Supporting Online Material for

The *Calyptogena magnifica* Chemoautotrophic Symbiont Genome

I. L. G. Newton, T. Woyke, T. A. Auchtung, G. F. Dilly, R. J. Dutton, M. C. Fisher, K. M. Fontanez, E. Lau, F. J. Stewart, P. M. Richardson, K. W. Barry, E. Saunders, J. C. Detter, D. Wu, J. A. Eisen, C. M. Cavanaugh*

*To whom correspondence should be addressed. E-mail: cavanaug@fas.harvard.edu

This PDF file includes:

Supplementary Online Material for:
A window into hydrothermal vent endosymbioses:
the *Calyptogena magnifica* chemoautotrophic symbiont genome
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], E. Saunders[2], J.C. Detter[2], D. Wu[4], J. A. Eisen[5], C.M. Cavanaugh[1]

METHODS:

Specimen collection and DNA extraction:

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

Shotgun library construction

3 kb library. Briefly, 3 µg of DNA was randomly sheared to 2-4 kb fragments using a HydroShear® (GeneMachines) and end-repaired using T4 DNA polymerase and DNA Polymerase I, Large (Klenow) Fragment (New England Biolabs). The DNA was agarose gel separated and gel-purified using the QIAquick Gel Extraction Kit (Qiagen). Approximately 200 ng of sheared DNA was then ligated into 100 ng of linearized and dephosphorylated pUC18 vector (Roche) at 24.5°C for 90 min using the Fast-Link™ DNA Ligation Kit (Epicentre). The ligation product was electroporated into ElectroMAX DH10B™ cells (Invitrogen) and plated on selective agar plates. Positive library clones were robotically picked using the Q-Bot multitasking robot (Genetix) and grown in selective media for sequencing.

8 kb library. Briefly, 10 µg of HMW DNA was randomly sheared to 6-8 kb fragments and end-repaired as described above. The DNA was agarose gel separated and filter tip gel-purified. Approximately 200 ng of DNA was blunt-end ligated into 100 ng of pMCL200 vector O/N at 16ºC using T4 DNA ligase (Roche Applied Science) and 10% (vol/vol) polyethylene glycol (Sigma). The ligation was phenol-chloroform extracted, ethanol precipitated, and resuspended in 20 µl TE. According to the manufacturers instructions, 1 µl of ligation product was electroporated into ElectroMAX DH10B™ Cells and processed as described above.

Fosmid library. The fosmid library was constructed using the CopyControl™ Fosmid Library Production Kit (Epicentre). DNA (~20 µg) was randomly sheared using a HydroShear, blunt-end repaired as described above and separated on an agarose pulsed-field gel O/N at 4.5 V/cm. The 40 kb fragments were excised, gel-purified using AgarACE™ (Promega) digestion followed by phenol-chloroform

