Discovery of an Antarctic ascidian-associated uncultivated Verrucomicrobia that encodes antimelanoma palmerolide biosynthetic capacity

Alison E. Murray[a*], Chien-Chi Lo[b], Hajnalka E. Daligault[b], Nicole E. Avalon[c], Robert W. Read[a], Karen W. Davenport[b], Mary L. Higham[a], Yuliya Kunde[b], Armand E.K. Dichosa[b], Bill J. Baker[c*], Patrick S.G. Chain[b*]

[a] Division of Earth and Ecosystem Science, Desert Research Institute, Reno, Nevada, USA.

[b] Bioscience Division, Los Alamos National Laboratory, Los Alamos, New Mexico, USA.

[c] Department of Chemistry, University of South Florida, Tampa, Florida, USA.

Corresponding authors:

Alison E. Murray, Division of Earth and Ecosystem Sciences, Desert Research Institute, Reno, NV 89512, USA. +1 775-673-7361. alison.murray@dri.edu.

Bill J. Baker, Department of Chemistry, University of South Florida, Tampa, FL 33620, USA. +1-813-974-1967, bjbaker@usf.edu

Patrick S.G. Chain, Bioscience Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA. +1-505-665-4019. pchain@lanl.gov

Author Contributions: The team that came together to conduct this research included A.E.M. B.J.B. and P.S.G.C. who designed this research. A.E.M., C-C.L., H.E.D., N.E.A., R.W.R., K.W.D., M.L.H., Y.K., A.E.K.D. performed the research. P.S.G.C., C-C.L. and Y.K. contributed analytic tools and reagents. A.E.M., C-C.L., H.E.D., N.E.A., and K.W.D., analyzed the data and A.E.M., C-C.L., H.E.D., N.E.A., A.E.K.D., B.J.B. and P.S.G.C. wrote the paper.

Competing Interest Statement: The authors have no competing interests

Classification: Biological sciences-environmental sciences

Keywords: Antarctic, palmerolide, anticancer, microbiome metagenome, *Synoicum adareanum*

This PDF file includes:

Main Text
Figures 1 to 4
Tables 1

Abstract

The Antarctic marine ecosystem harbors a wealth of biological and chemical innovation that has risen in concert over millennia since the isolation of the continent and formation of the Antarctic circumpolar current. Scientific inquiry into the novelty of marine natural products produced by Antarctic benthic invertebrates led to the discovery of a bioactive macrolide, palmerolide A, that has specific activity against melanoma and holds considerable promise as an anticancer therapeutic. While this compound was isolated from the Antarctic ascidian *Synoicum adareanum*, its biosynthesis has since been hypothesized to be microbially mediated, given structural similarities to microbially-produced hybrid non-ribosomal peptide-polyketide macrolides. Here, we describe a metagenome-enabled investigation aimed at identifying the biosynthetic gene cluster (BGC) and palmerolide A-producing organism. A 74kb candidate BGC encoding the multimodular enzymatic machinery (hybrid Type I-*trans*-AT polyketide synthase-non-ribosomal peptide synthetase and tailoring functional domains) was identified and found to harbor key features predicted as necessary for palmerolide A biosynthesis. Surveys of ascidian microbiome samples targeting the candidate BGC revealed a high correlation between palmerolide-gene targets and a single 16S rRNA gene variant (R=0.83 – 0.99). Through repeated rounds of metagenome sequencing followed by binning contigs into metagenome-assembled genomes, we were able to retrieve a near-complete genome (10 contigs) of the BGC organism, a novel verrucomicrobium within the *Opitutaceae* family that we propose here as *Candidatus* Synoicihabitans palmerolidicus. The refined genome assembly harbors five highly similar BGC copies, along with structural and functional features that shed light on the host-associated nature of this unique bacterium.

Significance Statement

Palmerolide A has potential as a chemotherapeutic agent to target melanoma. We interrogated the microbiome of the Antarctic ascidian, *Synoicum adarenum,* using a cultivation-independent high-throughput sequencing and bioinformatic strategy. The metagenome-encoded biosynthetic machinery predicted to produce palmerolide A was found to be associated with the genome of a member of the *S. adareanum* core microbiome. Phylogenomic analysis suggests the organism represents a new deeply-branching genus, *Candidatus* Synoicihabitans palmerolidicus, in the *Opitutaceae* family of the Verrucomicrobia phylum. The *Ca*. S. palmerolidicus 4.29 Mb genome encodes a repertoire of carbohydrate-utilizing and transport pathways enabling its ascidian-associated lifestyle. The palmerolide-producer's genome also contains five distinct copies of the large palmerolide biosynthetic gene cluster that may provide structural complexity of palmerolide variants.

Main Text

Introduction

Across the world's oceans, marine benthic invertebrates harbor a rich source of natural products that serve metabolic and ecological roles in situ. These compounds provide a multitude of medicinal and biotechnological applications to science, health and industry. The organisms responsible for their biosynthesis are often not clear (1, 2). Increasingly, the products, especially in the polyketide class (*trans*-AT in particular), are found to be produced by microbial counterparts associated with the invertebrate host (3-5). Invertebrates including sponges, corals, and ascidians for example, are increasingly being recognized to harbor a wealth of diverse microbes, few of which have been cultivated (e.g., 6-8). Genomic tools, in particular, are revealing biochemical pathways potentially critical in the host-microbe associations (9). Microbes that form persistent mutualistic (symbiotic) associations provide key roles in host ecology, such as provision of

metabolic requirements, production of adaptive features such as photoprotective pigments, bioluminescence, or antifoulants, and biosynthesis of chemical defense agents.

Antarctic marine ecosystems harbor species-rich macrobenthic communities (10-12), which have been the subject of natural products investigations over the past 30 years resulting in the identification of > 600 metabolites (13). Initially, it was not known whether the same selective pressures (namely predation and competition, e.g. (14)) that operate in mid and low latitudes would drive benthic organisms at the poles to create novel chemistry (15). However, this does appear to be the case, and novel natural products have been discovered across algae, sponges, corals, nudibranchs, echinoderms, bryozoans, ascidians, and increasingly amongst microorganisms (16) for which the ecological roles have been deduced in a number of cases (13, 17). Studies of Antarctic benthic invertebrate-microbe associations however, pale in comparison to studies at lower latitudes, yet the few studies that have been reported suggest these associations (i) harbor an untapped reservoir of biological diversity (18-21) including fungi (22), (ii) are host species-specific (23, 24), (iii) provide the host with sources of nitrogen and fixed carbon (25) and (iv) have biosynthetic functional potential (26, 27).

This study was specifically motivated by our desire to understand the biosynthetic origins of a natural product, palmerolide A, given its potent anticancer activity (28), that is found to be associated with the polyclinid Antarctic ascidian, *Synoicum adareanum* (Fig. 1 a and b). Ascidians are known to be rich sources of bioactive natural products (9). They have been found to harbor polyketide, terpenoid, peptide, alkaloid and a few other classes of natural products, of which the majority have cytotoxic and/or antimicrobial activities. In addition to palmerolide A, a few other natural products derived from Antarctic ascidians have been reported (29-31). Ascidian-associated microbes responsible for natural product biosynthesis have been shown to be affiliated with bacterial phyla including Actinobacteria (which dominates the recognized diversity), Cyanobacteria, Firmicutes, Proteobacteria (both Alphaproteobacteria and Gammaproteobacteria) and Verrucomicrobia in addition to many fungi (32, 33). Metagenome-enabled studies have been key in linking natural products to the organisms producing them in a number of cases, e.g. patellamide A and C to Cyanobacteria-affiliated *Prochloron* spp. (4), the tetrahydroisoquinoline alkaloid ET-743 to Gammaproteobacteria-affiliated *Candidatus* Endoecteinascidia frumentensis (34), patellazoles to Alphaproteobacteria-affiliated *Candidatus* Endolissoclinum faulkneri (35), and mandelalides to Verrucomicrobia-affiliated *Candidatus* Didemnitutus mandela (36). However, this is most certainly an under-representation of the diversity of ascidian-associated microorganisms with capabilities for synthesizing bioactive compounds, given the breadth of ascidian biodiversity (37). These linkages have been yet to be investigated for Antarctic ascidians.

Palmerolide A has anticancer properties with selective activity against melanoma when tested in the National Cancer Institute 60 cell-line panel (28). This result is of particular interest, as there are few natural product therapeutics for this devastating form of cancer. Palmerolide A inhibits vacuolar ATPases, which are highly expressed in metastatic melanoma. Given the current level of understanding that macrolides often have microbial biosynthetic origins, that the holobiont metagenome has biosynthetic potential (26) and a diverse, yet persistent core microbiome is found in palmerolide A-containing *S. adareanum* (27) we have hypothesized that a microbe associated with *S. adareanum* is responsible for the biosynthesis of palmerolide A.

The core microbiome of the palmerolide A-producing ascidian *S. adareanum* in samples collected across the Anvers Island archipelago (n=63 samples; (27)) is comprised of five bacterial phyla including Proteobacteria (dominating the microbiome), Bacteroidetes, Nitrospirae, Actinobacteria and Verrucomicrobia. A few candidate taxa in particular, were suggested to be likely palmerolide A producers based on relative abundance and biosynthetic potential determined by analysis of lineage-targeted biosynthetic capability (genera: *Microbulbifer*, *Pseudovibrio*, *Hoeflea*, and the family *Opitutaceae* (27). This motivated interrogation of the *S. adarenum* microbiome metagenome, with the goals of determining the metagenome-encoded

biosynthetic potential, identifying candidate palmerolide A biosynthetic gene cluster (BGC)(s) and establishing the identity of the palmerolide A producing organism.

