# Use of Large, Immunosignature Databases to Pose New Questions

# About Infection and Health Status

by

Lu Wang

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved January 2018 by the
Graduate Supervisory Committee:

Stephen Albert Johnston, Chair
Phillip Stafford
Kenneth Buetow
Grant McFadden

ARIZONA STATE UNIVERSITY

May 2018

# ABSTRACT

Immunosignature is a technology that retrieves information from the immune system. The technology is based on microarrays with peptides chosen from random sequence space. My thesis focuses on improving the Immunosignature platform and using Immunosignatures to improve diagnosis for diseases. I first contributed to the optimization of the immunosignature platform by introducing scoring metrics to select optimal parameters, considering performance as well as practicality. Next, I primarily worked on identifying a signature shared across various pathogens that can distinguish them from the healthy population. I further retrieved consensus epitopes from the disease common signature and proposed that most pathogens could share the signature by studying the enrichment of the common signature in the pathogen proteomes. Following this, I worked on studying cancer samples from different stages and correlated the immune response with whether the epitope presented by tumor is similar to the pathogen proteome. An effective immune response is defined as an antibody titer increasing followed by decrease, suggesting elimination of the epitope. I found that an effective immune response usually correlates with epitopes that are more similar to pathogens. This suggests that the immune system might occupy a limited space and can be effective against only certain epitopes that have similarity with pathogens. I then participated in the attempt to solve the antibiotic resistance problem by developing a classification algorithm that can distinguish bacterial versus viral infection. This algorithm outperforms other currently available classification methods. Finally, I worked on the concept of deriving a single number to represent all the data on the immunosignature platform. This is in resemblance to the concept of temperature, which is an approximate measurement of whether an individual is healthy.  The measure of Immune Entropy was found to work best as a single measurement to describe the immune system information derived from the immunosignature. Entropy is relatively invariant in healthy population, but shows significant differences when comparing healthy donors with patients either infected with a pathogen or have cancer.

# DEDICATION

I dedicate this dissertation to my family and friends

# ACKNOWLEDGMENTS

I am very grateful to my advisor, Dr. Stephen Albert Johnston. Dr. Johnston has inspired me and provided support during my PhD study. He always challenges me with ideas and persuade me to follow up on unusual phenomenon that constitute the majority parts of my thesis.

I am highly thankful to Dr. Phillip Stafford. Dr. Stafford provided most support in bioinformatics and statistical analysis. He helped in my transition into computational biologist and guided me with the right technical skills.

I would like to thank my committee members, Dr. Kenneth Buetow and Dr. Grant McFadden for helping broaden the scope of my research and offer critical reviews for my work. I would also like to thank my previous committee member Dr. Seungchan Kim, who provided lots of computational guidance during my early PhD studies.

I wish to thank lab members in Center for Innovations in Medicine (CIM) at Biodesign Institute. Dr. Neal Woodbury and Dr. Chris Diehnelt provided feedbacks on my research. Dr. Luhui Shen and Jian Zhang helped a lot with biological interpretations as well as array experiments. Dr. Kurt Whittemore had major contribution to the entropy study. Dr. Zhan-gong Zhao helped me to better understand Immunosignature array manufacturing. Dr. Josh Richer and Dr. Krupa Navalkar helped me with bioinformatics analysis. Dr. Zhao Dr. Zbigniew Cichacz and the Peptide Array Core team performed most of Immunosignature assays. Thank Pattie Madjidi and Penny Gwynne for their administrative support.

## TABLE OF CONTENTS

Page

CHAPTER Page

# LIST OF TABLES

TABLE Page

# LIST OF FIGURES

FIGURE Page

FIGURE Page

FIGURE Page

# CHAPTER 1

# INTRODUCTION

## Research overview

The immune system is rich with information. Immunosignature diagnostics is a technology that can retrieve the antibody information from the immune system. The platform is composed of peptides chosen from random sequence space that is able to bind complex mixtures of antibodies. My thesis is focusing on improving the immunosignature platform, using immunosignatures to characterize the immune system and improving current diagnosis both for pathogen infection and cancer.