47 extraction and ethanol precipitation. DNA fragments were ligated into the
48 pCC1Fos™ Vector and the ligation packaged using MaxPlax™ Lambda
49 Packaging Extract and used to transfect TransforMax™ EPI300 *E. coli*.
50 Transfected cells were plated on selective agar plates and fosmid clones picked
51 using the Q-Bot multitasking robot and grown in selective media for sequencing.
52
53 **End-sequencing**
54 The pUC library was sequenced using using DyEnamic ET Terminators and
55 resolved on MB4500 (MolecularDynamics/GeneralElectric). The pMCL and
56 pCC1Fos libraries were sequenced with BigDye Terminators v3.1 and resolved
57 with ABI PRISM 3730 (ABI) sequencers.
58
59 **Processing and Assembly of Shotgun Data**
60 A total of 22.15 Mb of phred Q20 sequence was generated from the three
61 libraries; 9.43 Mb from 13755 reads from the small insert pUC library, 8.79 Mb
62 from 13824 reads from the medium insert pMCL library, and 3.93 Mb from 9216
63 reads from the fosmid library. The DNA sequences derived from the *Ruthia*
64 *magnifica* libraries were estimated to be 20% contaminated with the *Calyptogena*
65 *magnifica* host genome. Although this level of contamination can confound
66 finishing efforts, the bacterial genome was readily identifiable in our study. The
67 36,795 sequencing reads were blasted against a database containing all mollusk
68 sequence available at NCBI and the 4X draft sequence available at the JGI for
69 *Lottia gigantea*. A total of 498 reads were removed based on hits to this mollusk
70 database. The remaining 24,595 reads were base called using phred version
71 0.990722.g, vector trimmed using crossmatch SPS-3.57, and assembled using
72 parallel phrap compiled for SUNOS, version SPS - 4.18. One large, bacterial
73 scaffold containing the *Ruthia magnifica* 16S rRNA gene resulted. The *Ruthia*
74 *magnifica* scaffold consisted of only 2 contigs spanned by 33 fosmid clones,
75 contained 17,307 reads, 1,156,121 consensus bp, was covered by an average
76 read depth of 14X, and had a G+C content of 34%. The next largest scaffold was
77 only 29 kb long, with an average read depth of ~7X and an average G+C content
78 of 55%. BLASTn indicated that this scaffold encoded ribosomal genes closely
79 related to those of *Caenorhabditis briggsae* and its binning (based on GC content
80 and read depth) with a small scaffold containing the *Calyptogena magnifica* 18S
81 rRNA gene confirmed its eukaryotic host origin.
82
83 **Annotation and pathway reconstruction**
84 The large bacterial scaffold sequence was first loaded into The Institute for
85 Genomic Research's (TIGR) auto-annotation pipeline before being imported into
86 MANATEE (http://manatee.sourceforge.net/), a web-based interface for manual
87 annotation.   Only after putative genes were computationally and manually
88 validated were they assigned names and gene symbols. The TIGR guidelines
89 for manual annotation based on annotator confidence in computational evidence
90 were followed.
91
92   **The *R. magnifica* Endosymbiont Genome**

93
94 Although first discovered at hydrothermal vents, chemosynthetic symbioses exist
95 at mud flats, seagrass beds, and hydrocarbon seeps(*1*). In each case it is clear
96 that these symbioses play major roles in community structuring and sulfur and
97 carbon cycling. However, despite the widespread occurrence of these
98 partnerships, little is known of the intricacies of host-symbiont interaction or
99 symbiont metabolism due to their inaccessibility and our inability to culture either
100 partner separately.
101
102 The *R. magnifica* genome has revealed striking differences between the
103 chemosynthetic endosymbiont genomes and those of other obligate mutualistic
104 symbionts for which data are available (Fig. S1). The genome has a low G+C
105 content (34%) compared to free-living relatives (Table S1). In addition, the
106 coding density (81.4%) and mean gene length (975 bp), though lower than
107 commonly seen in free-living bacteria, are consistent with that in other
108 endosymbiont genomes (*2*). These common features of endosymbionts are
109 likely the result of genome reduction and mutation accumulations that occur over
110 evolutionary time across diverse symbiont species. This trend is evident in
111 relatively recent symbioses such as the insect endosymbionts (30-250 Ma), as
112 well as in chloroplasts (~1,800-2,100 Ma). Upon closer examination however, *R.*
113 *magnifica* stands out in that its genome is large for a maternally transmitted
114 endosymbiont (1.2 Mb). For example, the genomes of the γ-proteobacterial
115 *Buchnera* species, which are endosymbionts of aphids, are some 80% smaller
116 than closely related free-living species like *E. coli.* In contrast, *R. magnifica*'s
117 genome is half the size of its relative's, *Thiomicrospira crunogena,* a free-living, γ-
118 proteobacterial, sulfur-oxidizing chemoautotroph.
119
120 We propose that the limited genome reduction in *R. magnifica* is due to a
121 fundamental difference in its biology compared to other nutritional endosymbionts
122 characterized so far. Insect endosymbionts typically supplement the diet of their
123 hosts, e.g., *Buchnera* provide essential amino acids that are missing in the
124 phloem sap diet of aphids (*2*). Similarly, the γ-proteobacteria *Baumannia* and
125 *Sulcia* together provide amino acids and vitamins for their sharpshooter hosts,
126 but apparently not much more (*3*). These symbionts acquire much of what they
127 need (e.g., sugars) from their host and thus can still survive with very small
128 genomes (*4*). In contrast, the retention of the large number of biosynthetic
129 pathways argues that the *Ruthia* chemoautotrophic symbionts provide their host
130 clam with the majority of its nutrition.
131
132 In the following sections we discuss detailed aspects of the metabolic
133 reconstruction of *R. magnifica* and what this might mean for the biology of its
134 host.
135
136 The Calvin cycle in *R. magnifica* is unconventional. The genome lacks homologs
137 of sedoheptulose 1,7-bis-phosphatase (SBPase, EC 3.1.3.37) and fructose 1,6-
138 bis-phosphatase (FBPase, EC 3.1.3.11), suggesting that the regeneration of

