Single-cell Genomics-Facilitated Read-first Binning of Candidate Phylum EM19 Genomes from Geothermal Spring Metagenomes

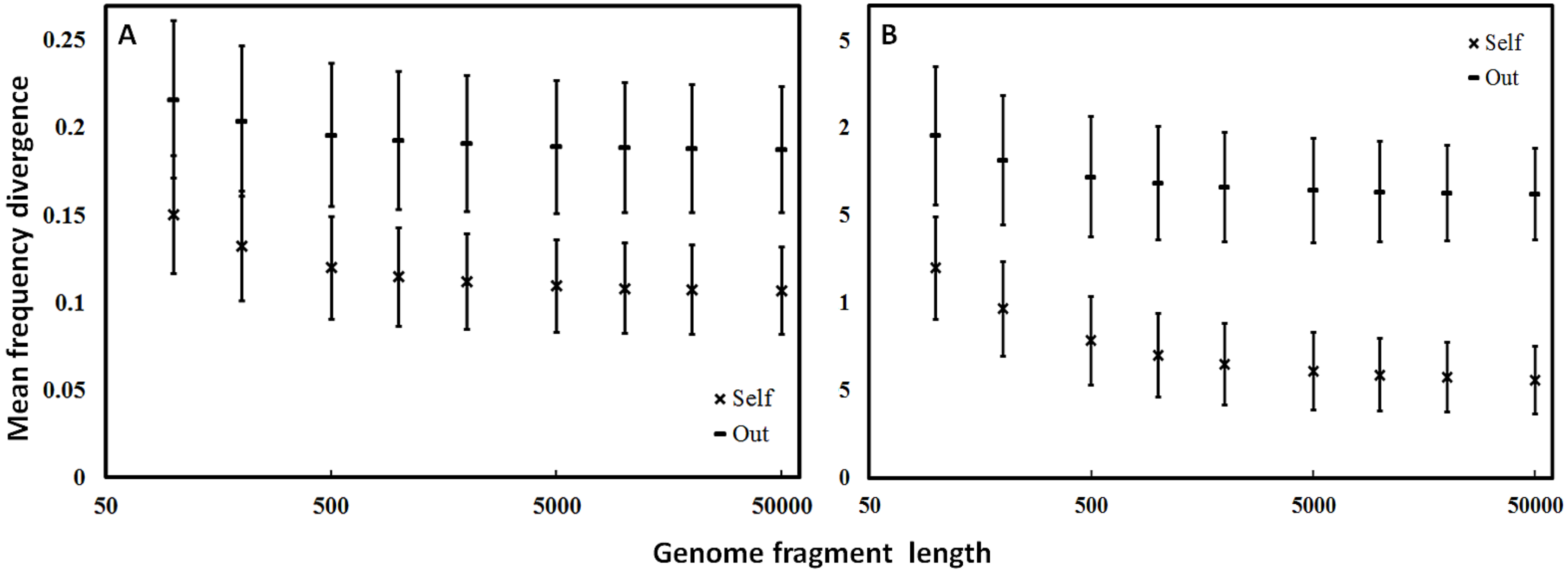
Supplemental Section I: Calescamantes Assembled Genome Alignment

Calescamantes SAG co-assembly from GBS was aligned with Progressive MAUVE to the metagenome assemblies obtained from GBS and Gongxiaoshe in Tengchong, China (Supplemental Figure 2). Contigs were matched between assemblies according to best BLASTN hit, and contigs were arranged by start position relative to the largest contig between assemblies. Contigs without a best BLASTN hit were added to the end of the alignment.

