J. Microbiol. Biotechnol. (2006), **16**(11), 1773–1777

JOURNAL OF MICROBIOLOGY AND BIOTECHNOLOGY

Evaluation of DNA Microarray Approach for Identifying Strain-Specific Genes

HWANG, KEUM-OK[†] AND JAE-CHANG CHO*

Institute of Environmental Science and Engineering, and Department of Environmental Science, Hankuk University of Foreign Studies, Kyoung-Ki Do 449-791, Korea

Received: June 11, 2006
Accepted: July 24, 2006

Abstract We evaluated the usefulness of DNA microarray as a comparative genomics tool, and tested the validity of the cutoff values for defining absent genes in test genomes. Three genome-sequenced *E. coli* strains (K-12, EDL933, and CFT073) were subjected to comparative genomic hybridization with DNA microarrays covering almost all ORFs of the reference strain K-12, and the microarray results were compared with the results obtained from *in silico* analyses of genome sequences. For defining the K-12 ORFs absent in test genomes (reference strain-specific ORFs), we applied and evaluated the cutoff level of −1. The average sequence similarity between ORFs, to which corresponding spots showed a log-ratio of >−1, was 96.9±4.8. The numbers of spots showing a log-ratio of <−1 (P<0.05, t-test) were 90 (2.5%) and 417 (10.6%) for the EDL933 genome and the CFT073 genome, respectively. Frequency of false negatives (FN) was *ca.* 0.2, and the cutoff level of −1.3 was required to achieve the FN of 0.1. The average sequence similarity of the false negative ORFs was 77.8±14.8, indicating that the majority of the false negatives were caused by highly divergent genes. We concluded that the microarray is useful for identifying missing or divergent ORFs in closely related prokaryotic genomes.

Key words: Comparative genomics, microarray, prokaryotic genome

Using whole genome sequences, comparisons of various organisms at the genome level are now available. Comparisons between distantly or intermediately related genomes provide information on the core set of genes (proteins) for life, and the genes (DNA sequences) showing signatures of purifying selection. On the other hand, comparisons between closely related genomes (*e.g.*, genomes from different strains belonging to the same species or genus) are useful for obtaining information on DNA sequences that account for the unique features of organisms tested.

However, when studying many strains, it is costly for individual researchers to carry out whole genome sequencings for every test strain. An alternative method is DNA microarray-based comparative genomics. Recently, several research groups applied the DNA microarray-based approach to reveal gene-specific differences between closely related microbial genomes [6, 8–10, 15, 17]. The approach uses competitive hybridization between differently labeled genomic DNAs. The relative extent of hybridization of target genes to probes on the microarray (*e.g.*, hybridization signal ratio =[Cy3-test signal/Cy5-reference signal]) provides information on whether the DNA sequences complementary to the ORFs of the reference genome are present or absent in test genomes. Whereas ORF probes showing a high signal ratio indicate that the corresponding DNA sequences are present (or highly similar) in both test and reference strains, ORF probes showing a low signal ratio indicate that the corresponding genes are absent in test strains, and hence, unique genes may be present in the reference strain. The major criterion to identifying whether the genes of a reference strain are present or absent in the test strains is based on the hybridization signal ratio. However, the results should vary according to the cutoff values in the signal ratio that are applied to define the absent genes, and researchers have used arbitrary values.

The purpose of this study was to provide insight into a technically feasible cutoff value in terms of sequence similarity. Here, we describe results from the DNA microarray-based comparative genomic hybridizations of *E. coli* strains. We compared the microarray results and *in silico* analysis results from genome sequences, and estimated the accuracy of defining the unique genes in a reference genome, as

*Corresponding author
Phone: 82-31-330-4350; Fax: 82-31-330-4529;
E-mail: chojc@hufs.ac.kr

[†]Present Address; Protein Therapeutics Research Center, Korea Research Institute of Bioscience and Biotechnology, Dae-Jeon 305-333, Korea

well as missing genes in a test genome, with the cutoff values applied.

