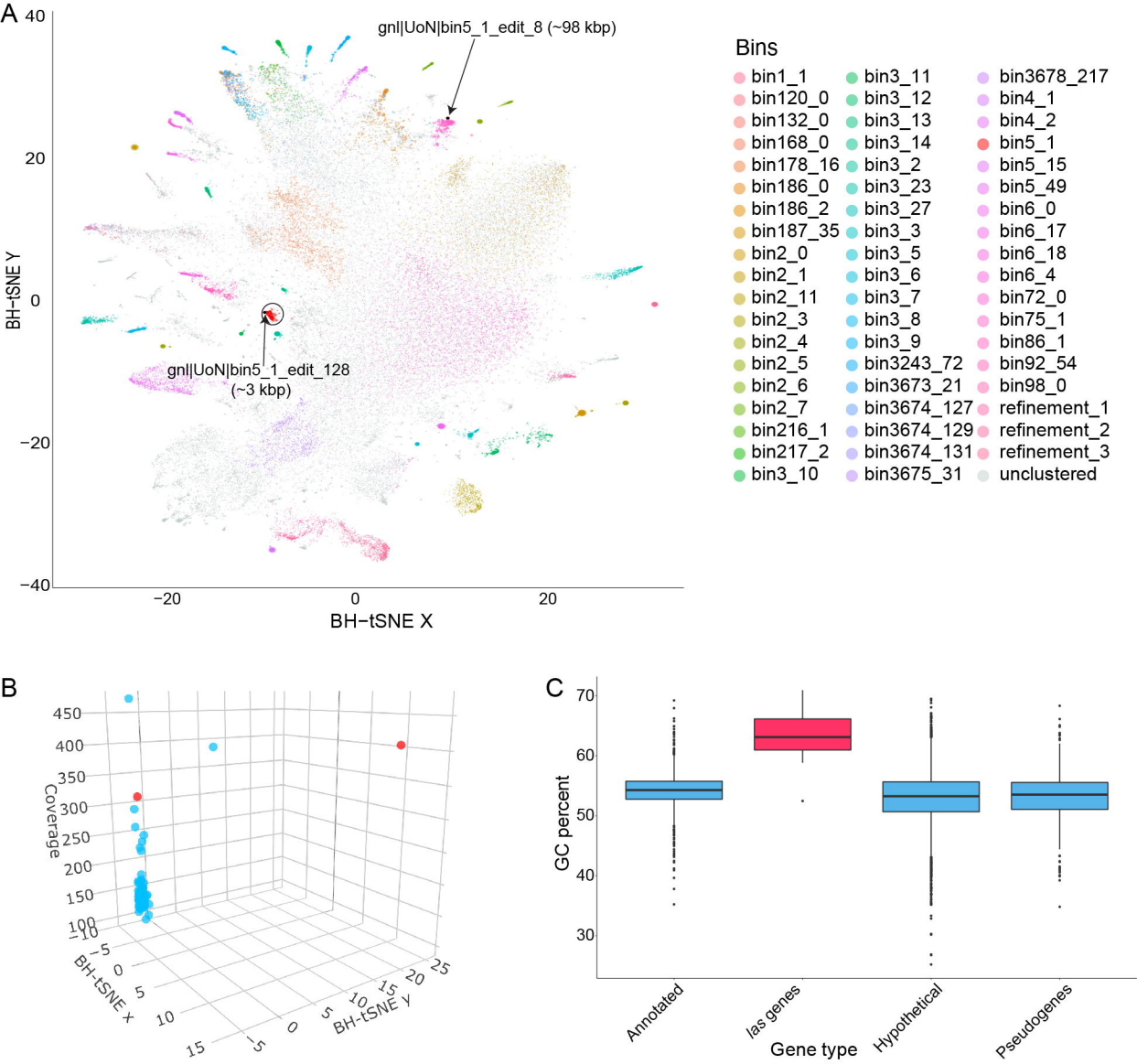

A
Florida
Gulf of Mexico
The Bahamas
Latitude
Longitude
31°N
30°N
29°N
28°N
27°N
26°N
25°N
24°N
86°W
84°W
82°W
80°W
78°W
B
OH
35
O
21
13
11
β-branch
Tetrahydropyran
C
Relative abundance
1.00
0.75
0.50
0.25
0.00
Phylum
Acidobacteriota
Actinobacteriota
Bdellovibrionota
Chloroflexota
Desulfobacterota_B
Desulfobacterota_D
Gemmatimonadota
Latescibacterota
Nitrospirota
Patescibacteria
Poribacteria
Proteobacteria
Tectomicrobia
Verrucomicrobiota
D
BGC Type
hglE−KS
hglE−KS − T1PKS
NRPS
other
RiPP
RiPP− terpene
T1PKS
T3PKS
terpene
trans-AT PKS
Length (kbp)
200
150
100
50
0
Acidobacteriota
Bdellovibrionota
Chloroflexota
Desulfobacterota_D
Gemmatimonadota
Latescibacterota
Nitrospirota
Poribacteria
Proteobacteria
Tectomicrobia
Verrucomicrobiota
Unluctered
Phylum
E
lasA
lasC
lasE
lasG
