Evaluating the Performance of Oligonucleotide Microarrays for Bacterial Strains with Increasing Genetic Divergence from the Reference Strain

Seungdae Oh, Deborah R. Yoder-Himes, James Tiedje, Joonhong Park and Konstantinos T. Konstantinidis

Appl. Environ. Microbiol. **2010, 76(9):2980. DOI: 10.1128/AEM.02826-09.**
Published Ahead of Print 12 March 2010.

APPLIED AND ENVIRONMENTAL MICROBIOLOGY, May 2010, p. 2980–2988 Vol. 76, No. 9
0099-2240/10/$12.00 doi:10.1128/AEM.02826-09

Evaluating the Performance of Oligonucleotide Microarrays for Bacterial Strains with Increasing Genetic Divergence from the Reference Strain▽

Seungdae Oh,[1]† Deborah R. Yoder-Himes,[3]† James Tiedje,[4] Joonhong Park,[5] and Konstantinos T. Konstantinidis[1,2]*

School of Civil and Environmental Engineering[1] and School of Biology,[2] Georgia Institute of Technology, Atlanta, Georgia; Department of Microbiology and Molecular Genetics, Harvard Medical School, Boston, Massacusetts[3]; Center for Microbial Ecology, Michigan State University, East Lansing, Michigan[4]; and Department of Civil and Environmental Engineering, Yonsei University, Seoul, South Korea[5]

Received 23 November 2009/Accepted 25 February 2010

DNA oligonucleotide microarrays (oligoarrays) are being developed continuously; however, several issues regarding the applicability of these arrays for whole-genome DNA-DNA strain comparisons (genomotyping) have not been investigated. For example, the extent of false negatives (i.e., no hybridization signal is observed when the amino acid sequence is conserved but the nucleotide sequence has diverged to a level that does not allow hybridization) remains speculative. To provide quantitative answers to such questions, we performed competitive DNA-DNA oligoarray (60-mer) hybridizations with several fully sequenced (tester) strains and a reference strain (whose genome was used to design the oligoarray probes) of the genus *Burkholderia* and compared the experimental results obtained to the results predicted based on bioinformatic modeling of the probe-target pair using the available sequences. Our comparisons revealed that the fraction of the total probes that provided experimental results consistent with the predicted results decreased substantially with increasing divergence of the tester strain from the reference strain. The fractions were 90.8%, 84.3%, and 77.4% for tester strains showing 96% 89%, and 80% genome-aggregate average nucleotide identity (ANI) to the reference strain, respectively. New approaches to determine gene presence or absence based on the hybridization signal, which outperformed previous approaches (e.g., 92.9% accuracy versus 86.0% accuracy) and to normalize across different array experiments are also described. Collectively, our results suggest that the performance of oligoarrays is acceptable for tester strains showing >90% ANI to the reference strain and provide useful guidelines for using oligoarray applications in environmental gene detection and gene expression studies with strains other than the reference strain.

DNA microarrays are common tools in the modern molecular laboratory (34). Oligoarrays, which are typically comprised of one or more short, 30- to 60-nucleotide probes for each gene in the genome, have gained popularity compared with microarrays made with PCR products of genes for expression studies due to their greater specificity during hybridization, flexibility of design, and potential for further technological development (10, 23). More recently, oligoarrays have been used for genetic (or DNA-DNA) comparisons of strains. For this application, the array is typically built using the available genomic sequence of one strain (the reference strain) and is used for competitive hybridization with genomic DNA of closely related strains (the tester strains) (3, 8, 24). The objective is to reveal the differences in genes between the reference strain and tester strains (genomotyping) that could explain the unique characteristics of the strains. Such DNA-DNA oligoarray experiments are also a promising and cost-effective means to uncover the gene content of natural microbial assemblages and thereby track their dynamics over time and during environmental fluctuations. Several studies of this type have been performed recently (4, 29, 32).

One cornerstone principle of DNA-DNA experiments is that the intensity of the hybridization signal of a probe depends directly on the degree of relatedness between the probe and target sequences. The relationship between signal intensity and sequence identity has been studied extensively previously, based on both simple sequence comparison models and more sophisticated models that incorporate the thermodynamic interactions of the probe-target sequence pair. Accordingly, we have a relatively good understanding of the maximum divergence of the target sequence from the probe sequence that provides a hybridization signal greater than the background signal, which indicates the presence of the corresponding target sequence in the hybridization mixture based on the array results (6, 13, 19, 21, 30). However, most, if not all, of the previous studies have been performed using a small set of selected probes and target sequences (13, 21). When thousands of target sequences (e.g., a whole genome) are hybridized, additional complications that cannot be accounted for by the simple experiments described above can occur. Examples of these complications include stereochemical interactions among the many targets and probes present (e.g., nonspecific hybridization), competition between (partially) identical target sequences (cross-hybridization), and nonuniform hybridization

* Corresponding author. Mailing address: School of Civil and Environmental Engineering, Georgia Institute of Technology, 311 Ferst Dr., Atlanta, GA 30332-0512. Phone: (404) 385-3628. Fax: (404) 894-8266. E-mail: kostas@ce.gatech.edu.
† S.O. and D.R.Y.-H. contributed equally to this work.
▽ Published ahead of print on 12 March 2010.

behavior of targets due to differences in melting hybridization temperature (T_m) and G+C content. Therefore, the applicability of the results of the previous studies that were based on a few probes and target sequences to the whole-genome level remains somewhat speculative. More importantly, it has been established that DNA-DNA experiments provide meaningful results with tester strains belonging to the same species as the reference strains (i.e., strains whose genomes show more than 95 to 96% nucleotide identity [9]) but not with tester strains related to the reference strain at levels below the sequence threshold for obtaining a significant signal greater than the background signal (usually around 80% nucleotide identity for oligoarrays [21]). However, we do not have a robust quantitative understanding of how strains with intermediate genetic relatedness (e.g., 80 to 95%) behave in such experiments. Dealing with these issues is important for DNA-DNA microarray applications, especially applications for environmental samples, which are characterized by high levels of species and genomic heterogeneity. For instance, natural populations sometimes show intrapopulation genomic diversity (expressed as levels of nucleotide identity) ranging from 90 to 100%, as revealed by recent metagenomic surveys (15, 31). The issues described above are also relevant to the use of oligoarrays for gene expression studies with strains other than the reference strain.

Here, we quantitatively assessed the performance of oligoarrays by conducting competitive hybridizations between a reference strain and selected tester strains that were fully sequenced and showed increasing genetic divergence from the reference strain. We then compared the experimental oligoarray results to bioinformatically predicted results based on whole-genome sequence comparisons. Our analyses allowed precise estimation of the numbers of false negatives (FN) (no hybridization signal observed when the amino acid sequence is conserved but the nucleotide sequence has diverged enough that there is no hybridization) and false positives (FP) for strains belonging to the same species as the reference strain or closely and moderately related species. We also describe a new simple method to normalize arrays from different experiments and determine the presence of a gene based on the hybridization signal that outperforms previously described approaches.

