DNA/DNA hybridization to microarrays reveals gene-specific differences between closely related microbial genomes

A. E. Murray*†, D. Lies‡, G. Li§, K. Nealson‡, J. Zhou§, and J. M. Tiedje¶

*Earth and Ecosystem Sciences, Desert Research Institute, Reno, NV 89512; ‡Department of Geology and Planetary Sciences, Jet Propulsion Laboratory and California Institute of Technology, Pasadena, CA 91109; §Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831; and ¶Center for Microbial Ecology, Michigan State University, East Lansing, MI 48824

DNA microarrays constructed with full length ORFs from *Shewanella oneidensis*, MR-1, were hybridized with genomic DNA from nine other *Shewanella* species and *Escherichia coli* K-12. This approach enabled visualization of relationships between organisms by comparing individual ORF hybridizations to 164 genes and is further amenable to high-density high-throughput analyses of complete microbial genomes. Conserved genes (*arc*A and ATP synthase) were identified among all species investigated. The *mtr* operon, which is involved in iron reduction, was poorly conserved among other known metal-reducing *Shewanella* species. Results were most informative for closely related organisms with small subunit rRNA sequence similarities greater than 93% and *gyr*B sequence similarities greater than 80%. At this level of relatedness, the similarity between hybridization profiles was strongly correlated with sequence divergence in the *gyr*B gene. Results revealed that two strains of *S. oneidensis* (MR-1 and DLM7) were nearly identical, with only 3% of the ORFs hybridizing poorly, in contrast to hybridizations with *Shewanella putrefaciens*, formerly considered to be the same species as MR-1, in which 63% of the ORFs hybridized poorly (log ratios below −0.75). Genomic hybridizations showed that genes in operons had consistent levels of hybridization across an operon in comparison to a randomly sampled data set, suggesting that similar applications will be informative for identification of horizontally acquired genes. The full value of microbial genomic hybridizations lies in providing the ability to understand and display specific differences between closely related organisms providing a window into understanding microheterogeneity, bacterial speciation, and taxonomic relationships.

The influx of microbial genome sequence data and analysis thereof are providing information that will help define genotypic, and the resulting phenotypic, differences between microorganisms, thus influencing the concept of the bacterial species and our understanding of the underlying processes affecting speciation and evolution. Currently, the bacterial species definition relies on a coordinated evaluation of DNA/DNA hybridization (1), small subunit (SSU) phylogeny, and phenotypic data (2). Microbial genome sequences provide information at an unprecedented level of detail such that an understanding of the differences between microorganisms at a genomic-wide scale is possible. Patterns of sequence similarity and variability indicate conservation of function, physiological plasticity, and evolutionary processes.

Microbial comparative genomics is a rapidly emerging field that is beginning to reveal highly ordered information concerning the differences between organisms at the population level. Genome sequences for more than one strain per species are accumulating for several groups of pathogens (*Chlamydia pneumoniae*, *Chlamydia trachomatis*, *Escherichia coli*, *Helicobacter pylori*, and *Mycobacterium tuberculosis*). Some nonpathogenic sequences such as *Prochlorococcus marinus* will soon follow. One of the emerging views from comparative genomics studies is the

Selfish Operon Model (3), which suggests that gene clusters involved in a single metabolic function provide a fitness benefit enabling them to be readily transferred. The implications of this theory are that lateral gene transfer promotes bacterial speciation by enriching the genome with new metabolic capabilities and hence the ability to survive in new ecological niches (4, 5). Detailed studies of the *E. coli* genome suggest that phenotypic differences between *E. coli* and *Salmonella enterica* are because of horizontally acquired operons rather than cumulative point mutations (4). This theory will be tested as more genome sequence information accrues and the field of comparative genomics develops. In addition to providing a hypothesis regarding gene transfer, the Selfish Operon Model can also be interpreted to suggest that genes in operons might have similar levels of DNA relatedness between species across the operon, a hypothesis tested in the present study.

To date, in the absence of full genome sequence information, it is still difficult to understand the differences between closely related microorganisms. To address this issue, we have explored the potential for using DNA microarray technology to view genome diversity and relatedness from a comparative genomics perspective. This approach circumvents the need for sequencing multiple closely related genomes, while using data from those sequences available. Although most studies using DNA microarrays have focused on comparative gene expression, several have been directed at population level differences (6–10). In this study, we have used genome sequence information from *Shewanella oneidensis* strain MR-1 to determine the relatedness between other bacterial strains in the *Shewanella* genus of the γ-proteobacteria and MR-1. This genus is characterized by a broad diversity of microorganisms isolated from a wide range of habitats, including marine and freshwater sediments, the deep sea, and oil brines. Many species in this genus are capable of dissimilatory metal reduction, nitrate, and sulfur reduction, can be opportunistic pathogens, and are associated with fish spoilage. The phylogeny of the *Shewanella* (recently revised; ref. 11) supports two distinct clusters that group according to ecological niche. One cluster has representatives characterized as halotolerant, mostly mesophilic, and metabolically versatile, whereas the other cluster has halophilic, piezophilic, and mostly psychrophilic species, which are often more fastidious.

