Untangling Genomes from Metagenomes: Revealing an Uncultured Class of Marine Euryarchaeota

Vaughn Iverson, Robert M. Morris, Christian D. Frazar, Chris T. Berthiaume, Rhonda L. Morales, E. Virginia Armbrust*

Ecosystems are shaped by complex communities of mostly unculturable microbes. Metagenomes provide a fragmented view of such communities, but the ecosystem functions of major groups of organisms remain mysterious. To better characterize members of these communities, we developed methods to reconstruct genomes directly from mate-paired short-read metagenomes. We closed a genome representing the as-yet uncultured marine group II *Euryarchaeota*, assembled de novo from 1.7% of a metagenome sequenced from surface seawater. The genome describes a motile, photo-heterotrophic cell focused on degradation of protein and lipids and clarifies the origin of proteorhodopsin. It also demonstrates that high-coverage mate-paired sequence can overcome assembly difficulties caused by interstrain variation in complex microbial communities, enabling inference of ecosystem functions for uncultured members.

Abundant and diverse populations of microbes shape every marine environment on Earth (*1*, *2*). Most environmental microbes are uncultured and resistant to laboratory-based physiological and genomic investigations (*3*, *4*). The importance of this uncultured majority cannot be overstated because microbial life plays a critical role in the biogeochemical cycling of elements, maintaining the chemical state of the oceans (*5*). How these communities may respond to remediate or exacerbate human impacts is unclear because their ecosystem function remains poorly understood.

Glimpses into the genomes of uncultured microbes are provided by environmental genomic approaches such as single-cell genomics (*6*) and the sequencing of long segments [10 to 60 kilo–base pairs (kbp)] of DNA cloned directly from environmental samples (*7*). Shotgun metagenomics bypasses manual isolation of individual sequences and instead yields large catalogs of randomly sampled environmental genes (*8*–*11*). Connecting these genes to specific taxonomic groups has been problematic except for groups previously sequenced or those with relative abundances exceeding ~20% (*12*, *13*). The greater sequencing depth provided by next-generation technologies (*14*) has recently been exploited to partially assemble multiple genomes from cow rumen microbiomes (*15*). Several of these assemblies were estimated to exceed 90% completeness, although all originated from taxonomic orders of Bacteria with cultured and sequenced representatives. We used mate-paired next-generation sequencing of two marine metagenomic samples to reveal the genome sequences of organisms constituting small minorities (<10%) of the total populations, including an uncultured marine group II *Euryarchaeote* genome.

School of Oceanography, University of Washington, Box 357940, Seattle, WA 98195, USA.

*To whom correspondence should be addressed. E-mail: armbrust@uw.edu

We collected cells (<0.8 μm) from surface waters of Puget Sound in October 2008 and May 2009 (*16*). Each sample harbored communities of marine bacteria and archaea present at ~10^6 cells per ml (table S1). Massively parallel sequencing (*14*) with a SOLiD DNA analyzer (Life Technologies Corporation, Carlsbad, California) produced a total of 58.5 gigabases of 50-base mate-paired reads (table S2). Analysis of metagenomic 16*S* ribosomal DNA (rDNA) sequence revealed diverse communities, with most family-level taxonomic groups present at less than 10% [figs. S1 to S9 and supporting online material (SOM) text]. Mismatch-tolerant alignments with the Global Ocean Sampling (GOS) (*10*) and RefSeq (*17*) databases recruited about 10 and 2% of reads, respectively (tables S2 and S3), highlighting the substantial proportion of previously unsequenced populations in our samples.