Results and Discussion

Identification of a putative palmerolide A biosynthetic gene cluster. Microbial-enriched fractions of *S. adareanum* metagenomic DNA sequence from 454 and Ion Proton next generation sequencing (NGS) libraries (almost 18 billion bases in all) were assembled independently, then merged, resulting in ~ 145 MB of assembled bases distributed over 86,387 contigs (referred to as CoAssembly 1; *SI Appendix*, Table S1). As the metagenome sequencing effort was focused on identifying potential BGCs encoding the machinery to synthesize palmerolide A, the initial steps of analysis specifically targeted those contigs in the assembly that were > 40 kb, as the size of the macrolide ring with 24 carbons would require a large number of polyketide modules to be encoded. This large fragment subset of CoAssembly 1 was submitted to antiSMASH v.3 (38) and more recently to v.5 (39). The results indicated a heterogeneous suite of BGCs, including a bacteriocin, two non-ribosomal peptide synthetases (NRPS), two hybrid NRPS -Type I PKS, two terpenes, and three hybrid *trans*-AT-PKS hybrid NRPS clusters (*SI Appendix,* Table S2).

We predicted several functional characteristics of the BGC that would be required for palmerolide A biosynthesis which aided our analysis (see (40) for details). This included evidence of a hybrid nonribosomal peptide-polyketide pathway and enzymatic domains leading to placement of two distinct structural features of the polyketide backbone, a carbamoyl transferase that appends a carbamate group at C-11 of the macrolide ring, and an HMGCoA synthetase that inserts a methyl group on an acetate C-1 position of the macrolide structure (C-25). The antiSMASH results indicated that two of the three predicted hybrid NRPS *trans*-AT-Type I PKS contained the predicted markers. Manual alignment of these two contigs suggested near-identical overlapping sequence (36,638 bases) and, when joined, the merged contig resulted in a 74,672 Kb BGC (Fig. 1c). The cluster size was in the range of other large *trans*-AT PKS encoding BGCs including pederin (54 Kb; 41), leinamycin (135.6 Kb; 42) as well as a *cis*-acting AT-PKS, jamaicamide (64.9 Kb; 43). The combined contigs encompassed what appeared to be a complete BGC that was flanked at the start with a transposase and otherwise unlinked in the assembly to other contiguous DNA. The cluster lacked phylogenetically informative marker genes from which putative taxonomic assignment could be attributed.

The antiSMASH results suggested that the BGC appears to be novel with the highest degree of relatedness to pyxipyrrolone A and B (encoded in the *Pyxidicoccus* sp. MCy9557 genome (44), to which only 14% of the genes have a significant BLAST hit to genes in the metagenome-encoded cluster. The ketosynthase (KS) sequences (13 in all) fell into three different sequence groups (40). One was nearly identical (99% amino acid identity) to a previously reported sequence from a targeted KS study of *S. adareanum* microbiome metagenomic DNA (26). The other two were most homologous to KS sequences from *Allochromatium humboldtianum,* and *Dickeya dianthicola* in addition to a number of hypothetical proteins from environmental sequence data sets.

Taxonomic inference of palmerolide A BGC. Taxonomic attribution of the BGC was inferred using a real time PCR strategy targeting three coding regions of the putative palmerolide A BGC spanning the length of the cluster (acyltransferase, AT1; hydroxymethylglutaryl Co-A synthase, HCS, and the condensation domain of the non-ribosomal peptide synthase NRPS, Fig. 1c) to assay a *Synoicum* microbiome collection of 63 samples that have been taxonomically classified using Illumina SSU rRNA gene tag sequencing (27). The three gene targets were present in all samples ranging within and between sites at levels from ~7 x 10^1 – 8 x 10^5 copies per gram of host tissue (Fig. 2a). The three BGC gene targets co-varied across all samples (r^2 > 0.7 for all pairs), with the NRPS gene copy levels slightly lower overall (mean: 0.66 and 0.59 copies per ng host tissue for NRPS:AT1 and NRPS:HCS respectively, n=63). We investigated the relationship between BGC gene copies per ng host tissue for each sample and palmerolide A levels determined for the same samples using mass spectrometry, however no correlation was found (R<0.03, n=63; (27)). We

then assessed the semi-quantitative relationship between the occurrence of SSU rRNA amplicon sequence variants (ASV, n=461; (27)) and the abundance of the three palmerolide A BGC gene targets. Here, we found a robust correlation (R=0.83 – 0.99) between all 3 gene targets and a single amplicon sequence variant (ASV15) in the core microbiome (*SI Appendix*, Fig. S1). This ASV is affiliated with the *Opitutaceae* family of the Verrucomicrobium phylum. The *Opitutaceae* family ASV (SaM_ASV15) was a member of the core microbiome as it was detected in 59 of the 63 samples surveyed at varying levels of relative abundance and displayed strong correlations with the abundances of BGC gene targets (Fig. 2b, r^2 =0.68 with AT1, 0.97 with HCS, and 0.69 with NRPS, n=63 for all). The only other correlations R > 0.5 were ASVs associated within the "variable" fraction of the microbiome, e.g., one low abundance ASV was present in 24 of 63 samples (SI Appendix, Fig. S1).

This result supports the finding of Murray et al. (27) in which gene abundance and natural product chemistry do not reflect a 1:1 ratio in this host-associated system. Neither the semi-quantitative measure of ASV copies nor the real time PCR abundance estimates of the three biosynthetic gene targets correlated with the mass-normalized levels of palmerolide A present in the same samples. As discussed (27), this is likely a result of bioaccumulation in the ascidian tissues. This result provided strong support that the genetic capacity for palmerolide A production was associated with a novel member of the *Opitutaceae*, a taxonomic family with representatives found across diverse host-associated and free-living ecosystems. Although the biosynthetic capacity of this family is not well known (45), recent evidence (36) suggests this family may be a fruitful target for cultivation efforts and natural product surveys.

Assembly of the palmerolide BGC-associated *Opitutaceae*-related metagenome assembled genome (MAG). With metagenomes some genomes come together easily – while others present compelling puzzles to solve. Assembly of the pal BGC-containing *Opitutaceae* genome was the result of a dedicated effort of binning contigs, gene searches, additional sequencing of samples with high BGC titer, and manual, targeted assembly. Binning efforts with CoAssembly 1 did not result in association of the *pal* BGC with an associated metagenome assembled genome (*SI Appendix* text). Therefore, a further round of metagenome sequencing using long read technology (Pacific Biosciences Sequel Systems technology; PacBio) ensued.

The 16S rRNA gene ASV occurrence (27) and real time PCR data were used to guide *S. adarenum* sample selection for sequencing. Two ascidian samples (Bon-1C-2011 and Del-2B-2011) with high *Opitutaceae* ASV occurrences (ASV_015; > 1000 sequences each – relative abundance of ~ 13.3-15.3 % compared to an overall average of 1.3 ± 2.77% across the 63 samples respectively, (27)) and high BGC gene target levels (> 6.9 x 10^5 and > 2.0 x 10^5 for the NRPS respectively; *SI Appendix,* Table S3) were selected for PacBio sequencing. This effort generated 28 GB of data that was used to create a new hybrid CoAssembly 2 which combined all three sequencing technologies. Similar to the assembly with the *Mycale hentscheli*-associated polyketide producers (46), the long-read data set improved the assembly metrics, and subsequent binning resulted in a highly resolved *Opitutaceae*-classified bin (*SI Appendix*, Fig. S2, Table S4). Interestingly however, the palmerolide BGC contigs still did not cluster with this bin, which we later attributed to binning reliance on sequence depth.

We used PacBio circular consensus sequence (CCS) reads to generate and manually edit the assembly for our *Opitutaceae* genome of interest. The resulting 4.3 MB genome (Fig. 3a) had a GC content of 58.7% and was resolved into a total of 10 contigs. Five of the contigs were unique and the other five contigs represented highly similar repeated units of the *pal* BGC (labeled *pal* BGC 1, 2, 3, 4, and 5) with broken ends resulting in linkage gaps. Nucmer alignment of contigs to the longest palmerolide-containing BGC revealed a long (36,198 kb) repeated region that was shared between all 5 contigs with some substantial differences at the beginning of the cluster and only minor differences at the end, indicating 3 full length, and 2 shorter palmerolide BGC-containing contigs (Fig. 1 and 3). This was consistent with coverage estimates based on read-mapping that suggested lower depth at the beginning of the cluster (Fig. 3b). BGC 1 and 3 are nearly identical (over 86,135 bases) with only 2 single nucleotide polymorphisms (SNPs) and an additional 1,468 bases in BGC1 (237 bases at the 5' end and 1231 bases at the 3' end). BGC 4 is 13,470 bases

shorter than BGC1 at the 5' end, and 5 bases longer than BGC 1 at the 3' end. Alignment of the real time PCR gene targets to the 5 *pal* BGCs provided independent support for the different lengths of the 5 BGCs, as the region targeted by the NRP primers was missing in two of the *pal* BGCs, thus explaining lower NRP: AT or NRP:HCS gene dosages reported above.

