Using DNA microarrays to detect multiple pathogen threats in water

T.M. Straub*, M.D. Quinonez-Diaz*, C.O. Valdez*, D.R. Call** and D.P. Chandler***

* Environmental Microbiology Group, Pacific Northwest National Laboratory, 902 Battelle Blvd., MS P7-50, Richland, Washington 99352, USA (E-mail: timothy.straub@pnl.gov; maria.quinonez-diaz@pnl.gov; catherine.valdez@pnl.gov)

** Department of Vet. Micro. & Path., Washington State University, 402 Bustad Hall, P.O. Box 647040, Pullman, Washington 99164-7040, USA (E-mail: drcall@wsu.edu)

*** Biochip Technology Group, Argonne National Laboratory, 9700 Cass Avenue, 202 Building, A249, Argonne, Illinois, 60439 USA (E-mail: dchandler@anl.gov)

Abstract We present four studies that illustrate the use of DNA microarrays for the detection and subsequent genotyping of waterborne pathogens. A genotyping array targeting four virulence factor genes in enterohemorrhagic Escherichia coli (EHEC) was tested. The arrays were clearly able to differentiate between E. coli O157:H7 genotypes and E. coli O91:H2. Non-pathogenic E. coli and non-target organisms were not detected on this array. In the second study, an hsp70 gene single nucleotide polymorphism (SNP) array for specific Cryptosporidium parvum detection was constructed to differentiate between principle genotypes. SNPs, and hence differences between genotypes, were easily detected on this type of array. In the third study an array for Helicobacter pylori was tested for simultaneous SNP discrimination and presence or absence of virulence factor genes. Results from this study showed that both SNP discrimination for some conserved genes, and the presence or absence of virulence factor genes was possible. In the fourth study, multiplexing was achieved by direct hybridization and detection of mRNA to the array. For highly expressed genes, visible signal was detected at 312.5 ng of total RNA, indicating that these new methods may have sufficient environmental sensitivity without the need to perform PCR.

Keywords DNA microarrays; multiplexed pathogen detection; single nucleotide polymorphisms

Introduction Ensuring that drinking water is microbiologically safe continues to be a worldwide goal. Several highly publicized waterborne disease outbreaks have increased concern, including the Cryptosporidium parvum outbreak in Milwaukee, Wisconsin USA in 1993, and E. coli O157:H7 outbreak in Walkerton, Ontario Canada in 2000. The United States Congress mandated several important revisions to the Safe Drinking Water Act in 1996, in response to the Milwaukee incident and expert testimony on the threat of emerging waterborne pathogens. One key revision requires the United States Environmental Protection Agency (USEPA) to submit a Candidate Contaminant List (CCL) that includes both emerging microbiological and chemical contaminants in drinking water (USEPA, 1998). New methods must be developed to determine the presence of these microbiological contaminants in drinking water and their potential public health risk. These new methods, in addition to methods that are employed for existing contaminant threats, pose a challenge to water utilities that may be required to monitor their supplies for these contaminants. This is because of the amount of time some of these methods require, and the technical expertise that is required to perform them (Allen et al., 2000).

Detection of multiple pathogens in a sample using a single assay is a technical challenge regardless of the chosen platform or molecular recognition element (nucleic acids and proteins). DNA microarrays may provide a potential solution to this problem. Several commercial vendors sell instrumentation for fabrication of microarrays. These instruments
are essentially robotic devices designed to transfer nanolitre volumes of probe solution from 96 or 384-well microtiter plates to a suitable substrate, usually a common glass microscope slide (Guo et al., 1994; Bavykin et al., 2001). Probe spots produced by these arrayers are generally 50 to 100 micrometres in diameter. The distance between each spot is typically 25–50 µm, but is dependent on the users’ program for producing the arrays, and the resolution of the array scanning instruments used. As shown in Figure 1 above, the researcher can array any kind of DNA sequence as a short, single-stranded DNA oligonucleotide or full-length cDNA. Arrays allow unparalleled multiplexed pathogen detection in that 1) multiple sequences within a given gene, for a given organism, 2) multiple genes for a given organism, and 3) multiple organisms can be surveyed.

