A Species Independent Universal Bio-detection Microarray for Pathogen Forensics

Shamira J. Shallom

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
In
Genetics Bioinformatics and Computational Biology

Harold 'Skip' Garner, Committee Chair
David R. Bevan
Reinhard Laubenbacher
Christopher Lawrence

May 4th, 2012
Blacksburg, VA

Keywords: (Array Surveillance Pathogen Detection)

A Species Independent Universal Bio-detection Microarray for Pathogen Forensics

Shamira J. Shallom

ABSTRACT

The detection and identification of bio-threat agents and study of host-pathogen interactions require a high-resolution detection platform capable of discerning closely related species. This dissertation addresses the completion of the development of an array based platform and provides a robust pipeline for the discovery of unique bio-signatures for pathogens and their host. Our collection (library) of host and pathogen signatures has been greatly expanded to improve robustness and identification accuracy of an 'unknown' sample. The library containing measured bio-signatures for each species/isolate is complemented with computational methodologies to resolve the identity of the unknown sample as well as a mixture of organisms or a pathogen in a host background.

Current approaches for pathogen detection rely on prior genomic sequence information. This research targets use of a broad based platform for identification of pathogens from field or laboratory samples on a high density Universal Bio-signature Detection Array (UBDA). This array is genome independent and contains all possible (4^9 combinations) 9-mer probes which are mathematically computed and genome independent. It works by comparing signal intensity readout to a library of readouts established by interrogating a wide spectrum of organisms. Each genome has a unique pattern of signal intensities corresponding to each of these probes. These signal intensities were used to generate un-biased cluster analysis patterns that can easily distinguish organisms into accepted and known phylogenomic relationships.

Classification methods such as hierarchical clustering, Pearson's correlation matrix, principal component analysis and curve fitting regression methods were tested for pathogen specific use cases. Hierarchical clustering and Pearson's correlation matrix methods can establish phylogenomic relationships between highly diverse genomes. However, in order to assign a given sample to one or more groups, such as a pure isolate of a single species or composite mixture of multiple species, principal component analysis (PCA) was used. The test cases included identification of mixed samples, case study of field samples from state diagnostic labs and finally a surveillance method for viral and parasite carrying insect host vectors. Completion of these application challenges is meant to demonstrate the power and confirm confidence in the Universal Bio-signature Detection Array.

This work was supported by a graduate fellowship from the Virginia Bioinformatics Institute, Virginia Tech and the Southern Regional Education Board (SREB) to S. Shallom and U.S. Department of Homeland Security through the national Center of Excellence for Foreign Animal and Zoonotic Disease Defense at Texas A&M University to Dr. Garner, PI.

Dedication

To my family with love and gratitude for their patience and support

Acknowledgements

I would like to express my sincere gratitude and appreciation to Dr. Harold 'Skip' Garner who provided mentorship, insight and vision for this research. I would also like to thank my PhD committee members: Drs. David R. Bevan, Reinhard Laubenbacher and Christopher Lawrence for their encouragement during my time at Virginia Tech. I would like to express my sincere gratitude to Dr. Gary Adams from Texas A and M University for guiding me in various aspects related to the study of bacterial pathogenesis, and critically reading my manuscripts. I would like to express my deep gratitude and appreciation to Ms. Dennie Munson from the Graduate school at Virginia Tech for guiding me and keeping me on track through the GBCB graduate program requirements.

Table of Contents

Chapter 1

1. Introduction

1.1 Array based approaches in pathogen forensics

The ability to identify a bio-terror attack or an accidental release of a research pathogen from a naturally occurring disease event is crucial to the safety and security of this nation, enabling an appropriate and rapid response. Micro-organisms that have been developed for an attack maybe altered, selected or engineered for enhanced survival outside the host with increased virulence. It is critical in an infected animal, environmental or laboratory sample to quickly and accurately identify the precise pathogen. This includes variants from natural or engineered genetic drift and to classify new pathogens in relation to those that are known. Rapid, accurate and sensitive detection of bio-threat agents requires a broad-spectrum assay capable of discriminating between closely related microbial or viral pathogens. In cases where a biological agent release has been identified, forensic analysis demands detailed genetic signature data for accurate strain identification and attribution. Identification of genetic signatures for detection, coupled with identification of pathogenic phenotypes, would provide a robust means of discriminating pathogens from closely related species [1].

Traditional strategies for detecting pathogens have used the following approaches. The first approach uses universal PCR to amplify one or more universal genes such as 16S ribosomal RNA, 18Sribosomal RNA, 23S ribosomal RNA and screen for pathogen specific polymorphisms[2]. The probes on the array are derived from a combination of ribosomal RNA genes from a given set of organisms of high priority. One of the challenges of this approach is the frequent and unexpected amplification of contaminating template DNA, as observed in control reactions[3]. Another potential problem that targets 16S ribosomal RNA pathogen specific sequences is unexpected polymorphisms. Hence naturally occurring variants may not be represented on the microarray and failure to detect the variants would represent false negatives [2]. The second approach uses detection by amplifying a specific set of genetic markers that are detected on an array that has several probes for genes from a set of organisms. Such tests have been used for food-borne bacteria such as *E. coli* O157:H7 [4], viruses [5] and mixtures of pathogens [6]. The drawback of using this approach with multiple PCR primers sets is the generation of spurious products [2]. The third strategy is the use of 70-mer oligonucleotide derived from pathogen specific genes which are spotted on the array. Strategies for viral detection have used a microarray composed of 1600 unique viral oligonucleotides (70-

mers) derived from 140 distinct viral genomes[7]. The drawback of this strategy is that only the group of pathogen specific genes will be queried. Information will not be obtained if the strain has undergone a genetic drift or has been engineered differently. Detection and identification of bio-threat agents and study of host-pathogen interaction requires a high-resolution detection system capable of discerning closely related species. Given the enormous spectrum of genetic possibilities, only a highly parallel field deployable, robust technology which is universal in nature has near-term potential to address these needs. This research addresses the development of an array-based platform and provides a robust pipeline for the discovery of unique signature patterns for pathogens and their host.

1.2 Universal Bio-signature Detection Array (UBDA): A species independent pathogen forensics platform

The initial vision for a universal DNA microarray was a matrix of oligonucleotides containing features with unique n-mer probes [8]. This requires constructing an array that requires 4^n features. Larger values of *n* infuse greater specificity into the arrayed probes, but as *n* increases the number of required features grows rapidly. This universality is obtained by synthesizing a combinatorial n-mer array containing all 4^n possible sequences of length n [9]. The key issue was to find a value of n that is large

enough to afford sufficient hybridization specificity and small enough to be practically fabricated and analyzed. The initial prototype of universal arrays used oligonucleotide probe lengths of 12 and 13 bases. A subset of 14,283 probes from 4^{12} possible probes were synthesized by in situ synthesis technology using digital optical chemistry (DOC) [10],[11], [12]. Subsequently a high density oligonucleotide microarray with 370k elements called Universal Bio-signature Detection Array (UBDA) has been designed by the Garner Laboratory and commercially produced by Roche-Nimblegen (Madison, WI).The main feature of this array is that the probes are computationally derived and sequence independent. There is one probe for each possible 9-mer sequence, thus 4^9 (262,144) probes. These probes were synthesized on the array using light-directed photolithographic synthesis of high density oligonucleotide arrays [13], [8].

The random 9-mer probes are comprised of a core of 9 nucleotides and flanked on both sides by three nucleotides, selected to maximize sequence coverage of these basic 15-mers. Probes with a low GC content were padded with additional bases at their termini to equalize melting temperatures, with most probes ranging from 15-21 nucleotides in total length. There are 262,144 random 9-mer probes and 20,000 of them are replicated 3 times in total. All sequences were uniquely mapped to integers

by representing them as the base-10 equivalents of base-4 representations of DNA sequences in which A=0, C=1, G=2 and T=3 where for example AAACTATAG = 000130302(base 4) = 1842 (decimal). Hence all 4^9 9-mers have corresponding unique decimal integer values. These integers will double as unique IDs and array indices. UBDA is a two-color array, where two independent samples can be labeled with either fluorescent dyes Cy3 and Cy5. Each dye fluoresces at a different wavelength and hence they do not interfere with signal intensity for each of their spectra.

This unique strategy uses the robustness of patterns generated from hybridization of any DNA or cDNA to a very high-density species independent oligonucleotide microarray. Each genome hybridized on this array has a unique pattern of signal intensities corresponding to each of these probes.

This platform is highly attractive because it has multiplex capacity where knowledge can be drawn from the various probe sets available on the array without prior knowledge of the sample's genomic composition. This dissertation addresses the development of broad based methods for identification of pathogens including variant strains from infected animals, environmental or laboratory samples on a Universal Bio-defense Array (UBDA). This platform has commercial applications for the development of

cost effective reliable platform for accurately screening of large number of samples for bio-threat agents in forensic analysis, pathogens that routinely infect animals of farm value, food borne pathogens and as a molecular diagnostic of micro-organisms in a clinical environment. It is also applicable to un-sequenced genomes or strains of microbes whose sequence may have drifted or been intentionally engineered. The spectrum of organisms chosen for hybridization on the UBDA array were based on priority, primarily bio-threat agents, micro-organisms infecting farm animals and organism of molecular diagnostic importance in a clinical setting.

1.3. Comparative genome hybridization analysis pipeline for UBDA array

This platform is highly attractive because it has multiplex capacity where knowledge can be drawn from the various probe sets available on the array without prior knowledge of the sample's genomic composition. These probe sets are available in a repository of bio-signatures that future users of this technology can compare and draw inferences related to the sample under study. Most of the genomic DNA from this collection of organisms belongs to the category of select agents NIAID A-C. The library of bio-signatures for these organisms and their hosts is complemented by computational

methodologies comprised of data parsing, clustering and classification algorithms and regression techniques to resolve mixed samples.

Clustering is one of the data mining processes for discovery and identifying patterns in the underlying data. Clustering algorithms partition data into subsets based on similarity and dissimilarity. Clustering methods follow three steps: pattern recognition, use of a clustering algorithm and similarity measure matrix [14]. For pattern recognition, pair wise comparisons will be used between samples to select the features on which the clustering is to be performed. A wide range of clustering algorithms have been developed for analysis of genomic expression data sets such as hierarchical clustering, k-means clustering, self-organizing maps and principal component analysis. K-means, self-organizing maps and principal component analysis are frequently used in gene expression studies to group related gene clusters. The initial approach utilized the hierarchical clustering algorithm to determine phylogenomic relationships between organisms. This method could be used as an initial approach to cluster a large number of arrays. Hierarchical clustering [15] transforms a distance matrix of pair-wise similarity measurements between all items into a hierarchy of nested groupings. The hierarchy is represented with a binary tree like dendrogram that shows the nested grouping of patterns and the similarity levels at which

groupings change. Hierarchical clustering algorithm then follows either agglomerative procedures or divisive procedures. In agglomerative hierarchical clustering, a similarity distance matrix is constructed by calculating the pair wise distance between all patterns [14]. Agglomerative procedures or rules that govern this distance or similarity calculation are classified under linkage methods: single, average, complete and centroid. In this study the centroid or complete linkage were used to generate relationship dendrogram by hierarchical clustering. Centroid linkage is an un-weighted pair group method, the distance between two clusters is the Euclidean distance between their centroids, calculated by arithmetic mean. Complete linkage clustering or the farthest neighbor method is a method of calculating distance between clusters in hierarchical cluster analysis. Similarity measure that is frequently used is Euclidean distance. Euclidean distance metric is a measure of the geometric distance between two components.

Hierarchical clustering algorithm was found to be not robust enough to handle multiple arrays due to the large number of data points on the array. Further, Pearson's correlation coefficient was used to develop a matrix of associations for a given set of samples and k-nearest distance [16] measure was used to generate a phylogenomic tree. Hierarchical clustering and

Pearson's correlation coefficient methods were not robust enough to distinguish between closely related isolates of *Brucella*. Hence Principal component analysis (PCA) [17] was used in classification of *Brucella* isolates into their respective species. Given *m* observations on *n* variables, the goal of PCA is to reduce the dimensionality of the data matrix by *r* new variables, where *r* is less than *n*. The most robust method for UBDA classification of bio-signatures signal intensities from a given sample was a regression analysis using the best least squares fit measure. This algorithm was originally applied to quantitate and attribute fluorescent intensities from a mixture of microspheres with different colors and comparing this to individual spectra generated from each colored microsphere. The standard linear curve fitting algorithm was used to determine the contribution of each individual dye to the measured emission spectrum [18]. This algorithm was developed further and recently has been applied in a hyper spectral imaging platform that can precisely identify and quantify the amount of a specific marker or a group in a given tumor sample. This was determined by comparing the spectra generated by the tumor sample to individual spectra generated from each of the fluorescently labeled molecular markers [REF]. Further this method has been used to determine the constituent proteins in a cell lysate on quantum dot lysate arrays using fluorescently labeled

antibodies. The algorithm compared the data spectra to a linear combination of standard spectra [19]. The identity or a close match to the unknown sample has been determined using regression analysis and classification approaches. The development of this technology has resulted in the creation of an integrated bio-signature, multiple select agent specific detection system.

1.4 Organization of publications and manuscripts

This research addresses the development of a pipeline for comparative genome analysis and creation of a data repository of bio-signatures specific for organisms under study. Our library contains over 70 pathogen and host 'patterns', and expands and increases in resolving power as more samples are processed. Hybridization patterns on these probes are unique to a genome, and potentially to different isolates and to a mixture of organisms. Identification of a new or emerging species can be classified on the similarity of its pattern to similar patterns found within a library of known samples. These probe sets were translated into a knowledge base repository of bio-signatures. Examples of unique hybridization signal intensity patterns are presented for different *Brucella* species as well as relevant host species and other pathogens. These results demonstrate the utility of the UBDA array as a diagnostic tool in pathogen forensics.

1.4.1 A species independent universal bio-detection microarray for pathogen forensics and phylogenetic classification of unknown microorganisms.

The first publication[20] addresses the development of the basic UBDA technology. Several arrays were tested to determine hybridization conditions and image scanning parameters. UBDA array sensitivity and specificity of probe hybridization and signal intensities were determined. UBDA technology was used to differentiate between an initial collection of samples that spanned eukaryotes, prokaryotes and virus clades. Hierarchical clustering algorithm and Pearson's correlation was used to determine phylogenetic relationships between organisms.

1.4.2 Comparison of genome diversity of *Brucella* spp. field isolates using Universal Bio-signature Detection Array and whole genome sequencing reveals limitations of current diagnostic methods.

The second manuscript addresses the utilization of this array technology to resolve closely related species and isolates of *Brucella* field isolates obtained from a state diagnostic lab "The Texas Animal Health Commission". This study was an in-depth study of the *Brucella* genome, for it involved real world field analysis of samples with unusual profiles from standard protocols (serology and biochemical typing).

The UBDA method was used to establish the identity of the species diversity and phylogenomic relationships between field isolates, and was shown to be sensitive to species variants of the type seen here. We demonstrate the use of signal intensities from UBDA to generate a principal component analysis and assign a given sample to one of more groups. Principal component analysis and Euclidean distance mapping to reference *B. abortus* 2308 and *B. suis* 1330 genomes provides a quantitative approximation to the composite species identity of the field isolate. Samples from bovine milk or tissue determined to be *B. suis* in biochemical or serological tests were found to be a mixed composite of *Brucella* species. To validate this, nine field isolates were sequenced and their sequencing reads were mapped to the *B. suis* 1330 and *B. abortus* 9-941 genome sequences. Hence, we determined that they were not pure *B. suis* isolates and presumably are the result of mixed or dual infections. The analysis of closely related strains and species by microarray-based comparative genomics provides a measure of genetic variability within natural populations.

1.4.3 Development of molecular diagnostics using Universal Bio-signature Detection Array technology in host pathogen forensics.

The third manuscript addresses advancing the development of the UBDA array as a molecular diagnostic and a robust pipeline for the discovery of unique nucleic acid signatures for pathogens and their host. Further the development of UBDA as a surveillance method for host insect vectors as carriers of viral and parasitic diseases was explored. This study provides a quantitative assessment of amounts of a given pathogen present in a host background or in a mixed organism population. Thus bio-signatures from the pathogen and host are simultaneously captured and analyzed.

Chapter 2

A species independent universal bio-detection microarray for pathogen forensics and phylogenetic classification of unknown microorganisms

Shamira J Shallom[1], Jenni N Weeks[2], Cristi L Galindo[1], Lauren McIver[1], Zhaohui Sun[1], John McCormick[1], L Garry Adams[3], Harold R Garner[1§]

[1]Virginia Bioinformatics Institute, Virginia Tech, Blacksburg, VA, USA; [2] St. Jude Children's Research Hospital, 262 Danny Thomas Place, Memphis, TN, USA; [3]Department of Veterinary Pathobiology, College of Veterinary Medicine, Texas A & M University, College Station, TX, USA.

[1§]Corresponding author

Harold R. Garner

Executive Director

Virginia Bioinformatics Institute

Virginia Tech

Washington Street, MC0477

Blacksburg, VA 24061-0477, USA

Email : garner@vbi.vt.edu

Phone: 540.231.2582

Fax: 540.231.2606

2.1 Abstract

Background

The ability to differentiate a bioterrorist attack or an accidental release of a research pathogen from a naturally occurring pandemic or disease event is crucial to the safety and security of this nation by enabling an appropriate and rapid response. It is critical in samples from an infected patient, the environment, or a laboratory to quickly and accurately identify the precise pathogen including natural or engineered variants and to classify new pathogens in relation to those that are known. Current approaches for pathogen detection rely on prior genomic sequence information. Given the enormous spectrum of genetic possibilities, a field deployable, robust technology, such as a universal (any species) microarray has near-term potential to address these needs.

Results

A new and comprehensive sequence-independent array (Universal Bio-Signature Detection Array) was designed with approximately 373,000 probes. The main feature of this array is that the probes are computationally derived and sequence independent. There is one probe for each possible 9-mer sequence, thus 4^9 (262,144) probes. Each genome hybridized on this array has a unique pattern of signal intensities corresponding to each of these probes. These signal intensities were used to generate an un-biased cluster

analysis of signal intensity hybridization patterns that can easily distinguish species into accepted and known phylogenomic relationships. Within limits, the array is highly sensitive and is able to detect synthetically mixed pathogens. Examples of unique hybridization signal intensity patterns are presented for different *Brucella* species as well as relevant host species and other pathogens. These results demonstrate the utility of the UBDA array as a diagnostic tool in pathogen forensics.

Conclusions

This pathogen detection system is fast, accurate and can be applied to any species. Hybridization patterns are unique to a specific genome and these can be used to decipher the identity of a mixed pathogen sample and can separate hosts and pathogens into their respective phylogenomic relationships. This technology can also differentiate between different species and classify genomes into their known clades. The development of this technology will result in the creation of an integrated biomarker-specific bio-signature, multiple select agent specific detection system.

2.2 Background

Rapid, accurate and sensitive detection of bio-threat agents requires a broad-spectrum assay capable of discriminating between closely related microbial or viral pathogens. In cases where a biological agent release has

been identified, forensic analysis demands detailed genetic signature data for accurate strain identification and attribution. Identification of genetic signatures for detection coupled with identification of pathogenic phenotypes would provide a robust means of discriminating pathogens from closely related but benign species [1].

Current forensics methods based on bacteriological, serological, biochemical and genomic strategies have been used to detect pathogens using serological methods [2], PCR [3], real time PCR [4, 5] and Multi-loci VNTR (variable-number tandem repeats) or MLVA [6-9]. Although bacteriological culture of *Brucella* spp. from blood, milk, fetal fluids and tissues, or other host tissues remain the 'gold standard' for diagnosis, bacteriologic culture has reduced sensitivity, is labour intensive, time consuming, typically requiring two weeks, and is a risk for laboratory personnel [5]. Serological assays, such as Rose Bengal, a rapid plate agglutination diagnostic test, is currently used for diagnosing infection with *Brucella* species in the field [2], however serological tests frequently have reduced specificity due to cross reactivity with other bacteria. Specific antibodies are required to be present at sufficiently high level and may require several weeks to develop before they are detectable. PCR based methods are used for epidemiological trace back and strain specific

identification [3]. Although rapid in nature, specific primers are required for specific genes from these genomes or 16S rRNA genes or VNTR (variable-number tandem repeats) in a given genome. Real time PCR based methods have been used to identify *Brucella* species using IS711, bcsp31 and per target genes [4, 5]. In addition, assays based on single-nucleotide polymorphisms have been developed for identification of *Brucella* isolates at the species level. These SNPs have been used to classify isolates into known *Brucella* species [10]. Recently MLVA or multi-loci VNTR (Variable-number tandem repeats) a genotype-based typing method and has been used as an epidemiological classification and SNP identification method for *Brucella* isolates in a field population [6-9]. MLVA method is used to understand the genetic diversity in polymorphic loci and to establish taxonomic relationships between different biovars of *Brucella*. It is used for microbial typing and epidemiologic studies by amplifying loci which are specific to a given genome and sequencing these regions. This is a powerful approach and is being used to create phylogenetic relationships and discovery of single nucleotide polymorphisms in independent loci from different *Brucella* isolates [7].

Array based approaches for forensic detection utilizes genome specific ribosomal RNA genes, genome specific PCR markers or

oligonucleotide probes. Arrays from rRNA are derived from a combination of rRNA genes from a given set of organisms of high priority. Universal PCR is used to amplify one or more universal genes, including 16S, 18S and 23S as well as screen for pathogen-specific polymorphisms [11]. One of the challenges of this approach is the frequent and unexpected amplification of contaminating template DNA, as observed in control reactions. Another potential problem with targeting 16S rRNA pathogen specific sequences is unexpected polymorphisms. Hence, naturally occurring variants may not be represented on the microarray, and failure to detect the variants would represent false negatives [11]. Another common PCR based approach detects pathogen type by amplification of a specific set of genetic markers that are measured on an array that has several probes for genes from a set of organisms. Such tests have been used for food-borne bacteria such as *E. coli* O157:H7 [12], viruses [13] and mixtures of pathogens [14]. The drawback of using this approach with multiplex PCR primers sets is the generation of spurious products [11]. Array based technologies using 70-mer oligonucleotide probes derived from pathogen specific genes have similar factors that require consideration. For instance, viral detection using a microarray composed of 1,600 unique viral oligonucleotides (70-mers) derived from 140 distinct viral genomes has been previously demonstrated

[15] as a powerful viral detection mechanism, but the drawback of this strategy is that only the group of known pathogen-specific genes will be queried.

Given the enormous spectrum of genetic possibilities, only a highly parallel field deployable technology that is universal in nature has near-term potential to address these needs. The initial vision for a universal DNA microarray was a matrix of oligonucleotide containing features with unique n-mer probes [16]. This matrix could in theory be used to query a biological sample for the presence of any nucleic acid sequence. This technique requires constructing an array that contains 4^n features. Larger values of n infuse greater specificity into the arrayed probes, but as n increases the number of required features grows rapidly. This universality is obtained by synthesizing a combinatorial n-mer array containing all 4^n possible sequences of length n [17]. The key issue is to find a value of n that is large enough to afford sufficient hybridization specificity, yet small enough to be practically fabricated and analyzed.

We have previously demonstrated the utility of a genome sequence-independent microarray for identifying genetic differences [18, 19]. The initial prototype of universal arrays used oligonucleotide probe lengths of 12 and 13 bases. From 4^{12} possible probes, a subset of 14,283 probes was

synthesized using *in situ* synthesis technology and digital optical chemistry (DOC) [20-22]. Fluorescently labelled genomic DNA was hybridized to produce unique informative patterns (i.e. bio-signatures) on a test set of pathogens and host (*Bacillus subtilis, Yersinia pestis, Streptococcus peumoniae, Bacillus anthracis,* and *Homo sapiens*). In addition, we have shown that a custom microsatellite microarray can be used to demonstrate global differences between species by measuring hybridization intensities for every possible repetitive nucleotide motif from 1-mers to 6-mers [19]. Further we have used genome sequence independent microsatellites to identify global differences in the genomes of 93 cancer, cancer-free and high risk patient cell line samples [23]. This paper describes a larger high density oligonucleotide microarray with 370,000 elements, called Universal Bio-signature Detection Array (UBDA), designed by our laboratory and commercially produced by Roche-Nimblegen (Madison, WI) using light-directed photolithography [16, 24]. The platform design which consists mainly of probes, that are tailored to be genome independent, is mathematically derived and therefore unbiased (Additional file 1, Table S1). This strategy exploits the unique signature of a sample in the form of a pattern generated from hybridization of any unknown genome (DNA or cDNA) to a very high-density species-independent oligonucleotide

microarray. *Brucella* species and several other pathogens were used as examples to demonstrate this forensics technology platform. Hybridization patterns are unique to a genome, and potentially to different isolates or a mixture of organisms. These techniques may be especially useful in evaluating and differentiating species whose genome has not yet been sequenced.

2.3 Results

2.3.1 UBDA array sensitivity and specificity of probe hybridization

DNA microarrays using oligonucleotides are widely used in biological research and are usually sequence specific. Two primary types of parameters are required to evaluate the robustness and sensitivity of DNA microarray experiments- labelling and hybridization [16]. Sensitivity of a given array platform is often defined as the minimum signal detected by the array scanning system [25]. In our case we have used labelling controls, where specified DNA molecules (70-mer oligonucleotides) are spiked into experimental human genomic DNA samples prior to fluorescent labelling. A set of six synthetic 70-mer oligonucleotides (Additional file 2, Table S2) was designed to be spiked into each labelling reaction and hybridized to a constellation of 361 probes that were replicated five times on the array. We compared signal intensity values from control probes on the array hybridized

with human genomic DNA and 70-mer oligonucleotides spiked into a separate sample of human genomic DNA. Each spike-in concentration was added on an individual array. We measured sensitivity of the array as a decrease in the correlation coefficient R^2 value in the signal intensity from human genomic DNA spiked with 70-mer oligonucleotides when compared to the un-spiked human genomic DNA sample. The sensitivity of the UBDA was examined by the addition of 70-mer synthetic oligonucleotides to the labelling reaction of human genomic DNA sample (Cy-3 label). Spike-in control synthetic 70-mer oligonucleotides were added at varying concentrations; 4.5 picomolar, 41 picomolar, 121 picomolar and 364 picomolar respectively. Figure 1 elucidates that the sensitivity range of detection for the UBDA is between 364 picomolar and 121 picomolar as seen by the decreased R^2 values of 0.84 and 0.92 respectively for perfect match probes for these two concentrations when compared to the un-spiked human genomic DNA sample. The sensitivity of detection is estimated between a concentration of 364 picomolar and 121 picomolar. At concentrations lower than 121 picomolar, the R^2 value for perfect match probes is 0.96 which is within the ability to resolve samples statistically and confirms that there was no detectable variation at the lower oligonucleotide spike-in at these concentrations. This evaluation demonstrates the variability

of signal intensities contributed by differences in oligonucleotide concentrations spiked into the human DNA sample compared to the un-spiked human DNA sample. Regression analysis of probe signal intensity values from the mis-matched probes in the data set are in Additional file 3, Figures S1A-S1D. We have assessed array variability over several arrays using a common human DNA sample in the reference channel. We obtained an R^2 value of 0.94 ±0.06.

The specificity of the computationally derived 9-mer probes on the UBDA array was studied using the selectivity of the middle nucleotide in each probe. We hypothesized that DNA strands generally will not hybridize efficiently to any probe for which there are multiple mismatches in proximity to the center most base. The array design was based upon the prediction that the use of relatively short probes (15-21 mers) would result in the middle approximately 9 bases dominating hybridization kinetics. Probes on the UBDA that contained the StuI site (AGG^CCT) were located and classified by the nucleotide position of the cut point, relative to the center of the probe on the microarray by a custom computer code. DNA was digested to completion with StuI, and compared to matched DNA that was not digested. Each of the 9-mer probes with StuI restriction enzyme sites were binned depending on the nucleotide position of the StuI restriction site

relative to the center of the probe. Thus probes with the StuI restriction enzyme site were binned in terms of base location according to the position of the StuI restriction enzyme cut site with respect to the center of the probe. As expected, probes with restriction enzyme site in the center of the probe displayed the highest degree of specificity demonstrated by a reduction in signal. A $\log_2$ fold change of -0.23 was obtained when comparing digested DNA to undigested DNA, averaged over microarray probes with the restriction enzyme site at the center of the probe. Microarray probes with the StuI site located at the center demonstrated reduced intensity, confirming specificity of genomic DNA to hybridize to the center of the probe. The trend of the $\log_2$ fold change increased as the StuI restriction enzyme site moved away from the center of the probe with the average results increasing towards zero (Additional file 4, Figure S2). Thus, confirming that the center nucleotide is the most selective in the hybridization complexes.

2.3.2 Identification of synthetically mixed pathogen sample

To establish the ability to decipher a synthetically mixed sample on the UBDA array, *Lactobacillus plantarum* [GenBank accession number ACGZ00000000, genome size 3,198,761 bases] and *Streptococcus mitis*[26] [Genbank accession number FN568063, genome size 2,146,611 bases] genomic DNA were mixed in a ratio of 4:1 (2.53×10^8 copies of *L.*

plantarum to 0.57 x 10^8 copies of *S. mitis* genomes) for a total of 1 µg of DNA, and thus adjusted for copy number of each of the two genomes and hybridized to the array. In addition, pure genomic DNA samples from *L. plantarum* and *S. mitis* were also hybridized individually on separate arrays. The minimum amount of sample required to be detected by hierarchical clustering was determined by an assumption that the mixed sample would cluster under the same node with known samples. As seen from Figure 2, the mixed sample comprising of *Lactobacillus plantarum* and *Streptococcus mitis* groups with pure samples from *L. plantarum* and *S. mitis* (as shown in Figure 2, lane 1, 2 and 3). These results show that if 25 % of the sample is from a second genome, it will group with the higher copy genome on the dendrogram heat map generated from the hierarchical clustering algorithm. A sample with *Lactobacillus plantarum* and *Streptococcus mitis* genomic DNA in a 4:1 ratio (2.53 x 10^8 copies of *L. plantarum* to 0.57 x 10^8 copies of *S. mitis* genomes) was spiked-in with 50 ng (1.54 x 10^{10} copies) of pBluescript plasmid (3,000 bases) [27] . However the node for this sample (Figure 2, lane 4) did not cluster with pure samples from *Lactobacillus plantarum* and *Streptococcus mitis*, instead it clustered closest to a pure sample of pBluescript (Figure 2, lane 5). Spike-in from a low complexity plasmid genome with a high copy number genome such as pBluescript can

dominate the signature pattern. The alteration of the signature pattern is so great that the sample cannot be distinguished on the dendrogram from pure bacterial genomes. Therefore, we are in the process of developing algorithms which will produce a similarity score for a given genome in a mixed genome sample by comparing it to a wide spectrum of species in our genome signature repository.

