1 **Salvaging complete and high-quality genomes of novel microbial species from a**
2 **meromictic lake using a workflow combining long- and short-read sequencing**
3 **platforms**

4 Yu-Hsiang Chen^{1,2,3}, Pei-Wen Chiang³, Denis Yu Rogozin^{4,5}, Andrey G.

5 Degermendzhy⁴, Hsiu-Hui Chiu³, Sen-Lin Tang^{2,3#}

6

7 ¹Bioinformatics Program, Taiwan International Graduate Program, National Taiwan
8 University, Taipei, Taiwan

9 ²Bioinformatics Program, Institute of Information Science, Taiwan International
10 Graduate Program, Academia Sinica, Taipei, Taiwan

11 ³Biodiversity Research Center, Academia Sinica, Taipei, Taiwan

12 ⁴Institute of Biophysics, Siberian Division of Russian Academy of Sciences,
13 Krasnoyarsk, Russia

14 ⁵Siberian Federal University, Krasnoyarsk, Russia

15

16 # Corresponding author

17 Biodiversity Research Center, Academia Sinica

18 E-mail: sltang@gate.sinica.edu.tw

19 Tel: +886-2-27893863

20 Fax: +886-2-27890844

21 **Abstract**

22 **Background**

23 Most of Earth's bacteria have yet to be cultivated. The metabolic and functional
 24 potentials of these uncultivated microorganisms thus remain mysterious, and the
 25 metagenome-assembled genome (MAG) approach is the most robust method for
 26 uncovering these potentials. However, MAGs discovered by conventional
 27 metagenomic assembly and binning methods are usually highly fragmented genomes
 28 with heterogeneous sequence contamination, and this affects the accuracy and
 29 sensitivity of genomic analyses. Though the maturation of long-read sequencing
 30 technologies provides a good opportunity to fix the problem of highly fragmented
 31 MAGs as mentioned above, the method's error-prone nature causes severe problems
 32 of long-read-alone metagenomics. Hence, methods are urgently needed to retrieve
 33 MAGs by a combination of both long- and short-read technologies to advance
 34 genome-centric metagenomics.

35 **Results**

36 In this study, we combined Illumina and Nanopore data to develop a new workflow to
 37 reconstruct 233 MAGs—six novel bacterial orders, 20 families, 66 genera, and 154
 38 species—from Lake Shunet, a secluded meromictic lake in Siberia. Those new MAGs

39 were underrepresented or undetectable in other MAGs studies using metagenomes
 40 from human or other common organisms or habitats. Using this newly developed
 41 workflow and strategy, the average N50 of reconstructed MAGs greatly increased
 42 10–40-fold compared to when the conventional Illumina assembly and binning
 43 method were used. More importantly, six complete MAGs were recovered from our
 44 datasets, five of which belong to novel species. We used these as examples to
 45 demonstrate many novel and intriguing genomic characteristics discovered in these
 46 newly complete genomes and proved the importance of high-quality complete MAGs
 47 in microbial genomics and metagenomics studies.

48 **Conclusions**

49 The results show that it is feasible to apply our workflow with a few additional long
 50 reads to recover numerous complete and high-quality MAGs from short-read
 51 metagenomes of high microbial diversity environment samples. The unique features
 52 we identified from five complete genomes highlight the robustness of this method in
 53 genome-centric metagenomic research. The recovery of 154 novel species MAGs
 54 from a rarely explored lake greatly expands the current bacterial genome
 55 encyclopedia and broadens our knowledge by adding new genomic characteristics of
 56 bacteria. It demonstrates a strong need to recover MAGs from diverse unexplored
 57 habitats in the search for microbial dark matter.

58 **Keywords:** Metagenome-assembled genome, metagenomics, saline lake,
59 microbiology, and Nanopore sequencing.

60

61 **Background**

62 Rapid developments in bioinformatics and sequencing methods enable us to
63 reconstruct genomes directly from environmental samples using a culture-independent
64 whole-genome-shotgun metagenomic approach. These genomes, also called
65 metagenome-assembled genomes (MAGs), have become a crucial information source
66 to explore metabolic and functional potentials of uncultivated microorganisms [1-4].
67 Mining MAGs quickly expands our knowledge of microbial genome, diversity,
68 phylogeny, evolution, and taxonomy [1-4]. For example, 18,365 MAGs were
69 identified out of a total of 410,784 microorganisms in the Genomes OnLine Database
70 (GOLD) [5]. A total of 52,515 MAGs were assembled from diverse habitats, and the
71 MAG collection contains 12,556 potentially novel species and expands the known
72 phylogenetic diversity in bacterial and archaeal domains by 44% [3].

73 Although genome-resolved metagenomics has revolutionized research in
74 microbiology, significant challenges need to be overcome to make MAGs more
75 accurate, reliable, and informative [1]. First, most MAGs are derived from the
76 metagenomic assembly of short reads [1, 6], and these short-read-derived MAGs

usually comprise numerous short contigs rather than complete or nearly-complete genomic sequences, and thus important information on genomic characters is missed, such as operons, gene order, gene synteny, and promoter/regulatory regions. As of March 2021, only 177 out of 84,768 MAGs released in NCBI were complete. Second, fragmented MAGs usually miss some gene sequences and comprise unknown contaminant sequences, mistakenly assembled into the contigs [1]. Hence, low contiguity, high fragmentation, and unwanted contamination in short-read MAGs greatly affect further analyses in a variety of microbial genome-related studies.

The emergence of long-read sequencing platforms (also called third-generation sequencing platforms) such as Nanopore and PacBio provides an opportunity to improve the contiguity of MAGs and even reconstruct complete MAGs from extremely complex microbial communities [7, 8]. Recently, researchers started to develop new assemblers to reconstruct microbial genomes with high accuracy and long contiguous fragments from long-read metagenomic datasets. In 2019, Nagarajan *et al.* developed a hybrid assembler called OPERA-MS [9]. The assembler yielded MAGs with 200 times higher contiguity than short-read assemblers used on human gut microbiomes. In October 2020, Pevzner *et al.* developed metaFlye, a long-read metagenome assembler that can produce highly accurate assemblies (approximately 99% assembly accuracy) [10, 11]. The success of these newly developed assemblers

96 becomes an important stepping-stone for reconstruction of complete MAGs with high
 97 accuracy. However, there is still much room to improve the procedures around data
 98 processing and assembling MAGs with long reads. The current study presents a new
 99 workflow for this purpose.

100 Our workflow combines Illumina sequencing reads and Nanopore long
 101 sequencing reads to recover many novel high-quality and high-contiguity prokaryotic
 102 MAGs from Lake Shunet, southern Siberia, one of only four meromictic lakes in all
 103 of Siberia. The lake contains stratified water layers, including a mixolimnion layer at
 104 3.0 m, chemocline at 5.0 m, and monimolimnion at 5.5 m. From our previous 16
 105 rRNA amplicon survey, we know that the lake comprises at least hundreds of
 106 unknown bacteria and archaea [12], highlighting the importance of mining microbial
 107 MAGs from this rarely explored lake. However, though we attempted to recover
 108 MAGs from these layers using deep Illumina sequencing with approximately 150 Gb,
 109 only one high-quality but still fragmented MAG was obtained [12]. Hence, in this
 110 study, we developed a new workflow combining Illumina and Nanopore sequencing
 111 reads by integrating several cutting-edge bioinformatics tools to recover and
 112 reconstruct MAGs with high contiguity and accuracy. We demonstrate that our newly
 113 built workflow can be used to reconstruct hundreds of complete high-quality MAGs
 114 from environmental samples in a high-complexity microbial community.

Results and Discussions

Reconstruction of metagenome-assembled genomes with high contiguity from

Lake Shunet by combining Nanopore and Illumina sequences

To recover novel metagenome-assembled genomes (MAGs) with high contiguity without compromising accuracy, 3.0-, 5.0-, and 5.5-m deep Lake Shunet samples were sequenced by Nanopore machines individually, and the resulting long reads (LRs) were analyzed together with short reads (SRs) using a workflow we developed for this study (Fig. 1a). Originally, we only used metaFlye, a state-of-art long-read metagenome assembler that can provide 99% accuracy [10, 11], to assemble the LRs. However, recent studies found that assemblies from long reads contain numerous in-del errors, leading to erroneous predictions of open reading frames and biosynthetic gene clusters [1, 10]. Incorrectly predicting open reading frames also affects the estimation of genome completeness by single-copy marker gene method, such as checkM [13]. Hence, we used SRs from Illumina sequencing to correct the contigs generated by LRs.