Interestingly, precedent for naturally occurring multi-copy BGCs to our knowledge, has only been found in another ascidian (*Lissoclinum* sp.)-associated *Opitutaceae*, *Candidatus* Didemnitutus mandela which have been linked to cytotoxic mandelalides (36). Likewise, we can invoke a rationale similar to (36) that multiple gene clusters may be linked to biosynthesis of different palmerolide variants, see Avalon et al. (40) for retro-biosynthetic predictions of these clusters. Gene duplication, loss, and rearrangement processes over evolutionary time, likely explain the source of the multiple copies. At present we do not yet understand the regulatory controls, whether all five are actively transcribed, if there is a producing-organism function and how this may vary amongst host microbiomes.

Phylogenomic characterization of the Opitutaceae-related MAG. The taxonomic relationship of the *Opitutacae* MAG to other Verrucomicrobiota was assessed using distance-based analyses with 16S rRNA and average amino acid identity (AAI). Then it was classified using the GTDB-Tk tool (47), and a phylogenomic analysis based on concatenated ribosomal protein markers. Comparison of 16S rRNA gene sequences amongst other Verrucomicrobia with available genome sequences (that also have 16S rRNA genes; *SI Appendix,* Fig. S3) suggests that the nearest relatives are *Cephaloticoccus primus* CAG34 (similarity of 0.9138), *Optitutus terrae* PB90-1 (similarity of 0.9132) and *Geminisphaera coliterminitum* TAV2 (similarity of 0.9108). The *Opitutaceae*-affiliated MAG sequence is identical to a sequence (uncultured bacterium clone Tun-3b A3) reported from the same host (*S. adareanum*) in a 2008 study (26); bootstrapping supported a deep branching position in the *Opitutaceae* family.

When characterizing the MAG using AAI metrics (average nucleotide identity, ANI, found no closely related genomes) the closest genomes were environmental metagenome assemblies from the South Atlantic TOBG_SAT_155 (53.08 % AAI) and WB6_3A_236 (52.71 % AAI); and the two closest isolate type genomes were *Nibricoccus aquaticus* str. NZ CP023344 (52.82 % AAI) and *Opitutus terrae* str. PB90 (52.75 % AAI). The Microbial Genome Atlas (MiGA) support for the MAG belonging in the *Opitutaceae* family was weak (p-values of 0.5). Attempts to classify this MAG using GTDB-Tk (47) were hampered by the fact we have no real representative in the genome databases, resulting in low confidence predictions at the species or genus levels (see the *SI Appendix* text for details).

Verrucomicrobia exhibit free-living and host-associated lifestyles in a multitude of terrestrial and marine habitats on Earth. We performed a meta-analysis of Verrucomicrobia genomes, with an emphasis on marine and host-associated *Opitutaceae*, to establish more confidence in the phylogenetic position of the *Opitutaceae* MAG. The analysis was based on 24 conserved proteins – 21 ribosomal proteins and three additional conserved proteins (InfB, lepA, pheS). The diversity of the *Opitutaceae* family, and of Verrucomicrobia in general, is largely known from uncultivated organisms in which there are 20 genera in GTDB (release 05-RS95), 2 additional genera in the NCBI taxonomy database, and numerous unclassified single amplified genomes (SAGs); in all, only eight genera have cultivated representatives. Given the uneven representations of the 24 proteins across all (115) genomes assessed (MAGs and SAGs are often incomplete), we selected a balance of 16 proteins across 48 genomes to assess phylogenomic relatedness across the *Opitutaceae* (Fig 4). Here too, as seen with the 16S rRNA gene phylogenetic tree, the *S. adarenum*-*Opitutaceae* MAG held a basal position compared to the other *Opitutaceae* genomes in the analysis.

***Opitutaceae-related MAG* relative abundance estimates and ecological inference.** The relative abundance of *Opitutaceae* bin 8 was estimated in the shotgun metagenomic samples by mapping the NGS reads back to the assembled MAG across the four *S. adareanum* samples collected. This indicated varying levels of genome coverage in the natural samples, with the two samples selected based on real time PCR-quantified high BGC copy number being clearly enriched

in this strain (44.70 % of reads mapped to Bon-1C-2011 and 36.78 % to Del-2B-2011, Table 1). These levels are higher than estimates of relative abundance derived from the 16S rRNA gene amplicon surveys (estimated at 13.33 and 15.34 % respectively) for the same samples. This is likely a result of the single-copy nature of the ribosomal operon in *Opitutaceae* bin 8 vs. other taxa with multiple rRNA operon copies that could thus be over-represented in the core microbiome library (e.g., *Pseudovibrio* sp. str. PSC04-5.I4 has 9 and *Microbulbifer* sp. is estimated at 4.1 ± 0.8 based on 9 finished *Microbulbifer* genomes available at the Integrated Microbial Genomes Database). All host *S. adarenum* lobes surveyed (n=63) in the Anvers Island regional survey contained high levels (0.49 – 4.06 mg palmerolide A x g^{-1} host dry weight) of palmerolide A (27), and variable, yet highly concordant levels of the *pal* BGCs and 16S rRNA ASV levels (Fig. 2). Despite the natural population structure sampled here (four single host lobes), the bin-level sequence variation was low (ranging from 72-243 SNPs) when the PacBio reads were mapped back to the Opitutaceae bin 8 (Table 1). This suggests maintenance of a relatively invariant population at the spatial and temporal scales of this coastal Antarctic region while highlighting our limited understanding of the biogeographical extent of the *S. adareanum*-symbiont-palmerolide relationship across a larger region of the Southern Ocean.

Several questions remain with regard to the in situ function of palmerolide A (a eukaryotic V-ATPase inhibitor in human cell line assays (28)) in this cryohabitat: how and why is it bioaccumulated by the host? Overall, the study of natural products in high latitude marine ecosystems is in its infancy. This palmerolide producing, ascidian-associated, *Opitutaceae* provides the first Antarctic example in which a well-characterized natural product has been linked to the genetic information responsible for its biosynthesis. Gaining an understanding of environmental and biosynthetic regulatory controls, establishing integrated transcriptomic, proteomic, and secondary metabolome expression in the environment will also reveal whether the different clusters are expressed in situ. In addition to ecological pursuits, the path to clinical studies of palmerolide will require genetic or cultivation efforts. At present, we hypothesize that cultivation of *Opitutaceae* bin 8 may be possible, given the lack of genome reduction or of other direct evidence for host-associated dependencies.

***Candidatus* Synoicihabitans palmerolidicus genome attributes**. The Antarctic ascidian, *Synoicum adareanum*, harbors a dense community of bacteria that has a conserved core set of taxa (27). The near complete ~4.30 Mbp *Opitutaceae* bin 8 metagenome assembled genome (Fig. 3) represents one of the core members. This MAG is remarkable in that it encodes for five 36-74 kb copies of the candidate BGCs that are implicated in biosynthesis of palmerolide A and possibly other palmerolide compounds. Intriguingly, this genome does not seem to show evidence of genome reduction as found in *Candidatus* Didemnitutus mandela (36); the other ascidian-associated *Opitutaceae* genome currently known to encode multiple BGC gene copies. This is the first *Opitutaceae* genome characterized from a permanently cold, ~ -1.8 - 2 °C, often ice-covered ocean ecosystem. This genome encodes one rRNA operon, 45 tRNA genes, and an estimated 5058 coding sequences. Based on the low (< 92%) SSU rRNA gene identity and low (< 54% AAI) values to other genera in the *Opitutaceae*, along with the phylogenomic position of the *Opitutaceae* bin 8, the provisional name "*Candidatus* Synoicihabitans palmerolidicus" (*Ca*. S. palmerolidicus) is proposed for this novel verrucomicrobium. The genus name *Synoicihabitans* (Syn.o.i.ci.ha'bitans. N.L. neut. N. *Synoicum* a genus of ascidians; L. pres. part *habitans* inhabiting; N.L. masc. n.) references this organism as an inhabitant of the ascidian genus *Synoicum*. The species name *palmerolidicus* (pal.me.ro.li'di.cus. N.L. neut. n. *palmerolidum* palmerolide; N.L. masc. adj.) designates the species as pertaining to palmerolide.

The GC content of 58.7% is rather high compared to other marine *Opitutaceae* genomes (ave. 51.49 s.d. 0.02, n=12), yet is ~ average for the family overall (61.58 s.d. 0.06, n=69; *SI Appendix,* Table S5). MetaERG includes metagenome assembled genomes available in the GTDB as a resource for its custom GenomeDB that new genomes are annotated against. This was a clear advantage in annotating the *Ca.* S. palmerolidicus genome as Verrucomicrobia genomes are widely represented by uncultivated taxa. Likewise, antiSMASH was an invaluable tool for *pal* BGC

identification and domain structure annotation. This formed the basis to derive a predicted step-wise mechanism of *pal* biosynthesis (40).

***Ca*. S. palmerolidicus genome structure, function and host-associated features.** Beyond the pal BGCs, the *Ca*. S. palmerolidicus genome encodes a variety of additional interesting structural and functional features that provide a window into its lifestyle. Here we will only provide a brief synopsis. In addition to the repeated BGCs, three additional repeats with two nearly identical copies each (15.3 Mb, 17.0 Mb, 27.4 Mb) were identified during the assembly process (Fig. 3a, *SI Appendix,* Table S6). These coded for 20, 25 and 41 CDs respectively, were in some cases flanked by transposase/integrases (both internal and proximal) and had widespread homology with Verrucomicrobia orthologs. The contents of the three repetitive elements were unique.

