

Lawrence Berkeley National Laboratory

LBL Publications

Title

Global metagenomic survey reveals a new bacterial candidate phylum in geothermal springs

Permalink

<https://escholarship.org/uc/item/0pw5c72r>

Journal

Nature Communications, 7(1)

ISSN

2041-1723

Authors

Eloe-Fadrosh, Emiley A

Paez-Espino, David

Jarett, Jessica

et al.

Publication Date

2016

DOI

10.1038/ncomms10476

Peer reviewed

1Global metagenomic survey reveals a new bacterial candidate phylum in

2geothermal springs

3

4Emiley A. Eloë-Fadrosh¹, David Paez-Espino¹, Jessica Jarett¹, Peter F. Dunfield², Brian

5P. Hedlund³, Anne E. Dekas⁴, Stephen E. Grasby⁵, Allyson L. Brady⁶, Hailiang Dong⁷,

6Brandon R. Briggs⁸, Wen-Jun Li⁹, Danielle Goudeau¹, Rex Malmstrom¹, Amrita Pati¹,

7Jennifer Pett-Ridge⁴, Edward M. Rubin^{1,10}, Tanja Woyke¹, Nikos C. Kyrpides¹, Natalia N.

8Ivanova^{1*}

9

10Affiliations:

11¹Joint Genome Institute, Walnut Creek, CA 94598, USA.

12²University of Calgary, Calgary, AB T2N 1N4, Canada.

13³University of Nevada, Las Vegas, Las Vegas, NV 89154, USA.

14⁴Lawrence Livermore National Laboratory, Livermore, CA 94550, USA.

15⁵Geological Survey of Canada, Calgary, AB, Canada.

16⁶McMaster University, Hamilton, ON L8S 4L8, Canada.

17⁷Miami University, Oxford, OH, 45056, USA.

18⁸University of Alaska-Anchorage, Anchorage, AK, USA.

19⁹Sun Yat-Sen University, Guangzhou, 510275, China.

20¹⁰Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA.

21

22*Correspondence to: nnivanova@lbl.gov.

23

24Abstract

25Analysis of the increasing wealth of metagenomic data collected from diverse

26environments can lead to the discovery of novel branches on the tree of life. Here we

27analyze 5.2 Tb of metagenomic data collected globally to discover a novel bacterial

28phylum ('*Candidatus* Kryptonion') found exclusively in high-temperature pH-neutral

29geothermal springs. This lineage had remained hidden as a taxonomic "blind spot" due

30to mismatches in the primers commonly used for ribosomal gene surveys. Genome

31reconstruction from metagenomic data combined with single-cell genomics results in

32several high-quality genomes representing four genera from the new phylum. Metabolic

33reconstruction indicates a heterotrophic lifestyle with conspicuous nutritional

34deficiencies, suggesting the need for metabolic complementarity with other microbes.
35Co-occurrence patterns identifies a number of putative partners, including an uncultured
36*Armatimonadetes* lineage. The discovery of *Kryptonia* within previously studied
37geothermal springs underscores the importance of globally sampled metagenomic data
38in detection of microbial novelty, and highlights the extraordinary diversity of microbial
39life still awaiting discovery.

40
41
42
43
44
45
46

47**Introduction**

48Molecular environmental surveys have provided a sizeable snapshot of microbial
49phylogenetic diversity. Sequencing of small subunit ribosomal RNA (SSU rRNA) genes
50directly from the environment has expanded the known microbial tree of life from
51Woese's original twelve phyla to more than 70 bacterial phyla^{1,2}. Advances in cultivation-
52independent methods for examining uncultured microbes, including single-cell genomics
53and deep sequencing of environmental samples, have begun yielding complete or near-
54complete genomes from many novel lineages³⁻¹⁰. These approaches have already led to
55the recovery of genomic information from a wealth of candidate lineages (phylogenetic
56lineages for which a cultured representative is not available), notably the
57Lokiarchaeota¹¹, Pacearchaeota and Woesearchaeota¹⁰, and members of the Candidate
58Phyla Radiation³. These lineages, previously recognized only through SSU rRNA data
59and residing in poorly sampled habitats, are providing a more complete topology of the
60tree of life.

61 More recently, it has been suggested that a wealth of novel bacterial and
62archaeal clades exist that are systematically under-represented (the 'rare biosphere') or
63missed altogether in classical surveys, leaving significant taxonomic "blind spots"¹².
64Compared to many of the proposed candidate phyla for which SSU rRNA gene
65information exists, these taxonomic "blind spots" are uncharted lineages with potentially
66important ecological and evolutionary implications. Further, these lineages may be
67highly abundant and hold important metabolic or functional roles within the community,
68yet have been overlooked thus far in ecological surveys. Metagenome sequencing is
69uniquely suited for uncovering taxonomic "blind spots" since it does not suffer from
70biases introduced during PCR amplification, and has limitations only with insufficient
71resolution of minor populations within a community. However, an exploration of the
72complete compendium of available metagenomic sequences for the presence of
73taxonomic "blind spots" has yet to be performed¹³. Here, we report the results of large-
74scale mining of metagenomic data and single cell genomics, which led to the discovery
75of a new bacterial phylum in geographically distinct geothermal springs.

76

77**Results**

78***Identification of a novel bacterial candidate phylum***

79To cast a global net for the discovery of novel microbial lineages in the absence of
80biases introduced via PCR amplicon-based surveys, we collected long assembled
81contigs (≥ 100 kbp) from a comprehensive collection of 4,290 metagenomic datasets
82available through the Integrated Microbial Genomes with Microbiome Samples (IMG/M),
83a database containing a total of more than 5 Tb of sequence data¹⁴. From these data,
8431,955 assembled contigs were identified and 744 contigs were further selected that
85contained SSU rRNA gene fragments greater than 100 bp (Fig. 1A). The SSU rRNA

86gene sequences were then aligned and phylogenetically placed on a reference tree
87consisting of high-quality SSU rRNA sequences from Bacteria and Archaea^{15,16}.
88Exploration of the constructed SSU rRNA tree for novel phylogenetic branches led to
89the identification of a distinct lineage consisting of a full-length SSU rRNA sequence. A
90subsequent search against all assembled metagenomic data identified three additional
91full-length SSU rRNA sequences. The four SSU rRNA gene sequences were from four
92geographically distant, high-temperature, pH-neutral, geothermal springs in North
93America and Asia (Fig. 1). These sequences shared an average 97.4% identity (\pm
941.97% s.d.), and showed a maximum identity of only 83% to SSU rRNA genes (such as
95the one in GenBank ID: AP011715) in NCBI's Non-Redundant (NR) database. In line
96with the notion of taxonomic "blind spots"¹², a comparison of 'universal' SSU rRNA
97primer sets typically used for full-length and hypervariable region amplification with the
98four novel sequences indicated numerous mismatches, explaining why members of this
99lineage likely eluded detection in previous microbial diversity surveys (Supplementary
100Fig. 1; Supplementary Table 1).

101 Phylogenetic analysis of the four SSU rRNA genes placed the newly discovered
102lineage into a monophyletic branch within the *Fibrobacteres-Chlorobi-Bacteroidetes*
103(FCB) superphylum^{9,17} (Supplementary Fig. 2). Based on suggested thresholds for SSU
104rRNA sequence identity to distinguish new phyla^{2,18}, we propose that this lineage
105represents a new bacterial candidate phylum (Supplementary Table 2).

