APPLIED AND ENVIRONMENTAL MICROBIOLOGY, Feb. 2006, p. 1688–1691 Vol. 72, No. 2
0099-2240/06/$08.00+0 doi:10.1128/AEM.72.2.1688–1691.2006

Improvement of Oligonucleotide Probe Design Criteria for Functional Gene Microarrays in Environmental Applications†

Jost Liebich,[1,2] Christopher W. Schadt,[1] Song C. Chong,[1] Zhili He,[1] Sung-Keun Rhee,[1,3] and Jizhong Zhou[1]*

Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831[1]; Agrosphere Institute (ICG IV), Forschungszentrum Jülich GmbH, 52425 Jülich, Germany[2]; and Department of Microbiology, Chungbuk National University, Cheungju 361-763, Korea[3]

Received 9 June 2005/Accepted 18 November 2005

To optimize oligonucleotide probe design criteria, PCR products with different similarities to probes were hybridized to a functional gene microarray designed to detect homologous genes from different organisms. In contrast to more restrictive probe designs based on a single criterion, simultaneous consideration of the percent similarity (≤90%), the length of identical sequence stretches (≤20 bases), and the binding free energy (≥−35 kcal mol^{-1}) was found to be predictive of probe specificity.

The application of microarrays to environmental samples presents many technical challenges not encountered in the study of model laboratory cultures (2, 11, 12, 14, 15, 16, 18). One of the most challenging problems results from the need to specifically and reliably distinguish between homologous genes from many different organisms that may share a high degree of sequence similarity. Apart from the stringency of the hybridization conditions, hybridization specificities may be affected by a variety of probe design factors, including the overall sequence similarity, the distribution and positions of mismatching bases (5, 6), and the amount of free energy of the DNA duplexes formed by the probe and target sequences (4, 8, 12). However, most probe design programs and strategies rely on only one or two of the factors mentioned above to assess probe specificity. This may be satisfactory for the design of probes for pure culture studies of global gene expression patterns, as rather divergent genes of a single organism are represented on an array. However, for microarrays targeting homologous genes from diverse microbial communities, it is necessary to design and maximize the number of oligonucleotide probes originating from a set of highly similar sequences. By simultaneous consideration of multiple probe-target characteristics, it is possible to relax each single criterion while also ensuring more accurate predictions of probe-target hybridization behavior.

Microarray design and experiment. PCR products from selected genes obtained from environmental clone libraries (9, 10) were labeled with fluorescent dyes and hybridized to a functional gene microarray containing 50-mer oligonucleotides. The PCR products chosen had different overall similarities (>85%, according to base-to-base comparisons after pairwise alignment) and/or shared identical sequence stretches of >15 bases for multiple probes. The oligonucleotide microarray used in this study followed the design of Rhee et al. (11). Briefly, probes were designed from sequence information for dissimilatory sulfite reductase genes (*dsrA* and *dsrB*), nitrite reductase genes (*nirS* and *nirK*), and an ammonium monooxygenase gene (*amoA*), which were downloaded from the NCBI database or from our own clone libraries using a modified version of PRIMEGENS software (17). Based on global optimal alignments, segments of 50 bases which had <85% nucleotide identity to the corresponding aligned regions of any of the BLAST hit sequences were selected as potential probes, also with consideration of the predicted probe-target melting temperature and probe self-complementarities. For these tests, a set of probe-target combinations with different similarities to several gene sequences was selected in order to cover a broad range of possible probe-target characteristics, such as the length of continuous stretches of matching bases (Fig. 1) and the predicted free energy of hybridization. The oligonucleotide probes were synthesized by MWG Biotech (High Point, NC) and were printed in duplicate spots (13) onto aminopropyl silane-coated glass slides (UltraGAPS; Corning, Corning, NY). The cross-hybridization patterns of 19 selected genes obtained from the clone library collection at Oak Ridge National Laboratory (9, 10) were analyzed. In total, we analyzed 516 specific and nonspecific potential hybridizations of target DNAs to probes, with sequence similarities between <86% and 100% or with a common identical sequence stretch of at least 16 bases. This was done in 71 hybridization experiments and resulted in 1,681 single observations, with two technical replicates (in duplicate spots) of each probe (see the supplemental material for further details).

