EXPLORING THE HABITAT DISTRIBUTION, METABOLIC DIVERSITIES AND POTENTIAL ECOLOGICAL ROLES OF CANDIDATE PHYLA "AMINICENANTES" (OP8) AND "LATESCIBACTERIA"(WS3)

By
IBRAHIM F. FARAG

Bachelor of Science in Microbiology
Ain Shams University
Cairo, Egypt
2006

Master of Science in Biotechnology
American University in Cairo
Cairo, Egypt
2012

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
DOCTOR OF PHILOSOPHY
December, 2017

EXPLORING THE HABITAT DISTRIBUTION, METABOLIC DIVERSITIES AND POTENTIAL ECOLOGICAL ROLES OF CANDIDATE PHYLA "AMINICENANTES" (OP8) AND "LATESCIBACTERIA"(WS3)

Dissertation Approved:

Dr. Mostafa S. Elshahed

Dissertation Adviser

Dr. Noha Youssef

Dr. Marianna A. Patrauchan

Dr. Wouter Hoff

Dr. Michael Anderson

Acknowledgements

I would like to express my sincere gratitude to my advisor Prof. Mostafa Elshahed for his continuous support during my Ph.D study and related research, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this dissertation. I could not have imagined having a better advisor and mentor for my Ph.D study.

Besides my advisor, I would like to thank the rest of my committee: Prof. Wouter Hoff, Dr. Noha Youssef, Dr. Marianna Patrauchan, and Dr. Michael Anderson, not only for their insightful comments and encouragement, but also for the hard question which incented me to widen my research from various perspectives. I thank my fellow lab mates Radwa Hanafy and Shelby Calkins for their continuous encouragement and support. I would like to thank the Department of Microbiology and Molecular genetics for their support.

I would like to thank my family: my parents and sisters for supporting me spiritually throughout PhD Journey and my life in general.

Finally, and most importantly, I would like to thank my wife Radwa and my sons Yahia and Yassin. Their support, encouragement, patience and love were the base of the past five years of my life.

Name: Ibrahim Farag

Date of Degree: December, 2017

Title of Study: EXPLORING THE HABITAT DISTRIBUTION, METABOLIC DIVERSITIES AND POTENTIAL ECOLOGICAL ROLES OF CANDIDATE PHYLA "AMINICENANTES" (OP8) AND "LATESCIBACTERIA"(WS3)

Major Field: MICROBIOLOGY

Abstract:

The overall goal of this study is to investigate the ecological distribution patterns, metabolic capabilities and physiological preferences of two yet uncultured bacterial phyla: The "Aminicinantes (previously called OP8) and the "Latescibacteria" (previously called WS3"). To this end, three different research projects were conducted. In the first project, we utilized 16S rRNA gene sequences available in public databases to explore the global patterns of abundance, diversity, and community structure of members of the "Aminicenantes". Our analyses revealed that "Aminicenantes" exhibits highest levels of relative abundance in hydrocarbon-impacted environments, followed by marine habitats, and aquatic, non-marine habitats. Notable preferences of members of the "Aminicenantes" to hypoxic/anoxic, as well as non-saline/low salinity habitats were also observed. Distinct patterns of "Aminicenantes" community structures were observed; and such patterns appear to be driven by habitat variations rather than prevalent environmental parameters. In the second project, a detailed genomic analysis and metabolic reconstruction effort was conducted to investigate the metabolic potential and ecological roles of four single cell derived genomes that belonging to the Latescibacteria". Metabolic reconstruction suggested that these cells possess an anaerobic fermentative metabolism, as well as the capability to degrade multiple polysaccharides and glycoproteins that are components of green (Charophyta and Chlorophyta) and brown (Phaeophycaea) algae cell walls. Further, the analyzed genomes suggest the ability to produce bacterial microcompartment (BMC) to sequester toxic intermediate produced during fucose and rhamnose metabolism. As well, genes for the formation of gas vesicles, flagella, type IV pili, and oxidative stress response were also identified. In the third project, we investigated the pangenomic diversity of the candidate phylum "Latescibacteria" (WS3) in a wide range of metagenomic data sets using a fragment recruitment strategy. We identified 68.9 Mb of "Latescibacteria"-affiliated contigs in publicly available metagenomic data sets comprising 73,079 proteins. Metabolic reconstruction of this "Latescibacteria" metagenome suggests a prevalent saprophytic lifestyle in all "Latescibacteria" orders, with marked capacities for the degradation of proteins, lipids, and polysaccharides predominant in plant, bacterial, fungal/crustacean and eukaryotic algal cell walls. Interestingly, genes and domains suggestive of the production of a cellulosome were identified in genomic fragments recovered from four anoxic aquatic habitats; hence extending the cellulosomal production capabilities in Bacteria beyond the Gram-positive Firmicutes. In addition to fermentative pathways, a complete electron transport chain with the capacity to operate under high oxygen as well as low oxygen tension was identified in fragments recovered from oxygenated and partially/seasonally oxygenated aquatic habitats. Overall, this work expanded our knowledge regarding the ecology, physiology and metabolic capabilities of two yet-uncultured microbial phyla.

Table of Contents

List of Figures

List of tables

Preface

During the last quarter century, culture-independent 16S rRNA gene diversity surveys have been extensively utilized to characterize microbial communities in a wide range of habitats. Collectively, these studies have demonstrated that the scope of phylogenetic diversity within the microbial world is much broader than previously implied by culturing-based approaches. Many of the recovered 16S rRNA gene sequences were phylogenetically unrelated to any known microbial phyla, hence necessitating coining the term candidate division, or candidate phylum (CP) to describe such sequences. Elucidating the metabolic capacities, physiological preferences, and ecological roles is one of the current grand challenges in microbial ecology.

Fortunately, recent experimental and bioinformatics advantages are allowing for the recovery of genomes of these novel CP without the need for culture enrichment and isolation procedures. Two distinct approaches could be outlined here: single cell genomics and genome recovery from metagenomics datasets. In single cell genomics, microfluidics or flow cytometry is used to physically separate a single microbial cell into a sterile microcompartment. The cell is lysed, and its genome is subsequently amplified and sequenced. Using this approach, scientists at the Joint Genome Institute (Walnut Creek, CA) have managed to recover 201 distinct single cell genomes all of which belong to various groups of uncultured bacterial and archaeal phyla from nine diverse habitats (Rinke et al 2013) [1]. These seminal studies (and subsequent efforts building on it

currently underway) represent a great step towards accessing the genomes of yet uncultured phyla.

An alternative approach to single cell genomics is utilizing various computational methods for genome recovery from metagenomics datasets. Two distinct genome recovery approaches could be highlighted here: fragment recruitment and genome-resolved metagenomics. In fragment recruitment, a customized database composed of reference genomes of the organism of interest is used as a bait to map reads from different metagenomic datasets and then extracting the reads showing primary affiliation to the targeted organisms. Then, the recruited fragments are assessed using both phylogenetic based and sequence composition based approaches through investigating single copy marker genes, tetranucleotide signatures, codon usage and GC content. This method allows for a pangenome global analysis of a target organism in a wide range of habitats and relies on already existing metagenomics datasets. However, due to its dependence on sequence similarity for target sequence recovery, the recovered sequences belonging to the target lineage are usually highly fragmented, and rarely a complete or near complete genome could not be recovered using fragment recruitment. This approach has been successfully utilized to recover and analysis fragments belonging to the yet-uncultured candidate division “Latescibacteria”, and analysis of the recovered sequences have yielded important insights regarding the metabolic abilities of this phylum (highlighted in chapter 3 in this thesis)

Genome-resolved metagenomics involves the implementation of three main steps (assembly, binning and genome quality checking) to recover partial or complete genomes from metagenomics datasets. Starting from short high throughput sequencing reads, large

genomic contigs are first constructed using general assembly tools. Since the assembly process often yield short contigs, these genomic contigs need to be grouped into genomic bins. Multiple softwares and pipelines were developed to facilitate the critical process of contigs into genomes using multiple diagnostic criteria e.g. tetranucleotide frequency, GC content, and coverage (e.g. MaxBin, MetaBAT, and GROOPM). Finally, multiple quality control programs are utilized to check on the accuracy of the assembled genomes, usually by identifying the frequency of occurrence and phylogenetic affiliation of marker genes to identify potentially contaminant fragments in a genome assembly. Using this approach, Brown et al (Nature 2015) have recovered hunderds of genomes phylogenetically affiliated with a large and enigmatic cluster of yet-uncultured bacterial candidate phyla: The candidate phyla radiation or CPR [2]. This study reconstructed 8 complete and 789 draft genomes from CPR superphylum and the reconstructed genomes comprised >35 phyla within the superphylum, representing ~15% of the phyla present in bacterial domain. Similarly, (Parks et al, 2017) have applied this approach to extract genomes from 1,550 publicly available metagenomes, to provide a global assessment of genomic diversity using genome-resolved metagenomics [3]. This effort resulted in the recovery of ~8000 bacterial and archaeal genomes, expanding the known bacterial and archaeal lineages with 17 and 3 phyla, respectively.

The work presented in this Ph.D. thesis is a contribution to our knowledge on the global distribution patterns, ecological roles, and metabolic capacities of two yet-uncultured CP: OP8 (recently called “Aminicinantes”) and WS3 (recently called “Latescibacteria”) using a wide range of approaches. In chapter one, I was interested in resolving the global abundance, diversity and community structure of the

"Aminicinantes". To this end, I conducted a detailed *in-silico* analysis, in which 16S rRNA in public databases is scoured, and the presence and phylogenetic diversity of "Aminicinantes" 16S rRNA gene sequences were analyzed. My work shows that "Aminicenantes" members are part of the rare fraction of microbial community, ubiquitous and present in all types of the habitats. Moreover, Members of the "Aminicenantes" exhibit a distinct community structure patterns across various datasets, and these patterns appear to be, mostly, driven by habitat variations rather than prevalent environmental parameters. This work has been published in PLoS ONE.

In chapter two and three I focused on analysis of genomes and genomic fragments belonging to the "Latescibacteria" to understand their metabolic capacities and ecological roles. In chapter two, I analyzed the genomes recovered from four different "Latescibacteria" cells. I show from this detailed analysis that they are mediating" the turnover of multiple complex organic polymers of algal origin that reach deeper anoxic/microoxic habitats in lakes and lagoons. This work has been published in PLoS ONE

In chapter three, I present a global fragment-recruitment based survey in which genomic fragments belonging to the "Latescibacteria" were recovered from a wide range of habitats. Detailed analysis of these fragments show that while they are all involved in different polymer degradation (e.g. proteins, sugars and fatty acids), generally having a fermentative lifestyle, however few members showed the capability to respire oxygen at low and high concentrations. Interestingly, genes and domains suggestive of the production of a cellulosome have been identified within members of "Latescibacteria". This work has been published in Applied and Environmental Microbiology.

References

1. Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng JF, et al. Insights into the phylogeny and coding potential of microbial dark matter. Nature. 2013;499(7459):431-7. doi: 10.1038/nature12352. PubMed PMID: 23851394.

2. Brown CT, Hug LA, Thomas BC, Sharon I, Castelle CJ, Singh A, et al. Unusual biology across a group comprising more than 15% of domain Bacteria. Nature. 2015;523(7559):208-11. doi: 10.1038/nature14486. PubMed PMID: 26083755.

3. Parks DH, Rinke C, Chuvochina M, Chaumeil PA, Woodcroft BJ, Evans PN, et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat Microbiol. 2017. doi: 10.1038/s41564-017-0012-7. PubMed PMID: 28894102.

Chapter I

Global patterns of abundance, diversity and community structure of the "*Aminicenantes*" (candidate phylum OP8)

Abstract

We investigated the global patterns of abundance, diversity, and community structure of members of the *"Aminicenantes"* (candidate phylum OP8). Our aim was to identify the putative ecological role(s) played by members of this poorly characterized bacterial lineages in various ecosystems. Analysis of near full-length 16S rRNA genes identified four classes and eight orders within the "*Aminicenantes"*. Within 3,134 datasets comprising ~1.8 billion high throughput-generated partial 16S rRNA genes, 47,351 *Aminicenantes*-affiliated sequences were identified in 913 datasets. The *"Aminicenantes"* exhibited the highest relative abundance in hydrocarbon-impacted environments, followed by marine habitats (especially hydrothermal vents and coral-associated microbiome samples), and aquatic, non-marine habitats (especially in terrestrial springs and groundwater samples). While the overall abundance of the *"Aminicenantes"* was higher in low oxygen tension as well as non-saline and low salinity habitats, it was encountered in a wide range of oxygen tension, salinities, and temperatures. Analysis of the community structure of the *"Aminicenantes"* showed distinct patterns across various datasets that appear to be, mostly, driven by habitat variations rather than prevalent environmental parameters. We argue that the detection of the *"Aminicenantes"* across environmental extremes and the observed distinct community structure patterns reflect a high level of intraphylum metabolic diversity and adaptive capabilities that enable its survival and growth in a wide range of habitats and environmental conditions.

Introduction

During the last quarter century, culture-independent diversity surveys have been extensively utilized to investigate bacterial diversity in almost all accessible habitats on earth [1-5]. These surveys have collectively demonstrated that the scope of bacterial diversity is much broader than previously expected based on culture-based assessments [6,7], with a large fraction of the 16S rRNA gene sequences encountered not belonging to known cultured bacterial phyla. The term candidate phylum (CP) was thus proposed to describe such lineages [2].

One of the most important challenges facing microbial ecologists is to elucidate the putative metabolic capabilities and ecological roles of these candidate phyla, as well as the underlying ecological factors controlling their observed patterns of abundance, diversity, and community structure on a global scale. Various environmental genomics approaches have been utilized to obtain genomic fragments and partial genome assemblies from these lineages. These include construction and screening of large insert (Fosmid and BAC) libraries [8-11], direct metagenomic surveys and subsequent implementation of novel binning approaches to reconstruct genomes from metagenomic sequence data [12-15], and single cell genomics [16-19]. Collectively, these efforts have yielded valuable insight regarding the genomic characteristics and putative metabolic capabilities of multiple novel candidate phyla. Further, in several incidents, these insights were successfully utilized as a stepping-stone for enrichment and isolation of some of these lineages [20-22].

Genomic approaches are extremely valuable for deciphering putative metabolic capabilities of uncultured bacterial lineages. However, information from genomic studies

is derived from a single sampling event in a single environment, and often from a single cell within the sample [17,19]. Extrapolation of such information to imply similar capabilities and genomic features to all members and lineages within an entire bacterial phylum is hence inappropriate. This is especially true since a single bacterial phylum could exhibit a bewildering array of metabolic capabilities.

A complementary approach that has previously been utilized on an ecosystem level [23-27], but rarely utilized in a global phylocentric context, relies on using *in silico* database mining approaches to examine patterns of distribution of members of a specific candidate phyla in 16S rRNA gene diversity surveys. This approach could clarify the patterns of abundance, diversity, and community structure of the targeted lineage. This phylocentric strategy could greatly benefit from the dramatic increase in the number and size of publicly available 16S rRNA gene datasets; brought about by utilizing next generation sequencing technologies in recent ambitious initiatives to catalogue 16S rRNA gene diversity on a global scale [28-30].

Here, we describe a comprehensive examination of the global distribution of members of the *"Aminicenantes"* (candidate phylum OP8) using *in silico* database mining approaches. Our aim was to understand the putative ecological role(s) played by members of this poorly characterized bacterial lineages in various ecosystems and to demonstrate the utility of in silico database mining approaches in extracting meaningful ecological patterns from high throughput 16S rRNA gene datasets. Candidate phylum OP8 was first identified in sediments from the Obsidian Pool in Yellowstone National Park [2]. Since then, it has subsequently been identified in a wide range of terrestrial and marine habitats [31-34]. A recent study has described two near candidate phylum OP8

genome assemblies from 38 partial single cell genomes obtained from deep sediments of a brackish lake (Sakinaw lake, British Columbia, Canada), and the name *"Aminicenantes"* was proposed for this candidate phylum to highlight the high proportion of genes encoding aminolytic enzymes identified in both assemblies [16]. Our results highlight the ubiquitous nature of the "*Aminicenantes"*, and identify various environmental conditions impacting its global abundance and distribution in various habitats. We argue that these observed patterns suggest that, collectively, members of the *"Aminicenantes"* exhibit a high level of intra-phylum metabolic and adaptive diversities, and are hence capable of survival, and growth in a wide range of environmental extremes.

Materials and Methods

1. A taxonomic outline of the *Aminicenantes*.

While the candidate phylum *"Aminicenantes"* (CD-OP8) is recognized in several curated taxonomic outlines e.g. Greengenes [35] and SILVA [36], only a fairly low number of *"Aminicenantes"* sequences are deposited in these databases (109, and 12, in Greengenes and SILVA, respectively). The continuous deposition of new near full-length 16S rRNA gene sequences in GenBank database repository, coupled to the sporadic updates of curated taxonomic schemes, raises the prospect that additional *"Aminicenantes"* 16S rRNA sequences putatively representing novel high rank (class/ order) lineages have been deposited in GenBank but have yet to be included in taxonomic schemes. Therefore, as a preliminary step, we aimed to identify and classify all GenBank-deposited *"Aminicenantes"* 16S rRNA gene sequences and produce an updated and comprehensive taxonomic outline of this phylum. To this end, we queried GenBank NR database using BlastN [37], to identify the closest relatives of each of the 109 *"Aminicenantes"* sequences currently recognized in Greengenes and SILVA databases. The 500 closest relatives of each sequence were downloaded; and duplicates, sequences shorter than 800 bp, and chimeric sequences, identified using Galaxy [38], were removed. The remaining sequences (n=2955) were aligned to a collection of reference sequences representing all *"Aminicenantes"* sequences, as well as sequences from a collection of 17 phyla, and 8 candidate phyla using ClustalX [39]. The phylogenetic positions of putative *"Aminicenantes"* sequences were evaluated using Distance, Parsimony, Maximum likelihood, and Bayesian approaches as previously described [40]. Sequences were deemed representative of a new class/order within the *"Aminicenantes"* if two or more

distinct sequences remained reproducibly monophyletic and formed a bootstrap-supported independent clade upon varying the composition and size of the data set used for phylogenetic analysis [41].

2. Identification of *"Aminicenantes"* members in next Generation 16S rRNA gene datasets.

Publicly available datasets generated using high throughput sequencing technologies (Pyrosequencing and Illumina) were downloaded from MG-RAST [42], VAMPS (http://vamps.mbl.edu/index.php), and GenBank SRA[43] (through the mirror web interface of DNA Databank of Japan http://www.ddbj.nig.ac.jp) in December 2012. Preliminary analysis indicated the absence of the *"Aminicenantes"* in human and metazoan microbiome samples and hence these datasets were excluded from further analysis (with the notable exception of rumen samples which were included). In total, 3,141 datasets from 110 different studies with 1,820,857,401 distinct 16S rRNA sequences were included in the analysis. All datasets were quality screened to filter all the sequences with lengths less than 50 base pairs, sequences with ambiguous nucleotides, and sequences with hompolymer stretches more than 8 bps. Sequences were classified using classify.seqs commands package in MOTHUR v.1.29.0 [44], using Silva alignment and Greengenes classification scheme. Sequences were identified as members of the *"Aminicenantes"* using a cutoff of 70% confidence threshold, as well as by confirmation of such assignment by sporadic manual insertion of putative *"Aminicenantes"* sequences into reference phylogenetic trees as described above. The subphylum level affiliation of all high throughput *"Aminicenantes"* sequences identified were determined using the updated taxonomic scheme produced in this study using near

full length 16S rRNA gene sequences as described above. All analyses were conducted on a the HPC Cowboy super computer, a 252 compute nodes with dual six core CPUs and 32 GB RAMs server, 2 fat nodes with 256 GB RAM, GPU cards and 120 TB very fast disk storage at the OSU High Performance Computing Center at Oklahoma State University.

3. Classification of next-generation datasets according to habitat type and prevalent environmental conditions.

All datasets included in this study were classified according to two different classification schemes: habitat type as well as prevalent environmental conditions. These classifications were used to determine the ecological prevalence and distribution patterns of various members of the *Aminicenantes*. Habitat-based classification scheme involved binning all 3,141 datasets into five major habitat types: Marine, aquatic non-marine, soil, hydrocarbon-impacted, and rumen/other (dust, animal-associated habitats and air). Due to the heterogeneity of geochemical and environmental conditions observed in marine, aquatic non-marine, and soil habitats, these three habitats were further sub-classified into multiple sub-habitat types, determined through the analysis of the projects' available metadata (Table 1-1). For classification of datasets according to prevalent environmental conditions, three different classification schemes using temperature, oxygen tension**,** and salinity were utilized (Table 1-2). Classification based on prevalent pH conditions was not feasible due to the frequent absence of accurate pH metadata in a large proportion of the datasets, as well as the exceedingly low number of datasets that appear to originate from environments with preeminently low (e.g. <3) or high (e.g. >9) pH.

4. Deciphering ecological preferences and patterns of distribution of the *Aminicenantes*.

The distribution and preferences of *"Aminicenantes"* were identified by correlating *"Aminicenantes"* relative abundance (% of sequences affiliated with *"Aminicenantes"* in the dataset), diversity, and community structure to its distribution in various habitats and sub-habitats, as well as across various environmental conditions. Rarefaction curves were used to compare diversities of *"Aminicenantes"* community in different datasets as previously described [45]. We chose rarefaction curve analysis since it provides a sample size unbiased estimate of diversity and is hence useful in comparing datasets with wide variations in the numbers of sequences examined. In brief, rarefaction curves were constructed for all datasets with more than 50 sequences belonging to the *Aminicenantes*. Rarefaction curve plots were used to rank the datasets in order of diversity. Datasets with intersecting rarefaction curves were given the same rank. The datasets were ranked from one (least diverse) to 198 (most diverse) and subsequently binned into diversity categories as follows; "very low" (ranks 1-40), "low" (41-80), "medium" (81-120), "high" (121-160), and "very high" (161-198) categories. The ranks were then used to correlate *"Aminicenantes"* diversity to specific environmental factors using Spearman rank correlation and the significance of these correlations were tested in R [46].

The community structure profiles i.e. the proportion of various *"Aminicenantes"* lineages in various datasets were examined to reveal overall patterns of community structure in different habitats and under different environmental conditions. In addition, to zoom in on the patterns of *"Aminicenantes"* community structure in datasets where

"Aminicenantes" represents a significant fraction of the overall bacterial community, the community structures in datasets with more than 50 *"Aminicenantes"* sequences (n=198) were compared using principal-component analysis (PCA) and biplots were constructed using the R statistical package. In this analysis, the relative position of datasets is indicative of the level of their similarity, the directions of the class/subclass arrows are indicative of their respective maximal abundances, and the lengths of the arrows are proportional to the differential abundances of such lineages.

Results

1. A revised taxonomic outline of the *Aminicenantes*.

A total of 142 near-full length 16S rRNA *"Aminicenantes"* gene sequences were identified in GenBank NR database. Detailed phylogenetic analysis grouped the *"Aminicenantes"* sequences into four candidate classes: OP8-1, OP8-2, OP8-3 and OP8-unclassified. Candidate class OP8-1 has the largest number of near full-length *"Aminicenantes"* sequences and is comprised of five distinct orders (OP8-1_HMMV, OP8-1_SHA-124, OP8-1_OPB95, OP8-1_unclassified, and OP8-1_YNP) (Figure 1). In contrast, classes OP8-2, OP8-3, and OP8-unclassified have a lower number of near full-length sequences and are not further sub-classified into candidate orders. This revision of *"Aminicenantes"* phylogeny hence increased the number of recognized near full-length 16S rRNA gene sequences by 30.3%, and added one candidate class (OP8-3) and one candidate order (OP8-1_YNP) to the Greengenes taxonomic outline, the most detailed *"Aminicenantes"* classification scheme in curated databases.

2. Identification of members of *"Aminicenantes"* in next generation 16S rRNA gene datasets.

We used pyrosequencing- and Illumina-generated 16S rRNA gene datasets available in three publicly available gene repositories (VAMPS, GenBank, and MG-RAST) [42,43] to identify the patterns of relative abundance, diversity, and community structure of members of the *"Aminicenantes"*. Within 3,141 datasets comprising ~1.8 billion 16S rRNA gene sequences, 47,315 (0.0026%) from 918 (29.2%) different datasets were affiliated with the *"Aminicenantes"*.

3. Patterns of *"Aminicenantes"* abundance.

Overall relative abundance of *"Aminicenantes"* varied widely between various datasets, and ranged between 0 and 10.2% (encountered in MG-RAST dataset number 4455892, obtained from groundwater heavily contaminated by arsenic in the Ganges-Brahmaputra Delta region of Bangladesh, [47] (Table 1-1). Although *"Aminicenantes"* has been identified in a substantial fraction (29.2%) of examined datasets, it invariantly constituted a minor fraction of the bacterial community identified, and rarely exceeded 5% in all datasets (Table 1-1).

Based on incidence of occurrence (i.e. percentages of datasets in which sequences affiliated with the *"Aminicenantes"* were identified), and relative abundance of *"Aminicenantes"* in various datasets (Table 1-1), members of the *"Aminicenantes"* appear to be most abundant in hydrocarbon-impacted habitats, being identified in 71.4% of the datasets (10/14), with an average abundance of 0.321%. The *"Aminicenantes"* was also frequently identified in marine (21.5% of datasets) and aquatic non-marine (38.74% of datasets) habitats, with average relative abundances of 0.275%, 0.146%, respectively (Table 1-1). On the other hand, the *"Aminicenantes"* were rarely identified in soils and rumen habitats (Table 1-1).

"Aminicenantes" abundance also demonstrated distinct patterns in relation to oxygen tension, temperature, and salinity (Table 2). The *"Aminicenantes"* were most abundant in anaerobic habitats (58% of datasets, average 0.46%) e.g. Mai Po mangrove marshes in Hong Kong, heavy metal contaminated ground water in Bangladesh [47], active hydrothermal vent sediments from the Mid-Atlantic Ridge [48], anoxic sulfide and sulfur-rich terrestrial spring in southwestern Oklahoma (Zodletone spring) [40], and anoxic sediments from the Guaymas [3] and Cariaco Basins [49]. However, the

"Aminicenantes" were also identified in much lower abundance in few oxic habitats e.g. water and sediments from coastal and open ocean sites surveyed from South Atlantic to the Caribbean seabed, coastal water of western English channel [50], and soils and sediments of hypersaline lake, La Sal del Rey's in southern Texas, USA [51]. Temperature profile of *"Aminicenantes"* abundance indicated an extremely rare occurrence in low temperature terrestrial and marine habitats (e.g. in datasets from the Canadian, Alaskan and European tundra and arctic soils, as well as the Amundsen sea [50,52]), and a slightly higher preference (based on incidence of occurrence) to habitats with temperate, medium, elevated, and extremely elevated temperatures (Table 1-2). Salinity wise, *"Aminicenantes"* was present at all levels of salinities, with slightly higher relative abundances in non-saline, and low salinity habitats (Table 1-2).

4. Patterns of *"Aminicenantes"* community structure.

Examination of patterns of *"Aminicenantes"* community composition across habitats revealed several distinct patterns. For example, order OP8-1_HMMV appears to be prevalent in marine environments, where it represented 53.5% of the total *"Aminicenantes"* sequences identified in marine datasets (Figure 1-2a). Class OP8-1_unclassified appeared to be the prominent lineage in aquatic non-marine environments, where it represented 77% of the total number of *"Aminicenantes"* sequences (Figure 1-2a). Order OP8-2 was the prevalent *"Aminicenantes"* lineage in hydrocarbon-impacted environments where it represented 66% of the total number of sequences. Although extremely rare in the rumen, the *"Aminicenantes"* sequences identified in a single dataset from this habitat belonged to order OP8-1_OPB95. PCA analysis conducted on datasets with more than 50 *"Aminicenantes"* sequences (n=198, Figure 1-2b) confirmed such

patterns where most of the environments from marine origins clustered along the OP8-1_HMMV species arrow (circles in Figure 1-2b), most of the environments from aquatic non-marine origins clustered along the OP8-1_unclassified species arrow (stars in Figure 2b), and the majority of the hydrocarbon-impacted environments clustered in the direction of the OP8-2 species arrow (diamonds in Figure 1-2b).

Sub-classification of habitats (Figures 1-2c-h) further revealed additional patterns at the sub-habitat level, especially in marine, soil, and aquatic non-marine habitats systems. Within marine environments, the prevalence of OP8-1_HMMV was more pronounced in coral-associated, pelagic, and deep marine datasets (Figure 1-2c). Indeed, in marine datasets with >50 *"Aminicenantes"* sequences, OP8-1_HMMV represents the majority (more than 80%) of the total *"Aminicenantes"* sequences in all coral-associated and pelagic datasets, as well as in the majority (10 out of 13) of deep sediment datasets. OP8-1_HMMV also represented the majority of *"Aminicenantes"* sequences in a few of the coastal (three out of 15) and hydrothermal (one out of six) datasets. Accordingly, those samples clustered together along the OP8-1_HMMV species arrow in the PCA biplot (red circles representing one vent sample, green circles representing five pelagic samples, yellow circles representing ten deep sediment samples, black circles representing five coral-associated samples, and blue circles representing three coastal samples in Figure 1-2d). In the remaining marine samples, the majority of *"Aminicenantes"* datasets has a mixed community of OP8-1_HMMV and other lineages, and so had an intermediary position between species arrows in the PCA biplot. In rare cases, some datasets did not contain any OP8-1_HMMV sequences. For example, all *Aminicenantes*-affiliated sequences from three hydrothermal vent samples belonged to

the newly proposed candidate class OP8-3, and hence clustered in the direction of OP8-3 species arrow in the PCA biplot (red circles in Figure 1-2d).

Within aquatic non-marine habitats, the overall majority of *"Aminicenantes"* sequences belonged to subclass OP8-1_unclassified (Figure 1-2e). The majority (85.1% of datasets originating from the two non-saline aquatic non-marine sub-habitats (temperate freshwater lakes, and spring and groundwater samples) showed >70% of *Aminicenantes*-affiliated sequences belonging to the order OP8-1_unclassified and were hence clustered along the OP8-1_unclassified arrow in the PCA biplot (black and red stars, Figure 1-2f). However, two notable exceptions to this pattern were observed: 1. In several datasets, a mixed community of OP8-1_ unclassified with other lineages was observed (e.g. 11 freshwater lake samples had a mixed community of OP8-1_unclassified (53.6±2.4%), OP8-1_OPB95 (34.2±3.2%), and OP8-1_SHA-124 (11.3±1.9%), and one sample from a sinkhole had a mixed *"Aminicenantes"* community of OP8-1_OPB95 (43.7%), OP8-1_unclassified (32.1%), and OP8-2 (22.4%). 2. In few datasets, OP8-1_unclassified order was absent e.g. sewage samples with high abundance (> 90%) of OP8-1_OPB95 (Figure 1-2f).

While the *"Aminicenantes"* class OP8-1_unclassified was the prevalent lineage in the majority of aquatic non-marine habitats originating from temperate freshwater lakes, as well as spring and groundwater datasets; a distinct community structure was observed in aquatic non-marine habitats with low to moderate salinity (Figure 1-2e). Within these habitats, e.g. three samples from the Amazon-Guianas estuaries, and a salt marsh samples from Cabo Rojo, PR, the majority of "*Aminicenantes"*-affiliated sequences belonged to

order OP8-1_HMMV (83.7±7.98%). Accordingly, those samples clustered along the HMMV species arrow in various PCA plot (Figure 1-2f).

Finally, a relatively small number of *Aminicenantes*-affiliated sequences were present in soil samples. Those were mainly affiliated with orders OP8-1_unclassified, OP8-1_OPB95, and OP8-1_SHA-124. Some unique patterns were observed at the sub-habitat level e.g. the prevalence of OP8-1_unclassified order in samples from permafrost soils (Figure 1-2g). However, it is important to note that the *"Aminicenantes"* exhibited an extremely rare distribution in all soil datasets examined, being only identified in 14 out of 276 datasets, with an extremely low average relative abundance (0.07%). Therefore, the significance of the observed patterns, given their extreme rarity, and doubtful ecological role in soil habitats, is questionable.

We also studied the effect of environmental conditions (O_2 tension, temperature, and salinity) on *"Aminicenantes"* community structure in various datasets. When environments were classified based on their salinity, we observed a shift in the prevalence of various *"Aminicenantes"* lineages, with order OP8-1_unclassified representing the majority of *Aminicenantes*-affiliated sequences in non-saline habitats, as opposed to order HMMV in low and moderate salinity environments, and class OP8-2 in hypersaline environments (Figure 1-3a). We also observed an effect of temperature on the pattern of *"Aminicenantes"* community structure changes, where order OP8-1_unclassified and class OP8_2 dominated in low temperature and psychrophilic habitats, as opposed to orders OP8-1_OPB95 and HMMV in thermophilic and hyperthermophilic habitats (Figure 1-3b). However, the uneven number of samples belonging to each

category (Table 1-3) could possibly skew these results. Finally, no remarkable effect of O_2 tension on *"Aminicenantes"* community structure was observed (Figure 1-3c).

5. Patterns of *"Aminicenantes"* diversity.

One hundred and ninety-eight datasets with more than 50 *Aminicenantes*-affiliated sequences were included in the diversity analysis. Due to the underrepresentation of hydrocarbon-impacted sites and soils, comparison of diversities was restricted to the marine and aqueous non-marine habitats and their subcategories. Within all habitats, the levels of diversity varied widely, but marine habitats showed higher diversity than freshwater habitats (Student t-test p-value=0.037), with most of the marine environments (72%) showing medium to very high *"Aminicenantes"* diversity (Table 3). Within marine habitats, a higher average diversity rank was observed in coastal samples, and a lower average diversity was observed in hydrothermal vent samples. Indeed, coastal samples *"Aminicenantes"* diversities were significantly higher than those in all other marine environments (p-value ranging from 0.0004 to 0.041). Hydrothermal vent samples *"Aminicenantes"* diversities were significantly lower than those in coastal, and deep marine sediment samples (p-values 0.0004, and 0.04, respectively).

"Aminicenantes" diversity within aquatic non-marine environments varied, with high diversities observed in spring/groundwater samples and the single sample from a salt marsh. Significantly lower diversities were observed in samples from freshwater temperate environments (p-value=0.002).

We also correlated diversity rankings to environmental conditions including temperature, salinity, and oxygen tension (Table 3). Interestingly, while no clear correlation was identified between temperature, or salinity and diversity levels of

"Aminicenantes" at $OTU_{0.03}$, a positive highly significant correlation existed between the dataset diversity rank and the environment's oxygen tension (Spearman rank correlation coefficient=0.4, p-value =5.3E-9).

Table 1-1. Classification and overll patterns of "Aminicenantes" relative abundance in various habitats and sub-habitats.