De novo assembly of the metagenomic reads produced ~300 megabases of contigs that recruited about threefold more metagenomic reads (~30%) than the GOS database (table S2). High-coverage mate-pairing information was exploited to link the assembled contigs (spanning short unassembled sequences), creating metagenomic assembly graphs (Fig. 1 and figs. S10 and S11). These connection graphs were split heuristically by using mate-pairing bit scores, nucleotide composition, and read-coverage statistics, producing parsimonious linear scaffolds. Scaffolds were binned by using tetra-nucleotide statistics into candidate genomes, which were taxonomically profiled by using mate-pair connections to informative 16*S* rDNA regions. Fourteen candidate genomes were produced, each representing 4 to 10% of a sampled community, including representatives of *Euryarchaeota*, *Thaumarchaeota*, *Flavobacteria*, and alpha-, beta-, and gamma-*Proteobacteria*. We selected one candidate genome from each sample for more detailed analysis.

A nearly complete genome closely related to the cultured and sequenced *Rhodobacterales* bacterium strain HTCC2255 was reconstructed entirely de novo from a population composing ~6.3% of the October sample (see *Thalassobacter*, fig. S6).

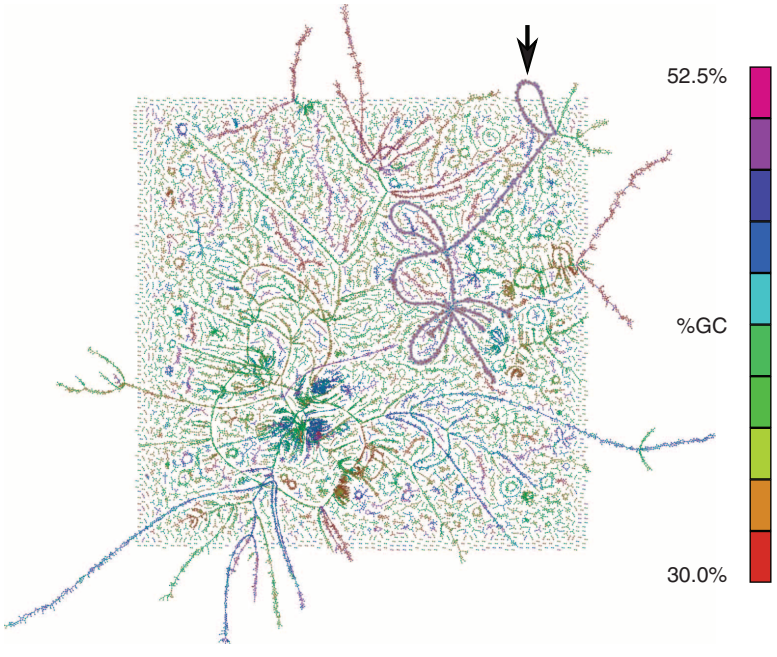

Fig. 1. Mate-pair connection graph illustrating the May 2009 metagenome de novo assembly. Lines represent contigs with mate-pair connections scoring greater than 750 bits (n = 30,945). Long strands represent prokaryote genome sequences, and small circular strands show likely virus or plasmid sequences. Contigs aligning with the MG-II genome assembly are indicated (black arrow, gray shading).

Nucleotide alignments of the binned scaffolds to the HTCC2255 reference genome show all 41 scaffolds aligning, with gaps reflecting the lack of metagenomic reads from the October sample recruiting to those regions of the reference genome (figs. S10 and S12). The remarkable level of agreement between the assembled and reference genomes demonstrates the robustness of our approach.

A closed genome (designated herein as MG-II) representing the marine group II *Euryarchaeota* was reconstructed via de novo assembly of 11 mate-pair connected scaffolds with highly correlated sequence statistics (Fig. 1 and fig. S13) from a population composing ~7.5% of the May sample (fig. S9). A small hypervariable region (HVR) was insufficiently covered for complete automated assembly. Two variants with coverage of ~20× were partially reconstructed, and one of these (~14 kbp) was finished by using Sanger sequencing to fill remaining gaps. The resulting circular chromosome of 2.06 megabases recruited ~5 million aligned reads (1.7% of the May metagenome), yielding 118-fold read coverage and 2194-fold physical coverage by mate-pair inserts (figs. S14 and S15). The proportion of coverage of the two HVR variants to the genome average and analysis of 16*S* rDNA sequences (fig. S16) each predict that five or more strains of marine group II euryarchaeota were present in the May sample. Our assembly therefore represents a "majority rules" consensus, with possible genomic rearrangements at scaffold boundaries in some strains.

