A window into hydrothermal vent endosymbioses: the *Calyptogena magnifica* chemoautotrophic symbiont genome

Authors: I.L.G. Newton[1], T. Woyke[2], T.A. Auchtung[1], G.F. Dilly[1], R.J. Dutton[3], M.C. Fisher[1], K. M. Fontanez[1], E. Lau[1], F.J. Stewart[1], P.M. Richardson[2], K.W. Barry[2], J.C. Detter[2], D. Wu[4], J. A. Eisen[5], C.M. Cavanaugh[1*]

*corresponding author

The *Calyptogena magnifica* symbiont is the most metabolically capable intracellular endosymbiont, able to oxidize sulfur, fix carbon dioxide, assimilate nitrogen, and synthesize vitamins, cofactors, and 20 amino acids.

Keywords: chemosynthesis, maternal transmission, deep-sea, symbiosis

None of this material has been published or is under consideration elsewhere, including the Internet.

18 Mailing Addresses:

19 1 –Harvard University, 16 Divinity Ave, Biolabs 4080, Cambridge, MA 02138
20 Telephone number: 617-495-1138
21 Fax number: 617-496-6933
22
23 2 - DOE Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598,
24 USA
25 Telephone number: 530 752 3498
26 Fax number: 925-296-5850
27
28 3 – Harvard Medical School, Department of Microbiology and Molecular
29 Genetics, 200 Longwood Ave., Boston, MA 02115
30 Telephone number: 617-432-1788
31 Fax number: 617-738-7664
32
33 4 – The Institute for Genomic Research, 9712 Medical Center Drive, Rockville,
34 MD 20850
35 Telephone number: 301-795-7000
36 Fax number: 301-838-0208
37
38 5 – University of California, Davis Genome Center
39 Genome and Biomedical Sciences Facility Room 5311
40 451 East Health Sciences Drive
41 Davis, CA 95616-8816
42 Telephone number: 530 752 3498
43

44 Email addresses:

45 I.L.G. Newton – Garcia@fas.harvard.edu

46 T. Woyke – twoyke@lbl.gov

47 T.A. Auchtung – auchtung@fas.harvard.edu

48 G.F. Dilly – gdilly@fas.harvard.edu

49 R.J. Dutton – Rachel_dutton@hms.harvard.edu

50 M.C. Fisher – mfisher@oeb.harvard.edu

51 K. M. Fontanez – kfontanez@oeb.harvard.edu

52 E. Lau – elau@oeb.harvard.edu

53 F.J. Stewart - fstewart@fas.harvard.edu

54 P.M. Richardson - PMRichardson@lbl.gov

55 K.W. Barry – kwbarry@lbl.gov

56 J.C. Detter - cdetter@lanl.gov

57 D. Wu - dwu@tigr.org

58 J. A. Eisen – jaeisen@ucdavis.edu

59 C.M. Cavanaugh – cavanaug@fas.harvard.edu

60

61

62 Chemosynthetic endosymbionts are the metabolic cornerstone of hydrothermal
63 vent communities, providing invertebrate hosts with nearly all of their nutrition.
64 The *Calyptogena magnifica* (Bivalvia: Vesicomyidae) symbiont, *Candidatus*
65 *Ruthia magnifica,* is the first intracellular chemosynthetic endosymbiont to have
66 its genome sequenced, revealing an enormous suite of metabolic capabilities.
67 The genome encodes the major chemosynthetic pathways as well as pathways
68 for biosynthesis of vitamins, cofactors, and all 20 amino acids required by the
69 host, indicating the host is entirely nutritionally dependent on *Ruthia*. This
70 genome sequence will be invaluable in the study of these enigmatic associations
71 and provides insights into the origin and evolution of autotrophic endosymbioses.
72

72 Miles below the surface of the ocean, where tectonic plates meet,
73 the food-limited habitat of the deep-sea is punctuated by diverse communities of
74 invertebrates and bacteria. Metazoans at these hydrothermal vents flourish
75 thanks to the chemoautotrophy of symbiotic bacteria (*1*). Seawater here
76 percolates into the crust, is heated as it reacts with oceanic basalt, and becomes
77 enriched in the reduced sulfur and carbon dioxide that sulfur oxidizing
78 chemoautotrophs require (*1*). The symbiotic bacteria use the energy gained in
79 oxidation of these reduced sulfur compounds for carbon fixation. Analogous to
80 photosynthetic chloroplasts, which are derived from cyanobacterial ancestors and
81 use light energy to fix carbon for their plant and algal hosts, these
82 chemosynthetic endosymbionts use chemical energy to provide their hosts with
83 not only carbon but also a large array of additional necessary nutrients. The
84 metazoan hosts, in turn, bridge the oxic-anoxic interface to provide their bacteria
85 with the inorganic substrates necessary for chemosynthesis. Hosts often betray
86 their nutritional dependence on these bacteria through their diminished or absent
87 digestive systems. Although first discovered at hydrothermal vents, similar
88 associations exist at mud flats, seagrass beds, and hydrocarbon seeps. In each
89 case it is clear that these symbioses play major roles in community structuring
90 and sulfur and carbon cycling. However, despite the widespread occurrence of
91 these partnerships, little is known of the intricacies of host-symbiont interaction or
92 symbiont metabolism due to their inaccessibility and our inability to culture either
93 partner separately.