To recover more MAGs and improve contiguity, the assemblies from SRs and LRs were combined before binning. The contiguity of MAGs generated by combining two sequencing reads was dramatically higher than that from the Illumina assembly alone. The average N50 of MAGs from SRs only were 12.4 kb, 6.0 kb, and 7.2 kb in

the 3.0, 5.0, and 5.5-m dataset, respectively. Average N50 increased to 476.5 kb, 269.5 kb, and 91.2 kb (Fig. 1b), respectively, when assembling with a combination of the two sequencing methods. A previous study showed that the qualities of MAGs can be improved by reassembly [14], so the step was incorporated into our workflow. When the MAGs were reassembled and selected, the average N50 increased from 476.5 kb to 530.0 kb in the 3.0 m dataset and 91.2 kb to 107.3 kb in the 5.5 m datasets (Fig. 1b).

The correlations between read coverages and contiguity were determined (Fig. 1c, d). The results revealed that the N50 values were more correlated with the Nanopore read coverage (Spearman's $r=0.7$) than the Illumina coverage (Spearman's $r=0.33$). This is consistent with the previous observation that contiguity plateaued when the coverage of SRs reached a certain point because the assembly of SRs cannot solve repetitive sequences [9]. Nevertheless, LR can address the issue by spanning repetitive regions. We also found that using SR assembly only, we cannot obtain MAGs with $N50 > 100\text{kbp}$. By comparison, using our workflow, we can obtain 73 MAGs with $N50 > 100\text{kbp}$. The mean SR coverage of these MAGs was 187 times, and mean LR coverage of them was only 67. Additionally, our data size of LRs is about 1/3 that of SRs. Taken together, it represents that the contiguity of MAGs can be greatly improved with one-third LRs. The results highlighted that 1) combining

153 two sequencing methods yield significant improvements in the qualities of MAGs that
154 are recovered from high-complexity metagenomic datasets, and 2) With only extra
155 one-third LR, we could retrieve genome information, such gene order, from previous
156 SR-derived MAG collections.

157 Using our workflow, a total of 233 MAGs with completeness > 50% and
158 contamination < 10% were reconstructed. For Genome Taxonomy Database (GTDB)
159 species representatives, the genome quality index, defined as
160 $\text{completeness} - 5 \times \text{contamination}$, should be larger than 50. To meet the GTDB
161 standard, the MAGs were filtered by this criterion, and the MAGs with low SR
162 coverages (<80%) were discarded, resulting in 187 MAGs (Dataset S1). All the
163 MAGs satisfied or surpassed the MIMAG standard for a medium-quality draft [15].
164 The median completeness of MAGs was 81%, and the median contamination was
165 1.1% (Fig. 1e). Moreover, 45.3% of the MAGs contained 16S rRNA gene sequences,
166 and 34.5% of MAGs had 23S, 16S, and 5S rRNA gene sequences (Fig. 1f). The
167 median GC ratio of MAGs from 3.0, 5.0, and 5.5 m were 52.75, 44.1, and 46.4%,
168 respectively (Fig. 1g). We also used OPERA-MS to retrieve MAGs from SRs and LR.
169 However, only 26 medium-quality or high-quality MAGs were recovered, indicating
170 that the method is suboptimal in our case.

171

172 **Phylogenetic diversity and novelty of MAGs**

173 To explore the diversity of MAGs, we clustered and de-duplicated the genomes
174 based on a 95% ANI cutoff for bacterial species demarcation [16], since identical
175 microbial species may be detected and assembled from the three different layers. The
176 procedure led to 165 species-level non-redundant MAGs (Dataset S1). The majority
177 (93%) of the species-level MAGs could not be assigned to any known species after
178 taxonomy annotation by the GTDB-Tk, revealing that a great deal of novel MAGs at
179 the species and higher taxonomic ranks were detected (Dataset S2). The novel MAGs
180 comprised six unknown bacterial orders, 20 families, and 66 genera (Fig. 2a).

181 To examine the phylogenetic diversity in the novel MAGs, a phylogenomic tree
182 was reconstructed using all these bacterial MAGs and representative bacterial
183 genomes in GTDB (Fig. 2b). The result demonstrated that the MAGs widely span the
184 bacterial phylogeny. The MAGs were distributed across 24 phyla, including unusual
185 and poorly-characterized phyla, such as *Armatimonadota*, *Margulisbacteria*,
186 *Bipolaricaulota*, *Cloacimonadota*, and *Caldatribacteriota*. The phylum frequencies
187 differed between the genome collections of the standard database and the Shunet
188 datasets (Fig. 2c). The GTDB mainly comprised *Proteobacteria*. In contrast, in
189 genome collections from the Shunet datasets, the phylum frequency was enriched in
190 the *Desulfobacterota*, *Verrucomicrobiota*, *Bacteroidota*, and *Omnitrophota*. The

191 difference can also be seen by comparing Shunet datasets with a genomic catalog of
192 Earth's microbiomes and 8,000 MAGs recovered by Tyson *et al.* [17, 18], which also
193 had a higher proportion of *Proteobacteria*, but limited in the other four phyla enriched
194 in Lake Shunet. The results suggest that, to gain a comprehensive picture of the
195 microbial genomes on earth, there is a strong need for future studies to explore
196 microbiomes from various habitats, especially overlooked or understudied habitats
197 [17, 19].

198

199 **Novel predicted secondary-metabolite biosynthetic clusters and** 200 **carbohydrate-active enzymes from newly recovered MAGs**

201 Here we demonstrate a) the value of recovering MAGs from rarely investigated
202 habitats to mine novel microbial function potentials and b) the advantage of
203 combining SRs and LRIs using two examples: secondary metabolite biosynthetic gene
204 clusters (BGCs) and carbohydrate-active enzymes (CAZymes). Secondary
205 metabolites are usually unique in one or a few species, and not related to the normal
206 growth of the organisms [20]. The secondary metabolites, associated with ecological
207 interactions, can serve as toxins, factors participating in symbiosis with other hosts,
208 defense mechanisms [20, 21]. Identifying novel secondary metabolites enables us to
209 understand the ecological interactions among the microbes. The majority of bacteria

210 remain uncultivated, so mining novel BGCs in metagenomes provides the opportunity
211 to discover new secondary metabolites [22, 23].

212 In our MAG collection, we identified 414 putative BGCs from 140 MAGs (Fig.
213 S1a). Among them, 134 BGCs were annotated as terpenes and 64 BGCs as
214 bacteriocins. To determine the novelty of these BGCs, the BGCs were searched
215 against the NCBI database using the cutoffs of 75% identity and 80% query coverage
216 based on a previous study [17]. The results demonstrated that 384 BGCs (92%) could
217 not be matched with these thresholds, indicating that most of these could be novel
218 BGCs. Comparably, only 83% of BGCs were predicted to be novel BGCs in the
219 recently-published Genomes from Earth's Microbiome catalog (GEM) [17].

220 Complete BGCs are important because they help us identify the metabolites that
221 these BGCs produce using molecular approaches [21]. 72% of BGCs identified from
222 MAGs in the 3.0 m dataset were not on the edge of the contigs, suggesting that the
223 majority of BGCs may be complete. However, only 22% of BGCs in the 5.0 and 5.5
224 m datasets were not on the edges, which could be because the MAGs from 3.0 m were
225 more contiguous because they had a 10-fold larger median N50 (Fig. 1b). In total, 213
226 BGCs (51%) we recovered were not on the edges. By comparison, only 34% BGCs
227 predicted in the GEM MAG collection were not on the edge. In the 414 BGCs, 552
228 core biosynthetic genes, 1,224 additional biosynthetic genes, 205 regulator genes, and

185 transporter genes were identified. This information will enable us to examine the products of the BGC, the function of these genes, and the roles of the products in the individual bacterium. On the other hand, the results also showed that the increased contiguity of MAGs by LRs enables us to obtain more complete BGCs.