The mesophilic marine group II euryarchaeota have a cosmopolitan distribution and are relatively abundant during summer months (*18*, *19*). The metabolism of this uncultured group has been mysterious, with only a handful of sequenced environmental clones containing definitive phylogenetic markers (16*S* and 23*S* rDNA) available to infer its biogeochemical role (fig. S16) (*20*–*23*). To evaluate the correspondence of the MG-II genome to known environmental sequences and identify additional sequences not previously ascribed to this group, we performed six-frame translated alignments of GenBank environmental sequences to the MG-II genome. Sixty-seven environmental clones and 2894 metagenomic assemblies were identified that covered 36 and 96%, respectively, of predicted MG-II protein coding sequences (Fig. 2A). Clones derived from globally distributed marine samples revealed significant conservation

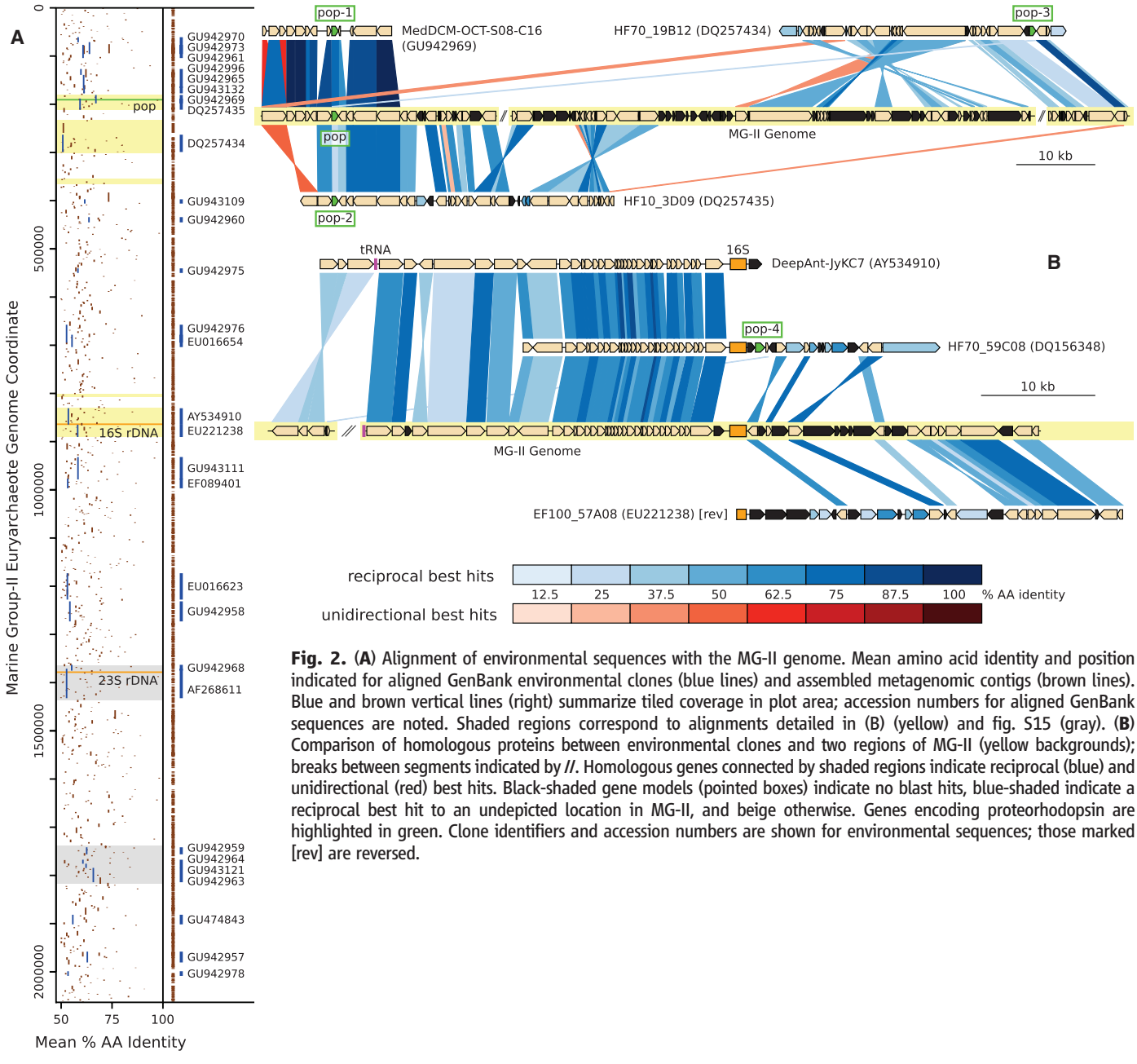

Fig. 2. (**A**) Alignment of environmental sequences with the MG-II genome. Mean amino acid identity and position indicated for aligned GenBank environmental clones (blue lines) and assembled metagenomic contigs (brown lines). Blue and brown vertical lines (right) summarize tiled coverage in plot area; accession numbers for aligned GenBank sequences are noted. Shaded regions correspond to alignments detailed in (B) (yellow) and fig. S15 (gray). (**B**) Comparison of homologous proteins between environmental clones and two regions of MG-II (yellow backgrounds); breaks between segments indicated by //. Homologous genes connected by shaded regions indicate reciprocal (blue) and unidirectional (red) best hits. Black-shaded gene models (pointed boxes) indicate no blast hits, blue-shaded indicate a reciprocal best hit to an undepicted location in MG-II, and beige otherwise. Genes encoding proteorhodopsin are highlighted in green. Clone identifiers and accession numbers are shown for environmental sequences; those marked [rev] are reversed.

