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

Microarray Genomic Fingerprinting

Barry N. Ford, Yimin Shei, Janice Bamforth
DRDC Suffield

Technical Memorandum
DRDC Suffield TM 2005-244
December 2005

Canada

Microarray Genomic Fingerprinting

Barry N Ford, Yimin Shei, Janice Bamforth
DRDC Suffield

Defence R&D Canada – Suffield

Technical Memorandum

DRDC Suffield TM 2005-244

December 2005

BFord Author

Barry N. Ford

Approved by

for

Les Nagata

Head Chemical and Biological Defence Section

Approved for release by

P. D'Agostino

Paul D'Agostino

Chair Document Review Panel

Abstract

Using current molecular biology, it is possible to create novel microbial forms which have never existed in nature. Current methods for identifying and characterizing microbes depend on some prior assumptions about the genetic content of the organism. Recombinant organisms containing novel genetic material by definition fall outside of the range of prior sequence data. Thus new methods which do not require such assumptions are required to screen and identify the strains, and to detect the presence of novel genetic material. Using an oligonucleotide microarray, we have executed a proof of concept of an array-based genomic fingeprinting technology. The test organisms for this work were *Escherichia coli* (4 strains), *Bacillus anthracis* (2 strains), and *Yersina enterocolitica*. Using standard molecular biology methods, we isolated genomic DNA, digested the DNA to reduce complexity, labelled it with fluorescent dyes, and hybridized the labelled DNA to microarrays containing 21,000 unique olignucleotide features. From a single grid of 484 features, or from a filtered subset of the hybridization data, species could be readily discriminated with high confidence. Strain differentiation may require analysis of the entire feature map, or refinement of the array sequences features, the analysis model, or the analysis software. The next phase of this work will be a test system for rapid species identification in addition to the oligonucleotide fingerprint, leading to a finalized design for a prototype genomic fingerprinting microarray.

Résumé

La biologie moléculaire actuelle permet de créer des nouvelles formes microbiennes qui n'ont jamais existé naturellement. Les méthodes actuelles qui identifient et caractérisent les microbes dépendent de certaines hypothèses précédentes concernant le contenu génétique d'un organisme. Les organismes recombinants qui contiennent des matériaux génétiques ne sont par définition pas contenus dans l'éventail des données précédentes de séquences. Par conséquent, il faut trouver des nouvelles méthodes qui ne requièrent pas de telles hypothèses pour calibrer et identifier les souches et pour détecter la présence de nouveau matériel génétique. Au moyen d'un microréseau oligonucléotide, nous avons exécuté une validation de principe de la technologie de la dactyloscopie génomique à base de réseau. Les organismes du test pour ces travaux étaient *Escherichia coli* (4 souches), *Bacillus anthracis* (2 souches), et *Yersina enterocolitica*. En utilisant les méthodes standard de biologie moléculaire, on a isolé l'ADN génomique, digéré l'ADN pour réduire la complexité, marqué avec des teintures fluorescentes et hybridé l'ADN marqué avec des microréseaux contenant 21 000 caractéristiques oligonucléotides uniques. À partir d'une seule grille de 484 caractéristiques ou bien à partir d'un sous-ensemble de données d'hybridation, les espèces pouvaient être discriminées facilement avec haute fiabilité. La différentiation des souches peut exiger de faire l'analyse du tableau des caractéristiques au complet ou bien de raffiner les caractéristiques des séquences de réseaux, du modèle des analyses, ou du logiciel d'analyse. La prochaine phase de ces travaux consistera en un système de tests pour l'identification rapide des espèces en plus de la dactyloscopie oligonucléotide ce qui aboutit à un concept finalisé du prototype d'un microréseau de la dactylotechnie génomique.

This page intentionally left blank.

Executive summary

Introduction: Identification of microbes to the species level using conventional microbiology is a well established technology. Molecular techniques based on direct genomic sequence information have enabled the rapid resolution of strain or substrain differences, even on many nonviable samples. However, the ease with which recombinant organisms can be created leads to a new problem. How can we identify and characterize recombinant species and strains with novel genetic content? Conventional microbiology will, in general, miss recombinants, and most molecular methods require a priori knowledge of the genomic structure of the suspect organism. Recombinants may easily contain genetic sequences either not frequently encountered in library strains, or not readily detected by existing techniques.

A possible approach demonstrated here is to use a large panel of non-specific oligonucleotides in a microarray hybridization format, to interrogate genomic DNA from unknowns. The assay output is a reproducible pattern of hybridization, a “fingerprint”, on the microarray slide. Comparing the hybridization of an unknown to a library of known strains allows for the direct examination or computational comparison of hybridization patterns.

Results: Different species of bacteria, including *Escherichia coli*, *Bacillus anthracis*, and *Yersinia enterocolitica* produce qualitatively different hybridization patterns on the microarrays. Different modes of comparison ranging from individual hybridization spots to a global assessment of hybridization pattern, highlight species differences in different ways. Strain-specific differences can also be detected by inspection of array patterns, albeit with lower confidence.

Significance: This technique requires no prior knowledge of the organism's genetic content. Additionally, strains containing novel genetic sequences relative to library strains, should be detectable as hybridization patterns different from those observed for libraries. In this proof on concept, it is demonstrated that differential hybridization patterns are indeed produced by different species. Preliminary data suggest that, given appropriate software, strain differences may also be detectable.

Future Plans: Further refinement of the technology, including a rapid species-specific genotyping array, combined with the fingerprinting approach, should enable a detailed assessment of the value of this technology. Software to facilitate quantitative comparisons will be developed.

Ford, B.N., Shei, Y., Bamforth, J. 2005. Microarray genomic fingerprinting. DRDC Suffield TM 2005-244. Defence R&D Canada – Suffield.

Sommaire

Introduction : L’identification des microbes au niveau de l’espèce au moyen de la microbiologie classique est une technologie bien établie. Des techniques moléculaires basées sur l’information tirée directement des séquences génomiques ont permis la résolution rapide des différences de souches ou sous-souches, même sur beaucoup d’échantillons non viables. La facilité avec laquelle on peut créer des organismes recombinants donne cependant naissance à un nouveau problème. Comment est-il possible d’identifier et de caractériser des espèces et des souches recombinantes avec un nouveau contenu génétique? La microbiologie traditionnelle ne reconnaît normalement pas les recombinants et la plupart des méthodes moléculaires exigent une connaissance préalable de la structure génomique de l’organisme suspect. Des recombinants peuvent facilement contenir des séquences génétiques qui ne sont pas fréquemment trouvées dans les couches documentées ou bien qui ne sont pas facilement détectées par les techniques existantes.

On démontre ici une méthode potentielle qui consiste à utiliser un grand éventail d’oligonucléotides non spécifiques sous le format d’hybridation de microréseaux afin d’interroger le DNA génomique à partir des inconnus. Le résultat du biotest est un modèle d’hybridation reproductible, une « empreinte », sur la lame du microréseau. Comparer l’hybridation d’une souche inconnue à une bibliothèque de souches connues permet d’examiner ou de calculer directement la comparaison des modèles d’hybridation.

Résultats : Différentes espèces de bactéries dont *Escherichia coli*, *Bacillus anthracis* et *Yersinia enterocolitica* produisent des modèles d’hybridation qualitativement différents sur les microréseaux. Des différents modes de comparaison allant de l’évaluation de taches individuelles d’hybridation à celle d’une évaluation globale des modèles d’hybridation, soulignent de différentes façons les différences chez les espèces. Les différences spécifiques aux souches peuvent aussi être détectées en inspectant les modèles de réseaux, ceci avec moins de fiabilité.