lasH
lasI
lasJ
lasK
lasL
lasM
2.5kb
lasN
lasO
lasP
lasR
lasT
V1
V2
lasB
lasD
lasF
lasQ
lasS
lasU

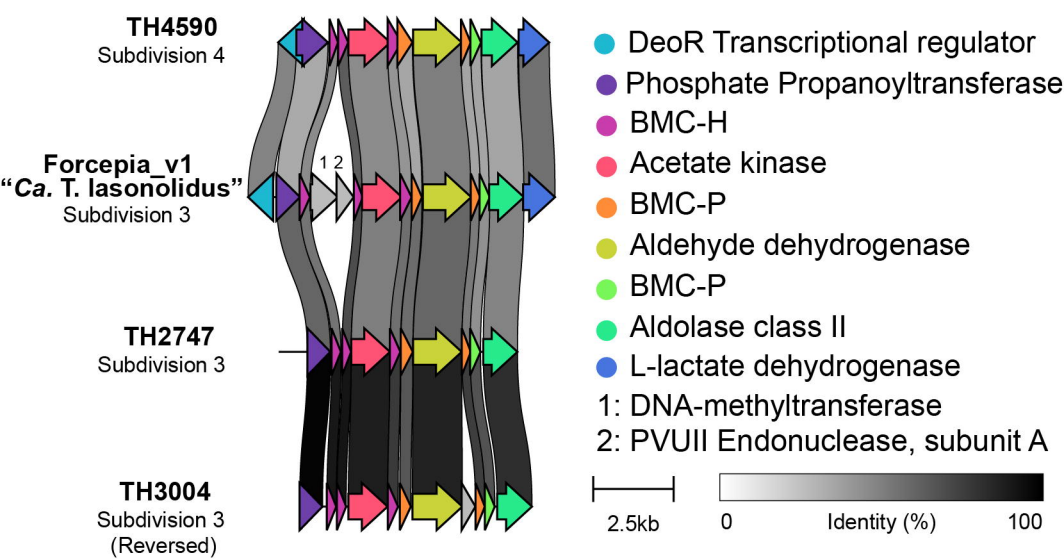

2-18
5-16
5-41
2-13
4-77
lasA lasB lasD lasE lasF
lasI
lasK
lasP lasQ lasR lasS
KS AT lasH lasJ lasL lasM lasN lasO ER HMGS
0 kb 10 kb 20 kb 30 kb 40 kb 50 kb 60 kb 70 kb 80 kb 90 kb 100 kb
lasC AH lasG
MOX MOX
P450 ECH1 lasT lasU
10 kb
lasH
Starter
β-branch
lasC
ACP
ACP KS ECH1 HMGS
lasCDRS
lasJ
β-keto
MOX
lasI
lasL
KS0
β-hydroxyl or
β-hydroxyl with
subsituted THP
lasQ
ER
β-hydroxyl
reduced
KS0
α-β
double bond
MOX
lasK
lasM
α-methyl β-hydroxyl
β-methyl or
β-γ double bond
Subsituted
THP
KS0
α-β
double bond
lasN
β-methyl or
β-γ double bond
lasO
α-β
double bond
β-branch
β-hydroxyl
Subsituted
THP
α-β
double bond
KS0
α-β
double bond
ACP KS ECH1 HMGS
lasCDRS
lasP P450
Lasonolide A