For high throughput applications, bead-based arrays are now becoming widely available. In this format, modified oligonucleotides or cDNAs (modifications are usually a 5' terminal amine or biotinylated functional group) are covalently attached to carboxylated or streptavidin coated fluorescent beads. In the Luminex® system, each bead population is prepared as a defined combination of red and infrared dyes. Currently, there are approximately 100 different bead population types (= 100 different probes) available. DNA (or RNA) in solution is labeled with a fluorescent dye (R-phycoerythrin, Molecular Probes Alexa 532, or Cy3 are all compatible with the Luminex® system and then hybridized to the immobilized probes on the beads. Results are obtained using a method analogous to flow cytometry. Red and green lasers are then used to classify the probes and hybridization events on these probes, respectively. Figure 2 below further illustrates the bead array concept.

Currently, several biotechnology companies now sell whole genome arrays for E. coli, Pseudomonas aeruginosa, yeast, Drosophila, rat, and humans. However a “water quality” microarray has not yet been developed and validated with the primary goal of multiplexed pathogen detection using a single assay. Nonetheless, there are a growing number of reports illustrating how microarrays can be used for environmental biodetection purposes. Ye et al. (2001), discusses applications of DNA arrays for microbial systems, and cites several studies involving detection of clinically important pathogens on microarrays. Guschin et al. (1997), describe a gel-pad array for the detection of 16S rRNA from pure cultures of environmentally relevant bacteria. Bavykin et al. (2001), described a total method from sample processing and labeling 16S rRNA, to readout on a portable microarray reader. Small et al. (2001), described the direct detection of intact rRNA on oligonucleotide microarrays, including the hybridization and detection of target organisms directly from an

Figure 1 Conceptual diagram of a DNA microarray for detection of multiple pathogens in a water source. The user defines what probes will be contained on the array. In this figure, surveying multiple sequences within a given gene for a given organism (Cryptosporidium parvum hsp70), multiple virulence factors for a given organism (E. coli O157:H7), and multiple waterborne pathogens are all possibilities for this array.
unpurified soil extract with limited sample preparation. The significance of the Bavykin and Small studies is the direct detection of RNA without a PCR amplification step. This allows higher order multiplexed organism detection and may circumvent inhibitors and common biases of multiplexed PCR for multi-species detection (Tebbe and Vahjen, 1993; Polz and Cavanaugh, 1998).

We present results of four studies on how we are developing and applying microarray technology for the detection of waterborne pathogens.

Methods and results

Multiplexed detection of *Escherichia coli* O157:H7 virulence factor genes

Work initially focused on genotyping *E. coli* O157:H7 isolates (Call et al., 2001). A simple four gene array examined the following virulence factor genes: intimin (*eaeA*), shiga-like toxin 1 (*stx1*), shiga-like toxin 2 (*stx2*), and hemolysin A (*hlyA*). Multiplex PCR was optimized with biotin labeled primers and used to generate labeled probes to hybridize to the array. Signal was developed using streptavidin-alkaline phosphatase enzyme linked fluorescence chemistry (Molecular Probes, Eugene, OR). Alternatively, we optimized PCR using Cy5 labeled PCR primers. This latter method allows direct detection of the hybridized probes without the need for enzymatic signal amplification (Figure 3). Potential gain in time savings, however, may be offset by decreased sensitivity of detection.
Several *E. coli* O157:H7 isolates were tested that either lacked one or both shiga-like toxin genes or contained both shiga-like toxin genes. We also tested another pathogenic strain, *E. coli* O91:H2, that only contains shiga-like toxin 2 and hemolysin A, and DH5 alpha, a non-pathogenic strain of *E. coli* (Figure 3). For water quality analysis, we tested this array with other members of the family *Enterobacteriaceae* and other non-target organisms. This array was shown to do the following: differentiate between different O157:H7 isolates, indicate the presence of other potential EHEC isolates, and non-pathogenic *E. coli* and non-target organisms do not hybridize to this array. Because it has the ability to target specific virulence factor genes, the array may be able to determine potential point sources and predict the severity of waterborne disease caused by O157:H7.

