The ISME Journal (2009) 3, 1420–1424

www.nature.com/ismej

SHORT COMMUNICATION

Abundant transposases encoded by the metagenome of a hydrothermal chimney biofilm

William J Brazelton and John A Baross
School of Oceanography and Center for Astrobiology and Early Evolution, University of Washington, Seattle, WA, USA

The carbonate chimneys of the Lost City Hydrothermal Field on the Mid-Atlantic Ridge are coated in thick microbial biofilms consisting of just a few dominant species. We report a preliminary analysis of a biofilm metagenome that revealed a remarkable abundance and diversity of genes potentially involved in lateral gene transfer (LGT). More than 8% of all metagenomic reads showed significant sequence similarity to transposases; all available metagenomic data sets from other environments contained at least an order of magnitude fewer transposases. Furthermore, the sequence diversity of transposase genes in the biofilm was much greater than that of 16S rRNA genes. The small size and high sequencing coverage of contigs containing transposases indicate that they are located on small but abundant extragenomic molecules. These results suggest that rampant LGT among members of the Lost City biofilm may serve as a generator of phenotypic diversity in a community with very low organismal diversity.
The ISME Journal (2009) **3,** 1420–1424; doi:10.1038/ismej.2009.79; published online 2 July 2009
Subject Category: microbial ecology and functional diversity of natural habitats
Keywords: lateral gene transfer; biofilm; transposase

In most natural environments, microorganisms are predominantly found in surface-associated, matrix-enclosed communities known as biofilms (Costerton *et al.*, 1995). At the Lost City Hydrothermal Field on the Mid-Atlantic Ridge, biofilms coat mineral surfaces of the highly porous carbonate chimneys venting $<90\,^{\circ}C$, pH 9–11 fluids (Schrenk *et al.*, 2004). Chimney fluids contain abundant hydrogen and methane, but very little carbon dioxide due to the high pH (Kelley *et al.*, 2005). Owing to extreme conditions, very little animal biomass is present at Lost City; instead, the thick, mucilaginous biofilms (containing up to 10^9 cells per gram of carbonate chimney) are the dominant life forms. Previous studies have highlighted the extremely low microbial diversity in carbonate chimneys. Although a single phylotype belonging to the *Methanosarcinales* order of methane-cycling archaea constitutes $>80\%$ of all active cells in the hottest, anoxic zones of the chimney (Schrenk *et al.*, 2004), a few species of aerobic and microaerophilic bacteria dominate the cooler, oxygenated zones (Brazelton *et al.*, 2006). As the continuous mixing of anoxic hydrothermal fluid with oxygenated seawater creates micro-scale redox gradients within the chimneys, these anaerobic archaea and aerobic bacteria live in close proximity to each other.

The low diversity and high cell density of biofilms make them attractive targets for metagenomic sequencing. We obtained 35 Mb of DNA sequence from 46 361 shotgun reads of two pUC18 libraries constructed by the DOE Joint Genome Institute (Walnut Creek, CA, USA) with DNA extracted from $\sim$1 kg of a single carbonate chimney sample. Initial characterization of the metagenome with BLAST (Basic Local Alignment Search Tool) (Altschul *et al.*, 1997) searches against the GenBank non-redundant database revealed a large number of hits to genes encoding transposases, enzymes involved in the transposition and integration of mobile genetic elements (that is, DNA that can be transferred within or between genomes). To test whether the apparently high abundance of transposases in the Lost City biofilm metagenome is unusual, we developed a simple method of quantifying numbers of transposase sequences in metagenomic datasets (Table 1). We only compared sets of unassembled reads, not assembled contigs, because it is not straightforward to make quantitative comparisons

Correspondence: W Brazelton, School of Oceanography, Center for Astrobiology and Early Evolution, University of Washington, Box 357940, Seattle, WA 98195, USA.
E-mail: braz@u.washington.edu
Received 16 March 2009; revised 27 May 2009; accepted 4 June 2009; published online 2 July 2009

Table 1 Abundance of transposases in the Lost City chimney biofilm metagenome compared with other metagenomes in the CAMERA database

