Biothreat Detection by Random Oligomer-Based Microarray

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

By

James C. Diggans
Bachelor of Science
University of Florida, 1999

Director: Dr. Jennifer Weller, Assoc. Professor
College of Computing and Informatics

Fall Semester 2008
George Mason University
Fairfax, VA

DEDICATION

This is dedicated to my grandfather, Roy Diggans, who would have enjoyed calling me ‘doctor’, and to my grandmother, Betty Diggans, for her patient support and encouragement.

ACKNOWLEDGEMENTS

I would like to express my deep gratitude to a talented group of people at the MITRE Corporation for their hard work and solid counsel as this work progressed. Haley Smith, Felicia Sutton, David Walburger, Olivia Peters, Mojdeh Mohtashemi, John Dileo and Ann Jones have all helped me more than they know.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

ABSTRACT

BIOTHREAT DETECTION BY RANDOM OLIGOMER-BASED MICROARRAY
James C. Diggans, PhD
George Mason University, 2008
Dissertation Director: Dr. Jennifer Weller

Current biosensors are primarily based upon previous observations: they detect organisms known to be pathogenic. Future biowarfare agents, however, are likely to contain completely novel or re-engineered proteins and nucleic acid sequences intended either to make previously harmless organisms pathogenic, to increase the pathenogenicity of existing agents or expressly to render the agent undetectable by conventional serotype- or PCR-based methods. The present work describes the creation and validation of a nucleic-acid microarray-based biosensor for the detection of putative biohazards present in environmental air samples. The prototype array consists of 15,200 pseudo-random 25bp oligonucleotide probes whose sequences were generated using variable-length Markov chain models trained on sequence from pathogenic prokaryotic genomes. Classifiers constructed on organism-specific patterns of hybridization were then applied to unknown or mixed samples to determine a likelihood of detection. With this approach, the ability to estimate the presence of a novel or engineered threat then requires only the

characterization of the binding pattern of the agent’s amplified genomic DNA to the array.

CHAPTER 1: INTRODUCTION

The term ‘weapons of mass destruction’ is usually defined such that it includes nuclear, radiological, chemical and biological weapons. The ability to detect weaponized biological organisms and emerging naturally occurring pathogens is critical to the missions of military, public health and safety bureaucracies. Release of a biological weapon, infectious or not, in an urban center could cause thousands of casualties and overwhelm the normal emergency response system; the tactical use of biological weapons against military targets carries similar consequences[1].

Important factors to consider in such an event include: organism identification, infectious dose of the weaponized organism, rate of occurrence of critical illness in those exposed, incubation time, and time at which a patient is infectious prior to the occurrence of definitive symptoms, as well as the severity and type of symptoms. Unlike chemical and nuclear weapons, some bioweapons are contagious and thus the impact can grow and spread far beyond initial levels of exposure. Ideally the characterization of the agent upon detection would also suggest the best public health measures to take to limit the larger-scale consequences of the event.

Detection technologies must not only be sensitive to particular agents but be specific; even with a small rate of false positive detection, a technology can quickly be rendered useless if it constantly raises false alarms. While the cost of managing false positives can be high both politically and economically, the danger of false negative results can be tragic. The trade-off faced by policy-makers is not unlike that for vaccination for smallpox – a vaccine which, while offering protection from disease, causes a fatal allergic reaction in a small percentage of cases. Until that rate of negative consequences is low enough to be politically acceptable, the positive effects cannot be realized on a wide scale. Striking an appropriate balance between Type I (false positive) and Type II (false negative) error rates poses a significant policy challenge in addition to the technical hurdle of detection. Within the scope of the current effort, characterization of error rates is of primary interest.

Most biosensors can be considered closed systems; that is, they are built to respond to one or a small number of response elements and are unable to respond in the absence of those exact elements (whether the elements change by natural genetic drift or by intentional engineering of antigens[2]). While an effective approach when the threat remains static, this design is not particularly robust or efficient[3], as it requires creation of new sensor capability whenever a novel weaponized organism or emerging infectious disease is discovered. An open system would provide data regardless of whether the particular measurement was expected, thus allowing new events to be recognized, characterized and managed in short order. The current work proposes the design and testing of a nucleic acid-based sensor that makes use of pseudo-random oligomer microarray probes paired with pattern recognition and classification algorithms, to provide likelihood estimates for the identity of putative biological agents within a sample. Using this approach, rapidly enabling a detection capability for a novel engineered- or emerging infectious organism then requires only the characterization of that organism's unique pattern of hybridization to the biosensor.

CHAPTER 2: BACKGROUND

Biosensors require two primary components: a mechanism to detect and capture the biological species of interest in the local environment and a signal transduction method to communicate the occurrence of a detection event in a human-readable way. Much recent work has gone into developing increasingly exotic transduction approaches including microelectromechanical systems[4], surface plasmon resonance[5] and quantum dot phosphors[6]. These enhance the ability to detect rare events. Regardless of the signal transduction approach used, it is often the primary interaction in the detection mechanism that limits the utility of a biosensor. These detection strategies fall largely into one of three approaches: use of antibodies, aptamers or complementarity-based nucleic-acid approaches.

Antibody Capture

Antibodies have a long track record of use in binding assays and offer very high specificity to known targets of interest[7]. Antibody-based sensors are designed to detect particular agents based upon proteins, sugars or other molecules those agents manufacture. However, the epitope, the binding target for a given antibody, must maintain a particular conformation, charge distribution and chemical modification state in order to be detected and, depending upon the assay, must also be located on the external cell membrane.

Small changes to these epitopes, whether naturally-occurring or intentionally engineered, can result in diminished or complete absence of antibody binding. For example, a bad actor may seek to engineer a strain of *Bacillus anthracis* (the causative agent of Anthrax) to lack the surface molecule targeted by antibodies known to be used in deployed sensors[2]. Precedence exists in the natural world for just such an epitope alteration strategy in, e.g., the causative agent of malaria, *Plasmodium falciparum*[8], as well as HIV[9]. If successful, such an engineered alteration would result in a strain of *B. anthracis* against which such sensors would be ‘blind’. Even in theoretical, highly-parallel systems[10], antibody-based approaches still rely on a single, binary (i.e. the target organisms of interest is either bound or unbound) event to detect an organism of interest.

While suffering this single point of failure, antibody-based sensors do, however, show a high degree of sensitivity and specificity. The CANARY biosensor[11] makes use of antibodies expressed by B-cell lines paired with a bioluminescent aequorin-based reporter derived from jellyfish. This strategy was shown to allow detection of *Yersinia pestis* (the causative agent of bubonic plague) down to 20 CFU/ml in less than three minutes with a true positive rate of 67%, ranging to 99% for 200+ CFU/ml. While the CANARY approach offers speed, accuracy and sensitivity, it relies upon a living reagent (the B-cells) thus limiting its potential for long-term field-deployed monitoring and raising requirements for size, weight and power of the device.

Approaches that do not rely upon fluorescent labeling and detection often offer better quantitation[12] of detected species. By functionalizing the surface of a microcantilever with antibodies, recent work has shown that the resultant movement in the cantilever due to antibody binding can be measured using a metal-oxide semiconductor field-effect transistor (MOSFET)[4]. An additional family of label-free methods, surface acoustic wave (SAW) methods, makes use of the deformation caused by a generated acoustic wave on a substrate. If that substrate is functionalized with antibodies and those antibodies bind to an epitope, this binding event changes the nature of the surface deformation, which can then be measured[13].

Many label-free methods utilize reflective and refractive properties of incident light for detection including refractive index (RI) detection, in which a substrate is functionalized with antibodies and, upon binding of those antibodies to a pathogen of interest, a change in the refractive index of the substrate can be measured[12]. Other methods that exploit the impact of a binding event on the physical properties of a detection scaffold include surface plasmon resonance, waveguides, fiber gratings, ring resonators and even photonic crystals[12].

Aptamer Capture

Aptamers are short nucleic acid sequences designed to bind to proteins or other biological molecules of interest due to the secondary structure assumed by the aptamer and the shape, charge distribution and so on of the target molecule. They are usually created via a laborious, iterative process known as SELEX (systematic evolution of ligands by exponential enrichment) by which enormous pools of synthesized, random candidate aptamers sequences are screened against a target ligand of interest to determine binding capacity[14]. Aptamers show great promise for applications in which antibodies are traditionally difficult to create. For example, toxins or prions are harmful or even deadly to animals usually used to raise antibodies; creating aptamers via SELEX obviates this need for live animals. Aptamers, however, still suffer from the same lack of flexibility in the face of changing targets in the organisms of interest as do antibodies, due to their specificity.

Signal transduction applications like those discussed above using antibodies have also been applied to aptamers e.g. surface plasmon resonance[15], quartz crystal acoustic wave sensing[14], and even aptamer-functionalized microcantilevers[16]. In some cases, aptamers can remain stable in ambient conditions longer than can antibodies but the increased overhead required by their evolutionary design and selection often outweighs this advantage.

Nucleic Acid Complementarity

Use of nucleic acids for their complementary binding properties (the predictable binding of A to T and G to C as opposed to their secondary structure used in aptamers) is the third common approach used in binding event-based biodetection. Nucleic-acid based sensors exploit the uniqueness of primary DNA or RNA sequences within organisms of interest[6]. If genomic DNA is targeted then the full, static genome of the organism is investigated resulting in reproducible hybridization regardless of the status of the organism. If mRNA is analyzed, then only those genes dynamically expressed in the host at the time of sampling are queried. This can aid in determining whether an organism is actually actively producing a toxin at the time of capture.

Several fielded biosensors make use of a PCR-based detection method using a small number of oligonucleotide probes designed to match specific regions of the genomic DNA of target organisms[17] including the Autonomous Pathogen Detection System (APDS) from Lawrence Livermore National Labs[18,19], and the JBAIDS system[20], used by the Department of Defense for identification of putative bioweapons in the field. ADPS is a self-contained system capable of measuring the presence of up to 100 agents and controls in a single sample, using PCR to amplify targets from the sample milieu. JBAIDS, the Joint Biological Agent Identification and Diagnostic System, is a real-time PCR-based system, originally developed by Idaho Technology as the R.A.P.I.D. (Ruggedized Advanced Pathogen Identification Device) system. JBAIDS can identify up to 32 samples in one hour from a panel of PCR primers for known biowarfare agents.

Signal transduction methods most often used in nucleic acid complementarity-driven assays include dye fluorescence[21] and molecular beacons[21]. A molecular beacon nucleotide probe has a photon-emitting fluorophore at one end and a photon-absorbing quencher at the other. Left to itself, it will form a stem and loop bringing the fluor into close proximity with the quencher resulting in very little detectable signal. In the presence of complementary sequence, the probe will hybridize, moving the two interacting ends away from one another, as a consequence of which the photon released by the fluor is emitted into the solutions and can be detected.

Fluorescence resonance energy transfer (FRET) has also been applied to label-free nucleic acid-based detection. FRET approaches make use of two separate dye molecules. The donor dye fluoresces in response to a laser and, when in close proximity to an acceptor dye (1 – 10nm), transfers energy to the acceptor dye which fluoresces, usually in a different color from that of the donor dye. One analytical approach[22] designed two separate probes per sequence of interest, one labeled with a donor dye and the other an acceptor. Both were mixed with the sample of interest in a microfluidics channel. If both probes hybridized to the sample, the two fluors were positioned close enough to one another for FRET to take place without any direct fluorescent labeling of the sample itself.

PCR-based approaches, however, all require that probes be designed to be universally specific to a region of the DNA of a particular organism of interest, thus limiting the potential number of organisms detected by any one set of discrete assays. This approach also creates the same weakness as previous antibody- and aptamer-based approaches: a single point of failure, in which no signal is detected when there should be a novel signal indicating the presence of a novel substrate. That is, if the targeted sequence against which PCR primers are designed is changed, the assay will fail to detect the engineered organism.

Microarray-Based Approaches

The current work seeks to move beyond this single point of failure by leveraging cheap, highly-parallel oligonucleotide synthesis capabilities along with a high-density microarray platform. This permits deployment of many thousands of parallel sensors in a format affordable enough for wide-scale deployment. With such a high-density sensing platform, engineering an organism to avoid detection would require a wholesale re-engineering of the pathogen's genome—a feat likely to remain beyond the capability of even the most advanced bioweapons development efforts. Under such an attempted sensor avoidance engineering effort, the very absence of signal from a few probes in the context of several positive signals may provide a warning that further investigation is warranted; i.e. that something very similar to a known pathogen hybridization pattern is present on the sensor and may represent an engineered threat.

A microarray, in its most common form, consists of a glass or silicon substrate onto which nucleotide probes of uniform length are attached in a regular grid. Nucleic acid samples to be analyzed are prepared and labeled with, in most cases, a fluorescent dye. The sample is then applied to the array and allowed time to hybridize to complementary probes[23]. Sample that did not hybridize to any of the probes is then removed in a wash step. A laser is swept over the array and fluorescence resulting from a probe hybridized to labeled sample is recorded by a photomultiplier tube or a charge-coupled device (CCD), akin to those found in commercially-available digital cameras. The resulting image can then be processed for spot location and fluorescent intensity estimations used subsequent analyses. Recent work[24] has even attempted to integrate a CCD directly into the array, removing the need for complex optics in the detection step and increasing fluorescence detection sensitivity. While novel, the cost associated with this approach is likely prohibitive for use in a widely-deployed field biosensor.

Microarray-based detection and identification approaches often consist of a series of probes designed with particular genomes in mind, such that if a probe hybridizes, the analyst can be reasonably sure the organism represented by that probe is present in the original sample. In some cases, multiple probes can be used to create ‘fingerprints’ representative of particular organisms, but this requires a great deal of complex, up-front probe design effort[25]. This approach has been used previously[26-28] to detect viruses; in one example by designing 70-mer probes unique to each of more than 100 viral species[26]. Microarrays with species- or strain-specific probes have also been designed to differentiate between strains of *Staphylococcus aureus*, by generating lists of thermodynamically-favorable probes from regions of sequence unique to particular strains[29]. Additional efforts have also constructed systems for the design of probes specific at the level of individual gene families[30], recognizing that some of these families will be specific for related pathogens.

A further series of microarray-based efforts has focused on detection and differentiation at the species level based upon ribosomal DNA (rDNA) or RNA (rRNA). rRNA is extremely abundant so pre-detection amplification is required although rRNA must be purified away from the protein component of the ribosome[31]. In most cases, rRNA and corresponding rDNA are distinct at the species or even the strain level allowing sets of probes to be designed to detect either rRNA or rDNA at various operational taxonomic unit (OTU) levels – i.e. family, genus, species or strain. Detection of organisms then is based upon which pattern of representative OTU probes hybridized[32].

While these approaches achieve an increase in robustness by using multiple, parallel measurements for each target organism, they still rely upon *a priori* knowledge of agent sequence (which in a novel or heavily engineered agent is unlikely to be the case). They are also limited in the scope of intended detection capability to only those organisms for which the individual arrays have been explicitly designed. Direct analysis of probe hybridization on such arrays may enlarge the universe of detectable organisms. However, the constraints placed on probes generated to match unique regions of sequence in a family of organisms, by definition, limits the capacity for these probes to hybridize to distinct novel or engineered organisms.

