Defence Research and Development Canada
Recherche et développement pour la défense Canada

# Detection/identification of microbes using microarrays

C.C. Ruttan and D.C.W. Mah
Canada West Biosciences Inc.

CSA: B.N. Ford, DRDC Suffield, 403-544-4612



**Defence R&D Canada**
Contract Report
DRDC Suffield CR 2012-018
July 2012

Canada

# Detection/Identification of microbes using microarrays

C.C. Ruttan and D.C.W. Mah

Prepared By:
Canada West Biosciences Inc.
5429 - 60th Street
Camrose AB T4V 4G9

PWGSC Contract Number: W7702-07R160

CSA: B.N. Ford, DRDC Suffield, 403-544-4612



## Defence R&D Canada – Suffield

Contract Report
DRDC Suffield CR 2012-018
July 2012

Principal Author

D.C.W. Mah
Project Director, Canada West Biosciences Inc.

Approved by

B.N. Ford
Contract Scientific Authority

Approved for release by

R.G. Clewley
A/DRP Chair

# Abstract

Microarray technology provides the potential for rapid identification of an unknown pathogen in a single test. In order to explore the utility of this tool, a custom Affymetrix DNA microarray chip was designed to recognize a wide variety of bacteria to the species and strain level. Antigenomic probes were included on the lower portion of the chip to investigate the potential of rapid visual assessment of organisms' footprints. Since pre-existing protocols and analysis software were unavailable for this work, protocols were optimized and a variety of data analysis options were investigated. Experiments were conducted on a number of pure DNA samples and mixed samples with equal quantities of input DNA. The effect of DNA amplification and sample dilution on signal intensity was also investigated.

The chips were readily able to identify the correct pathogen from pure samples but signal levels dropped with decreased input DNA. Amplified mixed DNA, both pre-amplification and post-amplification sample mixtures, produced different results from mixed unamplified DNA requiring more complex analysis technique or a better balancing of the number of probes designed for each organism. The dilution series demonstrated that signal intensity reduction corresponds to the decrease in template DNA (i.e. 1/10 DNA has 1/10 signal intensity). In all, 3,119 probes were selected that are strongly specific and sensitive for the organisms tested and these provide an excellent basis for future work with microarrays in the identification of bacterial pathogens.



# Résumé

La technologie des biopuces permet l'identification rapide d'un agent pathogène inconnu un moyen d'un seul test. Pour étudier l'utilité de cet outil, nous avons utilisé une puce à ADN Affymetrix conçue sur mesure pour reconnaître une grande variété d'espèces et de souches de bactéries. Nous avons intégré des sondes antigéniques dans la partie inférieure de la puce, pour étudier le potentiel de l'évaluation visuelle rapide de l'empreinte des organismes. Puisqu'il n'existait pas encore de protocoles et de logiciel d'analyse permettant de réaliser ce travail, les protocoles existants ont été optimisés et diverses options d'analyse de données ont été évaluées. Nous avons analysé un certain nombre d'échantillons D'ADN pur et d'échantillons mixtes qui contenaient la même quantité d'ADN recherché. En outre, nous avons étudié l'effet de l'amplification de l'ADN et de la dilution de l'échantillon sur l'intensité du signal.

Les puces ont permis de facilement identifier le bon agent pathogène dans le cas des échantillons purs, mais l'intensité du signal diminuait avec la diminution de la quantité d'ADN recherché. Les résultats obtenus pour les échantillons mixtes d'ADN amplifié (mélanges préamplification et postamplification) ont été différents de ceux obtenus pour les échantillons mixtes d'ADN non amplifié, ce qui indique que pour ces derniers, il faut faire appel à une technique d'analyse plus complexe ou équilibrer le nombre de sondes spécifiques à chaque organisme. L'analyse de la série de dilutions a montré que l'intensité du signal diminue en fonction de la diminution de l'ADN matrice (l'intensité du signal pour 1/10 d'ADN est de 1/10). Au total, nous avons

sélectionné 3 119 sondes qui étaient très spécifiques et sensibles à l'égard des organismes des essais. Ces sondes constituent un excellent point de départ pour des travaux futurs sur l'utilisation de biopuces permettant l'identification de bactéries pathogènes.

# Executive summary

## Detection/identification of microbes using microarrays

**C.C. Ruttan and D.C.W. Mah; Canada West Biosciences Inc.; DRDC Suffield CR 2012-018; Defence R&D Canada – Suffield; July 2012.**

**Background:** Over a dozen bacterial species have been considered viable biological warfare agents; even more species could be used as bioterrorism agents. The effectiveness of medical countermeasures for personnel exposed to such agents would depend to a large degree on rapid and robust identification of the species and strain employed. Historically, this would have difficult to achieve, as it would have required prolonged, iterative laboratory assays of samples typically containing a variety of naturally occurring bacterial species in addition to the agent employed.

The relatively recent availability of complete or near-complete DNA sequences for many microorganisms has allowed unique genetic markers to be identified at the species level in most cases and often at the strain level. The presence of these sequences in samples can in turn be used to confirm the presence of their host bacteria. In one such technology — a DNA "microarray" — thousands of unique genetic probes are attached to a solid surface in a defined two-dimensional pattern and exposed to DNA fragments extracted from a sample, allowing the determination of whether genetic sequences of interest are present and to what extent, regardless of what species may have been present in the original sample.

**Results:** To explore the utility of microarrays for biodefence purposes, a custom DNA microarray chip capable of recognizing a wide variety of bacterial agents of concern (through use of published genetic information) was developed and tested against various bacterial challenges. The microarray system correctly identified the presence of multiple genetic targets from each pathogen tested. Results from various preparation and purification methods varied in predictable ways, while samples containing DNA from multiple bacteria species yielded results which were a simple combination of those for the individual species.