139 ribulose 1,5-bisphosphate may not follow conventional routes. Instead, the *R.*
140 *magnifica* genome contains a reversible pyrophosphate-dependent
141 phosphofructokinase (EC 2.7.1.90) homolog that may be used to generate
142 fructose 6-phosphate (*5*).

143

144 Sulfur oxidation in *R. magnifica* proceeds via *sox* and *dsr* homologs. Homologs
145 of the *sox* genes are located in two positions in the *R. magnifica* genome with
146 *soxXYZA* located in a single operon while *soxB* is elsewhere. The symbiont's *dsr*
147 genes were contained in a single cluster, *dsrABEFHCMKLOP*, missing *dsrJNRS*.
148 As these latter proteins are not well characterized, it is not known how symbiont
149 sulfur metabolism may be affected.

150

151 Central intermediary metabolism in *R. magnifica* produces all of the
152 intermediates necessary for synthesis of amino acids and vitamins/cofactors for
153 the host. Interestingly, the symbiont is missing homologs of fumarate reductase,
154 succinyl-coa synthase, and succinate dehydrogenase. However, the genome
155 encodes isocitrate lyase, part of the glyoxylate shunt, and could produce
156 succinate from isocitrate.

157

158 Although able to synthesize 10 vitamins/cofators, the cobalamin (B_{12})
159 biosynthesis pathway is conspicuously absent. However, since cobalamin is a
160 cofactor for methionine synthase (*6*), and since *R. magnifica* encodes a
161 cobalamin-independent methionine synthase, the host is unlikely to require
162 cobalamin.

163

164 Although no substrate specific transporters were found in the *R . magfinica*
165 genome, transporters involved in chemoautotrophy (sulfate exporters), nitrogen
166 assimilation (ammonium and nitrate importers), inorganic compounds (TrkAH,
167 MgtE family, CaCA family and PiT family), and heavy metals (ZnuABC, RND
168 superfamily, iron permeases) were identified.

169

170 Interestingly, the *R. magnifica* genome lacked the key cell division gene, *ftsZ*.
171 FtsZ, a tubulin homolog, assembles as a ring within the bacterial cell, recruits the
172 remaining cell division proteins and constricts to initiate cytokinesis (*7*). It is
173 puzzling that *R. magnifica* lacked FtsZ given that it is almost universally
174 conserved in bacteria, with the notable exception of the obligately intracellular
175 pathogens in the Chlamydia division (*8*). In addition to the absence of *ftsZ, R.*
176 *magnifica* and Chlamydia both lack the *murI* gene (*9*), required for the synthesis
177 of D-glutamate, an essential component of the bacterial cell wall. The potential
178 similarities in cell division and cell wall machinery between *R. magnifica* and
179 Chlamydia may be responsible for the "elementary body" cell morphologies
180 observed in both organisms inside the host cell (*10*). In Chlamydia these bodies
181 are the infectious, propagating form (*11*); their appearance in *R. magnifica* may
182 reflect common mechanisms for adaptation to an obligately intracellular lifestyle.