La portée des résultats : Cette technique n’exige pas une connaissance préalable du contenu génétique de l’organisme. De plus, les couches contenant des séquences génétiques nouvelles liées aux souches documentées devraient être détectables comme modèles d’hybridation différents de ceux observés dans la bibliothèque. Dans cette validation de principe, on démontre que les espèces différentes produisent en fait des modèles différentiels d’hybridation. Les données préliminaires suggèrent que, avec un logiciel approprié, les différences dans les souches peuvent aussi être détectables.

Plans futurs : Un raffinement plus approfondi de la technologie, dont un réseau de typage génique rapide spécifique à des espèces, combiné à la méthode de dactylotechnie, devrait permettre une évaluation détaillée de la valeur de cette technologie. Des logiciels visant à faciliter les comparaisons quantitatives seront mis au point.

Ford, B.N., Shei, Y., Bamforth, J. 2005. Microarray genomic fingerprinting. DRDC Suffield TM 2005-244. R & D pour la défense Canada – Suffield.

Table of contents

List of figures

Acknowledgements

The authors gratefully acknowledge the assistance of Dr. John Cherwonogrodzky and Ms. Nicole Stady in the culturing of and preparation of genomic DNA extracts from *Bacillus anthracis* and *Yersinia enterocolitica.* Ms. Janice Bamforth completed many of the hybridizations in the data set during her tenure as a summer Co-Op student at DRDC Suffield.

This page intentionally left blank.

Introduction

Novel pathogenic organisms containing toxin genes or other alterations are a poorly understood threat in biological warfare or bioterrorism. The potential to create organisms which are more virulent [1] or have altered host specificity means that tools to detect and characterize recombinants are needed. Current technology to identify species and strain utilize known genetic or phenotypic data, based on the assumption that these represent all of the information relevant to the organism's biology. Estimation of the relatedness of a novel organism to known ones, including new combinations of host with toxin or other genes, may involve a battery of assays including conventional microbiology, DNA sequence analysis, or in vivo infectivity assays. Spontaneous natural acquisition or deliberate insertion of novel genetic material which may confer novel phenotypes is, by definition, beyond the assumptions of existing identification techniques. In the case of novel pathogens this data is either not available, or will be misleading due to confusion with typical organisms.

Methods in current use, including comparative genomic hybridization, multi-locus sequence analysis, RFLP, and PCR can detect the presence or absence and variability of genetic sequences which are already known and included in the reference genomic sequences [2, 3]. Such approaches are much less useful for detecting replaceable sequences or large fragments (*e.g.* episomes, plasmids) which are not included in the reference genome nor in the genome of any related species [4-6]. The size of the replaceable genome can be substantially greater than the size of the reference genome [6]. Methods for creating recombinants in organisms which already exist on lists of potential biowarfare agents are now available or published in the open literature [6-11].

Recent concerns about the ease with which novel genetically modified or chimeric microbes can be produced, highlight the deficiencies of current tools. These deficiencies include, but are not confined to:

- novel strains may be misidentified as existing ones
- lack of simultaneous testing for multiple strains/species/recombinants
- specificity of current tests traded off against sensitivity
- DNA-based tests require knowledge of DNA sequences of the microbe
- no simple assessment of the degree of threat for recombinants/chimeras

A primary source of the deficiencies described above is the limitation in the number of assays that can be performed on a single sample. Since sample source material is usually limited, the ability to test for many features on a single test platform in one pass could be a tremendous advance. Microarrays, would permit the screening of literally thousands of DNA sequences for comparison to known laboratory strains, and could provide rapid identification of virtually any known microbial/viral threats.

Current methods for microbial identification rely either on conventional microbiology or species and strain-specific identification tools. Fieldable kits for the rapid detection and identification of select organisms are robust, useable by minimally trained personnel, and simple to interpret. However, these approaches assume that a given organism is present, and are not useful as screening tests or investigative tools. Multiple iterations or multiplex tests can be a partial solution to this problem, but false positives or false negatives can occur, and there is no simple way to survey for possible novel organisms containing components from other microbes (*e.g.* antibiotic resistance, toxins).

In order to complement existing techniques, we are attempting to develop an assumption-free species/strain identification tool based on microarrays. The penultimate design will integrate microarray chips which carry known sequences derived from several hundred known genetic targets, representing known toxin, virulence, and host strain genes as well as several thousand "random" defined oligonucleotides which do not correspond to known genetic targets in microbes. In the first case, this is an extension of existing technology, which is based upon the assumption that an arbitrary microbial sample will be related to some species or strain which is already known. Systems based on this approach are in development for clinical diagnostics. But this approach is limited by assumptions that explicitly ignore the possibility that any given pathogen may in fact contain atypical (natural or deliberate) modifications within its genetic material.

In the second case of defined random oligonucleotides, any given organism should hybridize to the oligos in a unique way, which we will call the fingerprint. This approach has been previously applied using strain-specific short oligonucleotide features to discriminate closely related strains within a species [11-14]. Unlike conventional fingerprint analysis which uses a few dozen reference points, microarray fingerprints contain thousands of reference points in the hybridization features. Across thousands of features, such a pattern (unless highly erratic) would be difficult to interpret by examination. Using statistical or cluster analysis, however, the hybridization patterns (fingerprints) of unknown microbial samples across thousands of features can be digitized, and compared to a library of known fingerprints. Each hybridization may contain multiple subpatterns which can be discerned and clustered according to relative intensity. Thus fingerprinting on microarray platforms is an informatics problem rather than a data collection problem.

In this experiment, a proof of concept is presented of genomic fingerprinting by hybridizing labelled genomic DNA to heterospecific oligonucleotides, exploiting existing technologies and platforms. Various modes of presenting and visualizing the data are presented. Species and strains can be differentiated by visual or clustering comparison of fingerprint patterns at different levels of resolution and effort.

Materials and Methods

Genomic DNA Extraction

Genomic DNA from *Bacillus anthracis* (RP42 and Thraxol), and *Yersinia enterocolitica* were obtained from the DRDC Suffield BSL3 laboratory, prepared by Dr. John Cherwonogrodzky and Nicole Stady. The method used in genomic DNA extraction from Yersinia enterocolitica was essentially the same as for the *E. coli* DNA.

Bacillus anthracis

Bacillus anthracis Thraxol (Sterne) was taken from a stock culture and used to seed a nutrient agar plate (incubated 35 ºC, 5% CO_2, 90% humidity). *Bacillus anthracis* RP42 taken from a stock original culture from France (on a small paper disk), was used to seed a nutrient agar plate, which was incubated overnight at 35 ºC. Cells were scraped off (by rolling with glass beads), and suspended in 10 ml (per plate) of TRIS-EDTA buffer (50 mM TRIS, 50 mM EDTA pH 9.71). Approximately 100 ml of cell suspension (from 10 plates in total) was harvested, then left overnight at -70 ºC. For each genomic DNA preparation, a 1/10th volume (10 ml) of 250 mM TRIS (pH 7.2) with 100 mg of lysozyme was added to the frozen suspension. The suspension was then thawed at room temperature. When thawed, the suspension was placed on ice for 45 min with occasional gentle swirling. Twenty ml of extraction buffer (0.5% SDS, 50 mM TRIS, 0.4 M Na_2EDTA, pH 7.32, 1 mg/ml proteinase K) was added. The mixtures were placed in a 50 ºC water bath for 60 min. The cell suspensions were then extracted with an equal volume of TRIS saturated phenol (room temp, gentle stirring with magnetic stir bar for 10 min), then the suspensions were centrifuged at 10,000 x g for 15 minutes at 4 ºC. The upper (aqueous) layer was transferred to a fresh bottle, and a 1/10 volume of 3 M sodium acetate was added with gentle mixing, followed by 2 volumes of 95% ethanol.