MATERIALS AND METHODS

Strains used in the study and their genetic relatedness. The following six strains were used in this study: *Burkholderia cenocepacia* J2315 (reference strain), *B. cenocepacia* AU1054, *B. cenocepacia* HI2424, *Burkholderia ambifaria* AMMD, *Burkholderia vietnamiensis* G4, and *Burkholderia xenovorans* LB400. The whole-genome sequences of these six strains were obtained from the National Center for Biotechnology Information website (ftp://ftp.ncbi.nih.gov/). The genetic relatedness between the reference strain and a tester strain was measured based on the genome-aggregate average nucleotide identity (ANI) and 16S rRNA gene sequence identity. ANI values were calculated based on all reciprocal best-match conserved genes for the two strains (pairwise), as previously described (17). To calculate 16S rRNA gene sequence identities, all 16S rRNA gene copies in the genome of the reference strain were queried against the genome of the tester strain using the BLASTN algorithm (2). The average nucleotide identities for all copies, as determined by BLAST, are shown in Table 1.

DNA microarray fabrication. The arrays used in this study were *B. cenocepacia* 60-mer Agilent oligoarrays with one probe per gene, which have been described previously (18). Here, we focused on the 6,308 probes of this array that were designed based on the gene sequences of the *B. cenocepacia* strain J2315 genome (reference).

TABLE 1. Tester strains used in this study and their levels of relatedness to the reference strain

Tester strain	% Relatedness to J2315[a]	
	ANI	16S rRNA gene sequence identity
B. cenocepacia HI2424	95.1	99.7
B. cenocepacia AU 1054	95.1	99.7
B. ambifaria AMMD	90.3	99.4
B. vietnamiensis G4	89.1	99.4
B. xenovorans LB400	79.8	96.0

[a] ANI was calculated as described previously (17); 16S rRNA gene sequence identity was calculated as described in Materials and Methods.

DNA microarray template preparation, hybridization, and signal processing. Genomic DNA from all strains was purified from 5-ml cultures grown in LB medium overnight at 37°C by using a Wizard genomic DNA purification kit (Promega, Madison, WI) according to the manufacturer's instructions. Two micrograms of genomic DNA from each strain was sonicated using a Heat Systems Ultrasonics W-225 20-kHz, 200-W cup sonicator (Misonix, Farmingdale, NY) to generate sheared genomic DNA from 0.5 to 5 kb long. For each sample, 300 ng of sheared DNA was labeled with either Cy3 or Cy5 dye using the methods described by Wick et al. (33). Equal amounts of tester and reference (J2315) samples with similar specific activities were mixed in 4 μl 10 mM EDTA (pH 8.0) and heated to 95°C for 5 min. Samples were mixed with 45 μl of SlideHyb buffer 1 (Ambion, Austin, TX) by pipetting. Preparation of hybridization mixtures, loading procedures, and slide washing were performed by using Agilent's recommended protocols (1) with 16 h of hybridization at 65°C and with agitation and the optional stabilization and drying solution final wash. Two or more hybridizations were performed for each tester strain with at least one dye swap. Microarrays were scanned using an Axon GenePix 4000B scanner (Molecular Devices, Sunnyvale, CA), and probe intensities were extracted using GenePix Pro 5.0 software. The log hybridization signal ratio for each probe was calculated using the ratio of the background-subtracted mean probe signal of the tester to the background-subtracted mean probe signal of the reference strain with the following equation: log hybridization ratio = $\log[(MS - BS)_{tester}/(MS - BS)_{reference}]$, where MS and BS are the mean probe and median background signals, respectively.

Bioinformatics sequence analysis. To bioinformatically determine the degree of relatedness between a probe sequence and the target sequence in a tester genome, the BLAST score approach was used, essentially as described previously (26). In brief, the 60-mer oligonucleotide probe sequences were searched against the whole genome of a tester strain using the BLASTN algorithm (2) with the following settings: $X = 150$ (drop-off value for gapped alignment), $q = -1$ (penalty for nucleotide mismatch), $F = F$ (filter for repeated sequences), $W = 7$ (word size), and the default settings for the rest of the parameters. These settings are more robust for identifying moderately related short matches (i.e., sequences showing 60 to 80% nucleotide identity) than the default settings, which target sequences with high levels of identity (>90% nucleotide identity) (17). The BLAST score of a probe against the tester genome was calculated by subtracting the number of internal mismatches from the total length of the alignment for only the best BLASTN match (for multiple matching targets, see below). Thus, the BLAST score for probes that had a perfect match was 60. The BLAST score for a probe against a tester genome was divided by 60 (the score for a perfect match of the probe against the reference genome) to obtain the BLAST score ratio (BSR) of the probe against the tester genome, as follows (similar to calculation of the experimental hybridization signal ratio): BSR (%) = $100 \times (\text{BLAST score})_{tester}/60$.

To bioinformatically estimate the expected hybridization signal of a probe that had multiple matching targets in a tester genome, the following procedure was used. First, the multiple matches of a probe that had a BSR less than 83% were removed from the analysis because oligoarray hybridization experiments indicated that probes with BSR less than 83% did not typically provide hybridization signals greater than the background signal (Fig. 1). For the remaining matches of a probe, the bioinformatically expected hybridization signal of an individual match was assumed to equal the experimentally derived average hybridization signal of all single-match probes with the same BSR for the match in question (based on the equation shown in Fig. 1, inset). Finally, the expected signals for each match of a probe were added to obtain the bioinformatically predicted cumulative signal for the probe due to the multiple matches. This procedure was

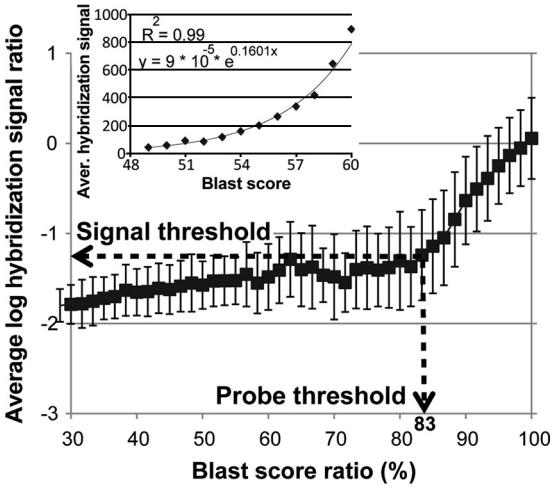

FIG. 1. Relationship between microarray hybridization signal and BLAST score ratio (BSR). Competitive DNA-DNA oligoarray hybridization was performed with *B. vietnamiensis* G4 (tester strain) and *B. cenocepacia* J2315 (reference strain). The experimental log hybridization signal ratios of probes (*y* axis) are plotted against the BSR of the probes (*x* axis); a log ratio of zero indicates that the hybridization signal intensity of G4 is equal to that of J2315. Data points indicate the mean log hybridization ratios of all probes with the same BSR; the error bars indicate one standard deviation from the mean. A total of 6,308 probes were included in the analysis; the number of probes used per unit of BSR is shown in Fig. 3. The dashed arrows indicate bioinformatic probe and experimental signal thresholds corresponding to the point of inflection. (Inset) Linear regression analysis of the mean log hybridization signal (*y* axis) plotted against the BLAST score, defined as described in Materials and Methods, for the same data set and BLAST scores in the range from 50 to 60, which corresponded to BSR of 83 to 100%.

used for both the tester and reference strains, as follows: predicted hybridization signal for individual match = 0.00009 × exp(0.1601 × BLAST score) and predicted logarithmic signal ratio for probe = log(sum of predicted hybridization signals for tester/sum of predicted hybridization signals for reference).