While, in a controlled laboratory setting, longer probes tailored to be specific to unique sequences in known genomes provide the specificity necessary for identification, the reality is more complex. As shown in Figure 1, in a sample consisting of a mixture of genomes, longer probes suffer from cross-hybridization to small regions of homology in organisms to which those probes were not designed to hybridize thus complicating sample analysis for complex mixtures. Tiling multiple probes per sequence of interest can help to address this issue but results in a tradeoff by reducing the total number of organisms any one array can be designed to detect. Nonetheless, this particular approach has already shown utility[33] in a clinical setting, detecting the presence of a known but uncommon pathogenic virus in a seriously ill patient.

To field a new detection capability for novel or engineered organisms, genome-specific approaches like those above require design, validation, construction and deployment of a redesigned sensor to enable detection of new or newly significant variant pathogens. This is in addition to the time required to acquire enough of the new species' genome to enable analysis of unique sequence regions and concomitant array probe design. For an array using 70-mers, probe length sacrifices specificity for sensitivity, and may not allow detection of small but important sequence variations of interest.

As an alternative to using specifically designed probes, a microarray was created containing *all possible* hexamers[34] (4^6 or 4,096 probes), and fingerprints for organisms were created based upon the hybridization patterns of mRNA transcripts from pure isolates. While similar in spirit to the pseudo-random probe approach presented here, the use of hexamers limits the universe of probes available and the complexity of resulting hybridization patterns. Use of such short probes also increases the propensity for non-specific hybridization, resulting in a decrease in the signal-to-noise ratio. While potentially useful for identification of organisms in pure culture, this approach certainly would quickly saturate in the presence of multiple genomes from e.g. an environmental air sample.

A further *n*-mer microarray-based pathogen identification system was designed using only 391 9-mer probes, drawn at random from the genome of *E. coli* and screened for thermodynamic suitability[35]. These probes were found, on average, to appear in the genome of *E. coli* 35 times with "nearly equal probability of hybridizing to each strand of the genome."[36] A second group made use of this 'universal' microarray approach to characterize distinct hybridization patterns occurring between ten closely related *Bacillus* species and three non-*Bacillus* species, using a binary bar code approach[37]. This method was later improved upon by using pattern matching and classification techniques to enable differentiation between closely-related strains of *B. anthracis* when using purified genomic DNA samples amplified by REP-PCR (use of repetitive extragenic palindromic consensus primers)[38]. This method too, however, is likely to be quickly overwhelmed by the presence of background genomic DNA in environmental sampling due to the reliance of the classification on only a very few discriminatory features (approximating the single point of failure found in antibody, aptamer and PCR-based assays).

An effort to build a microarray with longer, carefully selected 12- and 13-mer probes from a pool of random sequences resulted in an array comprised of 14,283 probes, each selected to have roughly 50% GC content and to differ by at least 4-5 bases from any other probe in the set[39]. While constrained by the biophysical criteria mentioned, the resulting probe set is otherwise a random sampling of the remaining 12- and 13-mer probe space, without regard to sequences actually present in the target class of organisms. This array was capable of distinguishing between purified genomic DNA from *Bacillus subtilis, Yersinia pestis, Streptococcus pneumonia, Bacillus anthracis* and *Homo sapiens* by a simple comparison of differentially hybridizing probes between pairs of organisms. Comparison of the hybridization intensities from unknown samples to those from a reference database of known hybridization intensities was used to provide putative identity for the unknown.

While similar in spirit to the current effort, this method was not tested against samples comprised of mixtures of purified genomic DNA from multiple organisms. Given the simplicity of the comparison method used to provide identity, mixtures of hybridization patterns would not be unlikely to provide reliable direct similarity to individual known hybridization patterns in the reference database.

Additional purely computational work[40] has analyzed the potential for a 'universal' *n*-mer array by studying the potential for degenerate hybridization to expressed genes in yeast and mice. The authors' mathematical model determined that an array with all possible 13-bp probes (67,108,864 probes) should, given perfectly stringent hybridization conditions, be capable of differentiating the known universe of expressed mRNAs (assuming a particular, limited degree of degenerate hybridization). Little consideration, however, was taken for the effect on the resolution of the proportion of thermodynamically inappropriate probes that could not be legitimately used in an experimental array. As with previously discussed efforts, this array, with relatively few (after pruning for biophysical behavior), short probes, would be overwhelmed quickly by a mix of organisms in an environmental sample, making this approach inappropriate for biological threat monitoring applications.

The array-based efforts discussed so far, while making use of varying degrees of probe density and probe length, all suffer from a similar shortcoming: in the presence of a single genome, the probes on the array are capable of providing detailed information on the unique pattern of hybridization of that genome. However, as a second or third genome is added to the hybridization mix, the amount of additional, novel information provided by a hexamer or nonamer array is limited. With each genome added, fewer and fewer probes fail to light up, reducing the overall information provided by any one probe on the array. In addition, extremely short probes, like those found on hexamer arrays, provide less independent information. When tiling the universe of hexamer probes, the information about genomes present in the sample provided by ACTGTC is not particularly distinct in a helpful way from that provided by ACTGTA. When using longer probes, selection strategies can maximize the orthogonality of the sequences represented, thereby attempting to maximize the information content provided by the array.

Further research into microarray platform optimization has sought to obviate the need to label samples with fluorescent dyes for assays, as this is a time- and reagent-consuming process. One recent approach[41], from Lawrence Berkeley National Lab, utilized charged silica microspheres. By covering a post-hybridization microarray with a layer of these microspheres, it was shown that, for probes hybridized to sample DNA compared to those without a complement, the increase in total negative charge (as single stranded DNA is negatively charged – hybridization occurs via hydrogen bonds) resulted in a repulsion of the negatively charged microspheres. This repulsion occurs to a much lesser degree for probes without hybridized sample. By imaging this array using only dark field microscopy, the investigators successfully related the relative height of the 200-300 microspheres per spot on the array to the level of sample hybridization at that spot (confirmed by standard fluorescence).

Alternative Detection Strategies

In addition to those techniques discussed above, a flood of distinctly novel detection strategies for general biosensing of threat agents has emerged in recent years. These include simple, first-pass approaches like differentiation based upon spore size[42] as well as more complex methods including mass spectrometry[43] and flow cytometry[44]. As the pace of advancement in high-speed genome sequencing increases, future sensors may in fact be able to simply sequence[45] the entire set of microbial flora in the local environment and make sensing decisions based upon this much more complete picture of biological threat agent presence, rather than relying upon a minimal sampling of those genomes using probe-based approaches.

CHAPTER 3: METHODS

In considering design of probes for a microarray-based biosensor, the search space, or set of available probes, for even a twenty-five base pair nucleotide probe is enormous (4^{25} or 1,125,899,906,842,624 unique sequences from which we must select some useful minority). Given that creation of a truly comprehensive array is not achievable, array-based sensors must, by necessity, make trade-offs in selecting probes. Perhaps the simplest probe selection strategy, selection of purely random sequences from a pool of such a size, is far more likely to give rise to uninformative probes than not for any given set of target organisms. This strategy, then, can be viewed as an optimization problem, in which the set of selected probes of a given sequence length, given the constraint of a particular array size, provides optimal discrimination among the universe of organisms (and sequences) to be monitored.

To address this shortcoming, and to maximize the value of probes on the array, popular selection strategies previously mentioned have included design of probes specific to a list of targeted organisms, or design of a series of probes specific to sequences shared within a phylogenetic tree of interest. These strategies, by their very definition, limit the number of species that fall within the detection purview of the resulting platform.

As described above, other efforts have sought to limit the necessity of probe selection entirely by tiling all or most of a probe search space using shorter length probes (in all cases < 13 base pairs). While this strategy allows the resulting array to hybridize to nearly any organism, the lack of relative probe specificity will result in any single organism hybridizing to many, if not most, of the probes. A mixture of genomes in a sample will very quickly exhaust the capacity for specificity of such an array.

In general, for an array intended for application as a flexible, 'universal' human pathogen biosensor, probe sequences must be selected to maximize the amount of information that can be inferred from a positive signal (the presence of a pathogen) and minimize the amount of noise arising from non-specific or competitively hybridizing sequences (non-pathogenic, environmental organisms) in a sample.

Probe Design

The current work makes use of variable-length Markov chains[46] (VLMCs), trained on bacterial sequences of interest, to generate probes. Like standard Markov chain models, VLMCs are trained on a set of sequences from organisms of interest, each containing a specific distribution of individual bases and codon utilization rules. Probes emitted by these models then share much of the sequence-space distributional characteristics of pathogenic genomic sequence. Given the clear distinction in sequence character of prokaryotic genomic DNA from that of random sequence[47,48], this strategy seeks to reduce the total search space for probe selection to a useful subset generally capable of hybridizing to the universe of prokaryotic pathogens and distinct from other genomes present in environmental air samples, e.g. plants, fungi, etc.

VLMCs hold an advantage over standard Markov chains in that they are not constructed with a single ‘order,’ i.e. the number of previous bases the model considers when calculating probabilities for the next base is not fixed but rather depends upon context. This offers increased flexibility, in that many possible VLMC models exist between any two orders of classical Markov models applied to the same training data. For example, within the coding region for a protein, a second order Markov model is likely an appropriate model due to the codon-centric nature of RNA encoding in the genome. However, within intergenic or regulatory regions, lower or higher (respectively – intergenic sequence has few next-base constraints while bases within promoter sequences may be reliably predicted by some higher number of prior bases) order models may fit the local sequence character with greater accuracy. Fitting a second-order model to the entire sequence would likely result in a higher degree of overall model error than would fitting a variable-length model.

We refer to the probes generated by these models as ‘pseudo-random’ in that no top-down design strategy was employed in their creation, yet they are far from random because constraints are applied and the entire space of possible sequences is not covered. In creating pseudo-random oligonucleotides, probes are intended to be capable of binding, to some degree, in aggregate, to any prokaryotic sequence found in a particular environmental sample. To address this challenge, full genomic sequence from a selected group of pathogenic prokaryotes drawn from previous work[49], listed in Table 1, were collected from GenBank and loaded into R using *seqinr*[50].

Table 1: Sequences used to train prokaryotic VLMC models

Species	Pathogenicity	Size	GenBank ID
Bacillus anthracis (Ames strain)	Anthrax	5.2 Mb	NC_003997
Yersinia pestis (CO92)	Bubonic plague	4.7Mb	NC_003143
Francisella tularensis (Schu 4)	Tularemia	1.9 Mb	NC_006570
Brucella suis	Brucellosis	2.1 Mb	NC_004310
Burkholderia mallei	Glanders	3.5 Mb	NC_006348
Burkholderia pseudomallei	Melioidosis	4.1 Mb	NC_006350
Escherichia coli O157 H7 str. Sakai	Hemolytic uremic syndrome	5.5 Mb	NC_002695

While VLMC models can be used to generate oligos of any length, a length of twenty-five bases was selected. Previous work[40] has shown that, for yeast and mouse transcriptomes, theoretically complete coverage can be accomplished by use of all thermodynamically appropriate probes of length ten to sixteen bases. As we wish to address a much larger group of organisms, increasing the oligomer length above sixteen bases is required. However, if longer probes are used, the space of total probes available grows exponentially and our array samples less of the total available probe space. Additional work[51], comparing yeast and human transcripts, showed 20-30bp as a ‘sweet spot’ of sorts (see Figure 1) in that these are long enough to minimize non-specific hybridization yet short enough to present a realistic probe selection search space from which to select probes to tile on an array.

Combined with concerns for commercial availability of arrays in specific probe lengths, 25-mers were chosen to balance exploring the available probe space adequately with addressing a theoretically large universe of organisms.

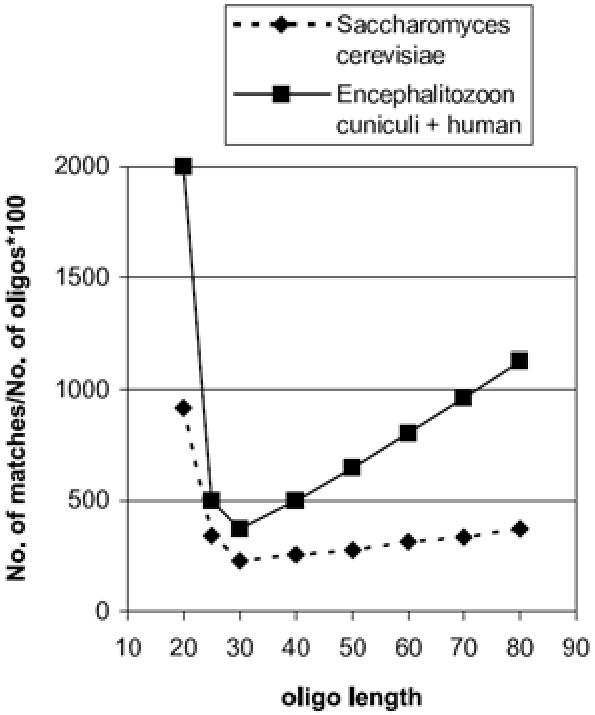


Figure 1: Probe specificity by size. The number of database matches by BLAST by oligo length from CDSs in the respective species against databases of within-species sequence. Source: Rimour et al, 2005.

Training a VLMC requires only a single parameter: the cutoff *K*, a threshold used to determine whether to prune a particular context branch by comparing the deviance between the pre- and post-trimmed trees. Larger values for *K* result in more pruning and smaller, less-complex trees. To determine an optimal value for *K*, a bootstrap-based validation approach, like that described in Mächler et al., was carried out. Five hundred base pair segments were selected at random from each organism in Table 1 and concatenated together to create a training sequence of length *n* = 3,500. This approach assumes that interruptions in sequence character at the borders between individual sequence samples are overwhelmed by the total volume of sequence available in the training set. Under this assumption, the concatenated sequence selection was used to train a VLMC model at each of six initial values for *K*, termed K_0 = 0, 0.5, 1.0, 1.6, 2.0, and 2.6.

Each of these initial, K_0-pruned VLMC models was then used to emit $n + 1$ base pairs (after discarding the first 10,000 base pairs to allow the Markov chain simulation to stabilize). Subsequent VLMC models were then created for values of *K* ranging from 1 to 3 in increments of 0.1 and used to predict the $n + 1^{\text{th}}$ base pair from the K_0 VLMC output. This process was iterated 1,000 times for each value of K_0 and the number of correct predictions recorded. A final value for *K* was chosen to be that K_0 that maximized correct next-base prediction across a range of values for *K*.

Once a value for *K* was selected, a final VLMC model was trained as above and used to generate 100,000 pseudo-random 25-mer oligonucleotides. For each oligomer, melting temperature was calculated using *oligotm* from Primer3[52] and eliminated if not found to be between 58°C and 68°C. Propensity for secondary structure was estimated by calculating ΔG for self-hybridization using *UNAFold*[53]. Probes with internal duplex $\Delta G \leq -1.1$, were removed as previous work has shown this level of self-hybridization to negatively impact probe hybridization[54].

Remaining probes were ranked in order of decreasing ΔG for self-hybridization and the top 12,600 were selected for inclusion on the array. Fifteen percent of the total real estate on the array was tiled in duplicate, resulting in 15,200 total probes on each array, the maximum number of non-control probes allowed on the selected microarray platform and configuration. The probes were then synthesized and placed on substrate by Agilent Technologies on their 8 x 15k Custom DNA Microarray platform. In this form factor, each individual glass slide contains 8 identical sub-arrays per slide and all eight are hybridized concurrently to 8 distinct samples using a gasketed hybridization slip-cover to maintain separation between individual samples.