183

184   Endosymbiont intracellular lifestyles have severe effects on genome evolution
185   including genome reductions, skewed base compositions, and elevated rates of
186   gene evolution (*2*). As noted above, *R. magnifica* does exhibit skewed
187   composition and genome reduction, although these are minor shifts compared to
188   those seen in insect endosymbionts. Previous studies have shown, however,
189   that *R. magnifica* also exhibits faster nucleotide substitution rates than those of
190   both free-living bacteria and environmentally transmitted chemosynthetic
191   symbionts (*12*). The factors that contribute to these features of endosymbiont
192   evolution are believed to be a combination of a relatively stable environment,
193   population bottlenecks (where few individuals contribute to the next generation),
194   and sequestration from free-living bacteria all of which likely occur in *R.*
195   *magnifica*. In addition, as with some but not all other endosymbionts, *R.*
196   *magnifica* has lost key genes in DNA repair processes that likely enhance
197   genome reduction and mutation rates. For example, it is missing genes involved
198   in induction of the SOS repair system and in recombinational repair, including the
199   exonuclease complex genes *recB,C,D* and the highly conserved recombinase
200   *recA* . Perhaps most importantly, it is also missing genes that could encode
201   homologs of the MutSLH proteins, which, in other species greatly limit mutation
202   rates by carrying out post-replication mismatch repair (*13*).
203
204   The *R. magnifica* genome is only ½ as large as that of its free-living relative,
205   *Thiomicrospira crunogena.* For comparison, chloroplast genomes have lost over
206   90% of their content since their cyanobacterial ancestor entered endosymbiosis,
207   with many of their genes having been transferred to the host nuclear genome
208   (*14*). The more modern insect endosymbioses have lost between 70-80% of
209   their genomes over a much shorter evolutionary time, and it is unknown if any of
210   these pathways are encoded by the nucleus (*4, 15*). In contrast, the *R. magnifica*
211   genome is the largest maternally transmitted endosymbiont genome sequenced
212   to date. It is the extent of this dependency that may be preventing the loss of
213   metabolic pathways in the *R. magnifica* genome. This selective pressure might
214   be great enough to counter the forces of genome reduction seen in other
215   endosymbionts and provides a novel framework for the study of endosymbiont
216   evolution.
217
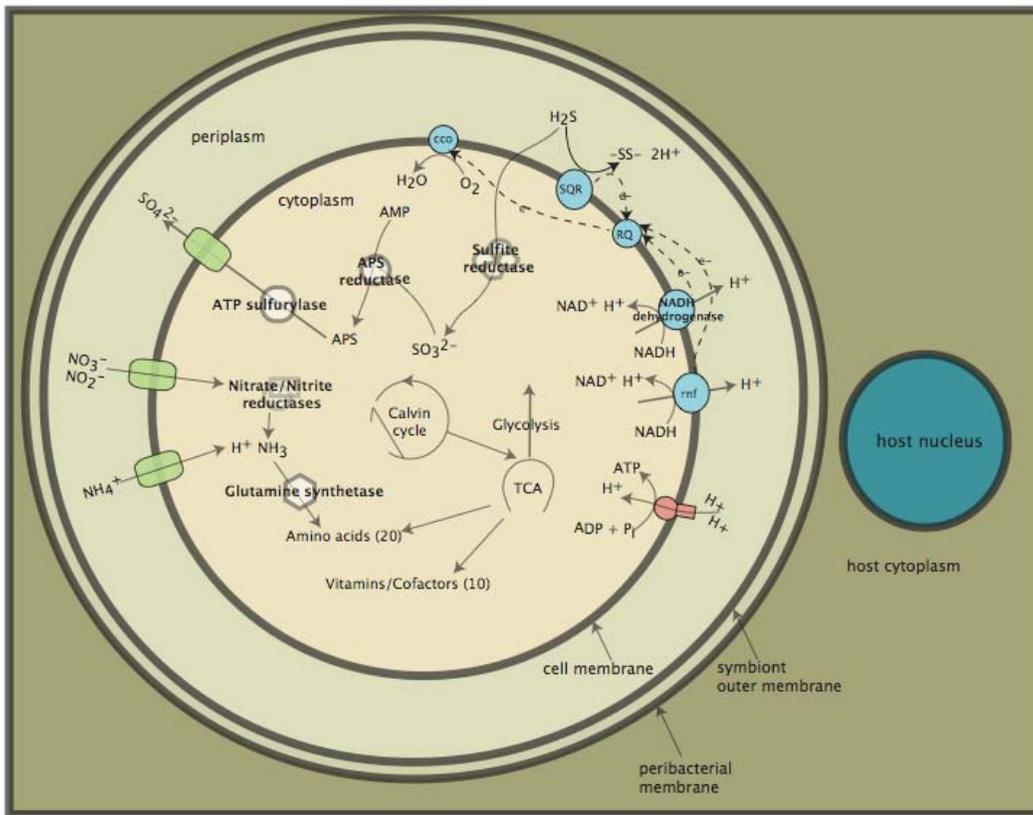
Figure S1