of gene order with the MG-II genome assembly (Fig. 2B and fig. S17), reaffirming the group's cosmopolitan distribution.

Of particular interest is the single-copy proteorhodopsin (*pop*) gene present in the MG-II genome, which encodes key active-site residues shared with the light-driven proton-pumping rhodopsin found in *Exiguobacterium sibiricum* (*24*). Environmental marine group II euryarchaeal *pop* genes appear in at least three MG-II genetic contexts: the position originally detected by Frigaard *et al*. (*22*) (pop-4, Fig. 2B); that of the MG-II *pop* gene (pop, Fig. 2B), confirmed by two environmental clones (pop-1 and pop-2, Fig. 2B); and that found in an environmental clone aligning with a third region of the MG-II genome (pop-3, Fig. 2B). These five *pop* genes cluster phylogenetically into two distinct clades that also contain 42 *pop* genes identified in metagenomic assemblies aligning with positions in the MG-II genome. Clade A shares a common ancestor with proteorhodopsins from *Proteobacteria* (*22*), whereas clade B is diverged from all known bacterial proteorhodopsins (*25*) (Fig. 3 and fig. S18). The only synteny shared by *pop* genes from both clades ($n = 22$; in red, Fig. 3 and fig. S18) is the genetic context seen in the MG-II genome (pop, pop-1, and pop-2; Fig. 2B), indicating that the most parsimonious origin of clade A is a euryarchaeal ancestor shared with clade B. Thus, the proteorhodopsin of marine *Proteobacteria* appears to have originated in marine group II *Euryarchaeota*.