Carbohydrate-active enzymes have a range of applications. For instance, CAZymes are used for food processing and food production [24-27]. Exploring novel CAZymes in the metagenome can benefit food industries [24, 25]. On the other hand, identifying novel CAZymes modules enables us to produce novel bioactive oligosaccharides that can be used to develop new drugs and supplements [26, 27]. From the MAGs reconstructed in this study, we identified 8,750 putative CAZymes: 3,918 glycosyltransferases, 3,304 glycoside hydrolases, 738 carbohydrate esterases, and 92 polysaccharide lyases (Fig. S1b). Previous studies indicated that 60~70% protein identity can be used as a threshold for the conservation of the enzymatic function [28-30]. Among the CAZymes we identified, 1,745 (44%) glycosyltransferases, 1,456 (44%) glycoside hydrolases, 267 (36%) carbohydrate esterases, and 57 (62%) polysaccharide lyases shared less than 60% protein identity with their closest homologs in the NCBI nr database (Fig. S1c). This indicates that these CAZymes could have novel carbohydrate-active functions, which future research efforts can explore further.

248

249 **Novel candidate archaeal families identified from Lake Shunet**

250 From the 5.5 m dataset, we identified two MAGs belonging to candidate novel
 251 families under *Methanomassiliicoccales* and *Iainarchaeales* (MAG ID: M55A1 and
 252 M55A2, respectively) and one MAG belonging to a potential novel species under
 253 *Nanoarchaeota*, according to the GTDB taxonomy assignment based on the
 254 phylogenomic tree and relative evolutionary divergence (RED) (Dataset S2). In the
 255 archaeal phylogenomic tree, M55A1 formed a clade basal to the clade containing
 256 species within the *Methanomethylophilaceae* family, a group of host-associated
 257 methanogens, and the branch was supported by a 94.7% UFBoot value (probability
 258 that a clade is true) [31]. The M55A1 and *Methanomethylophilaceae*-related clade
 259 formed a superclade that is adjacent to *Methanomassiliicoccaceae*-related clade, a
 260 group of environmental methanogens [32]. These clades formed the order
 261 *Methanomassiliicoccales*, the hallmark of which is the ability to produce methane.
 262 However, M55A1 did not contain genes encoding for a methane-producing key
 263 enzyme complex (Fig. S2). For example, genes encoding methyl-coenzyme M
 264 reductase alpha (*mcrA*), beta (*mcrB*), and gamma subunit (*mcrG*), a key enzyme
 265 complex involved in methane production, were absent in the M55A1. On the other
 266 hand, we did not find *Methanomassiliicoccaceae*-related *mcrA*, *mcrB*, or *mcrG* genes

267 in the other bins and unbinned sequences in the 5.5 m dataset. Furthermore, M55A1
268 lacks most of the core methanogenesis marker genes identified in
269 *Methanomassiliicoccales*.

270 The absence of these methanogenesis marker genes implies that the archaea may
271 have lost their methane-producing ability. If this is true, then a phylogenetic group of
272 *Methanomassiliicoccales* may have lost the ability to perform methanogenesis after its
273 ancestor evolved the ability to produce methane. The results not only showed the
274 potential functional diversity in this clade but also highlighted how much such a
275 little-studied environment can reveal about functional diversity in known microbial
276 lineages.

277

278 **Five complete MAGs of a candidate novel genus and species from Lake Shunet**

279 The assemblies of Shunet datasets yielded six complete and circulated bacterial
280 genomes. Among these six complete MAG, two belonged to a novel *Simkaniaceae*
281 genus, and there were classified as novel *Cyanobium* species, *Thiocapsa* species, and
282 species under GCA-2401735 (an uncharacterized genus defined previously based on
283 phylogeny), according to the GTDB taxonomy inference based on ANI and
284 phylogenomic analyses (Dataset S1 and S2). The following are individual
285 descriptions of their unique taxonomic and metabolic features. The nitrogen, carbon,

286 sulfur, and energy metabolisms are described in Figure S3.

287 ***Candidate novel Simkaniaceae genus.*** According to GTDB-TK, there were two
288 complete MAGs—M30B1 and M30B2—assigned as an unclassified genus under
289 *Simkaniaceae*, a family in the class *Chlamydiia*, based on the topology of the
290 phylogenetic tree. M30B1 and M30B2 formed a monophyletic group and shared
291 72.48% percentage of conserved protein (PCOP), above the genus boundary of 50%
292 PCOP [33]. The genomes shared 77% ANI, below the 95% cutoff for the same species
293 [16], and the identity of their rRNA gene sequences was 98.45%. Together, the results
294 showed that the two MAGs were two different new species under a novel genus.
295 Therefore, we propose a new genus, *Candidatus Andegerimia*, to include the two
296 MAGs, and renamed the two MAGs as *Candidatus Andegerimia shunetia* M30B1 and
297 *Candidatus Andegerimia siberian* M30B2, abbreviated as M30B1 and M30B2,
298 respectively.

299 *Simkaniaceae*, like all *Chlamydia*, are obligately intracellular bacteria that live in
300 eukaryotic cells [34]. Validated natural hosts include various multicellular eukaryotic
301 organisms like vertebrates. That some *Simkaniaceae* PCR clones were identified from
302 drinking water implies that *Simkaniaceae* may also live in unicellular eukaryotes [35].
303 Our samples were collected from saline water, and a membrane was used to filter
304 large organisms. Hence, *Ca. A. shunetia* and *Ca. A. siberian* may be derived from tiny

305 or unicellular eukaryotic organisms.

306 The reconstruction of complete MAGs enables us to compare genomes in a
307 precise and comprehensive manner by avoiding contamination caused by binning.
308 The two *Simkaniaceae* MAGs we recovered contained five KEGG orthologues that
309 were not present in known *Simkaniaceae* genomes (Table 1). First, the genomes have
310 cold shock protein genes, and the genes were highly conserved (93% amino acid
311 identity) between the two *Simkaniaceae* genomes. Cold shock proteins are used to
312 deal with the sudden drop in temperature [36]. The proteins are thought to be able to
313 bind with nucleic acids to prevent the disruption of mRNA transcription and protein
314 translation caused by the formation of mRNA secondary structures due to low
315 temperature [36]. The existence of the genes in the genomes may confer cold
316 resistance on the *Simkaniaceae* bacteria in Lake Shunet, allowing them to withstand
317 extremely cold environments.

318 Besides the cold shock protein genes, the two *Simkaniaceae* also had glutamate
319 decarboxylase (GAD) genes. GAD is an enzyme that catalyzes the conversion of
320 glutamate into γ -aminobutyric acid (GABA) and carbon dioxide. Many bacteria can
321 utilize the GAD system to tolerate acidic stress by consuming protons during a
322 chemical reaction [37]. The system usually accompanies glutamate/GABA antiporters,
323 responsible for coupling the efflux of GABA and influx of glutamate. The antiporter

324 can also be found in the two novel *Simkaniaceae* genomes, indicating that the bacteria
325 can use the system to tolerate acidic environments.

326 Along with the unique features in the genus, we identified a difference between
327 the two MAGs in terms of metabolism. Taking sulfur metabolism as an example, the
328 M30B2 had all the genes for assimilatory sulfate reduction (ASR), except for *cysH*,
329 and contained the sulfate permease gene (Fig. S3). On the contrary, M30B1 did not
330 contain ASR or the sulfate permease gene. This indicates that M30B2 can take up and
331 use sulfate as a sulfur source, but M30B1 cannot. In summary, the recovery of these
332 genomes broadens our knowledge of the metabolic versatility in *Simkaniaceae*.

333 ***Candidate novel Cyanobium species.*** The MAG M30B3 was classified as a
334 novel *Cyanobacteria* species genome under the genus *Cyanobium*, based on
335 phylogenomic tree and 84.28% ANI shared with the *Cyanobium_A* sp007135755
336 genome (GCA_007135755.1), its closest phylogenetic neighbor. We named the
337 genome *Candidatus Cyanobium* sp. M30B3, abbreviated as M30B3. The M30B3 is
338 the predominant bacterium in Lake Shunet at 3.0 m and plays a pivotal role in
339 providing organic carbon in the lake ecosystem [12].

340 Our analysis of the M30B3 genome revealed that the bacterium harbors an
341 anti-phage system that its known relatives lack. In the novel cyanobacterial genome
342 under the *Cyanobium* genus, we found that the genome harbored several

343 CRISPR-associated (Cas) protein genes that were not in other *Cyanobium* genomes
 344 (Table 1). The CRISPR-Cas system is a prokaryotic immune system that enables
 345 prokaryotic cells to defend against phages [38]. The system can be classified into six
 346 types and several subtypes according to protein content. The signature protein of type
 347 I is Cas3, which has endonuclease and helicase activities [38]. *cas3* genes can be
 348 found in the novel *Cyanobium* genome but not in other known *Cyanobium* genomes.
 349 Furthermore, the genome also had *cse1* and *cse2* proteins, signature proteins for the
 350 I-E subtype. Our results show that the novel genome harbors a type I-E CRISPR
 351 system and that this system is absent in its phylogenetic-close relatives.

