doi:10.1038/nature07660

LETTERS

A single regulatory gene is sufficient to alter bacterial host range

Mark J. Mandel[1], Michael S. Wollenberg[1], Eric V. Stabb[2], Karen L. Visick[3] & Edward G. Ruby[1]

Microbial symbioses are essential for the normal development and growth of animals[1–3]. Often, symbionts must be acquired from the environment during each generation, and identification of the relevant symbiotic partner against a myriad of unwanted relationships is a formidable task[4]. Although examples of this specificity are well-documented, the genetic mechanisms governing it are poorly characterized[5]. Here we show that the two-component sensor kinase RscS is necessary and sufficient for conferring efficient colonization of *Euprymna scolopes* squid by bioluminescent *Vibrio fischeri* from the North Pacific Ocean. In the squid symbiont *V. fischeri* ES114, RscS controls light-organ colonization by inducing the Syp exopolysaccharide, a mediator of biofilm formation during initial infection. A genome-level comparison revealed that *rscS*, although present in squid symbionts, is absent from the fish symbiont *V. fischeri* MJ11. We found that heterologous expression of RscS in strain MJ11 conferred the ability to colonize *E. scolopes* in a manner comparable to that of natural squid isolates. Furthermore, phylogenetic analyses support an important role for *rscS* in the evolution of the squid symbiosis. Our results demonstrate that a regulatory gene can alter the host range of animal-associated bacteria. We show that, by encoding a regulator and not an effector that interacts directly with the host, a single gene can contribute to the evolution of host specificity by switching 'on' pre-existing capabilities for interaction with animal tissue.

Genomic technologies are facilitating major advances in understanding the relationships between metazoans and their bacterial symbionts. The analysis of unculturable endosymbionts has revealed complex genetic interdependence between host and bacteria amid patterns of genome reduction in endosymbiotic lineages[3]. Similarly, members of the human microbiota are being identified through metagenomic analysis[1,2], and the molecular communication between host and microbe has begun to be interpreted through transcriptional profiling[6]. Despite these advances, the mechanisms by which host–symbiont specificity develops in animal–bacterial interactions are not clear. Many animals, including humans, are born devoid of symbionts and must recruit their microbiota from the environment[7]. The process by which hosts and symbionts find each other to initiate a mutualism must be sensitive enough to identify the correct partner even when the symbiont is a minority constituent of the microbial community, and specific enough to exclude interlopers from gaining access to the host. The basis of species specificity is also poorly understood for pathogenic interactions, as similar congeneric bacteria often have distinct host ranges[8–10].

In this study we used a comparative genomics approach to reveal how bacteria–host specificity is established in the *E. scolopes*–*V. fischeri* mutualism. We took advantage of the fact that *V. fischeri* strain MJ11, which was isolated from the Japanese pinecone fish, *Monocentris japonica*[11], is unable to colonize *E. scolopes* efficiently. As such, comparison of MJ11 with natural squid symbionts provided a valuable system for examining the genomic basis of host specificity in an animal symbiont.

Although the genome sequence of squid-symbiotic *V. fischeri* strain ES114 is known[12,13], we determined here the sequence of the fish symbiont MJ11. Genome assembly of MJ11 was based on the ES114 model using a combination of PCR- and fosmid-based approaches. Genome sequencing also revealed a 179-kilobase (kb) circular plasmid in MJ11 that we term pMJ100, in which 82% of the open reading frames are annotated as hypothetical proteins, and which is distinct from the plasmid carried by ES114. Alignment of the assembled chromosomes revealed two circular MJ11 chromosomes that are co-linear to those in ES114 (Fig. 1a). Over 90% of ES114 open reading frames are shared by MJ11, and the orthologues have a median amino-acid identity of 98.8%. One exception to the high level of conservation was significant divergence observed specifically in the LuxR quorum-sensing system (Supplementary Fig. 1 and Supplementary Discussion).

Examination of ES114 genes for those that could facilitate specific recognition identified *rscS* as a promising candidate because its product acts during symbiotic initiation[14,15], and we discovered it to be absent in the MJ11 genome (Fig. 1b). RscS is a membrane-bound two-component sensor kinase that acts upstream of the response-regulator SypG[16]. SypG, a σ^{54}-dependent transcriptional activator, facilitates transcription of the 18-gene exopolysaccharide locus *sypA-R*[17]. Production of the Syp exopolysaccharide enables *V. fischeri* aggregation in squid-derived mucus during colonization of *E. scolopes*. During growth in culture, *syp* genes are expressed at low levels but can be induced by the plasmid-borne *rscS1* overexpression allele[15,17,18], leading to the production of robust biofilms. Because the MJ11 genome revealed an intact *syp* locus, we asked whether signal transduction downstream of *rscS* was maintained in MJ11 by introducing *rscS1*. As shown in Fig. 1c, d, *rscS1* in MJ11 induced multiple biofilm phenotypes, suggesting that the *syp* locus of MJ11 was functional. We therefore examined whether *rscS* was sufficient to allow MJ11 to colonize *E. scolopes* efficiently.

We tested ES114 and MJ11 for their ability to colonize aposymbiotic *E. scolopes* hatchlings in a 3-h inoculation assay. ES114 colonized successfully, whereas MJ11 failed to initiate colonization, even if present at a tenfold higher inoculum concentration (Fig. 1e). However, when provided with *rscS*$^{+}$ *in trans*, MJ11 was competent to colonize *E. scolopes* to levels comparable to those seen with the natural symbiont ES114 (Fig. 1e). Furthermore, the luminescence emitted by MJ11/*rscS*$^{+}$-colonized animals was 100-fold greater than that from animals colonized by ES114. The increased luminescence is consistent with that of the brighter fish symbiont[19], and was not influenced by plasmid carriage (Supplementary Fig. 2). This result argues that the carrying capacity of the juvenile squid light organ is specified by symbiont cell

[1]Department of Medical Microbiology and Immunology, University of Wisconsin School of Medicine and Public Health, 1550 Linden Drive, Madison, Wisconsin 53706, USA. [2]Department of Microbiology, University of Georgia, 828 Biological Sciences, Athens, Georgia 30602, USA. [3]Department of Microbiology and Immunology, Loyola University Chicago, Maywood, Illinois 60153, USA.