The MG-II genome sequence reveals 1781 predicted proteins, single copies of each of the four ribosomal RNA genes, and a typical number of tRNAs and noncoding RNAs (table S4). The MG-II 16*S* rDNA gene clusters with the group II.a clade of marine group II *Euryarchaeota* (*26*) (fig. S16). An analysis using 31 conserved archaeal proteins (table S5) places the marine group II *Euryarchaeota* in the most deeply rooted position of a clade containing primarily thermoacidophiles, including *Aciduliprofundum boonei* (*27*) and members of order *Thermoplasmata* (fig. S19A). The MG-II genome shares the same complement of ribosomal proteins found in *A. boonei* with complete conservation of gene order within multigene clusters (table S6), indicating *A. boonei* is the most closely related sequenced organism to marine group II euryarchaeota.

Core MG-II archaeal metabolic genes include those encoding glycolysis and gluconeogenesis pathways, a complete tricarboxylic acid cycle, and oxidative phosphorylation complexes. Genes dedicated to performing autotrophy or methanogenesis using known pathways are absent. MG-II encodes proteins potentially requiring and transporting cobalamin (vitamin B_{12}) and biotin (vitamin B_7) but does not appear to possess pathways to synthesize these compounds de novo. Additionally it does not appear to possess enzymes to reduce nitrate or sulfate, implying a dependence on reduced forms of N and S. An operon containing genes for archaeal flagellar proteins similar in gene order to those identified in *A. boonei* (*27*) strongly suggests the capacity for motility. We interpret these observations to indicate a motile heterotrophic lifestyle.

Among the MG-II genes are many large putative peptidases (table S7), which are overrepresented in the genome relative to those in the known protein degrader *A. boonei* (*27*) (5.7 versus 4.8% of coding sequence; table S4). This adds considerable new evidence to previous speculation (*20*, *21*) that consumption of protein may be an important metabolic activity shared in common with the peripherally related *Thermoplasmata* (*28*). Unexpectedly, we identified enzymes most similar to bacterial proteins that form a complete fatty acid degradation pathway (table S8), suggesting that marine group II euryarchaeota may catabolize straight chain lipids. The degradation of protein and lipid suggests that particles may be an important growth substrate, and MG-II codes for proteins with a variety of adhesion domains, as well as type II/IV secretion systems to transport such proteins to the cell surface.

We identified enzymes necessary to synthesize precursors for archaeal ether-linked lipids and also identified several putative acyl-carrier-protein fatty-acid synthesis enzymes and enzyme homologs for glycerolipid biosynthesis, most similar to bacterial proteins (table S8) and without homologs among the sequenced Archaea. The exclusive use of ether-linked isoprenoid lipids is considered a signature trait of nearly all Archaea (*29*), so the possible production and function of Bacteria-like ester-linked lipids requires further examination. Overall, MG-II contains 332 genes (~18.6%) with nearest homologs in the genomes of Bacteria, higher than the 119 genes (~7.7%) similarly classified in *A. boonei* (figs. S19 and S20, tables S9 and S10, and SOM text). These genes (including those for ester-linked lipid metabolism) are distributed across all of the well-supported scaffolds in the MG-II assembly (fig. S15), suggesting that marine group II euryarchaeota share many genetic adaptations with Bacteria.

Marine group II euryarchaeota are likely motile photo-heterotrophs focused on protein and lipid degradation. Proteorhodopsin, which appears to be a euryarchaeal innovation, may provide beneficial supplemental energy for propulsion to cells traversing relatively large distances in search of food particles (*30*). De novo assembly and analysis of the MG-II genome from high-coverage mate-paired metagenomic sequence demonstrates the power of this approach to provide insights into the roles uncultured microbes play in a changing global environment.