We have conducted a survey of nine *Shewanella* species and assessed their genomic relatedness to *S. oneidensis*, strain MR-1, to understand better what the differences are between these

MICROBIOLOGY

Table 1. Strains used in DNA:DNA microarray hybridization analysis

Bacterium	Culture collection	Strain origin	SSU rRNA accession no.	gyrB accession no.
S. oneidensis MR-1	ATCC 700550 TS	Lake Oneida, NY, sediments	AF005251	AF005694
S. oneidensis DLM7		Lower Green Bay sediments, Lake Michigan, Michigan	AF387347*	AF005697
S. oneidenisis MR-4		Black Sea water column	AF005252	AF005695
Shewanella sp. CL 256/73	NCTC 12093	Human cerebrospinal fluid	AF387346*	AF378351*
S. putrefaciens str. 95	ATCC 8071 TS	Butter	U91550	AF005669
S. baltica str. 63	NCTC 10735, TS	Oil brine, Japan	AJ000214	AF387353*
Shewanella sp. str. 184	ATCC 8073	Butter	AF387349*	AF387354*
S. woodyi MS32	ATCC 51908, TS	Persian Gulf seawater	AF003548	AF014944
Environmental isolate W3 18-1		Pacific Ocean marine sediments (630 m, 5–6 cm in core)	AF387350*	AF387355*
Environmental isolate PV-4		Pacific Ocean seawater, Naha Vents, Hawaii	AF387348*	AF387352*
E. coli K-12	ATCC 10798	Human feces	AE000452	AE000447

Type strains (TS) indicated. *, accession nos. for sequences determined in this study.

organisms. This approach demonstrates the utility and breadth of DNA microarray technology. DNA/DNA hybridization, at the gene level where numerous genes can be surveyed independently, reveals valuable information concerning the relatedness among microorganisms at a high level of resolution. Furthermore, the results indicated that most genes in operons had high levels of DNA relatedness and suggest that this type of approach might be useful in whole genome arrays to identify genes or operons that are suspected to be horizontally acquired.

Materials and Methods

Strains and Nucleic Acid Extraction. Strains used in this study and related data are listed in Table 1. Most strains were from the culture collection in the lab of K.N., the two environmental strains (PV-4 and W3 18–1) were from the culture collection in J.Z.'s laboratory, and *E. coli* K-12 was from the American Type Culture Collection. All strains were cultured in nutrient broth (Difco) at 30°C, except for *Shewanella woodyi*, which was grown in marine broth (Difco) at 25°C, and *E. coli*, which was grown in LB (10 g of tryptone/5 g of yeast extract/10 g of NaCl per liter) at 37°C. DNA was typically prepared from 2 ml of exponentially growing culture by using a standard method for bacterial genomic DNA (gDNA) preparation and quantification (12, 13).

DNA Sequencing of 16S rDNA and gyrB Genes and Phylogenetic Analysis. DNA sequences for the SSU rDNA and DNA topoisomerase subunit B (*gyr*B) genes were determined for strains whose sequences had not yet been determined (Table 1). SSU rDNA from strains to be sequenced was amplified by PCR. Each reaction contained (100 μl): 20 ng of DNA, 1× GIBCO/ BRL buffer, 20 mM MgCl₂, 0.2 mM dNTPs, 0.5 units *Taq* DNA polymerase (GIBCO/BRL), and 0.5 μM each primer. Standard bacterial specific primers were used in SSU rRNA PCR and sequencing with some modifications from the original published sequences (14–20). Thermalcycling conditions included: initial denaturation at 94°C for 3 min, then 30 cycles of 94°C for 30 sec, 55°C for 30 sec, and 72°C for 1 min, then a final incubation at 72°C for 5 min. Previously published degenerate PCR primers (UP-1S and UP-2R), thermalcycling conditions, and sequencing primers (UP-1S and UP-2SR) were used to amplify the *gyr*B gene (21). An internal *Shewanella*-specific oligonucleotide to the *gyr*B gene (positions 556–570; 5′ GATGGTGGTACTCAC 3′) was designed to facilitate sequencing. All PCR-amplified *gyr*B and SSU products were purified on Qiagen (Chatsworth, CA) PCR purification columns. Each amplicon was bidirectionally sequenced on an ABI 377 automatic sequencer (Applied Biosystems) with Big Dye chemistry, following the manufacturer's instructions (Michigan State University DNA sequencing facility).

The SSU rDNA sequence contigs were assembled in SEQUENCHER v. 3.0 (Gene Codes, Ann Arbor, MI), then imported into ARB (22), and aligned by comparison with other γ-proteobacteria sequences (>10,000). The *gyr*B sequences were assembled in SEQUENCHER v. 3.0, aligned in GDE (v. 2.2) with other γ-proteobacteria sequences, including many from the *Shewanella* genus that were previously reported (11). The evolutionary relationships for the SSU and *gyr*B sequences were determined by maximum likelihood analysis, including bootstrap replicates (100) calculated by using FASTDNAML v. 1.1 (23, 24). The bootstrap values are superimposed on the optimum tree. Percent nucleotide sequence similarity between full-length SSU rRNA and *gyr*B sequences was calculated in ARB. Nucleotide sequence accession numbers are given in Table 1.

Microarray Construction. DNA microarrays were constructed with 192 ORFs selected from the *S. oneidensis* MR-1 (MR-1) genome sequence, currently in progress, at the Institute for Genome Research. The selected ORFs were biased to include those involved in energy metabolism (specifically anaerobic metabolism and electron transport, 40%) and regulatory function (20%). Other genes involved in nucleic acid and protein metabolism, cellular transport, and stress were also included, amounting to 40% of the genes selected. The preliminary gene identifications were assigned by BLAST (25). The gene list with ORF designations and preliminary gene identifications is published as supplemental data on the PNAS web site, www.pnas.org (Table 2).

The preparation of DNA microarrays was essentially identical to a previous report (26). Amplified products ranged between 242 and 2,900 bases (mean of 1,041 bases) in length. Two to three 100 μl of PCR reactions were pooled and then purified by using a Qiagen 96-well column purification kit. The purified PCR products were suspended in a final concentration of 3 × SSC and transferred to 384-well microtiter plates for arraying. In addition to the MR-1 PCR products, there were 5 ORFs (*act*1, *mfa*1, *mfa*2, *ste*3, and *ras*1) from *Saccharomyces cerevisiae* that were prepared with the same amplification and purification protocol that were used as controls.