352 ***Candidate novel Thiocapsa species M50B4.*** Lake Shunet features the extremely
 353 dense purple sulfur bacteria (PSB) in its chemocline (5.0 m) layer, and the density of
 354 these PSB is comparable to that of Lake Mahoney (Canada), renowned for containing
 355 the most purple sulfur bacteria of any lake in the world [39]. A complete MAG of
 356 *Thiocapsa* species, the predominant PSB in the 5.0 m layer, was recovered from the
 357 5.0 m dataset. The MAG was classified as a candidate novel species because it shared
 358 90.71% ANI with the genome of *Thiocapsa rosea*. Therefore, we propose the creation
 359 of a new species, *Candidatus Thiocapsa halobium*, abbreviated as M50B4.

360 The complete genome of the predominant PSB M50B4 will help us understand
 361 carbon, nitrogen, and sulfur cycling in Lake Shunet. *Thiocapsa* can perform

362 photosynthesis by reducing sulfur as an electron donor, and *Thiocapsa* can fix
 363 nitrogen [40, 41]. M50B4 contained genes for bacteriochlorophyll synthesis and the
 364 Calvin cycle for carbon fixation. A previous study revealed that *Thioflavicoccus*
 365 *mobilis*, a bacterium close to *Thiocapsa*, can utilize rTCA and the Calvin cycle to fix
 366 carbon [42]. In M50B4, all genes for the reverse TCA cycle (rTCA), except for the
 367 ATP citrate lyase gene, were identified. Whether the M50B4 can use both rTCA and
 368 the Calvin cycle like *T. mobilis* needs to be determined. For sulfur metabolism, the
 369 MAG carried intact gene sets involved in SOX system dissimilatory sulfate
 370 reduction/oxidation. The sulfate importer gene was also seen in the MAG, which
 371 equipped the bacterium with the ability to import extracellular thiosulfate and sulfate
 372 and to use them as sulfur sources. In terms of nitrogen metabolism, like other
 373 *Thiocapsa*, the bacterium had a gene cluster to conduct nitrogen fixation and a urea
 374 transporter and urease gene cluster to utilize urea. Besides nitrogen fixation, currently
 375 available *Thiocapsa* have all genes for denitrification, and some *Thiocapsa* have
 376 genes to convert nitrite to nitrate. However, the genes were not seen in our MAG.

377 There are currently five cultured *Thiocapsa* species. Two of these, *T. rosea* and *T.*
 378 *pendens*, contain gas vesicles. Our genomic analysis revealed gas vesicle structure
 379 protein genes in M50B4. These genes are also present in *T. rosea* and *T. pendens*, but
 380 not in any other *Thiocapsa* genomes, indicating that the genes are critical for vesicles

381 to exist in *Thiocapsa*, and therefore the novel species have gas vesicles. Gas vesicles
 382 enable *T. sp.* M50B4 cells to modulate their buoyancy so they can move to the
 383 locations with optimal light intensity or oxygen concentration [43]. Environmental
 384 conditions of Lake Shunet are known to be dynamic and change with seasons, and
 385 this function could be critical for M50B4.

386 We found that the novel *Thiocapsa* complete MAG have genes that encode
 387 dimethyl sulfoxide (DMSO) reductase subunits A and B (Table S2). DMSO reductase
 388 is an enzyme that catalyzes the reduction of DMSO into dimethyl sulfide (DMS). The
 389 reductase enables bacteria to use DMSO as terminal electron acceptors instead of
 390 oxygen during cellular respiration [44]. The DMSO reduction reaction could impact
 391 the environment. DMS, the product of the reaction, can be emitted into the
 392 atmosphere and be oxidized into sulfuric acid [45]. Sulfuric acid can act as a cloud
 393 condensation nucleus and leads to cloud formation, blocking radiation from the sun.
 394 The flux of the anti-greenhouse gas DMS is mainly investigated and discussed in
 395 oceanic environments [46, 47]. The flux and role of DMS in lake ecosystems are
 396 overlooked and rarely documented [48]. Our finding that the extremely dense PSB in
 397 Lake Shunet carried DMS metabolism shows the need to investigate the impact and
 398 importance of DMS from bacteria in lake ecosystems and sulfur cycling.

399 ***Candidate novel Methylophilaceae species M30B5.*** A complete MAG, named

M30B5, was classified as a novel *Methylophilaceae* species under a genus-level lineage, called GCA-2401735, which was defined based on phylogenetic placement [49]. The GCA-2401735 lineage currently only comprises two genomes—GCA-2401735 sp006844635 and GCA-2401735 sp002401735—neither of which meet high-quality genome standards due to their low completeness and lack of 16S rRNA gene sequence. The novel complete genome can serve as a representative species of the genus and can be used to infer the capability of the genus (Table S2). Here, we propose the genus *Candidatus* Methylofavorus to include the three GCA-2401735 genomes, and the M30B5 was renamed as *Candidatus* Methylofavorus khakassia.

The isolation locations of the three genomes imply that their habitats were distinct from those of other *Methylophilaceae*. The three “Methylofavorus” genomes were isolated from a cold subseafloor aquifer, shallow marine methane seep, and saline lake, indicating that the bacteria can live in saline environments. By comparison, most other *Methylophilaceae* members live in soil and freshwater or are associated with plants (except for the OM43 lineage) [50]. This indicates that the ancestor of “Methylofavorus” gained the ability to live in saline habitats and diverged from the ancestor of the genus *Methylophilus*, its closest phylogenetic relatives.

The complete genome of M30B5 enables us to comprehensively study metabolic

419 potentials. *Methylophilaceae* is a family of *Proteobacteria* that can use methylamine
420 or methanol as carbon or energy sources [51, 52]. In our analysis, methanol
421 dehydrogenase gene existed in our genome, and methylamine dehydrogenase gene
422 was absent, indicating that the bacteria use methanol as a carbon source instead of
423 methylamine. For motility, flagella are found in some *Methylophilaceae*. Interestingly,
424 flagella- and chemotaxis-related genes were not identified in the MAG but were
425 identified in the other two “*Methylofavorus*” species, suggesting that M30B5 lacks
426 mobility comparing to the other two “*Methylofavorus*” species (Fig. S4).

427 The comparative analysis of M30B5 and other “*Methylofavorus*” species
428 revealed that the bacteria use different types of machinery to obtain nitrogen (Fig. S4).
429 The formamidase, urease, and urea transporters were present in M30B5 but not the
430 other two “*Methylofavorus*” species. Instead, the two “*Methylofavorus*” species had
431 nitrite reductase, which was not in our MAG. The results indicate that M30B5 can
432 convert formamide into ammonia and formate, and take up extracellular urea as a
433 nitrogen source. On the contrary, the other two “*Methylofavorus*” can use nitrite as
434 nitrogen resources. Our analysis revealed that “*Methylofavorus*” is metabolically
435 heterogeneous.

436

437 **Conclusions**

438 In this study, we successfully developed a workflow to recover MAGs by
439 combining SRs and LR. This workflow reconstructed hundreds of high-quality and
440 six complete MAGs—including six candidate novel bacterial orders, 20 families, 66
441 genera, and 154 species—from water samples of Lake Shunet, a meromictic lake with
442 a highly complex microbial community. It demonstrates that with extra less LR, we
443 can salvage important genome information from previous SR metagenomes. Using
444 comparative genomics, unique and intriguing metabolic features are identified in
445 these complete MAGs, including two predominant novel species: *Thiocapsa* sp, and
446 *Cyanobium* sp. [12]. The findings show that it is advantageous to apply this method in
447 studies of microbial ecology and microbial genomics by revising and improving the
448 shortcomings of SRs-based metagenomes. Additionally, we show that the MAGs
449 contain a high proportion of potential novel BGCs and CAZymes, which can be
450 valuable resources to validate and examine the metabolic flexibility of various
451 microbial lineages through further experimental approaches and comparative
452 genomics. Finally, this study found a high ratio of poorly detectable taxa in the public
453 databases, suggesting that the investigation into rarely explored environments is
454 necessary to populate the genomic encyclopedia of the microbial world, explore
455 microbial metabolic diversity, and fill the missing gaps in microbial evolution.
456

457 **Materials and Methods**

458 **Sample collection**

459 Water samples at 3.0, 5.0, and 5.5 m deep were collected from Lake Shunet (54°
460 25'N, 90° 13'E) on July 21, 2010. The collection procedure was described in our
461 previous research [12]. Briefly, water was pumped from each depth into sterile
462 containers. Part of the water was transferred into sterile 2.0-ml screw tubes (SSIbio®)
463 and stored at -80°C until DNA extraction. The rest of the water was filtered through
464 10-µm plankton and concentrated using tangential flow filtration (TFF) system
465 (Millipore) with 0.22-µm polycarbonate membrane filters. The bacteria in the
466 retentate were then retained on cellulose acetate membranes (0.2 µm pore size;
467 ADVANTEC, Tokyo, Japan) and stored at -80°C until DNA extraction.