To bioinformatically determine the genes that were shared by the reference genome and a tester genome, the sequences of all reference genes were searched against the whole genome of the tester strain for reciprocal best-match conserved genes. Conserved genes were defined using the following two cutoffs: (i) for nucleotide-level conservation, BLASTN was used for the search with the settings described above for the probe sequences and a minimum cutoff for a match of at least 70% identity and 70% of the length of the reference gene covered in the corresponding alignment; and (ii) for amino-acid-level conservation, BLASTX was used for the search with default settings and a cutoff for a match of at least 50% identity and 70% of the length of the reference gene in the corresponding alignment (a less stringent cutoff).

Signal threshold methods for examining gene presence. Two of the most commonly used methods for determining gene presence based on experimentally derived hybridization signals were compared to our method (see the Results) for the same purposes. In one of these methods (the SNR^1 method) the ratio of the background-subtracted signal divided by the standard deviation of the background signal was used (7, 11, 34): $SNR^1 = (MS - BS) \times SBS^{-1}$, where SNR is the signal-to-noise ratio, MS and BS are the mean probe and background signals, respectively, and SBS is the standard deviation of background signal. In the other method (the SNR^2 method) the ratio of the mean signal to the mean background signal was used, taking the signals of negative controls into account, as described by Loy et al. (22): $SNR^2 = [MS - (MSN - BSN)] \times BS^{-1}$, where MSN and BSN are the mean probe signal and background signal of the negative controls, respectively. The most frequently employed SNR cutoffs for these methods are 2.0 and 3.0 (7, 11, 34); an SNR value of 2.0 was used in this study for both methods.

GACK analysis was performed as described by Kim et al. (14). In brief, the GACK algorithm examines the shape of the experimentally derived hybridization signal ratio distribution of an individual microarray experiment to determine a signal threshold for gene presence or absence based on a normal probability density function (expected distribution) and a user-defined estimated probability of presence (EPP). EPP indicates the number of genes that are expected to be absent (i.e., the number of genes that are expected to not give a hybridization signal greater than the background signal) due to sequence divergence of the tester strain from the reference strain. We employed two values for EPP, 0 and 50, in our microarray experiments, as previously suggested (14). We report here only the results obtained using an EPP value of 50 since this standard was more relevant for our experiments that were performed with strains divergent from the reference strain.

Microarray data accession number. The raw microarray intensity data have been deposited in the GenBank Gene Expression Omnibus (GEO) database under accession number GSE20096.

RESULTS

Relatedness of the tester strains to the reference strain. Our DNA-DNA oligoarray experiments were performed with strains of the genus *Burkholderia* due to the availability of a previously tested 60-mer Agilent oligoarray for an important member of this genus, *B. cenocepacia* J2315 (18), an epidemic isolate from a cystic fibrosis patient (12), and the availability of several sequenced strains that show different levels of divergence from J2315 (Table 1). The 16S rRNA gene sequences of the six *Burkholderia* strains used in this study showed 96 to 99.7% sequence identity, which is consistent with their high level of genetic relatedness (they belong to the same or closely related species). When the genetic relatedness was measured by using genome-aggregate average nucleotide identity (ANI), a more sensitive parameter for measuring evolutionary relatedness of closely related genomes (17), a gradient of genetic relatedness was observed. In particular, strain J2315 showed ~95% ANI to *B. cenocepacia* strains AU1054 and HI2424. These values match the 95% ANI that corresponds to the 70% DNA-DNA hybridization standard frequently used for species demarcation (9). Hence, these pairs of genomes sample the subspecies level. J2315 showed ~91% and 89% ANI to *B. ambifaria* AMMD and *B. vietnamiensis* G4, respectively, which represent different levels of genetic relatedness within the genus. Finally, J2315 showed ~80% ANI to *B. xenovorans* LB400, which represents the most divergent species sampled within the genus. This genetic gradient provided the opportunity to precisely estimate the number of false positives and false negatives in oligoarray DNA-DNA experiments when J2315 (reference strain) was compared with the five tester strains.

Relating the oligoarray signal to probe target sequence relatedness and determining the signal threshold cutoff. To bioinformatically predict the probes that should provide signals greater than the background signal based on their sequence similarity to the target sequences in the genomes of the tester strains, we used the BLAST score-based approach because of its simplicity and satisfactory accuracy (26). We determined the BLAST score for each probe compared with the top matching gene sequence in the whole genome of a tester strain. Subsequently, we calculated the BLAST score ratio (BSR) of the probe (i.e., the BLAST score with the tester divided by the BLAST score with the reference, which was always 60 due to the perfectly matching probe sequence) and plotted it against the experimentally derived oligoarray hybridization signal ratio for the same probe. The analysis was restricted to probes that had only one match in the genome to avoid the complications resulting from multiple matching tar-

gets (see below). Our results revealed that the relationship between the signal ratio and the BSR was clearly biphasic. In particular, a linear dynamic decrease in the oligoarray hybridization signal was observed for BSR of 83 to 100%; at values below the 83% inflection point, the slope of the regression line decreased significantly, and the line became almost flat (Fig. 1). Hence, significant hybridization signals greater than the background signal were obtained only for probes whose BSR with the tester strain (relative to the reference strain) was at least 83%.

The strong linear regression between the signal ratio and the BSR in the 83 to 100% BSR range was attributable to the almost perfect ($R^2 = 0.99$) correlation between the average hybridization signal intensity (based on all available probes) and the BLAST score in the BLAST score range from 50 to 60 (Fig. 1, inset). The latter results also revealed that the hybridization signal can, in general, be predicted robustly from the BLAST score, which is consistent with previous findings (26).

The slope of the regression line for BSR in the 30 to 83% range (Fig. 1) was due to an increase in the hybridization signal from the targets in the reference strain. This is most likely attributable to less competition from the targets of the tester strain for the available probe molecules on the chip in this BSR range due to an increased number of mismatches or no sequence match of the targets with the probe sequence (i.e., little or no hybridization). Hence, more probe molecules were available for the reference targets to bind. These results were reproducible independent of the genetic relatedness of the strain tested to the reference strain.