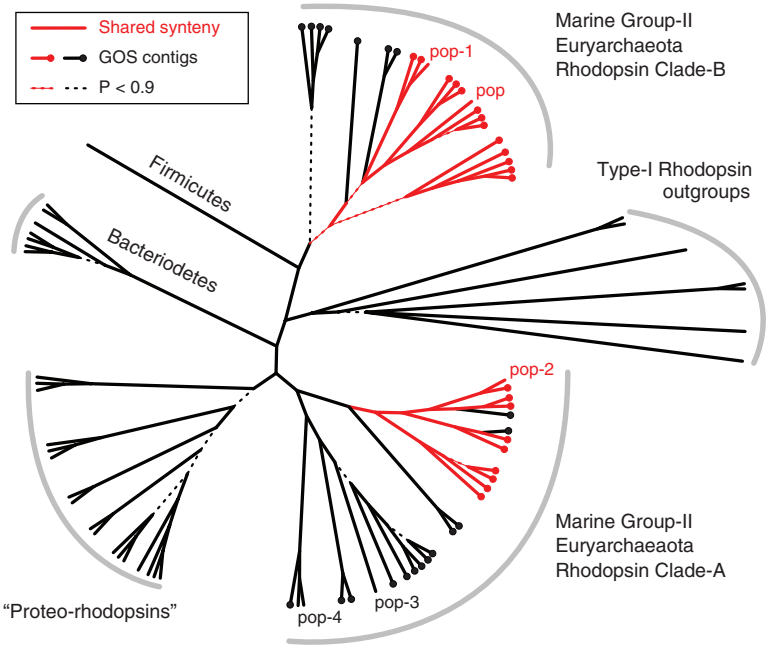

Fig. 3. Unrooted cladogram depicting Bayesian phylogeny of selected rhodopsin proteins. Major clades are labeled taxonomically, with "Proteo-rhodopsins" denoting those found in marine *Proteobacteria* (genomes and environmental clones). Out groups include bacterial and archaeal "bacteriorhodopsins," archaeal halorhodopsins, and sensory rhodopsins. The MG-II rhodopsin (pop), and four proteins from marine group II *Euryarchaeota* environmental clones (pop-1 to pop-4) are identified.

References and Notes

1. D. M. Karl, *Nat. Rev. Microbiol.* **5**, 759 (2007).
2. F. Azam, F. Malfatti, *Nat. Rev. Microbiol.* **5**, 782 (2007).
3. J. T. Staley, A. Konopka, *Annu. Rev. Microbiol.* **39**, 321 (1985).
4. M. S. Rappé, S. J. Giovannoni, *Annu. Rev. Microbiol.* **57**, 369 (2003).
5. P. G. Falkowski, T. Fenchel, E. F. Delong, *Science* **320**, 1034 (2008).
6. R. Stepanauskas, M. E. Sieracki, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 9052 (2007).
7. O. Béjà, *Curr. Opin. Biotechnol.* **15**, 187 (2004).
8. J. C. Venter *et al*., *Science* **304**, 66 (2004); 10.1126/science.1093857.