Annotations were assigned to a little more than half of the CDSs in the 15.3 Mb repeat in which support for xylose transport, two sulfatase copies, two endonuclease copies and a MacB-like (potential macrolide export) periplasmic core domain were encoded. Xylan might be sourced from seaweeds (48) or the even ascidian as it is a minor component of the tunic cellulose (49). Related to this, an endo-1,4-beta-xylanase which has exoenzyme activity in some microorganisms (50) was identified elsewhere in the genome. Altogether, eight sulfatase copies were identified in this host-associated organism (four in the 15.5 Mb repeat elements). These may be involved in catabolic activities of sulfonated polysaccharides, and possibly as *trans*-acting elements in palmerolide biosynthesis (40). In addition to the MacB-like CDSs found in this repeat, 13 different MacB-homologs were present in the genome – none of which were associated with the *pal* BGCs (*SI Appendix,* Fig. S4). MacB is a primary component of the macrolide tripartite efflux pump that operates as a mechanotransmission system which is involved both in antibiotic resistance and antibiotic export depending on the size of the macrolide molecule (51). However, two additional elements required for this pump to be functional, an intramembrane MacA and an outer membrane protein TolC, were not co-located in the genome. MacA may be missing, as hits to two other verrucomicrobia-associated MacA CDS were not identified using BLAST (Peat Soil MAG SbV1 SBV1_730043 and *Ca*. Udaeobacter copiosis KAF5408997.1; (52)). At least nine MacB CDS were flanked by a FstX-like permease family protein; the genomic structure of which were quite complex including several with multiple repeated domains. Detailed transporter modeling is beyond the scope of this work, but it is likely that these proteins are involved in signaling of cell division machinery rather than macrolide transport (53).

Predicted CDSs in the 17.0 Mb repeat included sugar binding and transport domains, as well as domains encoding rhamnosidase, arabinofuranosidase, and other carbohydrate catabolism functions. About half proteins encoded in the 27.4 Mb repeat were unknown in function, and those characterized suggested diverse potential functional capacities. For example, a zinc carboxypeptidase (1 of 3 in the genome), multidrug and toxic compound transporter (MatE/NorM), and an exodeoxyribonuclease were identified.

The *Ca.* S. palmerolidicus MAG has a number of features that suggest it is adapted to a host-associated lifestyle, several of these features were reported recently for two related sponge-associated Opitutales metagenome bins (*Petrosia ficiformis*-associated bins 0 and 01, Fig. 4; (54)). These include identification of a bacterial microcompartment (BMC) ‘super locus’. Such loci were recently reported to be enriched in host-associated Opitutales genomes when compared to free-living relatives. The structural proteins for the BMC were present as were other conserved Planctomyces-Verrucomicrobia BMC genes (55). As in the sponge *Pectoria ficiformis* metagenome bins, enzymes for carbohydrate (rhamnose) catabolism and modification were found adjacent to the BMC locus (*SI Appendix,* Fig. S5), in addition to the two that were found in the 27.4 Mb repeat. The genome did not appear to encode the full complement of enzymes required for fucose metabolism, though a few alpha-L-fucosidases were identified. Further evidence for carbohydrate metabolism was supported through classification of the genome using the CAZY database (56), including 7 carbohydrate binding modules, a carbohydrate esterase, 14 glycoside hydrolases, 6 glycosyl transferases and a polysaccharide lyase. In addition, three bacterial cellulases (PF00150, cellulase family A; glycosyl hydrolase family 5) were identified, all with a canonical conserved glutamic acid residue. These appear to have different evolutionary histories in which each variant

has nearest neighbors in different bacterial phyla (*SI Appendix*, Fig. S6) matching between 68% identity for Protein J6386 03765 to *Lacunisphaera limnophila*, 57.5% identity for Protein J6386 22340 with a cellulase from a shipworm symbiont *Alteromonadaceae* (*Terridinibacter* sp.), and 37.5% sequence identity to a Bacteroidetes bacterium. This suggests the potential for cellulose degradation – which is consistent with ascidians being the only animals known to produce cellulose where it acts as a skeletal structure (49). In addition to the BGCs, the enzymatic resources in this genome (e.g., xylan and cellulose hydrolysis) are a treasure trove rich with biotechnological potential.

Other indicators of host-association in the Opitutales include T-A domains, which were prevalent in the *Petrosia ficiformis*-associated bins 0 and 01 (54). The *Ca*. S. palmerolidicus genome encoded at least 22 TA-related genes including multiple MazG and AbiEii toxin type IV TA systems, AbiEii-Phd_YefM type II toxin-antitoxin systems, along with genes coding for PIN domains, Zeta toxin, RelB, HipA, MazE and MraZ. This analysis also resulted in identifying a putative AbiEii toxin (PF13304) with homology to SyrD, a cyclic peptide ABC type transporter that was present in all 5 BGCs (Fig. 1c; BLAST percent identity 52.7% to a *Desulfamplus* sp. homolog over the full length of the protein, and a variety of other bacteria including an *Opitutaceae*-related strain at similar levels of identity). These genes are encoded downstream of the BGC following the acyl transferase domains and precede the predicted *trans*-acting domains at the 3' end of the BGC. Given the proximity adjacent to the primary biosynthetic gene clusters, this protein is a candidate for palmerolide transport. Further research is needed however to discern the details of Ca. S. palmerolidicus' cellular biology, localization of palmerolide production, transport, and resistance mechanisms to the potent vacuolar ATPase as well as products made by others in the *S. adareanum* microbiome. Along these lines, in addition to the MatE (found in the 27.4 Mb repeat) two other multidrug export systems with homology to MexB and MdtB were identified.

Unlike in *Ca*. D. mandela (36), there does not appear to be ongoing genome reduction, which may suggest that the *S. adareanum-Ca*. S palmerolidicus relationship is more recent, and/or that the relationship is commensal rather than interdependent. Likewise, we suspect that the pseudogene content may be high as several CDS appear to be truncated, in which redundant CDS of varying lengths were found in several cases (including the MacB). There is evidence of lateral gene transfer acquisitions of cellulase and numerous other enzymes that may confer ecological advantages through the evolution of this genome. Likewise, the origin of the *pal* BGCs and how recombination events play out in the success of this Antarctic host-associated system in terms of adaptive evolution (57), not to mention the ecology of *S. adareanum* is a curiosity. This phylum promises to be an interesting target for further culture-based and cultivation-free studies – particularly in the marine environment.

Together, it appears that the genome of *Ca*. S. palmerolidicus is equipped for life in this host-associated interactive ecosystem that stands to be one of the first high latitude marine invertebrate-associated microbiomes with a genome-level understanding – and one that produces a highly potent natural product, palmerolide A. This system holds promise for future research now that we have identified the producing organism and *pal* BGC. We still have much to learn about the ecological role of palmerolide A – if is it involved predation avoidance, antifouling, antimicrobial defense or some other yet to be recognized aspect of life in the frigid, often ice-covered and seasonally light-limited waters of the Southern Ocean.

Materials and Methods

Sample Collection. *S. adareanum* lobes were collected in the coastal waters off Anvers Island, Antarctica and stored at -80 °C until processing (*SI Appendix*, Table S1). See the *SI Appendix* text and (27) for details of sample collection, microbial cell preparation and DNA extraction.

Metagenome sequencing. Three rounds of metagenome sequencing were conducted, the details of which are in the *SI Appendix* text. This included an initial 454 pyrosequencing effort with a bacterial-enriched metagenomic DNA preparation from *S. adareanum* lobe (Nor2c-2007). Next,

an Ion Proton System was used to sequence a metagenomic DNA sample prepared from *S. adareanum* lobe Nor2a-2007. Then two additional *S. adareanum* metagenome DNA samples (Bon-1C-2011 and Del-2b-2011) selected based on high copy numbers of the palmerolide A BGC (see real time PCR Methods, *SI Appendix* text) and sequenced using Pacific Biosciences Sequel Systems technology.

Metagenome assembly, annotation and binning. Raw 454 metagenomic reads (1,570,137 single end reads, 904,455,285 bases) were assembled by Newbler (58) v2.9 (Life Technologies, Carlsbad, CA, flags: -large -rip -mi 98 -ml 80), while Ion Proton metagenomic reads (89,330,870 reads, 17,053,251,055 bases) were assembled using SPAdes (59) v3.5 (flags: --iontorrent). Both assembled datasets were merged with MeGAMerge (60) v1.2 and produced 86,387 contigs with a maximum contig size 153,680 and total contig size 144,953,904 bases (CoAssembly 1). To achieve more complete metagenome coverage and facilitate metagenome assembled genome assembly, a Circular Consensus Sequence (CCS) protocol (PacBio) was used to obtain high quality long reads on two samples Bon-1C-2011 and Del-2b-2011. The 5,514,426 PacBio reads were assembled with aforementioned assembled contigs (CoAssembly 1) on EDGE Bioinformatics using wtdbg2 (61), a fast and accurate long-read assembler. The contigs were polished with three rounds of polishing by Racon (62) into a second Coassembly (CoAssembly 2) which has 4,215 contigs with a maximum contig size 2,235,039 and total size 97,970,181 bases. Lastly, A manual approach was implemented to arrive at assembly of the MAG of interest the details of which are described in the *SI Appendix* text.

