Silva Sandra (Orcid ID: 0000-0002-4763-0662)
Costa Rodrigo (Orcid ID: 0000-0002-5932-4101)

Comparative genomics reveals complex natural product biosynthesis capacities and carbon metabolism across host-associated and free-living *Aquimarina* (*Bacteroidetes, Flavobacteriaceae*) species

Sandra G. Silva[1], Jochen Blom[2], Tina Keller-Costa[1], Rodrigo Costa[1,3*]

[1]Institute for Bioengineering and Biosciences (iBB), Instituto Superior Técnico (IST), Universidade de Lisboa, Lisbon, Portugal.

[2]Bioinformatics and Systems Biology, Justus-Liebig-University Giessen, 35392 Giessen, Germany.

[3]Centre of Marine Sciences (CCMAR), Algarve University, 8005-139, Faro, Portugal.

[*]**Corresponding author:** Institute for Bioengineering and Biosciences (iBB), Instituto Superior Técnico (IST), Universidade de Lisboa. Av. Rovisco Pais 1, South Tower, Room 8.6-22, 1049-001 Lisbon, Portugal. Tel: +351 21 841 7339 E-mail: rodrigoscosta@tecnico.ulisboa.pt

Keywords: bacterial evolution; biosynthetic gene clusters (BGCs); chitinases; host-microbe interactions; phylogenomics; secondary metabolism

Running head: Comparative genomics of *Aquimarina* species

Originality-Significance statement: Previously underestimated, complex secondary metabolism is revealed for the recently described bacterial genus *Aquimarina*, an emerging keystone taxon mediating carbon and nitrogen cycling and host-microbe interactions across multiple marine microniches.

Summary

This study determines the natural product biosynthesis and full coding potential within the bacterial genus *Aquimarina*. Using comprehensive phylogenomics and functional genomics, we reveal that phylogeny instead of isolation source (host-associated vs. free-living habitats) primarily shape the inferred metabolism of A*quimarina* species. These can be coherently organized into three major functional clusters, each presenting distinct natural product biosynthesis profiles suggesting that evolutionary trajectories strongly underpin their secondary metabolite repertoire and presumed bioactivities. *Aquimarina* spp. are highly versatile bacteria equipped to colonize host-associated and free-living microniches, eventually displaying opportunistic behavior, owing to their shared ability to produce multiple glycoside hydrolases from diverse families. We furthermore uncover previously underestimated, highly complex secondary metabolism for the genus by detecting 928 biosynthetic gene clusters across all genomes, grouped in 439 BGC families, with polyketide synthases (PKSs), terpene synthases and non-ribosomal peptide synthetases (NRPSs) ranking as the most frequent BGCs encoding drug-like candidates. We demonstrate that the recently described cuniculene (*trans*-AT PKS) BGC is conserved among, and specific to, the here delineated *A. megaterium-macrocephali-atlantica* phylogenomic clade. Our findings provide a timely and in-depth perspective of an under-explored yet emerging keystone taxon in the cycling of organic matter and secondary metabolite production in marine ecosystems.

Introduction

Compelling evidence presently exists for the involvement of species in the genus *Aquimarina* (*Bacteroidetes*, *Flavobacteriaceae*) in manifold processes of relevance in marine biomes, ranging from regulation of harmful microbial blooms through mediation of carbon and nitrogen cycling to the emergence of infectious diseases (Chen *et al.*, 2012; Lin *et al.*, 2012; Zhou *et al.*, 2015; V. Kumar *et al.*, 2016). Such processes are of acknowledged importance for the conservation of marine ecosystems and the development of a modern, blue biotechnology (Pinhassi *et al.*, 2004; Alonso *et al.*, 2007; Unfried *et al.*, 2018), with implications to the development of several sectors such as fisheries, aquaculture, tourism and human health. However, our current understanding of the full coding potential within the genus and the roles *Aquimarina* spp. play in holobiont and ecosystem functioning remains limited regardless of their presumed importance and biotechnological potential.

Despite its quite recent discovery in 2005 (Nedashkovskaya *et al.*, 2005), a wide range of bioactivities of putative relevance in the cycling of matter and establishment of host-microbe and microbe-microbe relationships has lately been reported for a few *Aquimarina* strains. Besides showing a versatile carbon metabolism typical of members of the family *Flavobacteriaceae* and, in a broader sense, of the phylum *Bacteroidetes* (Alonso *et al.*, 2007; Boone *et al.*, 2010; Fernández-Gómez *et al.*, 2013; Unfried *et al.*, 2018), these organisms have been matter of current research attention owing to metabolic features of both ecological and biotechnological interest. For instance, *Aquimarina salinaria* (Chen *et al.*, 2012) was shown to possess antagonistic activity towards the toxic cyanobacterium *Microcystis aeruginosa* (Ming *et al.*, 2011; Chen *et al.*, 2012). Furthermore, at least three *Aquimarina* species have

already been described to possess agarolytic activity, namely *A. aggregata*, *A. agarivorans* and *A. agarilytica* (Lin *et al.*, 2012; Zhou *et al.*, 2015; Wang *et al.*, 2016). *Aquimarina* species have as well been implicated in the emergence of epizootic shell disease (ESD) in the American lobster (*Homarus americanus*). ESD is a dermal disease that is causing major losses to the lobster fishery in Southern New England (Zhou *et al.*, 2015). The yellowish lesions on the lobster carapaces have been reported as the isolation source of *Aquimarina* sp. I32.4 (Quinn *et al.*, 2012; Ranson *et al.*, 2018) and *Aquimarina muelleri* (Chistoserdov *et al.*, 2005), which have been suggested to play major roles in ESD development. These strains are also able to degrade crude chitin, which is thought to be directly correlated with the disease (Chistoserdov *et al.*, 2005). In fact, chitinase-encoding genes or chitin degradation ability have already been reported for various *Aquimarina* species (Donadio *et al.*, 2007; Hudson *et al.*, 2019). Thus, chitin breakdown may be a key feature to better understand the ecology of these organisms since it bears relevance not only for their potential opportunistic behaviour but also for biogeochemical cycling in the oceans. In line with this observation, *Aquimarina* strains AD1 and BL5 have been identified by Kumar et al. (V. Kumar *et al.*, 2016) as etiological agents of bleaching in the red alga *Delisea pulchra.* Their broad polysaccharide degradation profile has been suggested as a pivotal feature in their opportunistic interaction with the host (Hudson *et al.*, 2019).

Little was known about the secondary metabolite biochemistry of *Aquimarina* spp. until the presence of polyketide synthase (PKS)-encoding genes was for the first time reported, based on gene homology (PCR) screenings, in 10 out of 11 strains isolated from Irciniidae sponges in the North Atlantic (Esteves *et al.*, 2013). The retrieved PKS gene sequences moderately resembled those involved in the biosynthesis of the anti-tumor type-I polyketides onnamide

and bryostatins. This finding indicated possible cytotoxic capacities for an otherwise unsuspected bacterial taxon given its placement in the *Bacteroidetes* phylum, most renowned for its role in carbon cycling and whose genomes had been considered to present low numbers of PKS and non-ribosomal peptide synthetase (NRPS)-encoding genes (Donadio *et al.*, 2007). In recent years, the presence of such genes is being consistently detected in different *Aquimarina* strains via genome sequencing (Keller-costa *et al.*, 2016; Ranson *et al.*, 2018; Hudson *et al.*, 2019). Moreover, structure elucidation of cuniculene, the first *trans*-AT polyketide ever reported from *Aquimarina*, has just been achieved (Helfrich *et al.*, 2019). In this context, this study benefits from recent advances in computational biology and *in silico* genome mining (Medema *et al.*, 2015; Blin *et al.*, 2017; Medema, 2018; Navarro-muñoz *et al.*, 2018) to examine the abundance, diversity, phylogenetic relationships and co-occurrence networks of secondary metabolite biosynthetic gene clusters (BGCs) across all publicly available *Aquimarina* genomes as of February 25 2019, substantially enlarging the known genetic space underlying natural product biosynthesis within the *Bacteroidetes* phylum. Moreover, we use a suite of phylogenomics measures to deliver the first comprehensive genome-wide phylogeny for the genus, determine core- and pan-genomes, and link evolutionary trajectories with functional genomics and the secondary metabolite machinery within the group.

Results

Genome overview

The 26 non-redundant *Aquimarina* genome assemblies examined in this study have been retrieved from marine sponges (n=9), red algae (n=5), gorgonian coral (n=1), lobster shell

(n=1), seawater (n=8) and sediments (n=2) (Table 1). Thirteen have been taxonomically identified at the species level upon submission by authors, while the remaining 13 strains have been classified at genus level. Whereas 16 strains were isolated from eukaryotic hosts (host-associated source - HA), 10 strains were isolated from seawater or sediments (free-living source - FL). Genome sizes ranged from 4.07 Mb (*Aquimarina agarivorans*) to 6.50 Mb (*Aquimarina* sp. AU119), with an average length of 5,60 Mb. The GC content of all strains averaged 32.72% and ranged from 31.4% (*Aquimarina muelleri*) to 35.9% (*Aquimarina spongiae*). The number of coding sequences (CDSs) per genome ranged between 3,624 (*Aquimarina agarilytica)* and 5,800 (*Aquimarina* sp. AU119) (Table 1).

The core and pan genome of the 26 *Aquimarina* strains consisted of 1,226 and 21,211 CDSs, respectively, with a core/pan genome ratio of 0.058 (Fig. 1). While the observed pan-genome did not reach an obvious plateau, the observed core genome, in contrast, stabilized after the 10^{th} genome iteration. It is therefore reasonable to argue that the core genome captured in this study may be a close approximation of the common features of the genus, even though not all formally described species could be included in this assessment due to the lack of reference genomes.

Phylogeny and Taxonomy

A high degree of 16S rRNA gene relatedness was revealed between the species *A. atlantica* and *A. megaterium*, with *A. macrocephali* clustering with the former species to form a well-supported clade within the genus (Supporting Information Fig. S1 and Dataset S1). With exception of *Aquimarina* sp. Aq107, closely related to *A. latercula* and *A. versatilis*, all *Aquimarina* strains retrieved from marine invertebrates (sponges and corals) in the Northeast

Atlantic (Esteves *et al.*, 2013; Keller-costa *et al.*, 2016; Keller-Costa *et al.*, 2017) affiliated with the *A. atlantica-megaterium-macrocephali* 16S rRNA gene clade depicted here, which also included several strains retrieved from seawater (Fig. S1). The tree suggests closest affiliation of reported pathogenic strains I32.4 (lobster, (Quinn *et al.*, 2012; Ranson *et al.*, 2018)), AD1 and BL5 (red alga, (V. Kumar *et al.*, 2016)) to *A. macrocephali*, *A. latercula/versatilis* and *A. amphilecti*, respectively (Fig. S1). *A. agarilytica* and *A. agarivorans* formed a well-supported cluster apart from all other *Aquimarina* species. The 26 genomes targeted in our study, although not representing all 28 validly described species, were distributed throughout the full extent of the current *Aquimarina* phylogeny, with each major 16S rRNA gene clade within the genus being represented by at least one genome (Fig. S1).

A Maximum Likelihood phylogenomic tree was inferred using the nucleotide sequences of all core CDSs (n = 1,226) found throughout the 26 genomes, showing overall congruency with patterns of phylogenetic relationships depicted via sole 16S rRNA gene-based analyses (Fig. S1), albeit higher accuracy in establishing finer evolutionary divergencies between strains (Fig. 2A). For instance, the close relatedness between *A. megaterium*, *A. atlantica* and *A. macrocephali* species and several sponge- and coral-associated *Aquimarina* strains is reinforced here, now with consistent evidence for closer proximity between strains AU58, Aq349 and EL33 to *A. megaterium* than *A. atlantica*. Likewise, the relatedness between strains Aq135 (marine sponge) and I32.4 (lobster pathogen) suggested by 16S rRNA gene phylogeny was also depicted via phylogenomics, which further revealed that both strains form their own unique phylogenomic cluster, apart from *A. macrocephali*. Congruent with 16S rRNA gene phylogeny, pathogenic strains AD1 and BL5 were closely related to *A. latercula*

and *A. amphilecti*, respectively, and *A. agarilytica* and *A. agarivorans* presented the highest evolutionary divergencies to all other strains in the genus, consistent with ANI estimates (Supporting Information Dataset S2). Indeed, both these species shared less than 70% ANI with all the other strains in the genus, while ANI values for all other pairwise comparisons were always and clearly above this threshold (Supporting Information Dataset S2).

A suite of five phylogenomic measures was employed to attempt species-level classification of non-type strains present in the dataset. Table 2 provides an overview of results (1) obtained for reportedly pathogenic strains AqI32.4, AD1 and BL5, (2) where affiliations to formally-described species were deemed plausible and (3) where close relatedness within strains currently not classifiable as a valid species was evident, based on consensual thresholds of ANI > 95% and DDH probabilities ≥ 70% (Klappenbach *et al.*, 2007; Richter and Rossello-Mora, 2009; Meier-Kolthoff *et al.*, 2013). The strongest correlation between genomes across the data was found between *Aquimarina* strains Aq349 and EL33. Strain AU58 was also close to both Aq349 and EL33, with phylogenomic metrics indicating that all three host-associated strains could belong to the same species. Although ANI values retrieved between these strains and type strain *A. megaterium* XH134^{T} were borderline to their inclusion as members of the same species, in all cases the estimated DDH probabilities were far below the threshold acceptable for same-species diagnosis. Phylogenomic measures firmly affiliated strain AD10 with the species *A. aggregata* and strain AD1 to *A. latercula* (Table 2). Besides the comparisons highlighted above, solid affiliations of strains to validated species/type strains using *Aquimarina* genomes currently available were not possible (for details, see “Extended Discussion” in Supporting Information).

Functional genomics

Overall, about 52 to 58% of the CDSs predicted within *Aquimarina* genomes could be assigned a function on the RAST annotation platform (Table 1), supporting the notion of *Aquimarina* species as understudied organisms with yet uncharted metabolic pathways. The full genetic machinery required for gliding motility among the *Flavobacteriaceae* (genes *gldA* through *gldN*) was a remarkable feature of the core *Aquimarina* genome (Supporting Information Dataset S3). Mutations in *gld* genes are known to lead to deficient chitin utilization among *Flavobacteriaceae* spp. (Braun *et al.*, 2005), an observation supported by the recent discovery of a type IX secretion system (T9SS) confined to the Fibrobacteres-Chlorobi-Bacteroidetes superphylum, which is responsible for both chitinase export and gliding motility regulation (Kharade and McBride, 2014; Lauber *et al.*, 2018). In line with these findings, inspection of the predicted proteome of the 26 *Aquimarina* strains revealed the shared presence of several proteins containing T9SS C-terminal domains involved in targeted protein translocation (Supporting Information Dataset S4).

