



## Identification of Molecular Markers That Are Specific to the Class *Thermoleophilia*

### Danyu Hu<sup>1,2†</sup>, Yang Zang<sup>1,2†</sup>, Yingjin Mao<sup>1,2</sup> and Beile Gao<sup>1\*</sup>

<sup>1</sup> CAS Key Laboratory of Tropical Marine Bio Resources and Ecology, Guangdong Key Laboratory of Marine Materia Medica, South China Sea Institute of Oceanology, Chinese Academy of Sciences, Guangzhou, China, <sup>2</sup> University of Chinese Academy of Sciences, Beijing, China

The class Thermoleophilia is one of the deep-rooting lineages within the Actinobacteria phylum and metagenomic investigation of microbial diversity suggested that species associated with the class Thermoleophilia are abundant in hot spring and soil samples. However, very few species of this class have been cultivated and characterized. Our understanding of the phylogeny and taxonomy of Thermoleophilia is solely based on 16S rRNA sequence analysis of limited cultivable representatives, but no other phenotypic or genotypic characteristics are known that can clearly discriminate members of this class from the other taxonomic units within the kingdom bacteria. This study reports phylogenomic analysis for 12 sequenced members of this class and clearly resolves the interrelationship of not yet cultivated species with reconstructed genomes and known type species. Comparative genome analysis discovered 12 CSIs in different proteins and 32 CSPs that are specific to all species of this class. In addition, a large number of CSIs or CSPs were identified to be unique to certain lineages within this class. This study represents the first and most comprehensive phylogenetic analysis of the class Thermoleophilia, and the identified CSIs and CSPs provide valuable molecular markers for the identification and delineation of species belonging to this class or its subordinate taxa.

Keywords: Thermoleophilia, phylogeny, molecular signatures, conserved signature indels, conserved signature proteins

## INTRODUCTION

The class *Thermoleophilia* is one of the deep-rooting lineages within the *Actinobacteria* phylum and it has only recently been recognized as independent from the class *Rubrobacteria* (Zhi et al., 2009; Gao and Gupta, 2012b; Ludwig et al., 2012; Suzuki and Whitman, 2012). This class encompasses two recognized orders *Thermoleophilales* and *Solirubrobacteria* (and Whitman, 2015). A deep branching order *Gaiellales* within the phylum *Actinobacteria* (Albuquerque et al., 2011) has been proposed as an order of this class based on phylogenetic position, signature nucleotides of 16S rRNA, and physicochemical characteristics (Foesel et al., 2016). However, only one type strain *Gaiella occulta* F2-233 from this order was included in the analyses and its position in the phylogenetic tree is between the boundary of other *Thermoleophilia* orders and *Rubrobacteria*.

### **OPEN ACCESS**

### Edited by:

Haiwei Luo, The Chinese University of Hong Kong, China

#### Reviewed by:

Juan Antonio Ugalde, uBiome, Chile Bärbel Ulrike Fösel, Helmholtz Center Munich, Germany

\*Correspondence:

Beile Gao gaob@scsio.ac.cn

<sup>†</sup>These authors have contributed equally to this work

### Specialty section:

This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Microbiology

Received: 25 January 2019 Accepted: 09 May 2019 Published: 24 May 2019

#### Citation:

Hu D, Zang Y, Mao Y and Gao B (2019) Identification of Molecular Markers That Are Specific to the Class Thermoleophilia. Front. Microbiol. 10:1185. doi: 10.3389/fmicb.2019.01185

1

The order Thermoleophilales only contains one family Thermoleophilaceae with a single genus Thermoleophilum. Species of this genus are small regular rods, moderately thermophilic, and obligately aerobic (Suzuki and Whitman, 2012). Their distinct feature is growth restriction to substrate n-alkanes (Zarilla and Perry, 1986), thus these species are named as heat- and oil-loving microbes, "Thermoleophilum." While Thermoleophilum species are generally isolated from hot springs, members of the second order Solirubrobacterales are mainly detected in soil samples, and they exhibit more species diversity and different phenotypic characteristics. According to the most updated description of the taxonomic framework of the Actinobacteria phylum (Salam et al., 2019), the order Solirubrobacterales is composed of four families including Solirubacteraceae, Conexibacteraceae, Parviterribacteraceae and Patulibacteraceae. Currently described species of this order are mostly mesophilic with some psychrotolerant (Suzuki and Whitman, 2012). For example, metagenomic surveys of microbial diversity of soil samples from Antarctica revealed the presence of Thermoleophilia organisms, which can reach 15% abundance in some samples (Ji et al., 2016; Pulschen et al., 2017). Moreover, their preferred carbon sources are more diverse, including complex proteinaceous substrates, many sugars and a few other compounds (Foesel et al., 2016).

Several microbial diversity investigations suggest that Thermoleophilia species are abundant and diverse in nature (Joseph et al., 2003; Janssen, 2006), and they play an important role in geochemical recycling (Almeida et al., 2013; Ji et al., 2017; Li et al., 2018). However, similar to other deep-rooting classes with the phylum Actinobacteria, such as Acidimicrobiia, Rubrobacteria, Nitriliruptoria, etc., the cultivated isolates of Thermoleophilia are very limited (Ludwig et al., 2012; Suzuki and Whitman, 2015). Therefore, phenotypic characteristic descriptions of higher taxonomic ranks (e.g., class, order, family, and genus) within these classes are either lacking or speculative, which may not represent other yet uncultivated members belonging to these groups. In addition, our understanding of the phylogeny or taxonomy of the class Thermoleophilia is solely based on 16S rRNA sequence analysis, including their branching patterns in the phylogenetic trees or taxonspecific 16S rRNA signature nucleotides (Foesel et al., 2016; Salam et al., 2019). Except these two standards, no other molecular, biochemical or physiological characteristics are known that can clearly distinguish Thermoleophilia species from other Actinobacteria. Consequently, the bioprospecting or utilization of this group of bacteria is limited by our lack of knowledge of them. In the recent years, efforts have been made such as the "Genomic Encyclopedia of Bacteria and Archaea" (GEBA) project to sequence a diverse collection of the underrepresented phylogenetic lineages (Mukherjee et al., 2017), or to reconstruct genomes from metagenomic data for not yet cultivated species (Parks et al., 2017; Cabello-Yeves et al., 2018; Woodcroft et al., 2018). At the time of January 2018, there are 6 complete genomes and 10 genome assemblies for the class Thermoleophilia, providing great resource to explore phenotypic and genomic features of these microbes.

Two kinds of molecular markers have been described to define or delineate different higher taxa (e.g., genus level and above) for different prokaryotic phyla (Gupta and Gao, 2010; Gao and Gupta, 2012a). One kind of these molecular markers are conserved signature indels (CSIs) that are uniquely found in the genes/proteins homologs of a certain group of organisms, but absent in species outside of this group. The other kind of molecular markers are conserved signature proteins (CSPs) that are specifically present in a monophyletic prokaryotic group. These two molecular markers represent highly reliable characteristics of specific groups of organisms, and they provide novel methods for the identification or delineation of prokaryotic taxonomic units in clear molecular terms (Gao and Gupta, 2012b; Ho et al., 2016; Zhang et al., 2016; Alnajar and Gupta, 2017). We recently identified these molecular markers for Acidimicrobiia, another deep-branch class within the phylum Actinobacteria, which proved very useful for defining the whole class or different lineages within it and also provide interesting targets for functional studies of these microbes (Hu et al., 2018).

Here, we constructed a phylogenomic tree for 12 sequenced members of the class *Thermoleophilia* based on concatenation of 54 widely distributed conserved proteins. This tree clearly resolved the interrelationship of not yet cultivated species with reconstructed genomes and known type species. More importantly, by analyzing the sequenced *Thermoleophilia* species, we discovered 12 CSIs in different proteins and 32 CSPs that are specific to all members of this class. In addition, a large number of CSIs or CSPs were identified to be unique to certain lineages within this class. This study represents the first and most comprehensive phylogenetic analysis of the class *Thermoleophilia*, and the identified CSIs and CSPs provide valuable molecular markers for the identification and delineation of species belonging to this class or its subordinate taxa.

## MATERIALS AND METHODS

## **Phylogenetic Analysis**

A phylogenomic tree for 6 completely sequenced species and 6 metagenome-assembled genomes (MAGs) of the class Thermoleophilia (Supplementary Table 1) was constructed. These 6 MAGs were selected for phylogenomic analysis since most single copy orthologous proteins as proposed by Na et al. (2018) can be retrieved from these genomes while other MAGs lack many of these orthologs which will reduce the robustness of the phylogenetic analysis. The deep-branching order Gaiellales only has one species sequenced, Gaiella occulta F2-233, which was also added to the analyses. The final tree was based on the concatenation of 54 protein sequence alignments (Supplementary Table 2). In addition, sequences from 3 Rubrobacter species was used as outgroup to root the tree. Multiple sequence alignments for each protein were performed using the Clustal X 2.1 program (Larkin et al., 2007) and concatenated to produce a single alignment. Gblocks 0.91b program was applied to remove the poorly aligned regions (Talavera and Castresana, 2007) and the resulting alignment composed of 13,132 amino acids was used for phylogenetic analysis. A maximum-likelihood (ML) tree was constructed by MEGA 6.0 with the Whelan and Goldman substitution model based on 1000 bootstrap replicates (Tamura et al., 2013).

An ML tree based on 16S rRNA gene sequences was constructed for the representative strains of *Thermoleophilia* and deep-branching order *Gaiellales*, but no full length 16S rRNA sequences are available for the 6 MAGs. All the 16S rRNA sequences were obtained from Ribosomal Database Project (Cole et al., 2014) or NCBI GenBank, and accession number of each 16S rRNA sequences from 8 *Rubrobacter* species were used as outgroup to root the tree. The tree was constructed by MEGA 6.0 using the General Time Reversible model with 1000 bootstrap replicates.

### **Identification of CSIs**

CSIs were identified following the detailed method description by Gupta (Gupta, 2014). Briefly, BLASTP searches were performed on all protein sequences from the genome of Thermoleophilum album ATCC 35263 (Yakimov et al., 2003) against all sequences in the NCBI non-redundant protein sequences (nr) database, during the period from January to April, 2018. The general parameters used for BLASTP searches were default as shown in the NCBI website. Multiple sequence alignments were created for homologs of all available Thermoleophilia species and a few other bacteria by the Clustal X 2.1 program using default parameters. These sequence alignments were inspected for any conserved insertions or deletions that were restricted to Thermoleophilia species only and also flanked by at least 5-6 identical or conserved residues in the neighboring 30~40 amino acids on each side. The indels with non-conserved flanking regions were not considered. To verify the specificity of the identified indels, another round of BLASTP searches were performed with a short indel-containing fragment (60-100 amino acids long) against the GenBank database. To further confirm that the identified signatures are restricted to Thermoleophilia homologs, the top 500 BLAST hits with the highest similarity to the query sequence were inspected for the presence or absence of these CSIs. Final alignment files were generated by two softwares Sig\_Create and Sig\_Style1 (Gupta, 2014). Due to page limitation, indels-containing sequence alignment in all figures and Supplementary Figures only include those that are found in all Thermoleophilia sequences and few sequences from representative strains of other bacteria.