Reference Library Generation

To enable identification of organisms against a complex environmental background, a reference library of hybridization patterns of purified genomic DNA from known organisms must first be generated. Pathogenic strains are not commonly available and, in fact, are tightly regulated under the United States government's Critical Reagents Program. Accepting this, the technical goal in producing the reference set in the current work was to demonstrate discrimination between simulants (i.e. non-pathogenic organisms very similar to known pathogens) of the same genera (*B. subtilis* and *B. cereus* as within-genera stand-ins for *B. anthracis*) and across genera (using *Pantoea agglomerans* as a gram-negative stand-in for *Yersinia pestis*, the causative agent of bubonic plague, as both are members of the family *Enterobacteriaceae*).

Use of genomic DNA is preferable in this application due to the temporal fluctuation of mRNA: by using genomic DNA, the assay is not dependent on an organism expressing a particular set of genes at the time of sampling. The result of interest is a single hybridization pattern, consistent over time, for each organism in the reference library.

Isolation of genomic DNA from spores is left for future work but is crucial to the utility of microarrays (and any other assay method) in biological threat identification. Past work has shown that spore disruption can be accomplished in a small, deployable form factor[55], so this is not an insuperable problem; the intent of the current work is to demonstrate efficacy of pseudo-random probe-based identification strategy rather than to provide an end-to-end biosensing solution.

While, in a worst case scenario, a biosensor would face very high concentrations of pathogens in the local environment for detection, a fielded sensor must be capable of detection at much lower total pathogen concentrations in environmental air samples. As microarrays normally require a large amount (~1µg) of DNA relative to that recovered from this type of environmental monitoring (~10 - 100 ng), amplification of the genomic DNA prior to hybridization on the array is essential if the sensor is to be reliable.

Several methods are available for unbiased whole genome amplification. These include multiple displacement amplification (MDA), in which random hexamer primers are utilized, together with a highly processive polymerase derived from the bacteriophage φ29. This polymerase, once primed, can incorporate >70,000nt on average[56] before dissociating from the template. It acts by displacing the non-template strand and is capable of starting a new copy from still-elongating copies, resulting in 'hyperbranched' amplification[57] (hence 'multiple displacement'). MDA has been applied to the identification of biodefense-relevant pathogens as an enabling technology for sequencing such pathogens in the overwhelming presence of host DNA[58]. While this method is carried out isothermically (thus obviating the need for a thermocycler in a deployed system), it also requires 6 – 24 hours to generate suitable levels of amplified template (~250 ng), a time frame much too long for effective rapid response to some types of biological threats.

An alternative to multiple displacement amplification, the fragmentation/PCR-based GenomePlex® method (Sigma-Aldrich), carries out a random, non-enzymatic fragmentation of the input genomic DNA followed by ligation of adapter sequences onto each end of the resulting fragments, to construct a fragment library. This library is then amplified via standard PCR, using universal primers matching the ligated adapter sequences. The average length of individual sequences in the resulting library ranges from 200 to 1,000 base pairs. While the standard protocol used here requires 10ng of starting genomic DNA (roughly 1.8×10^6 copies for a 5.2Mb genome like *B. anthracis*), Sigma has demonstrated effective amplification from even single genome copies[59].

Table 2: Reference library experimental design

Genomic DNA	# Arrays	gDNA
B. subtilis gDNA	10	250 ng
B. cereus gDNA	10	250 ng
P. agglomerans gDNA	10	250 ng
B. subtilis / B. cereus gDNA	4	125 ng/species
B. subtilis / E. coli gDNA	4	125 ng/species
B. cereus / E. coli gDNA	4	125 ng/species
B. cereus / B. subtilis / P. agglomerans gDNA	4	84 ng/species
Oligo spike-ins	2	2.5 ng and 25 ng

Purified genomic DNA for reference library simulant organisms was obtained from the Biodefense and Emerging Infections Research Resource Repository (BEI), a Critical Reagents Program provider, and amplified via GenomePlex Whole Genome Amplification (WGA2) kit, as discussed above, using 10 ng of starting material for each genome and yielding 5-10 µg of resulting DNA after amplification and purification in Zeba spin columns (Thermo Scientific). This amplified genomic DNA was then labeled with ULYSIS™ Alexa Fluor® 546 (Invitrogen) and spun down in KREApure purification columns (Bioke) to remove excess dye.

For each experiment in Table 2, 250 ng total of DNA was used in accordance with Agilent Technologies CGH microarray protocols[60]. While the current investigation is not making use of CGH arrays, its use of genomic DNA (instead of cDNA libraries produced via *in vitro* transcription) makes the CGH protocol more appropriate than the single channel mRNA gene expression array protocol. The oligomer spike-in experiment indicated in Table 2 consisted of a set of 20 complements to probes on the array, chosen at random and sent for synthesis by a commercial vendor. These synthesized oligos, known to hybridize perfectly to their matched probes on the array, were then mixed together, labeled and hybridized onto two separate arrays at two levels of total oligomer DNA, 1% and 10% of 250 ng total DNA. These arrays were then used as negative controls in the follow-on classification analysis as no classifier trained on reference simulant genomic DNA should classify these arrays as members of simulant classes.

While the majority of dye was removed in the spin column, enough remained in the hybridization mix (due to the relatively high molecular weight of the dye used) to result in near universal, non-specific dye intercalation by the probes themselves. While Agilent's hybridization protocol does incorporate a blocking reagent, initial experiments using Agilent-recommended blocking reagent levels (4.5 μl) resulted in a high degree of non-specific dye incorporation on array probes (rendering the array useless). An additional aliquot of 11 μl of the blocking agent, KREAblock (Bioke), was added to ensure complete blockage of excess dye.

Prepared samples were then applied to the previously described Agilent 8 x 15k Custom microarray, containing 15,200 pseudo-random oligomer probes, and hybridized for 17 hours at 42°C in the recommended hybridization solution. Array washing was carried out in accordance with standard Agilent protocols[60]. Slides were then scanned with a 532nm laser using a Molecular Devices GenePix Professional 4200A microarray scanner, and signal acquisition and integration was carried out with GenePix 6.0 software, resulting in high-resolution TIFF images.

Image feature data was extracted using Agilent Feature Extraction software v9.5.3.1, without local or global background correction. Local background correction, calculated on a region surrounding each spot, has been shown for the Agilent platform to add noise to resulting data. Agilent protocols suggest, instead, making use of global negative control probes for background correction[61]. The current work does not perform any background correction, primarily due to the reliance of Agilent background methods on spike-in control probes. These probes, while tiled on the custom microarray, are not used in the current work, as they hybridize to RNA targets after in-vitro transcription and cDNA amplification, and are therefore not negative controls in this context. As the

GenomePlex® kit amplifies only genomic DNA, Agilent control probes were ignored in all downstream analysis.

Resulting median spot intensities were then normalized via quantile normalization using the *limma*[62] package in R from the Bioconductor project[63]. The advantage in selection of quantile normalization for creation of a reference database of array data lies in forcing the distributions of underlying probe intensities to be the same across arrays[64] allowing for direct comparison across arrays. While adjustment of probe intensities to quantiles may have a leveling effect on probes in the tails of intensity distributions, this is not of concern in the current application since it seeks only to recognize large-scale patterns of hybridization rather than to study detailed fold changes for individual probes between sample conditions (e.g. mRNA levels between normal vs. diseased tissue states).

Data quality was assessed using the *arrayQualityMetrics* package[65] in Bioconductor to produce MvA plots, to analyze probe density distributions and to explore sample-to-sample relationships using distance metric-based hierarchical clustering and visualization.

In addition to the 30 simulant training arrays and 2 oligo spike-in experiments, a series of mixed-genome arrays was run, to serve as first-order test cases for constructed classifiers. All mixed-genome arrays listed in Table 2 were hybridized with equal ratios of constituent genomic DNA, where the mixture led to a total of 250 ng of total gDNA on each array.

Classifier Construction

Once all training and preliminary test data had been generated, a series of classifiers was constructed using the *CMA* package[66] in R. Each classifier was assessed under an iterative five-fold cross-validation to determine approaches resulting in the most robust reference array class prediction performance. Algorithm families evaluated for use included: linear discriminant analysis (with and without preliminary dimensional reduction by partial least squares), diagonal linear discriminant analysis, shrunken centroids discriminant analysis, support vector machines (making use of a linear kernel), random forests (again with and without preliminary dimensional reduction by partial least squares), penalized logistic regression and a component-wise boosting method.

Due to the limited number of arrays available for analysis under each experimental condition, each classification technique was evaluated by an iterative 5-fold cross validation approach. For each simulant organism, two arrays (n = [N / fold] = 10 / 5) from each class were set aside as test cases, the classifier was trained on the remaining data and then used to classify the holdout arrays. Test samples are then replaced and train/test sets redrawn. This processed was repeated ten times to generate stable performance statistics on classifier success.

Once performance data was collected for classifiers on reference data, classifier instances were then trained on the entire reference set (n = 30) and tested against the mixed genome samples (n = 2 for each of four mixed-genome conditions as in Table 2), to determine whether classifiers trained on pure genomic DNA hybridization patterns could classify unknown samples comprised of mixtures of genomic material, based upon whether those samples contained the target organism.

CHAPTER 4: RESULTS

Selection of *K* for VLMC

When determining a suitable value for *K* by bootstrap analysis, the first question presented is one of sample size: what is the ideal number of bases to sample from each genome to maximize prediction success? A series of analyses was carried out using 50, 500 and 5,000 base pair selections from each of the seven test genomes in Table 1, using K_0 = 0 to prevent passing premature judgment on model pruning. As seen in Figure 2, a sequence selection size of 500 base pairs consistently, though not dramatically, outperformed, on next-base prediction, smaller and larger orders of magnitude in selection size. For this reason, a selection size of 500 base pairs was used throughout the remainder of the analysis. Input and output files and code are included in Appendix A.

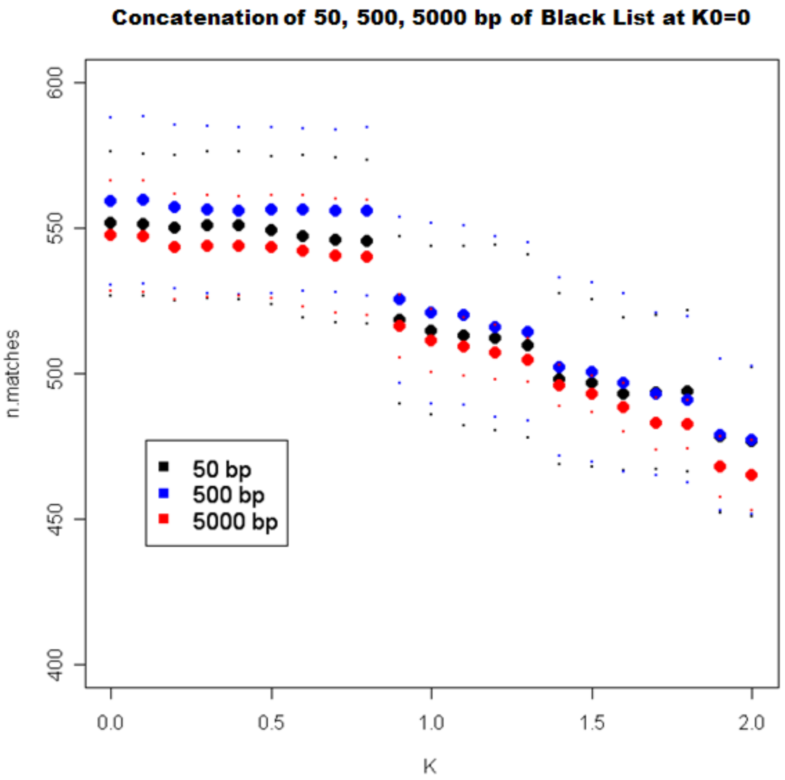


Figure 2: Sample size vs. next-base prediction for K_0 = 0. Bounding dots are 95% confidence intervals on 10 iterations of each sampling strategy.

Once a selection size was chosen, bootstrapping was performed for K_0 = 0, 0.5, 1.0, 1.6, 2.0 and 2.6. For each value of K_0, 1,000 VLMCs were constructed for each value of *K*, ranging from 0 to 3 by a step size of 0.1. Each of these VLMCs predicted the next base generated by the K_0-trained VLMC, and success rates were calculated; the results are shown in Figure 3.

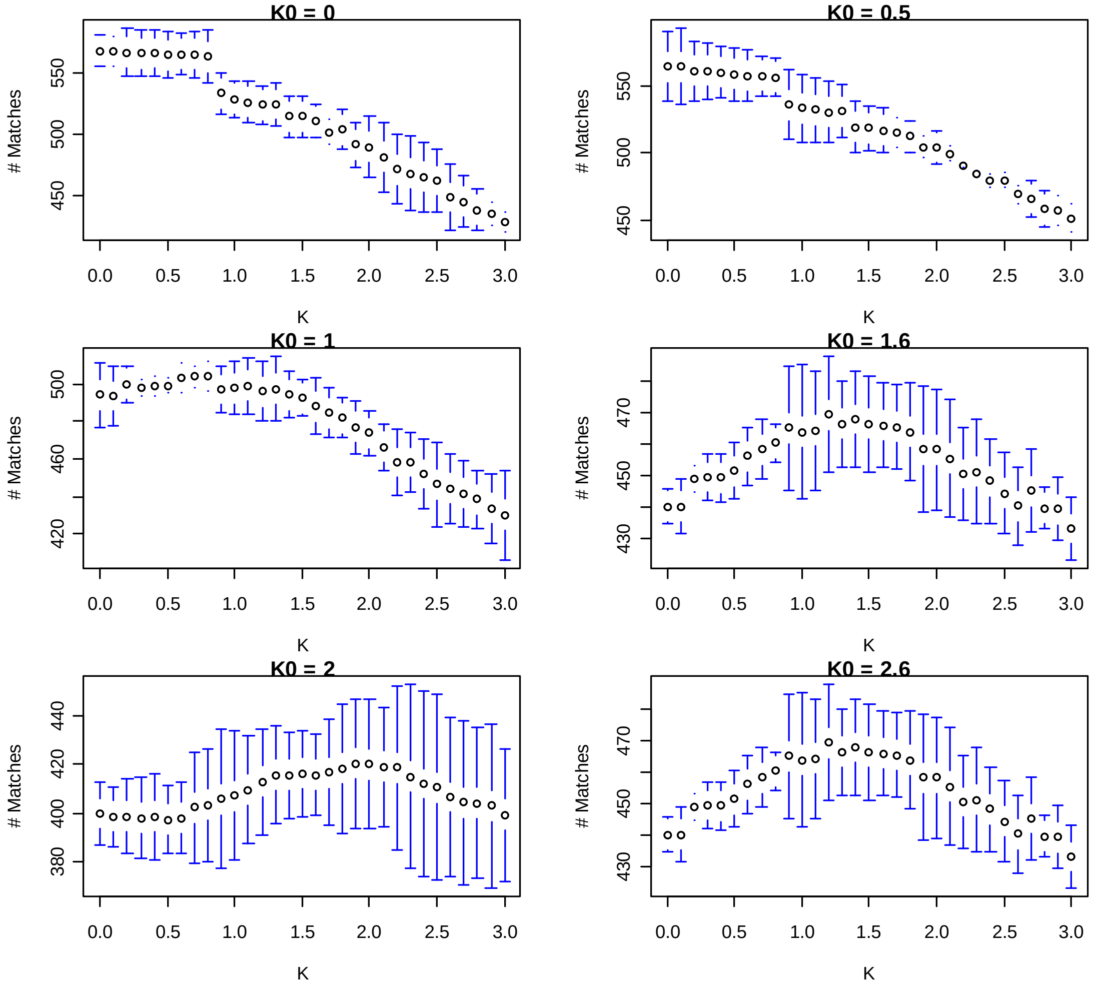


Figure 3: VLMC next-base prediction accuracy over a range of values for K. For each graph, a VLMC is trained on the value shown for K_0. For each value of K along the x-axis, 1,000 VLMCs are constructed and the number of correct next-base predictions (out of 1,000) graphed on the y-axis. 250 correct predictions would be expected by random chance. Error bars represent the 95% confidence interval around next-base prediction accuracy across five distinct iterations.