FP and FN as a function of the relatedness of the tester strain to the reference strain. To evaluate oligoarray performance, we employed a BSR of 83% as the cutoff for predicting bioinformatically which probes should have been expected to provide signals greater than the background signal and the average experimental hybridization signal ratio of all probes that showed a BSR of exactly 83% in each hybridization experiment (e.g., ~−1.3 in the example shown in Fig. 1) as the signal threshold for determining which probes provided hybridization signals greater than the background signal. The latter probes represented the genes that would have been considered present in the tester strain based on the experimental hybridization results and the previous signal threshold. Our evaluation revealed that most (>80%) probes provided experimental hybridization results consistent with the bioinformatic prediction using the standards described above (good probes [GP]). However, the fraction of GP decreased with greater divergence of the tester strain from the reference strain (Fig. 2), and there was a corresponding increase in false positives (FP) (i.e., probes that bioinformatically were not supposed to provide a signal greater than the background signal because they matched the tester genome at BSR less than 83%) (Fig. 1). For instance, GP constituted 90.8%, 88.5%, 84.3%, and 77.4% of all probes for tester strains with 95%, 91%, 89%, and 80% ANI to J2315, while the percentages of FP were 4.7%, 5.4%, 9.1%, and 16.7% for the same strains, respectively. The probes that contributed disproportionally to the pool of FP probes were the probes whose BSR were close to the BSR threshold (83%); in fact, there were very few FP probes among the probes that had a BSR less than about 70% (Fig. 3). The number of false-negative (FN) probes (i.e., probes with BSR

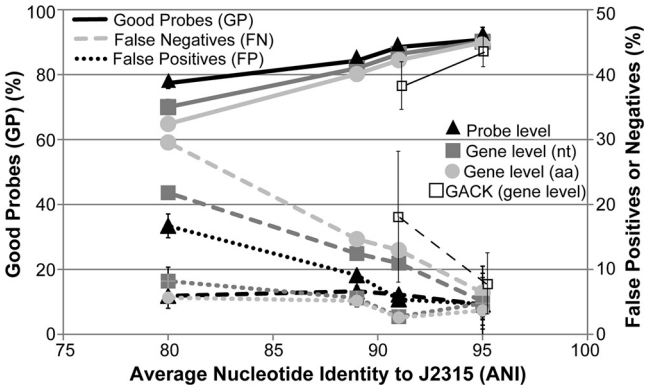

FIG. 2. Assessment of oligoarray performance with increasing relatedness of the tester strain to J2315. The average numbers of good probes (GP), false-positive probes (FP), and false-negative probes (FN) for J2315 and a tester strain (*y* axis) are plotted against the ANI of the tester strain to J2315 (*x* axis). Three independent assessments were performed (see text for details), a probe-level assessment, a nucleotide gene-level assessment (nt), and an amino acid gene-level assessment (aa). GP, FP probes, and FN probes for each assessment level were defined as shown in Table 2. The results obtained with the GACK algorithm (14) and the same data sets are also shown. GACK was performed using an estimated 50% probability of presence (for details, see Materials and Methods). A total of 6,308 probes were analyzed in each microarray hybridization experiment; four to eight experiments, including dye swaps, were performed for each tester strain. The error bars indicate one standard deviation from the mean for the experiments analyzed for each tester strain.

greater than 83% that provided hybridization signals that were less than the background signal) was relatively constant with the ANI of the tester strain, and these probes constituted about 5 to 6% of all of the probes (Fig. 2).

To perform an assessment of oligoarray performance for genomotyping that is more relevant for practical use, we also compared the probes that provided experimental hybridization signals greater than the background signal to the actual genes that were shared by the tester and reference genomes based on whole-genome nucleotide sequence comparisons. We found that the number of GP, defined in this case as the probes that gave hybridization signals greater than the background signal regardless of their BSR and corresponded to reference genes present in the tester genome (each probe is specific for one reference gene), also decreased with decreasing ANI of the tester strain to J2315, similar to the probe-level assessment described above (Table 2 shows the definitions for GP, FP, and FN for each assessment). However, the decrease was more dramatic for the assessment of gene versus probe level, as the percentages of GP in the former assessment were 90.1%, 86.3%, 82.0%, and 70.0% for tester strains according to their ANI rank (Fig. 2). These findings were due to a higher number of FN probes, defined in this case as the probes that corresponded to shared genes based on the bioinformatic sequence analysis and provided experimental hybridization signals less than the background signal regardless of their BSR, in the assessment of gene versus probe level. The increased number of FN probes in the gene-level assessment was attributable mostly to two factors. One of these factors was tester genes whose levels of nucleotide sequence identity to their orthologs in the reference genome were in the range from 70 to 83%;

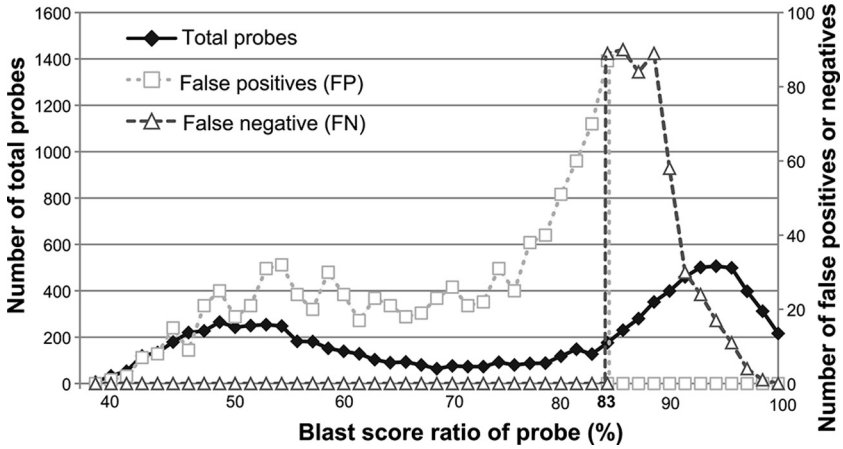

FIG. 3. False positives and false negatives as a function of the degree of sequence similarity between the probe and the target. Competitive DNA-DNA oligoarray hybridization was performed with *B. vietnamiensis* G4 and J2315. The numbers of all probes (*y* axis on the left) that provided false-positive (*y* axis on the right) and false-negative (*y* axis on the right) signals were plotted against the BSR (*x* axis) of the corresponding probes (probe-level assessment; see Table 2 and the text for details). The analysis was based on using a BSR of 83% as the cutoff for bioinformatically predicting the probes that were expected to provide signals greater than the background signal and the average log hybridization signal ratio of all single-match probes for a BSR of 83% as the threshold for identifying the probes that experimentally provided signals greater than the background signal, according to the results shown in Fig. 1. A total of 6,308 probes were included in the analysis. Similar results were obtained with other tester strains.

hence, these genes were considered genes that were shared by the tester and reference genomes (since the level of identity was more than the cutoff for considering a gene present), but the corresponding probe typically provided a hybridization signal less than the background signal (since it had <83% nucleotide identity to the target in the tester genome). This type of FN probes was more abundant for comparisons with distantly related genomes (i.e., genomes with more orthologs with sequence identities in the 70 to 83% range). The second factor was the fact that the FN probes for the gene-level assessment included most, if not all, of the FN probes for the probe-level assessment (i.e., probes the showed BSR of >83% and provided signals less than the background signal). These findings also suggest that DNA-DNA microarray studies can underestimate substantially the actual levels of gene content similarity between the genomes compared. In contrast, the number of FP probes was slightly lower for the assessment of gene versus probe. This was due to the fact that several FP probes for the probe-level assessment (i.e., probes that showed BSR less than 83% and provided hybridization signals greater than the background signal) corresponded to J2315 genes that were actually shared by the tester genome. Such probes were included in the GP category for the gene-level assessment instead of the FP category for the probe-level assessment; thus, the number of FP probes in the gene-level assessment was typically less than the number of FP probes in the probe-level assessment.