### **Identification of CSPs**

BLASTP searches were performed on individual proteins from the genome of *T. album* ATCC 35263 to identify proteins that are restricted to species of the class *Thermoleophilia* or the order *Thermoleophilales*. For CSPs that are specific to the order *Solirubrobacterales* or its subgroups at different taxonomic levels, the proteins from the genome of *Patulibacter americanus* DSM 16676 (Reddy and Garcia-Pichel, 2009) were selected as query sequences to do the BLASTP searches against all available sequences in the NCBI non-redundant protein sequences (nr) database. The parameters used for BLASTP searches were generally default except that "Max target sequences" were set to

<sup>1</sup>Gleans.net

be 500. The BLAST results were manually examined for putative *Thermoleophilia* -specific proteins based on Expected values (*E*-values) (Altschul et al., 1997). Only proteins with significant hits (*E*-values less than 0.01) merely from *Thermoleophilia* genomes while no other hits or hits from non-*Thermoleophilia* genomes generally with *E*-value higher than 1.0 were considered as CSPs in this work (Gao et al., 2006; Gao and Gupta, 2012b).

## **RESULTS AND DISCUSSION**

# Phylogenomic Analysis of the Class *Thermoleophilia*

Two recent comprehensive phylogenetic analyses of the Actinobacteria phylum have both applied phylogenomic methods to re-examine the evolutionary relationships or taxonomic framework of species within this phylum (Nouioui et al., 2018; Salam et al., 2019). However, both studies aimed at the entire phylogenetic structure of the phylum, only type species/strains were considered in their analyses. For the poorly represented Thermoleophilia, there are only 5~6 species included in both studies (Nouioui et al., 2018; Salam et al., 2019). Therefore, a comprehensive phylogenomic analysis of the Thermoleophilia class is still lacking in spite of the availability of reconstructed genomes for not yet cultivated species of this class. In addition, for these assembled genomes, their exact phylogenetic relationship with type species or taxonomic assignment need to be examined although their association with this class has been suggested (Cabello-Yeves et al., 2018; Woodcroft et al., 2018). Here, we constructed a phylogenetic tree for 6 completely sequenced species and 6 MAGs of this class, for which more single-copy ortholog sequences can be retrieved for a robust phylogenomic analysis (Supplementary Table 1). Finally, 54 orthologous protein sequences that mainly belong to the functional category "translation and transcription" were extracted for the above 12 genomes (Supplementary Table 2) and ML analysis was carried out for the concatenated protein dataset. To our knowledge, this is the most comprehensive phylogenetic analysis for the class Thermoleophilia (Figure 1A). In comparison with the current taxonomic framework, we also constructed a phylogenetic tree based on 16S rRNA sequences for this class (Figure 1B). However, surprisingly no complete 16S rRNA sequence were available for the incomplete genome assemblies selected for the above phylogenomic analyses (except that genome assembly of Solirubrobacter sp. URHD0082 contained a partial 643 bp fragment of 16S rRNA).

Overall, the combined protein tree showed a very similar branching pattern to the 16S rRNA tree. All species belonging to *Thermoleophilia* formed a robust cluster, separated from the class *Rubrobacteria*. The position of the deep branching order *Gaiellales* is between the boundary of other *Thermoleophilia* orders and the class *Rubrobacteria* in both trees. The single genome-sequenced species *G. occulta* F2-233 clusters with other *Thermoleophilia* orders with a very high bootstrap score 100% in the phylogenomic tree while showing a lower score 57% in the 16S rRNA tree, which is similar to the previous 16S rRNA analyses using the same *G. occulta* 



values (%) are shown at each node and different clusters that are consistently observed in both phylogenetic trees are marked.

strain (Foesel et al., 2016). Within *Thermoleophilia*, species of the two orders *Thermoleophilales* and *Solirubrobacterales* also formed distinctive clusters in the phylogenomic tree, supporting the current order assignment based on 16S rRNA analyses (Reddy and Garcia-Pichel, 2009; Suzuki and Whitman, 2012, 2015). Compared to the diverse soil-source *Solirubrobacterales*, only one cultivable species *T. album* ATCC 35263 from the order *Thermoleophilales* has been genome sequenced (Yakimov et al., 2003). Our phylogenomic tree revealed that MAG "bacterium HR41" clusters together with *T. album*. The genome of HR41 is reconstructed from metagenomic DNA from

high-temperature bioreactors, for which the initial samples were collected from an ammonia-rich geothermal groundwater stream in Japan (Kato et al., 2018). In view of their clustering pattern in the phylogenetic tree and common hot spring isolation environment, it is very likely that HR41 represents a species belonging to the family *Thermoleophiliaceae* or the order *Thermoleophilales*.

Notably, 3 MAGs- "Actinobacteria bacterium 13\_1\_20CM\_3\_68\_9" from grassland (Butterfield et al., 2016), "*Solirubrobacterales* bacterium 67-14" and "*Solirubrobacterales* bacterium 70-9" from bioreactors (Kantor et al., 2015) form a

|--|

Protein name	GI no. <sup>a</sup>	Figure number	Indel size	Indel position <sup>b</sup>	Specificity
Quinolinate synthase NadA	1225101978	Figure 2	4aa ins <sup>c</sup>	138–180	All Thermoleophilia
30S ribosomal protein S10	1093219170	Supplementary Figure S2	1aa ins	72-105	All Thermoleophilia
Glutamate-1-semialdehyde-2,1-aminomutase	1225102988	Supplementary Figure S3	2aa del	172-209	All Thermoleophilia
D-tyrosyl-tRNA(Tyr) deacylase	1225105696	Supplementary Figure S4	6aa del	100-135	All Thermoleophilia
Vitamin B12-dependent ribonucleotide reductase	1225104123	Supplementary Figure S5	1aa ins	746–793	All Thermoleophilia
DNA-directed RNA polymerase subunit beta	1225103324	Supplementary Figure S6	2aa ins	215-256	All Thermoleophilia
PspA/IM30 family protein	654611971	Supplementary Figure S7	3aa del	184–227	All Thermoleophilia
Glutamine-hydrolyzing GMP synthase	1225105599	Supplementary Figure S8	1aa ins	406-450	All Thermoleophilia
Elongation factor P	1225104642	Supplementary Figure S9	1aa ins	127-176	All Thermoleophilia
Replicative DNA helicase	1225103017	Supplementary Figure S10	2aa ins	15–55	All Thermoleophilia
Phenylalanine–tRNA ligase subunit alpha	654610443	Supplementary Figure S11	2–10aa ins	244–285	All Thermoleophilia
DNA polymerase III alpha subunit	1225105080	Supplementary Figure S12	1aa ins	84–128	All Thermoleophilia
Arginine-tRNA ligase	1225101858	Figure 3	7aa ins	314–367	Thermoleophiliaceae
LytR family transcriptional regulator	1225102507	Supplementary Figure S13	2aa ins	155-190	Thermoleophiliaceae
DNA gyrase subunit A	1225102941	Supplementary Figure S14	8aa ins	250-298	Thermoleophiliaceae
Chaperonin GroEL	1225103134	Supplementary Figure S15	3aa ins	459-497	Thermoleophiliaceae
Short chain dehydrogenase	1225103641	Supplementary Figure S16	2aa ins	222-264	Thermoleophiliaceae
Type II secretion system F family protein	1225104607	Supplementary Figure S17	1aa ins	299–342	Thermoleophiliaceae
Leucyl-tRNA synthetase	1093217654	Supplementary Figure S18	1aa ins	429-469	Thermoleophiliaceae
NADH-quinone oxidoreductase subunit B	551309834	Figure 4	1aa del	137–181	Conexibacteraceae,
					Patulibacteraceae
4-hydroxy-3-methylbut-2-enyl diphosphate reductase	739551922	Supplementary Figure S19	1aa ins	44–91	Conexibacteraceae, Solirubrobacteraceae, Patulibacteraceae
Pyruvate kinase	652636441	Supplementary Figure S20	5aa del	189–227	Conexibacteraceae, Solirubrobacteraceae, Patulibacteraceae
tRNA guanosine (34) transglycosylase Tgt	654594575	Supplementary Figure S21	1aa ins	312–357	Conexibacteraceae, Solirubrobacteraceae, Patulibacteraceae
Excinuclease ABC subunit UvrB	654612298	Supplementary Figure S22	1aa ins	215–263	Conexibacteraceae, Solirubrobacteraceae, Patulibacteraceae
Transcription antitermination factor NusB	494847549	Supplementary Figure S23	6aa ins	62–102	Conexibacteraceae, Solirubrobacteraceae, Patulibacteraceae
Thioredoxin-disulfide reductase	916615184	Figure 5	1aa ins	40-82	Conexibacteraceae
Trigger factor	917589205	Supplementary Figure S24	5aa ins	169-217	Conexibacteraceae
		Supplementary Figure S25	1aa ins	215-255	Conexibacteraceae
Glutamate-5-semialdehyde dehydrogenase	652642436	Supplementary Figure S26	5aa del	150-196	Conexibacteraceae
Glutamine amidotransferase	654598081	Figure 6	3aa ins	170-211	Solirubrobacteraceae
7,8-didemethyl-8-hydroxy-5-deazariboflavin synthase subunit CofH	654594367	Supplementary Figure S27	4aa del	152–192	Solirubrobacteraceae
methionine-tRNA ligase	654600348	Supplementary Figure S28	5aa ins	267-310	Solirubrobacteraceae
Asp-tRNA(Asn)/Glu-tRNA(Gln) amidotransferase subunit GatC	654597239	Supplementary Figure S29	1aa ins	20–65	Solirubrobacteraceae
CTP synthase	921290543	Supplementary Figure S30	2aa ins	264–308	Solirubrobacteraceae
DNA-directed RNA polymerase subunit beta'	494853285	Figure 7	8aa ins	376-420	Patulibacteraceae
SDR family NAD(P)-dependent	494848053	Supplementary Figure S31	2aa ins	149–198	Patulibacteraceae
oxidoreductase					
Dihydrolipoyl dehydrogenase	551307243	Supplementary Figure S32	1aa del	355–396	Patulibacteraceae
Methylmalonyl-CoA epimerase	551310266	Supplementary Figure S33	2aa ins	1–48	Patulibacteraceae
Acetyl-CoA carboxylase biotin carboxylase subunit	551309981	Supplementary Figure S34	2aa ins	224–268	Patulibacteraceae

(Continued)

### TABLE 1 | Continued

Protein name	GI no. <sup>a</sup>	Figure number	Indel size	Indel position <sup>b</sup>	Specificity
GTPase HflX	1225104795	Supplementary Figure S35	1aa ins	282–322	Patulibacteraceae
1-deoxy-D-xylulose-5-phosphate reductoisomerase	551310630	Supplementary Figure S36	6–8aa ins	146–188	Patulibacteraceae
Tryptophan–tRNA ligase	494851195	Supplementary Figure S37	4–12aa ins	152-191	Patulibacteraceae
Endopeptidase La	551309049	Supplementary Figure S38	1aa ins	228-266	Patulibacteraceae
7,8-didemethyl-8-hydroxy-5-deazariboflavin synthase subunit CofH	494847285	Supplementary Figure S39	4aa ins	481–522	Patulibacteraceae
NADH-quinone oxidoreductase subunit I	1113228917	Supplementary Figure S40	1aa ins	72–125	New cluster
Adenylosuccinate synthase	1113229450	Supplementary Figure S41	17–23aa ins	154-204	S.67-14 and S.70-9 <sup>d</sup>
GTPase Era	1113226493	Supplementary Figure S42	1–2aa ins	38–88	S.67-14 and S.70-9
Heme-copper oxidase subunit III	1113215223	Supplementary Figure S43	1–4aa ins	121–167	S.67-14 and S.70-9

<sup>a</sup> The GI number represents the GenBank identification number of the protein sequence from one Thermoleophilia species that contain the specific CSI. <sup>b</sup> The indel region indicates the region of the protein where the described CSI is present. <sup>c</sup>ins, insertion; del, deletion. <sup>d</sup>S.67-14 and S.70-9 are abbreviations for MAG "Solirubrobacterales bacterium 67-14" and "Solirubrobacterales bacterium 70-9."