In general, values for K_0 less than 0.75 consistently resulted in much more coherent trending in the resulting number of correct next-base predictions. Since trending and overall next-base prediction accuracy falls off for higher values of K_0, we can interpret this to mean that higher degrees of model pruning result in a poorer fit, indicating that the appropriate model for pathogenic prokaryotic genomic DNA is quite complex. As any value from $0 \leq K \leq 0.75$ performs at a similar next-base prediction level, 0.75 was selected to provide the most extensively trimmed, parsimonious model while maintaining performance.

By random chance the emitted base would be expected to match 250 times (given an alphabet of four symbols, ATGC, all equally likely) in 1,000 trials. However, prediction accuracy rates during bootstrap analysis were on average 550 or more correct trials out of 1,000, indicating that, more than half of the time, the VLMC model correctly predicted the 3,501st base pair in a simulated sequence.

It is this enrichment of probability over that of random sequence that the current effort seeks to exploit in using such models to emit 25 base pair oligomer probes. To demonstrate the impact of this effect on probe design, VLMCs were trained on random, 500bp selections from the genomes of several pathogens. For each pathogen, a second VLMC was trained on the same total length of purely random sequence (with uniform base utilization).

Both VLMCs were trained at $K = 0.75$ and the two VLMC models were then used to emit 100,000 25 bp oligos which were then aligned to the genomes of the respective organisms via mpiBLAST[67]. Any alignment of at least 16 contiguous base pairs was considered to be a 'hit' and hit rates per 1,000 bp were calculated for both sets of oligomers; the results are shown for each of the organisms in Figure 4. Note that, in almost every case, enrichment in hit rate for the VLMC trained on genomic sequence over that trained on random sequence was at least two-fold.

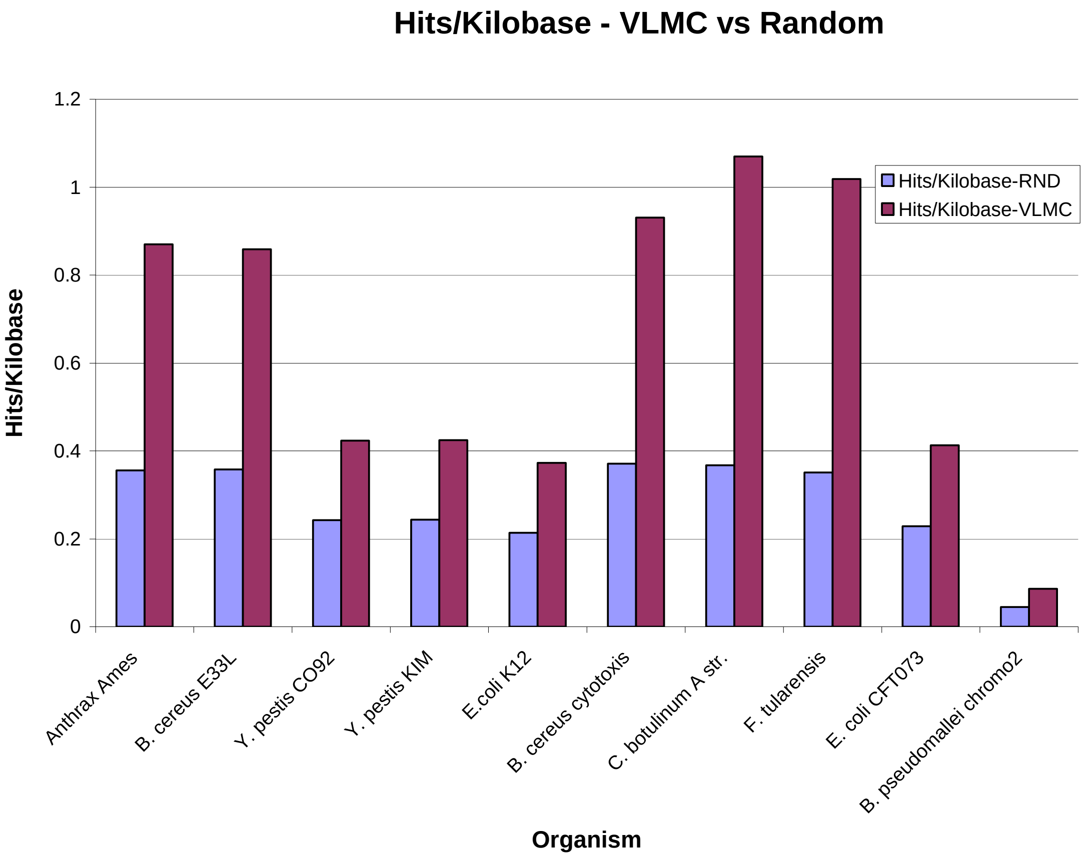


Figure 4: Enrichment in 'hit rate' in genomic vs. random sequences. VLMCs trained on genomic or random DNA. Here 'hit rate' is defined as a contiguous alignment greater than 16 bp to the named target genome.

Also note that the overall 'hit rate' is not consistent across organisms and can differ quite significantly e.g. the genome of *Burkholderia pseudomallei* has a per-kilobase hit rate less than a tenth that of *Francisella tularensis* even though both show relative enrichment between random and VLMC-derived probes.

Table 3: Genome sizes and GC content for genomes presented graphically in Figure 4

Genome	Size	% GC
B. anthracis (Ames)	5.23 Mb	35.37%
B. cereus E33L	5.30 Mb	35.35%
Y. pestis CO92	4.83 Mb	47.64%
Y. pestis KIM	4.95 Mb	47.65%
E. coli K12	4.63 Mb	50.78%
B. cereus cytotoxis	4.09 Mb	36.50%
C. botulinum str A	3.86 Mb	28.20%
F. tularensis	1.90 Mb	32.15%
E. coli CFT073	5.23 Mb	50.47%
B. pseudomallei chr. 2	3.17 Mb	67.97%

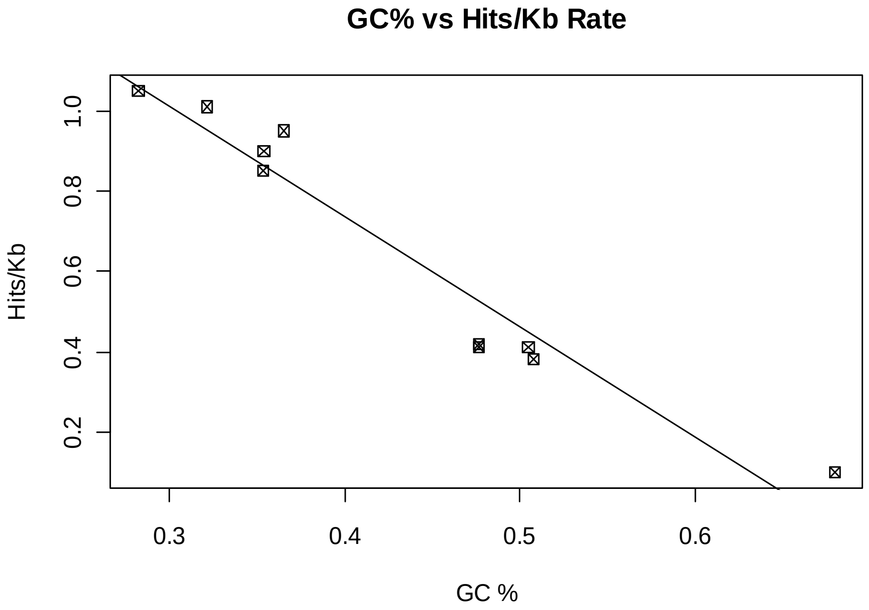


Figure 5: GC% versus hit/Kb rate for the species in Table 3 including calculated linear regression line. Correlation between the two data sets was 0.98.

A linear regression fit to the relationship between GC% and hits-per-kilobase of genomic sequence resulted in an R^2 value of 0.98; this relationship is depicted in Figure 5. The correlation can be explained by the T_m constraints on probes used on the array and concomitant limitation on maximum GC content. As genomes are not subject to such a constraint, the divergence in relative GC content results in the linear relationship shown. Even in the extreme case of *B. pseudomallei*, a hit rate of 1.5 hits/Kb still results in ~4,800 locations in the genome that align to probes on the array with an alignment of at least 16 contiguous base pairs likely leaving plenty of capacity for hybridization to the array and pattern classification.

Array Normalization

MvA plots are a common technique[68] used in the interpretation of bias in array data. This type of plot was originally used to analyze 2-color microarrays, in order to assess dye bias but can be repurposed for use in one-color microarrays by comparing individual probe intensity to the median probe intensity for that array, acting as a stand in for the second dye channel.

The y-axis, M, is defined as:

$$M = \log_2(\text{intensity}) - \log_2(\mathit{median}(\text{intensity}))$$

representing, for a given value of M, roughly the relative fold change in intensity of a given probe compared to the median intensity of the array as a whole.

Table 4: Array IDs for specific reference species arrays

Species/Class	Array ID #s
B. cereus	1,6,7,10,16,18,20,21,28,30
B. subtilis	2,3,8,9,12,13,19,22,26,27
P. agglomerans	4,5,11,14,15,17,23,24,25,29
Oligomer spike-ins	31,32

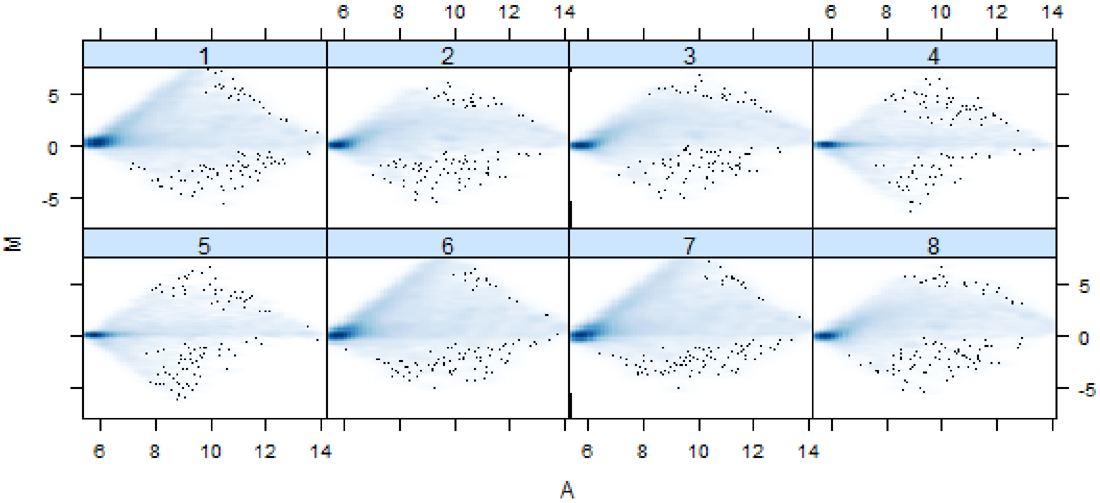


Figure 6: MvA plot of log-transformed, non-normalized data from the first eight reference arrays. M = log_2(intensity) – log_2(median intensity), A = 1/2(log_2(intensity)+log_2(median intensity)))

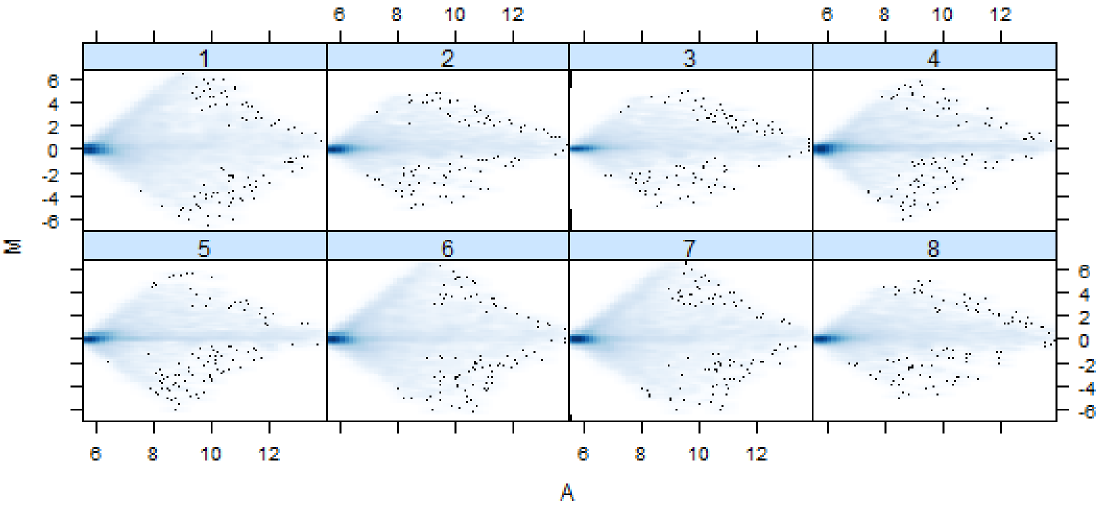


Figure 7: MvA plot of the first eight (see Appendix for all 40 MvA plots) reference arrays after quantile normalization. Notice the improvement over plots in Figure 6 in linearity and symmetry around M=0.

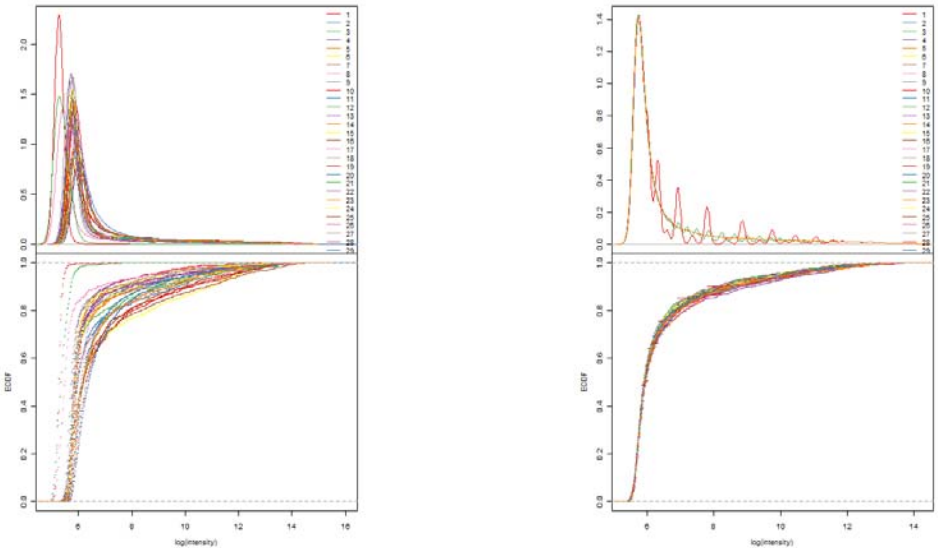


Figure 8: Probability densities (top) and cumulative distribution function plots (bottom) for log2(intensity) across all reference arrays pre- (left) and post-normalization (right). Note the increased overlap in intensity distributions after normalization.