DNAs from environmental clones were amplified either with the vector-specific universal primer set M13F (−20)/M13R (*nirS*, *nirK*, and *amoA* genes) or with the *dsr*-specific primer pair SR946F/SR2352R (7), and PCR products (25 to 200 pg) were indirectly labeled with Cy5- or Cy3-dUTP (Amersham Pharmacia, Piscataway, NJ) and subsequently hybridized for 16 h at 50°C in 50% formamide to the microarray according to the protocol of Schadt et al. (13). The slides were immersed briefly in buffer 1 (2× SSC [1× SSC is 0.15 M NaCl plus 0.015 M sodium citrate], 0.1% sodium dodecyl sulfate, 50°C), washed for 5 min in fresh buffer 1, for 10 min in buffer 2 (0.1× SSC,

* Corresponding author. Present address: Institute for Environmental Genomics, University of Oklahoma, Stephenson Research & Technology Center, 101 David L. Boren Blvd., Norman, OK 73019. Phone: (405) 325-6073. Fax: (405) 325-3442. E-mail: jzhou@ou.edu.

† Supplemental material for this article may be found at http://aem.asm.org/.

```
CTGATCGTGCACTCGCAGGCCAAC CGCGACAGCCGTCCGCATCTGATC GGCGGTCATGGCG   NKTT11_target
-----------CTCGCAGGCtAAC CGCGACAcCCGcCCGCAcCTGATC GGCGGcCATGGCG   probe M305039
tTGATCGTGCAtTCGCAGGCCAAt CGCGACAGCCGTCCGCATCTGATC aa-----------   probe NKFF17
```

FIG. 1. Illustration of the "stretch" problem (cross-hybridization to probes with low similarities to the target sequence but with an identical sequence stretch of >20 bases). NKTT11_target is a partial *nirK* sequence from our clone library. Upon hybridization with two probes, M305039 and NKFF17, only the latter resulted in positive signals, although both probes have an overall similarity of 90% and similar free binding energies for hybridization to the target sequence NKTT11 (−61.2 kcal mol^{-1} and −66.4 kcal mol^{-1}, respectively). However, NKFF17 and NKTT11 share an identical sequence stretch of 24 bases.

0.1% sodium dodecyl sulfate, room temperature), four times for 1 min each in buffer 3 (0.1× SSC, room temperature), and for 1 min in buffer 4 (0.01× SSC, room temperature), and immediately air dried using compressed air.

Microarrays were scanned with a ScanArray Express (Perkin-Elmer Life Sciences, Boston, MA) at 570 nm (Cy3) or 670 nm (Cy5) with a pixel resolution of 10 μm. The slides were scanned with 100% laser power and a photo multiplier tube gain of 80% or higher. The scanned images were saved as 16-bit TIFF files, and each spot was quantified using ImaGene 5.0 (Biodiscovery, El Segundo, CA). From the mean signal intensities, the signal-to-noise ratio (SNR) was calculated for each spot according to the following formula: SNR = (signal intensity − local background intensity)/standard deviation of the background. Signals with SNRs of >2 were considered positive hybridizations and further verified by visual inspection. Unless otherwise noted, data presented are mean values calculated from the two technical replicates on each slide. For unexpected hybridizations (cross-hybridizations) between the target and a mismatch probe, the free energy change of the probe-target hybrids was calculated after alignment (1), using the web-based Mfold software for nucleic acid folding and hybridization prediction (19), with the following theoretical hybridization conditions: 1 μM template DNA, 1 M Na^+, 1 mM Mg^{2+} at 37°C.

Effects of similarity, identical sequence stretches, and free energy on hybridization specificity. The overall similarities between probe and target sequences had a clear impact on hybridization specificities under the described experimental conditions (Fig. 2). From these data, a logarithmic relationship between probe-target similarity and the percentage of hybridized probes could be established, as follows: $f(x) = -44.47\ln(x) + 103.09$ ($R^2 = 0.9303$). The percentage of probes that gave positive hybridization signals was considerably lower for probes with up to 90% similarity to the target sequences (~10%) than for those with at least 92% similarity (~40%) (Fig. 2), suggesting that a 90% similarity of probes to target sequences could be an appropriate cutoff value for probe design. However, 8.9% (27 of 302) of the probe-target combinations tested with overall similarities of ≤85% still had positive hybridization signals with SNRs of >2, and 6.3% (19 probe-target combinations) were positive with SNRs of >3. The average SNRs of these hybridizations were 10.9 and 14.9, respectively, which are far above the background intensity. These results indicated that probe design criteria that use sequence similarity alone could not reliably eliminate all nonspecific probes.