Microarrays were printed by using an Omnigrid microarrayer (GeneMachines, San Carlos, CA) arraying robot at a controlled humidity between 50–55% at the Michigan State University *Arabidopsis* Functional Genomics Consortium microarray facility on polyL-lysine-coated slides (27) and SuperAldehyde substrates (TeleChem, Sunnyvale, CA). Standard postprocessing protocols (27, 28) and manufacturer's instructions (TeleChem) were followed. Two replicates of the PCR product set were spotted on each microarray.

Microarray Hybridizations. Genomic DNA was labeled with fluorescent dyes by random priming by using random octamers and the Klenow fragment of DNA polymerase following a protocol available at http://cmgm.stanford.edu/pbrown/protocols/4 genomic.html, with slight modifications. Briefly, for each labeling reaction, 2.25 μg of gDNA was sheared by beadbeating 4 $\times$ 20 sec at high speed. Manufacturer's instructions were followed for incorporation of fluorescent dyes (Cy3-dCTP and Cy5-dCTP, Amersham Pharmacia Pharmacia Biotech) by using the Bioprime DNA Labeling System (GIBCO/BRL), except that a different dNTP mix was used (10$\times$ mix: 1.2 mM each dATP, dGTP, and dTTP/0.6 mM dCTP/10 mM Tris, pH 8.0/1 mM EDTA), and 2 μl of fluorescent dye was added.

To normalize the two channels for label incorporation, DNA concentration differences, and variation in slide scanning, equal amounts of two to five *S. cerevisiae* (act1, *mfa*2, *ste*3, and *ras*1) genes were spiked in at concentrations ranging from 0.01 to 10 ng to each labeling reaction. Reactions were purified with Qiagen Qia-Quick PCR purification columns. Specific activity (SA) of dye incorporation, defined as the number of nucleotides divided by the dye-labeled nucleotides, was determined for the postlabeling of most reactions by measuring Abs_{260}, Abs_{550}, and Abs_{650}, then using the following calculation: SA = [labeled target (ng) $\times$ 1,000] $\times$ (dye incorporated (pmol) $\times$ 324.5)$^{-1}$]. Picomols Cy3 = Abs_{550} $\times$ volume (μl) $\times$ (0.15)$^{-1}$, and pmol Cy5 = Abs_{650} $\times$ volume (μl) $\times$ (0.25)$^{-1}$. Reactions with SAs less than 75 were used in these experiments; the mean SA for all experiments ($n = 54$) was 58 and 71 for Cy3 and Cy5, respectively. Labeling reaction volumes were reduced on a vacuum centrifuge (Savant) to volumes less than 9 μl each.

Each dual fluor-labeled experiment consisted of a test strain and reference strain (MR-1) DNA. Two microarrays were run for each test strain with both labeling combinations (i.e., test = Cy3, reference = Cy5, and visa versa). Five replicate experiments were performed for the environmental isolate W3 18–1. Hybridization reactions were prepared and washed as described in the gDNA labeling protocol mentioned above. Thirty-microliter hybridizations were run for 12 h at 65°C. Microarrays were then washed sequentially in (*i*) 1 $\times$ SSC/0.2% SDS for 5 min, (*ii*) 0.1 $\times$ SSC/0.2% SDS for 5 min, and (*iii*) 0.1 $\times$ SSC briefly, then dried by centrifugation at 500 rpm on a Sorvall T6000D for 5 min. A ScanArray 4000 Microarray Analysis System (Packard BioScience, Billerica, MA) was used for scanning microarrays.

Data Analysis. Signal intensity and local median background for each spot were determined by using SCANALYZE (M. B. Eisen, available at http://rana.lbl.gov/). Signal intensities were background corrected. To be included in further analysis, spots had (*i*) 55% of pixels in the spot greater than 1.5 times the local background (CH1 and CH2 GTB2 quality control parameter from SCANALYZE), and (*ii*) signal intensities greater than 2$\times$ the standard deviation of the mean background signal. The MR-1 hybridization signal (HS) was normalized by multiplying the raw MR-1 background HS by the inverse of the slope determined for the *S. cerevisiae* spiking controls. The theoretical slope (Cy5 $\times$ Cy3^{-1}) of the yeast spiking controls was 1; deviations from 1 were because of variation in labeling efficiency for the two fluorescent dyes. The variance between duplicate spots on each array was determined, and experiments with high variance (>25%) were repeated. The mean variance for all experiments reported in this study was 13.3%. The log ratio (ratios represent the test strain signal $\times$ reference strain ORF signal^{-1}) was calculated, and the mean of two duplicate spots was determined and used in subsequent data analysis. Relationships between different genes and microbial strains from the microarray hybridizations (log ratios) were determined by using hierarchical cluster analysis (CLUSTER) and visualized with TREEVIEW (29). Two-dimensional complete linkage analysis was performed by using both parametric (Pearson) and nonparametric (Spearman) correlations.