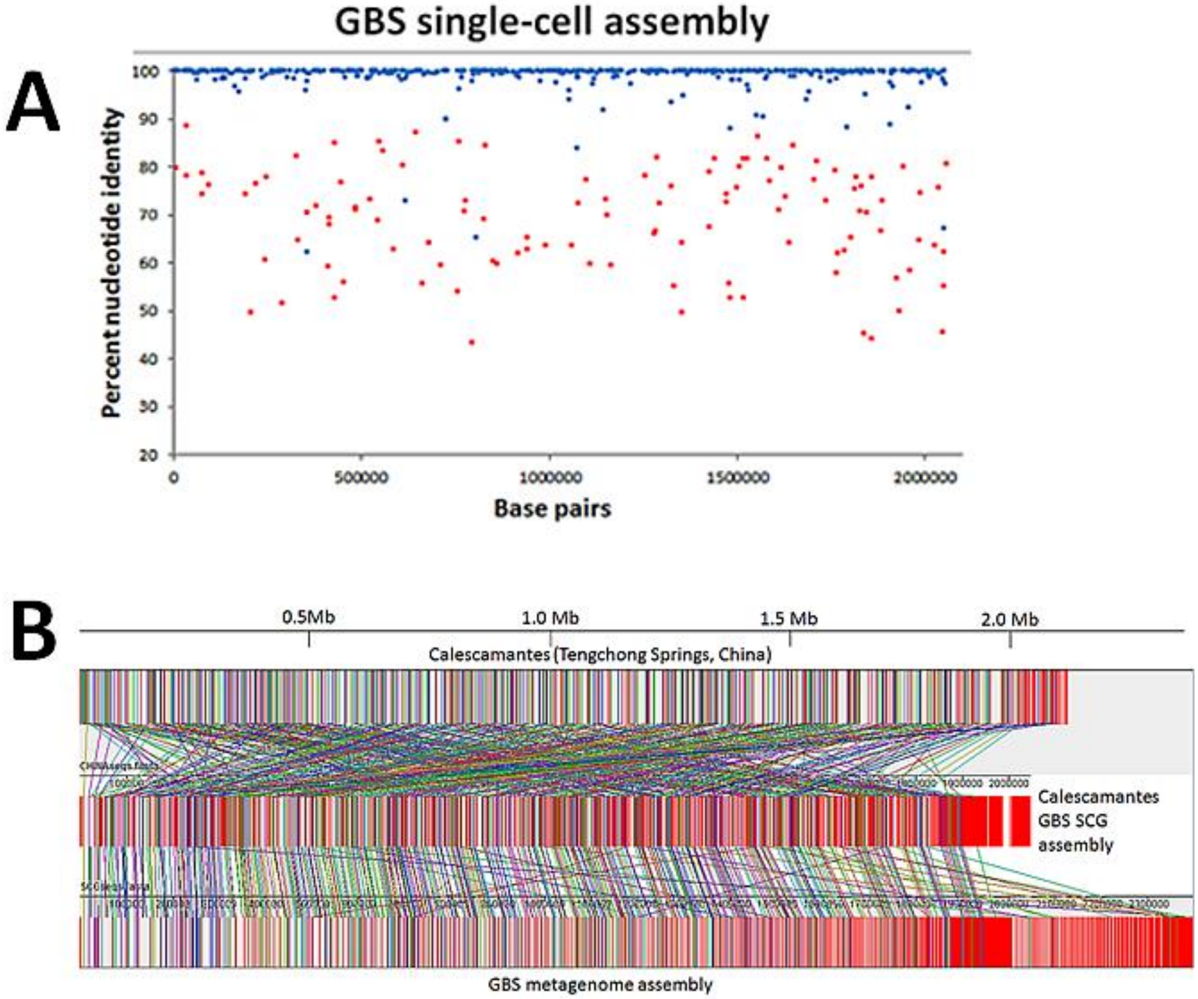
The GBS SAG co-assembly and metagenome assembly were largely syntenous along shared regions of the genome, and most variation between assemblies were hypothetical coding regions that did not have BLASTN hits, and were placed at the end of the alignment. In contrast, the SAG co-assembly was largely non-syntenous compared to the metagome assembly from Gongxiaoshe and the genomes only contained a few semi-syntenous regions.

Supplemental Section II: 16S rRNA gene recovery from metagenome datasets

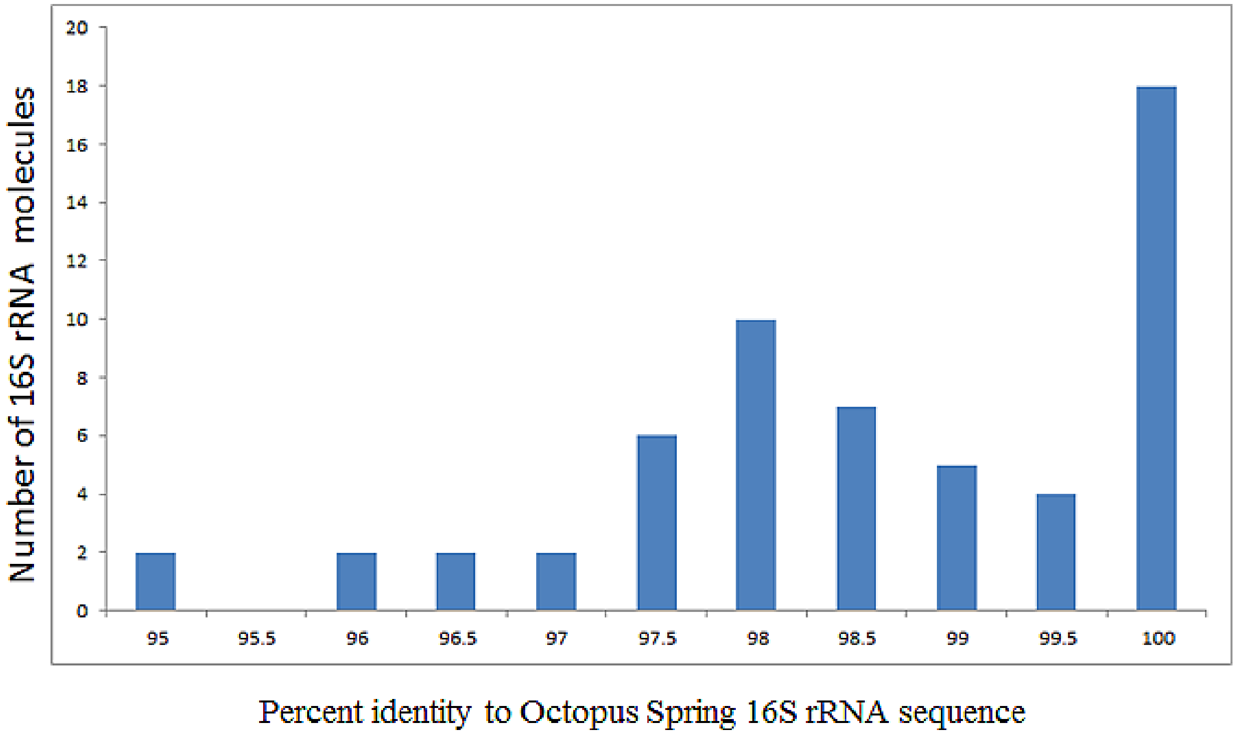
The co-assembled SAG and MLP GBS metagenome assemblies contained one full-length 16S rRNA and one full-length 23S rRNA gene sequence. The binned GXS and Octopus Spring assemblies, on the other hand, contained neither locus, likely because rRNA regions have different selection pressures on their nucleotide word frequencies (1). The recovery of these regions was accomplished using BLASTN with SAG 16S and 23S rRNA gene sequences as queries against the unassembled metagenome. Recovered reads were assembled, yielding full-length 16S and 23S rRNA sequences. Assembled sequences from GBS were 100% identical to the SAG co-assembly rRNA genes (Figure 3).