We delved into COG and Pfam-based annotations to further explore specificities and commonalities among the *Aquimarina* genomes. Totals of 2,320 COG entries were retrieved across all 26 genomes with 1,024 (core) COGs common to all strains, while 4,187 Pfam entries were assigned in the final dataset, with 1,130 (core) Pfams common to all genomes (Supporting Information Dataset S5-S7). All genomes displayed an even pattern of CDSs allocation across coarse metabolic functions, with unexpectedly high numbers of CDSs assigned to the COG class “Secondary metabolites biosynthesis, transport and catabolism” (Class Q, Supporting Information Dataset S5). Inspection of Pfam profiles revealed up to 24

glycoside hydrolase (GH) families common to all *Aquimarina* strains. For instance, potential chitinase (GH18), cellulase (GH5), agarase, carrageenase or porphyranase (GH16), xylanase (GH10, GH11) and N-acetylglucosaminidase activities were revealed, uncovering a truly versatile carbon metabolism for the group particularly with respect to the degradation of polysaccharides (Supporting Information Dataset S7). Exploration of Pfam and COG profiles enabled us to further identify numerous "symbiosis factors" in the core *Aquimarina* genome. These included a range of eukaryotic-like protein motifs (ELPs) such as WD-40 repeats, leucine-rich repeats (LRR), tetratricopeptide repeats (TRPs) and ankyrin repeats, known to play fundamental roles in host-microbe interactions, which could be identified across all genomes in varying degrees of abundance and diversity depending on the database inspected (Table 3, see also Supporting Information Dataset S6 and S7 to access all results). Likewise, manifold genomic features underpinning the biosynthesis of vitamins such as thiamin (B1 - COG0352), riboflavin (B2 - COG0054, COG0307, COG1985), nicotinic acid (B3 – COG1057), pyridoxin (B6 – COG0259), biotin (B7 – COG0340, COG0502, COG0511) and cobalamin (B12 – COG0368) were identified in the core *Aquimarina* genome, and could be involved in host nutrient provision processes.

In agreement with a recent permutational analysis in which no signature CDSs could be unequivocally identified for host-associated *Aquimarina* strains (Díez-Vives *et al.*, 2018), ordination of both COG (Fig. 3) and Pfam profiles in our study revealed no correlation between isolation source (free-living vs. host-associated) and functional genome clustering among the 26 strains examined. Instead, three robust genome groups could be clearly identified via both cluster analysis (node support ≥ 99% or all three groups, Fig. 2B) and ordination of COG profiles (one-way PERMANOVA F=7.638, p=0,0001; pairwise group

divergencies: $p < 0.0014$, Fig. 3A). The identified genome groups were found to display varying degrees of congruency with the taxonomic position / phylogenetic relationship (Fig. 2A) and with the inferred secondary metabolism (Fig. 4) of the strains. Hereafter, we refer to these groups as the *A. muelleri-longa* (Group 1), *A. megaterium-macrocephali-atlantica* (Group 2) and *A. aggregata-amphilecti-latercula* (Group 3) functional genome groups to examine their distinguishing features and the correspondence between COG-based functional annotation and secondary metabolism prediction. Although the species *A. agarivorans* and *A. agarilytica* clearly formed a separate genome cluster (Figs. 2B and 3), for the purposes of this study they were not considered a major "functional group" as Groups 1-3 due to their low sample size (n=2 genomes), which precludes proper use of multivariate statistical assessments to test for differences between groups.

Two COG entries involved in natural product biosynthesis were among the top-ten COGs most contributing to distinguish between functional genome groups 1-3 according to SIMPER analysis (Fig. 3B). The most differentiating COG entry (COG3321) was an acyl transferase domain of polyketide synthase (PKS) enzymes, found to be enriched in the *A. megaterium-macrocephali-atlantica* group (Group 2), whereas COG1020 (non-ribosomal peptide synthetase component F) ranked as the sixth most differentiating COG entry, being enriched in the *A. muelleri-longa* group (Group 1). An eukaryotic-like leucine-rich repeat (LRR) protein motif (COG4886), a genomic feature usually enriched in the marine sponge microbiome (Karimi *et al.*, 2017), was the second most differentiating COG entry among groups, being enriched in *A. megaterium-macrocephali-atlantica* Group 2. More abundant in this group was also the two-component sensor histidine kinase LytS (COG3275) / DNA-binding response regulator LytR (COG3279) system, a major regulator of the cell's

physiological status. Entries related with carbon metabolism were as well present in the list of top-differentiating COGs, including a beta-glucanase (COG2273) (least abundant in Group 2) and a chitodextrinase (COG3979) (most abundant in Group 2), involved in the catabolism of complex macromolecules. Finally, most abundant in the *A. aggregata-amphilecti-latercula* group (Group 3) was a COG classified as "Fasciclin 1 (Fas1)" (COG2335), implicated in the process of cell adhesion and signaling, while an uncharacterized protein of the type VI secretion system (T6SS) (COG3501), known for its participation in a variety of functions such as virulence, antibacterial activity and metal ion uptake, was enriched in *A. muelleri-longa* Group 1.

Identification of Biosynthetic Gene Clusters

A total of 928 BGCs from the 26 *Aquimarina* genomes was predicted with antiSMASH. These included 54 terpene, 13 type 1 PKS-NRPS, 21 *trans*-AT PKS, 24 type 3 PKS, 39 NRPS, 12 siderophore and 16 bacteriocin BGCs, among others, along with 657 putative BGCs documented with the ClusterFinder (cf) function (400 "cf putative", 154 "cf saccharide", and 121 "cf fatty acid", Supporting Information Dataset S8). Of the 928 identified BGCs, 180 BGCs corresponded to MIBiG entries (Supporting Information Dataset S9) while 748 did not display enough homology to MIBiG entries, implying that much of the inferred secondary metabolism within the genus remains uncharted and awaits experimental verification. Figure 4 provides an overview of BGC counts across the *Aquimarina* genomes - excluding saccharide, fatty acid and putative BGCs for the sake of simplicity and to give emphasis on BGCs more likely to underpin the biosynthesis of drug-like candidates -, illustrating overall congruency between coarse-level BGC profiling as assessed via

antiSMASH and COG-based genome clustering. Overall, the presence of terpene BGCs was a common feature of all the *Aquimarina* genomes, ranging from 1 to 3 per strain, while siderophore BGCs were present in almost all strains (24 out of 26). Likewise, polyketide synthase (PKS) BGCs, represented in the classes type I PKS-NRPS, *trans*-AT PKS and type III PKS, were present in all the *Aquimarina* genomes except in *A. agarilytica* and *A. agarivorans*. Again, these latter species stood out from the rest due to their poorer diversity of classifiable BGCs - saccharide and fatty acid BGCs aside (Fig. 4; for more details on other BGCs, including putatives, see Supporting Information Dataset S8). Also, a high number of NRPS BGCs could be detected across the genomes, especially among strains in the "*A. muelleri-longa*" group (Group 1) where five NRPS BGCs per genome were detected on average. The "*A. megaterium-macrocephali-atlantica*" group (Group 2) possessed diverse BGCs in general, being enriched in bacteriocin BGCs in comparison with Groups 1 and 3.

The BiG-SCAPE algorithm grouped the BGCs predicted with antiSMASH into 108 saccharide, 48 NRPS, 19 ribosomally synthesized and post-translationally modified peptides (RiPPs, which include bacteriocins and lantipeptides), 19 "PKSother", 13 terpene, 13 type I PKS, 10 PKS/NRPS and 209 "other" and non-classifiable 'Gene Cluster Families' (GCFs) (Fig. 5). Most of these families were constituted by one single BGC ("singletons", Fig. 5A). Hereafter, only GCFs composed by ≥ 2 BGCs are further considered (Supporting Information Dataset S10), and patterns of occurrence of GCFs in the class "PKSother" across the genomes are exemplarily illustrated (Figs. 5B and 5C). "PKSother" families (19 GCFs in total) were grouped into five "gene cluster clans" (GCCs), with clans "A" and "B" being the largest both in number of families and of individual BGCs. Particularly interesting was the partitioning of GFCs within clan A among members of *Aquimarina* functional groups. The

BGCs from FAM_00755 (marked in red within "clan A"), all belonged to strains within the "*A. megaterium-macrocephali-atlantica*" group (Group 2) whereas BGCs from FAM_00570 (marked in pale blue within "clan A") were prevalently detected in genomes within the *A. aggregata-amphilecti-latercula* Group (Group 3) (Fig. 5B). All BGCs within "PKSother" clan A possessed similar sizes and resembled type III PKSs as classified by antiSMASH, but no significant homologies with known BGCs were found. Within PSKother clan B, two closely related families were identified, with the largest family (FAM00536 in Data S10, marked in light blue in Fig. 5B) containing seven BGCs exclusive to strains within the *A. megaterium-macrocephali-atlantica* group. These BGCs underly the assembly of the newly-discovered *trans*-AT polyketide cuniculene (Helfrich *et al.*, 2019), whose bioactivity is yet to be established. Taken together, results on GFC partitioning within "PKSother" clans A and B provide evidence for specialized natural product biosynthesis and conservation of BGCs within a narrow, phylogenetically coherent range of species within the *Aquimarina* genus. Strengthening this observation are several other examples of fidelity in patterns of GCF distribution across functional genome groups 1-3, which were highly consistent for other compound classes such as NRPSs and RiPPs (Data S10). For instance, although the "*A. muelleri-longa*" group (Group 1) was numerically enriched in NRPS BGCs, especially in *Aquimarina longa* (n=9) and *Aquimarina muelleri* (n=7) (Fig. 4), in terms of diversity we found four NRPS clans each presenting clear-cut patterns of fidelity to one of the functional groups 1, 2 or 3 (two clans confined to Group 3, one clan to Group 2 and one clan to Group 1, Data S10). Although none of the NRPS BGCs from the clans above displayed considerable homology with known NRPS BGCs (Data S10), a wealth of tentative structures could be retrieved via antiSMASH (Fig. 4B, Supporting Information Fig. S2). Likewise, four RiPPs

clans (B through E) underlying the biosynthesis of bacteriocins or lantipeptides showed fidelity to either functional groups 2 or 3 (Data S10). Again, no BGC described so far shared significant homology with the RiPPs clans above. Regardless of levels of homology with already known BGCs, ClusterBlast searches revealed that similar clusters to those in "PKSother", NRPS and RiPPs clans (here studied in more detail) exist among the *Bacteroidetes* and/or *Flavobacteriaceae*. All closest hits obtained for *Aquimarina* BGCs in the PKSother class originated from *Flavobacteriaceae* species. A similar pattern was observed for BGCs in the RiPPs class, except for BGCs in clan B where closest hits derived from a cyanobacterium (*Microcystis aeruginosa*). In contrast, NRPS BGCs exhibited the most promiscuous spread across unrelated taxa, as closest hits to three out of four *Aquimarina* NRPS clans derived from *Gammaproteobacteria* / *Firmicutes* genomes (Data S10).

Discussion

This study revealed an expanding pan-genome in contrast with a well-defined and "stable" core genome for the set of *Aquimarina* genome sequences currently available. At the sub-species level, "open" pan-genomes are considered indicators of high horizontal gene transfer (HGT) rates, especially if the studied organisms live in multiple environments with complex microbial communities (Tseng and Tang, 2014). Although our results suggest aptitude of *Aquimarina* spp. to thrive in multiple and complex microniches (see below), it is presently not possible to examine pan vs. core genome dynamics within individual *Aquimarina* species due to the lack of available genomes, making assumptions on gene transfer amenability within species difficult. Instead, our analyses placed focus on patterns of gene gain and loss above the species level, suggesting that the *Aquimarina* pan-genome is certainly to grow if genomes

from e.g. the remaining 16 validly described *Aquimarina* species are included, while the extent to which intra-species genotypic variation contributes to pan-genome expansion remains to be revealed. Similarly, core / pan genome ratios and several phylogenomic metrics (ANI, AAI, DDH) have been used recently as a molecular-based aid in the delineation of species boundaries within prokaryotes (Klappenbach *et al.*, 2007; Richter and Rossello-Mora, 2009; Meier-Kolthoff *et al.*, 2013; Caputo *et al.*, 2015) but have been rarely employed to inspect or delineate thresholds at higher taxonomic ranks. Core / pan genome ratios within species may vary widely according to sample size (i.e. number of genomes), degree of genotypic relatedness of the strains in the analytical dataset and taxon-intrinsic factors such as HGT amenability, with values ranging from 0.94 for closely-related strains of the human pathogen *Klebsiella pneumoniae* (Caputo *et al.*, 2015), 0.26 for uncultivated *Synechococcus spongiarum* (*Cyanobacteria*) symbionts of marine sponges (Burgsdorf *et al.*, 2015), and 0.11 for a diverse panel of 20 strains of the promiscuous human commensal *Escherichia coli* (Touchon *et al.*, 2009). Consistent with genome-wide variability above the species level, the core / pan genome ratio observed here for *Aquimarina* spp. (0.058) was below the abovementioned intra-species values and above ratios calculated for more diversified genome groups such as uncultivated marine sponge symbionts of the *Rhodospirillalles* order, found to span different genera and to present a core / pan genome ratio of 0.026 (Karimi *et al.*, 2018). Phylogenomics metrics were further used in this study to attempt the identification of *Aquimarina* strains to the species level, revealing cases where the emergence of novel species could either be proposed or become matter of debate. The latter is apparently the case of host-associated *Aquimarina* strains Aq349, AU58 and EL33. Although it could be hypothesized that an evolutionary process is in place involving the emergence of a novel species or

subspecies within *Aquimarina megaterium*, eventually related with a specialization event (e.g. adaptation to a particular host-associated microniche), much larger genome collections would be needed to solidly identify the genotypic traits underpinning this process. Finally, we gathered compelling molecular-based evidence, from 16S rRNA gene phylogeny through phylogenomics, functional genomics and secondary metabolism inference, for a solid differentiation between *Aquimarina agarilytica* and *Aquimarina agarivorans* and the remaining *Aquimarina* strains / species, suggesting that the former two species may constitute a different/new genus on its own, or else that comprehensive phenotypic and DNA-DNA hybridization studies (Kim *et al.*, 2014) are necessary to firmly corroborate their placement in the *Aquimarina* genus. Altogether, besides providing a timely snapshot of the phylogenetic relationships within *Aquimarina* species, the genome-wide approach employed in this study indicates that much improvements to our understanding of the genotypic diversity, evolution and taxonomy of the group are to be made as more genome sequences populate the *Aquimarina* tree both at the intra- and inter-species levels.