Linkage disequilibrium was investigated by determining which ORFs used in this study were grouped in operons. Twenty-one operons represented by two to eight genes per operon were present in this 192-ORF dataset. The fluorescent hybridization signals for the genes in operons were correlated in a pairwise fashion (Pearson correlation coefficient, STATVIEW v. 5.0). The mean correlation value (X_R) for ORFs associated with 21 operons was then compared with an X_R determined from a randomly sampled population from the complete dataset ($n = 72$ pairwise correlations, which was the number of pairwise correlations performed for the 21 different operons). The X_Rs were compared by using Welch's approximate *t* test for unequal variance (30). Nucleotide sequence similarity between the MR-1 ORFs and their homologs in *E. coli* was estimated by BLAST score by our collaborators at the Institute for Genome Research.
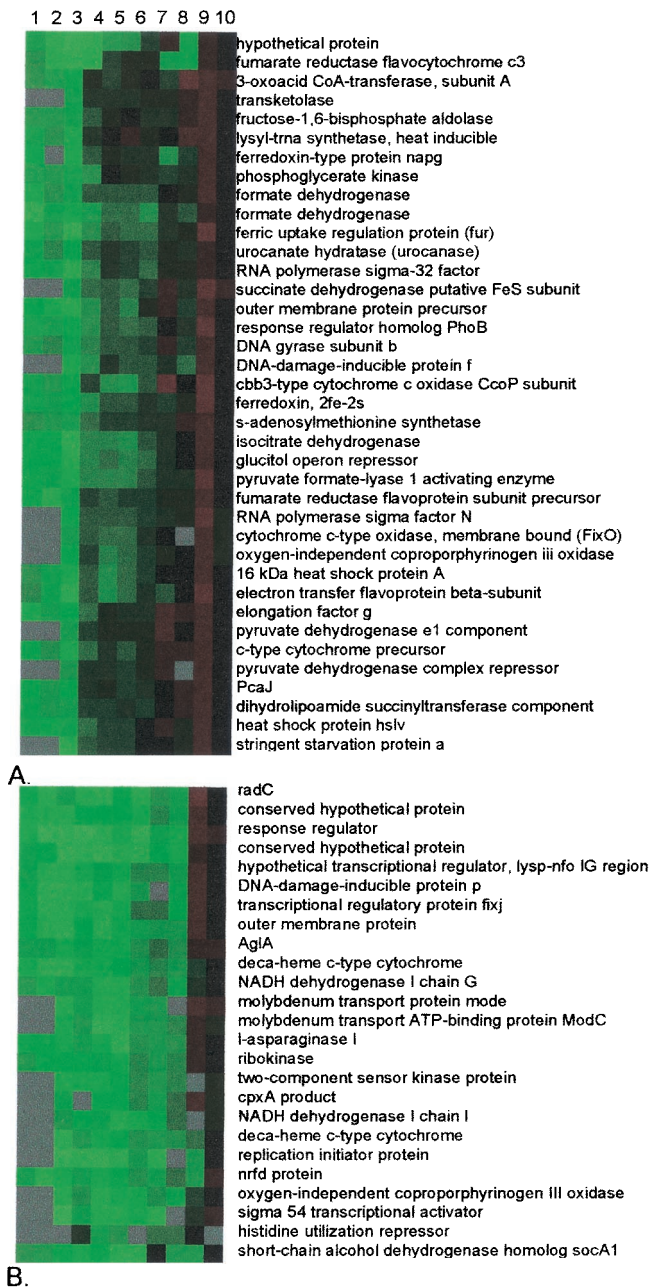
Results

DNA Microarray Analysis. To address the relatedness between *S. oneidensis* MR-1 and other strains in the *Shewanella* genus and *E. coli*, DNA microarrays constructed with genes primarily involved in energy metabolism and regulatory function from MR-1 were hybridized with labeled DNA from 10 different strains. Microarrays were constructed with 192 genes selected from the MR-1 genome. A total of 164 PCR products were determined to provide high-quality spots for most microarrays, although because of low printing volumes for a few (four microarrays, *E. coli*, and *S. woodyi*), only 126 products were spotted. A threshold for detection was determined to be 0.1 ng, although weak but highly variable signals were detected for the 0.01 ng of spiked DNAs.

A highly conserved group of genes including aerobic respiration control protein (*arc*A), ATP synthase (α subunit), and ribosomal protein S12 clustered in all analyses (mean log ratio of 0.00, −0.05 and 0.02, respectively, for all 11 bacteria examined). A second large cluster of conserved genes common to the halotolerant branch of the *Shewanella* genus was readily discriminated (Fig. 1*a*) and corresponded to the phylogenetic association of these organisms (Fig. 2). Several other common clusters of ORFs were identified that supported the close distance between *S. oneidensis* strain MR-4 and MR-1 or between *Shewanella* strain 184 and MR-1 (Fig. 4, which is published as supplemental data on the PNAS web site, www.pnas.org). Conversely, the majority of these ORFs had much lower log ratios (mean of −0.67 and −0.43 for strain MR-4 and strain 184, respectively). Only five of 162 genes between MR-1 and DLM7 (highly related *S. oneidensis* strains) had log ratios < −0.58, indicating poor conservation (Fig. 1*b*). Four of these [replication initiator protein, nitrite reductase subunit (nrfD), short-chain alcohol dehydrogenase, and a decaheme cytochrome] clustered in a group with other poorly conserved ORFs. The mean log ratios were low (−0.97) for the organisms examined in this cluster (excepting MR-1 and DLM7). Cluster analysis for the complete dataset can be found in the supplemental data (Fig. 4, www.pnas.org).