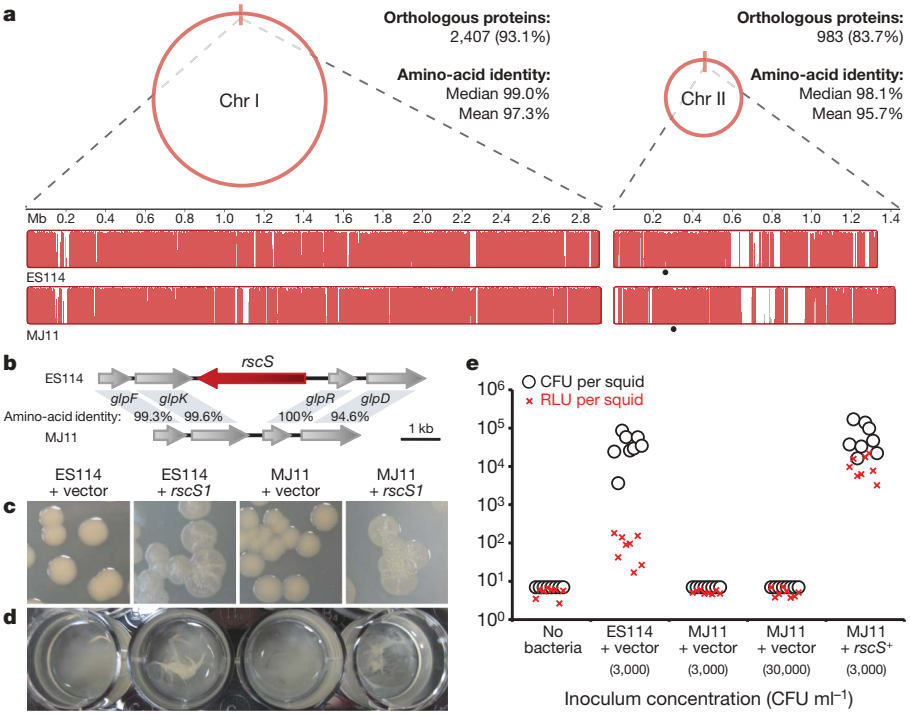


Figure 1 | *rscS* is sufficient to confer efficient colonization of *E. scolopes*. **a**, Mauve output shows each chromosome (chr) as one locally co-linear block. **b**, *rscS* is absent in MJ11 despite a high level of conservation in the surrounding locus. **c**, **d**, RscS-controlled biofilm phenotypes, including colony wrinkling (**c**) and pellicle formation (**d**) of strains harbouring either the vector control (pKV69) or the vector carrying *rscS1* (pKG11). **e**, Squid-colonization assay of strains harbouring either the vector control or the vector carrying *rscS*$^{+}$ (pLMS33), after 3 h inoculation and assayed at 48 h post-inoculation. Data points are individual animals. CFU, colony-forming units; RLU, relative light units.

number and not by the amount of luminescence emitted, provided that a minimum threshold of light production is achieved[20,21].

We next asked whether *rscS* was present in a collection of *V. fischeri* squid and fish isolates from the North Pacific Ocean to determine whether the gene's host distribution was consistent with a functional role in nature. All *V. fischeri* in the analysis revealed the presence of three representative *syp* genes (Fig. 2a). In contrast, although all of the squid isolates encoded *rscS*, regardless of geography, only five of the ten fish isolates encoded *rscS*; in four of these five the allele was significantly divergent. PCR amplification of the divergent alleles produced a band that was distinctively larger than the allele in ES114 and the rest of the squid isolates, whereas MJ12 was the only fish isolate that had this smaller squid–symbiont band (Fig. 2a).

We term the allele encoded by the smaller band $rscS_A$, and that encoded by the larger band $rscS_B$. $RscS_A$ was found in all assayed North Pacific squid isolates, and fish isolate MJ12, whereas $RscS_B$ was identified only in four fish isolates. Sequencing revealed that, within each type, the alleles for RscS are highly conserved (amino-acid identity at least 96%), but that divergence between the types was greater (84–86%; Supplementary Fig. 3a). The presence of an identical domain structure in all *V. fischeri* RscS proteins (Supplementary Fig. 3b) led us to ask whether there was detectable functional significance to this level of divergence.

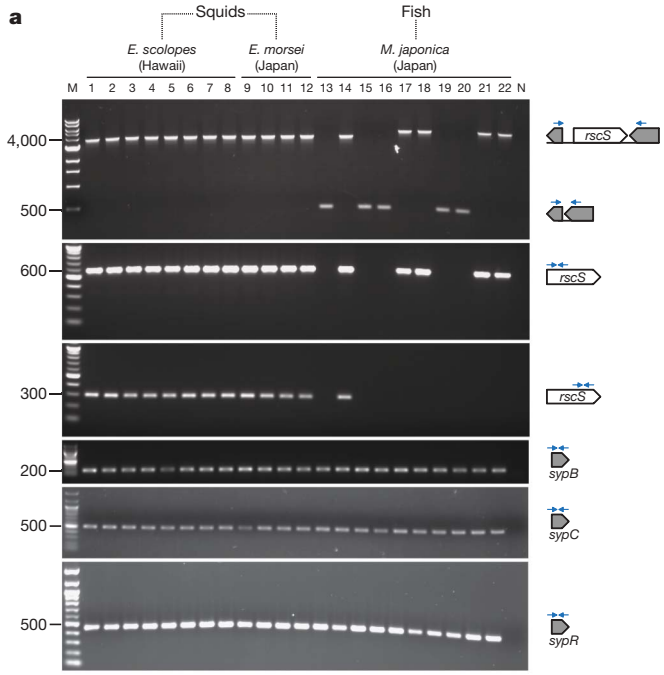
All of the $rscS_A$ strains were competent to colonize *E. scolopes* squid efficiently (Fig. 2b). In contrast, strains lacking *rscS* or encoding the divergent $rscS_B$ were unable to colonize consistently. The defect appeared to be due to RscS function and not to the *syp* locus or other differences; introduction of *rscS*$^{+}$ from ES114 (A-type) into $rscS_B$-containing *mjapo*.8.1 conferred 100% colonization efficiency (Fig. 2b). The only fish-symbiotic strain that was competent to colonize *E. scolopes* reproducibly (MJ12) was also the only one with the conserved $rscS_A$ allele. Interruption of *rscS* in MJ12 abolished its ability to colonize *E. scolopes* (Fig. 2b), confirming that $rscS_A$ is both sufficient and necessary to colonize the squid host in these populations.

To understand the evolution of *rscS* and its role in determining specificity in nature, we reconstructed the phylogeny of *V. fischeri* strains (Fig. 3 and Supplementary Fig. 4) using three well-characterized loci. Strains encoding *rscS* formed a monophyletic group within *V. fischeri* that was statistically well supported. Parametric bootstrapping rejected the alternative hypothesis of non-monophyletic origin for the *rscS*-encoding strains, at a significance level of $P < 0.01$.