**Significance:** This project demonstrated the utility of microarray technology for the identification of biological warfare/bioterrorism agents. Multiple agents can now be screened for rapidly in a single assay. Given the nature of the technology, it is unlikely that it could be employed in a fieldable, standalone detection/identification system, but microarray technology will almost certainly become a critical, (fixed site or field) laboratory-based tool for use in identifying unknown pathogens in samples and for forensic purposes.

**Future plans:** While work on this specific project at DRDC Suffield is complete, the Canadian Food Inspection Agency is expanding the existing array design to include genetic probes for a large number of food-borne and domesticated animal pathogens. The new design could serve as a broad spectrum detection/identification system for public health/public security purposes.

# Sommaire

## Detection/identification of microbes using microarrays

**C.C. Ruttan and D.C.W. Mah; Canada West Biosciences Inc.; DRDC Suffield CR 2012-018; R & D pour la défense Canada – Suffield; juillet 2012.**

**Introduction :** Plus d'une douzaine d'espèces de bactéries sont considérées comme des agents de guerre biologique potentiels, et encore plus d'espèces pourraient être utilisées comme agents de bioterrorisme. L'efficacité des contre-mesures médicales offertes au personnel exposé à ces agents dépendrait grandement de la rapidité et de la fiabilité de l'identification de l'espèce et de la souche utilisées. Il aurait été difficile de procéder à une telle identification rapide dans le passé, car il aurait fallu procéder à de longues analyses itératives en laboratoire d'échantillons renfermant généralement diverses espèces de bactéries présentes dans la nature en plus de l'agent utilisé.

Depuis relativement peu de temps, nous disposons des séquences d'ADN complètes ou presque complètes pour de nombreux microorganismes, ce qui nous permet d'utiliser des marqueurs génétiques uniques pour identifier les espèces et souvent même les souches. La présence de ces séquences dans les échantillons peut ainsi servir à confirmer la présence des bactéries qui y sont associées. Une puce à ADN comporte des milliers de sondes génétiques uniques fixées à sa surface solide bidimensionnelle; la puce est exposée à des fragments d'ADN extraits d'un échantillon, ce qui permet de déterminer si les séquences génétiques recherchées sont présentes dans l'échantillon et dans quelle proportion, peu importe le nombre d'espèces présentes dans l'échantillon d'origine.

**Résultats :** Pour évaluer l'utilité des biopuces pour la défense contre les armes biologiques, nous avons créé et mis à l'essai une puce à ADN capable de reconnaître une grande variété d'agents biologiques préoccupants (au moyen des renseignements génétiques publiés). Le système de biopuce a correctement détecté la présence de diverses cibles génétiques de chacun des agents pathogènes mis à l'essai. Les résultats obtenus pour diverses méthodes de préparation et de purification ont varié de façon prévisible, alors que les résultats obtenus pour les échantillons contenant l'ADN de nombreuses espèces de bactéries correspondaient à la combinaison des résultats obtenus pour chaque espèce analysée séparément.

**Importance et Perspectives :** Le présent projet a montré l'utilité de la technologie des micropuces pour l'identification d'agents de guerre biologique ou de bioterrorisme. Il est maintenant possible de dépister rapidement de multiples agents en un seul essai. Étant donné la nature de cette technologie, il est peu probable qu'on puisse en faire un système d'identification et de détection autonome utilisable sur le terrain, mais il est presque certain qu'elle deviendra un outil de laboratoire essentiel (site fixe ou terrain) pour l'identification des agents pathogènes inconnus dans les échantillons et qu'elle sera utilisée à des fins médico-légales.

Bien que le projet réalisé à RDDC Suffield sur le sujet soit terminé, l'Agence canadienne d'inspection des aliments travaille actuellement à ajouter aux modèles de puces existants des sondes génétiques pour un grand nombre d'agents pathogènes d'origine alimentaire et d'agents pathogènes des animaux domestiques. Ce nouveau modèle pourrait être utilisé comme système de détection et d'identification à large spectre, à des fins de santé et de sécurité publiques.

This page intentionally left blank.

# Table of contents

# List of tables

# Acknowledgements

The authors would like to thanks Melissa Crichton and Kevin Ruttan for their technical assistance, Yimin Shei for initial training and *in vitro* chip analysis and Douglas Bader, Chad Stratilo, Glen Fisher and Dr. Kingsley Amoako for providing samples.

This page intentionally left blank.

# 1 Introduction

Microarray technology, with its associated analysis software, represents an enormous technical advantage for problems in genome analysis, ranging from genotyping microbial species to fingerprinting microbial species and strains. Identification of microbes at the species level has already been established using conventional microbiology and then with the use of molecular biology strain genomic differences have also been determined. Although these methods require prior knowledge of the suspect organism, they do not address the issue of recombinant species. As the potential to create organisms with altered host specificity increases and ultimately develops an unknown or misleading genetic profile the current methods in use are deficient and limited.

Microarray Technology, on the other hand, would permit for rapid single test identification. The technology would offer a comprehensive single step test to simultaneously identify genetic fingerprints of numerous bacterial and viral species in replace of a dozen of present tests. As well on a robust system there would be the ability to detect DNA of a deadly pathogen inserted into a harmless host or detect whether or not a gene is present that would influence a microbe's resistance to antibiotics.

The biotechnology behind microarray systems does not vary according to the analysis endpoint; techniques applicable to one model system are applicable to most others.

# 2 Project progress

## 2.1 February – March, 2008

Equipment and reagent requirements assessed and ordered. Laboratory cupboards, shelves and benchtops were ordered and assembled.