106

107 **Comparative genomics and cell morphology of novel FCB lineage**

108 Reassembly of the metagenomic data combined with tetranucleotide-based

109binning methods using the initial contigs containing the SSU rRNA genes yielded near-
110complete recovery of four distinct genomes, each from one of the four spring samples

111(Supplementary Fig. 3; Supplementary Table 3). Phylogenetic analysis of conserved
112marker genes supported its placement as a sister phylum to the *Ignavibacteria* with
113100% bootstrap support (Fig. 2A; Supplementary Fig. 4). Three of the genomes
114reconstructed from metagenomes (GFMs) from Dewar Creek Spring, Canada¹⁹, Great
115Boiling Spring, Nevada^{20,21}, and Gongxiaoshe pool, Yunnan Province, China²² had an
116average 95.8% estimated coverage, while the genome from Jinze pool, Yunnan
117Province, China²² had a lower estimated coverage of 68% (Supplementary Table 4). The
118high genomic sequence coverage across the four metagenomes (average 31.2x
119coverage; Supplementary Table 3) suggested that this novel lineage might exist at
120sufficient cell abundance to be captured by single cell technology. We therefore
121employed high-throughput single-cell isolation, whole-genome amplification and SSU
122rRNA screening of single amplified genomes (SAGs) in search for the novel lineage
123(Fig. 1). We successfully recovered a total of 18 SAGs from three of the four samples,
124corresponding to the novel phylum-level clade with an estimated average genome
125completeness of 67.2% (± 20.1 s.d.) (Supplementary Table 3). We designate this new
126candidate phylum '*Candidatus* Kryptonia,' from the Greek word '*krupton*' meaning
127hidden or secret since it has hitherto eluded detection due to SSU rRNA primer biases
128(Supplementary Table 4).

129 The average nucleotide identity (ANI) based metric, Microbial Species Identifier
130(MiSI), was used to compare the four '*Ca. Kryptonia*' genomes reconstructed from
131metagenomes (GFMs) and the 18 SAGs²³. This analysis revealed that almost all of the
132genotypes extracted from the same sample belonged to a single species
133(Supplementary Data 1). For example, the GFM reconstructed from Dewar Creek ('*Ca.*
134*Kryptonium thompsoni*' JGI-4) and the thirteen SAGs ('*Ca. Kryptonium thompsoni*' JGI-5

135– JGI-17) collected from the same site shared an ANI of 99.67% (\pm 0.15 s.d.) and
136represent a single coherent species²³. A single exception to the above observations was
137the recovery of a divergent ‘Ca. Kryptonion’ SAG (‘Ca. Chrysopegis kryptomonas’ JGI-
13823) from the Jinze pool, Yunnan Province, China representing a population distinct from
139the other two SAGs recovered from this site (‘Ca. Kryptobacter tengchongensis’ JGI-24
140and JGI-25) (Supplementary Data 1). Across the four geothermal springs, the GFMs
141and SAGs collectively share average ANIs of only 78.86% (\pm 1.42 s.d.), suggesting that
142they represent different genera of ‘Ca. Kryptonion’. Further support for genus-level
143designations is evident from nuanced functional and metabolic differences across the
144genomes, as described below.

145 In addition to recovering single cells of ‘Ca. Kryptonion’ for genome amplification,
146we designed a SSU rRNA-targeted fluorescence *in situ* hybridization (FISH) probe to
147visualize cell morphology (Fig. 2B). The targeted ‘Ca. Kryptonion’ cells appeared
148filamentous, and exhibited morphological heterogeneity ranging from short to elongated
149filaments. These findings are consistent with numerous reports describing filamentous
150thermophilic bacteria, most notably cultivated members of the sister phylum
151*Ignavibacteria* that range in length from 1 μ m to greater than 15 μ m^{24,25}.

152
153**Unique CRISPR-Cas fusion and limited biogeographic distribution**

154 CRISPR (clustered regularly interspaced short palindromic repeats) elements
155and cas (CRISPR-associated) genes across the ‘Ca. Kryptonion’ genomes were
156recovered, and are suggestive of defense against viral attack. A novel fusion between
157two different CRISPR-Cas types (type I and III; subtypes I-B and III-A) was identified in
158all genomes. This unusual fusion contained the full gene set for components
159responsible for the multistep CRISPR processes for spacer acquisition, CRISPR locus

160transcription and maturation, and final nucleic acid interference^{26,27} (Supplementary Fig.
1615). This observation represents the first report of a type I-B/type III-A CRISPR-Cas
162fusion and expands the known genetic diversity of CRISPR-Cas loci. Based on
163reconstruction of repeat-spacer arrays, the 'Ca. Kryptonium thompsoni' genomes
164appear to represent a clonal CRISPR population without active spacer acquisition, while
165the 'Ca. Kryptobacter tengchongensis' genomes are considerably dynamic in terms of a
166mosaic spacer collection (Supplementary Note 1; Supplementary Data 2 and 3). These
167findings suggest that the CRISPR-Cas encoded by 'Ca. Kryptobacter tengchongensis' is
168highly active, while the 'Ca. Kryptonium thompsoni' genomes are not actively acquiring
169spacers through the CRISPR-Cas system.

170 To verify the limited biogeographic distribution of 'Ca. Kryptonia,' we
171systematically surveyed the collection of 640 Gb of assembled metagenomic data from
1724,290 environmental samples (including 169 samples from geothermal springs and
173hydrothermal vents) for the presence of a genomic signature beyond our initial search
174using SSU rRNA fragments from 100 kbp contigs (Fig. 3; Supplementary Data 4).
175Further, we searched against all available SSU rRNA data from the SILVA database¹⁶ for
176additional 'Ca. Kryptonia' phylotypes and did not recover a highly similar match. Using
177this expanded search, we found evidence for 'Ca. Kryptonia' in a total of twenty
178metagenomes, which included only three additional geographic sites compared to our
179initial SSU rRNA survey (Supplementary Data 4). The environments where this phylum
180was found were similar to the settings where we first discovered the genomic presence
181of 'Ca. Kryptonia': all were high-temperature ($\geq 70^{\circ}\text{C}$), pH-neutral (6.4 – 8.0) settings. In
182sum, the limited range of 'Ca. Kryptonia' is reflected in the observation that genomic
183signatures were found in nine unique geographical locations from a total of twenty-three

184pH neutral hot springs currently sampled by metagenomics, and absent from the 1,614

185unique locations represented by 4,290 metagenomic samples.

186 Additional metagenomic searches specific for all CRISPR repeat-spacer arrays

187collected from the 'Ca. Kryptonion' genomes resulted in a similar pattern of limited

188biogeographic distribution (Fig. 3; Supplementary Data 3). We identified shared spacers

189across 'Ca. Kryptonion' populations in geographically distinct geothermal springs. For

190example, shared spacers were identified between the 'Ca. Kryptobacter

191tengchongensis' JGI-2 and JGI-3 genomes despite sampling from separate geothermal

192pools in China. Further, shared spacers were identified across exceptionally wide

193geographic distances including Canada and Nevada ('Ca. Kryptonion thompsoni' JGI-4

194and the Great Boiling Springs metagenome), and China and Nevada ('Ca. Kryptobacter

195tengchongensis' JGI-2 and the Great Boiling Springs metagenome) (Fig. 3).