2.3.3 Identification of genetic signatures from closely related *Brucella* species

The spectrum of organisms chosen for hybridization on this array, were primarily bio-threat zoonotic agents infecting farm animals. Our initial studies were based on the ability of the 9-mer probe signal intensities to distinguish between different *Brucella* species. Currently, there are nine recognized species of *Brucella* based on host preferences and phenotypic preferences. Six of those species are *Brucella abortus* (cattle), *Brucella canis* (dogs), *Brucella melitensis* (sheep and goat), *Brucella neotomae* (desert wood rats), *Brucella ovis* (sheep) and *Brucella suis* (pigs) [28]. All of these species are zoonotic except *B. neotomae* and *B. ovis*. Raw signal values from the pair data files for the Cy3 channel were background corrected and quantile normalized [29]. Signal intensities related to the 9-mer data set were parsed from the data file using a PERL script. These files

were imported into the GeneSpring GX (Agilent, Santa Clara, CA) program. Data from these files was clustered using the hierarchical clustering algorithm to generate a heat map and identify a pattern within the underlying data.

The dendrogram of this heat map which runs vertically along the left side of the heat map in Figure 3 shows the unique bio-signature patterns from 9-mer probes obtained from *Brucella suis* 1330, *Brucella abortus* RB51, *Brucella melitensis* 16M, *Brucella abortus* 86-8-59and *Brucella abortus* 12. Normalized data from the 9-mer data set were filtered for intensity signals greater than the 20^{th} percentile. Only intensity signals with a fold change of 5 or greater were included. These 2,267 elements were subjected to a hierarchical clustering algorithm with Euclidean distance being used as a similarity measure. Centroid linkage rule was applied in the clustering algorithm. The signal intensity values were represented as a $\log_2$ scale. One of the array features was pathogen specific probes designed for independent validation. These probes are species specific to a small set of pathogens including Avian Influenza Virus, Rift Valley Fever Virus, Foot and Mouth Disease Virus, *Brucella melitensis* 16M, *Brucella suis* 1330 and *Brucella abortus* biovar 1 strain 9-941 (Additional file 1, Table S1).

The genomes of *B. melitensis* and *B. suis* have been completely sequenced (28, 29). Comparative genome analysis for these genomes shows that the two genomes are extremely similar. The sequence identity for most open reading frames (ORFs) was 99% or higher [30]. We computationally evaluated the published genome sequences for *B. suis* 1330 [30] and *B. melitensis* 16M [31] to determine the specific instances in the genome sequence of each 9 base core probe sequence from the array. Normalized signal intensity for each of the 262,144 9-mer probes represented on the array were divided by the corresponding counts of 9-mer probe occurrences for both *B. suis* and *B. melitensis*. The resulting values for a set of 32,000 probes were then plotted as illustrated in Figure 4, with *B. melitensis* and *B. suis* (signal intensity/counts) on the ordinate and abscissa, respectively. Pearson's correlation coefficient was subsequently calculated ($\rho = 0.93$ as shown). This correlation value indicates that the 9-mer probe signal intensities are in agreement with 'known' genome sequence similarity scores for *B. melitensis* and *B. suis*.

2.3.4 Taxonomic phylogenetic relationships between organisms hybridized on the UBDA array

Phylogenetic trees are used as a tool in comparative sequence analysis to illustrate the evolutionary relationships among sequences. To create a

phylogenetic tree based on 9-mer signal intensities, genomes listed in (Additional file 5, Table S3) were compared pair-wise, using the Pearson correlation measure (Figure 5). In this study, we demonstrate the use of signal intensities generated from 9-mer probe data to clearly cluster hosts and pathogens into to their 'known' phylogenetic relationships. We have previously shown that a custom microsatellite microarray can be used to demonstrate global microsatellite variation between species as measured by array hybridization signal intensities. This correlated with established taxonomic relationships [19]. Data obtained from the UBDA arrays (normalized signal intensity values) and computational analysis ($\log_2$ transformed, computed counts within sequenced genomes), for all 262,144 9-mer probes, were treated identically for the purposes of tree building. All 262,144 9-mer data points for each sample were first normalized using GeneSpring (percentile shift normalization followed by baseline to median normalization). A Pearson's correlation matrix was subsequently produced and then converted to a taxonomic tree using the neighbour-joining program within the PHYLIP software suite and TreeView program [32]. Trees were not rooted to any specific organism. The lower branches of the phylogenetic tree as shown in Figure 5 display the segregation and differentiation of the various *Brucella* species. The mixed sample comprising of *L. Plantarum*

and *S. Mitis* (4:1 ratio) was found to be closer to the *L. Plantarum*(ρ= 0.974) versus *S. mitis*(ρ= 0.957) on the phylogenetic tree since there was a higher copy number of this genome in the sample (Figure 5). The tree illustrates that the 9-mer probe intensities can be used in species differentiation. The taxonomic tree is an approximate visualization estimation, using a distance matrix which successfully separated mammalian, bacterial and viral clades.

2.3.5 Samples subjected to DNA amplification are comparable to unamplified samples

In preparation for the UBDA becoming not only a detection assay but also a diagnostic test for the identification of numerous pathogens, it was recognized that pathogens may be present in a given sample at very low copy numbers and may be further diluted by genetic material recovered from the host. Microarrays require 0.5 - 1 µg of high-purity genomic DNA, which may be difficult to obtain from all samples. To overcome this limitation the potential for DNA amplification, artefacts that may significantly alter hybridization to the microarray were examined. To analyze for this possible limitation, a 10 ng (4.89×10^6 copies) aliquot of *Francisella tularensis* LVS strain genomic DNA [Accession number NC_007880, genome size 1,895,994 bases] was amplified using the whole genome amplification method (GenomiPhi V2, GE Healthcare). A total of 1

µg of the resulting amplified DNA was hybridized to the UBDA array and compared to the hybridization pattern resulting from the hybridization of 1 µg of unamplified DNA from the same source. Figure 6 shows a linear regression of the two samples (all 262,144 probes) which resulted in an R^2 value of 0.91, well within the $R^2 = 0.94 +- 0.06$ reproducibility found for the custom microsatellite microarray [19]. This confirms that whole genome amplification of pathogen material in small amounts is comparable to the unamplified genomic sample. We obtained these results using the standard protocol with 10 ng of starting material without optimization. We are targeting a 1-2 nanogram sample size as a starting amount of material in an optimized robust, field sample evaluation.

2.4 Discussion

This is a new forensics array based technology to identify any species. This unique strategy of using patterns generated from hybridization of any unknown genome (DNA or cDNA) to a very high-density species independent oligonucleotide microarray and comparing those patterns to a library of patterns of known samples can be used to identify unknown organisms. Figure 5 shows the grouping of the different genomes into bacterial, viral and eukaryotic genomes. Further the *Brucella* species grouping pattern obtained from the phylogenomic analysis using the

Pearson's correlation matrix shown in Figure 5 are in agreement with *Brucella* species showing hierarchical clustering represented as a similarity matrix shown in Figure 3. The UBDA hybridization patterns are unique to a genome, and potentially to different isolates and to a mixture of organisms. In the future, this forensics method will work by comparing signal intensity readout to a library of readouts established by interrogating a wide spectrum of species which will be available at our website http://discovery.vbi.vt.edu/ubda/. The phylogenetic tree illustrates the ability of 9-mer probes to differentiate among *Brucella* species. Pair-wise comparisons between different genomes can be used as a measure to classify bacterial, viral or mammalian genomes into their respective clades. We have begun to amass a library of 'signatures' to facilitate accurate identification and classification of "unknown" samples. We are currently expanding the repository of available bio-signatures to several hundred genomes including field isolates from bacteria, viruses, host genomes and vectors infected with pathogens. Some of the genomes in this repository are classified in the select agent category. UBDA forensics application has the potential to be compatible with micro-machine based front end sample processing and preparation platforms, thus enabling the production of a highly automated, fast and accurate field-deployable detection system.

Other diagnostic techniques such as PCR or RT-PCR require several primers to be designed which are specific for each genome- bacterial, viral or host. There may be spurious products for primers binding at low specificity. The processing costs should also be taken into consideration for these methodologies. The current cost for the UBDA array is approximately $350 per sample which includes reagents and processing costs. The current turnaround time for this forensics technology is less than 24 hours. This is a single experimental procedure compared to other technologies which involve a series of methods such as serological, biochemical and genomic based. Genome specific arrays are in the similar price range as the UBDA array; however researchers can only assay a single genome or a small subset of them. Currently the UBDA platform requires a turnaround time approximately one day from hybridization on the array to data analysis. A diagnostic laboratory in the field requires proximately two weeks before the identity of a given infectious agent can be determined. These methods usually require several standard serological and biochemical tests that are usually selected and based on the clinical symptoms observed in the field. Serology test results are usually available after 48 hours. Although each of these tests is cost effective in nature, they must be fine-tuned to be pathogen specific.

The UBDA approach can be applied to any genome, even in the presence of background contamination (usually host DNA) for which, the unique pattern will be known. The patterns generated from an unknown sample (secretion, tissue culture, environmental sample, etc) with minimal specimen processing can be identified or at least the most similar related species will be predicted by comparison to a library or a repository of patterns. These techniques may be especially useful in evaluating and differentiating species whose genome has not yet been sequenced. Along with a repository of unique hybridization signatures from the genomes of pathogens and their hosts, this array has the ability to rapidly and adequately identify biological threat agents and newly emerging infectious pathogens that are high risk priorities in bio-defense. Application of this technology has the potential to extend to other areas such as food and environmental microbial monitoring and basic research including, (a) speciation and evolution, (b) human/animal disease biomarker discovery, (c) measurement of the genomic response to a chemical, radiation or other exposure, but most important, (d) pathogen forensics and characterization of natural or engineered variants that may confound other species-specific approaches.

2.5 Conclusions

Genetic signature discovery and identification of pathogenic phenotypes will provide a robust means of discriminating pathogens that are closely related. This array has high sensitivity as demonstrated by the detection of low amounts of spike-in oligonucleotides. Hybridization patterns are unique to a specific genome and these can be used to de-convolute and thus identity the constituents of a mixed pathogen sample. In addition it can distinguish hosts and pathogens by their divergent phylogenomic relationships as captured in their respective 9-mer hybridization signatures. This platform has potential for commercial and government agency applications as a cost effective reliable platform for accurately screening large numbers of samples for bio-threat agents in forensic analysis, screening for pathogens that routinely infect animals and humans, and as a molecular diagnostic of micro-organisms in a clinical environment. This platform is highly attractive, because it has multiplex capacity where knowledge can be drawn from the array hybridization patterns without prior explicit information of the genomes in the samples. These hybridization patterns are being translated into a knowledge base repository of bio-signatures so that future users of this technology can compare and draw inferences related to the sample under study. The data

from these experiments and the array design are located on our web site at http://discovery.vbi.vt.edu/ubda/.

2.6 Methods

2.6.1 Array design details

A custom microarray was designed by this laboratory and manufactured by Roche-Nimblegen (Madison, WI) as a custom 385K (385,000 probe platform) chip to include the following sets of probes; 9-mer, pathogen specific probes; rRNA gene specific, microsatellite and control 70-mer oligonucleotide probes. There were 262,144 9-mer probes, and 20,000 of them were replicated 3 times in total (Additional file 1, Table S1). The 9-mer probes were comprised of a core 9-mer nucleotide and flanked on both sides by three nucleotides, selected to maximize sequence coverage of these basic 15-mers. Probes with low GC content were padded with additional bases at their termini to equalize melting temperatures, with most probes ranging from 15-21 nucleotides in total length. For the 9-mer design, the length of the probes was adjusted to match a melting temperature of 54° C. The array design was based upon the prediction that the use of relatively short probes (15-21 mers) would result in the middle 9 bases dominating hybridization kinetics.

rRNA probes were included in the design to serve as positive controls and confirmation of the 9-mer probes power for differentiating genomes. The rRNA probes were selected from the green gene data (http://greengenes.lbl.gov/cgi-bin/nph-show_probes_2_otu_alignments.cgi), utilizing the complete list of 8,935 OTUs (Operational Taxonomic Unit). One probe was selected for each OTU and probe length was adjusted to a T_m equal to 54° C, as was done for 9-mer design. A mis-match probe (1 mis-match, MM) for each OTU probe was included in the design. Perfect match (PM) 8,935 probes and 8,935 one mis-match MM probes were included in the microarray design. All probes are replicated 3 times on the array. Genome specific probes for *Brucella* spp., Avian Influenza Virus (AIV), Foot and Mouth Disease Virus (FMDV), and Rift-Valley Fever Virus (RVFV) were designed and included on the microarray as an independent test when the array is used to analyze these species. Sequence alignments were performed to determine the similar and unique regions of the pathogens, with probes selected from the unique regions of each pathogen species or sub-type, and excluding sequences similar to host genomes. In total, 1,062 unique probes were selected and are replicated 3 times. Probes dedicated to surveying microsatellite content were designed for every 1- to 6-mer repetitive sequence. For each 1- to 5-mer repetitive sequence, single

mis-match (1MM) probes were also designed. A total of 3,557 unique microsatellite probes were generated and replicated at total of 3 times. Microsatellite probes were included on this array to anchor the results to previous experiments and to aid in the de-convolution of the contribution of host genomic DNA. For higher life forms typically have many microsatellite loci, whereas bacteria and viruses have none or very few in their genome. Gene-specific probes were designed to target important metabolic pathways, such as alcohol dehydrogenase, glucose-6-phosphate isomerase and SHV-like β-lactamase, by using the highly conserved sequences. In total, 432 probes were designed and replicated a total of 3 times.

For labelling controls, a set of six synthetic 70-mer oligonucleotides were designed to be spiked into each labelling reaction and hybridized to a constellation of 361 dedicated probes on the array comprising of perfect match probes (34 probes), 1 mis-match (100 probes), 2 mis-match (137 probes) and 3 mis-match probes (90 probes), ranging from 15-19 nucleotides. The set of 361 probes are replicated 5 times total (Additional file 2, Table S2). Figure 1 shows a comparison of signal intensity values of perfect match control probes on the array generated from human genomic DNA without spike of oligonucleotides to samples with a spiked-in. Regression analysis of signal intensity values from the mis-matched probes

on the data set is in Figures S1A-S1D (Additional file 3). The array design files for each feature category on the UBDA array are in Additional file 6 (9-mer probes) and Additional file 7 (all other probes) and also available at http://discovery.vbi.vt.edu/ubda/.

2.6.2 Microarray procedure

Human genomic DNA was extracted from blood samples collected from a volunteer by the McDermott Center for Human Growth and Development Genetics Clinical Laboratory in accordance with Institutional Review Board at UT Southwestern Medical Center (Dallas, TX). Genomic DNA from *Bos taurus*, *Gallus gallus*, *Meleagris gallopavo*, *Ovis aries*, *Capra hircus* and *Equus caballus* was obtained from Zyagen (San Diego, CA). *Brucella* species, *Cryptosporidium parvum*, *Lactobacillus plantarum*, *Streptococcus mitis*, *Escherichia coli* and Influenza virus genomic DNA was obtained from BEI resources and ATCC (Manassas, VA). The spectrum of organisms chosen for hybridization on the UBDA array was primarily bio-threat zoonotic agents, agents infecting farm animals.

DNA concentration (260nm) and purity (260/280 and 260/230 nm) was assessed by the spectrophotometer and quality by agarose gel electrophoresis. Samples with 260/230 nm ratios greater than 1.8 were used following established protocols for array comparative genomic hybridization

(CGH). Hybridization conditions were optimized to ensure specificity and sensitivity. All DNA test samples (1 µg) were labelled with Cy3 and co-hybridized with the same Cy5-labeled human reference (Promega, Inc, Madison, WI), according to Roche Nimblegen standard microarray labelling procedures. For each microarray, human genomic DNA (Promega, Madison, WI) was labelled with Cy-5 and used as a reference channel in each experiment. DNA labelling, hybridization and data acquisition were performed by Mogene (St. Louis, MO). We tested hybridization temperatures ranging from 30°C to 50°C.For microarray hybridization, a custom buffer (0.5% Triton X-100, 1 M NaCl, and 100 mM Tris-HCl pH 7.5, filtered with a 0.2 micron nitrocellulose filter, prepared fresh) was used at 38°C, and microarrays were washed following Roche Nimblegen's CGH standard techniques (available at www.nimblegen.com). Hybridization conditions were standardized for the UBDA array to minimize any errors that could lead to bias resulting after processing the slides and image scanning on an array scanner. Signals from probes complementary to labelling controls indicate that the post-DNA preparation process, from labelling to hybridization, washing and scanning, were successful. Hybridization, scanning, and data extraction were performed following

Roche NimbleGen standard protocol for CGH arrays, and the resulting raw data were provided via secure web link.

2.6.3 Array data processing and organism classification

A Robust Multi-chip Average (RMA) normalization procedure was performed across all arrays. The procedure included background subtraction and quantile normalization using Nimblescan Software (Roche NimbleGen). After normalization, all 262,144 9-mer probes were extracted from the 370K array using PERL scripts and averaged across the replicate probes. Subsequent statistical analysis was performed using GeneSpringGX 11.0 (Agilent Technologies, Santa Clara, CA). All signal intensity values were $\log_2$ transformed for further analysis. Data were also filtered by intensity values (lower cut off percentile of 20 % for raw signals), and subsequent pair-wise comparisons were performed on the sample data set. Clustering is one of the data mining processes for discovery and identifying patterns in the underlying data. Clustering algorithms partition data into subsets based on similarity and dissimilarity. Clustering methods follow three steps: pattern recognition, use of a clustering algorithm and similarity measure matrix [33]. For pattern recognition, pair-wise comparisons are used between samples to select the features on which the clustering is to be performed. Our experimental platform is comparative genome hybridization

for which hierarchical clustering is used to determine phylogenomic relationships between organisms. Hierarchical clustering [34] transforms a distance matrix of pair-wise similarity measurements between all items into a hierarchy of nested groupings. The hierarchy is represented with a binary tree-like dendrogram. Hierarchical clustering was performed on the resulting data sets, using the Euclidian matrix and centroid linkage to classify various organisms. Data sets were analyzed for *Brucella* species. A cut-off of 5-fold change in hybridization intensity for a given probe was used to reduce the data set to only those meaningful probes that showed a difference between at least one of the pair-wise comparisons.

2.6.4Phylogenetic taxonomic tree based on array intensity

Data obtained from the Universal Bio-Detection Array (normalized signal intensity values that were $\log_2$ transformed) and computational analysis for all 262,144 9-mer probes were treated identically for the purpose of tree building. All 262,144 data points for each of the 20 samples were first RMA normalized. For each sample, a Pearson's correlation matrix was created which included self-similarity and similarity to the remaining 19 samples from all the 262,144 data points of each sample. The resulting distance matrix was used to produce a phylogenetic tree, using the neighbour-joining method within the PHYLIP software suite and TreeView.

2.6.5Whole genome amplification

Francisella tularensis LVS strain genomic DNA, starting material, 10 nanogram was amplified using whole genome amplification method as defined (GenomiPhi V2, GE Healthcare). We obtained 2-3 µg of whole genome amplified DNA from 10 ng of starting genomic DNA.

2.7 Acknowledgements

This work was funded by Department of Homeland Security through the FAZD Center (National Center of Excellence for Foreign Animal and Zoonotic Disease Defense) at Texas A & M University and Virginia Bioinformatics Institute director's funds. SJS received support from a trans-disciplinary fellowship from Virginia Tech and Virginia Bioinformatics Institute. We would like to extend a special thanks to Angela George and Dale Preston of the Texas Animal Health Commission, Austin, Texas for assistance with sample preparation. We thank Dr. Abey Bandara and Dr. Tom Inzana at Virginia Tech for providing the *Francisella tularensis* LVS strain genomic DNA. We would like to extend a special thanks to Greg Thorne and Shaukat Rangwala with MoGene their valuable technical assistance. We appreciate the assistance of Linda Gunn, Renee Nester, Traci Roberts and Laurie Spotswood for administrative assistance**.** We also appreciate Zyagen and BEI resources for providing genomic DNA.

2.8 Attribution

SJS oversaw the project, coordinated the study design, carried out the analysis and subsequent parsing and data interpretation and drafted the manuscript. JNW initiated the project, participated in preliminary technical analyses. CLG participated in manuscript editing. LM participated in manuscript editing, created the UBDA website and provided computation expertise. ZS designed the array and provided computation expertise. JM provided useful discussions and technical assistance. LGA provided DNA samples, data interpretation and participated in manuscript editing. HRG conceived of the study, participated in the study design and mentored in drafting the manuscript. All authors have agreed to all the content in the manuscript, including the data as presented.

2.9 Bibliography

1. Pannucci J, Cai H, Pardington PE, Williams E, Okinaka RT, Kuske CR, Cary RB: **Virulence signatures: microarray-based approaches to discovery and analysis**. *Biosens Bioelectron* 2004, **20**(4):706-718.
2. Ruiz-Mesa JD, Sanchez-Gonzalez J, Reguera JM, Martin L, Lopez-Palmero S, Colmenero JD: **Rose Bengal test: diagnostic yield and use for the rapid diagnosis of human brucellosis in emergency departments in endemic areas**. *Clin Microbiol Infect* 2005, **11**(3):221-225.

3. Bricker BJ: **PCR as a diagnostic tool for brucellosis**. *Vet Microbiol* 2002, **90**(1-4):435-446.

4. Bounaadja L, Albert D, Chenais B, Henault S, Zygmunt MS, Poliak S, Garin-Bastuji B: **Real-time PCR for identification of Brucella spp.: a comparative study of IS711, bcsp31 and per target genes**. *Vet Microbiol* 2009, **137**(1-2):156-164.

5. Hinic V, Brodard I, Thomann A, Holub M, Miserez R, Abril C: **IS711-based real-time PCR assay as a tool for detection of Brucella spp. in wild boars and comparison with bacterial isolation and serology**. *BMC Vet Res* 2009, **5**:22.

6. Her M, Kang SI, Kim JW, Kim JY, Hwang IY, Jung SC, Park SH, Park MY, Yoo H: **A genetic comparison of Brucella abortus isolates from animals and humans by using an MLVA assay**. *J Microbiol Biotechnol* 2010, **20**(12):1750-1755.

7. Whatmore AM, Perrett LL, MacMillan AP: **Characterisation of the genetic diversity of Brucella by multilocus sequencing**. *BMC Microbiol* 2007, **7**:34.

8. Abril C, Thomann A, Brodard I, Wu N, Ryser-Degiorgis MP, Frey J, Overesch G: **A novel isolation method of Brucella species and**

molecular tracking of Brucella suis biovar 2 in domestic and wild animals. *Vet Microbiol* 2011.

9. De Santis R, Ciammaruconi A, Faggioni G, D'Amelio R, Marianelli C, Lista F: **Lab on a chip genotyping for Brucella spp. based on 15-loci multi locus VNTR analysis**. *BMC Microbiol* 2009, **9**:66.

10. Scott JC, Koylass MS, Stubberfield MR, Whatmore AM: **Multiplex assay based on single-nucleotide polymorphisms for rapid identification of Brucella isolates at the species level**. *Appl Environ Microbiol* 2007, **73**(22):7331-7337.

11. Call DR: **Challenges and opportunities for pathogen detection using DNA microarrays**. *Crit Rev Microbiol* 2005, **31**(2):91-99.

12. Call DR, Brockman FJ, Chandler DP: **Detecting and genotyping Escherichia coli O157:H7 using multiplexed PCR and nucleic acid microarrays**. *Int J Food Microbiol* 2001, **67**(1-2):71-80.

13. Chizhikov V, Wagner M, Ivshina A, Hoshino Y, Kapikian AZ, Chumakov K: **Detection and genotyping of human group A rotaviruses by oligonucleotide microarray hybridization**. *J Clin Microbiol* 2002, **40**(7):2398-2407.

14. Wilson WJ, Strout CL, DeSantis TZ, Stilwell JL, Carrano AV, Andersen GL: **Sequence-specific identification of 18 pathogenic**

microorganisms using microarray technology. *Mol Cell Probes* 2002, **16**(2):119-127.

15. Wang D, Coscoy L, Zylberberg M, Avila PC, Boushey HA, Ganem D, DeRisi JL: **Microarray-based detection and genotyping of viral pathogens**. *Proc Natl Acad Sci U S A* 2002, **99**(24):15687-15692.

16. Pease AC, Solas D, Sullivan EJ, Cronin MT, Holmes CP, Fodor SP: **Light-generated oligonucleotide arrays for rapid DNA sequence analysis**. *Proc Natl Acad Sci U S A* 1994, **91**(11):5022-5026.

17. Royce TE, Rozowsky JS, Gerstein MB: **Toward a universal microarray: prediction of gene expression through nearest-neighbor probe sequence identification**. *Nucleic Acids Res* 2007, **35**(15):e99.

18. Belosludtsev YY, Bowerman D, Weil R, Marthandan N, Balog R, Luebke K, Lawson J, Johnston SA, Lyons CR, Obrien K *et al*: **Organism identification using a genome sequence-independent universal microarray probe set**. *Biotechniques* 2004, **37**(4):654-658, 660.

19. Galindo CL, McIver LJ, McCormick JF, Skinner MA, Xie Y, Gelhausen RA, Ng K, Kumar NM, Garner HR: **Global microsatellite**

content distinguishes humans, primates, animals, and plants. *Mol Biol Evol* 2009, **26**(12):2809-2819.

20. Luebke KJ, Balog RP, Mittelman D, Garner HR: **Digital optical chemistry: A novel system for the rapid fabrication of custom oligonucleotide arrays**. *Microfabricated Sensors* 2002, **815**:87-106.

21. Luebke KJ, Balog RP, Garner HR: **Prioritized selection of oligodeoxyribonucleotide probes for efficient hybridization to RNA transcripts**. *Nucleic Acids Research* 2003, **31**(2):750-758.

22. Balog R, Hedhili MN, Bournel F, Penno M, Tronc M, Azria R, Illenberger E: **Synthesis of Cl-2 induced by low energy (0-18 eV) electron impact to condensed 1,2-C2F4Cl2 molecules**. *Physical Chemistry Chemical Physics* 2002, **4**(14):3350-3355.

23. Galindo CL: **Sporadic breast cancer patient's germline DNA exhibit an AT-rich microsatellite signature**. *Genes, Chromosomes and Cancer* 2010, **accepted**.

24. McGall GH, Fidanza JA: **Photolithographic synthesis of high-density oligonucleotide arrays**. *Methods Mol Biol* 2001, **170**:71-101.

25. Kane MD, Jatkoe TA, Stumpf CR, Lu J, Thomas JD, Madore SJ: **Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays**. *Nucleic Acids Res* 2000, **28**(22):4552-4557.

26. Denapaite D, Bruckner R, Nuhn M, Reichmann P, Henrich B, Maurer P, Schahle Y, Selbmann P, Zimmermann W, Wambutt R *et al*: **The genome of Streptococcus mitis B6--what is a commensal?***PLoS One* 2010, **5**(2):e9426.

27. Alting-Mees MA, Short JM: **pBluescript II: gene mapping vectors**. *Nucleic Acids Res* 1989, **17**(22):9494.

28. Morgan WJ: **Brucella classification and regional distribution**. *Dev Biol Stand* 1984, **56**:43-53.

29. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP: **Exploration, normalization, and summaries of high density oligonucleotide array probe level data**. *Biostatistics* 2003, **4**(2):249-264.

30. Paulsen IT, Seshadri R, Nelson KE, Eisen JA, Heidelberg JF, Read TD, Dodson RJ, Umayam L, Brinkac LM, Beanan MJ *et al*: **The Brucella suis genome reveals fundamental similarities between animal and plant pathogens and symbionts**. *Proc Natl Acad Sci U S A* 2002, **99**(20):13148-13153.

31. DelVecchio VG, Kapatral V, Redkar RJ, Patra G, Mujer C, Los T, Ivanova N, Anderson I, Bhattacharyya A, Lykidis A *et al*: **The**

genome sequence of the facultative intracellular pathogen Brucella melitensis. *Proc Natl Acad Sci U S A* 2002, **99**(1):443-448.

32. Page RD: **TreeView: an application to display phylogenetic trees on personal computers**. *Comput Appl Biosci* 1996, **12**(4):357-358.

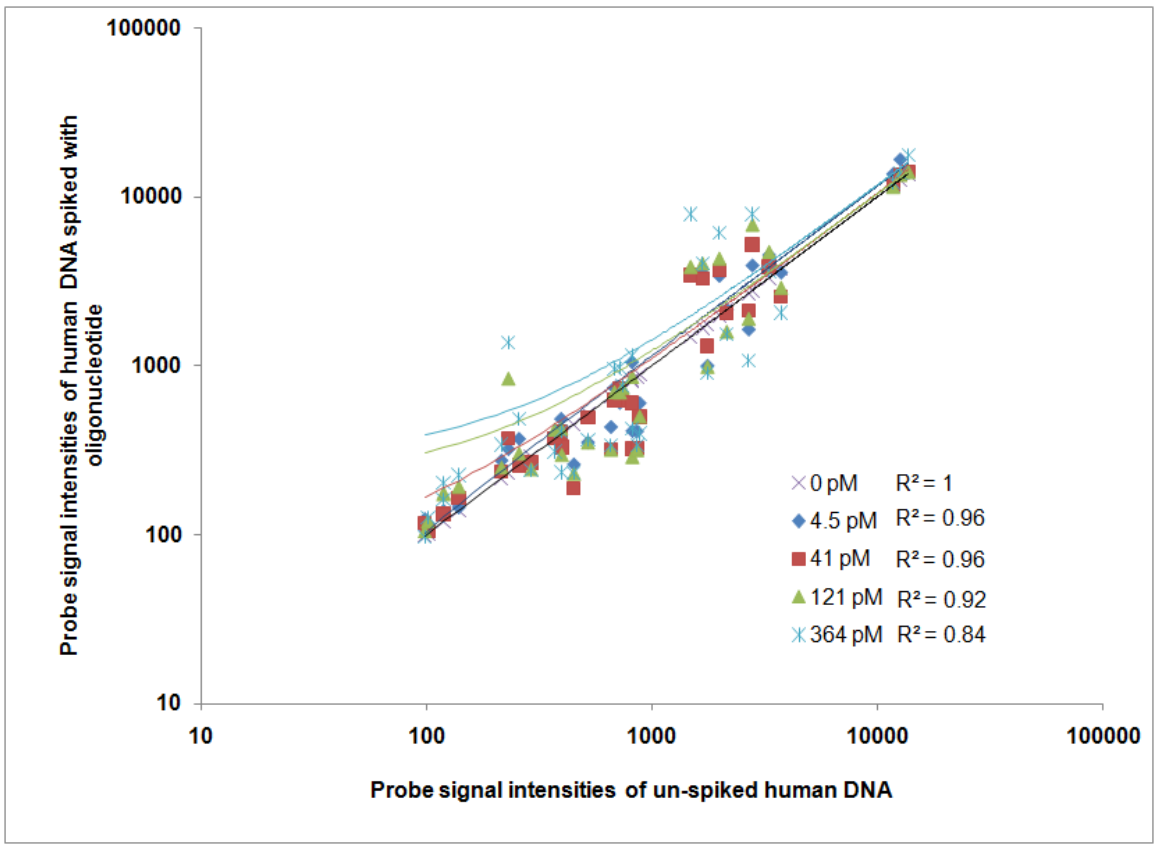
33. Frades I, Matthiesen R: **Overview on techniques in cluster analysis**. *Methods Mol Biol* 2010, **593**:81-107.

34. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns**. *Proc Natl Acad Sci U S A* 1998, **95**(25):14863-14868.

2.10 Figures

2.10.1 Figure 1 - Array sensitivity determined by control probe signal intensity values

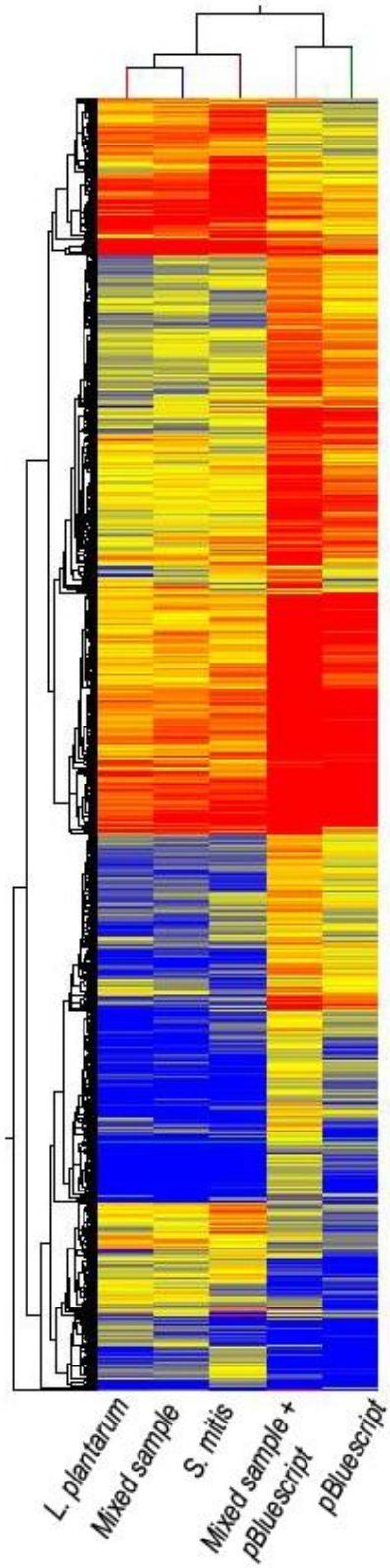
Human genomic DNA spiked with 70-mer oligonucleotides at different concentrations was compared against the same sample without oligonucleotides. Normalized signal intensity values from the Cy3 channel were plotted on a log scale and compared using linear regression from human genomic DNA samples with and without 70-mer oligonucleotides spiked into the labelling reaction. The probes being assessed on this scatter plot are perfect matches to the 70-mer oligonucleotide sequence. Each notation on the graph represents a specific concentration of spiked-in 70-mer oligonucleotides on an individual array. The oligonucleotides were spiked into the labelling reaction at a concentration range from 4.5 pM to 364 pM. The divergence of R^2 value from that with no spike-in was used to measure the sensitivity of detection on the array.



2.10.2 Figure 2 - Hierarchical clustering of mixed samples demonstrates the resolution capabilities of the UBDA array

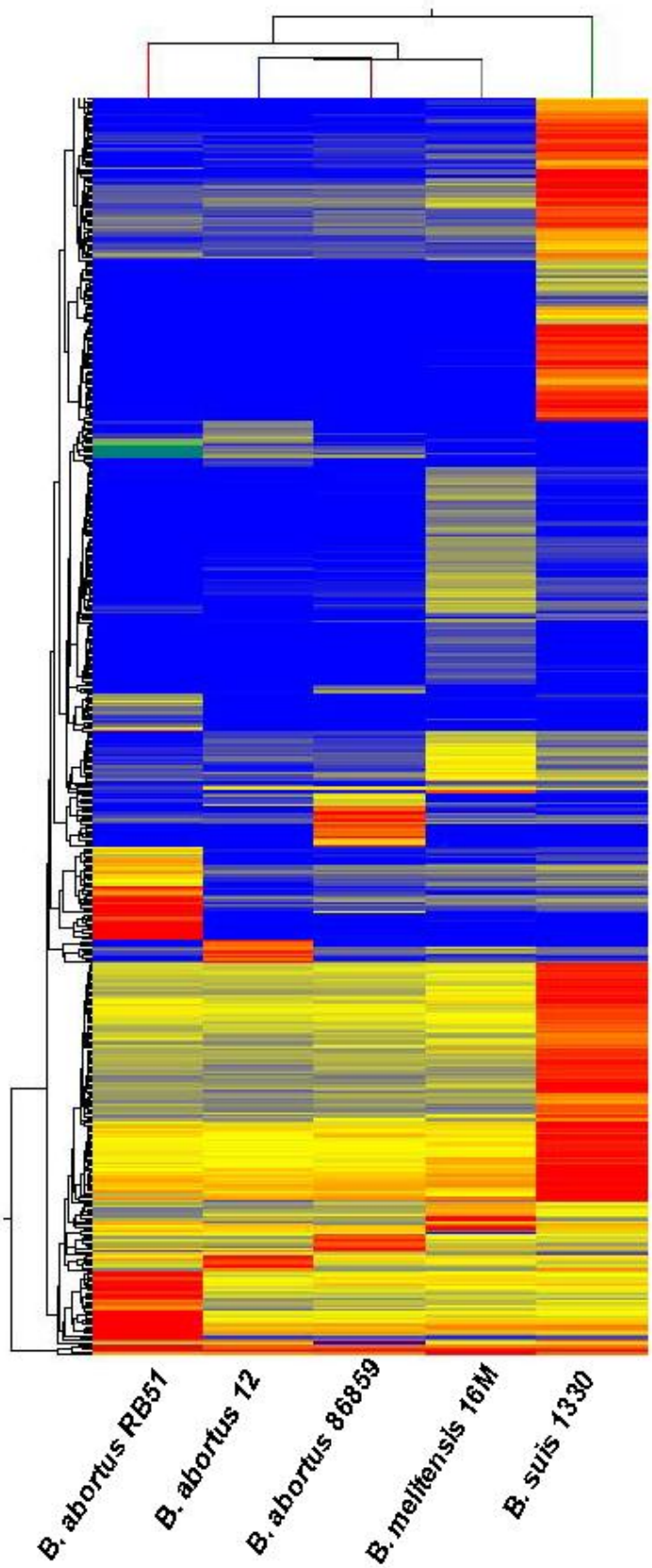
This dendogram and heat map illustrates a unique bio-signature pattern obtained from *Lactobacillus plantarum*, mixed sample (synthetic mixture in a 4:1 ratio of *L. plantarum and Streptococcus mitis*), *S. mitis*, mixed sample (a synthetic mixture of *L. plantarum* and *S. mitis* genomic DNA in a ratio of 4:1 with a spike-in of pBluescript plasmid at 50 ng) and pBluescript plasmid. Normalized data from the 9-mer data set were filtered for intensity signals greater than the 20th percentile. Only intensity signals with a fold change of

5 or greater were included. These 36,059 elements were subjected to hierarchical clustering with Euclidean distance being used as a similarity measure. The signal intensity values were represented on a $\log_2$ scale and range from 8.4 to 13.4.


L. plantarum
Mixed sample
S. mitis
Mixed sample + pBluescript
pBluescript

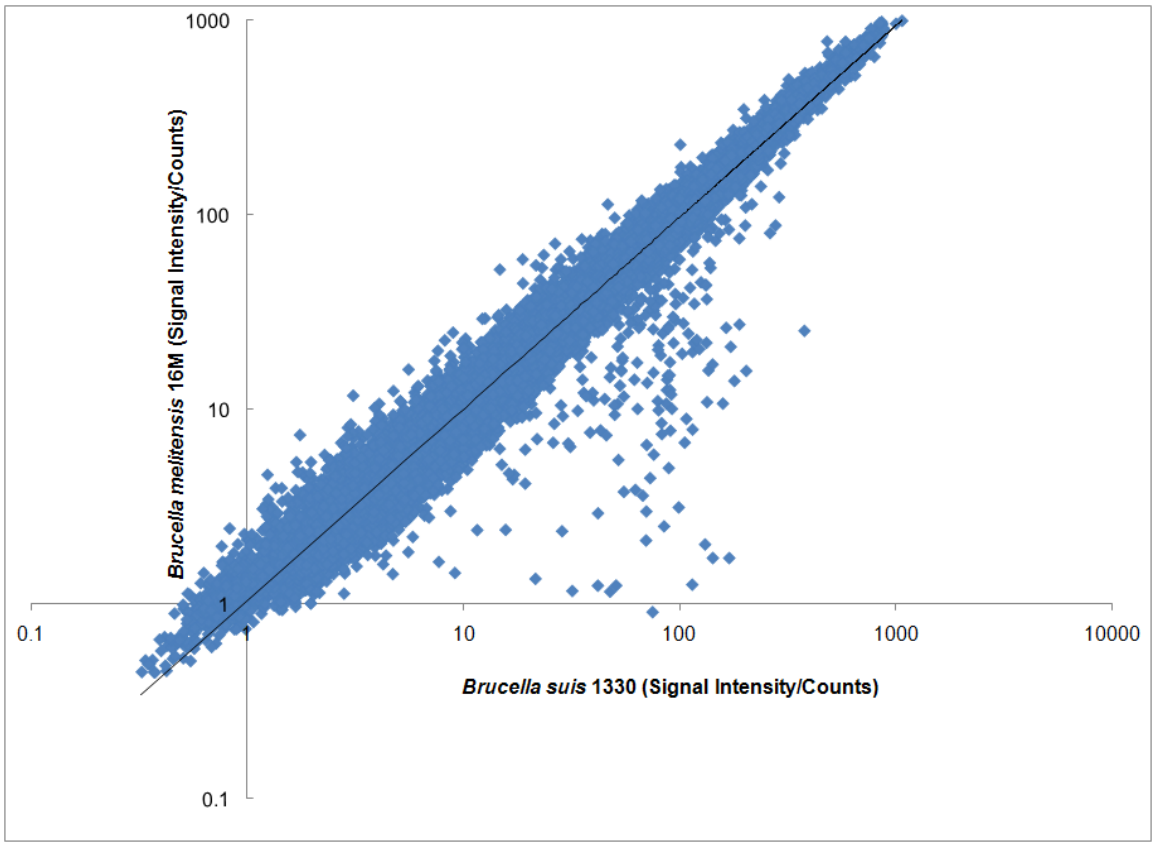
2.10.3 Figure 3 - Unique 9-mer probe bio-signatures from hybridization of *Brucella* genomes demonstrates ability to resolve highly similar genomes

This dendogram illustrates the unique bio-signature obtained from *Brucella abortus RB51*, *Brucella abortus* 12, *Brucella abortus* 86-8-59, Brucella *melitensis* 16M and Brucella *suis* 1330. Normalized data from the 9-mer data set were filtered for intensity signals greater than the 20^{th} percentile. Only intensity signals with a fold change of 5 or greater were included. These 2,267 elements were subjected to hierarchical clustering with Euclidean distance being used as a similarity measure. The signal intensity values were represented as a $\log_2$ scale. The range of $\log_2$ values are from 7.2 to 13.


B. abortus RB51
B. abortus 12
B. abortus 86859
B. melitensis 16M
B. suis 1330

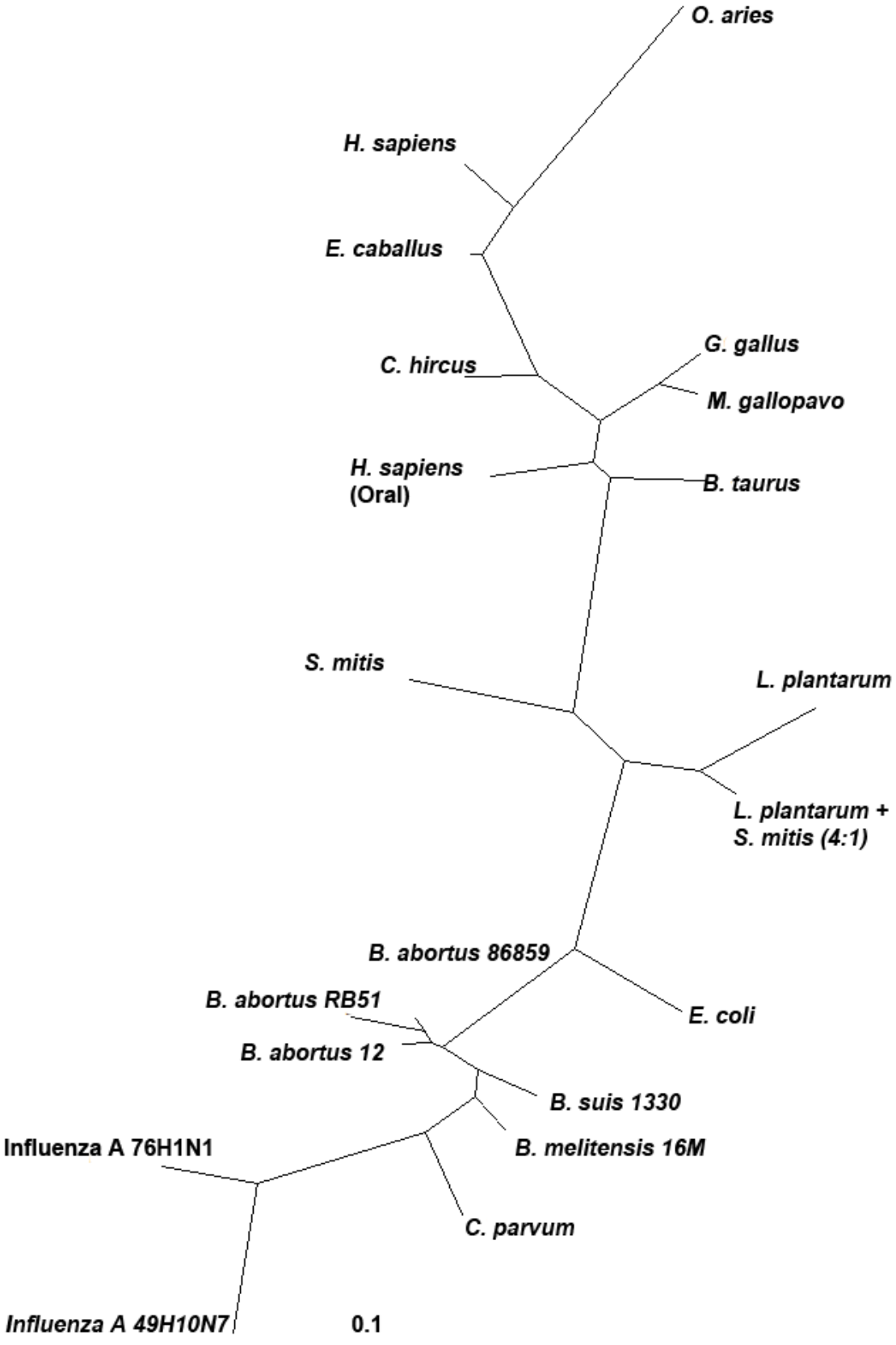
2.10.4 Figure 4 - Correlation of *Brucella Suis 1330* and *Brucella melitensis 16M* was computed by a ratio of signal intensity divided by counts of 9-mer probe occurrences in the respective genomes

Normalized signal intensity for each of the 262,144 9-mer probes represented on the array were divided by the corresponding counts of 9-mer probe occurrences in the respective genome sequences for both *B. suis* and *B.melitensis*. The resulting values for a set of 32,000 probes were then plotted, with *B. melitensis* and *B. suis* (signal intensity/counts) on the ordinate and abscissa, respectively. Pearson's correlation coefficient was subsequently calculated ($\rho = 0.93$ as shown).



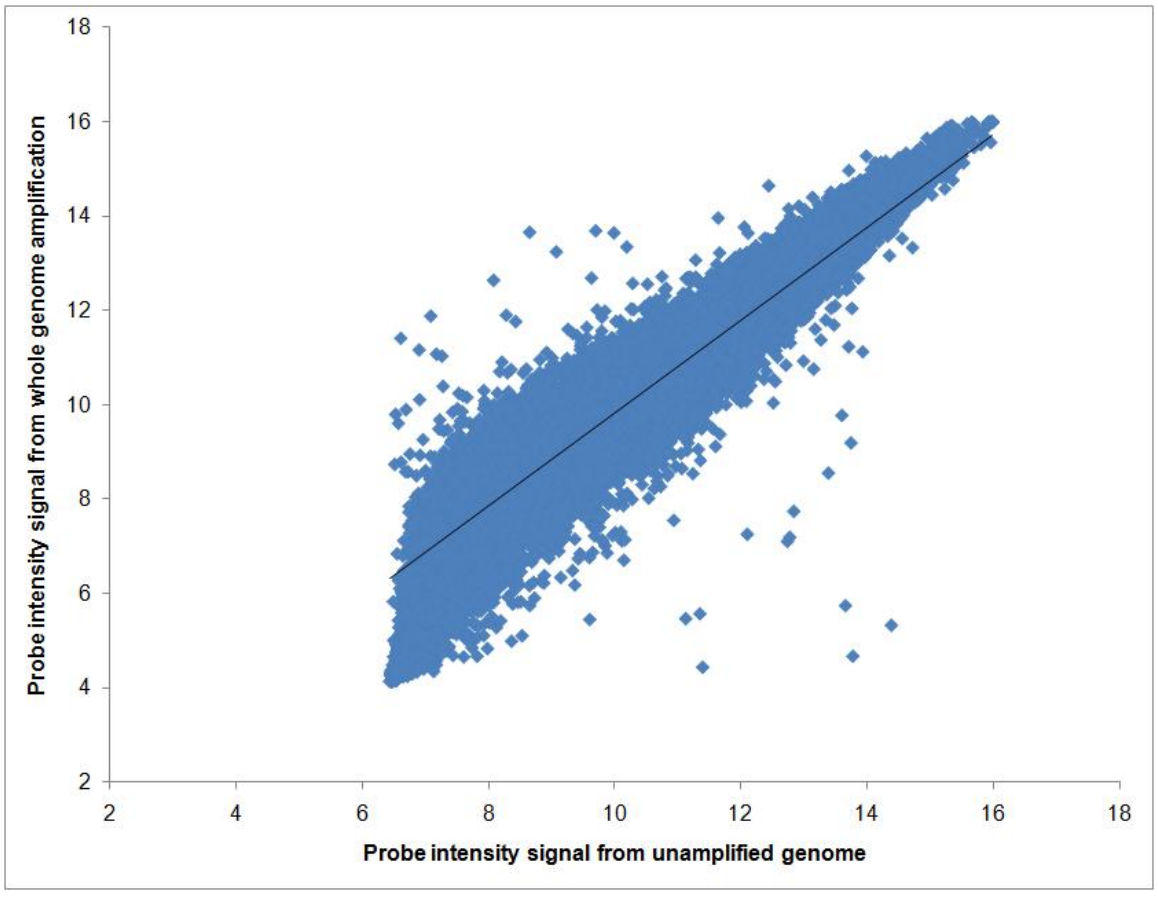
2.10.5 Figure 5 - Phylogenetic relationships from the 9-mer probe set between organisms hybridized on the UBDA array

All 262,144 9-mer data points for each of the 20 samples were RMA normalized and $\log_2$ transformed. A Pearson correlation matrix was created by comparing each sample against all other samples. The values were used to generate a taxonomic relationship tree using the PHYLIP software. The taxonomic tree, as visualized in the Treeview program, shows the separation between mammalian, bacterial and viral genomes.


O. aries
H. sapiens
E. caballus
G. gallus
C. hircus
M. gallopavo
H. sapiens
(Oral)
B. taurus
S. mitis
L. plantarum
L. plantarum +
S. mitis (4:1)
B. abortus 86859
B. abortus RB51
E. coli
B. abortus 12
B. suis 1330
Influenza A 76H1N1
B. melitensis 16M
C. parvum
Influenza A 49H10N7
0.1

2.10.6 Figure 6 - Bivariate Fit of *Francisella tularensis* whole genome amplified genomic DNA (log_2 values) by unamplified genomic DNA (log_2 values).

A linear regression of the two samples resulted in an R^2 value of 0.91, confirming that whole genome amplification of pathogen material such as *Francisella tularensis* LVS genomic DNA in small amounts (10 ng starting material) is comparable to the unamplified genomic sample.



2.11 Additional files

2.11.1 Table S1 Distribution of probe types included in the UBDA design

The table describes the different data set features on the array.

Feature Category	Feature count with replicates	Description
9-mer (4^9 probes equivalent to 262, 144 probes)	302,144	Sequence independent including every 9bp combination (20,000 probes replicated 3x total)
rRNA	35,756	Probes designed from 16s rRNA sequences (replicated 3x total)
Gene specific probes	1,296	Probes derived from alcohol dehydrogenase, glucose-6-phosphate isomerase and SHV-like β-lactamase (replicated 3x total)
Pathogen specific probes	3,186	Specific to *Brucella* spp., Avian Influenza virus, Rift Valley Fever Virus and Foot and Mouth Disease Virus (replicated 3x total)
Microsatellites probes	10,671	Every 1-mer to 6-mer repetitive sequences (replicated 3x total)
70 mer oligonucleotide probes	1,805	Measure and calibrate labelling and hybridization efficiency and specificity (replicated 5x total)
Total probes	**354,858**	

2.11.2 Table S2 Sequence of labelling control oligonucleotide probes

Sequence information of the 70-mer oligonucleotides used in the spike-in study to determine the sensitivity of the UBDA array.

Probe	**Sequence**
1	CTACCTCCGATCGCGATACAGAATGAATCATGGGATTCATATTGAGACAGTTGTTCTGTCTTGGCTGGAC
2	ACCGACTAAAGGTAATGACCATTGGTGAATTGATACCGTCTACAACCCTCCAATGTTACAAGAGACTAAC
3	AATGGAAAAGTTGGCTCCGGGTCTTACACCTGCGTGCCTCGATGCTAACAGACCCCAGGGCGACCGATAT
4	TGTCAGACCGTAGCGTTGCAGCTTCAGTCACACAGCTTTGGCTTAGAGATTCCGCCAAAAGAACCATCCT
5	ATGCGTATGCTGCAACCAACGATTAATCCGGTCTCCTATAGGACATCGCGATAAGATCGTCTAACGTAGC
6	GGACCGCTAGTTGTCGGACCATACATTGATGTTGGAATATGCGGATACCCAGGCAATCATTTCACCTTTT

2.11.3 Figures S1A – S1D Regression analysis of signal intensity values generated from spike in of different concentrations of 70-mer oligonucleotides to human genomic DNA versus the un-spiked sample.

Average Cy3 signal intensity values were plotted on a log scale. Normalized signal intensities from the Cy3 channel, which were human genomic DNA samples with and without the addition of 6 spike-in 70-mer oligonucleotides, were compared by linear regression. Each notation on the graph represents

an individual control probe spot on the array. The R^2 value is displayed in the lower right quadrant of the graph. Purple x represent perfect match probes (PM), blue diamonds represent 1 mis-match (MM) probes, red squares represent probes with 2 mis-matches and green triangles represent 3 mis-matches. (A) At 4.5 picomolar of oligonucleotide spike-in, an R^2 value of 0.96 was obtained for probes with a PM, 0.93 for 1 MM, 0.95 for 2MM and 0.92 for 3MM. (B) At 41 picomolar of oligonucleotide spike-in, an R^2 value of 0.96 was obtained for probes with a PM, 0.87 for 1 MM, 0.94 for 2MM and 0.86 for 3MM. (C) At 121 picomolar of oligonucleotide spike-in, an R^2 value of 0.92 was obtained for probes with a PM (perfect match), 0.85 for 1 MM, 0.90 for 2MM and 0.83 for 3MM. (D) At 364 picomolar of oligonucleotide spike-in, an R^2 value of 0.84 was obtained for probes with a PM (perfect match), 0.81 for 1 MM, 0.90 for 2MM and 0.75 for 3MM. Blast searches were done for all 70 mer probe combinations to the human genome sequence. The 2 MM 70-mer oligonucleotide probes were highly similar to the human genome and hence are not considered informative and do not show any variation as represented by the linear regression value.

Figure S1A

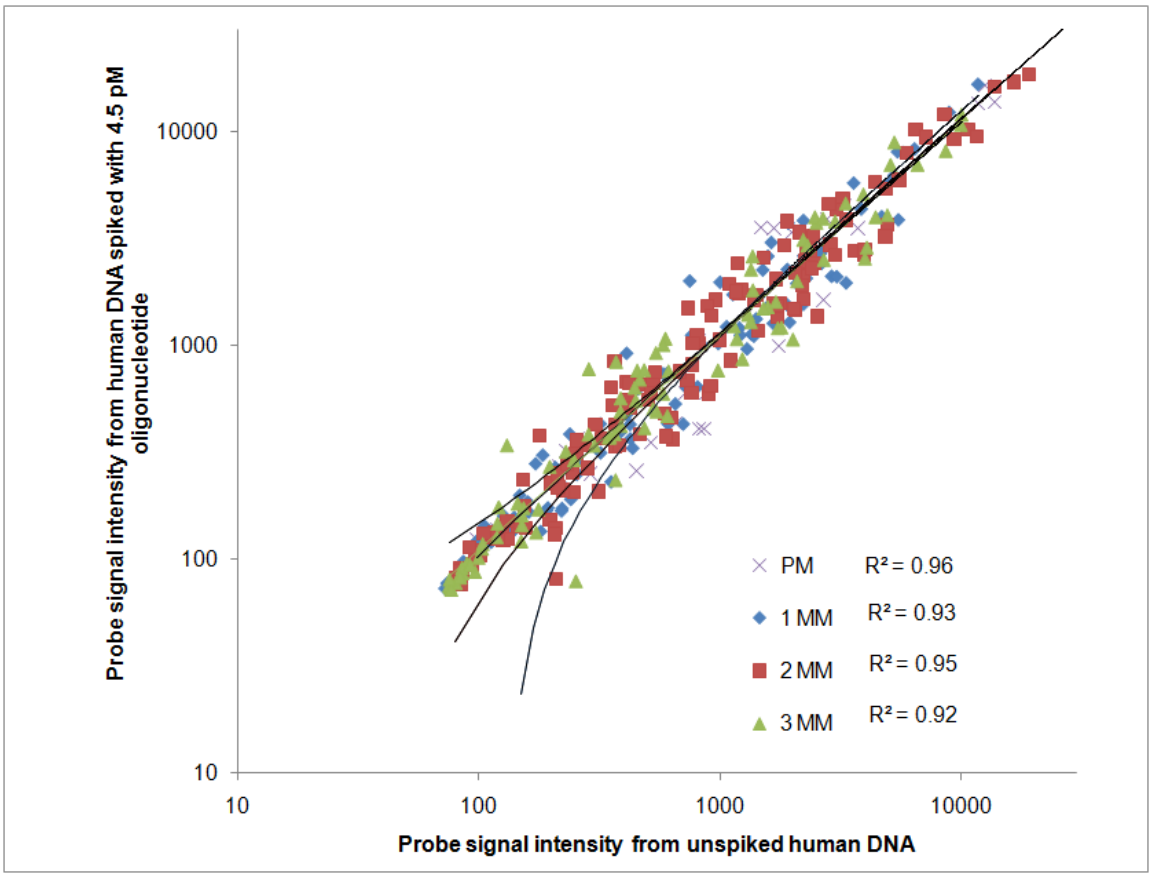

Probe signal intensity from human DNA spiked with 4.5 pM oligonucleotide
10000
1000
100
10
10
100
1000
10000
Probe signal intensity from unspiked human DNA
PM
R² = 0.96
1 MM
R² = 0.93
2 MM
R² = 0.95
3 MM
R² = 0.92