			138	180
]	Thermoleophilum album	1225101978	VKAESDYCCTSSNAVQVVEHVR	NTYG EDVEILFGPDMWLGAYV
	bacterium HR41	1286951167		REH- D
	Solirubrobacterales bacterium 67-14	1113228771	IKQ-IW	AEHTF-
	Solirubrobacter sp. URHD0082	654594689	TVIY	A-HTF
	Solirubrobacterales bacterium URHD0059	1175306363	ILTVKIL	AEH- P-T
ALL	Solirubrobacterales bacterium 70-9	1113216885	TVSIW	REH- D-TF-
Thermoleophilia	Solirubrobacter soli	739638916	TVAI-K-IY	D-H- D-TF
-	Patulibacter medicamentivorans	494847271	LTVPLY	REHT-VFSF-
	Patulibacter minatonensis	916864665	LTVPRIY	RDHTFSF-
	Patulibacter americanus	655313328	LTVPRIY	RDH- P-TFSF-
	Conexibacter woesei	502700743	TVRIL	REH- P-TF
	Actinobacteria bacterium SCGC AG-212-D09	1028408900	F	REHAF
	Saccharopolyspora flava	1225403411	T-IDRSIP	R-VCQFH-
	Corynebacterium camporealensis	1175413662	LT-VDKSLP	KFH-
	Crossiella equi	1197850840	T-IDASIP	R-VCQFH-
011	Saccharomonospora marina	496444309	T-IDASIP	S-T-VCQFH-
Other	Nakamurella multipartita	506228592	T-IERSIP	A-TPVQFH-
Actinobacteria	Mycobacterium duvalii	1264392727	T-IDASIP	A-R-VQFH-
	Actinomyces radicidentis	1180915638	T-VERSIP	QCQFFH-
	Tsukamurella tyrosinosolvens	1327722296	LT-IEQSIP	TLQFH-
	Amycolatopsis thermoflava	654464395	T-IDASIP	A-R-VQFH-
	Microlunatus soli	1223485061	ES-P	A-Q-VLF-A-H-
	Alicyclobacillus acidocaldarius	504278386	QSIP	REFS
	Deltaproteobacteria bacterium TMED126	1200454574	SIA	K-KLF
	Armatimonadetes bacterium JP3_11	1232016804	QAIP	KEE
Other Bacteria _	Bacillus niameyensis	1011190324	DI-RSIP	KEFSF-
	Gracilimonas tropica	521070624	TLD-INSIP	K-KKFS-I
	Oceanobacillus damuensis	1056242062	DI-RSIP	KEFSF-
	Tuberibacillus calidus	654150363	LDI-RSIP	QYSF-
FIGURE 2   CSI specific to all 11	<i>nermoleophilia</i> species. Partial sequence alignment c	of the protein quinc	plinate synthase NadA showing	g a 4 amino acid insertion in a

FIGURE 2 CSI specific to all *Thermoleophila* species. Partial sequence alignment of the protein quinolinate synthase NadA showing a 4 amino acid insertion in a conserved region that is specific for members of the class *Thermoleophilia*. The dashes in this alignment as well as all other alignments indicate identity with the amino acid on the top line. The GenBank identification numbers of the protein sequences are shown, and the topmost numbers indicate the position of this sequence in the species shown on the top line.

distinct cluster in the phylogenomic tree, more closely related with other *Solirubacteraceae* families than *Thermoleophilales* (**Figure 1A**). In view of the branching pattern of these 3 MAGs, it is likely that they represent species of a novel family within the order *Solirubrobacterales*. Alternatively, the phylogenetic position of these MAGs is very similar to the two *Parviterribacter* species in the 16S rRNA tree, raising the possibility that they might be members of the *Parviterribacteraceae* family. However, neither the 16S rRNA of the 3 MAGs nor the genome information from the two *Parviterribacter* species is available at the moment, which preclude further analyses. Future new 16S rRNA or genome sequences from closely related species of either the 3 MAGs or the *Parviterribacteraceae* family are needed to define their relationship. In addition, assembled genomes for two monoisolates from the same study of grassland rhizosphere branched differently in our phylogenomic tree. "Solirubrobacterales bacterium URHD0059" clusters together with the type species Conexibacter woesei DSM 14684 (Pukall et al., 2010), indicating that it might be a new species belonging to the family Conexibacteraceae; while "Solirubrobacter sp. URHD0082" clusters with *S. soli* DSM 22325 with 100% bootstrap support, demonstrating its affiliation with the family Solirubacteraceae. The later association is also confirmed by the 16S rRNA tree based on partial sequence alignment (Supplementary Figure S1). Taken together, these phylogenomic analyses based on a concatenated protein dataset support current

taxonomic structure of the class *Thermoleophilia* based on 16S rRNA analyses. In addition, it revealed a new cluster composed of not yet cultivated species that might be a novel family within the order *Solirubrobacterales*.

## Molecular Markers Unique to the Class *Thermoleophilia*

The main purpose of this work is to identify genomic characteristics that are unique to the class *Thermoleophilia* or its subordinate taxa, which can be used to define their taxonomic ranks and also provide targets for functional studies.

**TABLE 2** Conserved Signature Proteins that are uniquely found in the *Thermoleophilia* class.

Protein product	Length	Specificity	Function
(A) CSPs uniquely pr	esent in All The	rmoleophilia species (29)	a
WP_093115104.1	242	All thermoleophilia	Unknown
WP_093115134.1	90	All thermoleophilia	Unknown
WP_093115673.1	127	All thermoleophilia	Unknown
WP_093115681.1	103	All thermoleophilia	Unknown
WP_093115745.1	166	All thermoleophilia	Unknown
WP_093115827.1	993	All thermoleophilia	Unknown
WP_093116216.1	151	All thermoleophilia	Unknown
WP_093116230.1	213	All thermoleophilia	Unknown
WP_093116634.1	159	All thermoleophilia	Unknown
WP_093116636.1 <sup>b</sup>	64	All thermoleophilia	Unknown
WP_093116642.1	114	All thermoleophilia	Unknown
WP_093116769.1	130	All thermoleophilia	Unknown
WP_093116819.1	167	All thermoleophilia	Unknown
WP_093116917.1	120	All thermoleophilia	Unknown
WP_093116997.1	185	All thermoleophilia	Unknown
WP_093117023.1	151	All thermoleophilia	Unknown
WP_093117047.1	572	All thermoleophilia	Unknown
WP_093117060.1	247	All thermoleophilia	Unknown
WP_093117260.1	72	All thermoleophilia	Unknown
WP_093117458.1 <sup>b</sup>	142	All thermoleophilia	Unknown
WP_093117523.1	269	All thermoleophilia	Unknown
WP_093118104.1	79	All thermoleophilia	Unknown
WP_093118304.1 <sup>b</sup>	132	All thermoleophilia	Unknown
WP_093118364.1 <sup>b</sup>	257	All thermoleophilia	Unknown
WP_093118537.1	154	All thermoleophilia	Unknown
WP_093118589.1	178	All thermoleophilia	Unknown
WP_093118635.1	120	All thermoleophilia	Unknown
WP_093118833.1	82	All thermoleophilia	Unknown
WP_093119001.1	187	All thermoleophilia	Unknown
(B) CSPs unique to 7	hermoleophilia	class but not found in ne	w cluster (3)
WP_093116803.1	141	<i>Thermoleophilia</i> except new cluster	Unknown
WP_093118036.1	211	Thermoleophilia except new cluster	Unknown
WP_093116745.1	226	Thermoleophilia except new cluster	Unknown

**TABLE 3** Conserved Signature Proteins that are uniquely found in the subgroups of *Thermoleophilia* class.

Accession no.	Length	Specificity
(A) CSPs uniquely p	present in famil	y Thermoleophilaceae (29) <sup>a</sup>
WP_093115090.1	197	Thermoleophilaceae
WP_093115144.1	179	Thermoleophilaceae
WP 093115294.1	164	Thermoleophilaceae
WP 093115296.1	180	Thermoleophilaceae
WP 093115479.1	319	Thermoleophilaceae
WP 093115661.1	93	Thermoleophilaceae
WP 093115901.1	156	Thermoleophilaceae
	202	Thermoleophilaceae
WP 093116532.1	154	Thermoleophilaceae
WP 093116727.1	429	Thermoleophilaceae
WP_093116780.1	110	Thermoleophilaceae
WP_093116825.1	68	Thermoleophilaceae
WP_093116919.1	83	Thermoleophilaceae
WP_093117092.1	264	Thermoleophilaceae
WP_093117483.1	93	Thermoleophilaceae
WP_093117587.1	83	Thermoleophilaceae
WP_093117642.1	114	Thermoleophilaceae
WP_093117817.1	157	Thermoleophilaceae
WP_093117827.1	199	Thermoleophilaceae
WP_093117877.1	136	Thermoleophilaceae
WP_093118281.1	403	Thermoleophilaceae
WP_093118340.1	146	Thermoleophilaceae
WP_093118436.1	119	Thermoleophilaceae
WP_093118524.1	170	Thermoleophilaceae
WP_093118569.1	80	Thermoleophilaceae
WP_093118679.1	148	Thermoleophilaceae
WP_093118731.1	93	Thermoleophilaceae
WP_093118750.1	573	Thermoleophilaceae
WP_093118752.1	195	Thermoleophilaceae
(B) CSPs uniquely	present in Cone	exibacteraceae, Solirubrobacteraceae,
and Patulibacterac	eae (24)	
WP_022926981.1	246	Conexibacteraceae, Solirubrobacteraceae, Patulibacteraceae
WP_022926986.1	115	Conexibacteraceae, Solirubrobacteraceae, Patulibacteraceae
WP_022927172.1	216	Conexibacteraceae, Solirubrobacteraceae, Patulibacteraceae
WP_022927347.1	417	Conexibacteraceae, Solirubrobacteraceae, Patulibacteraceae
WP_022927380.1	114	Conexibacteraceae, Solirubrobacteraceae, Patulibacteraceae
WP_022927389.1	468	Conexibacteraceae, Solirubrobacteraceae, Patulibacteraceae
WP_022927525.1	461	Conexibacteraceae, Solirubrobacteraceae, Patulibacteraceae
WP_022927538.1	181	Conexibacteraceae, Solirubrobacteraceae, Patulibacteraceae
WP_022927665.1	153	Conexibacteraceae, Solirubrobacteraceae, Patulibacteraceae
WP_022927703.1	253	Conexibacteraceae, Solirubrobacteraceae, Patulibacteraceae
WP_022927703.1	253	Conexibacteraceae, Solirubrobacteraceae, Patulibacteraceae
WP_022927792.1	224	Conexibacteraceae, Solirubrobacteraceae, Patulibacteraceae

<sup>a</sup> The number in brackets represents the total number of CSPs unique to the specific group. <sup>b</sup>Four CSPs are also present in the genome of Gaiella occulta F2-233 (GenBank accession: GCA\_003351045.1).