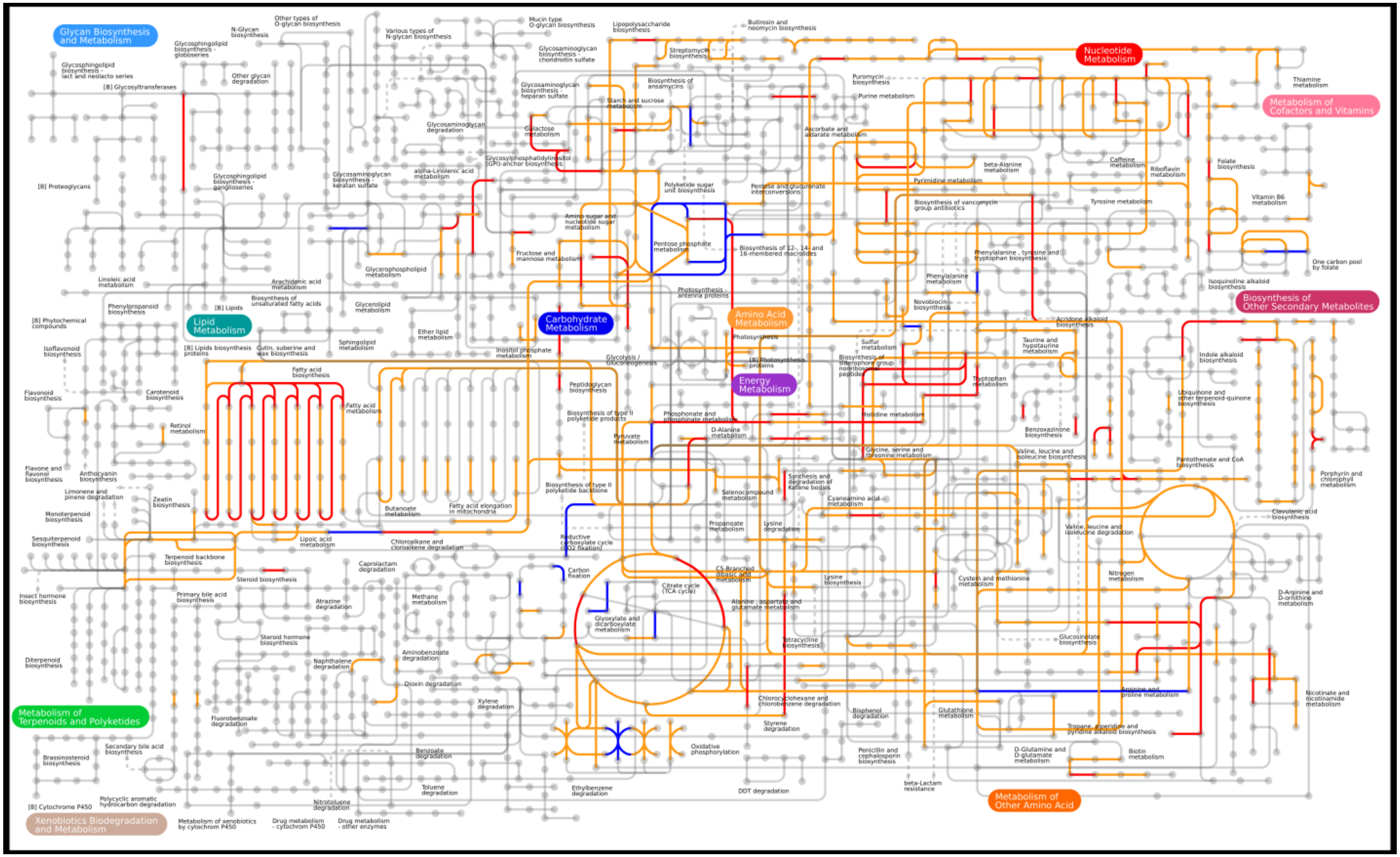
Supplemental Figure 1. Mean divergence of trimer frequencies of Illumina-sized (100 bp, panel A) and 454-sized (500 bp, panel B) genomic fragments randomly selected from single-amplified genomes (SAGs) and other genomes from Genbank for multilayer Perceptron (MLP) training. Model Illumina and 454 fragments were scored by their Euclidean distance to increasingly sized reference fragments selected from the all genomes and separated into self vs. self and self vs. outgroup scores. The stabilization of the mean divergence of trimer frequencies and the standard deviation separation between self and genome out groups appeared to occur when using between 1000 and 5000 bp genomic fragments, indicating that these sized fragements would be optimal for the MLP training algorithm. Error bars represent standard deviation from the mean.