We propose a model for *rscS* evolution in the symbioses of North Pacific Ocean squids and fish. This model represents a parsimonious synthesis of the colonization and genomics data, within the phylogenetic framework. Specifically, we hypothesize that an acquisition event introduced *rscS* into the *V. fischeri* lineage before the expansion of this species into squid hosts in the North Pacific Ocean (Figs 3 and 4). An initial acquisition, followed by vertical transmission of *rscS* among *V. fischeri*, would predict both a similar guanine–cytosine (GC) content among all *rscS* alleles in the species, and a single conserved genomic location for the gene in all extant *V. fischeri* genomes. We confirmed these predictions, as the *rscS* alleles from *V. fischeri* have similar GC-content (Supplementary Fig. 3a) and are present in the same genomic position (Fig. 2a).

Because the fish isolates that contain *rscS* fall within the same clade as squid isolates (Fig. 3), we argue that the fish- and squid-symbiotic populations in Japan are, indeed, sympatric and that the *rscS*-containing fish isolates are descendents of squid symbionts. We hypothesize that $rscS_A$ diverged significantly in the fish host to generate the $rscS_B$ allele, which is not sufficient to allow these strains to colonize the squid host niche. Further, because $RscS_B$ maintained its reading frame and domain structure—despite significant amino-acid divergence and loss of function for squid colonization—we hypothesize that $RscS_B$ is fish-adapted, and may play a role in activating *syp*, and/or other targets, under fish-specific conditions. The identification of a fish isolate encoding $RscS_A$ provides strong evidence that the $rscS_A$ locus does not preclude successful fish colonization by *V. fischeri*, despite the low frequency of this allele among the fish isolates examined. Unfortunately, *M. japonica* eggs do not fully develop in the laboratory, so we are unable to test this aspect of our model by investigating *V. fischeri* colonization of fish[22].

There are two formal possibilities for how *rscS* first entered the *V. fischeri* lineage. Either a gene duplication/translocation event within *V. fischeri* led to the initial generation of *rscS*, or *rscS* was generated outside *V. fischeri* and then acquired by horizontal gene transfer. We have found no DNA sequences paralogous to either of the *rscS* alleles in the full genomes of ES114 or MJ11. These data are also the most compelling, if indirect, evidence supporting the proposal of horizontal gene transfer. Nonetheless, it is difficult to reconstruct the event that introduced *rscS* to *V. fischeri* based on the usual criteria that define larger genomic islands (for example, direct repeats or insertion elements in flanking DNA, or aberrant codon usage or GC content within



b

Lane	Strain	Host	*rscS* type	Number of *E. scolopes* colonized/attempted
1	ES114	*Es*	A	29/29 0/8 ...[*rscS*$^-$]
2	KB1A97	*Es*	A	16/16 0/8 ...[*rscS*$^-$]
3	KB2B1	*Es*	A	8/8
4	KB5A1	*Es*	A	8/8
5	ES213	*Es*	A	8/8
6	MB11B1	*Es*	A	8/8
7	MB13B1	*Es*	A	8/8
8	MB14A1	*Es*	A	8/8
9	EM17	*Em*	A	8/8
10	EM18	*Em*	A	8/8
11	EM24	*Em*	A	8/8
12	EM30	*Em*	A	8/8
13	MJ11	*Mj*	–	0/21
14	MJ12	*Mj*	A	20/20 0/8 ...[*rscS*$^-$]
15	*mjapo*.2.1	*Mj*	–	1/8
16	*mjapo*.3.1	*Mj*	–	0/8
17	*mjapo*.4.1	*Mj*	B	0/8
18	*mjapo*.5.1	*Mj*	B	0/8
19	*mjapo*.6.1	*Mj*	–	1/8
20	*mjapo*.7.1	*Mj*	–	0/8
21	*mjapo*.8.1	*Mj*	B	0/8 4/4 ...[+*rscS*$_A$]
22	*mjapo*.9.1	*Mj*	B	1/8

Figure 2 | The presence of *rscS* is correlated to a natural association with squid and to the ability to colonize *E. scolopes* experimentally. **a**, PCR assay for the presence of *rscS* and *syp* genes with primer sets (top to bottom) *rscS*-flank, *rscS*-internal1, *rscS*-internal2, *sypB*-internal, *sypC*-internal or *sypR*-internal. M, markers; N, negative control. **b**, Colonization competence of strains in **a**, colonized as described in Fig. 1, at a concentration of approximately 5×10^3 colony-forming units per millilitre. Entries in red indicate *rscS*$^-$ derivatives; entry in blue indicates a strain carrying a plasmid (pLMS33) encoding *rscS*$^+$ from ES114. *Em*, *Euprymna morsei*; *Es*, *E. scolopes*; *Mj*, *M. japonica*.

rscS)[23]. Furthermore, the only convincing orthologue of *rscS* outside *V. fischeri* is *V. shiloi* AK1 VSAK1_16757. If a horizontal transfer event were responsible for *rscS* transmission into *V. fischeri*, the *V. shiloi* orthologue is unlikely to be the source: the GC content of ES114 *rscS* is 31.7%, or 6.6% below the ES114 genome average (38.3%). The *V. shiloi* orthologue has a GC content (41.0%) that is even higher than this average, and 9.3% higher than that of the ES114 *rscS* allele.

Attempts to understand the molecular basis of host specificity have been unsuccessful in many pathogen–host animal interactions. *Salmonella enterica* serovar Typhi can infect only humans, whereas serovar Typhimurium has a broad host range that includes mice[9].