Metagenome	*Reference*	*Size (Mb)*	*Reads*	*Transposases*[a]	*% Reads*
Lost City chimney	This study	35	46 361	3735	8.06%
Bioreactor sludge	Garcia Martin *et al.*, (2006)	221	224 516	1855	0.83%
Acid mine drainage	Tyson *et al.*, (2004); Lo *et al.*, (2007)	326	19 166	2281	0.71%
Gutless worm consortium	Woyke *et al.*, (2006)	315	313 773	2226	0.71%
Deep Mediterranean plankton	Martin-Cuadrado *et al.*, (2007)	7	9048	59	0.65%
Farm Soil	Tringe *et al.*, (2005)	154	138 347	765	0.55%
Saltern crystallizer	Legault *et al.*, (2006)	2	2947	16	0.54%
Whale Fall	Tringe *et al.*, (2005)	118	117 326	518	0.44%
Ocean ALOHA station	DeLong *et al.*, (2006)	64	65 675	258	0.39%
Hot springs virome	Schoenfeld *et al.*, (2008)	31	30 624	79	0.26%
Alvinella epibiont	CAMERA database	290	293 065	684	0.23%
Global ocean Sampling	Rusch *et al.*, (2007)	10 635	10 133 846	5352	0.05%
Chesapeake Bay virioplankton	Bench *et al.*, (2007)	4	5641	2	0.04%
Marine viromes	Angly *et al.*, (2006)	179	1 768 297	430	0.02%
Sargasso Sea bacterioplankton	CAMERA database	127	606 285	35	0.01%

[a]Number of reads with a TBLASTN hit of 10^{-5} or better to one of 852 protein sequences representing 29 transposase protein families in the PFAM database (Finn *et al.*, 2008). By making use of the PFAM seed sequences as the query, this approach is more exhaustive and less error prone than simply counting keywords in BLAST (Basic Local Alignment Search Tool) results, which relies on the accuracy of the BLAST hits' annotations.

among assembled metagenomes containing contigs and scaffolds of various numbers, sizes and sequence coverages.

Our results show that the Lost City biofilm contains an unprecedented abundance of transposases. Over 8% of all reads in the metagenome matched one of the transposase protein families with an E value of 10^{-5} or better (hereafter referred to as significant hits). Similar results were achieved with an E value cutoff of 10^{-10}, but control searches revealed that this cutoff resulted in some false negatives. Very few reads contained significant hits to more than one transposase family with the 10^{-5} cutoff, indicating that our searches were family specific and unlikely to yield non-transposase sequences.

We conducted the same search against all collections of unassembled metagenomic reads in the CAMERA database (Seshadri *et al.*, 2007). Each metagenome contained at least an order of magnitude fewer transposases per read than did the Lost City metagenome (Table 1). Interestingly, the four metagenomes in Table 1 with the highest proportion of reads containing transposases are those derived from biofilms. In contrast, the four metagenomes in Table 1 with the fewest transposases ($200\times - 800\times$ fewer per read than the Lost City metagenome) are from water samples with little or no biomass contribution from biofilms. Furthermore, the three viral metagenomes were among those with the fewest transposases, suggesting that the abundance of transposases in the biofilm metagenomes is not easily explained by the presence of viruses.

A further analysis revealed that the transposases in the Lost City biofilm's metagenome are very diverse as well as abundant. We detected 21 different transposase families with significant hits to a Lost City metagenomic read (Table 2). Two of the families (retroviral integrase and transposase 11) were present in 2353 Lost City reads, comprising 63% of all significant hits. To examine whether the 2353 reads represent just a few genes present in many copies or a large diversity of genes, we constructed multiple sequence alignments with the MUSCLE aligner (Edgar, 2004), including the nucleotide sequence region of each read with a significant TBLASTN alignment with a retroviral integrase or transposase 11 sequence. The Lost City sequences were clustered into operational taxonomic units on the basis of a 3% sequence difference threshold using DOTUR (Schloss & Handelsman, 2005). The results (shown in Figure 1) show that the Lost City retroviral integrase and transposase 11 sequences each include >100 operational taxonomic units. (Equivalent results were achieved by aligning amino acid sequences.) A similar analysis of 16S rRNA sequences in the Lost City metagenome yielded just 83 reads representing 22 operational taxonomic units (Figure 1). Therefore, genes encoding transposases are much more abundant and diverse than are 16S rRNA genes in the Lost City biofilm.