94

95 The giant clam, *Calyptogena magnifica* Boss and Turner (Bivalvia:
96 Vesicomyidae) was one of the first organisms described following the discovery
97 of hydrothermal vents (*2*). The vesicomyids are relatively old, with fossil records
98 and phylogenies dating them at 50-100 Ma (*3*). *C. magnifica* grows to a large size
99 (>26 cm in length), despite having a reduced gut and ciliary food groove (*2*),
100 presenting a conundrum regarding how it acquires sufficient nutrients. The
101 mystery of this clam's nutrition was solved when chemosynthetic, γ-
102 proteobacterial symbionts, here named *Candidatus Ruthia magnifica* (in memory
103 of Prof. Ruth Turner), were discovered within its gill bacteriocytes (*4, 5*) (Figure
104 1). The host depends largely on these symbionts for its carbon, as indicated by
105 its anatomy and by stable carbon isotopic ratios (*6, 7*). However, how the host
106 satisfies the rest of its nutritional needs remains unknown.
107
108 *R. magnifica* is the first intracellular chemosynthetic symbiont to have its genome
109 sequenced. Here we describe analysis of this finished sequence. In particular
110 we discuss how, despite a relatively small genome, the symbiont is predicted to
111 convey a striking diversity of nutritional capabilities on the host. In addition, we
112 consider how this symbiont's genome differs in fundamental ways from those of
113 other nutritional endosymbionts.
114

115 Although, in some ways, the *R. magnifica* genome resembles that of other
116 obligate mutualistic symbionts for which data are available, surprising differences
117 were found. The genome has a low G+C content (34%) compared to free-living
118 relatives (Table 1). In addition, the coding density (81.4%) and mean gene
119 length (975 bp), though lower than commonly seen in free-living bacteria, are
120 consistent with that in other endosymbiont genomes (*8*). These common
121 features of endosymbionts are likely the result of genome reduction and
122 degradation (rampant gene loss and mutation rate increases, respectively) that
123 occur over evolutionary time across diverse symbiont species. This trend is
124 evident in relatively recent symbioses such as the insect endosymbionts (30-250
125 Ma), as well as in chloroplasts (~1,800-2,100 Ma). Upon closer examination
126 however, *R. magnifica* stands out in that its genome is large for a maternally
127 transmitted endosymbiont (1.2 Mb). For example, the genomes of the γ-
128 proteobacterial *Buchnera* species, which are endosymbionts of aphids, are some
129 80% smaller than closely related free-living species like *E. coli.* In contrast, *R.*
130 *magnifica*'s genome is half the size of its relative's, *Thiomicrospira crunogena,* a
131 free-living, γ-proteobacterial, sulfur-oxidizing chemoautotroph.
132
133 We propose that the limited genome reduction in *R. magnifica* is due to a
134 fundamental difference in its biology compared to other nutritional endosymbionts
135 characterized so far. Insect endosymbionts typically supplement the diet of their
136 hosts, e.g., *Buchnera* provide essential amino acids that are missing in the

137 phloem sap diet of aphids. Similarly, the γ-proteobacteria *Baumannia* and *Sulcia*
138 together provide amino acids and vitamins for their sharpshooter hosts, but
139 apparently not much more (*9*). These symbionts acquire much of what they need
140 (e.g., sugars) from their host and thus can still survive with very small genomes
141 (*10*). In contrast, and most strikingly, *R. magnifica* is predicted to encode all the
142 metabolic pathways one would expect in free-living chemoautotrophs including
143 carbon fixation, sulfur oxidation, nitrogen assimilation, and amino acid and
144 cofactor/vitamin biosynthesis (Figure 2). Thus we conclude it provides the clam
145 with the majority of its nutrition. In the following sections we discuss different
146 aspects of the metabolic reconstruction of *R. magnifica* and what this might mean
147 for the biology of its host. For simplicity, we refer to these reconstructions as
148 though the pathways have been validated, although it should be emphasized that
149 these are predictions.
150
151 *R. magnifica*'s genome is largely dedicated to biosynthesis and energy
152 metabolism, highlighting the importance of these pathways in the symbiosis
153 (Figure 2). The *R. magnifica* genome also encodes enzymes for carbon fixation,
154 sulfur oxidation, nitrogen assimilation and energy conservation. Genes encoding
155 enzymes specific to the Calvin Cycle, a form II ribulose-1,5- bisphosphate
156 carboxylase/oxygenase (RubisCO) and phosphoribulokinase (*11, 12*), were
157 found in the *R. magnifica* genome (Figure 3). This pathway synthesizes
158 phosphoglyceraldehyde from carbon dioxide and is the dominant form of carbon

159 fixation in vent symbioses (*13*). However, the genome lacks homologs of
160 sedoheptulose 1,7-bis-phosphatase (SBPase, EC 3.1.3.37) and fructose 1,6-bis-
161 phosphatase (FBPase, EC 3.1.3.11), suggesting that the regeneration of ribulose
162 1,5-bisphosphate may not follow conventional routes. Instead, the *R. magnifica*
163 genome contains a reversible pyrophosphate-dependent phosphofructokinase
164 (EC 2.7.1.90) homolog that may use to generate fructose 6-phosphate (*14*).
165
166 Energy generation for carbon fixation in *R. magnifica* can result from sulfur
167 oxidation via the *sox* (sulfur oxidation) and *dsr* (dissimilatory sulfite reductase)
168 genes (Figure 3). The *R. magnifica sox* genes resemble those of the γ-
169 proteobacteria *Thiobacillus denitrificans* and *Allochromatium vinosum,* and the
170 green sulfur bacterium *Chlorobium tepidum* (*15-17*). Homologs of the *sox* genes
171 are located in two positions in the *R. magnifica* genome with *soxXYZA* located in
172 a single operon while *soxB* is elsewhere. The symbiont genome also contains
173 homologs for many of the *dsr* genes which catalyze the oxidation of intracellularly
174 stored sulfur in both *A. vinosum* and *Chlorobium limicola* (*16, 18*). Indeed, sulfur
175 granules observed within *R. magnifica* cells may be a source of reduced sulfur
176 when external sulfide is lacking (*19*). The symbiont's *dsr* genes were contained
177 in a single cluster, *dsrABEFHCMKLOP*, missing *dsrJNRS*. As these latter
178 proteins are not well characterized, it is not known how symbiont sulfur
179 metabolism may be affected. Homologs encoding both a sulfide:quinone
180 oxidoreductase and rhodanese are present, and along with the *dsr* and *sox*