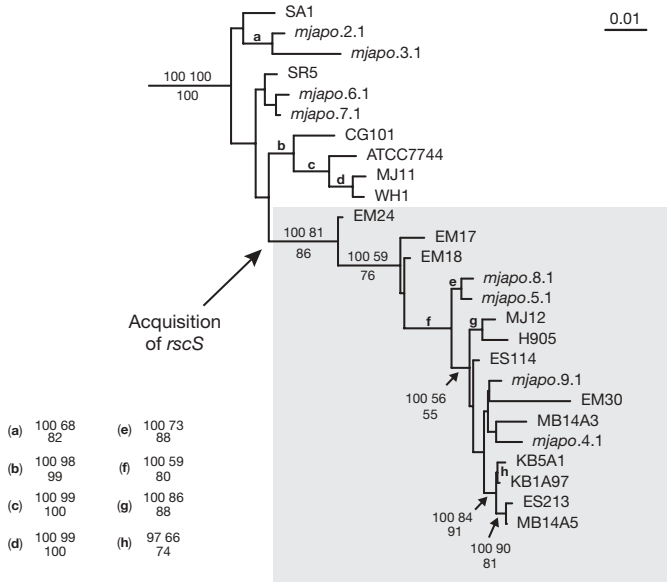


Figure 3 | *rscS* entered the *V. fischeri* lineage before colonization of squid in the North Pacific Ocean. Bayesian phylogeny of 26 *V. fischeri* strains using the concatenation of three loci (*recA*, *mdh*, *katA*). The tree was rooted with *Aeromonas salmonicida* subsp. *salmonicida* A449. Support for branches is indicated in cases where support is greater than 50% for all three: Bayesian posterior probabilities (top left), maximum likelihood bootstrap (top right), and maximum parsimony bootstrap (bottom). *rscS* DNA was detected in all strains in the grey box and in none of the strains outside the box. Bar: 0.01 expected changes per site.

Although the conserved regions of the genomes of these two strains are over 97% identical, efforts to account for this differential host specificity have not succeeded. Similarly, different *Brucella* species share over 98% identity across 90% of their genes, yet exhibit strict host specificity; the molecular basis of this specificity remains unclear[10]. In contrast, the study of mutualisms is providing insight into how specificity develops. In plant-associated bacteria, work from many laboratories has established nitrogen-fixing, nodulating rhizobia as the best-understood system for the development and evolution of host specificity[24]. Bacteria secrete Nod factors—lipo-chitooligosaccharide signals—to the plant host, and host-strain-specific backbone modifications encoded by the bacteria lead to relationship specificity. Recently, in an animal–bacterial mutualism, the *nilABC* genes of *Xenorhabdus nematophila* were characterized as sufficient for colonization of *Steinernema carpocapsae* worms by congeneric *Xenorhabdus* bacteria[25].

In contrast to rhizobia and *Xenorhabdus*, in which specificity comes either from the modification of a secreted signal or from structural proteins in the cell envelope, respectively, *rscS*-mediated specificity in *V. fischeri* is novel because the immediate effect of cytoplasmic RscS is on bacterial gene expression, which only subsequently has an effect on the interaction with the host. Because RscS is a signal transduction protein, the evolutionary consequence of the introduction of *rscS* appears to be a re-programming of inherent *V. fischeri* capabilities to expand the host range into squid populations that *V. fischeri* could not previously colonize, or could colonize only inefficiently. It remains a mystery as to why the *syp* genes are conserved in *V. fischeri* strains that are naive to *rscS* (for example, MJ11). That such *syp* clusters are functional, and ancestral to *rscS* in *V. fischeri*, strongly suggests that regulation of *syp* in these strains may be achieved in a manner independent of *rscS*. In support of this, there are *V. fischeri* isolated from the Mediterranean Sea that lack *rscS*, yet have *syp* genes and colonize squid hosts of a different genus through morphological structures that are conserved with those of *E. scolopes*[26].

Our study indicates that a regulatory gene is sufficient to alter host range in an animal–bacterial mutualism. The fundamental biological question of how animal–bacterial partnerships are established has been difficult to access through investigations of pathogenic interactions. In

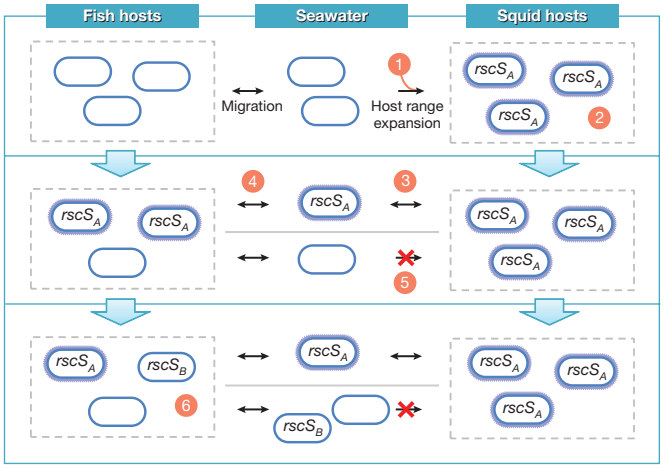


Figure 4 | Reconstruction of the evolution of *V. fischeri* symbioses in the North Pacific Ocean. All strains are genotypically *syp*$^+$; strains that are phenotypically Syp$^+$ in squid are illustrated with the grey mottled outline. The model reconstructs host range expansion into squid by (1) acquisition of *rscS*$_A$, and (2) subsequent developmentally appropriate expression of the ancestral *syp* polysaccharide genes. (3) Gene flow into planktonic strains and (4) into fish symbionts accounts for the presence of *rscS*$_A$ in fish isolates, but (5) *rscS*$^-$ strains maintain a colonization incompatibility for squid. (6) We postulate that *rscS*$_B$ evolved from *rscS*$_A$ in fish.

contrast, mutualism evolves to confer joint benefits to its partners, and relies on a strict specificity for this outcome: that is, entry of only a few appropriate symbionts and exclusion of the many non-specific interlopers. The evolution of developmental mechanisms to winnow the appropriate partner(s) is a hallmark of all horizontally acquired mutualisms. The binary squid–*Vibrio* system thus represents a valuable model in which to interrogate the mechanisms that underlie the development of bacteria–host specificity.

METHODS SUMMARY

The previously deposited draft genome of *V. fischeri* MJ11 was assembled into the final scaffold by comparing contigs to the ES114 assembly using Mauve[27]. Hypotheses were tested by PCR across the contig gaps or by sequencing of fosmids spanning the gap. In a few cases, no PCR product was produced, and the model was refined by rearranging contigs and retesting. In this manner, all of the contigs were arranged relative to ES114, and contig gaps that could be spanned by PCR were sequenced to complete the gap sequence. Three gaps on chromosome I contained tandem (at least two) *rrn* operons; in these cases, the sequence flanking the gap was PCR-amplified through the first ribosomal RNA (rRNA) gene at each end of the *rrn* array so that the completed genome was expected to contain all predicted open reading frames in *V. fischeri* MJ11.

We identified and corrected frameshift and nonsense mutations in the genome model[12], and the final sequences were annotated by the J. Craig Venter Institute (JCVI). To identify ES114–MJ11 orthologues, we performed reciprocal BLASTP searches between the predicted proteomes. Percentage identity was used to score results, and at least 60% coverage of each protein was demanded.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 18 August; accepted 20 November 2008.
Published online XX 2008.