The contigs from both co-assemblies 1 and 2 were submitted initially to the EDGE bioinformatics platform (63) for sequence annotation using Prokka (64) v1.13 and taxonomy classification using BWA (65) mapping to NCBI RefSeq (Version: NCBI 2017 Oct 3). Bioinformatic predictions of natural product potential was performed using the antibiotics and secondary metabolite analysis shell (antiSMASH, bacterial versions 3.0, 4.0 and 5.0 (45, 70). This tool executed contig identification, annotation and analysis of secondary metabolite biosynthesis gene clusters on both CoAssemblies 1 and 2 (> 1 kb and > 40 kb data sets). As most of our attention was focused on analysis the *Ca*. S. palmerolidicus assembled metagenome, we also used MetaERG (66) as the primary pipeline for metagenome annotation of the ten final contigs in addition to NCBI's PGAP pipeline. There were 5186 coding sequences predicted in the MetaERG annotation and 5186 in NCBI's PGAP annotation.

MaxBin (67) and MaxBin2 (68) were used to form metagenome bins for both CoAssembly 1 and 2. CheckM v1.1.11 (69) and v.1.1.12, and GTDB-Tk v.1.0.2 (47) were used to verify bin quality and taxonomic classification. See *SI Appendix* text for details. In order to assess the representation of assembled *Opitutaceae* genome across the 4 environmental samples used for metagenome sequencing (resulting from MaxBin2 binning of CoAssembly 2), we used BWA to map the CCS reads to each metagenome data set.

Real time PCR. Gene targets (non-ribosomal peptide synthase, acyltransferase, and 3-hydroxymethylglutaryl coenzyme A synthase) were selected at different positions along the length of the candidate BGC. *SI Appendix,* Table S7 lists the primer and the GBlocks synthetic positive control sequence. Metagenomic DNA extracts from a large *S. adareanum* sample set (n=63 *S. adareanum* lobes from 21 colonies), all containing high levels of palmerolide A (27), were screened with the real time PCR assays on a Quant Studio 3 (Thermo Fisher Scientific, Inc.; see *SI Appendix* text for details of controls and analysis).

Phylogenomic analyses. A phylogenomic analysis of the assembled *Opitutaceae* MAG was conducted based on shared ribosomal RNA and ribosomal proteins amongst 46 and 48 reference genomes respectively, out of 115 genomes in total, mined from various databases (NCBI, GTDB and IMG) for uncultivated and cultivated microorganisms identified in the Verrucomicrobia phylum (*SI Appendix*, Table S8). The details of these analyses are described in the *SI Appendix* text. In addition, we used MiGA (NCBI Prokaryotic taxonomy and the environmental TARA Oceans

(Tully) databases; accessed August 2020) and GTDB-Tk (ver. 1.3.0) tools for MAG taxonomic classification.

Phylogenetic analysis of the MacB CDS sequences were retrieved from MetatERG annotated *Ca*. S. palmerolidicus contigs and homologs were retrieved from the NCBI based on BLAST results. Maximum likelihood analysis was conducted on 994 aligned (MUSCLE) positions using RAxML v.8.2.12 using the PROTGAMMALG model and 550 bootstrap replicates. For the phylogenetic analysis of the cellulase CDS, homologs were retrieved from the NCBI based on BLAST results resulting in 19 sequences and 496 aligned positions (ClustalOmega) was also conducted using RAxML v.8.2.12 under the PROTGAMMALG model of evolution with 1000 bootstraps.

Acknowledgments

Support for this research was provided in part by the National Institute of Health award (CA205932) to A.E.M., B.J.B., and P.S.G.C., with additional support from National Science Foundation awards (OPP-0442857, ANT-0838776, and PLR-1341339) to B.J.B., and Desert Research Institute (Institute Project Assignment) to A.E.M..

The assistance of several collaborators and students are acknowledged including C. Amsler, M. Amsler, L. Bishop, J. Cuce, B. Dent, N. Ernster, C. Gleasner, A. Maschek, A. Shilling, L. Siao, S. Thomas, and the Palmer Station science support staff. We especially would like to acknowledge the help of A. Oren in assisting with bacterial nomenclature proposed here. Likewise, we thank J.T. Hollibaugh and A.L. Reysenbach for comments on previous drafts of the manuscript.

References

1. T. L. Simmons *et al.*, Biosynthetic origin of natural products isolated from marine microorganism-invertebrate assemblages. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 4587-4594 (2008).
2. M. Morita, E. W. Schmidt, Parallel lives of symbionts and hosts: chemical mutualism in marine animals. *Nat. Prod. Rep.* **35**, 357-378 (2018).
3. J. Piel, Metabolites from symbiotic bacteria. *Nat. Prod. Rep.* **21**, 519-538 (2004).
4. E. W. Schmidt *et al.*, Patellamide A and C biosynthesis by a microcin-like pathway in *Prochloron didemni*, the cyanobacterial symbiont of *Lissoclinum patella*. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 7315-7320 (2005).
5. J. Piel, Metabolites from symbiotic bacteria. *Nat. Prod. Rep.* **26**, 338-362 (2009).
6. S. Sunagawa, C. M. Woodley, M. Medina, Threatened corals provide underexplored microbial habitats. *Plos One* **5**, e9554 (2010).
7. L. Behrendt *et al.*, Microbial diversity of biofilm communities in microniches associated with the didemnid ascidian Lissoclinum patella. *ISME J.* **6**, 1222-1237 (2012).
8. N. S. Webster, M. W. Taylor, Marine sponges and their microbial symbionts: love and other relationships. *Environ. Microbiol.* **14**, 335-346 (2012).
9. M. McFall-Ngai *et al.*, Animals in a bacterial world, a new imperative for the life sciences. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 3229-3236 (2013).
10. J. Gray, Antarctic marine benthic biodiversity in a world-wide latitudinal context. *Polar Biol.* **24**, 633-641 (2001).
11. A. Clarke, Antarctic marine benthic diversity: patterns and processes. *J. Exp. Mar. Biol. Ecol.* **366**, 48-55 (2008).

12. D. Piepenburg *et al.*, Towards a pan-Arctic inventory of the species diversity of the macro- and megabenthic fauna of the Arctic shelf seas. *Mar. Biodivers.* **41**, 51-70 (2011).
13. J. von Salm, K. Schoenrock, J. McClintock, C. Amsler, B. Baker, "The status of marine chemical ecology in Antarctica: form and function of unique high-latitude chemistry" in Chemical Ecology: The ecological impacts of marine natural products. (CRC Press, Boca Raton, 2018), 10.1201/9780429453465, pp. 69.
14. J. B. McClintock, C. D. Amsler, B. J. Baker, Introduction to the symposium: Antarctic marine biology. *Am. Zool.* **41**, 1-2 (2001).
15. G. J. Bakus, G. Green, Toxicity in sponges and holothurians - geographic pattern. *Science* **185**, 951-953 (1974).
16. S. Soldatou, B. J. Baker, Cold-water marine natural products, 2006 to 2016. *Nat. Prod. Rep.* **34**, 585-626 (2017).
17. C. Avila, S. Taboada, L. Nunez-Pons, Antarctic marine chemical ecology: what is next? *Mar. Ecol.-Evol. Persp.* **29**, 1-71 (2008).
18. N. S. Webster, A. P. Negri, M. Munro, C. N. Battershill, Diverse microbial communities inhabit Antarctic sponges. *Environ. Microbiol.* **6**, 288-300 (2004).
19. N. S. Webster, D. Bourne, Bacterial community structure associated with the Antarctic soft coral, *Alcyonium antarcticum*. *FEMS Microbiol. Ecol.* **59**, 81-94 (2007).
20. A. E. Murray *et al.*, Microbiome composition and diversity of the ice-dwelling sea anemone, *Edwardsiella andrillae*. *Integr. Comp. Biol.* **56**, 542-555 (2016).
21. S. Rodriguez-Marconi *et al.*, Characterization of bacterial, archaeal and eukaryote symbionts from Antarctic sponges reveals a high diversity at a three-domain level and a particular signature for this ecosystem. *Plos One* **10**, e0138837 (2015).
22. V. M. Godinho *et al.*, Diversity and distribution of hidden cultivable fungi associated with marine animals of Antarctica. *Fungal Biol.-UK* **123**, 507-516 (2019).
23. G. Steinert *et al.*, Prokaryotic diversity and community patterns in Antarctic continental shelf sponges. *Front. Mar. Sci.* **6**, 297 (2019).
24. O. Sacristan-Soriano, N. P. Criado, C. Avila, Host species determines symbiotic community composition in Antarctic sponges (Porifera: *Demospongiae*). *Front. Mar. Sci.* **7**, 474 (2020).
25. M. Moreno-Pino, A. Cristi, J. F. Gillooly, N. Trefault, Characterizing the microbiomes of Antarctic sponges: a functional metagenomic approach. *Sci. Rep.* **10**, 645 (2020).
26. C. S. Riesenfeld, A. E. Murray, B. J. Baker, Characterization of the microbial community and polyketide biosynthetic potential in the palmerolide-producing tunicate, *Synoicum adareanum*. *J. Nat. Prod.* **71**, 1812-1818 (2008).
27. A. E. Murray *et al.*, Uncovering the core microbiome and distributions of palmerolide in *Synoicum adareanum* across the Anvers Island archipelago, Antarctica. *Mar. Drugs* **18**, 298 (2020).
28. T. Diyabalanage, C. D. Amsler, J. B. McClintock, B. J. Baker, Palmerolide A, a cytotoxic macrolide from the Antarctic tunicate *Synoicum adareanum*. *J. Am. Chem. Soc.* **128**, 5630-5631 (2006).
29. L. H. Franco *et al.*, Indole alkaloids from the tunicate *Aplidium meridianum*. *J. Nat. Prod.* **61**, 1130-1132 (1998).
30. Y. Miyata *et al.*, Ecdysteroids from the antarctic tunicate *Synoicum adareanum*. *J. Nat. Prod.* **70**, 1859-1864 (2007).