Dataset type	Total datasets	Datasets with *"Aminicenantes"* (%)	Average *"Aminicenantes"* abundance (%)	Maximum relative abundance
Total datasets	3,141	918 (29.22%)	0.20%[1]	10.20%
Total 16S rRNA sequences	1,820,857,401	47,315	0.0026%	
Marine datasets	1,154	248 (21.50%)	0.28%	5.28%
Deep marine sediments	32	30	0.50%	2.89%
Coral associated microbiome	19	10	0.89%	4.67%
Pelagic	390	40	0.20%	2.46%
Hydrothermal vents	101	60	0.23%	5.28%
Coastal	612	107	0.20%	1.87%
Aquatic non-marine datasets	1,665	645 (38.74%)	0.15%	10.20%
Spring and ground water	25	10	2.80%	10.20%
Temperate freshwater	1569	587	0.11%	2.50%
Salt marshes	71	48	0.03%	0.67%
Soil datasets	276	14 (5.072%)	0.07%	0.80%
Agriculture	28	2	0.03%	0.06%
Grassland	140	10	0.00%	0.00%
Heavy metal/hydrocarbon contaminated	8	1	0.00%	0.01%
Arid and Semi-arid	46	0	0%	0%
Permafrost	54	1	0.01%	0.01%
Hydrocarbon-impacted datasets	14	10 (71.43%)	0.32%	0.95%
Herbivorous gut and other datasets[2]	32	1 (3.125%)	0.02%	0.02%

[1]Average abundance values in datasets where *"Aminicenantes"* sequences were identified.
[2] 26 Datasets were designated "other"; these datasets originated from dust, air and animal associated habitat.

Table 1-2 Patterns of "Aminicenantes" relative abundance in datasets classified by prevalent environmental conditions.

Dataset type	Total datasets	Datasets with *"Aminicenantes"* (%)	Average *"Aminicenantes"* abundance (%)[1]	Maximum relative abundance
Oxygen Tension				
Oxic	2,787	735(26.4%)	0.10%	2.50%
Hypoxic	101	35 (34.65%)	0.17%	2.90%
Anoxic	253	148 (58.5%)	0.46%	10.20%
Temperature[2]				
Low	317	4 (1.26%)	0.004%	0.01%
Temperate	2657	807 (30.372%)	0.19%	10.20%
Medium	53	48 (90.56%)	0.02%	0.06%
Elevated	11	6 (54.55%)	0.06%	0.20%
Extremely elevated	103	53 (51.46%)	0.24%	5.28%
Salinity[3]				
Non-Saline	1,863	575 (30.86%)	0.16%	10.20%
Low Salinity	1,179	274 (23.24%)	0.26%	5.30%
Moderate salinity	77	51 (66.23%)	0.02%	0.20%
Hypersaline	22	18 (81.81%)	0.07%	0.68%

[1]Average abundance values in datasets where *"Aminicenantes"* sequences were identified.
[2] Temperature classifications: Low: Arctic, Antarctic, subarctic, and permafrost marine and terrestrial conducive to the growth of psychrophilic microorganisms; temperate: Habitats in temperate ecosystems e.g. lakes, soils in continental settings; Medium: Habitats with temperatures around 37^0C e.g. rumen; Elevated: habitats with temperatures conducive to the growth of thermophiles (50-80^0C) e.g. Alberta oil sands tailings pond; Extremely elevated: habitats conducive to the growth of hyperthermophiles (>80^0C degrees) e.g. Hydrothermal vents.
[3] Salinity classifications: Non-saline: Environments with <1% salinity; Low salinity: Marine environments, and environments with comparable salinities; Moderate salinities: Environments with salinities around 5-15% e.g. Alberta oil sands tailings pond and Huabei Oilfield in China; Hypersaline: Environments with >15% salinity.

Table 1-3 Diversity rankings of all datasets classified according to habitat and prevalent environmental conditions.

Habitat/ Environmental parameter	Average diversity rank±SD	Number of samples belonging to this diversity rank				
		Very low	Low	Medium	High	Very high
Marine	115±68.5	11	1	6	10	12
Pelagic	89.9±63.3	1	0	2	1	0
Coastal	169.1±39.6	0	1	0	1	7
Coral	115.4±57.3	1	0	1	3	0
Deep_sed	106.4±65.9	5	0	3	4	5
Hyd_vent	34.3±59.1	4	0	0	1	0
Non-marine	96±52.8	26	38	34	31	21
Freshwater	93.7±51.6	26	38	34	30	18
Spring/GW	187.5±8.8	0	0	0	0	3
Salt marsh	148	0	0	0	1	0
Hydrocarbon-Impacted Soil	96.5±83.4	3	0	0	0	4
Soil	174.5	0	0	0	0	1
Salinity						
Non-saline	95.8±52.4	29	38	34	29	24
Low-salinity	115.8±69.6	11	0	6	10	14
Hypersaline	115.3±47.4	0	1	0	2	0
Temperature						
Temperate	102.6±56.1	33	39	40	40	38
Elevated	24.3±47	7	0	0	1	0
O2 tension						
Anoxic	88.3±74.5	12	1	1	6	7
Hypoxic	84.6±44.1	26	36	37	27	5
Oxic	157.4±49.5	2	2	2	7	26

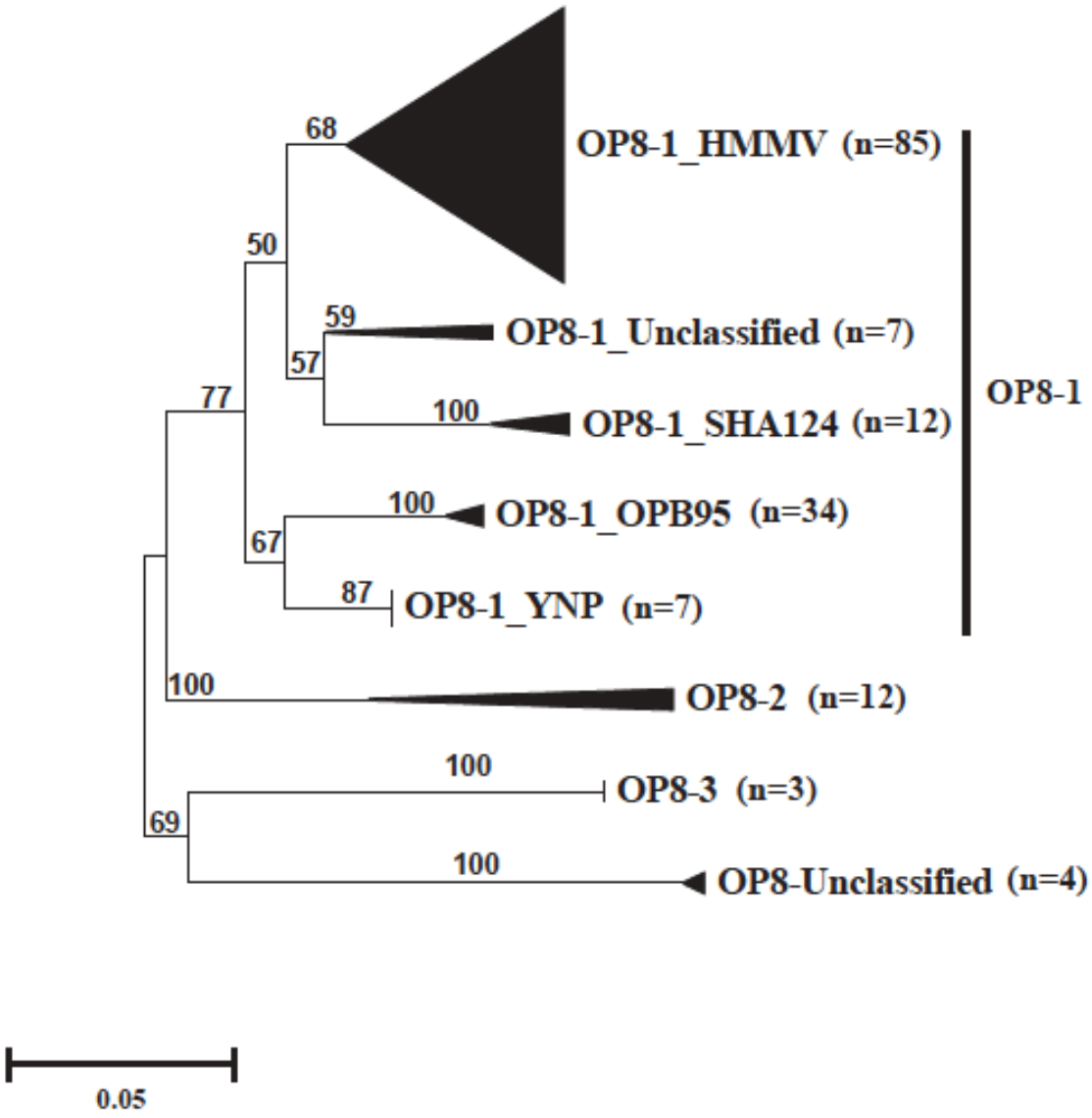


Figure 1 An updated taxonomic outline of "Aminicenantes"

The Distance NJ tree was constructed using Jukes-Cantor corrections in MEGA5 [64]. Bootstrap values (in percent) are based on 1000 replicates and are shown for branches with more than 50% bootstrap support. Numbers in parentheses represent the number of sequences in each OP8 sub-phylum.

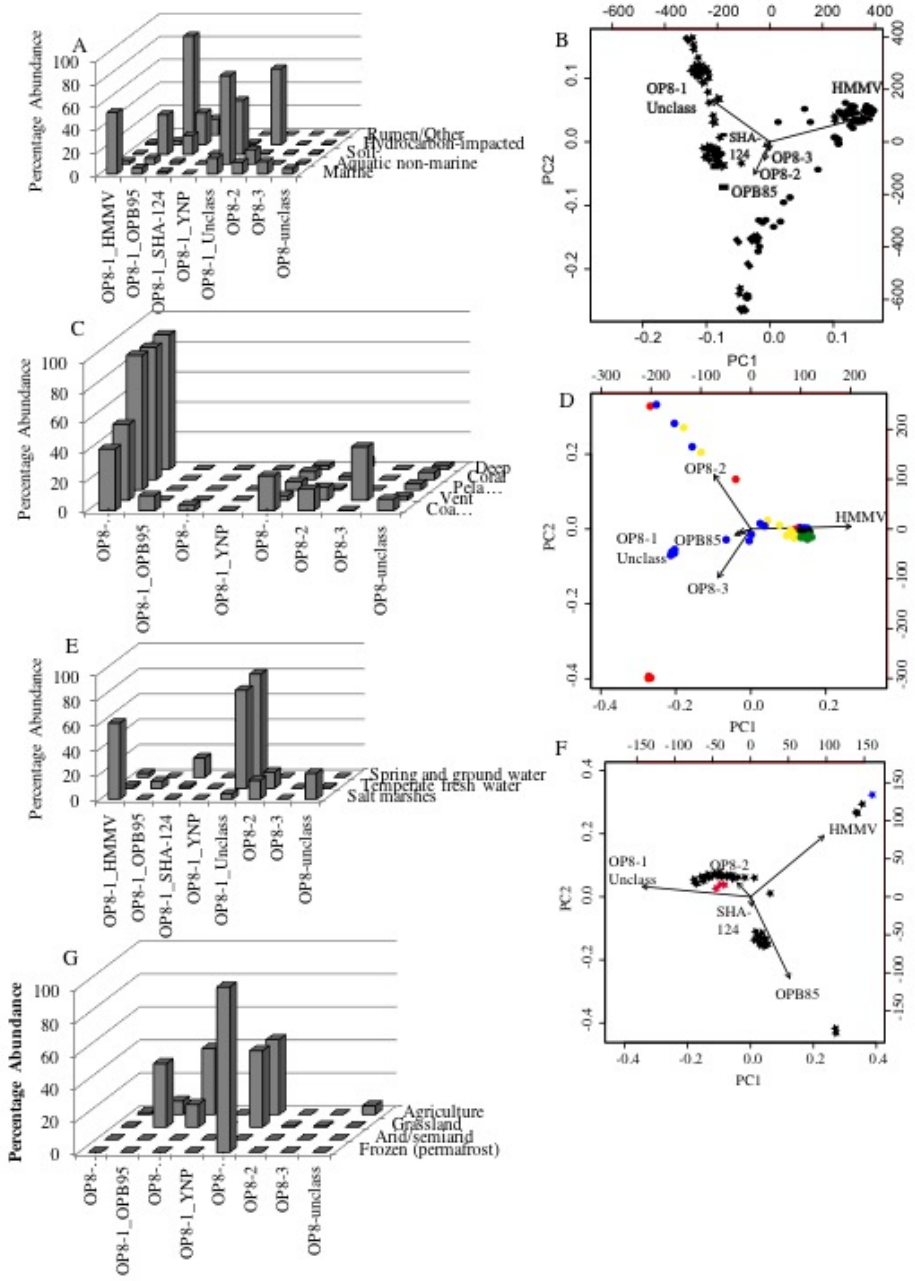

A
Percentage Abundance
Rumen/Other
Hydrocarbon-impacted
Soil
Aquatic non-marine
Marine
OP8-1_HMMV
OP8-1_OPB95
OP8-1_SHA-124
OP8-1_YNP
OP8-1_Unclass
OP8-2
OP8-3
OP8-unclass
B
PC1
PC2
OP8-1
Unclass
HMMV
SHA-
124
OP8-3
OP8-2
OPB85
C
Deep
Coral
Pela…
Vent
Coa…
D
OP8-2
HMMV
OP8-1
Unclass
OPB85
OP8-3
E
Spring and ground water
Temperate fresh water
Salt marshes
F
HMMV
OP8-1
Unclags
OP8-2
SHA-
124
OPB85
G
Agriculture
Grassland
Arid/semiarid
Frozen (permafrost)

Figure 1-2. "Aminicenantes" relative abundance and community structure

(A) Relative abundance of Aminicenantes-affiliated sequences in marine, aquatic non-marine, soil, hydrocarbon-impacted, and rumen/other habitats. (B) PCA biplot of the community structure of Aminicenantes in datasets belonging to marine (●), aquatic non-marine (★), soil (n), hydrocarbon-impacted (u), and rumen (⬢) with >50 Aminicenantes sequences. The biplot was generated in R using the prcomp and biplot functions in library labdsv. The first 2 axes explained 73% of the variance. There are two sets of axis scales on the biplot; the ones on the right and top correspond to the axis scores for samples, and the bottom and left axes correspond to the loadings of the variables (in this case, OP8 subphyla). (C) Relative abundance of *Aminicenantes*-affiliated sequences in various marine subhabitats. (D) PCA biplot of the community structure of *Aminicenantes* in marine datasets classified as coastal (blue), pelagic (green), hydrothermal vent (red), coral-associated (black), and deep sediment (yellow). There are two sets of axis scales on the biplot; the ones on the right and top correspond to the axis scores for samples, and the bottom and left axes correspond to the loadings of the variables (in this case, OP8 subphyla). (E) Relative abundance of *Aminicenantes*-affiliated sequences in environments originating from aquatic non-marine habitats. (F) PCA biplot of the community structure of *Aminicenantes* in aquatic non-marine datasets classified as freshwater (black), spring and groundwater (red), and salt marshes (blue). There are two sets of axis scales on the biplot; the ones on the right and top correspond to the axis scores for samples, and the bottom and left axes correspond to the loadings of the variables (in this case, OP8 subphyla). (G) Relative abundance of *Aminicenantes*-affiliated sequences in environments originating from soil habitats. Since only one soil dataset contained >50 *Aminicenantes* sequence, a PCA soil biplot is not feasible.

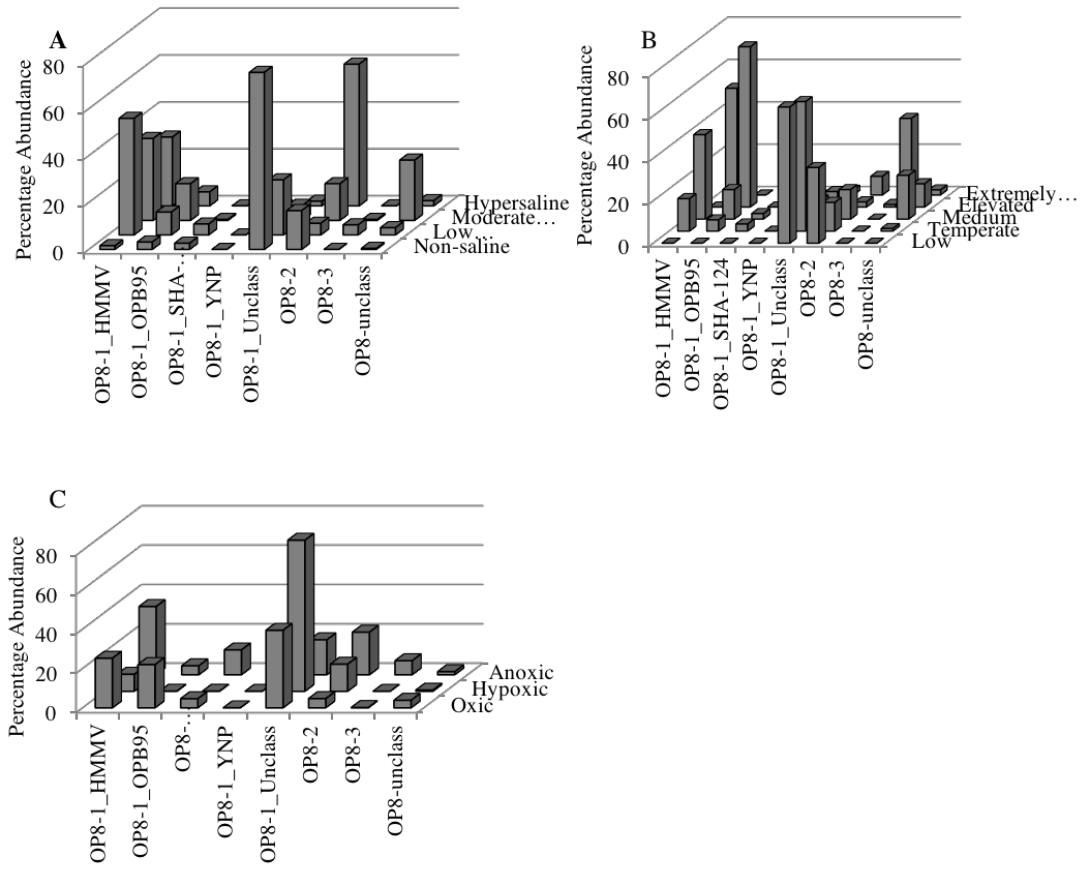


Figure 1-3. Relative abundance of "Aminicenantes"-affiliated sequences in different environments sub-classified according to different parameters

(A) Temperature, (B) oxygen tension, and (C) salinity.

Discussion

In this study, we utilized *in silico* database mining approaches to provide an updated and expanded taxonomic outline of the candidate phylum *"Aminicenantes"* using near full-length 16S rRNA gene sequences, as well as to examine the global patterns of *"Aminicenantes"* distribution using high throughput (Pyrosequencing and Illumina) generated 16S rRNA gene datasets. We report that: 1. Members of the *"Aminicenantes"* are present in a substantial fraction (918 out of 3,141) of high throughput-generated datasets examined, where they represent a minor/rare fraction of the community, with very few exceptions. 2. Members of the *"Aminicenantes"* are ubiquitous, being encountered in all different types of habitats and across all spectra of environmental parameters (temperature, salinity, and oxygen tension) examined. 3. Distinct differences exist between the relative abundance of the *"Aminicenantes"* across different habitats and environmental conditions. 4. Members of the *"Aminicenantes"* exhibit a distinct community structure patterns across various datasets, and these patterns appear to be, mostly, driven by habitat variations rather than prevalent environmental parameters.

Utilizing high throughput-generated datasets of partial 16S rRNA gene sequences in dedicated sequence repositories (VAMPS, MG-RAST, and GenBank SRA) for analyzing patterns of prokaryotic diversity represents an extremely valuable, yet largely overlooked, resource. Next generation sequencing datasets are often deposited with a single accession number per dataset, often with inadequate metadata, and, unlike Sanger-generated sequences, these datasets are not readily amendable to online search queries. Nevertheless, when properly exploited, these datasets represent an excellent resource for

testing specific ecological hypothesis. Examining *"Aminicenantes"* diversity in 3,141 distinct datasets, comprising a total of ~1.8 billion partial sequences clearly demonstrates the presence of members of this candidate phylum in a large number (29.2% of datasets examined) of habitats. However, the *"Aminicenantes"* always represented a minor fraction of the overall community and often exhibited an extremely rare distribution: The relative abundance of the *"Aminicenantes"* was less than 0.01% of the total community in 70.1% of datasets examined, 0.01-0.1% in 16.1% of datasets examined, 0.1-1% in 12.9% of datasets, and more than 1% in only 0.9% of datasets examined. The reason for the occurrence, survival, and retention of various lineages as members of the rare biosphere (e.g. less than 0.1%) in various environments is an issue that has previously been thoroughly debated [6,53,54]. Possible reasons explaining this phenomenon vary and range between filling very specialized niches, acting as a backup system that readily responds to seasonal variations encountered in various ecosystem, exhibiting extremely slow growth or dormancy, introduction to the ecosystem through recent immigration of these rare phylotypes to the sampling site, or introduction to the dataset through contamination during sampling, DNA extraction, or amplification. Indeed, several of these explanations are plausible to elucidate the role of extremely rare members of the *"Aminicenantes"* in their respective ecosystems. Regardless, it is reasonable to assume that the detection of the *"Aminicenantes"* above a certain empirical threshold (e.g. 1%, equivalent to 10^5 cells/gram or ml in a community with a cell count of 10^7) reflects its successful colonization and propagation in a specific habitat, and suggests its importance in fulfilling vital ecosystem services that justifies its retention in that habitat. Therefore, examination of the few datasets in which the *"Aminicenantes"* are present in relatively

higher abundances could offer a window on what factors are conducive for *"Aminicenantes"* survival and propagation *in-situ*. Datasets with more than 1% *"Aminicenantes"* relative abundance (0.9% of the total number of datasets) were not restricted to one habitat type or one environmental condition, but occurred within the majority of the five habitats examined, and across a wide range of environmental conditions. Therefore, it is improbable that a single, specific, environmental condition e.g. hypersalinity or extreme temperature represents the only scenario for eliciting a competitive niche for the *Aminicenantes*. Rather, we argue that conditions at which *"Aminicenantes"* propagates appear to be induced by other types of natural or anthropogenic stressors, which effectively preclude a large fraction of the population, opening the window for *"Aminicenantes"* to propagate. This is apparent from the fact that many of the datasets with >1% *"Aminicenantes"* relative abundance came from environments with variable types of environmental stressors e.g. high levels of hydrocarbons (e.g. Alberta oil sands tailing ponds, Petroleum reservoirs in Huabei, China, and north slope oil facility) [55-57], or high levels of metal (arsenic) contamination in Araihazar, Bangladesh [47].

Overall relative abundance of the *"Aminicenantes"* appeared to vary widely across various habitats, as well as across specific environmental conditions. The *"Aminicenantes"* appear to be most abundant in hydrocarbon-impacted environments, being encountered in 71.4% of the datasets (10/14), with an average abundance of 0.321%. The association of specific lineages and phylotypes with hydrocarbon-impacted environments regardless of its origin (natural or anthropogenic), or chemical composition (natural gas, petroleum, enrichments on a single substrate) has previously been noted

([58,59]. This prevalence in hydrocarbon-impacted settings is in agreement with the notion that success and propagation of members of the *"Aminicenantes"* in a specific environment is contingent on the occurrence of specific environmental stressors (hydrocarbon contamination and possibly associated anaerobiasis and high sulfide levels in such habitats) that partially alleviates competition, allowing for successful propagation of members of the *Aminicenantes*. The *"Aminicenantes"* were also identified in a considerable fraction of marine and aquatic non-marine habitats (Table 1). However, the complexity and variability of geochemical parameters encountered in these heterogeneous ecosystems prevents us from deciphering what exact environmental characteristics, or combination thereof, within these habitats favored *"Aminicenantes"* propagation. Correlating *"Aminicenantes"* abundance to environmental conditions (temperature, salinity, and oxygen tension) revealed that while members of the *"Aminicenantes"* could be encountered in a wide range of environmental conditions, it appears to exhibit significantly higher abundances in anoxic (compared to oxic and microoxic) habitats and a significantly lower abundance in low temperature (compared to temperate and elevated temperature) habitats. The relatively higher abundance of the *"Aminicenantes"* in anoxic environments suggests a prevalent anaerobic/facultative mode of metabolism within the *Aminicenantes*. Indeed, the majority of studies where the *"Aminicenantes"* represented more than 1% the total bacterial community originated from seemingly anaerobic habitats (e.g. arsenic contaminated ground water from Bangladesh, Guayamas methane seeps, and hypoliminion sites in Lake Mendota).

Analysis of the *"Aminicenantes"* community structure was conducted by: 1. Utilizing the classification of all (47,351) next generation sequences identified to

examine the *"Aminicenantes"* community structure in various types of habitats and across various environmental conditions, and 2. PCA analysis of the *"Aminicenantes"* community structure in datasets where they exhibited relatively higher abundances (n>50). Overall, it appears that factors impacting *"Aminicenantes"* community structure are mostly habitat-driven (i.e. similar community structure observed in similar habitats), rather than driven by prevalent environmental conditions (temperature, salinity, oxygen tension) within an ecosystem. For example, class OP8-3 was exclusively identified in hydrothermal vent habitats; order OP8-1_ HMMV represented the majority of *"Aminicenantes"* sequences encountered in coral associated, pelagic, and deep marine habitats; OP8-1_unclassified represented the majority of sequences in aqueous non-marine habitats; and OP8-2 represented the majority of sequence in hydrocarbon-impacted habitats. The role of prevalent environmental condition in shaping the *"Aminicenantes"* microbial community is less certain, mostly due to the inadequate representation of special categories e.g. normal (body) temperature, elevated temperature, and hypersaline environments. However, one notable exception in which an environmental parameter appears to play a clear role in shaping the *"Aminicenantes"* microbial community is the distinct prevalence of order OP8-1_ HMMV in multiple low salinity datasets regardless of their habitat.

Finally, it is interesting to note that *"Aminicenantes"* sequences were identified across all ranges of salinity and temperatures including those conducive to the growth of obligate halophiles and hyperthermophiles, respectively. Collectively, the detection of *Aminicenantes*-affiliated sequences across environmental extremes, coupled to their observed ubiquitous distribution on a global scale and the distinct patterns of community

structure exhibited argues for a high level of intraphylum metabolic and adaptive diversity within the *Aminicenantes*. Therefore it is probable that *"Aminicenantes"* cells in nature exhibit multiple distinct metabolic capabilities, wide array of survival weapons, and various adaptive strategies. This, in turn, highlights the importance of obtaining multiple genomic assemblies that adequately represents the broad phylogenetic diversity of this phylum, as well as its wide environmental distribution to truly gauge the pangenomic diversity within the *Aminicenantes*. The recently acquired genomic information from single cell-based efforts from Sakinaw Lake represents admirable effort to investigate this understudied and yet-uncultured lineage. However, information from such assemblies should not be extrapolated to describe all members of the *Aminicenantes*. Indeed, the discovery of novel capabilities within well-establish lineages e.g. phototrophy amongst *Acidobacteria* [60], methane oxidation amongst the *Verrucomicrobia* [61,62], anaerobic oxidation of ammonia amongst the *Planctomycetes* [63] highlights the importance of continued efforts to decipher and expand genomic diversity within various bacterial phyla.

Acknowledgments. We would like to thank Dana Brunson at the OSU high performing computer center for technical assistance. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. This work was supported by the National Science Foundation Microbial Observatories Program (grant EF0801858).

Reference

1. Dojka MA, Hugenholtz P, Haack SK, Pace NR (1998) Microbial diversity in a hydrocarbon- and chlorinated-solvent-contaminated aquifer undergoing intrinsic bioremediation. Appl Environ Microbiol 64: 3869-3877.

2. Hugenholtz P, Pitulle C, Hershberger KL, Pace NR (1998) Novel division level bacterial diversity in a Yellowstone hot spring. J Bacteriol 180: 366-376.

3. Teske A, Hinrichs KU, Edgcomb V, de Vera Gomez A, Kysela D, et al. (2002) Microbial diversity of hydrothermal sediments in the Guaymas Basin: evidence for anaerobic methanotrophic communities. Appl Environ Microbiol 68: 1994-2007.

4. Roesch LF, Fulthorpe RR, Riva A, Casella G, Hadwin AK, et al. (2007) Pyrosequencing enumerates and contrasts soil microbial diversity. ISME J 1: 283-290.

5. Wang Y, Sheng HF, He Y, Wu JY, Jiang YX, et al. (2012) Comparison of the levels of bacterial diversity in freshwater, intertidal wetland, and marine sediments by using millions of illumina tags. Appl Environ Microbiol 78: 8264-8271.

6. Sogin ML, Morrison HG, Huber JA, Mark Welch D, Huse SM, et al. (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". Proc Natl Acad Sci U S A 103: 12115-12120.

7. Hugenholtz P (2002) Exploring prokaryotic diversity in the genomic era. Genome Biol 3: REVIEWS0003.

8. Vergin KL, Urbach E, Stein JL, DeLong EF, Lanoil BD, et al. (1998) Screening of a fosmid library of marine environmental genomic DNA fragments reveals four clones related to members of the order Planctomycetales. Appl Environ Microbiol 64: 3075-3078.

9. Treusch AH, Kletzin A, Raddatz G, Ochsenreiter T, Quaiser A, et al. (2004) Characterization of large-insert DNA libraries from soil for environmental genomic studies of Archaea. Environ Microbiol 6: 970-980.

10. Elshahed MS, Najar FZ, Aycock M, Qu C, Roe BA, et al. (2005) Metagenomic analysis of the microbial community at Zodletone Spring (Oklahoma): insights into the genome of a member of the novel candidate division OD1. Appl Environ Microbiol 71: 7598-7602.

11. Kielak AM, van Veen JA, Kowalchuk GA (2010) Comparative analysis of acidobacterial genomic fragments from terrestrial and aquatic metagenomic libraries, with emphasis on acidobacteria subdivision 6. Appl Environ Microbiol 76: 6769-6777.

12. Narasingarao P, Podell S, Ugalde JA, Brochier-Armanet C, Emerson JB, et al. (2012) De novo metagenomic assembly reveals abundant novel major lineage of Archaea in hypersaline microbial communities. ISME J 6: 81-93.

13. Albertsen M, Hugenholtz P, Skarshewski A, Nielsen KL, Tyson GW, et al. (2013) Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. Nat Biotechnol 31: 533-538.

14. Wrighton KC, Thomas BC, Sharon I, Miller CS, Castelle CJ, et al. (2012) Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. Science 337: 1661-1665.

15. Pelletier E, Kreimeyer A, Bocs S, Rouy Z, Gyapay G, et al. (2008) "Candidatus Cloacamonas acidaminovorans": genome sequence reconstruction provides a first glimpse of a new bacterial division. J Bacteriol 190: 2572-2579.

16. Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, et al. (2013) Insights into the phylogeny and coding potential of microbial dark matter. Nature 499: 431-437.

17. Youssef NH, Blainey PC, Quake SR, Elshahed MS (2011) Partial genome assembly for a candidate division OP11 single cell from an anoxic spring (Zodletone Spring, Oklahoma). Appl Environ Microbiol 77: 7804-7814.

18. Campbell JH, O'Donoghue P, Campbell AG, Schwientek P, Sczyrba A, et al. (2013) UGA is an additional glycine codon in uncultured SR1 bacteria from the human microbiota. Proc Natl Acad Sci U S A 110: 5540-5545.

19. McLean JS, Lombardo MJ, Badger JH, Edlund A, Novotny M, et al. (2013) Candidate phylum TM6 genome recovered from a hospital sink biofilm provides genomic insights into this uncultivated phylum. Proc Natl Acad Sci U S A 110: E2390-2399.

20. Tourna M, Stieglmeier M, Spang A, Konneke M, Schintlmeister A, et al. (2011) Nitrososphaera viennensis, an ammonia oxidizing archaeon from soil. Proc Natl Acad Sci U S A 108: 8420-8425.

21. Konneke M, Bernhard AE, de la Torre JR, Walker CB, Waterbury JB, et al. (2005) Isolation of an autotrophic ammonia-oxidizing marine archaeon. Nature 437: 543-546.

22. Girguis PR, Cozen AE, DeLong EF (2005) Growth and population dynamics of anaerobic methane-oxidizing archaea and sulfate-reducing bacteria in a continuous-flow bioreactor. Appl Environ Microbiol 71: 3725-3733.

23. Bates ST, Clemente JC, Flores GE, Walters WA, Parfrey LW, et al. (2013) Global biogeography of highly diverse protistan communities in soil. ISME J 7: 652-659.

24. Freitas S, Hatosy S, Fuhrman JA, Huse SM, Welch DB, et al. (2012) Global distribution and diversity of marine Verrucomicrobia. ISME J 6: 1499-1505.

25. Bergmann GT, Bates ST, Eilers KG, Lauber CL, Caporaso JG, et al. (2011) The under-recognized dominance of Verrucomicrobia in soil bacterial communities. Soil Biol Biochem 43: 1450-1455.

26. Jones RT, Robeson MS, Lauber CL, Hamady M, Knight R, et al. (2009) A comprehensive survey of soil acidobacterial diversity using pyrosequencing and clone library analyses. ISME J 3: 442-453.

27. Buckley DH, Huangyutitham V, Nelson TA, Rumberger A, Thies JE (2006) Diversity of Planctomycetes in soil in relation to soil history and environmental heterogeneity. Appl Environ Microbiol 72: 4522-4531.

28. Gilbert JA, Meyer F, Jansson J, Gordon J, Pace N, et al. (2010) The Earth Microbiome Project: Meeting report of the "1 EMP meeting on sample selection and acquisition" at Argonne National Laboratory October 6 2010. Stand Genomic Sci 3: 249-253.

29. Huse SM, Ye Y, Zhou Y, Fodor AA (2012) A core human microbiome as viewed through 16S rRNA sequence clusters. PLoS One 7: e34242.

30. Knight R, Jansson J, Field D, Fierer N, Desai N, et al. (2012) Unlocking the potential of metagenomics through replicated experimental design. Nat Biotechnol 30: 513-520.

31. Chouari R, Le Paslier D, Daegelen P, Ginestet P, Weissenbach J, et al. (2005) Novel predominant archaeal and bacterial groups revealed by molecular analysis of an anaerobic sludge digester. Environ Microbiol 7: 1104-1115.

32. Joynt J, Bischoff M, Turco R, Konopka A, Nakatsu CH (2006) Microbial community analysis of soils contaminated with lead, chromium and petroleum hydrocarbons. Microb Ecol 51: 209-219.

33. Losekann T, Knittel K, Nadalig T, Fuchs B, Niemann H, et al. (2007) Diversity and abundance of aerobic and anaerobic methane oxidizers at the Haakon Mosby Mud Volcano, Barents Sea. Appl Environ Microbiol 73: 3348-3362.

34. Dhillon A, Teske A, Dillon J, Stahl DA, Sogin ML (2003) Molecular characterization of sulfate-reducing bacteria in the Guaymas Basin. Appl Environ Microbiol 69: 2765-2772.

35. McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, et al. (2012) An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J 6: 610-618.

36. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, et al. (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res 41: D590-596.