Host-associated microorganisms, particularly obligate symbionts, usually have reduced genome sizes due to the loss of non-essential genes in the host-associated habitat (McCutcheon and Moran, 2012; Nayfach *et al.*, 2019). Consistent with generalist, "bi-modal" host/free-living life-strategies (Van Elsas *et al.*, 2011; Karimi *et al.*, 2019) and a complex nutrient and secondary metabolism, *Aquimarina* genomes were prevalently large usually exceeding 5.0 Mb (Table 1). Further, genomes from host-associated habitats did not present significantly different size or number of predicted genes in comparison with genomes from free-living habitats. In general, *Aquimarina* genome sizes are larger than those reported for other marine *Flavobacteriales* genera such as *Polaribacter*, *Dokdonia*, *Gramella*,

Leeuwenhoekiella or *Formosa*, usually in the 2.9 – 4.2 Mb range (Fernández-Gómez *et al.*, 2013; Mann *et al.*, 2013), and the extent to which this trend could result from higher BGC densities within *Aquimarina* genomes remains to be determined. Our assessment of functional traits across the studied genomes unveiled several core functions presumably conferring *Aquimarina* spp. the ability to colonize and persist in both host- and particle ("free-living")-associated settings. Simultaneously, a signal for functional divergence among phylogenetic groups, including the emergence of different secondary metabolism profiles across these groups, was depicted that could underlie specialization of *Aquimarina* clades or species to (yet undetermined) distinct microniches.

Among the shared functions, noteworthy was the presence of several glucoside hydrolases (GH) families in the core *Aquimarina* genome, highlighting their likely importance as major mediators of carbon cycling in the oceans. This is consistent with accumulating, genome-based evidence for a primary role of marine *Bacteroidetes* in the breakdown of high molecular weight biopolymers, particularly in surface-dense microhabitats (Fernández-Gómez *et al.*, 2013; Hudson *et al.*, 2019). Hudson et al. (Hudson *et al.*, 2019) have recently provided a mechanistic understanding of the carbohydrate degradation capacities of strains AD1, AD10 and BL5 and the likely roles of carbon metabolism in the association with their algal host. Here we reveal that the ability to degrade major compounds usually present in the cell wall of, or exuded by, ubiquitous marine hosts such as phytoplankton/algae/ and zooplankton/crustaceans, or present in marine snow, recently-lysed cells, detritus or excretions of live organisms, is a common attribute of the *Aquimarina* genus being shared by all species and strains inspected so far. We therefore posit that this shared attribute functions as a pivotal mechanism underpinning the generalist pattern of occurrence of these organisms

in the seas as reflected by their several source habitats. The presence of the full genetic machinery required for T9SS-dependent chitinase export and gliding motility across all *Aquimarina* genomes suggests furthermore a conserved substrate-specific mechanism of sensing, scavenging and breakdown of chitin by *Aquimarina* spp. that emerges as a likely delineating feature of the genus, making these species important mediators of ocean carbon and nitrogen cycling. Although the ability to degrade chitin has been reported in some *Aquimarina* species descriptions (Yu *et al.*, 2014; Ranson *et al.*, 2018) and might be correlated with pathogenicity in specific cases (Quinn *et al.*, 2012; Ranson *et al.*, 2018), dedicated studies of the diversity of chitinase-encoding genes, chitinase kinetics under varying physical-chemical conditions and chitinase purification and description from *Aquimarina* strains remain wanting. Such studies may be of interest in biotechnology-oriented research owing to the multiple potential applications of chitinases in several sectors such as the food industry and biomedicine (Younes and Rinaudo, 2015; Ilangumaran *et al.*, 2017).

Besides their ability to degrade multiple carbon sources, other features shared by *Aquimarina* species may decisively contribute to their overall particle- and host-associated aptitude. Notably, diverse eukaryotic-like proteins including most-studied WD-40 repeats, tetratricopeptide repeats (TRPs), ankyrin repeats and leucine-rich repeats (LRR) could be found in all genomes. These ELPs have been consistently diagnosed as enriched genomic signatures of the marine sponge microbiome (Thomas *et al.*, 2010; Ilangumaran *et al.*, 2017; Karimi *et al.*, 2017), and as such they may play a deterministic role, involving host-microbe and/or microbe-microbe molecular interactions, in the establishment of marine sponge-microbiome associations. Particularly, ankyrin repeats are known to enable their carrying bacteria to evade phagocytosis by amoeba *in vitro*, functioning as a possible key mechanism

contributing to the persistence of symbionts within marine sponges (Nguyen *et al.*, 2014) and, eventually, other eukaryotic hosts. It is noteworthy that the abundance of ankyrin repeats detected across genome Groups 1 to 3 in this study (COG0666: average of > 8.5 CDSs per genome, ranging from 3 up to 17 CDSs) by far exceeded that observed previously for a diverse group of culturable, sponge-associated *Alphaproteobacteria* (COG0666: average of 1 CDS per genome, ranging from 0 up to 3 CDSs) (Karimi *et al.*, 2019), indicating that *Aquimarina* spp. are well-equipped to thrive in association with filter-feeding invertebrates such as sponges or in other densely-populated habitats such as sediments where ELP abundances are higher than in seawater (Karimi *et al.*, 2017). Further, the biosynthesis of inhibitory secondary metabolites by microorganisms is considered a decisive factor in microbe-microbe competitive interactions or host-microbe beneficial relationships whereby the symbiotic producer aids the host with chemical defence against natural enemies (Lopanik *et al.*, 2004). Coarse-level BGC assignment with antiSMASH enabled us to ascertain the potential biosynthesis of several structurally diverse compound classes such as polyketides, terpenoids, non-ribosomal peptides and bacteriocins as a common feature across all (or nearly all) *Aquimarina* genomes, which again would confer these species with competitive capacities in highly dense microbiomes such as those from high abundance marine sponges, sediments, or phytoplankton blooms.

However, in-depth and fine-grained homology-based analysis of BGCs unveiled differences in secondary metabolite profiles among strains beyond the limited granularity of BGC detection methods employed in most studies. Several gene cluster families or clans within the major compound classes carefully examined here (PKSother, NRPSs and RiPPs) were confined to specific functional genome groups 1-3 as defined by COG profiling, and COG

entries involved in secondary metabolism ranked among the most differentiating functions between groups. We argue that such a pattern of differentiation in secondary metabolite biosynthesis - among other co-varying factors including predatory and competitive interactions, substrate affinity of catalytic enzymes, genetic drift and habitat heterogeneity at the microscale (Stocker and Seymour, 2012; Teeling *et al.*, 2012), has implications to the diversity and coexistence of *Aquimarina* species in marine ecosystems.

PKSother clan A, for instance, possessed three GCFs of which one was found exclusively in members of functional group 2 (*A. megaterium-macrocephali-atlantica*). All BGCs in the clan encode for type III PKSs, well-known for their immense potential to synthesize small compounds with a high structural and bioactivity diversity (46). Interestingly, within PKSother clan B are the BGCs involved with the biosynthesis of a novel polyketide named cuniculene, recently described for *Aquimarina* sp. Aq349 (Helfrich *et al.*, 2019). Cuniculene has a *trans*-AT PKS and ladderane-type structure, the same class that was found, in this study, in the antiSMASH predictions for *Aquimarina* strains Aq349, AU58 and EL33. Highly similar BGCs were, moreover, found for the other strains in functional group 2, namely *Aquimarina* sp. Aq78 and type strains *A. megaterium* XH134^{T} and *A. atlantica* 22II-S11-z7^{T}, hinting at the existence of conserved BGCs encoding for cuniculene and possibly cuniculene-like compounds within the group. Given its recent structural elucidation and purification in the laboratory, it is likely that bioactivities for cuniculene and structurally related compounds will soon be revealed. As polyketides, non-ribosomal peptides (NRPs) are of utmost interest for the field of drug discovery and their potential biosynthesis has been often reported from symbiotic microbial communities (Agrawal *et al.*, 2016; Raimundo *et al.*, 2018). Among NRPs are more than 20 marketed drugs with antimicrobial, antitumoral and

immunosuppressant activity (47), encouraging further studies of compounds from diverse microorganisms and isolation sources. Bacteriocins, in their turn, are ribosomally synthesized peptides produced by bacteria which can kill or inhibit bacterial strains closely related or non-related to the producer. This class is thus considered a possible target for the development of new antibacterial compounds (Cotter *et al.*, 2013). In fact, antibacterial (i.e. *Staphylococcus aureus*) activity was previously described by our team for a dataset of 11 *Aquimarina* strains cultured from *Irciniidae* marine sponges (Esteves *et al.*, 2013). These results suggest that further mining of the secondary metabolism of *Aquimarina* species may result in novel antibacterial activities, with implications to research on alternative modes of action in the current scenario of bacterial pathogens resistant to multiple antibiotics presently in use. At least two scenarios could be evoked to explain the pattern of occurrence of BGC families and clans across specific *Aquimarina* genomes unveiled in this study: either a true phylogenetic signal in which the presence of conserved BGCs across *Aquimarina* species results from consistent evolutionary trajectories within the *Flavobacteriaceae*, which is apparently the case of PKSother clans and nearly all (except one) RiPPs clans documented here; or HGT events involved in BGC acquisition from (or delivery to) unrelated species, which seems plausible for BGCs from most NRPS clans here delineated.

Natural products and natural product-derived small molecules are still the most important source and inspiration for modern chemotherapeutics. One of the most preeminent ecosystems as a reservoir of new compounds is the marine environment where bacteria, often associated with eukaryotic hosts such as sponges, corals and algae usually display a striking secondary metabolism (Raimundo *et al.*, 2018). In this context, this study showcases unprecedented, intriguing biosynthetic potential for the *Aquimarina* genus, highlighting its likely use as a

future source of biotechnological appliances and widening the natural product coding spectrum of the *Bacteroidetes*, a quintessential bacterial phylum in marine ecosystems. We uncovered several examples of clade-specific natural product biosynthesis capacities which underpin a tight coupling between phylogeny, ecology and secondary metabolism within the group, providing a glimpse on much novel structural diversity that may be harnessed through dedicated metabolome mining of *Aquimarina* spp. Our findings furthermore highlight the importance of symbiont cultivation attempts as a pivotal platform to the discovery of novel microbial-derived molecules and the genetic machinery dictating their biosynthesis. Potential for polyketide and NRP production from symbiotic consortia, particularly, has been deeply investigated by cultivation-independent attempts for the past 20 years or so, leading to numerous and exciting new findings (Piel, 2002; Piel *et al.*, 2004; Wilson *et al.*, 2014; Trindade *et al.*, 2015). However, the general uncultivability of the studied symbionts have been a major hindrance to effective metabolite harvesting from these complex communities. The ability to cultivate understudied or less-represented bacteria opens new opportunities to the study of novel metabolites and their biosynthetic genes as demonstrated by the recent discovery of cuniculene from *Aquimarina* spp. (25).

In conclusion, our comprehensive genome-driven strategy unveiled key functional traits underlying the generalist pattern of occurrence of *Aquimarina* spp. across multiple marine habitats including microbial, plant and animal hosts besides seawater and sediments. It sheds light on the carbon turnover versatility and complex secondary metabolism of a recently described taxon of increasing relevance to our understanding of the functioning of marine ecosystems.

Experimental Procedures

This study comprises the comparative analysis of 26 *Aquimarina* genomes available in the NCBI database (NCBI Resource Coordinators, 2016) on February 25th 2019. Of these genomes, five belong to an in-house collection of symbionts of octocorals and marine sponges: *Aquimarina* sp. EL33, Aq78, Aq107, Aq135 and Aq349 (Esteves *et al.*, 2013; Keller-costa *et al.*, 2016; Keller-Costa *et al.*, 2017) and the remainder originate from seawater, marine sediments and algae (Table 1). Strain Aq349 is the source of the recently-described *trans*-AT polyketide cuniculene (Helfrich *et al.*, 2019). Other notable strains analysed in this study are the above-mentioned ESD causing agent strain I32.4 (Quinn *et al.*, 2012; Ranson *et al.*, 2018) and macroalgal bleaching agents *Aquimarina* sp. AD1 and BL5 (V. Kumar *et al.*, 2016). Fasta files containing the contig sequences of each genome were downloaded from NCBI along with their accompanying metadata. *Aquimarina* strains were further labeled as free-living (FL) or host-associated (HA) depending on their isolation source. Prediction of coding sequences (CDSs) was performed using the RAST (Rapid Annotation using Subsystem Technology) prokaryotic genome annotation server (version 2.0) using the "classic RAST" algorithm (Overbeek *et al.*, 2014) and the basic genome features of all strains were compiled (Table 1).

Phylogenetic and phylogenomic assessments

A 16S rRNA gene phylogeny was inferred for the *Aquimarina* genus. Briefly, the tree comprises 16S rRNA gene sequences of the 26 strains thoroughly analysed in this study, of their closest 16S rRNA gene relatives identified by BLAST and of all type strains of accepted *Aquimarina* species (Supporting Information Table S1). All sequences were downloaded from

NCBI and aligned using ClustalW within the software package MEGA7 (S. Kumar *et al.*, 2016). Evolutionary distances were calculated using the GTR+G+I model, predicted as the best-fitting model to the dataset, and the Maximum Likelihood method was employed for tree construction. Bootstrapping tests of phylogeny were performed with 1,000 repetitions.

For genome-wide assessments of phylogeny, all 26 genome sequences were uploaded to the web server "Efficient Database framework for comparative Genome Analyses using BLAST score Ratios" (EDGAR) 2.0 (Blom *et al.*, 2016), where singleton genes, core and pan genomes, average amino-acid and nucleotide sequence identities (AAI and ANI, respectively) were determined following standard procedures (Blom *et al.*, 2016). The web server JSpeciesWS (Richter *et al.*, 2015) was also used to determine ANI values as a means of comparison. The JSpeciesWS platform allows pairwise ANI comparisons based on BLAST+ (ANIb) and on MUMmer (ANIm) algorithms (Kurtz *et al.*, 2004; Klappenbach *et al.*, 2007). In addition, digital DNA-DNA hybridization values were calculated for each pair of strains using the Genome-to-Genome Distance Calculator (GGDC) (Meier-Kolthoff *et al.*, 2013). Finally, a phylogenomic tree was constructed using FastTree on EDGAR based on the nucleotide sequences from all protein-encoding genes common to the 26 genomes (core CDSs, n = 1,226). Sequence alignments were performed using MUSCLE and concatenated to one large multiple alignment, after which evolutionary distances were calculated with the Kimura-2 parameter model and tree construction was carried out with the Maximum Likelihood method, with local support values calculated with 250 iterations using the Shimodaira-Hasegawa test.