## 2.2 April – June, 2008

Lab set-up completed. Decision was made to design a custom Affymetrix genotyping microarray as a commercial chip from Affymetrix was not available. List of organisms to be included on the chip was determined by Dr. Barry Ford (DRDC), Doug Bader (DRDC) and Dr. Kingsley Amoako (CFIA). NIAID Category A pathogens were prioritized. Probe design focused on obtaining sequences that would clearly differentiate between different pathogens and between stains from the same species. Yimin Shei provided training on the Affymetrix system using Test3 chips.

## 2.3 July – September, 2008

Probe selection for custom chip completed and chips ordered. A total of 19,643 unique sequences (31 to 31,737 bp in length) were selected then submitted to Affymetrix. The Affymetrix software identified up to five 25-mer probes within each sequence that would provide optimal results on the Affymetrix system. Degenerate probes were removed to leave at least 1 probe for 11,516 of the original 19,643 sequences for a total of 81,721 probes. The remaining features of the 226,576-probe chip were filled with Affymetrix controls and randomly generated probes. Probe tally by pathogen is shown in Table 1.

## 2.4 October – December, 2008

The shipment of 128 chips was received on Sept. 29th. DNA for this project was obtained from 4 sources. Dr. Barry Ford provided *B. anthracis*, *E. coli*, *Y. pestis* and *Y. pseudotuberculosis* samples on Sept. 26/08. Doug Bader provided *B.* anthracis, *Burkholderia*, *Francisella* and *Yersinia* species on Nov. 14/08. Dr. Kingsley Amoako from Canada Food Inspection Agency provided *B. cereus*, *E. faecalis*, *E. coli*, *L. monocytogenes*, *P. aeruginosa*, *S. typhimurium*, *S. dysentrine*, *S. aureus*, *S. pyogenes*, *V. vulnificus* , *Y. enterocolytica* and *Y. pseudotuberculosis* species which were received Oct. 24/08. In addition, genomic DNA for *B. cereus*, *A. baumannii*, *Y. pseudotuberculosis* and *H. influenzae* were ordered from Cedarlane and arrived Oct. 23/08. The full list of samples received is shown in Table 2. A DNA tracking system was set-up and sample concentration was verified prior to chip optimization. Due to the large quantity of sample provided, *E. coli* JM109 was used for the majority of optimization runs. Initial runs had few probes with an intensity >10,000 although most of these matched probes designed to recognize *E.coli* or *Shigella*. Reaction conditions were altered in an attempt to increase maximum signal intensity. A total of 23 chips were run during this quarter.

*Table 1: Summary of probes by pathogen.*

| **Pathogen** | **All probes** |
|---|---|
| *Acinetobacter baumannii* | 334 |
| *Bacillus* species | 4665 |
| *Bartonella* species | 1020 |
| *Bordetella* species | 1443 |
| *Borrelia afzelii* | 5 |
| *Brucella* species | 1374 |
| *Burkholderia* species | 1765 |
| *Campylobacter* species | 2570 |
| *Chaetomium* species | 209 |
| *Chlamydia* species | 780 |
| *Clostridium* species | 2682 |
| *Corynebacter* species | 439 |
| *Coxiella* | 282 |
| *Escherichia coli* | 3019 |
| *Enterococcus faecalis* | 165 |
| *Francisella* species | 11045 |
| *Haemophilus* species | 2544 |
| *Helicobacter* species | 1833 |
| *Homo sapiens* | 100 |
| *Klebsiella pneumonia* | 5 |
| *Lactobacillus delbrueckii* | 5 |
| *Legionella pneumophila* | 1894 |
| *Listeria* species | 7663 |
| multispecies/antibiotic resistance | 34 |
| *Mycobacterium* species | 5460 |
| *Mycoplasma* species | 1085 |
| *Neisseria meningitidis* | 1117 |
| Plasmid pBC16 , pSL1, pIP1202 | 24 |
| *Pseudomonas* species | 11326 |
| *Ricinus communis* | 20 |
| *Rickettsia* species | 245 |
| *Salmonella enterica* | 2620 |
| *Shigella* species | 2308 |
| *Staphylococcus* species | 1603 |
| *Streptococcus* species | 3041 |
| *Treponema pallidum* | 45 |
| *Ureaplasma parvum* | 5 |
| *Vibrio* species | 4058 |
| *Xanthomonas axonopodis* | 5 |
| *Yersinia* species | 2884 |
| **Total specific probes** | **81721** |
| Affymetrix probes | 144855 |
| **Total probes on chip** | **226576** |