1. Aas, J. A., Paster, B. J., Stokes, L. N., Olsen, I. & Dewhirst, F. E. Defining the normal bacterial flora of the oral cavity. *J. Clin. Microbiol.* **43,** 5721–5732 (2005).
2. Gill, S. R. *et al.* Metagenomic analysis of the human distal gut microbiome. *Science* **312,** 1355–1359 (2006).
3. Moran, N. A. Symbiosis as an adaptive process and source of phenotypic complexity. *Proc. Natl Acad. Sci. USA* **104** (suppl. 1), 8627–8633 (2007).
4. Mazmanian, S. K., Round, J. L. & Kasper, D. L. A microbial symbiosis factor prevents intestinal inflammatory disease. *Nature* **453,** 620–625 (2008).
5. Groisman, E. A. & Casadesús, J. The origin and evolution of human pathogens. *Mol. Microbiol.* **56,** 1–7 (2005).
6. Sonnenburg, J. L. *et al.* Glycan foraging *in vivo* by an intestine-adapted bacterial symbiont. *Science* **307,** 1955–1959 (2005).
7. Xu, J. & Gordon, J. I. Honor thy symbionts. *Proc. Natl Acad. Sci. USA* **100,** 10452–10459 (2003).
8. Brinig, M. M., Register, K. B., Ackermann, M. R. & Relman, D. A. Genomic features of *Bordetella parapertussis* clades with distinct host species specificity. *Genome Biol.* **7,** R81 (2006).
9. Edwards, R. A., Olsen, G. J. & Maloy, S. R. Comparative genomics of closely related salmonellae. *Trends Microbiol.* **10,** 94–99 (2002).
10. Rajashekara, G., Glasner, J. D., Glover, D. A. & Splitter, G. A. Comparative whole-genome hybridization reveals genomic islands in *Brucella* species. *J. Bacteriol.* **186,** 5040–5051 (2004).
11. Ruby, E. G. & Nealson, K. H. Symbiotic association of *Photobacterium fischeri* with the marine luminous fish *Monocentris japonica*; a model of symbiosis based on bacterial studies. *Biol. Bull.* **151,** 574–586 (1976).
12. Mandel, M. J., Stabb, E. V. & Ruby, E. G. Comparative genomics-based investigation of resequencing targets in *Vibrio fischeri*: focus on point miscalls and artefactual expansions. *BMC Genomics* **9,** 138 (2008).
13. Ruby, E. G. *et al.* Complete genome sequence of *Vibrio fischeri*: a symbiotic bacterium with pathogenic congeners. *Proc. Natl Acad. Sci. USA* **102,** 3004–3009 (2005).
14. Visick, K. L. & Skoufos, L. M. Two-component sensor required for normal symbiotic colonization of *Euprymna scolopes* by *Vibrio fischeri*. *J. Bacteriol.* **183,** 835–842 (2001).
15. Yip, E. S., Geszvain, K., DeLoney-Marino, C. R. & Visick, K. L. The symbiosis regulator *rscS* controls the *syp* gene locus, biofilm formation and symbiotic aggregation by *Vibrio fischeri*. *Mol. Microbiol.* **62,** 1586–1600 (2006).
16. Hussa, E. A., Darnell, C. L. & Visick, K. L. RscS functions upstream of SypG to control the *syp* locus and biofilm formation in *Vibrio fischeri*. *J. Bacteriol.* **190,** 4576–4583 (2008).
17. Yip, E. S., Grublesky, B. T., Hussa, E. A. & Visick, K. L. A novel, conserved cluster of genes promotes symbiotic colonization and σ^{54}-dependent biofilm formation by *Vibrio fischeri*. *Mol. Microbiol.* **57,** 1485–1498 (2005).
18. Geszvain, K. & Visick, K. L. Multiple factors contribute to keeping levels of the symbiosis regulator RscS low. *FEMS Microbiol. Lett.* **285,** 33–39 (2008).
19. Stabb, E. V., Schaefer, A., Bose, J. L. & Ruby, E. G. in *Chemical Communication among Bacteria* (eds Winans, S. C. & Bassler, B. L.) 233–250 (ASM Press, 2008).
20. Bose, J. L., Rosenberg, C. S. & Stabb, E. V. Effects of *luxCDABEG* induction in *Vibrio fischeri*: enhancement of symbiotic colonization and conditional attenuation of growth in culture. *Arch. Microbiol.* **190,** 169–183 (2008).
21. Visick, K. L., Foster, J., Doino, J., McFall-Ngai, M. & Ruby, E. G. *Vibrio fischeri lux* genes play an important role in colonization and development of the host light organ. *J. Bacteriol.* **182,** 4578–4586 (2000).
22. Haygood, M. G. Light organ symbioses in fishes. *Crit. Rev. Microbiol.* **19,** 191–216 (1993).
23. Hacker, J. & Kaper, J. B. Pathogenicity islands and the evolution of microbes. *Annu. Rev. Microbiol.* **54,** 641–679 (2000).
24. Long, S. R. Genes and signals in the rhizobium–legume symbiosis. *Plant Physiol.* **125,** 69–72 (2001).
25. Cowles, C. E. & Goodrich-Blair, H. The *Xenorhabdus nematophila nilABC* genes confer the ability of *Xenorhabdus* spp. to colonize *Steinernema carpocapsae* nematodes. *J. Bacteriol.* **190,** 4121–4128 (2008).
26. Foster, J. S., Von Boletzky, S. & McFall-Ngai, M. J. A comparison of the light organ development of *Sepiola robusta* Naef and *Euprymna scolopes* Berry (Cephalopoda: Sepiolidae). *Bull. Mar. Sci.* **70,** 141–153 (2002).
27. Darling, A. C., Mau, B., Blattner, F. R. & Perna, N. T. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* **14,** 1394–1403 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank: P. Dunlap for sharing bacterial strains; N. Perna, J. Glasner, K. Geszvain, D. Baum, J. Johnson, M. Sarmiento and S. Ferriera for technical assistance; A. Wier, N. Bekiares, R. Gates and the Hawaii Institute of Marine Biology for animal facilities and care; J. McCosker and the Steinhart Aquarium for access to fish specimens; M. McFall-Ngai, H. Goodrich-Blair, C. Brennan and J. Troll for discussions; and L. Proctor for project support. MJ11 genome sequencing was funded by the Gordon and Betty Moore Foundation Marine Microbial Genome Sequencing Project; E.G.R. and co-workers are funded by the National Institutes of Health–National Center for Research Resources and the National Science Foundation Division of Integrative Organismal Systems; E.V.S. is funded by a National Science Foundation CAREER Award; K.L.V. is funded by the National Institute of General Medical Sciences; M.J.M. is funded by a National Institute of General Medical Sciences National Research Service Award Postdoctoral Fellowship; M.S.W. is funded by a National Science Foundation Predoctoral Fellowship and a National Institutes of Health Molecular Biosciences Training Grant to the University of Wisconsin.