Annotation and functional genomics

Amino acid fasta files retrieved from RAST were used as input data for functional annotations based on Clusters of Orthologous Groups of Proteins (COGs) and Protein families (Pfams), performed with the webserver WebMGA (e-value = 0.001) (Sitao Wu, Zhengwei Zhu, Liming Fu *et al.*, 2011). Using custom scripts written in R, contingency tables containing all COG / Pfam entries per genome, with the respective number of CDSs per genome assigned to each entry, were created. Quantitative functional comparisons between the genomes were carried out using COG and Pfam annotations after Hellinger transformation of the profiles (i.e. square root calculation of the relative abundance of each COG / Pfam entry in a given genome). To test whether functional genome groups exist and whether they are determined by the organisms' source habitat (host-associated or free-living) or phylogenetic relationships, ordination of COG and Pfam profiles was performed using Principal Components Analysis (PCA) on Bray-Curtis dissimilarity matrices calculated with STAMP (Parks *et al.*, 2014). One-way permutational analysis of variance (PERMANOVA) was employed thereafter with Past v3.0 (Hammer *et al.*, 2001) to determine whether groups visualized after PCA (if any) were statistically significant. In addition, cluster analysis based on functional profiles was carried out with the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) algorithm using 5,000 bootstraps to test for robustness of genome functional groups eventually depicted by PCA. Finally, COG profiles of all 26 *Aquimarina* strains were subjected to SIMPER analysis in Past v3.0 to determine the most differentiating COG entries among functional genome groups.

Genome mining for secondary metabolite biosynthetic gene clusters

Identification of Biosynthetic Gene Clusters (BGCs) in all genomes was performed on the online server "Antibiotics and secondary metabolite analysis shell" (antiSMASH) v4.0 using the ClusterFinder algorithm and all other extra features available except for the detection of TTA codons (Blin *et al.*, 2017). In brief, antiSMASH identifies known and putatively novel BGCs underlying the possible biosynthesis of major compound classes such as polyketides, terpenoids, non-ribosomal peptides, saccharides, among others, using the "Minimum information about a biosynthetic gene cluster" (MIBiG) repository (Medema *et al.*, 2015), which possess experimentally characterized reference gene clusters, as a framework for BGC classification. The tool also delivers chemical structure predictions, when possible, and amino acid sequence files for each predicted BGC. All structure predictions obtained with antiSMASH were inventoried, while the retrieved amino acid sequences, in GenBank format, were used as input data for downstream analyses with the "Biosynthetic Genes Similarity Clustering and Prospecting Engine" (BiG-SCAPE)(Navarro-muñoz *et al.*, 2018). This tool uses the Pfam database and the hmmscan algorithm, from the HMMER suite (Eddy, 2011), to predict Pfam entries in each sequence, thus using hidden Markov models to summarize each BGC as a linear string of Pfams. For every pair of BGCs in the set, the pairwise distance between them is calculated as the weighted combination of the Jaccard Index (JI), Adjacency Index (AI) and Domain Sequence Similarity (DSS) (Cimermancic *et al.*, 2014). This way, a distance matrix between gene clusters is created based on a comparison of their protein domain content, order, copy number and sequence identity (Navarro-muñoz *et al.*, 2018). An upper distance cut-off value of 0.3 was used to define 'Gene Cluster Families' (GCFs) for BGCs identified by antiSMASH based on their homologies as explained above. GFCs were further interconnected in broader groups denominated "Gene Cluster Clans" (GCCs) using

default parameters (upper distance cut-off between BGCs: 0.7). BGCs belonging to the most representative GFCs identified in this study were subjected to ClusterBlast (21) to search for related BGCs on NCBI and thus inspect their distributions across the tree of life.

Author contributions

RC and SGS conceived the study. RC and JB provided resources and materials. SGS and TKC analysed the data. RC and SGS wrote the first manuscript draft. All authors revised and improved the manuscript prior to submission.

Acknowledgements

This study was financed by the Portuguese Science and Technology Foundation (FCT) through the research grant PTDC/xx/. SGS is the recipient of a PhD scholarship conceded by FCT in the framework of the Applied and Environmental Microbiology PhD Programme of the Institute of Bioengineering and Biosciences, Instituto Superior Técnico, Lisbon University.

Supporting Information

Supporting Information accompanies the submission of this manuscript. It encompasses one Supporting Table, three Supporting Figures (submitted in one single file along with an "extended discussion" section) and ten Supporting Datasets (submitted as separated sheets in one single Excel file). In this first submission, the respective captions and legends are shown only next to the supporting features.

Legends to Figures

Figure 1. Core- vs pan-genome rarefaction plot (A) and core / pan genome ratio representation (B) of the 26 *Aquimarina* genomes analysed in this study.

Figure 2. Current phylogenomic and functional genomics status of the *Aquimarina* genus. (A) Maximum Likelihood phylogenomic tree based on average nucleotide identity (ANI) values calculated from CDSs common to the 26 genomes (n = 1,226). The tree is drawn to scale, with the scale bar representing residue substitutions per site. Local support values (250 repetitions) are shown on tree nodes. Genomes representing type strains of species with standing in nomenclature are highlighted with ^{T}. (B) Cluster analysis of *Aquimarina* strains based on COG functional annotation. Cluster analysis of COG profiles was performed within

Past v3 using the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) algorithm on a Bray-Curtis similarity matrix calculated from Hellinger-transformed data. Cluster robustness was tested with 5,000 permutations and bootstrap values are shown next to dendrogram nodes.

Figure 3. Ordination of *Aquimarina* genomes based on COG profiles. (A) Principal Components Analysis (PCA) of genomes from host-associated (HA, orange) and free-living *Aquimarina* (FL, blue) strains was carried out using COG profiles as genome descriptors. Values displayed on the diagram axes refer to the percentage variation in the total dataset explained by the correspondent axis (that is, principal components 1 and 2). Samples (genomes) are plotted in the ordination diagram following pair-wise Bray-Curtis (dis)similarity values calculated from Hellinger-transformed COG abundance profiles. Functional genome Groups 1, 2 and 3 are highlighted with ellipses and considered significantly different from one another based on One-Way permutational analysis of variance (PERMANOVA). Strains belonging to each functional group are listed on the right panel. Three strains found not to belong to any functional group are identified in the ordination diagram. (B) Top-ten most differentiating COG categories across functional genome groups 1-3 according to SIMPER analysis. Values correspond to Hellinger-transformed counts of CDSs in COGs per genome group. "% Contrib." accounts for how much each COG entry contributes to the differentiation among Groups 1, 2 and 3.

Figure 4. Distribution of Biosynthetic Gene Clusters (BGCs) across *Aquimarina* genomes. BGC counts per compound class were obtained using antiSMASH and are shown for the 26 *Aquimarina* genomes. Fatty acid, saccharide and putative BGC counts are excluded (for details on these BGCs, see Supporting Data S8) for better visualization and to give emphasis on BGCs more likely involved in the biosynthesis of drug-like candidates. Strains belonging to functional genome groups 1, 2 and 3 (Figure 3) are highlighted in boxes. (B) Examples of NRPS-derived structures predicted with antiSMASH. The putative products of BGC #14 from *A.* megaterium XH134^{T} (NRPS clan B) and of BGC #19 from *Aquimarina* sp. AD1 (NRPS clan C) are shown. For more details, see Supporting Figure S2.

Figure 5. Genetic space underlying the secondary metabolism of *Aquimarina* species. (A) Gene Cluster Families (GFCs, n=439) identified across seven major compound classes are represented after homology-based analysis of BGCs with BiG-SCAPE. The majority of the GCFs ("Others", n=209) could not be classified using current BiG-SCAPE BGC nomenclature. The number of GCFs per compound class is shown next to each entry. Most GCFs were composed by only one BGC (singletons), with GCFs containing two or more BGCs represented through BGC networks inferred from protein sequence homology. (B) Distribution of Biosynthetic Gene Clusters (BGCs) across *Aquimarina* Gene Cluster Families (GCFs) in the class "PKSother". This class was composed by 19 GCFs in total, with eight GFCs containing more than two BGCs. These latter GFCs were further categorized into five clans A through E, shown in greater detail here. Two major clans A and B were identified, encompassing three and two GCFs, respectively, while three smaller clans each composed by one single GCF were found: clans D and E, each containing 2 BGCs, and clan C, containing 3 BGCs. BGCs belonging to different GFCs within clan A are marked in pale blue, brown and red, while BGCs belonging to different GFCs in clan B are marked in grey and light blue. (C) AntiSMASH structure prediction of metabolites derived from different GCFs in "PKSother" clan B. Further details are provided as supplements (Supporting Figure S3 and Data S10).

Table 1. Basic features of the *Aquimarina* genomes analyzed in this study.

Organism	Isolation source	Isolate source	GC content (%)	Nr contigs	Total length (Mb)	Nr CDSs	CDSs with function (%)	Nr RNAs	GenBank Accession number	References
Aquimarina sp. EL33	Gorgonian coral *Eunicella labiata*	HA	32.9	20	6.27	5530	3009 (54.6)	45	GCA_900089995.1	(Keller-costa *et al.*, 2016)
Aquimarina sp. Aq78	Marine sponge *Sarcotragus spinosulus*	HA	32.9	37	5.97	5257	2919 (55.5)	43	GCA_900299475.1	(Esteves *et al.*, 2013)
Aquimarina sp. Aq107	Marine sponge *Sarcotragus spinosulus*	HA	31.9	20	5.41	4730	2667 (56.4)	56	GCA_900299505.1	(Esteves *et al.*, 2013)
Aquimarina sp. Aq135	Marine sponge *Ircinia variabilis*	HA	32.3	64	4.94	4297	2232 (47.4)	40	GCA_900299535.1	(Esteves *et al.*, 2013)
Aquimarina sp. Aq349	Marine sponge *Sarcotragus spinosulus*	HA	32.9	22	6.17	5331	2766 (51.9)	43	GCA_900299485.1	(Esteves *et al.*, 2013)
Aquimarina sp. AU58	Marine sponge *Tedania* sp.	HA	33.0	16	6.19	5505	2945 (53.5)	45	GCA_900312745.1	(Esteves *et al.*, 2016)
Aquimarina sp. AU119	Marine sponge *Tedania* sp.	HA	32.4	39	6.50	5800	3072 (53.9)	49	GCA_900312735.1	(Esteves *et al.*, 2016)
Aquimarina sp. AU474	Marine sponge *Tedania* sp.	HA	33.0	11	4.84	4292	2449 (57.8)	42	GCA_900312815.1	(Esteves *et al.*, 2016)
Aquimarina sp. I32.4	Shell of lobster *Homarus americanus*	HA	32.4	244	4.99	4608	2322 (50.4)	39	GCA_002924285.1	(Quinn *et al.*, 2012; Ranson *et al.*, 2018)
Aquimarina sp. MAR_2010_214	Seawater	FL	32.8	1	5.98	5354	2868 (53.6)	62	GCA_002846555.1	(Hahnke and Harder, 2013)
Aquimarina sp. BL5	Marine red alga *Delisea pulchra*	HA	32.9	1	6.01	5129	2834 (55.3)	67	GCA_003443675.1	(V. Kumar *et al.*, 2016)
Aquimarina sp. AD1	Marine red alga *Delisea pulchra*	HA	32.1	1	5.46	4753	2678 (56.3)	66	GCA_003443695.1	(V. Kumar *et al.*, 2016)
Aquimarina sp. AD10	Marine red alga *Delisea pulchra*	HA	32.4	1	6.05	5205	2865 (55.0)	63	GCA_003443715.1	(V. Kumar *et al.*, 2016)
A. agarilytica ZC1^T	Marine red alga *Porphyra haitanensis*	HA	32.8	131	4.26	3624	1917 (52.9)	35	GCA_000255455.1	(Lin *et al.*, 2012)
A. agarivorans HQM9^T	Marine red alga *Gelidium amansii*	HA	33.2	183	4.07	3734	1940 (52.0)	41	GCA_000218485.2	(Zhou *et al.*, 2015)
A. aggregata RZW4-3-2^T	Seawater	FL	32.3	60	6.19	5373	2704 (50.3)	53	GCA_001632745.1	(Wang *et al.*, 2016)
A. amphilecti 92V^T	Marine sponge *Amphilectus fucorum*	HA	32.0	16	5.31	4715	2643 (56.1)	60	GCA_900109375.1	(Kennedy *et al.*, 2014)
A. atlantica 22II-S11-z7^T	Surface seawater	FL	33.0	39	5.69	5079	2630 (51.8)	43	GCA_000626715.1	(Li *et al.*, 2014)
A. latercula SIO-1^T	Seawater aquarium outflow	FL	32.2	31	6.24	5516	2916 (52.9)	55	GCA_000430645.1	(Nedashkovskaya *et*

										al., 2006)
A. longa SW024^T	Seawater	FL	31.5	90	5.50	4945	2502 (50.6)	42	GCA_001401755.1	(Yu *et al.*, 2013)
A. macrocephali JAMB N27^T	Sediment adjacent to sperm whale carcasses	FL	32.9	169	6.10	5509	2685 (48.7)	43	GCA_000520995.1	(Miyazaki *et al.*, 2010)
A. megaterium XH134^T	Surface Seawater	FL	32.9	170	6.21	5499	2769 (50.3)	43	GCA_000520975.1	(Yu *et al.*, 2014)
A. muelleri DSM 19832^T	Seawater	FL	31.3	106	4.90	4155	2253 (54.2)	46	GCA_000430665.1	(Nedashkovskaya *et al.*, 2005)
A. pacifica SW150^T	Surface Seawater	FL	33.5	145	5.26	4458	2203 (49.4)	43	GCA_000520955.1	(Zhang *et al.*, 2014)
A. sediminis W01^T	Marine sediment	FL	33.3	40	5.84	5158	2654 (51.5)	53	GCA_002895435.1	(Zhou *et al.*, 2018)
A. spongiae A6^T	Marine sponge *Halichondria oshoro*	HA	35.9	28	5.35	4803	2480 (51.6)	46	GCA_900141785.1	(Yoon *et al.*, 2011)