Sterility checks were performed on the *B. anthracis* ethanol suspensions. One ml of the suspension was added to each of 3 nutrient broth flasks which in turn were used to inoculate each of 3 blood agar plates. Each plate was incubated was for 1 week at 35 ºC, in a 5% CO_2 atmosphere with 90% humidity. When sterility was verified by absence of growth, the suspensions were transferred from the BSL3 suite, centrifuged at 10,000 x g to precipitate genomic DNA, then washed with 70% ethanol and dried briefly at room temperature before resuspending for 20 minutes in 8mM NaOH. Finally the solutions were neutralised with TE pH 8.0 (10 mM TrisHCl, 1 mM EDTA).

Escherichia coli / Yersinia enterocolitica

Except where otherwise noted, reagents used were obtained from SIGMA (Oakville, ON). DNA from *Escherichia coli* (strains JM108, JM109, DH5α, TOP10, see Annex for genotypes) grown overnight at 37 °C in 50 ml of Luria broth (LB) was extracted by first pelleting the bacteria at 1500 x g, and treating with 1 ml lysozyme solution (5 mg/ml lysozyme in 50 mM glucose, 10 mM EDTA, and 25 mM Tris HCl, pH 8.0). This was followed by 19 ml of DNAzol (Invitrogen, Carlsbad, CA) for 20 minutes with gentle mixing. Insoluble polysaccharides, proteins, and lipids were pelleted at 10,000 x g for 45 minutes, and the supernatant was collected. DNA was precipitated from the DNAzol by addition of ethanol to a final 30% and centrifugation at 5000 x g for 35 minutes. The DNA pellet was washed twice with 20 ml of 75% ethanol and dried for 5 minutes. DNA was resuspended in 250 µl of 8 mM NaOH at 30 °C for 5 minutes, followed by neutralization with an equal volume of TE (10 mM TrisHCl, 1 mM EDTA, pH 7.5). Undissolved DNA was pelleted by centrifugation at 5000 x g the next day and discarded. DNA concentration and quality in the supernatant was estimated by optical density measurements at OD_{260}/OD_{280}.

DNA Restriction Digest and RNase Treatment

DNA was digested by EcoR I (GE Healthcare, Baie dUrfe, PQ), followed by enzyme inactivation at 65 °C, and RNase treatment at 37 °C. Digested, RNA-free DNA was precipitated with 2.5 volumes of isopropyl alcohol and a 1/10 volume of 3M sodium acetate (pH 5.5), followed by two 75% ethanol washes. The DNA pellet was dissolved in 8 mM NaOH and then an equal volume of TE was added. The DNA was quantified by the SYBR Green I assay in 96-well microtitre plates, adapted from the Quant-iT™ DNA Assay Kit (Invitrogen). Sheared salmon testes DNA (Sigma) was calibrated against a DNA fluorescence standard (GE Healthcare) to 0.1 µg/µl, and serially diluted in TE that was supplemented with 2x SYBR Green I to generate a concentration gradient of 100, 50, 25, 12.5, and 6.25 ng/ml. DNA samples were diluted in the same TE-SYBR Green I solution as the standards. The standards and the unknowns were assayed at excitation 485 nm and emission 538 nm. DNA concentration was calculated from the standard curve fluorescence readings. DNA quality was also assayed by running a 1 % agarose gel (Sigma).

DNA Labelling / Hybridization

Following final cleanup, digested DNA was labelled with Cy3 (GE Healthcare) using the Random Primed DNA Labelling Kit (Roche, Laval, PQ). 0.5 µg DNA (assayed by the SYBR Green I method) in 12.4 µl water was denatured at 95 °C for 10 minutes and snap cooled on ice. 4.6 µl of Cy3-dNTP (4 nmoles Cy3-dCTP, 4 nmoles dCTP, 5nmoles dATP, 5nmoles dTTP, and 5 nmoles dGTP) was added to the DNA, together with 2 µl hexanucleotide primer mix and 1 µl Klenow enzyme (2 Units). The DNA was labelled at 37 °C for 1 hr, and stored at -20 °C until hybridization.

Microarray slides were spotted with oligonucleotides from the Operon Genetics Human Genome Oligo Set Version 2.0 (Huntsville, AL). Slides were prepared under contract at the Prostate Cancer Research Centre Microarray Facility, at Vancouver General Hospital (Vancouver, BC), under the direction of Dr. Colleen Nelson. For hybridization to the microarray, 80 µl of hybridization buffer (10% dextran sulfate, 1 M NaCl, 1% SDS) were added to the labelled DNA, and heated to 95 °C for 5 minutes. After cooling to 65 °C, 80 µl of the hybridization mixture were applied to the microarray slide under a LifterSlip (VWR, Mississauga, ON). Hybridization was conducted at 45 °C for 18 hours in a humidified chamber. After hybridization, slides were immediately dipped into 1x SSC (0.15 M NaCl, 0.015 M sodium citrate) with 0.1% SDS at room temperature. After removing the cover slip, slides were washed in 0.5 x SSC with 0.05% SDS and spin dried for 10 minutes at 250 x g in 50 ml tubes.

Microarray Digitization

Dried microarray slides were scanned at 550 nm excitation with 580 nm emission in the DNAScope IV array scanner (Biomedical Photometrics, Waterloo, ON). Microarray images obtained from the scanner were digitized using GenePix Pro (TM) software from Axon Genetics (Palo Alto, CA). Images were also saved as jpeg files for manual examination. Within Genepix Pro, the user specifies the number of spots (features) in the array, the number of subgrids, feature dimensions and shape, and upper or lower boundaries for maximum or background values. The software automatically aligns the defined grid to the imaged microarray. The alignments were inspected and manually refined by the user. Digitized values for each feature were then extracted and exported as an Excel-readable flat text file for computer analysis. For each experiment, data were organized in the spreadsheet by sample (columns) and feature intensity (rows). For replicates within each species or strain, an averaged dataset was created by averaging each row of intensites in a new column.

Image Analysis

Images of the hybridized microarrays were compared directly by examination, including a grid-based comparison of hybridization patterns. The digitization software yields intensity values (average intensity over a fixed number of pixels) to three decimal places in a range from 0 to 65353. In Microsoft Excel data were initally processed by removing zeros, pruning to integer values, and binning. Cells containing zero (0) were replaced with one (1) to simplify analysis and statistical comparisons. Pruning to integer values is a built in Excel function, applied simply to reduce data complexity. Intensity values to three decimal places are not justified by the precision of the experimental system.

A data visualization Excel tool developed in house called `chromaBlast` was used. `chromaBlast` determines a chip-dependent minimum and maximum value, then sorts that chip data into user-defined subsets (bins) of equal size. The number of bins (and the heat map colour assigned to the bins) is determined by the user. For example, if the range of values on a chip is from 0 to 1000, the user might set 5 bins. Each bin is of equal size in terms of the total range, thus the bins in this case are: 1 to 200, 201 to 400, 401 to 600, 601 to 800, and 801 to 1000. Asymmetric visualizations may be created by assigning the same color to more than one bin range. One can use asymmetric displays to minimize the effect of low intensity (noisy) features, and differentially consider high-value (less prone to random variation) features. Using the heat-map coloured display, it is straightforward to visually compare two intensity distributions for differences. Using the "data-filter" tool in Excel, one can remove non-differentiating data from the display to simplify analysis. An important characteristic of `chromaBlast` is that the binning process effectively normalizes the data, so that numerical manipulations are not required for basic analysis. For large scale automated analysis (e.g. comparison of thousands of features to a library of species and strains), numerical normalization would be required.

Results

A preliminary library of genomic DNA samples was prepared and archived for future use. This library will grow as strains are isolated and DNA becomes available, serving as a reference for future work.