When the data were analyzed in relation to the length of a common identical sequence stretch, a similar relationship to that for percent overall similarity was found (Fig. 3). The percentage of probes that gave positive hybridization signals was considerably lower for probes with up to 20-bp identical sequence stretches to the target sequences (~10%) than those with at least 22-bp identical sequence stretches (~25%) (Fig. 3), suggesting that a 20-bp identical sequence stretch could be an appropriate cutoff value for probe design. Interestingly, considerable differences in the percentages of probes giving positive hybridization signals were observed between the probes with <35-bp identical sequence stretches and those with longer identical sequence stretches (Fig. 3). He et al. (3) recently used similar findings to establish group probes for sequences with high similarities. However, 3.4% (8 of 236) of the probes still hybridized to targets sharing a <20-base stretch. If the more common threshold of SNRs of >3 was used to determine positive hybridizations, there were still four probes (1.7%) that showed cross-hybridization.

Although most unspecific hybridizations can be eliminated

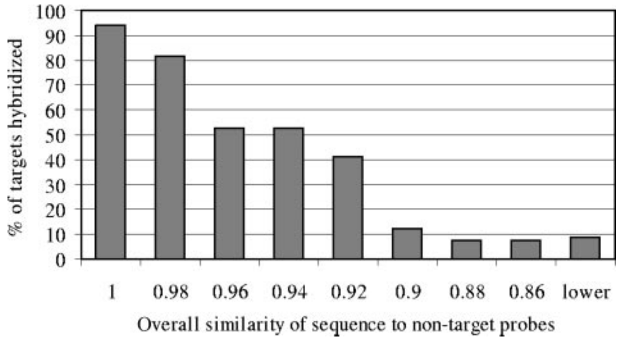

FIG. 2. Relationship between percentage of observed hybridizations with SNRs of >2 and overall sequence similarity for the 50-mer oligonucleotide probes tested in this study.

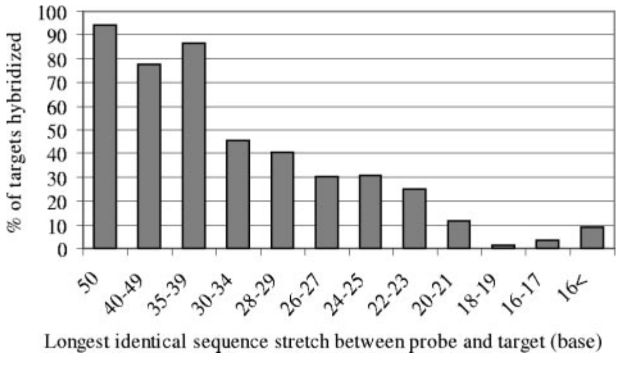

FIG. 3. Relationship between percentage of observed hybridizations with SNRs of >2 and length of the longest identical stretch of base pairs for the 50-mer oligonucleotide probes tested in this study.