181 proteins, these enzymes can oxidize both thiosulfate ($S_2O_3^{2-}$) or sulfide (HS^-) to
182 sulfite (SO_3^{2-}) (Figure 3). Sulfite can then be oxidized to sulfate (SO_4^{2-}) by the
183 actions of APS reductase (AprAB) and ATP sulfurylase (Sat) before being
184 exported from the cell via a sulfate transporter. This genomic evidence is
185 supported by ATP sulfurylase activity detected in *C. magnifica* gill tissue (*7*),
186 carbon dioxide uptake when sulfide or thiosulfate are provided to the clam (*20,*
187 *21*) and sulfide binding, zinc-containing lipoprotein in the host blood stream (*22*).
188 Thus through the activities of the *sox* and *dsr* genes, the *R. magnifica* symbiont
189 can generate energy from the oxidation of sulfide and thiosulfate.
190
191 Energy conservation, which involves creating a charged membrane, proceeds in
192 *R. magnifica* through NADH dehydrogenase, a sulfide:quinone oxidoreductase,
193 and an *rnf* complex, which in other bacteria has been shown to possess NADH
194 and FMN:quinone oxidoreductase activity (*23*). The genome encodes a
195 straightforward electron transport chain, thus the reduced quinone in the
196 symbiont membrane could transfer electrons to cytochrome c via a bc_1 complex
197 and a terminal cytochrome c oxidase could then transfer these electrons to
198 oxygen.
199
200 Nitrogen assimilation is as important as carbon fixation in the context of this
201 symbiosis as *Ruthia* appears to provide the majority if not all of the host's amino
202 acids. In the predicted pathways, nitrate and ammonia enter the cell via a

203 nitrate/nitrite (NarK) transporter and two ammonium permeases (AmtB1/2) and
204 are then reduced via nitrate (NarB) and nitrite (NirA) reductase, and assimilated
205 via glutamine synthetase (GlnA) and glutamate synthase (GltB/D), respectively
206 (Figure 3). Although nitrate is the dominant form of nitrogen present at vents (*24*)
207 and likely the source of nitrogen for the symbiosis, the symbiont may also
208 assimilate ammonia via recycling of the host's amino acid waste products.
209

210 In keeping with the nutritional role of the symbionts, *R. magnifica*'s inferred
211 intermediary metabolism can produce all necessary biosynthetic intermediates.
212 The genome encodes a complete glycolytic pathway with a pyrophosphate-
213 dependent phosphofructokinase homolog and the non-oxidative branch of the
214 pentose phosphate pathway. The symbiont genome encodes a "horseshoe
215 shaped" tricarboxylic acid (TCA) cycle, lacking alpha-ketoglutarate
216 dehydrogenase. For other chemosynthetic bacteria, the lack of this enzyme has
217 been suggested as an indicator of obligate autotrophy (*25*). Interestingly, the
218 symbiont is also missing homologs of fumarate reductase, succinyl-coA
219 synthase, and succinate dehydrogenase. However, the genome encodes
220 isocitrate lyase, part of the glyoxylate shunt, and could produce succinate from
221 isocitrate. Carbon fixed via the Calvin cycle can enter the TCA cycle through
222 phosphoenolpyruvate and here could follow biosynthetic routes either to fumarate
223 or alpha-ketoglutarate. All of the pathways for biosynthetic reagents required to

224 support the metabolic capabilities of *R. magnifica* are thus encoded in the

225 symbiont genome.

226

227 Unlike any other sequenced endosymbiont genome, *R. magnifica* encodes

228 complete pathways for the biosynthesis of 20 amino acids. This full complement

229 suggests that the symbiont can supply its host with the 9 essential amino acids or

230 their precursors. However, while *E. coli* has 16 essential amino acid

231 biosynthesis regulatory genes (*26*), *metR* (involved in regulating methionine

232 biosynthesis) is the only regulatory gene present in the *R. magnifica* genome.

233 This lack of regulatory genes may be the result of the stability experienced by *R.*

234 *magnifica* in its intracellular environment.

235

236 Animals are dependent on external sources for many of their vitamins and

237 cofactors and bacterial symbionts often provide these nutrients (*10, 27*). The *R.*

238 *magnifica* genome appears to have complete biosynthetic pathways for the

239 majority of vitamins and cofactors *(39).* The only pathway conspicuously absent

240 is that for cobalamin (B_{12}), a cofactor for methionine synthase (*27, 28*). Since *R.*

241 *magnifica* encodes a cobalamin-independent methionine synthase, it is able to

242 provide the host with methionine and the host is unlikely to require cobalamin.