31. A. M. Seldes, M. F. Rodrıguez Brasco, L. H. Franco, J. A. Palermo, Identification of two meridianins from the crude extract of the tunicate *Aplidium meridianum* by tandem mass spectrometry. *Nat. Prod. Rep.* **21**, 555–563 (2007).
32. L. Chen, J. S. Hu, J. L. Xu, C. L. Shao, G. Y. Wang, Biological and chemical diversity of ascidian-associated microorganisms. *Mar. Drugs* **16**, 362 (2018).
33. X. Dou, B. Dong, Origins and Bioactivities of Natural Compounds Derived from Marine Ascidians and Their Symbionts. *Mar. Drugs* **17**, 670 (2019).
34. C. M. Rath *et al.*, Meta-omic characterization of the marine invertebrate microbial consortium that produces the chemotherapeutic natural product ET-743. *ACS Chem. Biol.* **6**, 1244-1256 (2011).
35. J. C. Kwan *et al.*, Genome streamlining and chemical defense in a coral reef symbiosis. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 20655-20660 (2012).
36. J. Lopera, I. J. Miller, K. L. McPhail, J. C. Kwan, Increased biosynthetic gene dosage in a genome-reduced defensive bacterial symbiont. *mSystems* **2**, e00096-00017 (2017).
37. N. Shenkar, B. J. Swalla, Global diversity of Ascidiacea. *PLoS One* **6**, e20657 (2011).
38. T. Weber *et al.*, antiSMASH 3.0-a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Res.* **43**, W237-W243 (2015).
39. K. Blin *et al.*, antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res.* **47**, W81-W87 (2019).
40. N. Avalon *et al.*, Palmerolide PKS-NRPS gene clusters characterized from the microbiome of an Antarctic ascidian. *bioRxiv*, doi.org/10.1101/2021.1104.1105.438531.
41. J. Piel, A polyketide synthase-peptide synthetase gene cluster from an uncultured bacterial symbiont of *Paederus* beetles. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 14002-14007 (2002).
42. Y. Q. Cheng, G. L. Tang, B. Shen, Identification and localization of the gene cluster encoding biosynthesis of the antitumor macrolactam leinamycin in *Streptomyces atroolivaceus* S-140. *J. Bacteriol.* **184**, 7013-7024 (2002).
43. D. J. Edwards *et al.*, Structure and biosynthesis of the jamaicamides, new mixed polyketide-peptide neurotoxins from the marine cyanobacterium *Lyngbya majuscula*. *Chem. Biol.* **11**, 817-833 (2004).
44. L. Kjaerulff *et al.*, Pyxipyrrolones: structure elucidation and biosynthesis of cytotoxic myxobacterial metabolites. *Angew. Chem. Int. Ed. Engl.* **56**, 9614-9618 (2017).
45. K. Blin *et al.*, The antiSMASH database version 2: a comprehensive resource on secondary metabolite biosynthetic gene clusters. *Nucleic Acids Res.* **47**, D625-D630 (2019).
46. M. A. Storey *et al.*, Metagenomic exploration of the marine sponge *Mycale hentscheli* uncovers multiple polyketide-producing bacterial symbionts. *Mbio* **11**, e02997-02919 (2020).
47. P.-A. Chaumeil, A. Mussig, P. Hugenholtz, D. Parks, GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925–1927 (2020).
48. F. I. Qeshmi *et al.*, Xylanases from marine microorganisms: A brief overview on scope, sources, features and potential applications. *BBA-Proteins Proteom.* **1868**, 140312 (2020).
49. Y. Zhao, J. Li, Excellent chemical and material cellulose from tunicates: diversity in cellulose production yield and chemical and morphological structures from differnet tunicate species. *Cellul. Chem. Technol.* **21**, 3427-3441 (2014).

50. V. Juturu, J. C. Wu, Microbial exo-xylanases: a mini review. *Appl. Biochem. Biotechnol.* **174**, 81-92 (2014).
51. N. P. Greene, E. Kaplan, A. Crow, V. Koronakis, Antibiotic resistance mediated by the MacB ABC transporter family: a structural and functional perspective. *Front. Microbiol.* **9**, 950 (2018).
52. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *J. Mol. Biol.* **215**, 403-410 (1990).
53. A. Crow, N. P. Greene, E. Kaplan, V. Koronakis, Structure and mechanotransmission mechanism of the MacB ABC transporter superfamily. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 12572-12577 (2017).
54. S. Sizikov *et al.*, 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 (2020).
55. O. Erbilgin, K. L. McDonald, C. A. Kerfeld, Characterization of a planctomycetal organelle: a novel bacterial microcompartment for the aerobic degradation of plant saccharides. *Appl. Environ. Microbiol.* **80**, 2193-2205 (2014).
56. Lombard V, H. G. Ramulu, E. Drula, P. Coutinho, B. Henrissat, The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Ress* **42**, D490–D495 (2014).
57. M. G. Chevrette *et al.*, Evolutionary dynamics of natural product biosynthesis in bacteria. *Nat. Prod. Rep.* **37**, 566-599 (2020).
58. M. J. Chaisson, P. A. Pevzner, Short read fragment assembly of bacterial genomes. *Genome Res.* **18**, 324-330 (2008).
59. S. Nurk *et al.*, Assembling single-cell genomes and mini-metagenomes from chimeric MDA products. *J. Comput. Biol.* **20**, 714-737 (2013).
60. M. Scholz, C. C. Lo, P. S. G. Chain, Improved assemblies using a source-agnostic pipeline for metagenomic assembly by merging (MeGAMerge) of contigs. *Sci. Rep-UK* **4**, 6480 (2014).
61. J. Ruan, L. Li, Fast and accurate long-read assembly with wtdbg2. *Nat. Methods* **17**, 155-158 (2019).
62. R. Vaser, I. Sovic, N. D. Nagarajan, M. Sikic, Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737-746 (2017).
63. P.-E. Li *et al.*, Enabling the democratization of the genomics revolution with a fully integrated web-based bioinformatics platform. *Nucleic Acids Res.* **45**, 67-80 (2017).
64. T. Seemann, Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068-2069 (2014).
65. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).
66. X. L. Dong, M. Strous, An integrated pipeline for annotation and visualization of metagenomic contigs. *Front. Genetics* **10**, 999 (2019).
67. Y. W. Wu, Y. H. Tang, S. G. Tringe, B. A. Simmons, S. W. Singer, MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. *Microbiome* **2**, 26 (2014).
68. Y. W. Wu, B. A. Simmons, S. W. Singer, MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* **32**, 605-607 (2016).

693 69. D. H. Parks, M. Imelfort, C. T. Skennerton, P. Hugenholtz, G. W. Tyson, CheckM:
694 assessing the quality of microbial genomes recovered from isolates, single cells, and
695 metagenomes. *Genome Res.* **25**, 1043-1055 (2015).
696 70. T. Weber *et al.*, antiSMASH 3.0 - a comprehensive resource for the genome mining of
697 biosynthetic gene clusters. *Nucleic Acids Res.* **43**, W237-W243 (2015).