A preliminary assembly of the Lost City metagenome (assembled by the DOE Joint Genome Institute) is consistent with the carriers of the transposases being small molecules of extragenomic DNA. Approximately half of all 41 393 reads assembled into 6324 contigs of 2 or more reads, including 49 contigs >7 kb in length (Figure 2). The largest contigs had a high sequence similarity to the 16S rRNA gene and to many of the open reading frames in the genome of *Thiomicrospira crunogena* XCL-2 (Scott *et al.*, 2006). (Nearly half of all bacterial 16S rRNA clones sequenced from this sample showed a similarity to species belonging to the *Thiomicrospira* genus; see Supplementary information for more detail.) The large contigs had a similar %GC and sequencing coverage ($\sim$38% GC and

Table 2 Abundance of each transposase protein family in the Lost City metagenome

PFAM	*Description*	*Insertion sequence/transposon families*	*Additional information*	*Lost City reads*[a]	*% Lost City reads*
PF00665	Retroviral integrase	Retroviral genomes and bacterial elements (*eg.* IS30)	RNase H clan	1224	2.640
PF01609	Transposase 11	IS4, IS421, IS5377, IS427, IS402, IS1355, IS5	RNase H clan	1129	2.435
PF00872	Transposase, mutator	IST2, IS256, IS1201, IS1081, ISRm3	RNase H clan	611	1.318
PF01610	Transposase 12	IS204, IS1001, IS1096, IS1165	RNase H clan	457	0.986
PF01527	Transposase 8	IS3	HTH clan	80	0.173
PF00589	Phage integrase		Found in phages, integrons, genomic islands, conjugative elements	64	0.138
PF02371	Transposase 20	IS116, IS110, IS902		38	0.082
PF01710	Transposase 14	IS5, IS4, IS630		31	0.067
PF01385	Transposase 2	IS891, IS1136, IS1341		18	0.039
PF01548	Transposase 9	IS111A, IS1328, IS1533		14	0.030
PF01526	Transposase 7	Tn3, Tn21, Tn1721, Tn2501, Tn3926		13	0.028
PF01797	Transposase 17	IS200		11	0.024
PF02899	Phage integrase N-terminal		See PF00589	10	0.022
PF05717	Transposase 34	IS66		7	0.015
PF07592	Transposase 36			6	0.013
PF03050	Transposase 25	IS66		6	0.013
PF04754	Transposase 31		PDDEXK clan	5	0.011
PF04986	Transposase 32	IS1294, IS801		4	0.009
PF03400	Transposase 27	IS1		3	0.006
PF02914	Mu transposase	Bacteriophage Mu	RNase H clan	2	0.004
PF02316	Mu DNA-binding	Bacteriophage Mu		2	0.004
	Total			3735	8.06

[a]Number of reads with a TBLASTN hit of 10^{-5} or better to a protein sequence belonging to the PFAM protein family indicated.

5–8 × coverage), indicating that they represent the genome of a *Thiomicrospira* species that previous studies have shown to be widespread in Lost City carbonate chimneys (Brazelton *et al.*, 2006). The 689 contigs containing transposases, by contrast, have a wide range in their %GC, and many have high sequence coverage and are less than 5 kb (Figure 2). These contrasting patterns indicate that most of the transposases do not belong to the same genome as do the *Thiomicrospira*-like genes. Instead, they are likely to be located on small, extragenomic molecules such as viral genomes, plasmids or extracellular DNA fragments.

To test these possibilities, similar TBLASTN searches to those described above for transposases were conducted with PFAM protein families representing plasmid and viral proteins. None of these searches resulted in many significant hits: only 13 reads contained plasmid replication proteins, 6 reads contained viral capsid proteins and 10 reads contained reverse transcriptases (data not shown). Therefore, if plasmids or viruses are carrying the abundant transposases found in the Lost City biofilm, they do not show significant sequence similarity with previously published plasmids and viruses. We also found little evidence of the presence of integrons and genomic islands; only 1.7% of all transposases in the Lost City biofilm (Table 2) belonged to the phage integrase family (which is associated with integrons and genomic islands), and few transposases were found on the same contig as that of tRNA genes (data not shown), frequent insertion sites for these elements. (See Supplementary information for further analyses.) It is also possible that transposase genes are present as multiple copies in regions of cellular genomes that are not conducive to assembly into large contigs, but this scenario cannot easily explain the high diversity of transposase sequences. In conclusion, the most likely carriers of transposases in Lost City biofilms are small extracellular fragments of DNA. Many biofilms are known to contain extracellular DNA, and in some cases, it has been shown that the presence of extracellular DNA is required for biofilm formation (Whitchurch *et al.*, 2002).