Comparisons of images of microarrays (subgrid 17, metacolumn 1, metarow 5) are shown in Figures 1-3. In Figure 1, *B. anthracis* RP42 is compared to *E. coli* JM108. Features (oligonucleotide spots) are displayed in the same relative position in the image, with some small variations due to enlargement. In Figure 1, areas where the two species have essentially the same hybridization patterns are indicated in grey-green boxes, while areas where the hybridization patterns are different are shown in red boxes.

In Figure 2, a qualitative comparison for presence versus absence of detectable hybridization signal is shown. For *B. anthracis*, features with hybridization signal are marked in the overlying grid with yellow squares. For *E. coli* JM108, signals are marked with blue squares. In the overlap image in the lower frame, concordant signals are hatched grey-green, *B. anthracis*-only signals are yellow, and JM108-only signals are blue. Signals for which the feature is completely clean and which have background which could be interpreted as having unambiguous signal, are given a high trust scoring. Thus for this subgrid of 484 (out of >23,000), 22 signals have high trust for discrimination of the two species, at the qualitative level.

In Figure 3, comparative hybridizations following the same scheme as Figure 2, are shown for *B. anthracis* RP42 and Thraxol. Out of the 484 features in the subgrid shown, 9 are discriminating with high trust. One of the differential signals (column 1, row 9, indicated with an asterisk in Figure 3) with high trust between RP42 and Thraxol also is highly discriminant between RP42 and *E. coli* JM108, but appears to be common between JM108 and Thraxol. It should be noted that the qualitative signal comparisons at low replication are susceptible to minor chip defects. Thus replicate comparisons will need to be performed to verify any discriminating feature.

Figure 4 represents a summary comparison of pairwise X-Y scatter plots of the *E. coli* strains, *B. anthracis* RP42, and *Y. enterocolitica*, to the averaged *E. coli* array data across all 23,232 features. These plots are derived from the intensity scores of replicate hybridizations, at least two for each sample. The differences in the distributions are apparent. A linear pattern is apparent in all of the plots, caused by shared signals (background) on many features. Adjustment of intensities by scaling does not change the pattern of distribution, although it may shift the "cloud" of intensities relative to the X-Y midline, indicated in red.

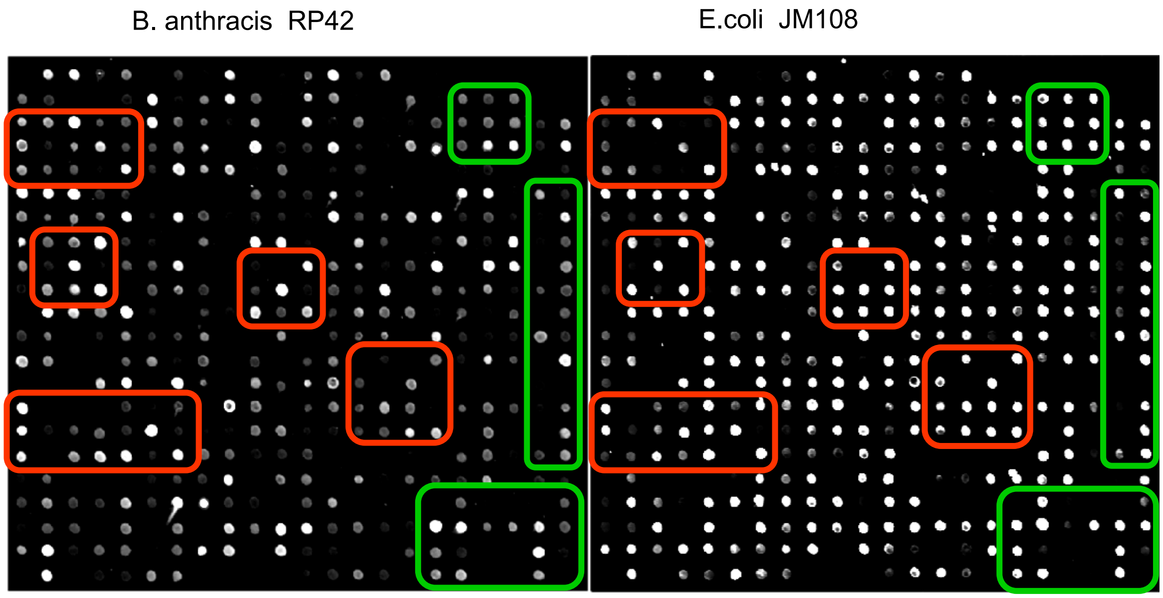


Figure 1: *Grid details from hybridizations of strains of B. anthracis and E. coli. Example concordant regions are marked with green boxes, nonconcordant are marked with red boxes. At this qualitative level considering positive signals, it is apparent that patterns of hybridization differ.*