Hierarchical clustering between microarray experiments revealed the relationships between different bacterial hybridization profiles. Duplicate microarray experiments clustered as nearest neighbors in all cases, as did the five replicates of environmental isolate W3 18–1 (data not shown). Branching order for the profiles represented as individual experiments or as means of duplicates was consistent, as was the branching order between *S. putrefaciens* strain 95 and isolate W3 18–1, which are highly related by phylogenetic analysis. However, the overall branching order was somewhat variable, depending on which correlation coefficient was used. The results of the Spearman correlation demonstrate a neighboring association between

1 2 3 4 5 6 7 8 9 10

hypothetical protein
fumarate reductase flavocytochrome c3
3-oxoacid CoA-transferase, subunit A
transketolase
fructose-1,6-bisphosphate aldolase
lysyl-trna synthetase, heat inducible
ferredoxin-type protein napg
phosphoglycerate kinase
formate dehydrogenase
formate dehydrogenase
ferric uptake regulation protein (fur)
urocanate hydratase (urocanase)
RNA polymerase sigma-32 factor
succinate dehydrogenase putative FeS subunit
outer membrane protein precursor
response regulator homolog PhoB
DNA gyrase subunit b
DNA-damage-inducible protein f
cbb3-type cytochrome c oxidase CcoP subunit
ferredoxin, 2fe-2s
s-adenosylmethionine synthetase
isocitrate dehydrogenase
glucitol operon repressor
pyruvate formate-lyase 1 activating enzyme
fumarate reductase flavoprotein subunit precursor
RNA polymerase sigma factor N
cytochrome c-type oxidase, membrane bound (FixO)
oxygen-independent coproporphyrinogen iii oxidase
16 kDa heat shock protein A
electron transfer flavoprotein beta-subunit
elongation factor g
pyruvate dehydrogenase e1 component
c-type cytochrome precursor
pyruvate dehydrogenase complex repressor
PcaJ
dihydrolipoamide succinyltransferase component
heat shock protein hslv
stringent starvation protein a

A.

radC
conserved hypothetical protein
response regulator
conserved hypothetical protein
hypothetical transcriptional regulator, lysp-nfo IG region
DNA-damage-inducible protein p
transcriptional regulatory protein fixj
outer membrane protein
AglA
deca-heme c-type cytochrome
NADH dehydrogenase I chain G
molybdenum transport protein mode
molybdenum transport ATP-binding protein ModC
l-asparaginase I
ribokinase
two-component sensor kinase protein
cpxA product
NADH dehydrogenase I chain I
deca-heme c-type cytochrome
replication initiator protein
nrfd protein
oxygen-independent coproporphyrinogen III oxidase
sigma 54 transcriptional activator
histidine utilization repressor
short-chain alcohol dehydrogenase homolog socA1

B.

Fig. 1. Hierarchical cluster analysis of *Shewanella* DNA/DNA microarray hybridizations. Columns represent the mean log ratios for dual-labeled DNA/DNA hybridizations for: (1) *E. coli*, (2) *S. woodyi*, (3) env. isolate PV-4, (4) *S. oneidensis*, MR-4, (5) *S. putrefaciens* (strain 95), (6) env. isolate W3 18–1, (7) *Shewanella* sp. (strain 184), (8) *Shewanella baltica* (strain 63), (9) *Shewanella* sp. (strain CL256/73), (10) *S. oneidensis* (strain DLM7), and (11) *S. oneidensis* (strain MR-1). Black represents log ratios of 0, and bright green represents log ratios approaching −1.5. (*A*) Conserved cluster of genes. Hybridizations in columns 4–11 are all strains in the halotolerant branch of the *Shewanella* genus. (*B*) Variant cluster of genes. Note that *S. oneidensis* strains DLM7 and MR-1 are very similar, even in the most variable cluster of genes surveyed.

Shewanella strains MR-4 and CL256, which are highly related according to the phylogenetic analysis (SSU rDNA similarity 98.9%, *gyr*B similarity 96.0%). Differences between the correlation coefficients are suspected to result from the differences in yeast normalization correction factors between the experiments, thus suggesting that the nonparametric estimation supports a more conservative, and potentially more reliable, estimation of the relationship between hybridization profiles.
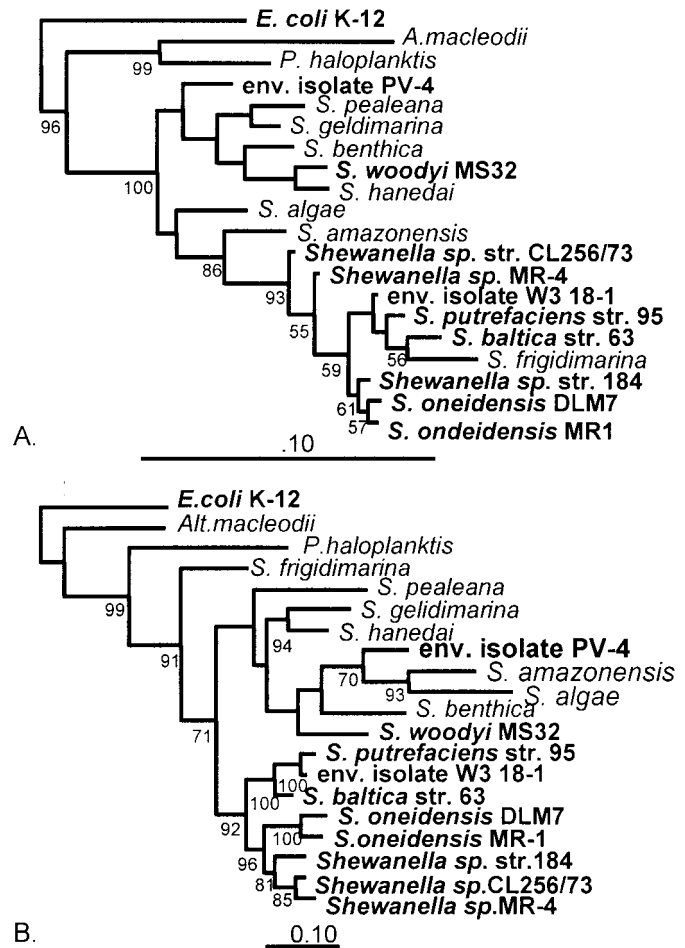
Fig. 2. Phylogenetic tree showing strains of *Shewanella* based on (*A*) SSU rRNA and (*B*) *gyr*B nucleotide sequence. Relationships calculated by using maximum-likelihood analysis. Species in bold were used in this study. Numbers at nodes represent bootstrap values (of 100 replicates).