Although it is possible that many of the transposase genes detected in this study may not be expressed (Ram *et al*, 2005), their unprecedented abundance and diversity strongly suggest that lateral gene transfer (LGT) is a frequent occurrence within Lost City biofilms. Considering the low organismal diversity of these biofilms, LGT may be an important source of phenotypic diversity. Mathematical modeling suggests that LGT among genomes within a

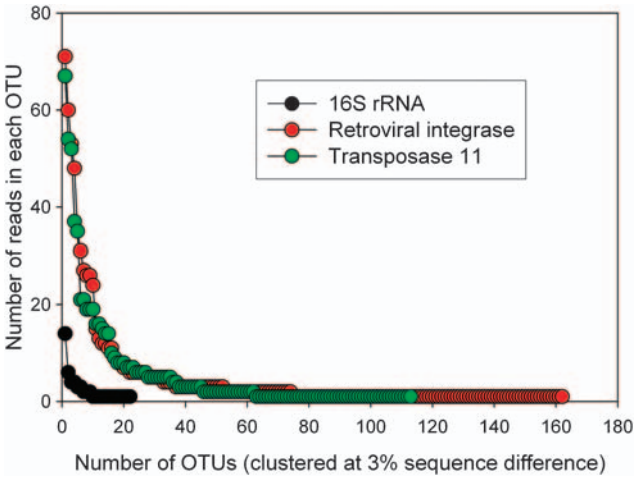

Figure 1 The diversity of transposase sequences in the Lost City metagenome is much greater than 16S rRNA diversity. Each operational taxonomic unit (OTU) cluster of sequences (defined as 3% nucleotide sequence difference) is listed on the X-axis, and the number of reads with sequences in each OTU are plotted on the Y-axis. The two transposase families, retroviral integrase (red points) and transposase 11 (green), each have >100 OTUs, whereas there are only 22 16S rRNA OTUs (black) representing a small number of reads.

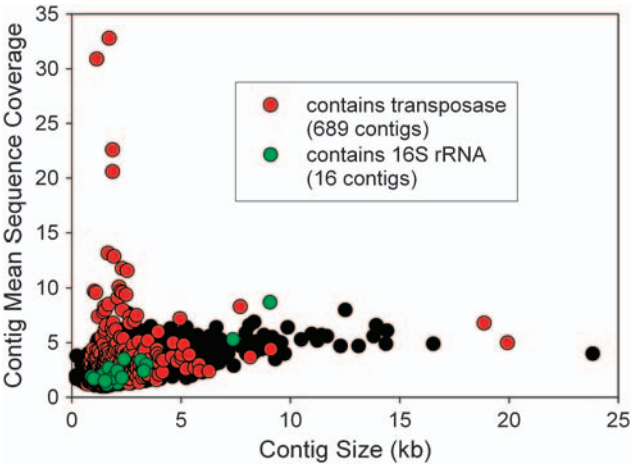

Figure 2 Size in kilobases and sequencing coverage of each contig (black points) in a preliminary assembly of the Lost City metagenome. The 689 contigs with transposases (red) are smaller and have higher coverage than do the 16 contigs with 16S rRNA sequences (green).

biofilm can stabilize the coexistence of multiple phenotypes and therefore contribute to the overall fitness of the biofilm community (Chia *et al.*, 2008). Biofilms in other systems are known to generate physiological diversity in response to environmental stresses and gradients, despite limited genetic diversity (Boles *et al.*, 2004; Stewart and Franklin, 2008), and many biofilm communities exhibit highly structured networks of interactions requiring interspecies communication and cooperation (Shapiro, 1998; Stoodley *et al.*, 2002; West *et al.*, 2006). Therefore, further work should investigate whether similar kinds of collective interactions are operating in the low-diversity biofilms of Lost City chimneys, where survival in extreme conditions may require the stable coexistence of multiple phenotypes enabled by LGT.