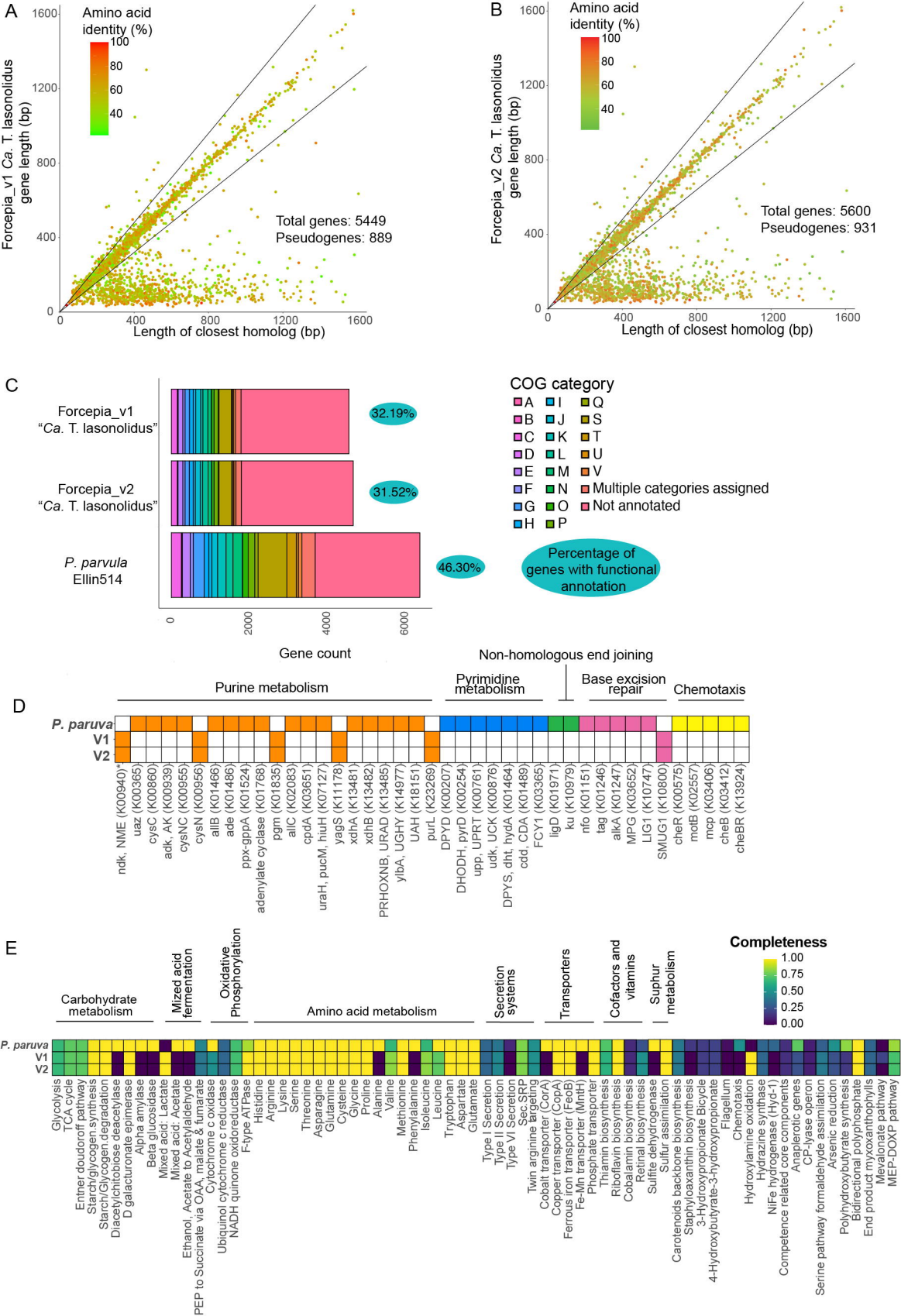

Lasonolide F
Lipid droplet
associated hydrolase
Lasonolide A
lasP
P450
Lasonolide B
P450
lasP
Lasonolide C
CH3CH2OH
Trans-esterification
Lasonolide E
CH3(CH2)10COOH
Trans-esterification
Lasonolide G


contig 5
lasD
contig 7
5'
contig 8
contig 6
About 3 kbp of
las BGC (2 copies)
lasC
ACP
About 98 kbp of
las BGC (3 copies)
contig 1
10 kb
3'
contig 2
contig 3
contig 4


Insertion 2, 24 bp
DH
GGGGTCGGATGGGGGGTCGGATGG
KR
Insertion 3, 54 bp
KS
GCGGCGGTTGAGGCGGAGGCGGCGGTTGAGGCGGAGGCGGCGGTTGAGGCGGAG
KR
lasD
lasE
lasF
lasG
lasH
lasI
lasJ
lasK
lasL
lasM
lasN
lasO
lasP
lasQ
lasR
lasS
lasT
lasU
0 kb
10 kb
20 kb
30 kb
40 kb
50 kb
60 kb
70 kb
80 kb
90 kb
A
G
SNP 1
SNP 2
10 kb
KS
GGAGGATGGGGTGGAGGATGGGGTGGAGGATGGGGTGGAGGATGGGGTGGAGGATGGGGT
ACP
Insertion 1, 60 bp


A
gnl|UoN|bin5_1_edit_8 (~98 kbp)
gnl|UoN|bin5_1_edit_128
(~3 kbp)
BH−tSNE Y
BH−tSNE X
Bins
bin1_1
bin120_0
bin132_0
bin168_0