When shared genes were assessed at the amino acid level (which provided a less stringent but probably more realistic estimate of shared protein-encoding genes), an even smaller number of GP with decreasing ANI of the tester strain to J2315 was observed. For instance, the percentages of GP were 89.9%, 84.4%, 80.2%, and 64.8% with decreasing ANI at the amino acid level, compared with 90.8%, 88.5%, 84.3%, and 77.4% at the nucleotide gene level, respectively (Fig. 2). The latter results were attributable to the fact that the amino acid level was more conserved than the nucleotide level; thus, more shared genes were found for J2315 and the tester strains in the amino acid comparisons, and more genes in this gene set had related nucleotide sequences with BSR values below the 83% cutoff (contributing FN probes). Finally, the FN and FP probes in the gene- or probe-level assessments were typically tester strain

TABLE 2. Definitions of good probes, false-negative probes, and false-positive probes for the assessments performed

Case	Results for the following conditions used to determine gene presence[a]:				Probe defined by bioinformatics and exptl signal		
	Bioinformatics			Exptl signal			
	Amino acid	Nucleotide (gene)	Nucleotide (probe)		Amino acid	Nucleotide (gene)	Nucleotide (probe)
1	1	1	1	1	GP	GP	GP
2	1	1	1	0	FN	FN	FN
3	1	1	0	1	GP	GP	FP
4	1	1	0	0	FN	FN	GP
5	1	0	1	1	GP	FP	GP
6	1	0	1	0	FN	GP	FN
7	1	0	0	1	GP	FP	FP
8	1	0	0	0	FN	GP	GP
9	0	1	1	1	FP	GP	GP
10	0	1	1	0	GP	FN	FN
11	0	1	0	1	FP	GP	FP
12	0	1	0	0	GP	FN	GP
13	0	0	1	1	FP	FP	GP
14	0	0	1	0	GP	GP	FN
15	0	0	0	1	FP	FP	FP
16	0	0	0	0	GP	GP	GP

[a] 1, present (or signal greater than the background signal) based on the cutoffs used (see Materials and Methods); 0, absent (or signal less than the background signal) based on the cutoffs used (see Materials and Methods). For example, in case 3, the bioinformatic assessment indicated that the target gene (i.e., the ortholog in the corresponding reference strain) was present in the tester genome at the gene level at both the amino acid and the nucleotide levels, but the BSR of the probe with the target gene sequence was less than 83%. Accordingly, since the experimental hybridization signal for the probe was greater than the background signal (as indicated by the experimental signal entry), the probe was defined as a GP for the gene-level assessments but as an FP probe for the probe-level assessment.

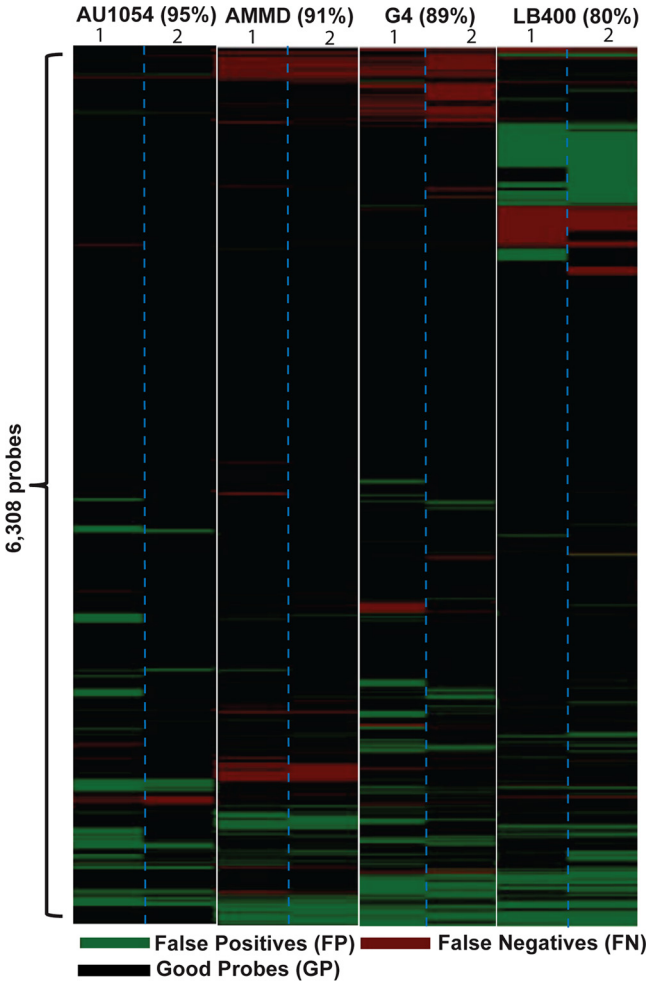

FIG. 4. False-positive and false-negative probes are typically tester strain specific. A heat map shows an overview of the good (black), false-positive (green), and false-negative (red) probes for four tester strains with increasing ANI to J2315. The results for a total of 6,308 probes are represented by horizontal lines. Data for two independent (i.e., not dye swap) competitive microarray hybridizations for each tester strain are shown in columns 1 and 2.

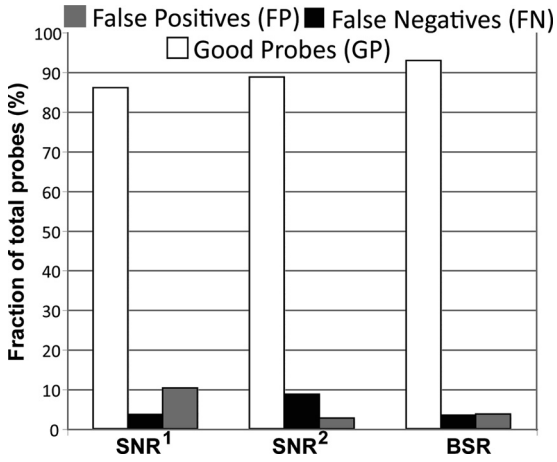

FIG. 5. Comparison of different methods for estimating the signal threshold ratio. The presence of a J2315 gene in *B. cenocepacia* HI2424 was determined based on the experimental hybridization signal ratio of the corresponding probe using three different methods for defining the signal threshold ratio for gene presence (see Materials and Methods for details), the SNR^1 (7, 11, 34), SNR^2 (22), and BSR (this study) methods. The genes were then checked against the genes shared by J2315 and HI2424 based on bioinformatic whole-genome sequence comparisons (nucleotide gene-level assessment), and the results are shown. The bars indicate the average values for two microarray experiments (dye swaps).

specific, although a few probes consistently showed FP or FN signals independent of the tester genome used (Fig. 4). However, the latter probes were frequently attributed to specific characteristics of the corresponding genes. For instance, the probes that systematically yielded FN signals were frequently associated with genes that evolved faster than the average gene in the genome, such as genes encoding membrane-associated proteins, and thus showed lower levels of nucleotide sequence identity to their J2315 orthologs than to the rest of the genes.