MATERIALS AND METHODS

Bacterial Strains and DNA Extraction

The three genome-sequenced *Escherichia coli* strains used in this study were nonpathogenic *E. coli* K-12 (MG1655) (ATCC 700926), enterohemorrhagic *E. coli* EDL933 (O157:H7) (ATCC 700927), and uropathogenic *E. coli* CFT073 (ATCC 700928). All strains were purchased directly from American Type Culture Collection (ATCC) and routinely cultivated at 37°C in Luria broth (Difco, Detroit, MI, U.S.A.). Genomic DNAs from the strains were extracted and purified using a Wizard Genomic DNA Purification kit (Promega, Madison, WI, U.S.A.). The concentration of DNA was determined with a fluorometric DNA Quantitation kit (Sigma, St. Louis, MO, U.S.A.). Other details were followed according to the manufacturer's protocols, and previously described by Lee *et al.* [16] and Hwang *et al.* [12].

E. coli Genome Microarray

DNA microarrays (IntelliGene *E. coli* CHIP ver. 2.0) were purchased from Takara Bio Inc. (Otsu, Japan). The microarrays were immobilized with 4,155 PCR-amplified DNA fragments, which cover 94.6% of the 4,390 annotated open reading frames (ORFs) of *E. coli* K-12 (W3110) (http://ecoli.aist-nara.ac.jp).

Genomic DNA Labeling and Hybridization

Genomic DNAs (1 μg) from all the strains listed were labeled with FluoroLink Cy3-dCTP (Amersham Pharmarcia, Piscataway, NJ, U.S.A.) by random priming (HighPrime; Roche, Indianapolis, IN, U.S.A.) and used as test DNAs. Genomic DNA (1 μg) from *E. coli* K-12 was labeled with FluoroLink Cy5-dCTP (Amersham Pharmarcia) and used as reference DNA for the hybridization signal ratio calculation (Cy3-test/Cy5-reference). Unincorporated fluorescent nucleotides were removed using a Sephadex MicroSpin G-50 column (Amersham Pharmarcia).

The microarrays were prehybridized in prehybridization buffer (3.5×SSC, 0.1% SDS, 10 mg/ml bovine serum albumin) for 20 min at 65°C, hybridized with approximately 1 g of Cy3- and Cy5-labeled DNA mixture (1:1) in hybridization buffer (3×SSC, 0.1% SDS, 0.5 mg/ml herring sperm DNA) at 65°C for 16 h, and then washed once with primary wash buffer (0.1×SSC, 0.1% SDS) at room temperature for 5 min and twice with secondary wash buffer (0.1×SSC) for 5 min. Other details were previously described by Cho and Tiedje [4, 5].

Scanning and Data Processing

Hybridized arrays were scanned with a GenePix 4000B laser scanner (Axon, Foster City, CA, U.S.A.). Laser lights of wavelengths 532 nm and 635 nm were used to excite the Cy3 dye and the Cy5 dye, respectively. Fluorescent images were captured as a multi-image tagged image file format (TIFF) and analyzed with GenePix Pro 3.0 software (Axon) according to the manufacturer's protocol. Subsequent data analyses were conducted using Microsoft Excel and Acuity 3.0 software (Axon).

The spots showing a signal-to-noise (S/N) ratio (foreground hybridization signal/background hybridization signal) of <2 were excluded from further analysis. The ratio (R) of the extent of hybridization between test DNAs and reference DNAs was derived from a median value of pixel-by-pixel ratios. Using this approach to calculate R, nonspecific signals (which appear in both wavelength images) had less of an effect than when the mean values of a whole spot were used. The hybridization ratios (R) were $\log_2$ transformed and normalized using the mean $\log_2$ ratios of all spots as zero (global normalization). The normalized $\log_2$ ratios (Log_2 R') from quadruplicate experiments were averaged and used as the final values for each ORF tested.

To determine the consistency of ratios across replicate hybridizations, a *t*-test was applied. We used only those ratios having a 95% confidence interval as determined by the *t*-test.