Key metabolic pathways inferred from the genomic content in *R. magnifica* are shown
and include sulfur metabolism, carbon fixation, carbon metabolism, and energy
metabolism. The Sox proteins act in the periplasm to oxidize thiosulfate while sulfide
may be oxidized intracellularly by the reversible dissimilatory sulfate reductase (dsr)
system. Electrons from this oxidation and from that of NADH (by NADH
dehydrogenase and the *rnf* complex) are carried by reduced quinone (RQ) to cytochrome
c oxidase (cco) and ultimately to oxygen. Energy derived from this oxidation fuels the
Calvin cycle, used by the symbiont for carbon fixation. Although missing fructose 1,6-
bisphosphatase (FBPase) and sedoheptulose 1,7-bisphosphatase (SBPase), the symbiont
could use a reversible phosphofructokinase to regenerate ribulose 5-phosphate. Nitrogen
assimilation pathways via both ammonia and nitrate are present in the symbiont genome
and incorporated nitrogen may be used in the synthesis of 20 amino acids and 10
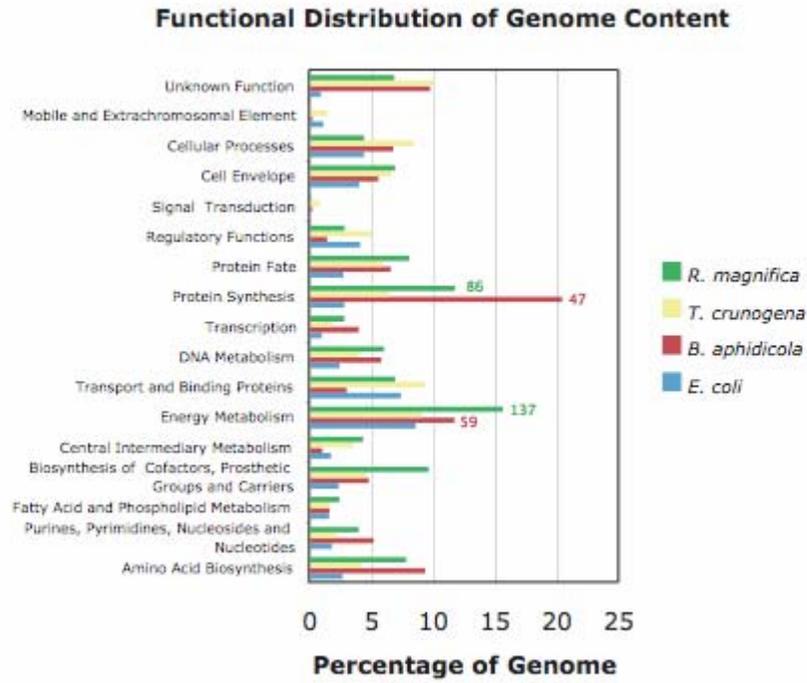vitamins and cofactors.

6
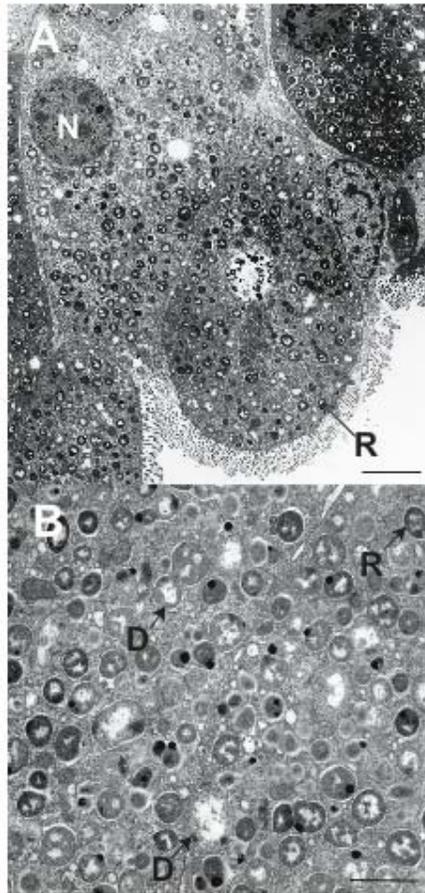
Figure S2

Electron micrographs of *Ruthia magnifica* within host bacteriocytes. (A) Bacteriocyte containing many small (0.3 m) coccoid-shaped symbionts. Scale bar = 5 m (B) Higher magnification of *R. magnifica* showing the electron dense granules suggestive of *Chlamydia's* "elementary bodies." Scale bar = 2 m *D.* symbiont in putative degradative state, *N*, bacteriocyte nucleus, *R, R. magnifica.*