The x-axis, A, is defined as:

$$A = \frac{1}{2}\left(\log_2(\text{intensity}) + \log\ \left(\mathit{median}(\text{intensity})\right)\right)$$

representing, for a given value of A, the intensity of a given probe relative to the median intensity of the array (as the log-transformed geometric mean of the two values). These projections into log space result in a plot along the horizontal M = 0 (as opposed to a plot along y = x in unlogged space) that, in the ideal case, is symmetric and linear along the length of M = 0. Any point density trending away from M = 0 can indicate yet-to-be addressed intensity-dependent bias. Array images, feature extraction results, normalized data and all MvA plots are available in Appendix B.

MvA plots (see Figure 6 and Figure 7) of pre- and post-normalization show a marked improvement in linearity around M = 0 and symmetry when compared to non-normalized, log-transformed data. Density and empirical cumulative distribution function plots (see Figure 8) also demonstrate dramatic improvement in the alignment of distributions between reference arrays due to normalization.

Figure 9 and Figure 10 provide box-plot representations of the distribution of intensity values for individual arrays. Note again that for post-normalization plots the distributions have all been essentially homogenized.

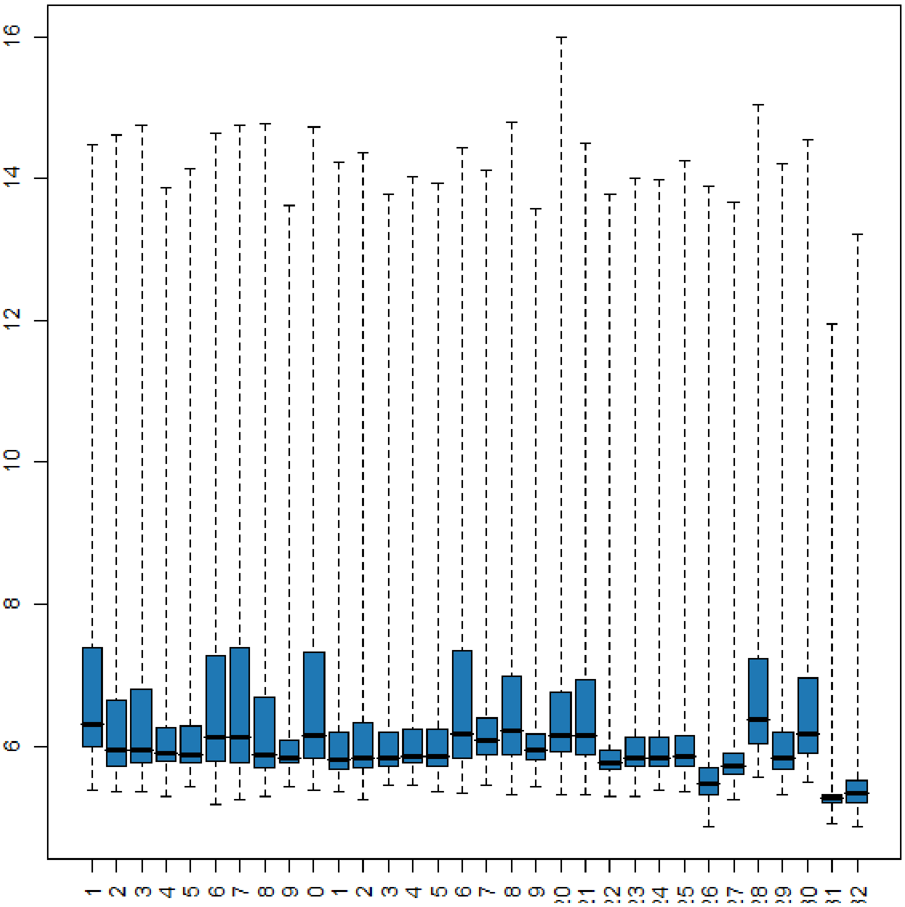


Figure 9: Log-transformed distributions for reference arrays. Samples 1-30 are reference database training arrays. Samples 31 and 32 are oligomer spike-in arrays.

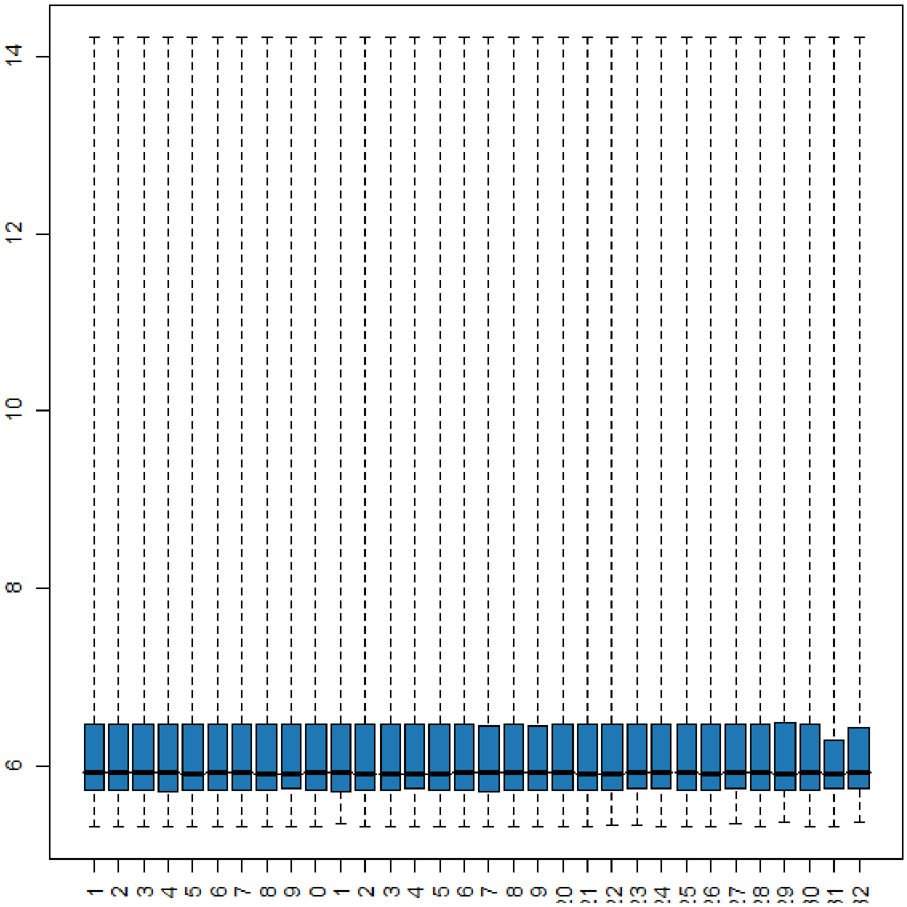


Figure 10: Quantile normalized distributions for reference arrays. Samples 1-30 are reference database training arrays. Samples 31 and 32 are oligomer spike-in arrays.

In preparation for construction of classifiers, and to explore global relationships between samples across all probes, a hierarchical clustering routine was performed using a simple distance metric, $d_{xy} = \text{median}|M_{xi} - M_{yi}|$, defined as the median difference between matched probes across any two arrays x and y. The resulting distance matrix was then used in *arrayQualityMetrics* to construct a heatmap with matched dendrograms (see Figure 11 and Figure 12 for pre- and post-normalization result graphs).

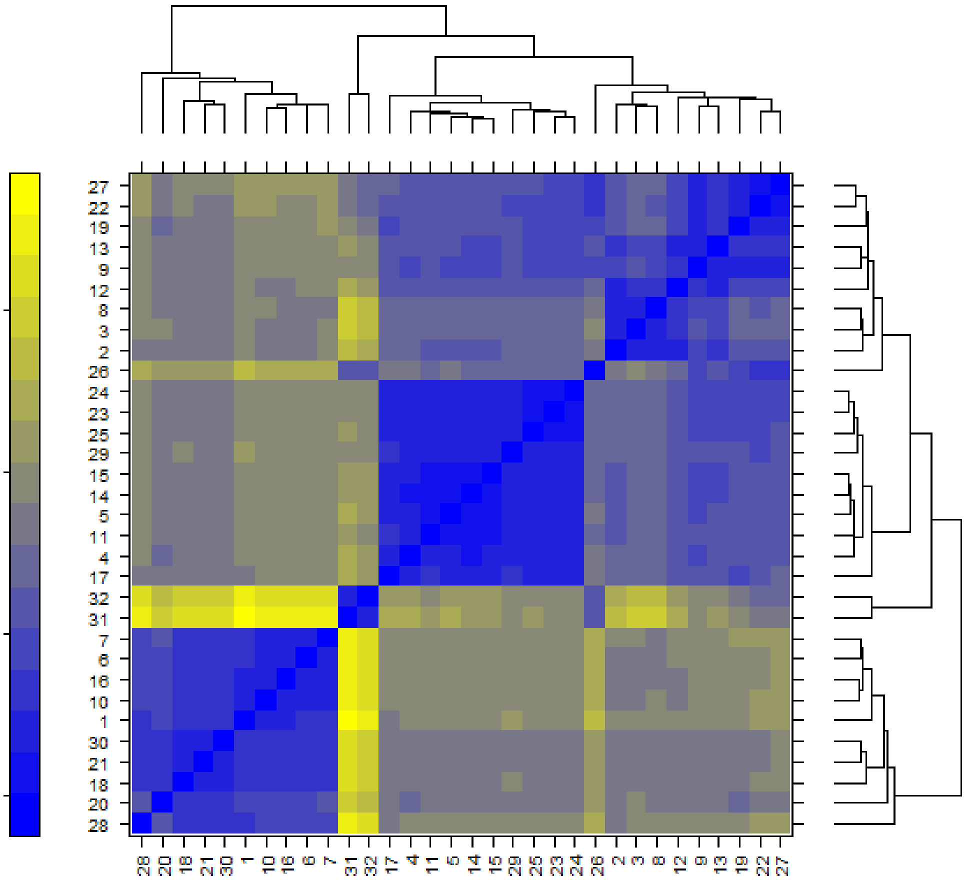


Figure 11: Sample/sample comparative heatmap and accompanying dendrogram for pre-normalization, log$_2$ transformed intensity data. Visualization was constructed from a distance matrix as d_{xy} = median$|M_{xi} – M_{yi}|$. Classes for sample IDs are found in Table 4.

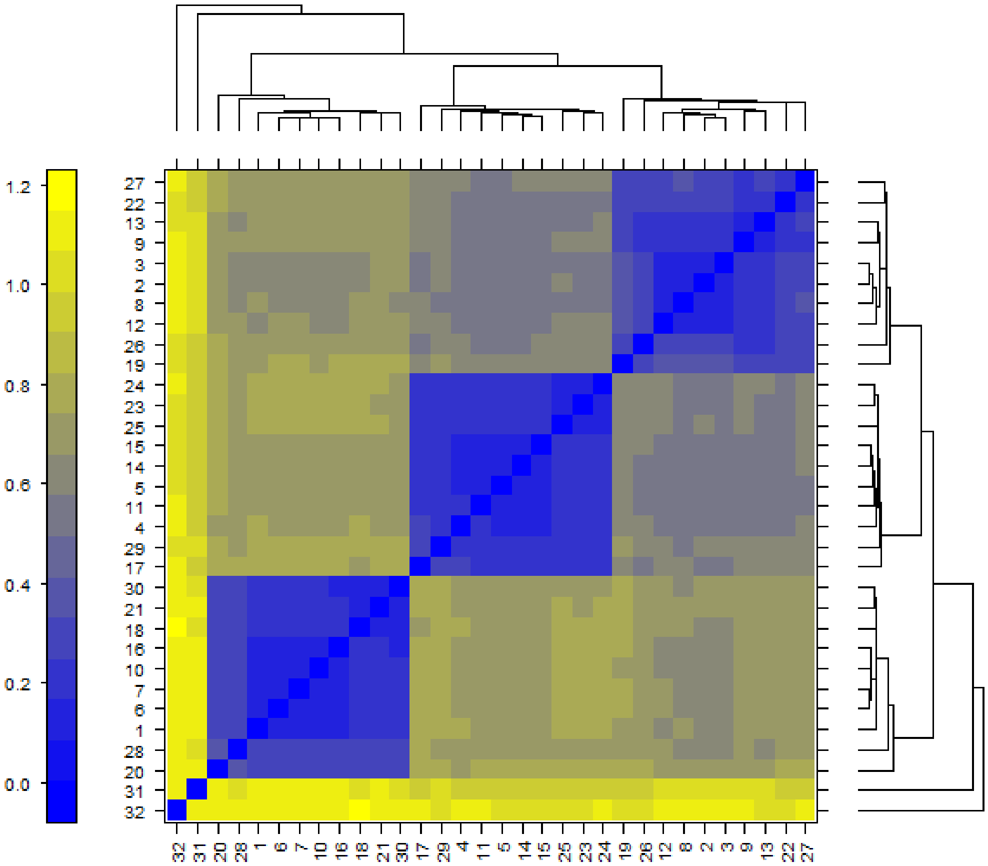


Figure 12: Sample/sample comparative heatmap and accompanying dendrogram for post-normalization intensity data. Visualization was constructed from a distance matrix as d_{xy} = median$|M_{xi} - M_{yi}|$. Classes for sample IDs are found in Table 4.

Note that in the pre-normalization visualization, only all ten *B. cereus* samples cluster together in an isolated way. All ten *P. agglomerans* samples cluster together but did so within the scattered *B. subtilis* samples. In the post-normalization visualization results, class members remain within clearly-defined class cluster boundaries and the two oligomer spike-in arrays are clearly identified as significantly distinct from the simulant genome arrays. The dendrogram relationships parallel the visual representation of class membership in the heatmap.

Of particular interest is the separability of samples even within closely-related species at this high level of analysis. Even without application of sophisticated clustering algorithms, the sample classes are clearly distinct from one another.

Reproducibility of Whole Genome Amplification

In building a reference database, a key assumption lies in the comparability of data in the reference database to that being compared against the reference data. As a source of potential variability, the whole genome amplification method used is of concern due to the random nature of the non-enzymatic digestion of precursor DNA.

To explore this potential source of noise in the resulting system, four amplifications were carried out, all starting with 10 ng of *B. subtilis* genomic DNA. Each amplification reaction was run in duplicate, following the same labeling and array protocol as previously discussed. Spot intensities were estimated as described above, and Pearson's correlation coefficients were calculated for all sample pairs (eight choose two, or 28 total correlations) and a distribution of resulting correlations was calculated. This distribution was compared with distributions of pair-wise correlations (ten choose two, or 45 total correlations) calculated for each simulant in the reference database and graphed in Figure 13. The scanned array image, feature extraction results, normalized intensity matrix and all calculated correlation coefficients are available in Appendix C.