Regression analysis of the SSU rRNA and *gyr*B sequence percent similarities resulted in a good fit ($r^2 = 0.93$), suggesting a consistent relationship between the sequence differences for the two molecules. The relationship between the hybridization value (log ratio) and the percent nucleotide sequence similarity for *gyr*B indicated a good concordance ($r^2 = 0.6$, $P = 0.008$). Results from microarray hybridization profiles were regressed against the SSU rRNA and *gyr*B nucleotide sequence similarity data. Pairwise percent similarity (P) was calculated for microarray hybridization profiles between the test strain and MR-1 [$P = \Sigma_i$ minimum (p_{1i}, p_{2i}), where P = percentage similarity between the test strain ($p\,1$) and MR-1 ($p\,2$), p_{1i} = percentage of the hybridization signal for gene i in $p\,1$, and p_{2i} = percentage of the hybridization signal for gene i in $p\,2$ (31)]. The percent similarity calculation for the microarray hybridizations was a good predictor of phylogenetic distance for those microorganisms with percent nucleotide similarities > 0.93 or 0.80, resulting in regressions with $r^2 = 0.75$ and 0.97 for the SSU rRNA and *gyr*B genes, respectively (Fig. 3). For more distantly related organisms with low hybridization ratios, there was no relationship.

Operons in which two or more genes are under the same regulatory control were detected by (*i*) proximity in the MR-1 genome (<300 nucleotides between each gene), and (*ii*) common gene function, pathway, or role. Functional assignments were completed by using BLAST. The suite of 164 genes producing
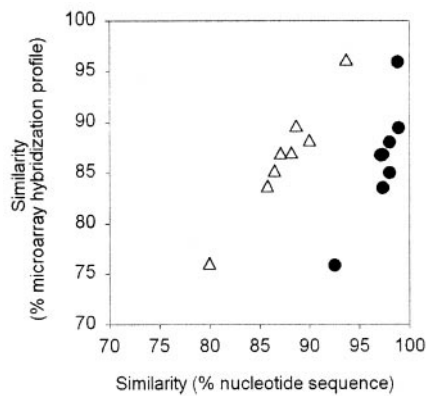
Fig. 3. Relationship between percent similarity in nucleotide sequences for the SSU rRNA (circles) and *gyr*B (triangles) genes (calculated as percent similarity of test strain relative to the *S. oneidensis*, MR-1 strain) and percent similarity (31) determined for microarray hybridization profiles between test strains and *S. oneidensis*, MR-1. Note that the data were not included for *E. coli* and for env. isolate PV-4, the two most distantly related organisms tested in this study, as the relationship falls off at greater distances.

quality hybridization results in this study contained genes representing 21 different operons. Relatedness between organisms investigated at the operon level revealed patterns of DNA relatedness that were consistent within most, although not all, of the operons (Fig. 5, which is published as supplemental data on the PNAS web site, www.pnas.org). The mean pairwise correlation (X_r) between all ORFs in the dataset tested suggests that they are significantly correlated (Welch's approximate t) at a higher level ($X_r = 0.92$, SE 0.01, $n = 72$) than a randomly sampled dataset with the same number of pairwise correlations ($X_r = 0.80$, SE 0.02, $n = 72$).

Array hybridizations by using labeled MR-1 and *E. coli* gDNA revealed the limits of hybridization between distantly related species (SSU rRNA genes are 88.1% similar). Similarity between the MR-1 genome and *E. coli* K12 was assessed by using FASTA3. Approximately one-third of the MR-1 genes had homologs in *E. coli* determined as those genes with E values <0.00001 (J. Eisen, personal communication). Results from the DNA/DNA hybridization experiments indicated that seven genes (*atp*A, *rec*A, *met*X, *arc*A, *gyr*B, *ccm*C, and *pgm*) had significant log ratios (ranging from −0.33 to −0.89) and mean raw hybridization signals (ranging from 6,000 to 12,800) in comparison to overall mean hybridization signals of 3,450 and 3,900 for the two experiments. These genes had pairwise nucleotide sequence similarities (sequences aligned by using CLUSTALX) above 60% between MR-1 and *E. coli*, suggesting that this might be a minimum limit for detection. The overall hybridization of *E. coli* gDNA to the array was not significantly different from the hybridization signals of *S. woodyi*, or environmental isolate PV-4.

Phylogenetic Relationships. Estimation of the evolutionary relationships between the *Shewanella*-related microorganisms of interest in this study was carried out by phylogenetic analysis of the 1.45-kb nucleotide sequences of the SSU rRNA gene (base positions 29–1460 *E. coli* numbering) and for the 1.15-kb nucleotide sequences of the *gyr*B genes. Two variable regions in the *gyr*B sequence (totaling 30 positions) were masked out because of poor alignment. Maximum likelihood phylogenetic trees (Fig. 2) indicate slightly different topologies for the two gene sets, although two distinct clusters are present in both phylogenies. SSU rDNA nucleotide sequence similarity was equal to or greater than 97% for 8 of 10 sequences representing organisms used in this study (referred to as the halotolerent cluster). The faster-evolving gyrB gene had nucleotide sequence similarities

(%) for the same group of sequences ranging between 85.3 and 97.4. The sequences of *S. woodyi* and env. isolate PV-4 were associated with the other cluster, which has a number of organisms isolated from marine environments (referred to as the halophilic cluster). Bootstrap scores support both the SSU rRNA and *gyr*B gene phylogenies. The values on several SSU rRNA branches were low because of a lack of variable positions between the species included in the analysis. The bootstrap scores for the *gyr*B phylogeny are higher for the same branches, supporting the relationships predicted in the maximum likelihood analysis.