Table 2. Overview of phylogenomic comparisons between selected pairs of strains.[1]

		DGGC		EDGAR	JSpeciesWS	
Genome 1	**Genome 2**	**Prob. DDH >= 70%**	**ΔG+C (%)**	**ANI (mean) %**	**ANIb %**	**ANIm %**
Aquimarina sp. AU58	*A. megaterium* XH134 ^{T}	58.83	0.09	95.51	94.21	95.44
Aquimarina sp. Aq349	*A. megaterium* XH134 ^{T}	43.24	0.05	94.53	93.42	94.53
Aquimarina sp. EL33	*A. megaterium* XH134 ^{T}	43.01	0.07	94.57	93.30	94.54
Aquimarina sp. Aq349	*Aquimarina* sp. EL33	97.27	0.02	99.80	99.19	99.48
Aquimarina sp. Aq349	*Aquimarina* sp. AU58	70.31	0.04	96.37	95.42	96.07
Aquimarina sp. AU58	*Aquimarina* sp. EL33	69.15	0.02	96.37	95.32	96.02
Aquimarina sp. AD1	*A. latercula* DSM2041 ^{T}	94.69	0.14	99.41	98.27	98.27
Aquimarina sp. Aq107	*A. latercula* DSM2041 ^{T}	46.55	0.32	94.61	93.62	94.71
Aquimarina sp. AD10	*A. aggregata* RZW4-3-2 ^{T}	96.10	0.15	99.31	98.56	99.04
Aquimarina sp. Aq78	*A. macrocephali* JAMBN27 ^{T}	68.23	0.00	96.08	94.66	95.84
Aquimarina sp. I32.4	*Aquimarina* sp. Aq135	62.43	0.03	95.29	94.61	95.53
Aquimarina sp. I32.4	*A. macrocephali* JAMB N27^{T}	26.50	0.02	79.05	79.11	84.69
Aquimarina sp. BL5	*A. amphilecti* 92V^{T}	28.40	0.05	81.00	80.89	85.62

[1]The tool Genome-to-Genome Distance Calculator (DGGC) was used to calculate the probability of DNA-DNA hybridization (DDH) being over 70% and the G+C difference (%). In the EDGAR platform, the mean ANI was retrieved. Finally, the webservice JSpeciesWS was used to calculate two other ANI values: ANI based on BLAST+ (ANIb) and ANI based on MUMmer (ANIm).

Table 3. Number of coding sequences (CDSs) assigned to Pfam entries related with eukaryotic-like proteins (ELPs) and glycosyl hydrolases (GHs).[1]

	Name	Description	*Aquimarina* spp. Group 1	*Aquimarina* spp. Group 2	*Aquimarina* spp. Group 3	*Aquimarina agarilytica* ZC1^T	*Aquimarina agarivorans* HQM9^T	*Aquimarina pacifica* SW150^T
ELPs	**Ank**	Ankyrin repeats	18	26	57	36	21	19
	WD-40	WD-40 repeats	10	10	18	4	5	94
	TPRs	Tetratricopeptide repeats	113	112	119	50	71	61
	LRRs	Leucine-rich repeats	27	23	29	14	12	9
GHs	**GH5, GH9**	Glycosyl hydrolase families 5 & 9 (cellulases)	10	17	10	6	8	6
	GH8	Glycosyl hydrolases family 8 (chitosanases)	2	2	1	9	2	2
	GH10, GH11	Glycosyl hydrolase families 10 & 11 (xylanases)	7	11	7	5	3	7
	GH16	Glycosyl hydrolases family 16 (b-agarases, k-carrageenases)	4	3	3	1	2	2
	GH18, GH19	Glycosyl hydrolases family 18 & 19 (chitinases)	6	12	4	3	2	3
	GH25	Glycosyl hydrolases family 25 (lysozymes)	2	1	3	4	1	1
	GH26	Glycosyl hydrolase family 26 (mannanases)	1	2	2	14	1	1
	GH31, GH57	Glycosyl hydrolase families 31 & 57 (a-amylases)	4	5	5	1	3	3
	GH71	Glycosyl hydrolase family 71 (endoglucanases)	1	2	1	1	9	1

[1]For groups 1-3, the average number of coding sequences (CDSs) assigned to each Pfam entry is shown.

References

Agrawal, S., Adholeya, A., and Deshmukh, S.K. (2016) The pharmacological potential of non-ribosomal peptides from marine sponge and tunicates. *Front Pharmacol* **7**: 1–21.

Alonso, C., Warnecke, F., Amann, R.I., and Pernthaler, J. (2007) High local and global diversity of *Flavobacteria* in marine plankton. *Environ Microbiol* **9**: 1253–1266.

Blin, K., Wolf, T., Chevrette, M.G., Lu, X., Schwalen, C.J., Kautsar, S.A., et al. (2017) AntiSMASH 4.0 - improvements in chemistry prediction and gene cluster boundary identification. *Nucleic Acids Res* **45**: W36–W41.

Blom, J., Kreis, J., Spänig, S., Juhre, T., Bertelli, C., Ernst, C., and Goesmann, A. (2016) EDGAR 2.0: an enhanced software platform for comparative gene content analyses. *Nucleic Acids Res* **44**: W22–W28.

Boone, D.R., Bergey, D.H., Castenholz, R.W., and Garrity, G.M. (2010) Bergey's Manual of Systematic Bacteriology: The Bacteroidetes, Spirochaetes, Tenericutes (Mollicutes), Acidobacteria, Fibrobacteres, Fusobacteria, Dictyoglomi, Gemmatimonadetes, Lentisphaerae, Verrucomicrobia, Chlamydiae, and Planctomycetes, 2nd ed. Athens, USA: Springer.

Braun, T.F., Khubbar, M.K., Saffarini, D.A., and McBride, M.J. (2005) *Flavobacterium johnsoniae* gliding motility genes identified by *mariner* mutagenesis. *J Bacteriol* **187**: 6943–6952.

Burgsdorf, I., Slaby, B.M., Handley, K.M., Haber, M., Blom, J., Marshall, C.W., et al. (2015) Lifestyle Evolution in Cyanobacterial Symbionts of Sponges. *MBio* **6**: 1–14.

Caputo, A., Merhej, V., Georgiades, K., Fournier, P.E., Croce, O., Robert, C., and Raoult, D. (2015) Pan-genomic analysis to redefine species and subspecies based on quantum discontinuous variation: The *Klebsiella* paradigm. *Biol Direct* **10**: 1–12.

Chen, W.M., Sheu, F.S., Sheu, S.Y., Sheu, W.C.F.S.S., Chen, W.M., Sheu, F.S., et al. (2012) *Aquimarina salinaria* sp . nov ., a novel algicidal bacterium isolated from a saltpan. *Arch Microbiol* **194**: 103–112.

Chistoserdov, A.Y., Smolowitz, R., Mirasol, F., and Hsu, A. (2005) Culture-dependent characterization of the microbial community associated with epizootic shell disease lesions in american lobster, *Homarus americanus*. *J Shellfish Res* **24**: 741–747.

Cimermancic, P., Medema, M.H., Claesen, J., Kurita, K., Wieland Brown, L.C., Mavrommatis, K., et al. (2014) Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell* **158**: 412–421.

Cotter, P.D., Ross, R.P., and Hill, C. (2013) Bacteriocins-a viable alternative to antibiotics? *Nat Rev Microbiol* **11**: 95–105.

Díez-Vives, C., Esteves, A.I.S., Costa, R., Nielsen, S., and Thomas, T. (2018) Detecting signatures of a sponge-associated lifestyle in bacterial genomes. *Environ Microbiol Rep* **10**: 433–443.

Donadio, S., Monciardini, P., and Sosio, M. (2007) Polyketide synthases and nonribosomal peptide synthetases: The emerging view from bacterial genomics. *Nat Prod Rep* **24**: 1073–1079.

Eddy, S.R. (2011) Accelerated profile HMM searches. *PLoS Comput Biol* **7**: 1–16.

Van Elsas, J.D., Semenov, A. V., Costa, R., and Trevors, J.T. (2011) Survival of *Escherichia coli* in the environment: Fundamental and public health aspects. *ISME J* **5**: 173–183.

Esteves, A.I.S., Amer, N., Nguyen, M., and Thomas, T. (2016) Sample processing impacts the viability and cultivability of the sponge microbiome. *Front Microbiol* **7**: 1–17.

Esteves, A.I.S., Hardoim, C.C.P., Xavier, J.R., Gonçalves, J.M.S., and Costa, R. (2013) Molecular richness and biotechnological potential of bacteria cultured from Irciniidae sponges in the north-east Atlantic. *FEMS Microbiol Ecol* **85**: 519–536.

Fernández-Gómez, B., Richter, M., Schüler, M., Pinhassi, J., Acinas, S.G., González, J.M., and Pedrós-Alió, C. (2013) Ecology of marine bacteroidetes: A comparative genomics approach. *ISME J* **7**: 1026–1037.

Hahnke, R.L. and Harder, J. (2013) Phylogenetic diversity of *Flavobacteria* isolated from the North Sea on solid media. *Syst Appl Microbiol* **36**: 497–504.

Hammer, Ø., Harper, D.A.T., and Ryan, P.. (2001) PAST: Paleontological statistics software package for education and data analysis. *Palaeontol Electron* **4**: 1–9.

Helfrich, E.J.N., Ueoka, R., Dolev, A., Rust, M., Meoded, R.A., Califano, G., et al. (2019) Automated structure prediction of trans-acyltransferase polyketide synthase products. *Nat Chem Biol*.

Hudson, J., Kumar, V., and Egan, S. (2019) Comparative genome analysis provides novel insight into the interaction of *Aquimarina* sp. AD1, BL5 and AD10 with their macroalgal host. *Mar Genomics* 0–8.

Ilangumaran, G., Stratton, G., Ravichandran, S., Shukla, P.S., Potin, P., Asiedu, S., and Prithiviraj, B. (2017) Microbial degradation of lobster shells to extract chitin derivatives for plant disease management. *Front Microbiol* **8**: 1–14.

Karimi, E., Keller-Costa, T., Slaby, B.M., Cox, C.J., da Rocha, U.N., Hentschel, U., and Costa, R. (2019) Genomic blueprints of sponge-prokaryote symbiosis are shared by low abundant and cultivatable Alphaproteobacteria. *Sci Rep* **9**: 1–15.

Karimi, E., Ramos, M., Gonçalves, J.M.S., Xavier, J.R., Reis, M.P., and Costa, R. (2017) Comparative metagenomics reveals the distinctive adaptive features of the *Spongia officinalis* endosymbiotic consortium. *Front Microbiol* **8**: 1–16.

Karimi, E., Slaby, B.M., Soares, A.R., Blom, J., Hentschel, U., Costa, R., et al. (2018) Metagenomic binning reveals versatile nutrient cycling and distinct adaptive features in alphaproteobacterial symbionts of marine sponges. *FEMS Microbiol Ecol* **94**: 1–18.

Keller-Costa, T., Eriksson, D., Gonçalves, J.M.S.S., Gomes, N.C.M., Lago-Lestón, A., Costa,

R., et al. (2017) The gorgonian coral *Eunicella labiata* hosts a distinct prokaryotic consortium amenable to cultivation. *FEMS Microbiol Ecol* **93**: 1–14.

Keller-costa, T., Silva, R., and Lago-lestón, A. (2016) Genomic Insights into *Aquimarina* sp. Strain EL33, a Bacterial Symbiont of the Gorgonian Coral *Eunicella labiata*. *Genome Announc* **4**: 2–3.

Kennedy, J., Margassery, L.M., O'Leary, N.D., O'Gara, F., Morrissey, J., and Dobson, A.D.W. (2014) *Aquimarina amphilecti* sp. nov., isolated from the sponge *Amphilectus fucorum*. *Int J Syst Evol Microbiol* **64**: 501–505.

Kharade, S.S. and McBride, M.J. (2014) *Flavobacterium johnsoniae* chitinase ChiA is required for chitin utilization and is secreted by the type IX secretion system. *J Bacteriol* **196**: 961–970.

Kim, M., Oh, H.-S., Park, S.-C., and Chun, J. (2014) Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. *Int J Syst Evol Microbiol* **64**: 346–351.

Klappenbach, J.A., Goris, J., Vandamme, P., Coenye, T., Konstantinidis, K.T., Tiedje, J.M., et al. (2007) DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *Int J Syst Evol Microbiol* **57**: 81–91.

Kumar, S., Stecher, G., and Tamura, K. (2016) MEGA7 : Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol Biol Evol* **33**: 1870–1874.

Kumar, V., Zozaya-Valdes, E., Kjelleberg, S., Thomas, T., and Egan, S. (2016) Multiple opportunistic pathogens can cause a bleaching disease in the red seaweed *Delisea pulchra*. *Environ Microbiol* **18**: 3962–3975.

Kurtz, S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C., and Salzberg, S.L. (2004) Versatile and open software for comparing large genomes. *Genome Biol* **5**: 1–9.

Lauber, F., Deme, J.C., Lea, S.M., and Berks, B.C. (2018) Type 9 secretion system structures reveal a new protein transport mechanism. *Nature* **564**: 77–82.

Li, Guizhen, Lai, Q., Sun, F., Liu, X., Xie, Y., Du, Y., et al. (2014) *Aquimarina atlantica* sp. nov., isolated from surface seawater of the Atlantic Ocean. *Antonie van Leeuwenhoek, Int J Gen Mol Microbiol* **106**: 293–300.

Lin, B., Lu, G., Zheng, Y., Xie, W., Li, S., and Hu, Z. (2012) *Aquimarina agarilytica* sp. nov., an agarolytic species isolated from a red alga. *Int J Syst Evol Microbiol* **62**: 869–873.

Lopanik, N., Lindquist, N., and Targett, N. (2004) Potent cytotoxins produced by a microbial symbiont protect host larvae from predation. *Oecologia* **139**: 131–139.

Mann, A.J., Hahnke, R.L., Huang, S., Werner, J., Xing, P., Barbeyron, T., et al. (2013) The Genome of the Alga-Associated Marine Flavobacterium *Formosa agariphila* KMM 3901 T Reveals a Broad Potential for Degradation of Algal Polysaccharides. *Appl Environ Microbiol* **79**: 6813–6822.

McCutcheon, J.P. and Moran, N.A. (2012) Extreme genome reduction in symbiotic bacteria. *Nat Rev Microbiol* **10**: 13–26.