## Current status of healthcare system

Increasing healthcare expenditure is a major burden for every citizen. In the US, healthcare expenditures always increase at faster rate than GDP and now accounts for 17.8% of GDP in year 2015 (Martin, Hartman et al. 2016). One major reason for this is the primary focus on treating patients with late-stage diseases. Hundreds of thousands of dollars can be spent to extend life for a few months for one late-stage cancer patient. The new checkpoint inhibitor and CAR-T treatments are estimated to cost over $200,000. These new treatments may be much more effective but are also much more expensive. To help change the situation, focus should be shifted to diagnosis of diseases early and treatment of patients early. Research has shown that if diagnosed early, breast cancer patients would have low mortality rate (Tabar, Gad et al. 1985). So if we can have a diagnosis technology that is both cheap and accurate in identifying patients with early stage diseases, we can lower the overall cost of treatment and potentially increase the survival rate. As a result, diagnosis should be the focus of future healthcare system. This thesis illustrates in detail that the Immunosignature technology may be the diagnosis platform that could totally change the paradigm of healthcare system through its ability in performing diagnosis accurately for various diseases.

## Biomarkers used in diagnosis

There are lots of biomarkers being studied in research labs and being used in clinical settings to help with diagnosis of diseases. They can be classified into several groups. DNA, RNA, protein and carbohydrate biomarkers (Mishra and Verma 2010). DNA and RNA biomarkers are generally used for non-infection diseases, including cancer, auto-immune diseases and Alzheimer disease et al (Wang, Fan et al. 1998, Begovich, Carlton et al. 2004, Zhao, Li et al. 2004, Li, Wetten et al. 2008), although they are also used as pathogen biomarkers (Periyannan Rajeswari, Soderberg et al. 2017). Single nucleotide polymorphism (SNP) is a major type of DNA biomarkers (Hueber, Utz et al. 2002) and miRNA is one class of RNA biomarkers (Uhlmann, Brinckmann et al. 2002, Gunderson, Steemers et al. 2005, Raghavan, Lillington et al. 2005, Mitchell, Parkin et al. 2008, Duttagupta, Jiang et al. 2011, Pritchard, Kroh et al. 2012). Carbohydrate biomarkers are changes in glycoproteins, glycolipids or proteoglycans. They are generally very stable and can be used as biomarkers for pathogen and chronic diseases (Liang, Wu et al. 2008, Packer, von der Lieth et al. 2008, Lawrence, Brown et al. 2012).

Of these biomarkers, proteins are the most important, because proteins are the major functional bio-molecules in any organism (Rifai, Gillette et al. 2006). As a result, they are more closely related to disease initiation and progression. Protein biomarkers can be used for infectious diseases and chronic diseases like autoimmune diseases, Alzheimer's diseases and cancer (Qiu, Madoz-Gurpide et al. 2004, Georganopoulou, Chang et al. 2005, Haab 2005, Keating 2005, Lueking, Huber et al. 2005, Kingsmore 2006). There are currently various technologies using proteins as probes including mass spectrometry (MS), protein or peptide microarrays and bead based immunoassay (Tanaka, Waki et al. 1988, Choi, Oh et al. 2002, Templin, Stoll et al. 2002, Aebersold and Mann 2003, Angenendt, Glökler et al. 2003, Angenendt 2005, Yang, Lien et al. 2008).

Antibody biomarkers are the most important type of protein biomarker. There are several advantages of antibodies as biomarkers. First, an antibody response can be elicited towards any disease. This allows the use of antibodies as a universal biomarker for any diseases (Andresen and Grotzinger 2009, Ballew, Murray et al. 2013). Second, the antibody response can be magnified in titer. Upon encountering foreign molecules, the immune system will generate antibodies in extremely high amount, making antibodies a better biomarker than the foreign molecules. Third, antibodies can be measured in serum and are very stable. This allows the process of diagnosis to be simpler and more accurate (Cole, DeNardo et al. 1987, Geijersstam, Kibur et al. 1998).

### Current microarray based diagnostic technologies

Microarrays are widely used as a diagnostic technology. They can be low in cost and can analyze thousands of proteins in single assay (Navalkar 2014). Microarrays can be customized to a specific disease according to the need of researchers (Russo, Zegar et al. 2003). As a result, there are various types of microarrays, including DNA, RNA, protein and peptide microarrays.