37. Johnson M, Zaretskaya I, Raytselis Y, Merezhuk Y, McGinnis S, et al. (2008) NCBI BLAST: a better web interface. Nucleic Acids Res 36: W5-9.

38. Goecks J, Nekrutenko A, Taylor J, Galaxy T (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol 11: R86.

39. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, et al. (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23: 2947-2948.

40. Youssef N, Steidley BL, Elshahed MS (2012) Novel high-rank phylogenetic lineages within a sulfur spring (Zodletone Spring, Oklahoma), revealed using a combined pyrosequencing-sanger approach. Appl Environ Microbiol 78: 2677-2688.

41. Dalevi D, Hugenholtz P, Blackall LL (2001) A multiple-outgroup approach to resolving division-level phylogenetic relationships using 16S rDNA data. Int J Syst Evol Microbiol 51: 385-391.

42. Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, et al. (2008) The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics 9: 386.

43. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, et al. (2013) GenBank. Nucleic Acids Res 41: D36-42.

44. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, et al. (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol 75: 7537-7541.

45. Youssef NH, Elshahed MS (2009) Diversity rankings among bacterial lineages in soil. ISME J 3: 305-313.

46. Team RDC (2011) R: A Language and Environment for Statistical Computing. Reference Index. Vienna, Austria: R Foundation for Statistical Computing.

47. Legg TM, Zheng Y, Simone B, Radloff KA, Mladenov N, et al. (2012) Carbon, metals, and grain size correlate with bacterial community structure in sediments of a high arsenic aquifer. Front Microbiol 3: 82.

48. Flores GE, Campbell JH, Kirshtein JD, Meneghin J, Podar M, et al. (2011) Microbial community structure of hydrothermal deposits from geochemically different vent fields along the Mid-Atlantic Ridge. Environ Microbiol 13: 2158-2171.

49. Madrid VM, Taylor GT, Scranton MI, Chistoserdov AY (2001) Phylogenetic diversity of bacterial and archaeal communities in the anoxic zone of the Cariaco Basin. Appl Environ Microbiol 67: 1663-1674.

50. Zinger L, Amaral-Zettler LA, Fuhrman JA, Horner-Devine MC, Huse SM, et al. (2011) Global patterns of bacterial beta-diversity in seafloor and seawater ecosystems. PLoS One 6: e24570.

51. Hollister EB, Engledow AS, Hammett AJ, Provin TL, Wilkinson HH, et al. (2010) Shifts in microbial community structure along an ecological gradient of hypersaline soils and sediments. ISME J 4: 829-838.

52. Chu H, Fierer N, Lauber CL, Caporaso JG, Knight R, et al. (2010) Soil bacterial diversity in the Arctic is not fundamentally different from that found in other biomes. Environ Microbiol 12: 2998-3006.

53. Elshahed MS, Youssef NH, Spain AM, Sheik C, Najar FZ, et al. (2008) Novelty and uniqueness patterns of rare members of the soil biosphere. Appl Environ Microbiol 74: 5422-5428.

54. Pedros-Alio C (2012) The rare bacterial biosphere. Ann Rev Mar Sci 4: 449-466.

55. Stevenson BS, Drilling HS, Lawson PA, Duncan KE, Parisi VA, et al. (2011) Microbial communities in bulk fluids and biofilms of an oil facility have similar composition but different structure. Environ Microbiol 13: 1078-1090.

56. Saidi-Mehrabad A, He Z, Tamas I, Sharp CE, Brady AL, et al. (2013) Methanotrophic bacteria in oilsands tailings ponds of northern Alberta. ISME J 7: 908-921.

57. Li H, Yang SZ, Mu BZ, Rong ZF, Zhang J (2006) Molecular analysis of the bacterial community in a continental high-temperature and water-flooded petroleum reservoir. FEMS Microbiol Lett 257: 92-98.

58. Elshahed MS, Senko JM, Najar FZ, Kenton SM, Roe BA, et al. (2003) Bacterial diversity and sulfur cycling in a mesophilic sulfide-rich spring. Appl Environ Microbiol 69: 5609-5621.

59. Davis JP, Struchtemeyer CG, Elshahed MS (2012) Bacterial communities associated with production facilities of two newly drilled thermogenic natural gas wells in the Barnett Shale (Texas, USA). Microb Ecol 64: 942-954.

60. Bryant DA, Costas AM, Maresca JA, Chew AG, Klatt CG, et al. (2007) Candidatus Chloracidobacterium thermophilum: an aerobic phototrophic Acidobacterium. Science 317: 523-526.

61. Pol A, Heijmans K, Harhangi HR, Tedesco D, Jetten MS, et al. (2007) Methanotrophy below pH 1 by a new Verrucomicrobia species. Nature 450: 874-878.

62. Dunfield PF, Yuryev A, Senin P, Smirnova AV, Stott MB, et al. (2007) Methane oxidation by an extremely acidophilic bacterium of the phylum Verrucomicrobia. Nature 450: 879-882.

63. Strous M, Fuerst JA, Kramer EH, Logemann S, Muyzer G, et al. (1999) Missing lithotroph identified as new planctomycete. Nature 400: 446-449.

64. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, et al. (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28: 2731-2739.

Chapter 2

In silico analysis of the metabolic potential and niche specialization of candidate phylum "*Latescibacteria*" (WS3)

Abstract

The "*Latescibacteria*" (formerly WS3), member of the Fibrobacteres–Chlorobi–Bacteroidetes (FCB) superphylum, represents a ubiquitous *candidate phylum found in* terrestrial, aquatic, and marine ecosystems. Recently, single-cell amplified genomes (SAGs) representing the "*Latescibacteria*" were obtained from the anoxic monimolimnion layers of Sakinaw Lake (British Columbia, Canada), and anoxic sediments of a coastal lagoon (Etoliko lagoon, Western Greece). Here, we present a detailed *in-silico* analysis of the four SAGs to gain some insights on their metabolic potential and apparent ecological roles. Metabolic reconstruction suggests an anaerobic fermentative mode of metabolism, as well as the capability to degrade multiple polysaccharides and glycoproteins that represent integral components of green (Charophyta and Chlorophyta) and brown (Phaeophycaea) algae cell walls (pectin, alginate, ulvan, fucan, hydroxyproline-rich glycoproteins), storage molecules (starch and trehalose), and extracellular polymeric substances (EPSs). The analyzed SAGs also encode dedicated transporters for the uptake of produced sugars and amino acids/oligopeptides, as well as an extensive machinery for the catabolism of all transported sugars, including the production of a bacterial microcompartment (BMC) to sequester propionaldehyde, a toxic intermediate produced during fucose and rhamnose metabolism. Finally, genes for the formation of gas vesicles, flagella, type IV pili, and oxidative stress response were found, features that could aid in cellular association with algal detritus. Collectively, these results indicate that the analyzed *"Latescibacteria"* mediate the turnover of multiple complex organic polymers of algal origin that reach deeper anoxic/microoxic habitats in lakes and lagoons. The implications

of such process on our understanding of niche specialization in microbial communities mediating organic carbon turnover in stratified water bodies are discussed.

Introduction

Over the past few decades, small subunit ribosomal RNA (SSU or 16S rRNA) gene-based surveys have prompted a drastic reevaluation of the scope of phylum level diversity within the domain Bacteria. Current taxonomic outlines indicate that the majority of recognized bacterial phyla (54.1% using SILVA database [1], 65.48% using Greengenes database [2]) have no pure culture representatives (candidate phyla). Many of these candidate phyla, so-called microbial dark matter (MDM) are globally distributed and display significant levels of intra-phylum level diversity [3-7]. Recent advances in cell sorting and whole genome amplification and assembly have facilitated the acquisition of single-cell amplified genomes (SAGs) derived from numerous candidate phyla [8-16]. Metabolic reconstruction with these SAGs provides a unique opportunity to uncover the ecological and biogeochemical roles played by these enigmatic microbial groups.

One such candidate phylum is WS3 (Wurtsmith aquifer Sequences-3), whose members were first identified in a 16S rRNA gene-based survey of anoxic sediments obtained from a hydrocarbon- and chlorinated-solvents-contaminated aquifer in northern Michigan, USA in 1998 [17]. Since then, their presence has been documented across a wide range of habitats including marine hydrothermal vents, gas hydrate-bearing habitats, cold methane seeps, cave rock walls, marine sediments, soils, wastewater treatment bioreactors, deep sea hypersaline anoxic lakes, and oil-exposed microbial mats [18-28]. Recently, as part of an extensive single cell genomic study of 9 different habitats, Rinke et al. [29] reported on the recovery of four SAGs from WS3 single cells. Phylogenomic-based analysis using conserved marker genes indicated the monophyletic nature of WS3

as part of the Fibrobacteres–Chlorobi–Bacteroidetes (FCB) superphylum together with "Marinimicrobia" (SAR406), "Cloacimonetes" (WWE1), Gemmatimonadetes, and Caldithrix. The name "*Latescibacteria*" (hiding small rods) was suggested for the candidate phylum.

However, little is known about the biological capabilities of this phylum, and no systematic attempts have been made to reconstruct its metabolic potential. Thus, we here present a detailed analysis of the metabolic and physiological capabilities, and putative ecological roles of four "*Latescibacteria*" SAGs obtained from two different aquatic environments. Our analysis suggests that the *"Latescibacteria"* recovered from Sakinaw Lake and Etoliko lagoon transform algal detritus sinking from sunlit surface waters into fermentation products with the potential to contribute to microbial food webs in anaerobic waters below.

Materials and Methods

Origin of "*Latescibacteria*" SAGs. *"Latescibacteria"* SAGs analyzed in this study were obtained from two different locations [29]: Three SAGs originated from a single sample obtained from the anaerobic monimolimnion of Sakinaw lake (British Columbia, Canada) at 49°40'30"N, 124°2'2.4"W coordinates, and a depth of 120m (Gies et al 2014). A fourth SAG was obtained by sampling anaerobic sediments in Etoliko Lagoon, a coastal lagoon in the south of Aetolia-Acarnania, Greece, at the deepest point (~27.5 m) at 38°28'59.54"N, 21°19'17.44"E. Single cell sorting and lysis, whole genome amplification, identification via 16S rRNA gene sequencing of amplified genomes, as well as SAG sequencing, assemblies and estimates of genome completion were previously described [29]. The four *"Latescibacteria"* SAGs were deposited under Genbank assembly IDs: NZ_ASMB00000000.1, NZ_AQSL00000000.1, ASWY00000000.1, and AQRO00000000.1, and in Integrated Microbial Genomics (IMG) under SAG IDs: SCGC AAA252-D10, SCGC AAA252-B13 and SCGC AAA252-E07 for Sakinaw lake SAGs, and SCGC AAA257-K07 for the Etoliko lagoon SAG. These SAGs will henceforth be referred to as S-D10, S-B13, and S-E07 for Sakinaw Lake SAGs, and E-K07 for Etoliko Lagoon SAG. The type species for "*Latescibacteria*" is S-E07, for which the name *Candidatus* "Latescibacter anaerobius" has been proposed [29].

Detailed analysis was conducted on S-E07, which has the highest estimated genome completion (73.02%) among the *"Latescibacteria"* SAGs. The closely related S-B13 (94% 16S rRNA gene sequence similarity to SAG S-E07) with 57.1% estimated genome completion was used to confirm shared gene content and fill pathway holes when

needed. Only general metabolic features for SAG S-D10 (94% 16S rRNA gene sequence similarity to S-E07, and 96% to S-B13) are discussed, given its low percentage of estimated genome completeness (38.2%). Due to the observed differences between the 3 Sakinaw Lake SAGs, and the Etoliko lagoon SAG E-K07 (85-86% 16S rRNA gene sequence similarity to Sakinaw Lake SAGs), as well as its low estimated genome completion (23.02 %), analysis of SAG E-K07 was restricted to identifying variation in conserved genes or pathways between "*Latescibacteria*" SAGs from two distinct locations.

Genome annotation, general genomic features, and metabolic reconstruction. The IMG platform (http://img.jgi.doe.gov) was used for genome functional annotation. Detailed metabolic reconstruction of relevant pathways was performed with both KEGG [30] and Metacyc [31] databases. As part of the IMG annotation pipeline, CRISPR elements are detected with CRT [32] and PILERCR [33]. Predictions from both methods are concatenated and in case of overlapping elements, the shorter one is removed. Overall annotation followed procedures outlined in [13]: In brief, proteases, peptidases, and protease inhibitors were identified with Blastp against the Merops database [34]. Transporters were identified with the transporter classification database (TCDB) [35]. dbCAN HMMs [36] were used to identify carbohydrate active enzymes (CAZymes) including glycoside hydrolases (GH), polysaccharide lyases (PL), and carboxyl esterases (CE) following the classification scheme of the Carbohydrate active enzyme (CAZy) database [37].

Results

Phylogenetic affiliation and general genomic features of "*Latescibacteria*" SAGs. All four SAGS were affiliated with the candidate order PBS_III_9 based on phylogenetic analysis of the candidate phylum "*Latescibacteria*" using 1198 near-full length 16S rRNA gene sequences (Fig. 2-1A). Sakinaw lake SAGs belonged to family I, while Etoliko lagoon SAG E-K07 belonged to family VI within this order (Fig. 2-1B). General genomic features for each SAG are shown in Table 1.

Metabolic potential of Sakinaw Lake SAGs. Anabolic pathways identified in S-E07 and S-B13 include machinery for the production of amino acids, cofactors, fatty acids, purines and pyrimidines, terpenoid unit backbone, and glycerophospholipids. In addition, the SAGs encode near-complete replication, transcriptional, and translational machineries. The presence of genes for lipopolysaccharide (LPS) biosynthesis and pathway for LPS insertion in the outer membrane suggests a Gram-negative cell wall.

Catabolical pathways identified in S-E07 and S-B13 indicate a heterotrophic lifestyle. Moreover, the apparent absence of a respiratory chain suggests sole dependence on fermentative pathways and substrate level phosphorylation for coupled energy release and conservation. Both S-E07 and S-B13 encode a diverse array of carbohydrate active enzymes (CAZymes), with a conspicuous enrichment (Genes/Mbp), and diversity (number of different families) of polysaccharide lyases (PLs) (Fig. 2-2). In contrast, the SAGs are relatively depauperate in genes encoding glycoside hydrolases (GHs) including enzymes involved in the degradation of cellulose (1 putative endoglucanase (GH5), 3 putative β-glucosidases (GH3, GH116, and GH9), and no putative cellobiohydrolase), and enzymes involved in the degradation of xylans (xylanases, and β-xylosidases).

Interestingly, many of the polymers that S-E07 and S-B13 are predicted to degrade are integral components of cell walls of the green algal phyla Charophyta (most commonly encountered in freshwater habitats), and Chlorophyta (widely distributed in freshwater, marine, and terrestrial habitats), as well as the brown algal Class Phaeophyceae. Green and brown algal cell walls are complex, with a diverse array of structural fibrillar polymers enmeshed in complex matrices with crystalline polymer components (Fig. 2-3). Both S-E07 and S-B13 encode genes necessary for the conversion of these cell wall components, including pectin, alginate, ulvans, fucans, hydroxyproline-rich glycoproteins (HRGP), e.g. arabinogalactan proteins (AGP) and extensins, and xyloglucan (Table 2-2). Moreover, the SAGs also encode pathways mediating the conversion of soluble organic compounds commonly utilized for storage in algae (e.g. starch and trehalose). A more in depth description of these capabilities follows.

Algal cell wall degradation potential.

1. Pectins. Pectins are components of the amorphous matrix and outer lattice of Charophyta cell wall (Fig. 2-3A) [38]. Both S-E07 and S-B13 encode machinery for depolymerizing the pectic polysaccharide homogalacturonan (HG) (Table 2-2). They encode carboxyl esterases (CE8 and CE12) for the removal of accessory acetyl and methyl groups attached to the backbone, pectin lyase and pectate lyase (PL1, and PL10) to breakdown the backbone to oligosaccharides with 4-deoxy-α-D-galact-4-enuronosyl groups at their non-reducing ends, exopolygalacturonate lyase (PL9) to cleave digalacturonate unit, and oligogalacturonide lyase (PL22) to degrade the digalacturonate units to 5-dehydro-4-deoxy-D-glucuronate and galacturonic acid as the final end products of HG degradation [39, 40]. In addition to HG, S-E07 and S-B13 encode all the necessary

machinery to degrade rhamnogalacturonan I (RGI) (Table 2-2). These include carboxyl esterases (CE8 and CE12), rhamnogalacturonan endolyase (PL11) that attack the backbone to produce oligosaccharides with L-rhamnopyranose at the reducing end and 4-deoxy-4,5-unsaturated D-galactopyranosyl uronic acid at the non-reducing end, rhamnogalacturonan exolyase (PL11) that attacks those oligosaccharides to release the disaccharide 2-O-(4-deoxy-beta-L-threo-hex-4-enopyranuronosyl)-alpha-L-rhamnopyranose from the reducing end, and d-4,5-unsaturated β-glucuronyl hydrolase (GH88) that degrades those disaccharides to rhamnose and 5-dehydro-4-deoxy-D-glucuronate. The SAGs also encode β-galactosidase (GH42) for removal of galactosyl sugar substitutions [39, 40].

2. Alginate. Alginates are present in the brown algal cell walls enmeshing fibrillar cellulose and also in the interfibrillar layers with fucans (Fig. 2-3C) [41]. Both S-E07 and S-B13 encode PLs for the complete degradation of alginate (Table 2-2). These PLs include alginate lyases (PL6, PL15, PL17) that break down the alginate backbone producing oligosaccharides with 4-deoxy-α-L-*erythro*-hex-4-enopyranuronosyl groups at their non-reducing ends, as well as oligoalginate lyase (PL15, and PL17) that exolytically cleave these oligosaccharides into monosaccharides and releases 4-deoxy-α-L-erythro-hex-4-enopyranuronose from the non-reducing end. The produced 4-deoxy-α-L-erythro-hex-4-enopyranuronose is spontaneously converted into 5-dehydro-4-deoxy-D-glucuronate as the final end product of alginate degradation [42].

3. Fucans. In addition to pectin and alginate, S-E07, and S-B13 also encode machinery for fucan degradation. Fucans are present, together with alginates, in brown algal cell walls interfibrillar matrix (Fig. 2-3C) [41]. Fucans exhibit wide variations in chemical

structures, ranging from the highly sulfated homofucan polymers to the highly branched high-uronic-acid, low-sulfate-containing polymers (xylofucoglucan, xylofucogalactan, xylofucomannan, xylofucoglucuronan) [41]. However, mechanistic details on the degradation of fucans are still in their infancy. Genomic analysis of "*Latescibacteria*" SAGs that S-E07, and S-B13 have the capacity to transform several fucans including homofucans, sulfated-xylofucoglucan, and sulfated-xylofucoglucoronan. Indeed, a potential homofucan-degrading enzyme with sequence similarity to *Mariniflexile fucanivorans* fucoidan lyase could attack the backbone releasing unsaturated, non-sulfated fucan di- and tetrasaccharides. The SAGs also encode many α-fucosidases (GH29, and GH95), that could attack those oligosaccharides and release fucosyl residues from the reducing end. Genomic evidence for the degradation of the highly branched high-uronic-acid, low-sulfate-containing polymers include many α-fucosidases (GH29, and GH95), and one α-glucuronidase (GH67).

4. Ulvans. Ulvans are present in the amorphous interfibrillar matrix of Chlorophyta cell walls (Fig. 2-3B) [43-45]. Ulvan backbones are made of a few repeating disaccharides. However, the exact composition of ulvans is largely unknown. One important characteristic of ulvans is the presence of unusual sugars, e.g. iduronic acid, in its backbone [44]. Iduronic acid is also an important constituent of mammalian glycosaminoglycans (GAGs), e.g. heparan sulfate, dermatan sulfate, heparin [46]. "*Latescibacteria*" SAGs harbor several PLs annotated as heparin and heparan lyase (PL12 and PL21). Structural similarity in sugar composition between ulvans and mammalian GAGs such as heparin suggest that those polysaccharide lyases (annotated as PL12 and PL21 with heparinase activity) might be potential ulvan lyases responsible for

ulvan backbone cleavage to produce di- and tetrasaccharides with an unsaturated β-glucuronyl residue located at the non-reducing end [47]. SAGs also harbor several copies of unsaturated glucuronyl hydrolases (GH88) that could potentially act on the oligosaccharides produced and release 5-dehydro-4-deoxy-D-glucuronate and other sugar residues, e.g. rhamnose, and xylose, as end products.

5. Xyloglucan. Xyloglucan is a component of Charophyta and Chlorophyta cell wall usually present in association with cellulose microfibrils (Figs. 2-3A and 2-3B) [48-50]. Both S-E07 and S-B13 encode machinery to degrade xyloglucan, a component of Charophyta and Chlorophyta cell walls usually present in association with cellulose microfibrils (Figs. 2-3A and 2-3B) [48-50], including endo-β-1,4-glucanases (GH74), that cleave the xyloglucan backbone at locations of unsubstituted glycosyl moieties and give rise to a mixture of oligosaccharides, α-1,2-fucosidase (GH95), and β-galactosidases (GH2, GH42) that attack those oligosaccharides to give rise to XXXG xyloglucans. The latter oligosaccharide can be attacked by oligoxyloglucan β-glycosidase (GH3) generating isoprimeverose (Xyl-α(1,6)-Glu), and glucose. However, no homologs of oligoxyloglucan β-glycosidase were identified.

6. Hydroxyproline-rich, other O-linked, and N-linked glycoproteins. Hydroxyproline-rich glycoproteins (HRGP) are minor components in green algal cell walls (Fig. 2-3) [51, 52]. Both S-E07, and S-B13 SAGs encode β-L-arabinofuranosidase (GH127) that specifically targets arabinose residues attached to hydroxyproline in extensins [53] and release the sugar monomer arabinose. The SAGs also encode machinery for arabinogalactan protein (AGP) degradation including endo-β-1,6-galactanases (GH30) that hydrolyses the β-1,6-galactan side chains and gives rise to galactan oligosaccharides,

β-galactosidases (GH2, GH42), β-glucuronidase (GH79), α-fucosidase (GH29, GH95), and α-rhamnosidase (GH28, GH78, GH106) that attack the produced oligosaccharides and release substituting sugar monomers, e.g. galactose, glucuronic acid, fucose, and rhamnose [54]. In addition to HRGP degradation potential, the SAGs encode several α-N-acetylgalactosaminidases (GH109) that specifically release N-acetylgalactosaminyl residues from O-linked glycoproteins [55], as well as several α-mannosidases (GH38) that could potentially release mannosyl residues from N-linked glycoproteins [56]. Recently, sialic acid (neuraminic acid), a 9-carbon sugar acid was identified in green algal N-linked glycoproteins [57]. While a sialidase (GH33) homologue was not identified in the SAGs, they do encode for all the enzymes required for sialic acid degradation, including sialate O-acetylesterase, N-acetylneuraminate lyase, and N-acyl-D-glucosamine 2-epimerase that will collectively degrade sialic acid into pyruvate and N-acetyl-glucosamine (NAG).

7. Degradation of cell wall proteins. Both S-E07 and S-B13 SAGs encode multiple peptidases that can attack the peptide moiety of glycoproteins in algal cell walls (Table 2-3). The majority of these peptidases (~66% in S-E07, and 63.4% in S-B13) are thought to be nutritional, where they non-specifically break down proteins into oligopeptides (protease families C25, M06, M10, M20, M41, M48, M50, S01, S08, S09, S41, S54, and U62), dipeptides (protease family M19), and free amino acids (protease families M24, M28, S49, T03).

8. Sulfatase activity on sulfated polysaccharide. Both S-E07 and S-B13 encode multiple sulfatases (n = 14 in S-E07 and n = 3 in S-B13) belonging to the family of arylsulfatases (pfam 00884). Many of the polymers in marine algal cell walls are sulfated,

e.g. ulvans, homofucans, sulfated-xylofucoglucan, and sulfated-xylofucoglucoronan [58]. Removal of the sulfate groups from such polysaccharides prior to their degradation facilitates access of GHs and PLs to side chains and backbones [59]. The SAGs also harbor the essential anaerobic sulfatase maturation enzyme-coding gene [60] for post-translational modification of a critical Cys or Ser in the active site to a C-α-formylglycine [61].

Degradation of algal storage compounds and additional polymers of non-algal origin. In addition to algal cell wall components, both S-E07 and S-B13 encode GHs that could potentially target algal intracellular carbon storage compounds, or secreted polysaccharides sourced from other organisms. The SAGs encode GHs specific for starch (α-amylase belonging to GH119, GH57, GH13, and α-glucosidase belonging to GH97), as well as for trehalose (trehalase/maltase belonging to GH65) degradation. Starch is recognized as an important intracellular storage compound in green algae and green plants [62], while trehalose is an intracellular storage compound in brown algae [41]. In addition, S-E07 and S-B13 encode β-fructofuranosidase (GH32) specific for sucrose, and endo1,4-poly-D-galactosaminidase (GH114) specific for poly-D-galactosamine (Table 2-2).

Extracellular polymeric substance (EPS) as additional potential source of energy for the "*Latescibacteria*". EPS forms extensive mucilaginous sheath outside the algal cell wall and function in adhesion, gliding motility, biofilm formation, and protection. Although the exact chemistry of EPS is not entirely known, EPS was shown to be composed mainly of polysaccharides (up to 75%), with minor protein content (2-10%). The polysaccharide fraction is rich in uronic acids, as well as monosaccharides, mainly

glucose, galactose, mannose, xylose, arabinose, fucose, and rhamnose [63, 64]. As mentioned above, "*Latescibacteria*" SAGs harbor genes involved in the uptake and catabolism of all such components.

"*Latescibacteria*" SAGs harbor extensive transport systems for sugars, and amino acids/oligopeptides uptake. Both S-E07 and S-B13 encode several non-specific porins for transport of substrates across the outer membrane, coupled to specialized transporters in the inner membrane, including multiple secondary (symport), ABC (ATP-binding cassette), and phosphotransferase system (PTS) transporters for the uptake of a wide array of monomers, e.g. those putatively produced from the degradation of all polymers described above (Fig. 2-4, Table 2-2). Uronic acids and uronic acid derivatives are potentially imported using a single common transporter (a sugar phosphate permease transporter of the major facilitator superfamily similar to ExuT transporter of *Ralstonia solanacearum* [65]). Fucose, rhamnose, as well as xylose are potentially imported via dedicated proton symporters, while glucose and galactose are potentially imported via dedicated sodium symporters. Moreover, the SAGs encode components of dedicated ABC transporters for arabinose, ribose, and oligopeptides and dipeptides as well as components of the PTS specific for N-acetylgalactosamine, fructose, and mannose import. The SAGs also encode a complete two-component signal transduction system for sensing di/tricarboxylates, e.g. malate, citrate, (DctBD), as well as a tripartite ATP-independent di/tricarboxylate transport system (TRAP) (DctPQM) [66].

Catabolism of imported sugars. Both S-E07 and S-B13 encode extensive pathways for the catabolism of a wide array of sugars, sugar acids, amino sugars, amino acids, as well as citrate and malate. Monomer degradation pathways in the SAGs are predicted to

converge on one of three central metabolic routes, (i) feeding into the EMP pathway (for glucose, galactose, mannose, fructose, sugar acids, amino sugars, aspartate, and citrate and malate), (ii) feeding into PPP (for xylose, ribose, and arabinose), or (iii) the special fucose and rhamnose degradation pathways to propionate and propanol.

Monomer catabolism is depicted in (Fig. 2-5). Briefly, the genomes encode a complete glycolytic pathway for metabolism of various C6 sugars to pyruvate, including glucose, galactose, mannose, and fructose and the amino sugars N-acetylgalactosamine, N-acetylglucosamine, and D-galactosamine. The genomes also encode the necessary enzymes for channeling the C6 sugar acids galacturonic acid, glucuronic acid, and 5-dehydro-4-deoxy-D-glucuronate to the central metabolite 2-dehydro-3-deoxy-D-gluconate (KDG), which is subsequently converted to pyruvate and glyceraldehyde-3-phosphate (GAP), that feed into the EMP. In addition, the amino acid aspartate, as well as dicarboxylates (malate) and tricarboxylates (citrate) that could potentially serve as C and energy source are catabolized via conversion to oxaloacetate and subsequently to phosphoenolpyruvate (PEP). On the other hand, the C5 sugars xylose, ribose, and arabinose are metabolized via the non-oxidative branch of the pentose phosphate pathway by first conversion to xylulose-5-P. Collectively, the metabolism of these compunds via the EMP or the PPP results in the production of pyruvate. Pyruvate could potentially be converted to acetyl-CoA via the action of pyruvate:ferredoxin oxidoreductase. Indeed, as indicated previously, the SAGs encode the machinery necessary for substrate-level phosphorylation including acetyl CoA synthase, as well as propanediol transacetylase and acetate kinase, both of which convert acetyl-CoA to acetate with concomitant ATP production (Fig. 2-5).

Fucose and rhamnose metabolism requires a different catabolic pathway and partially occurs in an intracellular bacterial microcompartment (BMC) to protect against cellular damage by containing the reactive intermediate propionaldehyde [67, 68]. Both S-E07 and S-B13 encode a dedicated pathway for the degradation of fucose and rhamnose to lactaldehyde and dihydroxyacetone-phosphate. Several genes encoding for BMC structural shell proteins with BMC domains (pfam 00936, as well as pfam 03319) were identified in the SAGs consistent with a recent observation by Axen and colleagues exploring the taxonomic distribution of BMCs across bacterial phyla [69]. Inside the BMC, lactaldehyde is converted to 1,2-propanediol (1, 2-PD). Although homologues for 1,2-PD dehydratase, the enzyme responsible for conversion of 1,2-PD to propionaldehyde, were not identified in S-E07 and S-B13, both SAGs harbor NAD-dependent aldehyde dehydrogenase, and NADH-dependent alcohol dehydrogenase for conversion of propionaldehyde to propionyl-CoA, and propanol, respectively. Propionyl-CoA can then be converted to propionate with the concomitant production of 1 mole of ATP per propionate produced.

Additional genomic features. Both S-E07 and S-B13 encode machinery for pili and flagella production, enabling potential attachment to surfaces [70], as well as gas vesicles production for maintaining a position in the water column with the most favorable growth conditions [71]. In addition, the SAGs encode multiple oxidative stress enzymes that counter harmful effects of changing oxygen tension caused by vertical migration in the stratified water column while in pursuit of decaying algal cells or other food particles. These include rubrerythrin, rubredoxin, rubredoxin oxidoreductase, superoxide reductase (desulfoferredoxin), ferritin-like protein, NADPH-dependent alkyl hydroperoxide

reductase, and glutathione peroxidase [72], as well as machinery for bacillithiol biosynthesis, a thiol implicated in peroxide sensing [72-75].

General features of Etoliko lagoon SAG E-K07. While the Etoliko lagoon SAG E-K07 shared similar metabolic potential with respect to algal cell wall polymer degradation to the Sakinaw Lake SAGs several unique features were apparent. In addition to harboring a large genome (estimated size 7.7 Mbp, Table 1) E-K07 encodes machinery for the following: (1) Degradation of the amino acids Thr, D-Cys, Glu, and Met, (2) Neuraminidase (GH33) gene for cleavage of sialic acid residues from N-linked glycoproteins, and endo-β-1,4-glucuronan lyase (PL20) [76], that targets β-(1→4)-glucuronan, a minor polysaccharide present in green algal cell walls [77], and (3) A papain (peptidase family C01), and a hyicolysin-like peptidase (family M30), possibly involved in matrix degradation. Also, E-K07 SAG encodes several stress response pathways, signal transduction, and defense mechanisms that were not identified in Sakinaw Lake SAGs. These include (1) oxidative stress enzymes catalase and ferroxidase, (2) CRISPR-associated genes including the 6 core *cas* genes (*cas1-cas6*), as well as the CRISPR-associated *csn1* gene [78], and (3) type VI secretion system including ten of the thirteen core *tss* genes [79].

Table 2-1. General genomic features of "Latescibacteria" SAGs.

	SAGs from Sakinaw Lake			SAG from Etoliko Lagoon
	SCGC AAA252-E07	SCGC AAA252-B13	SCGC AAA252-D10	SCGC AAA257-K07
Genome size, Mb	2.3	1.49	0.5	1.77
Estimated genome completeness, %	73.02	57.09	38.17	23.02
Estimated size, Mb	3.15	2.61	1.31	7.69
GC %	42.07	40.86	40.86	42.12
% Non coding DNA	14.6	15.1	16.6	13.8
Average gene length, bp	988	947	762	980
RNA genes				
5S rRNA Count	1	1	1	1
16S rRNA Count	1	0*	1	1
23S rRNA Count	1	1	1	1
tRNA Count	27	18	10	19
Number of CDS	1951	1558	647	1534
with function prediction	1451	1158	433	1073
without function prediction	500	400	214	461

* The S-B13 16S rRNA couldn't be retrieved via the whole genome shotgun approach, however the affiliation of S-B13 to CP-"*Latescibateria"* was confirmed through analyzing the amplified and Sanger-sequenced full-length 16S rRNA gene.

Table 2-2. Polymers potentially targeted by "Latescibacteria", their distribution and occurence in algae, structure, degradation enzymes encoded in "Latescibacteria" SAGs, potential degradation products, their transport systems encoded in the SAGs and pathways.