Medema, M.H. (2018) Computational Genomics of Specialized Metabolism: from Natural Product Discovery to Microbiome Ecology. *mSystems* **3**: 1–3.

Medema, M.H., Kottmann, R., Yilmaz, P., Cummings, M., Biggins, J.B., Blin, K., et al. (2015) Minimum Information about a Biosynthetic Gene cluster. *Nat Chem Biol* **11**: 625–631.

Meier-Kolthoff, J.P., Auch, A.F., Klenk, H.P., and Göker, M. (2013) Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics* **14**: 1–14.

Ming, W., Sian, F., and Yi, S. (2011) Novel l -amino acid oxidase with algicidal activity against toxic cyanobacterium *Microcystis aeruginosa* synthesized by a bacterium *Aquimarina* sp . *Enzyme Microb Technol* **49**: 372–379.

Miyazaki, M., Nagano, Y., Fujiwara, Y., Hatada, Y., and Nogi, Y. (2010) *Aquimarina macrocephali* sp. nov., isolated from sediment adjacent to sperm whale carcasses. *Int J Syst Evol Microbiol* **60**: 2298–2302.

Navarro-muñoz, J.C., Selem-mojica, N., Mullowney, M.W., Kautsar, S., Abubucker, S., Roeters, A., et al. (2018) A computational framework for systematic exploration of biosynthetic diversity from large-scale genomic data. *bioRxiv*.

Nayfach, S., Shi, Z.J., Seshadri, R., Pollard, K.S., and Kyrpides, N. (2019) Novel insights from uncultivated genomes of the global human gut microbiome. *Nature* **568**: 505–510.

NCBI Resource Coordinators (2016) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **44**: D7–D19.

Nedashkovskaya, O.I., Kim, S.B., Lysenko, A.M., Frolova, G.M., Mikhailov, V. V., Lee, K.H., and Bae, K.S. (2005) Description of *Aquimarina muelleri* gen. nov., sp. nov., and proposal of the reclassification of [*Cytophaga*] *latercula* Lewin 1969 as Stanierella latercula gen. nov., comb. nov. *Int J Syst Evol Microbiol* **55**: 225–229.

Nedashkovskaya, O.I., Vancanneyt, M., Christiaens, L., Kalinovskaya, N.I., Mikhailov, V. V., and Swings, J. (2006) Aquimarina intermedia sp. nov., reclassification of Stanierella latercula (Lewin 1969) as Aquimarina latercula comb. nov and *Gaetbulimicrobium brevivitae* Yoon et al. 2006 as Aquimarina brevivitae comb. nov. and amended descript. *Int J Syst Evol Microbiol* **56**: 2037–2041.

Nguyen, M.T.H.D., Liu, M., and Thomas, T. (2014) Ankyrin-repeat proteins from sponge symbionts modulate amoebal phagocytosis. *Mol Ecol* **23**: 1635–1645.

Overbeek, R., Olson, R., Pusch, G.D., Olsen, G.J., Davis, J.J., Disz, T., et al. (2014) The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Res* **42**: 206–214.

Parks, D.H., Tyson, G.W., Hugenholtz, P., and Beiko, R.G. (2014) STAMP: Statistical

analysis of taxonomic and functional profiles. *Bioinformatics* **30**: 3123–3124.

Parvez, A., Giri, S., Giri, G.R., Kumari, M., Bisht, R., and Saxena, P. (2018) Novel Type III Polyketide Synthases Biosynthesize Methylated Polyketides in *Mycobacterium marinum*. *Sci Rep* **8**: 1–13.

Piel, J. (2002) A polyketide synthase-peptide synthetase gene cluster from an uncultured bacterial symbiont of *Paederus* beetles. *Proc Natl Acad Sci* **99**: 14002–14007.

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

Pinhassi, J., Sala, M.M., Havskum, H., Peters, F., Guadayol, O., Malits, A., and Marrase, C. (2004) Changes in bacterioplankton composition under different phytoplankton regimens. *Appl Environ Microbiol* **70**: 6753–6766.

Quinn, R.A., Metzler, A., Smolowitz, R.M., Tlusty, M., and Chistoserdov, A.Y. (2012) Exposures of *Homarus americanus* Shell to Three Bacteria Isolated from Naturally Occurring Epizootic Shell Disease Lesions. *J Shellfish Res* **31**: 485–493.

Raimundo, I., Silva, S.G., Costa, R., and Keller-Costa, T. (2018) Bioactive Secondary Metabolites from Octocoral-Associated Microbes—New Chances for Blue Growth. *Mar Drugs* **16**: 1–25.

Ranson, H.J., Laporte, J., Spinard, E., Chistoserdov, A.Y., Gomez-Chiarri, M., Rowley, D.C., et al. (2018) Draft Genome Sequence of the Putative Marine Pathogen *Aquimarina* sp. Strain I32.4. *Genome Announc* **6**: 1–2.

Richter, M. and Rossello-Mora, R. (2009) Shifting the genomic gold standard for the prokaryotic species definition. *Proc Natl Acad Sci* **106**: 19126–19131.

Richter, M., Rosselló-Móra, R., Oliver Glöckner, F., and Peplies, J. (2015) JSpeciesWS: A web server for prokaryotic species circumscription based on pairwise genome comparison. *Bioinformatics* **32**: 929–931.

Sitao Wu, Zhengwei Zhu, Liming Fu, B.N. and W.L., Wu, S., Zhu, Z., Fu, L., Niu, B., Li, W., and Sitao Wu, Zhengwei Zhu, Liming Fu, B.N. and W.L. (2011) WebMGA : a Customizable Web Server for Fast Metagenomic Sequence Analysis. *BMC Genomics* **12**: 1–9.

Stocker, R. and Seymour, J.R. (2012) Ecology and Physics of Bacterial Chemotaxis in the Ocean. *Microbiol Mol Biol Rev* **76**: 792–812.

Süssmuth, R.D. and Mainz, A. (2017) Nonribosomal Peptide Synthesis—Principles and Prospects. *Angew Chemie - Int Ed* **56**: 3770–3821.

Teeling, H., Fuchs, B.M., Becher, D., Klockow, C., Gardebrecht, A., Bennke, C.M., et al. (2012) Substrate-Controlled Succession of Marine Bacterioplankton Populations Induced by a Phytoplankton Bloom. *Science (80-)* **336**: 608–611.

Thomas, T., Rusch, D., DeMaere, M.Z., Yung, P.Y., Lewis, M., Halpern, A., et al. (2010) Functional genomic signatures of sponge bacteria reveal unique and shared features of symbiosis. *ISME J* **4**: 1557–1567.

Touchon, M., Hoede, C., Tenaillon, O., Barbe, V., Baeriswyl, S., Bidet, P., et al. (2009) Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLoS Genet* **5**:.

Trindade, M., van Zyl, L.J., Navarro-Fernández, J., and Elrazak, A.A. (2015) Targeted metagenomics as a tool to tap into marine natural product diversity for the discovery and production of drug candidates. *Front Microbiol* **6**: 1–14.

Tseng, C.H. and Tang, S.L. (2014) Marine microbial metagenomics: From individual to the environment. *Int J Mol Sci* **15**: 8878–8892.

Unfried, F., Becker, S., Robb, C.S., Hehemann, J.H., Markert, S., Heiden, S.E., et al. (2018) Adaptive mechanisms that provide competitive advantages to marine bacteroidetes during microalgal blooms. *ISME J* **12**: 2894–2906.

Wang, Y., Ming, H., Guo, W., Chen, H., and Zhou, C. (2016) *Aquimarina aggregata* sp. nov., isolated from seawater. *Int J Syst Evol Microbiol* **66**: 3406–3412.

Wilson, M.C., Mori, T., Rückert, C., Uria, A.R., Helf, M.J., Takada, K., et al. (2014) An environmental bacterial taxon with a large and distinct metabolic repertoire. *Nature* **506**: 58–62.

Yang, S.C., Lin, C.H., Sung, C.T., and Fang, J.Y. (2014) Antibacterial activities of bacteriocins: Application in foods and pharmaceuticals. *Front Microbiol* **5**: 1–10.

Yoon, B.J., You, H.S., Lee, D.H., and Oh, D.C. (2011) *Aquimarina spongiae* sp. nov., isolated from marine sponge *Halichondria oshoro*. *Int J Syst Evol Microbiol* **61**: 417–421.

Younes, I. and Rinaudo, M. (2015) Chitin and chitosan preparation from marine sources. Structure, properties and applications. *Mar Drugs* **13**: 1133–1174.

Yu, T., Yin, Q., Song, X., Zhao, R., Shi, X., and Zhang, X.-H.H. (2013) *Aquimarina longa* sp. nov., isolated from seawater, and emended description of *Aquimarina muelleri*. *Int J Syst Evol Microbiol* **63**: 1235–1240.

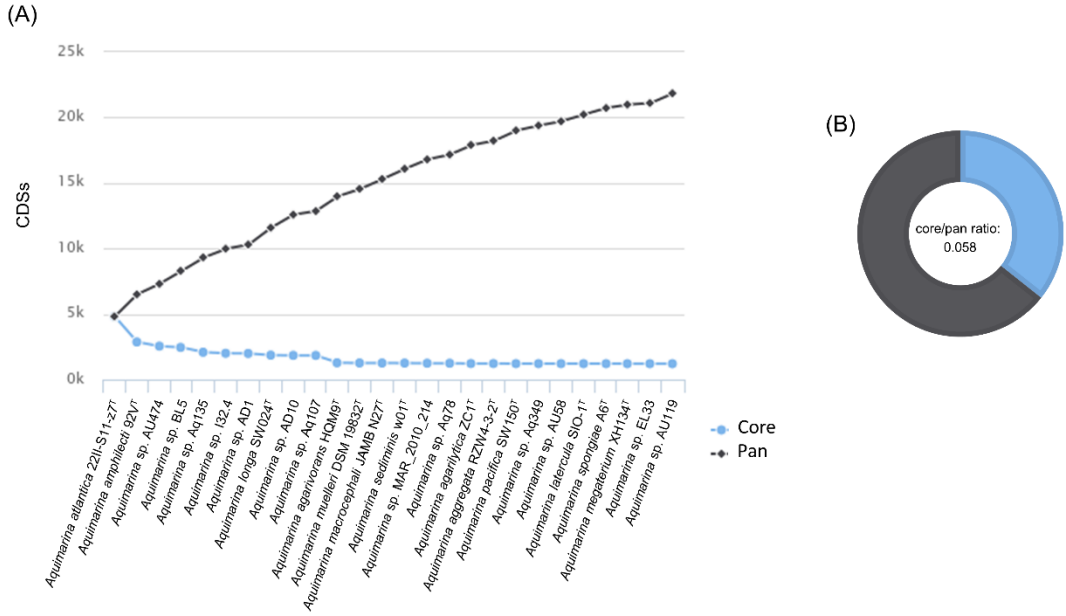
Yu, T., Zhang, Z., Fan, X., Shi, X., and Zhang, X.-H. (2014) *Aquimarina megaterium* sp. nov., isolated from seawater. *Int J Syst Evol Microbiol* **64**: 122–127.

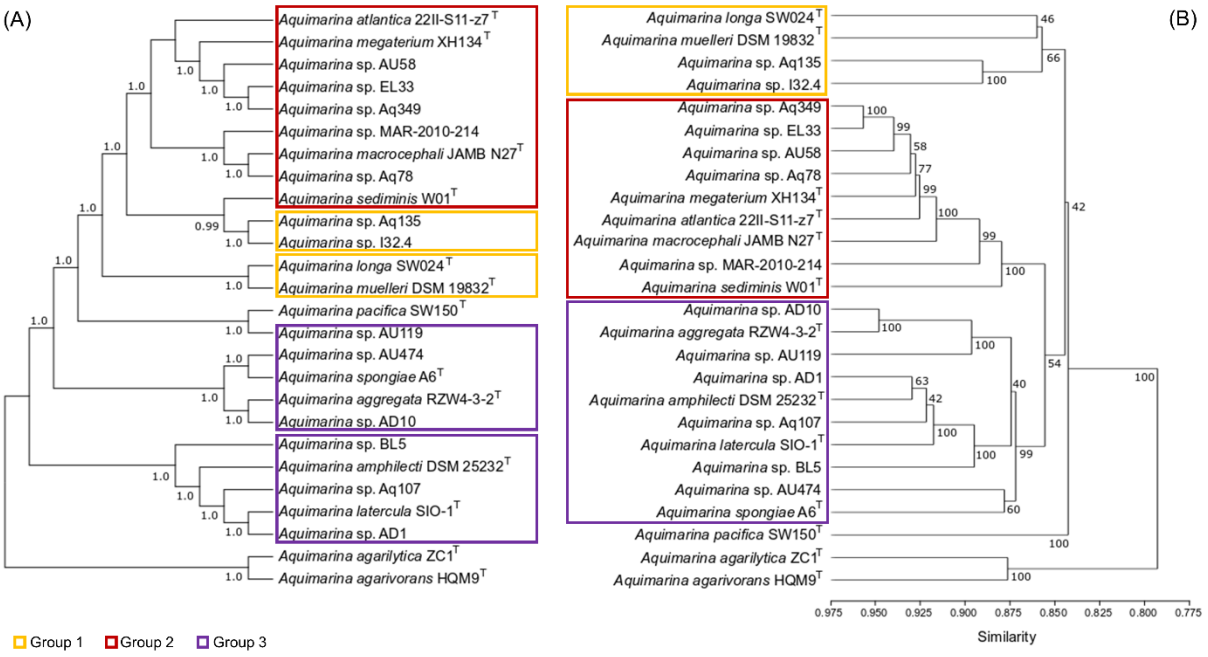
Zhang, Z., Yu, T., Xu, T., and Zhang, X.-H. (2014) *Aquimarina pacifica* sp. nov., isolated from seawater. *Int J Syst Evol Microbiol* **64**: 1991–1997.

Zhou, N.W.L., Du, Y.L.Z., Wang, N.-N., Zhou, L.-Y., Li, Y.-X., and Du, Z.-J. (2018) *Aquimarina sediminis* sp. nov., isolated from coastal sediment. *Antonie Van Leeuwenhoek* **111**: 2257–2265.