(Continued)

#### TABLE 3 | Continued

Accession no.	Length	Specificity
WP_022927799.1	564	Conexibacteraceae, Solirubrobacteraceae, Patulibacteraceae
WP_022927801.1	265	Conexibacteraceae, Solirubrobacteraceae Patulibacteraceae
WP_022928134.1	160	Conexibacteraceae, Solirubrobacteraceae Patulibacteraceae
WP_022928438.1	136	Conexibacteraceae, Solirubrobacteraceae, Patulibacteraceae
WP_022928438.1	136	Conexibacteraceae, Solirubrobacteraceae, Patulibacteraceae
WP_022929183.1	133	Conexibacteraceae, Solirubrobacteraceae Patulibacteraceae
WP_022929536.1	104	Conexibacteraceae, Solirubrobacteraceae Patulibacteraceae
WP_022929558.1	227	Conexibacteraceae, Solirubrobacteraceae Patulibacteraceae
WP_022930026.1	369	Conexibacteraceae, Solirubrobacteraceae Patulibacteraceae
WP_022930484.1	604	Conexibacteraceae, Solirubrobacteraceae, Patulibacteraceae
WP_028721853.1	100	Conexibacteraceae, Solirubrobacteraceae Patulibacteraceae
WP_051160538.1	289	Conexibacteraceae, Solirubrobacteraceae Patulibacteraceae
(C) CSPs uniquely p	resent in fan	nilv Patulibacteraceae (31)
WP 022926969.1	211	Patulibacteraceae
WP 022926970.1	304	Patulibacteraceae
WP 022927005.1	421	Patulibacteraceae
WP 022927132.1	338	Patulibacteraceae
WP 022927548.1	105	Patulibacteraceae
WP 022927557.1	100	Patulibacteraceae
WP 022927572.1	162	Patulibacteraceae
WP 022928009.1	773	Patulibacteraceae
WP 022928045.1	170	Patulibacteraceae
WP 022928129.1	165	Patulibacteraceae
WP 022928139 1	176	Patulibacteraceae
WP 022928142.1	174	Patulibacteraceae
WP 022928143 1	155	Patulibacteraceae
WP 022928333.1	248	Patulibacteraceae
WP 022928557.1	67	Patulibacteraceae
WP 022928588.1	110	Patulibacteraceae
WP 022928655.1	62	Patulibacteraceae
	242	Patulibacteraceae
	236	Patulibacteraceae
WP_022929154.1	209	Patulibacteraceae
	417	Patulibacteraceae
	66	Patulibacteraceae
WP_022929735.1	411	Patulibacteraceae
WP_022929823.1	153	Patulibacteraceae
WP_022929914.1	269	Patulibacteraceae
WP_022929990.1	171	Patulibacteraceae
WP_022930081.1	281	Patulibacteraceae
WP_022930294.1	190	Patulibacteraceae
WP_022930374.1	124	Patulibacteraceae
WP_022930538.1	472	Patulibacteraceae
WP_022930714.1	206	Patulibacteraceae

<sup>a</sup> The number in brackets represents the total number of CSPs unique to the specific group.

The complete genome sequences of type species and recently reported MAGs of *Thermoleophilia* are great resources to explore group-specific molecular markers. We focused on two molecular markers as noted earlier: CSIs and CSPs (Gao et al., 2009; Gupta and Gao, 2009; Zhang et al., 2016). Both have been identified for various prokaryotic phyla or other taxonomic ranks higher than genera in the past two decades, and proved to be very useful for phylogenetic and evolutionary studies (Gao and Gupta, 2012b; Ho et al., 2016; Alnajar and Gupta, 2017; Hu et al., 2018).

Comparative genomic analyses of species of the class Thermoleophilia and other taxonomic units within the kingdom bacteria led to the identification of 12 CSIs in various conserved universal proteins that are only found in Thermoleophilia species but not in other bacteria (Table 1). For example, a 4 amino acids (aa) insertion in a very conserved region of quinolinate synthase NadA was specifically shared by Thermoleophilia species (Figure 2). NadA is a widely distributed protein in both Archaea and Bacteria and highly conserved due to its important role in nicotinamide adenine dinucleotide (NAD) de novo biosynthesis (Ollagnier-De Choudens et al., 2005). A 4aa insertion that is located in a surface loop region of the 3D structure (Volbeda et al., 2016) is only found in homologs from Thermoleophilia but not from species outside this class. Therefore, this 4-aa insertion is a distinctive characteristic of the Thermoleophilia class. Sequence information for additional 11 CSIs that are specific to all members of this class including assembled genomes of not yet cultivated species is provided in Supplementary Figures S2–S12. In view of their specificity, these CSIs can serve as molecular markers to define and distinguish species belonging to the Thermoleophilia class. In addition, none of these 12 CSIs are found in the genome of Gaiella occulta F2-233, which is the only genome recently available from the deep-branching order Gaiellales.

Except the CSIs, we performed BLASTp searches for each protein from the type species T. album ATCC 35263 to identify CSPs that are specific to the Thermoleophilia class. In total, 32 proteins are uniquely shared by almost all sequenced Thermoleophilia genomes but not found in any other bacterial taxa except 4 present in G. occulta F2-233 (Table 2). Foesel et al. have proposed that Gaiellales is a deeply branching order within the class Thermoleophilia based on 16S rRNA analyses and some shared phenotypic features of one single strain G. occulta F2-233 and other Thermoleophilia/Rubrobacteria species (Foesel et al., 2016). The presence of 4 CSPs in the same G. occulta strain could be derived from the common ancestor of Gaiellales with the other Thermoleophilia orders or due to lateral gene transfer, which awaits confirmation from more genomes of the Gaiellales. Additionally, 3 proteins are missing in the MAGs from the newly defined potential family based on our phylogenomic analysis presented in Figure 1A but found in the other members of the class, which is possibly due to incomplete genome information. Indeed, the assembly qualities of MAGs varies as indicated by the summary of Contig-N50 statistic values in Supplementary Table 1. Therefore, it is very likely that the identified 3 CSPs are present in the species of the newly defined cluster, while the MAGs did not cover the sequence region. Together with the identified CSIs, these

			314		367
Thermoleophiliaceae	Thermoleophilum album	1225101858	GEYTYFAADIAYHADKLER	LARAEEP	GGARAIDVLGADHHGYVARMRAAWQALG
1	bacterium HR41	1286951175	A-	A-SPE	¥GG
]	Solirubrobacterales bacterium 67-14	1113228701	VGVVI		-ND-LPAI-S-E
	Solirubrobacterales bacterium 70-9	1113216195	PLMPW		-FD-L-NLSIE
	Patulibacter minatonensis	916865331	-RPS-VHRV-		-YD-VIEFE
	Conexibacter woesei	652639730	HSAQN-R		-YD-MVYIGFS
Other	Solirubrobacterales bacterium URHD0059	654612039	HSAQN-R		-FD-MVYIGYE
Thermeleenhilie	Solirubrobacter sp. URHD0082	739553939	-SSVDR-		-FD-L-NG-VK-SIA
Inermoreophiria	Patulibacter medicamentivorans	494850325	-RPS-VYRV-		-FD-VIH-LKFV
	Solirubrobacter soli	921289701	-APVEN-VQ-		-YD-LIGKGVME
	Patulibacter americanus	551309280	-RPSS-VHRV-		-YD-I-NIEFV
	Actinobacteria bacterium 13 1 20CM 3 68 9	1125519411	PR		-AE-MPVPIA
	Acidimicrobiaceae bacterium TMED224	1200577077	PVPHE-F		LVPAM-M
	marine actinobacterium MedAcidi-G3	745856093	PVPHE-F		LVPAM-M
	Citricoccus massiliensis	1330257762	-GLTPQN		-FDKI-N-WIPI
	Dietzia cinnamea	1028235075	-SLLPK		-FEKL-NIWIPKIE
	Actinobacteria bacterium 66_15	973088847	-SFD-		-FDLVIWKES-VA
Other	Acidimicrobiaceae bacterium TMED130	1200463057	FLTPKN-F		SDWL-N-WIKM
	Atopobium vaginae	493347989	WFD-		-AQQSLWIVTNVCD
Actinobacteria	Actinobacteria bacterium IMCC26207	918748109	FLLPRFD-		-FDLL-N-WKV-S
	Coriobacterium glomerans	503474332	-DS-VWFQ-		VDHIW-SIG-VCD
	Olsenella sp. An290	1199585886	S-VWFQ-		VDHVIWIQ-VQS-CE
	Atopobium	488626704	S-VWFQ-		QDYSLWIV-N-CD
	Enorma massiliensis	517958342	-DSSWN-FQ-		VD-VL-IWIVK-VCD
l	- Mycobacterium abscessus	1119054200	-SLTPKND-		-FEKL-NIWIPKIE
	Bacillus shacheensis	973243049	-SLTPSLFD-		-NT-LPI
	Thiorhodococcus drewsii	494101699	-QASMM		-FD-VIWIP-VKL
	Desulfurobacterium atlanticum	1219307349	YI		-YDKGIWIP-VKIE
	Methyloterricola oryzae	771615812	-QTSL		-FD-VIWIP-VKL
Other Basteria	Sedimenticola thiotaurini	821538298	-QTSM		-FE-VWP-VKMV-
Other Bacterra	Thioflavicoccus mobilis	505093553	-QTSM		-FVWP-VKLT
	Thioalkalivibrio denitrificans	1148530547	-QTSM		-YE-VVWP-VKL
	Thermovibrio ammonificans	503303206	К-		-YDVWIP-VELK
	Sedimenticola selenatireducens	1321751653	-QTSM		-FE-VWP-VKMV-

FIGURE 3 | CSI specific to *T. album* and MAG HR41. Partial sequence alignment of arginine–tRNA ligase showing a 7 amino acid insertion that is uniquely shared by *T. album* and MAG HR41.

			137	181
	Patulibacter americanus	551309834	MLEPKYAIAMGACASSMGVFNNYAVVP	ADKFMPVDVHIPGCPPRP
Conquibestances	Patulibacter medicamentivorans	494844423	I	
Conexidacteraceae	Patulibacter minatonensis	1180795370	I	
Patulibacteraceae	Solirubrobacterales bacterium URHD0059	654608822	DI	v
Solimphrobactoracoac	Conexibacter woesei	652641364	DI	LV
Solliubrobacteraceae	Solirubrobacter soli	921289541	WSI	LI
	Solirubrobacter sp. URHD0082	916717095	WSQI	L-V
	Solibacteraceae bacterium SbA6	1346293128	-PWST-TL	VNQVI-IFV
Other	Solirubrobacterales bacterium 67-14	1113229265	WSSG-M-QQ (	GVV
Other	Solirubrobacterales bacterium 70-9	1113216860	WSSG-M-AQ (	;I-L
Thermoleophilia	Thermoleophilum album	1225102635	WV-SA-ML-A (	GR-L-IYV
-	bacterium HR41	1286949636	WV-SA-ML-A (	GR-L-IYV
	Streptosporangium subroseum	1216215008	-AWVVG-MI-Q (	G V-HVVIYL
	Thermobispora bispora	502895570	-AWVVG-MQ (	G V-HIVYL
	Spirillospora albida	663122931	-TWVVG-MI-Q (	G V-HIVIYL
Other	Actinomadura kijaniata	1056372406	-AWVVG-MQ (	G V-HIVYL
other	Rhodococcus triatomae	494801652	-VWVLVG-MI-Q (	G V-HLVIYL
Actinobacteria	Thermomonospora curvata	502617842	-TWVVG-MQ (	G V-HIVIYL
	Thermobifida fusca	510815716	-AWVVG-MI-Q (	G V-HIVIYL
	Gordonia westfalica	1124569337	-VWVLVG-MI-Q (	G V-HVVIYL
	Mycobacterium goodii	907630117	-VWVLVG-MQ (	G V-HVVIYL
	Serinibacter salmoneus	1267709596	-SWV-SG-MQ (	G C-HVVIYL
	Planctomycetes HGW-Planctomycetes-2	1308511053	-AWV-STGDTQ (	G C-Q-IYV
	Escherichia coli	1124136448	WV-SN-G-MYDI-SQ (	G VIY
	Nephila clavipes	1355916818	WV-SSN-G-MYDI-SQ (	G VLY
	Salmonella enterica	646664994	WV-SN-G-MYDI-SQ (	3 VIY
Other Bactoria	Estrella lausannensis	1267161906	WVDGYI-Q (	G VIVIYVA
Other Bacteria	Shigella boydii	1129010287	WV-SN-G-MYDI-SQ (	3IY
	Klebsiella pneumoniae	1319911046	WV-SN-G-MYDI-SQ (	G VIY
	cyanobacterium TDX16	1213740860	-AWV-STGDTQ (	G V-Q-IYV
	Desulfobacca sp. 4484_104	1162405805	-TWVG-I-DSQ (	G I-LYLYV
	Yersinia pestis	514988300	WV-SN-G-MYDI-SQ (	G VLY
FIGURE 4   CSI specific to the families	s Conexibacteraceae. Solirubrobacteraceae and	Patulibacteracea	e. Partial alignment of the protein N	JADH-auinone

FIGURE 4 CSI specific to the families Conexibacteraceae, Solirubrobacteraceae and Patulibacteraceae. Partial alignment of the protein NADH-quinone oxidoreductase subunit B showing a 1 amino acid deletion that is uniquely shared by 3 families Conexibacteraceae, Solirubrobacteraceae and Patulibacteraceae.