196Remarkably, we also found spacer matches to a set of metagenomic contigs that we

197assigned as viral due to their linkage to known viral genes, from these same samples

198and metagenome samples collected from Yellowstone National Park²⁸ (Fig. 3;

199Supplementary Note 1; Supplementary Fig. 5; Supplementary Data 5). These genomic

200recruitment and spacer signature data suggest that 'Ca. Kryptonion' is present in

201additional geothermal spring sites and that viruses which appear to infect 'Ca. Kryptonion'

202circulate across wide geographic space as revealed from the conserved infection

203vestiges.

204

205***Metabolic potential of 'Candidatus Kryptonion'***

206 The availability of multiple nearly complete 'Ca. Kryptonion' genomes from both

207GFMs and SAGs enabled metabolic and putative functional predictions for this novel

208candidate phylum, as well as insights into some of the unique properties and notable

209absence of function for the individual genera. Approximately 50% of the predicted
210composite proteome for the 'Ca. Kryptonia' genomes showed similarity to a diverse
211array of FCB superphylum members, with 11.3% and 1.96% best matches to
212thermophilic members of the phylum *Ignavibacteria* and *Caldithrix abyssi*, respectively
213(Supplementary Fig. 6). The conserved Por secretion system C-terminal sorting domain
214(TIGR04183), found exclusively in members of the FCB superphylum⁹, was recovered in
215all GFMs and SAGs, and altogether totaled 811 predicted proteins across the 'Ca.
216Kryptonia' genomes. Reverse gyrase, the presumptive gene indicator for the extreme
217thermophilic and hyperthermophilic lifestyle in bacteria and archaea²⁹, was found in all
218'Ca. Kryptonia' genomes, which suggests that most, if not all members, of this lineage
219are extreme thermophiles or hyperthermophiles. Further, we found evidence for
220horizontal gene transfer of the reverse gyrase from the crenarchaeal order
221*Thermoproteales* (Supplementary Note 2; Supplementary Fig. 7) and hypothesize that
222'Ca. Kryptonia's' thermophilic traits might have been acquired via lateral gene transfer
223rather than ancestral inheritance.
224 'Ca. Kryptonia' is a motile heterotroph with a complete tricarboxylic acid cycle
225and key metabolic enzymes for Embden-Meyerhof glycolysis and the pentose
226phosphate pathway. We found evidence for a complex oxidative phosphorylation
227pathway, which points towards aerobic respiration (Fig. 4; Supplementary Data 6). An
228elaborate and unique respiratory pathway for the redox transformation of iron is
229encoded in the 'Ca. Kryptonia' genomes with similar, yet non-homologous components
230to the well-characterized Mtr-like respiratory pathway³⁰ (Supplementary Fig. 8).
231Altogether, 'Ca. Kryptonia' has the machinery to carry out ferric iron respiration under

232thermophilic conditions and likely vies with archaeal community members to impact
233metal biogeochemistry in these geothermal springs.
234 'Ca. Kryptonium' hosts the genomic potential for aromatic hydrocarbon degradation
235via oxidation to catechol, and subsequent catechol meta-cleavage (Fig. 4). Further, the
236'Ca. Kryptonium thompsoni' genomes encode a putative gene complement for the
237anaerobic degradation of aromatic amino acids or similar compounds, notably
238represented by a phenylacetyl-CoA oxidoreductase homologous to the
239hyperthermophilic archaeon *Ferroglobus placidus*³¹. This feature appears to be the first
240example of an extremely thermophilic or hyperthermophilic bacterium with the
241presumptive capacity to completely mineralize aromatic compounds, and holds
242biotechnological potential as well as implications for carbon cycling within geothermal
243springs³².

244

245***Unexpected metabolic deficiencies identified in 'Ca. Kryptonium'***

246An unexpected observation was that all 'Ca. Kryptonium' genomes had conspicuous
247nutritional deficiencies, displaying gene loss for many biosynthetic pathways, including
248thiamine, biotin and amino acids, such as the evolutionarily conserved histidine
249biosynthesis³³ (Fig. 4; Supplementary Data 6). While obligately host-dependent
250microbes and some free-living organisms with reduced genomes are known to omit a
251suite of anabolic pathways^{34,35}, the 'Ca. Kryptonium' genomes do not appear to have
252signatures of either lifestyle. An analysis of 759 high-quality FCB superphylum genomes
253indicate the near-complete 'Ca. Kryptonium' genomes are distinct from free-living
254microbes in terms of amino acid pathway coverage and genome size, yet are not highly
255reduced compared to obligate symbionts (Supplementary Fig. 9). These findings

256 suggest that 'Ca. Kryptonina' has potentially evolved functional dependency on other
257 microbes in order to acquire necessary metabolic requirements.
258 To explore the existence of possible microbial partners, we performed a co-
259 occurrence analysis of SSU rRNA sequences retrieved through their targeted assembly
260 from an expanded set of 22 geothermal springs metagenomes (Supplementary Note 3;
261 Supplementary Table 5). An analysis of co-occurrence patterns for clusters of
262 taxonomically coherent groups (clustered at 90% sequence identity) revealed a subset
263 of taxonomically clustered groups (phylotypes) highly correlated with the abundance of
264 'Ca. Kryptonina' (Supplementary Table 6). These clusters included an *Armatimonadetes*
265 lineage, which had the highest correlation value, three separate lineages of *Chloroflexi*,
266 and *Thermus* spp. (Fig. 5). For the twelve metagenomes in which 'Ca. Kryptonina's' SSU
267 rRNA was reconstructed, the *Armatimonadetes* lineage was found to co-occur in seven
268 of those metagenomes at similar sequence coverage to the 'Ca. Kryptonina' genomes,
269 and was conspicuously absent across all other metagenomes surveyed. To explore the
270 potential of the *Armatimonadetes* lineage to complement the metabolic deficiencies
271 identified in 'Ca. Kryptonina,' we reconstructed three nearly complete genomes of
272 *Armatimonadetes* (Fig. 2; Supplementary Table 3; Supplementary Data 7) to infer
273 metabolic potential and signatures of possible metabolic exchange and interaction.
274 Analysis of the reconstructed genomes identified metabolic features complementary to
275 those of 'Ca. Kryptonina,' such as histidine, cysteine and methionine, proline, aspartic
276 acid, and thiamine biosynthesis, and degradation of pentoses (Fig. 5B; Supplementary
277 Note 4; Supplementary Data 7). Furthermore, in the reconstructed *Armatimonadetes*
278 genomes we also identified a CsgG family protein, which forms transmembrane
279 channels for secretion of "functional amyloids," a class of bacterial proteins capable of

280assembling highly stable fibers through a nucleation-precipitation mechanism³⁶.
281“Functional amyloids” play major roles in adhesion to surfaces and biofilm formation in
282diverse bacteria including *Escherichia coli*, *Caulobacter crescentus* and *Bacillus*
283*subtilis*³⁷. Further, the CsgG-like transporter was located in a six-gene conserved cluster
284containing a predicted subtilase-family peptidase and a putative secreted protein with
285four copies of a “carboxypeptidase regulatory-like domain” (Pfam13620)
286(Supplementary Fig. 10). This domain is a member of the transthyretin clan and has
287been found to form amyloid in physiological conditions³⁸. We hypothesize that this
288cluster in the *Armatimonadetes* genomes encodes for synthesis, secretion and
289assembly of “functional amyloid,” in which other members of the community may be
290embedded. On the other hand, the ‘Ca. Kryptonia’ genomes encode many proteases
291and peptidases, which may be responsible for remodeling and digestion of this
292extracellular matrix.