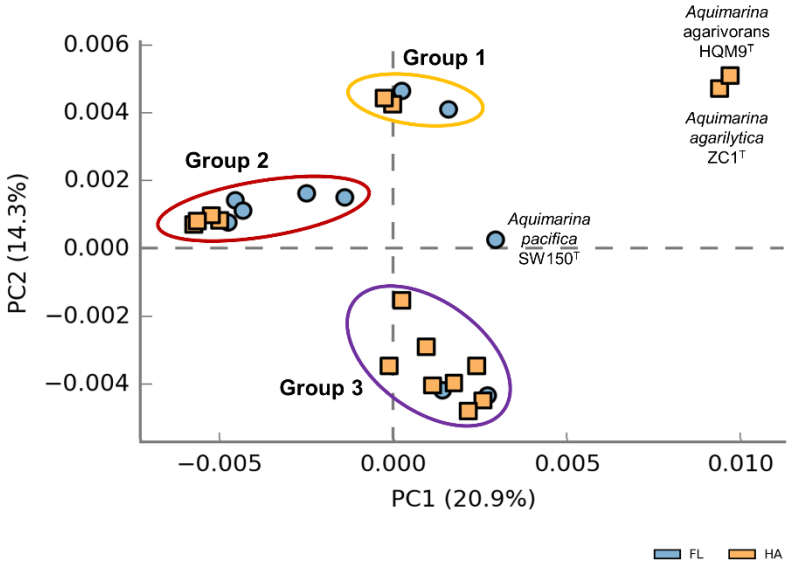
Zhou, Y.X., Wang, C., Du, Z.J., and Chen, G.J. (2015) *Aquimarina agarivorans* sp. Nov., a genome-sequenced member of the class *Flavobacteriia* isolated from *Gelidium amansii*.

Int J Syst Evol Microbiol **65**: 2684–2688.





(A)



Group 1
Aquimarina sp. I32.4
Aquimarina longa SW024^T
Aquimarina muelleri DSM 19832^T
Aquimarina sp. Aq135

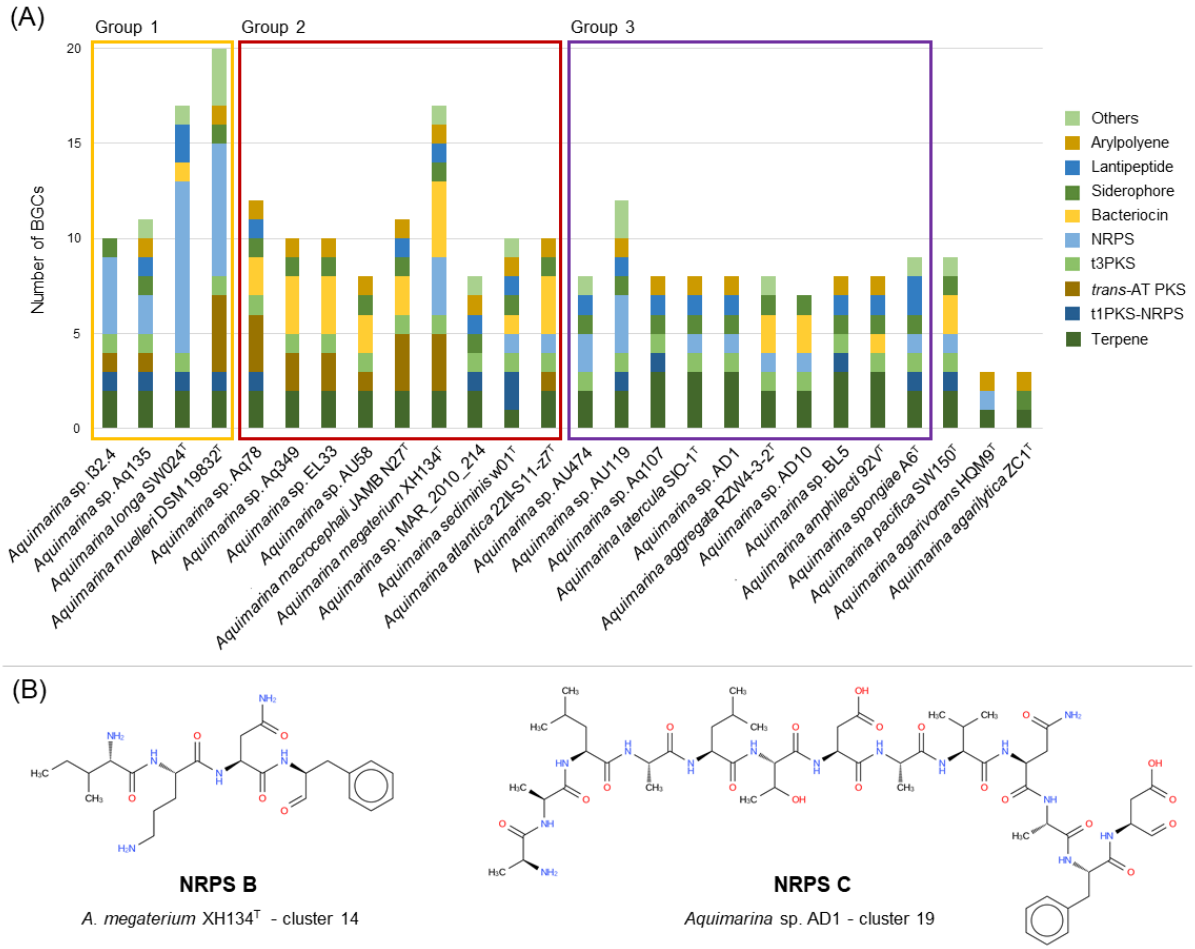
Group 2
Aquimarina sediminis w01^T
Aquimarina sp. MAR_2010_214
Aquimarina atlantica 22II-S11-z7^T
Aquimarina macrocephali JAMB N27^T
Aquimarina sp. Aq349
Aquimarina sp. AU58
Aquimarina sp. EL33
Aquimarina sp. Aq78
Aquimarina megaterium XH134^T

Group 3
Aquimarina sp. AU474
Aquimarina sp. AU119
Aquimarina spongiae A6^T
Aquimarina sp. Aq107
Aquimarina aggregata RZW4-3-2^T
Aquimarina latercula SIO-1^T
Aquimarina sp. BL5
Aquimarina sp. AD10
Aquimarina sp. AD1
Aquimarina amphilecti 92V^T

(B)

Taxon	Average dissim.	Contrib. %	Annotation	Mean Group 1*	Mean Group 2	Mean Group 3
COG3321	0,05777	0,3766	Acyl transferase domain in polyketide synthase (PKS) enzymes	0,0464	0,0606	0,00896
COG4886	0,04966	0,3238	Leucine-rich repeat (LRR) protein	0,0242	0,079	0,0457
COG2273	0,04762	0,3105	Beta-glucanase, GH16 family	0,00435	0,00335	0,0452
COG2207	0,04713	0,3073	AraC-type DNA-binding domain and AraC-containing proteins	0,11	0,169	0,155
COG3275	0,04656	0,3036	Sensor histidine kinase, LytS/YehU family	0,0615	0,115	0,088
COG1020	0,04599	0,2999	Non-ribosomal peptide synthetase component F	0,0885	0,0476	0,0541
COG3279	0,04131	0,2693	DNA-binding response regulator, LytR/AlgR family	0,0835	0,131	0,109
COG3979	0,03851	0,251	Chitodextrinase	0,0704	0,0923	0,0579
COG3501	0,03683	0,2401	Uncharacterized conserved protein, implicated in type VI secretion and phage assembly	0,0535	0,013	0,0341
COG2335	0,03568	0,2327	Uncharacterized surface protein containing fasciclin (FAS1) repeats	0,0192	0,0238	0,0523

*Values correspond to Hellinger-transformed counts of CDSs in COGs per genome group.



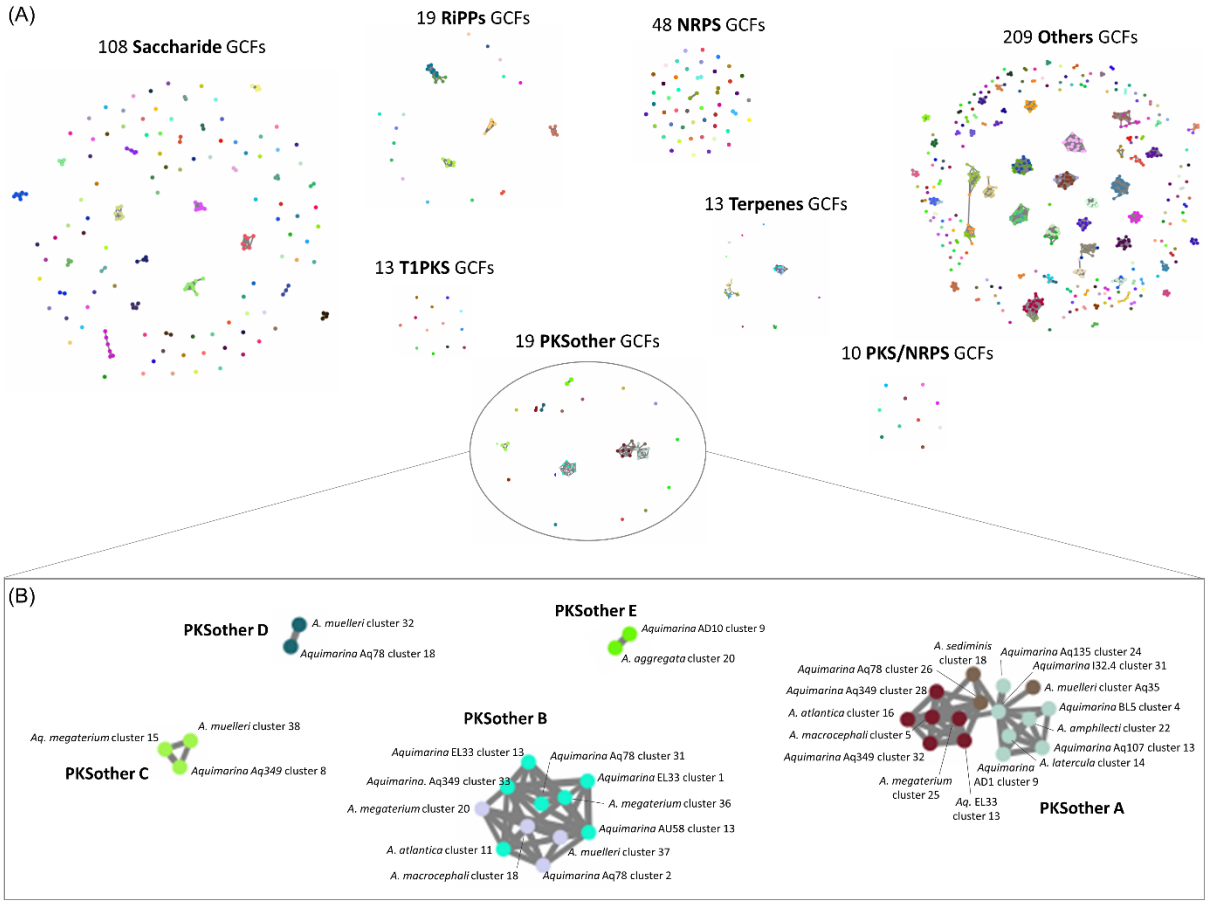
(B)