**Differentiation of Cryptosporidium parvum genotypes using single nucleotide polymorphism mismatch discrimination**

The *Cryptosporidium* array is a single nucleotide polymorphism (SNP) array that investigates seven variable positions within a 190 bp fragment of the 70 kilodalton heat shock-protein (*hsp70*). Differences between the principle genotypes primarily occur at the nucleotide level. Several assays based on sequencing and restriction fragment length polymorphism analysis have been proposed (Sulaiman *et al.*, 2000; Xiao *et al.*, 1999, 2000, 2001). For each variable position, a perfectly matched probe, and 6 to 13 additional mismatch probes were designed and arrayed. The goal for SNP analysis is to demonstrate that the perfectly matched probe, for a given variable position, will have the greatest hybridization signal compared with the mismatch probes for that position. PCR using Cy3-labeled PCR primers was used to generate labeled probes to hybridize to the array. Three genotype I and two genotype II isolates were ultimately tested on this array. The results displayed in Figure 4 show that hybridization patterns between the two genotypes were clearly different on the array, and analysis of the hybridization data showed that SNP discrimination could be achieved for most of the seven positions within the 190 bp fragment (Straub *et al.* 2002).

**SNP discrimination and the presence or absence of virulence factor in Helicobacter pylori**

*H. pylori* is one of the organisms on the United States Environmental Protection Agency’s Candidate Contaminant List (CCL). The *H. pylori* arrays are a combination of SNP analysis

---

**Figure 4** SNP hybridization patterns for one variable position for genotype I and genotype II isolates of *Cryptosporidium parvum*. The table to the right of the photograph indicates whether the perfectly matched probes for each variable position had the greatest hybridization intensity compared to the mismatch probes designed for each variable position (*p* < 0.05). In some cases SNP discrimination was not achieved due to failed hybridization of any probes for a given variable position (variable positions 1479 and 1542). In other cases, SNP discrimination could be achieved for some probes by increasing stringency, but the trade-off was loss of SNP discrimination for the other variable positions. Data from Straub *et al.* (2002).
and analysis of the presence or absence of virulence factor genes. Two toxin genes were examined on this array, \textit{vacA} and \textit{cagA}. Approximately 50 to 60\% of all \textit{H. pylori} isolates have the \textit{cagA} gene (Covacci et al., 1993). Expression of this gene has been implicated in patients who ultimately develop gastric cancer (Blasser, 1995). Capture probe sequences for the arrays were based on van Doorn et al. (1998). However, van Doorn’s capture probe sequences for both genes had several degenerate bases (e.g. R = A or G, W = A or T, etc.). For the arrays, we split these probes into their respective, non-degenerate probes, and used these as SNP probes to test on the array. Multiplex PCR with biotin labeled PCR primers was optimized to amplify two regions of the \textit{vacA} gene and the \textit{cagA} gene (if present) based on the van Doorn et al. (1998) study. Two ATCC strains of \textit{H. pylori} were evaluated: ATCC 43504 and ATCC 700392. Both of these strains have the s1a/m1 genotype for the \textit{vacA} gene. ATCC 700392 does not contain the \textit{cagA} gene. Within the s1a genotype there were four possible probe combinations based on the two degenerate sequence positions reported (Table 1). Likewise, one of the \textit{cagA} gene sequences had four possible combinations. As shown in Figure 5, the arrays were clearly able to detect the presence or absence of the \textit{cagA} gene. In addition, SNP discrimination was possible for some of the reported degenerate probes.

For the ATCC 43504 isolate, the second \textit{cagA} gene was split into 4 SNP variants (\textit{cagA}/A – \textit{cagA}/D, Figure 5 above). Two of these probes were hybridized. However the fourth variant (\textit{cagA}/D) of the probe consistently showed the strongest hybridization signal of all probes in this suite, and thus is the likely correct sequence for this isolate. For the s1a probes, the A and B variants were hybridized with equal success, but the C and D variants failed to hybridize. Table 1 shows the sequence for the s1a61 probes on the array. The probe sequences for the s1a52 probes are merely shifted 9 bases upstream (e.g. overlapping sequences), placing these same degenerate locations at the 3’ end of the probe. While perfect SNP discrimination could not be achieved for these probes, the results suggest that the first SNP is clearly a guanine, and the second SNP could either be a guanine or

Table 1  SNP discrimination within the s region of the vacA gene of \textit{Helicobacter pylori}. The two genotypes that were tested were s1a/m1/cagA+ (ATCC 43504) and s1a/m1/cagA− (ATCC 700392). The s1a52 probes contain the same degenerate bases. However, those probe sequences start 9 bases upstream of the s1a61 probes. Despite this, identical hybridization results were achieved (see Figure 5 below). This indicates that the first “R” is truly a guanine, but the second “R” remained indeterminate (A or G).