698

Figures and Tables

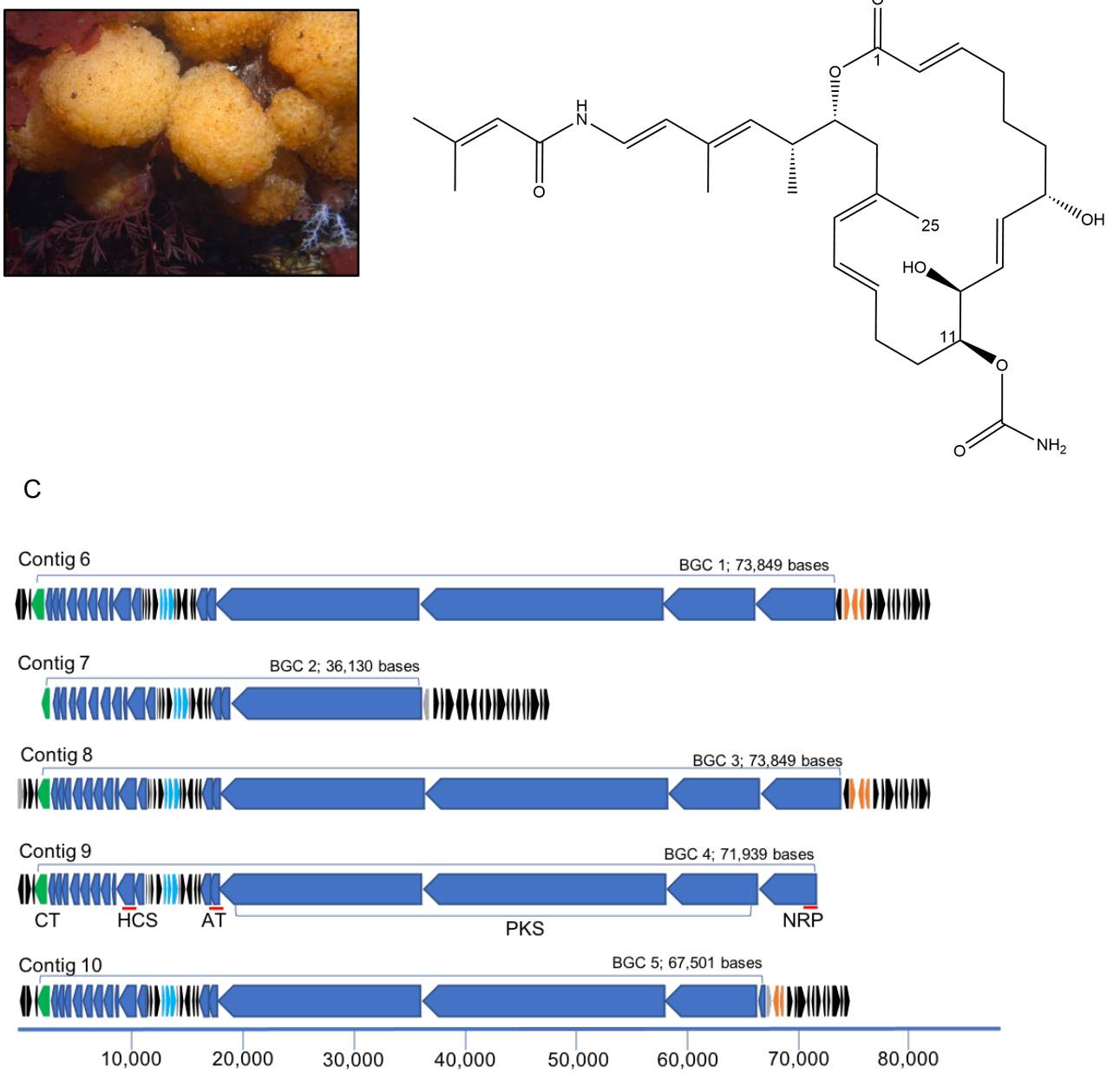
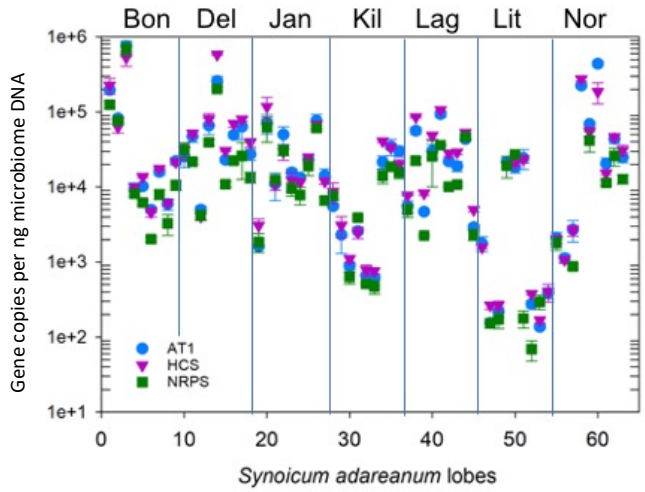


Figure 1. Palmerolide A, a cytotoxic, macrolide with anti-melanoma activity is found in the tissues of *Synoicum adareanum* in which a candidate biosynthetic gene cluster has been identifed. (A) *S. adareanum* occurs on rocky coastal seafloor habitats in the Antarctic; this study focused on the region off-shore of Anvers Island in the Antarctic Peninsula. (B) Palmerolide A, is the product of a hybrid PKS-NRPS system in which biosynthesis begins with a PKS starter unit followed by incorporation of a glycine subunit by an NRPS module. Subsequent elongation, cyclization and termination steps follow. Two additional features of the molecule include a methyl group on C-25

716 and a carbamate group on C-11. (C) Five repeats encompassing candidate palmerolide
717 biosynthetic gene clusters were identified. The BGC (in blue) is defined as starting with the NRP
718 unit and ending at the carbamoyltransferase (green). Candidate palmerolide A biosynthetic gene
719 cluster BGC4 was identified from initial metagenome library assemblies. The other four clusters
720 were identified following a third round of sequencing, assembly and manual finishing. Primary
721 BGC coding sequences (CDS) and a conserved tailoring cassette are in blue. Light blue CDS are
722 an ATP transporter with homology to an antibiotic transporter, SyrD. All black CDS are repeated
723 amongst the BGCs. Orange CDS are transposase/integrase domains. Gray CDS are unique,
724 non-repeated; and in BGC2 and 5, the unique CDS encode transposases, distinct from the
725 predicted amino acid sequences of those in orange. The red lines associated with Contig 9
726 indicate targeted quantitative PCR regions.
727

A



B

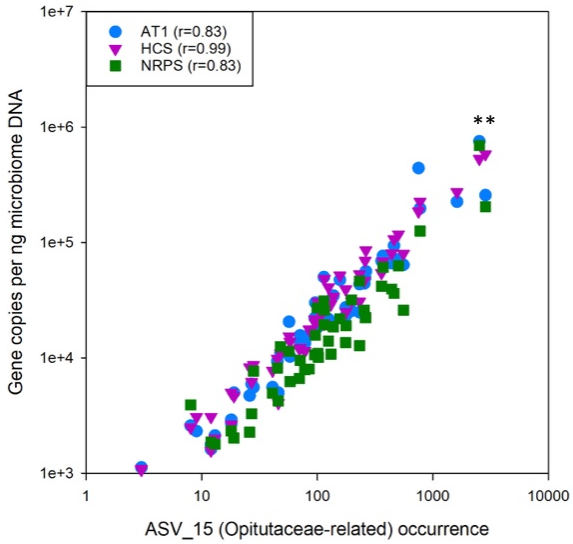
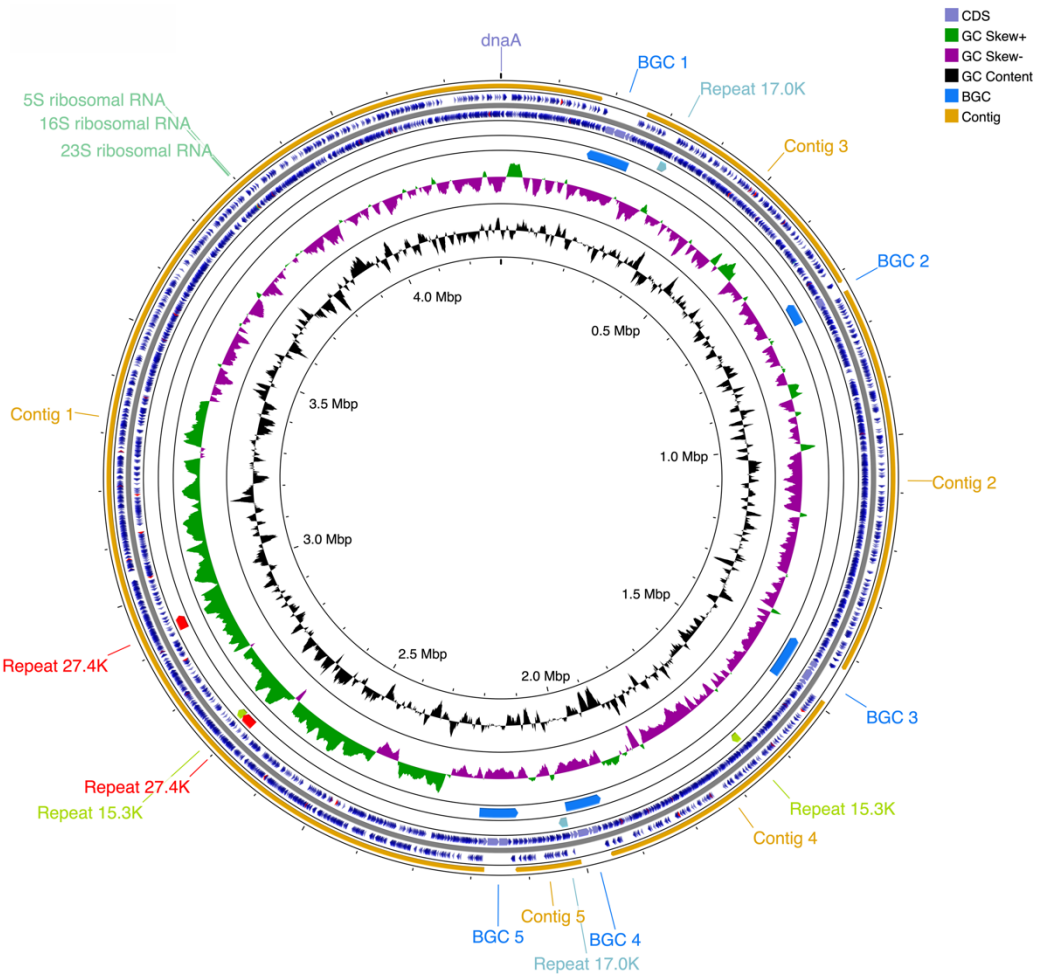
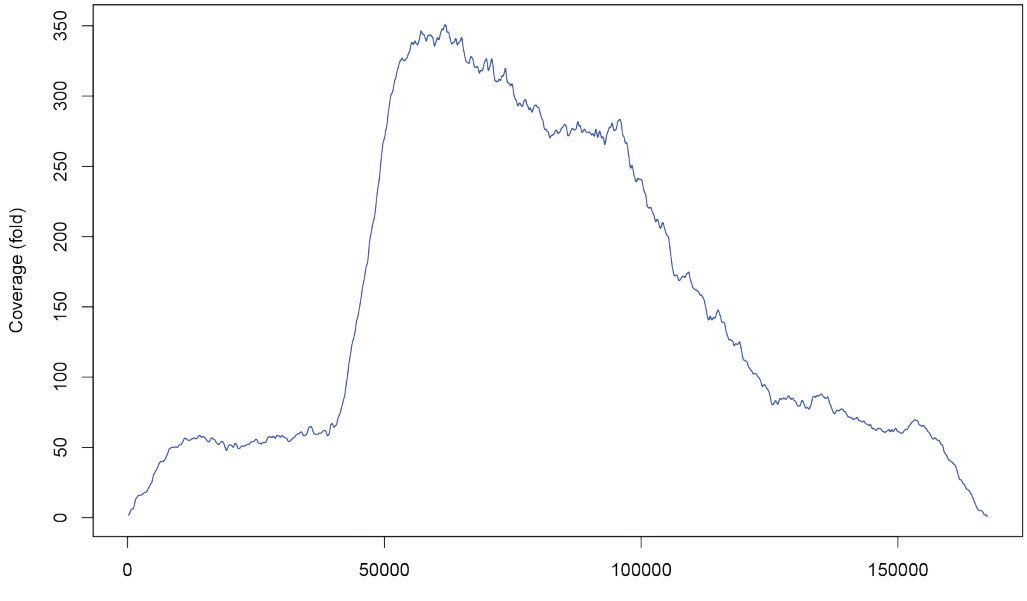