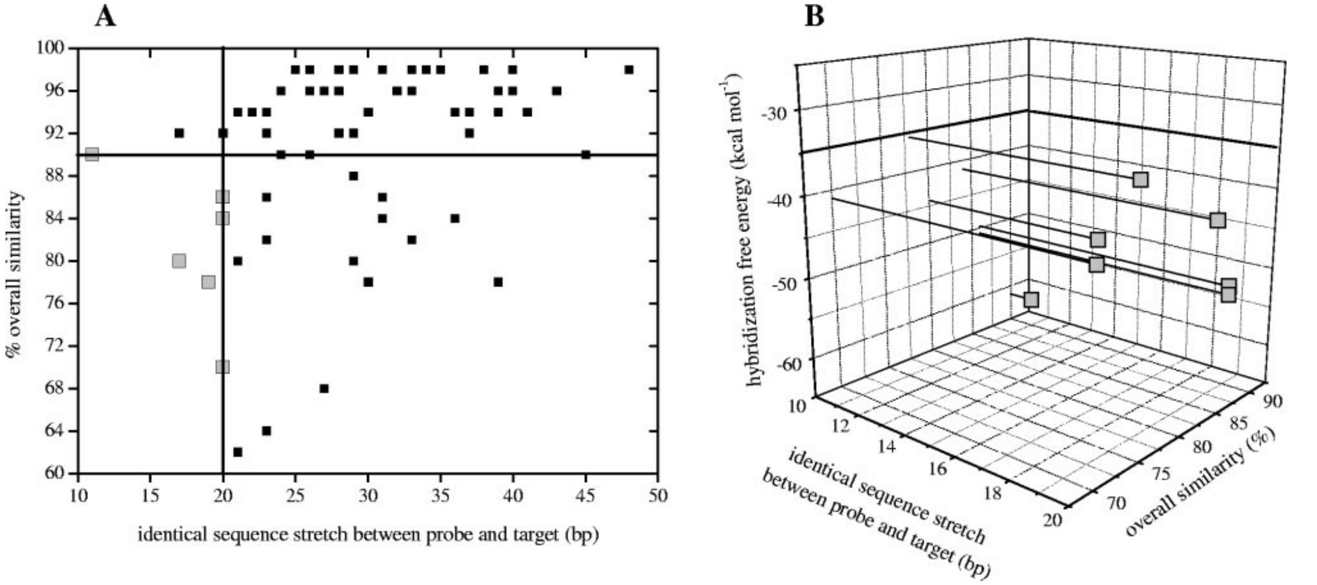

FIG. 4. Relationship between overall sequence similarity, identical sequence length, free energy, and nonspecific probe hybridization behavior. (A) All hybridizations of mismatched probes with SNRs of >2. The gray squares highlight those probes that hybridized with overall similarities of ≤90% and identical sequence stretches of ≤20 bp. (B) Three-dimensional graph of the characteristics of the seven nonspecific probes highlighted in gray in panel A, with the additional free energy calculation on the third axis.

by considering the overall similarity or the length of continuous sequence stretch for probe design, the above results indicate that neither criterion alone is adequate to eliminate all potential cross-hybridizations. By combining both criteria (Fig. 4A), a much clearer picture for predicting probe specificity can be obtained. Below 90% overall sequence similarity, only a few cross-hybridizations occurred. This is consistent with the previous findings of Tiquia et al. (15) and Rhee et al. (11), who proposed threshold values of 85 and 88%, respectively, for specific probe design. However, upon closer examination, our data revealed that specificity problems still exist even at more conservative threshold values. Cross-hybridizations often occurred with sequences of low similarity but long identical sequence stretches, as reported earlier by Kane et al. (6). If a threshold of <16-base identical stretches was used, we could avoid many nonspecific hybridizations. However, these cutoff values restrict probe design to the most divergent sequences. This shortcoming can be overcome by simultaneous consideration of the above values along with the hybridization free energy.

Using threshold criteria of ≤90% similarity and continuous stretches of ≤20 bases, 67 of 74 nonspecific probes were excluded, but 7 were not (Fig. 4A). The detailed characteristics of these seven probes are given in Table 1. However, when an additional free energy criterion of ≥-35 kcal mol^{-1} was applied, these seven nonspecific probes were eliminated (Fig. 4B). This free energy value is close to 50% of the average value (-69.4 kcal mol^{-1}) for the hybridization free energy of all perfectly matched 50-mer probes used in this study. This is similar to the criterion applied by Taroncher-Oldenburg et al. (14) for the design of a microarray consisting of 70-mer oligonucleotide probes. They used a cutoff value of 56% of the free energy released by the perfectly matching probe, together with a maximum overall similarity of 87%, to reject probes with the potential for cross-hybridizations. Kane et al. (6) suggested even more conservative threshold values concerning overall similarities and identical sequence stretches, but they did not consider the change of free energy upon hybridization as a design criterion. The focus of their study also lay in the design of cross-hybridization-safe criteria for expression analyses with single model organisms. These highly stringent hybridization conditions certainly would have a tradeoff in terms of the number of probes that can be designed for homologous genes of related organisms. Additionally, one must be careful in making comparisons of absolute values for probe design parameters such as those discussed above, as each of these studies used different hybridization conditions (e.g., different formamide concentrations and temperatures).

By simultaneously considering probe sequence similarity, identical sequence stretches, and free energy, specific 50-mer probes can be designed with sequence similarity to <90% of the target sequences. This level of sequence similarity will provide the species level of resolution (15). Based on the data set obtained in this study, the sequence similarity criterion can be further relaxed if the other two criteria are applied. Specific hybridization was also observed for some probes with <98% similarity to the target sequences (data not shown), suggesting that the strain level of resolution could possibly be achieved with some 50-mer probes under the hybridization conditions examined.