*Table 2: Samples received for custom chip project.*

| Pathogen | Strain(s) | Supplier |
|---|---|---|
| *Acinetobacter baumannii* | ATCC 17978 | CR |
| *Bacillus anthracis* | Thraxol, RP42, ThSA | BF/YS |
| *Bacillus anthracis* | 94188c, ACB, Ames PLG 6, NH, RP42, Vollum | GF/DB |
| *Bacillus cereus* | ATCC 10987 | CR |
| *Bacillus cereus* | ATCC 11778 | KA |
| *Bartonella henselae* | ATCC 49882 | CR |
| *Bordetella pertussis* | ATCC BAA-589 | CR |
| *Burkholderia mallei* | ATCC 1053, ATCC 23344 | GF/DB |
| *Burkholderia pseudomallei* | Env-81, Env-FB20 | GF/DB |
| *Campylobacter jejuni* | ATCC 700819 | CR |
| *Clostridium perfringens* | ATCC 13124 | CR |
| *Enterococcus faecalis* | ATCC 29212 | KA |
| *Escherichia coli* | JM108, JM109 + digested DNA | BF/YS |
| *Escherichia coli* | ATCC 25922 | KA |
| *Escherichia coli* | O157:H7 EDL933 | KA |
| *F. tularensis holarctica* | FT-65-4, FT-67-4, Swed3, Swed9, Swed10, Swed 4Q, Swed 6Q | GF/DB |
| *F. tularensis mediasiatica* | Swed 8-6 | GF/DB |
| *F. tularensis tularensis AI* | Swed 1, Swed 2 (IDI), Swed 2Q | GF/DB |
| *F. tularensis tularensis AII* | Swed 7 Q | GF/DB |
| *Haemophilus influenzae* | ATCC 51907 | CR |
| *Listeria monocytogenes* | ATCC 15313, NTCC 7933 | KA |
| *Mycobacterium BCG* | ATCC 19015 | CR |
| *Mycoplasma pneumoniae* | ATCC 15531 | CR |
| *Pseudomonas aeruginosa* | ATCC 27853 | KA |
| *Salmonella typhimurium* | 71-471 | KA |
| *Shigella dysentrine* | ATCC 11835 | KA |
| *Staphylococcus aureus* | Z1 | KA |
| *Streptococcus pyogenes* | ATCC 19615 | KA |
| *Vibrio vulnificus* | Z86 | KA |
| *Yersinia aldovae* | ATCC 35237 | GF/DB |
| *Yersinia bercovieri* | CCRI 14920 | GF/DB |
| *Yersinia enterocolytica* | CCUG 31436 | GF/DB |
| *Yersinia enterocolytica* | CCUG 33553 | GF/DB |
| *Yersinia enterocolytica* | #7 field strain, #14 field strain, ATCC 23715 | KA |
| *Yersinia enterocolytica* | genomic DNA, YE-D1, YE-D3, YE-D4 | BF/YS |
| *Yersinia frederiksenii* | CCRI 14915 | GF/DB |
| *Yersinia intermedia* | ATCC 33648 | GF/DB |
| *Yersinia kristensenii* | ATCC 33638 | GF/DB |
| *Yersinia mollaretii* | ATCC 43969 | GF/DB |
| *Yersinia pestis* | CO92, GM33-1, N5151-1, SRY 6.3, px14-3 pp1964 , Y1088 + digests | BF/YS |
| *Yersinia pestis* | C12, CO92, GB | GF/DB |
| *Yersinia pseudotuberculosis* | ATCC 13979 | CR |
| *Yersinia pseudotuberculosis* | ATCC 6902, ATCC 29833 | GF/DB |
| *Yersinia pseudotuberculosis* | ATCC 29833 | KA |
| *Yersinia rohdei* | CCRI 14919 | GF/DB |
| *Yersinia ruckeri* | ATCC 29473 | GF/DB |

## 2.5 January – March, 2009

Optimization continued with runs adjusting the hybridization temperature. Overall, regardless of the conditions used, average signal intensity remained low. The chip protocol was finalized and is shown in Annex A. Since the protocol requires 5.0 ug of DNA, an amplification step was required for most samples. The Qiagen Repli-G Mini-kit was selected for the amplification procedure, following the instructions provided in the kit. Amplification yields were satisfactory and this product was used for all amplified samples. 20 chips were run during this quarter (12 unamplified optimization and 8 amplified samples). 2 chips failed due to air leakage in the fluidics system. Chip runs were postponed from Feb. 13 to Mar. 12 until fluidics station was repaired. Scanner gain settings were also checked in an attempt to increase signal intensity but were within setting parameters and were not adjusted. Additional genomic DNA from Cedarlane arrived Mar. 13/09 (*B. henselae*, *B. pertussis*, *C. jejuni, C. perfringens*, *Mycobacterium* sp. *BCG* and *M. pneumonia*). Kevin Ruttan trained on chip protocol to fill in for Melisssa Crichton during maternity leave.

## 2.6 April – June, 2009

A total of 79 chips were run using amplified samples. All samples were amplified and run twice to provide insight into the reproducibility of the protocol. Good correlation between replicates was observed but not quantified at this time. A new batch of chips and reagents was ordered. 84 chips were received Jun. 23/09.

## 2.7 July – September, 2009

New reagents arrived Jul. 3/09. Scanner failed during July $4^{th}$-$5^{th}$ chip run delaying scanning of 8 chips until scanner was repaired July 17/09. A total of 36 custom chips were run during this quarter. Of these, 29 were on amplified samples and 7 were mixed samples. Mixed sample experiments combined 2 samples (Table 3) at 10:90, 50:50 and 90:10 ratios to examine the effect of sample concentration on data. In addition, training on and run of 12 Tessarae chips was completed in Sept.

## 2.8 October – December, 2009

Mixed sample experiments continued with a total of 48 chips run this quarter. Four chips were lost due to scanner failure and runs were repeated. Scanner was down from Nov. 13 to Dec. 10. Data analysis on DRDCpath chips commenced in earnest. All cel files produced to date were verified for content and compiled into a master file. Data from chip 38355 (*E. coli*) determined to be unreliable. Various strategies for data pruning were examined.