Supplemental Figure 2. (A) Recruitment plot of Great Boiling Spring (GBS; blue dots) and Gongxiaoshe Spring (red dots) predicted protein regions to the GBS SAG co-assembly. (B) MAUVE alignment of Calescamantes Gongxiaoshe assembly (top), and the GBS metagenome assembly (bottom), to the GBS single-amplified genome (SAG) co-assembly (middle). Contigs were ordered by best BLASTN hit and position relative to the largest contig between assemblies. Contigs without BLASTN hits were added to the end of the alignment.

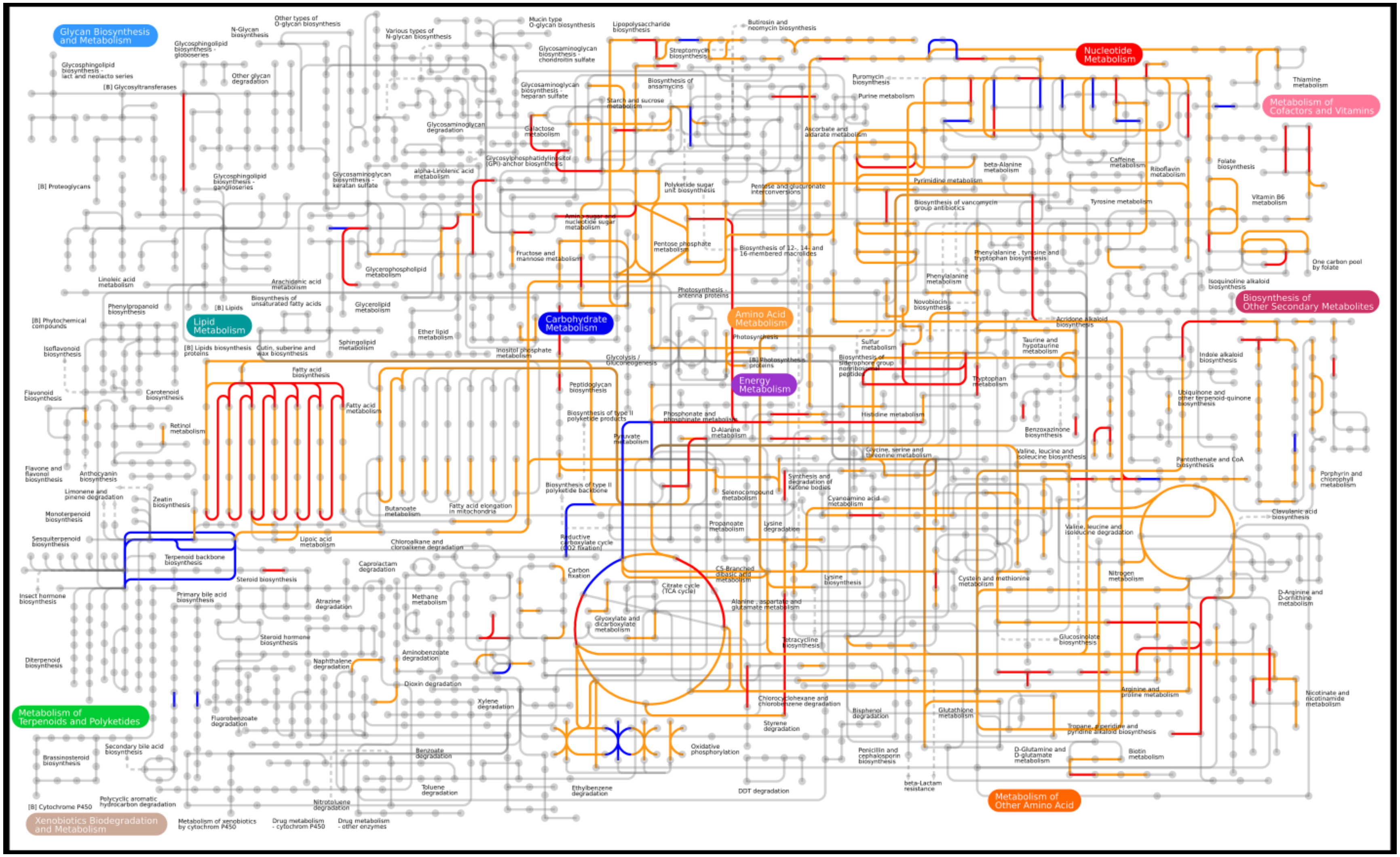


Supplemental Figure 3. 16S rRNA gene metagenomic reads in Octopus Spring aligned to the previously identified 16S rRNA sequence (2).