Sequence Comparison

Genome sequences of strains K-12 [3], EDL933 [18], and CFT073 [21] were downloaded from The Institute for Genome Research (TIGR) Web site (http://www.tigr.org). Sequence similarity (% identity) between ORFs of the reference and the test genomes was calculated after the global alignment using a default scoring matrix used in BLAST [1], with penalties for opening and extension gaps of 16 and 7, respectively.

RESULTS AND DISCUSSION

Since the microarrays were fabricated with ORFs from *E. coli* K-12, the genomic DNA from strain K-12 was used as the reference genome and cross-hybridized to genomes from strains EDL933 and CFT073, respectively. Results from strain K-12 self-hybridization (K-12 *vs.* K-12) and the two cross-hybridizations (K-12 *vs.* EDL933 and K-12 *vs.* CFT073; hereafter, EDL933 hybridization and CFT073 hybridization, respectively) are shown in Fig. 1. In the quadruplicate experiments for each strain, 4,052 (97.5%), 3,610 (86.9%), and 3,922 (94.4%) spots in self-hybridization, EDL933 hybridization, and CFT073 hybridization, respectively, passed the spot-quality control criteria. Because we included only spots that passed the quality control criteria in all four hybridizations, the number of the criteria-passed spots for each strain would be arbitrary and not affect subsequent analyses. After global normalization, self-hybridization

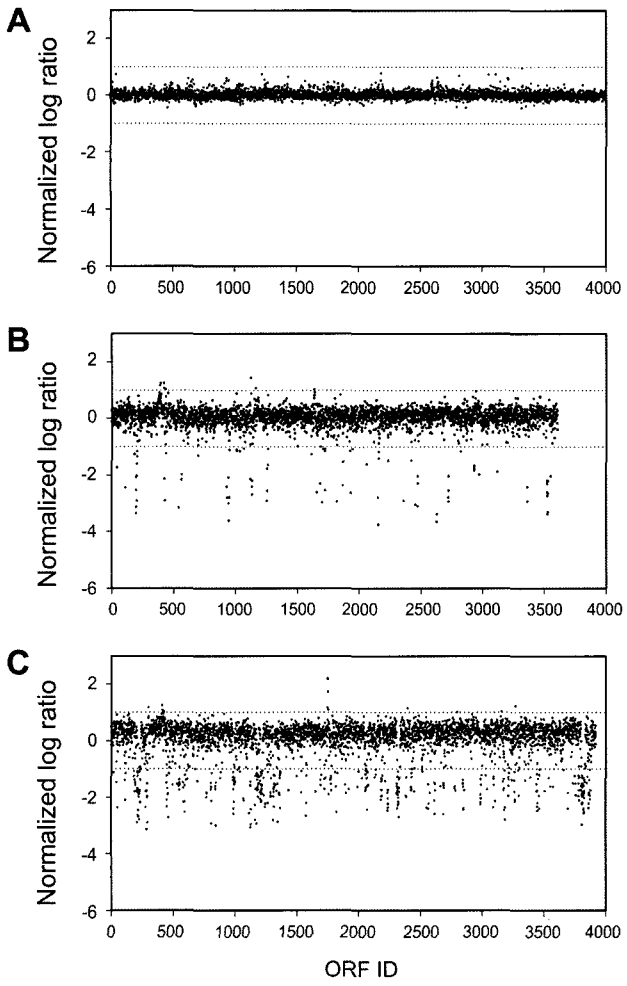

Fig. 1. Scatterplot diagram of microarray hybridization profiles of *E. coli* K-12 self-hybridization (**A**), *E. coli* EDL933 (**B**), and *E. coli* CFT073 (**C**) cross-hybridizations.

E. coli K-12 ORFs are arbitrarily shown on the x axis, and normalized $\log_2$ ratios (Cy3-test genome/Cy5-reference genome) for each ORF are shown on the y axis. Dotted lines indicate the log-ratio of +1 and −1, respectively.