*Table 3: Pathogen pairs used in mixed sample experiments.*

| | |
|---|---|
| *Bacillus anthracis* Vollum | *Francisella tularensis* Swed 7Q |
| *Yersina pestis* CO92 | *Yersinia pseudotuberculosis* ATCC 6902 |
| *Bacillus anthracis* NH | *Escherichia coli* 0157:H7 EDL933 |
| *Yersina pestis* C12 | *Escherichia coli* JM109 |
| *Francisella tularensis* Swed1 | *Escherichia coli* JM109 |
| *Burkholderia mallei* ATCC 1053 | *Escherichia coli* JM109 |
| *Burkholderia pseduomallei* Env. FM20 | *Escherichia coli* JM109 |
| *Listeria monocytogenes* NTCC 7937 | *Escherichia coli* JM109 |
| *Shigella dysentrine* ATCC 11835 | *Escherichia coli* JM109 |
| *Staphylococcus aureus* Z1 | *Escherichia coli* JM109 |
| *Streptococcus pyogenes* ATCC 19615 | *Escherichia coli* JM109 |
| *Yersina enterocolitica* CCUG 33553 | *Escherichia coli* JM109 |
| *Campylobacter jejuni* ATCC 700819 | *Escherichia coli* JM109 |
| *Clostridium perfringens* ATCC 13124 | *Escherichia coli* JM109 |
| *Haemophilus influenzae* ATCC 51907 | *Escherichia coli* JM109 |
| *Enterococcus faecalis* ATCC 29212 | *Escherichia coli* JM109 |
| *Salmonella typhimurium* 71-471 | *Escherichia coli* JM109 |
| *Vibrio vulnificus* Z28 | *Escherichia coli* JM109 |

## 2.9 January – March, 2010

Tessarae chips completed. Met with Dr. Christoph Sensen's team in Calgary to discuss methods of pruning and analyzing data. Data given to Paul Gordon who will determine a set of specific probes that provide useable data at 1/10 signal. Training and technology demonstration on IdahoTech film arrays was held in the lab during the week of Feb. 22. Contract extended until Mar. 31/11. Chip runs on hold awaiting additional reagents from Affymetrix.

## 2.10 April – June, 2010

Work was started on a manual approach to identifying a set of specific and sensitive probes from the custom chip. The background probe intensity was decided to be 325 (approximately $1/200^{th}$ of maximum obtainable signal). Looking at data obtained from all pure and mixed sample chips, any probe that did not register a signal above 325 on any of the chips was excluded from further investigation in this data set. Of the original 226576 probes on the chip, 162299 probes remained. It must be noted that many of the 64277 probes removed were Affymetrix negative controls, SNP mismatch sequences or designed to recognize a pathogen that was not available to us during this project's

timeframe and therefore cannot be excluded from future analyses. Four chips were run during this quarter.

## 2.11 July – September, 2010

Probes were further separated on the basis of scoring >325 intensity on at least one chip containing the pathogen the probe was designed to recognize. All probes that scored >325 for at least one chip with the appropriate pathogen were sorted into the "Pruned species +ve probe" worksheet. The remaining probes, having at least one signal >325 but not matching the pathogen the probe was designed to bind, were sorted into the "Pruned species -ve probes" worksheet. Due to common co-representation, probes designed to recognize *E.coli* and *Shigella* were considered species positive for either pathogen. Further specificity parameters for these two organisms would be examined at a later date. Of the 81721 specific probes, 26552 were positive for the designated organism. Again, many of the probes that were species negative were SNP mismatch or designed to recognize a pathogen that was not run during this project. Amplification reactions and cDNA quantification were completed for the 22 samples.

## 2.12 October – December, 2010

Amplification bias experiments are being completed, with triplicate runs of five 50:50 pathogen mixtures. For each pathogen pair, 1 chip was run on sample that were mixed pre-amplification, 1 chip for samples mixed post-amplification and 1 chip for a mixture of unamplified DNA. In addition, one set of chips will be run with a mixture of 5 different samples and another of all 10 samples used for these experiments (Table 4). During this quarter 16 chips were run.

*Table 4: Pathogen combinations used in amplification bias experiments.*

| | |
|---|---|
| *Acinetobacter* ATCC 17978 | *Bacillus cereus* ATCC10987 |
| *Bartonella henselae* ATCC 49882 | *Bordetella pertussis* ATCC BAA-589 |
| *Campylobacter jejuni* ATCC 700819 | *Clostridium perfringens* ATCC 13124 |
| *Haemophilus influenzae* ATCC 51907 | *Mycobacterium* sp. *BCG* ATCC 19015 |
| *Yersinia pseudotuberculosis* ATCC 13979 | *Listeria monocytogenes* ATCC 15313 |
| *Bacillus anthracis* RP42-A | *Yersinia enterocolitica* YE-D3 |
| 5 sample mix 1 | Aba/Bhe/Cje/Hin/Yps |
| 5 sample mix 2 | Bce/Bpe/Cpe/MBCG/Lmo |
| 10 sample mix | Aba/Bhe/Cje/Hin/Yps/Bce/Bpe/Cpe/MBCG/Lmo |

## 2.13 January – March, 2011

The amplification bias chips were completed and the effect of template dilution (limit of detection) was examined. Amplified and non-amplified *B*. *anthracis* strain RP42 was used for the dilution series, with serial 10-fold dilutions to 1/10000 (5.0 to 0.0005 ug DNA). DNA was isolated from castor bean samples and run on chips to determine whether the ricin gene could be detected. The species +ve probes were further sorted to identify those that are species specific (no other pathogen scores >1000 against that probe). Future work will examine whether the signal can be differentiated at 1/10 sample dilution.

# 3 Results

## 3.1 Optimization

Alterations in the proportions of the reaction mixtures did not have a significant impact on the resulting signal intensity values. The desired outcome of obtaining saturated signal for positive probes was not obtained with any reaction modifications attempted. The finalized protocol is shown in Appendix A

## 3.2 Pure samples

It was immediately apparent that probe set used for this custom chip could clearly identify the species of pathogens in all of the pure samples that were run. In the attached data files, three colour conditional formatting was applied to the intensity values to illustrate the specificity. All pure samples that were run during this project (including *Ricinus communis*) produced high intensity signals in a portion of probes that were designed to recognize that pathogen.