Polymer	Distribution	Degradation	Products	Transport system	Central pathway
1. Pectin					
Homogalacturonan	Land plants and Charophyta green algae	Pectin methylesterase	Pectinate/ Pectate (demethylated)		EMP
		Pectin acetylesterase	Deacetylated polymer		
		Pectin lyase, pectate lyase	Oligosaccharides with 4-deoxy-α-D-galact-4-enuronosyl groups at their non-reducing ends		
		Exopolygalacturonate lyase	digalacturonate		
		Oligogalacturonide lyase	5-dehydro-4-deoxy-D-glucuronate	ExuT symporter	
			galacturonic acid		
Rhamnogalacturonan I		Pectin methylesterase	Demethylated RGI		
		Pectin acetylesterase	Deacetylated RGI		
		Rhamnogalacturonan endolyase	Oligosaccharides with L-rhamnopyranose at the reducing end and 4-deoxy-4,5-unsaturated D-galactopyranosyl uronic acid at the non-reducing end		
		Rhamnogalacturonan exolyase	disaccharide 2-O-(4-deoxy-beta-L-threo-hex-4-enopyranuronosyl)-alpha-L-rhamnopyranose		
		d-4,5-unsaturated β-glucuronyl hydrolase	rhamnose	Rhamnose:proton symporter	BMC
			5-dehydro-4-deoxy-D-glucuronate	ExuT symporter	EMP
		β-galactosidase	Galactose	Galactose:sodium symporter	EMP
Alginates	Brown Algae	Poly β-D-mannuronate lyase	Oligosaccharides with 4-deoxy-α-L-*erythro*-hex-4-enopyranuronosyl groups at their non-		

			reducing ends		
		Oligoalginate lyase	4-deoxy-α-L-erythro-hex-4-enopyranuronose spontaneously converted to 5-dehydro-4-deoxy-D-glucuronate	ExuT symporter	EMP
2. Fucans					
Homofucan	Brown Algae	Sulfatase	Unsulfated homofucans		
		Fucoidan lyase-like protein (hypothetical protein)	Unsaturated, non-sulfated di- and tetrasaccharides		
		α-L-fucosidase	Fucose	Fucose:proton symporter	BMC
Xylofucogalactan/ xylofucomannan		Sulfatase	Unsulfated fucans		
		α-L-fucosidase	Fucose	Fucose:proton symporter	BMC
		β-glucuronidase	Glucuronic acid	ExuT symporter	EMP
Xylofucoglucuronan		Sulfatase	Unsulfated fucans		
		α-L-fucosidase	Fucose	Fucose:proton symporter	BMC
3. Ulvans	Chlorophyta	Sulfatase	Unsulfated ulvans		
		Heparin lyase	Unsaturated, non-sulfated di- and tetrasaccharides		
		d-4,5-unsaturated β-glucuronyl hydrolase	5-dehydro-4-deoxy-D-glucuronate	ExuT symporter	EMP
			Rhamnose	Rhamnose:proton symporter	BMC
			Xylose	Xylose:proton symporter	PPP
4. Xyloglucan	Land plants, some green algae	endo-β-1,4-glucanase	A mixture of oligosaccharides		
		α-1,2-fucosidase	Fucose	Fucose:proton symporter	BMC
		β-galactosidase	Galactose	Galactose:sodium symporter	EMP
		β-glucosidase	Glucose	Glucose:sodium symporter	EMP
5. Hydroxyproline-rich glycoprotein (HRGP)					
Extensin	Land plants, some green algae	β-L-arabinofuranosidase	Arabinose	ABC transporter	PPP
Arabinogalactan protein (AGP)		endo-β-1,6-galactanase (?)	galactan oligosaccharides		
		β-glucuronidase	Glucuronic acid	ExuT symporter	EMP
		α-Fucosidase	Fucose	Fucose:proton symporter	BMC
		α-rhamnosidase	Rhamnose	Rhamnose:proton symporter	BMC
		β-galactosidase	Galactose	Galactose:sodium symporter	EMP
6. Others					
Extracellular proteins	All organisms	Non-specific	Oligopeptides	ABC transporter	

		endopeptidases			
		Dipeptidases	Dipeptides	ABC transporter	
		Dipeptide peptidase, aminopeptidases, or Carboxypeptidases	Free amino acids	Symporters for Pro, Ala, Asp, Glu, Gly, cationic aa ABC transporter for Pro	EMP (Asp and Glu)
Starch	Land plants, and green algae storage compounds	α-amylase	Oligosaccharides		
		α-glucosidase	Glucose	Glucose:sodium symporter	EMP
Trehalose	Brown algae storage compound	Trehalase	Glucose	Glucose:sodium symporter	EMP
			Glucose-1-P		
Poly-D-galactosamine	Some fungi such as *Aspergillus,* and *Neurocrassa*	Endo1,4-poly-D-galactosaminidase	Galactosamine	PTS	EMP

Table 2-3. Number of peptidases belonging to various Merops peptidase families identified in "Latescibacteria" genomes and their possible physiological roles.

Merops Family	Genomes		Annotation	Possible physiological function
	S-E07	S-B13		
A08	1	1	Signal peptidase II [EC: 3.4.23.36]	Protein activation
A24	1	0	Type IV prepilin peptidase 1 [EC: 3.4.23.43]	Protein activation
A31	2	1	Hydrogenase 3 maturation protease [EC: 3.4.23.51]	Protein activation
C14	2	0	Apoptosis caspase	Protein activation
C25	1	0	Gingipain	Matrix degradation
C39	0	2	Bacteriocin processing*	Activation and transport of peptide AB
C45	1	0	Isopenicillin-N N-acyltransferase [EC: 2.3.1.164]*	Protein modification
M06	2	0	Metalloprotease	Possibly nutritional, non-specific.
M10	0	1	Matrixin	Matrix degradation
M16	3	4	Signal peptidase	Protein activation
M19	1	2	Membrane dipeptidase [EC: 3.4.13.19]	Possibly nutritional
M20	2	2	Metalloprotease	Hydrolysis of the late products of protein degradation so as to complete the conversion of proteins to free amino acids. Possibly nutritional, non-specific.
M22	2	1	Hydrogenase maturation protease	Protein activation
M23	7	3	Membrane-bound metallopeptidase	Bacterial cell wall lysis. Possibly defensive or feeding mechanism
M24	1	1	Methionyl aminopeptidase [EC: 3.4.11.18]	Removal of the initiating methionine of many proteins
M28	0	5	Predicted aminopeptidase, Iap family	Removal of amino acids from N-terminus. Possibly nutritional
M41	1	1	Membrane protease FtsH catalytic subunit [EC: 3.4.24.-]	Degrading unneeded or damaged membrane proteins
M48	0	1	Endopeptidase	Degradation of abnormal proteins
M50	2	1	Intra-membrane protease	Protein activation or possibly nutritional
M56	2	3	Potential penicillin-binding protein required for induction of beta-lactamase	Antirepressor regulating drug resistance
S01	2	5	Trypsin-like serine proteases	Possibly nutritional, non-specific proteolysis
S08	4	1	Subtilisin-like serine proteases	Possibly nutritional, non-specific proteolysis
S09	2	0	Non-specific metalloprotease	Degradation of biologically active peptides. Possibly nutritional
S24	0	2	RecA-mediated autopeptidases	SOS-response transcriptional repressors

S26	1	0	Signal peptidase I [EC: 3.4.21.89]	Protein activation
S41	1	1	C-terminal processing peptidase-3	Degradation of incorrectly synthesized proteins
S49	0	1	Signal peptide peptidase A.	Degrade the signal peptide cleaved by signal peptidases. Possibly nutritional
S54	1	1	Rhomboid intra-membrane protease	Protein activation or possibly nutritional
T01	1	1	ATP-dependent protease HslVU, peptidase subunit	Turnover of intracellular proteins
T03	1	0	Gamma-glutamyltransferase.	Degradation of glutathione by cleavage of the gamma-glutamyl bond
U62	3	0	Predicted Zn-dependent protease	Possibly nutritional, non-specific proteolysis
Total	47	41		

*: Biosynthetic genes for the related antibiotic were not identified in the genome. Possibly performing a different function.

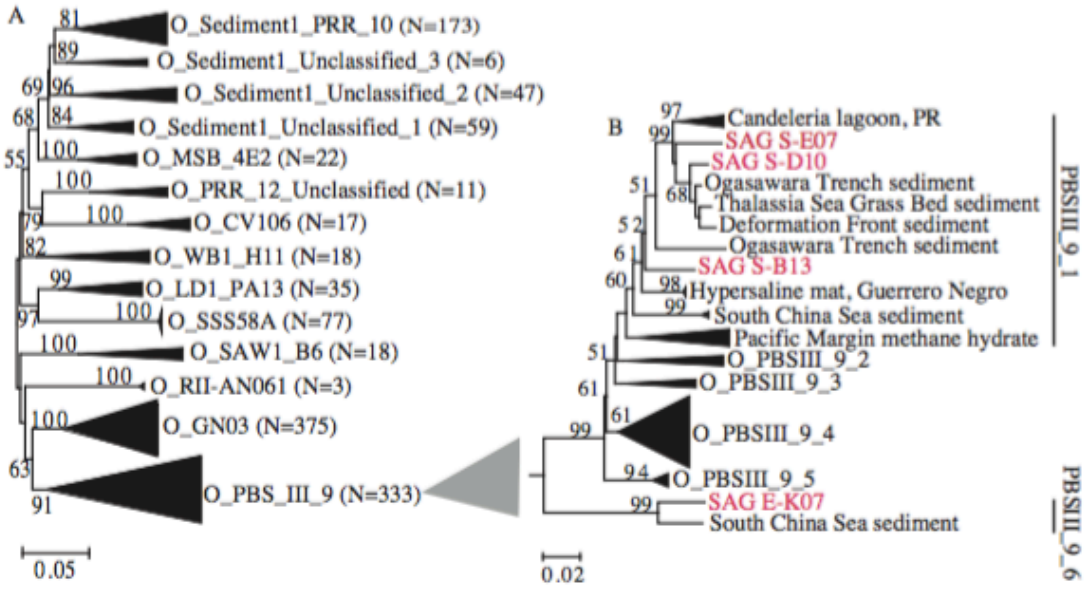


Fig. 2-1. Updated taxonomic outline for candidate phylum "*Latescibacteria*" (A), and for the candidate order PBSIII_9 (B). Neighbor joining trees were constructed using Jukes-Cantor corrections in MEGA6-Beta2 [100]. Bootstrap values (in percent) are based on 1000 replicates and are shown for branches with more than 50% bootstrap support. Numbers in parentheses represent the number of sequences in each WS3 candidate order.

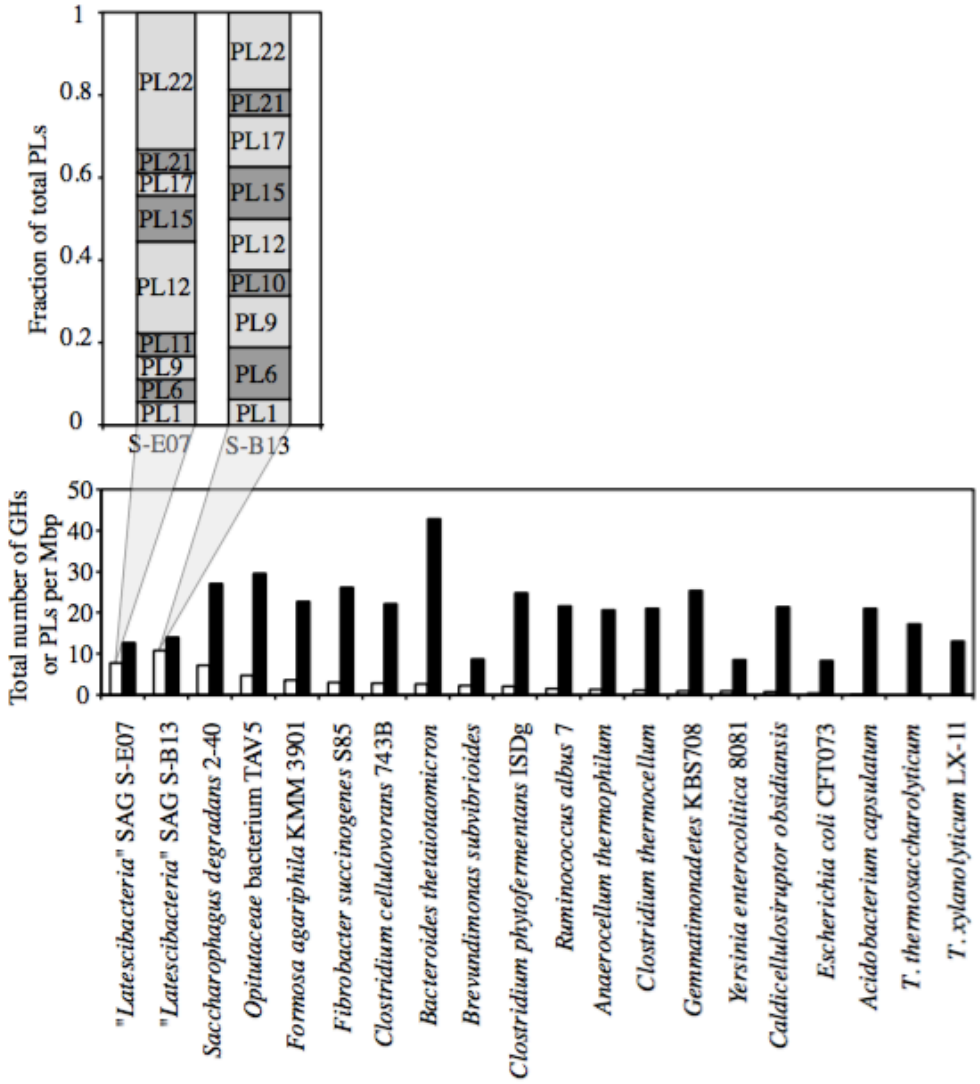


Fig. 2-2. Total number of PLs (white columns) and GHs (black columns) per Mbp of various pectinolytic and lignocellulolytic microorganisms' genomes. Note that, compared to other genomes, "*Latescibacteria*" SAGS are enriched in PLs as opposed to GHs. The inset shows SAGs S-E07 and S-B13 different PL families as a fraction of total PLs.

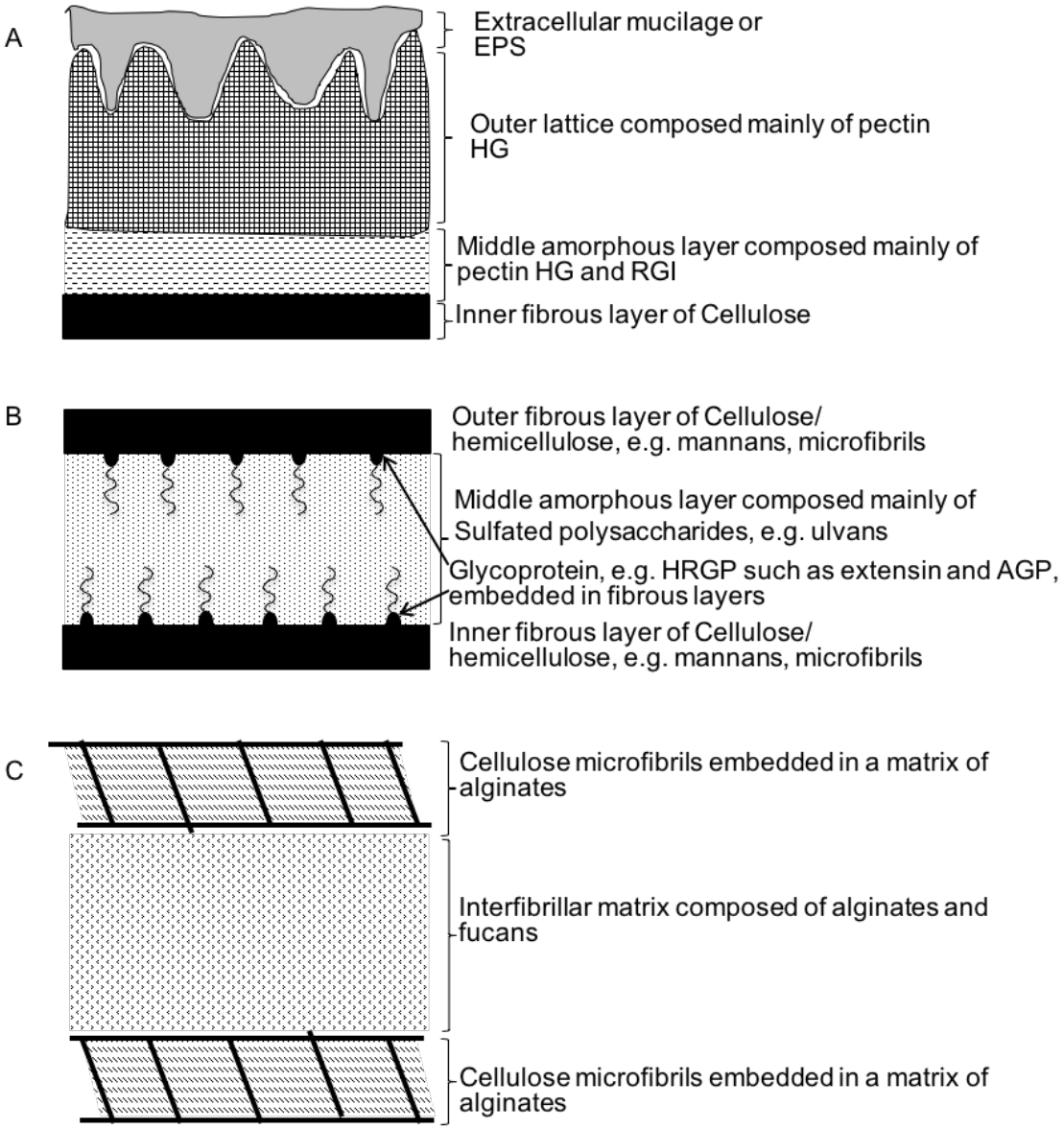

A
Extracellular mucilage or EPS
Outer lattice composed mainly of pectin HG
Middle amorphous layer composed mainly of pectin HG and RGI
Inner fibrous layer of Cellulose
B
Outer fibrous layer of Cellulose/ hemicellulose, e.g. mannans, microfibrils
Middle amorphous layer composed mainly of Sulfated polysaccharides, e.g. ulvans
Glycoprotein, e.g. HRGP such as extensin and AGP, embedded in fibrous layers
Inner fibrous layer of Cellulose/ hemicellulose, e.g. mannans, microfibrils
C
Cellulose microfibrils embedded in a matrix of alginates
Interfibrillar matrix composed of alginates and fucans
Cellulose microfibrils embedded in a matrix of alginates

Fig. 2-3. Schematic representation of algal cell walls. The cell wall composition differs between various algal groups [43]. Within the Charophyta (A), the wall is formed of an inner fibrillar layer made of cellulose microfibrils. The fibrillar layer is enmeshed in and surrounded by a middle amorphous matrix of pectin (homogalacturonan, HG, and rhamnogalacturonan I, RGI) that anchors the inner fibrillar cellulose layer to an outer lattice of homogalacturonan. Extracellular polymeric substances or mucilages are also present outside the outer lattice [38, 43, 101]. Similarly, cell walls of Chlorophyta (B) contain skeletal polysaccharides enmeshed in a matrix. However, the skeletal polysaccharides in Chlorophyta cell walls form double fibrillar layers (inner layer and outer layer) with an amorphous matrix in between. The fibrillar layers vary in composition between cellulose, β-1,3-xylans or β-1,4-mannans or complex heteropolymers, and are rich in hydroproline-rich glycoprotein such as extensins and AGPs. The amorphous matrix polysaccharides are generally in the form of ulvans (e.g. in Ulva species). Brown algal cell walls (C) consist of a fibrillar framework of cellulose microfibrils present in layers parallel to the cell surface but with no clear orientation within each layer. Two such layers are depicted in the figure. All cellulose layers are enmeshed in acidic polysaccharides, e.g. alginates. The interfibrillar matrices are composed of alginates and fucans [41, 43]**.**

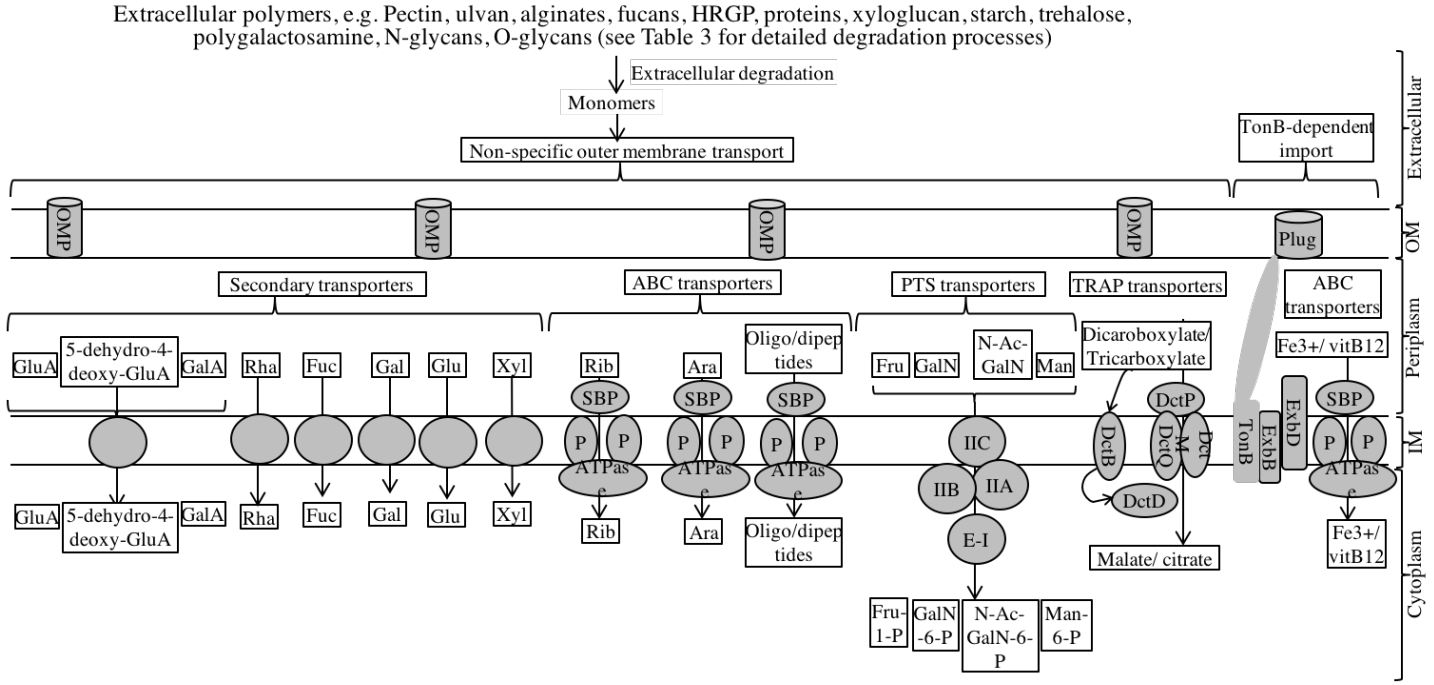

Extracellular polymers, e.g. Pectin, ulvan, alginates, fucans, HRGP, proteins, xyloglucan, starch, trehalose, polygalactosamine, N-glycans, O-glycans (see Table 3 for detailed degradation processes)
Extracellular degradation
Monomers
Non-specific outer membrane transport
TonB-dependent import
Extracellular
OM
Periplasm
IM
Cytoplasm
OMP
Plug
Secondary transporters
ABC transporters
PTS transporters
TRAP transporters
ABC transporters
GluA
5-dehydro-4-deoxy-GluA
GalA
Rha
Fuc
Gal
Glu
Xyl
Rib
Ara
Oligo/dipeptides
Fru
GalN
N-Ac-GalN
Man
Dicaroboxylate/ Tricarboxylate
Fe3+/ vitB12
SBP
P
ATPase
IIC
IIB
IIA
E-I
DctB
DctP
DctQ
DctM
Dct
DctD
TonB
ExbB
ExbD
Malate/ citrate
Fru-1-P
GalN-6-P
N-Ac-GalN-6-P
Man-6-P
Fe3+/ vitB12

Fig. 2-4. Import systems in "*Latescibacteria*" predicted from the SAGs. Extracellular degradation of polymers, as detailed in Table 2, results in the production of monomers that could potentially be transported across the outer membrane (OM) of "*Latescibacteria*" cell wall through non-specific outer membrane porins (OMP). In the periplasm, those monomers are then transported across the inner membrane (IM) via dedicated transporters including (1) Secondary transporters: glucosamine (GluA), galactosamine (GalA), and 5-dehydro-4-deoxy-glucosamine (5-dehydro-4-deoxy-GluA) are potentially imported using a single common transporter ExuT. Fucose (Fuc), rhamnose (Rha), and xylose (Xyl) are imported via dedicated proton symporters, while glucose (Glu), and galactose (Gal) are imported via dedicated sodium symporters. (2) ATP-binding cassette (ABC) transporters: ribose (Rib) and arabinose (Ara) sugars, as well as oligopeptides and dipeptides have dedicated ABC transporters with specific periplasmic substrate binding protein (SBP), two membrane permeases (P), and an ATPase. And (3) Phosphotransferase system (PTS) transporters: mannose (Man), fructose (Fru), galactosamine (GalN), and N-acetyl galactosamine (N-Ac-GalN) are imported via dedicated PTS transporters with cytoplasmic enzyme-I component (E-I) and membrane associated enzyme II components (IIA, IIB, and IIC). Sugars are phosphorylated during this kind of transport. The SAGs also encode a dedicated signal transduction system, and a tripartite ATP-independnent transporter (TRAP) for sensing, and importing, respectively, dicarboxylates, e.g. malate, and tricarboxylates, e.g. citrate, across the inner membrane. The signal transduction system is composed of the sensor histidine kinase DctB, and the cytoplasmic response regulator DctD, while the TRAP transporter is composed of the periplasmic solute receptor (DctP), the membrane small

permease component (DctQ), and the membrane large permease component (DctM). TonB-dependent import of vitamin B12 and iron complexes is also predicted from the SAGs. Several proteins with Plug domains could potentially act as the outer membrane receptor protein for vitamin B12 and iron complexes. Binding of the ligand to the receptor activates TonB-dependent import across the outer membrane via three proteins TonB, ExbB, and ExbD, that couple proton motive force to ligand transport across the outer membrane. In the periplasm, vitamin B12 or iron complexes are then transported across the inner membrane via a dedicated ABC transporter.

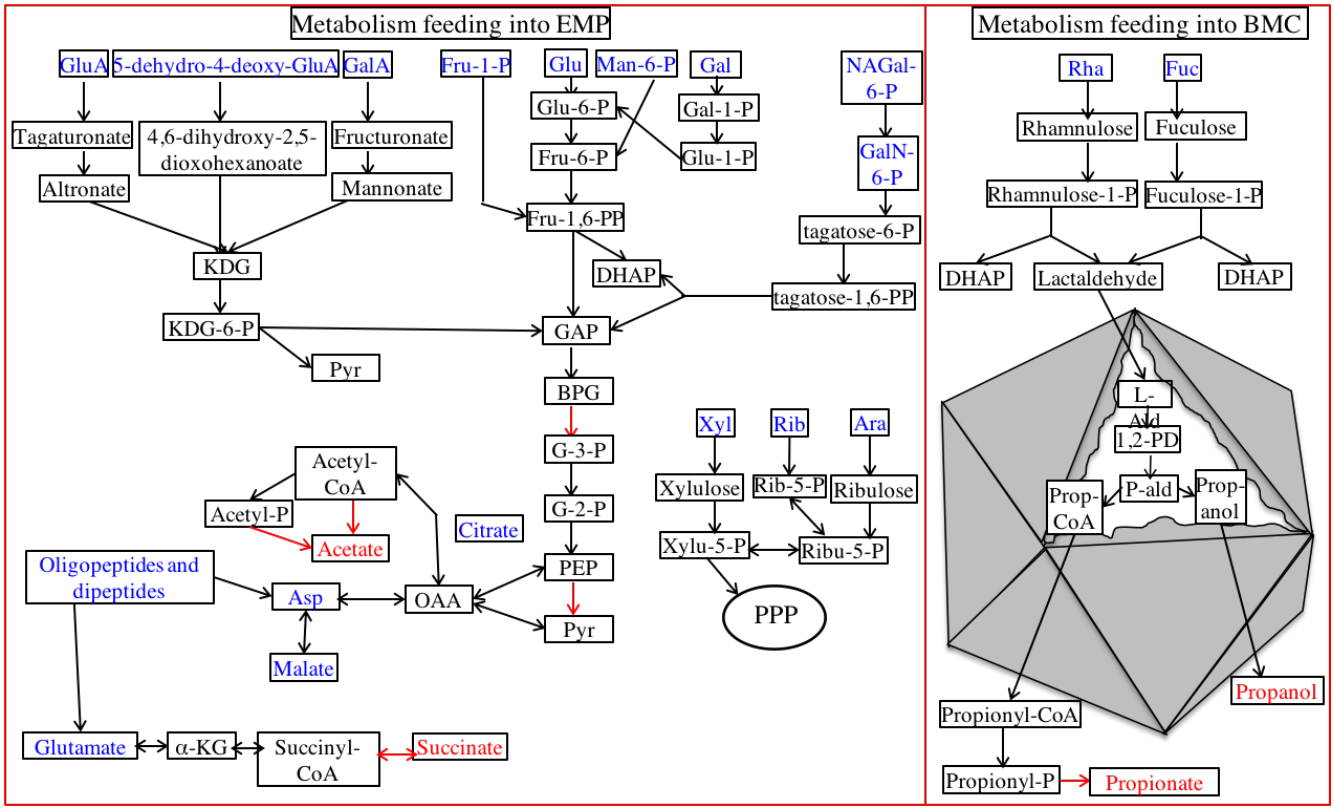

Metabolism feeding into EMP
GluA
5-dehydro-4-deoxy-GluA
GalA
Fru-1-P
Glu
Man-6-P
Gal
NAGal-6-P
Tagaturonate
4,6-dihydroxy-2,5-dioxohexanoate
Fructuronate
Glu-6-P
Gal-1-P
GalN-6-P
Altronate
Mannonate
Fru-6-P
Glu-1-P
Fru-1,6-PP
tagatose-6-P
KDG
DHAP
tagatose-1,6-PP
KDG-6-P
GAP
Pyr
BPG
Xyl
Rib
Ara
G-3-P
Acetyl-CoA
Xylulose
Rib-5-P
Ribulose
Acetyl-P
Citrate
G-2-P
Acetate
Xylu-5-P
Ribu-5-P
Oligopeptides and dipeptides
Asp
OAA
PEP
PPP
Pyr
Malate
Glutamate
α-KG
Succinyl-CoA
Succinate
Metabolism feeding into BMC
Rha
Fuc
Rhamnulose
Fuculose
Rhamnulose-1-P
Fuculose-1-P
DHAP
Lactaldehyde
DHAP
L-Ald
1,2-PD
Prop-CoA
P-ald
Prop-anol
Propanol
Propionyl-CoA
Propionyl-P
Propionate

Fig. 2-5. Metabolic reconstruction deduced from "Latescbacteria SAGs".

Metabolism is shown for the monomers produced during extracellular degradation of polymers (Table 2) followed by their transport across the outer and inner membranes as shown in Fig. 3. Three major routes are shown (depicted by red boxes) for the degradation of those monomers, Embden-Meyerhof-Paranas (EMP) pathway, Pentose phosphate pathway (PPP), and bacterial microcompartment (BMC) pathway. The BMC is depicted by an octahedral structure showing all reactions thought to occur inside of the BMC. All possible substrates potentially supporting growth are shown in blue, predicted final products are shown in red, and reactions with substrate level phosphorylations are shown by red arrows. Abbreviations (other than those mentioned in Fig. 3 legend): KDG, 2-dehydro-3-deoxy-D-gluconate; Pyr, pyruvate; Asp, aspartic acid; OAA, oxaloacetate; α-KG, α-ketoglutarate; Glu, glucose; Fru, fructose; Fru-1,6-PP, fructose-1,6-bisphosphate; DHAP, dihydroxyacetone phosphate; GAP, glyceraldehyde-3-phosphate; BPG, bisphosphoglycerate; G-3-P, 3-phosphoglycerate; G-2-P, 2-phosphoglycerate; PEP, phosphoenolpyruvate; Man, mannose; Gal, galactose; NAG, N-acetylglucosamine; NAGal, N-acetylgalactosamine; GluN, glucosamine; GalN, galactosamineRib, ribose; Ribu, ribulose; Xyl, xylose; Xylu, xylulose; Ara, arabinose; Rha, rhamnose; Fuc, fucose; L-Ald, lactaldehyde; 1,2-PD, 1,2-propanediol; P-ald, propionaldehyde; Prop-CoA, propionyl-CoA.

Discussion

Our analysis of four "*Latescibacteria*" SAGs obtained from the anaerobic monimolimnion water column of Sakinaw Lake, and the anaerobic sediments of Etoliko lagoon revealed extensive saccharolytic and proteolytic capabilities, with preference for specific polysaccharides and glycoproteins such as pectins, alginates, fucans, ulvans, xyloglucans, starch, extensins, and arabinogalactan protein originating from algal cell walls and EPS. While the degradation of some of these polymers (e.g. pectins and alginates) have been fairly well characterized at the genomic, enzymatic, and organismal levels [39, 40, 42], limited information is available regarding the pathways, genes, and microorganisms mediating the degradation of others (e.g. fucans, ulvans, extensins and arabinogalactan proteins) [44, 46, 47, 53, 54, 80]. More importantly, our knowledge of the degradation of many of these compounds is based on the study of model aerobic organisms with little knowledge of such pathways in anaerobes.

We argue that the observed patterns of polymer degradation, and monomer/oligomer transport and catabolism reflect niche specialization within *"Latescibacteria"* for survival and substrate acquisition in aquatic ecosystems. Specifically, we hypothesize that *"Latescibacteria"* SAGs analyzed in Sakinaw lake and Etoliko lagoon are involved in the degradation of a considerable fraction of algal cell wall polysaccharides and glycoprotein, algal EPS, and algal storage molecules within the detritus of green and brown algae originating at the oxic and photic zones and sinking to the anoxic and aphotic zones through sedimentation. Primary productivity is an important source for organic matter deposited in lakes [81, 82]. Algal cells represent up to 90% of such sinking organic matters, especially in stratified lakes like Sakinaw. Prior studies

have demonstrated that CO_2 fixation by algae represents the major source of organic carbon input in Sakinaw Lake, with the water column being the main site for the degradation of fixed organic carbon [83]. The stratified nature and lack of upwelling within meromictic lakes results in greater accumulation of organic matter into the lake's deeper anoxic layers [84]. The overall contribution of algal detritus to lacustrine sediments is often enhanced by the frequent occurrence of algal blooms, an ecological phenomenon predicted to increase due to global warming trends, and the progressive increase in fertilizers usage [85]. This has been reported in the lagoon systems of Western Greece, where the occurrence of algal blooms and subsequent sedimentation of organic matter represent one of the driving forces for the observed progressive eutrophication and anoxia within this ecosystem [86, 87].

It should also be noted that, in addition to polymers putatively degraded "*Latescibacteria*", algal cells are known to produce considerable quantities of oils (up to 60% of their weight), especially under unfavorable conditions (e.g. N and P starvation, temperature, salinity, or pH shifts, or heavy metal accumulation) [88, 89]. Interestingly, the analyzed SAGs lack all enzymes of the fatty acid degradation pathway to acetyl CoA. Similarly, cellulose represents an important constituent of green and brown algal cell wall [43], but the analyzed "*Latescibacteria*" SAGs display an extremely sparse cellulose degradation capacity. We reason that readily degradable components within algal detritus, e.g. cellular lipids and fatty acids, free proteins, and cellulose, are promptly utilized by microorganisms in the algal phycosphere [90-92], as well as by aerobic and anaerobic copiotrophs in the surrounding water column during the sedimentation process. Thus *"Latescibacteria"* residing in the deeper anaerobic layers of Sakinaw lake and

Etoliko lagoon sediments have evolved to specialize in the degradation of the more recalcitrant substrates that accumulate as algal detritus descends to deeper anoxic layers in stratified aquatic ecosystems. Indeed, studies in meromictic lakes have demonstrated that degradation of algal blooms occurs during sedimentation leading to biomass loss and chemical structure alteration of the algal blooms with depth [81, 82].

The proposed ecological role for members of the "*Latescibacteria*" strongly suggests cellular attachment to sinking algal detritus. "*Latescibacteria*" SAGs encode genes for flagella and pili production, and formation of gas vesicles; traits that could enhance cellular capacity for tracking and attachment to particulate organic matter. A recent survey of microbial communities in the oxygen starved Black Sea with considerable primary productivity within the upper oxic zone, shows higher relative abundance of "*Latescibacteria*" in particulate-associated samples derived from the deep anoxic zone when compared to water samples from the same location [24].