9. E. F. DeLong *et al.*, *Science* **311**, 496 (2006).
10. D. B. Rusch *et al.*, *PLoS Biol.* **5**, e77 (2007).
11. S. Yooseph *et al.*, *PLoS Biol.* **5**, e16 (2007).
12. D. A. Walsh *et al.*, *Science* **326**, 578 (2009).
13. D. B. Rusch, A. C. Martiny, C. L. Dupont, A. L. Halpern, J. C. Venter, *Proc. Natl. Acad. Sci. U.S.A.* **107**, 16184 (2010).
14. E. R. Mardis, *Trends Genet.* **24**, 133 (2008).
15. M. Hess *et al.*, *Science* **331**, 463 (2011).
16. Materials and methods are available as supporting material on *Science* Online.
17. K. D. Pruitt, T. Tatusova, D. R. Maglott, *Nucleic Acids Res.* **35**, D61 (2007).
18. A. Pernthaler, C. M. Preston, J. Pernthaler, E. F. DeLong, R. Amann, *Appl. Environ. Microbiol.* **68**, 661 (2002).
19. L. Herfort *et al.*, *FEMS Microbiol. Ecol.* **62**, 242 (2007).
20. O. Béjà *et al.*, *Environ. Microbiol.* **2**, 516 (2000).
21. D. Moreira, F. Rodríguez-Valera, P. López-García, *Environ. Microbiol.* **6**, 959 (2004).
22. N.-U. Frigaard, A. Martinez, T. J. Mincer, E. F. DeLong, *Nature* **439**, 847 (2006).
23. A.-B. Martin-Cuadrado *et al.*, *ISME J.* **2**, 865 (2008).
24. L. E. Petrovskaya *et al.*, *FEBS Lett.* **584**, 4193 (2010).
25. R. Ghai *et al.*, *ISME J.* **4**, 1154 (2010).
26. P. E. Galand, C. Gutiérrez-Provecho, R. Massana, J. M. Gasol, E. O. Casamayor, *Limnol. Oceanogr.* **55**, 2117 (2010).
27. A.-L. Reysenbach, G. E. Flores, *Geobiology* **6**, 331 (2008).
28. A. Ruepp *et al.*, *Nature* **407**, 508 (2000).
29. Y. Koga, H. Morii, *Microbiol. Mol. Biol. Rev.* **71**, 97 (2007).
30. J. M. Walter, D. Greenfield, C. Bustamante, J. Liphardt, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 2408 (2007).

Acknowledgments: We thank D. Kingsbury and T. Hunkapiller for helpful conversations about experimental design; W. Martens-Habbena, D. Stahl, and J. Baross for reviewing draft protein annotations; Q. Wang for providing the RDP classifier training set data; D. Schruth for technical support; and A. Ingalls and J. Baross for valuable comments on the draft manuscript. We are grateful to Life Technologies, Incorporated, for SOLiD sequencing of the October 2008 sample. This study was supported by a Gordon and Betty Moore Foundation Marine Microbiology Investigator Award and NSF grant OCE-0723866. Sequences reported in this study have been deposited in the National Center for Biotechnology Information (NCBI) Sequence Read Archive under accession no. SRA047968 and in GenBank under accession nos. JF965490 to JF965492 and JN591771 to JN592031. The Whole Genome Shotgun project has been deposited at DNA Data Bank of Japan/European Molecular Biology Laboratory/GenBank under the accession AHCG00000000. The version described in this paper is the first version, AHCG01000000. Information regarding the custom software developed for use in this research is available at http://armbrustlab.ocean.washington.edu/software.

Supporting Online Material

www.sciencemag.org/cgi/content/full/335/6068/587/DC1
Materials and Methods
SOM Text
Figs. S1 to S20
Tables S1 to S10
References (*31*–*65*)

15 August 2011; accepted 14 December 2011
10.1126/science.1212665

Sequential Signaling Crosstalk Regulates Endomesoderm Segregation in Sea Urchin Embryos

Aditya J. Sethi,[1] Radhika M. Wikramanayake,[2] Robert C. Angerer,[1] Ryan C. Range,[1] Lynne M. Angerer[1]*

The segregation of embryonic endomesoderm into separate endoderm and mesoderm fates is not well understood in deuterostomes. Using sea urchin embryos, we showed that Notch signaling initiates segregation of the endomesoderm precursor field by inhibiting expression of a key endoderm transcription factor in presumptive mesoderm. The regulatory circuit activated by this transcription factor subsequently maintains transcription of a canonical Wnt (cWnt) ligand only in endoderm precursors. This cWnt ligand reinforces the endoderm state, amplifying the distinction between emerging endoderm and mesoderm. Before gastrulation, Notch-dependent nuclear export of an essential β-catenin transcriptional coactivator from mesoderm renders it refractory to cWnt signals, insulating it against an endoderm fate. Thus, we report that endomesoderm segregation is a progressive process, requiring a succession of regulatory interactions between cWnt and Notch signaling.