CSPs are additional molecular markers for *Thermoleophilia*. It should be mentioned that all these identified CSPs are hypothetical proteins with unknown function. Since they are restricted to species of *Thermoleophilia*, functional studies on them may uncover biochemical or physiological features that are unique to this class.

# Molecular Signatures for Major Lineages Within *Thermoleophilia*

As described earlier, the order *Thermoleophilales* or its sole family *Thermoleophiliaceae* only have two genomes available, including *T. album* ATCC 35263 and MAG "bacterium HR41." Our analyses identified 7 CSIs in different proteins

ConexibacteraceaeConexibacter voesi9161518GCULQUTTEVENFRETS GGIDETLINTEMEQADFOSFRTTSolirubrobacterales bacteriuu NED0059801rubrobactera p, URH000821175138173		_		40	82
OtherSolirubrobacter age, URB0082654611834Y-EY-E	Conevibacteraceae	Conexibacter woesei	916615184	GGLLQQTTEVENFPGFPA	GIDGPTLMTKMREQAEDFGSRFIT
Other    Solirubrobacter sp. URHD0082    1175138179	conexibacteraceae	Solirubrobacterales bacterium URHD0059	654611834	YG -	AEQ-L-DTL-
Other ThermoleophiliaSolirubrobacterisoli Actinobacteria bacterium 13 1 200M 3.68 9 1125520470	[	Solirubrobacter sp. URHD0082	1175138179	У-Е	-VTEM-QQL-DRT
Other IntermoleophiliaActinobacteria bacterium 13_1_20CM_3_66.9 bacterium 1841 Solirubrobacterales bacterium 67-14 patulibacter marianensis Patulibacter marianensis Bolirubrobacterales bacterium 67-14 patulibacter marianensis Patulibacter marianensis Patulibacter Patulibacter marianensis Patulibacter Patulibacter Patulibacter marianensis Patulibacter Patulibacter marianensis Patulibacter Patulibacte		Solirubrobacter soli	921290520	У-К	MEM-QDL-DRT
Other Thermoleophiliabacterium HR41III1286951328L-DYT-D-VQQM-SDF-RR-T-T-SThermoleophiliaSolirubrobacterales bacterium 67-14 Patulibacter minatonensis1113226685 652518571DYT-DD-M-QDL-DRA-LKPatulibacter meinatonensis652518571D-DYYYY		Actinobacteria bacterium 13 1 20CM 3 68 9	1125520470	QNDYG	MEQHF-ARTS
ThermoleophiliaSolirubrobacterales bacterium 67-141113226685QN-DYWEM-EMRF-DR-TV- G2518571Patulibacter minatonensis652518571DYYL-DM-QDL-DAR-A-LQ- Patulibacter medicamentivoransPatulibacter medicamentivorans494850811DYYL-DM-QDL-DAR-A-LQ- DY-YEPatulibacter medicamentivorans1225104347Y-YM-EM-SRF-ARTVSolirubrobacterales bacterium 70-91113215192QN-DY-YEM-EM-SRF-AR-TVCryptosporangium arvum1223939415A-MNCGD-DNKTKAELV-Mycetocola miduiensis1222661519A-MN	Other	bacterium HR41	1286951328	YY-D	-VQQM-SDF-RRTS
IntermoticopinitiesPatulibacter minatonensis652518571D	Thermeleenhilis	Solirubrobacterales bacterium 67-14	1113226685	QNDYЕ	MEMRF-DRTV-
OtherPatulibacter americanus55130771D	Inermoreophiria	Patulibacter minatonensis	652518571	У-К	LDM-QDL-DRA-LK-
Patulibacter medicamentivorans494850811DY-KMDM-QDL-DR-A-LK-Thermoleophilum album1225104347L-DYY-D-VQQM-ADF-RR-T-LSSolirubrobacteriales bacterium 70-911321519QN-DYY-D-VQQM-ADF-RR-T-VPropionibacterium cyclohexanicum1223939415A-MNDYY-EMD-DS-SR-AR-T-VMicrobrobacterial122261519A-MNDYY-EMQN-TKAELMicrobrobacterial122261519A-MNYEMD-DNKRAL-Y-Tessaracoccus bendigoensis1120321519A-MNYEMD-DNKRAELLongispora albida517160181A-MTEMD-DNKRAELPresiaracoccus aquimaris1149110729A-MNYEMD-DNKRAELActinobacteriaMicromonopora pattaloongensis1223515108A-MTDMEDSKRAELActinobacteria bacterium1272478442A-MNYEMEDSKRAELActinobacteria bacterium breve1307865310Q-VNDDKRAEL-DBifidobacterium breve1307865310Q-VNDDRRAEL-DBifidobacterium breve1307865310Q-VNDDRRAEDVerrucosispora maris653092629A-MT		Patulibacter americanus	551307771	У-К	LDM-QDL-DARA-LQ-
OtherItermoleophilum album1225104347L-DYY-DVQQM-ADF-RR-T-T-LSSolirubrobacterales bacterium 70-91113215192QN-DDYY-EM-EW-SRF-AR-T-V-Solirubrobacterium cyclohexanicum1223939415A-MNDDM-QN-TYAELV-Cryptosporangium arvum737893215A-MTDM-QN-ARAELMycetocola miduiensis1222661519A-VNMD-DN-KRAELV-Iongispora albida517160181A-MNMD-DN-KRAELTessaracoccus aquimaris1149110729A-MN		Patulibacter medicamentivorans	494850811	У-К	MDM-QDL-DRA-LK-
Other  solirubrobacterales bacterium 70-9  1113215192 QN-DY-Y-E MEM-SRF-AR-TV-    Propionibacterium cyclohexanicum  1223939415 A-MNDDY-Y-E MQN-T-TK-AELV-    Mycetocola miduiensis  1222661519 A-MTDD M-D-DNKT-AELV-    Mycetocola miduiensis  1222661519 A-MND M-D-DNKT-AELV-    Mycetocola miduiensis  1120321519 A-NNE M-D-DNKR-AELV-    Longispora albida  517160181 A-MNE M-D-DNKR-AELV-    Micromonospora pattaloongensis  1122315108 A-MNE M-D-DNKR-AELV-    Actinobacteria  Micromonospora pattaloongensis  122315108 A-MTID ME-OSKR-AELV-    Actinobacteria bacterium  1272478442 A-MTID ME-OSKR-AELV-    Kibdelosporangium aridum  1181020341 A-MTD-D-NKSDKR-AELV-    Kibdelosporangium aridum  1181020341 A-MTD-DDD		Thermoleophilum album	1225104347	ҰҰ-Д	-VQQM-ADF-RRTLS
Other ActinobacteriaPropionibacterium cyclohexanicum (Cryptosprangium arvum Mycetocola miduiensis Tessaracoccus bendigoensis Longispora albida1223939415 737893215 7-A-MT	l	-Solirubrobacterales bacterium 70-9	1113215192	QNDYY-Е	MEM-SRF-ARTV-
OtherCryptosporangium arvum737893215A-MTGD-DNKT-AELV-Mycetccla midniensis1222661519A-VNDM-D-DN-KRA-L-YTessaracoccus bendigoensis1120321519A-VNED-DNKRA-L-YLongispora albida517160181A-WTED-DNKRA-L-YTessaracoccus aquimaris1149110729A-WNED-DNKRAEActinobacteriaMicromonospora pattaloongensis1223515108A-MTIActinobacteria bacterium1272478442A-MNDMEDSKRAEIActinobacteria bacterium1272478442A-MNSQ-DNARAUITABifidobacteriu breve1307885310-Q-VNDMEDSDKTQABifidobacterium breve1307885310Q-VNDMDNKRAEIActinocispora maris653092629A-MTDAVM-E-DQRRAEIBifidobacterium breve503501331A-MTDMDNKTQABifidobacterium preve503501331A-MTDMEDNKRAEIAlicyclobacillus ponorum652571869Q-TIDM-ESNQRAMAlicyclobacillus ponorum652571869Q-TI		Propionibacterium cyclohexanicum	1223939415	A-MND	MQNTKAEL
Other ActinobacteriaMycetocola miduiensis1222661519 1120321519A-VND A-WNPM-D-DN-KR-A-L-Y D-AN-AR-AEL D-DN-KR-AEL D-DN-KR-AEL DD-N-KR-AEL 		Cryptosporangium arvum	737893215	A-MTG	DDNKT-AELV-
Other ActinobacteriaTessaracoccus bendigoensis1120321519 517160181A-MNYE A-MNEDANARAEL M-D-DDN-RRAEL M-D-DDN-RRALL-SActinobacteriaTessaracoccus aquimaris1149110729 1223515108A-MNE A-MNEDNNARAEL A-MNActinobacteriaMicromonospora pattaloongensis1223515108 1223515108A-MTID ME-DSKRAEL D-DN-KRAELL- Actinobacteria bacterium1272478442 1272478442 A-MNDDN		Mycetocola miduiensis	1222661519	A-VND	MDDNKRA-L-Y
Other ActinobacteriaLongispora albida517160181A-MTEM-D-DN-KR-AE Cohnella laeviribosiActinobacteriaMicromonospora pattaloongensis1149110729A-MTIED-D-N-KR-AE Cohnella laeviribosiMicromonospora pattaloongensis1223515108A-MTIDM-E-DS-KR-AEL Actinocatenispora seraActinobacteriabacterium1272478442A-MTIDM-E-DS-DRTKM Cohnella laeviribosiKibdloaporangium aridum1181020341A-MTDDM-E-DS-DRTKM CO-VNDCother BacteriaBifidobacteria bacterium1272478442A-MNDDBifidobacteriaCother Bacteria10785310Q-VNDCDKTQA Streptonyces sp. XY51192547939Paenibacillus piniumi653102262A-MTDAVM-E-DQ-RRAE-L- Verucosispora marisAVM-E-DQ-RRAE-L- Verucosispora marisOther BacteriaPaenibacillus piniumi Alicyclobacillus pomorum652571869Q-TIDL-E-DNKR-AE-L- PAE-DN-KR-AE-L-Paenibacillus castaneae133269283Q-TTDM-ESNQRAM- PAE-DN-KKR-AE-V- DM-E-DN-KR-AE-V-V Paenibacillus castaneae503259880Q-TIDDehalobacter sp. FTH1 Cohnella laeviribosi517834684Q-TTEM-EANKR-AE-V- PAE-EAN-KR-AE-V-V DD-AD-ANRAE-V-V		Tessaracoccus bendigoensis	1120321519	A-MNYE	DANARAEL
OtherTessaracoccus aquimaris1149110729A-MNYEDNARAILVSActinobacteriaMicromonospora pattaloongensis1223515108A-MTDM-E-DSKRAEActinobacteria66360948A-MTDM-E-DSKRAEActinobacteria bacterium1272478442A-MNDM-E-DSDNKRAELActinobacteria bacterium1272478442A-MNDM-E-DSDRTKMKibdelosporangium aridum1181020341A-MDSQ-DAARAVLTABifidobacterium breve1307885310-Q-VNDM-D-D-DKRAELStreptomyces sp. XY511925479393-S-TTDAVME-DQRKAEM-DHamadaea tsunoensis653092629A-MTDAVME-DQRKAEV-Verrucosispora maris503501331A-MTDM-ESNQRAL-Alicyclobacillus pomorum652571869-Q-TLDMESNQRAVDAlicyclobacillus pomorum652571869-Q-TTEMESNQRAVDPaenibacillus castaneae1332692983-Q-TTEMESNQRAVDDehalobacter sp. FTH1736355551A-MNMESNQRAVDCohnella laeviribosi517834684-Q-TTEMEANKR-AEL-		Longispora albida	517160181	A-MTE	MDDNKRAE
ActinobacteriaMicromonospora pattaloongensis122351518A-MTIDME-DSKRAEActinocatenispora sera66366048A-MTDD-DDN-KRAELActinobacteria bacterium1272478442A-MTDD-DDN-KRAELKibdelosporangium aridum1181020341A-MTDME-DSD-RTKMBifidobacterium breve130786531Q-VNDDMD-DR-DRTKMStreptomyces sp. XV51192547939-S-TTDMD-DR-DKTQAHamadaea tsunoensis65309262A-MTDAVM-EDQRRAE-V-Verrucosispora maris503501931A-MTDM-ESNQRAAlicyclobacillus pinihumi65517169-Q-TTDM-ESNQ-RAV-Paenibacillus castaneae1332692983-Q-TTDM-ESNQ-R-AV-Paenibacillus castaneae1332692983-Q-TTDMESNQ-R-AV-Paenibacillus castaneae50325980-Q-TTDMESNQ-R-AVDDehalobacter sp. FTH1736355551A-MNDLENRAE-V-Cohnella laeviribosi517834684-Q-TT	Other	Tessaracoccus aquimaris	1149110729	A-MNYE	DNARAILVS
Actinocatenispora sera663660948A-MTDD-DNKRAELL-Actinobacteria bacteria1272478442A-MNDD-DNKRAELL-Actinobacteria bacteria1181020341A-MNDD-DNKRALLAKibdelosporangium aridum1181020341A-MNDDDRTKMBifidobacterium breve1307865310Q-VNDD-D-DRDKTQAStreptomyces sp. XY511925479939S-TTDDLNAKAEM-DHamadaea tsunoensis653092629A-MTDAVM-E-DQRRAE-V-Verrucosispora maris503501931A-MTDE-DNKRAE-V-Verrucosispora maris655112291Q-TTDESNQRAM-Alicyclobacillus pomorum652571869Q-TLDM-ESNQRAM-Alicyclobacillus castaneae1332692983Q-TTDM-EENKR-AE-V-Dehalobacter sp. FTH173635551A-MNDL-DARAR-AE-V-Cohnella laeviribosi517834684Q-TTEM-EANKR-AQ-R-	Actinobacteria	Micromonospora pattaloongensis	1223515108	A-MTID	MEDSKRAE
Actinobacteria bacterium  1272478442 A-MN		Actinocatenispora sera	663660948	A-MTD	DDNKR-AELL-
Kibdelosporangium aridum  1181020341 A-MDS Q-DA-ARXUTA    Bifidobacterium breve  1307885310 Q-VND M-D-D-DNAA    Streptomyces sp. XY511  92547939 S-TTD M-D-D-DN		Actinobacteria bacterium	1272478442	A-MNS	MEDSDRTKM
Bifidobacterium breve    1307885310   Q-VND   MD-DRDK-TQA      Streptomyces sp. XY511    925479939   S-TTD   DLNAKAEM-D      Hamadaea tsunoensis    653092629   A-MTD    AVM-E-C-DNKRAE-L-      Verrucosispora maris    50351931   A-MTD   MEDNKRAE-L-      Paenibacillus pinihumi    655112291    -Q-TTD   MEDNKRAE-L-      Alicyclobacillus pomorum    652571669   Q-TLD   MEDNKKAK-VA      Paenibacillus castaneae    132692983   Q-TTD   MEENNKRAE-L-      Dehalobacter marianensis    503259860   Q-TLD   MEENNKRAE-L-      Dehalobacter sp. FTH1    736355551   A-MND   MEENNKRAE-LV-      Cohnella laeviribosi    517834684   Q-TTE   MEANNKRAE-IV-		Kibdelosporangium aridum	1181020341	A-MDS	QDAARAVLTA
Streptomyces sp. XY511    925479939   S-TTD   DLNAKAEM-D      Hamadaea tsunoensis    653092629   A-MTD    AVME-DQRRAE-V-      Verrucosispora maris    503501931   A-MTD   L-E-DDKRAE-V-      Paenibacillus pinihumi    655112291   Q-TTD   MESNQRAM-      Alicyclobacillus pomorum    652571869   Q-TLD   MESNQRAM-      Paenibacillus castaneae    1332692983   Q-TTD   MESNQR-AM-      Dehalobacter marianensis    503259880   Q-TTD   MESNQRA-A-VD      Dehalobacter sp. FTH1    736355551   A-MND   L-DA-RARAETV-      Cohnella laeviribosi    517834684   Q-TTE   MEANKRAQ-R-		Bifidobacterium breve	1307885310	Q-VND	MDDRDKTQA
Hamadaea tsuncensis    653092629   A-MTD    AVMEDQRRAE-V-      Verrucosispora maris    503501931   A-MTD   L-EDNKR-AE-V-      S03501931   A-MTD   L-E-DQRR-AE-V-   L-E-DNKR-AE-V-      Alicyclobacillus piniumi    655112291   Q-TTD   L-E-SN-QRAM-      Alicyclobacillus pomorum    652571869   Q-TLD   M-E-SN-QRAM-      Paenibacillus castaneae    1332692983   Q-TTE   M-E-ENN-CR-AA-TV-      Thermerbacter marianensis    503259880   Q-MIDDL-D-AR-AV-D   L-D-AR-ARAETV-      Ochnella laeviribosi    517834684   Q-TTE   ME-ANN-KRAETV-		Streptomyces sp. XY511	925479939	S-TTD	DLNAKAEM-D
Verrucosispora maris    503501931   A-MTAD   L-E-D-N-KRAE-L-      Paenibacillus pinihumi    655112291   Q-TTD   M-ESN-QR-AL-      Alicyclobacillus pomorum    652571669   Q-TLD   M-EDN-KRAE-L-      Paenibacillus castaneae    1332692983   Q-TLD   M-EDN-KRAE-L-      Thermaerobacter marianensis    50325983   Q-TLD   M-EDN-KRAE-VD      Dehalobacter sp. FTH1    73635551   A-MND   L-DAR-QRAVD      Cohnella laeviribosi    517834684   Q-TTE   MEAN-KRAQ-R-		Hamadaea tsunoensis	653092629	A-MTD	AVMEDQRRAE-V-
Paenibacillus pinihumi    655112291   Q-TTD   MESNQRAM-      Alicyclobacillus pomorum    62571869   Q-TLD   MEDNKKAK-VA      Paenibacillus castaneae    132692983   Q-TLD   MEDNKKAAVA      Thermaerobacter marianensis    50325980   Q-MLD   L-DARQRA-AVD      Dehalobacter sp. FTH1    73635551   A-MNYTE   MDLNARA-VP      Cohnella laeviribosi    517834684   Q-TTE   MEANKR-AQ-R-	l	Verrucosispora maris	503501931	A-MTAD	LEDNKRAE-L-
Alicyclobacillus pomorum  652571869 Q-TLD MEDNKKAK-VA    Other Bacteria  Paenibacillus castaneae  1332692983 Q-TLE MEENKRAT    Thermaerobacter marianensis  503259880 Q-HLD LDARQU-RA-AVD    Dehalobacter sp. FTH1  736355551 A-MNE MEANKR-AEIV-    Cohnella laeviribosi  517834684 Q-TTE MEANKR-AQ-R-	ĺ	Paenibacillus pinihumi	655112291	Q-TTD	MESNQRAM-
Other Bacteria  Paenibacillus castaneae  1332692983 Q-TTE MEENAT    Thermaerobacter marianensis  503259880 Q-MLD L-DARQRA-AVD    Dehalobacter sp. FTH1  736355551 A-MNE MEANRAEIV-    Cohnella laeviribosi  517834684 Q-TTE MEANKRAQ-R-		Alicyclobacillus pomorum	652571869	Q-TLD	MEDNKKAK-VA
Thermaerobacter marianensis    503259880   Q-MLD   LDARQRA-AVD      Dehalobacter sp. FTH1    736355551   A-MNTE   MDLNARAEIV-      Cohnella laeviribosi    517834684   Q-TTE   MEANKRAQ-R-	Other Bacteria	Paenibacillus castaneae	1332692983	Q-TTE	MEENKRAT
Dehalobacter sp. FTH1    736355551   A-MNTE   MDLNARAEIV-      Cohnella laeviribosi    517834684   Q-TTE   MEANKRAQ-R-		Thermaerobacter marianensis	503259880	Q-MLD	LDARQRA-AVD
Cohnella laeviribosi 517834684Q-TTEMEANKRAQ-R-		Dehalobacter sp. FTH1	736355551	A-MNYTE	MDLNARAEIV-
	l	- Cohnella laeviribosi	517834684	Q-TTE	MEANKRAQ-R-