In addition to the major contribution to sinking organic matter in water bodies, algal biomass degradation under anaerobic conditions has recently received additional attention due to its potential use for biogas production [93-99]. Surprisingly, little is currently known regarding the microbial community involved in algal biomass degradation under anaerobic conditions [93]. Thus analysis of *"Latescibacteria"* SAGs directly contributes to our understanding of potential bacterial lineages involved in the anaerobic turnover of algal cell components.

Finally, the *"Latescibacteria"* SAGs encode numerous biosynthetic capabilities and a rich repertoire of catabolic enzymes and transporters with the potential to promote growth on a large number of substrates. Such capabilities are in contrast to multiple

recently obtained genomes of several uncultured bacterial and archaeal CP, where sparse anabolic capabilities, small genome size, and apparent dependence on syntrophic interactions for growth were observed [9, 13]. As such, the reported physiological properties (anaerobic nature and predicted slow growth rate due to possession of a relatively large genome size and a single rRNA operon), metabolic capabilities (distinct preference to specific polymers and sugars/sugar acids, auxotrophy to specific amino acids), and ecological distribution (preference to anaerobic and eutrophic habitats) should be considered when designing strategies for the isolation of members of the *"Latescibacteria"*.

Acknowledgements. This work was supported by the National Science Foundation Microbial Observatories Program (Grant EF0801858); the Tula Foundation, Natural Sciences and Engineering Research Council (NSERC) of Canada, Canada Foundation for Innovation (CFI), and the Canadian Institute for Advanced Research (CIFAR) through grants awarded to S.J.H. The work conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported under Contract No. DE-AC02-05CH11231.

References

1. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 2013;41:D590-6.

2. McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, et al. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J. 2012;6:610-8.

3. Winsley TJ, Snape I, McKinlay J, Stark J, van Dorst JM, Ji M, et al. The ecological controls on the prevalence of candidate division TM7 in polar regions. Front Microbiol. 2014;5:345.

4. Farag IF, Davis JP, Youssef NH, Elshahed MS. Global patterns of abundance, diversity and community structure of the Aminicenantes (candidate phylum OP8). PloS one. 2014;9:e92139.

5. Portillo MC, Sririn V, Kanoksilapatham W, Gonzalez JM. Differential microbial communities in hot spring mats from Western Thailand. Extremophiles. 2009;13:321-31.

6. Ferrari B, Winsley T, Ji M, Neilan B. Insights into the distribution and abundance of the ubiquitous candidatus *Saccharibacteria* phylum following tag pyrosequencing. Sci Rep. 2014;4:3957.

7. Ohkuma M, Sato T, Noda S, Ui S, Kudo T, Hongoh Y. The candidate phylum 'Termite Group 1' of bacteria: phylogenetic diversity, distribution, and endosymbiont members of various gut flagellated protists. FEMS Microbiol Ecol. 2007;60:467-76.

8. Kamke J, Rinke C, Schwientek P, Mavromatis K, Ivanova N, Sczyrba A, et al. The candidate phylum Poribacteria by single-cell genomics: new insights into phylogeny,

cell-compartmentation, eukaryote-like repeat proteins, and other genomic features. PloS one. 2014;9:e87353.

9. Kantor RS, Wrighton KC, Handley KM, Sharon I, Hug LA, Castelle CJ, et al. Small genomes and sparse metabolisms of sediment-associated bacteria from four candidate phyla. mBio. 2013;4:e00708-13.

10. McLean JS, Lombardo MJ, Badger JH, Edlund A, Novotny M, Yee-Greenbaum J, et al. Candidate phylum TM6 genome recovered from a hospital sink biofilm provides genomic insights into this uncultivated phylum. Proc Natl Acad Sci USA. 2013;110:E2390-9.

11. Wrighton KC, Castelle CJ, Wilkins MJ, Hug LA, Sharon I, Thomas BC, et al. Metabolic interdependencies between phylogenetically novel fermenters and respiratory organisms in an unconfined aquifer. ISME J. 2014;8:1452-63.

12. Takami H, Noguchi H, Takaki Y, Uchiyama I, Toyoda A, Nishi S, et al. A deeply branching thermophilic bacterium with an ancient acetyl-CoA pathway dominates a subsurface ecosystem. PloS one. 2012;7:e30559.

13. Youssef NH, Rinke C, Stepanauskas R, Farag I, Woyke T, Elshahed MS. Insights into the metabolism, lifestyle and putative evolutionary history of the novel archaeal phylum *'Diapherotrites'*. ISME J. 2015;9:447-60.

14. Kamke J, Sczyrba A, Ivanova N, Schwientek P, Rinke C, Mavromatis K, et al. Single-cell genomics reveals complex carbohydrate degradation patterns in poribacterial symbionts of marine sponges. ISME J. 2013;7:2287-300. Epub 2013/07/12.

15. Campbell JH, O'Donoghue P, Campbell AG, Schwientek P, Sczyrba A, Woyke T, et al. UGA is an additional glycine codon in uncultured SR1 bacteria from the human microbiota. Proc Natl Acad Sci. 2013;110:5540-5.

16. Wilson MC, Mori T, Ruckert C, Uria AR, Helf MJ, Takada K, et al. An environmental bacterial taxon with a large and distinct metabolic repertoire. Nature. 2014;506:58-62.

17. Dojka MA, Hugenholtz P, Haack SK, Pace NR. Microbial diversity in a hydrocarbon- and chlorinated-solvent-contaminated aquifer undergoing intrinsic bioremediation. Appl Environ Microbiol. 1998;64:3869-77.

18. Pereira AD, Leal CD, Dias MF, Etchebehere C, Chernicharo CA, de Araujo JC. Effect of phenol on the nitrogen removal performance and microbial community structure and composition of an anammox reactor. Bioresour Technol. 2014;166:103-11.

19. Schabereiter-Gurtner C, Saiz-Jimenez C, Pinar G, Lubitz W, Rolleke S. Phylogenetic diversity of bacteria associated with Paleolithic paintings and surrounding rock walls in two Spanish caves (Llonin and La Garma). FEMS Microbiol Ecol. 2004;47:235-47.

20. Ikenaga M, Guevara R, Dean AL, Pisani C, Boyer JN. Changes in community structure of sediment bacteria along the Florida coastal everglades marsh-mangrove-seagrass salinity gradient. Microb Ecol. 2010;59:284-95.

21. Reed AJ, Lutz RA, Vetriani C. Vertical distribution and diversity of bacteria and archaea in sulfide and methane-rich cold seep sediments located at the base of the Florida Escarpment. Extremophiles. 2006;10:199-211.

22. Hernandez-Raquet G, Budzinski H, Caumette P, Dabert P, Le Menach K, Muyzer G, et al. Molecular diversity studies of bacterial communities of oil polluted microbial mats from the Etang de Berre (France). FEMS Microbiol Ecol. 2006;58:550-62.

23. Briggs BR, Pohlman JW, Torres M, Riedel M, Brodie EL, Colwell FS. Macroscopic biofilms in fracture-dominated sediment that anaerobically oxidize methane. Appl Environ Microbiol. 2011;77:6780-7.

24. Fuchsman CA, Kirkpatrick JB, Brazelton WJ, Murray JW, Staley JT. Metabolic strategies of free-living and aggregate-associated bacterial communities inferred from biologic and chemical profiles in the Black Sea suboxic zone. FEMS Microbiol Ecol. 2011;78:586-603.

25. Kormas KA, Meziti A, Dahlmann A, GJ DEL, Lykousis V. Characterization of methanogenic and prokaryotic assemblages based on mcrA and 16S rRNA gene diversity in sediments of the Kazan mud volcano (Mediterranean Sea). Geobiology. 2008;6:450-60.

26. Lee OO, Yang J, Bougouffa S, Wang Y, Batang Z, Tian R, et al. Spatial and species variations in bacterial communities associated with corals from the Red Sea as revealed by pyrosequencing. Appl Environ Microbiol. 2012;78:7173-84.

27. Carbonetto B, Rascovan N, Alvarez R, Mentaberry A, Vazquez MP. Structure, composition and metagenomic profile of soil microbiomes associated to agricultural land use and tillage systems in Argentine Pampas. PloS one. 2014;9:e99949.

28. Yakimov MM, La Cono V, Slepak VZ, La Spada G, Arcadi E, Messina E, et al. Microbial life in the Lake Medee, the largest deep-sea salt-saturated formation. Sci Rep. 2013;3:3554.

29. Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng JF, et al. Insights into the phylogeny and coding potential of microbial dark matter. Nature. 2013;499:431-7.

30. Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M. Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res. 2014;42:D199-205.

31. Caspi R, Altman T, Billington R, Dreher K, Foerster H, Fulcher CA, et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic Acids Res. 2014;42:D459-D71.

32. Bland C, Ramsey TL, Sabree F, Lowe M, Brown K, Kyrpides NC, et al. CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. BMC Bioinformatics. 2007;8:209.

33. Anonymous. PILER Genomic repeat analysis software. 2009.

34. Rawlings ND, Waller M, Barrett AJ, Bateman A. MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. Nucleic Acids Res. 2014;42:D503-9.

35. Saier MH, Jr., Reddy VS, Tamang DG, Vastermark A. The transporter classification database. Nucleic Acids Res. 2014;42:D251-8.

36. Yin Y, Mao X, Yang J, Chen X, Mao F, Xu Y. dbCAN: a web resource for automated carbohydrate-active enzyme annotation. Nucleic Acids Res. 2012;40:W445-51.

37. Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B. The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res. 2014;42:D490-5.

38. Domozych DS, Sorensen I, Popper ZA, Ochs J, Andreas A, Fangel JU, et al. Pectin metabolism and assembly in the cell wall of the charophyte green alga *Penium margaritaceum*. Plant Physiol. 2014;165:105-18.

39. Abbott DW, Boraston AB. Structural biology of pectin degradation by Enterobacteriaceae. Microbiol Mol Biol Rev. 2008;72:301-16.

40. Benoit I, Coutinho P, Schols H, Gerlach J, Henrissat B, de Vries R. Degradation of different pectins by fungi: correlations and contrasts between the pectinolytic enzyme sets identified in genomes and the growth on pectins of different origin. BMC Genomics. 2012;13:321.

41. Michel G, Tonon T, Scornet D, Cock JM, Kloareg B. The cell wall polysaccharide metabolism of the brown alga *Ectocarpus siliculosus*. Insights into the evolution of extracellular matrix polysaccharides in Eukaryotes. New Phytol. 2010;188:82-97.

42. Kabisch A, Otto A, Konig S, Becher D, Albrecht D, Schuler M, et al. Functional characterization of polysaccharide utilization loci in the marine Bacteroidetes '*Gramella forsetii*' KT0803. ISME J. 2014;8:1492-502.

43. Domozych DS. Algal cell walls. eLS. Chichester: John Wiley & Sons, Ltd; 2011.

44. Chiellini F, Morelli A. Ulvan: A Versatile Platform of Biomaterials from Renewable Resources: INTECH Open Access Publisher; 2011.

45. Jiao G, Yu G, Zhang J, Ewart HS. Chemical structures and bioactivities of sulfated polysaccharides from marine algae. Mar Drugs. 2011;9:196-223.

46. Ferro DR, Provasoli A, Ragazzi M, Casu B, Torri G, Bossennec V, et al. Conformer populations of L-iduronic acid residues in glycosaminoglycan sequences. Carbohydr Res. 1990;195:157-67.

47. Nyvall Collen P, Sassi JF, Rogniaux H, Marfaing H, Helbert W. Ulvan lyases isolated from the Flavobacteria *Persicivirga ulvanivorans* are the first members of a new polysaccharide lyase family. J Biol Chem. 2011;286:42063-71.

48. Domozych DS, Sorensen I, Willats WG. The distribution of cell wall polymers during antheridium development and spermatogenesis in the Charophycean green alga, *Chara corallina*. Ann Bot. 2009;104:1045-56.

49. Ikegaya H, Hayashi T, Kaku T, Iwata K, Sonobe S, Shimmen T. Presence of xyloglucan-like polysaccharide in *Spirogyra* and possible involvement in cell–cell attachment. Phycol Res. 2008;56:216-22.

50. Lahaye M, Jegou D, Buleon A. Chemical characteristics of insoluble glucans from the cell wall of the marine green alga *Ulva lactuca* (L.) Thuret. Carbohydr Res. 1994;262:115-25.

51. Estevez JM, Fernandez PV, Kasulin L, Dupree P, Ciancia M. Chemical and in situ characterization of macromolecular components of the cell walls from the green seaweed *Codium fragile*. Glycobiology. 2009;19:212-28.

52. Domozych DS, Ciancia M, Fangel JU, Mikkelsen MD, Ulvskov P, Willats WG. The cell walls of green algae: a journey through evolution and diversity. Front Plant Sci. 2012;3:82.

53. Fujita K, Sakamoto S, Ono Y, Wakao M, Suda Y, Kitahara K, et al. Molecular cloning and characterization of a beta-L-arabinobiosidase in *Bifidobacterium longum* that belongs to a novel glycoside hydrolase family. J Biol Chem. 2011;286:5143-50.

54. Knoch E, Dilokpimol A, Geshi N. Arabinogalactan proteins: focus on carbohydrate active enzymes. Front Plant Sci. 2014;5:198.

55. Bakunina I, Nedashkovskaya O, Balabanova L, Zvyagintseva T, Rasskasov V, Mikhailov V. Comparative analysis of glycoside hydrolases activities from phylogenetically diverse marine bacteria of the genus *Arenibacter*. Mar Drugs. 2013;11:1977-98.

56. Cobucci-Ponzano B, Conte F, Strazzulli A, Capasso C, Fiume I, Pocsfalvi G, et al. The molecular characterization of a novel GH38 alpha-mannosidase from the crenarchaeon *Sulfolobus solfataricus* revealed its ability in de-mannosylating glycoproteins. Biochimie. 2010;92:1895-907.

57. Mamedov T, Yusibov V. Green algae *Chlamydomonas reinhardtii* possess endogenous sialylated N-glycans. FEBS Open Bio. 2011;1:15-22.

58. Popper ZA, Michel G, Herve C, Domozych DS, Willats WG, Tuohy MG, et al. Evolution and diversity of plant cell walls: from algae to flowering plants. Ann Rev Plant Biol. 2011;62:567-90.

59. Gerken HG, Donohoe B, Knoshaug EP. Enzymatic cell wall degradation of *Chlorella vulgaris* and other microalgae for biofuels production. Planta. 2013;237:239-53.

60. Benjdia A, Leprince J, Guillot A, Vaudry H, Rabot S, Berteau O. Anaerobic sulfatase-maturating enzymes: radical SAM enzymes able to catalyze in vitro sulfatase post-translational modification. J Am Chem Soc. 2007;129:3462-3.

61. Dierks T, Miech C, Hummerjohann J, Schmidt B, Kertesz MA, von Figura K. Posttranslational formation of formylglycine in prokaryotic sulfatases by modification of either cysteine or serine. J Biol Chem. 1998;273:25560-4.

62. Deschamps P, Haferkamp I, d'Hulst C, Neuhaus HE, Ball SG. The relocation of starch metabolism to chloroplasts: when, why and how. Trends Plant Sci. 2008;13:574-82.

63. Bellinger BJ, Gretz MR, Domozych DS, Kiemle SN, Hagerthey SE. Composition of extracellular polymeric substances from periphyton assemblages in the Florida everglades. J Phycol. 2010;46:484-96.

64. Mishra A, Kavita K, Jha B. Characterization of extracellular polymeric substances produced by micro-algae *Dunaliella salina.* Carbohydr Pol. 2011;83:852-7.

65. Gonzalez ET, Allen C. Characterization of a *Ralstonia solanacearum* operon required for polygalacturonate degradation and uptake of galacturonic acid. Mol Plant-Micr Int. 2003;16:536-44.

66. Valentini M, Storelli N, Lapouge K. Identification of C(4)-dicarboxylate transport systems in *Pseudomonas aeruginosa* PAO1. J Bacteriol. 2011;193:4307-16.

67. Havemann GD, Bobik TA. Protein content of polyhedral organelles involved in coenzyme B12-dependent degradation of 1,2-propanediol in *Salmonella enterica* serovar *Typhimurium* LT2. J Bacteriol. 2003;185:5086-95.

68. Petit E, LaTouf WG, Coppi MV, Warnick TA, Currie D, Romashko I, et al. Involvement of a bacterial microcompartment in the metabolism of fucose and rhamnose by *Clostridium phytofermentans*. PloS One. 2013;8:e54337.

69. Axen SD, Erbilgin O, Kerfeld CA. A taxonomy of bacterial microcompartment loci constructed by a novel scoring method. PLoS Comput Biol. 2014;10:e1003898.

70. Dunne WM. Bacterial adhesion: Seen any good biofilms lately? Clin Microbiol Rev. 2002;15:155-66.

71. Walsby AE. Gas vesicles. Microbiol Rev. 1994;58:94-144.

72. Briukhanov AL, Netrusov AI. Aerotolerance of strictly anaerobic microorganisms and factors of defense against oxidative stress: a review. Prikladnaia Biokhimiia i Mikrobiologiia. 2007;43:635-52.

73. Liu R, Ochman H. Stepwise formation of the bacterial flagellar system. Proc Natl Acad Sci USA. 2007;104:7116-21.

74. Pfeifer F. Distribution, formation and regulation of gas vesicles. Nat Rev Microbiol. 2012;10:705-15.

75. Takhar HK, Kemp K, Kim M, Howell PL, Burrows LL. The platform protein is essential for type IV pilus biogenesis. Journal Biol Chem. 2013;288:9721-8.

76. Konno N, Ishida T, Igarashi K, Fushinobu S, Habu N, Samejima M, et al. Crystal structure of polysaccharide lyase family 20 endo-β-1,4-glucuronan lyase from the filamentous fungus *Trichoderma reesei*. FEBS Lett. 2009;583:1323-6.

77. Redouan E, Cedric D, Emmanuel P, Mohamed EG, Bernard C, Philippe M, et al. Improved isolation of glucuronan from algae and the production of glucuronic acid oligosaccharides using a glucuronan lyase. Carbohydr Res. 2009;344:1670-5.

78. Horvath P, Barrangou R. CRISPR/Cas, the immune system of bacteria and archaea. Science. 2010;327:167-70.

79. Coulthurst SJ. The Type VI secretion system – a widespread and versatile cell targeting system. Res Microbiol. 2013;164:640-54.

80. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421.

81. Meyers PA. Preservation of elemental and isotopic source identification of sedimentary organic matter. Chem Geol. 1994;114:289-302.

82. Meyers PA, Ishiwatari R. Lacustrine organic geochemistry—an overview of indicators of organic matter sources and diagenesis in lake sediments. Org Geochem. 1993;20:867-900.

83. Perry KA. The chemical limnology of two meromictic lakes with emphasis on pyrite formation. Vancouver: University of British Columbia; 1990.

84. Bresciani M, Bolpagni R, Laini A, Matta E, Bartoli M, Giardino C. Multitemporal analysis of algal blooms with MERIS images in a deep meromictic lake. Eur J Remote Sens. 2013;46:445-58.

85. Paerl HW, Scott JT. Throwing fuel on the fire: synergistic effects of excessive nitrogen inputs and global warming on harmful algal blooms. Environ Sci Technol. 2010;44:7756-8.

86. Kormas KA, Nicolaidou A, Reizopoulou S. Temporal variations of nutrients, chlorophyll A and particulate matter in three coastal lagoons of Amvrakikos Gulf (Ionian Sea, Greece). Mar Ecol. 2001;22:201-13.

87. Diapoulis A, Haritonidis S. Marine algae of West Greek Coasts. Acta Adriatica. 1987;28:85-101.

88. Chisti Y. Biodiesel from microalgae. Biotechnol Adv. 2007;25:294-306.

89. Sharma KK, Schuhmann H, Schenk PM. High lipid induction in microalgae for biodiesel production. Energies. 2012;5:1532-53.

90. Buchan A, LeCleir GR, Gulvik CA, Gonzalez JM. Master recyclers: features and functions of bacteria associated with phytoplankton blooms. Nat Rev Microbiol. 2014;12:686-98.

91. Sapp M, Schwaderer AS, Wiltshire KH, Hoppe HG, Gerdts G, Wichels A. Species-specific bacterial communities in the phycosphere of microalgae? Microb Ecol. 2007;53:683-99.

92. Mann AJ, Hahnke RL, Huang S, Werner J, Xing P, Barbeyron T, et al. The genome of the alga-associated marine flavobacterium *Formosa agariphila* KMM 3901T reveals a broad potential for degradation of algal polysaccharides. Appl Environ Microbiol. 2013;79:6813-22.

93. Ward AJ, Lewis DM, Green FB. Anaerobic digestion of algae biomass: A review. Alg Res. 2014;5:204-14.

94. Chisti Y. Biodiesel from microalgae beats bioethanol. Trends Biotechnol. 2008;26:126-31.

95. Sialve B, Bernet N, Bernard O. Anaerobic digestion of microalgae as a necessary step to make microalgal biodiesel sustainable. Biotechnol Adv. 2009;27:409-16.

96. Sutherland DL, Turnbull MH, Broady PA, Craggs RJ. Effects of two different nutrient loads on microalgal production, nutrient removal and photosynthetic efficiency in pilot-scale wastewater high rate algal ponds. Water Res. 2014;66c:53-62.

97. Prajapati SK, Kaushik P, Malik A, Vijay VK. Phycoremediation and biogas potential of native algal isolates from soil and wastewater. Bioresource Technol. 2013;135:232-8.

98. Vergara-Fernández A, Vargas G, Alarcón N, Velasco A. Evaluation of marine algae as a source of biogas in a two-stage anaerobic reactor system. Biomass Bioenergy. 2008;32:338-44.

99. Mahadevaswamy M, Venkataraman LV. Bioconversion of poultry droppings for biogas and algal production. Agric Wastes. 1986;18:93-101.

100. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol Biol Evol. 2013;30:2725-9.

101. Eder M, Lutz-Meindl U. Analyses and localization of pectin-like carbohydrates in cell wall and mucilage of the green alga *Netrium digitus*. Protoplasma. 2010;243:25-38.

Chapter III

Global distribution patterns and pangenomic diversity of the candidate phylum "Latescibacteria" (WS3)

Abstract

We investigated the global distribution patterns and pangenomic diversity of the candidate phylum *"*Latescibacteria" (WS3) in 16S rRNA gene as well as metagenomic datasets. We document distinct distribution patterns for various "Latescibacteria" orders in 16S rRNA gene datasets, with prevalence of orders sediment_1 in terrestrial, PBSIII_9 in groundwater and temperate freshwater, and GN03 in pelagic marine, saline-hypersaline, and wastewater habitats. Using a fragment recruitment approach, we identified 68.9 Mb of "Latescibacteria"-affiliated contigs in publicly available metagenomic datasets comprising 73,079 proteins. Metabolic reconstruction suggests a prevalent saprophytic lifestyle in all "Latescibacteria" orders, with marked capacities for the degradation of proteins, lipids, and polysaccharides predominant in plant, bacterial, fungal/crustacean, and eukaryotic algal cell walls. As well, extensive transport and central metabolic pathways for the metabolism of imported monomers were identified. Interestingly, genes and domains suggestive of the production of a cellulosome, e.g. protein-coding genes harboring dockerin I domains attached to a glycosyl hydrolase, and scaffoldin-encoding genes harboring cohesin I and CBM37 domains, were identified in orders PBSIII_9, GN03, and MSB-4E2 fragments recovered from four anoxic aquatic habitats; hence extending the cellulosomal production capabilities in Bacteria beyond the Gram-positive Firmicutes. In addition to fermentative pathways, a complete electron transport chain with terminal cytochrome C oxidases Caa3 (for operation under high oxygen tension), and Cbb3 (for operation under low oxygen tension) were identified in PBSIII_9, and GN03 fragments recovered from oxygenated, and partially/seasonally oxygenated aquatic habitats. Our metagenomic recruitment effort hence represents a

comprehensive pangenomic view of this yet-uncultured phylum, and provides broader and complimentary insights to those gained from genome recovery initiatives focusing on a single or few sampled environments.

Importance

Our understanding of the phylogenetic diversity, metabolic capabilities, and ecological roles of yet-uncultured microorganisms is rapidly expanding. However, recent efforts mainly been focused on recovering genomes of novel microbial lineages from a specific sampling site, rather from a wide range of environmental habitats. To comprehensively evaluate the genomic landscape, putative metabolic capabilities, and ecological roles of yet-uncultured candidate phyla, efforts that focus on the recovery of genomic fragments from a wide range of habitats and that adequately sample the intra-phylum diversity within a specific target lineage are needed. Here, we investigated the global distribution patterns and pangenomic diversity of the candidate phylum *"*Latescibacteria". Our results document the preference of specific "Latescibacteria" orders to specific habitats, the prevalence of plant polysaccharide degradation abilities within all "Latescibacteria" orders, the occurrence of all genes/domains necessary for the production of cellulosome within three "Latescibacteria" orders (GN03, PBSIII_9, and MSB-4E2) in datasets recovered from anaerobic locations, and the identification of the components of an aerobic respiratory chain, as well as occurrence of multiple O_2 dependent metabolic reactions in "Latescibacteria" orders GN03 and PBSIII_9 recovered from oxygenated habitats. The results demonstrate the value of phylocentric pangenomic surveys for understanding the global ecological distribution and pan-metabolic abilities of yet-uncultured microbial lineages since they provide broader and complimentary insights to those gained from single cell genomic and/or metagenomics-enabled genome recovery efforts focusing on a single sampling site.

Introduction

Our understanding of the phylogenetic diversity, metabolic capabilities, and ecological roles of yet-uncultured microorganisms is rapidly expanding. Multiple studies have recently reported on the recovery and analysis of genomes belonging to yet-uncultured bacterial and archaeal lineages (1-4). The recent proliferation of such efforts could be attributed to the timely convergence of multiple experimental and computational advances, such as the development of low cost high throughput sequencing technologies (5-7), the increased utilization and access to super computing capacity (8), the development of multiple bioinformatics tools for fast sequence assembly and genome recovery from metagenomes (9-12), and the development of reliable protocols for single cell sorting and genome amplification (13).

So far, many of these recently published studies represent the first reported analysis of genomes belonging to a specific phylum (1, 14-16). While extremely valuable, such studies should be regarded as a first step towards comprehensive evaluations of the pangenomic diversity and metabolic potential of such lineages. This is especially true when examining uncultured phyla (candidate phyla); given the often observed wide scope of intra-phylum (class, order, and family) level phylogenetic diversity (17), and the global distribution patterns of many of these candidate phyla (18). Therefore, to comprehensively evaluate the genomic landscape, putative metabolic capabilities, and ecological roles of yet-uncultured candidate phyla, efforts that focus on the recovery of genomic fragments from a wide range of habitats and that adequately sample the intra-phylum diversity within a specific target lineage are needed.

Members of the candidate phylum “Latescibacteria” (previously known as WS3) were first discovered in a hydrocarbon and chlorinated solvents contaminated aquifer (19). Since then, they have been detected in culture-independent surveys in a wide range of terrestrial and marine habitats, e.g. soil, marine sediments, hydrothermal vents, anoxic lakes, hydrocarbon-impacted environments, and wastewater treatment bioreactors (20-29). We recently reported on the genomic features of four “Latescibacteria” single cell amplified genomes (SAGs) recovered from the anoxic hypolimnion layers of two water bodies (Sakinaw lake, British Colombia-Canada, and Etoliko lagoon, Gulf of Patras, Greece) (1, 16). The analyzed genomes suggested a heterotrophic, strictly fermentative lifestyle, with predicted dedicated saccharolytic and proteolytic enzymes and transporters for the degradation and uptake of pectin, ulvan, fucan, alginate, and hydroxyproline rich polymers. Based on these peculiar substrate utilization patterns, we proposed that the “Latescibacteria” play an important role in the turnover of algal detritus that reaches the anoxic/microxic layers in these stratified water bodies.

However, given the global distribution of the “Latescibacteria” and the fact that all analyzed genomes in this prior study belonged to a single “Latescibacteria” lineage (family PBSIII_9_1 in order PBSIII_9), it is improbable that such specific findings represent a comprehensive depiction of the global metabolic abilities and ecological roles of the “Latescibacteria”. Here, we conducted an extensive survey of amplicon-generated and metagenomic datasets to identify the global distribution patterns and pangenomic capabilities of the “Latescibacteria” in a wide range of biotopes. The results provide novel insights into the ecological distribution, habitat preferences, and metabolic capabilities of various orders within this ubiquitous yet-uncultured bacterial phylum, and

demonstrate the value of *in-silico* data mining surveys in providing a comprehensive pangenomic overview of a target microbial lineage.

Materials and Methods

A taxonomic outline of the "**Latescibacteria**". We aimed to produce an updated and comprehensive taxonomic outline of the "Latescibacteria" for utilization as the basis for our subsequent community structure analysis and metabolic reconstruction efforts. The "Latescibacteria" is represented by ten orders and 310 sequences in the May 2013 version of Greengenes database (http://greengenes.secondgenome.com/downloads/database/13_5) (30). We used these 310 curated reference "Latescibacteria" sequences to query the GenBank nr database for each of their 50 closest relatives(30), and from these, we extracted non-chimeric sequences that are longer than 900 bp and exhibited >75% sequence identity to at least one of the "Latescibacteria" reference sequences. The phylogenetic affiliation of the identified sequences (n=8,752) was assessed by aligning the sequences in ClustalW (31) (using a gap opening score 15.0 and gap extension penalty of 6.66), end trimming alignments, and using the trimmed alignment (positions 70-1170, *E. Coli* 16S rRNA gene) for stepwise insertion into distance and maximum likelihood reference trees with multiple "Latescibacteria" sequence representatives, as well as representatives of additional 15 phyla and candidate phyla using MEGA7 (32, 33). Sequences were identified as belonging to the "Latescibacteria" if they continuously remained monophyletic within the reference "Latescibacteria" clade regardless of the tree building algorithm and taxa included in the analysis. Order-level affiliation within the "Latescibacteria" was also assessed using the same criteria described above; and novel orders were proposed if more than one sequence remained reproducibly unaffiliated with

all recognized orders regardless of the tree building algorithm and the taxa included in the analysis.

Ecological distribution and community structure of the "Latescibacteria" in high throughput 16S rRNA gene datasets. To assess "Latescibacteria" global ecological distribution patterns and community structure, we extracted "Latescibacteria"-affiliated 16S rRNA gene sequences from 16S rRNA gene datasets generated using high throughput (454 pyrosequencing, Illumina) sequencing technologies. Queried datasets were downloaded from the GenBank SRA archive (through the mirror web interface of DNA Data Bank of Japan http://www.ddbj.nig.ac.jp) (34), MG-RAST (35), and VAMPS (36) (December 2014). A total of ≈2.2 billion (2,205,192,887) distinct high throughput 16S rRNA gene sequences from 4041 distinct datasets belonging to 131 different studies were analyzed. We excluded human microbiome studies from our analysis since our preliminary screening failed to identify any "Latescibacteria"-affiliated sequence in these datasets. All identified sequences were classified using classify.seqs command in Mothur (v.1.33.0) (37) and the Greengenes "Latescibacteria" taxonomy manually curated to match our updated taxonomic outline. We applied a confidence threshold cutoff of 80% for the classification.

Following the IMG/M habitat classification scheme (38), each of the datasets was assigned into one of six major habitats (marine, freshwater, terrestrial, bioremediation, host-associated, and engineered). The marine, freshwater, and terrestrial habitats were further sub-classified into 15 different sub-habitats (38). The relative abundances of the phylum "Latescibacteria", as well as each of the "Latescibacteria" orders in every dataset, habitat, and sub-habitat were used to deduce the order/suborder-level ecological

distribution and habitat preferences of the "Latescibacteria". Only datasets, with > 0.1% "Latescibacteria" abundance, were included in subsequent analyses. The relative abundance values of "Latescibacteria" orders/suborders in the different habitats/sub-habitats were used in a principal component analysis (PCA) using the prcomp command in the R statistical package Labdsv (http://ecology.msu.montana.edu/labdsv/R/labs/), and the results were visualized in a biplot.

Identification of "Latescibacteria" genomic fragments in metagenomic datasets using a fragment recruitment approach. We utilized a fragment recruitment approach to identify and recover "Latescibacteria" genomic fragments by mapping metagenomics contigs to reference genomes.

The fragment recruitment approach is broadly similar to that previously utilized in reference (1); and is based on blast similarity to a reference database, followed by examining contigs identified as putatively belonging to target lineage using Phylosift v1.0.1 (39) to extract housekeeping genes present within the contigs and confirm their phylogenetic affiliation. A customized database encompassing 194 genomes belonging to a diverse array of cultured and uncultured bacterial and archaeal lineages, including the four previously reported "Latescibacteria" single amplified genomes (GenBank accession numbers: ASMB00000000, AQSL00000000, ASWY00000000, and AQRO00000000) was constructed and utilized for fragment recruitment.

Prior to its implementation for identifying "Latescibacteria" contigs in metagenomics datasets, we sought to benchmark the performance of this approach using well-sampled bacterial phyla and mock microbial communities. We examined the sensitivity of this procedure by quantifying the recovery percentage of three target

organisms belonging to the phylum Firmicutes (*Bacillus anthracis* str Ames GenBank accession number: AE016879.1, *Paeniclostridium sordellii* GenBank accession number: BDJI00000000.1*,* and *Megasphaera elsdenii* GenBank accession number: HE576794.1*)* from a mock microbial community. These organisms have strain level relatedness, family level relatedness, or only class level relatedness to organisms in the custom database (*Bacillus anthracis* str Ames to *Bacillus anthracis 3154*, *Paeniclostridium sordellii* to *Clostridium difficile CD131*, family Peptostreptococcaceae, and *Megasphaera elsdenii*, are representing a distinct order in the class Negativicutes for which *Veillonella atypica KON, ATCC 17744* are representatives). Each genome were fragmented to N50 of ≈20Kb and added to a mock microbial community of 94 bacterial, non-Firmicutes genomes. Recovery percentages of target (sensitivity) and non-target (specificity) organisms were determined at different empirical e-values cutoffs (10^{-5}, 10^{-10} and 10^{-15}). In general, the percentage recovery of a specific genome was dependent on its level of relatedness to Firmicutes genomes in the utilized database, with highest recovery for *B. anthracis* and lowest for *Megasphaera elsdenii*. The recovery percentage was also dependence on the stringency of e-value cutoff utilized. However, the utilization of less stringent e-values resulted in lower specificity i.e. slightly higher percentage recovery of non-specific sequences, even after utilization of quality control criteria described above. Based on these results, we opted for an e-value of 10^{-10} and to minimize non-target recruitment.

Using the criteria described above, we identified "Latescibacteria" contigs in a total of 1589 metagenomic datasets from 86 different studies (approximately 2.45 X 10^{11} bp) available on the IMG/M database (38). We excluded datasets from human

microbiome studies due to the expected absence of "Latescibacteria" in these datasets based on our analysis of 16S rRNA gene datasets (as outlined above). (39)

All analyses were conducted on the HPC Cowboy supercomputer, a 252 compute node with dual six core CPUs and 32 GB RAM server, 2 fat nodes with 256 GB RAM, GPU cards and 120 TB very fast disk storage, at the OSU High Performance Computing Center at Oklahoma State University.