DNA microarrays are a common type of microarray. They are mostly used to profile human gene expression in different diseases (Heller 2002). The idea is that diseases can cause differential gene expression compared to healthy individuals and this can be used as diagnosis or be treated as risk factor. Pathogen DNA can also be printed onto the array to directly monitor for diagnosis of specific pathogens (Leinberger, Schumacher et al. 2005, Cleven, Palka-Santini et al. 2006). DNA microarrays have been used for diagnosis of infectious diseases, cancer and other chronic diseases (Cummings and Relman 2000, Chen, Liu et al. 2001, Chizhikov, Rasooly et al. 2001, Li, Chen et al. 2001, Petrik 2001). There are several commercially available DNA microarray platforms from Affymetrix (High-Density microarrays), Nanogen (Microelectronic array), and new technologies are being developed (Diehl, Grahlmann et al. 2001, Degliangeli, Kshirsagar et al. 2014, Li, Zhao et al. 2014, Rödiger, Liebsch et al. 2014, Moran, Arribas et al. 2016).

RNA microarrays are relatively less used because RNA is less stable (Scherrer, Latham et al. 1963, Salser, Janin et al. 1968, Auer, Lyianarachchi et al. 2003). RNA is usually reverse transcribed into cDNA followed by printing on to a cDNA microarray for diagnosis (Hegde, Qi et al. 2000, Seki, Ishida et al. 2002). RNA microarrays generally have similar usage to DNA microarrays and have shown applications in various diseases (Zhou, Thompson et al. 2002, Gottardo, Liu et al. 2007).

Protein microarrays are becoming more and more important in diagnosis because they can be used to study interaction between proteins, peptides and other molecules (Ge 2000, Angenendt, Glökler et al. 2003). There are several types of protein arrays: detection (analytical), functional and reverse phase microarrays, with analytical microarray being the most common one (Bertone and Snyder 2005, Hall, Ptacek et al. 2007). Antibody microarrays are a type of detection microarray. Protein microarrays, especially antibody microarrays, have been applied in diagnosis in various infectious and chronic diseases (Davies, Liang et al. 2005, Zhong, Hidalgo et al. 2005, Zhu, Hu et al. 2006, Kwon, Lee et al. 2008, Hartmann, Roeraade et al. 2009, Bilek 2014, Hu, Niu et al. 2015, Werner, Chen et al. 2015, Borrebaeck 2017, Lessa-Aquino, Lindow et al. 2017).

Peptide microarrays are highly similar to protein microarrays, except here a relatively short peptide is used instead of long protein (Cretich, Damin et al. 2006). There are several advantages for using peptide array compared with protein array. First, peptides are shorter than proteins, which makes it possible to partition and identify specific region of the binding interaction. Second, peptide microarrays only require very small amount of sample and minimal preparation steps. Peptide microarrays that measures antibody binding are the most commonly used peptide array. Peptide microarrays are widely used to diagnose pathogen infections, cancer and other chronic diseases (Duburcq, Olivier et al. 2004, Gaseitsiwe, Valentini et al. 2008, Maksimov, Zerweck et al. 2012, Stafford, Cichacz et al. 2014).

## Immunosignature

Immunosignature is a technology developed at the Center for Innovations in Medicine at Biodesign Institute, Tempe, AZ. It is a peptide microarray technology that incorporates advantages of both using antibodies and peptide as probes on the platform. The unique feature of immunosignature is that instead of printing or synthesizing biological peptide sequences, non-biological peptides selected from random sequence space are used. Since immunosignature is not using sequences from any specific organism, it can be used to perform diagnosis on any diseases including infections, cancer and other chronic diseases (Restrepo, Stafford et al. 2011, Restrepo, Stafford et al. 2012, Legutki, Zhao et al. 2014, Navalkar, Magee et al. 2014, Stafford, Cichacz et al. 2014, Richer, Johnston et al. 2015).

The general workflow of how immunosignature works starts from a drop of blood. Less than 1 µl of sample is needed in each assay (Chase, Johnston et al. 2012). Serum is incubated on the immunosignature array to allow interaction between antibodies in the serum with peptides on the array. Theoretically, a specific antibody will bind to peptides that are similar to the antibody's original epitope. After incubation, serum is washed off, leaving only antibodies that are bound to peptides on the array. Fluorescent secondary antibodies are used to visualize the binding of primary antibodies (Stafford, Cichacz et al. 2014). The basic premise of immunosignatures is that different diseases will reproducibly elicit the same antibodies that can be detected on the array. As a result, diagnosis can be performed by comparing disease sample versus healthy sample and analyze the differential antibody binding pattern.