468 **DNA extraction and sequencing**

469 Reads from Illumina and Nanopore sequencing platforms were used in this study.
470 The sequencing reads from Illumina were described in our previous study [12] (Table
471 S1). DNA for Illumina sequencing was extracted from a TFF-concentrated sample
472 using the cetyltrimethylammonium bromide (CTAB) method [53]. In terms of
473 Nanopore sequencing for 3.0-m samples, the same DNA batch used for Illumina
474 sequencing of 3.0-m was sent to Health GeneTech Corp. (Taiwan) for Nanopore
475 sequencing. For 5.0- and 5.5-m samples, there was no DNA remaining after Illumina

sequencing, so in 2020 the DNA was extracted again from frozen water samples using the CTAB method by retaining the bacteria on cellulose acetate membranes without TFF concentration. The amounts of DNA were still insufficient for Nanopore sequencing, so the DNA samples were mixed with the DNA of a known bacterium, *Endozoicomonas* isolate, at a 1:2 ratio. No *Endozoicomonas* was detected in the water samples according to our 16S rRNA amplicon survey [12]. The mixed DNA was then sent to the NGS High Throughput Genomics Core at Biodiversity Research Center, Academia Sinica for Nanopore sequencing. To remove reads that had originated from the *Endozoicomonas* isolate, Kaiju web server [54] and Kraken 2 [55] were used to assign the taxonomy for each read; reads that were classified as *Endozoicomonas* by Kaiju or Kraken were removed from our sequencing results. The Nanopore sequencing and processing yielded 13.83, 12.57, and 4.79 Gbp of reads from the 3.0, 5.0, and 5.5 m samples, respectively (Table S1).

Metagenome assembly

MAG assembly was performed by combining short reads (SRs) from Illumina sequencing and long reads (LRs) from Nanopore sequencing; this workflow is described in Figure 1a. First, the LRs from 3.0, 5.0, and 5.5 m datasets were individually assembled by metaFlye v2.8 [11] with default settings, and the assemblies were polished with corresponding SRs using Pilon v1.23 [56]. On the

other hand, SRs were also assembled by MEGAHIT v1.2.9 with k-mer of 21, 31, 41, and 51 [57]. The assemblies from SRs and LRAs were then merged by quickmerge v0.3 with parameters -ml 7500 -c 3 -hco 8 [58]. The merge assemblies were then binned using MaxBin2 [59], MetaBAT2 [60], and CONCOCT [61] in metaWRAP v1.3 [14]. The bins from the three bin sets were then refined by the bin refinement module in metaWRAP v1.3. The resulting bins were then polished again by Pilon v1.23 five times. To reassemble the bin, sorted reads that belonged to individual bin were extracted by BWA-MEM v0.7.17 [62] for SRs and by minimap2 for LRAs [63]. The extracted long reads were assembled by Flye v2.8, or metaFlye v2.8 if the assembly failed using Flye v2.8 [11, 64]. The bins were then reassembled individually using Unicycler v0.4.8 using the extracted reads and reassembled long-read contigs which were used as bridges [65]. To determine whether the original or reassembled bin was better, the bin with higher value of genome completeness - $2.5 \times$ contamination, estimated by checkM v1.1.3 [13], was chosen and retained. Contigs labeled as circular by Flye or metaFlye, >2.0 Mb in size, and completeness >95% were considered “complete” MAGs. The complete MAGs were visualized using CGView Server [66].

While we were preparing this manuscript, Damme *et al.* published a hybrid assembler, called MUFFIN, that also integrates metaFlye and metaWRAP to recover MAGs and Unicycler for reassembly [67]. However, our workflow has a step to

merge the assemblies from SRs and LR to increase the contiguity and assembly size.

Moreover, for the reassembly, we use contigs from metaFlye, instead of default setting: miniasm, as the bridge, which we found can produce a better quality reassembly.

Annotation of metagenome-assembly genomes

The completeness, contamination, and other statistics on metagenome-assembled genomes (MAGs) were evaluated using CheckM v1.1.3 [13]. The genome statistics were processed in R [68] and visualized using the ggplot2 package [69]. The taxonomy of MAGs was inferred by GTDB-Tk v1.3.0 [70]. Average Nucleotide Identities (ANIs) between MAGs were determined by FastANI v1.32 [16]. MAGs were annotated using Prokka v1.14.5 with ‘rfam’ options [71]. To annotate MAGs with KEGG functional orthologs (K numbers), putative protein sequences predicted by Prodigal v2.6.3 [72] were annotated using EnrichM v0.6.0 [73]. The K number annotation results were then used to reconstruct the transporter systems and metabolic pathways using KEGG mapper [74], and the completeness of KEGG modules was evaluated using EnrichM. Secondary metabolite biosynthetic gene clusters in each MAG were identified using antiSMASH v5.0 [75]. Ribosomal RNA sequences were inferred by barrnap v0.9 [76].

Phylogenetic analysis

Bacterial and archaeal phylogenomic trees were inferred by a *de novo* workflow in GTDB-Tk v1.3.0 [70]. All species-level non-redundant MAGs recovered in this study were analyzed together with the reference genomes in Genome Taxonomy Database (GTDB) [49]. In the *de novo* workflow, marker genes in each genome were identified using HMMER 3.1b2 [77]. Multiple sequence alignments based on the bacterial or archaeal marker sets were then generated and masked with default settings. Trees of bacteria and archaea were then inferred from the masked multiple sequence alignment using FastTree with the WAG+GAMMA models and 1,000 bootstraps [78]. The trees were visualized with the interactive Tree of Life (iTOL) v4 [79].

542

543

544

545 **Declarations**

546 *Ethics approval and consent to participate*

547 Not applicable

548

549 *Consent for publication*

550 Not applicable

551

552 ***Availability of data and materials***

553 All sequencing data and assembled genomes are available through National Center for
554 Biotechnology Information (NCBI) repositories under BioProject ID: PRJNA721826.
555 Sequence reads of metagenomes from samples at 3.0, 5.0, and 5.5 m deep can be
556 found under SRA accession numbers SRR14300307, SRR14300308, SRR14300309,
557 SRR14307495, SRR14307795, and SRR14307796. The accession numbers of MAGs
558 can be found in dataset S1 and S2.

559

560 ***Conflict of Interest***

561 The authors declare that they have no conflict of interest.

562

563 ***Funding***

564 The study was supported by the Ministry of Science and Technology in Taiwan
565 through the Taiwan–Russia Joint Project Grant NSC 102-2923-B-001-004-MY3 and
566 MOST 105-2923-B-001-001-MY3.

567

568 ***Author's contributions***

569 Y.H.C. and S.L.T. conceived the idea for this study. Y.H.C. assembled the genomes,
570 performed the bioinformatics analysis, and wrote the manuscript. P.W.C. and H.H.C.

571 prepared the DNA samples. D.R. and A.D. collected water samples. S.L.T. supervised
572 the overall study. All authors read and approved the manuscript.

573

574 *Acknowledgements*

575 This study was supported by funding from the Ministry of Science and Technology,
576 Taiwan. Y.H.C. would like to acknowledge the Taiwan International Graduate
577 Program (TIGP) for its fellowship towards his graduate studies. We would like to
578 thank Noah Last of Third Draft Editing for his English language editing.