(Fig. 1A) showed that the standard deviation of the log-ratios was 0.13, and none of the spots had log-ratios of >1 or <−1, indicating high reproducibility of the microarray experiments. On the other hand, the standard deviations of EDL933 hybridization and CFT073 hybridization were 0.46 and 0.69, respectively. Because the log-ratios of spots corresponding to the missing or duplicated ORFs are expected to be scattered from zero, the larger standard deviation of log-ratios implies a more divergent genome from the reference genome. We surmised from the standard deviation of the log-ratios that the CFT073 genome was more divergent than the EDL933 genome from the reference K-12 genome.

For defining the K-12 ORFs absent in test genomes, we applied and evaluated the cutoff level of −1. The numbers of spots showing a log-ratio of <−1 (P<0.05) were 90 (2.5%) and 417 (10.6%) for the EDL933 genome and the CFT073 genome, respectively. Although there have been no empirical or theoretical guidelines for cutoff values defining missing genes in the test strains, the most frequently used cutoff value in publications is −1 on the $\log_2$ scale [2, 9, 13, 14, 19]. We supposed that the value of −1 might be derived from microarray-based gene expression analysis, where ORFs showing a $\log_2$ ratio of <−1 are considered to be underexpressed.

To determine the validity of the cutoff level of <−1 for defining K-12 ORFs absent in the test genomes, we compared the list of K-12 ORFs on the reference genome array with the list of annotated ORFs of test genomes. If the K-12 ORFs with log-ratios below the cutoff level of <−1 were present in the ORF list of the test genomes, such K-12 ORFs were recorded as false negatives. Log-ratios were divided into log-ratio classes with intervals of 0.1, and the cumulative number of K-12 ORFs on the reference genome array (OM) and the cumulative number of K-12 ORFs present in test genome sequences (OS) were calculated. The ratio of OS to OM (OS/OM) in each log-ratio class obtained from EDL933 hybridization and CFT073 hybridization were averaged and plotted against the log-ratio classes (Fig. 2). Sigmoidal regression analysis resulted in an equation, $\text{OS/OM ratio}=0.01+(0.81/(1+e^{-(\text{log-ratio}-0.01)/0.14})^{0.21})$, in which the coefficient of determination (r^2) was 1.00. At the log-ratio class of −1, the OS/OM ratio, which indicates the frequency of the false negatives (FN), was *ca.* 0.2. To achieve the FN of 0.1, a cutoff level of −1.3 was required.

For inferring the reason why 20% of the K-12 ORFs present in test genomes showed a log-ratio of <−1 in

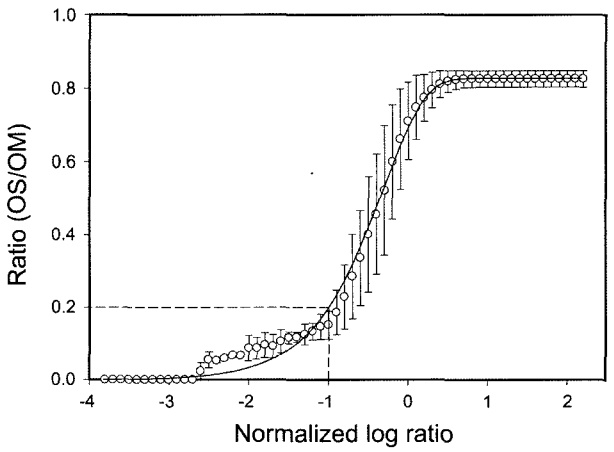

Fig. 2. The frequency of false negatives in comparative genomic hybridization using microarray.