## 3.3 Mixed samples

The pathogen pairs shown in Table 3 were each combined in 10:90, 50:50 and 90:10 mixtures of amplified DNA. Data indicated that, in general, addition of $1/10^{th}$ of the starting material resulting in approximately $1/10^{th}$ of the signal intensity thus limiting the usefulness of low intensity specific probes. Although high signals can be observed from probes for both pathogens in some chips, algorithm development or use of minimum cut-off values would facilitate data analysis of mixed samples.

## 3.4 Amplification bias

Equal amount of DNA of the pathogens shown in Table 4 were combined unamplified, prior to amplification or after samples were individually amplified to examine the effect of the amplification procedure on the resulting data. This set of experiments demonstrated that amplification definitely affect the results obtained for each sample. Not surprisingly, combining samples post-amplification resulting in the best representation of high intensity signals for both pathogens. Non-amplified and pre-amplification mixed samples tended to favour one pathogen over the other in the 2 sample experiment. It must also be noted that the same pathogen was not necessarily favoured in both cases. As additional samples were added, some of the pathogens were not represented at all in the high intensity probes. Some of the loss of sensitivity can be attributed to the wide disparity in the number of probes designed for each pathogen.

## 3.5 Dilution bias

*Bacillus anthracis* sample RP42 was chosen for the dilution series due to the abundant supply of DNA on hand. Unfortunately, this sample proved to be a poor choice as the maximum signal intensity obtained for the undiluted samples were generally quite low. This resulted in a significant signal drop-off at 1/10 dilution and background signal for all further dilutions. Again, 1/10 sample produced roughly 1/10 signal intensity.

# 4 Data analysis

Raw data obtained from the GeneChip scanner was saved into subfolders sorted by experiment type, pathogen and chip number. For each chip, total intensity data for all 226,576 spots was obtained using Affymetrix power tools .cel extraction application. The resulting text file provided the probe #, probe x and y coordinates and intensity value. This information was copied into a master Excel file (e-copy on disk attached).

To facilitate data analysis, all probes that produced intensity values less than 326 for all chips run were excluded from further analysis. Since some of these probes were designed for organisms that were not run in this experimental set, they cannot be excluded from all future chip experiments.

A number of different approaches to data analysis were investigated. Most were very time consuming based on manual examination of probe data and individual comparisons of signal values obtained versus the organism expected to be recognized. These approaches were determined to be unfeasible for any regular application of this technology.

The approach utilized in the accompanying data files identifies probes that are specific and sensitive for the organisms run. This approach provides a good selection of probes that could be used as a baseline for future chip work by applying a mask to forthcoming data to select only these probes. One drawback of this approach is data from the Affymetrix controls and random probe set are filtered out.

Multiple data files were constructed to separate the probe data at various levels. The first contains all chip data with all probes. The second contains those probes remaining after the <326 pruning is applied, leaving 162,299 probes. At this stage, data from *E.coli* chip 38355 were removed due to poor correlation with other *E.coli* data. This changed the selection of probes in the original <326 pruning and resulted in 162,751 probes. From this stage, several columns were added. The first column identifies probes produced signal above background for at least one chip hybridized to the pathogen that the probe was designed to recognize (species positive). Pathogen recognition was based on the genus or species level depending on the organism. This procedure removed the Affymetrix and random probe sets. Due to sequence commonalities, E. coli and Shigella

were grouped together in this sorting. In a number of cases, probes registered very high intensity values for a different organism than it was designed to recognize and were in some cases reclassified. This change is recorded in the "Description" column as xxx originally yyy. A total of 55,839 probes remained after this step.

To identify species specific probes, several classifications were used. Probes were categorized as species specific (y) if the signal for designated organism was greater than 1000, the signal for non-designated organisms was less than 1000 and difference between designated and non-designated organisms was less than background (325). Probes were negative (n) if the intensity value for a non-designated organism exceeded that of the designated organism or the difference between the two was less than background. Some probes were designated (m) if the signal for the both the designated and a low number of undesignated organisms was greater than 1000. This category can be further refined to identify differential probes. Finally, probes that are expected to hit multiple organisms (e.g. antibiotic resistance sequences) were designated N/A and carried onto the next level of analysis. A total of 15,635 remained after this step.

To determine the best probe set for species specificity, the species specific data was used construct the set of probes having a maximum specific signal >5000. In the data file, conditional formatting used to show probes with signal between 1000-4999 (green), 5000-9999 (orange) and >9999 (yellow). Probes with signal <1000 were not coloured. The final tally of probes was 3119. Data analysis indicated that it is probable that the *Pseudomonas aeruginosa* samples were contaminated with *Bacillus cereus* sample ATCC 11778 due to their probe co-expression. By taxonomy, *Pseudomonas aeruginosa* would be expected to more closely resemble *Acinetobacter* than *bacillus.*

The breakdown of probes by level of pruning and pathogen is shown in Table 5.