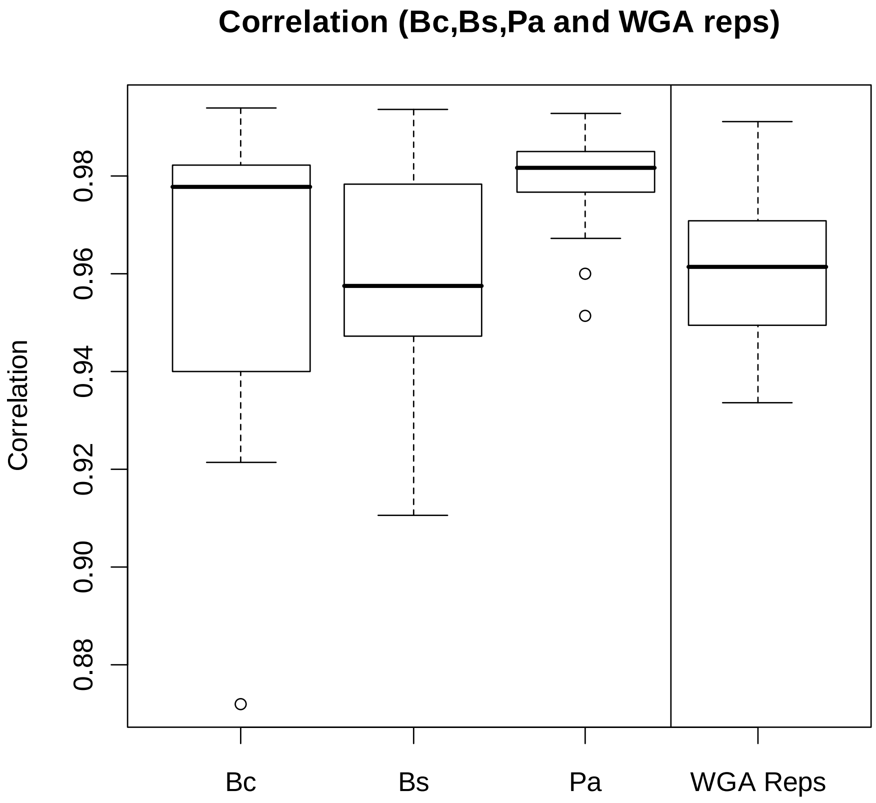
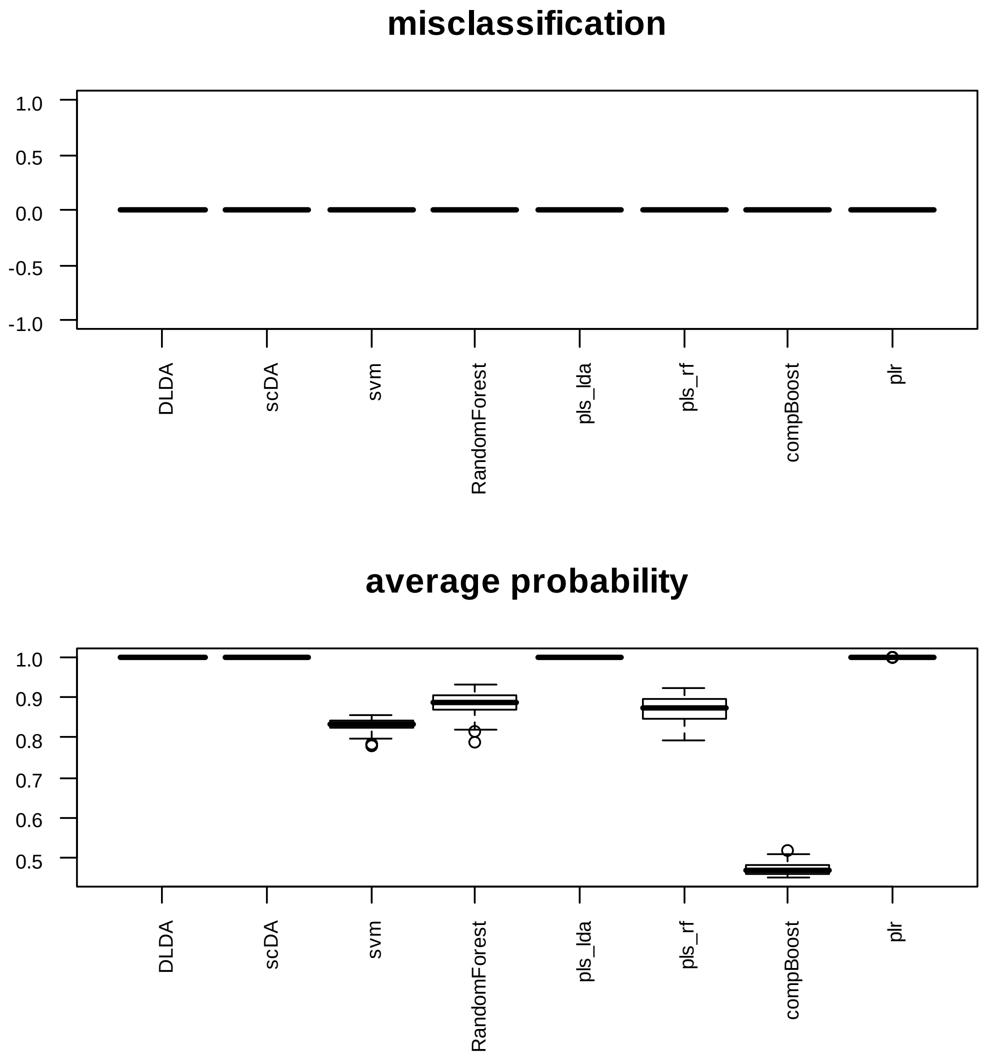


Figure 13 - Distributions of correlations calculated on pair-wise samples from three simulant classes in the reference database (n = 10 each) and for the WGA reproducibility study (n = 8).

The distribution of pair-wise correlations for the WGA study samples falls well within that for the correlations for the reference simulants. As the reference database samples were each derived from a pool of 2-3 amplifications (depending upon total yield for individual simulants), the degree of sample similarity across both individual WGA and pooled WGA samples provides ample reassurance that the WGA process itself is not a source of significant error in the classification process and that, while randomly fragmenting sample DNA, the stochastic nature of this fragmentation still leads to consistent hybridization patterning.

Reference Classification

As previously mentioned, classification was carried out using *CMA*, an R package wrapping various popular classifier implementations, cross-validation strategies, and methods for evaluation of classifier robustness. An initial broad survey of classifier performance on the post-normalization intensities from the reference simulant data was carried out to determine which families of methods to select for additional analysis by classification of mixed-genome samples. Results of this survey are depicted in Figure 14.



misclassification

average probability

Figure 14: Cross-classifier comparison of predictive accuracy and class assignment probability across eight classifiers. From left to right, classifiers tested are diagonal linear discriminant analysis, shrunken centroids discriminant analysis, support vector machines, random forests, linear discriminant analysis after dimensional reduction via partial least squares, random forests after dimensional reduction with partial least squares, component-wise boosting and penalized logistic regression.

Two measures of performance are indicated in Figure 14: misclassification and the average classification probability. Misclassification graphs the percentage of samples classified as a class other than that sample's true class. In this case, no samples were misclassified by any of the classification methods. While encouraging, this gives no measure of how overfit the various models may be to the reference training data, which represents a best-case detection scenario (a lone pathogen genome, in a vacuum with respect to host DNA or other sources of competition) rather than the real-world complexity within which any fielded identification method must work.

The average probability graph shows, for those methods that provide more detail than a simple 1/0 classification score, the distribution of scores representing how 'sure' each method was of the final classification of each sample calculated as[66]:

$$n^{-1}\sum_{i=1}^{n}\sum_{k=0}^{K-1} I(y_i \in k)\hat{P}(y_i \in k|x)$$

where n is the number of samples, K is the total number of classes, $I(y_i \in k)$ is an indicator function as 1/0 depending upon whether the assigned class is correct and $\hat{P}(y_i \in k|x)$ is the probability that sample y is of class k, conditional on x, the vector of array intensity data for sample y.

Ideal methods will show a high probability of classification for true positive results. Beyond the LDA-based methods (which do not provide probabilities of classification beyond a 1/0 indicator of predicted class), only the competitive boosting method showed relatively low confidence in predicted sample classes, with a median probability of classification of only 0.45.

Drilling down further into predictions on individual samples, voting plots were constructed for each classifier (see Figure 15 and Figure 16). In each plot, the 100 observations on samples from each class (5-fold, 2 samples per test group, iterated 10 times for 100 total observations per class) are grouped horizontally and divided by vertical lines while scores from each classifier for each predicted class are color-coded. For example, in Figure 15 for DLDA, the first 100 dots represent predictions on the 10 *B. cereus* samples, the second 100 dots on *B. subtilis* and the final 100 dots on *P. agglomerans*.

The y-axis represents the probability associated by the classifier with membership of that sample in the predicted class. For LDA-based methods, this is a 0/1 value but for methods like SVM, penalized logistic regression, random forests or component-wise boosting, more information on the separation between predicted classes is provided. For example, in Figure 16 for the random forests classifier, there is some degree of variability in prediction probability from sample to sample; although at no time does any predicted score from any other class approach that of the correctly predicted class.

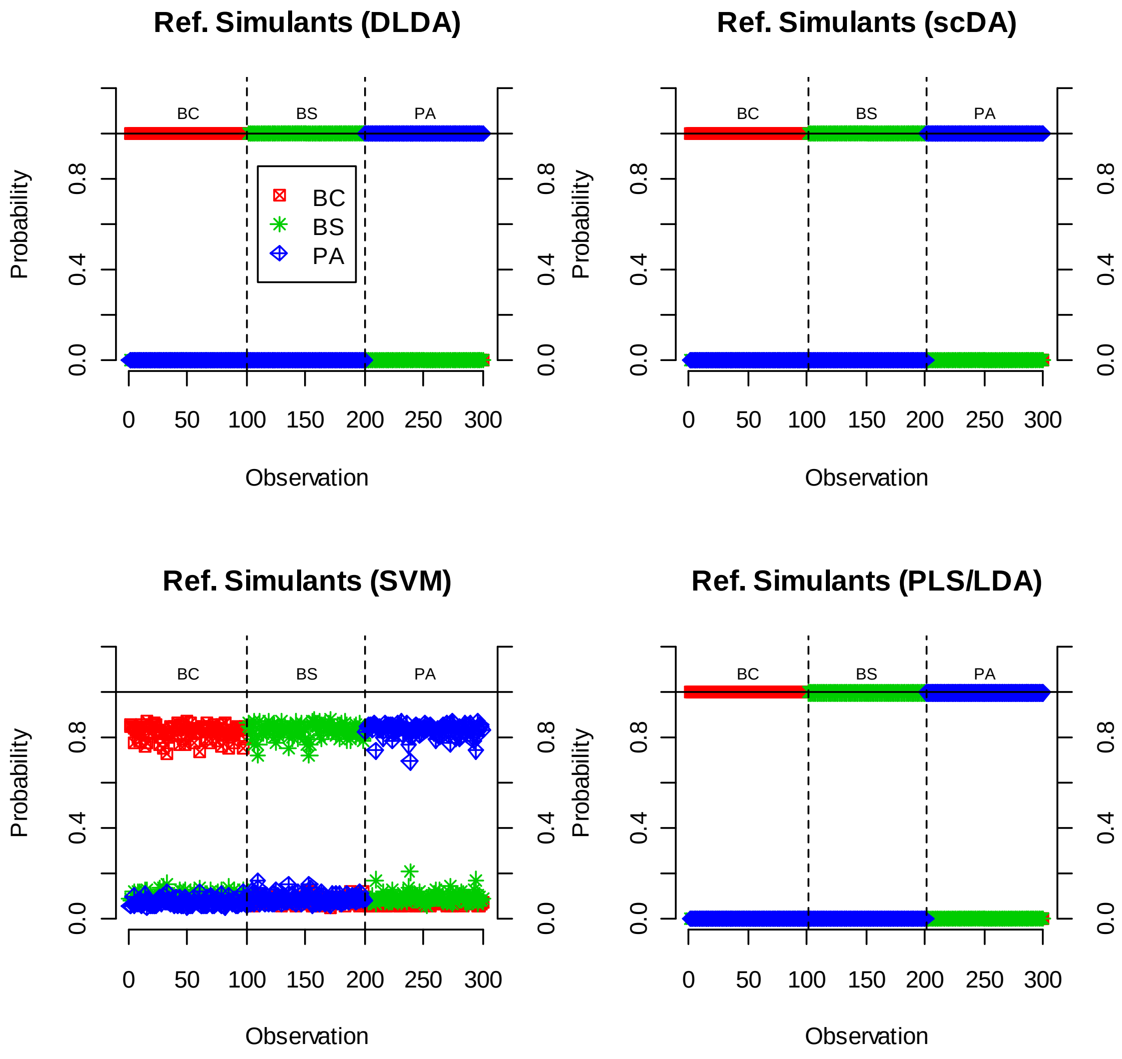


Figure 15: Voting plots across all class/cross-validation observations for the first four classifiers. Classifiers shown are diagonal linear discriminant analysis, shrunken centroids discriminant analysis, support vector machines and linear discriminant analysis after dimensional reduction with partial least squares. For each class, 100 total classifications (5-fold cross-validation x two test samples per fold x ten iterations) were made for each species. Observations are divided by true species horizontally with class-specific voting probabilities plotting on the y-axis color-coded by predicted species.

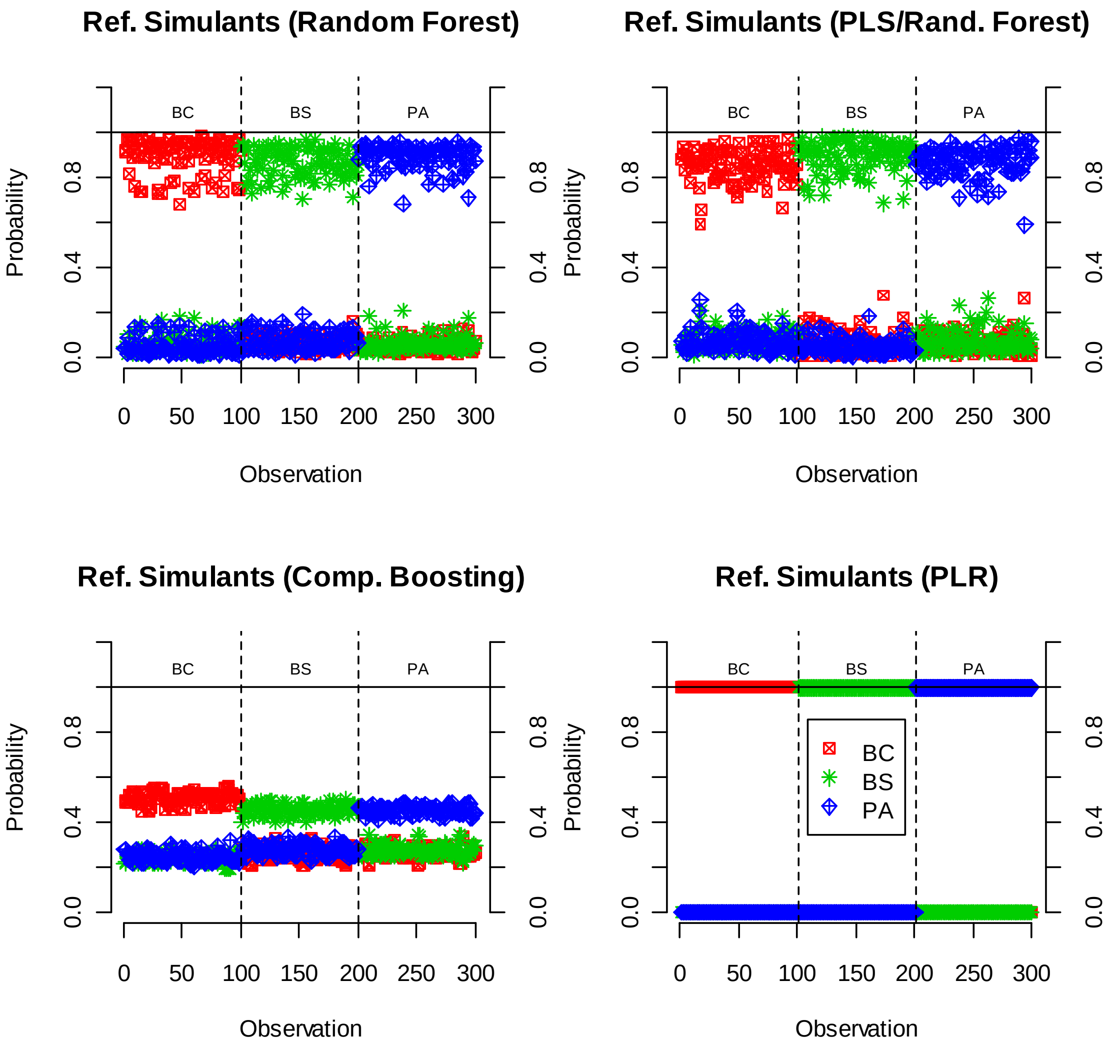


Figure 16: Voting plots across all class/cross-validation observations for the second four classifiers. Classifiers shown are random forests, random forests after dimensional reduction with partial least squares, component-wise boosting and penalized logistic regression.