"Latescibacteria" contigs order level classification. A two-step approach was utilized for order-level assignment of the identified "Latescibacteria" contigs. First, identified contigs harboring 16S rRNA genes were readily assigned to various "Latescibacteria" orders by insertion of the rRNA gene into a reference 16S rRNA tree as described above. Following, housekeeping genes in these 16S rRNA gene-harboring contigs were extracted using phylosift (39), and utilized as the corresponding "Latescibacteria" order-specific markers. We used these markers to query the remaining contigs that lacked 16S rRNA genes. Contigs that were identified as first hits to any of the markers were assigned to that marker specific order. A second round of housekeeping genes extraction and screening of the remaining unassigned contigs, based on the results of the first round, was subsequently conducted.

"**Latescibacteria**" **pangenomes functional annotation and metabolic reconstruction.** Gene prediction in all "Latescibacteria" contigs identified in the metagenomic datasets (thereafter "Latescibacteria" pangenome) was achieved using Prodigal (40). Detailed metabolic reconstruction was performed using KEGG (41) and Metacyc (42) databases. All potential carbohydrate active enzymes (CAZymes) were identified and classified using dbCAN-fam-HMMs (v4) database (43) with an e-value of 10^{-04} as the cutoff score.

Cellular localization of predicted proteins were identified using SignalP 4.1 (44). Proteases, peptidases, and peptidase inhibitors were identified using BlastP against the MEROPS database release 9.10 (45). Transporters were identified using BlastP against the TCDB database (46) with the criteria 40% sequence identity and e-value $\leq 10^{-04}$ used as the cutoff.

The occurrence of predicted dockerin domains in multiple identified "Latescibacteria" CAZyme-encoding genes promoted our search for various domains and motifs suggestive of cellulosomal production in the "Latescibacteria". The "Latescibacteria" pangenomes were queried for the occurrence of dockerin I and II domains (pfam PF00404), cohesin I and II domains (pfam PF00963), scaffoldin modules (using BLASTP), and surface layer proteins (SLP, pfam PF00395). The identified "Latescibacteria" predicted dockerin and cohesin domains were aligned to reference sequences using Clustal omega (v1.2.3) (47) and the alignments were used for tree construction using distance neighbor joining (NJ) approach with Jukes-Cantor corrections in MEGA7 (32). The tertiary structure model of the identified dockerin domains was predicted using homology modeling by Iterative Threading Assembly Refinement (I-TASSER suite) (48).

Results

Global patterns of "Latescibacteria" distribution and community structure in 16S rRNA gene databases. We identified 1,167 near full-length 16S rRNA gene sequences belonging to the "Latescibacteria" within the GenBank nr database (Fig. 3-1). These sequences clustered into 11 distinct orders: Sediment_1, PBSIII_9, GN03, MSB-4E2, PRR_12, CV_106, SSS_58A, LD1_PA13, WB1_H11, SAW1_B6, and RII_AN061 (Fig. 3-1). Strong bootsrap support for all order and phylum level branches was obtained regardless of the tree-building algorithm utilized (Fig. 3-1). The average and maximum pairwise divergence values within and between various "Latescibacteria" orders, as well as to their closest relatives outside the phylumdemonstrate a clear delineation was observed for all 11 "Latescibacteria" orders as well as the entire phylum. Collectively, this argues for the validity of the phylogenetic outline of the "Latescibacteria" presented in this study.

In high throughput generated 16S rRNA gene datasets, we identified 149,754 "Latescibacteria"-affiliated 16S rRNA gene sequences (Table 3-1). Within these datasets, the "Latescibacteria" invariably represented a minor fraction of the overall microbial community. The highest "Latescibacteria" relative abundance (4.37%) was identified in no-tillage cultivated soil samples from La Pampa Ondulada region, Argentina (28). "Latescibacteria" sequences were identified in only four habitat types: marine, freshwater, terrestrial, and bioremediation habitats.

Generally, broad agreements were observed when comparing the "Latescibacteria" community structures in near full-lengths and high throughput 16S rRNA gene datasets (Fig. 3-2A). Three "Latescibacteria" orders: Sediment_1, PBSIII_9,

and GN03 invariably represented the majority of "Latescibacteria" sequences in all habitats (Fig. 3-2A). Terrestrial habitats, in general, exhibited a limited "Latescibacteria" order level diversity and were dominated by members of the "Latescibacteria" order Sediment_1 (Fig. 3-2A). The community was highly similar between various sub-classifications of terrestrial habitats (Fig. 3-2B), and they all clustered in PCA biplot based on the overrepresentation of order Sediment_1 (Fig. 3-2C). In contrast, marine habitats exhibited a more even distribution of various "Latescibacteria" orders, with multiple orders identified in relatively high abundance in several sub-habitats (Fig. 3-2B). The majority of marine sub-habitats showed a similar structural composition defined by overrepresentation of order GN03, with few exceptions (e.g. relative overrepresentation of orders CV106 in coral reef, SAW1-B6 in pelagic, and SSS_58A in hydrothermal vent habitats) (Fig. 3-2B). Within freshwater habitats, a wide variation in community structure was observed across various sub-habitats examined. Order PBSIII_9 dominated temperate freshwater and ground water habitats, while order GN03 dominated saline/hypersaline and wastewater treatment habitats (Fig. 3-2B). The wider variation in community composition is reflected by the disparate position of various freshwater datasets in the PCA biplot (Fig. 3-2C). Finally, the community structure of bioremediation habitats revealed that, in addition to the three major "Latescibacteria" orders described above, members of the order SSS58A were also frequently encountered (Fig. 3-2B).

Global patterns of "Latescibacteria" distribution and community structure in metagenomic datasets. A total of 68.9 Mb of "Latescibacteria"-affiliated sequences were identified in all examined datasets. These fragments comprised 39,137 contigs

ranging in size between 1-287 Kbp (N50=17,072 Kbp, L50= 625bp), and contained 153 16S rRNA genes, and 73,079 protein-coding genes (Table 3-2). Detailed analysis of the largest ten contigs demonstrated significant overlap with available Latescibacteria genomes, and the extraction and phylogenetic analysis of these housekeeping genes using Phylosift (39) confirmed the affiliation of these fragments to the "Latescibacteria"; further confirming the specificity of the utilized approach for targeting the "Latescibacteria". Large (>10 Kbp contigs) represented 94.37%, while short (<1 Kbp) contigs represented only 5.63% of the total size (in bp) of identified "Latescibacteria". The majority of "Latescibacteria" fragments identified originated from marine (28.0 Mb, 35,974 protein-coding genes), freshwater (14.0 Mb, protein-coding 16,443 genes), terrestrial (12.1 Mb, 15,157 protein-coding genes), and bioremediation habitats (4.7 Mb, 5,505 protein-coding genes) (Tables 3-2). Since relatively limited information was obtained from host-associated, and engineered habitats, where none of the "Latescibacteria" identified contigs harbored a 16S rRNA gene, and since none of the examined air datasets contained "Latescibacteria"-affiliated contigs, we excluded these habitats from the analysis and focused our subsequent metabolic reconstruction efforts on the other four habitats; marine, freshwater, terrestrial, and bioremediation habitats. We successfully assigned 83.6% of the total "Latescibacteria" contigs (87.0% in marine, 76.8 % in freshwater, 84.8% in terrestrial, and 78.8% in bioremediation habitats) into orders. The community composition of these metagenomic bins (Fig. 3-3) was in broad agreement with that obtained using 16S rRNA gene analysis (Fig. 3-2A) with one notable exception: while relatively rare in 16S rRNA datasets, members of order MSB-4E2 represented a significant component of "Latescibacteria" metagenomic bins (Fig. 3-3).

"**Latescibacteria" polysaccharide degradation capabilities.** A large number of glycoside hydrolases (GHs, 717 genes), polysaccharide lyases (PLs, 100 genes), and carbohydrate esterases (CEs, 266 genes) were observed in "Latescibacteria" pangenomes (Fig. 3-4A). Within various "Latescibacteria" orders examined, order PBSIII_9 harbored the highest GH (21.6 GH/ Mb) and PL (3.5 PL/ Mb) densities (Fig. 3-4B), values that are comparable to those observed in the genomes of model lignocellulolytic (e.g. *Clostridium thermocellum,* 20.8 GHs/ Mb (33)), and pectinolytic (e.g. *Formosa agariphilia* (49)) bacteria. On the other hand, order Sediment_1 had the smallest repertoire and the lowest CAZyme density (11.37 GH and 1.76 PL/Mb, Fig. 3-5B) amongst all examined "Latescibacteria" orders.

Complete machineries for the degradation of fourteen different polysaccharides (cellulose, xylan, xyloglucan, mannan, pectin, starch, chitin, Poly-N-Acetyl- α -D-galactosamine, agar, porphyran, carrageenan, arabinogalactan, alginate, and fucans) were collectively identified within "Latescibacteria" pangenomes (Fig. 3-5). The absolute majority (13/14) of these polysaccharides degradation abilities were identified in all four "Latescibacteria" orders (Fig. 3-5A), with the exception of alginate that appears to be solely degraded by a single order (PBSIII_9). PCA analysis demonstrated the enrichment of specific polysaccharide degradation capabilities across "Latescibacteria" orders. For example, order PBSIII_9 pangenome was relatively enriched in porphyran, carrageenan, xylan, and fucans degradation capabilities (Fig. 3-5B), order GN03 pangenome was relatively enriched in agar and starch degradation capabilities (Fig. 3-5B), order MSB-4E2 pangenome was enriched in chitin, arabinogalactans, and xyloglucan degradation

capabilities (Fig. 3-5B), and order Sediment_1 pangenome was relatively enriched in pectin degradation capability (Fig. 3-5B)

Analysis of GH and PL distribution patterns across habitats demonstrated that the degradation capacities for seven polymers (cellulose, starch, mannan, xyloglucan, fucans, arabinogalactan and poly N-acetyl-galactosamine) were present in all four habitats, while the degradation capacities for three polymers (xylan, pectin, and chitin) were present in only three of the four habitats (marine, freshwater, and terrestrial). The machineries for degradation of agar, porphyran, and carrageenan were only observed in aquatic (freshwater and marine) habitats, and that for alginate was observed only in terrestrial habitats (Fig. 3-5C).

PCA analysis (Fig. 3-5D) indicated that marine habitats were relatively enriched in xylan, and sulfated galactans (agar, porphyran, and carrageenan) degradation capabilities. Sulfated galactans represent integral components of the cell wall of red algae (Rhodophyta) commonly encountered in marine habitats. Pangenomes from freshwater habitats were relatively enriched in pectin degradation capabilities. Finally, pangemomes from bioremediation habitats exhibited a relatively higher proportion in enzymes mediating starch degradation.

Of special interest is the observation that all “Latescibacteria” pangenomes were highly enriched in GH109 (Fig. 3-4A), a glycosyl hydrolase family harboring highly specific N-acetylgalactosaminidases that mediate the breakdown of poly-N-acetyl-galactosamine (50). Poly-N-acetyl-galactosamines are unique polymers that are present in the cell membrane of the zoospores of the green algus *Ulva* (51, 52). The wide distribution of GH109 within “Latescibacteria” pangenomes derived from various

habitats, and within all "Latescibacteria" orders is in stark contrast to the relatively narrow distribution of these enzymes in nature: CAZy database (September 2016) lists only 308 bacterial GH109 genes belonging to 9 phyla. Further, the density of GH109 in "Latescibacteria" pangenomes is significantly higher than those observed in genomes harboring these capabilities.

Cellulosomal components in "Latescibacteria" pangenomes. Cellulosomes are cell-bound extracellular structures harboring extracellular enzymes bound to scaffoldins (53). Cellulosomal supramolecular structure greatly enhances polymer-degrading capacities of cellulosome-harboring organisms (54, 55). Within the "Latescibacteria" pangenomes analyzed, we identified 27 genes harboring dockerin I domains, and 24 genes harboring cohesin I domains (Table 3-3). These genes belonged to metagenomes of four distinct anaerobic marine and freshwater environments: anoxic sediments of Sakinaw lake in British Columbia, CA (SL) (56, 57), anoxygenic microbial mats from Octopus pool, a thermal alkaline spring in the Lower Geyser Basin of Yellowstone National Park (YNP) (58), samples from expanding oxygen minimum zone (120 and 150 m depth) in Saanich Inlet in British Columbia, Canada (SI) (59), and samples from expanding oxygen minimum zones in the Northeastern Subarctic Pacific Ocean (NPO) (1). The identified genes belonged to contigs classified as members of the orders GN03, PBSIII_9, and MSB-4E2 (Table 3-3). Within the 27 identified dockerin-domain-containing genes, 25 also harbored a GH domain. These included GHs mediating cellulose (GH9 and GH124 endoglucanase), arabinoxylan (GH10 endoxylanase, and GH127 β-L-arabinofuranosidase), mannan (GH76 α-1,6-mannanase) and pectin (GH105 rhamnogalacturonyl hydrolase) degradation. The remaining two dockerin-domain-

containing genes encode the production of the extracellular enzymes phytase and lipase (Table 3).

Phylogenetically, “Latescibacteria” dockerin modules formed distinct phylogenetic clusters, with their closest relative being dockerin domains identified in *Clostridium acetobutylicum* (Fig. 3-6A). Dockerin type I domains typically contain two antiparallel α-helices (H1 and H3 in Fig. 3-6B, C) corresponding to the repeated amino acid sequences characteristic of the dockerin type 1 domain, linked by a short third alpha helix (H2) as well as two F hand motifs that confer the module with a dual-binding ability necessary for the flexible interactions between the type I dockerins and cohesins (60, 61). Sequence alignment of “Latescibacteria” dockerin I domains to those from model cellulosome-producing organisms (e.g. *Ruminiclostridium thermocellum*, *Clostridium cellulyticum*, *Ruminococcus* sp. and *Acetovibrio cellulyticus*) (Fig. 3-6B), as well as homology modeling by Iterative Threading ASSEmbly Refinement (I-TASSER) (48) using *Ruminiclostridium thermocellum* dockerin I domain three-dimensional model (Fig. 3-6C) identified conserved residues and characteristic alpha helix secondary structures and F hands in the “Latescibacteria” dockerin I domains suggestive of a functional domain.

In cellulosomes, dockerin-domain-containing enzymes are usually bound to a central cohesin-harboring scaffoldin through interactions between the type-I dockerin domains and the cohesin domains. Within the “Latescibacteria” pangenomes, we identified multiple genes encoding scaffoldin proteins that harbor cohesin domains (Table 3-3). These identified genes co-occurred with the dockerin domain-harboring genes in the same four anaerobic environments described above, belonged to orders PBSIII_9, GN03

and MSB-4E2, and were often located on the same contig with dockerin-domain-harboring genes. Further, within many of these cohesin-harboring scaffoldin-encoding genes, the occurrence of CBM37 was also identified (Tables 3-3). CBM37 domain is a characteristic component of cellulosomal scaffoldin proteins and is implicated in binding substrates to the cellulosome. Similar to dockerin domains, phylogenetic analysis demonstrated that "Latescibacteria" cohesin I domains exhibited significant sequence divergence when compared to those from other cultured cellulosome-harboring bacteria.

"Latescibacteria" proteolytic capabilities. A large number of aspartic peptidases, cysteine peptidases, metallo-peptidases, and serine peptidases (635 total, 9.22/ Mb) were identified in all "Latescibacteria" pangenomes (Table 3-4). Notably, the metallo-peptidase family M23 was overrepresented in freshwater, and marine habitats, corresponding to 9.52%, and 19.25% of the total peptidases in these habitats, respectively. M23 peptidase family enzymes are extracellular beta-lytic endopeptidases that lyse N-acylmuramoyl-alanine bonds between peptidoglycan and the cross-linking peptides in bacterial cell walls (62). Their abundance in "Latescibacteria" freshwater, and marine pangenomes suggests that bacterial cell lysis represents a potential nutrient acquisition strategy for marine and freshwater "Latescibacteria". In addition, we identified multiple collagenases (U32 family of peptidases) in terrestrial "Latescibacteria" pangenomes belonging to order Sediment_1. We speculate that these collagenases could initiate the degradation of extensins or hydroxyproline rich glycoproteins (HRGP) present in plant cell walls, and/or collagens from nematodes' carcasses (63).

Nutrients transport in "Latescibacteria" pangenomes. Collectively, the "Latescibacteria" pangenomes possessed a wide range of transporters for the uptake of

sugars, amino acid monomers and oligomers, and fatty acids (Fig. 3-7A). Briefly, genes encoding the following transporters were identified in all "Latescibacteria" order and habitat bins: phosphotransferase system (PTS) transporters for mannose, fructose, and N-acetyl galactosamine, hexose uniporters for galactose, ABC as well as Na^+ symporters for glucose, ExtU transporters for the uptake of various uronic acids, ABC transporters for amino acids, di-/oligo-peptide, and fatty acids, as well as transporters for mono-, di-, and tri-carboxylic acids.

In addition, multiple transporters were identified only within a specific habitat and/or "Latescibacteria" order. For example, proton symporters for rhamnose and fucose uptake were only identified in the GN03 and PBSIII_9 pangenomes, and ABC transporters specific for the osmo- and cryo-protectant compatible solutes glycine betaine and trehalose were identified only in marine "Latescibacteria" contigs belonging to the orders GN03 and MSB-4E2. Interestingly, 2-oxaloglutarate: malate antiporters were identified in soil "Latescibacteria" belonging to order Sediment_1. Malate has previously been shown to be secreted by plant roots to recruit beneficial microbes (64). The presence of such transporters might suggest that malate could potentially serve as a carbon source for members of "Latescibacteria" order Sediment_1 in soil, where malate will be converted to pyruvate by the oxaloacetate decarboxylating malate dehydrogenase.

Central metabolic pathways in "Latescibacteria" pangenomes. Multiple central metabolic pathways were identified in the "Latescibacteria" pangenomes, a reflection of the phylum pervasive capacity for polymers degradation. Genes for the glycolytic (EMP) pathway for glucose metabolism, for channeling galactose, fructose, mannose, fucose, rhamnose and uronic acids to the EMP pathway, as well as for the pentose phosphate

pathway for xylose and arabinose metabolism were identified in all "Latescibacteria" habitat and order bins (Fig. 3-7B).

Both fermentative and respiratory abilities were encountered in the "Latescibacteria". This is in contrast to the strict anaerobic fermentative capabilities previously reported from the analysis of single cell "Latescibacteria" genomes (16). Genes for pyruvate conversion to lactate, acetate, formate, propanoate, acetoin, and ethanol were widely distributed in all "Latescibacteria" habitats and orders bins (Fig. 3-7B). On the other hand, respiratory pathways were only identified in datasets obtained from seemingly oxygenated habitats. These included: 1. Pyruvate oxidative decarboxylation to acetyl-coA using pyruvate dehydrogenase (PD) complex, which was identified in multiple freshwater and marine "Latescibacteria" fragments affiliated with orders PBSIII_9 and GN03. 2. A complete TCA pathway, a strong indicator of respiratory activity, identified in multiple contigs derived from multiple pelagic marine habitats, and affiliated with orders PBSIII_9 and GN03, and 3. A complete aerobic respiratory chain identified only in a few studies from oxygenated and partially/seasonally oxygenated habitats, e.g. surface layers of Sakinaw Lake, and multiple pelagic marine locations within the global ocean sampling survey (Fig. 3-7B). Such respiratory chain identified was composed of four major subunits, NADH:quinone oxidoreductase (complex I), succinate dehydrogenase (complex II), cytochrome bc1 complex (complex III), and cytochrome C oxidase (complex IV). Interestingly, two types of cytochrome C oxidases were identified: 1. A Cbb3 type (Cbb3-cox) that functions optimally under microxic conditions (65, 66), identified in both freshwater and marine "Latescibacteria" in contigs affiliated with order PBSIII_9, and 2. A Caa3 type

cyctochrome C oxidase that functions under high oxygen tensions, was identified solely in marine "Latescibacteria" contigs affiliated with orders GN03 and PBSIII_9 (Fig. 3-7B).

In addition, in contrast to the BMC-dependent anaerobic lactaldehyde degradation (an intermediate produced during fucose and rhamnose metabolism) that was identified in the previously analyzed single cell "Latescibacteria" genomes (16), a complete pathway for fucose and rhamnose degradation through lactaldehyde oxidation to lactate by lactaldehyde dehydrogenase (LD), an oxygen-requiring NAD^+-dependent oxidoreductase, was identified in contigs belonging to order PBSIII_9 from the surface layers of Sakinaw Lake (56), which further extends the phylum capability beyond these identified through single cell genomes analysis (Fig. 3-7B). Additionally, a full pathway for the transport and metabolism of glycine betaine, a compatible solute commonly present in marine environments, was identified in habitats and orders pangenomes, including the oxygen-dependent enzyme sarcosine oxidase for the conversion of sarcosine to glycine, attesting to the phylum aerobic potential.

Finally, it is interesting to note that a complete Wood-Ljungdahl (WL) pathway for acetogenesis was detected in pangenomes of "Latescibacteria" inhabiting bioremediation habitats (Fig. 3-7). "Latescibacteria" contigs encoding the WL enzymes were exclusively affiliated with orders PBSIII_9 and MSB-4E2. The detection of WL pathway encoding genes in "Latescibacteria"-affiliated contigs suggest the potential capacity of "Latescibacteria" for utilization of acetogenesis as an alternative mechanism, for example when favorable polysaccharide and proteinaceous substrates commonly degraded by the "Latescibacteria" communities are scarce.

Table 3-1 "Latescibacteria" habitat and sub-habitat level distribution based on 16S rRNA genes in high throughput generated datasets.

Habitat	Sub-habitat	Number of datasets	Datasets with "Latescibacteria" >0.1%	Total number of 16S rRNA gene sequences analyzed	Number of "Latescibacteria" 16S rRNA gene sequences
Marine	Deep marine sediments	87	65	1.90 $X10^6$	12,369
	Coral associated microbiome	25	1	1.93 $X10^6$	323
	Pelagic	289	15	4.03 $X10^7$	6,538
	Hydrothermal vents	288	10	3.88 $X10^6$	3,200
	Coastal/Estuary	609	34	3.99 $X10^7$	21,621
	Total	**1,298**	**125**	**8.80$X10^7$**	**44,051**
Freshwater	Spring and ground water	26	2	4.47 $X10^5$	165
	Saline/ hypersaline environments	118	29	3.02 $X10^6$	14,567
	Temperate freshwater	1,572	38	8.16 $X10^7$	26,744
	Waste water	236	9	6.10 $X10^5$	2,221
	Total	**1,952**	**78**	**8.57 $X10^7$**	**43,697**
Terrestrial	Agriculture soil	76	25	1.55 $X10^7$	24,916
	Contaminated soil	59	5	6.14 $X10^5$	974
	Forest	194	10	1.76 $X10^5$	1,985
	Permafrost	264	3	3.18 $X10^5$	1,612
	Grassland	25	23	2.12 $X10^7$	18,082
	Soil/Other	106	5	1.12 $X10^6$	13,789
	Total	**724**	**71**	**4.33 $X10^7$**	**61,358**
Bioremediation[a]	**Total**	**24**	**11**	**1.12 $X10^6$**	**648**
Host-Associated	**Total**	**43**	**0**	**2.26$X10^7$**	**0**
Total		**4,041**	**285**	**2.2$X10^8$**	**149,754**

[a] Within all 24 bioremediation datasets examined, the "Latescibacteria" sequences (648 total sequences) never exceeded 0.1% of relative abundance.

Table 3-2. "Latescibacteria" sub-habitat level distribution of metagenomic datasets analyzed in this study.

Habitat	Habitat subclass	Datasets analyzed		Number of Datasets harboring "Latescibacteria" contigs	"Latescibacteria" contigs identified	
		Total number	Total size (Mbp)		Total number	Total size (Mbp)
Marine	Estuary	40	$8.17X10^{9}$	34	3,286	4.45
	Deep Sediment	92	$9.98X10^{9}$	80	1,735	1.95
	Hydrothermal vents	13	$6.80X10^{9}$	12	1,191	2.2
	Pelagic	259	$5.05X10^{10}$	219	8,038	19.4
	Total	**404**	$7.55X10^{10}$	**345**	14,250	**28.0**
Freshwater	Freshwater	159	$1.28X10^{10}$	140	1,694	1.76
	Groundwater/ Thermal springs	116	$1.93X10^{10}$	64	4,899	6.84
	Non-marine saline	40	$1.04X10^{10}$	32	2,263	5.38
	Total	**315**	$4.25X10^{10}$	**189**	8,856	**14.0**
Terrestrial	Agriculture	202	$1.14X10^{8}$	53	537	0.0244
	Arid/semi arid	6	$1.18X10^{9}$	6	210	0.0431
	Forest	198	$1.59X10^{10}$	83	510	0.489
	Grassland	82	$2.30X10^{10}$	38	2,657	1.25
	Contaminated/Treated	76	$2.03X10^{10}$	66	4,062	7.35
	Permafrost	53	$9.35X10^{9}$	52	1,146	2.4
	Rhizosphere	49	$6.98X10^{9}$	31	759	0.547
	Total	**666**	$7.69X10^{10}$	**329**	9,881	**12.1**
Bioremediation	Bioremediation	48	$1.12X10^{10}$	32	1,716	**4.7**
Host-associated	Animal	22	$6.86X10^{9}$	9	214	2.42
	Insects	17	$1.67X10^{10}$	13	1,082	1.27
	Plants	13	$2.08X10^{9}$	10	227	0.112
	Total	**52**	$2.57X10^{10}$	**31**	1,523	**3.8**
Engineered	Solid waste/compost	16	$4.65X10^{9}$	9	350	0.417
	Lab enrichment	13	$7.63X10^{8}$	12	95	0.269
	Bioreactor	46	$7.17X10^{9}$	24	1,810	5.61
	Total	**75**	$1.26X10^{10}$	**41**	2,255	**6.3**
Air	Air	29	$3.59X10^{8}$	0	0	0
Total		**1,589**	$2.44X10^{11}$	**907**	**39,137**	**68.9**

Table 3-3. Potential cellulosomal elements identified in the "Latescibacteria" pangenomes[a]

Structure	Function	Total number	Dataset source[b]				Order Level Classification				
			SI	NPO	SL	YNP	Sediment 1	GN03	PBSIII_9	MSB-4E2	NA[c]
Dockerin	Structural element in cellulosomes	29	21	2	5	1	0	5	12	12	0
Cohesin	Structural element in cellulosomes	24	21	2	0	1	0	4	14	6	0
Scaffoldin	Structural element in cellulosomes	46	42	4	0	0	0	8	26	12	0
CAZymes											
CE1	Acetyl xylan esterase	1	1	0	0	0	0	1	0	0	0
CE7	Acetyl xylan esterase	1	0	0	0	1	0	0	1	0	0
GH10	Endo- 1,4 β-xylanase	1	0	0	0	1	0	0	1	0	0
GH105	Unsaturated rhamnogalacturonyl hydrolase	19	17	2	0	0	0	4	10	5	0
GH124	Endoglucanase	22	17	2	3	0	0	4	10	8	0
GH127	β-L-arabinofuranosidase	19	17	2	0	0	0	4	10	5	0
GH76	α-1, 6-mannanase	19	17	2	0	0	0	4	10	5	0
GH9	Endoglucanase	19	17	2	0	0	0	4	10	5	0
CBM37	Carbohydrate binding Module with broad binding specificity	4	4	0	0	0	0	0	3	1	0
Non-CAZymes											
Lipase	Lipid-degrading enzyme	1	1	0	0	0	0	0	0	1	0
Phytase	hysrolizing P-containing compounds in grains and oil seeds	1	1	0	0	0	0	0	1	0	0

[a] A complete list of contigs harboring these genes/domains is presented in table S3.
[b] SI= Saanich inlet, NPO=North Pacific Ocean, SL: Sakinaw lake, YNP: Yellowstone National Park.
[c] Unassigned to any of the four major "Latescibacteria" lineages.

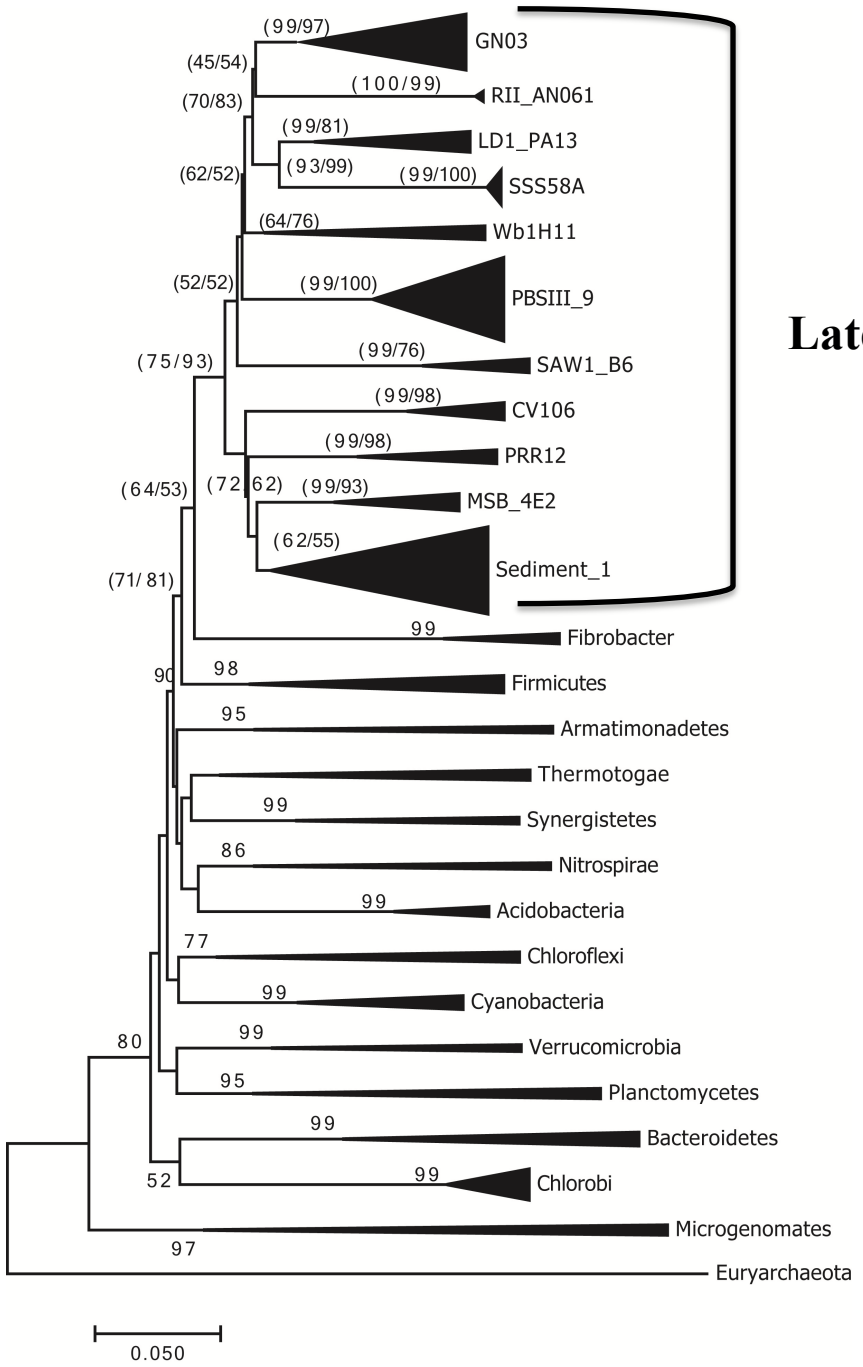
Table 3-4. List of peptidases potentially encoded by "Latescibacteria" pangenomes.

Class	Total	Habitat				Phylogenetic Affiliation				
		F	M	T	B	Sediment_1	PBSIII_9	GN03	MSB-4E2	NA[b]
1. Aspartic (A) Peptidases										
A24A	8	2	7	0	1	0	0	0	2	8
2.Cysteine (C) Peptidases										
C01A	2	2	0	0	0	0	0	1	0	1
C26	53	10	32	8	3	6	32	9	5	1
C40	3	0	3	0	0	0	2	0	1	0
C44	14	5	7	2	0	0	4	2	4	4
3.Metallo (M) Peptidases										
M01	25	2	23	0	0	0	12	5	8	0
M03A	6	1	3	2	0	0	5	0	1	0
M03B	3	1	1	1	0	0	3	0	0	0
M10B	2	0	1	1	0	0	1	0	1	0
M12B	1	0	1	0	0	0	1	0	0	0
M14A	5	3	0	1	1	2	1	1	1	0
M14B	1	1	0	0	0	1	0	0	0	0
M15B	2	0	0	0	2	0	0	2	0	0
M16B	25	7	6	11	1	2	17	1	4	1
M17	11	0	11	0	0	0	5	5	1	0
M18	2	1	1	0	0	0	0	1	1	0
M19	3	0	2	0	1	1	2	0	0	0
M20A	2	1	1	0	0	0	2	0	0	0
M20B	4	0	1	3	0	0	1	0	0	3
M20C	2	1	0	0	1	0	1	0	0	1
M20D	11	5	4	1	1	0	7	2	1	1
M23B	86	14	62	4	6	12	25	12	26	11
M24A	48	8	24	13	3	3	26	8	8	3
M24B	12	5	7	0	0	0	6	2	2	2
M28A	5	0	0	4	1	0	0	0	3	2
M28D	8	0	6	2	0	0	4	0	3	1
M28E	2	2	0	0	0	0	2	0	0	0
M38	15	8	2	3	2	0	5	6	1	3
M41	28	0	27	0	1	0	13	4	3	8
M48B	25	7	2	15	1	8	1	1	8	7
M50	14	7	3	4	0	2	5	1	1	5
M79	10	1	0	9	0	0	9	0	1	0
4.Serine (S) Peptidases										
S01C	67	13	46	2	6	1	30	10	17	9
S08A	11	5	1	3	2	2	1	4	3	1
S09	20	2	8	9	1	0	0	0	4	16
S11	3	0	1	1	1	2	1	0	0	0

S12	7	1	6	0	0	0	4	1	1	1
S13	5	1	0	0	4	0	2	2	1	0
S14	19	7	4	2	6	3	4	6	4	2
S16	20	10	4	2	4	0	4	5	6	5
S33	12	3	7	1	1	1	4	3	3	1
S41A	19	8	7	0	4	1	10	1	7	0
S49	9	3	1	4	1	1	5	1	2	0
5.Peptidases of Unknown Catalytic Type										
U32	3	0	0	3	0	0	3	0	0	0

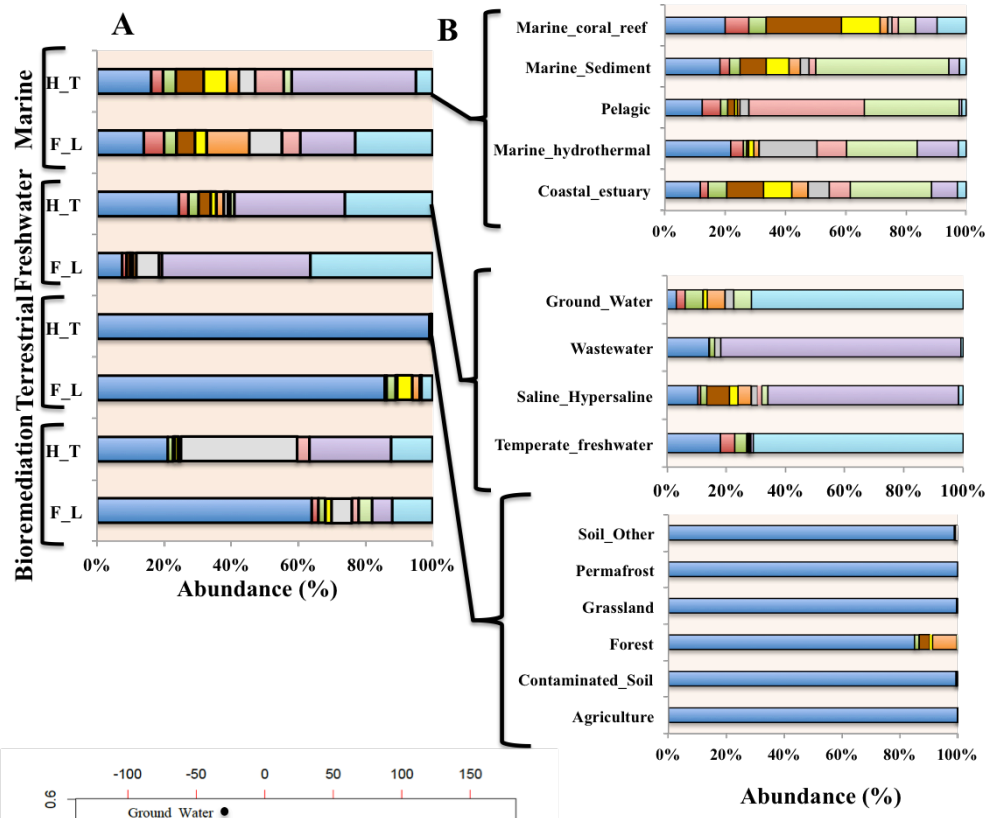
[a] F: freshwater, M: marine, T: terrestrial, B: bioremediation

[b] Unassigned to any of the four major “Latescibacteria” lineages.