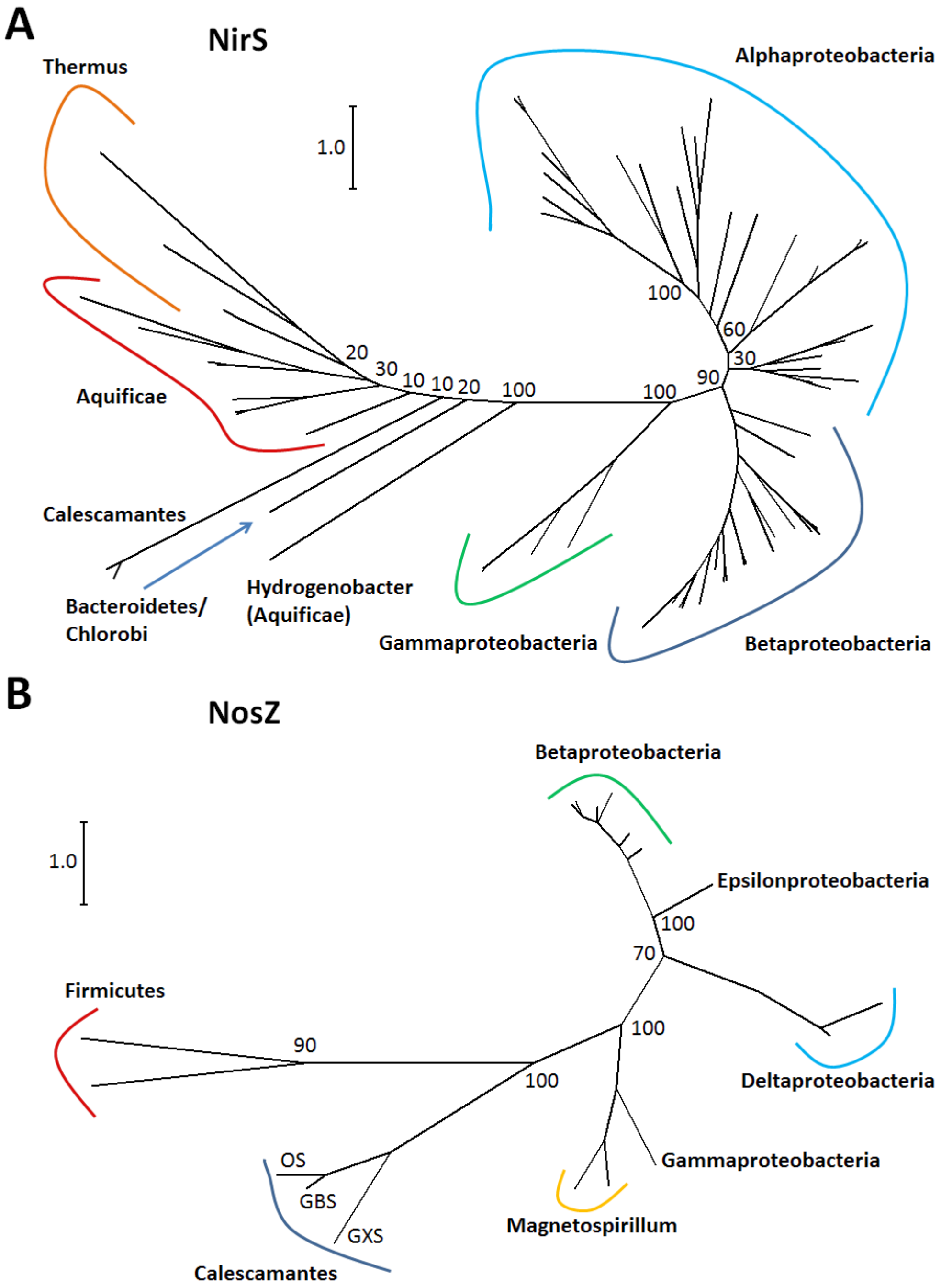


Glycan Biosynthesis and Metabolism
Nucleotide Metabolism
Metabolism of Cofactors and Vitamins
Biosynthesis of Other Secondary Metabolites
Lipid Metabolism
Carbohydrate Metabolism
Amino Acid Metabolism
Energy Metabolism
Metabolism of Terpenoids and Polyketides
Xenobiotics Biodegradation and Metabolism
Metabolism of Other Amino Acid
N-Glycan biosynthesis
Other types of O-glycan biosynthesis
Various types of N-glycan biosynthesis
Mucin type O-glycan biosynthesis
Lipopolysaccharide biosynthesis
Butirosin and neomycin biosynthesis
Streptomycin biosynthesis
Glycosphingolipid biosynthesis - globoseries
Glycosphingolipid biosynthesis - lact and neolacto series
Other glycan degradation
[B] Glycosyltransferases
Glycosaminoglycan biosynthesis - chondroitin sulfate
Glycosaminoglycan biosynthesis - heparan sulfate
Biosynthesis of ansamycins
Starch and sucrose metabolism
Puromycin biosynthesis
Purine metabolism
Thiamine metabolism
Glycosaminoglycan degradation
Galactose metabolism
Ascorbate and aldarate metabolism
[B] Proteoglycans
Glycosphingolipid biosynthesis - ganglioseries
Glycosaminoglycan biosynthesis - keratan sulfate
alpha-Linolenic acid metabolism
Glycosylphosphatidylinositol (GPI)-anchor biosynthesis
Polyketide sugar unit biosynthesis
Pentose and glucuronate interconversions
Pyrimidine metabolism
beta-Alanine metabolism
Caffeine metabolism
Riboflavin metabolism
Folate biosynthesis
Vitamin B6 metabolism
Biosynthesis of vancomycin group antibiotics
Tyrosine metabolism
Amino sugar and nucleotide sugar metabolism
Fructose and mannose metabolism
Pentose phosphate metabolism
Biosynthesis of 12-, 14- and 16-membered macrolides
Phenylalanine , tyrosine and tryptophan biosynthesis
One carbon pool by folate
Linoleic acid metabolism
Arachidonic acid metabolism
Glycerophospholipid metabolism
Photosynthesis - antenna proteins
Phenylalanine metabolism
Isoquinoline alkaloid biosynthesis
Phenylpropanoid biosynthesis
[B] Lipids
Biosynthesis of unsaturated fatty acids
Glycerolipid metabolism
Novobiocin biosynthesis
Acridone alkaloid biosynthesis
[B] Phytochemical compounds
Ether lipid metabolism
Photosynthesis
Sulfur metabolism
Taurine and hypotaurine metabolism
Isoflavonoid biosynthesis
[B] Lipids biosynthesis proteins
Cutin, suberine and wax biosynthesis
Sphingolipid metabolism
Inositol phosphate metabolism
Glycolysis / Gluconeogenesis
[B] Photosynthesis proteins
Biosynthesis of siderophore group nonribosomal peptides
Indole alkaloid biosynthesis
Fatty acid biosynthesis
Tryptophan metabolism
Flavonoid biosynthesis
Carotenoid biosynthesis
Peptidoglycan biosynthesis
Fatty acid metabolism
Ubiquinone and other terpenoid-quinone biosynthesis
Retinol metabolism
Biosynthesis of type II polyketide products
Phosphonate and phosphinate metabolism
Histidine metabolism
Benzoxazinone biosynthesis
D-Alanine metabolism
Pyruvate metabolism
Flavone and flavonol biosynthesis
Anthocyanin biosynthesis
Glycine, serine and threonine metabolism
Valine, leucine and isoleucine biosynthesis