579

- 580 1. Chen LX, Anantharaman K, Shaiber A, Eren AM, Banfield JF: **Accurate and**
581 **complete genomes from metagenomes**. *Genome Res* 2020, **30**(3):315-333.
- 582 2. Che Y, Xia Y, Liu L, Li AD, Yang Y, Zhang T: **Mobile antibiotic resistome in**
583 **wastewater treatment plants revealed by Nanopore metagenomic**
584 **sequencing**. *Microbiome* 2019, **7**(1):44.
- 585 3. Nayfach S, Shi ZJ, Seshadri R, Pollard KS, Kyrpides NC: **New insights from**
586 **uncultivated genomes of the global human gut microbiome**. *Nature* 2019,
587 **568**(7753):505-510.
- 588 4. Dong XY, Greening C, Rattray JE, Chakraborty A, Chuvochina M, Mayumi D,
589 Dolfing J, Li C, Brooks JM, Bernard BB *et al*: **Metabolic potential of uncultured**
590 **bacteria and archaea associated with petroleum seepage in deep-sea**
591 **sediments**. *Nature Communications* 2019, **10**.
- 592 5. Mukherjee S, Stamatis D, Bertsch J, Ovchinnikova G, Sundaramurthi JC, Lee J,
593 Kandimalla M, Chen IA, Kyrpides NC, Reddy TBK: **Genomes OnLine Database**
594 **(GOLD) v.8: overview and updates**. *Nucleic Acids Res* 2021,
595 **49**(D1):D723-D733.
- 596 6. Ayling M, Clark MD, Leggett RM: **New approaches for metagenome assembly**
597 **with short reads**. *Brief Bioinform* 2020, **21**(2):584-594.
- 598 7. Moss EL, Maghini DG, Bhatt AS: **Complete, closed bacterial genomes from**
599 **microbiomes using nanopore sequencing**. *Nature Biotechnology* 2020,

- 600 **38(6):701-+.**
- 601 8. Liu L, Wang YL, Che Y, Chen YQ, Xia Y, Luo RB, Cheng SH, Zheng CM, Zhang T:
602 **High-quality bacterial genomes of a partial-nitrification/anammox system by**
603 **an iterative hybrid assembly method.** *Microbiome* 2020, **8(1).**
- 604 9. Bertrand D, Shaw J, Kalathiyappan M, Ng AHQ, Kumar MS, Li C, Dvornic M,
605 Soldo JP, Koh JY, Tong C *et al*: **Hybrid metagenomic assembly enables**
606 **high-resolution analysis of resistance determinants and mobile elements in**
607 **human microbiomes.** *Nat Biotechnol* 2019, **37(8):937-944.**
- 608 10. Latorre-Perez A, Villalba-Bermell P, Pascual J, Vilanova C: **Assembly methods**
609 **for nanopore-based metagenomic sequencing: a comparative study.** *Sci Rep*
610 2020, **10(1):13588.**
- 611 11. Kolmogorov M, Bickhart DM, Behsaz B, Gurevich A, Rayko M, Shin SB, Kuhn K,
612 Yuan J, Polevikov E, Smith TPL *et al*: **metaFlye: scalable long-read**
613 **metagenome assembly using repeat graphs.** *Nat Methods* 2020,
614 **17(11):1103-1110.**
- 615 12. Wu YT, Yang CY, Chiang PW, Tseng CH, Chiu HH, Saeed I, Baatar B, Rogozin D,
616 Halgamuge S, Degermendzhi A *et al*: **Comprehensive Insights Into**
617 **Composition, Metabolic Potentials, and Interactions Among Archaeal,**
618 **Bacterial, and Viral Assemblages in Meromictic Lake Shunet in Siberia.**
619 *Frontiers in Microbiology* 2018, **9.**
- 620 13. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW: **CheckM:**
621 **assessing the quality of microbial genomes recovered from isolates, single**
622 **cells, and metagenomes.** *Genome Res* 2015, **25(7):1043-1055.**
- 623 14. Uritskiy GV, DiRuggiero J, Taylor J: **MetaWRAP-a flexible pipeline for**
624 **genome-resolved metagenomic data analysis.** *Microbiome* 2018, **6(1):158.**
- 625 15. Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy
626 TBK, Schulz F, Jarett J, Rivers AR, Elie-Fadrosh EA *et al*: **Minimum information**
627 **about a single amplified genome (MISAG) and a metagenome-assembled**
628 **genome (MIMAG) of bacteria and archaea.** *Nat Biotechnol* 2017,
629 **35(8):725-731.**
- 630 16. Jain C, Rodriguez RL, Phillippy AM, Konstantinidis KT, Aluru S: **High**
631 **throughput ANI analysis of 90K prokaryotic genomes reveals clear species**
632 **boundaries.** *Nat Commun* 2018, **9(1):5114.**
- 633 17. Nayfach S, Roux S, Seshadri R, Udway D, Varghese N, Schulz F, Wu D,
634 Paez-Espino D, Chen IM, Huntemann M *et al*: **A genomic catalog of Earth's**
635 **microbiomes.** *Nat Biotechnol* 2020.
- 636 18. Parks DH, Rinke C, Chuvochina M, Chaumeil PA, Woodcroft BJ, Evans PN,
637 Hugenholtz P, Tyson GW: **Recovery of nearly 8,000 metagenome-assembled**

- 638 **genomes substantially expands the tree of life.** *Nat Microbiol* 2017,
639 **2(11):1533-1542.**
- 640 19. Obbels D, Verleyen E, Mano MJ, Namsaraev Z, Sweetlove M, Tytgat B,
641 Fernandez-Carazo R, De Wever A, D'hondt S, Ertz D *et al*: **Bacterial and**
642 **eukaryotic biodiversity patterns in terrestrial and aquatic habitats in the Sor**
643 **Rondane Mountains, Dronning Maud Land, East Antarctica.** *Fems Microbiol*
644 *Ecol* 2016, **92(6).**
- 645 20. Demain AL, Fang A: **The natural functions of secondary metabolites.** *Adv*
646 *Biochem Eng Biotechnol* 2000, **69**:1-39.
- 647 21. Beedessee G, Hisata K, Roy MC, Van Dolah FM, Satoh N, Shoguchi E:
648 **Diversified secondary metabolite biosynthesis gene repertoire revealed in**
649 **symbiotic dinoflagellates.** *Sci Rep-Uk* 2019, **9.**
- 650 22. Medema MH, Fischbach MA: **Computational approaches to natural product**
651 **discovery.** *Nat Chem Biol* 2015, **11(9):639-648.**
- 652 23. Chavali AK, Rhee SY: **Bioinformatics tools for the identification of gene**
653 **clusters that biosynthesize specialized metabolites.** *Brief Bioinform* 2018,
654 **19(5):1022-1034.**
- 655 24. Gong G, Zhou SS, Luo RB, Gesang Z, Suolang S: **Metagenomic insights into the**
656 **diversity of carbohydrate-degrading enzymes in the yak fecal microbial**
657 **community.** *Bmc Microbiol* 2020, **20(1).**
- 658 25. Sathya TA, Khan M: **Diversity of Glycosyl Hydrolase Enzymes from**
659 **Metagenome and Their Application in Food Industry.** *J Food Sci* 2014,
660 **79(11):R2149-R2156.**
- 661 26. Nakamura AM, Nascimento AS, Polikarpov I: **Structural diversity of**
662 **carbohydrate esterases.** *Biotechnology Research and Innovation* 2017,
663 **1(1):35-51.**
- 664 27. Alagawany M, Elnesr SS, Farag MR: **The role of exogenous enzymes in**
665 **promoting growth and improving nutrient digestibility in poultry.** *Iran J Vet*
666 *Res* 2018, **19(3):157-164.**
- 667 28. Espadaler J, Eswar N, Querol E, Aviles FX, Sali A, Marti-Renom MA, Oliva B:
668 **Prediction of enzyme function by combining sequence similarity and protein**
669 **interactions.** *Bmc Bioinformatics* 2008, **9.**
- 670 29. Addou S, Rentzsch R, Lee D, Orengo CA: **Domain-Based and Family-Specific**
671 **Sequence Identity Thresholds Increase the Levels of Reliable Protein**
672 **Function Transfer.** *Journal of Molecular Biology* 2009, **387(2):416-430.**
- 673 30. Radivojac P, Clark WT, Oron TR, Schnoes AM, Wittkop T, Sokolov A, Graim K,
674 Funk C, Verspoor K, Ben-Hur A *et al*: **A large-scale evaluation of**
675 **computational protein function prediction.** *Nature Methods* 2013,