There are several versions of immunosignatures. The earliest version consisted of 10,000 peptides whose sequences were generated by a random amino acid selection process. The peptides were synthesized commercially and printed onto glass slides. The later versions consist of in situ synthesized arrays of 120,000~330,000 peptides. Programs were developed to choose the peptides from random sequence space that maximize chemical diversity. The boost in peptide number enables better distinguishing power because more peptides will allow for more precise binding and can better stratify the antibodies. The use of in-situ synthesis is an important advance for immunosignature. It can improve the quality of peptides because peptides will be synthesized in batch compared with printing each peptide individually. Thus the in-situ method can produce low variability in the quality between peptides. And in-situ synthesis gives the ability to synthesize much higher number of peptides at lowered cost. Purchasing peptides individually can be expensive compared with in-situ synthesis. Synthesizing more peptides in-situ only has minimal effect on cost. The only limitation of how many peptides to synthesize is space on the array.

Early immunosignature tests used glass microscope slides (Stafford, Halperin et al. 2012). The newer in-situ synthesized immunosignature are manufactured on silicon wafers (Donnell, Maurer et al. 2015). The manufacturing process is through photolithography that is similar to how Intel synthesizes CPUs and will be elaborated in Chapter 2 (Baidya, Dandekar et al. 2016). Briefly, peptides to be synthesized are re-coded into photomasks. Then each photomask is used in sequential order. At each step, a specific amino acid will be added onto the sequence at locations specified by the photomask. This process is repeated until all photomask are used and the desired peptides are synthesized on the immunosignature array (Stafford, Cichacz et al. 2014, Donnell, Maurer et al. 2015). After synthesis, the wafer is cut into standard glass microscope slide size for further processing. Each wafer can be cut into 12 slides and each slide can process 24 independent assays simultaneously.

The immunosignature technology itself is still evolving. There have been improvements in synthesis as described above. New sample preparation methods are being developed for immunosignature (Chase, Johnston et al. 2012). More advanced analytical methods of the signatures are being applied and optimized for immunosignature (Brown, Stafford et al. 2011, Kukreja, Johnston et al. 2012, Whittemore 2014, Donnell, Maurer et al. 2015). ASU spinout company HealthTell (www.healthtell.com) was founded to explore commercial usage of the immunosignature technology.

Immunosignature has been used in performing diagnosis of various diseases. Navalkar et al used immunosignature to diagnose valley fever (Navalkar, Magee et al. 2014, Navalkar 2014, Navalkar, Johnston et al. 2015). Legutki et al used immunosignature to distinguish 6 types of pathogens and healthy individuals from each other (Legutki, Zhao et al. 2014). Richer et al used immunosignature to perform diagnosis on 7 types of infections and identified disease-specific epitopes (Richer, Johnston et al. 2015). Johnston et al showed immunosignature can be used to perform diagnosis for canine lymphoma (Johnston, Thamm et al. 2014). Stafford et al managed to distinguish 14 different diseases including various cancers and infectious diseases in parallel using immunosignature (Stafford, Cichacz et al. 2014). Restrepo et al showed immunosignature can be used to diagnosis Alzheimer’s disease (Restrepo, Stafford et al. 2011, Restrepo, Stafford et al. 2012). Singh et al showed that immunosignature can distinguish chronic fatigue patients from healthy controls (Singh, Stafford et al.). All these results show the feasibility of using immunosignature as a diagnosis platform for various diseases of human or other animals with antibody-based immune systems.

In addition to performing traditional diagnosis, Immunosignature is also a powerful research tool. The construct of using non-biological sequences enables unbiased study of various diseases at the same time. This allows researchers to find commonality and dissimilarity for diseases. For example, is it possible that all infectious disease share common signatures (Chapter 3)? Does the same cancer at different stages have different epitopes and how is it changed (Chapter 4)? Can our immune system itself distinguish bacterial versus viral infection (Chapter 5)?

My research relies on Immunosignature technology throughout this thesis. I used the platform to performed diagnosis and answer fundamental biological questions.

### The use of antibiotics and challenges of antibiotics overdose problem

Antibiotics are drugs used to treat or prevent bacterial infections. They can kill or inhibit growth of bacteria (Walsh 2003). One of the best well-known of antibiotics is penicillin, which can be dated back into 1920s (Abraham, Chain et al. 1941). Penicillin is best-known for its use during World War II that reduced mortality of wounded soldiers (Kardos and Demain 2011). After World War II, penicillin was quickly made available to the public for civilian use in multiple countries (Ligon 2004). The penicillin drug group itself has seen several major developments, from ampicillin that offered broader spectrum in 1961, to carbenicillin that offers protection against Gram-negative bacteria (Knudsen, Rolinson et al. 1967, Anderl, Franklin et al. 2000).