Pantothenate and CoA biosynthesis
Limonene and pinene degradation
Biosynthesis of type II polyketide backbone
Synthesis and degradation of ketone bodies
Selenocompound metabolism
Porphyrin and chlorophyll metabolism
Zeatin biosynthesis
Cyanoamino acid metabolism
Monoterpenoid biosynthesis
Butanoate metabolism
Fatty acid elongation in mitochondria
Propanoate metabolism
Lysine degradation
Clavulanic acid biosynthesis
Valine, leucine and isoleucine degradation
Sesquiterpenoid biosynthesis
Lipoic acid metabolism
Chloroalkane and chloroalkene degradation
Reductive carboxylate cycle (CO2 fixation)
Terpenoid backbone biosynthesis
Caprolactam degradation
Carbon fixation
C5-Branched dibasic acid metabolism
Steroid biosynthesis
Citrate cycle (TCA cycle)
Lysine biosynthesis
Cystein and methionine metabolism
Nitrogen metabolism
Insect hormone biosynthesis
Primary bile acid biosynthesis
Atrazine degradation
Methane metabolism
Alanine , aspartate and glutamate metabolism
D-Arginine and D-ornithine metabolism
Glyoxylate and dicarboxylate metabolism
Steroid hormone biosynthesis
Tetracycline biosynthesis
Glucosinolate biosynthesis
Diterpenoid biosynthesis
Naphthalene degradation
Aminobenzoate degradation
Dioxin degradation
Arginine and proline metabolism
Nicotinate and nicotinamide metabolism
Xylene degradation
Chlorocyclohexane and chlorobenzene degradation
Fluorobenzoate degradation
Bisphenol degradation
Glutathione metabolism
Styrene degradation
Tropane, piperidine and pyridine alkaloid biosynthesis
Secondary bile acid biosynthesis
Oxidative phosphorylation
Penicillin and cephalosporin biosynthesis
D-Glutamine and D-glutamate metabolism
Biotin metabolism
Brassinosteroid biosynthesis
Benzoate degradation
Toluene degradation
Ethylbenzene degradation
DDT degradation
beta-Lactam resistance
[B] Cytochrome P450
Polycyclic aromatic hydrocarbon degradation
Nitrotoluene degradation
Metabolism of xenobiotics by cytochrom P450
Drug metabolism - cytochrom P450
Drug metabolism - other enzymes

Supplemental Figure 4. KEGG metabolic map comparing the predicted pathways of the single amplified genome (SAG) (red) and the Great Boiling Spring (GBS) Multi-layer Perceptron (MLP) metagenome assembly (blue) and shared pathways identified in both (orange). The map was generated using iPath 2.0 (3).



Supplemental Figure 5. KEGG metabolic map comparing the predicted pathways of the Great Boiling Spring (GBS) multilayer perceptron (MLP) assembled genome (blue) and the Gonxiaoshe Spring (GXS) MLP metagenome assembly (red) and shared pathways identified in both (orange). The map was generated using iPath 2.0 (3).



Supplemental Figure 6. Maximum-likelihood phylogenies of the top 100 BLASTP hits in Genbank to the Calescamantes A) NirS and B) NosZ proteins. Bootstrap values are reported for phylum-level nodes.

Supplemental Table 1. IMG IDs and Genbank accession numbers for Calescamantes single assembled genomes (SAGs) and metgenomic data sets. All individual SAG data are also located at http://microbialdarkmatter.org/index.php/mdm-project/4-single-cell-data.