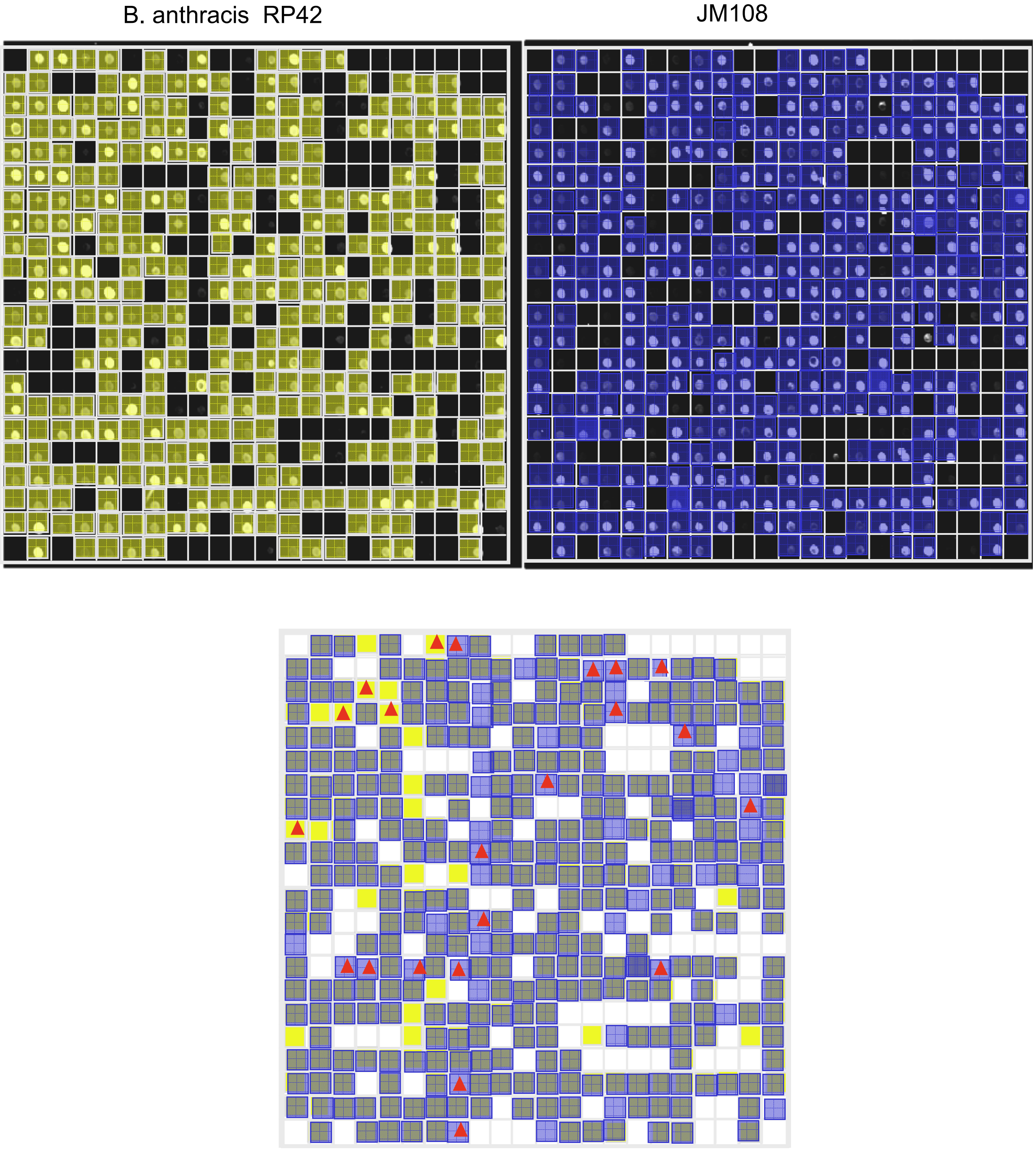


Figure 2: *Image comparison of B. anthracis RP42 versus JM108 on subgrid 17 (metacolumn 1 metarow 5) of the 21K array. The color in the lower image indicates features where the two samples have different (yellow, blue) or overlapping (grey-green) hybridization to the array. Red triangles indicate differences with high trust (22/484).*

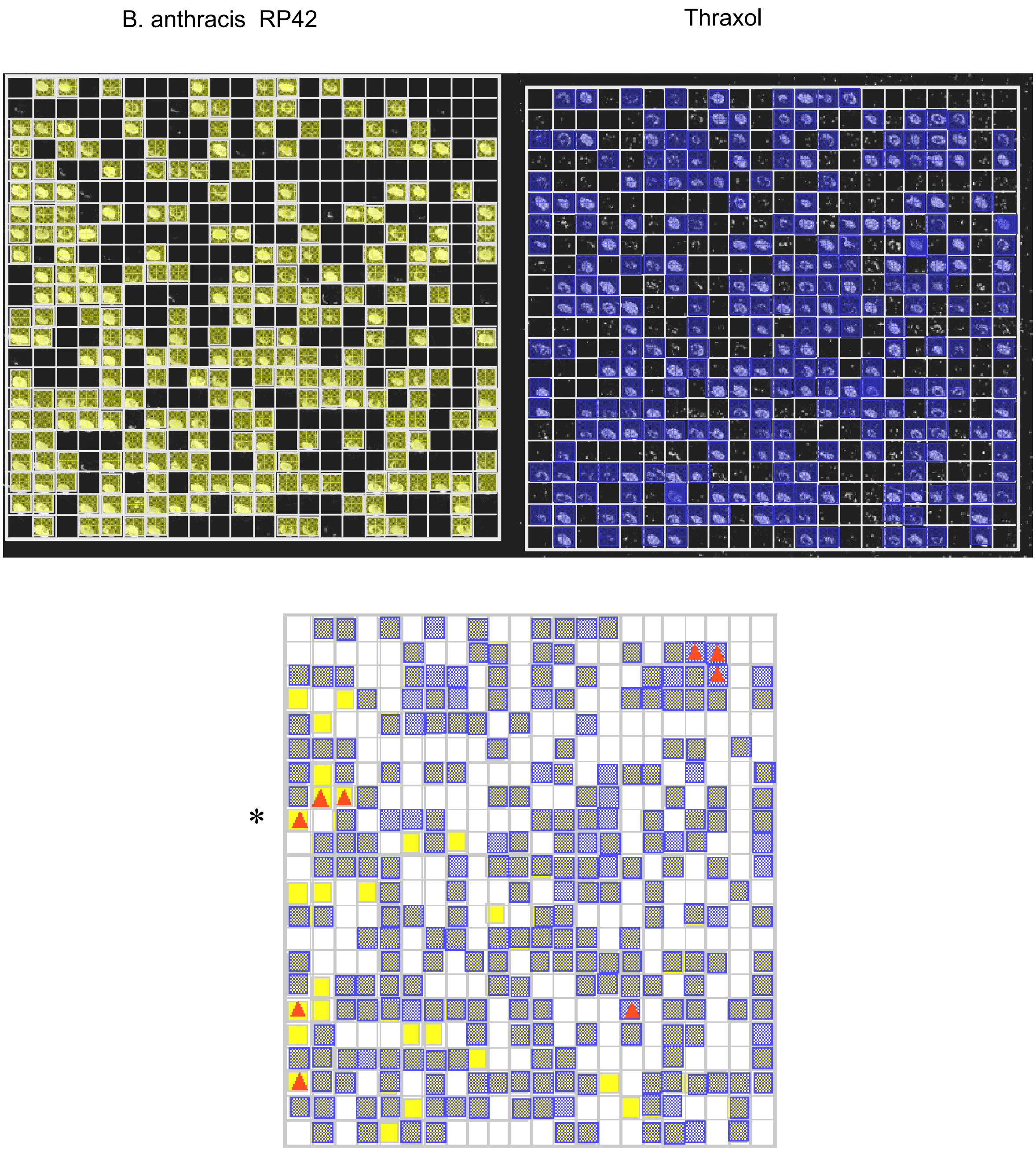


Figure 3: *Image comparison of B. anthracis RP42 versus B. anthracis Thraxol on subgrid 17 (metacolumn 1 metarow 5) of the 21K array. The color in the lower image indicates features where the two samples have different (yellow, blue) or overlapping (grey-green) hybridization to the array. Red triangles indicate differences with high trust (9/484).*