Early endomesoderm induction and subsequent segregation of endoderm from mesoderm are fundamental processes in animal development. Although initial endomesoderm specification has been studied extensively (*1*, *2*), its separation in deuterostomes is poorly understood. In several deuterostomes including vertebrates such as zebrafish and *Xenopus* and echinoderms such as sea urchins and sea stars (*3*–*7*), Notch signaling induces the expression of endoderm- or mesoderm-specific markers within an endomesoderm field. Although Notch might regulate endomesoderm segregation, it is unknown whether it alters the early endomesoderm signaling milieu, a change that is probably required to stabilize lineage identities. Canonical Wnt (cWnt) signaling probably establishes that precursor environment, given its ancestral role in specifying early endomesoderm (*8*). Elucidating the mechanisms underlying such a Notch-cWnt interaction would substantially advance our understanding of the progressive specialization of the endomesoderm.

Like other deuterostomes, sea urchin embryos are enriched asymmetrically in the cWnt signaling effector nuclear β-catenin (nβ-catenin) (*9*), which specifies endomesoderm precursors in three ways. First, it establishes an early endoderm regulatory state in a tier of vegetal blastomeres (veg_2, Fig. 1A) at cleavage stages (*9*–*11*). Second, in micromere descendants located immediately adjacent to the veg_2 tier (Fig. 1A), nβ-catenin induces expression of the ligand Delta (*12*–*14*), which signals through the Notch receptor in veg_2 blastomeres and activates mesoderm gene expression (*5*, *6*, *13*, *15*). Third, cWnt also makes veg_2 cells competent to receive the micromere Delta signal (*12*). Thus, specification of both endoderm and mesoderm is initiated by the blastula stage throughout veg_2 blastomeres (Fig. 1A) (*6*, *9*, *11*, *16*, *17*).

By the hatching blastula (HB) stage, veg_2 progeny form outer and inner rings of cells (Fig. 1A). Only inner veg_2 daughters adjacent to Delta-expressing micromere progeny can transduce the cell contact–dependent Notch signal and continue expressing mesoderm markers (*17*). Transcripts encoded by endoderm regulatory genes, in turn, are detected in outer veg_2 daughters by this time (*11*, *17*). Notch is required for this restriction because it inhibits expression of the endoderm markers *foxa*, *blimp1b*, and *dac* in inner veg_2 daughters (*17*–*19*). During cleavage and HB stages, nβ-catenin is detected throughout the veg_2 endomesoderm precursor field (*9*). By the mesenchyme blastula (MB) stage, 6 to 8 hours after endoderm marker expression first clears from inner veg_2 daughters, nβ-catenin is down-regulated in these mesoderm precursors through a process requiring Notch (*9*, *20*, *21*). Thus, Notch probably plays a substantial role in endomesoderm segregation beyond merely activating mesoderm regulatory genes. However, several major questions remain unanswered. First, how does Notch restrict endoderm fate to a subset of the endomesoderm progenitor field? Second, how does Notch inhibit endomesoderm-inducing cWnt in presumptive mesoderm (*9*, *21*) 6 to 8 hours after initial endoderm marker expression has disappeared? Third, are these Notch-dependent events mechanistically linked?

To understand initial Notch-dependent restriction of endoderm potential from mesoderm, we used Notch-deficient embryos to systematically assess the expression of each gene in the early endoderm gene regulatory network (GRN) (fig. S1A), which represents cWnt-induced endoderm specification until the HB stage (*10*, *11*). We found that *hox11/13b*, *brachyury*, *blimp1b*, and *foxA* transcripts accumulated ectopically in

[1]National Institute of Dental and Craniofacial Research, National Institutes of Health, Bethesda, MD, USA. [2]Department of Public Health, University of Oxford, Oxford, UK.

*To whom correspondence should be addressed. E-mail: langerer@mail.nih.gov