Figure 2. Abundances of real time PCR-targeted coding regions in the candidate pal biosynthetic gene cluster in Antarctic ascidian samples. (A) Gene copies estimated for three targeted coding regions (Acyltransferase, AT1; 3-hydroxy-methyl-glutaryl coenzyme A synthase, HCS; and the condensation domain of a non-ribosomal peptide synthase; NRPS) in the candidate *pal*A biosynthetic gene cluster surveyed over 63 DNA extracts derived from microbial cell preparations enriched from the Antarctic ascidian *Synoicum adarenum.* Nine samples were collected at each of seven sites: Bon, Bonaparte Point; Del, Delaca Island; Jan, Janus Island, Kil, Killer Whale Rocks; Lag, Laggard Island, Lit, Litchfield Island; Nor, Norsel Point (27). (B) Relationship between gene copy number for the three gene targets and the 16S rRNA gene ASV occurrences of Opitutaceae-related ASV_15 across a 63 *S. adareanum* microbial DNA sample set. * indicates samples Bon-1C-2011 and Del-2b-2011 that were selected for PacBio sequencing.

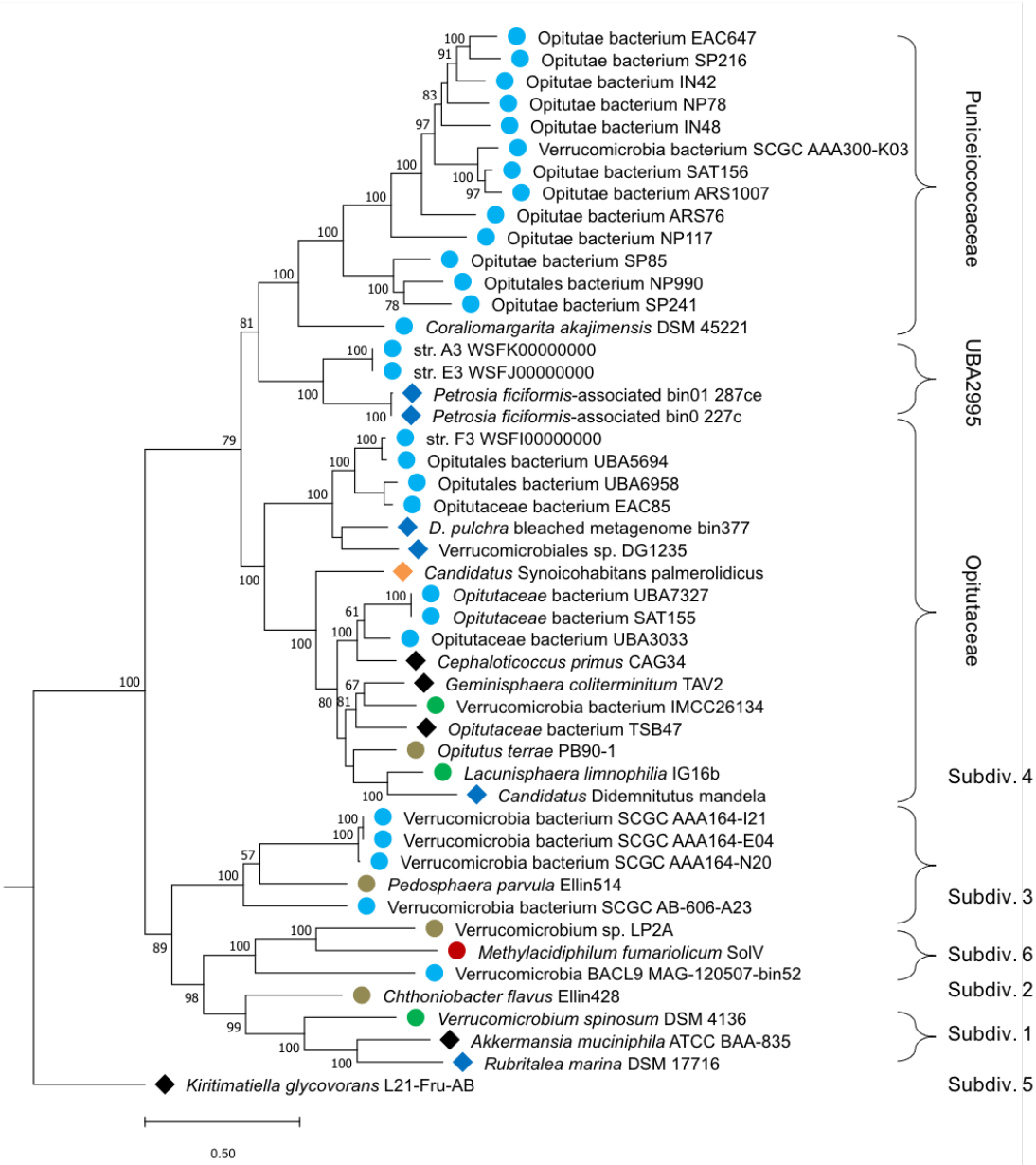
A



B



779 **Figure 3**. Genome maps of assembled MAG, *Candidatus* Synoicohabitans palmerolidicus and
780 evidence of muti-copy biosynthetic gene clusters. (a) The 4,297,084 bp gene map is oriented to
781 dnaA at the origin. One possible assembly scenario of the *Ca.* S. palmerolidicus genome is
782 shown as the order of the contigs and palmerolide BGC's are not currently known. In addition to
783 the 5 BGCs, three other internally repetitive regions were identified (15.3, 17.0 and 27.4 kb). The
784 genes and orientation are shown with blue and tRNA indicated in red. (b) To demonstrate depth
785 of coverage outside and inside the BGC regions, CCS reads from sample Bon-1C-2011 and Del-
786 2b-2011 were mapped to a 167.6 Kb region. The profile extends 40 kb into the genome on either
787 side of the BGC where depth of coverage averages 60 fold, while in the BGC depth of coverage
788 varies across the BGC given differences in cover across the BGC, the highest cover is 5X, or ~
789 300 fold, supporting the finding of 5 repeats encoding the BGC.
790



791

792 **Figure. 4.** Maximum likelihood phylogenomic tree showing 48 verrucomicrobia genomes
793 Phylogenomic relationship of *Candidatus* Synoicohabitans palmerolidicus (*Opitutaceae* bin 8)
794 with respect to other mostly marine, and host-associated verrrucomicrobia subdivision 4 and
795 other genomes. The tree is based on 16 concatenated ribosomal proteins (5325 amino acids)
796 common across 48 Verrucomicrobia genomes. Distance was estimated with RAxML with 300
797 bootstrap replicates. Symbols designate environmental origins of the organisms: free-living are
798 represented by circles: light blue - marine, green– freshwater, red – hydrothermal mud, brown –
799 soils. Host-associated taxa from marine systems with blue diamonds, and from terrestrial systems
800 with black diamonds.

Table 1. Metagenomic reads from 4 different samples were mapped back to the *Ca.* S. palmerolidicus MAG.

	Synoicum adareanum samples			
	Bon-1c-2011	Del-2b-2011	Nor-2c-2007	Nor-2a-2007
Technology	PacBio CCS reads (1 cell)	PacBio CCS reads (3 cells)	454	Ion Torrent Proton
Number of reads	48,298	9,576	1,570,126	89,330,870
Mapped reads	21,591	3,522	23,993	15,979,084
Mapped reads (%)	44.70	36.78	1.53	17.89
Base coverage	99.98	99.89	90.15	99.98
Average fold	58.38	8.43	2.79	708.79
Gaps	2	3	1,734	8
Gap bases	644	4,618	422,870	774
SNPs	72	196	168	243
Indels	126	64	17	68