FIGURE 5 | CSI specific to Conexibacteraceae. A 1 amino acid insertion in the protein thioredoxin-disulfide reductase that is uniquely shared by C. woesei and associated MAG.

			170	211
Solirubrobacteraceae	Solirubrobacter soli	654598081	VLKGHGNDGRSGFEGIR GG	P EGTVVGTYLHGPLLPKNSWFAD
Dorradionacteraceae	Solirubrobacter sp. URHD0082	916716971	v	- R-NA
Ì	Conexibacter woesei	652640894	VI	K-N-IA
	Solirubrobacterales bacterium URHD0059	654610719	КVH	R-S-IA
	Patulibacter medicamentivorans	494845702	ST-AAA-	R-N-IA
0+1	Patulibacter americanus	551308866	AT-SHA-	T-N-IA
Other	Patulibacter minatonensis	738836408	st-vA-	S-NA
Thermoleophilia	Thermoleophilum album	1225104799	RNDRTVV	R-N-IV-L
	Actinobacteria bacterium 13_1_20CM_4_69_9	1125476587	-VA-FDC-	V-RAR-P
	Actinobacteria bacterium RBG_16_67_10	1082231316	F-FEC-	L-RAIR-P-L
	Solirubrobacterales bacterium 67-14	1113228176	-IHN-EDVK	-DNLIA-L
l	Actinobacteria bacterium 13_1_20CM_3_68_9	1125518739	-VRN-TD-LV-	RRNMIA-L
١	Coriobacteriaceae bacterium BV3Ac1	545610270	YKTCL	YKN-LIPGV
	Olegusella massiliensis	1057150058	YKTCL	YKN-LIPGV
	Atopobium sp. oral taxon 810	545384899	YKTCL	YKNIPGV
	Olsenella umbonata	1222819417	YEYCL	YKNVPGV
Other	Actinomyces radicidentis	987451697	-VT-DN-TD-SA-	AHN-ISPLV
Actinobactoria	Collinsella sp. An268	1199465925	-VA-TN-TD-YV	HHGLFV-SPAL
ACLINODACLEIIA	Actinomyces radicidentis	1180915828	-VT-DN-TD-SA-	AHN-ISPLV
	Dermacoccus sp. PE3	828383866	-RS-FN-ED-HA-	TNNSIA-PHL
	Actinomyces israelii	759876994	-RS-DN-KD-TA-	ADH-ISPAV
	Tetrasphaera australiensis	880971399	-RS-FN-EDQTA-	T-NIA-PA
	Collinsella aerofaciens	1176610803	-VA-TN-DD-LLI	YKG-IAPEL
ĺ	Alicyclobacillus macrosporangiidus	1124669995	AN-ED-QV-	HLN-FIPRL
	Caldicellulosiruptor acetigenus	1181399074	YN-KDLI	YKN-IPHI
	Firmicutes bacterium CAG:56	524306304	Y-SEYV	YKH-IPQVC-
Other Bastoria	Clostridium	493487041	Y-AEYVI	YKN-IPEIC-
Other Bacteria	Merdimonas faecis	1077993569	YTYVV	YKN-IPEVC-
	Blautia obeum	491570992	Y-SKYVV	YKN-IPQL
	Peptococcaceae bacterium BRH_c4b	780811628	ANDA-	YKN-FCSVL
l	Dielma fastidiosa	551318403	TYQ-AFY	N-Q-LMPEI

FIGURE 6 | CSI specific to Solirubrobacteraceae. A 3 amino acid CSI in the protein glutamine amidotransferase that is specific for S. soli and associated MAG.