Lost City carbonate chimneys have been discussed as models for the origin and early evolution of life because of the prevalence of ultramafic environments on early Earth (Grove and Parman, 2004) and Mars (Hamilton and Christensen, 2005), because of the exothermic generation of hydrogen and organic compounds by serpentinization (Proskurowski *et al.*, 2008), because of the potential of chimney pores to concentrate biochemicals (Baaske *et al.*, 2007) and because of the advantage of a high pH for prebiotic chemistry (Martin *et al.*, 2008). In addition, it has been suggested that a community of primitive precells (Baross and Hoffman, 1985) or progenotes (Woese, 1998) undergoing extensive gene transfer represented an early stage of evolution before the advent of free-living cells. Therefore, the biofilms of Lost City carbonate chimneys could serve as a model for this theory, considering the potential for generating phenotypic diversity with limited organismal diversity through rampant LGT.

Accession numbers

This Whole Genome Shotgun project has been deposited at DDBJ/EMBL/Genbank under the project accession ACQI00000000. The version described in this paper is the first version, ACQI01000000. All sequencing reads are deposited under accession numbers ACQI01006325–ACQI01026573, and assembled contigs are deposited under accession numbers ACQI01000001–ACQI01006324.

Acknowledgements

We thank Deborah Kelley and Bob Ballard, chief scientists of the 2005 Lost City expedition funded by NOAA Ocean Exploration, as well as the captain and crew of the R/V *Ronald H Brown* and crew of the ROV *Hercules*. We also appreciate technical assistance from the DOE Joint Genome Institute and helpful discussions with Susannah Green Tringe. This work was supported by the NASA Astrobiology Institute through the Carnegie Institution for Science.

References

Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W *et al.* (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389–3402.

Angly FE, Felts B, Breitbart M, Salamon P, Edwards RA, Carlson C *et al.* (2006). The marine viromes of four oceanic regions. *PloS Biol* **4**: 2121–2131.

Baaske P, Weinert FM, Duhr S, Lemke KH, Russell MJ, Braun D. (2007). Extreme accumulation of nucleotides

in simulated hydrothermal pore systems. *Proc Natl Acad Sci USA* **104**: 9346–9351.

Baross JA, Hoffman SE. (1985). Submarine hydrothermal vents and associated gradient environments as sites for the origin and evolution of life. *Orig Life Evol Biosph* **15**: 327–345.

Bench SR, Hanson TE, Williamson KE, Ghosh D, Radosovich M, Wang K *et al.* (2007). Metagenomic characterization of Chesapeake Bay virioplankton. *Appl Environ Microbiol* **73**: 7629–7641.

Boles BR, Thoendel M, Singh PK. (2004). Self-generated diversity produces 'insurance effects' in biofilm communities. *Proc Natl Acad Sci USA* **101**: 16630–16635.

Brazelton WJ, Schrenk MO, Kelley DS, Baross JA. (2006). Methane- and sulfur-metabolizing microbial communities dominate the Lost City hydrothermal field ecosystem. *Appl Environ Microbiol* **72**: 6257–6270.

Chia N, Woese CR, Goldenfeld N. (2008). A collective mechanism for phase variation in biofilms. *Proc Natl Acad Sci USA* **105**: 14597–14602.

Costerton JW, Lewandowski Z, Caldwell DE, Korber DR, Lappin-Scott HM. (1995). Microbial biofilms. *Annu Rev Microbiol* **49**: 711–745.

DeLong EF, Preston CM, Mincer T, Rich V, Hallam SJ, Frigaard NU *et al.* (2006). Community genomics among stratified microbial assemblages in the ocean's interior. *Science* **311**: 496–503.

Edgar RC. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**: 1792–1797.

Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, Hotz H-R *et al.* (2008). The Pfam protein families database. *Nucleic Acids Res* **36**: D281–D288.

Garcia Martin H, Ivanova N, Kunin V, Warnecke F, Barry KW, McHardy AC *et al.* (2006). Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. *Nat Biotechnol* **24**: 1263–1269.

Grove TL, Parman SW. (2004). Thermal evolution of the Earth as recorded by komatiites. *Earth Planet Sci Lett* **219**: 173–187.

Hamilton VE, Christensen PR. (2005). Evidence for extensive, olivine-rich bedrock on Mars. *Geology* **33**: 433–436.