Figure S1B

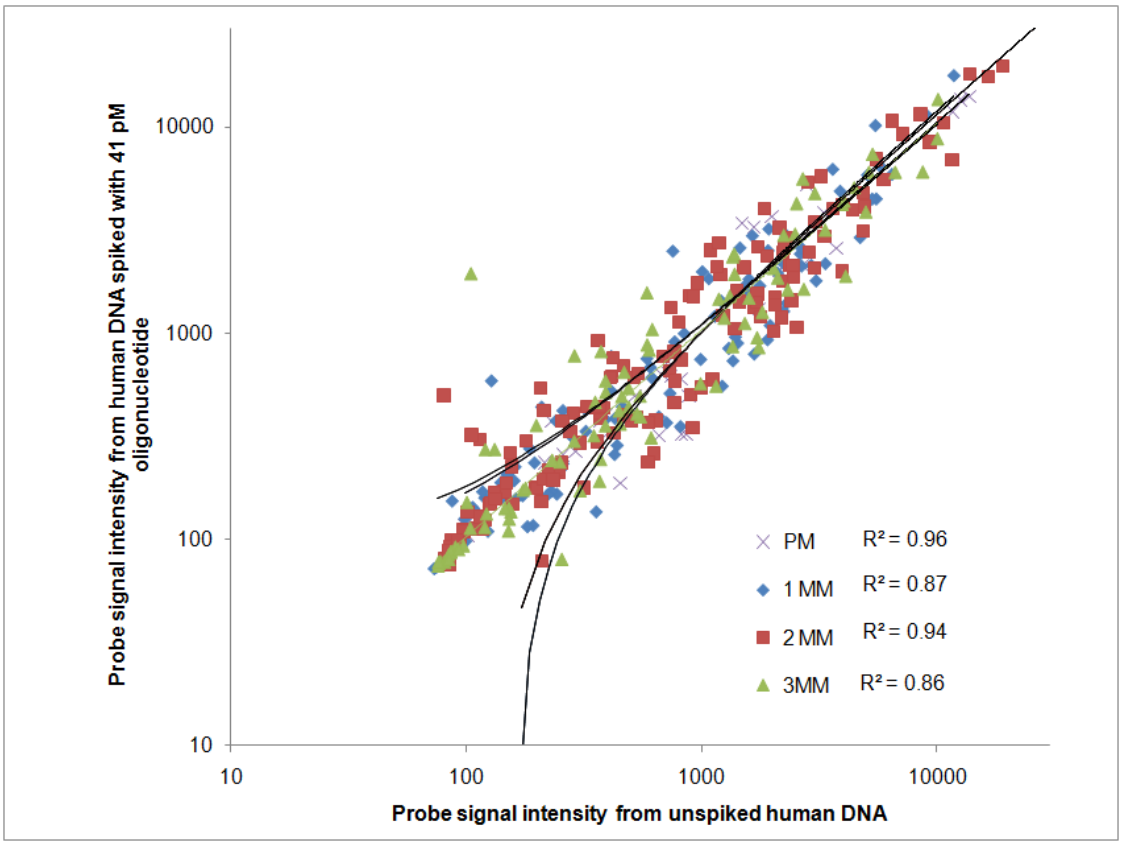

Probe signal intensity from human DNA spiked with 41 pM oligonucleotide
10000
1000
100
10
10
100
1000
10000
Probe signal intensity from unspiked human DNA
PM R² = 0.96
1 MM R² = 0.87
2 MM R² = 0.94
3MM R² = 0.86

Figure S1C

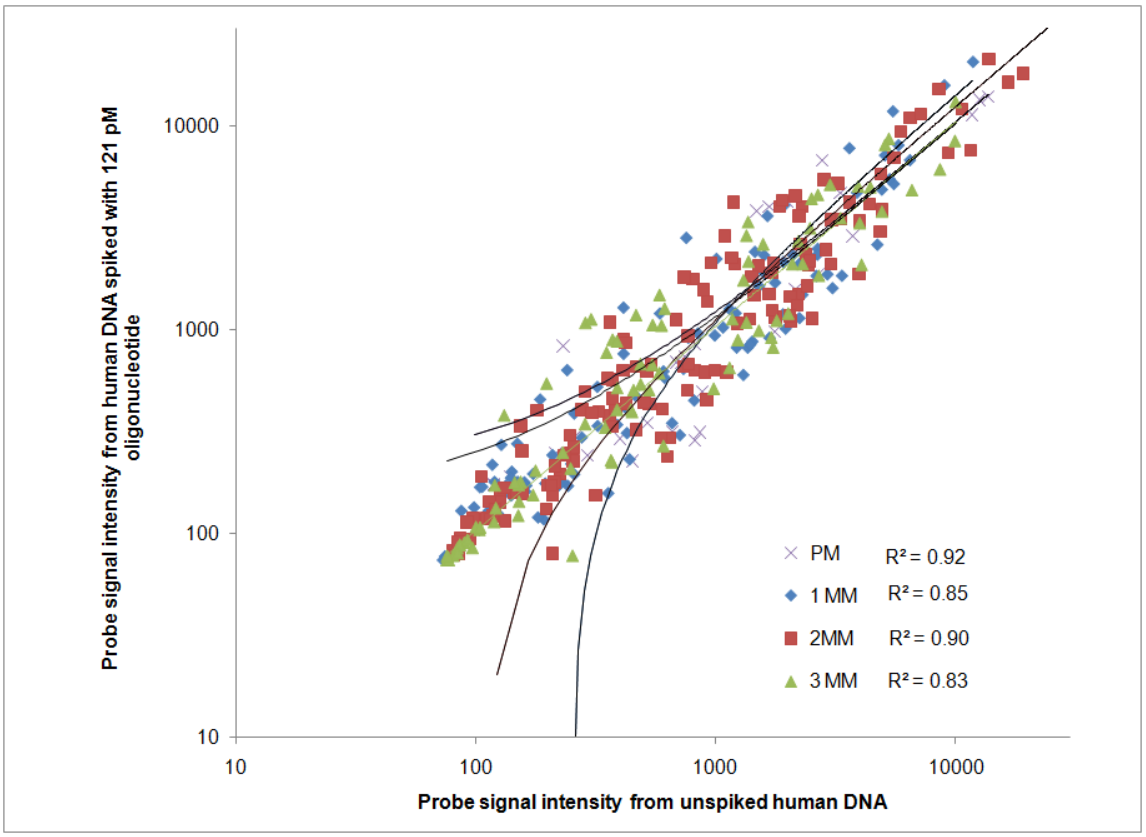
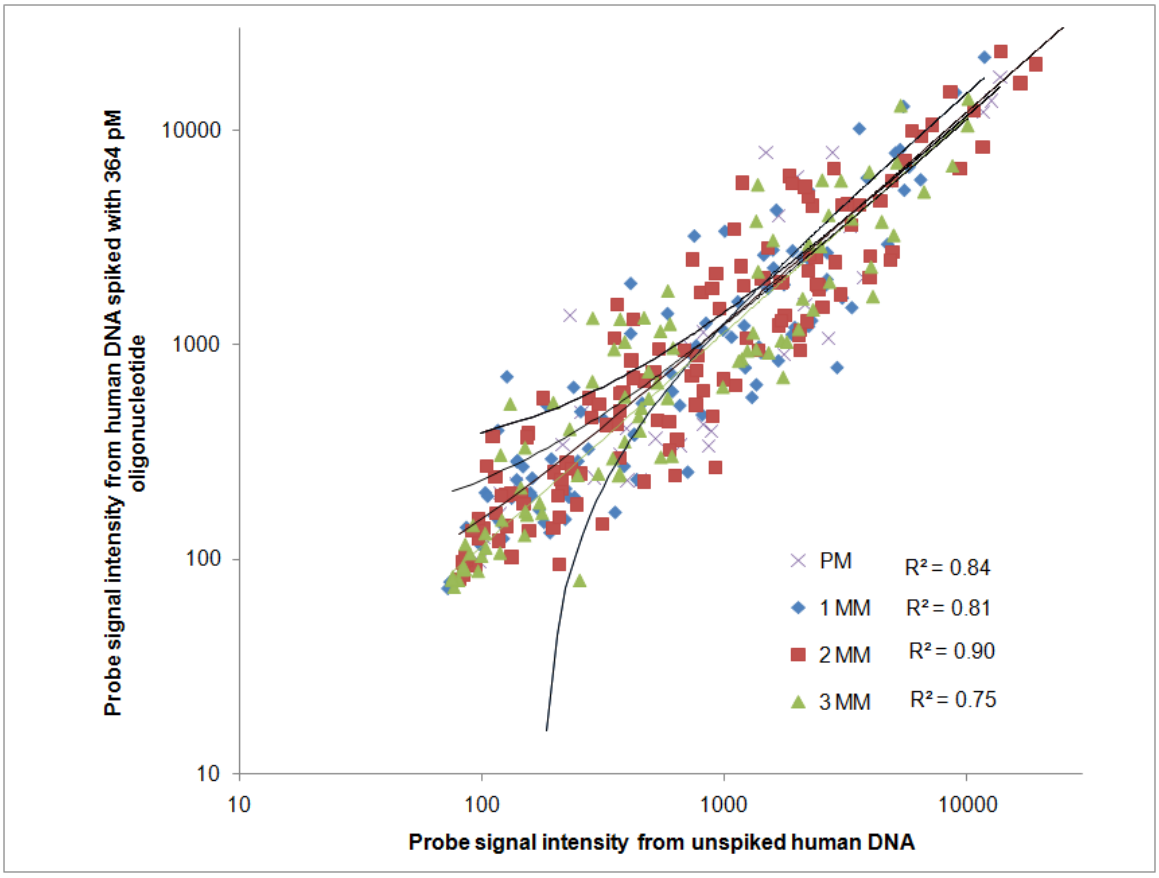

Probe signal intensity from human DNA spiked with 121 pM oligonucleotide
10000
1000
100
10
10
100
1000
10000
Probe signal intensity from unspiked human DNA
PM R² = 0.92
1 MM R² = 0.85
2MM R² = 0.90
3 MM R² = 0.83

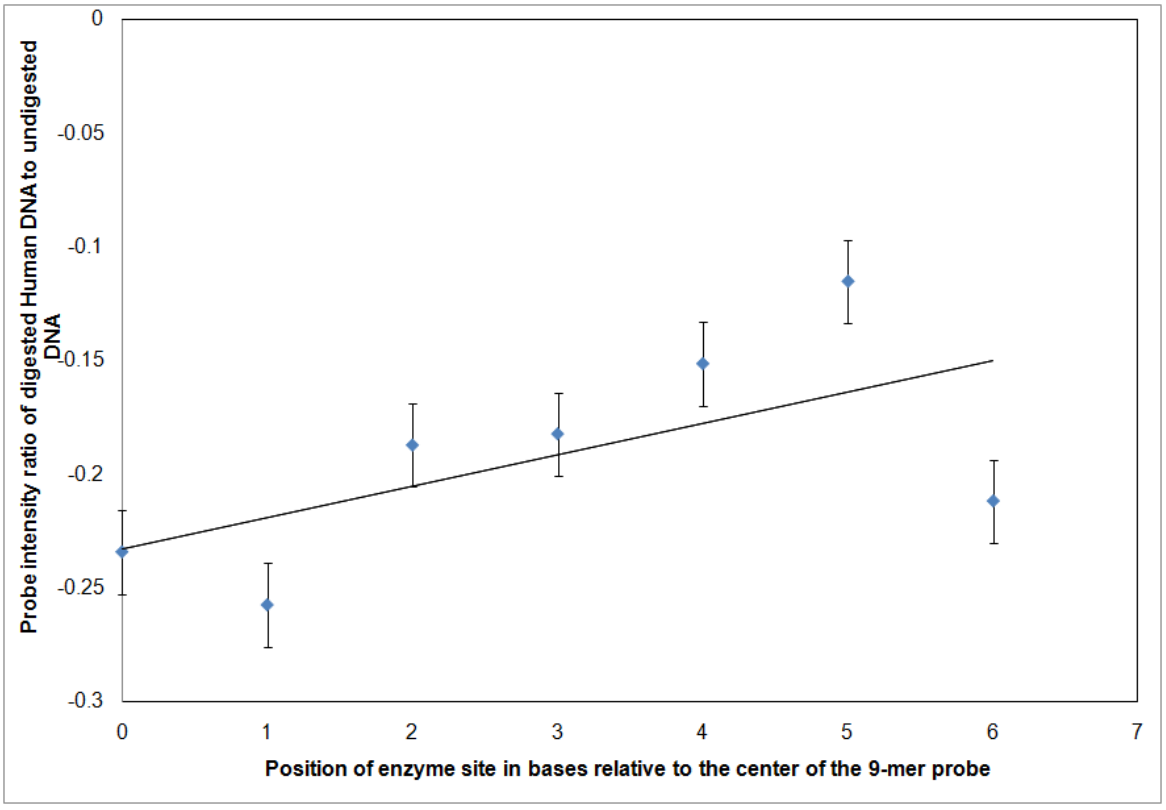
Figure S1D



2.11.4 Figure S2 Analysis of probe hybridization specificity on the UBDA array.

Human genomic DNA was digested with StuI (AGG^CCT) restriction enzyme, and then compared to undigested human genomic DNA from the same individual. The resulting values were plotted, with ratio of the human genomic DNA digested with StuI and undigested human genomic DNA as $\log_2$ fold change on the ordinate axis. The nucleotide position of the StuI restriction enzyme site relative to the center of the 9-mer probe is plotted on

the abscissa axis.Probe specificity analysis of individual 9-mer probes is confirmed by demonstrating that the center most base governs the hybridization kinetics. This is shown by a reduction in probe signal intensity values when the human genomic DNA sample was digested with StuI enzyme. The reduction in the probe intensity signal is greater when the restriction enzyme site is located at the center of the 9-mer probe. Therefore the center nucleotide of the probe is the most restrictive in determining the specificity of the probe hybridization complex.



2.11.5 Table S3 Genomes hybridized on the array

Genomic DNA from the following genomes was hybridized on the UBDA array.

Eukaryotes	**Prokaryotes**	**Viruses**
Homo sapiens (Human)	*Lactobacillus plantarum*	Influenza A 49H10N7
Bos taurus (Bull)	*Escherichia coli* K12	Influenza A 76H1N1
Gallus gallus (Chicken)	*Brucella abortus* RB51	
Meleagris gallopavo (Turkey)	*Brucella abortus* 12	
Ovis aries (Sheep)	*Brucella abortus* 86859	
Capra hircus (Goat)	*Brucella suis* 1330	
Equus caballus (Horse)	*Brucella melitensis* 16M	
Cryptosporidium parvum		

2.11.6 Annotation file for 9-mer probes on the UBDA array (available at http://innovation.vbi.vt.edu)

2.11.7 Annotation file for all other probes on the UBDA array (available at http://innovation.vbi.vt.edu)

Genomic DNA from the following genomes was hybridized on the UBDA array.

Chapter 3

Comparison of genome diversity of *Brucella* spp. field isolates using Universal Bio-signature Detection Array and whole genome sequencing reveals limitations of current diagnostic methods

Shamira J. Shallom[1§], HongSeok Tae[1§], Luciana Sarmento[2], Dale Preston[3], Lauren McIver[1], Christopher Franck[1,4], Allan Dickerman[1], L. Garry Adams[2] and Harold R. Garner[1*]

[1] Virginia Bioinformatics Institute, Virginia Tech, Blacksburg, VA, USA

[2] Department of Veterinary Pathobiology, College of Veterinary Medicine, Texas A&M University, College Station, TX, USA

[3] Texas Animal Health Commission, State-Federal Diagnostic Laboratory, Austin, TX, USA

[4] Laboratory for Interdisciplinary Statistical Analysis (LISA), Department of Statistics of Virginia Tech, Blacksburg, VA, USA

[§] The first two authors contributed equally to this work

***Corresponding Author:**

Harold R. Garner

Executive Director, Virginia Bioinformatics Institute

Virginia Tech

Washington Street, MC0477

Blacksburg, VA 24061-0477, USA

E-mail: garner@vbi.vt.edu, phone: 540.231.2582, fax: 540.231.2606

3.1 Abstract

Diverse analysis methods are used to identify pathogens, specifically *Brucella* species or biovars. Diagnostic approaches included serology and biochemical tests, PCR assays, microarray analyses using a Universal Bio-signature Detection Array (UBDA) and whole genome 'deep' sequencing techniques for *Brucella* organisms including a number of field isolates. Although there was frequent agreement among different tests, some gave compound/contradictory results that were a consequence of species diversity as measured by UBDA and validated from whole genome sequence. The field isolates have clearly diverged from known *Brucella* reference genomes, such that they confound identification using serological and biochemical tests. This could imply that the *Brucella* isolates were from mixed or dual infections of both *Brucella abortus* and *Brucella suis* in the same animal, or infected by a chimera or variant of *Brucella suis* in cattle that evoked a false positive serological test. UBDA is sensitive in tracing genomic differences among the isolates.

Keywords: Genomics, *B. suis, B. abortus*, Bovine, Porcine, Diagnostics

3.2 Introduction

Brucellosis is an anthropo-zoonotic disease caused by small intracellular facultative, Gram negative cocco-bacilli belonging to the genus *Brucella*. Brucellosis is considered the world's most widespread zoonotic infection causing abortion, fetal death, and genital infections in animals [1]. In humans, this highly diverse illness initially presents as fever, malaise and myalgia and has the potential to develop into a chronic illness affecting various organs and tissues. The genus is further classified into nine species of *Brucella* based on host preference and phenotype. These are *B. abortus* (cattle), *B. canis* (dogs), *B. melitensis* (sheep and goat), *B. neotomae* (desert wood rats), *B. ovis* (sheep) and *B. suis* (pigs) [2], *B. microti* (voles) [3] and the recently described marine mammals infecting species *B. ceti* and *B. pinnipedialis.* Currently it takes approximately two weeks from collection of a clinical specimen to definitive identification of *Brucella*. Due to the zoonotic nature of most *Brucella*, the tests are complex and are a bio-hazard for laboratory personnel who handle this agent. Once an infected herd is identified, the infection is contained by quarantine and eliminating all infected and exposed animals until the disease is eradicated. In addition, federal and state animal health officials check neighboring herds and vendors that may have received animals from the infected herd

(www.aphis.usda.gov, *Facts about Brucellosis*, August 4, 2011). Diagnostic tests are used to identify infected cattle, however, there is no single test for the detection of *Brucella*. Instead, there are a series of tests comprised of growth characteristics, serology and bacteriological methods for classification of the species and then biovar subcategories for each [4]. The traditional test used in the field is the brucellosis milk ring test and/or the Rose Bengal test (RBT). In addition, buffered plate agglutination, complement fixation test (CFT), and indirect and competitive enzyme-linked immunoblotting (iELISA and cELISA) are widely used to detect serum antibodies for bovine brucellosis diagnosis. After a culture has been identified as a member of the genus *Brucella*, the species and biovar are established. For *B. melitensis*, *B. abortus* and *B. suis*, the identification is performed based on four tests: carbon dioxide (CO_2) requirement, production of hydrogen sulfide (H_2S), and dye (thionin and basic fuschin) sensitivity.

Most standard serological tests such as serum agglutination, complement fixation and enzyme-linked immuno-sorbent assays (ELISAs) use whole cell preparations [5], cell sonic extracts or lipopolysaccharide (LPS) fractions [6]. A strong antibody response is obtained with LPS, however, the immuno-dominant epitope of the *Brucella* O-polysaccharide is similar to that of

various bacteria such as *Yersinia enterocolitica* O9, *Salmonella* Urbana group N, *Vibrio cholera*, *Francisella tularensis* and *Escherichia coli* O:157 which results in cross-reactivity [7-9]. Further, there are separate tests for infected cattle in vaccinated herds that require additional testing with a combination of Rivanol and complement fixation test along with bacteriological examination of milk samples [4, 10]. In international eradication programs, it is time consuming to differentiate vaccinated from infected animals [11] hence, differential serological diagnosis in brucellosis remains a challenge.

Several PCR-based assays for typing of *Brucella* species biovars have been developed. These are mainly genus specific PCR assays targeted to genes such as the *BCP31*, *omp2A*, *omp2B* and the 16S and 23S rRNA operon genes [12]. Other PCR tests have also been developed for typing the *Brucella* species biovars, such as analysis of the *IS711* repetitive element [13]. Genotyping using variable number tandem repeats (VNTRs) has been used previously to provide a *Brucella* species level resolution [14, 15]. However, many of these loci result in homoplasy where the same variability is observed in different branches of the phylogenomic tree [16]. Single nucleotide polymorphism studies using real time PCR assays have been developed in specific house-keeping genes in the *Brucella* clade [17] .

In this study, we were approached by the Texas Animal Health Commission (TAHC), Austin, Texas, because they had a number isolates (n = 36) obtained from milk and tissues of cattle, horses and pigs suspected to have brucellosis. These samples had variable *Brucella* serology profiles and biochemical phenotypes. We sought to determine if these isolates originated from mixed samples or were possibly intermediate (chimeric) biovars indicating an evolving *B. suis* host preference using data from several technologies: serology tests, PCR amplification of the *IS711* element from *Brucella* species, UBDA technology and whole genome sequencing. The UBDA platform for detection and differentiation is an oligonucleotide array that contains all possible (4^9 combinations) 9-mer probes, and is therefore genome independent. [18]. This technology facilitates classification of new pathogens and profiling of the 36 field isolates in relation to a library of known standards acquired using the same array design. Finally, this study describes the analysis results of whole genome 'deep sequencing' for further validation of the identity of these field isolates.

3.3 Results

Samples cultured from milk and/or tissues of cattle, horses and pigs suspected to having brucellosis, were selected for analysis using the UBDA array and Illumina-based sequencing because of abnormal serological and

biochemical tests. The results generated from biochemical typing, PCR, UBDA PCA analysis and whole genome analysis have been summarized in Supplementary table 1A and 1B.The original biochemical tests of 36 field isolates indicated that 19 isolates were *B. abortus* and 17 isolates were *B. suis.* We sequenced nine TAHC field samples and an aliquot of the sample on which original *B. suis*1330 reference sequence was based using Illumina paired-end sequencing to validate the UBDA findings and to explain the anomalous PCR and other test results (Supplementary Table 2).

3.3.1 PCR assay on the *IS711* Element of *Brucella*

The PCR assay identifies polymorphisms arising from species specific localization of the genetic element *IS711* in the *Brucella* chromosomes. *IS711* is a repetitive element unique to *Brucella* species and for most species at least one copy of the element occurs at a unique species or biovar specific chromosomal locus. The *IS711* element is 281 base pairs with additional nucleotides flanking the 3' end of the element that are species specific. Using accepted primer sets, band of 498 base pairs was expected for *B. abortus* and 285 base pairs for *B. suis* [13]. The BLAST alignment of *IS711* element *Brucella* primers to five known *Brucella* genomes is shown using the Mauve alignment (Figure 1) [19] to have complementary PCR primer sequences for bio-assays of both *B. abortus* and *B. suis* species.

From serology and biochemical tests of the 36 field isolates, 19 isolates(1, 2, 3, 5, 6, 7, 11, 12, 14, 15, 18, 19, 20, 21, 23, 24, 25, 32, 33) and 17 isolates (4, 8, 9, 10, 13, 16, 17, 22, 26, 29, 30, 31, 34, 35, 36, 37, 40)were determined to be *B. abortus* and *B. suis* respectively. PCR assays targeting the traditionally used *IS711* element in *Brucella* showed ambiguity in the size of the PCR products obtained from these field isolates. All 19 isolates identified as *B. abortus* by the serology test produced the expected 498 base pair *B. abortusIS711* element PCR products (Supplementary Figure 1A through 1E). They were also positive for *B. suis* specific PCR primers but the sizes of products were approximately 900 bases and their sequences did not align to the *B. suis* 1330 genome. Hence, the products from the *B. abortus* isolates for the *B. suis* specific PCR test were determined to be non-specific and spurious. The only exception, isolate 18 (*B. abortus* strain 19), produced the expected sizes of products from both the *B. abortus* specific PCR primer set and *B. suis* primer set (Supplementary Figure 1B) and was suspected to contain both *B. abortus* and *B. suis* infections.

Most isolates which were identified as *B. suis* from the serology test, produced *B. abortus* specific PCR products (498 bases) as well as the *B. suis* specific products of 285 base pairs (Supplementary Figure 1A through 1E). The PCR products from these isolates were sequenced and compared to *B.*

abortus and *B. suis* reference genomes. The sequences mapped to the expected positions of PCR products in their specific genomes, which suggested that they could be mixed isolates of both *B. abortus* and *B. suis*. However, isolate 4, 36 and 37 which were also determined to be *B. suis* by the biochemical test showed unexpected results. While isolate 4 produced an expected size of the product from the *B. abortus* specific PCR test and an abnormal 900 bases non-specific product from the *B. suis* specific PCR test (Supplementary Figure 1A), Isolate 36 and 37 produced only *B. abortus* specific PCR products (Supplementary Figure 1E. These 36 field isolates did not produce an *IS711* product with the *B. melitensis* 16M primers.

PCR reactions for the control genomic DNA from *B. suis* 1330 were positive for the *B. suis* specific primer set and negative for *B. abortus* or *B. melitensis* specific primer set, as expected. *B. abortus* 2308 had a product specific to *B. abortus*. The *B. abortus RB*51 had the expected size of the PCR product for the *B. abortus* primer set but different size fragments including approximately 900, 1200 and 1500 bases for the *B. suis* primer set. *B. melitensis* 16M primers also produced the expected 700 base product for the *B. melitensis* primer set (Supplementary Figure 2).

3.3.2 Principal Component Analysis of UBDA array probe signal intensity values

Genomic DNA from each of the samples was hybridized on the UBDA array which contains probes to all possible 9-mers. To determine the closest similarity of the field isolates to the reference samples *B. abortus* and *B. suis* genomes, signal intensity values generated from UBDA array probes for each of the field isolates and reference samples (*B. sui*s 1330 and *B. abortus* 2308) were evaluated using Principal Component Analysis (PCA) (Supplementary Table 1B). PCA method provides a quantitative measure and helps assign the sample to one of more groups such as a pure isolate of a single species or composite mixture of multiple species. Samples from bovine milk or tissue determined to be *B. suis* from biochemical typing (4, 10, 13, 16, 17, 22, 26, 29, 34, 35, 36 and 37) were found to be similar to *B. suis* or a composite mixture of *B. suis* and *B. abortus* using PCA. The other *B. suis* typed isolates (8, 30, 31 and 40) were determined to be most similar to the *B. abortus* reference. These results would indicate that these isolates may have a higher proportion of *B. abo*rtus genomic DNA in the sample. Based on the PCA analysis and a discriminatory value of 0.6% between the genome sizes of *B. abortus* and *B. suis* (Supplementary Table 5), we determined that they were not pure *B. suis* isolates and they maybe the result of mixed or dual infections.

Further, the *B. abortus* biochemically typed isolates (1, 2, 7, 11, 15, 19, 20, 21, 23 and 25) were found to be similar to the *B. abortus* bio-signature. Isolate 18, which was biochemically identified as *B. abortus*, is predominantly composed *B. suis* based on PCA analysis of UBDA array data. Isolates (3, 5, 6, 13, 14, 32 and 33) were found to be more similar to *B. suis* from PCA analysis. This may be attributed to the limits of detection of the UBDA array and PCA-based analysis. The detection limit for de-convoluting the identity of these highly similar *Brucella* species is in the range of ~ 0.6% based on sequence similarity (Supplementary Table 2).

3.3.3 Comparison of species independent 9-mer probe signal intensity values from the UBDA of known *Brucella* species and field samples from the Texas Animal Health Commission (TAHC) using phylogenomic analysis

To understand and visualize the relative similarity among all samples, the hybridization signal intensities were then evaluated using phylogenomic analysis tools used to describe the evolutionary relationships among sequences [20]. This involves cluster analysis, an unsupervised learning technique used to organize samples into groups such that samples within a cluster are more similar to each other than to samples in other clusters. Un-

biased cluster analysis of hybridization patterns can easily distinguish species into accepted phylogenic relationships [21].

Our group has previously demonstrated a low resolution array design that focused on a subset of species independent random probes [22] and cluster analysis to distinguish species. In this study, we use signal intensities generated from 9-mer probe (262,144) data to create a neighbor-joining phylogenomic relationship tree of the *Brucella* field isolates. The tree was rooted to *F. tularensis* LVS UBDA data and the order of the samples were randomized to create the phylogeny relationship tree (Figure 2). A hierarchical clustering algorithm analysis of signal intensities from the UBDA array was used to establish nearest neighbor relationships between these isolates. A similar phylogenomic relationship tree (Supplementary Figure 3) was obtained when signal intensities from each sample were RMA normalized as a batch process using Nimblescan [23].

From the 9-mer phylogenomic tree, isolate33 appears to be a highly chimeric or contaminated isolate and appears as an out group on the phylogenomic tree. The Pearson's correlation distance to *E. coli* and *F. tularensis* LVS strain was $\rho = 0.936$ and $\rho = 0.915$, respectively. Whole genome sequencing of this isolate revealed that it had the highest proportion of unmapped reads (46,344 reads) with 25.2% of these reads being

completely novel with no alignment to the NT database (Supplementary Table 3).

B. suis isolates 29 and 34 clustered close to each other (Figure 2) and were originally obtained from bovine tissue derived from the same herd. Isolate 29 and 34 are highly similar with a Pearson's distance of $\rho = 0.997$. *B. suis* isolates 30 and 31 were also isolated from the same herd and are closely related with a Pearson's distance of $\rho = 0.993$. Although *B. suis* isolates 9, 10 and 17 were isolated from porcine sources, they are different in their UBDA signal intensity pattern from the *B. suis* 1330 isolate and are separated by Pearson correlation values of 0.926, 0.867 and 0.891, respectively. Isolate 9 appears to be the nearest neighbor to *B. suis* 1330 (Figure 2). Isolate 13 was the only isolate obtained from equine tissue and was found to be at a distance of $\rho = 0.864$ from *B. suis* 1330. Sequencing of this isolate revealed a large proportion (49.4%) of unmapped reads (66,035 reads) having no similarity (BLAST output < 1e-15) to sequences in the NT database (Supplementary Table 3). Additional analysis of the unmapped reads from contaminating organisms ("other" column in Supplementary Table 3) is described in Supplementary Table 4.

Two isolates, 36 and 37, did not cluster with any of the *B. suis* or *B. abortus* isolates, although they were identified as *B. suis* in biochemical

tests. On the phylogenomic tree generated from UBDA data, they separated by a distance of $\rho = 0.913$ (Figure 2). These isolates showed a band only with the *IS711* element PCR with *B. abortus* primers while no product was detected with the *B. suis* primers (Supplementary Figure 1E). This data again indicates significant discrepancy among the various analyses.

The reference genome samples such as *B. suis* 1330, *B. melitensis* 16M, *B. abortus* 2308 and *B. abortus* RB51, are in a subgroup away from the field isolates. We determined that unmapped reads from *B. suis* 1330 reference strain had a high proportion of reads (32,489) that mapped to the human genome (Supplementary Table 3). Hence the *B. suis* 1330 DNA sample was found to contain a high proportion of human genomic DNA sequences. Since the probes on the array are not species-specific, the array gives a spectrum of signal intensities that are derived from the genomic contents of a sample. We did not sequence genomic DNA from the other known *Brucella* such as *B. abortus* RB51, *B. melitensis* 16M and *B. abortus* 2308.