bin178_16
bin186_0
bin186_2
bin187_35
bin2_0
bin2_1
bin2_11
bin2_3
bin2_4
bin2_5
bin2_6
bin2_7
bin216_1
bin217_2
bin3_10
bin3_11
bin3_12
bin3_13
bin3_14
bin3_2
bin3_23
bin3_27
bin3_3
bin3_5
bin3_6
bin3_7
bin3_8
bin3_9
bin3243_72
bin3673_21
bin3674_127
bin3674_129
bin3674_131
bin3675_31
bin3678_217
bin4_1
bin4_2
bin5_1
bin5_15
bin5_49
bin6_0
bin6_17
bin6_18
bin6_4
bin72_0
bin75_1
bin86_1
bin92_54
bin98_0
refinement_1
refinement_2
refinement_3
unclustered
B
Coverage
BH-tSNE x
BH-tSNE y
C
GC percent
Annotated
las genes
Hypothetical
Pseudogenes
Gene type


TH4590
Subdivision 4
Forcepia_v1
"Ca. T. lasonolidus"
Subdivision 3
TH2747
Subdivision 3
TH3004
Subdivision 3
(Reversed)
1 2
DeoR Transcriptional regulator
Phosphate Propanoyltransferase
BMC-H
Acetate kinase
BMC-P
Aldehyde dehydrogenase
BMC-P
Aldolase class II
L-lactate dehydrogenase
1: DNA-methyltransferase
2: PVUII Endonuclease, subunit A
2.5kb
0
Identity (%)
100