<table>
<thead>
<tr>
<th>Reported sequence for s1a61 probe suite</th>
<th>G G A G C R T R G T C A G C A T C A C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Printed probe variation s1a61/A (hybridized)</td>
<td>G G A G C G T T G T C A G C A T C A C</td>
</tr>
<tr>
<td>Printed probe variation s1a61/B (hybridized)</td>
<td>G G A G C G T A G T C A G C A T C A C</td>
</tr>
<tr>
<td>Printed probe variation s1a61/C (not hybridized)</td>
<td>G G A G C A T T A G T C A G C A T C A C</td>
</tr>
<tr>
<td>Printed probe variation s1a61/D (not hybridized)</td>
<td>G G A G C A T T G T C A G C A T C A C</td>
</tr>
</tbody>
</table>

Figure 5  Hybridization patterns for two strains of \textit{Helicobacter pylori} that have the s1a/m1 genotype for the \textit{vacA} gene. The presence or absence of the \textit{cagA} gene is clearly seen.
adenine (both purines). These results show that it is possible to use an array to determine
the presence or absence of virulence factor genes and discern SNPs to further “fingerprint”
a pathogenic isolate.

**Direct hybridization of messenger RNA**

The three studies discussed above all rely on PCR. For studies targeting multiple virulence
factor genes and/or multiple organisms in a sample, multiplex PCR must be optimized.
This can be difficult if more than four PCR primer pairs are used in a multiplex reaction. In
order to take full advantage of the arrays, a better protocol for multiplexing is needed.

Direct hybridization of labeled messenger RNA (mRNA) to the arrays may provide a
potential solution to the multiplex PCR problem. For some genes there may be several
thousand mRNA copies per cell. There are new generation methods for labeling mRNA
that are commercially available, and the vendors claim that the sensitivity of the assay is
approximately one thousand to ten thousand copies. If these claims are correct, it is con-
ceivable that direct mRNA hybridization to the array may have equivalent sensitivity to
PCR. Because these new labeling methods are not gene-specific, the multiplexing problem
is greatly simplified.

Two commercially available kits that directly label RNA are the ULYSIS™ kits from
Molecular Probes (Eugene, OR), and the Digene HC ExpressArray™ Kit (Digene,
Gaithersburg, MD). The Molecular Probes kit works by labeling the N7 residue of guanine
with one of their proprietary ALEXA or Oregon Green fluorescent dyes via a platinum
catalyzed reaction. The labeled RNA is then hybridized to the array. The Digene kit works
by direct hybridization of RNA to DNA capture probes on the array. RNAse is used to
remove unduplexed RNA. RNA:DNA duplexes are then detected using antibodies specific
to RNA:DNA hybrids. Signal is generated using a Cy3 labeled secondary antibody. The
entire procedure from RNA extraction to readout on the array takes approximately 5 hours.

We have investigated the Digene kit for expression array analysis of the metal reducing
microorganism *Shewanella oneidensis*. Typical expression array protocols advise using
10 micrograms of total RNA as starting material. As shown in Figure 6, we were able to
determine that, for some genes, as little as 312.5 ng of total RNA could be detected on the
planar array. The applicability of these studies for the detection of waterborne pathogens is
that 1) if the correct genes (especially those present in very high copy number) are selected,
there will be sufficient sensitivity to detect them without using PCR, and 2) some of these
genes may be correlated with viability, thus allowing both detection and viability discrimi-
nation using a single assay.

**Conclusions**

These studies present the potential application of DNA microarrays for the detection of
multiple waterborne pathogens in a sample using a single assay. Further work is needed to
develop this method into a tool that can be useful for routine monitoring and risk assessment applications. This work should include 1) optimization of highly multiplexed gene and organism detection, 2) further characterization of specificity and sensitivity of the assay, 3) potential viability discrimination, and 4) enumeration. In addition, automation of this assay into a bead-based array format would allow for high throughput screening of water samples, making this an even more useful tool for routine surveillance of water sources.

Acknowledgements

This work was funded by the United States Environmental Protection Agency Science to Achieve Results (STAR) program. Other funding agencies include the Water Environment Research Foundation, the American Water Works Research Foundation, the United States Department of Energy (DOE) Chemical/Biological Non-proliferation Program, and the Biomolecular Sciences Initiative Laboratory Directed Research and Development Program funded by DOE at Pacific Northwest National Laboratory. The authors also wish to thank the Metropolitan Water District of Southern California for their scientific insights for designing the microarrays, and Tufts University for supplying \textit{C. parvum} genotype I and genotype II oocysts for testing on the arrays.

References