In addition, to assess the reproducibility of the array since independent samples were hybridized in the two channels on the array, a single sample of *B. abortus* RB51 was hybridized in both channels which showed an R^2 value of 0.99, demonstrating that there was no dye bias in the array hybridization.

3.3.4 Experimental confirmation of UBDA findings using next generation sequencing methodology

We used whole genome sequence data to determine whether the *B. suis* isolates were either truly mixed, or chimeric intermediate genotypes. We sequenced nine TAHC field samples and an aliquot of the original *B. suis*1330 reference DNA sample using Illumina paired-end sequencing (76 cycles for isolate 2, 3, 17, 22, 29, 34 and 35, and 101 cycles for isolate 13, 33 and *B. suis* 1330) to validate the UBDA findings and to attempt to explain the anomalous PCR and other test results. We obtained 42,000,000 ~ 49,000,000 reads per sample, and used BWA [24] to map the reads to the genomic sequences of the five completely sequenced *Brucella* species as references to measure divergence from those genomes.

The first three samples, *B. abortus* isolates (2, 3 and 33) had 99.9% similarity to *B. abortus* biovar 1 9-941, while *B. suis* isolates (13, 17, 22, 29, 34 and 35) had 99.9% base level identities to *B. suis* 1330 genomes (Supplementary Table 2). From the PCR assays using the *IS711* element and the UBDA analysis, six isolates (13, 17, 22, 29, 34 and 35) were suspected to be either mixtures of *B. abortus* and *B. suis,* or variants of *B. suis* producing PCR products for both *B. abortus* and *B. suis* primer sets. To address this question, we compared the average sequence coverage

between261,000 and 280,000 bases of *B. abortus* 9-941 genome, which is a long deletion in *B. suis* 1330 genome (Supplementary Table 5). The average coverage of isolates 13, 17, 22, 29, 34 and 35 on the region varied from 0.07 to 0.6x, which was significantly different from the average coverage on the *B. abortus* or *B. suis* whole genome (1,000x ~ 1,700x, Supplementary Table 2). The region also exists in several other *Brucella* species including *B. abortus* S19, *B. abortus* 2308, *B. melitensis* 16M, *B. ovis* ATCC 25840 and *B. suis* ATCC 23445. The results indicated that the samples were mixtures of different *Brucella* species and the mixture ratios were between 13,000:1 ~ 1,700:1, *B. suis* and other *Brucella*. This data indicate that minimum specificity limit for resolving these two very close genomes is a 19 kb region ((280,000-261,000)/3,300,000 = 0.006 = 0.6%) (Supplementary Table 5). In order to distinguish a closely related contaminant using the UBDA array, there must be more than one copy of the minor species in 1,700 copies of the major species. Hence the detection limit on the UBDA array was set at 0.6% for Principal Component Analysis of these highly similar *Brucella* species. In addition, the percentage identity of these 3 isolates to *B. suis* is ~ 99.4% compared to *B. abortus* biovar 1 9-941 at ~ 99.9%. The other sequenced isolates (13, 17, 22, 29, 34 and 35) had only about 80 variant loci to *B. suis* 1330 and were ~99.9% identical compared to ~99.2% identity to

B. abortus biovar 1 9-941 (Supplementary Table 2). The numbers of loci containing sequence differences for each sample against the *B. suis* 1330 genome were compared. Isolates 2, 3 and 33 were determined to be *B. abortus* or a close variant, although they had more than 7,900 loci containing sequence differences with respect to *B. suis* 1330. Even though the numbers of variants in the six isolates were almost the same, the comparison of common variants showed differences between the samples (Table 1). Each cell in table 1 represents the number of common variants between two samples when compared to the *B. suis* 1330 genome[25].

3.3.5 Phylogenomic tree built using amino acid sequence as translated from sequenced genomes of selected field isolates

A phylogenomic tree from publicly available sequenced *Brucella* genomes and Texas Animal Health Field *Brucella* field isolates was generated to visualize differences among isolates at the gene level. The results of a maximum likelihood phylogenomic analysis using translated amino acid sequences of the 13 *Brucella* genomes from the NCBI database and nine newly sequenced *Brucella* field isolates (2, 3, 13, 17, 22, 29, 33, 34 and 35) is shown in Figure 3. Representatives of the major groups of *Brucella* including multiple strains of *B. suis*, *B. abortus*, and *B. melitensis* and *B. ovis*, *B. ceti*, *B. pinnipedialis*, and *B. microti* were included in the tree

for context. The tree was rooted at *B. microti* based on prior analyses using more distant *Brucella* strains. The maximum likelihood tree was created based on 1,006 presumptively vertically inherited genes. The new genomes were localized to two locations on the tree: 6 nested nearest *B. suis* and 3 nested nearest *B. abortus*. The new *B. suis* genomes (isolates 13, 17, 22, 29, 34 and 35) are highly similar with per-site nucleotide differences on the order of 10^{-5} between pairs, and this cluster was supported by 100% of bootstraps while the distances to *B. suis 1330* or *B. suis* S2 are 3 to 7 x 10^{-4} mutations per site. The three new genomes, in the *B. abortus* clade are more divergent. The patristic distance between any of the *B. abortus* field isolates 2, 3, or 33 to *B. abortus* S19 are indicated to be about 10^{-3} mutations per site.

Note, that the phylogenomic tree generated using gene level sequence information emphasizes analysis of single copy genes of a highly curated data set (Figure 3). It provides a comparison of the open-reading frames (ORFs) of the predominant genome in a given sample and unmapped reads which have been derived from contaminating organisms in a given sample, are not considered. However, the phylogenomic analysis using the signal intensities from genome independent probes on the UBDA array (Figure 2) captured the mixed or chimeric genomes present in the *Brucella* field isolates.

3.4 Discussion

The detection and identification of bio-threat agents requires a high-resolution detection platform capable of discerning closely related species from a given organism. Using one "gold standard" test, the PCR test, we observed spurious PCR products from several samples due to non-specific binding of PCR primers under standard conditions. Since the reverse primers for both *B. abortus* and *B. suis* are common and target *IS711* transposon elements which are duplicated at the several loci, the primer can bind to multiple regions in the genomes and thus the specificity contribution by this primer is diminished. The *IS711* transposon element is known to continuously copy and insert itself in new positions in a genome which may generate non-specific PCR products [26]. PCR-based assays for typing of *Brucella* species biovar are ideal for high quality pure samples. However, since several field isolates had positive results for both *B. abortus* and *B. suis* in the PCR assay using *IS711* elements, which may be explained if the isolates are mixed, chimeric assemblies where *B. suis* has now merged with *B. abortus* genomic components, or a new variant of *B. suis* that has now moved into cattle thus evoking a false positive serological or biochemical test.

The UBDA method was used to establish the identity of the species diversity and phylogenomic relationships between field isolates, and was shown to be sensitive to species variants of the type seen here. We demonstrate the use of signal intensities from UBDA to generate a principal component analysis and assign a given sample to one of more groups. Principal component analysis and Euclidean distance mapping to reference *B. abortus* 2308 and *B. suis* 1330 genomes provides a quantitative approximation to the composite species identity of the field isolate. Samples from bovine milk or tissue determined to be *B. suis* in biochemical or serological tests were found to be a mixed composite of *Brucella* species. Hence, we determined that they were not pure *B. suis* isolates and presumably are the result of mixed or dual infections. The specificity limit for de-convoluting the identity of these highly similar species is in the range of ~ 0.6% sequence similarity of these field isolates to known reference genomes such as *B. abortus* biovar 1 9-941 and *B. suis* 1330 (Supplementary Table 2). Phylogenomic analysis using a nearest-neighbor joining algorithm for the signal intensities from UBDA revealed that standardized biochemically phenotyped *B. suis* isolates failed to cluster with known *B. suis* 1330, and instead clustered as mixed samples or an unknown intermediate species. Inspection of the dendrogram of nearest neighbors

between *Brucella* field isolates derived from the species-independent UBDA method shows that the field isolates have diverged such that they confound the identity given to the *Brucella* field isolates from serological and biochemical tests. To validate this, nine field isolates were sequenced and their sequencing reads were mapped to the *B. suis* 1330 and *B. abortus* 9-941 genome sequences. By comparing the ratio of sequencing reads mapped to *B. suis* 1330 and *B. abortus* 9-941 genome specific regions, it was confirmed that *B. suis* and *B. abortus* coexist in isolates 13, 17, 22, 29, 34 and 35 of which the predominant genome was *B. suis*.

In this study, the standard diagnostics including serology tests and PCR assays used to determine species of field isolates were found to have limitations when those isolates are potentially complex, leading to the identification of only predominant or targeted organisms or false conclusions as a consequence of species genomic evolution or minority contaminants. Deploying other more comprehensive techniques, including species-independent microarrays with superior speed and cost benefit and whole genome sequencing, which has superior comprehensive analysis, can lead to a greater confidence in the final interpretation. These techniques can also be revealing of an underlying reason for the success or failure of other

analytical technique through providing data which more completely describes the genomic complexity of real-world field samples.

3.5 Materials and Methods

3.5.1 Bacterial Isolates: Bacteriologic, serology and biochemical procedures

Culture and identification of *Brucella* spp., which had been previously described by Alton[4], was performed from 36 milk and tissue samples of bovine, porcine and equine animals at the Texas Animal Health Commission (Austin, TX). The card test was performed by standard procedures used by the Department of Agriculture. Samples identified positive with the card test were plated onto Farrell's media and selective *Brucella* media with ethyl violet and incubated at 37ºC under aerobic conditions in the presence of 5-10% CO_2 for five to seven days. The bacterial colonies demonstrating a gross morphology typical for smooth colonies of *Brucella* spp. were screened for catalase, oxidase and urease activity. Species and biovar identification were performed according to CO_2 requirement, production of H_2S, growth in the presence of basic fuchsin, thionin and slide agglutination test with monospecific anti-A and anti-M antigenic determinants of *Brucella* LPS (Lipolysaccharide) sera. Of the 36 TAHC samples, 19 were identified as *B. abortus* and 17 were identified as *B. suis* from these tests.

3.5.2 Genomic DNA sample preparation

Genomic DNAs of 36 samples were prepared from milk or tissue samples of cattle, horses and pigs suspected to have brucellosisbased on serology and biochemical diagnostic tests. The organisms were methanol inactivated at the Texas Animal Health Commission (Austin, TX) and genomic DNA was extracted at College of Veterinary Medicine, Texas A&M University (College Station, TX) as follows. Pellets of methanol inactivated cells were washed with 25 ml of J-buffer (0.1 M Tris pH 8.0; 0.1 M EDTA; 0.15 M NaCl) and then lysed in 1 ml of J-buffer containing 10% lysozyme solution (10 mg/ml in 0.25 M Tris, pH 8.0). After 10 min of incubation, DNA was released from the cells by sodium N-lauryl sarcosine (Sigma, St. Louis, MO) treatment followed by degradation of RNA by DNase-free RNase (Roche Applied Science, Indianapolis, IN) treatment and digestion of proteins with proteinase K (Roche Applied Science). The resulting solution was transferred to a dialysis bag and dialyzed against TE (10 mM Tris, pH 8.0 and 1 mM EDTA) overnight at 37°C. DNA was subsequently extracted twice using neutral water-saturated phenol (Ambion, Austin, TX) first and then ether (Sigma-Aldrich, St. Louis, MO) before dialyzing overnight against TE. DNA concentration was quantified by NanoDrop ND-1000 (Thermo Scientific, Wilmington, DE) and stored at 4°C

until used [27]. These DNA samples were then received at VBI, Virginia Tech (Blacksburg, VA). Of the 36 TAHC genomic DNAs, 19 samples had low DNA concentration and were whole genome amplified using 10 ng of starting material as specified by the manufacturer (GenomiPhi V2, GE Healthcare, Piscataway, NJ), resulting in 2-3 μg of whole genome amplified DNA from 10 ng of starting genomic DNA. *B. suis* 1330 genomic DNA was obtained from BEI resources (Manassas, VA).

3.5.3 PCR assay on the *IS711* Element of *Brucella* species and sequencing of PCR products

Primers were chosen for the *IS711* element of *B. abortus* and *B. suis* as described in [13] and synthesized by IDT (Integrated DNA technologies, Coralville, IA). Samples were analyzed by PCR using 25 ng of starting material in a total reaction volume of 50 μl containing 2x master mix (Promega, Madison, WI), template and 20 pmoles of primer (Integrated DNA Technologies, Coralville, IA). Reactions were performed using an initial three minutes denaturation step at 95ºC, followed by 35 cycles of 30 seconds at 95ºC, 50 seconds at 56ºC, 1 min at 72ºC, and a final extension step for 7 minutes at 72ºC. The samples (5 μl) were treated with 2μl of ExoSAP (Affymetrix, Santa Clara, CA) for 15 minutes at 37 ºC, and the reaction was inactivated by heating to 80 ºC for 15 minutes. The samples

were sequenced using ABI big dye terminator chemistry reactions and sequenced on the 3730 sequencer (ABI, Foster City, CA).

3.5.4 Species independent array design, preparation and hybridization and array data processing

DNA concentration (260 nm) and purity (260/280 and 260/230 nm) were assessed by the spectrophotometer and quality by agarose gel electrophoresis. Samples with 260/230 nm ratios greater than 1.8 were used following established protocols for array comparative genomic hybridization (CGH). We designed the UBDA microarray which was then manufactured by Roche-Nimblegen (Madison, WI) as a custom 373K probe chip and genomic DNAs(1 μg) were labeled and hybridized on the UBDA chip as previously described [18]. Data files from the UBDA arrays were imported individually into Nimblescan (Roche Nimblegen, Madison, WI,) and background corrected [28]. A parsing script written in Perl was used to extract 9-mer (262,144 probes and replicates) probe intensities from the 373K UBDA array and signal intensity values were $\log_2$ transformed. Signal intensity values from 9-mer probes are available in Supplementary Table 6.

3.5.5 Principal component analysis of UBDA array probe signal intensity values using singular value decomposition

Principal component analysis (PCA) was employed to determine the isolate's composite identity from the UBDA array data for the entire 9-mer probe set using a custom MATLAB (Natick, MA) script. For each sample the total sum of all distances for all probes was computed with respect to the reference sample *B. abortus* 2308 or *B. suis* 1330. The weighted distance was computed by dividing the total sum of distances by the number of probes which is 262,144. The statistical confidence for the detection limit was set at 0.6% and the species with the lower distance measure was used to determine the closest match to the field isolate. Samples for which similarity to the references were within 0.6% were designated as mixed or chimeric.

3.5.6 Phylogenomic relationship tree based on UBDA signal intensity values

For each sample a Pearson's correlation matrix, which included self-similarity and similarity to the remaining samples in the matrix for the 9-mer probes on the array, was created. Then, the distance matrix was input to the neighbor-joining method implemented in the PHYLIP software suite and TreeView [30] to produce a phylogenetic tree. For comparison, a phylogenomic tree was also created using RMA normalization method [23].

The files were normalized in a batch mode using Nimblescan (Roche NimbleGen, Madison, WI).

3.5.7 Sequence analysis using Illumina sequencer

Illumina paired-end sequencing protocols were used to sequence nine TAHC field isolates genomic DNA and the original *B. suis*1330 DNA (76 cycles for isolate 2, 3, 17, 22, 29, 34 and 35, and 101 cycles for isolate 13, 33 and *B. suis* 1330) on the Illumina GAIIx sequencer (San Diego, CA). For accurate analysis, all low quality bases (< 0.99 quality score) from the sequencing reads were trimmed. Then, these reads were mapped to the genomic sequences of the five completely sequenced *Brucella* species including *B. melitensis* 16M, *B. abortus* 9-941, *B. abortus* 2308, *B. suis* 1330 and *B. abortus* ATCC 23445 by BWA [24] to determine the optimum reference for each sample based on similarity. Consensus sequences from the reads mapped to the selected references were then generated by SAMtools [31]. Unmapped reads were also analyzed by BLASTN search [32, 33] against the nucleotide (NT) database (using options '-e 1e-15 -F F').

3.5.8 Phylogenomic analysis using protein sequences of field isolates

Single-copy genes were sought among annotated *Brucella* genomes obtained from the PATRIC (www.patricbrc.org, July 31 2011) resource and nine whole genome sequenced isolates from TAHC. Protein sequences from

all genomes were compared using BLAST (BLASTP with an –e parameters set to 1e-80). The BLAST results were clustered using MCL [34] with default parameters. The resulting clusters showed a peak size category at 42 with 1154 clusters of this size. Amino-acid sequences were grouped into files by cluster and aligned using MUSCLE [35] (version 3.7). Hidden Markov models were built from each protein alignment using hmmbuild [36]. The HMMs were then used to search the nucleotide genomes of the annotated *Brucella* genomes plus the new field isolate sequences (unannotated) using estwise of the WISE2 package [37]. The estwise output was parsed to align nucleotides by codon position along with the amino acid encoded. HMMs for which the worst estwise score was less than 0.9 of the average score were deleted (as potential mismatched homologs or incomplete sequences), leaving 1006 high-quality single-copy genes comprising 357,745 codons. The program raxmlHPC [38] was used to infer the maximum likelihood phylogeny under the GTRGAMMA rate model with different rates for first, second and third codon positions.

3.6 Acknowledgements

This project was funded by subaward 570636 from DHS 2007-ST-061-000002 from the U.S. Department of Homeland Security - National Center of Excellence for Foreign Animal and Zoonotic Disease Defense at Texas

A&M University and by the Director's funds at Virginia Bioinformatics Institute, Virginia Tech. S. Shallom received funding from SREB (Southern Regional Education Board) state doctoral scholar award. Dr. Nammalwar Sriranganathan and Hamzeh Al Qublan from Biomedical and Veterinary Sciences at Virginia Tech kindly provided *B. abortus* 2308genomic DNA. *Francisella tularensis* LVS genomic DNA was kindly provided by Dr. Abey Bandara and Dr. Tom Inzana from the Department of Biology at Virginia Tech. The following reagent was obtained through the NIH Biodefense and Emerging Infections Research Resources Repository, NIAID, NIH: Genomic DNA from *Brucella suis,* Strain 1330 (NCTC 10316), NR-2526, Genomic DNA from *Brucella melitensis*, Strain 16M (NCTC 10094), NR-2525, Genomic DNA from *Brucella abortus*, Strain RB51, NR-2553. We would like to extend special thanks to Greg Thorne and Shaukat Rangwala at MoGene for their valuable technical assistance. The field isolates were sequenced by the Virginia Bioinformatics Institute's Core Laboratory Facility at Virginia Tech.

3.7 Attribution

S. Shallom designed and carried out experiments, analyzed the data, developed principal component analysis algorithm and wrote the manuscript. H. Tae, assembled and aligned the *Brucella* genomes and wrote the

manuscript, C. Franck provided useful discussions, A. Dickerman contributed to the phylogenetic tree from protein coding regions on the *Brucella* genome, L. McIver provided computation expertise, D. Preston provided the *Brucella* field isolates, L. Sarmento extracted genomic DNA from the *Brucella* organism, G. Adams conceived of the study, reviewed the manuscript and provided useful discussions, H. Garner conceived of the study, participated in the study design and mentored in drafting the manuscript.

3.8 Bibliography

1. Godfroid J, Cloeckaert A, Liautard JP, Kohler S, Fretin D, Walravens K, Garin-Bastuji B, Letesson JJ: **From the discovery of the Malta fever's agent to the discovery of a marine mammal reservoir, brucellosis has continuously been a re-emerging zoonosis**. *Veterinary research* 2005, **36**(3):313-326.
2. Morgan WJ: **Brucella classification and regional distribution**. *Dev Biol Stand* 1984, **56**:43-53.
3. Scholz HC, Hubalek Z, Sedlacek I, Vergnaud G, Tomaso H, Al Dahouk S, Melzer F, Kampfer P, Neubauer H, Cloeckaert A *et al*: **Brucella microti sp. nov., isolated from the common vole Microtus**

arvalis. *International journal of systematic and evolutionary microbiology* 2008, **58**(Pt 2):375-382.

4. G.G A: **Techniques for the brucellosis laboratory**. In., 1988 edn. Paris: paris: Institut National de la Recherche Agronomique; 1988.
5. De Klerk E, Anderson R: **Comparative evaluation of the enzyme-linked immunosorbent assay in the laboratory diagnosis of brucellosis**. *J Clin Microbiol* 1985, **21**(3):381-386.
6. Lindberg AA, Haeggman S, Karlson K, Carlsson HE, Mair NS: **Enzyme immunoassay of the antibody response to Brucella and Yersinia enterocolitica 09 infections in humans**. *The Journal of hygiene* 1982, **88**(2):295-307.
7. Caroff M, Bundle DR, Perry MB, Cherwonogrodzky JW, Duncan JR: **Antigenic S-type lipopolysaccharide of Brucella abortus 1119-3**. *Infect Immun* 1984, **46**(2):384-388.
8. Caroff M, Bundle DR, Perry MB: **Structure of the O-chain of the phenol-phase soluble cellular lipopolysaccharide of Yersinia enterocolitica serotype O:9**. *European journal of biochemistry / FEBS* 1984, **139**(1):195-200.

9. Corbel MJ, Stuart FA, Brewer RA: **Observations on serological cross-reactions between smooth Brucella species and organisms of other genera**. *Dev Biol Stand* 1984, **56**:341-348.

10. Cloeckaert A KP, Limet JN: **Antibody response to *Brucella* ourter membrane proteins in bovine brucellosis: immunoblot analysis and competitive enzyme-linked immunosorbent assay using monoclonal antibodies.***Journal of Clinical Microbiology* 1992, **30**:3168-3174.

11. Samartino L, Gall D, Gregoret R, Nielsen K: **Validation of enzyme-linked immunosorbent assays for the diagnosis of bovine brucellosis**. *Vet Microbiol* 1999, **70**(3-4):193-200.

12. Bricker BJ: **PCR as a diagnostic tool for brucellosis**. *Vet Microbiol* 2002, **90**(1-4):435-446.

13. Bricker BJ, Halling SM: **Differentiation of Brucella abortus bv. 1, 2, and 4, Brucella melitensis, Brucella ovis, and Brucella suis bv. 1 by PCR**. *J Clin Microbiol* 1994, **32**(11):2660-2666.

14. Whatmore AM, Shankster SJ, Perrett LL, Murphy TJ, Brew SD, Thirlwall RE, Cutler SJ, MacMillan AP: **Identification and characterization of variable-number tandem-repeat markers for typing of Brucella spp**. *J Clin Microbiol* 2006, **44**(6):1982-1993.

15. Bricker BJ, Ewalt DR, Halling SM: **Brucella 'HOOF-Prints': strain typing by multi-locus analysis of variable number tandem repeats (VNTRs)**. *BMC Microbiol* 2003, **3**:15.

16. Bricker BJ, Ewalt DR: **Evaluation of the HOOF-Print assay for typing Brucella abortus strains isolated from cattle in the United States: results with four performance criteria**. *BMC Microbiol* 2005, **5**:37.

17. Foster JT, Okinaka RT, Svensson R, Shaw K, De BK, Robison RA, Probert WS, Kenefic LJ, Brown WD, Keim P: **Real-time PCR assays of single-nucleotide polymorphisms defining the major Brucella clades**. *J Clin Microbiol* 2008, **46**(1):296-301.

18. Shallom SJ, Weeks JN, Galindo CL, McIver L, Sun Z, McCormick J, Adams LG, Garner HR: **A species independent universal bio-detection microarray for pathogen forensics and phylogenetic classification of unknown microorganisms**. *BMC Microbiol* 2011, **11**:132.

19. Darling AC, Mau B, Blattner FR, Perna NT: **Mauve: multiple alignment of conserved genomic sequence with rearrangements**. *Genome Res* 2004, **14**(7):1394-1403.

20. Holder MT, Zwickl DJ, Dessimoz C: **Evaluating the robustness of phylogenetic methods to among-site variability in substitution processes**. *Philos Trans R Soc Lond B Biol Sci* 2008, **363**(1512):4013-4021.

21. Galindo CL, McIver LJ, McCormick JF, Skinner MA, Xie Y, Gelhausen RA, Ng K, Kumar NM, Garner HR: **Global microsatellite content distinguishes humans, primates, animals, and plants**. *Mol Biol Evol* 2009, **26**(12):2809-2819.

22. Belosludtsev YY, Bowerman D, Weil R, Marthandan N, Balog R, Luebke K, Lawson J, Johnston SA, Lyons CR, Obrien K *et al*: **Organism identification using a genome sequence-independent universal microarray probe set**. *Biotechniques* 2004, **37**(4):654-658, 660.

23. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP: **Exploration, normalization, and summaries of high density oligonucleotide array probe level data**. *Biostatistics* 2003, **4**(2):249-264.

24. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform**. *Bioinformatics* 2009, **25**(14):1754-1760.

25. Tae H, Shallom S, Settlage R, Preston D, Adams LG, Garner HR: **Revised Genome Sequence of Brucella suis 1330**. *J Bacteriol* 2011, **193**(22):6410.

26. Ocampo-Sosa AA, Garcia-Lobo JM: **Demonstration of IS711 transposition in Brucella ovis and Brucella pinnipedialis**. *BMC Microbiol* 2008, **8**:17.

27. Rossetti CA, Galindo CL, Lawhon SD, Garner HR, Adams LG: **Brucella melitensis global gene expression study provides novel information on growth phase-specific gene regulation with potential insights for understanding Brucella:host initial interactions**. *Bmc Microbiology* 2009, **9**:-.

28. Ritchie ME, Silver J, Oshlack A, Holmes M, Diyagama D, Holloway A, Smyth GK: **A comparison of background correction methods for two-colour microarrays**. *Bioinformatics* 2007, **23**(20):2700-2707.

29. Jackson EJ: **A User's Guide to Principal Components**. New Jersey: John Wiley & Sons, Inc.; 2003.

30. Page RD: **TreeView: an application to display phylogenetic trees on personal computers**. *Comput Appl Biosci* 1996, **12**(4):357-358.

31. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: **The Sequence Alignment/Map format and SAMtools**. *Bioinformatics* 2009, **25**(16):2078-2079.

32. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**. *Journal of molecular biology* 1990, **215**(3):403-410.

33. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Res* 1997, **25**(17):3389-3402.

34. van Dongen S, Abreu-Goodger C, Enright AJ: **Detecting microRNA binding and siRNA off-target effects from expression data**. *Nature methods* 2008, **5**(12):1023-1025.

35. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput**. *Nucleic Acids Res* 2004, **32**(5):1792-1797.

36. Eddy SR: **Multiple alignment using hidden Markov models**. *Proceedings / International Conference on Intelligent Systems for Molecular Biology ; ISMB International Conference on Intelligent Systems for Molecular Biology* 1995, **3**:114-120.

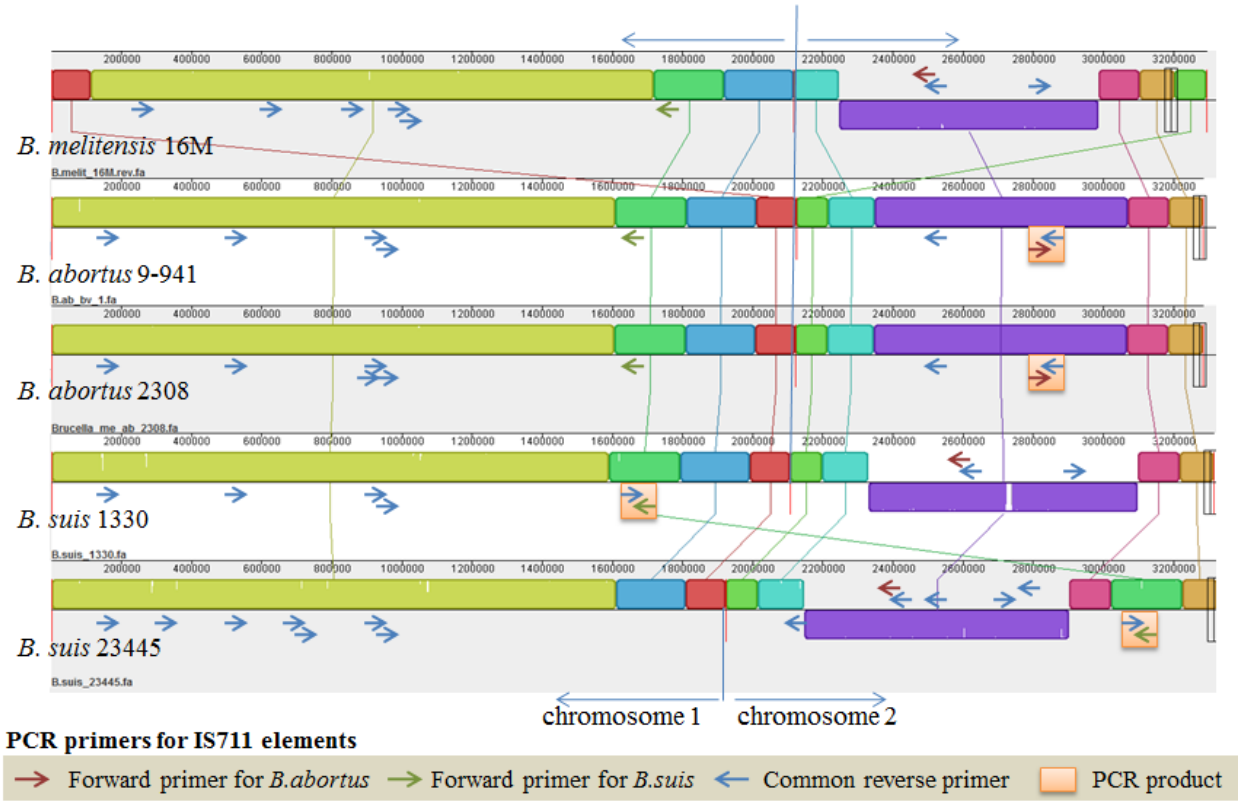
37. Birney E, Clamp M, Durbin R: **GeneWise and Genomewise**. *Genome Res* 2004, **14**(5):988-995.

38. Stamatakis A: **RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models**. *Bioinformatics* 2006, **22**(21):2688-2690.