Methods for determining the signal threshold for gene presence. Two of the most popular approaches for defining the presence of genes based on the difference between the probe hybridization signal and the background signal (22) revealed trends similar to those obtained with our method based on the average signal for probes showing a BSR of exactly 83% (data not shown). However, the two former methods performed substantially worse than our BSR method based on two independent experiments with strain HI2424 (~95% ANI to J2315). In particular, the number of FP and FN probes did not exceed 3.7% of all of total probes based on our method, while the FN and FP probes accounted for 8.7% and 10.3% of all of the probes based on the SNR^2 and SNR^1 methods, respectively (Fig. 5). Also, the percentages of GP were 86.0%, 88.7%, and 92.9% as determined by the SNR^1 and SNR^2 methods and our method, respectively. Thus, our approach clearly outperformed the approaches described previously. Using a threshold different than 83% is unlikely to further improve the processing of microarray data in practice. For instance, increasing the BSR threshold (i.e., being more stringent) resulted in a disproportionally increased number of FN probes and a proportionally decreased number of FP probes, while decreasing the threshold had the reverse effect, as suggested previously (11).

We also compared the results of our method to those of a tool commonly used in bacterial genomotyping, GACK (14). GACK does not employ an *a priori* determined signal threshold for gene presence (as SNR methods do) but, similar to our method, determines the threshold based on the shape of the signal ratio distribution for each microarray experiment individually (14). Our evaluation showed that GACK provided slightly worse results than our method for tester strains showing 95% ANI to the reference strain; e.g., the number of GP was smaller, while the performance declined dramatically for strains showing ~90% ANI to the reference strain. In the latter case, the percentage of GP for GACK was only ~75%, compared with ~88% for our method (Fig. 2). The decreased performance of GACK was also reflected by the greater variation (Fig. 2) in the number of GP or FN probes for different oligoarray experiments with the same tester strain (dye swaps). We were not able to perform GACK with more divergent tester strains because GACK requires the shape of the signal ratio distribution to approximate normal in order to determine

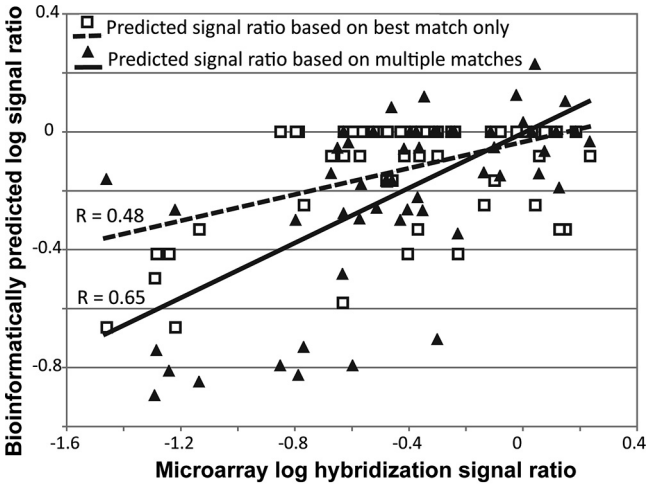

FIG. 6. Modeling the effect of multiple matching targets to the total hybridization signal of a probe. For 55 probes there were two or more matches in the *B. vietnamiensis* G4 genome with a BSR of 83% or higher. For each match of a probe, we calculated the predicted microarray hybridization signal based on the relationship shown in the inset in Fig. 1 (see Materials and Methods for details), and the predicted signals were added to obtain a bioinformatic estimate of the total predicted signal for the probe. The total predicted signal (expressed as log ratio on the *y* axis) is plotted against the actual experimental hybridization signal ratio of the probe (*x* axis). Much better Pearson correlation (R) between predicted and actual signal ratios was observed when the predicted signal was calculated using all multiple matches of a probe than when the predicted signal was calculated using only the best match.

the signal threshold and such a shape was not observed for these strains (because the majority of the hybridization signals were too similar to the background signal). These findings were also consistent with findings reported previously for divergent tester strains (14). Thus, GACK is not appropriate for tester strains that show less than ~95% ANI to the reference strain (i.e., strains in different species).

Identifying gene duplications based on microarray signals. To assess the contribution of multiple matching targets to the hybridization signal of a probe (cross-hybridization) and to determine whether a higher hybridization signal can be used to identify duplicated genes in the tester genome compared to the reference genome (or *vice versa*), the following analysis was performed. We identified the J2315 probes that matched two or more genes in the *B. vietnamiensis* G4 genome with BSR of ≥83%. Fifty-five probes in the total pool of about 6,308 probes met these criteria. Subsequently, we bioinformatically predicted the expected hybridization signals for these probes in the tester and reference genomes, as described in Materials and Methods, and compared the resulting ratios to the actual, experimentally derived hybridization signal ratios for the same probes. The comparisons indicated that the experimental hybridization signal ratio correlated significantly better with the predicted ratio based on all matches of a probe than with the predicted signal based on only the best match (Pearson correlation coefficients, 0.65 and 0.48, respectively) (Fig. 6). Hence, a higher hybridization signal may indicate multiple matching targets (e.g., duplicated genes) in the tester strain compared to the reference strain, and the equations described here could be used to identify such duplicated genes.

DISCUSSION

Oligoarrays represent a mature technology; however, all issues associated with the broad applications of oligoarrays have not been fully elucidated, as a plethora of recent studies have indicated (25, 27, 29). In this study, we used complete genomic sequences of several tester strains to validate microarray results and to obtain quantitative assessments of the performance of oligoarrays for whole-genome DNA-DNA competitive hybridization. Our findings reveal that the numbers of microarray FP and FN probes are not negligible even for tester strains of the same species; e.g., they may constitute >5% of all probes. Our findings also provide robust estimates of the number of FP and FN that should be expected based on the degree of genetic relatedness of the tester strain to the reference strain (Fig. 2). Our analysis showed that the performance of oligoarrays decreased gradually (as opposed to abruptly) with decreasing genetic relatedness of the tester strains in the 80 to 100% ANI range. Accordingly, some of the hybridization signal data, particularly the absence of a signal, which indicated that the corresponding gene was not shared, were reliable even with highly divergent tester strains (e.g., strains showing ~80% ANI to the reference strain). As a rule of thumb, however, we recommend that tester strains should not exhibit less than about 90% ANI to the reference strain, if it is expected that at least 90% of the probes will provide reliable signals (GP) (Fig. 2). We also did not observe any apparent biases in the number of FP or FN probes among the probes for core and noncore genes in the genome ("core" denotes genes shared by all genomes in a group), except for a few predictable genes. For instance, informational genes (e.g., genes encoding ribosomal proteins, DNA and RNA polymerases, etc.), which tend to show higher levels of nucleotide sequence conservation than the genome average level (15), had more GP and fewer FN probes than genes that tend to evolve faster than the average gene, such as genes encoding membrane and sensory proteins, transposases, and hypothetical proteins.