293 Other co-occurring lineages with ‘Ca. Kryptonia’ include the *Thermus* spp. cluster
294(Supplementary Table 6). Interestingly, ‘Ca. Kryptonia’ might complement an incomplete
295denitrification pathway in *Thermus* spp., which may be responsible for high rates of
296nitrous oxide production^{39,40}. *Thermus* spp. have been experimentally characterized to
297reduce nitrate to nitrous oxide but lack the capacity to subsequently produce
298dinitrogen^{39,40}. ‘Ca. Kryptonia’ encodes a nitrous oxide reductase (EC 1.7.2.4) but lacks
299other components of the denitrification pathway (Supplementary Note 5; Supplementary
300Table 7). Taken together, we hypothesize that ‘Ca. Kryptonia’ may participate in a
301partnership with other organisms, such as the *Armatimonadetes*, or might interact with a
302broader consortium of microbes within the geothermal spring environment.

303

304Discussion

305A comprehensive survey of a global set of assembled metagenomic data for novel
306microbial lineages has resulted in the discovery of a new bacterial candidate phylum in
307geothermal springs. The high-quality draft genome assemblies enabled by
308complementary approaches from metagenomic data and single-cell genomics data for
309‘Ca. Kryptonita’ facilitated delineation of the host-virus interaction across geographically
310distant sites. Further, we observed a novel fusion between two different CRISPR-Cas
311types, representing the first report of a type I-B/type III-A CRISPR-Cas fusion and
312expanded the known genetic diversity of CRISPR-Cas loci.

313 The metabolic capacity for ‘Ca. Kryptonita’ provides evidence for a unique
314heterotrophic lifestyle with the putative capacity for iron respiration within a consistent
315ecological niche in geothermal springs. An unexpected observation was that all ‘Ca.
316Kryptonita’ genomes had conspicuous nutritional deficiencies, which led to the
317hypothesis of a microbial partnership or interaction with a broader consortium of
318microbes. Subsequent genome reconstruction of genomes from a co-occurring
319*Armatimonadetes* lineage indicated potential complementarity for those metabolic
320features presumably absent in ‘Ca. Kryptonita.’ It is well recognized that certain marine
321microbes, such as SAR11 (ref.⁴¹) and SAR86 (ref.⁴²), lack a variety of anabolic pathways
322and likely rely on other microbial community members to supplement their
323requirements. Within geothermal springs, the growth of chlorophototroph *Candidatus*
324*Chloracidobacterium thermophilum* in the laboratory was shown to depend upon two
325heterotrophs, *Anoxybacillus* and *Meiothermus* spp., due to lack of biosynthetic
326pathways for branched-chain amino acids, lysine and cobalamin⁴³. Our study suggests
327that dependency on other organisms within the geothermal spring community might be

328a more common occurrence than previously appreciated, perhaps contributing to
329challenges in obtaining many of these lineages as isolated monocultures. Future efforts
330to delineate this hypothesized interaction, particularly utilizing microscopy methods to
331visualize these uncultivated cells *in situ*, will further contribute to our understanding of
332'Ca. Kryptonia' and its role within the environment.

333 Geothermal springs have been heavily surveyed as a rich source of novel
334microbial branches on the tree of life^{18,44}, yet our results indicate that additional
335phylogenetic novelty has yet to be captured from these environments. The discovery of
336a new candidate phylum emphasize that extraordinary microbial novelty is likely still
337awaiting discovery using the vast metagenomic data assembled from locations sampled
338globally.

339

340

341

342

343

344**Methods**

345**Metagenomes**

346All publicly available metagenome datasets from IMG/M were used in the study (data
347accessed September 8, 2014)¹⁴. The metagenomes can be accessed at

348<http://img.jgi.doe.gov> and associated metadata can be found in the GOLD database at

349<http://genomesonline.org>.

350**Metagenomic binning**

351Tetranucleotide-based binning methods were implemented as previously described to

352recover near-complete genomes from metagenomes⁴⁵. Both single metagenomes and

353combined metagenome assemblies were used to recruit additional contigs that

354harbored the same tetranucleotide signature, and the raw reads were subsequently re-

355assembled using SPAdes version 3.1.0 (ref.⁴⁶).

356**SAG generation**

357 Sediment samples were collected from Dewar Creek hot spring (49.9543667°,
358-116.5155000°) near the source of the hot spring on September 28, 2012, from the
359 Jinze pool (25.44138°, 98.46004°) on August 12, 2012, and from the Gongxiaoshe pool
360 (25.44012°, 98.44081°) on August 9, 2011. Samples were mixed with 4% DMSO in TE
361 buffer (1 mM EDTA, 10 mM Tris) for cryopreservation and stored at -80°C within 24
362 hours of sample collection. Single cells were isolated using FACS, lysed, and subjected
363 to whole-genome amplification (WGA) as previously described⁹ with the following
364 modifications: the alkaline lysis was preceded by a 20 min digest with lysozyme
365 (Epicentre) at 30°C; WGA was performed with a REPLI-g Single Cell Kit (Qiagen) with a
366 scaled-down reaction volume of 2 µl; and the amplification reaction was incubated for 6
367 hr at 30°C. WGA reactions were diluted 10-fold, then aliquots were further diluted 200-
368 fold for PCR screening targeting the V6-V8 regions (Forward primer: 926wF
369 (GAAACTYAAAKGAATTGRCGG) and Reverse primer: 1392R
370 (ACGGGCGGTGTGTRC)) of the SSU rRNA using a QuantiNova SYBR Green PCR kit
371 (Qiagen) for 45 cycles of amplification⁹. PCR products were purified and sequenced,
372 and SAGs matching 'Ca. Kryptonia' SSU rRNA sequences were selected for shotgun
373 sequencing.

374 **SAG sequencing, assembly and QC**

375 Draft genomes for the eighteen SAGs were generated at the DOE Joint Genome
376 Institute (JGI) using the Illumina MiSeq technology according to standard protocols
377 (<http://www.jgi.doe.gov/>). Assembly was performed using SPAdes version 3.1.0 (ref.⁴⁶)
378 using the --sc flag to denote MDA-derived data to account for uneven coverage of the
379 single-cell genomes. Quality control and contaminant removal from the resultant
380 assemblies was achieved using a two-step process. First, all assembled reads were

381used as input for a newly developed single-cell decontamination method (ProDeGe)⁴⁷,
382which uses both taxonomic and k-mer based decisions to flag putative non-target
383contigs. Since the taxonomic information was limited to phylum-level designations, we
384further supplemented this procedure with direct mapping to the genomes reconstructed
385from metagenomic data. For mapping, a combination of blast and blat were
386implemented to validate correct recruitment of the assembled SAG contigs to 'Ca.
387Kryptonia'-specific GFM scaffolds. This method was important for retaining
388CRISPR/Cas genetic regions since ProDeGe had the tendency to flag these contigs
389based on divergent k-mer frequencies. Gene annotation was performed within the
390Integrated Microbial Genomes (IMG) platform developed by the DOE Joint Genome
391Institute¹⁴.