*Table 5: Probe count by pathogen in different data sets.*

| Pathogen | Total pruned data set | Pruned data set exluding Eco 38355 | Species specific | # probes w/ max signal >5000 |
|---|---|---|---|---|
| *Acinetobacter baumannii* | 267 | 497 | 234 | 34 |
| Affymetrix | 106645 | 106912 | 0 | 0 |
| *Bacillus* sp. | 2911 | 3340 | 1203 | 180 |
| *Bartonella* sp. | 689 | 1031 | 552 | 227 |
| *Bordetella* sp. | 1297 | 1422 | 507 | 97 |
| *Borellia* | 3 | 2 | 0 | 0 |
| *Brucella* sp. | 1099 | 603 | 0 | 0 |
| *Burkholderia* sp. | 1577 | 4654 | 2048 | 79 |
| *Campylobacter* sp. | 1453 | 1292 | 501 | 120 |
| *Chaetomium* sp. | 147 | 106 | 0 | 0 |
| *Chlamydia* sp. | 424 | 307 | 0 | 0 |
| *Clostridium* sp. | 678 | 778 | 249 | 36 |
| *Corynebacter* sp. | 333 | 205 | 0 | 0 |
| *Coxiella* | 173 | 121 | 0 | 0 |
| *E.coli* | 2496 | 3516 | 1214 | 264 |
| *Enterococcus faecalis* | 99 | 169 | 55 | 12 |
| *Francisella* sp. | 8053 | 7838 | 1579 | 91 |
| *Haemophilus* sp. | 1600 | 1389 | 470 | 165 |
| *Helicobacter* sp. | 956 | 685 | 0 | 0 |
| *Homo sapiens* | 47 | 35 | 0 | 0 |
| *Klebsiella* | 4 | 1 | 0 | 0 |
| *Lactobacillus* | 4 | 4 | 0 | 0 |
| *Legionella* | 942 | 666 | 0 | 0 |
| *Listeria* | 2150 | 2176 | 565 | 158 |
| multispecies/antibiotic resistance | 31 | 27 | 0 | 0 |
| *Mycobacterium* sp. | 4963 | 3333 | 348 | 41 |
| *Mycoplasma* sp. | 539 | 600 | 111 | 26 |
| *Neisseria* | 846 | 507 | 0 | 0 |
| Plasmids | 19 | 43 | 4 | 0 |
| *Pseudomonas* sp. | 9891 | 5464 | 144 | 1 |
| *Ricinus* | 6 | 87 | 41 | 5 |
| *Ricketsia* | 93 | 65 | 0 | 0 |
| *Salmonella* sp. | 2126 | 2722 | 1084 | 434 |
| *Shigella* sp. | 1762 | 1280 | 153 | 26 |
| *Staphylococcus* sp. | 966 | 977 | 449 | 30 |
| *Streptococcus* sp. | 1648 | 1397 | 249 | 136 |
| *Treponema* | 34 | 20 | 0 | 0 |
| *Ureaplama* | 3 | 2 | 0 | 0 |
| *Vibrio* sp. | 2712 | 1817 | 41 | 2 |
| *Xanthomonas* | 5 | 4 | 0 | 0 |
| *Yersinia* sp. | 2608 | 6657 | 3834 | 955 |
| **TOTAL PROBE COUNTS** | **162299** | **162751** | **15635** | **3119** |

# Annex A GeneChip Protocol for DRDCpath Custom Chips (per reaction)

**dUTP Incorporation – Using ROCHE kit (no amplification protocol)**

1. Add 16 ul containing 5.0 ug of sample DNA to PCR tube. Incubate at 95 C for 10 min. (Program 95HOLD)
2. During the above incubation, prepare the following (Incorporation mix):

   0.8 ul $dH_20$
   0.8 ul of 1 mM dUTP
   1.6 ul of 0.5 mM dTTP
   2.0 ul of 0.5 mM dATP
   2.0 ul of 0.5 mM dCTP
   2.0 ul of 0.5 mM dGTP

3. Vortex the mixture briefly.
4. When the incubation is almost finished, add
   2.0 ul of the hexamer primer reaction mix
   1.0 ul of Klenow enzyme (keep on ice until use)
   Vortex the mixture briefly and keep on ice until use.

5. Add 12.2 ul (total of above mixture) to the cooled sample DNA. Vortex briefly.
6. Incubate the sample at 37 C for 2 hrs (Program AFFY1)

**Fragmentation – Using AFFY GeneChip WT Terminal Labelling Kit**

7. Incubate sample tube at 95 C for 10 min. (Program 95HOLD)
8. When this incubation is almost finished, prepare the following (Fragmentation mix):
   10.0 ul RNase-free $H_20$
   4.8 ul of Fragmentation buffer
   1.0 ul of 10 U/ul UDG enzyme
   1.0 ul of 1000U/ul APE1 enzyme
   Vortex briefly and keep on ice until needed.
9. Add 16.8 ul of Fragmentation mix (total of above mixture) to the cooled sample DNA. Total volume should be 45 ul.
10. Incubate mixture at 37 C for 1 hour (Program AFFY2)

**Labelling - Using the AFFY GeneChip WT Terminal Labelling Kit**

11. When this incubation is almost finished, prepare the following (Labelling mix):
    12.0 ul 5x TdT buffer
    2.0 ul TdT
    1.0 ul DNA labelling reagent
    Vortex briefly and keep on ice until used.

12. Add 15.0 ul (total of above mixture) to cooled sample mixture. Final volume should be 60 ul.
13. Incubate at 37 C for 1 hour (Program AFFY3).
14. Preset the heating blocks to 99 C and 45 C

**Hybridization – Using AFFY GeneChip Hybridization, Wash and Stain Kit**

15. When the above incubation is almost finished, incubate the 20x Eukaryote Hybridization Control at 65 C for 5 min.
16. Collect the sample mixture and prepare the following (hybe mix):
    60.0 ul sample mixture
    11.0 ul warmed 20x Eukaryote Hybridization Control
    3.7 ul B2 Oligo control
    15.4 ul DMSO
    110.0 ul 2X Hybe Buffer Mix
    20.0 ul $dH_20$
    Vortex briefly.

17. Heat the mixture at 99 C for 5 min. then cool to 45 C for 5 min.
18. Vortex then centrifuge mixture at maximum speed for 1 min.
19. Load 200 ul of mixture into Array. Cover loading holes with Tough spot stickers.
20. Record the array and sample numbers in all documentation.
21. Place array in hybe oven at 42°C, 60 rpm for 18 hrs.