Latescibacteria

Figure 3-1: Phylogenetic affiliation of the “Latescibacteria” and clades within to other bacterial phyla based on the sequences of the 16S rRNA gene. The tree was obtained using both Neighbor-Joining and maximum likelihood approaches with 157 sequences as the outgroup. The topology of both NJ and ML trees overlapped and bootstrap values (from 100 replicates) obtained are shown between parentheses (NJ/ ML) for nodes with more than 50% bootstrap support as. The analysis involved 164 nucleotide sequences. All ambiguous positions were removed for each sequence pair. There were a total of 1837 positions in the final dataset. Evolutionary analyses were conducted in MEGA7.


A
B
Bioremediation
Terrestrial
Freshwater
Marine
H_T
F_L
0%
20%
40%
60%
80%
100%
Abundance (%)
Marine_coral_reef
Marine_Sediment
Pelagic
Marine_hydrothermal
Coastal_estuary
Ground_Water
Wastewater
Saline_Hypersaline
Temperate_freshwater
Soil_Other
Permafrost
Grassland
Forest
Contaminated_Soil
Agriculture
Abundance (%)

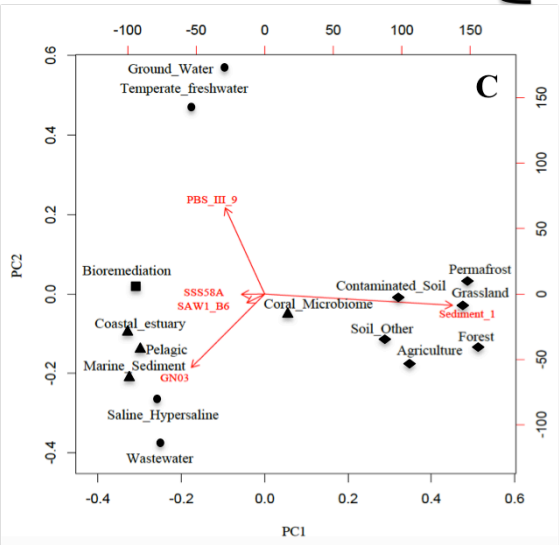

C
PC1
PC2
Ground_Water
Temperate_freshwater
PBS_III_9
Bioremediation
SSS58A
SAW1_B6
Coral_Microbiome
Contaminated_Soil
Permafrost
Grassland
Sediment_1
Coastal_estuary
Soil_Other
Forest
Pelagic
Agriculture
Marine_Sediment
GN03
Saline_Hypersaline
Wastewater

Figure 3-2. (A) Community structure of the "Latescibacteria" sequences identified in amplicon-generated near full-lengths (F_L), and high throughput (H_T) 16S rRNA gene datasets from marine, freshwater, terrestrial, and bioremediation habitats. (B**)** Sub-habitats level classification of "Latescibacteria" 16S rRNA gene sequences from marine (top), freshwater (middle), and terrestrial (bottom) habitats in high throughput datasets. The relatively limited number (1167 sequences) of near full-length sequences precluded conducting a similar analysis. (C) PCA biplot of the "Latescibacteria" community structure in different habitats. The plot was generated using the relative abundances of "Latescibacteria" orders in high throughput generated 16S rRNA datasets as an input. Top and right axes are the axis scores for the samples (no units). Bottom and left axes are the loadings of the variables (PC1 and PC2). The first two most important components are plotted. Sub-habitats are shown for marine (▲), freshwater (●), terrestrial (◆), and bioremediation (■) habitats.

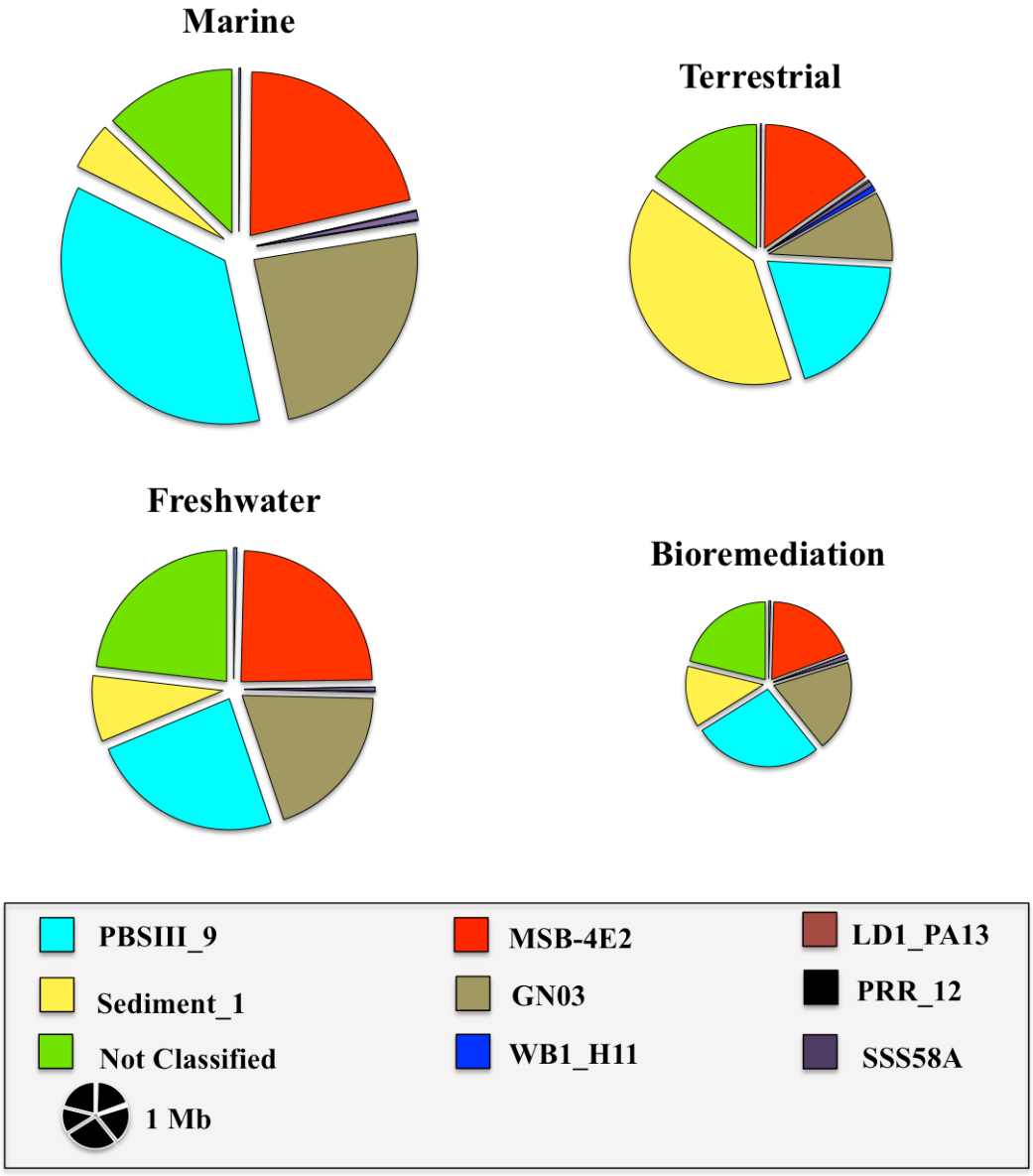


Figure 3-3. Order level classification of the different "Latescibacteria" pangenomes identified in metagenomic datasets analyzed. Circles are drawn to scale with 1 Mb shown in the legend.

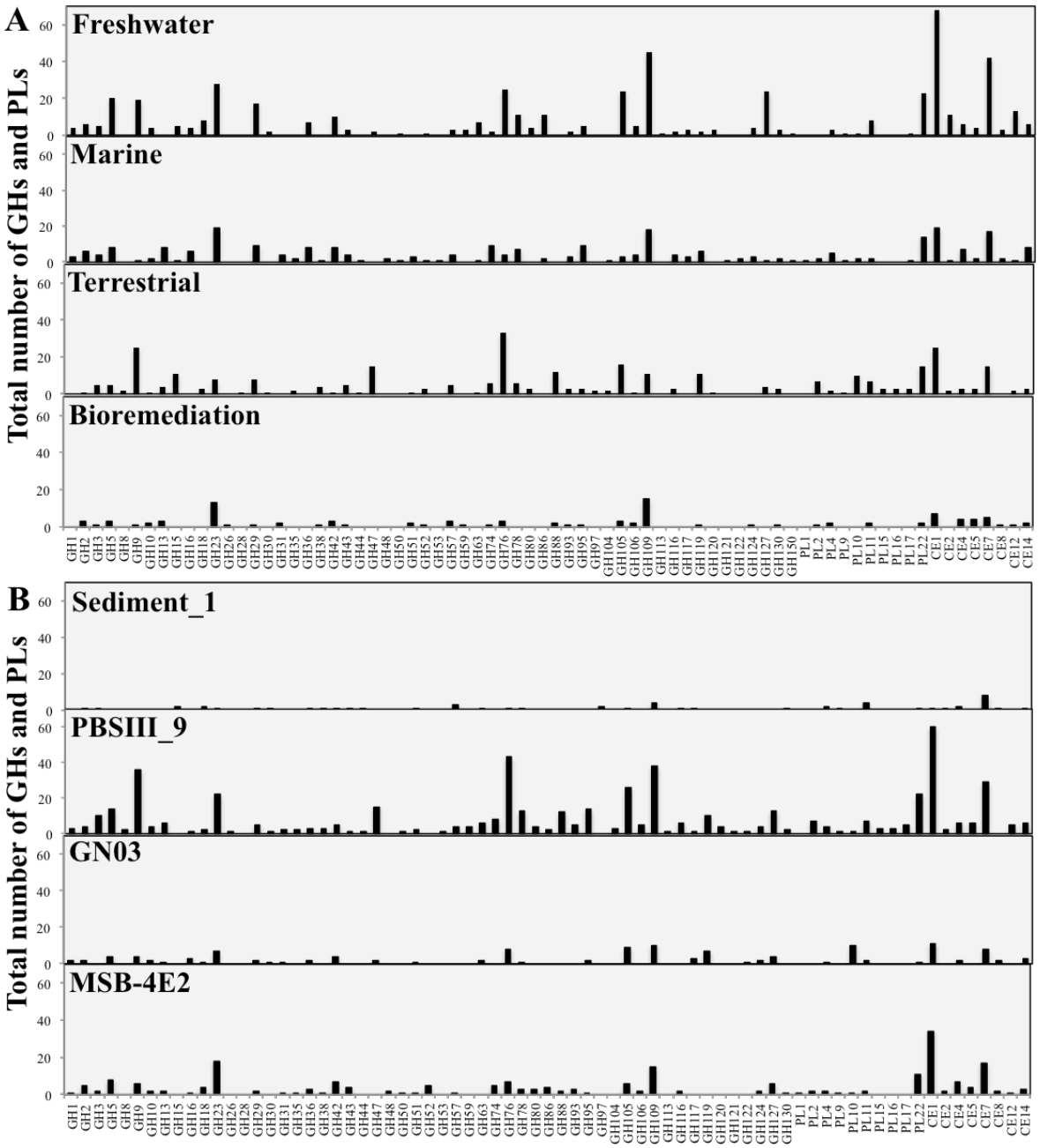


Figure 3-4. CAZyme family classification in "Latescibacteria" pangenomes based on habitat type (A), and order level classification (B).

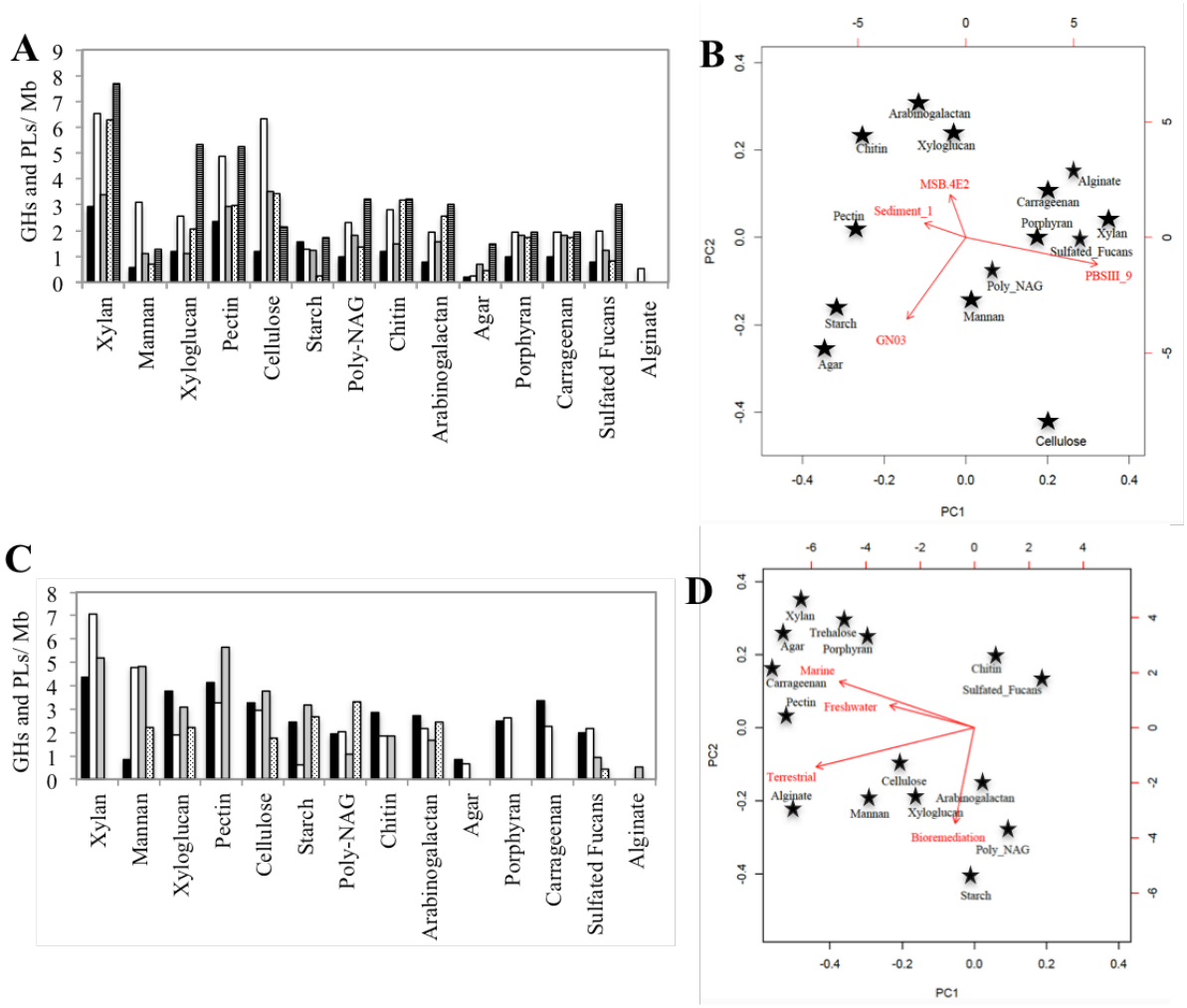


Figure 3-5. Relative density (number per 1 Mb) of CAZymes (GHs and PLs) targeting specific polysaccharides (A), and PCA biplot (B) depicting polysaccharide degradation patterns in "Latescibacteria" order pangenomes. Column symbols in panel A are black: Sediment 1, white: PBSIII_9, Grey: GN03, dotted: MSB-4E2, and horizontal stripes: other lineages. Relative density (number per 1 Mb) of CAZymes (GHs and PLs) targeting specific polysaccharides (C), and PCA biplot (D) depicting polysaccharide degradation patterns in "Latescibacteria" habitat pangenomes. Column symbols in panel C are black: Freshwater, white: marine, grey: terrestrial, and dotted:bioremediation. Top and right axes in the PCA plot (panel D) are the axis scores for the samples (no units). Bottom and left axes are the loadings of the variables (PC1 and PC2). The first two most important components are plotted.

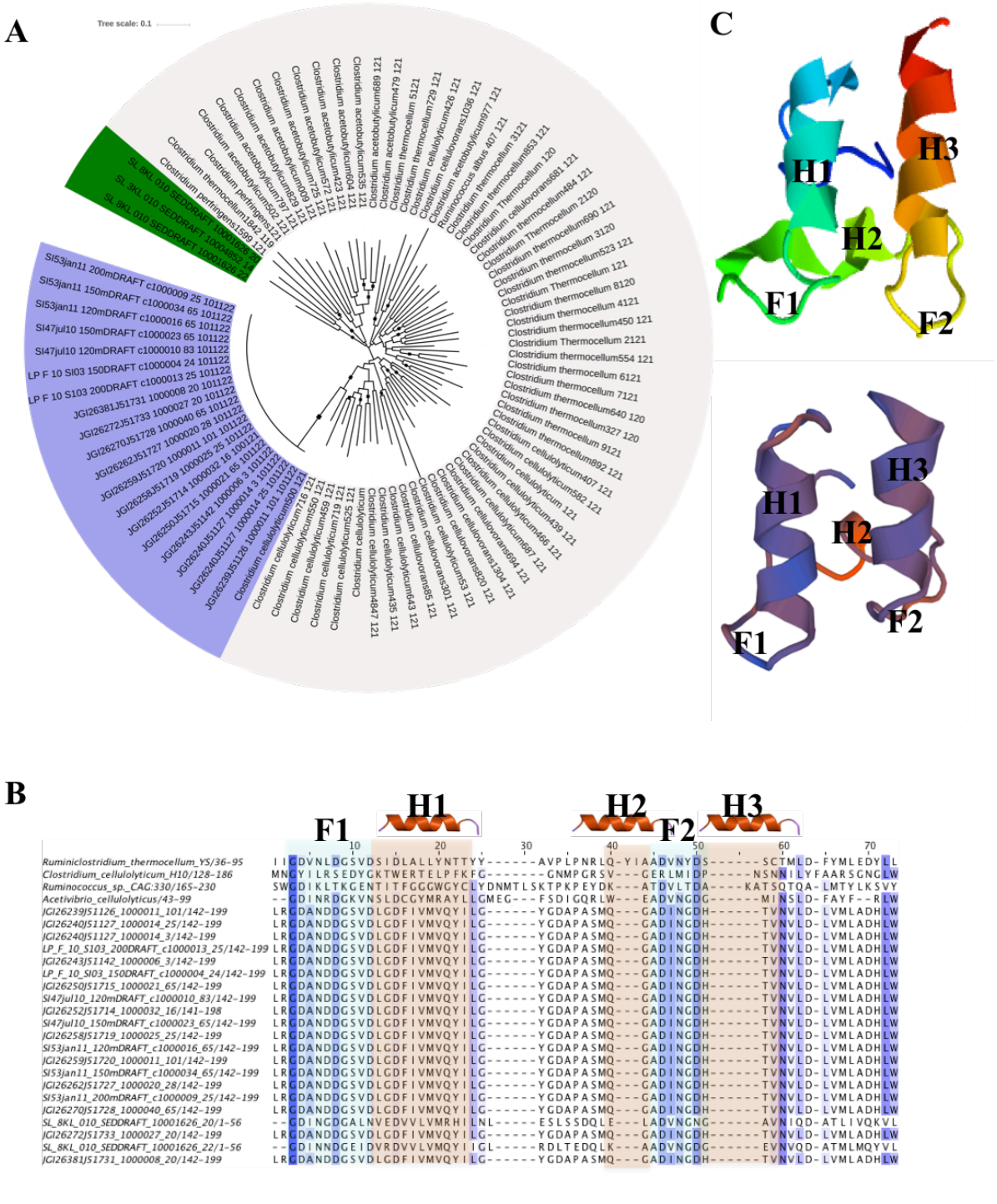
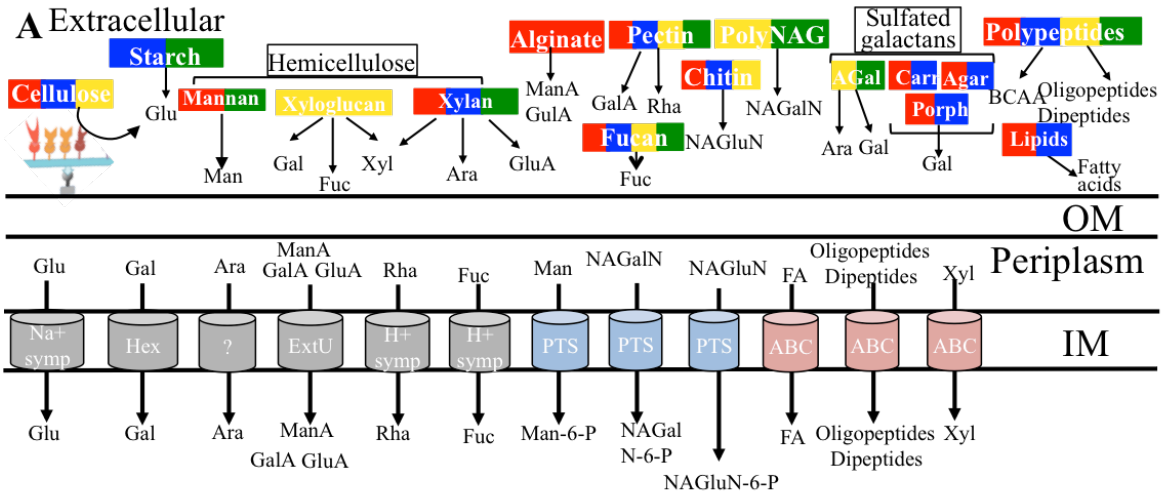

A
Tree scale: 0.1
SL 8KL 010 SEDDRAFT 10001626 22
SL 3KL 010 SEDDRAFT 10004552 21
SL 8KL 010 SEDDRAFT 10001626 21
SI53jan11 200mDRAFT c1000009 25 101122
SI53jan11 150mDRAFT c1000034 65 101122
SI53jan11 120mDRAFT c1000016 65 101122
SI47jul10 150mDRAFT c1000023 65 101122
SI47jul10 120mDRAFT c1000010 83 101122
LP F 10 SI03 150DRAFT c1000004 24 101122
LP F 10 S103 200DRAFT c1000013 25 101122
JGI26381J51731 1000008 20 101122
JGI26272J51733 1000027 20 101122
JGI26270J51728 1000040 65 101122
JGI26262J51727 1000020 28 101122
JGI26259J51720 1000011 101 101122
JGI26258J51719 1000025 25 101122
JGI26252J51714 1000032 16 101122
JGI26250J51715 1000021 65 101122
JGI26243J51142 1000006 3 101122
JGI26240J51127 1000014 25 101122
JGI26239J51126 1000011 101 101122
Clostridium cellulolyticum500 121
Clostridium cellulolyticum716 121
Clostridium cellulolyticum550 121
Clostridium cellulolyticum459 121
Clostridium cellulolyticum719 121
Clostridium cellulolyticum525 121
Clostridium cellulolyticum
Clostridium cellulolyticum4847 121
Clostridium cellulolyticum435 121
Clostridium cellulolyticum643 121
Clostridium cellulovorans885 121
Clostridium cellulolyticum53 121
Clostridium cellulovorans301 121
Clostridium cellulovorans820 121
Clostridium cellulovorans1304 121
Clostridium cellulolyticum687 121
Clostridium cellulolyticum466 121
Clostridium cellulolyticum439 121
Clostridium cellulolyticum121
Clostridium cellulolyticum582 121
Clostridium cellulolyticum407 121
Clostridium thermocellum892 121
Clostridium thermocellum9121
Clostridium thermocellum327 120
Clostridium thermocellum640 120
Clostridium thermocellum7121
Clostridium thermocellum6121
Clostridium thermocellum554 121
Clostridium Thermocellum 2121
Clostridium thermocellum450 121
Clostridium thermocellum4121
Clostridium thermocellum8120
Clostridium thermocellum121
Clostridium thermocellum523 121
Clostridium thermocellum3120
Clostridium thermocellum690 121
Clostridium thermocellum2120
Clostridium thermocellum484 121
Clostridium cellulovorans681 121
Clostridium Thermocellum 120
Clostridium thermocellum853 121
Clostridium thermocellum3121
Ruminococcus albus 407 121
Clostridium acetobutylicum977 121
Clostridium cellulovorans1036 121
Clostridium cellulolyticum426 121
Clostridium thermocellum729 121
Clostridium thermocellum 5121
Clostridium thermocellum479 121
Clostridium acetobutylicum689 121
Clostridium acetobutylicum535 121
Clostridium acetobutylicum604 121
Clostridium acetobutylicum572 121
Clostridium acetobutylicum423 121
Clostridium acetobutylicum625 121
Clostridium acetobutylicum029 121
Clostridium acetobutylicum791 121
Clostridium acetobutylicum902 121
Clostridium perfringens1842 119
Clostridium thermocellum1599 121
Clostridium perfringens1516 121
C
H1
H2
H3
F1
F2
H1
H2
H3
F1
F2
B
F1
H1
H2
F2
H3
10
20
30
40
50
60
70
Ruminiclostridium_thermocellum_YS/36-95 IIGDVNLDGSVDSIDLALLYNTTYY------AVPLPNRLQ-YIAADVNYDS------SCTMLD-FYMLEDYLL
Clostridium_cellulolyticum_H10/128-186 MNGYILRSEDYGKTWERTELPFKFG------GNMPGRSV---GERLMIDP-----NSNNILYFAARSGNGLW
Ruminococcus_sp._CAG:330/165-230 SWGDIKLTKGENTITFGGGWGYCLYDNMTLSKTPKPEYDK---ATDVLTDA----KATSQTQA-LMTYLKSVY
Acetivibrio_cellulolyticus/43-99 --GDINRDGKVNSLDCGYMRAYLLGMEG---FSDIGQRLW---EADVNGDG------MINSLD-FAYF--RLW
JGI26239J51126_1000011_101/142-199 LRGDANDDGSVDLGDFIVMVQYILG------YGDAPASMQ---GADINGDH------TVNVLD-LVMLADHLW
JGI26240J51127_1000014_25/142-199 LRGDANDDGSVDLGDFIVMVQYILG------YGDAPASMQ---GADINGDH------TVNVLD-LVMLADHLW
JGI26240J51127_1000014_3/142-199 LRGDANDDGSVDLGDFIVMVQYILG------YGDAPASMQ---GADINGDH------TVNVLD-LVMLADHLW
LP_F_10_S103_200DRAFT_c1000013_25/142-199 LRGDANDDGSVDLGDFIVMVQYILG------YGDAPASMQ---GADINGDH------TVNVLD-LVMLADHLW
JGI26243J51142_1000006_3/142-199 LRGDANDDGSVDLGDFIVMVQYILG------YGDAPASMQ---GADINGDH------TVNVLD-LVMLADHLW
LP_F_10_SI03_150DRAFT_c1000004_24/142-199 LRGDANDDGSVDLGDFIVMVQYILG------YGDAPASMQ---GADINGDH------TVNVLD-LVMLADHLW
JGI26250J51715_1000021_65/142-199 LRGDANDDGSVDLGDFIVMVQYILG------YGDAPASMQ---GADINGDH------TVNVLD-LVMLADHLW
SI47jul10_120mDRAFT_c1000010_83/142-199 LRGDANDDGSVDLGDFIVMVQYILG------YGDAPASMQ---GADINGDH------TVNVLD-LVMLADHLW
JGI26252J51714_1000032_16/141-198 LRGDANDDGSVDLGDFIVMVQYILG------YGDAPASMQ---GADINGDH------TVNVLD-LVMLADHLW
SI47jul10_150mDRAFT_c1000023_65/142-199 LRGDANDDGSVDLGDFIVMVQYILG------YGDAPASMQ---GADINGDH------TVNVLD-LVMLADHLW
JGI26258J51719_1000025_25/142-199 LRGDANDDGSVDLGDFIVMVQYILG------YGDAPASMQ---GADINGDH------TVNVLD-LVMLADHLW
SI53jan11_120mDRAFT_c1000016_65/142-199 LRGDANDDGSVDLGDFIVMVQYILG------YGDAPASMQ---GADINGDH------TVNVLD-LVMLADHLW
JGI26259J51720_1000011_101/142-199 LRGDANDDGSVDLGDFIVMVQYILG------YGDAPASMQ---GADINGDH------TVNVLD-LVMLADHLW
SI53jan11_150mDRAFT_c1000034_65/142-199 LRGDANDDGSVDLGDFIVMVQYILG------YGDAPASMQ---GADINGDH------TVNVLD-LVMLADHLW
JGI26262J51727_1000020_28/142-199 LRGDANDDGSVDLGDFIVMVQYILG------YGDAPASMQ---GADINGDH------TVNVLD-LVMLADHLW
SI53jan11_200mDRAFT_c1000009_25/142-199 LRGDANDDGSVDLGDFIVMVQYILG------YGDAPASMQ---GADINGDH------TVNVLD-LVMLADHLW
JGI26270J51728_1000040_65/142-199 LRGDANDDGSVDLGDFIVMVQYILG------YGDAPASMQ---GADINGDH------TVNVLD-LVMLADHLW
SL_8KL_010_SEDDRAFT_10001626_20/1-56 --GDINGDGALNVEDVVLVMRHILNL-----ESLSSDQLE---LADVNGNG------AVNIQD-ATLIVQKVL
JGI26272J51733_1000027_20/142-199 LRGDANDDGSVDLGDFIVMVQYILG------YGDAPASMQ---GADINGDH------TVNVLD-LVMLADHLW
SL_8KL_010_SEDDRAFT_10001626_22/1-56 --GDINNDGEIDVRDVVLVMQYIIGL-----RDLTEDQLK---AADVNGDG------EVNVQD-ATMLMQYVL
JGI26381J51731_1000008_20/142-199 LRGDANDDGSVDLGDFIVMVQYILG------YGDAPASMQ---GADINGDH------TVNVLD-LVMLADHLW

Figure 3-6. (A) Distance dendogram depicting the phylogenetic relationship between dockerin I domains (PF00404) extracted from "Latescibacteria" genes and those obtained from model cellulosome-harboring organisms. The tree was constructed using distance neighbor joining (NJ) approach with Jukes-Cantor corrections. Sequence names reflect their order level affiliation and study site. Domains extracted from genes in marine datasets are shaded in blue while those extracted from freshwater datasets are shaded in green. The tree is bootstrapped based on 100 replications, the bootstrap values are shown as squares for the clades with > 50% bootstrap support, where the square size is proportional to the bootstrap score. (B) Multiple sequence alignment of freshwater and marine "Latescibacteria" type I Dockerin domains (shown in A) to those of model Gram-positive cellulosome-producing organisms. Beige-shaded areas are predicted alpha helix domains while blue-shaded areas correspond to F hands. The sequences were aligned with an estimated TM-score 0.84±0.08. (C) 3D structure prediction of the "Latescibacteria" docekrin type I domain sequence (top) compared to the structure of docekrin type I domain of *Ruminiclostridium thermocellum* using I-TASSER three-dimensional model (bottom). The "Latescibacteria" domain was modeled with a confidence score of 0.95.