392**SSU rRNA phylogeny**

393Full-length SSU rRNA gene sequences from 'Ca. Kryptonia' were aligned using the
394SINA aligner¹⁵ to a comprehensive database of references (SILVA-NR version 119)¹⁶. A
395total of 187 full-length bacterial and archaeal reference sequences were selected based
396on taxonomic breadth from the SILVA database, and 1,354 distinct alignment patterns
397were used, and filtered using the *E. coli* positional mask. A maximum likelihood tree was
398calculated from the masked alignments with 100 bootstrap resamplings using the
399Generalized Time-Reversible model with G+I options in RAxML version 7.6.3
400(raxmlHPC-PTHREADS-SSE3 -f a -x 12345 -p 12345 -# 100 -T 5 -m GTRGAMMAI)⁴⁸.
401To resolve placement within the FCB superphylum, a subset of 77 FCB superphylum
402members and 37 archaeal references sequences were selected based on broad
403taxonomic representation within the FCB superphylum and phylogenies constructed
404using two separate algorithms with the GTR+G+I model: maximum likelihood

405(RAxML⁴⁸) and Bayesian inference (MrBayes⁴⁹). Node stability was evaluated using a
406rapid bootstrapping analysis (RAxML, 100 runs) and posterior probabilities (MrBayes,
4072.4 million generations, burnin of 25%). Alignments and phylogenetic trees are available
408in Supplementary Data 8 and 9, respectively.

409**Microscopy**

410An oligonucleotide probe specific for 'Ca. Kryptonia' (Kryp56; 5'-
411CCGTGTCCCTGACTTGCA-3') was designed in ARB (version 6.0.2)⁵⁰. The probe is a
412perfect match to 19 of the 22 'Ca. Kryptonia' SSU rRNA gene sequences recovered in
413this study, and contains two or more mismatches to all SSU rRNA gene sequences in
414the SILVA-NR database (version 123)¹⁶. The probe sequence was synthesized by
415Biomers.net (Ulm, Germany) with horseradish peroxidase (HRP) conjugated to the 5'
416end. Cells from Dewar Creek sediment were separated from particulates by brief
417vortexing followed by centrifugation (30 s, 1,300 x g). Suspended cells were preserved
418with dimethyl sulfoxide (4% DMSO) and stored at -80°C. The cells were permeabilized
419with lysozyme (10 mg/ml in TE buffer (1 mM EDTA, 10 mM Tris)) for 1 hr at 37°C and
420catalyzed reporter deposition fluorescence *in situ* hybridization (CARD-FISH) was
421performed based on the protocol of Pernthaler and colleagues⁵¹. Hybridization was
422carried out at 46°C with 20% formamide, and the amplification was performed with
423tyramides conjugated to Alexa 488 (Life Technologies, #T20948). The optimal
424formamide concentration and specificity was predicted using mathFISH⁵² and the
425DECIPHER ProbeMelt tool⁵³ (Supplementary Data 10), and confirmed empirically by
426performing CARD-FISH on the Dewar Creek cells over a gradient of formamide
427concentrations (10 – 35%). Samples were counterstained with 4',6-diamidino-2-
428phenylindole (DAPI) in VECTASHEILD Antifade Mounting Media (Vector Laboratories,

429#H-1200). Cells were visualized and imaged using a Leica DM6000B microscope using
430a HCX PL APO 100X oil immersion objective.
431 **Conserved single-copy and housekeeping gene phylogenetic inference**
432 A set of 56 universally conserved single copy proteins in the Bacteria and Archaea was
433 used for phylogenetic inference (Supplementary Data 11). Marker genes were detected
434 and aligned with hmmsearch and hmalign included in HMMER3 (ref.⁵⁴) using HMM
435 profiles obtained from phylsift (<http://phylsift.wordpress.com/>)⁵⁵. Alignments were
436 concatenated and filtered⁵⁶. Housekeeping genes were aligned using MAFFT with mafft-
437 linsi option⁵⁷. Best substitution model was selected using protest⁵⁸. Phylogeny was
438 inferred using maximum likelihood methods with RAxML (version 7.6.3)⁴⁸. Tree
439 topologies were tested for robustness using 100 bootstrap replicates with the LG+I+G
440 model (raxmlHPC-PTHREADS-SSE3 -f a -x 12345 -p 12345 -# 100 -m
441 PROTGAMMALG -T 5). Trees were visualized using Dendroscope⁵⁹. The concatenated
442 protein alignment and phylogenetic tree are available in Supplementary Data 12 and 13,
443 respectively.

444 **Phylogenetic distribution of predicted proteins**

445 The taxonomic distribution of all proteins across the genomes reconstructed from
446 metagenomic data along with the 'Ca. Kryptonia' SAGs was compiled based on best
447 matches to a comprehensive protein database of high-quality non-redundant bacterial
448 and archaeal isolate genomes¹⁴. This search was performed using usearch (version
449 7.0)⁶⁰, where a protein match was considered for proteins with $\geq 30\%$ sequence identity
450 across $\geq 50\%$ of the query alignment length. Phylogenetic affiliation at the phylum level
451 was assigned for top matches, while proteins lacking a match according to the above
452 criteria were noted as 'no match.'

453 **Biogeography of 'Ca. Kryptonia'**

454All genomic data for 'Ca. Kryptonia' was searched against the assembled metagenomic
455data from 4,290 environmental samples using blat with the -fastMap option⁶¹. Significant
456matches for non-ribosomal genomic regions were considered for sequences \geq 250 bp in
457length and with \geq 75% identity threshold. For metagenomic contigs mapping to the
458ribosomal operon, a 97% identity threshold was used to capture only high-quality
459matches to 'Ca. Kryptonia.' Visualization of metagenomic matches globally was
460performed using the R package 'maps'⁶². All genomic matches can be found in
461Supplementary Data 4.

462**CRISPR repeat-spacer arrays analysis**

463The CRISPR Recognition Tool (CRT)⁶³ was used to detect CRISPR repeat-spacer
464regions across all 'Ca. Kryptonia' assembled scaffolds. In the case of 'Ca.
465Thermokryptus mobilis' GFM JGI-1, we were unable to detect spacers, and therefore we
466additionally used the CRISPR assembler algorithm (Crass)⁶⁴ on the raw reads. Spacers
467were manually curated to cull false positives from the dataset that clearly did not
468represent authentic spacer regions (in sum, 38 false positives). Potentially active
469repeat-spacer arrays were inferred based on direct association with a *cas* gene locus.
470We also considered the isolated repeat-spacers arrays when they shared the same
471repeat sequence with associated *cas* genes. CRISPRmap^{65,66} was used to further
472characterize identified repeat regions. From a total of 1,031 trusted spacers, we next
473clustered these into 795 groups based on identity \geq 90% over the whole spacer length.
474Spacer groups were BLAST queried against distinct databases including 'Ca. Kryptonia'
475genomes, reference public plasmid and viral datasets (from NCBI), and across the
476broad available metagenomic space (IMG/M).