**Washing, Staining and Scanning – Using AFFY GeneChip Hybridization, Wash and Stain Kit**

22. Remove array from hybe oven, remove tough spot stickers and hybridization mixture from array and refill with 250 ul of WASH Buffer A.
23. Place array in Fluidics Node and run fluidics program FS450_0005.
24. Scan chips using the GeneChip 4200 Scanner.
25. Store chip data files in individually named subfolders in the format: "Chip#"_"Pathogen"_"Strain"_"Date scanned"_"Amp#"

# Annex B Attached Data Files (on disk)

See files entitled:

All probe data all chips FINAL (360 MB)
All chips pruned data FINAL (264 MB)
All chips species positive FINAL (263 MB)
All chips summary table (150 MB)

This page intentionally left blank.

## DOCUMENT CONTROL DATA

(Security classification of title, body of abstract and indexing annotation must be entered when the overall document is classified)

1. ORIGINATOR (The name and address of the organization preparing the document. Organizations for whom the document was prepared, e.g. Centre sponsoring a contractor's report, or tasking agency, are entered in section 8.)

DRDC Suffield

2. SECURITY CLASSIFICATION (Overall security classification of the document including special warning terms if applicable.)

UNCLASSIFIED
**(NON-CONTROLLED GOODS)**
**DMC A**
**REVIEW: GCEC JUNE 2010**

3. TITLE (The complete document title as indicated on the title page. Its classification should be indicated by the appropriate abbreviation (S, C or U) in parentheses after the title.)

Detection/Identification of Microbes Using Microarrays

4. AUTHORS (last name, followed by initials – ranks, titles, etc. not to be used)

C.C. Rattan, D.C.W. Mah

| 5. DATE OF PUBLICATION (Month and year of publication of document.) | 6a. NO. OF PAGES (Total containing information, including Annexes, Appendices, etc.) | 6b. NO. OF REFS (Total cited in document.) |
|---|---|---|
| July 2012 | 22 | 0 |

7. DESCRIPTIVE NOTES (The category of the document, e.g. technical report, technical note or memorandum. If appropriate, enter the type of report, e.g. interim, progress, summary, annual or final. Give the inclusive dates when a specific reporting period is covered.)

Contract Report

8. SPONSORING ACTIVITY (The name of the department project office or laboratory sponsoring the research and development – include address.)

DRDC Suffield, Box 4000, STN MN, Medicine Hat, Alberta T1A 8K6

| 9a. PROJECT OR GRANT NO. (If appropriate, the applicable research and development project or grant number under which the document was written. Please specify whether project or grant.) | 9b. CONTRACT NO. (If appropriate, the applicable number under which the document was written.) |
|---|---|
| | W7702-R16 |

| 10a. ORIGINATOR'S DOCUMENT NUMBER (The official document number by which the document is identified by the originating activity. This number must be unique to this document.) | 10b. OTHER DOCUMENT NO(s). (Any other numbers which may be assigned this document either by the originator or by the sponsor.) |
|---|---|
| DRDC CR 2012-018 | |

11. DOCUMENT AVAILABILITY (Any limitations on further dissemination of the document, other than those imposed by security classification.)

Unlimited

12. DOCUMENT ANNOUNCEMENT (Any limitation to the bibliographic announcement of this document. This will normally correspond to the Document Availability (11). However, where further distribution (beyond the audience specified in (11) is possible, a wider announcement audience may be selected.))

Unlimited

13. ABSTRACT (A brief and factual summary of the document. It may also appear elsewhere in the body of the document itself. It is highly desirable that the abstract of classified documents be unclassified. Each paragraph of the abstract shall begin with an indication of the security classification of the information in the paragraph (unless the document itself is unclassified) represented as (S), (C), (R), or (U). It is not necessary to include here abstracts in both official languages unless the text is bilingual.)

Microarray technology provides the potential for rapid identification of an unknown pathogen in a single test. In order to explore the utility of this tool, a custom Affymetrix DNA microarray chip was designed to recognize a wide variety of bacteria to the species and strain level. Antigenomic probes were included on the lower portion of the chip to investigate the potential of rapid visual assessment of organisms' footprints. Since pre-existing protocols and analysis software were unavailable for this work, protocols were optimized and a variety of data analysis options were investigated. Experiments were conducted on a number of pure DNA samples and mixed samples with equal quantities of input DNA. The effect of DNA amplification and sample dilution on signal intensity was also investigated.

The chips were readily able to identify the correct pathogen from pure samples but signal levels dropped with decreased input DNA. Amplified mixed DNA, both pre-amplification and post-amplification sample mixtures, produced different results from mixed unamplified DNA requiring more complex analysis technique or a better balancing of the number of probes designed for each organism. The dilution series demonstrated that signal intensity reduction corresponds to the decrease in template DNA (i.e. 1/10 DNA has 1/10 signal intensity). In all, 3,119 probes were selected that are strongly specific and sensitive for the organisms tested and these provide an excellent basis for future work with microarrays in the identification of bacterial pathogens.

14. KEYWORDS, DESCRIPTORS or IDENTIFIERS (Technically meaningful terms or short phrases that characterize a document and could be helpful in cataloguing the document. They should be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location may also be included. If possible keywords should be selected from a published thesaurus, e.g. Thesaurus of Engineering and Scientific Terms (TEST) and that thesaurus identified. If it is not possible to select indexing terms which are Unclassified, the classification of each should be indicated as with the title.)

microarray; microbial; microbial identification; molecular biology; Affymetrix

**Defence R&D Canada**

Canada's Leader in Defence and National Security Science and Technology

**R & D pour la défense Canada**

Chef de file au Canada en matière de science et de technologie pour la défense et la sécurité nationale


**www.drdc-rddc.gc.ca**