243

244 As with other intracellular species, *R. magnifica* encodes a limited repertoire of

245 transporters, however, those present reveal important details about the

246 movement of metabolites between host and symbiont. Of the 58 proteins
247 predicted to be involved in cell transport and binding in the *R. magnifica* genome,
248 transporters involved in chemosynthesis (sulfate exporters), nitrogen assimilation
249 (ammonium and nitrate importers), inorganic compounds (TrkAH, MgtE family,
250 CaCA family and PiT family), and heavy metals (ZnuABC, RND superfamily, iron
251 permeases) were identified. Surprisingly, few substrate-specific transporters and
252 only two ABC transporter proteins of unknown substrate were found. As it is
253 unlikely that these two ABC transporter proteins are translocating amino acids,
254 vitamins, and cofactors to the host, perhaps the symbionts are "leaky" or the host
255 is actively digesting symbiont cells. Indeed, the closest known relative to *Ruthia*,
256 the bathymodiolid mussel symbionts, are digested intracellularly by their host
257 (*29*). Although the vesicomyid clam and the bathymodiolid mussels are not
258 closely related, electron micrographs suggest the presence of putative
259 degradative stages of symbionts within *C. magnifica* bacteriocytes (Figure 1b).
260
261 Interestingly, the *R. magnifica* genome lacked the key cell division gene, *ftsZ*.
262 FtsZ, a tubulin homolog, assembles as a ring within the bacterial cell, recruits the
263 remaining cell division proteins and constricts to initiate cytokinesis (*30*). It is
264 puzzling that *R. magnifica* lacked FtsZ given that it is almost universally
265 conserved in bacteria, with the notable exception of the obligately intracellular
266 pathogens in the Chlamydia division (*31*). In addition to the absence of *ftsZ*, *R.*
267 *magnifica* and Chlamydia both lack the *murI* gene (*32*), required for the synthesis

268 of D-glutamate, an essential component of the bacterial cell wall. The potential
269 similarities in cell division and cell wall machinery between *R. magnifica* and
270 Chlamydia may be responsible for the "elementary body" cell morphologies
271 observed in both organisms inside the host cell (Figure 1b, *33*). In Chlamydia
272 these bodies are the infectious, propagating form (*34*); their appearance in *R.*
273 *magnifica* may reflect common mechanisms for adaptation to an obligately
274 intracellular lifestyle.
275
276 Endosymbiont intracellular lifestyles have severe effects on genome evolution
277 including genome reductions, skewed base compositions, and elevated rates of
278 gene evolution (*8*). As noted above, *R. magnifica* does exhibit skewed
279 composition and genome reduction, although these are minor shifts compared to
280 those seen in insect endosymbionts. Previous studies have shown, however,
281 that *R. magnifica* also exhibits faster nucleotide substitution rates than those of
282 both free-living bacteria and environmentally transmitted chemosynthetic
283 symbionts (*35*). The factors that contribute to these features of endosymbiont
284 evolution are believed to be a combination of a relatively stable environment,
285 population bottlenecks, and sequestration from free-living bacteria all of which
286 likely occur in *R. magnifica*. In addition, as with some but not all other
287 endosymbionts, *R. magnifica* has lost key genes in DNA repair processes that
288 likely enhance the speed of genome degradation. For example, it is missing
289 genes involved in induction of the SOS repair system and in recombinational

290 repair, including the exonuclease complex genes *recB,C,D* and the highly
291 conserved recombinase *recA* . Perhaps most importantly, it is also missing
292 genes that could encode homologs of the MutSLH proteins, which, in other
293 species greatly limit mutation rates by carrying out post-replication mismatch
294 repair (*36*).
295
296 Given the apparent defects in DNA repair and the likely population forces
297 pushing this organism's genome towards degradation it is particularly informative
298 that it has retained genes that encode a full suite of chemosynthesis processes.
299 For comparison, chloroplast genomes have lost over 90% of their content since
300 their cyanobacterial ancestor entered endosymbiosis, with many of their genes
301 having been transferred to the host nuclear genome (*37*). The more modern
302 insect endosymbioses have lost between 70-80% of their genomes over a much
303 shorter evolutionary time, and it is unknown if any of these pathways are
304 encoded by the nucleus (*10, 38*). *R. magnifica*, in contrast, has the largest
305 genome of any intracellular symbiont sequenced to date and may represent an
306 early evolutionary intermediate towards a chemoautotrophic "plastid". The broad
307 array of metabolic pathways encoded by *R. magnifica* expands prior knowledge
308 of host nutritional dependency based on stable carbon isotopic ratios and host
309 physiology and anatomy (*6, 7*). It is the extent of this dependency that may be
310 preventing the loss of metabolic pathways in the *R. magnifica* genome. This
311 selective pressure might be great enough to counter the forces of genome

312 reduction and degradation seen in other endosymbionts and provides a novel

313 framework for the study of endosymbiont evolution.

314

314

315 **Methods:**

316

317 **Specimen collection and DNA extraction:**

318 *Calyptogena magnifica* clams were collected using *DSV Alvin* at the East Pacific
319 Rise, 9°N, during a December 2004 cruise on the *R/V Atlantis*. The symbiont-
320 containing gills were dissected out of the clams, frozen in liquid nitrogen, and
321 kept at -80°C until processed. They were then ground in liquid nitrogen, placed
322 in lysis buffer (20 mM EDTA, 10 mM Tris-HCl, pH 7.9, 0.5 mg/ml lysozyme, 1%
323 Triton X-100, 500 mM guanidine-HCl, 200 mM NaCl) and kept at 40ºC for 2 hr.
324 After subsequent RNase (20 μg/ml, 37°C, 30 min) and proteinase K (20 μg/ml,
325 50ºC, 1.5 hr) treatments, the samples were centrifuged and the supernatant
326 loaded onto a Qiagen genomic tip column and processed according to
327 manufacturer's instructions.