477**SSU rRNA gene assembly and co-occurrence analysis**

478 Raw reads aligning to 16S and 18S rRNAs were collected for 22 metagenomes
479 (Supplementary Table 5) from geothermal environments using hmalign⁵⁴ against hmm
480 models representing bacterial, archaeal and eukaryotic sequences and also by BBmap
481 with default settings⁶⁷ against sequences from the SILVA database (version 119)¹⁶
482 dereplicated at 95% identity using UCLUST⁶⁰. Collected paired-end Illumina reads were
483 merged using BBmerge⁶⁷ and assembled using Newbler (v. 2.8)⁶⁸ with -ml 60 -mi 99 -rip
484 options. Resulting contigs and scaffolds were screened using cmalign from Infernal 1.1
485 package⁶⁹ and Rfam 16S and 18S rRNA models (RF00177.cm, RF01959.cm and
486 RF01960.cm)⁷⁰. 16S and 18S rRNA sequences longer than 300 nt were retained and
487 trimmed using cmalign against the best-matching model with '--matchonly' option to
488 remove introns. Reference sequences from the SILVA database were trimmed using
489 cmalign with a domain-specific model and '--matchonly' option, and clustered together
490 with 16S sequences extracted from shotgun metagenome data using UCLUST and
491 percent identity cutoffs of 94%, 92% and 90%. Clusters including sequences from at
492 least two metagenome samples were retained and their abundances in metagenome
493 samples were computed by multiplying the length of SSU rRNA sequence by the
494 average coverage. Taxonomy was assigned to the clusters as last common ancestor
495 (LCA) of SILVA reference sequences included in the cluster, or as LCA of SILVA
496 sequences in the larger cluster obtained by co-clustering SILVA and metagenome
497 sequences at 83% identity. Spearman's rank-order correlation of cluster abundances
498 was used to estimate co-occurrence of the clusters in metagenome data.

499

500 **References**

5011. Pace, N. R. Mapping the tree of life: Progress and prospects. *Microbiol. Mol. Biol.*
502 *Rev.* **73**, 565-576 (2009).
5032. Yarza, P. *et al.* Uniting the classification of cultured and uncultured bacteria and
504 archaea using 16S rRNA gene sequences. *Nature Rev. Microbiol.* **12**, 635-645
505 (2014).
5063. Brown, C. T. *et al.* Unusual biology across a group comprising more than 15% of
507 domain Bacteria. *Nature* **523**, 208-211 (2015).
5084. Castelle, C. J. *et al.* Extraordinary phylogenetic diversity and metabolic versatility
509 in aquifer sediment. *Nat. Commun.* **4**, 2120 (2013).
5105. Iverson, V. *et al.* Untangling genomes from metagenomes: Revealing an
511 uncultured class of marine euryarchaeota. *Science* **335**, 587-590 (2012).
5126. Sekiguchi, Y. *et al.* First genomic insights into members of a candidate bacterial
513 phylum responsible for wastewater bulking. *PeerJ* **3**, e740 (2015).
5147. Wrighton, K. C. *et al.* Fermentation, hydrogen, and sulfur metabolism in multiple
515 uncultivated bacterial phyla. *Science* **337**, 1661-1665 (2012).
5168. Kantor, R. S. *et al.* Small genomes and sparse metabolisms of sediment-
517 associated bacteria from four candidate phyla. *mBio* **4**, e00708-00713 (2013).
5189. Rinke, C. *et al.* Insights into the phylogeny and coding potential of microbial dark
519 matter. *Nature* **499**, 431-437 (2013).
52010. Castelle, C. J. *et al.* Genomic expansion of domain archaea highlights roles for
521 organisms from new phyla in anaerobic carbon cycling. *Curr. Biol.* **25**, 690-701
522 (2015).

52311. Spang, A. *et al.* Complex archaea that bridge the gap between prokaryotes and
524 eukaryotes. *Nature* **521**, 173-179 (2015).
52512. Lynch, M. D. J. & Neufeld, J. D. Ecology and exploration of the rare biosphere.
526 *Nature Rev. Microbiol.* **13**, 217-229 (2015).
52713. Woyke, T. & Rubin, E. M. Searching for new branches on the tree of life. *Science*
528 **346**, 698-699 (2014).
52914. Markowitz, V. M. *et al.* IMG/M 4 version of the integrated metagenome
530 comparative analysis system. *Nucleic Acids Res.* **42**, D568-D573 (2014).
53115. Pruesse, E., Peplies, J. & Glöckner, F. O. SINA: Accurate high-throughput
532 multiple sequence alignment of ribosomal RNA genes. *Bioinformatics* **28**, 1823-
533 1829 (2012).
53416. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data
535 processing and web-based tools. *Nucleic Acids Res.* **41**, D590-D596 (2013).
53617. Gupta, R. S. The phylogeny and signature sequences characteristics of
537 Fibrobacteres, Chlorobi, and Bacteroidetes. *Crit. Rev. Microbiol.* **30**, 123-143
538 (2004).
53918. Hugenholtz, P., Pitulle, C., Hershberger, K. L. & Pace, N. R. Novel division level
540 bacterial diversity in a Yellowstone hot spring. *J. Bacteriol.* **180**, 366-376 (1998).
54119. Sharp, C. E. *et al.* Humboldt's spa: microbial diversity is controlled by
542 temperature in geothermal environments. *ISME J.* **8**, 1166-1174 (2014).
54320. Costa, K. *et al.* Microbiology and geochemistry of great boiling and mud hot
544 springs in the United States Great Basin. *Extremophiles* **13**, 447-459 (2009).

54521. Cole, J. K. *et al.* Sediment microbial communities in Great Boiling Spring are
546 controlled by temperature and distinct from water communities. *ISME J.* **7**, 718-
547 729 (2013).
54822. Hou, W. *et al.* A comprehensive census of microbial diversity in hot springs of
549 Tengchong, Yunnan Province China using 16S rRNA gene pyrosequencing.
550 *PLoS One* **8**, e53350 (2013).
55123. Varghese, N. J. *et al.* Microbial species delineation using whole genome
552 sequences. *Nucleic Acids Res.* doi:10.1093/nar/gkv657 (2015).
55324. Iino, T. *et al.* *Ignavibacterium album* gen. nov., sp. nov., a moderately
554 thermophilic anaerobic bacterium isolated from microbial mats at a terrestrial hot
555 spring and proposal of *Ignavibacteria* classis nov., for a novel lineage at the
556 periphery of green sulfur bacteria. *Int. J. Syst. Evol. Microbiol.* **60**, 1376-1382
557 (2010).
55825. Podosokorskaya, O. A. *et al.* Characterization of *Melioribacter roseus* gen. nov.,
559 sp. nov., a novel facultatively anaerobic thermophilic cellulolytic bacterium from
560 the class *Ignavibacteria*, and a proposal of a novel bacterial phylum
561 *Ignavibacteriae*. *Environ. Microbiol.* **15**, 1759-1771 (2013).
56226. van der Oost, J., Westra, E. R., Jackson, R. N. & Wiedenheft, B. Unravelling the
563 structural and mechanistic basis of CRISPR-Cas systems. *Nature Rev. Microbio.*
564 **12**, 479-492 (2014).
56527. Makarova, K. S., Wolf, Y. I. & Koonin, E. V. The basic building blocks and
566 evolution of CRISPR-CAS systems. *Biochem. Soc. Trans.* **41**, 1392-1400 (2013).