Normalized log-ratios are divided into log-ratio classes with intervals of 0.1 and are shown on the x axis. The proportions of the cumulative number of K-12 ORFs present in test genome sequences (OS) to the cumulative number of K-12 ORFs on the reference genome array (OM) in each log-ratio class are shown on the y axis (ratio=OS/OM). Error bars indicate the range of data and a solid line indicates the sigmoidal regression curve.

the comparative genomic microarray hybridizations, we calculated sequence similarities between the K-12 ORFs and the ORFs of the test genomes. The average sequence similarity of the false negative ORFs was 77.8±14.8, indicating that the K-12 ORFs did not hybridize to the ORFs of test genomes, as the sequence similarities between them were below a critical point. We supposed that the majority of the false negatives were caused by highly divergent genes, since the sequence similarity for itself is not very predictive of biological function. For example, *yfcU*, which codes for an outer membrane protein, is present in both K-12 and CFT073 genomes, but the sequence similarity between K-12 *yfcU* and CFT073 *yfcU* was 44.7%, resulting in a log-ratio of −1.74 in the comparative genomic microarray hybridization. However, 12 (2.4%) false negative ORFs (5 EDL933 ORFs and 7 CFT073 ORFs) showed a sequence similarity of >90%. We first thought that the 12 false negatives were caused by gene-specific dye incorporation bias [7, 20], but the log-ratios of the 12 false negative ORFs in self-hybridization were not significantly different from zero, indicating no such gene-specific biases in our experiments. We performed PCRs to examine whether ORF deletions had occurred during the cultivation of the test strains, and all PCR tests gave positive results. Although the average log-ratio of the 12 false negative ORFs (−1.40±0.28) was marginal to the cutoff level, it was baffling as to why such highly similar ORFs did not hybridize to K-12 ORFs. However, except for the 12 false negatives (2.4%), the comparative microarray hybridization detected all of the absent K-12 ORFs in the test genomes [18, 21].

On the other hand, the great majority of spots showed log-ratios between −1 and +1 (Fig. 1), and the number of spots showing a log-ratio near zero (−1 SD <log-ratio <+1 SD) were 3,216 (89.1%) and 3,296 (84.0%), respectively. These spots were expected to be K-12 ORFs equally present in EDL933 and CFT073 genomes. Average sequence similarities between K-12 ORFs, whose corresponding spots showed a log-ratio of >−1, and the ORFs of test strains EDL933 and CFT073 were 97.9±4.0 and 95.7±5.4, respectively.

The numbers of spots showing a log-ratio of >1 (P<0.05) were eight and 11 for EDL933 and CFT073 hybridizations, respectively. The majority of these corresponding ORFs were coding for hypothetical proteins, and were considered to be K-12 ORFs possibly duplicated in the EDL933 and CFT073 genomes. A duplicated K-12 ORF common to the two test genomes was a prophage gene, *yeeU* [11]. Although the criteria for defining the duplicated genes is beyond the scope of this study, the determination of the cutoff level for defining a duplicated gene is much more difficult compared with the cutoff level for defining missing genes. Even if the duplicated ORFs in test genomes are expected to have increased log-ratios, the possibility of nonspecific hybridization should be considered. Examination of potential nonspecific hybridization between related sequences, such as those derived from one gene family, has revealed that ORF-type probes could not distinguish target DNAs with >*ca.* 80% sequence similarity [22]. Such nonspecific hybridization might cause overestimated log-ratios, subsequently resulting in many false positives when a low cutoff level is applied. A solution to nonspecific hybridization could be the use of oligonucleotide probes, which make the precise control of stringency possible. However, the oligonucleotide microarrays, which have probes that distinguish perfect matches from even one-base mismatches, may result in many false negatives compared with ORF-type microarrays [8].

In conclusion, our results suggested that the microarray is reliable and powerful for comparing prokaryotic genomes. Although we used genome-sequenced strains in this study, microarray-based comparative genomics could certainly be applicable to unsequenced strains. A cutoff level of <−1 on the $\log_2$ scale was adequate to detect absent and divergent ORFs in the test genome with only 2.4% of true false negatives. Considering the cost of genome sequencing, the microarray method provides a rapid and convenient tool for prokaryotic comparative genomics. One drawback of the microarray method is that it detects ORFs that are spotted on the microarray in test genomes. Hence, identification of only reference genome-specific ORFs are possible and the microarray, which is fabricated with ORFs from the reference genome, should be available. However, there is no need to sequence the reference genome for comparative genomics purposes, since random genome fragments from the reference strain can be cloned and spotted on the microarray (shotgun microarray) [4]. With shotgun microarray, only the genome fragments of interest require sequencing after hybridization. We concluded that the microarray is useful to detect strain-specific ORFs for a variety of topics, ranging from evolution to the search for new pharmaceuticals.