328

329 **Shotgun library construction**

330 *3 kb library*. Briefly, 3 µg of DNA was randomly sheared to 2-4 kb fragments
331 using a HydroShear® (GeneMachines) and end-repaired using T4 DNA
332 polymerase and DNA Polymerase I, Large (Klenow) Fragment (New England
333 Biolabs). The DNA was agarose gel separated and gel-purified using the
334 QIAquick Gel Extraction Kit (Qiagen). Approximately 200 ng of sheared DNA was
335 then ligated into 100 ng of linearized and dephosphorylated pUC18 vector

336 (Roche) at 24.5˚C for 90 min using the Fast-Link™ DNA Ligation Kit (Epicentre).
337 The ligation product was electroporated into ElectroMAX DH10B™ cells
338 (Invitrogen) and plated on selective agar plates. Positive library clones were
339 robotically picked using the Q-Bot multitasking robot (Genetix) and grown in
340 selective media for sequencing.
341 *8 kb library*. Briefly, 10 µg of HMW DNA was randomly sheared to 6-8 kb
342 fragments and end-repaired as described above. The DNA was agarose gel
343 separated and filter tip gel-purified. Approximately 200 ng of DNA was blunt-end
344 ligated into 100 ng of pMCL200 vector O/N at 16ºC using T4 DNA ligase (Roche
345 Applied Science) and 10% (vol/vol) polyethylene glycol (Sigma). The ligation was
346 phenol-chloroform extracted, ethanol precipitated, and resuspended in 20 µl TE.
347 According to the manufacturers instructions, 1 µl of ligation product was
348 electroporated into ElectroMAX DH10B™ Cells and processed as described
349 above.
350 *Fosmid library*. The fosmid library was constructed using the CopyControl™
351 Fosmid Library Production Kit (Epicentre). DNA (~20 µg) was randomly sheared
352 using a HydroShear, blunt-end repaired as described above and separated on an
353 agarose pulse-field gel O/N at 4.5 V/cm. The 40 kb fragments were excised, gel-
354 purified using AgarACE™ (Promega) digestion followed by phenol-chloroform
355 extraction and ethanol precipitation. DNA fragments were ligated into the
356 pCC1Fos™ Vector and the ligation packaged using MaxPlax™ Lambda
357 Packaging Extract and used to transfect TransforMax™ EPI300 *E. coli*.

358 Transfected cells were plated on selective agar plates and fosmid clones picked
359 using the Q-Bot multitasking robot and grown in selective media for sequencing.
360

361 **End-sequencing**

362 The pUC library was sequenced using using DyEnamic ET Terminators and
363 resolved on MB4500 (MolecularDynamics/GeneralElectric). The pMCL and
364 pCC1Fos libraries were sequenced with BigDye Terminators v3.1 and resolved
365 with ABI PRISM 3730 (ABI) sequencers.
366

367 **Processing and Assembly of Shotgun Data**

368 A total of 22.15 Mb of phred Q20 sequence was generated from the three
369 libraries; 9.43 Mb from 13755 reads from the small insert pUC library, 8.79 Mb
370 from 13824 reads from the medium insert pMCL library, and 3.93 Mb from 9216
371 reads from the fosmid library. The DNA sequences derived from the *Ruthia*
372 *magnifica* libraries were estimated to be 20% contaminated with the *Calyptogena*
373 *magnifica* host genome. Although this level of contamination can confound
374 finishing efforts, the bacterial genome was readily identifiable in our study. The
375 36,795 sequencing reads were blasted against a database containing all mollusk
376 sequence available at NCBI and the 4X draft sequence available at the JGI for
377 *Lottia gigantea*. A total of 498 reads were removed based on hits to this mollusk
378 database. The remaining 24,595 reads were base called using phred version
379 0.990722.g, vector trimmed using crossmatch SPS-3.57, and assembled using

380 parallel phrap compiled for SUNOS, version SPS - 4.18. One large, bacterial
381 scaffold containing the *Ruthia magnifica* 16S rRNA gene resulted. The *Ruthia*
382 *magnifica* scaffold consisted of only 2 contigs spanned by 33 fosmid clones,
383 contained 17,307 reads, 1,156,121 consensus bp, was covered by an average
384 read depth of 14X, and had a G+C content of 34%. The next largest scaffold was
385 only 29 kb long, with an average read depth of ~7X and an average G+C content
386 of 55%. BLASTn indicated that this scaffold encoded ribosomal genes closely
387 related to those of *Caenorhabditis briggsae* and its binning (based on GC content
388 and read depth) with a small scaffold containing the *Calyptogena magnifica* 18S
389 rRNA gene confirmed its eukaryotic host origin.
390

391 **Annotation and pathway reconstruction**

392 Assembled sequence was first loaded into The Institute for Genomic Research's
393 (TIGR) auto-annotation pipeline before being imported into MANATEE
394 (http://manatee.sourceforge.net/), a web-based interface for manual annotation.
395 Only after putative genes were computationally and manually validated were they
396 assigned names and gene symbols. The TIGR guidelines for manual annotation
397 based on annotator confidence in computational evidence were followed. The
398 *Ruthia magnifica* genome was finished at the Joint Genome Institute and the
399 assembly is currently being quality checked.
400