Author Contributions M.J.M. designed the experiments, performed all work not described below, and wrote the paper. M.S.W. conducted the phylogenetic studies. K.L.V. constructed plasmids and strains, and imaged biofilm phenotypes. M.J.M. planned and performed the genome assembly and analytics, and M.J.M., E.V.S. and E.G.R. analysed the bioinformatics results.

Author Information The *recA*, *mdh*, *katA* and *rscS* sequence data from the additional strains described in the article are deposited in GenBank under accession numbers EU907941–EU908017; MJ11 genome data are deposited under accession numbers CP001133, CP001134 and CP001139. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to M.J.M. (mmandel@wisc.edu).

METHODS

Bacterial growth, strains and plasmids. Standard microbial techniques were used to construct strains and plasmids[28]. Growth of *V. fischeri* was at 20–28 °C with aeration. Media for growth of *V. fischeri* was LBS[29], and for *E. coli* was LB[30] or brain heart infusion (Bacto, Becton Dickinson). Antibiotics used included: chloramphenicol (2.5 μg ml^{-1} for *V. fischeri*, 25 μg ml^{-1} for *E. coli*), erythromycin (5 μg ml^{-1} for *V. fischeri*, 150 μg ml^{-1} for *E. coli*), and tetracycline (5 μg ml^{-1} for *V. fischeri*).

Plasmids pKV69, pLMS33 and pKG11 were introduced into *V. fischeri* by triparental conjugation as described previously[28] with *E. coli* carrying the pEVS104 helper plasmid. Briefly, overnight cultures of the following strains were used for the reaction. One hundred microlitres from each of donor and helper *E. coli* strains were pelleted at 16,000*g* for 2 min in a microfuge tube, and the supernatant was aspirated. One hundred microlitres of recipient *V. fischeri* was added to the same tube, and pelleted as above. After aspiration of the supernatant, the pellet was re-suspended in 10 μl LBS, and the entire 10 μl was spotted onto an LBS agar plate, and incubated at 28 °C overnight. The spot was re-suspended in 500 μl LBS, and 50 μl were plated onto selective media (LBS–chloramphenicol) to select for plasmid transfer.

The *rscS* mutagenesis plasmid pKV188 was constructed by subcloning an approximately 700-base-pair (bp) internal *rscS* fragment from ES114, terminating in the internal *Pst*I site, into the *Kpn*I/*Pst*I sites of pEVS122 (ref. 28). Mutagenesis of *rscS* in the strains noted (Supplementary Table 1) was as described previously[31]; briefly, after triparental conjugation as above, integration of the suicide vector was identified by selection of the entire mating spot on LBS–erythromycin.

V. fischeri strain MJ11 (alias MJ101) was isolated by sterile expression of the light-organ sample from a live *M. japonica* at the Steinhart Aquarium, in February 1991. *M. japonica* symbiont strains denoted '*mjapo.#.#*' were shared by P. Dunlap.

Sources of other strains were as noted in Supplementary Table 1.

MJ11 genome assembly. The *V. fischeri* MJ11 draft genome was sequenced by the JCVI as part of the Gordon and Betty Moore Foundation Marine Microbial Genome Sequencing Project. Cloning and shotgun sequencing was performed at JCVI; draft coverage was obtained at 8.57-fold coverage, and the contigs were previously deposited into GenBank as a whole genome shotgun sequencing project, accession number ABIH00000000, project version 01.

Step 1: sealing contig gaps. *V. fischeri* MJ11 contigs from the above project were aligned to the ES114 genome using Mauve[27] and Projector2 (ref. 32). Alignment of all contigs that were from multiple reads (greater than 2 kb), excluding repetitive rDNA sequences, identified strong matches to the ES114 chromosome I and chromsome II, with the exception of 179-kb contig number 1101159000798. Owing to the high level of conservation between the strains, the corresponding ES114 sequences at all contig gaps were used as estimates of the gap lengths—and as guides for sequencing primers across long gaps—and primers were designed to amplify across each contig-gap, including at least an additional 200 bp of overlap with each adjacent contig, and extending beyond regions of repetitive DNA at either side of the gap boundary. PCR primers were designed with Primer3Plus[33] and sequencing primers were designed at the SGD website (http://seq.yeastgenome.org/cgi-bin/web-primer). In some cases, custom software was used to assist in the identification of probable-unique regions for primer-binding sites within extended repeat regions.

Contig number 1101159000798 was unique in bearing no homology to ES114 or to any extended sequence in GenBank at the start of this project (autumn 2006). We postulated that this represented a large (circular) plasmid in MJ11, and consistent with this hypothesis primers pointing outward from both ends of the contig together amplified a small fragment (less than 1 kb).

Step 2: PCR-walking into tandem rRNA operons. Three of the contig gaps contained tandem rRNA operons, and PCR across the entire gap was unsuccessful. We initiated PCR-walking into each gap by amplifying from one end of the gap to a conserved region that was distal to the 16S–23S spacer on that end of the gap, using primers listed in Supplementary Table 2. Using this approach, we identified the six unique spacer sequences that comprised the terminal rRNA spacers for the three gaps, leaving only additional rRNA (and potentially transfer RNA (tRNA)) sequences remaining to be sealed within each of these three regions. These contigs were still assembled into a scaffold and submitted for deposition so that information about the relative positions of DNA sequences within the molecule (chromosome I) was preserved.

Step 3: re-sequencing of select targets. We previously published methodologies to identify and correct sequencing errors in microbial genomes[12]. We applied this technology to the MJ11 genome and identified 15 high-priority re-sequencing targets. Nine of these sites were in fact in error, and we corrected these errors for inclusion in the final genome release. The PCR/sequencing primers used to target these regions are listed in Supplementary Table 2.

After assembly, the sequence was resubmitted to the JCVI Annotation Service and post-processing at JCVI and the National Center for Biotechnology Information, and deposited into GenBank as listed in Supplementary Table 3.