Acknowledgements

This work was supported by the Korea Research Foundation Grant (KRF-2005-041-C00458) funded by the Korean Government (MOEHRD). We thank James Tiedje for his helpful discussions.

References

1. Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.
2. Bjorkholm, B., A. Lundin, A. Sillen, K. Guillemin, N. Salama, C. Rubio, J. I. Gordon, P. Falk, and L. Engstrand. 2001.

Comparison of genetic divergence and fitness between two subclones of *Helicobacter pylori*. *Infect. Immun.* **69:** 7832–7838.

3. Blattner, F. R., G. Plunkett 3rd, C. A. Bloch, N. T. Perna, V. Burland, M. Riley, J. Collado-Vides, J. D. Glasner, C. K. Rode, G. F. Mayhew, J. Gregor, N. W. Davis, H. A. Kirkpatrick, M. A. Goeden, D. J. Rose, B. Mau, and Y. Shao. 1997. The complete genome sequence of *Escherichia coli* K-12. *Science* **277:** 1453–1474.
4. Cho, J. C. and J. M. Tiedje. 2001. Bacterial species determination from DNA-DNA hybridization by using genome fragments and DNA microarrays. *Appl. Environ. Microbiol.* **67:** 3677–3682.
5. Cho, J. C. and J. M. Tiedje. 2002. Quantitative detection of microbial genes by using DNA microarrays. *Appl. Environ. Microbiol.* **68:** 1425–1430.
6. Dobrindt, U., F. Agerer, K. Michaelis, A. Janka, C. Buchrieser, M. Samuelson, C. Svanborg, G. Gottschalk, H. Karch, and J. Hacker. 2003. Analysis of genome plasticity in pathogenic and commensal *Escherichia coli* isolates by use of DNA arrays. *J. Bacteriol.* **185:** 1831–1840.
7. Dombkowski, A. A., B. J. Thibodeau, S. L. Starcevic, and R. F. Novak. 2004. Gene-specific dye bias in microarray reference designs. *FEBS Lett.* **560:** 120–124.
8. Dong, Y., J. D. Glasner, F. R. Blattner, and E. W. Triplett. 2001. Genomic interspecies microarray hybridization: Rapid discovery of three thousand genes in the maize endophyte, *Klebsiella pneumoniae* 342, by microarray hybridization with *Escherichia coli* K-12 open reading frames. *Appl. Environ. Microbiol.* **67:** 1911–1921.
9. Fukiya, S., H. Mizoguchi, T. Tobe, and H. Mori. 2004. Extensive genomic diversity in pathogenic *Escherichia coli* and *Shigella* strains revealed by comparative genomic hybridization microarray. *J. Bacteriol.* **186:** 3911–3921.
10. Ge, H., Y. Y. Chuang, S. Zhao, J. J. Temenak, and W. M. Ching. 2003. Genomic studies of *Rickettsia prowazekii* virulent and avirulent strains. *Ann. NY Acad. Sci.* **990:** 671–677.
11. Hayashi, T., K. Makino, M. Ohnishi, K. Kurokawa, K. Ishii, K. Yokoyama, C. G. Han, E. Ohtsubo, K. Nakayama, T. Murata, M. Tanaka, T. Tobe, T. Iida, H. Takami, T. Honda, C. Sasakawa, N. Ogasawara, T. Yasunaga, S. Kuhara, T. Shiba, M. Hattori, and H. Shinagawa. 2001. Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. *DNA Res.* **8:** 11–22.
12. Hwang, K. O. and J. C. Cho. 2005. Diversity and genotypic structure of ECOR collection determined by repetitive extragenic palindromic PCR genome fingerprinting. *J. Microbiol. Biotechnol.* **15:** 672–677.