A Extracellular
Starch
Cellulose
Glu
Hemicellulose
Mannan
Xyloglucan
Xylan
Man
Gal
Fuc
Xyl
Ara
GluA
Alginate
ManA
GulA
Pectin
GalA
Rha
PolyNAG
NAGalN
Chitin
Fucan
NAGluN
Fuc
Sulfated galactans
AGal
Carr
Agar
Porph
Ara
Gal
Gal
Polypeptides
BCAA
Oligopeptides
Dipeptides
Lipids
Fatty acids
OM
Periplasm
IM
Glu
Gal
Ara
ManA
GalA
GluA
Rha
Fuc
Man
NAGalN
NAGluN
FA
Oligopeptides
Dipeptides
Xyl
Na+ symp
Hex
?
ExtU
H+ symp
H+ symp
PTS
PTS
PTS
ABC
ABC
ABC
Glu
Gal
Ara
ManA
GalA GluA
Rha
Fuc
Man-6-P
NAGal
N-6-P
NAGluN-6-P
FA
Oligopeptides
Dipeptides
Xyl

B

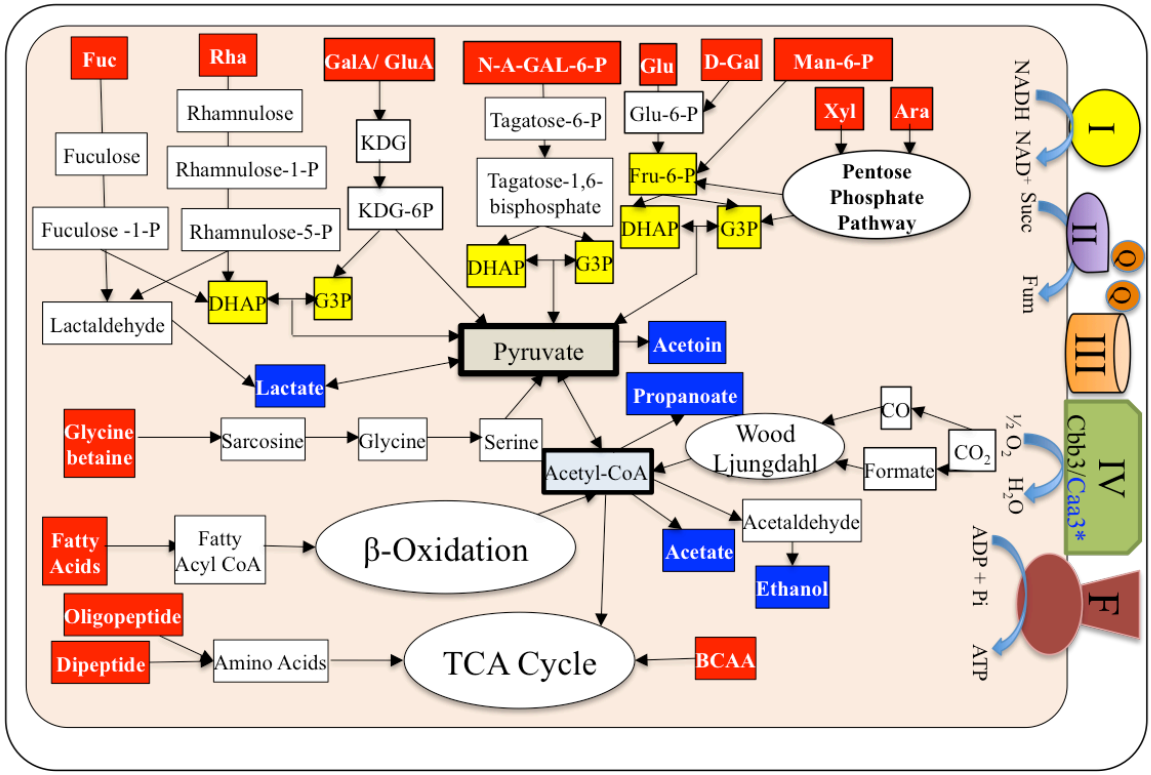

Fuc
Rha
GalA/ GluA
N-A-GAL-6-P
Glu
D-Gal
Man-6-P
Rhamnulose
KDG
Tagatose-6-P
Glu-6-P
Xyl
Ara
Fuculose
Rhamnulose-1-P
Tagatose-1,6-bisphosphate
Fru-6-P
Pentose Phosphate Pathway
KDG-6P
Fuculose -1-P
Rhamnulose-5-P
DHAP
G3P
DHAP
G3P
DHAP
G3P
Lactaldehyde
Pyruvate
Acetoin
Lactate
Propanoate
Glycine betaine
Sarcosine
Glycine
Serine
Acetyl-CoA
Wood Ljungdahl
CO
CO2
Formate
Fatty Acids
Fatty Acyl CoA
β-Oxidation
Acetaldehyde
Acetate
Ethanol
Oligopeptide
Dipeptide
Amino Acids
TCA Cycle
BCAA
NADH NAD+ Succ
Fum
Q
Q
½ O2 H2O
ADP + Pi
ATP
I
II
III
IV
Cbb3/Caa3*
F

Figure 3-7. Metabolic reconstruction of the “Latescibacteria” pangenome. (A) Extracellular polymer degradation, and monomer import. Different cell compartments are labeled (extracellular, outer membrane (OM), periplasm, and cytoplasmic membrane (IM)). The shading of the polymer rectangles correspond to the different “Latescibacteria” orders whose pangenomes suggest the polymer degradation capability, with color coding as follows; red: PBSIII_9; blue: GN03; green: sediment_1; and yellow: MSB-4E2. Transporters are color coded as follows; Secondary transporters, grey; Phosphotransferase system (PTS), blue; ABC transporter, red. (B) Central metabolic pathways. Substrates are shaded in red, fermentation end products are shaded in blue, and EMP intermediates are shaded in yellow. Components of the ETS are shown in the membrane. For complex IV, Cbb3 type cytochrome C oxidase was identified in both freshwater and marine “Latescibacteria” in contigs affiliated with order PBSIII_9, while Caa3 type cyctochrome C oxidase was identified solely in marine “Latescibacteria” contigs affiliated with orders GN03 and PBSIII_9. Abbreviations are as follows: FA, fatty acids; PolyNAG, Poly-N-acetylgalactosamine; AGal, arabinogalactan; Carr, carrageenan; Porph, porphyran; Fuc, fucose; Rha, rhamnose; GalA, galcturonate; GluA, glucuronate; ManA, mannuronate; NAGalN, N-acetyl galactosamine; NAGluN, N-acetyl glucosamine; Gal, galactose; Glu, glucose; Man, mannose; Xyl, xylose; Ara, arabinose; BCAA, branched chain amino acids; KDG, 2-dehydro-3-deoxy-D-gluconate; DHAP, dihydroxyacetone-phosphate; G3P, glyceraldehyde-3-phosphate; Q, quinone.

Discussion

Here, we present a detailed analysis of the global distribution patterns and putative metabolic capabilities of the candidate phylum "Latescibacteria" by analyzing publicly available near full-length and high throughput 16S rRNA sequences in public datasets, as well as recovering and analyzing "Latescibacteria" genomic fragments from a wide range of habitats using a fragment recruitment approach. Our results highlight: 1. The global distribution of the "Latescibacteria" in a wide range of ecosystems and the preference of specific "Latescibacteria" orders to specific habitats, e.g. Sediment_1 in terrestrial, PBSIII_9 in groundwater and temperate freshwater, and GN03 in pelagic marine, saline-hypersaline, and wastewater habitats (Fig. 1, Table 1), 2. The prevalence of polymer (polysaccharides, proteins, glycoproteins, lipids and fatty acids) degradation abilities within all "Latescibacteria" orders, with higher densities and specific abilities observed in specific orders, e.g. higher densities of GHs and PLs in order PBSIII_9, alginate degradation capability confined to order PBSIII_9 in soil, and agar, porphyran, and carrageenan degradation capability confined to aquatic (freshwater and marine) habitats (Figs. 3, 4, 6, S3, S4, and Tables 4, S6), 3. The occurrence of all genes/domains necessary for the production of cellulosome within three "Latescibacteria" orders (GN03, PBSIII_9, and MSB-4E2) in datasets recovered from four different anaerobic locations (Fig. 5, S5, Table 3, S7), and 4. The identification of the components of an aerobic respiratory chain, as well as occurrence of multiple O_2 dependent metabolic reactions in "Latescibacteria" orders GN03 and PBSIII_9 recovered from oxygenated habitats, suggesting the occurrence of both respiratory and fermentative modes of metabolisms within the "Latescibacteria" (Fig. 6, Table S9).

In spite of their ubiquitous distribution", members of the "Latescibacteria are invariably present in low abundance in all examined habitats (Table 1). Direct recovery of complete/near complete genomes of low abundance organisms using genome-resolved metagenomics approaches could be challenging, due to the relatively low representation level of target lineage(s) in assembled datasets. As well, individual reads and read coverage statistics, a necessary input in many genomic assembly programs, are often unavailable in public databases. Finally the processing power required for assembling genomes from thousands of datasets renders the utilization of such an approach in database mining studies a computationally daunting task.

Therefore, to recover genomic fragments belonging to the relatively low abundance (Table 1) phylum "Latescibacteria" from a large number of metagenomic datasets, we opted to utilize a fragment recruitment approach that maps pre-assembled metagenomes to the available "Latescibacteria" single amplified genomes (SAGs) (1, 67). This fragment recruitment approach hence allows for the screening of thousands of datasets for a target lineage within a reasonable time and computational capacity, and has been previously conducted to identify genomic fragments belonging to a wide range of yet-uncultured bacterial and archaeal phyla in available metagenomics datasets (1). We opted for conservative recruitment (first hit with an e-value of $1e^{-10}$ and a preset minimum 25% or 2.5 Kbp overlap), and strict quality control criteria (housekeeping gene extraction and phylogenetic analysis, as well as alignment comparisons for a selected number of large contigs, Table S5) to guard against false assignment of contigs to the "Latescibacteria". Indeed, our benchmarking efforts on mock microbial communities yielded only very few (<1%) non-target sequences (Table S2), demonstrating that high

specificity of the pipeline. Further, benchmarking efforts (Table S2) demonstrated that the sensitivity of the process is dependent on the size of the fragment as well as the level of relatedness between target and reference sequences. Therefore, it is likely that such strict recruitment criteria could have hindered the identification process, and it is entirely plausible that some "Latescibacteria" fragments within the analyzed datasets have escaped detection. As such, the presented data should not be regarded as exhaustive and comprehensive metabolic capacities of the "Latescibacteria", and additional genomic and metabolic insights could yet be gained in future studies.

Prior analysis of "Latescibacteria" single cell genomes belonging to family PBSIII_9_1 in order PBSIII_9 demonstrated their ability to metabolize pectin, ulvan, fucan, alginate and hydroxyproline rich polymers (16). The current study confirmed such abilities, and further expanded the "Latescibacteria" degradation capacities to include additional polysaccharides prevalent in cell walls of plants (cellulose, mannan, xylan, and xylofucan), bacteria (peptidoglycan), fungi/crustaceans (chitin), and various eukaryotic algae (porphyran, agar, carrageenan, and poly N-acetyl-galactosamine) (Fig. 3, 4, 6, S3, S4, Table 4, S8). Collectively, such broad polymer degradation portfolio suggests a global saprophytic strategy, where members of the "Latescibacteria" are involved in metabolizing dead cells and cell lysates of prokaryotes and eukaryotic lineages for carbon acquisition and energy conservation. This hypothesis is bolstered by the observation that the "Latescibacteria" are capable of degrading a wide range of intracellular constituents (e.g. fatty acids, proteins) in addition to polysaccharides (Fig. 6, Table 4, S9).

Perhaps the most unexpected finding in this study is the discovery of genes and domains suggestive of cellulosomal production in three "Latescibacteria" orders

(PBSIII_9, GN03, and MSB-4E2) recovered from four different anoxic habitats. This discovery renders the "Latescibacteria" the first Gram-negative bacterial lineage potentially capable of producing a cellulosome, and extends the occurrence of such process beyond the handful of Gram-positive genera *Rumnicoccus*, *Acetivibrio, Clostridium, and Pseudobacteroides cellulosolvens* (68-78) within the class Clostridia, and the anaerobic gut fungi (Neocallimastigomycota) (54). Dockerin type I modules were identified within twenty-seven different genes, all of which encode extracellular CAZymes, phytase, or lipase enzymes (Fig. 5, Table 3). It should be mentioned that in rare cases (6.18 %, pfam database (79) October 2016), dockerin I domains (pfam PF00404) were observed in genes derived from non-cellulosomal- producing organisms. However, in all these instances, the dockerin I domain was invariably attached to a non-CAZyme/non-polymer degrading genes (80), contrary to what was observed in the current study, as well as in all cellulosome-producing organisms (Table 3, S7).

Dockerin I domains in cellulosomes bind to cohesin domains that are present as part of a large scaffoldin protein (81). We identified 46 cohesin-containing scaffoldin protein-coding genes, all of which co-occurred in the same datasets with dockerin 1 domains, and were often co-located on the same contig (Table S7). In addition to cohesin I modules, scaffoldin proteins in cellulosome-producing organisms often contain additional domains such as CBM37, X domain of yet-unknown function, and dockerin II model for docking the entire cellulosomal structure onto the cell wall. "Latescibacteria" scaffoldin genes identified contained, in addition to cohesin I modules, multiple CBM37 domains (Table S7). However, we failed to identify X modules, or dockerin II modules in these genes. While puzzling, this could be a reflection of utilization of an alternative,

hitherto unknown strategy for cellulosomal attachment to cell wall, given the predicted Gram negative diderm cell wall of the "Latescibacteria" (16), as opposed to the relatively simpler Gram positive monoderm cell wall in other cellulosome-producing bacteria.

In conclusion, our work describes the global ecological distribution patterns of the "Latescibacteria", and provides a pangenomic view of this yet-uncultured phylum. We suggest that similar phylocentric pangenomic surveys targeting uncultured lineages could be extremely beneficial for understanding the global ecological distribution and pan-metabolic abilities of such lineages; and could provide broader and complimentary insights to those gained from single cell genomic and/or metagenomics-enabled genome recovery efforts focusing on a single sampling site.

Acknowledgments. We thank Radwa A. Hanafy and Fares Z. Najar for valuable technical assistance. This work was supported by the National Science Foundation Microbial Observatories Program (Grant EF0801858).

References

1. **Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng JF, Darling A, Malfatti S, Swan BK, Gies EA, Dodsworth JA, Hedlund BP, Tsiamis G, Sievert SM, Liu WT, Eisen JA, Hallam SJ, Kyrpides NC, Stepanauskas R, Rubin EM, Hugenholtz P, Woyke T.** 2013. Insights into the phylogeny and coding potential of microbial dark matter. Nature **499:**431-437.
2. **Seitz KW, Lazar CS, Hinrichs KU, Teske AP, Baker BJ.** 2016. Genomic reconstruction of a novel, deeply branched sediment archaeal phylum with pathways for acetogenesis and sulfur reduction. ISME J doi:10.1038/ismej.2015.233.
3. **Wrighton KC, Thomas BC, Sharon I, Miller CS, Castelle CJ, VerBerkmoes NC, Wilkins MJ, Hettich RL, Lipton MS, Williams KH, Long PE, Banfield JF.** 2012. Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. Science **337:**1661-1665.
4. **Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF.** 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature **428:**37-43.
5. **Mardis ER.** 2011. A decade's perspective on DNA sequencing technology. Nature **470:**198-203.
6. **Mardis ER.** 2013. Next-generation sequencing platforms. Annu Rev Anal Chem (Palo Alto Calif) **6:**287-303.

7. **Muir P, Li S, Lou S, Wang D, Spakowicz DJ, Salichos L, Zhang J, Weinstock GM, Isaacs F, Rozowsky J, Gerstein M.** 2016. The real cost of sequencing: scaling computation to keep pace with data generation. Genome Biol **17:**53.

8. **Buske FA, French HJ, Smith MA, Clark SJ, Bauer DC.** 2014. NGSANE: a lightweight production informatics framework for high-throughput data analysis. Bioinformatics **30:**1471-1472.

9. **Kang DD, Froula J, Egan R, Wang Z.** 2015. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ **3:**e1165.

10. **Wu YW, Tang YH, Tringe SG, Simmons BA, Singer SW.** 2014. MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. Microbiome **2:**26.

11. **Imelfort M, Parks D, Woodcroft BJ, Dennis P, Hugenholtz P, Tyson GW.** 2014. GroopM: an automated tool for the recovery of population genomes from related metagenomes. PeerJ **2:**e603.

12. **Langmead B, Salzberg SL.** 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods **9:**357-359.

13. **Rinke C, Lee J, Nath N, Goudeau D, Thompson B, Poulton N, Dmitrieff E, Malmstrom R, Stepanauskas R, Woyke T.** 2014. Obtaining genomes from uncultivated environmental microorganisms using FACS-based single-cell genomics. Nat Protoc **9:**1038-1048.

14. **Brown CT, Hug LA, Thomas BC, Sharon I, Castelle CJ, Singh A, Wilkins MJ, Wrighton KC, Williams KH, Banfield JF.** 2015. Unusual biology across a group comprising more than 15% of domain Bacteria. Nature **523:**208-211.

15. **Youssef NH, Blainey PC, Quake SR, Elshahed MS.** 2011. Partial genome assembly for a candidate division OP11 single cell from an anoxic spring (Zodletone Spring, Oklahoma). Appl Environ Microbiol **77:**7804-7814.

16. **Youssef NH, Farag IF, Rinke C, Hallam SJ, Woyke T, Elshahed MS.** 2015. In Silico Analysis of the Metabolic Potential and Niche Specialization of Candidate Phylum "Latescibacteria" (WS3). PLoS One **10:**e0127499.

17. **Winsley TJ, Snape I, McKinlay J, Stark J, Dorst JMv, Ji M, Ferrari BC, Siciliano SD.** 2014. The ecological controls on the prevalence of candidate division TM7 in polar regions. Front Microbiol **5:**345.

18. **Farag IF, Davis JP, Youssef NH, Elshahed MS.** 2014. Global Patterns of Abundance, Diversity and Community Structure of the Aminicenantes (Candidate Phylum OP8). PLoS ONE **9:**e92139.

19. **Dojka MA, Hugenholtz P, Haack SK, Pace NR.** 1998. Microbial diversity in a hydrocarbon- and chlorinated-solvent-contaminated aquifer undergoing intrinsic bioremediation. Appl Environ Microbiol **64:**3869-3877.

20. **Schabereiter-Gurtner C, Saiz-Jimenez C, Pinar G, Lubitz W, Rolleke S.** 2004. Phylogenetic diversity of bacteria associated with Paleolithic paintings and surrounding rock walls in two Spanish caves (Llonin and La Garma). FEMS Microbiol Ecol **47:**235-247.

21. **Hernandez-Raquet G, Budzinski H, Caumette P, Dabert P, Le Menach K, Muyzer G, Duran R.** 2006. Molecular diversity studies of bacterial communities of oil polluted microbial mats from the Etang de Berre (France). FEMS Microbiol Ecol **58:**550-562.

22. **Reed AJ, Lutz RA, Vetriani C.** 2006. Vertical distribution and diversity of bacteria and archaea in sulfide and methane-rich cold seep sediments located at the base of the Florida Escarpment. Extremophiles **10:**199-211.

23. **Kormas KA, Meziti A, Dahlmann A, GJ DEL, Lykousis V.** 2008. Characterization of methanogenic and prokaryotic assemblages based on mcrA and 16S rRNA gene diversity in sediments of the Kazan mud volcano (Mediterranean Sea). Geobiology **6:**450-460.

24. **Ikenaga M, Guevara R, Dean AL, Pisani C, Boyer JN.** 2010. Changes in community structure of sediment bacteria along the Florida coastal everglades marsh-mangrove-seagrass salinity gradient. Microb Ecol **59:**284-295.

25. **Briggs BR, Pohlman JW, Torres M, Riedel M, Brodie EL, Colwell FS.** 2011. Macroscopic biofilms in fracture-dominated sediment that anaerobically oxidize methane. Appl Environ Microbiol **77:**6780-6787.

26. **Fuchsman CA, Kirkpatrick JB, Brazelton WJ, Murray JW, Staley JT.** 2011. Metabolic strategies of free-living and aggregate-associated bacterial communities inferred from biologic and chemical profiles in the Black Sea suboxic zone. FEMS Microbiol Ecol **78:**586-603.

27. **Yakimov MM, La Cono V, Slepak VZ, La Spada G, Arcadi E, Messina E, Borghini M, Monticelli LS, Rojo D, Barbas C, Golyshina OV, Ferrer M,**

Golyshin PN, Giuliano L. 2013. Microbial life in the Lake Medee, the largest deep-sea salt-saturated formation. Sci Rep **3:**3554.

28. **Carbonetto B, Rascovan N, Alvarez R, Mentaberry A, Vazquez MP.** 2014. Structure, composition and metagenomic profile of soil microbiomes associated to agricultural land use and tillage systems in Argentine Pampas. PLoS One **9:**e99949.

29. **Pereira AD, Leal CD, Dias MF, Etchebehere C, Chernicharo CA, de Araujo JC.** 2014. Effect of phenol on the nitrogen removal performance and microbial community structure and composition of an anammox reactor. Bioresour Technol **166:**103-111.

30. **McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, Andersen GL, Knight R, Hugenholtz P.** 2012. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. The ISME J **6:**610-618.

31. **Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG.** 2007. Clustal W and Clustal X version 2.0. Bioinformatics **23:**2947-2948.

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

33. **Paye JM, Guseva A, Hammer SK, Gjersing E, Davis MF, Davison BH, Olstad J, Donohoe BS, Nguyen TY, Wyman CE, Pattathil S, Hahn MG, Lynd**

LR. 2016. Biological lignocellulose solubilization: comparative evaluation of biocatalysts and enhancement via cotreatment. Biotechnol Biofuels **9:**8.

34. **Tateno Y, Imanishi T, Miyazaki S, Fukami-Kobayashi K, Saitou N, Sugawara H, Gojobori T.** 2002. DNA Data Bank of Japan (DDBJ) for genome scale research in life science. Nucleic Acids Res **30:**27-30.

35. **Wilke A, Bischof J, Gerlach W, Glass E, Harrison T, Keegan KP, Paczian T, Trimble WL, Bagchi S, Grama A, Chaterji S, Meyer F.** 2015. The MG-RAST metagenomics database and portal in 2015. Nucleic Acids Research.

36. **Huse SM, Mark Welch DB, Voorhis A, Shipunova A, Morrison HG, Eren AM, Sogin ML.** 2014. VAMPS: a website for visualization and analysis of microbial population structures. BMC Bioinformatics **15:**41.

37. **Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ.** 2009. Introducing mothur: open source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol **75**.

38. **Markowitz VM, Chen IM, Chu K, Szeto E, Palaniappan K, Pillay M, Ratner A, Huang J, Pagani I, Tringe S, Huntemann M, Billis K, Varghese N, Tennessen K, Mavromatis K, Pati A, Ivanova NN, Kyrpides NC.** 2014. IMG/M 4 version of the integrated metagenome comparative analysis system. Nucleic Acids Res **42:**D568-573.

39. **Darling AE, Jospin G, Lowe E, Matsen FAt, Bik HM, Eisen JA.** 2014. PhyloSift: phylogenetic analysis of genomes and metagenomes. PeerJ **2:**e243.

40. **Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ.** 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics **11:**119.

41. **Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M.** 2016. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res **44:**D457-462.

42. **Caspi R, Billington R, Ferrer L, Foerster H, Fulcher CA, Keseler IM, Kothari A, Krummenacker M, Latendresse M, Mueller LA, Ong Q, Paley S, Subhraveti P, Weaver DS, Karp PD.** 2016. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res **44:**D471-480.

43. **Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B.** 2014. The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res **42:**D490-495.

44. **Petersen TN, Brunak S, von Heijne G, Nielsen H.** 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods **8:**785-786.

45. **Rawlings ND, Barrett AJ, Finn R.** 2016. Twenty years of the MEROPS database of proteolytic enzymes, their substrates and inhibitors. Nucleic Acids Res **44:**D343-350.

46. **Saier MH, Jr., Reddy VS, Tsu BV, Ahmed MS, Li C, Moreno-Hagelsieb G.** 2016. The Transporter Classification Database (TCDB): recent advances. Nucleic Acids Res **44:**D372-379.

47. **Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Soding J, Thompson JD, Higgins DG.** 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol **7:**539.

48. **Roy A, Kucukural A, Zhang Y.** 2010. I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc **5:**725-738.

49. **Mann AJ, Hahnke RL, Huang S, Werner J, Xing P, Barbeyron T, Huettel B, Stuber K, Reinhardt R, Harder J, Glockner FO, Amann RI, Teeling H.** 2013. The genome of the alga-associated marine flavobacterium Formosa agariphila KMM 3901T reveals a broad potential for degradation of algal polysaccharides. Appl Environ Microbiol **79:**6813-6822.

50. **Liu QP, Sulzenbacher G, Yuan H, Bennett EP, Pietz G, Saunders K, Spence J, Nudelman E, Levery SB, White T, Neveu JM, Lane WS, Bourne Y, Olsson ML, Henrissat B, Clausen H.** 2007. Bacterial glycosidases for the production of universal red blood cells. Nat Biotech **25:**454-464.

51. **Pitt MJ.** 1981. Rachitic and osteomalacic syndromes. Radiol Clin North Am **19:**581-599.

52. **Bakunina IY, Nedashkovskaya OI, Kim SB, Zvyagintseva TN, Mikhailov VV.** 2012. Distribution of α-N-acetylgalactosaminidases among marine bacteria of the phylum Bacteroidetes, epiphytes of marine algae of the Seas of Okhotsk and Japan. Microbiology **81:**373-378.

53. **Jobst MA, Milles LF, Schoeler C, Ott W, Fried DB, Bayer EA, Gaub HE, Nash MA.** 2015. Resolving dual binding conformations of cellulosome cohesin-dockerin complexes using single-molecule force spectroscopy. Elife **4**.

54. **Bayer EA, Belaich JP, Shoham Y, Lamed R.** 2004. The cellulosomes: multienzyme machines for degradation of plant cell wall polysaccharides. Annu Rev Microbiol **58:**521-554.

55. **Munir RI, Schellenberg J, Henrissat B, Verbeke TJ, Sparling R, Levin DB.** 2014. Comparative Analysis of Carbohydrate Active Enzymes in Clostridium termitidis CT1112 Reveals Complex Carbohydrate Degradation Ability. PLoS ONE **9:**e104260.

56. **Gies EA, Konwar KM, Beatty JT, Hallam SJ.** 2014. Illuminating microbial dark matter in meromictic Sakinaw Lake. Appl Environ Microbiol **80:**6807-6818.

57. **Zaikova E, Hawley A, Walsh DA, Hallam SJ.** 2009. Seawater sampling and collection. J Vis Exp doi:10.3791/1159.

58. **Inskeep WP, Jay ZJ, Tringe SG, Herrgard MJ, Rusch DB, Committee YNPMPS, Working Group M.** 2013. The YNP Metagenome Project: Environmental Parameters Responsible for Microbial Distribution in the Yellowstone Geothermal Ecosystem. Front Microbiol **4:**67.

59. **Walsh DA, Zaikova E, Howes CG, Song YC, Wright JJ, Tringe SG, Tortell PD, Hallam SJ.** 2009. Metagenome of a versatile chemolithoautotroph from expanding oceanic dead zones. Science **326:**578-582.

60. **Chen C, Cui Z, Xiao Y, Cui Q, Smith SP, Lamed R, Bayer EA, Feng Y.** 2014. Revisiting the NMR solution structure of the Cel48S type-I dockerin module from

Clostridium thermocellum reveals a cohesin-primed conformation. J Struct Biol **188:**188-193.

61. **Karpol A, Kantorovich L, Demishtein A, Barak Y, Morag E, Lamed R, Bayer EA.** 2009. Engineering a reversible, high-affinity system for efficient protein purification based on the cohesin-dockerin interaction. J Mol Recognit **22:**91-98.

62. **Odintsov SG, Sabala I, Marcyjaniak M, Bochtler M.** 2004. Latent LytM at 1.3A resolution. J Mol Biol **335:**775-785.

63. **Labadie J, Hebraud M.** 1997. Purification and characterization of a collagenolytic enzyme produced by Rathayibacter sp. strains isolated from cultures of Clavibacter michiganensis subsp. michiganensis. J Appl Microbiol **82:**141-148.

64. **Rudrappa T, Czymmek KJ, Paré PW, Bais HP.** 2008. Root-Secreted Malic Acid Recruits Beneficial Soil Bacteria. Plant Physiology **148:**1547-1556.

65. **Pitcher RS, Watmough NJ.** 2004. The bacterial cytochrome cbb3 oxidases. Biochim Biophys Acta **1655:**388-399.

66. **de Gier JW, Schepper M, Reijnders WN, van Dyck SJ, Slotboom DJ, Warne A, Saraste M, Krab K, Finel M, Stouthamer AH, van Spanning RJ, van der Oost J.** 1996. Structural and functional analysis of aa3-type and cbb3-type cytochrome c oxidases of Paracoccus denitrificans reveals significant differences in proton-pump design. Mol Microbiol **20:**1247-1260.

67. **Nobu MK, Dodsworth JA, Murugapiran SK, Rinke C, Gies EA, Webster G, Schwientek P, Kille P, Parkes RJ, Sass H, Jorgensen BB, Weightman AJ, Liu**

WT, Hallam SJ, Tsiamis G, Woyke T, Hedlund BP. 2016. Phylogeny and physiology of candidate phylum 'Atribacteria' (OP9/JS1) inferred from cultivation-independent genomics. Isme j **10:**273-286.

68. **Ding SY, Bayer EA, Steiner D, Shoham Y, Lamed R.** 1999. A novel cellulosomal scaffoldin from Acetivibrio cellulolyticus that contains a family 9 glycosyl hydrolase. J Bacteriol **181:**6720-6729.

69. **Pages S, Belaich A, Fierobe HP, Tardif C, Gaudin C, Belaich JP.** 1999. Sequence analysis of scaffolding protein CipC and ORFXp, a new cohesin-containing protein in Clostridium cellulolyticum: comparison of various cohesin domains and subcellular localization of ORFXp. J Bacteriol **181:**1801-1810.

70. **Ding SY, Bayer EA, Steiner D, Shoham Y, Lamed R.** 2000. A scaffoldin of the Bacteroides cellulosolvens cellulosome that contains 11 type II cohesins. J Bacteriol **182:**4915-4925.

71. **Ding SY, Rincon MT, Lamed R, Martin JC, McCrae SI, Aurilia V, Shoham Y, Bayer EA, Flint HJ.** 2001. Cellulosomal scaffoldin-like proteins from Ruminococcus flavefaciens. J Bacteriol **183:**1945-1953.

72. **Kakiuchi M, Isui A, Suzuki K, Fujino T, Fujino E, Kimura T, Karita S, Sakka K, Ohmiya K.** 1998. Cloning and DNA sequencing of the genes encoding Clostridium josui scaffolding protein CipA and cellulase CelD and identification of their gene products as major components of the cellulosome. J Bacteriol **180:**4303-4308.

73. **Kirby J, Martin JC, Daniel AS, Flint HJ.** 1997. Dockerin-like sequences in cellulases and xylanases from the rumen cellulolytic bacterium Ruminococcus flavefaciens. FEMS Microbiol Lett **149:**213-219.

74. **Lamed R, Naimark J, Morgenstern E, Bayer EA.** 1987. Specialized cell surface structures in cellulolytic bacteria. J Bacteriol **169:**3792-3800.

75. **Lamed R, Setter E, Bayer EA.** 1983. Characterization of a cellulose-binding, cellulase-containing complex in Clostridium thermocellum. J Bacteriol **156:**828-836.

76. **Nolling J, Breton G, Omelchenko MV, Makarova KS, Zeng Q, Gibson R, Lee HM, Dubois J, Qiu D, Hitti J, Wolf YI, Tatusov RL, Sabathe F, Doucette-Stamm L, Soucaille P, Daly MJ, Bennett GN, Koonin EV, Smith DR.** 2001. Genome sequence and comparative analysis of the solvent-producing bacterium Clostridium acetobutylicum. J Bacteriol **183:**4823-4838.

77. **Pohlschroder M, Canale-Parola E, Leschine SB.** 1995. Ultrastructural diversity of the cellulase complexes of Clostridium papyrosolvens C7. J Bacteriol **177:**6625-6629.

78. **Shoseyov O, Takagi M, Goldstein MA, Doi RH.** 1992. Primary sequence analysis of Clostridium cellulovorans cellulose binding protein A. Proc Natl Acad Sci U S A **89:**3483-3487.

79. **Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, Salazar GA, Tate J, Bateman A.** 2016. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res **44:**D279-285.

80. **Peer A, Smith SP, Bayer EA, Lamed R, Borovok I.** 2009. Noncellulosomal cohesin- and dockerin-like modules in the three domains of life. FEMS Microbiol Lett **291:**1-16.

81. **Smith SP, Bayer EA.** 2013. Insights into cellulosome assembly and dynamics: from dissection to reconstruction of the supramolecular enzyme complex. Curr Opin Struct, Biol **23:**686-694.

Conclusion

In this dissertation, I employed a wide range of sequence analysis approaches to study the distribution patterns, metabolic capabilities and ecological roles of two candidate phyla: The "Aminicinantes" and the "Latescibacteria". Overall, our analyses revealed the ubiquitous nature and delineated the distinct distribution patterns of members of these phyla along different habitats. The occurrence of a wide range of "Aminicinantes" and "Latescibacteria" inter-phylum in examined datasets underscore the importance of global diversity survey, rather than single habitat based studies, to accurately describe the ecology and metabolism of a target microbial lineage.

While multiple prior studies have utilized publicly available 16S rRNA gene datasets to explore the phylogenetic diversities of a specific lineage, these surveys have been mainly conducted for phyla with cultured representatives (e.g. Proteobacteria, Spain ISME 2007). Our work exploring the phylogenetic diversity of members of the "Aminicinantes" (chapter 2) using publicly available 16A rRNA gene data represents one of the few available studies conducting such approach on a yet-uncultured lineage. Therefore, we emphasize the value of phylogenetic surveys for understanding the global ecological distribution and panmetabolic abilities of yet-uncultured microbial lineages.

Similarly, the work on utilizing fragment-recruitment to explore the global functional diversity of the "Latescibacteria" pangenome (Chapter 2) represents one of the very few studies utilizing fragment recruitment for targeting a yet-uncultured microbial phylum. This approach allowed for access to "Latescibacteria" genomic fragments from a

wide range of habitats, thus providing a global view of the metabolic abilities of this phylum. The results suggest a highly diverse phylum, and adds to our knowledge regarding the metabolic capability of this phylum, specifically, the wide range of polysaccharides that could be metabolized by the "Latescibacteria", the occurrence of a cellulosome in this gram negative lineage, and the ability to conduct respiratory in addition to fermentative metabolism.

Finally, the work on analyzing single cell genomes (Chapter 3) has focused on a few near complete genome assemblies from a single habitat. This myopic view does not provide the global perspective gained from metagenomics fragment recruitment. However, when analysis of near complete genomes are coupled with geochemical data and ecological measurments, valuable insights could be gained towards understanding the role of a specific lineage in a specific habitat. In this case, coupling “Latescibacteria” genomic analysis with liminlogical data from Sakinaw lake have allowed us to propose a role for the “Latescibacteria” in the degradation of algal cell wall detritus in the lake’s hypolimnion.

In conclusion, this work has greatly enhanced our understanding of the ecology, physiology, and metabolism of two yet-uncultured ubiquitous bacterial phyla. The knowledge gained from this effort could be used to design sequence-guided strategy for the enrichment and isolation of these enigmatic and elusive bacterial lineages.

Reference

1. Spain AM, Krumholz LR, Elshahed MS. Abundance, composition, diversity and novelty of soil Proteobacteria. ISME J. 2009;3(8):992-1000. doi: 10.1038/ismej.2009.43. PubMed PMID: 19404326.

VITA

Ibrahim F. Farag
Candidate for the Degree of
Doctor of Philosophy

Thesis: EXPLORING THE HABITAT DISTRIBUTION, METABOLIC DIVERSITIES AND POTENTIAL ECOLOGICAL ROLES OF CANDIDATE PHYLA "AMINICENANTES" (OP8) AND "LATESCIBACTERIA"(WS3)

Major Field: Microbial Ecology

Education:

Completed the requirements for the Doctor of Philosophy in Microbiology at Oklahoma State University, Stillwater, Oklahoma in December 2017.

Completed the requirements for the Master of Science in Biotechnology at American University in Cairo, Cairo, Egypt in 2012.

Completed the requirements for the Bachelor of Science in Microbiology at Ain Shams University, Cairo, Egypt in 2006.

Experience:

Teaching Assistant and Research Assistant – Department of Microbiology and Molecular Genetics, Oklahoma State University, Stillwater, Oklahoma, August 20012 through May 2017.