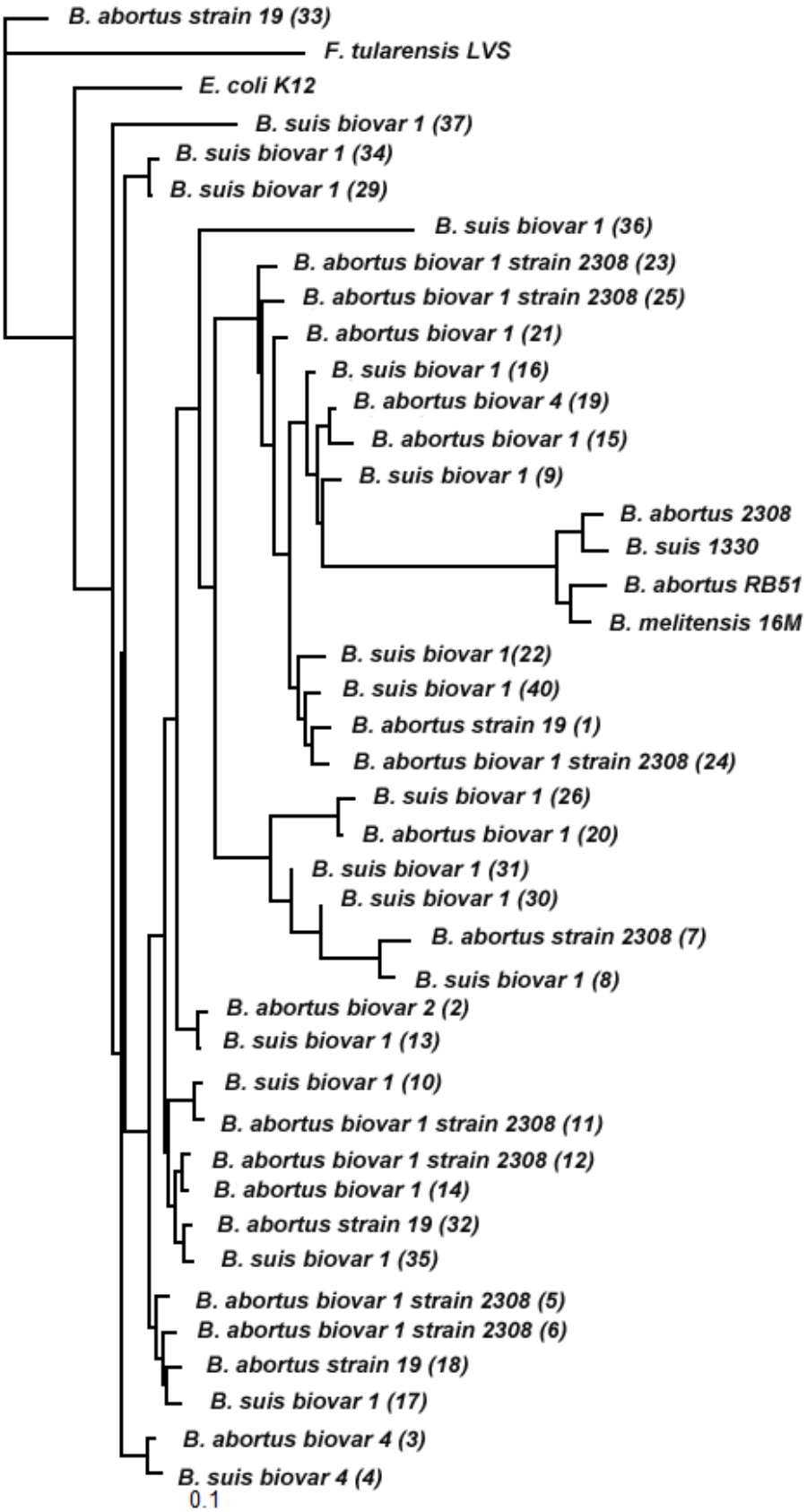
3.9. Figures

3.9.1 Figure 1: Locations of PCR primer sequences for *B. suis* and *B. abortus* in 5 completed *Brucella* genomes aligned by Mauve.

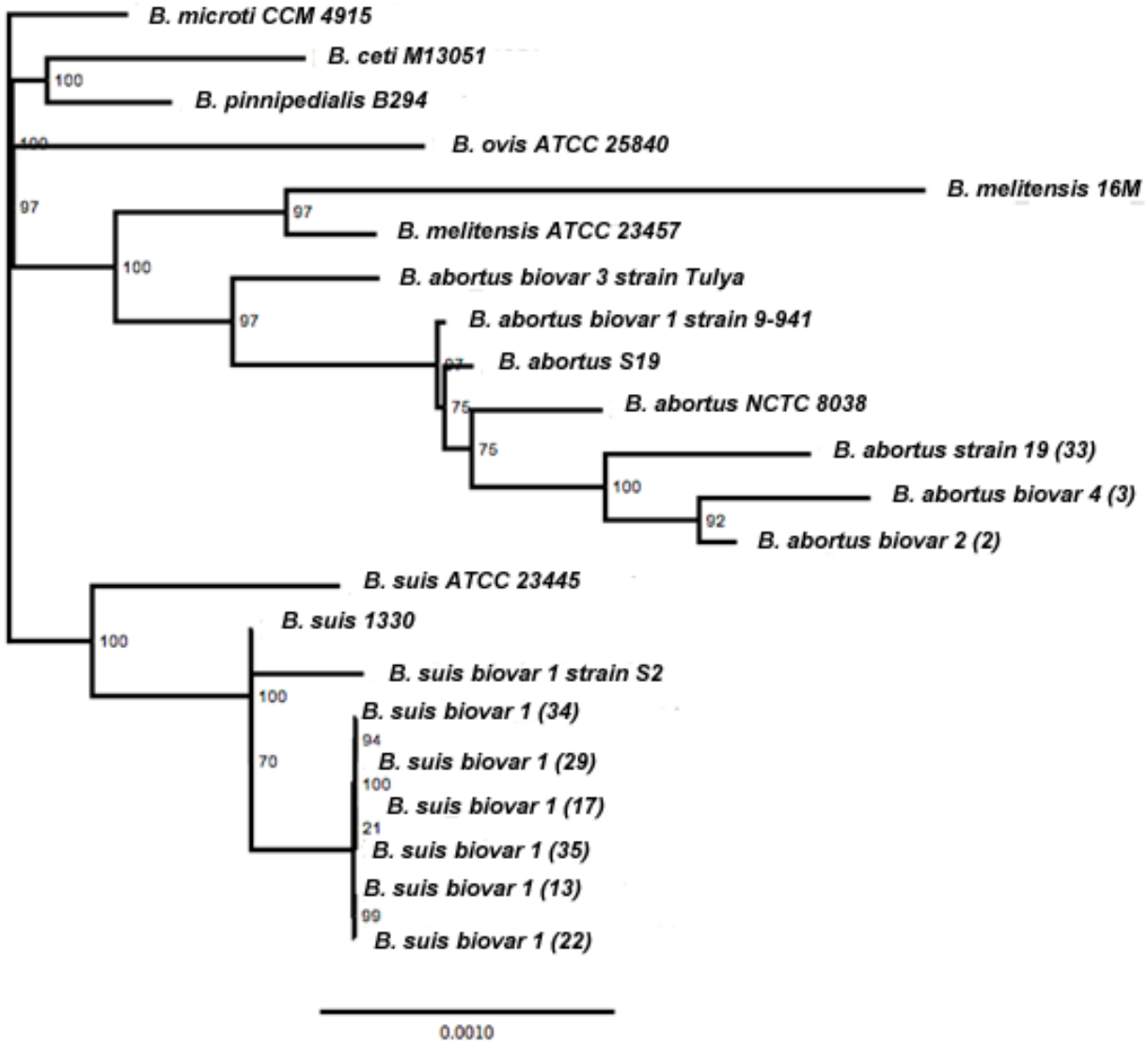
Five completed genomes including *B. melitensis* 16M, *B. abortus* biovar 1 9-941, *B. abortus*2308, *B. suis* 1330 and *B. suis* ATCC 23445 were aligned by Mauve [19]. Most genomes are very similar except a few rearranged regions. All five genomes have PCR primer sequences for bio-assays of both *B. abortus* and *B. suis* species, but produce PCR products for only their corresponding species.



3.9.2 Figure 2: Phylogenomic relationships from 9-mer probe set between *Brucella* field isolates and other known reference genomes. All 262,144 9-mer data points for each of the samples were $\log_2$ transformed. A Pearson correlation matrix was created by comparing each sample against all other samples to generate a taxonomic relationship tree using the PHYLIP software and visualized in the *Treeview* program.


B. abortus strain 19 (33)
F. tularensis LVS
E. coli K12
B. suis biovar 1 (37)
B. suis biovar 1 (34)
B. suis biovar 1 (29)
B. suis biovar 1 (36)
B. abortus biovar 1 strain 2308 (23)
B. abortus biovar 1 strain 2308 (25)
B. abortus biovar 1 (21)
B. suis biovar 1 (16)
B. abortus biovar 4 (19)
B. abortus biovar 1 (15)
B. suis biovar 1 (9)
B. abortus 2308
B. suis 1330
B. abortus RB51
B. melitensis 16M
B. suis biovar 1(22)
B. suis biovar 1 (40)
B. abortus strain 19 (1)
B. abortus biovar 1 strain 2308 (24)
B. suis biovar 1 (26)
B. abortus biovar 1 (20)
B. suis biovar 1 (31)
B. suis biovar 1 (30)
B. abortus strain 2308 (7)
B. suis biovar 1 (8)
B. abortus biovar 2 (2)
B. suis biovar 1 (13)
B. suis biovar 1 (10)
B. abortus biovar 1 strain 2308 (11)
B. abortus biovar 1 strain 2308 (12)
B. abortus biovar 1 (14)
B. abortus strain 19 (32)
B. suis biovar 1 (35)
B. abortus biovar 1 strain 2308 (5)
B. abortus biovar 1 strain 2308 (6)
B. abortus strain 19 (18)
B. suis biovar 1 (17)
B. abortus biovar 4 (3)
B. suis biovar 4 (4)
0.1

3.9.3 Figure 3: Phylogenomic tree from nine recently sequenced *Brucella* field isolates and thirteen known previously sequenced *Brucella* genomes. The maximum likelihood tree based on 1,006 presumptively vertically inherited genes places the new field isolates in two locations on the tree, 6 nested within *B. suis* and 3 nested within *B. abortus.* The number of the new isolates is in parenthesis.



3.10 Tables

3.10.1 Table 1: Comparison of common variations between nine field samples to the *B. suis* 1330 genome. Each cell represents a number of common variations between two samples. Diagonal cells represent self-comparison for each sample, which is identical with the number of variations in each sample with respect to the *B. suis* 1330 genome[25].

No.	2	3	33	13	17	22	29	34	35
2	7912	7730	7676	39	40	39	36	36	36
3	7730	7931	7676	39	40	39	37	37	36
33	7676	7676	8062	39	40	39	37	37	36
13	39	39	39	80	36	65	35	35	36
17	40	40	40	36	74	39	63	66	64
22	39	39	39	65	39	75	37	38	38
29	36	37	37	35	63	37	71	68	63
34	36	37	37	35	66	38	68	72	64
35	36	36	36	36	64	38	63	64	72

3.10 Supplementary Tables

3.10.1A Supplementary Table 1A: Comparison of Biochemical Typing, Universal Biosignature Detection Array, PCR and Genome Sequence Analysis.

Mixed infection is a combination of *B. abortus* and *B. suis*. Note that Genome sequence analysis indicated that some samples labeled with an "*" contained contamination of a minor species at between 1:1736 and 1: 13642. Statistical confidence measure was set at 0.6% for PCA analysis (Section 3.3.4. and Supplementary Table 2 and 5).

No.	Biochemical Typing	Source	*IS711* PCR	UBDA	Genome sequence analysis
1	*B. abortus* strain 19	Unknown	*B. abortus*	*B. abortus*	---
2	*B. abortus* biovar 2	Unknown	*B. abortus*	mixed infection	*B. abortus*
3	*B. abortus* biovar 4	Unknown	*B. abortus*	*B. suis*	*B. abortus*
4	*B. suis* biovar 4	Unknown	*B. abortus*	*B. suis*	---
5	*B. abortus* biovar 1 strain 2308	Bovine milk	*B. abortus*	*B. suis*	---
6	*B. abortus* biovar 1 strain 2308	Bovine milk	*B. abortus*	*B. suis*	---
7	*B. abortus* strain 2308	Bovine milk	*B. abortus*	*B. abortus*	---
8	*B. suis* biovar 1	Bovine milk, tissue	mixed infection	*B. abortus*	---
9	*B. suis* biovar 1	Porcine tissue	mixed infection	mixed infection	---
10	*B. suis* biovar 1	Porcine tissue	mixed infection	*B. suis*	---
11	*B. abortus* biovar 1 strain 2308	Bovine tissue	*B. abortus*	mixed infection	---
12	*B. abortus* biovar 1 strain 2308	Bovine tissue	*B. abortus*	*B. suis*	---
13*	*B. suis* biovar 1	Equine tissue	mixed infection	mixed infection	*B. suis*
14	*B. abortus* biovar 1	Bovine milk	*B. abortus*	*B. suis*	---
15	*B. abortus* biovar 1	Bovine milk	*B. abortus*	*B. abortus*	---
16	*B. suis* biovar 1	Bovine milk, tissue	mixed infection	mixed infection	---
17*	*B. suis* biovar 1	Porcine tissue	mixed infection	*B. suis*	*B. suis*
18	*B. abortus* strain 19	Bovine tissue	mixed infection	*B. suis*	---

19	***B. abortus* biovar 4**	**Bovine milk**	***B. abortus***	***B. abortus***	**---**
20	***B. abortus* biovar 1**	**Bovine milk, tissue**	***B. abortus***	**mixed infection**	**---**
21	***B. abortus* biovar 1**	**Bovine milk, tissue**	***B. abortus***	***B. abortus***	**---**
22*	***B. suis* biovar 1**	**Bovine milk**	**mixed infection**	**mixed infection**	***B. suis***
23	***B. abortus* biovar 1strain 2308**	**Bovine milk**	***B. abortus***	**mixed infection**	**---**
24	***B. abortus* biovar 1strain 2308**	**Bovine milk**	***B. abortus***	***B. abortus***	**---**
25	***B. abortus* biovar 1strain 2308**	**Bovine milk**	***B. abortus***	***B. abortus***	**---**
26	***B. suis* biovar 1**	**Bovine tissue**	**mixed infection**	***B. suis***	**---**
29*	***B. suis* biovar 1**	**Bovine tissue**	**mixed infection**	***B. suis***	***B. suis***
30	***B. suis* biovar 1**	**Bovine tissue**	**mixed infection**	***B. abortus***	**---**
31	***B. suis* biovar 1**	**Bovine tissue**	**mixed infection**	***B. abortus***	**---**
32	***B. abortus* strain 19**	**Bovine milk**	***B. abortus***	***B. suis***	**---**
33	***B. abortus* strain 19**	**Bovine milk**	***B. abortus***	***B. suis***	***B. abortus***
34*	***B. suis* biovar 1**	**Bovine tissue**	**mixed infection**	***B. suis***	***B. suis***
35*	***B. suis* biovar 1**	**Bovine tissue**	**mixed infection**	***B. suis***	***B. suis***
36	***B. suis* biovar 1**	**Bovine tissue**	***B. abortus***	***B. suis***	**---**
37	***B. suis* biovar 1**	**Bovine milk**	***B. abortus***	***B. suis***	**---**
40	***B. suis* biovar 1**	**Bovine milk**	**mixed infection**	***B. abortus***	**---**

3.10.1B Supplementary Table 1B: Principal component analysis of field isolates with *B. suis* 1330 and *B. abortus* 2308 using 9-mer (262,144) probes

No.	*B. suis* 1330 distance (all 9-mer probes)	*B. abortus* 2308 distance (all 9-mer probes)	*B. suis* 1330 (weighted distance)	*B. abortus* 2308 (weighted distance)	Identity (9-mer probes)
1	84826	81706	0.324	0.312	*B. abortus*
2	95434	95165	0.364	0.363	mixed infection
3	92150	94314	0.352	0.360	*B. suis*
4	90956	93132	0.347	0.355	*B. suis*
5	103382	105476	0.394	0.402	*B. suis*
6	192472	196927	0.734	0.751	*B. suis*
7	105010	102283	0.401	0.390	*B. abortus*
8	97791	95725	0.373	0.365	*B. abortus*
9	80377	80217	0.307	0.306	mixed infection
10	97508	98318	0.372	0.375	*B. suis*
11	94407	94715	0.360	0.361	mixed infection
12	98323	99559	0.375	0.380	*B. suis*
13	96078	96322	0.367	0.367	mixed infection
14	100938	101991	0.385	0.389	*B. suis*
15	83143	80367	0.317	0.307	*B. abortus*
16	83995	83832	0.320	0.320	mixed infection
17	96578	99461	0.368	0.379	*B. suis*
18	94542	97069	0.361	0.370	*B. suis*
19	84293	81169	0.322	0.310	*B. abortus*
20	93257	93683	0.356	0.357	mixed infection
21	86248	85737	0.329	0.327	*B. abortus*
22	84242	81365	0.321	0.310	mixed infection
23	84569	84207	0.323	0.321	mixed infection
24	82335	79106	0.314	0.302	*B. abortus*

25	**82841**	**82170**	**0.316**	**0.313**	***B. abortus***
26	**96425**	**96682**	**0.368**	**0.369**	***B. suis***
29	**93399**	**95673**	**0.356**	**0.365**	***B. suis***
30	**100962**	**99980**	**0.385**	**0.381**	***B. abortus***
31	**93537**	**92640**	**0.357**	**0.353**	***B. abortus***
32	**96808**	**97498**	**0.369**	**0.372**	***B. suis***
33	**93689**	**97874**	**0.357**	**0.373**	***B. suis***
34	**93007**	**95147**	**0.355**	**0.363**	***B. suis***
35	**99465**	**100384**	**0.379**	**0.383**	***B. suis***
36	**104003**	**104836**	**0.397**	**0.400**	***B. suis***
37	**100921**	**103455**	**0.385**	**0.395**	***B. suis***
40	**79899**	**76128**	**0.305**	**0.290**	***B. abortus***

3.10.2 Supplementary Table 2: Comparison of similarities among nine *Brucella* samples and two reference genomes; *B. abortus* biovar 1 9-941 *and B. suis* 1330.

Sequencing reads of all nine isolates were mapped to *B. abortus* biovar 1 9-941 and *B. suis* 1330 separately for comparison of similarities among nine *Brucella* samples and two references. The mapping results show that isolates 2, 3 and 33 are highly similar to *B. abortus* biovar 1 9-941 while isolates 13, 17, 22, 29, 34 and 35 are highly similar to *B. suis* 1330.

No.	Identification by antibody test	Identification by PCR test	Mapping on *B. abortus* biovar 1 9-941				Mapping on *B. suis* 1330			
			coverage	Base identity %	# of consensus blocks	# of variants	coverage	Base identity %	# of consensus blocks	# of variants
2	*B. abortus* biovar 2	*B. abortus*	1020	99.991	2	150	1002	98.676	248	7912
3	*B. abortus* biovar 4	*B. abortus*	1034	99.988	3	299	1015	98.197	113	7931
33	*B. abortus* S19	*B. abortus*	1797	99.964	4	261	1773	98.409	149	8062
13	*B. suis* biovar I	*B. abortus* *B. suis*	1438	99.104	79	7786	1460	99.993	2	80
17	*B. suis* biovar I	*B. abortus* *B. suis*	1084	99.272	144	7713	1094	99.996	2	74
22	*B. suis* biovar I	*B. abortus* *B. suis*	1104	99.094	107	7693	1114	99.995	3	75
29	*B. suis* biovar I	*B. abortus* *B. suis*	947	98.939	57	7690	955	99.996	3	71
34	*B. suis* biovar I	*B. abortus* *B. suis*	1081	99.023	89	7690	1091	99.996	2	72
35	*B. suis* biovar I	*B. abortus* *B. suis*	1019	98.979	69	7688	1029	99.996	2	72

3.10.3 Supplementary Table 3: Analysis of unmapped reads using BLAST program against NT database.

Each cell shows the number of reads identified as contamination by BLAST program. While the main contaminant components of the *B. suis* 1330 re-sequenced original sample is human DNA, contaminants of the other sample are either other *Brucella* strains, or other organisms. Detailed information for the other contaminant components is available in Supplementary table4.

No.	Map to Reference	Total Read	Total Unmapped (% for total reads)	Unmapped Reads (% for unmapped)					
				No Hit in NT DB	Other *Brucella*	Human	Bovine	Porcine	other
2	*B. abortus* 9-941	44,646,218	15,412 (0.03%)	4,858 (31.52%)	1,081 (7.01%)	5,135 (33.32%)	51 (0.33%)	21 (0.14%)	4,266 (27.68%)
3	*B. abortus* 9-941	45,344,854	14,103 (0.03%)	3,787 (26.85%)	3,101 (21.99%)	4,627 (32.81%)	44 (0.31%)	23 (0.16%)	2,521 (17.88%)
33	*B. abortus* S19	63,154,922	46,344 (0.07%)	11,673 (25.19%)	20,910 (45.12%)	1,825 (3.94%)	4 (0.01%)	16 (0.03%)	11,916 (25.71%)
13	*B. suis* 1330	49,912,388	66,035 (0.13%)	32,586 (49.35%)	5,953 (9.01%)	1,524 (2.31%)	2 (3-5%)	3 (4-5%)	25,967 (39.32 %)
17	*B. suis* 1330	57,946,186	8,095 (0.01%)	3,521 (43.50%)	2,268 (28.02%)	1,144 (14.13%)	19 (0.23%)	1 (0.01%)	1,142 (14.11%)
22	*B. suis* 1330	49,278,036	7,942 (0.02%)	3,229 (40.66%)	2,397 (30.18%)	1,234 (15.54%)	6 (0.08%)	3 (0.04%)	1,073 (13.51%)
29	*B. suis* 1330	42,295,608	22,681 (0.05%)	5,640 (24.87%)	1,429 (6.30%)	7,003 (30.88%)	18 (0.08%)	24 (0.11%)	8,567 (37.77 %)
34	*B. suis* 1330	48,359,626	10,603 (0.02%)	3,470 (32.73%)	1,662 (15.67%)	3,250 (30.65%)	21 (0.20%)	15 (0.14%)	2,185 (20.61%)
35	*B. suis* 1330	45,641,280	19,590 (0.04%)	6,770 (34.56%)	1,721 (8.79%)	7,597 (38.78%)	19 (0.10%)	137 (0.70%)	3,346 (17.08%)
***B. suis* 1330**	*B. suis* 1330	52,210,172	36,184 (0.07%)	1,458 (4.03%)	907 (2.51%)	32,489 (89.79%)	4 (0.01%)	2 (0.01%)	1,324 (3.66%)

3.10.4 Supplementary Table 4: Analysis of the unmapped reads from other contaminant micro-organisms listed in supplementary table 3.

The major genomes were from *Propionibacterium, Mycobacterium, Staphylococcus, Escherichia, Ochrobactrum and Pseudomonas*.

No.	*Propionibacterium*	*Mycobacterium*	*Staphylococcus*	*Escherichia*	*Ochrobactrum*	*Pseudomonas*	other
2	1,912	1,234	66	46	88	131	789
3	1,403	284	66	11	61	39	657
33	1,847	2	4,801	15	1,451	50	3,750
13	4,705	8	18,504	5	438	62	2,245
17	479	245	8	6	16	21	367
22	410	181	15	29	47	25	366
29	4,567	189	62	1,436	24	27	2,262
34	1,596	40	22	7	32	12	476
35	2,223	99	67	16	33	42	866
***B. suis* 1330**	124	0	0	1	18	8	1,173

3.10.5 Supplementary Table 5: Sequence coverage on a non *B. suis* 1330 region.

The read coverage at the sequence between 261,000th and 280,000th base (which does not exist in the *B. suis* 1330 genome) of *B. abortus* 9-941 genome was compared. While the coverage of the genome sequence of *B. suis* 1330 original sample was nil, the coverage of genome sequences of isolates 13, 17, 22, 29, 34 and 35 were approximately 0.07x to 0.63x, suggesting these field isolates were mixtures of *B. suis* and other *Brucella*. (Mixture ratio is estimated by comparing the read coverage at the whole genome to that at 261K~280Kof *B. abortus* 9-941.)

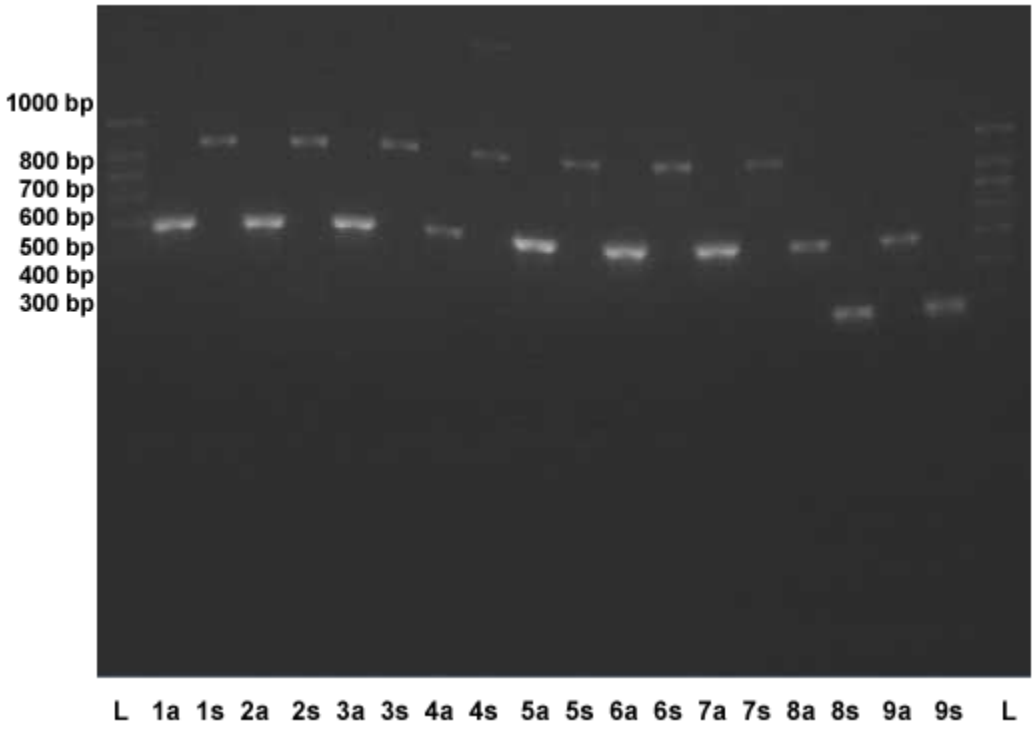
Analysis	**2**	**3**	**33**	**13**	**17**	**22**	**29**	**34**	**35**	***B. suis* 1330**
Coverage at 261K~280K of *B. abortus 9-941*	1066.8	1113.6	5709.3	0.35	0.63	0.28	0.07	0.17	0.11	0
Mixture ratio of *B. suis* and other *Brucella*	-	-	-	4171:1	1736:1	3978:1	13642:1	6417:1	9354:1	0

3.10.6 Supplementary Table 6: Universal Bio-signature Detection Array probe intensities from 9-mer with *Brucella* field isolates hybridized on the array (log_2 scale).

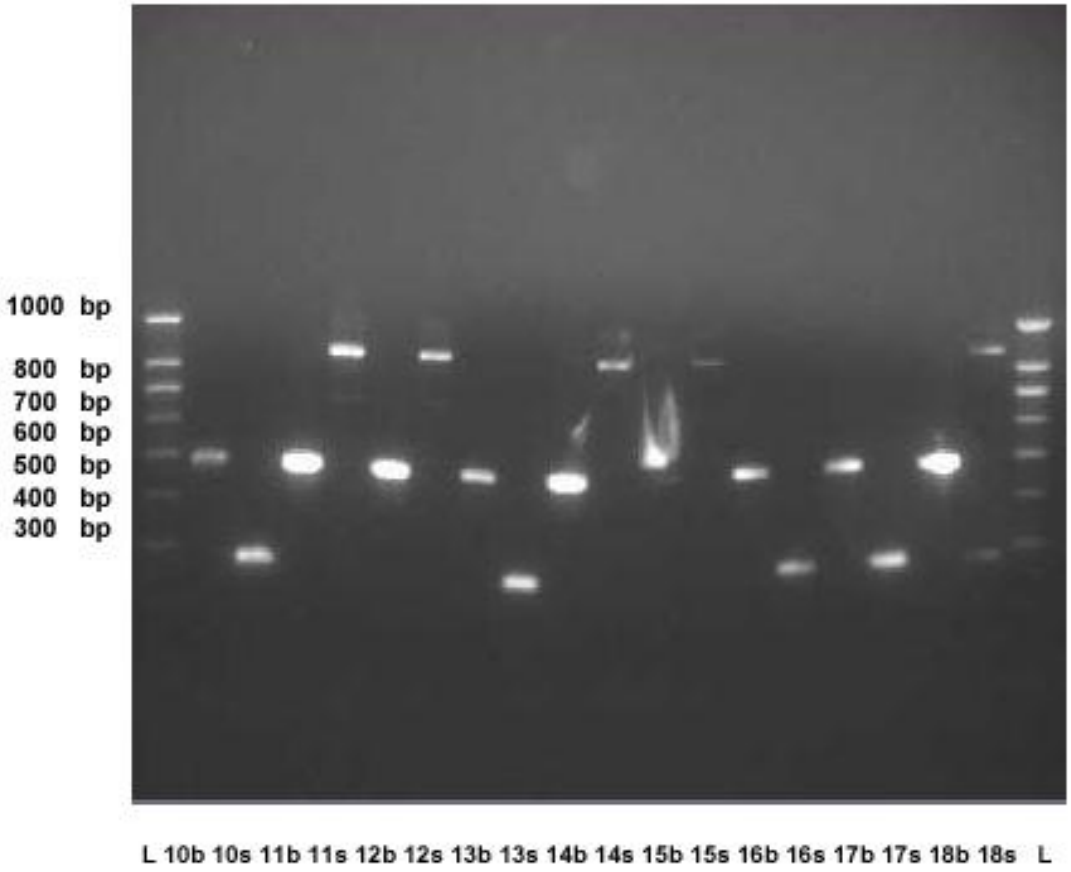
File name: TAHC_UBDA_log_2 (available at http://innovation.vbi.vt.edu)

3.11 Supplementary Figures

3.11.1A Supplementary Figure 1A: PCR of genetic element *IS711* from *Brucella* field isolates 1 through 9 with *IS711* element *B. abortus* (a) and *B. suis* (s) primers.