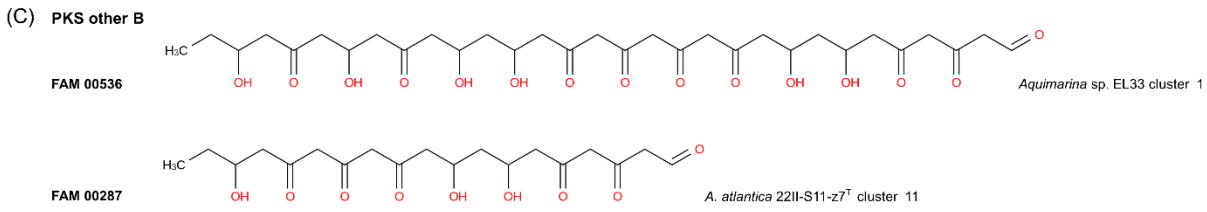
NRPS B

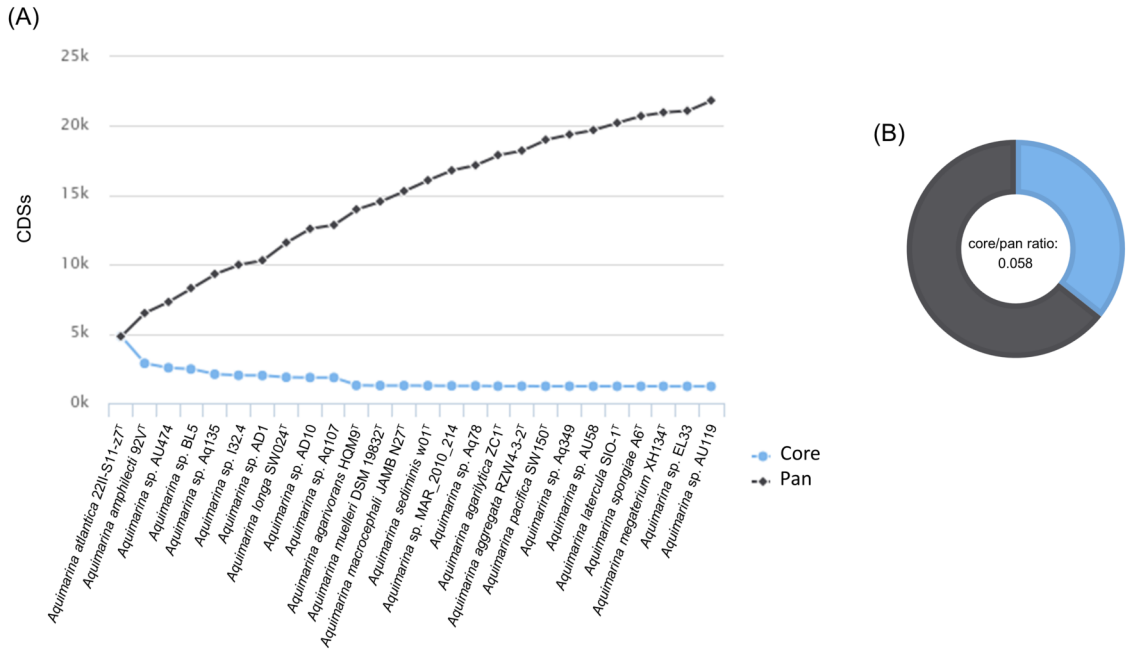
A. megaterium XH134^T - cluster 14

NRPS C

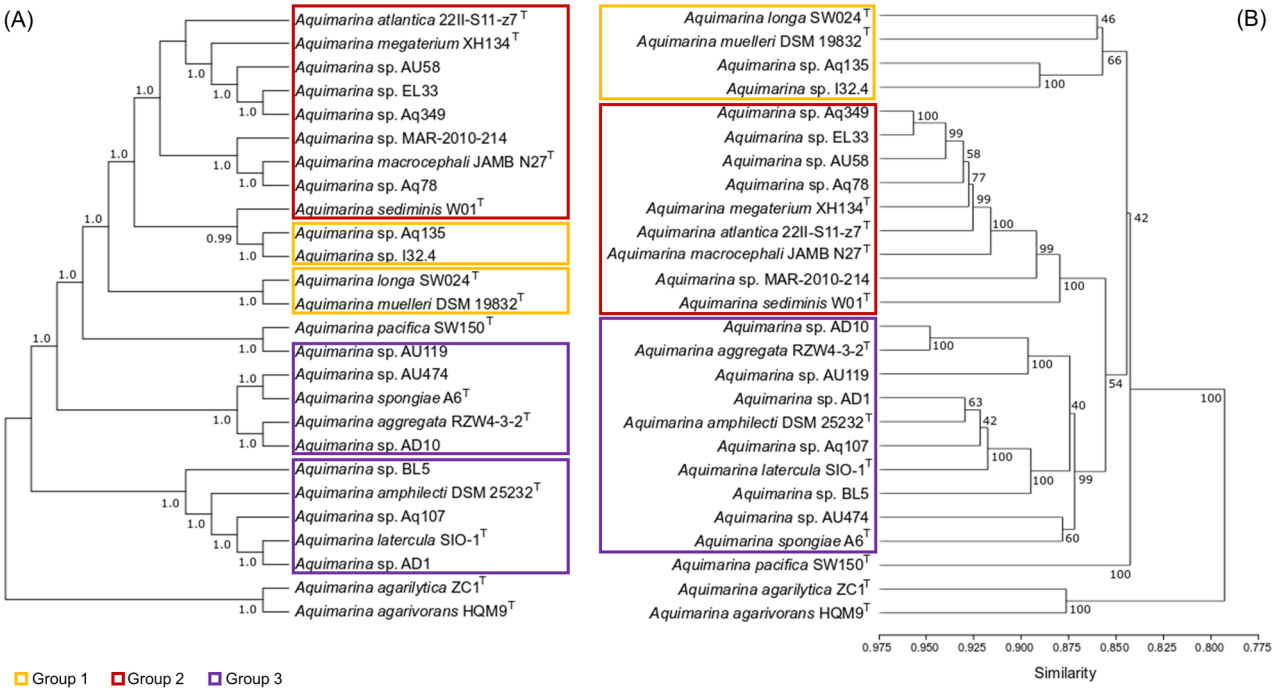
Aquimarina sp. AD1 - cluster 19





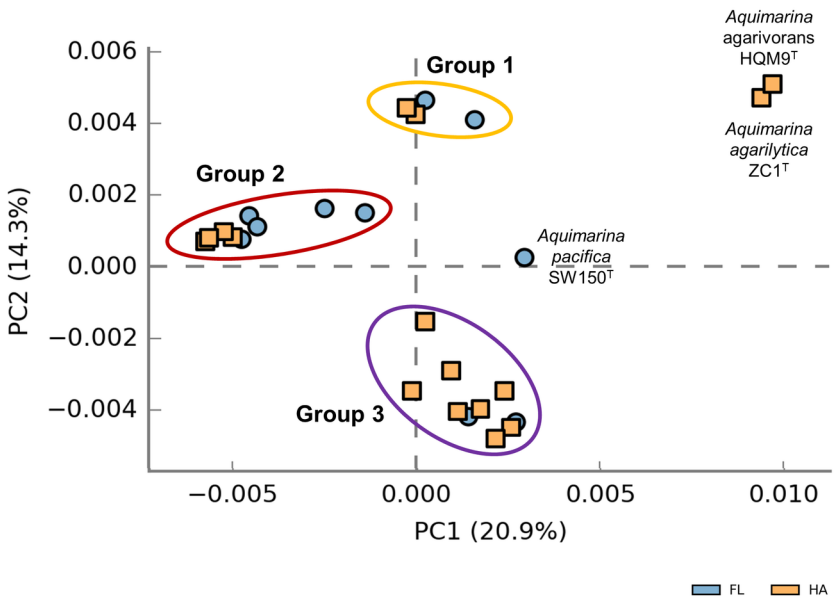


EMI_14747_Fig.1.tif



EMI_14747_Fig.2.tif

(A)



Group 1
Aquimarina sp. I32.4
Aquimarina longa SW024^T
Aquimarina muelleri DSM 19832^T
Aquimarina sp. Aq135

Group 2
Aquimarina sediminis w01^T
Aquimarina sp. MAR_2010_214
Aquimarina atlantica 22II-S11-z7^T
Aquimarina macrocephali JAMB N27^T
Aquimarina sp. Aq349
Aquimarina sp. AU58
Aquimarina sp. EL33
Aquimarina sp. Aq78
Aquimarina megaterium XH134^T

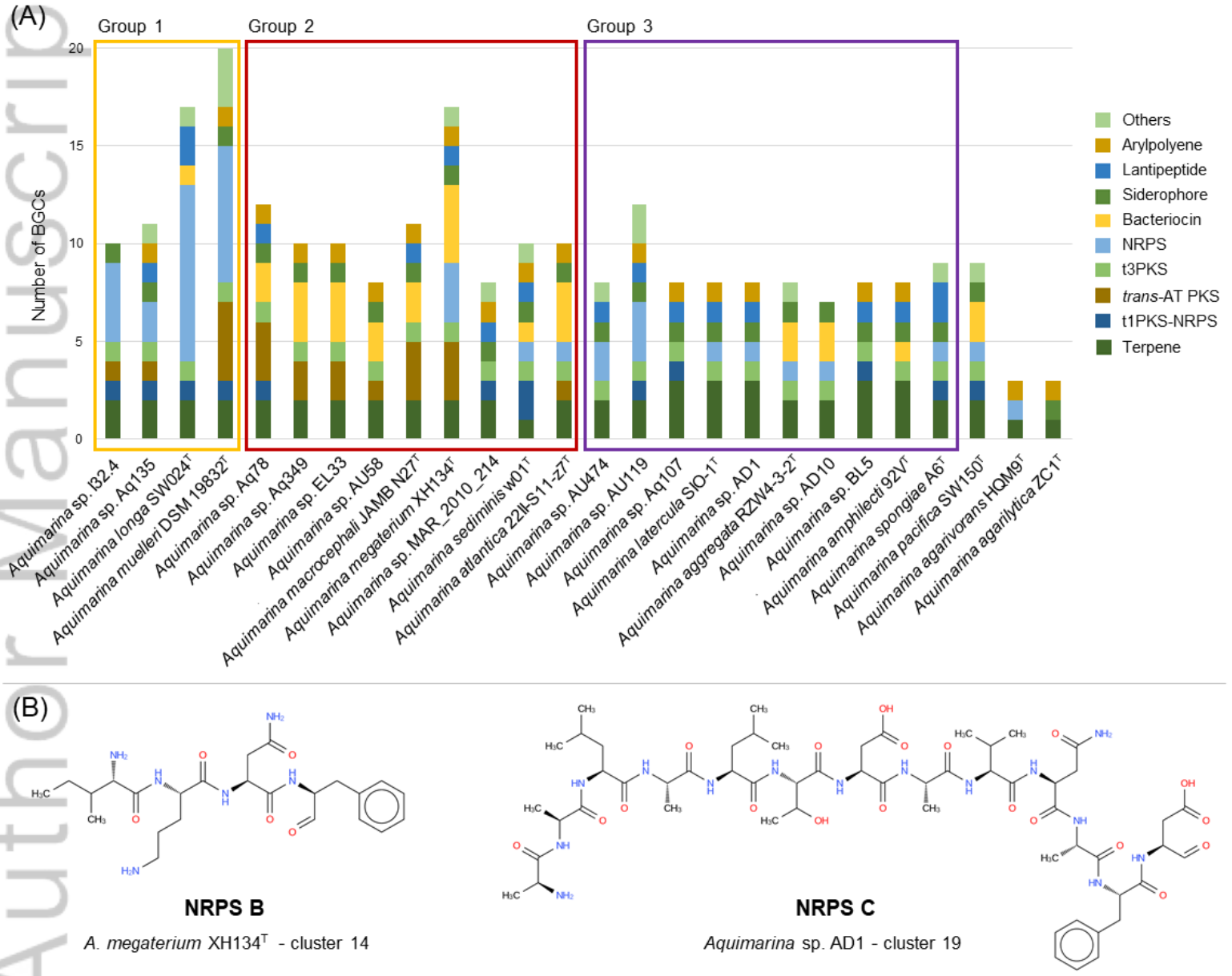
Group 3
Aquimarina sp. AU474
Aquimarina sp. AU119
Aquimarina spongiae A6^T
Aquimarina sp. Aq107
Aquimarina aggregata RZW4-3-2^T
Aquimarina latercula SIO-1^T
Aquimarina sp. BL5
Aquimarina sp. AD10
Aquimarina sp. AD1
Aquimarina amphilecti 92V^T

(B)

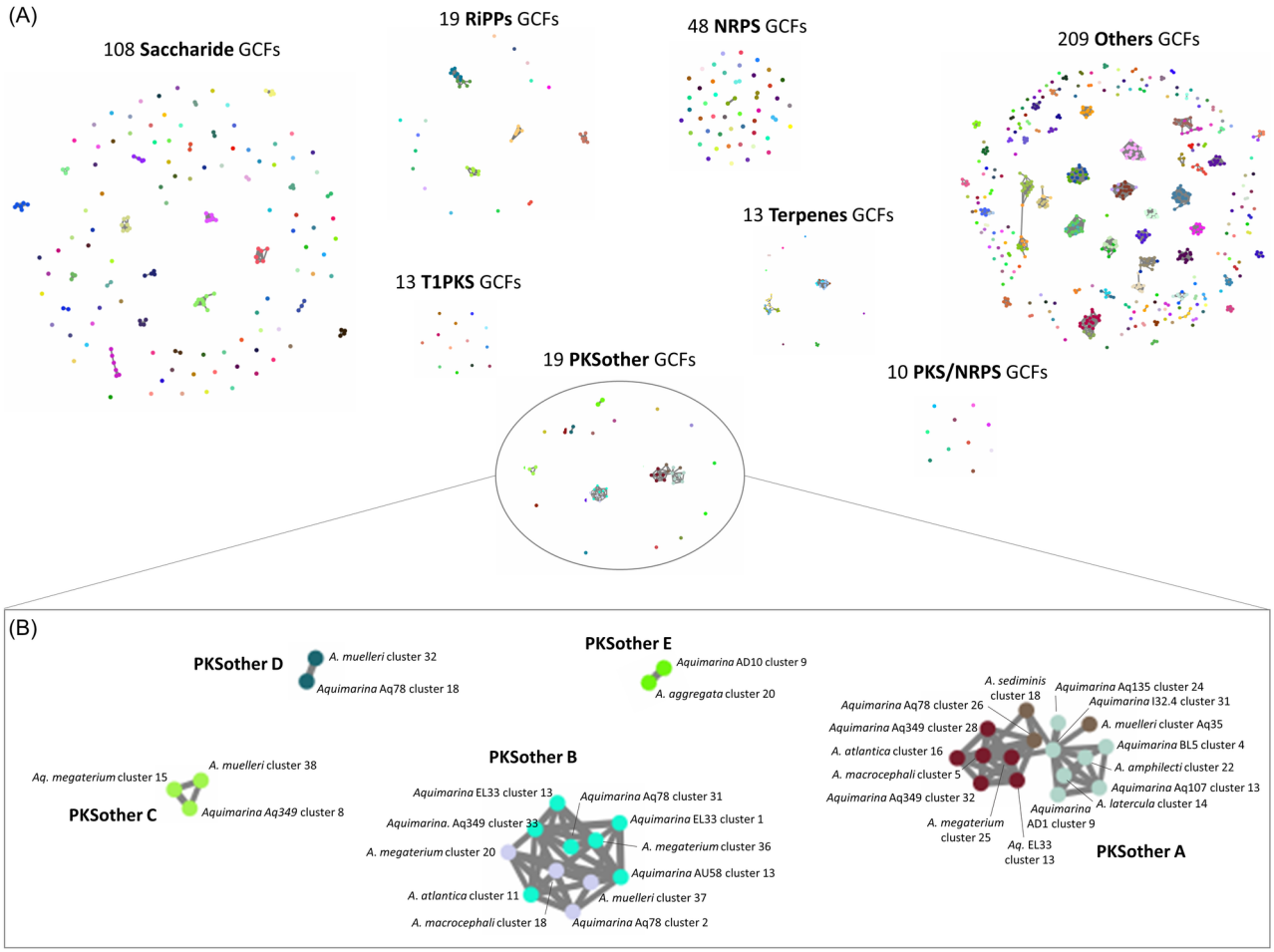
Taxon	Average dissim.	Contrib. %	Annotation	Mean Group 1*	Mean Group 2	Mean Group 3
COG3321	0,05777	0,3766	Acyl transferase domain in polyketide synthase (PKS) enzymes	0,0464	0,0606	0,00896
COG4886	0,04966	0,3238	Leucine-rich repeat (LRR) protein	0,0242	0,079	0,0457
COG2273	0,04762	0,3105	Beta-glucanase, GH16 family	0,00435	0,00335	0,0452
COG2207	0,04713	0,3073	AraC-type DNA-binding domain and AraC-containing proteins	0,11	0,169	0,155
COG3275	0,04656	0,3036	Sensor histidine kinase, LytS/YehU family	0,0615	0,115	0,088
COG1020	0,04599	0,2999	Non-ribosomal peptide synthetase component F	0,0885	0,0476	0,0541
COG3279	0,04131	0,2693	DNA-binding response regulator, LytR/AlgR family	0,0835	0,131	0,109
COG3979	0,03851	0,251	Chitodextrinase	0,0704	0,0923	0,0579
COG3501	0,03683	0,2401	Uncharacterized conserved protein, implicated in type VI secretion and phage assembly	0,0535	0,013	0,0341
COG2335	0,03568	0,2327	Uncharacterized surface protein containing fasciclin (FAS1) repeats	0,0192	0,0238	0,0523

*Values correspond to Hellinger-transformed counts of CDSs in COGs per genome group.

EMI_14747_Fig.3.tif



EMI_14747_Fig.4.tif



EMI_14747_Fig.5.tif