A
Amino acid identity (%)
100
80
60
40
Forcepia_v1 Ca. T. lasonolidus gene length (bp)
Length of closest homolog (bp)
Total genes: 5449
Pseudogenes: 889
B
Amino acid identity (%)
Forcepia_v2 Ca. T. lasonolidus gene length (bp)
Length of closest homolog (bp)
Total genes: 5600
Pseudogenes: 931
C
Forcepia_v1 "Ca. T. lasonolidus"
Forcepia_v2 "Ca. T. lasonolidus"
P. parvula Ellin514
32.19%
31.52%
46.30%
Gene count
COG category
A I Q
B J S
C K T
D L U
E M V
F N Multiple categories assigned
G O Not annotated
H P
Percentage of genes with functional annotation
Non-homologous end joining
Purine metabolism
Pyrimidine metabolism
Base excision repair
Chemotaxis
D
P. paruva
V1
V2
ndk, NME (K00940)*
uaz (K00365)
cysC (K00860)
adk, AK (K00939)
cysNC (K00955)
cysN (K00956)
allB (K01466)
ade (K01486)
ppx-gppA (K01524)
adenylate cyclase (K01768)
pgm (K01835)
allC (K02083)
cpdA (K03651)
uraH, pucM, hiuH (K07127)
yagS (K11178)
xdhA (K13481)
xdhB (K13482)
PRHOXNB, URAD (K13485)
ylbA, UGHY (K14977)
UAH (K18151)
purL (K23269)
DPYD (K00207)
DHODH, pyrD (K00254)
upp, UPRT (K00761)
udk, UCK (K00876)
DPYS, dht, hydA (K01464)
cdd, CDA (K01489)
FCY1 (K03365)
ligD (K01971)
ku (K10979)
nfo (K01151)
tag (K01246)
alkA (K01247)
MPG (K03652)
LIG1 (K10747)
SMUG1 (K10800)
cheR (K00575)
motB (K02557)
mcp (K03406)
cheB (K03412)
cheBR (K13924)
E
Completeness
1.00
0.75
0.50
0.25
0.00
Carbohydrate metabolism
Mixed acid fermentation
Oxidative Phosphorylation
Amino acid metabolism
Secretion systems
Transporters
Cofactors and vitamins
Suphur metabolism
P. paruva
V1
V2
Glycolysis
TCA cycle
Entner doudoroff pathway
Starch/glycogen.synthesis
Starch/Glycogen.degradation
Diacetylchitobiose deacetylase
D galacturonate epimerase
Alpha amylase
Beta glucosidase
Mixed acid: Lactate
Mixed acid: Acetate
Ethanol, Acetate to Acetylaldehyde
PEP to Succinate via OAA, malate & fumarate
Cytochrome c oxidase
Ubiquinol cytochrome c reductase
NADH quinone oxidoreductase
F-type ATPase
Histidine
Arginine
Lysine
Serine
Threonine
Asparagine
Glutamine
Cysteine
Glycine
Proline
Alanine
Valine
Methionine
Phenylalanine
Isoleucine
Leucine
Tryptophan
Aspartate
Glutamate
Type I Secretion
Type II Secretion
Type VI Secretion
Sec.SRP
Twin arginine targeting
Cobalt transporter (CorA)
Copper transporter (CopA)
Ferrous iron transporter (FeoB)
Fe-Mn transporter (MntH)
Phosphate transporter
Thiamin biosynthesis
Riboflavin biosynthesis
Cobalamin biosynthesis
Retinal biosynthesis
Sulfite dehydrogenase
Sulfur assimilation
Carotenoids backbone biosynthesis
Staphyloaxanthin biosynthesis
3-Hydroxypropionate Bicycle
4-Hydroxybutyrate-3-hydroxypropionate
Flagellum
Chemotaxis
Hydroxylamine oxidation
Hydrazine synthase
NiFe hydrogenase (Hyd-1)
Competence related core components
Anaplerotic genes
CP-lyase operon
Serine pathway formaldehyde assimilation
Arsenic reduction
Polyhydroxybutyrate synthesis
Bidirectional polyphosphate
End product myxoxanthophylls
Mevalonate pathway
MEP-DOXP pathway