400

401 1. C. M. Cavanaugh, Z. P. McKiness, I. L. G. Newton, F. Stewart, in *The*
402 *Prokaryotes*. (Springer-Verlag, Berlin, 2004).
403 2. K. J. Boss, R. D. Turner, *Malacologia* **20**, 161 (1980).
404 3. A. S. Peek, R. G. Gustafson, R. A. Lutz, R. C. Vrijenhoek, *Mar. Biol.* **130**,
405 151 (Dec, 1997).
406 4. H. Felbeck, G. Somero, *Trends Biochem. Sci.*, 201 (1982).
407 5. C. Cavanaugh, *Nature* **302**, 58 (1983).
408 6. J. J. Robinson, Ph.D., Harvard University (1997).
409 7. C. R. Fisher *et al.*, *Deep-Sea Res.* **35**, 1811 (1988).
410 8. J. J. Wernegreen, *Nature Rev. Genet.* **3**, 850 (Nov, 2002).
411 9. D. Wu *et al.*, *PLoS Biol.* **4**, e188 (Jun 6, 2006).
412 10. E. Zientz, T. Dandekar, R. Gross, *Microbiol. Mol. Biol. Rev.* **68**, 745 (Dec,
413 2004).
414 11. J. M. Shively, G. van Keulen, W. G. Meijer, *Annu. Rev. Microbiol.* **52**, 191
415 (1998).
416 12. J. J. Robinson, C.M. Cavanaugh, *Limnol. Oceanogr.* **40**, 1496 (1995).
417 13. C. M. Cavanaugh, J. J. Robinson, in *Microbial Growth on C^1 Compounds*
418 M. E. Lidstrom, F. R. Tabita, Eds. (Kluwer Academic Publishers,
419 Dordrecht, The Netherlands, 1996) pp. 285-292.
420 14. R. G. Kemp, R. L. Tripathi, *J. Bacteriol.* **175**, 5723 (Sep, 1993).
421 15. H. R. Beller *et al.*, *J. Bacteriol.* **188**, 1473 (Feb, 2006).
422 16. C. Dahl *et al.*, *J. Bacteriol.* **187**, 1392 (Feb, 2005).
423 17. J. A. Eisen *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 9509 (Jul 9, 2002).
424 18. F. Verte *et al.*, *Biochemistry* **41**, 2932 (Mar 5, 2002).
425 19. A. Fiala-medioni, C. Metivier, *Marine Biology* **90**, 215 (1986).
426 20. J. J. Childress, C. R. Fisher, J. A. Favuzzi, N. K. Sanders, *Physiol. Zool.*
427 **64**, 1444 (Nov-Dec, 1991).
428 21. D. Nelson, C. Fisher, in *Microbiology of deep-sea hydrothermal vents* D.
429 Karl, Ed. (CRC Press, Boca Raton, 1995).
430 22. F. Zal *et al.*, *Cah. Biol. Mar.* **41**, 413 (2000).
431 23. H. Kumagai, T. Fujiwara, H. Matsubara, K. Saeki, *Biochemistry* **36**, 5509
432 (May 6, 1997).
433 24. K. S. Johnson, J.J. Childress, R.R. Hessler, C.M. Sakamoto-Arnold, C.L.
434 Beehler, *Deap-Sea Res.* **35**, 1723 (1988).
435 25. A. P. Wood, J. P. Aurikko, D. P. Kelly, *FEMS Microbiol. Rev.* **28**, 335 (Jun,
436 2004).
437 26. P. S. Fritsch, M. L. Urbanowski, G. V. Stauffer, *J. Bacteriol.* **182**, 5539
438 (Oct, 2000).
439 27. I. M. Keseler *et al.*, *Nucleic Acids Res.* **33**, D334 (Jan 1, 2005).
440 28. C. J. Krieger *et al.*, *Nucleic Acids Res.* **32**, D438 (Jan 1, 2004).
441 29. M. E. Streams, C. R. Fisher, A. Fiala-Medioni, *Mar. Biol.* **129**, 465 (Sep,
442 1997).

443 30. N. W. Goehring, J. Beckwith, *Curr. Biol.* **15**, R514 (Jul 12, 2005).
444 31. W. J. Brown, D. D. Rockey, *Infect. Immun.* **68**, 708 (Feb, 2000).
445 32. A. J. McCoy, A. T. Maurelli, *Trends Microbiol.* **14**, 70 (Feb, 2006).
446 33. unpublished observation, C.M. Cavanaugh.
447 34. M. R. Hammerschlag, *Semin. Pediatr. Infect. Dis.* **13**, 239 (Oct, 2002).
448 35. A. S. Peek, R. C. Vrijenhoek, B. S. Gaut, *Mol. Biol. Evol.* **15**, 1514 (Nov,
449 1998).
450 36. J. A. Eisen, P. C. Hanawalt, *Mutat. Res.* **435**, 171 (Dec 7, 1999).
451 37. W. Martin *et al.*, *Nature* **393**, 162 (May 14, 1998).
452 38. N. A. Moran, P. H. Degnan, *Mol. Ecol.* **15**, 1251 (Apr, 2006).
453
454
455 39. Table S2 available on Science Online

456 **Acknowledgements**

457 This research was funded by a US Department of Energy grant to Cavanaugh
458 and Eisen and a Howard Hughes Medical Institute Predoctoral Fellowship to
459 Newton. Sequencing was carried out at the Joint Genome Institute and we thank
460 David Bruce and Eddy Rubin for project management. We thank Dan Fraenkel
461 for his review of metabolic pathways in *Ruthia* and Peter Girguis and Kathleen
462 Scott for their helpful comments on the manuscript.