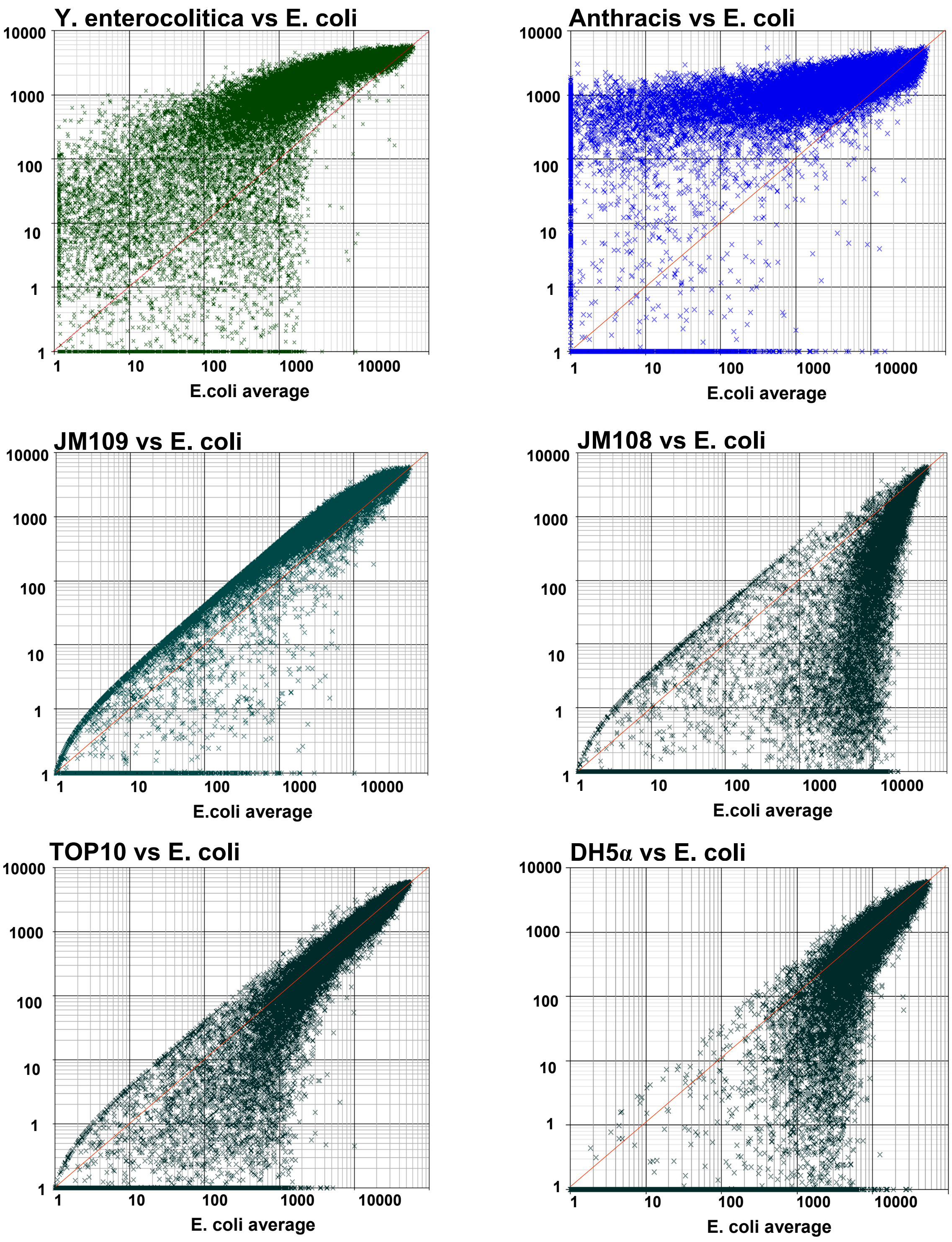


Figure 4: *Pairwise feature intensity comparison. The linear patterns along axes are due to shared intensity values on many features (e.g. no hybridization signal).*

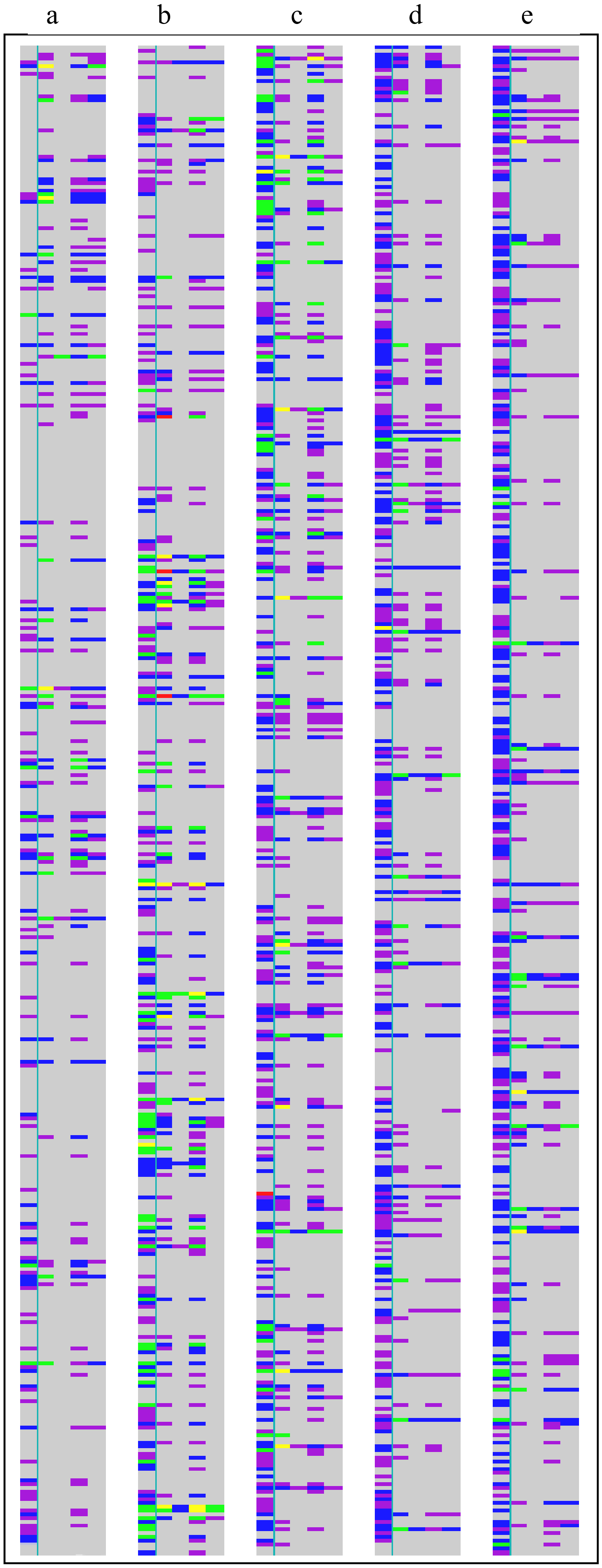
Clustering analysis offers another approach to differentiating microbial genomic DNA samples. Using `chromaBlast`, which normalizes different array data sets, and produces a heat map of the intensities, it is possible to readily visualize comparisons between samples. A detailed description of `chromaBlast` will be presented elsewhere. In Figure 5, the first 2000 features of the averaged data shown in Figure 4 are displayed.

In the color map, it is apparent that signal pattern differences exist between the species and strains being tested. While still at this level a qualitative comparison, the heat map allows the interpetation of differences in relative intensity (the data are normalized by the binning procedure). It is clear that *B. anthracis* hybridizes at many more features than *E. coli*, above the minimum color threshold (column d and e, Figure 5). The natural extension of this approach is to quantitate the differences. It is also apparent that the pattern of differences, not easily reduced to a single numerical representation, will also be useful in discriminating samples from each other.

Figure 5: Analysis by chromablast of first 2000 features as in Figure 4. Data are unfiltered.

Column	*Features*
a	*1-400*
b	*401-800*
c	*801-1200*
d	*1201-1600*
e	*1601-2000*

Withinin each column (a-e), the data represented are (left to right): B. anth. RP42, E.coli DH5α , E. coli JM108, E. coli JM109, E. coli TOP10.



color	**bin**	**max intensity**	**count**	**percentile**
🟥	**G**	**56672**	**34**	**99.8**
🟨	**F**	**48514**	**212**	**99**
🟩	**E**	**40472**	**1023**	**95**
🟦	**D**	**32377**	**3277**	**80**
🟪	**C**	**24287**	**5147**	**58**
⬜	**B**	**16189**	**6685**	**30**
⬜	**A**	**8094**	**6861**	

Discussion

Using microarray technology, one can assay thousands of features simultaneously, in a single sample. A microarray in this application is a microscope slide onto which are spotted several thousand individual DNA probe sequences, each one of which can detect unique fragments of DNA. Using DNA probes specific to known microbial sequences, one can identify with high confidence the species and probably the strain of organism under examination. Such a tool is a useful complement to existing PCR, RFLP, or AFLP technologies.

The very specificity of the hybridization, however, precludes the ready detection of the presence of atypical or novel genetic sequences. In order to relieve the requirement for prior knowledge of all possible genetic content, the application of non-specific probes may be useful. By the use of a selection of DNA features which do not explicitly correspond to known microbial sequences, but which have relatively fixed hybridization properties, one can detect the presence of novel or unexpected genetic sequences in the microbe. This panel of non-specific sequences in itself provides a unique “fingerprint” for any microbe, such that cultivated strains which have diverged from known laboratory strains will be detected, and a measure of the divergence calculated. Prior knowledge of the microbe origin, species, or DNA sequence is not required. In addition, by using DNA probes specific for known “threat factors” (for example antibiotic resistance, virulence, etc.) one can identify organisms that may have been genetically modified to include these traits. Any organism carrying unusual or unexpected resistance or virulence genes represents an increased threat. Inclusion of probes specific to existing cloning vector fragments has the added virtue of detecting recombinants which are likely to be synthetic rather than naturally occurring.