- 676 10(3):221-227.
- 677 31. Minh BQ, Nguyen MAT, von Haeseler A: **Ultrafast Approximation for**
678 **Phylogenetic Bootstrap.** *Mol Biol Evol* 2013, **30**(5):1188-1195.
- 679 32. Borrel G, Parisot N, Harris HMB, Peyretailade E, Gaci N, Tottey W, Bardot O,
680 Raymann K, Gribaldo S, Peyret P *et al*: **Comparative genomics highlights the**
681 **unique biology of Methanomassiliicoccales, a Thermoplasmatales-related**
682 **seventh order of methanogenic archaea that encodes pyrrolysine.** *Bmc*
683 *Genomics* 2014, **15**.
- 684 33. Qin QL, Xie BB, Zhang XY, Chen XL, Zhou BC, Zhou J, Oren A, Zhang YZ: **A**
685 **proposed genus boundary for the prokaryotes based on genomic insights.** *J*
686 *Bacteriol* 2014, **196**(12):2210-2215.
- 687 34. Everett KDE, Bush RM, Andersen AA: **Emended description of the order**
688 **Chlamydiales, proposal of Parachlamydiaceae fam. nov. and Simkaniaceae**
689 **fam. nov., each containing one monotypic genus, revised taxonomy of the**
690 **family Chlamydiaceae, including a new genus and five new species, and**
691 **standards for the identification of organisms.** *Int J Syst Bacteriol* 1999,
692 **49**:415-440.
- 693 35. Lienard J, Croxatto A, Gervais A, Levi Y, Loret JF, Posfay-Barbe KM, Greub G:
694 **Prevalence and diversity of Chlamydiales and other amoeba-resisting**
695 **bacteria in domestic drinking water systems.** *New Microbes New Infect* 2017,
696 **15**:107-116.
- 697 36. Keto-Timonen R, Hietala N, Palonen E, Hakakorpi A, Lindstrom M, Korkeala H:
698 **Cold Shock Proteins: A Minireview with Special Emphasis on Csp-family of**
699 **Enteropathogenic Yersinia.** *Front Microbiol* 2016, **7**:1151.
- 700 37. Feehily C, Karatzas KA: **Role of glutamate metabolism in bacterial responses**
701 **towards acid and other stresses.** *J Appl Microbiol* 2013, **114**(1):11-24.
- 702 38. Makarova KS, Haft DH, Barrangou R, Brouns SJJ, Charpentier E, Horvath P,
703 Moineau S, Mojica FJM, Wolf YI, Yakunin AF *et al*: **Evolution and classification**
704 **of the CRISPR-Cas systems.** *Nat Rev Microbiol* 2011, **9**(6):467-477.
- 705 39. Overmann J, Beatty JT, Hall KJ, Pfennig N, Northcote TG: **Characterization of a**
706 **Dense, Purple Sulfur Bacterial Layer in a Meromictic Salt Lake.** *Limnol*
707 *Oceanogr* 1991, **36**(5):846-859.
- 708 40. Caumette P, Guyoneaud R, Imhoff JF, Suling J, Gorlenko V: **Thiocapsa marina**
709 **sp. nov., a novel, okenone-containing, purple sulfur bacterium isolated from**
710 **brackish coastal and marine environments.** *Int J Syst Evol Microbiol* 2004,
711 **54**(Pt 4):1031-1036.
- 712 41. Schott J, Griffin BM, Schink B: **Anaerobic phototrophic nitrite oxidation by**
713 **Thiocapsa sp. strain KS1 and Rhodopseudomonas sp. strain LQ17.**

- 714 *Microbiology (Reading)* 2010, **156**(Pt 8):2428-2437.
- 715 42. Rubin-Blum M, Dubilier N, Kleiner M: **Genetic Evidence for Two Carbon**
- 716 **Fixation Pathways (the Calvin-Benson-Bassham Cycle and the Reverse**
- 717 **Tricarboxylic Acid Cycle) in Symbiotic and Free-Living Bacteria.** *mSphere*
- 718 2019, **4**(1).
- 719 43. Walsby AE: **Gas vesicles.** *Microbiol Rev* 1994, **58**(1):94-144.
- 720 44. Bilous PT, Weiner JH: **Dimethyl sulfoxide reductase activity by anaerobically**
- 721 **grown Escherichia coli HB101.** *J Bacteriol* 1985, **162**(3):1151-1155.
- 722 45. Veres PR, Neuman JA, Bertram TH, Assaf E, Wolfe GM, Williamson CJ,
- 723 Weinzierl B, Tilmes S, Thompson CR, Thames AB *et al*: **Global airborne**
- 724 **sampling reveals a previously unobserved dimethyl sulfide oxidation**
- 725 **mechanism in the marine atmosphere.** *P Natl Acad Sci USA* 2020,
- 726 **117**(9):4505-4510.
- 727 46. Andreae MO, Raemdonck H: **Dimethyl Sulfide in the Surface Ocean and the**
- 728 **Marine Atmosphere - a Global View.** *Science* 1983, **221**(4612):744-747.
- 729 47. Yoch DC: **Dimethylsulfoniopropionate: Its sources, role in the marine food**
- 730 **web, and biological degradation to dimethylsulfide.** *Appl Environ Microb*
- 731 2002, **68**(12):5804-5815.
- 732 48. Steinke M, Hodapp B, Subhan R, Bell TG, Martin-Creuzburg D: **Flux of the**
- 733 **biogenic volatiles isoprene and dimethyl sulfide from an oligotrophic lake.**
- 734 *Sci Rep-Uk* 2018, **8**.
- 735 49. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil PA,
- 736 Hugenholtz P: **A standardized bacterial taxonomy based on genome**
- 737 **phylogeny substantially revises the tree of life.** *Nat Biotechnol* 2018,
- 738 **36**(10):996-1004.
- 739 50. Salcher MM, Schaefle D, Kaspar M, Neuenschwander SM, Ghai R: **Evolution in**
- 740 **action: habitat transition from sediment to the pelagial leads to genome**
- 741 **streamlining in Methylophilaceae.** *Isme Journal* 2019, **13**(11):2764-2777.
- 742 51. Salcher MM, Schaefle D, Kaspar M, Neuenschwander SM, Ghai R: **Evolution in**
- 743 **action: habitat transition from sediment to the pelagial leads to genome**
- 744 **streamlining in Methylophilaceae.** *ISME J* 2019, **13**(11):2764-2777.
- 745 52. Kalyuzhnaya MG, Bowerman S, Lara JC, Lidstrom ME, Chistoserdova L:
- 746 **Methylotenera mobilis gen. nov., sp nov., an obligately**
- 747 **methylamine-utilizing bacterium within the family Methylophilaceae.** *Int J*
- 748 *Syst Evol Micr* 2006, **56**:2819-2823.
- 749 53. Wilson K: **Preparation of genomic DNA from bacteria.** *Curr Protoc Mol Biol*
- 750 2001, **Chapter 2**:Unit 2 4.
- 751 54. Menzel P, Ng KL, Krogh A: **Fast and sensitive taxonomic classification for**

752 **metagenomics with Kaiju**. *Nat Commun* 2016, **7**:11257.

753 55. Wood DE, Lu J, Langmead B: **Improved metagenomic analysis with Kraken 2**.
754 *Genome Biol* 2019, **20**(1):257.

755 56. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA,
756 Zeng Q, Wortman J, Young SK *et al*: **Pilon: an integrated tool for**
757 **comprehensive microbial variant detection and genome assembly**
758 **improvement**. *PLoS One* 2014, **9**(11):e112963.

759 57. Li D, Liu CM, Luo R, Sadakane K, Lam TW: **MEGAHIT: an ultra-fast single-node**
760 **solution for large and complex metagenomics assembly via succinct de**
761 **Bruijn graph**. *Bioinformatics* 2015, **31**(10):1674-1676.

762 58. Chakraborty M, Baldwin-Brown JG, Long AD, Emerson JJ: **Contiguous and**
763 **accurate de novo assembly of metazoan genomes with modest long read**
764 **coverage**. *Nucleic Acids Res* 2016, **44**(19):e147.

765 59. Wu YW, Simmons BA, Singer SW: **MaxBin 2.0: an automated binning**
766 **algorithm to recover genomes from multiple metagenomic datasets**.
767 *Bioinformatics* 2016, **32**(4):605-607.

768 60. Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, Wang Z: **MetaBAT 2: an**
769 **adaptive binning algorithm for robust and efficient genome reconstruction**
770 **from metagenome assemblies**. *PeerJ* 2019, **7**:e7359.