56728. Inskeep, W. P. *et al.* The YNP Metagenome Project: Environmental parameters
568 responsible for microbial distribution in the Yellowstone geothermal ecosystem.
569 *Front. Microbiol.* **4**, 67 (2013).
57029. Heine, M. & Chandra, S. C. The linkage between reverse gyrase and
571 hyperthermophiles: A review of their invariable association. *J. Microbiol.* **47**, 229-
572 234 (2009).
57330. Weber, K. A., Achenbach, L. A. & Coates, J. D. Microorganisms pumping iron:
574 anaerobic microbial iron oxidation and reduction. *Nature Rev. Microbiol.* **4**, 752-
575 764 (2006).
57631. Akujkar, M. *et al.* Anaerobic degradation of aromatic amino acids by the
577 hyperthermophilic archaeon *Ferroglobus placidus*. *Microbiology* **160**, 2694-2709
578 (2014).
57932. Holmes, D. E., Risso, C., Smith, J. A. & Lovley, D. R. Genome-scale analysis of
580 anaerobic benzoate and phenol metabolism in the hyperthermophilic archaeon
581 *Ferroglobus placidus*. *ISME J.* **6**, 146-157 (2012).
58233. Hernández-Montes, G., Díaz-Mejía, J. J., Pérez-Rueda, E. & Segovia, L. The
583 hidden universal distribution of amino acid biosynthetic networks: a genomic
584 perspective on their origins and evolution. *Genome Biol.* **9**, R95-R95 (2008).
58534. McCutcheon, J. P. & Moran, N. A. Extreme genome reduction in symbiotic
586 bacteria. *Nature Rev. Microbiol.* **10**, 13-26 (2012).
58735. Giovannoni, S. J., Thrash, J. C. & Temperton, B. Implications of streamlining
588 theory for microbial ecology. *ISME J.* **8**, 1553-1565 (2014).

58936. Cao, B. *et al.* Structure of the nonameric bacterial amyloid secretion channel.
590 *Proc. Natl. Acad. Sci. U.S.A.* **111**, E5439-E5444 (2014).
59137. Evans, M. L. & Chapman, M. R. Curli biogenesis: order out of disorder. *Biochim.*
592 *Biophys. Acta* **1843**, 1551-1558 (2014).
59338. Garcia-Pardo, J. *et al.* Amyloid formation by human carboxypeptidase D
594 transthyretin-like domain under physiological conditions. *J. Biol. Chem.* **289**,
595 33783-33796 (2014).
59639. Hedlund, B. P. *et al.* Potential role of *Thermus thermophilus* and *T. oshimai* in
597 high rates of nitrous oxide (N₂O) production in ~80°C hot springs in the U.S.
598 Great Basin. *Geobiology* **9**, 471-480 (2011).
59940. Murugapiran, S. K. *et al.* *Thermus oshimai* JL-2 and *T. thermophilus* JL-18
600 genome analysis illuminates pathways for carbon, nitrogen, and sulfur cycling.
601 *Stand. Genomic Sci.* **7**, 449-468 (2013).
60241. Giovannoni, S. J. *et al.* Genome streamlining in a cosmopolitan oceanic
603 bacterium. *Science* **309**, 1242-1245 (2005).
60442. Dupont, C. L. *et al.* Genomic insights to SAR86, an abundant and uncultivated
605 marine bacterial lineage. *ISME J.* **6**, 1186-1199 (2012).
60643. Garcia Costas, A. M. *et al.* Complete genome of *Candidatus Chloracidobacterium*
607 *thermophilum*, a chlorophyll-based photoheterotroph belonging to the phylum
608 Acidobacteria. *Environ. Microbiol.* **14**, 177-190 (2012).
60944. Barns, S. M., Fundyga, R. E., Jeffries, M. W. & Pace, N. R. Remarkable archaeal
610 diversity detected in a Yellowstone National Park hot spring environment. *Proc.*
611 *Natl. Acad. Sci. U.S.A.* **91**, 1609-1613 (1994).

61245. Dick, G. J. *et al.* Community-wide analysis of microbial genome sequence
613 signatures. *Genome Biol.* **10**, R85-R85 (2009).
61446. Nurk, S. *et al.* Assembling single-cell genomes and mini-metagenomes from
615 chimeric MDA products. *J. Comp. Biol.* **20**, 714-737 (2013).
61647. Tennessen, K. *et al.* ProDeGe: a computational protocol for fully automated
617 decontamination of genomes. *ISME J.*, doi:10.1038/ismej.2015.100 (2015).
61848. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic
619 analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688-
620 2690 (2006).
62149. Ronquist, F. & Huelsenbeck, J. P. MrBayes 3: Bayesian phylogenetic inference
622 under mixed models. *Bioinformatics* **19**, 1572-1574 (2003).
62350. Ludwig, W. *et al.* ARB: a software environment for sequence data. *Nucleic Acids*
624 *Res.* **32**, 1363-1371 (2004).
62551. Pernthaler, A., Pernthaler, J. & Amann, R. Fluorescence in situ hybridization and
626 catalyzed reporter deposition for the identification of marine bacteria. *Appl.*
627 *Environ. Microbiol.* **68**, 3094-3101 (2002).
62852. Yilmaz, L. S., Parnerkar, S. & Noguera, D. R. mathFISH, a web tool that uses
629 thermodynamics-based mathematical models for in silico evaluation of
630 oligonucleotide probes for fluorescence in situ hybridization. *Appl. Environ.*
631 *Microbiol.* **77**, 1118-1122 (2011).
63253. Yilmaz, L. S., Loy, A., Wright, E. S., Wagner, M. & Noguera, D. R. Modeling
633 formamide denaturation of probe-target hybrids for improved microarray probe
634 design in microbial diagnostics. *PLoS One* **7**, e43862 (2012).

63554. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive
636 sequence similarity searching. *Nucleic Acids Res.* **39**, W29-W37 (2011).
63755. Darling, A. E. *et al.* PhyloSift: phylogenetic analysis of genomes and
638 metagenomes. *PeerJ* **2**, e243 (2014).
63956. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for
640 automated alignment trimming in large-scale phylogenetic analyses.
641 *Bioinformatics* **25**, 1972-1973 (2009).
64257. Katoh, K. & Standley, D. M. MAFFT Multiple sequence alignment software
643 version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772-
644 780 (2013).
64558. Abascal, F., Zardoya, R. & Posada, D. ProtTest: selection of best-fit models of
646 protein evolution. *Bioinformatics* **21**, 2104-2105 (2005).
64759. Huson, D. H. *et al.* Dendroscope: An interactive viewer for large phylogenetic
648 trees. *BMC Bioinform.* **8**, 460-460 (2007).
64960. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST.
650 *Bioinformatics* **26**, 2460-2461 (2010).
65161. Kent, W. J. BLAT—The BLAST-Like Alignment Tool. *Genome Res.* **12**, 656-664
652 (2002).
65362. Brownrigg, R., Minka, T. P., Becker, R. A. & Wilks, A. R. maps: Draw
654 Geographical Maps. R package version 2.1-5. [http://CRAN.R-](http://CRAN.R-project.org/package=maps)
655 [project.org/package=maps](http://CRAN.R-project.org/package=maps) (2010).

65663. Bland, C. *et al.* CRISPR recognition tool (CRT): a tool for automatic detection of
657 clustered regularly interspaced palindromic repeats. *BMC Bioinform.* **8**, 209
658 (2007).
65964. Skennerton, C. T., Imelfort, M. & Tyson, G. W. Crass: identification and
660 reconstruction of CRISPR from unassembled metagenomic data. *Nucleic Acids*
661 *Res.* **41**, e105 (2013).
66265. Alkhnbashi, O. S. *et al.* CRISPRstrand: predicting repeat orientations to
663 determine the crRNA-encoding strand at CRISPR loci. *Bioinformatics* **30**, i489-
664 496 (2014).
66566. Lange, S. J., Alkhnbashi, O. S., Rose, D., Will, S. & Backofen, R. CRISPRmap:
666 an automated classification of repeat conservation in prokaryotic adaptive
667 immune systems. *Nucleic Acids Res.* **41**, 8034-8044 (2013).
66867. Bushnell, B. BBmap software package, <http://sourceforge.net/projects/bbmap/>
669 (2015).
67068. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre
671 reactors. *Nature* **437**, 376-380 (2005).
67269. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology
673 searches. *Bioinformatics* **29**, 2933-2935 (2013).
67470. Burge, S. W. *et al.* Rfam 11.0: 10 years of RNA families. *Nucleic Acids Res.* **41**,
675 D226-232 (2013).