Antibiotics can be classified into several groups based on mechanism of action, spectrum or structure (Schwalbe, Steele-Moore et al. 2007). By mechanism, it can be divided into Bactericides and Bacteriostatic agent (Finberg, Moellering et al. 2004). Bactericides directly kills bacteria while Bacteriostatic antibiotics prevent bacteria from dividing (Pankey and Sabath 2004). Bactericides can be further divided into antibiotics that target the cell wall, cell membrane or essential enzymes. Penicillin is an example of an antibiotic that targets the cell wall. By spectrum, antibiotics can be divided into broad-spectrum and narrow-spectrum (Rea, Dobson et al. 2011). As the name indicates, broad-spectrum antibiotics work against a wide range of bacteria, while narrow-spectrum antibiotics only target specific types of bacteria. Ampicillin, which belongs to the penicillin group, is an example of broad-spectrum antibiotic (Montecalvo, Horowitz et al. 1994). By chemical structure, antibiotics can be divided into over 20 types. Some of the major types include penicillin, peptide, aminoglycoside and glycopeptide (Cunha 2010).

The effectiveness of antibiotics makes society to rely more and more on them. However, the general usage of antibiotics is causing problems. Because bacteria are always evolving new antibiotics are needed (D'costa, King et al. 2011). An antibiotic can kill bacteria subtypes that are not resistant to it, while in the meantime promote the growth of a bacteria subtype that is resistant that antibiotic (Goossens, Ferech et al. 2005). As a result, antibiotics that used to be useful can stop being effective after years of clinical usage

(Hawkey and Jones 2009). New antibiotics need to be developed to counter this challenge. And the resistant bacterium can be more difficult to treat, especially if it is resistant to multiple antibiotics at the same time (Mitscher, Pillai et al. 1999). We can easily imagine the future where there are limited or no options to treat some bacteria. Many reviews have called attention to this serious crisis. (Bell, Schellevis et al. 2014, Camargo, García et al. 2014, Rossolini, Arena et al. 2014, Ghotaslou, Leylabadlo et al. 2015, Lainson, Fuenmayor et al. 2015, Teillant, Gandra et al. 2015, Gupta, Lainson et al. 2016, Sharma, Johnson et al. 2016, Gupta, Lainson et al. 2017).

As it has been described above, antibiotics can only be used to treat bacterial infection, with some examples of treating protozoa (Felsenfeld, Volini et al. 1950, Krupp and Madhivanan 2015, Park 2016). But they are not used to treat viral infections. The public generally does not know this is the case, and often requests doctors to prescribe antibiotics when they have flu, which is actually the major type of misuse of antibiotics (McNulty, Boyle et al. 2007). This misuse and overdose can cause resistance while at the same time do no good for the patient (Huttner, Goossens et al. 2010).

In addition to the mistaken opinion of the public, one major reason for antibiotics over-usage is the lack of accurate diagnosis. Bacterial and viral infections can have the same symptoms, which makes it hard for doctors to diagnosis the type of diseases. Using respiratory tract infection as an example, it refers to various infectious diseases involved in the respiratory tract. This includes bacterial infections like Bordetella pertussis, Mycoplasma pneumoniae, Streptococcus pneumoniae and Haemophilus influenza and viral infections like influenza, Adenovirus, Herpes simplex virus and respiratory syncytial virus (Eccles, Grimshaw et al. 2007, Ruuskanen, Lahti et al. 2011). Respiratory tract infections occurs more frequently in children and lower respiratory tract infections are actually the leading cause of death considering all infectious diseases (Organization 2004). Doctors are usually faced with the dilemma of without knowing the type of infection, whether antibiotics should be prescribed immediately to save the life of the child or be a little bit more cautious for the prescription. If an accurate diagnosis for distinguishing bacterial and viral infection existed, doctors do not need to make the choice and can handily decide on the correct treatment immediately.

In Chapter 5 I will describe a diagnosis test to distinguish between bacterial infection and viral infection using immunosignature.