Orthologue comparisons between ES114 and MJ11. We compared the predicted proteomes from both ES114 chromosomes against both MJ11 chromosomes. Reciprocal exhaustive BLASTP[34] searches were performed with an expect cutoff of 10. Results were filtered to demand that the query length and subject length each be a minimum of 60% of their respective total lengths. Among the remaining results for each query protein, best-hits were scored by percentage amino-acid identity, and additional results were included for analysis if they scored at least 70% of the maximum score for that query. ES114–MJ11 protein pairs included on reciprocal lists were candidate orthologues, and for pairs in which there was a duplicate of query or subject protein, manual assignment of orthology was curated using the parameters of percentage amino-acid identity, percentage of each protein aligned and the local genomic context (synteny) of the two proteins.

Biofilm phenotypes. The two biofilm phenotypes evaluated were colony morphology and pellicle formation. To assay for the ability to form wrinkled colonies, cells were streaked onto LBS–tetracycline agar and the plates were incubated at room temperature for two days. To assay for the ability to form pellicles, cells were inoculated into HEPES minimal medium[35] containing 0.3% casamino acids, 0.2% glucose and tetracycline at a final concentration of 30 μg ml^{-1}. After overnight growth at 28 °C with shaking, cells were diluted to an $OD_{600\,nm}$ of 0.1 in fresh HEPES minimal medium. Three millilitres of cell suspensions were introduced into the wells of a 12-well microtitre dish, and the cells were incubated statically at room temperature for five days. To facilitate visualization and imaging of the pellicles, the media surface was disturbed with a pipette tip, resulting in clumps of cells if a pellicle had formed.

Squid colonization assays. Juvenile *E. scolopes* hatchlings were collected aposymbiotically, and washed in Instant Ocean (Aquarium Systems) that was filter-sterilized through a 22-μm pore-sized Nalgene filter (FSIO: filter-sterilized Instant Ocean). Overnight cultures of bacteria in LBS were subcultured 1:40, grown for 70 min, assayed for $OD_{600\,nm}$, and then inoculated into 40 ml FSIO-squid at a volume equivalent to 1.25 μl per $OD_{600\,nm}$. The inoculum was plated onto LBS plates to confirm that the bacterial concentration was 2×10^3 to 10×10^3 colony-forming units per millilitre. Squid were washed with fresh FSIO at 3 and 24 h post-inoculation. Individual animal luminescence was recorded at 48 h post-inoculation before they were euthanized at 48 h by freezing at −80 °C. Symbiont colony-forming units per squid were determined by homogenizing thawed animals, and plating the homogenates onto LBS. For experiments involving pKV69-series plasmids, squid hatchlings were maintained in FSIO containing chloramphenicol at a final concentration of 2.5 μg ml^{-1}.

Luminescence of animals is reported as RLU (1 RLU $\approx 1.98 \times 10^4$ quanta per second) $= 24 \times lum$, where *lum* is the recorded luminescence of a single animal in a TD20/20 luminometer, with recordings performed in 4 ml FSIO in glass scintillation vials at 51.9% sensitivity with integration for 6 s. Animals with RLU > 25 were scored as colonized. Because the strains that failed to colonize *E. scolopes* were significantly brighter than ES114, this metric served as a conservative measure of colonization competency for the set of isolates examined in this study.

PCR amplification for MJ11 genome closure. PCR amplification was conducted using Platinum Taq DNA Polymerase High-Fidelity (Invitrogen). Fifty-microlitre reactions contained: 50 ng MJ11 genomic DNA, 1× reaction buffer, 0.2 mM of each dNTP, 2 mM $MgSO_4$, 0.25 μM of each primer, and 1 U DNA polymerase. At least three independent PCR reactions were combined for sequencing to minimize the effect of PCR error. Thermal cycling was conducted in a PTC-200 thermal cycler (MJ Research): 95 °C for 2 min; then 30 cycles of 95 °C for 30 s, 55 °C for 30 s, 68 °C for 30 s to 1 min per kilobase amplified; then 68 °C for 5 min.

Products greater than 5 kb were amplified with Platinum Pfx DNA Polymerase (Invitrogen) before sequencing. Fifty-microlitre reactions contained: 50 ng MJ11 genomic DNA, 2× reaction buffer, 0.3 mM of each dNTP, 1 mM $MgSO_4$, 0.30 μM of each primer, and 1 U DNA polymerase. Thermal cycling was conducted as above.

Primers for MJ11 genome closure are listed in Supplementary Table 2.

Diagnostic PCR amplification of rscS and syp genes. Conditions were as described above (MJ11 genome closure) with the following alterations. Template preparation consisted of bacterial strains grown overnight in LBS, diluted 1:100 in dH_2O, and then used as template in the PCR reactions at a dilution of 1:10. Annealing temperature for the *sypR*-internal primer set was 50 °C. Primers for diagnostic amplification of *rscS* and *syp* locus genes are listed in Supplementary Table 2.

PCR amplification for phylogenetic analyses. PCR amplification was conducted using GoTaq (Promega). Bacterial strains were grown overnight in

LBS, diluted 1:100 in dH_2O, and then used as template in the PCR reactions at a dilution of 1:10. Twenty-five-microlitre reactions contained: template preparation (2.5 μl), 1 × colourless reaction buffer, 0.2 mM of each dNTP, 0.9 μM of each primer and 1 U DNA polymerase. Thermal cycling was conducted in a PTC-200 thermal cycler (MJ Research): 95 °C for 3 min; then 26 cycles of 94 °C for 30 s, 55 °C for 30 s, 72 °C for 1 min; then 72 °C for 10 min. Primers for phylogenetic analyses are listed in Supplementary Table 2.

DNA sequencing. Sanger-type sequencing of PCR products for MJ11 genome assembly, and for phylogenetic analyses, was performed at the University of Washington High-Throughput Genomics Unit (Seattle, Washington) and the University of Wisconsin Biotechnology Center DNA Sequencing Facility (Madison, Wisconsin), with the primers listed in Supplementary Table 2.

Sequence data for *rscS* from *V. fischeri* ES114 and from *V. shiloi* AK1 are from GenBank accession numbers AF319618 and EDL55668, respectively. Sequence data for phylogenetic analysis of the following strains are as noted: *A. salmonicida* subsp. *salmonicida* (GenBank accession number CP000644), *Vibrio harveyi* BB120 (CP000789, CP000790), *V. parahaemolyticus* RIMD2210633 (BA000031, BA000032), *Photobacterium profundum* 3TCK (AAPH00000000) and *V. fischeri* ES114 (CP000020, CP000021).