(Table 1) and 29 CSPs (Table 3) that are only present in these two genomes but absent in other bacteria. Figure 3 shows one example of these CSIs. In the sequence alignment of arginine-tRNA ligase, a 7aa insertion flanked by highly conserved residues is uniquely found in homologs from both *T. album* and MAG "bacterium HR41." Sequence information for further 6 CSIs with the same specificity are shown in

**Supplementary Figures S13–S18**. Whether these identified CSIs and CSPs can constitute distinctive markers for the *Thermoleophiliaceae* family or even the *Thermoleophilales* order awaits confirmation from more sequences of other species belonging to this lineage. Nevertheless, these results provide additional evidence for the close relationship of MAG "bacterium HR41" and *T. album*.

specific for Patulibacteraceae

	·		376		420
Datulibaatamagaaa	Patulibacter medicamentivorans	494853285	GQKQQRGVKRLKVLSAF	YDPENAPY	YSQVNKPEMMVLDAVPVIPP
FaluiiDacleiaceae	Patulibacter americanus	551308296	A	P	
	Patulibacter minatonensis	652518781	A		-AI
	Solirubrobacter sp. URHD0082	739552113	V		LNSGI-E
	Solirubrobacter soli	921289100	V		LHSGI-E
	Solirubrobacterales bacterium 70-9	1113216523	A-AA		LHSE-RW-I
Other	Solirubrobacterales bacterium URHD0059	739542357	A-A		LHSDHE
Ocher	Solirubrobacterales bacterium 67-14	1113227231	A-AA		LKSEW-ICI
Thermoleophilia	Thermoleophilum album	1225103326	S-AV		LNSDY-I-E-IL
_	Conexibacter woesei	737138710	A-A		LHSDHE
	Actinobacteria bacterium RBG_16_68_12	1082278050	REIVN		IKSE-RWE
	Actinobacteria bacterium RBG_13_35_12	1082225837	KS-SIIS-		IDSDY-IIL
l	Actinobacteria bacterium 13_2_20CM_68_14	1125170944	V		VTSEWE
	Kribbia dieselivorans	1056988448	KT-ALVT		LTTDG
	Cutibacterium acnes	488503631	KTLVQ		LDSG-SAC
	Coriobacteriaceae bacterium EMTCatB1	1122514373	K-KAIV		RQSK-RW-I-E-I
	Knoellia sinensis	737949374	KT-ALV		MTTTAG
	Bombiscardovia coagulans	1233711393	RKA-ALVN		LTTEAI
Other	Propionicicella superfundia	655300311	RKT-ALV		LHTE-S-SG
Actinohactoria	Janibacter hoylei PVAS-1	404554692	KT-ALVT		LTTDAG
ACLINODACLEIIA	Luteococcus japonicus	1234381648	RKT-ALV		LDGR-S-AG
	Streptomyces yangpuensis	820828172	KT-ALV		LQTSKGC
	Kytococcus schroeteri	1325981133	KT-ALVT		LTTD-E-QG
	Granulicoccus phenolivorans	652534206	RKT-ALV		L-TQ-SGC
	Kitasatospora albolonga	1181453102	KT-ALV		LQTSKGC
	Anaerostipes hadrus	941901557	RA-IE-VE		LNSDW-I
	Clostridiales	495669326	RA-IE-VE		LNSDW-I
Other Bacteria	Eubacterium uniforme	1154858368	RS-IE-VE		RTSDW-IVI
Other Bacterra	Lachnospiraceae bacterium G11	1221739058	RA-VIE-VD		LESGW-IMI
	Firmicutes bacterium CAG:24	524040904	RA-IIE-VE		RESG-RW-I
FIGURE 7   CSI specific to Patuliba	cteraceae. Partial sequence alignment of DNA-direct	cted RNA poly	merase subunit beta' s	howing an	8 amino acid insertion that is

Within the order Solirubrobacterales, we have identified 6 CSIs that are specific to species of 3 families including Conexibacteraceae, Solirubacteraceae, and Patulibacteraceae, but no CSIs also shared by members of the new cluster (Table 1). One of these CSIs is illustrated in Figure 4, which is 1 aa deletion in a very conserved fragment of NADH-quinone oxidoreductase subunit B. Sequence information for other 5 CSIs that are uniquely shared by these 3 families are presented in Supplementary Figures S19-S23. Additionally, we discovered 24 CSPs that are only found in genomes of the named above 3 families but not in any other bacteria (Table 3). The shared presence of 6 CSIs and a number of CSPs indicate that Conexibacteraceae, Solirubacteraceae, and Patulibacteraceae are monophyletic. These two kinds of signature sequences were most likely introduced in the common ancestor of these three families and later on passed to all decedents. Moreover, if genome sequence of the fourth family Parviterribacteraceae becomes available in the future, it is worthwhile to examine whether some of these CSIs and CSPs are also shared by Parviterribacteraceae and actually constitute molecular markers of the Solirubrobacterales order.

As mentioned earlier, at family level within *Thermoleophilia*, only few cultivable strains are available and our current descriptions of some families such as *Conexibacteraceae* or *Solirubacteraceae* are only based on 1 or 2 strains. Here, we identified a number of CSIs that are specific to all genomesequenced members of each family of *Thermoleophilia* except *Parviterribacteraceae* that don't have genome sequence available (**Table 1**). For example, 4 CSIs were found to be unique to members of *Conexibacteraceae* (Figure 5 and Supplementary Figures S24–S26), 5 CSIs for *Solirubacteraceae* (Figure 6 and Supplementary Figures S27–S30), and totally 10 CSIs shared by 3 species of *Patulibacteraceae* (Figure 7 and Supplementary Figures S31-S39). We attempted to search for CSIs that are specific to the new cluster revealed by our phylogenomic analysis. Due to the incompleteness of the 3 genome assemblies, only 1 CSI is specifically shared by all three members of the new cluster (Supplementary Figure S40), while another 3 CSIs are only found in MAG "Solirubrobacterales bacterium 67-14" and "Solirubrobacterales bacterium 70-9" with two protein homologs missing in "Actinobacteria bacterium 13\_1\_20CM\_3\_68\_9" (Supplementary Figures S41-S43). Furthermore, since more genomes are sequenced for *Patulibacteraceae*, we also identified 31 CSPs that are restricted to the genomes of this family, which provide additional markers for them (Table 3).

## CONCLUSION

Although metagenomic studies suggest that species of the class *Thermoleophilia* are abundant in hot spring and soil samples and they play an important role in biogeochemical cycling, very few studies have been performed on the phylogeny of this deep branch of *Actinobacteria*. Our current understanding of their taxonomy and phylogeny based on few cultivated species needs to be updated to better serve our exploration of this class. In this work, we have carried out detailed phylogenomic analysis of sequenced *Thermoleophilia* species and assembled genomes. The constructed phylogenetic tree clearly demonstrated the close affiliation of not yet cultivated MAGs with culturable type species. A new robust cluster composed of not yet cultivated MAGs is revealed within this class that might be a novel family belonging to *Solirubrobacterales*. Moreover, we identified a large number

of CSIs and CSPs that are either specific to all species of this class or various lineages within it. These two types of signature sequences provide novel molecular markers that can be applied to define or distinguish the class *Thermoleophilia* or its affiliated taxa at higher taxonomic ranks, in addition to the 16S rRNA gene alone based standard.

In addition to their phylogenetic implications, these lineagespecific CSIs and CSPs can also be used to test the presence of *Thermoleophilia* species in different environmental samples. PCR primers could be designed for gene fragments that contain the above described CSIs or genes for CSPs, then we can detect the existence of certain lineages based on the presence or absence of the CSIs and CSPs. Furthermore, the functional significance of all CSIs and CSPs identified from this work are unknown. Due to their specificity to the *Thermoleophilia* class, functional studies on them might lead to identification of biochemical or physiological characteristics that are unique to this class of bacteria.

### **AUTHOR CONTRIBUTIONS**

DH carried out comparative analyses of the *Thermoleophilia* genomes to identify signatures reported here and constructed the phylogenetic trees. BG, DH, YZ and YM were responsible for the

### REFERENCES

- Albuquerque, L., Franca, L., Rainey, F. A., Schumann, P., Nobre, M. F., and Da Costa, M. S. (2011). *Gaiella occulta* gen. nov., sp. nov., a novel representative of a deep branching phylogenetic lineage within the class actinobacteria and proposal of *Gaiellaceae* fam. nov. and *Gaiellales* ord. nov. *Syst. Appl. Microbiol.* 34, 595–599. doi: 10.1016/j.syapm.2011.07.001
- Almeida, B., Kjeldal, H., Lolas, I., Knudsen, A. D., Carvalho, G., Nielsen, K. L., et al. (2013). Quantitative proteomic analysis of ibuprofen-degrading *Patulibacter* sp. strain 111. *Biodegradation* 24, 615–630. doi: 10.1007/s10532-012-9610-5
- Alnajar, S., and Gupta, R. S. (2017). Phylogenomics and comparative genomic studies delineate six main clades within the family *Enterobacteriaceae* and support the reclassification of several polyphyletic members of the family. *Infect. Genet. Evol.* 54, 108–127. doi: 10.1016/j.meegid.2017.06.024
- Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402. doi: 10.1093/nar/25.17. 3389
- Butterfield, C. N., Li, Z., Andeer, P. F., Spaulding, S., Thomas, B. C., Singh, A., et al. (2016). Proteogenomic analyses indicate bacterial methylotrophy and archaeal heterotrophy are prevalent below the grass root zone. *PeerJ* 4:e2687. doi: 10.7717/peerj.2687
- Cabello-Yeves, P. J., Zemskaya, T. I., Rosselli, R., Coutinho, F. H., Zakharenko, A. S., Blinov, V. V., et al. (2018). Genomes of novel microbial lineages assembled from the sub-ice waters of lake baikal. *Appl. Environ. Microbiol.* 84:e2132-17. doi: 10.1128/AEM.02132-17
- Cole, J. R., Wang, Q., Fish, J. A., Chai, B., Mcgarrell, D. M., Sun, Y., et al. (2014). Ribosomal database project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res.* 42, D633–D642. doi: 10.1093/nar/gkt1244
- Foesel, B. U., Geppert, A., Rohde, M., and Overmann, J. (2016). Parviterribacter kavangonensis gen. nov., sp. nov. and Parviterribacter multiflagellatus sp. nov., novel members of parviterribacteraceae fam. nov. within the order solirubrobacterales, and emended descriptions of the classes thermoleophilia and rubrobacteria and their orders and families. Int. J. Syst. Evol. Microbiol. 66, 652–665. doi: 10.1099/ijsem.0.000770

writing and editing of the manuscript. All of the work was carried out under the direction of BG.