676

677**Acknowledgments**

678We thank the DOE JGI production sequencing, IMG, and Genomes OnLine Database
679teams for their support, along with Steven Quake for metagenomic sequencing and

680assembly of the Jinze and Gongxiaoshe samples. We thank BC Parks and the Ktunaxa
681Nation for their cooperation on the Dewar Creek spring. This work was conducted by the
682U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User
683Facility, under Contract No. DE-AC02-05CH11231 and used resources of the National
684Energy Research Scientific Computing Center, which is supported by the Office of
685Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.
686This work was also supported by U.S. Department of Energy (DOE) grant DE-EE-
6870000716; the U.S. Department of Energy Joint Genome Institute (CSP-182); NASA
688Exobiology grant EXO-NNX11AR78G; U.S. National Science Foundation grant OISE
6890968421; Key Project of International Cooperation by the Chinese Ministry of Science
690and Technology (MOST, 2013DFA31980); B.P.H. acknowledges generous support from
691Greg Fullmer through the UNLV Foundation. Metagenome analysis of Dewar Creek was
692supported in part by funding from Genome Canada, Genome Alberta, Genome BC, and
693the Government of Alberta (GC Grant 1203). Work at LLNL was conducted under the
694auspices of DOE Contract DE-AC52-07NA27344 and supported by a Lawrence
695Fellowship to A. Dekas.

696

697**Author Contributions**

698E.A.E-F., N.C.K., and N.N.I. designed the project; P.F.D., B.P.H., S.E.G., A.L.B., H.D.,

699B.R.B., and W-J.L. provided the samples; J.J., D.G., R.M., and T.W. performed the

700single-cell experiments; A.E.D. and J.P-R. performed the CARD-FISH experiments;

701E.A.E-F., D.P-E., J.J., P.F.D., A.P., T.W., N.C.K., and N.N.I. analyzed the data; E.A.E-F.,

702D.P-E., N.C.K, and N.N.I. wrote the manuscript with significant input from P.F.D., B.P.H.,

703T.W., and E.M.R. All authors discussed the results and commented on the manuscript.

704

705**Competing interests**

706The authors declare no competing interests.

707

708 **Accession Codes**

709 Genome sequence data, assemblies and annotations have been deposited as Whole

710 Genome Shotgun projects at DDBJ/EMBL/GenBank with the accession codes

711 PRJEB11785 to PRJEB11788 (GFMs) and PRJEB11711 to PRJEB11728 (SAGs).

712

713 **Figure 1. New lineage identified using metagenomic and single-cell genomic**

714 **approaches.** Workflow used to **(A)** identify novel SSU rRNA gene sequences globally,

715 along with **(B)** single-cell genomics pipeline to screen and sequence single cells

716 isolated from geothermal springs samples. For the three geothermal spring

717 environments, we sequenced 13, 2, and 3 SAGs, respectively. SSU rRNA gene, small-

718 subunit ribosomal gene; MDA, multiple displacement amplification; QC, quality control;

719 SAG, single-amplified genome.

720

721 **Figure 2. Maximum likelihood concatenated protein phylogeny and cell imaging**

722 **for ‘Ca. Kryptonina.’** **(A)** Phylogeny was based on concatenation of 56 conserved

723 marker proteins, where at least 10 marker proteins were used to infer SAG phylogenetic

724 placement (with the exception of JGI-22 with only six marker proteins recovered).

725 Bootstrap support values $\geq 50\%$ are shown with small circles on nodes with robust

726 phylogenetic support. The *Fibrobacteres-Chlorobi-Bacteroidetes* (FCB) superphylum is

727 shown in the gray shaded region. Expanded phylogenetic tree for ‘Ca. Kryptonina’ shows

728 the placement of the proposed four genera represented by GFMs and SAGs, along with

729 the estimated genome completeness shown in parentheses. **(B)** A ‘Ca. Kryptonina’-

730 specific FISH (fluorescence *in situ* hybridization) probe was designed and used to

731 visualize cells from Dewar Creek Spring sediment samples. ‘Ca. Kryptonina’ cells

732 hybridizing with the probe are green, while other cells are visualized with 4',6-diamidino-
733 2-phenylindole (DAPI; blue). Scale bar, 5 μ m.

734

735 **Figure 3. Limited, yet widely dispersed biogeographic distribution of 'Ca.**

736 **Kryptonita' genomes and CRISPR spacers.** All genomic content from the 'Ca.

737 Kryptonita' GFM and SAGs was used to comprehensively search the collection of 640

738 Gb of assembled metagenomic data from 4,290 environmental samples, including 169

739 samples from geothermal springs and hydrothermal vents denoted by red triangles

740 (temperature $\geq 50^\circ\text{C}$). Marked circles are as follows: (A) Great Boiling Spring,

741 Nevada^{20,21}, (B) Dewar Creek Spring, Canada¹⁹, (C) Jinze pool, Yunnan Province,

742 China²², and (D) Gongxiaoshe pool, Yunnan Province, China²². Significant matches

743 were determined for sequences ≥ 250 bp in length and with $\geq 75\%$ identity threshold for

744 non-ribosomal genomic regions. For metagenomic contigs mapping to the 'Ca.

745 Kryptonita' ribosomal operon, a 97% identity threshold was used to capture only high-

746 quality matches to 'Ca. Kryptonita.' For CRISPR spacers, only significant matches

747 allowing for up to 3 bp mismatch along the entire length of the spacer were considered.

748 The 'Ca. Kryptonita' genomic hits can be found in Supplementary Data 4 and the

749 manually curated spacer hits can be found in Supplementary Data 3.

750

751

752 **Figure 4. Reconstructed metabolic capacity of 'Ca. Kryptonita.'** Key metabolic

753 predictions and novel features identified in 'Ca. Kryptonita' GFM and SAGs, with full

754 gene information available in Supplementary Data 6.

755

756 **Figure 5. Co-occurrence patterns and metabolic complementarity with 'Ca.**

757 **Kryptonita.'** (A) Spearman-rank correlation values were calculated based on

758 reconstructed SSU rRNA sequences across 22 geothermal spring metagenomes, and

759 led to the identification of a cluster of highly correlated phylotypes with 'Ca. Kryptonita.'

760 *Armatimonadetes* (cluster 3107) had the highest correlation value ($\rho = 0.82$) with 'Ca.

761 Kryptonita.' (B) Biosynthetic pathways present in the *Armatimonadetes* genome which

762 complement missing components in 'Ca. Kryptonita.' Full gene information for the

763 *Armatimonadetes* genome is available in Supplementary Data 7. Each arrow represents

764 an enzymatic component of the biosynthetic pathways; arrows highlighted in blue are

765 contributed by the *Armatimonadetes*, while arrows highlighted in dark orange are

766 contributed by 'Ca. Kryptonita.' Black arrows indicate enzyme was not recovered in

767 either.

768