Conclusions. Our results demonstrated that under defined experimental conditions, nonspecific hybridizations could be essentially avoided by implementing simultaneous threshold criteria for 50-mer oligonucleotide probes of ≤90% similarity, ≤20-base stretches, and a free energy release of ≥-35 kcal mol^{-1}. This level of sequence similarity should provide species-level resolution in most cases (15). By simultaneous consideration of all three characteristics, it seems that previously reported threshold values can be relaxed, thus allowing more predictable probe behavior and a larger number of specific probes to be designed.

This work was supported by a fellowship within the postdoctoral program of the German Academic Exchange Service (DAAD) and was performed at Oak Ridge National Laboratory within the following projects funded by the U.S. DOE Office of Science: the Biotechnology Investigations-Ocean Margins Program, the Carbon Sequestration Program (as part of the Consortium on Research To Enhance Carbon Sequestration in Terrestrial Ecosystems), the Natural and Accelerated Bioremediation Research Program, and the Genomics:GTL Program. Oak Ridge National Laboratory is managed by UT-Battelle, LLC, for the U.S. Department of Energy under contract DE-AC05-00OR22725.

REFERENCES

1. **Corpet, F.** 1988. Multiple sequence alignment with hierarchical clustering. Nucleic Acids Res. **16:**10881–10890.
2. **Dennis, P., E. A. Edwards, S. N. Liss, and R. Fulthorpe.** 2003. Monitoring gene expression in mixed microbial communities by using DNA microarrays. Appl. Environ. Microbiol. **69:**769–778.
3. **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.
4. **Held, G. A., G. Grinstein, and Y. Tu.** 2003. Modeling of DNA microarray data by using physical properties of hybridization. Proc. Natl. Acad. Sci. USA **100:**7575–7580.
5. **Hughes, T. R., M. Mao, A. R. Jones, J. Burchard, M. J. Marton, K. W. Shannon, S. M. Lefkowitz, M. Ziman, J. M. Schelter, J. R. Meyer, S. Kobayashi, C. Davis, H. Dai, Y. D. He, S. B. Stephaniants, G. Cavet, W. L. Walker, A. West, E. Coffey, D. D. Showmarker, R. Stoughton, A. D. Blanchard, S. H. Friend, and P. S. Linsley.** 2001. Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. Nat. Biotechnol. **19:**342–347.
6. **Kane, M. D., T. A. Jatkoe, C. R. Stumpf, J. Lu, J. D. Thomas, and J. M. Madore.** 2000. Assessment of the specificity and sensitivity of oligonucleotide (50mer) microarrays. Nucleic Acids Res. **28:**4552–4557.
7. **Karkhoff-Schweizer, R. R., D. P. Huber, and G. Voordou.** 1995. Conservation of the genes for dissimilatory sulfite reductase from *Desulfovibrio vulgaris* and *Archaeoglobus fulgidus* allows their detection by PCR. Appl. Environ. Microbiol. **61:**290–296.
8. **Li, F. G., and G. D. Stormo.** 2001. Selection of optimal DNA oligos for gene expression arrays. Bioinformatics **17:**1067–1076.
9. **Liu, X., C. E. Bagwell, L. Wu, A. H. Devol, and J. Zhou.** 2003. Molecular diversity of sulfate-reducing bacteria from two different continental margin habitats. Appl. Environ. Microbiol. **69:**6073–6081.
10. **Liu, X., S. M. Tiquia, G. Holguin, L. Wu, S. C. Nold, A. H. Devol, K. Luo, A. V. Palumbo, J. M. Tiedje, and J. Zhou.** 2003. Molecular diversity of denitrifying genes in continental margin sediments within the oxygen-deficient zone off the Pacific coast of Mexico. Appl. Environ. Microbiol. **69:**3549–3560.
11. **Rhee, S. K., X. D. Liu, L. Y. Wu, S. C. Chong, X. F. Wan, and J. Z. Zhou.** 2004. Detection of genes involved in biodegradation and biotransformation in microbial communities by using 50-mer oligonucleotide microarrays. Appl. Environ. Microbiol. **70:**4303–4317.
12. **Rouillard, J. M., C. J. Herbert, and M. Zuker.** 2002. OligoArray: genome-scale oligonucleotide design for microarrays. Bioinformatics **18:**486–487.
13. **Schadt, C. W., J. Liebich, S. C. Chong, T. J. Gentry, Z. He, H. Pan, and J. Z. Zhou.** 2005. Design and use of functional gene microarrays (FGAs) for the characterization of microbial communities. Methods Microbiol. **34:**329–365.
14. **Taroncher-Oldenburg, G., E. M. Griner, C. A. Francis, and B. Ward.** 2003. Oligonucleotide microarray for the study of functional gene diversity in the nitrogen cycle in the environment. Appl. Environ. Microbiol. **69:**1159–1171.
15. **Tiquia, S. M., L. Y. Wu, S. C. Chong, S. Passovets, D. Xu, Y. Xu, and J. Z. Zhou.** 2004. Evaluation of 50-mer oligonucleotide arrays for detecting microbial populations in environmental samples. BioTechniques **36:**664–675.
16. **Wu, L. Y., D. Thompson, G.-S. Li, R. Hurt, H. Huang, J. M. Tiedje, and J.-H. Zhou.** 2001. Development and evaluation of functional gene arrays for detection of selected genes in the environment. Appl. Environ. Microbiol. **67:**5780–5790.
17. **Xu, D., G. Li, L. Wu, J.-Z. Zhou, and Y. Xu.** 2002. PRIMEGENS: a computer program for robust and efficient design of gene-specific targets on microarrays. Bioinformatics **18:**1432–1437.
18. **Zhou, J.** 2003. Microarrays for bacterial detection and microbial community analysis. Curr. Opin. Microbiol. **6:**288–294.
19. **Zuker, M.** 2003. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. **31:**3406–3415.