240

Functional Distribution of Genome Content

241
242   **Figure S3**
243
244   The percentages of the genomes dedicated to different functional categories as
245   predicted by annotation are shown for γ-proteobacterial symbionts (*Ruthia*
246   *magnifica, Buchnera aphidicola*) and free-living relatives (*Thiomicrospira*
247   *crunogena* and *Escherichia coli,* respectively)*.* The number of genes in relevant
248   functional categories are shown next to bars for *Ruthia* and *Buchnera.*
249
250
251
252
253
254
255
256
257
258
259
260

8

Features	Ruthia magnifica	Thiomicrospira crunogena	Buchnera aphidicola	Escherichia coli
Chromosome, Mb	1.2	2.4	0.6	4.6
G+C content, %	34	43	26	50
Total gene number	1248	2199	608	4289
Protein-coding, %	81.4	97.8	86.5	97.9
Mean gene length, bp	975	948	991	800
Metabolism				
Glycolysis	+	+	-	+
TCA cycle	+	+	-	+
Glyoxylate shunt	Partial	Partial	-	+
Respiration	+	+	-	+
Pentose phosphate pathway	Partial	Partial	Partial	+
Fatty acid biosynthesis	+	+	-	+
Cell wall biosynthesis	Partial	+	+	+
Biosynthesis of 20 amino acids	+	+	Partial	+
Vitamin and Cofactor Biosynthesis				
Heme	+	+	-	+
Ubiquinone	+	+	-	+
Nicotinate and nicotinamide	+	+	+	+
Folate	+	+	-	+
Lipoate	+		+	+
Riboflavin	+	+	+	+
Pantothenate	+	+	-	+
Pyridoxine	+	+	-	+
Thiamine	+	+	-	+
Biotin	+	+	-	+
Cobalamin	-	Partial	-	+

262   *T. crunogena* and *R. magnifica* share 83.3% 16S rRNA sequence identity.

263   **Table S1.** General genome features and metabolic pathways of the
264   chemoautotrophic symbiont *Ruthia magnifica* compared with those of other γ-
265   proteobacteria, including the free-living chemoautotroph, *Thiomicrospira*
266   *crunogena,* and an obligately intracellular aphid symbiont, *Buchnera aphidicola.*
267   Complete pathways are indicated by a '+' while missing pathways are marked '-'

REFERENCES:

1. C. M. Cavanaugh, Z. P. McKiness, I. L. G. Newton, F. Stewart, in *The Prokaryotes*. (Springer-Verlag, Berlin, 2004).
2. J. J. Wernegreen, *Nat Rev Genet* **3**, 850 (Nov, 2002).
3. D. Wu *et al.*, *PLoS Biol* **4**, e188 (Jun 6, 2006).
4. E. Zientz, T. Dandekar, R. Gross, *Microbiol Mol Biol Rev* **68**, 745 (Dec, 2004).
5. R. G. Kemp, R. L. Tripathi, *J Bacteriol* **175**, 5723 (Sep, 1993).
6. M. E. Streams, C. R. Fisher, A. FialaMedioni, *Marine Biology* **129**, 465 (Sep, 1997).
7. N. W. Goehring, J. Beckwith, *Curr Biol* **15**, R514 (Jul 12, 2005).
8. W. J. Brown, D. D. Rockey, *Infect Immun* **68**, 708 (Feb, 2000).
9. A. J. McCoy, A. T. Maurelli, *Trends Microbiol* **14**, 70 (Feb, 2006).
10. u. o. C.M. Cavanaugh.
11. M. R. Hammerschlag, *Semin Pediatr Infect Dis* **13**, 239 (Oct, 2002).
12. A. S. Peek, R. C. Vrijenhoek, B. S. Gaut, *Molecular Biology and Evolution* **15**, 1514 (Nov, 1998).
13. J. A. Eisen, P. C. Hanawalt, *Mutat Res* **435**, 171 (1999).
14. W. Martin *et al.*, *Nature* **393**, 162 (May 14, 1998).
15. N. A. Moran, P. H. Degnan, *Mol Ecol* **15**, 1251 (Apr, 2006).