For component-wise boosting, on the other hand, the distance in probability space between the correct class and incorrect classes in some samples comes to less than 0.1

suggesting that component-wise boosting is not an appropriate method for the current application.

Mixed-Genome Classification

Correct classification of genomic DNA in isolation is certainly of interest but of no practical application for identification of bioweapons from environmental samples. As a next step, the same panel of classifier methods tested previously was trained on the entirety of the reference data set and the resulting classifiers used to predict the classification of two- and three-way mixed-genome samples, as described in Table 2. *Escherichia coli* (K12) was used as a possibly confounding background genome not represented in the classifier training set.

For each mixed genome condition, arrays were subjected to the same panel of classifier methods, this time trained on all reference samples. In Figure 17and Figure 18, results from the eight classifiers are shown. In almost all cases (component boosting again the performance exception), the presence of *B. cereus* in a background of *E. coli* was correctly classified. The presence of *B. subtilis* in a background of *E. coli* was also correctly classified by all but two of the classifier methods (component boosting and random forests on dimensionally reduced data).

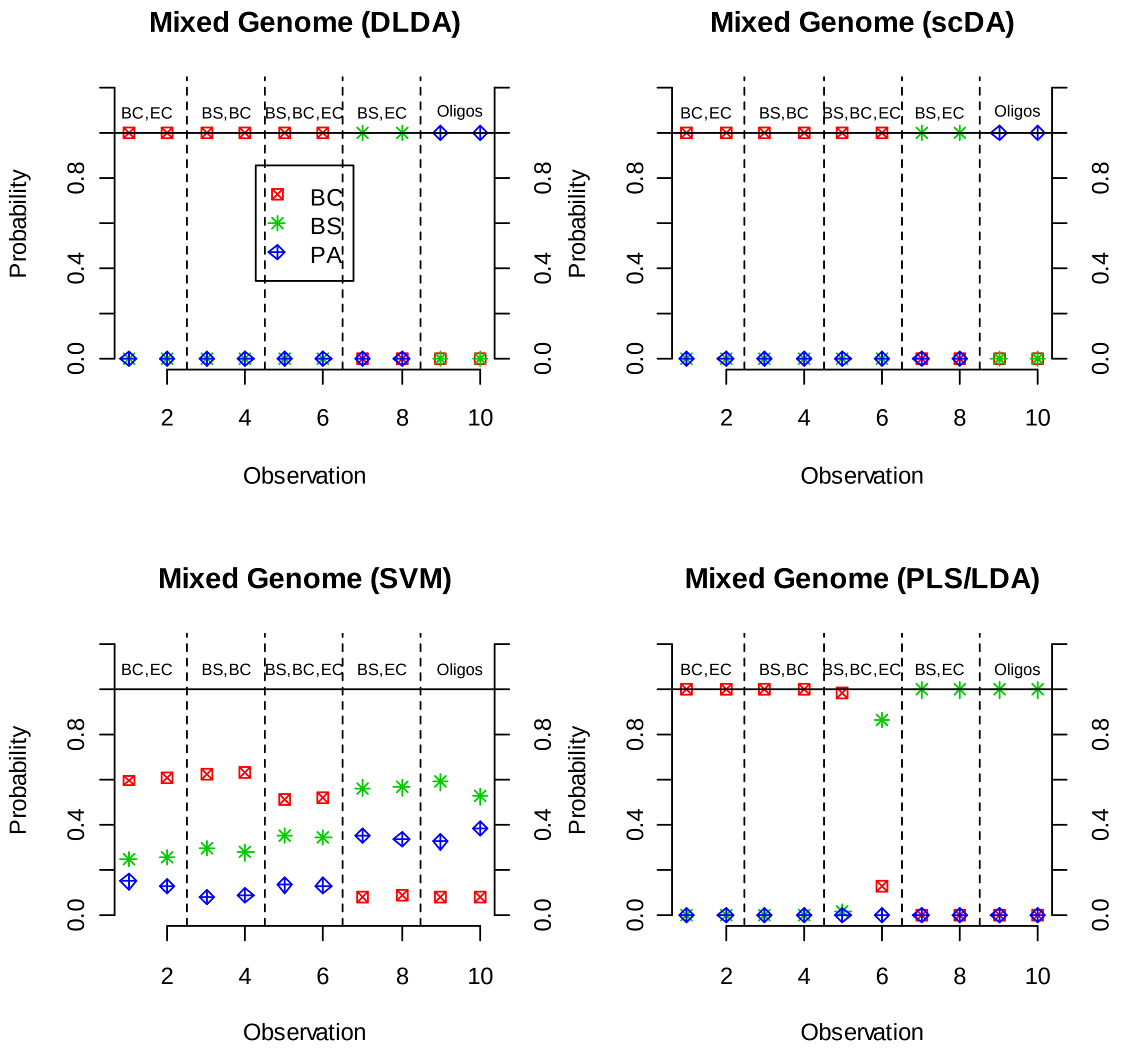


Figure 17: Voting plots across mixed-genome samples for the first four classifiers. Classifiers shown are diagonal linear discriminant analysis, shrunken centroids discriminant analysis, support vector machines and linear discriminant analysis after dimensional reduction with partial least squares. Two samples were available for each mixed-genome condition. Observations are divided and labeled by genomes hybridized horizontally with class-specific voting probabilities plotted on the y-axis color-coded by predicted species.

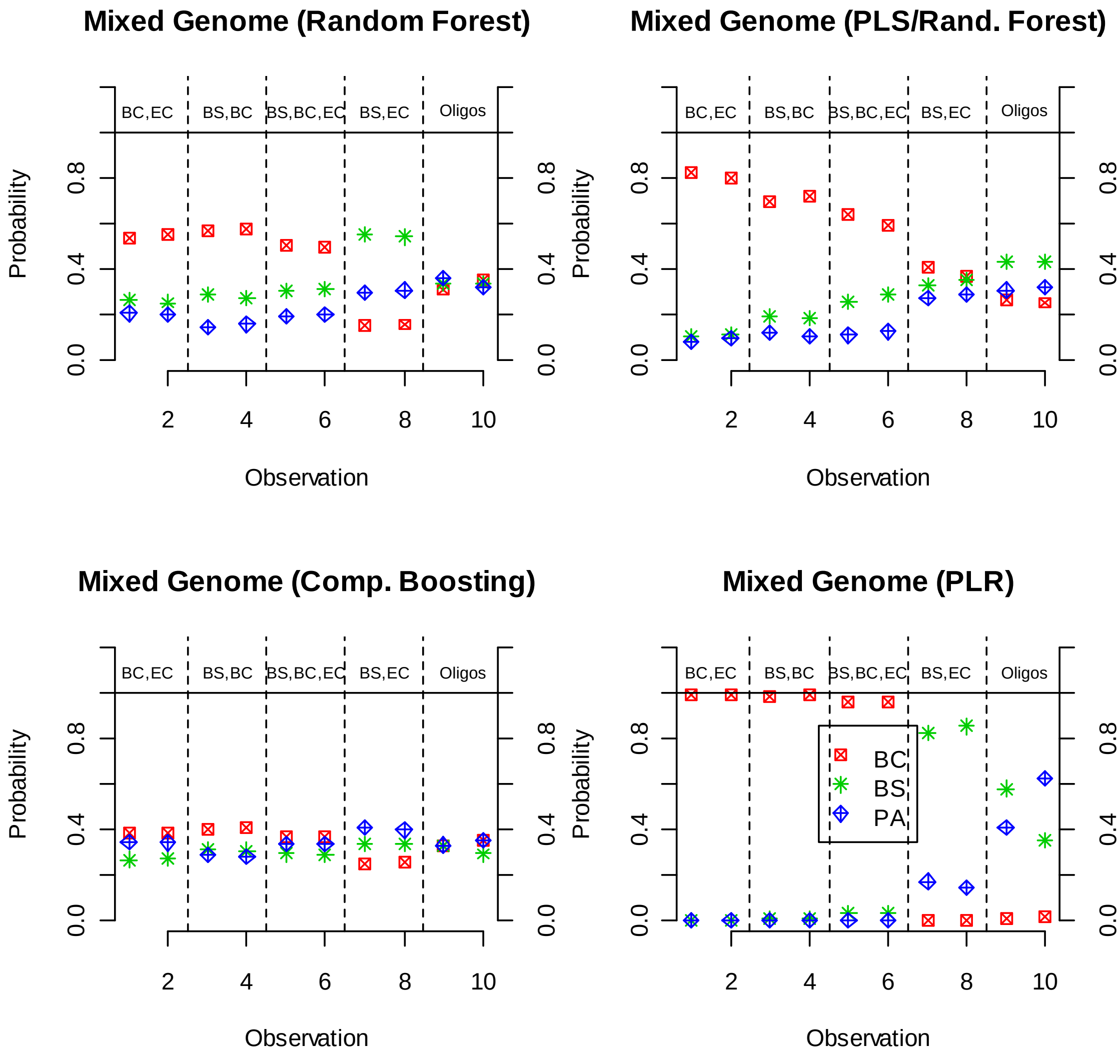


Figure 18: Voting plots across mixed-genome samples for the second four classifiers. Classifiers shown are random forests, random forests after dimensional reduction with partial least squares, component-wise boosting and penalized logistic regression.

This loss of fidelity likely represents the loss of information associated with dimensional reduction trading off the amount of information available on training conditions with the amount of data required for accurate model-fitting.

In all cases, whether against a background of *E. coli* or not, when both *B. cereus* and *B. subtilis* were mixed together in a sample, *B. cereus* was correctly classified but the corresponding *B. subtilis* signal was masked. This is likely due to several factors, including the overlap between reference hybridization patterns and the inability of many traditional classifier methods, often used in microarray analysis, to properly discriminate mixtures of training signals.

The LDA-based methods and support vector machines all had false-positive classifications of the negative control oligo arrays as either *B. subtilis* or *P. agglomerans*. In the case of the LDA-based methods, this is unsurprising given their lack of suitability when the number of observations greatly outnumbers the number of samples observed[69,70]. The random forests-based methods did a better job at quieting this oversensitivity but did so in a way that traded true-positive performance for a decrease in false-positives. Only the penalized logistic regression classifier had a discernable gulf in probability between the true positives and the false positive oligo array classifications. This gap could be exploited, with additional rigorous validation, to set thresholds for classification based upon PLR-derived predicted class probabilities.

Given these measures of relative performance, of the methods tested here, penalized logistic regression showed the most consistent separation of predicted classes and maintained high true-positive rates for *B. cereus* and *B. subtilis* (in isolation) although it, too, failed to correctly separate signals associated with *B. cereus* and *B. subtilis* when both are present in a sample.

CHAPTER 5: DISCUSSION

Initial results are encouraging for the use of pseudo-random probe arrays as a biological threat identification platform. A survey of classifier techniques used frequently in analysis of mRNA expression profiles found clear success in discriminating between species using purified genomic DNA. However, classifier performance was confounded by similar, within-genera simulants, which would likely result in a high rate of false positive alerts in a fielded sensor.

Classifier methods in microarray analysis normally look at individual up-and-down comparative gene expression levels between training conditions; the current work is instead interested in the overlap of hybridization patterns globally across the array. This likely requires re-thinking the approach to classification for many overlapping genomes like those found in environmental monitoring applications. Future efforts may explore the potential for applying mixture models to account for the overlapping of trained signals in the test data.

Fielding Sensors - Drawbacks

The current effort focuses only on genomic DNA. Many viruses that pose a threat to human health make use of RNA as their genetic material, so any implementation of this technique for environmental monitoring must be coupled to a secondary, RNAcentric system useful in detecting the presence of such viruses. This paired system would lack the flexibility in detection of the pseudo-random oligomer array but monitoring the much-noisier realm of environmental RNA (including not just the ‘genomic’ RNA of viruses but the universe of active cellular mRNA as well) is a tall task.

In addition, nucleic acid-based assays do not assess the viability or pathogenicity of a given organism; only that the genetic material of the organism is present in the environment. Since for many of the CDC Class A pathogens their presence alone is enough to cause concern, this may not be a significant issue. A front-line, tactical identification is still useful and follow-on, confirmatory secondary assays can be performed to assess viability and to inform larger-scale, high-cost decisions.

Next Steps

While the present work seeks only to demonstrate that model-driven probe design can result in effective classification of organisms of interest, addition of automated sample processing and a reagent-free approach to array readout represent excellent next steps. Micro-scale DNA isolation, amplification and separation by electrophoresis has been demonstrated in a single integrated unit[71] and a previously-mentioned method[41] has demonstrated label-free array hybridization detection. This confluence of technologies can greatly reduce the operational overhead burdening the current approach while maintaining the detection flexibility promised by the use of pseudo-random oligomer probes.

APPENDIX

The data distribution accompanying this document contains the following files and folders:

`Appendix A - VLMC Training\` - Files and folders relating to the design of the pseudo-random oligo array including training and validation studies for variable-length Markov chain models.

File	Description
`arrayProbes-Final.txt`	File of all 15,200 probe sequences tiled on the array (probes with a "_d" are duplicate probes)
`gc-hitskb-fit.emf`	Plot of BLAST hits/kb for VLMC/random probes
`hits-per-kb.jpg`	Oligo alignments per kb of genomic sequence for several species, random vs. VLMC-derived
`k0-ranging.emf`	Plot of VLMC accuracy over a range of K0 and K values
`vlmc.R`	R code to generate and bootstrap VLMC models
`generateOligos.pl`	Perl script to generate random oligos based upon ACTG frequency
`screenOligos.pl`	Perl script to calculate Tm and secondary structure propensity for oligos from a VLMC model
`rankOligos.pl`	Perl script to rank output of screenOligos.pl by secondary structure propensity
`selectOligos.pl`	Perl script to take N oligos from rankOligos.pl and duplicate a percentage at random
`sample-size-effects.jpg`	Graph of sequence sample size versus next-base accuracy for K0 = 0

`Appendix B - Reference Arrays\` - Files and folders relating to the generation of single-genome reference arrays as well as mixed-genome arrays. Files include array TIFF images, extracted data, and normalized data.