The importance of the problem of recombinants has been demonstrated in an open literature description of a modified mouse pox virus (related to human smallpox) containing an extra gene encoding interleukin-4 [1]. Mice vaccinated against mouse pox, or which are genetically resistant to the normal virus, exhibit mouse pox symptoms and high mortality when exposed to the modified virus. The publication of multiple methods by which potential biowarfare agents can by modified emphasizes the need for tools to detect such recombinants [7-11].

The combination of the “random” sequence fingerprint with the panels of known sequences (microbial genes, antibiotic resistance, etc.) will offer a rapid screen for identification of species, strain, and the presence of foreign genetic material. This work has demonstrated the success of the “random” probe concept in preliminary experiments. That such patterns are reproducible demonstrates that the hybridization of the genomic DNA to the features is sequence specific and not random. A further consideration is that the hybridizations are probably not context specific. That is, if a genome has been rearranged, it will in general contain the same sequence content, in altered context. RFLP, AFLP, or pulsed field techniques would correctly indicate that the organism is genetically different from library samples, even if functionally (*i.e.* microbiologically) similar. Using hybridization techniques that do not require positional stability (*e.g.* on multi-dimension gels or microarrays) means that we should be able to identify and compare genomic content without much interference from contextual alterations. This hypothesis has not yet been tested, and it is not known at this point what is the sampling density of the genomic DNA or whether it is sensitive to positional changes within the genome. Experiments to test this question are underway.

An advantage of the current approach is that identification and characterization of bacterial, fungal or parasitic agents could be done without DNA amplification, depending on sample quality. Viral agents will probably necessitate sample amplification but no other special work would be needed. Maximum benefit from this technique may require integration of other evolving technologies (field sample processing, PCR, random amplification of polymorphic DNA (RAPD), hi-speed hybridization and post-processing). The microarray approach will not be as rapid or technically trivial as the hand-held “tickets” used in some applications. Results will be obtained in a few hours, rather than minutes. On the other hand, a single microarray design, implemented in a robust platform, should identify a test sample as one of hundreds of strains, and give a quantitative estimate for similarity to known strains in the library. In addition, it would be possible to identify organisms with novel or unusual elements, such as toxin or resistance genes. Reliable data from such work will require good standardization of methods.

The microarray itself is non-hazardous and can be shipped or distributed without restriction. The protocols and equipment to use microarrays are commercially available. Thus the technique can be used for interlaboratory or international comparisons of isolated or unknown microbes, without the need to transport hazardous biological samples. The identification of bacterial strains, and the ability to detect the presence in the microbe of unusual or artificially constructed genetic elements is key to verifying the origin of the microbe, whether it constitutes a risk to humans, and possible therapeutic strategies. The microarray assay can be used to determine the relatedness of a given test sample to members of the library of known strains, and whether or not additional genetic material has been added to the organism, including antibiotic resistance, virulence or toxin genes, or other markers. From this information, rational decisions about treatment, personnel protection, or quarantine can be made. A primary application of this technology is forensic identification and tracking of recombinant organisms.

References

1. Jackson, R., Ramsay, A., Christensen, C., Beaton S., Hall D., Ramshaw, I. (2001). Expression of mouse interleukin-4 by a recombinant ectromelia virus suppresses cytolytic lymphocyte responses and overcomes genetic resistance to mouse pox. *Journal of Virology,* 75: 1205-1210.

2. Tettelin, H., Masignani, V., Cieslewicz, M. *et al.* (2005). Genome analysis of multiple pathogenic isolates of Strepotcoccus agalactiae: Implications for the microbial pan-genome. *Proceedings of the National Academy of Sciences,* 102:13950-13955.

3. Sofi Ibrahim, M., Kulesh, D., Saleh, S., Damon, I., Esposito, J., Schmaljohn, A., Jahrling, P. (2003). Real-time PCR assay to detect smallpox virus. *Journal of Clinical Microbiology*., 41: 3835-3839.

4. Jackson, R. Maguire, D., Hinds,L., Ramshaw, I. (1998). Infertility in mice induced by a recombinant ectromelia virus expressing mouse zona pellucida glycoprotein 3. *Biology of Reproduction*, 58: 152 - 159.

5. Lloyd, M., Shellam, G., Papadimitriou, J., Lawson, M. (2003). Immunocontraception is induced in BALB/c mice inoculated with murine cytomegalovirus expressing mouse zona pellucida 3. *Biology of Reproduction*, 68: 2024-2032.

6. Kwik, G., Fitzgerald, J., Inglesby, T., O'Toole, T. (2003). Biosecurity: responsible stewardship of bioscience in an age of catastrophic terrorism. *Biosecurity and Bioterrorism: Biodefense Strategy, Practice, and Science*, 1: 27- 35.

7. Tardif, C., Maamar, H., Balfin, M., Belaich, J. (2001). Electrotransformation studies in Clostridium cellulolyticum. *Journal of Industrial Microbiology and Biotechnology*, 27: 271-274.

8. Vertiev, Y., Zdanovsky, A., Shevelev, A., Borinskaya, S., Gening, E., Martin, T., Ivanov, P., Yankovsky, N. (2001) Recombinant Listeria strains producing the nontoxic L-chain of botulinum neurotoxin A in a soluble form. *Research in Microbiology*, 152:563-7.

9. Jennert, K., Tardif, C., Young, D., Young, M. (2000). Gene transfer to Clostridium cellulolyticum ATCC 35319. *Microbiology*, 146 Pt 12:3071-3080.

10. Nagahama, M., Michiue, K., Sakurai, J. (1996) Production and purification of Clostridium perfringens alpha-toxin using a protein-hyperproducing strain, Bacillus brevis 47. FEMS *Microbiology Letters*, 145:239-243.

11. Poppe, C., Ziebel,l K., Martin, L., Allen, K. (2002). Diversity in antimicrobial resistance and other characteristics among Salmonella typhimurium DT104 isolates. *Microbial Drug Resistance*, 8:107-122.

12. Reyes-Lopez, M., Mendez-Tenorio, A., Maldonado-Rodriguez, R., Doktycz, M., Fleming, J., Beattie, K. (2003). Fingerprinting of prokaryotic 16S rRNA genes using oligodeoxyribonucleotide microarrays and virtual hybridization. *Nucleic Acids Research*, 31: 779-789.

13. Call, D., Borucki, M., Besser, T. (2003). Mixed-genome microarrays reveal multiple serotype and lineage-specific differences among strains of Listeria monocytogenes. *Journal of Clinical Microbiology*, 41:632-69.

14. Kato-Maeda, M., Rhee, J., Gingeras, T., Salamon, H., Drenkow, J., Smittipat, N., Small, P. (2001). Comparing genomes within the species Mycobacterium tuberculosis. *Genome Research*, 1:547-554.