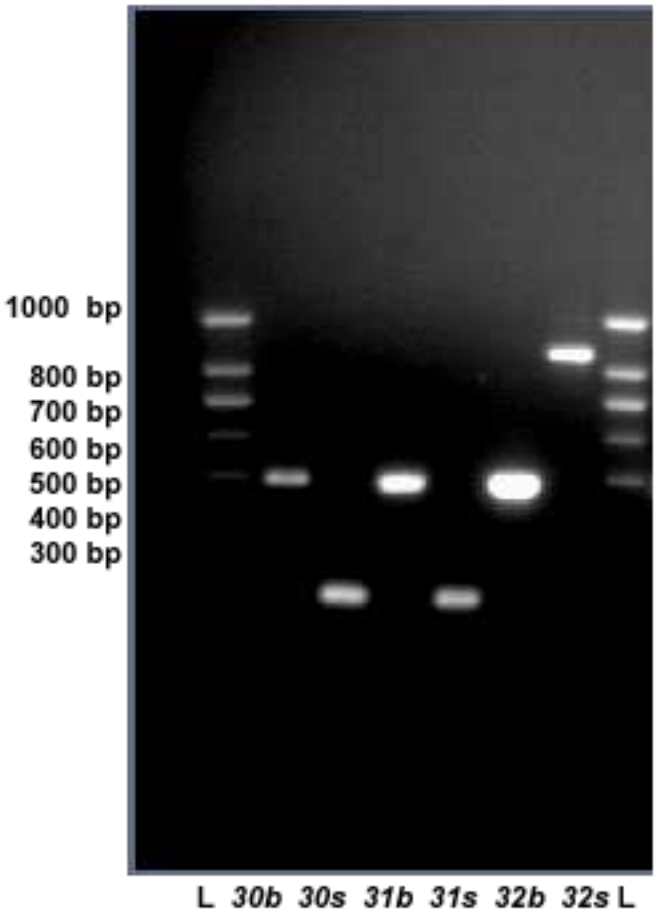


3.11.1B Supplementary Figure 1B: PCR of genetic element *IS711* from *Brucella* field isolates 10 through 18 with *IS711* element *B. abortus* (a) and *B. suis* (s) primers.

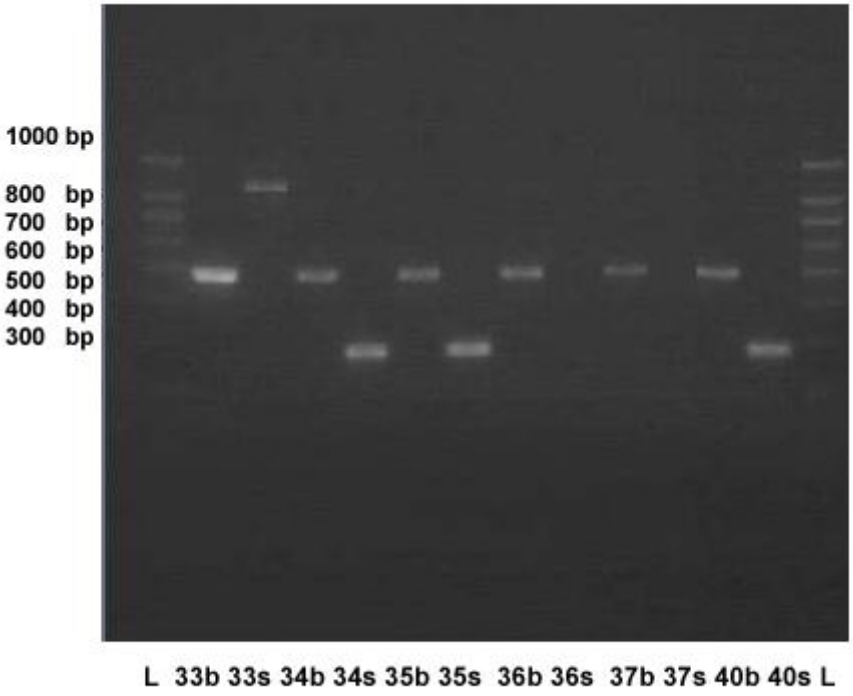


3.11.1C Supplementary Figure 1C: PCR of genetic element *IS711* from *Brucella* field isolates 19 through 26 and 29 with *IS711* element *B. abortus* (a) and *B. suis* (s) primers.

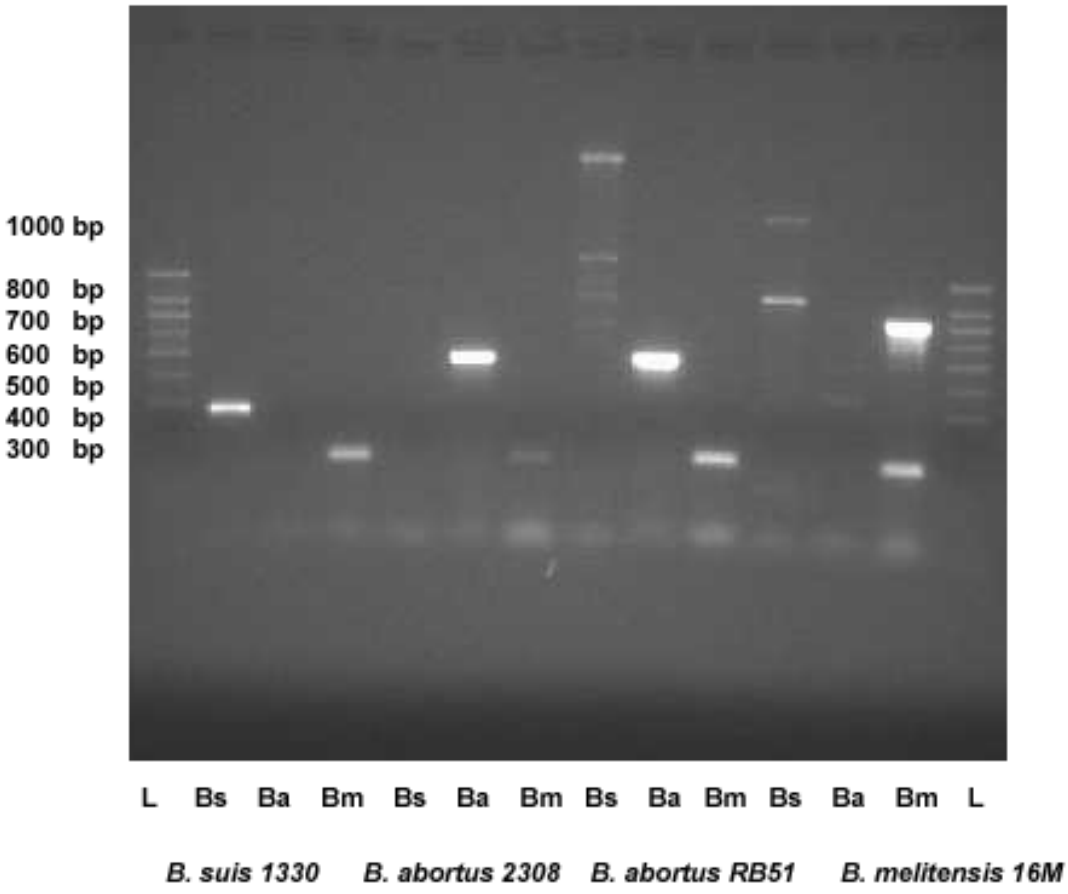
3.11.1D Supplementary Figure 1D: PCR of genetic element *IS711* from *Brucella* field isolates 30 through 32 with *IS711* element *B. abortus* (a) and *B. suis* (s) primers.



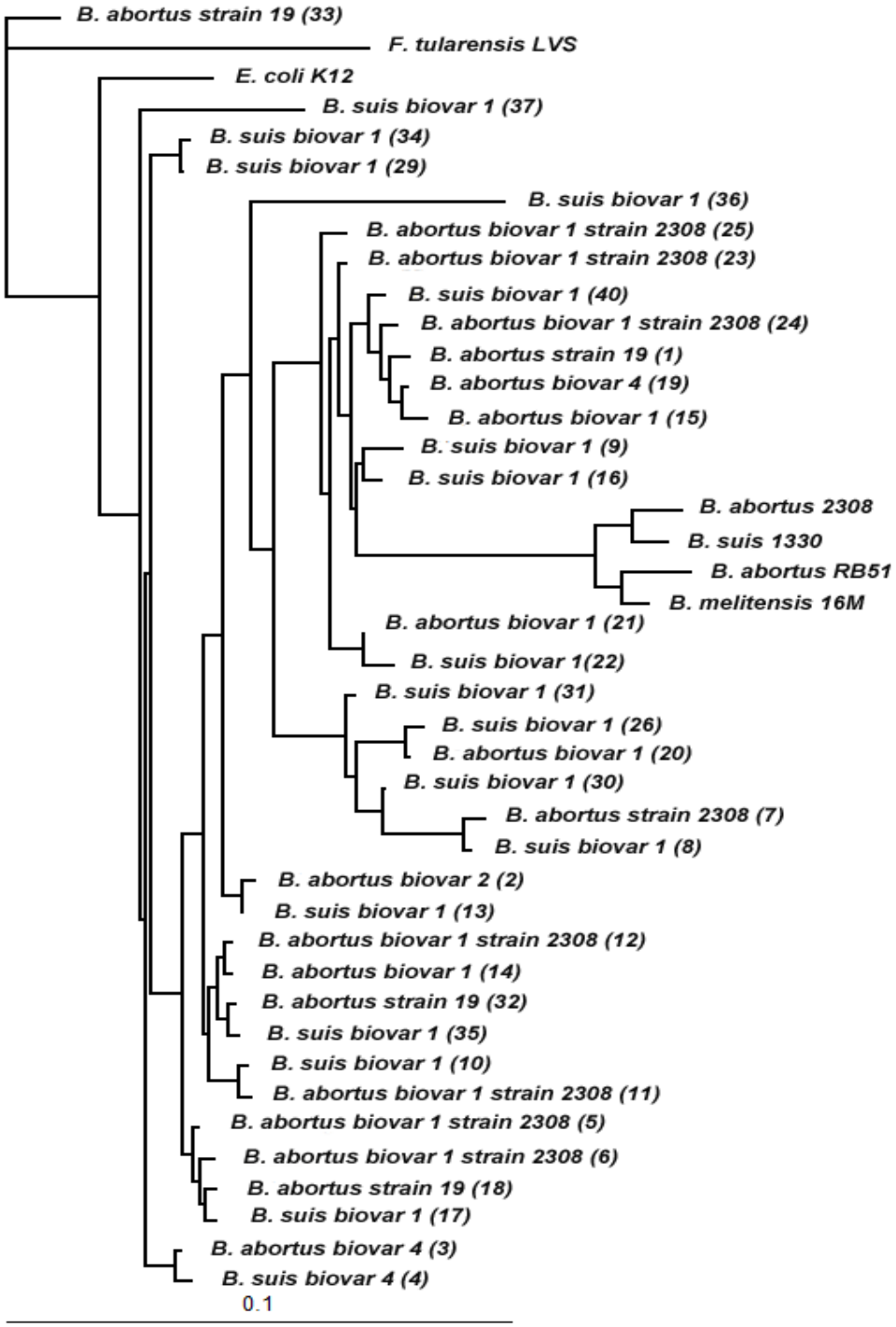
3.11.1E Supplementary Figure 1E: PCR of genetic element *IS711* from *Brucella* field isolates 33 through 37 and 40 with *IS711* element *B. abortus* (a) and *B. suis* (s) primers.



3.11.2 Supplementary Figure 2: PCR assay of *IS711* element primers from *Brucella* species *suis* (Bs), *abortus* (Ba) and *melitensis* (Bm) with *B. suis* 1330, *B. abortus* 2308, *B. abortus* RB51 and *B. melitensis* 16M for reference *Brucella* genomes.



3.11.3 Supplementary Figure 3: Phylogenomic relationships from 9-mer probe set between *Brucella* field isolates and other known reference genomes. All 262,144 9-mer data points for each of the samples were RMA normalized and $\log_2$ transformed. A Pearson correlation matrix was created by comparing each sample against all other samples. The values were used to generate a taxonomic relationship tree using the PHYLIP software and visualized in the *Treeview* program.


B. abortus strain 19 (33)
F. tularensis LVS
E. coli K12
B. suis biovar 1 (37)
B. suis biovar 1 (34)
B. suis biovar 1 (29)
B. suis biovar 1 (36)
B. abortus biovar 1 strain 2308 (25)
B. abortus biovar 1 strain 2308 (23)
B. suis biovar 1 (40)
B. abortus biovar 1 strain 2308 (24)
B. abortus strain 19 (1)
B. abortus biovar 4 (19)
B. abortus biovar 1 (15)
B. suis biovar 1 (9)
B. suis biovar 1 (16)
B. abortus 2308
B. suis 1330
B. abortus RB51
B. melitensis 16M
B. abortus biovar 1 (21)
B. suis biovar 1(22)
B. suis biovar 1 (31)
B. suis biovar 1 (26)
B. abortus biovar 1 (20)
B. suis biovar 1 (30)
B. abortus strain 2308 (7)
B. suis biovar 1 (8)
B. abortus biovar 2 (2)
B. suis biovar 1 (13)
B. abortus biovar 1 strain 2308 (12)
B. abortus biovar 1 (14)
B. abortus strain 19 (32)
B. suis biovar 1 (35)
B. suis biovar 1 (10)
B. abortus biovar 1 strain 2308 (11)
B. abortus biovar 1 strain 2308 (5)
B. abortus biovar 1 strain 2308 (6)
B. abortus strain 19 (18)
B. suis biovar 1 (17)
B. abortus biovar 4 (3)
B. suis biovar 4 (4)
0.1

Chapter 4

Development of molecular diagnostics using Universal Bio-signature Detection Array technology in host pathogen forensics

Shamira J Shallom[1], Lauren McIver[1], Amanda Rumore[1], Christopher Lawerence[1], L Garry Adams[2], Harold R Garner[1§]

[1]Virginia Bioinformatics Institute, Virginia Tech, Blacksburg, VA, USA;

[2]Department of Veterinary Pathobiology, College of Veterinary Medicine, Texas A & M University, College Station, TX, USA.

[1§]Corresponding author

Harold R. Garner

Director Medical Informatics and Systems

Virginia Bioinformatics Institute

Virginia Tech

Washington Street, MC0477

Blacksburg, VA 24061-0477, USA

Email : garner@vbi.vt.edu

Phone: 540.231.2582

Fax: 540.231.2606

4.1 Abstract

Background

This research describes the development of a broad based platform for identification of pathogens including variant strains from infected animals, environmental or laboratory samples on a high density universal biodefense oligonucleotide microarray. The detection and differentiation platform is an oligonucleotide array that contains all possible (4^9 combinations) 9-mer probes and is genome independent. In addition it has probes specific to bacteria, microsatellites, antibiotic resistance genes and control probes. The Universal Bio-signature Detection Array (UBDA) array is a 370K element 2-color array developed by the Garner laboratory and manufactured by Nimblegen. The development of this technology resulted in the creation an integrated biomarker-specific bio-signature, multiple select agent detection system. Currently, the repository comprises of approximately 100 pathogens and hosts hybridized to this array. The spectrum of organisms chosen for hybridization on this array were primarily bio-threat agents, micro-organisms infecting farm animals, food borne pathogens and organisms of molecular diagnostic importance in a clinical setting. We have expanded this method for viral and parasite detection in insect vectors.

Results

Regression analysis using curve fitting algorithm was used to de-convolute the bacterial organism from a complex mixture of genomic DNA extracted from the soil sample. A soil sample identified precisely the correct pathogen which was *Bacillus anthracis* Sterne strain spiked into the soil genomic DNA. Further this method has applications in detection of *Aspergillus fumigatus* fungal infection in human cells.

Further, a surveillance method using the UBDA array platform was developed. The two test cases were detection of Dengue virus strains in *Aedes aegypti* mosquitoes and parasite genomic DNA from *Leishmania* species were spiked into Sand fly vector genomes. Principal component analysis algorithm was used to differentiate between the Dengue viral strains in the *Aedes aegypti* host background and the *Leishmania species* in the Sand fly (*Phlebotomus papatasi*) and determine the percentage of probes attributed to each of these organisms.

Conclusions

This microarray-based assay will propel classification of new pathogens or mixtures in relation to those that have already been tested on the array. The UBDA has the potential to be fully compatible with micro-machine based front end sample processing and preparation platforms.

Along with a repository of unique hybridization signatures from several genomes of pathogens and their hosts, the UBDA array has the ability to rapidly identify biological threats and newly emerging infectious pathogens that are high priorities in biodefense. Application of the UBDA has the potential to be extended to food and environmental microbial monitoring and a surveillance of insect vectors as carriers of infectious disease.

4.2 Background

The rapid identification of bacteria in clinical sample is important for patient management and antimicrobial therapy. Further food borne illnesses are of prime importance in public health and in biodefense. However, detection of pathogenic microorganisms by traditional methods suffers from several limitations. Even after the detection of growth in cultured blood which usually takes 6 to 12 hours of incubation, conventional blood cultures require at least another 24 to 48 hours for the definitive identification of the pathogen and the assessment of resistance to antibiotics [1].

The existing techniques suffer from several limitations, i.e. low success rate (less than 50%) of cultivation-based assays in samples from patients previously treated with antibiotics; long detection time required for cultivation-based assays; and poor performance of serological and

immunological methods such as high false positive rates, poor reproducibility, lengthy processes and intensive labor. Array based platforms have been designed using the conserved and variable regions of 16s rRNA of specific pathogens [2], 23S rRNA and 16S-23S rRNA intergenic spacer region and gene segments from a group of pathogens [3].

The universal and species specific probes to be spotted on the array are designed based on common and unique sequences from a set of bacterial sequences by comparing sequences within this group and others using multiple sequence alignments and a BLAST search [4]. However most studies on 16S rRNA sequences offer very low sequence diversity, thus difficulties arise in discrimination of phylogenetically close bacteria or subspecies. The probe specificity was affected by the sequence mismatches between the capture probe and the target probe and this affected probe specificity [2]. Further selected pathogenic bacteria were used to construct 16s rRNA oligonucleotide microarray capture probes [5]. However some target species were difficult to discriminate by perfect match analysis due to nonspecific binding of conserved 16S rRNA derived capture probes with high sequence similarity. This group used pattern mapping statistical model using an artificial neural network algorithm trained on known pattern of a hybridization of a training set of organisms. UBDA research described in

this manuscript demonstrates biodiversity studies done on soil borne bacteria and *Bacillus anthracis* using the UBDA array.

Invasive aspergillosis is a major cause of morbidity and mortality in immune compromised and critically ill patients. Standard culture based methods for the diagnosis of *Aspergillus* infections have limited sensitivity and specificity and are time consuming [6]. In patients with pulmonary aspergillosis, cultures of bronchoalveolar lavage fluid are frequently negative [7] and by the time, that positive cultures are obtained, the disease is in its advanced stages [6]. Antibody detection tests in immuno-compromised patients is very limited because of unpredictable humoral responses [8]. BEAS-2B genomic DNA (Immortalized human lung epithelial cells) was spiked with *Aspergillus fumigates* and hybridized on the UBDA array. UBDA research described in this manuscript study demonstrates fungal infection bio-signatures on the UBDA array which could have a potential application in a clinical setting.

The Dengue virus is a member of the virus family *Flaviviridae* and is transmitted to humans through the bite of the mosquitoes *Aedes aegypti*. Dengue virus is now believed to be the most common arthropod-borne disease in the world [9].

In recent years Dengue viruses (serotypes 1, 2, 3 and 4) have spread tropical regions worldwide. Several regions have multiple Dengue virus serotypes are circulating concurrently, which may increase the risk for the more severe form of the disease, Dengue hemorrhagic fever. Current approaches rely on typing of Dengue viruses in clinical specimens (blood) and mosquitoes by Reverse Transcriptase PCR [10]. A rapid and sensitive microarray has been developed using 70-mer and 50-mer oligonucleotides from the most conserved region of several highly pathogenic viruses such as Chikunya virus, Japanese encephalitis virus, Yellow fever virus, Dengue virus, Hanta virus, SARS-CoV an H5N1 avian Influenza virus. Many of these viruses have been considered as potential biological warfare agents. Therefore accurate detection and identification of these pathogens is required for effective control of their transmission. These primers had to be carefully designed to avoid any contamination from the host genome [11]. In general 60-80% global sequence similarity between two sequences can cause substantial cross hybridization [12] . The UBDA array detected the Dengue virus strain spiked into the *Aedes aegypti* mosquitoes.

Leishmaniasis is a worldwide vector-borne zoonotic disease caused by several species of the genus *Leishmania*. By clinical symptoms, the disease is mainly classified into cutaneous and visceral Leishmaniasis. Cutaneous

Leishmaniasis is usually caused by *Leishmania major*, *Leishmania tropica* and other species. Visceral Leishmaniasis is caused by *Leishmania infantum* [13]**.** It is important to distinguish between the species of *Leishmania,* since the treatment is different for cutaneous and visceral Leishmaniasis. The PCR-RFLP (restriction fragment length polymorphism) analysis of the internal transcribed spacer 1 is used to distinguish *L. donovani* and *L. major* from other species [13]**.** Primers specific for the *Leishmania* minicircle [14] have been used in detection of Sand flies infected with *Leishmania*. In addition, 18S rRNA has small interspecies variability and is difficult to distinguish [15][16]. However the species of *Leishmania* cannot be distinguished. Further 7SL RNA gene sequences from *Leishmania*[17] has been successfully used in differentiating *Leishmania* species. *Leishmania* species are highly similar for 7SL RNA gene in the range of 81 to 99.3%. In this study, principal component analysis (PCA) was used in distinguishing *Leishmania* species in Sand fly genomic DNA sample.

4.3 Results

The sensitivity of detection on the UBDA array is estimated between a concentration of 1.5 to 5 ng determined from the spike in of 70-mer oligonucleotides into the human genomic DNA sample [18]. The first application described is the detection of *Bacillus anthracis* in a soil sample.

The second application is detection of a fungal infection in a human genomic DNA background. The third application is development of a surveillance platform for the detection of Dengue viral strains and *Leishmania* parasite species in a host insect vector background.

4.3.1 Use of the UBDA array in direct biodefense application: Detection of *Bacillus anthracis* Sterne strain contamination in a soil sample.

The first example presented in this case study is a biodefense application of detection of *Bacillus anthracis* in a soil sample. Regression analysis shows that the closest match is the *Bacillus anthracis* strain which lacks the pXO2 plasmid (Figure 2 and Table 1). Principal component analysis shows that *Bacillus anthracis* constitutes about 22% of the hybridization intensity compared to 78% from soil genomic DNA (Table 2).

4.3.2 Diagnostic utility in determining genomic signature of a fungal pathogen: *Aspergillus fumigatus* in BEAS B2B human cell line

As described in the methods section a synthetic genomic DNA mixture was created to simulate a fungal infection in a human patient. *Aspergillus fumigatus* 293 genomic DNA was spiked at a concentration of 10%. The closest match was *Candida albicans* which is a fungal sample and *Aspergillus fumigatus* 293 (Table 3). The soil sample is not considered since

it is a highly heterogeneous sample. Table 4 shows the percentage of probes attributed to *Aspergillus fumigatus.*

4.3.3 Surveillance method for vector borne disease

4.3.3.1 Detection of Dengue Virus in *Aedes aegypti* mosquitoes

Principal component analysis was used to distinguish the Dengue viral strain that was spiked into the *Aedes aegypti* host background. This method was also used to quantitate the number of probes attributed to the Dengue viral strain. Table 5 shows that Dengue virus strain 2 was spiked into the mosquito genomic DNA. The amount of spike in was 5% of the total sample hybridized to the array. The number of probes attributed the Dengue virus 2 was 16.9%. The number of probes attributed to Dengue virus stain 4 could be attributed to shared bio-signatures between the two viral strains.

4.3.3.2 Detection of *Leishmania* species in *Phlebotomus papatasi* Sand fly

Hierarchical clustering was used to determine the shows unique signature pattern in the Sand fly *Phlebotomus papatasi* versus the *Leishmania* species Figure 1. *Leishmania donovani* and *Leishmania tropica* species are similar to each other and *Leishmania infantum* and *Leishmania major* are similar in pattern. Principal component analysis was used to determine the number of probes attributed to each of the spiked *Leishmania* species. The UBDA array was able to correctly identify the 10% spiked in *L.*

major and *L. infantum* and the 1% spiked in *L. donovani* and *L. tropica*. The array was able to correctly identify *L. infantum* at the 0.1% spiked in level.

4.4 Discussion

The analysis of closely related strains and species by microarray-based comparative genomics provides a measure of genetic variability within natural populations and identifies crucial differences between pathogen and host.

Bacillus anthracis the microbial agent responsible for the disease anthrax [19]. To exhibit pathogenic characteristics, *B. anthracis* must carry pXO1 and pXO2, two virulence factor encoding plasmids. The Ames strain carries both these plasmids however, the Sterne strain has only the pXO1 plasmid. The size of the pXO1 plasmid is 181,677 bases and the size of the pXO2 plasmid is 94,830 bases. UBDA array can distinguish between these two closely related strains. The level of detection of *Bacillus anthracis* in a soil sample is 5 $x10^5$ cell copy number as determined from [20]. The UBDA array was able to detect the 8 x 10^6 copies of *B. anthracis* Sterne strain spiked into the sample.

The regression analysis shows a high similarity to *Candida albicans* in a human cell line genomic DNA spiked with *Aspergillus fumigatus* since both are fungal organisms. The UBDA array can accurately distinguish

between the Dengue viral strains spiked into the mosquito sample. The type of *Leishmania* species spiked into the Sand fly genome can be detected at the 1% (1.3×10^5) and 0.1 % (1.3×10^4) copies.

4.5 Conclusions

The UBDA array was used to de-convolute the identity of a spiked in pathogen such as *Bacillus anthracis* in a soil sample. Further it can quantitate the amount of pathogen present in the host background as shown with spike in of the *Aspergillus fumigatus* fungal genomic DNA into human genomic DNA sample. UBDA can be used as a potential surveillance detection system for investigating an insect population for viruses or parasites that cause disease in humans.

4.6 Methods

4.6.1 Extraction of genomic DNA from soil

Approximately 1g of sediment in 1x extraction buffer was put into a 50 ml corning tube and mixed. Sediment extraction buffer mix (1ml) was aliquoted into 10, 2ml tubes containing 0.5 g of sterile glass beads. Lysozyme stock 25 µl (20mg/l) (Sigma MO) was added. This was mixed and incubated for 4 hours at 37º C and later 50 µl of 20% SDS was added and mixed. The sample was incubated at room temperature for 20 minutes

and then 50ul of 350 mg/ml DTT was added. This was incubated at room temperature for 20 minutes and mixed for 15 seconds and held on ice in between vortex repetitions. Proteinase K (6.5 µl of stock 20 mg/ml) (Sigma MO) was added, mixed and incubated at 65º C for 30 minutes in a water bath. Liquid from all 10 tubes was pooled into one 50 ml corning tube. Equal volume of phenol, chloroform, isoamyl alcohol (25:24:1) was added, emulsified by mixing. The tubes were centrifuged at 7,500 g for 15 minutes in a large centrifuge. The aqueous layer was transferred to a fresh corning tube. The extraction procedure was repeated. Further, equal volume of chloroform isoamyl alcohol (24:1) was added and repeat extraction procedure by centrifugation was carried out. The aqueous layer was collected and the supernatant was measured. Sodium acetate (0.1 volume, pH 5.2) was added and mixed and one volume of isopropanol was added and was left at room temperature for 30 minutes. The tubes were then centrifuged at 10,000 g for 10 minutes. The supernatant was decanted and the pellet was washed with 1 ml of cold 70% ethanol. The tube was then centrifuged at 10,000 g for one minute. The alcohol wash was repeated and the tube was dried at room temperature. The DNA was dissolved in 1x Tris EDTA buffer (10mM Tris, 1mM EDTA) and stored at -20º C.

4.6.2 Sample preparation of genomic DNA from mosquitoes and Dengue viral cDNA

DNA was extracted from *Aedes aegypti* mosquitoes using Blood and Cell culture DNA mini kit (Qiagen, Valencia CA). Dengue viral RNA was converted to a double stranded cDNA using the SuperScript double stranded cDNA synthesis kit (Life Technologies, Invitrogen, Grand Island, NY).

4.6.3 Microarray procedure and array data processing

DNA concentration (260 nm) and purity (260/280 and 260/230 nm) were assessed by the spectrophotometer and quality by agarose gel electrophoresis. Samples with 260/230 nm ratios greater than 1.8 were used following established protocols for array comparative genomic hybridization (CGH). We designed the UBDA microarray which was then manufactured by Roche-Nimblegen (Madison, WI) as a custom 373K probe chip and genomic DNAs(1 μg) were labeled and hybridized on the UBDA chip as previously described [18]. Data files from the UBDA arrays were imported individually into Nimblescan (Roche Nimblegen, Madison, WI,) and background corrected. A parsing script written in Perl was used to extract 9-mer (262,144 probes and replicates) probe intensities from the 373K UBDA array and signal intensity values were $\log_2$ transformed.

4.6.4 Regression analysis using curve fit

Previously hyper spectral imaging using regression curve fit analysis has been used to discriminate multiple colors in a fluorescent sample labeled with multiple fluorophores [21]. Differentially colored fluorescent calibration standard microspheres were used. Custom code was written in the program IDL and contributions from each individual fluorophore were determined. This has been further applied to determining protein quantification in a dot blot assay [22]. Recently this algorithm has been used to determine the amount of a particular molecular marker present in a tumor sample [23]. A similar custom code was applied to signal intensities generated from the UBDA array. The sum of intensities for all probes was computed for each of the pure samples and this is divided by the total number of samples. The unknown sample is then compared by regression analysis to the library of values generated from the pure reference sample using the curve fit function in IDL code (Boulder, CO).

4.6.5 Quantification of pathogen in a host background using principal component analysis

Principal component analysis (PCA) was employed to determine the isolate's composite identity from the UBDA array data. Principal component analysis [24] was calculated for the entire 9-mer probe set using a custom MATLAB (Natick, MA) script. PCA was calculated in order to determine

the linear fit for a given 2-component data array. The Eigen vectors were used to generate a linear fit. The Euclidean distance measure was used to calculate the orthogonal distance from each data point (for each probe) generated from Cartesian coordinates between the two samples to the linear fit line. The score with the shortest distance was determined for each probe across multiple two by two comparisons between the unknown sample and the reference (pure) sample. These scores were ranked to create the percentage of probes attributed to a given reference or pure sample that are part of the unknown sample.