463

463

464 **Figure 1**. Electron micrographs of *Ruthia magnifica* within host bacteriocytes.

465 (A) Bacteriocyte containing many small (0.3 μm) coccoid-shaped symbionts.

466 Scale bar = 5 μm (B) Higher magnification of *R. magnifica* showing the electron

467 dense granules suggestive of *Chlamydia's* "elementary bodies." Scale bar = 2

468 μm *D.* symbiont in putative degradative state, *N*, bacteriocyte nucleus, *R, R.*

469 *magnifica.*

470

471 **Figure 2.** The percentages of the genomes dedicated to different functional

472 categories as predicted by annotation are shown for γ-proteobacterial symbionts

473 (*Ruthia magnifica, Buchnera aphidicola*) and free-living relatives (*Thiomicrospira*

474 *crunogena* and *Escherichia coli,* respectively).

475

476 **Figure 3.** Three major metabolic pathways are shown as inferred from the

477 genomic content in *R. magnifica*. Enzymes or pathways present in the genome

478 are colored while those not yet identified are either white or dashed. The Calvin

479 cycle is used by the symbiont for carbon fixation and although missing fructose

480 1,6-bisphosphatase (FBPase) and sedoheptulose 1,7-bisphosphatase (SBPase),

481 it could use a reversible phosphofructokinase to regenerate ribulose 5-

482 phosphate. The sulfur oxidation pathway appeared similar to that of *Chlorobium*

483 *tepidum*. The Sox proteins act in the periplasm to oxidize thiosulfate while sulfide

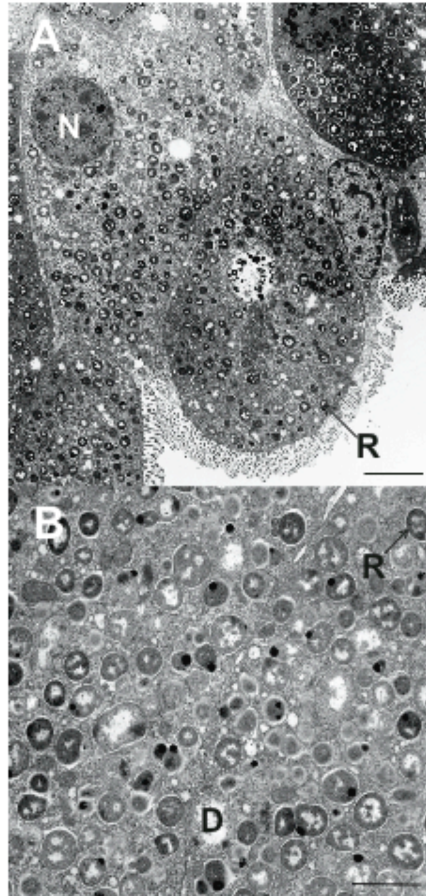
484 may be oxidized intracellularly by the reversible dissimilatory sulfate reductase

485 (dsr) system. Nitrogen assimilation pathways via both ammonia and nitrate are

486 present in the symbiont genome.

487

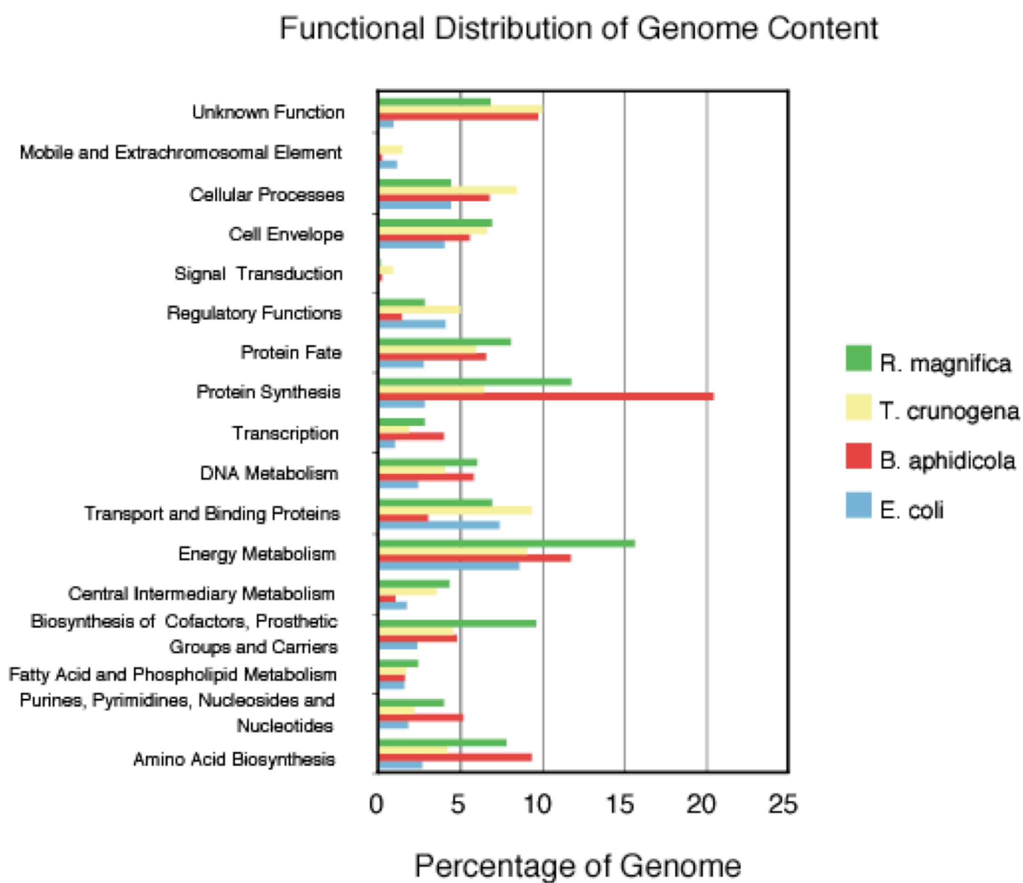
487 Figure 1.