Annex A: Genotypes of Escherichia coli strains

STRAIN	Source	Date	Genotype
DH5alpha F-	Bader/DRDC	20/9/2002	F$^-$ /*end*A1 *hsd*R17 ($r_K^- m_k^+$) *gln*V44 *thi*-1 *sup*E44 *rec*A1 *gyr*A (Nal^r) *rel*A1 *Δ*(*lacI*ZYA-*arg*F)U169 *deo*R
JM108	Bader/DRDC	20/9/2002	F-, recA1, endA1, gyrA96, thi-1, hsdR17(r_k^- , m $_k^+$), supE44, relA1, Δ(lac-proAB)
JM109	Bader/DRDC	20/9/2002	[F' *tra*D*36 pro*A$^+$*pro*B$^+$ *lacI*^{q} *Δ*(*lac*ZM15] *Δ*(*lac*-*pro*AB) *gln*V44 *e*14$^-$ *gyr*A96(Nal^r) *sup*E44 *rec*A1 *rel*A1 *end*A1 *thi*-1 *hsd*R17($r_K^- m_K^+$)
TOP10	Invitrogen	20/1/2003	F$^-$ *mcr*A Δ(*mrr-hsd*RMS-*mcr*BC) (ϕ80d*lac* Δ(*lac*Z)M15) Δ*lac*X74 *deo*R *rec*A1 *ara*D139 Δ(*ara-leu*)7697 *gal*U *gal*K *rps*L(Str^r) *end*A1 *nup*G

List of symbols/abbreviations/acronyms/initialisms

AFLP	amplified fragment length polymorphism
BSL3	Biosafety Level 3
dnTP	dATP, dCTP, dGTP, dTTP : deoxynucleotide triphosphates of DNA bases
DNA	deoxyribonucleic acid
EDTA	ethylenediamine tetraacetic acid
OD_{260}/OD_{280}	ratio of optical densities at 260 and 280 nm
PCR	polymerase chain reaction
RFLP	restriction fragment length polymorphism
RNA	ribonucleic acid
SDS	sodium dodecyl sulfate
TRIS	2-amino-2-hydroxymethyl-1,3-propanediol

Glossary

antibiotic resistance	a gene which confers the ability to a microbe, to survive and grow in the presence of an antibiotic which would normally inhibit its growth. Some resistance genes may be readily transferred from one microbe to another, creating novel organisms with new antibiotic resistance properties, a standard lab method.
chimera	an organism containing genetic material (part of one gene, up to many genes) from another organism, possibly having toxins, antibiotic resistance, pathogenicity, or host range differences from the original organism.
episome	a fragment of DNA, usually having its own replication origins, which is not a component of the normal cellular genome; episomes may be found integrated into host chromosomes.
fingerprint	a collection of signal intensity scores, digitized from an image of a hybridization of genomic DNA to a microarray spotted with DNA fragments (oligonucleotides, PCR products from genes, etc.). The fingerprint of a given species and strain is unique from that of other species or strains.
gene	a DNA sequence which encodes a single genetic trait which is inherited by offspring.
genomic DNA	the DNA which comprises the genetic material of an cell, and is inherited by the progeny of the cell. The so-called blueprint of life. The sequence of nucleotides in the genomic DNA comprises the genes, and determines the properties of the microbe. For many microbes, the entire sequence of the genomic DNA is completed and in the public domain.

host-range	the tendency of a microbe to infect a limited number of different hosts, usually determined by microbial genetic traits.
hybridization	sample DNA (or RNA) is tagged with a fluorescent dye, then applied to the surface of the microarray. Under controlled incubation temperature and solution conditions, sequences in the sample DNA which correspond to sequences in the microarray features, will bind to the features (hybridize). Hybridization often refers to the entire process from labelling to binding, to post incubation washing.
library	1) an orderly collection of DNA (or oligo-nucleotides) which is maintained in the lab as a source of genetic material. Libraries may be synthetic, derived from cloned genes, or PCR products. 2) a digital collection of data from related assays used for comparison.
microarray	a microscope slide, filter membrane or other solid surface, onto which DNA fragments have been spotted in an organized grid. The DNA may originate from cDNA libraries, PCR products, genomic DNA isolates, or synthetic oligonucleotides. Each spot is called a feature.
nucleotide	the components of DNA are the nucleotides deoxyadenosine monophosphate, deoxycytidine monophosphate, deoxyguanosine monophosphate, deoxythymidine monophosphate, and the chemical bonds which join them into long chains. Genetic information is encoded in the order in which the nucleotides occur in the DNA chain.

oligonucleotide (oligo)	a fragment of DNA (or RNA) chemically synthesized, and often representing some section of genetic material from which the sequence is already known. Oligos may also be “random” in sequence, such that the oligo sequence is not intentionally derived from known DNA sequences.
PCR	(polymerase chain reaction) A technique to amplify small amounts of genetic material using an enzyme reaction (DNA polymerase), in order to facilitate manipulation or analysis, a standard method.
recombinant	an organism containing modified or additional genetic material, possibly created in laboratory, but may also arise in nature.
restriction enzyme	an enzyme which cuts DNA at specific nucleotide sequences, a standard method.
species	the grouping of microbes according to major genetic differences (e.g. the ability to grow (or not) in an oxygen-free environment).
strain	a microbe which differs from other members of the same species by minor or additional genetic characters (e.g. resistance or sensitivity to penicillin).
toxin	a gene product which when expressed, is poisonous to the host organism or to other microbes. For example, certain antibiotics (e.g. penicillin) are produced by microbes in order to poison other microbes.
virulence gene(s)	a gene or set of genes which modify the growth or infectiousness of a microbe, enhancing its ability to cause disease.

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 who the document was prepared, e.g. Establishment sponsoring a contractor's report, or tasking agency, are entered in Section 8.)

Defence R&D Canada – Suffield
PO Box 4000, Station Main
Medicine Hat, AB   T1A 8K6

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

Unclassified

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).

Microarray Genomic Fingerprinting (U)

4. AUTHORS (Last name, first name, middle initial. If military, show rank, e.g. Doe, Maj. John E.)

Ford, Barry N., Shei, Y., Bamforth, J.

5. DATE OF PUBLICATION (month and year of publication of document)

December 2005

6a. NO. OF PAGES (total containing information, include Annexes, Appendices, etc) 31

6b. NO. OF REFS (total cited in document) 14

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.)

Technical Memorandum

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

DRDC Suffield/Chemical and Biological Defence Section, CFB Suffield, Ralston, Alberta

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.)

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.)

DRDC Suffield TM 2005-244

10b. OTHER DOCUMENT NOs. (Any other numbers which may be assigned this document either by the originator or by the sponsor.)

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

(x) Unlimited distribution
(   ) Distribution limited to defence departments and defence contractors; further distribution only as approved
(   ) Distribution limited to defence departments and Canadian defence contractors; further distribution only as approved
(   ) Distribution limited to government departments and agencies; further distribution only as approved
(   ) Distribution limited to defence departments; further distribution only as approved
(   ) Other (please specify):

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

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) or (U). It is not necessary to include here abstracts in both official languages unless the text is bilingual).

Using current molecular biology, it is possible to create novel microbial forms which have never existed in nature. Current methods for identifying and characterizing microbes depend on some prior assumptions about the genetic content of the organism. Recombinant organisms containing novel genetic material by definition fall outside of the range of prior sequence data. Thus new methods which do not require such assumptions are required to screen and identify the strains, and to detect the presence of novel genetic material. Using an oligonucleotide microarray, we have executed a proof of concept of an array-based genomic fingeprinting technology. The test organisms for this work were *Escherichia coli* (4 strains), *Bacillus anthracis* (2 strains), and *Yersina enterocolitica*. Using standard molecular biology methods, we isolated genomic DNA, digested the DNA to reduce complexity, labelled it with fluorescent dyes, and hybridized the labelled DNA to microarrays containing 21,000 unique olignucleotide features. From a single grid of 484 features, or from a filtered subset of the hybridization data, species could be readily discriminated with high confidence. Strain differentiation may require analysis of the entire feature map, or refinement of the array sequences features, the analysis model, or the analysis software. The next phase of this work will be a test system for rapid species identification in addition to the oligonucleotide fingerprint, leading to a finalized design for a prototype genomic fingerprinting microarray.

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. Identifies, 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
genomic fingerprint
recombinant
biowarfare
microorganism