File	Description
`Mixed-PhenoData.txt`	Mixed genome sample definitions
`mix-expr.txt`	Normalized mixed genome intensities
`Mixed-Targets.txt`	Mixed array definition file for *limma* methods
`norm-qc.R`	R code to read in, normalize and QC arrays
`PhenoData.txt`	Reference genome sample definitions
`ref-expr.txt`	Normalized reference genome intensities
`Targets.txt`	Reference array definition file for *limma* methods
`unnorm-expr.txt`	Unnormalized reference genome intensities
`data\`	Feature extractions; see PhenoData.txt/Mixed-PhenoData.txt
`mix-qc\`	Same file definitions as in `ref-qc\` but for mixed samples
`ref-qc\`	Output from *arrayQualityMetrics*
`ref-qc\boxplot.pdf`	Array intensity distribution boxplots
`ref-qc\boxplot.png`	Array intensity distribution boxplots (as PNG)
`ref-qc\density.pdf`	Array intensity densities
`ref-qc\density.png`	Array intensity densities (as PNG)
`ref-qc\heatmap.pdf`	Distance-based array comparison heatmap
`ref-qc\heatmap.png`	Distance-based array comparison heatmap (as PNG)
`ref-qc\MA1.pdf`	MvA plots for arrays 1-8
`ref-qc\MA1.png`	MvA plots for arrays 1-8 (as PNG)
`ref-qc\MA2.pdf`	MvA plots for arrays 9-16
`ref-qc\MA3.pdf`	MvA plots for arrays 17-24
`ref-qc\MA4.pdf`	MvA plots for arrays 25-32
`ref-qc\meanSd.pdf`	Mean vs. standard deviation of arrays
`ref-qc\meanSd.png`	Mean vs. standard deviation of arrays (as PNG)
`ref-qc\QMreport.html`	HTML report describing visualizations
`unnorm-qc\`	Same files as `ref-qc\` but for unnormalized ref. samples

`Appendix C - WGA Replicate Study\` - Files and folders relating to the whole genome amplification (WGA) inter-amplification study. Files include array TIFF image, extracted data and normalized data.

File	Description
`array-corr.emf`	Visualization of reference and inter-WGA correlation
`PhenoData-WGA.txt`	Inter-WGA study sample definitions
`expr-WGA.txt`	Inter-WGA study sample normalized probe intensities
`Targets-WGA.txt`	Inter-WGA study array definition file for *limma* methods
`wga-corr.R`	R code to read in, normalize and QC arrays
`corr\`	Files of raw correlation coefficients
`corr\BC-corr.txt`	*B. cereus* correlation coefficients
`corr\BS-corr.txt`	*B. subtilis* correlation coefficients
`corr\PA-corr.txt`	*P. agglomerans* correlation coefficients
`corr\Reference-Corr.txt`	All inter-WGA study sample correlations
`WGACorr.txt`	All inter-WGA study correlation coefficients
`data\`	Feature extractions; see PhenoData-WGA.txt
`qc\`	Output from *arrayQualityMetrics*

Appendix D – Classification\ - Files and folders relating to the classification of training and mixed genome arrays. Files include classification results and visualizations.

File	Description
class-5-10-Comparison.emf	Cross-classifier accuracy/prob. comparison
classifiers.R	R code to generate classifier results and visualizations
mixed-voting-1.emf	Voting plots for mixed genome classifiers part I
mixed-voting-2.emf	Voting plots for mixed genome classifiers part II
reference-yhat.txt	Predicted classes for all cross-validation observations on reference data
votes-1-510-h.emf	Voting plots for ref. genome classifiers part I
votes-2-510-h.emf	Voting plots for ref. genome classifiers part II
Class Probabilities\	Directory of files, one per classifier method, of predicted class membership probabilities across all cross-validation observations

REFERENCES

REFERENCES

1. Sidell, F., Takafuji, E. & Franz, D. *Medical Aspects of Chemical and Biological Warfare*. 771(Office of the Surgeon General, Department of the Army: 1997).

2. Pomerantsev, A.P. et al. Expression of cereolysine AB genes in Bacillus anthracis vaccine strain ensures protection against experimental hemolytic anthrax infection. *Vaccine* **15**, 1846-50(1997).

3. Sabelnikov, A., Zhukov, V. & Kempf, R. Probability of real-time detection versus probability of infection for aerosolized biowarfare agents: A model study. *Biosens Bioelectron* **21**, 2070-7(2006).

4. Shekhawat, G., Tark, S.H. & Dravid, V.P. MOSFET-Embedded microcantilevers for measuring deflection in biomolecular sensors. *Science* **311**, 1592-5(2006).

5. Deisingh, A.K. & Thompson, M. Biosensors for the detection of bacteria. *Can J Microbiol* **50**, 69-77(2004).

6. Lim, D.V. et al. Current and developing technologies for monitoring agents of bioterrorism and biowarfare. *Clin Microbiol Rev* **18**, 583-607(2005).

7. Hock, B. Antibodies for immunosensors a review. *Analytica Chimica Acta* **347**, 177-186(1997).

8. Voss, T.S. et al. A var gene promoter controls allelic exclusion of virulence genes in Plasmodium falciparum malaria. *Nature* **439**, 1004-1008(2006).

9. Brumme, Z.L. et al. Evidence of Differential HLA Class I-Mediated Viral Evolution in Functional and Accessory/Regulatory Genes of HIV-1. *PLoS Pathog* **3**, e94(2007).

10. Ligler, F.S. & Erickson, J.S. Bioengineering: diagnosis on disc. *Nature* **440**, 159-60(2006).

11. Rider, T.H. et al. A B Cell–Based Sensor for Rapid Identification of Pathogens. *Science* **301**, (2003).

12. Fan, X. et al. Sensitive optical biosensors for unlabeled targets: a review. *Anal Chim Acta* **620**, 8-26(2008).

13. Gronewold, T.M.A. Surface acoustic wave sensors in the bioanalytical field: Recent trends and challenges. *Analytica Chimica Acta* **603**, 119-128(2007).

14. Tombelli, S., Minunni, M. & Mascini, M. Analytical applications of aptamers. *Biosens Bioelectron* **20**, 2424-34(2005).

15. Wang, J. & Zhou, H.S. Aptamer-based Au nanoparticles-enhanced surface plasmon resonance detection of small molecules. *Anal Chem* **80**, 7174-8(2008).

16. Savran, C.A. et al. Micromechanical Detection of Proteins Using Aptamer-Based Receptor Molecules. *Anal. Chem.* **76**, 3194-3198(2004).

17. Ivnitski, D. et al. Nucleic acid approaches for detection and identification of biological warfare and infectious disease agents. *Biotechniques* **35**, 862-9(2003).

18. Hindson, B.J. et al. APDS: the autonomous pathogen detection system. *Biosens Bioelectron* **20**, 1925-31(2005).

19. Regan, J.F. et al. Environmental Monitoring for Biological Threat Agents Using the Autonomous Pathogen Detection System with Multiplexed Polymerase Chain Reaction. *Anal Chem* (2008).

20. Christensen, D.R. et al. Detection of biological threat agents by real-time PCR: comparison of assay performance on the R.A.P.I.D., the LightCycler, and the Smart Cycler platforms. *Clin Chem* **52**, 141-5(2006).

21. Borisov, S.M. & Wolfbeis, O.S. Optical Biosensors. *Chem. Rev.* **108**, (2008).

22. Chen, L. et al. DNA hybridization detection in a microfluidic channel using two fluorescently labelled nucleic acid probes. *Biosens Bioelectron* **23**, 1878-82(2008).

23. Schulze, A. & Downward, J. Navigating gene expression using microarrays - a technology review. *Nat Cell Biol* **3**, E190-E195(2001).

24. Martinelli, L. et al. Sensor-integrated fluorescent microarray for ultrahigh sensitivity direct-imaging bioassays: Role of a high rejection of excitation light. *Applied Physics Letters* **91**, (2007).

25. Satya, R.V. et al. A high-throughput pipeline for designing microarray-based pathogen diagnostic assays. *BMC Bioinformatics* **9**, 185(2008).

26. Wang, D. et al. Microarray-based detection and genotyping of viral pathogens. *Proc Natl Acad Sci U S A* **99**, 15687-92(2002).

27. Wang, D. et al. Viral discovery and sequence recovery using DNA microarrays. *PLoS Biol* **1**, E2(2003).

28. Urisman, A. et al. E-Predict: a computational strategy for species identification based on observed DNA microarray hybridization patterns. *Genome Biol* **6**, R78(2005).

29. Charbonnier, Y. et al. A generic approach for the design of whole-genome oligoarrays, validated for genomotyping, deletion mapping and gene expression analysis on Staphylococcus aureus. *BMC Genomics* **6**, 95(2005).

30. Feng, S. & Tillier, E.R. A fast and flexible approach to oligonucleotide probe design for genomes and gene families. *Bioinformatics* **23**, 1195-202(2007).

31. Theodore, M.L., Jackman, J. & Bethea, W.L. Counterproliferation with Advanced Microarray Technology. *JHU APL Tech. Digest* **25**, (2004).

32. DeSantis, T.Z. et al. Comprehensive aligned sequence construction for automated design of effective probes (CASCADE-P) using 16S rDNA. *Bioinformatics* **19**, 1461-8(2003).

33. Chiu, C.Y. et al. Microarray detection of human parainfluenzavirus 4 infection associated with respiratory failure in an immunocompetent adult. *Clin Infect Dis* **43**, e71-6(2006).

34. Luscombe, N.M. & Babu, M.M. GenCompass: a universal system for analysing gene expression for any genome. *Trends Biotechnol* **22**, 552-5(2004).

35. Chandler, D.P. et al. Diagnostic oligonucleotide microarray fingerprinting of Bacillus isolates. *J Clin Microbiol* **44**, 244-50(2006).

36. Willse, A. et al. Comparing bacterial DNA microarray fingerprints. *Stat Appl Genet Mol Biol* **4**, Article19(2005).

37. Willse, A. et al. Quantitative oligonucleotide microarray fingerprinting of Salmonella enterica isolates. *Nucleic Acids Res* **32**, 1848-56(2004).

38. Doran, M. et al. Oligonucleotide microarray identification of Bacillus anthracis strains using support vector machines. *Bioinformatics* **23**, 487-92(2007).

39. Belosludtsev, Y.Y. et al. Organism identification using a genome sequence-independent universal microarray probe set. *Biotechniques* **37**, 654-8, 660(2004).

40. van Dam, R.M. & Quake, S.R. Gene expression analysis with universal n-mer arrays. *Genome Res* **12**, 145-52(2002).

41. Clack, N.G., Salaita, K. & Groves, J.T. Electrostatic readout of DNA microarrays with charged microspheres. *Nature Biotechnology* (2008).

42. Carrera, M. et al. Difference between the spore sizes of Bacillus anthracis and other Bacillus species. *J Appl Microbiol* **102**, 303-12(2007).

43. Adams, K.L. et al. Reagentless Detection of Mycobacteria tuberculosis H37Ra in Respiratory Effluents in Minutes. *Anal. Chem.* **80**, 5350-5357(2008).

44. Rule, A.M. et al. Application of Flow Cytometry for the Assessment of Preservation and Recovery Efficiency of Bioaerosol Samplers Spiked with Pantoea agglomerans. *Environ. Sci. Technol.* **41**, 2467-2472(2007).

45. Whiteford, N. et al. An analysis of the feasibility of short read sequencing. *Nucleic Acids Res* **33**, e171(2005).

46. Mächler, M. & Bühlmann, P. Variable Length Markov Chains: Methodology, Computing, and Software. *Journal of Computational and Graphical Statistics* **13**, 435-455(2004).

47. Nakamura, Y., Gojobori, T. & Ikemura, T. Codon usage tabulated from international DNA sequence databases: status for the year 2000. *Nucleic Acids Res* **28**, 292(2000).

48. Hsieh, L.C. et al. Minimal model for genome evolution and growth. *Phys Rev Lett* **90**, 018101(2003).

49. Carrera, M. & Sagripanti, J.L. Design and engineering of a multi-target (multiplex) DNA simulant to evaluate nucleic acid based assays for detection of biological threat agents. (2006).

50. Charif, D. & Lobry, J. Seqin{R} 1.0-2: a contributed package to the {R} project for statistical computing devoted to biological sequences retrieval and analysis. *Structural approaches to sequence evolution: Molecules, networks, populations* 207-232(2007).

51. Rimour, S. et al. GoArrays: highly dynamic and efficient microarray probe design. *Bioinformatics* **21**, 1094-103(2005).

52. Rozen, S. & Skaletsky, H. *Primer3 on the WWW for general users and for biologist programmers*. (Humana Press: Totowa, NJ, 2000).

53. Markham, N.R. & Zuker, M. UNAFold: software for nucleic acid folding and hybridization. *Methods Mol Biol* **453**, 3-31(2008).

54. Matveeva, O.V. et al. Thermodynamic criteria for high hit rate antisense oligonucleotide design. *Nucleic Acids Res* **31**, 4989-94(2003).

55. Belgrader, P. et al. A minisonicator to rapidly disrupt bacterial spores for DNA analysis. *Anal Chem* **71**, 4232-6(1999).

56. Blanco, L. et al. Highly Efficient DNA Synthesis by the Phage 429 DNA Polymerase. *J. Biol. Chem.* **264**, 8935-8940(1989).

57. Dean, F.B. et al. Rapid amplification of plasmid and phage DNA using Phi 29 DNA polymerase and multiply-primed rolling circle amplification. *Genome Res* **11**, 1095-9(2001).

58. Xu, Y. et al. Rapid detection and identification of a pathogen's DNA using Phi29 DNA polymerase. *Biochem Biophys Res Commun* (2008).

59. Brueck, C. et al. Single Cell Whole Genome Amplification: Unleashing a World within a Cell. (2005).at <http://www.sigmaaldrich.com/sigma/general%20information/singlecellwga.pdf>

60. Agilent Technologies, Inc. Agilent 8x15K CGH Microarray Protocol for Processing. (2007).at <http://www.chem.agilent.com/Library/usermanuals/Public/G4427-90010.pdf>

61. Zahurak, M. et al. Pre-processing Agilent microarray data. *BMC Bioinformatics* **8**, (2007).

62. Smyth, G.K. Limma: linear models for microarray data. *Bioinformatics and Computational Biology Solutions using R and Bioconductor* 397-420(2005).

63. Gentleman, R.C. et al. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology* **5**, R80(2004).

64. Bolstad, B.M. et al. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185-93(2003).

65. Kauffmann, A. & Huber, W. *arrayQualityMetrics: Quality metrics on microarray data sets*.

66. Slawski, M. & Boulesteix, A. *CMA: Synthesis of microarray-based classification*.

67. Darling, A.E., Carey, L. & Feng, W. The Design, Implementation, and Evaluation of mpiBLAST. *In Proceedings of ClusterWorld 2003* (2003).

68. Dudoit, S. et al. Statistical methods for identifying genes with differential expression in replicated cDNA microarray experiments. *Statistica Sinica* **12**, 111-139(2002).

69. DiPillo, P.J. The application of bias to discriminant analysis. *Communications in Statistics - Theory and Methods* **5**, 843-854(1976).

70. Guo, Y., Hastie, T. & Tibshirani, R. Regularized linear discriminant analysis and its application in microarrays. *Biostatistics* **8**, 86(2007).

71. Huang, F.C., Liao, C.S. & Lee, G.B. An integrated microfluidic chip for DNA/RNA amplification, electrophoresis separation and on-line optical detection. *Electrophoresis* **27**, 3297-3305(2006).

CURRICULUM VITAE

James C. Diggans received his Bachelor of Science from the University of Florida in 1999 with majors in both Computer Science and Microbiology/Cell Science and minors in Chemistry and Spanish. He spent seven years working in commercial biotechnology before moving on to work in non-profit government-sponsored biotechnology research and development.