### FUNDING

This work was supported by National Science Foundation of China (31570011), Strategic Priority Research Program of the Chinese Academy of Sciences (XDA13020300 and XDA19060301), and Natural Science Foundation of Guangdong Province (2015A030306039). BG was also a scholar of the "100 Talents Project" of the Chinese Academy of Sciences.

### ACKNOWLEDGMENTS

We want to thank Dr. Radhey S. Gupta from McMaster University for generously providing the "SIG\_CREATE" and "SIG\_STYLE" programs.

### SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb. 2019.01185/full#supplementary-material

- Gao, B., and Gupta, R. S. (2012a). Microbial systematics in the post-genomics era. *Antonie Van Leeuwenhoek* 101, 45–54. doi: 10.1007/s10482-011-9663-1
- Gao, B., and Gupta, R. S. (2012b). Phylogenetic framework and molecular signatures for the main clades of the phylum actinobacteria. *Microbiol. Mol. Biol. Rev.* 76, 66–112. doi: 10.1128/MMBR.05011-11
- Gao, B., Mohan, R., and Gupta, R. S. (2009). Phylogenomics and protein signatures elucidating the evolutionary relationships among the gammaproteobacteria. *Int. J. Syst. Evol. Microbiol.* 59(Pt 2), 234–247. doi: 10.1099/ijs.0.002741-0
- Gao, B., Paramanathan, R., and Gupta, R. S. (2006). Signature proteins that are distinctive characteristics of actinobacteria and their subgroups. *Antonie Van Leeuwenhoek* 90, 69–91. doi: 10.1007/s10482-006-9061-2
- Gupta, R. S. (2014). "Identification of conserved indels that are useful for classification and evolutionary studies," in *Methods in Microbiology*, eds M. Goodfellow, I. Sutcliffe, and J. Chun (Oxford: Academic Press), 153–182. doi: 10.1016/bs.mim.2014.05.003
- Gupta, R. S., and Gao, B. (2009). Phylogenomic analyses of clostridia and identification of novel protein signatures that are specific to the genus clostridium sensu stricto (cluster I). *Int. J. Syst. Evol. Microbiol.* 59, 285–294. doi: 10.1099/ijs.0.001792-0
- Gupta, R. S., and Gao, B. (2010). "Recent advances in understanding microbial systematics," in *Microbial Population Genetics*, ed. J. P. Xu (Norfolk: Caister Academic Press).
- Ho, J., Adeolu, M., Khadka, B., and Gupta, R. S. (2016). Identification of distinctive molecular traits that are characteristic of the phylum "Deinococcus-Thermus" and distinguish its main constituent groups. *Syst. Appl. Microbiol.* 39, 453–463. doi: 10.1016/j.syapm.2016.07.003
- Hu, D., Cha, G., and Gao, B. (2018). A phylogenomic and molecular markers based analysis of the class acidimicrobiia. *Front. Microbiol.* 9:987. doi: 10.3389/fmicb. 2018.00987
- Janssen, P. H. (2006). Identifying the dominant soil bacterial taxa in libraries of 16S rRNA and 16S rRNA genes. *Appl. Environ. Microbiol.* 72, 1719–1728. doi: 10.1128/aem.72.3.1719-1728.2006
- Ji, M., Greening, C., Vanwonterghem, I., Carere, C. R., Bay, S. K., Steen, J. A., et al. (2017). Atmospheric trace gases support primary production in Antarctic desert surface soil. *Nature* 552, 400–403. doi: 10.1038/nature25014

- Ji, M., Van Dorst, J., Bissett, A., Brown, M. V., Palmer, A. S., Snape, I., et al. (2016). Microbial diversity at mitchell peninsula, eastern antarctica: a potential biodiversity "hotspot". *Polar Biol.* 39:13.
- Joseph, S. J., Hugenholtz, P., Sangwan, P., Osborne, C. A., and Janssen, P. H. (2003). Laboratory cultivation of widespread and previously uncultured soil bacteria. *Appl. Environ. Microbiol.* 69, 7210–7215. doi: 10.1128/aem.69.12.7210-7215. 2003
- Kantor, R. S., Van Zyl, A. W., Van Hille, R. P., Thomas, B. C., Harrison, S. T., and Banfield, J. F. (2015). Bioreactor microbial ecosystems for thiocyanate and cyanide degradation unravelled with genome-resolved metagenomics. *Environ. Microbiol.* 17, 4929–4941. doi: 10.1111/1462-2920.12936
- Kato, S., Sakai, S., Hirai, M., Tasumi, E., Nishizawa, M., Suzuki, K., et al. (2018). Long-term cultivation and metagenomics reveal ecophysiology of previously uncultivated thermophiles involved in biogeochemical nitrogen cycle. *Microbes Environ.* 33, 107–110. doi: 10.1264/jsme2.ME17165
- Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., Mcgettigan, P. A., Mcwilliam, H., et al. (2007). Clustal W and clustal X version 2.0. *Bioinformatics* 23, 2947–2948. doi: 10.1093/bioinformatics/btm404
- Li, H. Y., Wang, H., Wang, H. T., Xin, P. Y., Xu, X. H., Ma, Y., et al. (2018). The chemodiversity of paddy soil dissolved organic matter correlates with microbial community at continental scales. *Microbiome* 6:187. doi: 10.1186/s40168-018-0561-x
- Ludwig, W., Euzeby, J., Schumann, P., Busse, H., Trujillo, M. E., Kampfer, P., et al. (2012). "Road map of the phylum actinobacteria," in *Bergey's Manual of Systematic Bacteriology*, 2 Edn, Vol. 5, eds M. Goodfellow, P. Kampfer, H. J. Busse, M. E. Trujillo, K. Suzuki, W. Ludwig, et al. (New York, NY: Springer), 1–28. doi: 10.1007/978-0-387-68233-4\_1
- Mukherjee, S., Seshadri, R., Varghese, N. J., Eloe-Fadrosh, E. A., Meier-Kolthoff, J. P., Goker, M., et al. (2017). 1,003 reference genomes of bacterial and archaeal isolates expand coverage of the tree of life. *Nat. Biotechnol.* 35, 676–683. doi: 10.1038/nbt.3886
- Na, S. I., Kim, Y. O., Yoon, S. H., Ha, S. M., Baek, I., and Chun, J. (2018). UBCG: up-to-date bacterial core gene set and pipeline for phylogenomic tree reconstruction. *J. Microbiol.* 56, 280–285. doi: 10.1007/s12275-018-8014-6
- Nouioui, I., Carro, L., Garcia-Lopez, M., Meier-Kolthoff, J. P., Woyke, T., Kyrpides, N. C., et al. (2018). Genome-based taxonomic classification of the phylum actinobacteria. *Front. Microbiol.* 9:2007. doi: 10.3389/fmicb.2018.02007
- Ollagnier-De Choudens, S., Loiseau, L., Sanakis, Y., Barras, F., and Fontecave, M. (2005). Quinolinate synthetase, an iron-sulfur enzyme in NAD biosynthesis. *FEBS Lett.* 579, 3737–3743. doi: 10.1016/j.febslet.2005.05.065
- Parks, D. H., Rinke, C., Chuvochina, M., Chaumeil, P. A., Woodcroft, B. J., Evans, P. N., et al. (2017). Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.* 2, 1533–1542. doi: 10.1038/s41564-017-0012-7
- Pukall, R., Lapidus, A., Glavina Del Rio, T., Copeland, A., Tice, H., Cheng, J. F., et al. (2010). Complete genome sequence of conexibacter woesei type strain (ID131577). Stand. Genomic Sci. 2, 212–219. doi: 10.4056/sigs.751339
- Pulschen, A. A., Bendia, A. G., Fricker, A. D., Pellizari, V. H., Galante, D., and Rodrigues, F. (2017). Isolation of uncultured bacteria from antarctica using long incubation periods and low nutritional media. *Front. Microbiol.* 8:1346. doi: 10.3389/fmicb.2017.01346
- Reddy, G. S., and Garcia-Pichel, F. (2009). Description of *Patulibacter americanus* sp. nov., isolated from biological soil crusts, emended description of the genus patulibacter takahashi et al. 2006 and proposal of solirubrobacterales ord. nov.

and thermoleophilales ord. nov. Int. J. Syst. Evol. Microbiol. 59, 87–94. doi: 10.1099/ijs.0.64185-0

- Salam, N., Jiao, J., Zhang, X., and Li, W. (2019). Update in the classification of higher ranks in the phylum actinobacteria. *Int. J. Syst. Evol. Microbiol.* (in press).
- Suzuki, K., and Whitman, W. B. (2012). "Class VI. thermoleophilia class. nov," in Bergey's Manual of Systematic Bacteriology, 2nd Edn, eds M. Goodfellow, P. Kampfer, H. J. Busse, M. E. Trujillo, K. Suzuki, W. Ludwig, et al. (New York, NY: Springer), 2010–2028.
- Suzuki, K., and Whitman, W. B. (2015). "Thermoleophilia class. nov," in *Bergey's Manual of Systematics of Archaea and Bacteria*, eds W. B. Whitman, F. A. Rainey, P. Kampfer, M. Trujillo, J. Chun, P. Devos, et al. (Hoboken, NY: John Wiley & Sons, Inc.).
- Talavera, G., and Castresana, J. (2007). Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst. Biol.* 56, 564–577. doi: 10.1080/10635150701472164
- Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* 30, 2725– 2729. doi: 10.1093/molbev/mst197
- Volbeda, A., Darnault, C., Renoux, O., Reichmann, D., Amara, P., Ollagnier De Choudens, S., et al. (2016). Crystal structures of quinolinate synthase in complex with a substrate analogue, the condensation intermediate, and substrate-derived product. J. Am. Chem. Soc. 138, 11802–11809. doi: 10.1021/ jacs.6b05884
- Woodcroft, B. J., Singleton, C. M., Boyd, J. A., Evans, P. N., Emerson, J. B., Zayed, A. A. F., et al. (2018). Genome-centric view of carbon processing in thawing permafrost. *Nature* 560, 49–54. doi: 10.1038/s41586-018-0338-1
- Yakimov, M. M., Lunsdorf, H., and Golyshin, P. N. (2003). Thermoleophilum album and thermoleophilum minutum are culturable representatives of group 2 of the rubrobacteridae (actinobacteria). *Int. J. Syst. Evol. Microbiol.* 53, 377–380. doi: 10.1099/ijs.0.02425-0
- Zarilla, K. A., and Perry, J. J. (1986). Deoxyribonucleic-acid homology and other comparisons among obligately thermophilic hydrocarbonoclastic bacteria, with a proposal for thermoleophilum-minutum Sp-Nov. Int. J. Syst. Evol. Microbiol. 36, 13–16. doi: 10.1099/00207713-36-1-13
- Zhang, G., Gao, B., Adeolu, M., Khadka, B., and Gupta, R. S. (2016). Phylogenomic analyses and comparative studies on genomes of the bifidobacteriales: identification of molecular signatures specific for the order bifidobacteriales and its different subclades. *Front. Microbiol.* 7:978. doi: 10.3389/fmicb.2016.00978
- Zhi, X. Y., Li, W. J., and Stackebrandt, E. (2009). An update of the structure and 16S rRNA gene sequence-based definition of higher ranks of the class actinobacteria, with the proposal of two new suborders and four new families and emended descriptions of the existing higher taxa. *Int. J. Syst. Evol. Microbiol.* 59, 589–608. doi: 10.1099/ijs.0.65780-0

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Hu, Zang, Mao and Gao. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.