4.7 Acknowledgements

Soil sample extraction protocol was provided by Dr. Biswarup Mukhopadhay and Dwi Susanti at Virginia Bioinformatics Institute, Virginia Tech. Genomic DNA from *Apsergillus fumigatus* and BEAS 2B cells was kindly provide by Dr. Rumore and Dr. Lawrence at Virginia Bioinformatics Institute, Virginia Tech. Dengue virus RNA and *Aedes aegypti* mosquitoes were provided by Dr. Myles at Virginia Tech. All other genomic DNA listed was obtained from BEI resources or ATCC. *Leishmania* species detection in Sand fly was funded by the Department of the Army to Dr. Sriram Shanker (Lynntech) and Dr. Harold Garner (VBI). *Leishmania donovani* was kindly provided by Dr. Soong from University of Texas Medical branch. This

project was funded by subaward 570636 from DHS 2007-ST-061-000002 from the U.S. Department of Homeland Security - National Center of Excellence for Foreign Animal and Zoonotic Disease Defense at Texas A&M University to Dr. Harold Garner. S. Shallom received funding from SREB (Southern Regional Education Board) state doctoral scholar award.

4.8 Attribution

S. Shallom designed and carried out experiments, analyzed the data, developed principal component analysis algorithm in MATLAB and wrote the manuscript. L. McIver provided computation expertise. A. Rumore extracted fungal and human genomic DNA and participated in study design. C. Lawrence participated in study design. G. Adams provided useful discussions, H. Garner conceived of the study, participated in study design and mentored in drafting the manuscript.

4. 9 Bibliography

1. Beekmann SE, Diekema DJ, Chapin KC, Doern GV: **Effects of rapid detection of bloodstream infections on length of hospitalization and hospital charges**. *J Clin Microbiol* 2003, **41**(7):3119-3125.
2. Eom HS, Hwang BH, Kim DH, Lee IB, Kim YH, Cha HJ: **Multiple detection of food-borne pathogenic bacteria using a novel 16S**

rDNA-based oligonucleotide signature chip. *Biosensors & bioelectronics* 2007, **22**(6):845-853.

3. Palka-Santini M, Cleven BE, Eichinger L, Kronke M, Krut O: **Large scale multiplex PCR improves pathogen detection by DNA microarrays**. *BMC Microbiol* 2009, **9**:1.

4. Yoo SM, Lee SY, Chang KH, Yoo SY, Yoo NC, Keum KC, Yoo WM, Kim JM, Choi JY: **High-throughput identification of clinically important bacterial pathogens using DNA microarray**. *Molecular and cellular probes* 2009, **23**(3-4):171-177.

5. Hwang BH, Cha HJ: **Pattern-mapped multiple detection of 11 pathogenic bacteria using a 16s rDNA-based oligonucleotide microarray**. *Biotechnology and bioengineering* 2010, **106**(2):183-192.

6. Faber J, Moritz N, Henninger N, Zepp F, Knuf M: **Rapid detection of common pathogenic Aspergillus species by a novel real-time PCR approach**. *Mycoses* 2009, **52**(3):228-233.

7. Kahn FW, Jones JM, England DM: **The role of bronchoalveolar lavage in the diagnosis of invasive pulmonary aspergillosis**. *Am J Clin Pathol* 1986, **86**(4):518-523.

8. Young RC, Bennett JE: **Invasive aspergillosis. Absence of detectable antibody response**. *Am Rev Respir Dis* 1971, **104**(5):710-716.

9. Kuniholm MH, Wolfe ND, Huang CY, Mpoudi-Ngole E, Tamoufe U, LeBreton M, Burke DS, Gubler DJ: **Seroprevalence and distribution of Flaviviridae, Togaviridae, and Bunyaviridae arboviral infections in rural Cameroonian adults**. *The American journal of tropical medicine and hygiene* 2006, **74**(6):1078-1083.

10. Harris E, Roberts TG, Smith L, Selle J, Kramer LD, Valle S, Sandoval E, Balmaseda A: **Typing of dengue viruses in clinical specimens and mosquitoes by single-tube multiplex reverse transcriptase PCR**. *J Clin Microbiol* 1998, **36**(9):2634-2639.

11. Xiao-Ping K, Yong-Qiang L, Qing-Ge S, Hong L, Qing-Yu Z, Yin-Hui Y: **Development of a consensus microarray method for identification of some highly pathogenic viruses**. *Journal of medical virology* 2009, **81**(11):1945-1950.

12. Kane MD, Jatkoe TA, Stumpf CR, Lu J, Thomas JD, Madore SJ: **Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays**. *Nucleic Acids Res* 2000, **28**(22):4552-4557.

13. Katakura K: **Molecular epidemiology of leishmaniasis in Asia (focus on cutaneous infections)**. *Current opinion in infectious diseases* 2009, **22**(2):126-130.

14. Kato H, Uezato H, Gomez EA, Terayama Y, Calvopina M, Iwata H, Hashiguchi Y: **Establishment of a mass screening method of sand fly vectors for Leishmania infection by molecular biological methods**. *The American journal of tropical medicine and hygiene* 2007, **77**(2):324-329.

15. Hughes AL, Piontkivska H: **Phylogeny of Trypanosomatidae and Bodonidae (Kinetoplastida) based on 18S rRNA: evidence for paraphyly of Trypanosoma and six other genera**. *Mol Biol Evol* 2003, **20**(4):644-652.

16. Stevens JR, Noyes HA, Schofield CJ, Gibson W: **The molecular evolution of Trypanosomatidae**. *Advances in parasitology* 2001, **48**:1-56.

17. Zelazny AM, Fedorko DP, Li L, Neva FA, Fischer SH: **Evaluation of 7SL RNA gene sequences for the identification of Leishmania spp**. *The American journal of tropical medicine and hygiene* 2005, **72**(4):415-420.

18. Shallom SJ, Weeks JN, Galindo CL, McIver L, Sun Z, McCormick J, Adams LG, Garner HR: **A species independent universal bio-detection microarray for pathogen forensics and phylogenetic classification of unknown microorganisms**. *BMC Microbiol* 2011, **11**:132.

19. Mock M, Fouet A: **Anthrax**. *Annual review of microbiology* 2001, **55**:647-671.

20. Call DR, Borucki MK, Besser TE: **Mixed-genome microarrays reveal multiple serotype and lineage-specific differences among strains of Listeria monocytogenes**. *J Clin Microbiol* 2003, **41**(2):632-639.

21. Schultz RA, Nielsen T, Zavaleta JR, Ruch R, Wyatt R, Garner HR: **Hyperspectral imaging: a novel approach for microscopic analysis**. *Cytometry* 2001, **43**(4):239-247.

22. Rosenblatt KP, Huebschman ML, Garner HR: **Construction and hyperspectral imaging of quantum dot lysate arrays**. *Methods Mol Biol* 2012, **823**:311-324.

23. Uhr JW, Huebschman ML, Frenkel EP, Lane NL, Ashfaq R, Liu H, Rana DR, Cheng L, Lin AT, Hughes GA *et al*: **Molecular profiling of individual tumor cells by hyperspectral microscopic imaging**.

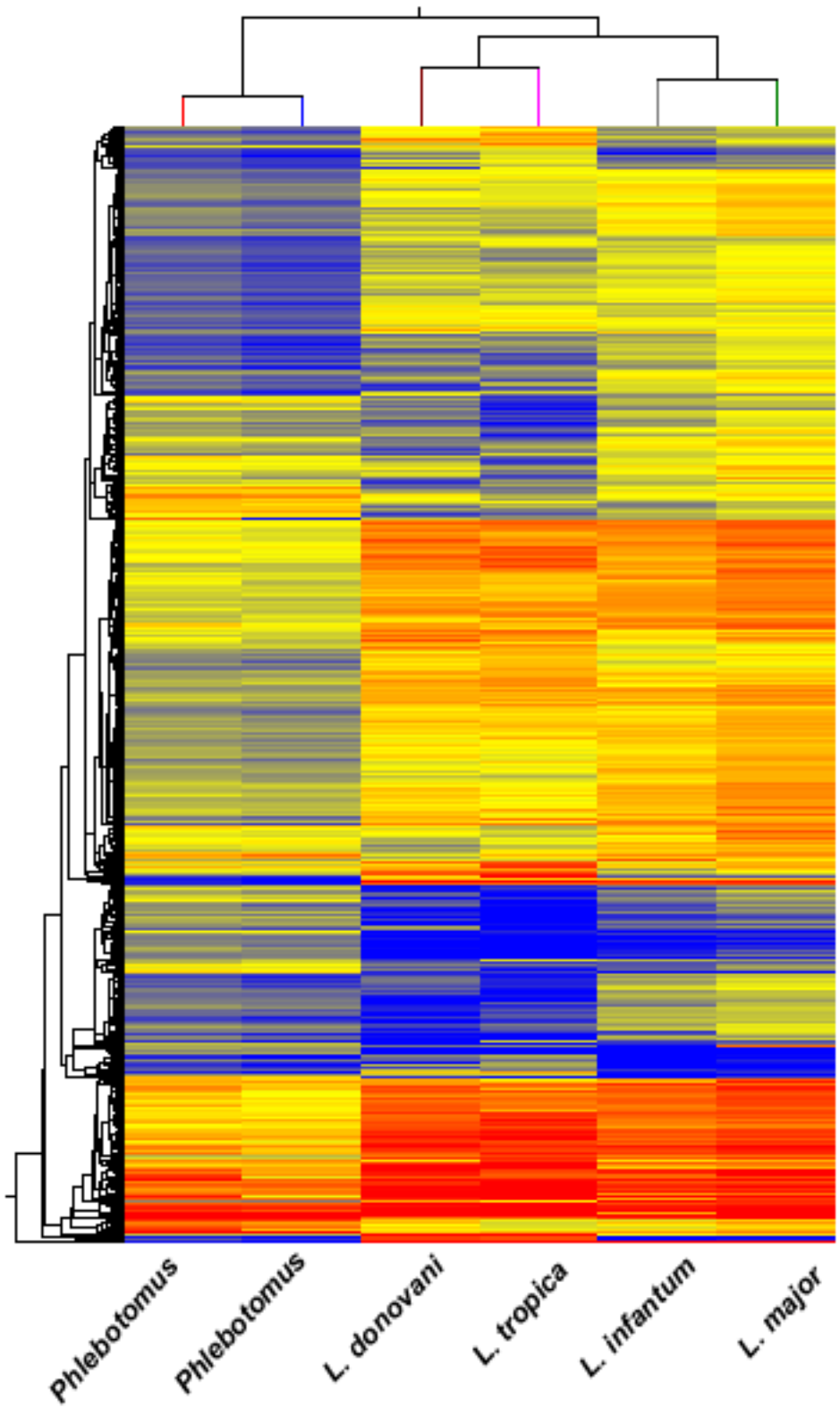
Translational research : the journal of laboratory and clinical medicine 2012, **159**(5):366-375.

24. Raychaudhuri S, Stuart JM, Altman RB: **Principal components analysis to summarize microarray experiments: application to sporulation time series**. *Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing* 2000:455-466.

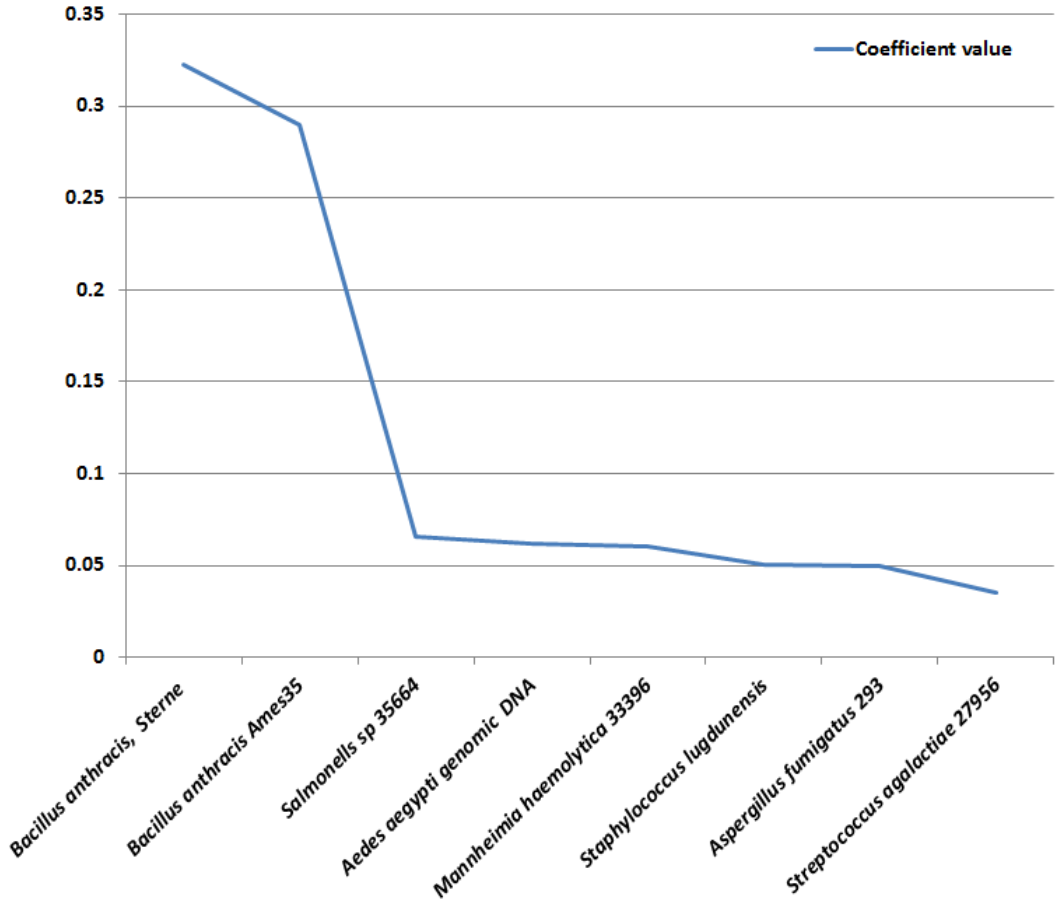
4.10 Figures

4.10.1 Figure 1: Comparison of *Phlebotomus (both channels)* and *Leishmania* species pure bio-signatures

All 9-mer data points for each of the samples were background corrected individually and $\log_2$ transformed. *Phlebotomus* signatures from both dye channels are represented. Only intensity signals with a fold change of 3 or greater with 27,574 elements were included. These elements were subjected to hierarchical clustering algorithm using the Euclidean distance being used as a similarity measure with centroid linkage. The signal intensity values were represented on a Log_2 scale. The range of $\log_2$ values ranged from 8.5 (blue) to 13.5 (red).


Phlebotomus
Phlebotomus
L. donovani
L. tropica
L. infantum
L. major

4.10.2 Figure 2: Regression analysis of soil sample spiked with *Bacillus anthracis* Sterne strain



4.11 Tables

4.11.1 Table 1: Regression analysis of Soil sample spiked with *Bacillus anthracis*

Organism	Soil sample spiked with *Bacillus anthracis*
***Bacillus anthracis*, Strain Sterne ΔGBAA1941**	**0.322813**
Bacillus anthracis Strain Ames35	**0.28986**
Salmonella sp 35664	**0.065335**
Aedes aegypti	0.061832
Mannheimia haemolytica 33396	0.06011
Staphylococcus lugdunensis	0.050337
Aspergillus fumigatus 293	0.04948
Streptococcus agalactiae 27956	0.03502
Streptococcus agalactiae 27956	0.033043
Influenza A Virus (H10N7)	0.022402
Clostridium difficile 193	0.015179
Staphylococcus xylosus 700404	1.34E-05
Clostridium botulinum VPI 4404	1.34E-05
Escherichia coli 25922	7.42E-06
Clostridium alkali	2.30E-06
Thermotoga composti	1.93E-06
Salmonella enterica serovar Braenderup BAA-664	1.70E-06
Rhodococcus equi 6936	8.57E-07
Brucella melitensis 16M	8.49E-07
Methanococcus jannaschii	8.15E-07
Rhodococcus equi 6939	6.83E-07
BEAS B2B human cell line	1.29E-07
Brucella abortus 2308	1.14E-07
Escherichia coli 35218	4.63E-08
Candida albicans 90028	-4.45E-09
Pseudomonas aeruginosa 27853	-1.93E-08
Streptococcus pneumoniae 49619	-7.17E-08
Dengue virus DNV4	-8.56E-08
Clostridium difficile NAP8	-1.25E-06
Influenza A virus A/New Jersey/11/76 (H1N1) Mutant, High (H) Yield	-1.64E-06
Staphylococcus epidermidis 12228	-1.69E-06
soil genomic DNA	**-3.45E-06**
Dengue virus DNV2	-5.21E-06
Brucella suis 1330	-2.67E-05
Brucella abortus RB51	-3.95E-05
Camphylobacter jejuni D3071	-6.12E-05
Clostridium difficile NAP7	-0.00041

4.11.2 Table 2: Quantification of probe signal intensity attributed to the pathogen spiked soil sample. In parenthesis is described the number of probes on the UBDA array attributed to the particular sample

Sample	*Bacillus anthracis* Sterne strain	Soil genomic DNA	Spiked in amount
Soil genomic DNA spiked with *Bacillus anthracis*	22.3% (58,668)	77.6% (203,476)	10%

4.11.3 Table 3: Regression analysis of Human genomic DNA spiked with *Aspergillus fumigatus*

Organism	Human cell line spiked with *Aspergillus fumigatus*
BEAS B2B human cell line	**0.373678**
soil genomic DNA	0.168418
Candida albicans 90028	**0.109529**
Aspergillus fumigatus 293	**0.108016**
Salmonella sp 35664	0.0975944
Clostridium difficile 193	0.057521
Streptococcus agalactiae 27956	0.0358861
Dengue virus DNV2	0.0293556
Aedes aegypti	0.0177858
Clostridium difficile NAP7	0.0129463
Clostridium difficile NAP8	0.00401249
Streptococcus agalactiae 27956	0.00336875
Influenza A virus A/New Jersey/11/76 (H1N1)	0.00174856
Brucella suis 1330	0.000912651
Camphylobacter jejuni D3071	0.000324419
Brucella abortus RB51	0.000242184
Thermotoga composti	0.000238669
Brucella melitensis 16M	0.000213086
Rhodococcus equi 6939	0.000155674
Pseudomonas aeruginosa 27853	0.000139877
Staphylococcus xylosus 700404	0.000105951
Influenza A Virus, A/chicken/Germany/N/49 (H10N7)	5.99E-05
Staphylococcus epidermidis 12228	5.15E-05
Escherichia coli 35218	2.91E-05
Bacillus anthracis, Strain Sterne ΔGBAA1941	-7.30E-09
Salmonella enterica serovar Braenderup BAA-664	-2.18E-05
Methanococcus jannaschii	-2.65E-05
Rhodococcus equi 6936	-4.09E-05
Escherichia coli 25922	-7.36E-05
Clostridium alkali	-7.44E-05
Dengue virus DNV4	-0.000135834
Clostridium botulinum VPI 4404	-0.000277169
Streptococcus pneumoniae 49619	-0.000297945
Bacillus anthracis Strain Ames35	-0.000392961
Staphylococcus lugdunensis	-0.000549802
Mannheimia haemolytica 33396	-0.00101925
Brucella abortus 2308	-0.00147753

4.11.4 Table 4: Quantification of probe signal intensity attributed to the fungal signature in a human host background. In parenthesis is described the number of probes on the UBDA array attributed to the particular sample

Sample	*Aspergillus fumigatus*	BEAS B2B	*Candida albicans*	Spiked amount
Human genomic DNA spiked with *Aspergillus fumigatus*	28.6% (75,059)	35.1% (92,221)	36.1% (94,864)	10%

4.11.5 Table 5: Quantification of probe signal intensity attributed to Dengue virus bio-signature in the *Aedes aegypti* mosquito host.

Sample	Dengue virus 2	Dengue virus 4	*Aedes aegypti*	Spiked in amount
Aedes aegypti genomic DNA spiked with Dengue virus 2	16.9% (44,352)	7.6% (20,100)	75.4% (205,234)	5%

4.11.6 Table 6: Quantification of probe signal intensity bio-signatures attributed to four species of *Leishmania* in a the Sand fly host *Phlebotomus papatasi*

Sample	*Phlebotomus*	*L. major*	*L. infantum*	*L. donovani*	*L. tropica*
10% *L. major* spike	101,563	41,500	38,071	40,669	40,338
10% *L. infantum* spike	94,185	39,948	44,181	43,151	40,679
1% *L. major* spike	108,428	37,067	33,423	42,931	40,296
1% *L. infantum* spike	143,636	26,683	27,986	31,795	32,044
1% *L. donovani* spike	74,700	41,219	35,586	60,062	50,575
1% *L. tropica* spike	131,740	24,347	23,848	34,464	47,744
0.1% *L. major* spike	111,314	33,522	37,109	41,319	38,878
0.1% *L. infantum* spike	98,317	36,694	43,888	41,471	41,772

Chapter 5

5. Outlooks and Perspectives

This research addressed the development of a pipeline for comparative genome analysis and creation of a data repository of bio-signatures specific for organisms under study. The library comprises of over 100 pathogen and host 'patterns' and expands and increases in resolving power as more samples are processed. The array is also very sensitive, for it can use whole genome amplified DNA at or below 10 nanograms, which has been demonstrated with a greater than 0.9 correlation with unamplified samples. Further the taxonomic tree generated using UBDA signal intensities from the mathematically derived genome independent probes was successful at distinguishing between mammalian, bacterial and viral genomes. It has the ability to identify intermediate, variant *Brucella* spp. (*suis* versus *abortus* or mixed) genotypes. It can detect the composition of a mixed bacterial sample and assign percentage scores to a given mixture. This is also applicable to a detection of a pathogen in a host background.

Another potential robust method that can be applied to the UBDA technology is Support Vector machines (SVM), a supervised learning algorithm that is used to solve many classification problems. An SVM algorithm classifies the data by finding the optimal hyper-plane between the

classes of data. The training data that lie on this optimal hyper-plane are called support vectors [21], which has been used to evaluate microarray data sets from different cancer types and normal tissues and other applications.

The development of the web user interface for the Universal Bio-signature Detection array platform is ongoing. The UBDA currently has a website which describes the project and allows users to view and sort through experimental data (http://discovery.vbi.vt.edu/ubda/). This site includes the UBDA array design, which is made available for download. In the future user login into the site will allow access to the UBDA analysis tools. The site was written in Python using the Django framework and runs on an Apache server which is backed up regularly. The custom UBDA analysis tools will allow the user to upload their raw microarray data for clustering computations using custom software written in Perl, MATLAB and IDL following the coding approaches used in many of our other projects (see http://innovation.vbi.vt.edu).

This platform has commercial applications for the development of a cost effective reliable platform for accurate screening of large number of samples for bio-threat agents in forensic analysis, pathogens that routinely infect animals of farm value, food borne pathogens and as a molecular diagnostic of micro-organisms in a clinical environment. This platform is

highly attractive because it has multiplex capacity where knowledge can be drawn from the various probe sets available on the array without prior knowledge of the sample. These probe sets will be translated into a knowledge base repository of bio-signatures that future users of this technology can compare and draw inferences related to the sample under study. In addition a scaled down version of the Universal Bio-signature Detection Assay (UBDA) oligonucleotide microarray will be fabricated and manufactured as a move towards a more cost effective, deployable version of the detection system.

This platform has significant advantages over other approaches. Antibody-based tests, biochemistry or PCR that are usually employed in testing laboratories are serial, and can be confounded by pathogen mixtures or genomic drift, although the cost for each stepwise test is reasonable, on the order of $25 per sample test. An emerging alternative is complete genome sequencing, which we propose to examine in this proposal, which has the ultimate resolution, but at a significant cost in money (~$1-2 thousand per sample) and time (days to a week to acquire and analyze the results). And although the cost per base of 'deep sequencing' has and is continuing to drop rapidly, the cost per sample because of the complex steps (isolation, library preparation, readout, assembly) may take some time, if at

all, before it can drop to the cost of the UBDA array ($100 per sample, now).The current cost for the UBDA array is approximately $350 per sample, which includes reagents and processing costs. The current turnaround time for this forensics technology is less than 24 hours. This is a single experimental procedure compared to other technologies, which involve a series of methods such as serological, biochemical and genomic based. Genome specific arrays are in the similar price range as the UBDA array; however researchers can only assay a single genome or a small subset of them.

At the conclusion of this study the analysis pipeline will be a 'complete' system, making it available to any user. The 'product' will be a finalized array design and analysis environment, including a large, high-resolution bio-signature pattern library, and released versions of both so that anybody can use the UBDA either in-house or via an established service. Depending on the results of our technological, economic and speed analysis, we may further produce an automated or pseudo-automated process for whole genome pathogen assembly and analysis from next generation sequence data such that it rapidly produces actionable data, even for unique pathogens that have drifted, naturally or intentionally. All findings, designs and tools will be released (opened) so that 'customers' can re-create the

entire process within their walls, necessary for certain secure applications, or use our existing established arrays or services. The most likely targeted customer will be USDA/APHIS/Veterinary services and State/Federal Diagnostic labs and also FDA/USDA for the Food Safety Network testing and field surveillance. Additional target customers include the major agro-business companies producing human and animal based food.

Additional Bibliography

1. Pannucci J, Cai H, Pardington PE, Williams E, Okinaka RT, Kuske CR, Cary RB: **Virulence signatures: microarray-based approaches to discovery and analysis**. *Biosens Bioelectron* 2004, **20**(4):706-718.
2. Call DR: **Challenges and opportunities for pathogen detection using DNA microarrays**. *Crit Rev Microbiol* 2005, **31**(2):91-99.
3. Warsen AE, Krug MJ, LaFrentz S, Stanek DR, Loge FJ, Call DR: **Simultaneous discrimination between 15 fish pathogens by using 16S ribosomal DNA PCR and DNA microarrays**. *Appl Environ Microbiol* 2004, **70**(7):4216-4221.
4. Call DR, Brockman FJ, Chandler DP: **Detecting and genotyping Escherichia coli O157:H7 using multiplexed PCR and nucleic acid microarrays**. *Int J Food Microbiol* 2001, **67**(1-2):71-80.
5. Chizhikov V, Wagner M, Ivshina A, Hoshino Y, Kapikian AZ, Chumakov K: **Detection and genotyping of human group A rotaviruses by oligonucleotide microarray hybridization**. *J Clin Microbiol* 2002, **40**(7):2398-2407.
6. Wilson WJ, Strout CL, DeSantis TZ, Stilwell JL, Carrano AV, Andersen GL: **Sequence-specific identification of 18 pathogenic**

microorganisms using microarray technology. *Mol Cell Probes* 2002, **16**(2):119-127.

7. Wang D, Coscoy L, Zylberberg M, Avila PC, Boushey HA, Ganem D, DeRisi JL: **Microarray-based detection and genotyping of viral pathogens**. *Proc Natl Acad Sci U S A* 2002, **99**(24):15687-15692.
8. Pease AC, Solas D, Sullivan EJ, Cronin MT, Holmes CP, Fodor SP: **Light-generated oligonucleotide arrays for rapid DNA sequence analysis**. *Proc Natl Acad Sci U S A* 1994, **91**(11):5022-5026.
9. Royce TE, Rozowsky JS, Gerstein MB: **Toward a universal microarray: prediction of gene expression through nearest-neighbor probe sequence identification**. *Nucleic Acids Res* 2007, **35**(15):e99.
10. Luebke KJ, Balog RP, Mittelman D, Garner HR: **Digital optical chemistry: A novel system for the rapid fabrication of custom oligonucleotide arrays**. *Microfabricated Sensors* 2002, **815**:87-106.
11. Luebke KJ, Balog RP, Garner HR: **Prioritized selection of oligodeoxyribonucleotide probes for efficient hybridization to RNA transcripts**. *Nucleic Acids Research* 2003, **31**(2):750-758.
12. Balog RP, de Souza YE, Tang HM, DeMasellis GM, Gao B, Avila A, Gaban DJ, Mittelman D, Minna JD, Luebke KJ *et al*: **Parallel assessment of CpG methylation by two-color hybridization with oligonucleotide arrays**. *Analytical Biochemistry* 2002, **309**(2):301-310.
13. McGall GH, Fidanza JA: **Photolithographic synthesis of high-density oligonucleotide arrays**. *Methods Mol Biol* 2001, **170**:71-101.
14. Frades I, Matthiesen R: **Overview on techniques in cluster analysis**. *Methods Mol Biol* 2010, **593**:81-107.

15. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns**. *Proc Natl Acad Sci U S A* 1998, **95**(25):14863-14868.
16. Huttenhower C, Flamholz AI, Landis JN, Sahi S, Myers CL, Olszewski KL, Hibbs MA, Siemers NO, Troyanskaya OG, Coller HA: **Nearest Neighbor Networks: clustering expression data based on gene neighborhoods**. *BMC bioinformatics* 2007, **8**:250.
17. Raychaudhuri S, Stuart JM, Altman RB: **Principal components analysis to summarize microarray experiments: application to sporulation time series**. *Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing* 2000:455-466.
18. Schultz RA, Nielsen T, Zavaleta JR, Ruch R, Wyatt R, Garner HR: **Hyperspectral imaging: a novel approach for microscopic analysis**. *Cytometry* 2001, **43**(4):239-247.
19. Rosenblatt KP, Huebschman ML, Garner HR: **Construction and hyperspectral imaging of quantum dot lysate arrays**. *Methods Mol Biol* 2012, **823**:311-324.
20. Shallom SJ, Weeks JN, Galindo CL, McIver L, Sun Z, McCormick J, Adams LG, Garner HR: **A species independent universal bio-detection microarray for pathogen forensics and phylogenetic classification of unknown microorganisms**. *BMC Microbiol* 2011, **11**:132.
21. Zhang C, Li P, Rajendran A, Deng Y, Chen D: **Parallelization of multicategory support vector machines (PMC-SVM) for classifying microarray data**. *BMC bioinformatics* 2006, **7 Suppl 4**:S15.