Notably, FP results were observed even with probes that had very low BSR (e.g., <70%) with the target sequence, and these probes accounted for about one-half of all FP probes (despite the fact that FP probes with BSR closer to 83% were more frequent [Fig. 3]). The former FP probes were likely attributable to nonspecific or nonorthologous target sequences that matched the probe sequences in short but identical segments, as hypothesized previously (11). However, we were unable to come up with simple and reliable rules for sequence similarity or the position of internal mismatches in the probe-target sequence pair that underlie the behavior of such FP probes since the target sequences that hybridized to the probes were not known. Further, our evaluations indicated that predicting *a priori* which probes are likely to provide FN (or FP) signals is not generally feasible, since these probes were typically tester strain specific (Fig. 4). Thus, it should be taken for granted that in DNA-DNA experiments and in expression experiments with strains other than the reference strain, a portion of the hybridization signal, which depends primarily on the degree of genetic relatedness of the tester strain to the reference strain

(Fig. 2), cannot be trusted to be a reliable proxy for gene presence and activity, respectively.

Traditional approaches for determining gene presence based on DNA-DNA microarray experiments have typically employed thresholds for the probe hybridization signal based on the background hybridization signal (20, 22) or based on the shape of the signal distribution for all probes (14). Although these approaches are intuitive and easy to use, they are based on arbitrary thresholds and/or are vulnerable to variable biases resulting from slide heterogeneities, including unequal DNA templates, different amounts of incorporated dye, uneven labeling and hybridization conditions etc., and from the degree of genetic divergence of the tester strains (5, 28). In contrast to these traditional methods, the BLAST score-based method developed in this study employs slide-specific thresholds based on the dynamic hybridization range of probes (i.e., the average signal corresponding to the inflection point [Fig. 1]) and outperforms the previously described approaches (Fig. 2 and 5). For the Agilent oligoarrays used in this study, the dynamic hybridization range corresponded consistently to BSR of 83 to 100%, regardless of the tester strain used (data not shown), while the average hybridization signal of probes with a BSR of 83% (i.e., the signal threshold for determining gene presence) varied slightly in different hybridization experiments, depending primarily on the dye incorporation efficiency and the amount of DNA labeled for each dye. The BSR of ~83% is similar to the 85% nucleotide sequence identity cutoff proposed previously for determining the 50-mer probes that were expected to cross-hybridize with the target sequences (11).

To apply our approaches to nonsequenced tester strains, we recommend obtaining the sequences of at least a few genes in the genome of each strain (e.g., genes that are sequenced as part of multilocus sequence typing [MLST]). These sequences could be used to determine the relationship between hybridization signal and BSR, as shown in Fig. 1. The curve of this relationship can subsequently be employed to make educated predictions about the signal threshold that corresponds to the inflection point. It can also help normalize the hybridization signal across different oligoarray experiments based on the shape of the curve for each experiment (e.g., by applying a normalization factor that would bring the inflection point to the same signal and BSR values for each slide). Our previous work also provided a means to estimate ANI values for any two strains for which MLST data are available (16).

Gene duplication and independent acquisition of highly similar genes are important evolutionary processes in bacteria and lower eukaryotes, and they frequently reflect ecological adaptation of a lineage, such as increased virulence. However, using competitive DNA-DNA oligoarray experiments to identify such "duplicated" genes in the tester strain compared to the reference strain or *vice versa* is technically challenging. We found that duplicated genes typically resulted in increased hybridization signals (Fig. 6); thus, an increased signal may in fact indicate duplicated genes. While the trend line shown in Fig. 6 would be useful for identifying such candidate genes, further improvements in modeling the expected hybridization signal based on sequence similarity beyond what is encompassed by the BSR (25) and/or employing several probes per gene are necessary for more accurate predictions.

Although we tested only one type of oligoarray platform, our preliminary work with other platforms, such as the oligo-spotted MWG arrays for *Escherichia coli* (MWG Biotech) and a custom-made 50-mer *in situ*-synthesized oligoarray from Biodiscoveries LLC (Ann Arbor, MI), provided results similar to those obtained with the Agilent oligoarrays (our unpublished observations). Nonetheless, small variations in performance between the different platforms were observed, and *in situ* oligoarrays typically performed better than spotted oligoarrays and had a clearer dynamic hybridization range. Therefore, our results provide useful practical guidelines for designing microarray experiments for a wide range of microarray platforms. Further, the genus *Burkholderia* targeted by our Agilent oligoarrays is one of the most metabolically versatile bacterial genera known, and its members have some of the largest genomes (~8 Mbp); it also includes both clinical (e.g., J2315) and nonclinical representatives involved in cycling organic matter in soils and sediments and controlling fungal diseases. Therefore, our findings obtained with *Burkholderia* strains have important implications for studying additional members of this important genus with oligoarrays and should be applicable to other bacterial groups as well. While bacterial genome sequencing is becoming less costly, genomotyping with microarrays will likely remain cost-effective for larger-scale strain comparisons and for environmental surveys of genes of complex microbial communities for some time, since adequately covering all members of a community with sequencing remains out of reach for current sequencing technologies.

ACKNOWLEDGMENTS

We thank Erick Cardenas and two anonymous reviewers for helpful discussions regarding the manuscript. We thank the Sanger Institute and the Joint Genome Institute for giving us early access to the genome sequences used in this study.

This work was supported by the NSF (award DEB-0516252 to J.T. and K.T.K.), the U.S. Department of Energy (award DE-FG02-07ER64389 to J.T. and K.T.K.), and the Cystic Fibrosis Foundation Therapeutics.