TABLE 1. Characteristics of seven probes with ≤90% sequence similarities to target sequences and continuous stretches of ≤20 bases but with cross-hybridizations to the target sequences

Target or probe	Sequence[a]	% Overall similarity	Longest identical stretch (no. of bases)	No. of positive replicates/total		Average SNR	ΔG [kcal mol^{-1}]
				SNR > 2	SNR > 3		
Target NKTT11	. . . GCACTCGCAGGCCAACCGCGACAGCCGTCCGCATCTGATCGGCGGTCATGGCGATT . . .	0.90	11	2/7	1/7	1.3	−61.2
Probe M305039	CTCGCAGGCTAACCGCGACACCCGCCCGCACCTGATCGGCGGCCATGGCG						
Target W301131	. . . GATATGATCACCCACCCACGCACGAACCCCTACATCTTCTTCGAGGACGAGCTCGA . . .	0.86	20	2/2	2/2	5.3	−50.7
Probe TPB16020A	ATGGTCAGCCACCCTCGCACGAACCCCTACATCTTTTTTCAGGACGAGAT						
Target M300308	. . . CCAACTGCGTCAAGTGCATGCACTGCATCAATGTGATGCCCAAGGCCCTCAAGCC . . .	0.86	20	3/4	3/4	8.4	−49.7
Probe M306130A	ACTGCGTGAAGTGCATGCACTGCATCAACCTGATGCCGAAAGCGCTCAAA						
Target W301131	. . . CCACCCACGCACGAACCCCTACATCTTCTTCGAGGACGAGCTCGAAGACGACGAC . . .	0.80	17	2/2	1/2	3.4	−44.3
Probe TPB16142A	CCCGCGCACGAACCCCTACATTTTCTTCAACGACGAGATCTTCGAAGACA						
Target W301131	. . . CGGAAGACCGATATCGGCCCCCCGCACTACGAGAAGTTCCTTCCGGAGATCATCAA . . .	0.84	20	2/2	2/2	3.2	−41.2
Probe FW015017B	AAGACCGATATCGGCCCCCCTCACTATCAGAAGCATCTCCCCGAGATCAT						
Target M318A25	. . . ATACCGCATCCCGGCCGCGGCGCCAATCTCGTGGATCCCCAACACGGACCCGTGTG . . .	0.70	20	4/4	4/4	17.7	−41.0
Probe S14	CCCCATCCCGGCCGCGGCGCCAACTTCACCCATCCGAAGTTCGGACCGGT						
Target W301131	. . . ATGTCAACGACCGAACGGAAGACCGATATCGGCCCCCCGCACTACGAGAAGTTCCT . . .	0.78	19	2/2	1/2	2.9	−35.5
Probe FW010031B	TCAGCAGCTGAACGCAAGACCGATATCGGCCCCCGCAATTACAAGCAGTT						

[a] Underlining indicates stretches of identical sequence.