Calescamantes SAG assemblies [1]	**IMG Genome ID**	**Genbank Accession**
Calescamantes bacterium Combined SAG Assembly [2]	2527291514	AWOA00000000.1
Calescamantes bacterium JGI 0000106-I5 (GBS-C_001_287) [3]	2264867083	ASNA00000000.1
Calescamantes bacterium JGI 0000106-I17 (GBS-C_001_286) [3]	2264867082	ASMX00000000.1
Calescamantes bacterium JGI 0000106-M22 (GBS-C_001_290) [3]	2264867085	ASMW00000000.1
Calescamantes bacterium JGI 0000106-N5 (GBS-C_001_291) [3]	2264867086	ASMT00000000.1
Calescamantes bacterium JGI 0000106-P5 (GBS-C_001_294) [3]	2264867088	ASMZ00000000.1
Calescamantes bacterium JGI 0000106-G12 (GBS-C_001_282) [3]	2264867080	AQTE00000000.1
Calescamantes bacterium SCGC AAA471-M6 (GBS-N_001_25) [4]	2264867079	AQST00000000.1
Calescamantes bacterium JGI 0000106-J16 (GBS-C_001_289) [3]	2264867084	ASMV00000000.1
Calescamantes bacterium JGI 0000106-H18 (GBS-C_001_283) [3]	2264867081	ASMY00000000.1
Calescamantes bacterium JGI 0000106-N7 (GBS-C_001_292) [3]	2264867087	ASMU00000000.1
Metagenomes		
Great Boiling Spring sediment metagenome	2053563014	n/a
Gongxiaoshe hot spring sediment metagenome	3300000865	n/a
Octopus hot spring sediment metagenome	3300001339	n/a
Bison Pool sediment metagenome	(1-5)_050719	n/a

[1] SAGs and coassembly from Rinke et al. 2013 (4).

[2] Coassembly included all SAGs except GBS-C_001_282 and GBS-N_001_25, based on ANI of >97% (4).

[3] Obtained from samples of the top ~1 cm of sediment taken from the main pool of Great Boiling Spring (N 40° 39.682' W 119° 21.973', corresponding to site C in (5) on 22 July 2010 (78 °C).

[4] Obtained from samples of the top ~1 cm of sediment taken from the main pool of Great Boiling Spring (N 40° 39.684' W 119° 21.973', corresponding to site B in (5) on 9 February 2010 (79.2 °C).

Supplemental Table 2. Pairwise comparison of average nucleotide identity (ANI) between Calescamantes Great Boiling Spring (GBS) single-amplified genome (SAG) co-assembly and multilayer perceptron (MLP) metagenome assemblies targeted in this study.

	GBS SAG	GBS MLP	Gxs MLP	Oct MLP	Bison MLP
GBS SAG	100				
GBS MLP	99.45	100			
Gxs MLP	76.37	low [1]	100		
Oct MLP	88.42	88.6	low	100	
Bison MLP	85.46	85.71	low	92.51	100

[1] low = too few hits to be accurately calculated (6).

Supplemental Table 3. Citric acid cycle (TCA) enzymes identified by KAAS in the Calescamantes single-amplified genome (SAG) and Multi-Layer Perceptron (MLP) assemblies from Great Boiling Spring.

Reaction	Enzyme	IMG number
PPP → oxaloacetate	phosphoenolpyruvate carboxylase	EM19COM1_02016
oxaloacetate → citrate	citrate synthase	EM19COM1_01336
citrate → isocitrate	accinitate hydratase [1]	EM19COM1_00283
isocitrate → oxalosuccinate	isocitrate dehydrogenase	EM19COM1_01479
oxalosuccinate → oxoglutarate	isocitrate dehydrogenase	EM19COM1_01479
oxoglutarate → succinyl-CoA	2-oxoglutarate ferredoxin oxidoreductase	EM19COM1_01888
succinyl-CoA → succinate	succinyl-CoA synthetase	EM19COM1_00047/48
succinate → fumerate	succinate dehydrogenase	EM19COM1_00693/482
fumerate → malate	fumarate hydratase [a]	EM19COM1_01222
malate → oxaloacetate	malate dehydrogenase	EM19COM1_00179

[1] Enzyme is only present in the SAG co-assembly.

Supplemental Table 4. Lipid markers typically observed in Gram-negative organisms for the Bison Pool, Octopus Spring, Great Boiling Spring (GBS) and Gongxiaoshe Spring (GXS) multi-layer perceptron (MLP) assemblies and the single-amplified genome (SAG) co-assembly identified by (7).

RAST gene number	Gene function	PFAM number
Bison_MLP		
fig|6666666.83684.peg.1651	Peptidase_A8	PF01252
fig|6666666.83684.peg.1654	Peptidase_A8	PF01252
fig|6666666.83684.peg.1569	LpxC	PF03331
fig|6666666.83684.peg.1408	Surf_Ag_VNR	PF07244
GBS_MLP		
fig|6666666.73513.peg.111	FlgH	PF02107
fig|6666666.73513.peg.110	FlgI	PF02119
fig|6666666.73513.peg.709	SecY	PF00344
fig|6666666.73513.peg.298	TatC	PF00902
fig|6666666.73513.peg.1552	LGT	PF01790
fig|6666666.73513.peg.1074	LGT	PF01790
fig|6666666.73513.peg.17	Peptidase_A8	PF01252
fig|6666666.73513.peg.107	Bac_surface_Ag	PF01103
fig|6666666.73513.peg.1151	Bac_surface_Ag	PF01103
fig|6666666.73513.peg.48	OEP	PF02321
fig|6666666.73513.peg.80	OEP	PF02321
fig|6666666.73513.peg.1337	Secretin	PF00263
fig|6666666.73513.peg.441	Secretin	PF00263
fig|6666666.73513.peg.1338	Secretin_N	PF03958
fig|6666666.73513.peg.1337	Secretin_N	PF03958
fig|6666666.73513.peg.1855	LpxC	PF03331
fig|6666666.73513.peg.2004	LpxC	PF03331
fig|6666666.73513.peg.1081	OstA	PF03968
fig|6666666.73513.peg.2187	Surf_Ag_VNR	PF07244
fig|6666666.73513.peg.107	Surf_Ag_VNR	PF07244
fig|6666666.73513.peg.1613	Surf_Ag_VNR	PF07244
fig|6666666.73513.peg.784	LolA	PF03548
GBS_SAG_co-assembly		
fig|6666666.85483.peg.520	FlgH	PF02107
fig|6666666.85483.peg.519	FlgI	PF02119
fig|6666666.85483.peg.1841	SecY	PF00344
fig|6666666.85483.peg.196	TatC	PF00902
fig|6666666.85483.peg.1129	LGT	PF01790
fig|6666666.85483.peg.117	LGT	PF01790
fig|6666666.85483.peg.516	Bac_surface_Ag	PF01103
fig|6666666.85483.peg.1545	Bac_surface_Ag	PF01103
fig|6666666.85483.peg.243	OEP	PF02321
fig|6666666.85483.peg.434	OEP	PF02321
fig|6666666.85483.peg.1798	Secretin	PF00263