Kelley DS, Karson JA, Fruh-Green GL, Yoerger DR, Shank TM, Butterfield DA *et al.* (2005). A serpentinite-hosted ecosystem: the Lost City hydrothermal field. *Science* **307**: 1428–1434.

Legault BA, Lopez-Lopez A, Alba-Casado JC, Doolittle WF, Bolhuis H, Rodriguez-Valera F *et al.* (2006). Environmental genomics of *Haloquadratum walsbyi* in a saltern crystallizer indicates a large pool of accessory genes in an otherwise coherent species. *BMC Genomics* **7**: 171.

Lo I, Denef VJ, VerBerkmoes NC, Shah MB, Goltsman D, DiBartolo G *et al.* (2007). Strain-resolved community proteomics reveals recombining genomes of acidophilic bacteria. *Nature* **446**: 537–541.

Martin-Cuadrado AB, Lopez-Garcia P, Alba JC, Moreira D, Monticelli L, Strittmatter A *et al.* (2007). Metagenomics of the deep Mediterranean, a warm bathypelagic habitat. *PLoS ONE* **2**: e914.

Martin W, Baross J, Kelley D, Russell MJ. (2008). Hydrothermal vents and the origin of life. *Nat Rev Microbiol* **6**: 805–814.

Proskurowski G, Lilley MD, Seewald JS, Fruh-Green GL, Olson EJ, Lupton JE *et al.* (2008). Abiogenic hydrocarbon production at lost city hydrothermal field. *Science* **319**: 604–607.

Ram RJ, Verberkmoes NC, Thelen MP, Tyson GW, Baker BJ, Blake RC *et al.* (2005). Community proteomics of a natural microbial biofilm. *Science* **308**: 1915–1920.

Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S *et al.* (2007). The Sorcerer II global ocean sampling expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biol* **5**: 398–431.

Schloss PD, Handelsman J. (2005). Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Appl Environ Microbiol* **71**: 1501–1506.

Schoenfeld T, Patterson M, Richardson PM, Wommack KE, Young M. (2008). Assembly of viral metagenomes from Yellowstone hot springs. *Appl Environ Microbiol* **74**: 4164–4174.

Schrenk MO, Kelley DS, Bolton SA, Baross JA. (2004). Low archaeal diversity linked to subseafloor geochemical processes at the Lost City Hydrothermal Field, Mid-Atlantic Ridge. *Environ Microbiol* **6**: 1086–1095.

Scott KM, Sievert SM, Abril FN, Ball LA, Barrett CJ, Blake RA *et al.* (2006). The genome of deep-sea vent chemolithoautotroph *Thiomicrospira crunogena* XCL-2. *PLoS Biol* **4**: e383.

Seshadri R, Kravitz SA, Smarr L, Gilna P, Frazier M. (2007). CAMERA: a community resource for metagenomics. *PLoS Biol* **5**: e75.

Shapiro JA. (1998). Thinking about bacterial populations as multicellular organisms. *Annu Rev Microbiol* **52**: 81–104.

Stewart PS, Franklin MJ. (2008). Physiological heterogeneity in biofilms. *Nat Rev Microbiol* **6**: 199–210.

Stoodley P, Sauer K, Davies DG, Costerton JW. (2002). Biofilms as complex differentiated communities. *Annu Rev Microbiol* **56**: 187–209.

Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW *et al.* (2005). Comparative metagenomics of microbial communities. *Science* **308**: 554–557.

Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM *et al.* (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**: 37–43.

West SA, Griffin AS, Gardner A, Diggle SP. (2006). Social evolution theory for microorganisms. *Nat Rev Microbiol* **4**: 597–607.

Whitchurch CB, Tolker-Nielsen T, Ragas PC, Mattick JS. (2002). Extracellular DNA required for bacterial biofilm formation. *Science* **295**: 1487.

Woese CR. (1998). The universal ancestor. *Proc Natl Acad Sci USA* **95**: 6854–6859.

Woyke T, Teeling H, Ivanova NN, Huntemann M, Richter M, Gloeckner FO *et al.* (2006). Symbiosis insights through metagenomic analysis of a microbial consortium. *Nature* **443**: 950–955.

Supplementary Information accompanies the paper on The ISME Journal website (http://www.nature.com/ismej)