REFERENCES

1. **Agilent Technologies, Inc.** 2007. Oligonucleotide array-based CGH for genomic DNA analysis, version 2.0 ed. Agilent Technologies, Inc., Santa Clara, CA.
2. **Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman.** 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25:**3389–3402.
3. **Behr, M. A., M. A. Wilson, W. P. Gill, H. Salamon, G. K. Schoolnik, S. Rane, and P. M. Small.** 1999. Comparative genomics of BCG vaccines by whole-genome DNA microarray. Science **284:**1520–1523.
4. **Bhaya, D., A. R. Grossman, A. S. Steunou, N. Khuri, F. M. Cohan, N. Hamamura, M. C. Melendrez, M. M. Bateson, D. M. Ward, and J. F. Heidelberg.** 2007. Population level functional diversity in a microbial community revealed by comparative genomic and metagenomic analyses. ISME J. **1:**703–713.
5. **Bilban, M., L. K. Buehler, S. Head, G. Desoye, and V. Quaranta.** 2002. Defining signal thresholds in DNA microarrays: exemplary application for invasive cancer. BMC Genomics **3:**19.
6. **Bozdech, Z., J. Zhu, M. P. Joachimiak, F. E. Cohen, B. Pulliam, and J. L. DeRisi.** 2003. Expression profiling of the schizont and trophozoite stages of Plasmodium falciparum with a long-oligonucleotide microarray. Genome Biol. **4:**R9.
7. **Coombes, K. R., J. Wang, and L. V. Abruzzo.** 2004. Monitoring the quality of microarray experiments, p. 159–163. *In* F. Johnson and S. M. Lin (ed.), Methods of microarray data analysis, vol. 3. Kluwer Academic Publishers, Norwell, MA.
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. **Goris, J., K. T. Konstantinidis, J. A. Klappenbach, T. Coenye, P. Vandamme, and J. M. Tiedje.** 2007. DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. Int. J. Syst. Evol. Microbiol. **57:**81–91.
10. **He, Z., L. Wu, X. Li, M. W. Fields, and J. Zhou.** 2005. Empirical establishment of oligonucleotide probe design criteria. Appl. Environ. Microbiol. **71:**3753–3760.
11. **He, Z., and J. Zhou.** 2008. Empirical evaluation of a new method for calculating signal-to-noise ratio for microarray data analysis. Appl. Environ. Microbiol. **74:**2957–2966.
12. **Holden, M. T., H. M. Seth-Smith, L. C. Crossman, M. Sebaihia, S. D. Bentley, A. M. Cerdeno-Tarraga, N. R. Thomson, N. Bason, M. A. Quail, S. Sharp, I. Cherevach, C. Churcher, I. Goodhead, H. Hauser, N. Holroyd, K. Mungall, P. Scott, D. Walker, B. White, H. Rose, P. Iversen, D. Mil-Homens, E. P. Rocha, A. M. Fialho, A. Baldwin, C. Dowson, B. G. Barrell, J. R. Govan, P. Vandamme, C. A. Hart, E. Mahenthiralingam, and J. Parkhill.** 2009. The genome of *Burkholderia cenocepacia* J2315, an epidemic pathogen of cystic fibrosis patients. J. Bacteriol. **191:**261–277.
13. **Kane, M. D., T. A. Jatkoe, C. R. Stumpf, J. Lu, J. D. Thomas, and S. J. Madore.** 2000. Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. Nucleic Acids Res. **28:**4552–4557.
14. **Kim, C. C., E. A. Joyce, K. Chan, and S. Falkow.** 2002. Improved analytical methods for microarray-based genome-composition analysis. Genome Biol. **3:**RESEARCH0065.
15. **Konstantinidis, K. T., and E. F. DeLong.** 2008. Genomic patterns of recombination, clonal divergence and environment in marine microbial populations. ISME J. **2:**1052–1065.
16. **Konstantinidis, K. T., A. Ramette, and J. M. Tiedje.** 2006. Toward a more robust assessment of intraspecies diversity, using fewer genetic markers. Appl. Environ. Microbiol. **72:**7286–7293.
17. **Konstantinidis, K. T., and J. M. Tiedje.** 2005. Genomic insights that advance the species definition for prokaryotes. Proc. Natl. Acad. Sci. U. S. A. **102:** 2567–2572.
18. **Leiske, D. L., A. Karimpour-Fard, P. S. Hume, B. D. Fairbanks, and R. T. Gill.** 2006. A comparison of alternative 60-mer probe designs in an in-situ synthesized oligonucleotide microarray. BMC Genomics **7:**72.
19. **Li, F., and G. D. Stormo.** 2001. Selection of optimal DNA oligos for gene expression arrays. Bioinformatics **17:**1067–1076.
20. **Li, X. M., J. Kim, J. Zhou, W. K. Gu, and R. Quigg.** 2005. Use of signal thresholds to determine significant changes in microarray data analyses. Genet. Mol. Biol. **28:**191–200.
21. **Liebich, J., C. W. Schadt, S. C. Chong, Z. He, S. K. Rhee, and J. Zhou.** 2006. Improvement of oligonucleotide probe design criteria for functional gene microarrays in environmental applications. Appl. Environ. Microbiol. **72:** 1688–1691.
22. **Loy, A., A. Lehner, N. Lee, J. Adamczyk, H. Meier, J. Ernst, K. H. Schleifer, and M. Wagner.** 2002. Oligonucleotide microarray for 16S rRNA gene-based detection of all recognized lineages of sulfate-reducing prokaryotes in the environment. Appl. Environ. Microbiol. **68:**5064–5081.
23. **Lucito, R., J. Healy, J. Alexander, A. Reiner, D. Esposito, M. Chi, L. Rodgers, A. Brady, J. Sebat, J. Troge, J. A. West, S. Rostan, K. C. Nguyen, S. Powers, K. Q. Ye, A. Olshen, E. Venkatraman, L. Norton, and M. Wigler.** 2003. Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation. Genome Res. **13:** 2291–2305.
24. **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. U. S. A. **98:**9853–9858.
25. **Naiser, T., J. Kayser, T. Mai, W. Michel, and A. Ott.** 2008. Position dependent mismatch discrimination on DNA microarrays—experiments and model. BMC Bioinformatics **9:**509.
26. **Palmer, C., E. M. Bik, M. B. Eisen, P. B. Eckburg, T. R. Sana, P. K. Wolber, D. A. Relman, and P. O. Brown.** 2006. Rapid quantitative profiling of complex microbial populations. Nucleic Acids Res. **34:**e5.
27. **Pritchard, L., H. Liu, C. Booth, E. Douglas, P. Francois, J. Schrenzel, P. E. Hedley, P. R. Birch, and I. K. Toth.** 2009. Microarray comparative genomic hybridisation analysis incorporating genomic organisation, and application to enterobacterial plant pathogens. PLoS Comput. Biol. **5:**e1000473.
28. **Quackenbush, J.** 2002. Microarray data normalization and transformation. Nat. Genet. **32:**496–501.
29. **Rich, V. I., K. Konstantinidis, and E. F. DeLong.** 2008. Design and testing of 'genome-proxy' microarrays to profile marine microbial communities. Environ. Microbiol. **10:**506–521.
30. **Rouillard, J. M., M. Zuker, and E. Gulari.** 2003. OligoArray 2.0: design of oligonucleotide probes for DNA microarrays using a thermodynamic approach. Nucleic Acids Res. **31:**3057–3062.
31. **Rusch, D. B., A. L. Halpern, G. Sutton, K. B. Heidelberg, S. Williamson, S. Yooseph, D. Wu, J. A. Eisen, J. M. Hoffman, K. Remington, K. Beeson, B. Tran, H. Smith, H. Baden-Tillson, C. Stewart, J. Thorpe, J. Freeman, C. Andrews-Pfannkoch, J. E. Venter, K. Li, S. Kravitz, J. F. Heidelberg, T. Utterback, Y. H. Rogers, L. I. Falcon, V. Souza, G. Bonilla-Rosso, L. E. Eguiarte, D. M. Karl, S. Sathyendranath, T. Platt, E. Bermingham, V. Gallardo, G. Tamayo-Castillo, M. R. Ferrari, R. L. Strausberg, K. Nealson, R. Friedman, M. Frazier, and J. C. Venter.** 2007. The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. PLoS Biol. **5:**e77.
32. **Wang, F., H. Zhou, J. Meng, X. Peng, L. Jiang, P. Sun, C. Zhang, J. D. Van Nostrand, Y. Deng, Z. He, L. Wu, J. Zhou, and X. Xiao.** 2009. GeoChip-based analysis of metabolic diversity of microbial communities at the Juan de Fuca Ridge hydrothermal vent. Proc. Natl. Acad. Sci. U. S. A. **106:**4840–4845.
33. **Wick, L. M., W. Qi, D. W. Lacher, and T. S. Whittam.** 2005. Evolution of genomic content in the stepwise emergence of *Escherichia coli* O157:H7. J. Bacteriol. **187:**1783–1791.
34. **Zhou, J., and D. K. Thompson.** 2004. DNA microarray technology, p. 141–176. *In* J. Zhou, D. K. Thompson, Y. Xu, and J. M. Tiedje (ed.), Microbial functional genomics. John Wiley & Sons, Hoboken, NJ.