13. Israel, D. A., N. Salama, C. N. Arnold, S. F. Moss, T. Ando, H. P. Wirth, K. T. Tham, M. Camorlinga, M. J. Blaser, S. Falkow, and R. M. Peek Jr. 2001. *Helicobacter pylori* strain-specific differences in genetic content, identified by microarray, influence host inflammatory responses. *J. Clin. Invest.* **107:** 611–620.
14. Israel, D. A., N. Salama, U. Krishna, U. M. Rieger, J. C. Atherton, S. Falkow, and R. M. Peek Jr. 2001. *Helicobacter pylori* genetic diversity within the gastric niche of a single human host. *Proc. Natl. Acad. Sci. USA* **98:** 14625–14630.
15. Koide, T., P. A. Zaini, L. M. Moreira, R. Z. Vencio, A. Y. Matsukuma, A. M. Durham, D. C. Teixeira, H. El-Dorry, P. B. Monteiro, A. C. da Silva, S. Verjovski-Almeida, A. M. da Silva, and S. L. Gomes. 2004. DNA microarray-based genome comparison of a pathogenic and a nonpathogenic strain of *Xylella fastidiosa* delineates genes important for bacterial virulence. *J. Bacteriol.* **186:** 5442–5449.
16. Lee, S. H., H. R. Oh, J. H. Lee, S. J. Kim, and J. C. Cho. 2004. Cold-seep sediment harbors phylogenetically diverse uncultured bacteria. *J. Microbiol. Biotechnol.* **14:** 906–913.
17. Murray, A. E., D. Lies, G. Li, K. Nealson, J. Zhou, and J. M. Tiedje. 2001. DNA/DNA hybridization to microarrays reveals gene-specific differences between closely related microbial genomes. *Proc. Natl. Acad. Sci. USA* **98:** 9853–9858.
18. Perna, N. T., G. Plunkett 3rd, V. Burland, B. Mau, J. D. Glasner, D. J. Rose, G. F. Mayhew, P. S. Evans, J. Gregor, H. A. Kirkpatrick, G. Posfai, J. Hackett, S. Klink, A. Boutin, Y. Shao, L. Miller, E. J. Grotbeck, N. W. Davis, A. Lim, E. T. Dimalanta, K. D. Potamousis, J. Apodaca, T. S. Anantharaman, J. Lin, G. Yen, D. C. Schwartz, R. A. Welch, and F. R. Blattner. 2001. Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature* **409:** 529–533.
19. Salama, N., K. Guillemin, T. K. McDaniel, G. Sherlock, L. Tompkins, and S. Falkow. 2000. A whole-genome microarray reveals genetic diversity among *Helicobacter pylori* strains. *Proc. Natl. Acad. Sci. USA* **97:** 14668–14673.
20. Tseng, G. C., M. K. Oh, L. Rohlin, J. C. Liao, and W. H. Wong. 2001. Issues in cDNA microarray analysis: Quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Res.* **29:** 2549–2557.
21. Welch, R. A., V. Burland, G. Plunkett 3rd, P. Redford, P. Roesch, D. Rasko, E. L. Buckles, S. R. Liou, A. Boutin, J. Hackett, D. Stroud, G. F. Mayhew, D. J. Rose, S. Zhou, D. C. Schwartz, N. T. Perna, H. L. Mobley, M. S. Donnenberg, and F. R. Blattner. 2002. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *Proc. Natl. Acad. Sci. USA* **99:** 17020–17024.
22. Wu, L., D. K. Thompson, G. Li, R. A. Hurt, J. M. Tiedje, and J. Zhou. 2001. Development and evaluation of functional gene arrays for detection of selected genes in the environment. *Appl. Environ. Microbiol.* **67:** 5780–5790.