488

489

490 Figure 2



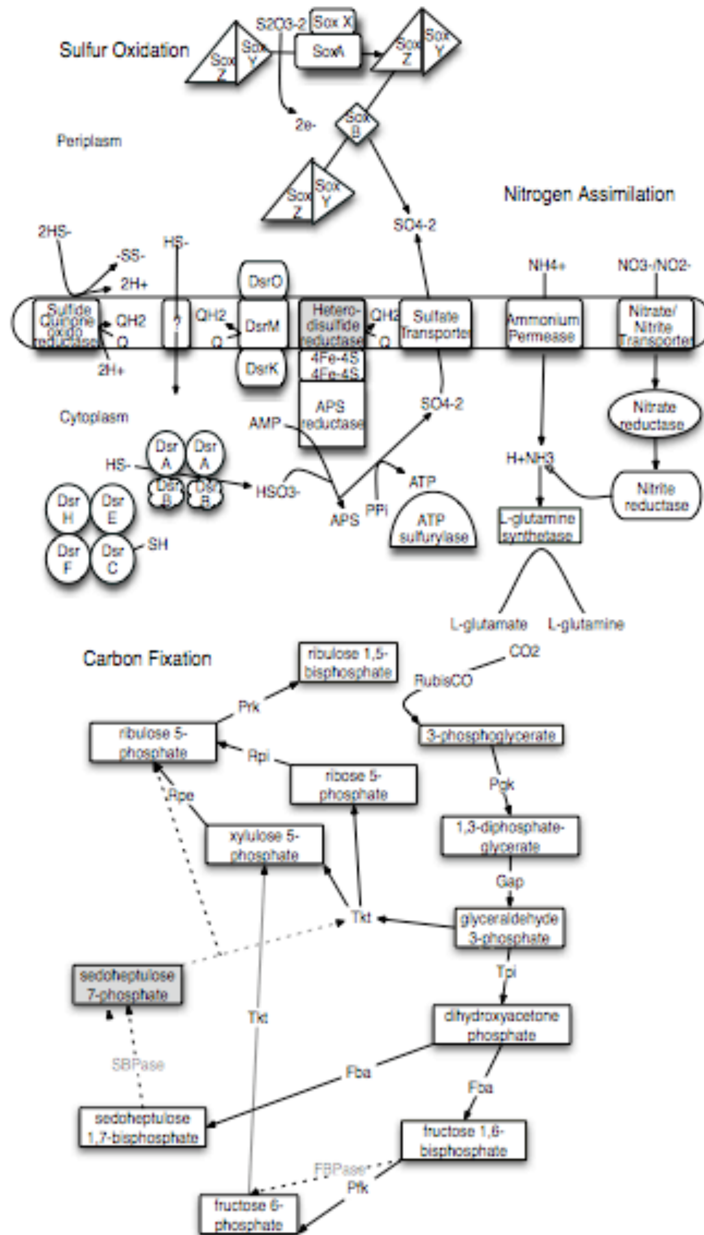
491

492

493

494

495



495 Figure 3

496
497 **Table 1.** General genome features of the chemoautotrophic symbiont *Ruthia*
498 *magnifica* compared with those of other γ-proteobacteria, including the free-living
499 chemoautotroph, *Thiomicrospira crunogena,* an obligately intracellular aphid
500 symbiont, *Buchnera aphidicola,* and a free-living relative of the aphid symbiont*,*
501 *Escherichia coli*.
502

Features	*Ruthia magnifica*	*Thiomicrospira crunogena*	*Buchnera aphidicola*	*Escherichia coli*
Chromosome, Mb	1.2	2.4	0.6	4.6
Plasmids	0	0	1	0
G+C content, %	34	43	26	50
Total gene number	1248	2199	608	4289
rRNAs	3	9	3	22
tRNAs	36	43	32	88
Protein-coding, %	81.4	97.8	86.5	97.9
Mean gene length, bp	975	948	991	800

503 • *E. coli* is closely related to *B. aphidicola,* with 87.2% sequence identity in
504 16S rRNA; *T. crunogena* and *R. magnifica* share 83.3% 16S rRNA
505 sequence identity.
506

506

507 Supporting online material

508

509 **Supplementary Table 2.** The *Ruthia magnifica* genome encodes pathways for

510 many metabolic processes and biosynthesis of important amino acids, vitamins

511 and cofactors. Complete pathways found in the genome are indicated by '+'

512 while absent pathways are indicated by '-'.

Pathway	Prediction
Glycolysis	+
TCA cycle	+
Glyoxylate shunt	Partial
Respiration	+
Pentose phosphate pathway	Partial
Fatty acid biosynthesis	+
Cell wall biosynthesis	Partial
Biosynthesis of all 20 amino acids	+
Vitamin and Cofactor Biosynthesis	
Heme	+
Ubiquinone	+
Nicotinate and nicotinamide	+
Folate	+
Lipoate	+
Riboflavin	+
Pantothenate	+
Pyridoxine	+
Thiamine	+
Biotin	+
Cobalamin	-

513
514