771 61. Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijaz UZ, Lahti L,
772 Loman NJ, Andersson AF, Quince C: **Binning metagenomic contigs by**
773 **coverage and composition**. *Nat Methods* 2014, **11**(11):1144-1146.

774 62. Li H, Durbin R: **Fast and accurate short read alignment with**
775 **Burrows-Wheeler transform**. *Bioinformatics* 2009, **25**(14):1754-1760.

776 63. Li H: **Minimap2: pairwise alignment for nucleotide sequences**.
777 *Bioinformatics* 2018, **34**(18):3094-3100.

778 64. Kolmogorov M, Yuan J, Lin Y, Pevzner PA: **Assembly of long, error-prone reads**
779 **using repeat graphs**. *Nature Biotechnology* 2019, **37**(5):540-+.

780 65. Wick RR, Judd LM, Gorrie CL, Holt KE: **Unicycler: Resolving bacterial genome**
781 **assemblies from short and long sequencing reads**. *PLoS Comput Biol* 2017,
782 **13**(6):e1005595.

783 66. Grant JR, Stothard P: **The CGView Server: a comparative genomics tool for**
784 **circular genomes**. *Nucleic Acids Research* 2008, **36**:W181-W184.

785 67. Van Damme R, Holzer M, Viehweger A, Muller B, Bongcam-Rudloff E, Brandt C:
786 **Metagenomics workflow for hybrid assembly, differential coverage binning,**
787 **metatranscriptomics and pathway analysis (MUFFIN)**. *Plos Computational*
788 *Biology* 2021, **17**(2).

789 68. Team RDC: **R: A Language and Environment for Statistical Computing**. In.: R

- Foundation for Statistical Computing; 2020.
69. Villanueva RAM, Chen ZJ: **ggplot2: Elegant Graphics for Data Analysis, 2nd edition**. *Meas-Interdiscip Res* 2019, **17**(3):160-167.
70. Chaumeil PA, Mussig AJ, Hugenholtz P, Parks DH: **GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database**. *Bioinformatics* 2019.
71. Seemann T: **Prokka: rapid prokaryotic genome annotation**. *Bioinformatics* 2014, **30**(14):2068-2069.
72. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ: **Prodigal: prokaryotic gene recognition and translation initiation site identification**. *BMC Bioinformatics* 2010, **11**:119.
73. Joel A Boyd, Ben J Woodcroft, Tyson. GW: **Comparative genomics using EnrichM**. *In preparation* 2019.
74. Kanehisa M, Sato Y: **KEGG Mapper for inferring cellular functions from protein sequences**. *Protein Sci* 2020, **29**(1):28-35.
75. Blin K, Shaw S, Steinke K, Villebro R, Ziemert N, Lee SY, Medema MH, Weber T: **antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline**. *Nucleic Acids Res* 2019, **47**(W1):W81-W87.
76. Seemann T: **barrnap 0.9: rapid ribosomal RNA prediction**. *Google Scholar* 2013.
77. Eddy SR: **Accelerated Profile HMM Searches**. *PLoS Comput Biol* 2011, **7**(10):e1002195.
78. Price MN, Dehal PS, Arkin AP: **FastTree 2--approximately maximum-likelihood trees for large alignments**. *PLoS One* 2010, **5**(3):e9490.
79. Letunic I, Bork P: **Interactive Tree Of Life (iTOL) v4: recent updates and new developments**. *Nucleic Acids Res* 2019, **47**(W1):W256-W259.

Figure legends

Figure 1. Recovery of genomes from Lake Shunet using long- and short-read sequencing. **a.** The workflow for assembling metagenome-assembled genomes (MAGs). **b.** The value of log (N50) at 3.0, 5.0, and 5.5 m deep using SRs only, combining SRs an LRs (Hybrid), and reassembly of bins using the hybrid method. **c.** The correlation between SR coverage and log (N50) in the final Shunet MAG

822 collection (Reassembly). **d.** The correlation between LR coverage and log (N50) in
823 the final Shunet MAG collection (Reassembly). **e.** The completeness and
824 contamination of recovered MAGs. **f.** Venn diagram from the ratio of MAGs,
825 containing 5S, 16S, and 23S rRNA gene sequences. **g.** The GC ratio of MAGs
826 recovered from the 3.0, 5.0, and 5.5 m deep datasets.

827

828 **Figure 2. Taxonomical and molecular phylogenetic analyses of recovered**
829 **bacterial MAGs. a.** The numbers of novel taxonomic ranks of MAGs assigned by
830 GTDB-Tk. **b.** The phylum frequencies in the MAG collection from the Shunet dataset
831 and GTDB representative genomes. **c.** A phylogenetic tree based on the concatenation
832 of 120 single-copy gene protein sequences. After masking, 5,040 amino acid sites
833 were used in the analysis. The phylogenetic tree includes 188 recovered bacterial
834 MAGs and 30,238 bacterial representative genomes in GTDB-r95. The blue points
835 represent the placement of MAGs that are classified as novel species, the green points
836 represent novel genera, the red points represent novel families, and the black points
837 represent novel orders. Scale bar represents changes per amino acid site.

838

839 **Figure 3. Molecular phylogenetic analysis of recovered archaeal MAGs.** The
840 phylogenetic tree was reconstructed based on the concatenation of 122 single-copy

gene protein sequences. After masking, 5,124 amino acid sites were used in the analysis. The phylogenetic tree including three MAGs from Lake Shunet and 1,672 archaeal representative genomes in GTDB-r95. The blue dot represents the placement of MAG M55A2, the red dot represents MAG M55A1, and the green dot represents MAG M55A3. Scale bar represents changes per amino acid site.

Figure 4. Representation of the six complete MAGs.

The rings from the inside to outside represent GC content (black), GC skew- (purple), GC skew + (green), coding sequence regions (blue), rRNA gene sequences (black), transfer-messenger RNA (red), and secondary metabolite gene clusters (light blue). MAG ID M30B6 is classified as *Alcanivorax* sp002354605, M30B1 and M30B2 are *Simkaniaceae* sp., M30B3 is *Cyanobium* sp., M30B5 is “*Methylofavorus khakassia*”, and M50B4 is “*Thiocapsa halobium*”.

Table 1. KEGG orthologues that are present in the novel MAGs but absent in their sister taxa

856

	KO number	Definition
<i>Simkaniaceae</i>		
	K00954	Pantetheine-phosphate adenylyltransferase
	K01580	glutamate decarboxylase
	K15736	L-2-hydroxyglutarate
	K01607	carboxymuconolactone decarboxylase
	K03704	cold shock protein
<i>Thiocapsa</i>		
	K07306	anaerobic DMSO reductase subunit A
	K07307	anaerobic DMSO reductase subunit B
	K01575	acetolactate decarboxylase
	K13730	internalin A
	K05793	tellurite resistance protein TerB
	K05794	tellurite resistance protein TerC
	K05791	tellurium resistance protein TerZ
<i>Cyanobium</i>		
	K07012	CRISPR-associated endonuclease/helicase
	K07475	Cas3
	K15342	CRISP-associated protein Cas1
	K19046	CRISPR system Cascade subunit CasB
	K19123	CRISPR system Cascade subunit CasA
	K19124	CRISPR system Cascade subunit CasC
	K19125	CRISPR system Cascade subunit CasD
	K19126	CRISPR system Cascade subunit CasE

857

a

b

c

Tree scale: 1

- 1. Firmicutes
- 2. Firmicutes_D
- 3. Firmicutes_A
- 4. Firmicutes_F
- 5. Armatimonadota
- 6. Actinobacteriota
- 7. Patescibacteria
- 8. Chloroflexota
- 9. Cyanobacteria
- 10. Margulisbacteria
- 11. Bipolaricaulota
- 12. Spirochaetota
- 13. Caldatibacteriota
- 14. Bacteroidia
- 15. Marinisomatota
- 16. Cloacimonadota
- 17. Verrucomicrobiota
- 18. Verrucomicrobiota_A
- 19. Planctomycetota
- 20. Omnitrophota
- 21. Campylobacterota
- 22. Desulfobacterota
- 23. Myxococcota
- 24. Proteobacteria

Ca. Andegerimia shunetia

Ca. Andegerimia siberian

Cyanobium sp.

Ca. Thiocapsa halobium

Ca. Methylofavorsus khakassia

Alcanivorax sp002354605