Discussion

The microarray platform facilitates analysis of many genes (or complete microbial genomes) in parallel and provides a way to view the relatedness among microorganisms that has not been possible, to our knowledge, for evolutionary biologists, physiologists, or ecologists to study. The ability to "see" the relationships between closely related microbes by using data visualization tools such as cluster analysis (29) provides an unmatched view of how different microorganisms are affiliated with each other.

The *Shewanella* genus comprises a geographically diverse assortment of organisms that have been isolated from nearly as many different habitats. The phylogenetic trees presented in Fig. 2 closely match the results presented in a comprehensive phylogeny of the *Shewanella* genus presented recently (11). Two distinct clades were present in both the SSU and *gyr*B phylogenies, which branch into three clades when more sequences are added to the analysis. The clinical strain tested in the present study (CL256/73) clustered with *S. oneidensis* MR-1, similar to some other clinical isolates presented previously (11).

In comparison to a phylogenetic evolutionary comparison, a microarray-based approach permits comparisons of organisms in a functional genomic framework. After hierarchical cluster analysis, a unique picture of the genetic mosaic of a microbial genome comprising genes ranging from various levels of conservation can be visualized. Those genes that were "universally" conserved could be used to identify new phylogenetic marker candidates. For example, *arc*A (aerobic respiration control), a key gene in the two-component system responsible for repressing transcription of "aerobic" genes under low oxygen conditions, appeared to be highly conserved in all species tested. The nucleotide percent similarity of *arc*A between MR-1 and *E. coli*, the two most distantly related organisms tested, was 73%. Two other well-characterized regulators, *etr*A (analogous to fnr in *E. coli*) and *nar*Q, another two-component regulatory system, appeared to be much less conserved with homologs between *Shewanella* sp. strain 184, DLM7, and MR-1 for *etr*A, and only between DLM7 and MR-1 for *nar*Q.

Under the conditions used in this study, the results of DNA microarray hybridizations for different microorganisms were most robust at the level of >97% rRNA gene similarity, or >90% gyrB similarity. The halotolerant cluster of the *Shewanella* genus contained eight microorganisms that were highly related according to these criteria. The results of the microarray hybridizations (interpreted by calculating pairwise percent similarity) approximated sequence divergence more for *gyr*B than the SSU rRNA gene for those organisms that were closely related, perhaps reflecting functional rather than structural gene characteristics.

Although all organisms tested were capable of anaerobic respiration, there was a large degree of difference in many of the key ORF hybridizations. Those ORFs that resulted in log ratios below −1.0 were considered to be not significant and represented either poor hybridizations because of low degrees of gene conservation or true gene deletions, because the raw signals were often less than 1,000 and approaching background. For example, the genes *mtr*A, B, and E involved in iron reduction (32) had a

mean log ratio of −1.15 for *S. woodyi* and −1.07 for *E. coli*, which are known not to reduce iron oxides and therefore would not be suspected to have these genes. However, other strains that are capable of iron reduction also had low mean log ratios, suggesting that this operon is not highly conserved. This observation was supported by independent experiments based on cloning and sequence analysis of the *mtr*B genes from *S. putrefaciens*, *S. oneidensis* MR-4, isolate W3 18–1, and Southern hybridizations to isolate PV-4. One point of caution for interpretation of the hybridization ratios is the inability to distinguish hybridization signals resulting from cross hybridization to paralogous genes, although in this partial array, we did not include genes with suspected paralogs in MR-1.

Although discrimination of results on the low end (poor hybridizations) is limited, hybridization ratios above −0.5 provide a reliable estimation of gene conservation that can readily identify differences in DNA relatedness between closely related microorganisms. For example, in the gene set investigated, few differences between MR-1 and DLM7 are apparent. Two genes with low similarity, such as nitrite reductase and TMAO reductase, may be useful for investigating further the gene-specific differences between these closely related strains. The array hybridization results are also corroborated in this study with the findings of Venkateswaran *et al.* (11), in which *S. putrefaciens* and MR-1, which had previously been considered to be the same species, are clearly divergent. The whole genome DNA/DNA hybridization results between the two organisms were estimated at 37.5% (11).

Analysis of the *Shewanella* microarray hybridization data set allowed us to test the null hypothesis that DNA relatedness (represented by hybridization ratio) would be evenly distributed between genomes. This was tested and dismissed by comparing the hybridization ratios for randomly sampled genes in the data set to the hybridization ratios for genes associated in operons. The results suggest that the hybridization ratios for genes associated in operons were more likely to be correlated between different microorganisms than a randomly sampled set of genes. These findings imply that operons harbor conserved hybridization patterns that can be identified by using DNA microarrays. Furthermore, it may be possible to identify horizontally transferred operons by using a similar approach; these gene clusters would have conserved levels of hybridization (and high log ratios) in comparison to infrequently transferred genes or operons.

The DNA/DNA hybridization approach applied in this study is an economically viable application of this technology, which maximizes and extends the information gained from one genome sequencing effort to other closely related strains within a reasonably close genetic distance. A comprehensive view of genomic relatedness between closely related strains or species can be determined rapidly, even for organisms that have not been fully characterized such as environmental or clinical isolates. The ability to produce a genome snapshot in terms of genomic relatedness will provide a first approximation of phenotypic potential in a rapid format. This approach can also provide the framework and starting point for further analysis and is therefore a hypothesis-generating tool. This technology will facilitate understanding of microbial evolution, speciation, and genome heterogeneity of organisms that are critical to everyday functions of the ecosystem and its inhabitants.