fig|6666666.85483.peg.1115	Secretin	PF00263
fig|6666666.85483.peg.1798	Secretin_N	PF03958
fig|6666666.85483.peg.1925	LpxC	PF03331
fig|6666666.85483.peg.1200	OstA	PF03968
fig|6666666.85483.peg.516	Surf_Ag_VNR	PF07244
fig|6666666.85483.peg.1545	Surf_Ag_VNR	PF07244
fig|6666666.85483.peg.1685	LolA	PF03548
GXS_MLP		
fig|6666666.80949.peg.634	FlgH	PF02107
fig|6666666.80949.peg.633	FlgI	PF02119
fig|6666666.80949.peg.1575	SecY	PF00344
fig|6666666.80949.peg.9	TatC	PF00902
fig|6666666.80949.peg.1563	LGT	PF01790
fig|6666666.80949.peg.581	LGT	PF01790
fig|6666666.80949.peg.387	Peptidase_A8	PF01252
fig|6666666.80949.peg.1746	Bac_surface_Ag	PF01103
fig|6666666.80949.peg.355	Bac_surface_Ag	PF01103
fig|6666666.80949.peg.856	OEP	PF02321
fig|6666666.80949.peg.2132	OEP	PF02321
fig|6666666.80949.peg.60	Secretin	PF00263
fig|6666666.80949.peg.640	Secretin	PF00263
fig|6666666.80949.peg.60	Secretin_N	PF03958
fig|6666666.80949.peg.1577	LpxC	PF03331
fig|6666666.80949.peg.629	OstA	PF03968
fig|6666666.80949.peg.1746	Surf_Ag_VNR	PF07244
fig|6666666.80949.peg.355	Surf_Ag_VNR	PF07244
fig|6666666.80949.peg.1797	LolA	PF03548
Octopus Spring_MLP		
fig|6666666.89741.peg.299	FlgI	PF02119
fig|6666666.89741.peg.1359	SecY	PF00344
fig|6666666.89741.peg.1247	TatC	PF00902
fig|6666666.89741.peg.1354	LGT	PF01790
fig|6666666.89741.peg.212	LGT	PF01790
fig|6666666.89741.peg.495	LGT	PF01790
fig|6666666.89741.peg.1874	Bac_surface_Ag	PF01103
fig|6666666.89741.peg.258	OEP	PF02321
fig|6666666.89741.peg.1557	Secretin	PF00263
fig|6666666.89741.peg.443	LpxC	PF03331
fig|6666666.89741.peg.1001	OstA	PF03968
fig|6666666.89741.peg.1874	Surf_Ag_VNR	PF07244

Supplemental references:

1. **Wang HC, Hickey DA.** 2002. Evidence for strong selective constraint acting on the nucleotide composition of 16S ribosomal RNA genes. Nucleic Acids Res **30:**2501-2507.
2. **Reysenbach AL, Wickham GS, Pace NR.** 1994. Phylogenetic analysis of the hyperthermophilic pink filament community in Octopus Spring, Yellowstone National Park. Appl Environ Microbiol **60:**2113-2119.
3. **Yamada T, Letunic I, Okuda S, Kanehisa M, Bork P.** 2011. iPath2.0: interactive pathway explorer. Nucleic Acids Research **39:**W412-W415.
4. **Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng J-F, Darling A, Malfatti S, Swan BK, Gies EA, Dodsworth JA, Hedlund BP, Tsiamis G, Sievert SM, Liu W-T, Eisen JA, Hallam SJ, Kyrpides NC, Stepanauskas R, Rubin EM, Hugenholtz P, Woyke T.** 2013. Insights into the phylogeny and coding potential of microbial dark matter. Nature **499:**431-437.
5. **Cole JK, Peacock JP, Dodsworth JA, Williams AJ, Thompson DB, Dong H, Wu G, Hedlund BP.** 2012. Sediment microbial communities in Great Boiling Spring are controlled by temperature and distinct from water communities. The ISME Journal **7:**718-729.
6. **Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, Tiedje JM.** 2007. DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. Int J Syst Evol Microbiol **57:**81-91.
7. **Sutcliffe IC.** 2010. A phylum level perspective on bacterial cell envelope architecture. Trends in Microbiology **18:**464-470.