Phylogenetic analyses. Sequences from the three loci (*recA*, *mdh*, *katA*) were aligned using ClustalX 1.83 (ref. 36), and trimmed and concatenated using custom Perl scripts (https://mywebspace.wisc.edu/wollenberg/web/science/scripts/scripts.html) and MEGA4 (ref. 37). The best-fit model of DNA substitution and parameter estimates used for tree reconstruction was chosen by performing hierarchical likelihood ratio tests on these data, as implemented in PAUP* 4.0b10 (ref. 38) and MODELTEST 3.7 (ref. 39); this model was SYM + I + G, a submodel under the GTR + I + G (general time reversible with gamma-rate distribution across sites and a proportion of invariant sites). Phylogenetic trees' clade topology and confidence were studied using three approaches. (1) MrBayes 3.1 (ref. 40), implementing the Markov chain Monte Carlo method with an evolutionary model set to GTR with gamma-distributed rate variation across sites and a proportion of invariable sites, was run for 5,000,000 generations using the CIPRES project portal (http://www.phylo.org/). Sample frequency was 1,000, creating a posterior probability distribution of 5,000 trees; when summarizing the substitution model parameters and trees, 1,250 trees were discarded as burn-in to address potential chain instability. (2) Maximum likelihood analyses were performed using the genetic algorithm approach of GARLI[41] as implemented in the CIPRES portal, with an evolutionary model set to GTR with gamma-distributed rate variation across sites, and a proportion of invariable sites. Bootstrap analysis of 1,000 replications was used to assess the support for internal nodes. (3) Unweighted maximum parsimony analysis and bootstrap were performed by PAUP* (1,000 replications) using heuristic searches implementing tree bisection and reconnection branch-swapping to find the shortest trees and assess the support for internal nodes. For maximum likelihood and Bayesian approaches, the process was independently repeated three times to ensure arrival at a similar, most-likely tree topology. Resulting trees were rooted with *A. salmonicida* subsp. *salmonicida* A449 as the outgroup (Supplementary Fig. 4).

Parametric bootstrap analysis. From our original data, the difference in likelihood scores between an unconstrained phylogeny and a constrained phylogeny with a non-monophyletic, *rscS*-containing clade was calculated. One hundred simulations of the data set were created using the constrained topology; likelihood scores were produced from these 100 simulated data sets both with and without the constraint of the non-monophyly of the *rscS*-containing clade using the software PAUP*. Our null hypothesis of the significance of the constraint of non-monophyly of the *rscS*-containing clade within our initial phylogeny was rejected based on analysis of the resulting likelihood ratio distribution ($P < 0.01$).

Codon usage. Whole genome codon usage for ES114 was analysed by the methods of Karlin and Mrázek[42], against a set of reference sequences from ES114, which included ribosomal proteins, chaperones and transcription/translation factors. Calculations were performed on the Computational Microbiology Laboratory server (http://www.cmbl.uga.edu/software/phxpa.html), and the codon usage of *rscS* was predicted to be neither highly expressed (PHX) nor alien (PA). All comparisons of codon bias performed placed *rscS* in the 95% confidence interval for the ES114 genome.

Bioinformatic notes and software used. In addition to the software noted above, Lasergene Seqbuilder and Seqman (DNASTAR) were used for sequencing and genome assembly. Mauve[27] was used extensively during the dynamic process of contig assembly and orientation. Analysis of RscS domain structure was assisted by PFAM[43] and Phobius[44]. Primer design was aided by Primer3Plus[33]. TreeView 1.6.6 (ref. 45) was used to view phylogenetic trees.

28. Stabb, E. V. & Ruby, E. G. RP4-based plasmids for conjugation between *Escherichia coli* and members of the Vibrionaceae. *Methods Enzymol.* **358,** 413–426 (2002).
29. Stabb, E. V. *et al.* in *Recent Advances in Marine Science and Technology 2000* (ed. Saxena, N.) 269–277 (PACON International, 2001).
30. Silhavy, T. J., Berman, M. L. & Enquist, L. W. *Experiments with Gene Fusions* (Cold Spring Harbor Laboratory Press, 1984).
31. Hussa, E. A., O'Shea, T. M., Darnell, C. L., Ruby, E. G. & Visick, K. L. Two-component response regulators of *Vibrio fischeri*: identification, mutagenesis, and characterization. *J. Bacteriol.* **189,** 5825–5838 (2007).
32. van Hijum, S. A., Zomer, A. L., Kuipers, O. P. & Kok, J. Projector 2: contig mapping for efficient gap-closure of prokaryotic genome sequence assemblies. *Nucleic Acids Res.* **33,** W560–W566 (2005).
33. Untergasser, A. *et al.* Primer3Plus, an enhanced web interface to Primer3. *Nucleic Acids Res.* **35,** W71–W74 (2007).
34. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–3402 (1997).
35. Ruby, E. G. & Nealson, K. H. Pyruvate production and excretion by the luminous marine bacteria. *Appl. Environ. Microbiol.* **34,** 164–169 (1977).
36. Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. & Higgins, D. G. The CLUSTAL X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **25,** 4876–4882 (1997).
37. Tamura, K., Dudley, J., Nei, M. & Kumar, S. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol. Biol. Evol.* **24,** 1596–1599 (2007).
38. Swofford, D. L. *PAUP*: Phylogenetic Analysis Using Parsimony (* and Other Methods)* 4th edn (Sinauer, 2003).
39. Posada, D. & Crandall, K. A. MODELTEST: testing the model of DNA substitution. *Bioinformatics* **14,** 817–818 (1998).
40. Ronquist, F. & Huelsenbeck, J. P. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19,** 1572–1574 (2003).
41. Zwickl, D. J. *Genetic Algorithm Approaches for the Phylogenetic Analysis of Large Biological Sequence Datasets under the Maximum Likelihood Criterion.* PhD. thesis, Univ. Texas at Austin (2006).
42. Karlin, S. & Mrázek, J. Predicted highly expressed genes of diverse prokaryotic genomes. *J. Bacteriol.* **182,** 5238–5250 (2000).
43. Finn, R. D. *et al.* Pfam: clans, web tools and services. *Nucleic Acids Res.* **34,** D247–D251 (2006).
44. Kall, L., Krogh, A. & Sonnhammer, E. L. Advantages of combined transmembrane topology and signal peptide prediction – the Phobius web server. *Nucleic Acids Res.* **35,** W429–W432 (2007).
45. Page, R. D. TreeView: an application to display phylogenetic trees on personal computers. *Comput. Appl. Biosci.* **12,** 357–358 (1996).