1. Wayne, L. G., Brenner, D. J., Colwell, R. R., Grimont, P. A. D., Kandler, O., Krichevsky, M. I., Moore, L. H., Moore, W. E. C., Murray, R. G. E., Stackebrandt, E., *et al.* (1987) *Int. J. Syst. Bacteriol.* **37,** 463–464.
2. Stackebrandt, E. & Goebel, B. M. (1994) *Int. J. Syst. Bacteriol.* **44,** 846–849.
3. Lawrence, J. G. & Roth, J. R. (1996) *Genetics* **143,** 1843–1860.
4. Lawrence, J. G. & Ochman, H. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 9413–9417.
5. Ochman, H., Lawrence, J. G. & Grolsman, E. A. (2000) *Nature (London)* **405,** 299–304.
6. Behr, M. A., Wilson, M. A., Gill, W. P., Salamon, H., Schoolnik, G. K., Rane, S. & Small, P. M. (1999) *Science* **284,** 1520–1523.
7. Gingeras, T. R., Ghandour, G., Wang, E., Berno, A., Small, P. M., Drobniewski, F., Alland, D., Desmond, E., Holodniy, M. & Drenkow, J. (1998) *Genome Res.* **8,** 435–448.
8. Hacia, J. G., Makalowski, W., Edgemon, K., Erdos, M. R., Robbins, C. M., Fodor, S. P. A., Brody, L. C. & Collins, F. S. (1998) *Nat. Genet.* **18,** 155–158.
9. Lashkari, D. A., McCusker, J. H. & Davis, R. W. (1997) *Proc. Natl. Acad. Sci USA* **94,** 8945–8947.
10. Troesch, A., Nguyen, H., Miyada, C. G., Desvarenne, S., Gingeras, T. R., Kaplan, P. M., Cros, P. & Mabilat, C. (1999) *J. Clin. Microbiol.* **37,** 49–55.
11. Venkateswaran, K., Moser, D. P., Dollhopf, M. E., Lies, D. P., Saffarini, D. A., MacGregor, B. J., Rinelberg, D. B., White, D. C., Nishijima, M., Sano, H., *et al.* (1999) *Int. J. Syst. Bacteriol.* **49,** 705–724.
12. Ausubel, F. M., Brent, R., Kingston, R. E., Moore, D. D., Seidman, J. G., Smith, J. A. & Struhl, K. (1991) *Current Protocols in Molecular Biology* (Wiley, New York).
13. Sambrook, J., Fritsch, E. F. & Maniatis, T. (1989) *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Lab. Press, Plainview, NY).
14. Amann, R. I., Ludwig, W. & Schleifer, K. H. (1995) *Microbiol. Rev.* **59,** 143–169.
15. Edwards, U., Rogall, T., Blocker, H., Emde, M. & Bottger, E. C. (1989) *Nucleic Acids Res.* **17,** 7843–7853.
16. Lane, D. J., Pace, B., Olsen, G. J., Stahl, D. A., Sogin, M. L. & Pace, N. R. (1985) *Proc. Natl. Acad. Sci. USA* **82,** 6955–6959.
17. Lane, D. J. (1991) in *Nucleic Acid Techniques in Bacterial Systematics*, eds. Stackebrandt, E. & Goodfellow, M. (Wiley, West Sussex, U.K.), pp. 115–175.
18. Rhodes, A. N., Urbance, J. W., Youga, H., Corlew-Newman, H., Reddy, C. A., Klug, M. J., Tiedje, J. M. & Fisher, D. C. (1998) *Appl. Environ. Microbiol.* **64,** 651–658.
19. Teske, A., Alm, E., Regan, J. M., Toze, S., Rittmann, B. E. & Stahl, D. A. (1994) *J. Bacteriol.* **176,** 6623–6630.
20. Weisburg, W. G., Barns, S. M., Pelletier, D. A. & Lane, D. J. (1991) *J. Bacteriol.* **173,** 697–701.
21. Yamamoto, S. & Harayama, S. (1995) *Appl. Environ. Microbiol.* **61,** 1104–1109.
22. Stunk, O. & Ludwig, W. *The ARB Project: A Software Environment for Sequence Data* (Munich).
23. Felsenstein, J. (1981) *J. Mol. Evol.* **17,** 368–376.
24. Olsen, G. J., Matsuda, H., Hagstrom, R. & Overbeek, R. (1994) *Comp. Appl. Biosci.* **10,** 41–48.
25. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *J. Mol. Biol.* **215,** 403–410.
26. Wilson, M., DeRisi, J., Kristensen, H. H., Imboden, P., Rane, S., Brown, P. O. & Schoolnik, G. K. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 12833–12838.
27. Eisen, M. B. & Brown, P. O. (1999) *Methods Enzymol.* **303,** 179–205.
28. DeRisi, J. L., Iyer, V. R. & Brown, P. O. (1997) *Science* **278,** 680–686.
29. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. (1998) *Proc. Natl. Acad. Sci. USA* **9**5, 14863–14868.
30. Zar, J. H. (1999) *Biostatistical Analysis* (Prentice–Hall, Englewood Cliffs, NJ).
31. Krebs, C. J. (1999) *Ecological Methodology* (Addison–Welsey, Reading, MA).
32. Beliaev, A. & Saffarini, D. (1998) *J. Bacteriol.* **180,** 6292–6297.