### Entropy as a measurement of system orderness

Immunosignature may be powerful in terms of performing diagnosis. However, the high-dimensional nature of immunosignature makes it hard for people without bioinformatics background to interpret (Stafford, Cichacz et al. 2014). It would be best to summarize an immunosignature result into one single measurement so that the result can be interpreted by anyone. This idea should work like the concept of temperature. If your number of temperature is within a specific range, then you are probably healthy. If your temperature is higher or lower than the specific range, then you may become ill and should take appropriate preventative measures. This measurement does not need to be perfectly accurate, but should be able to reflect the health status of individual with relatively good accuracy. Note that since we are collapsing high-dimensional data into one single measurement, this measurement will have less accuracy than the result by directly analyzing high dimensional data.

There are various measurements can be tested for the feasibility of application to Immunosignature. These measurements can be divided into three major groups: central tendency , dispersion measurements and shape measurements (Chandler 1987, Dodge 2006). Central tendency measurements aim at finding the “center” of the distribution. The most common types of central tendency measurement include mean and median. Dispersion measurements try to measure the stretchiness of a distribution. For example, variance measures how spread out is the distribution. Other common measurements include range, interquartile range (IQR), coefficient of variance (CV) and entropy. Shape measurements describe the shape of the distribution. Skewness and kurtosis are the major examples in this group (Mardia 1970). This means skewness will tell you whether most data are shifted to the left, to the right, or equally balanced for both ends. Skewness measures asymmetry of the distribution, while kurtosis measures the “tailness” of the distribution (Joanes and Gill 1998). All of these measurements have the potential to be used as the single statistical measure to describe the immunosignature distribution. However, as will be discussed in detail in chapter 6, entropy is found to be the best measurement.

Entropy measures the randomness or uncertainty of the distribution (Rényi 1961). In equivalence, it measures the information contained from the distribution content. More information equals less uncertainty. A coin toss is an example of high entropy, because the probability of the next result is totally unknown, with both sides having equal chance to appear. As a result, a coin toss contains minimal information and has maximum randomness. An opposite example is the English text (Shannon 1951). Even though we cannot predict with 100% accuracy which word will follow another one, we do know that certain characters are used more than others and certain words will have higher probability to follow a specific word. There is research that shows missing a small portion of words in a sentence or paragraph does not influence the understanding of the content (Honeyfield 1977, Beck, McKeown et al. 1983). As a result, English text contains lots of information and lower uncertainty, and is an example of low entropy. The equation of Shannon's entropy is written as follows:

$$\mathrm{H(X)} = -\sum_{i=1}^{n} P(Xi) \log_b P(Xi)$$

X is the random variable with possible values of X1… Xi. P(X) is the probability function. This concept was first introduced by Claude Shannon in 1948 (Shannon 1948). He tried to use entropy to measure the uncertainty in messages for the application of encoding information. However, the concept was quickly adopted by researchers from various fields for new tasks. For example, entropy has been used to describe diversity of species (Jost 2006). It has been used to calculate stochastic process information rate (Cover and Thomas 2012). And entropy has been used in improving financial decisions (Tang, Leung et al. 2006).

Entropy has also been used extensively in various aspects of biological research. It is applied to research on evolution (Gladyshev 1999). It is used in analyzing functional genomics (Butte and Kohane 2000). Neurologist performed research using entropy (Shaw, Seneff et al. 2014). However, no one has used entropy to do diagnosis using microarray data. In Chapter 6, I will describe the feasibility of using entropy to describe the health status using Immunosignature technology.

## Project description

This thesis focuses on using Immunosignature technology to answer various new questions about infection and health status.

In Chapter 2, I described the contributions I made in optimizing the Immunosignature technology. This improvement enables Immunosignature to represent much larger sequences space and potentially increasing disease distinguishing power. I developed various scoring metrics to evaluate the performance of different immunosignature versions, shed light on potential biases in sequence synthesis and helped to gain better understanding of the platform itself.

Chapter 3 describes an unusual phenomenon of all pathogens sharing the same signature. I first observed it and tested it on various datasets and various diseases. I then identified the epitopes behind the signature and proposed possible biological relevance of the common signature. The possible usages of this finding are in population monitoring for an unknown disease outbreak and broadly protective vaccines against a large group of pathogens.

The next Chapter investigates the cancer epitope evolution from early to late stage. I found the epitopes are different at different stages. The immune response is different to different epitopes and suggested the immune response is efficient towards pathogen-like epitopes, indicating the immune system might have limitation and can only work against specific epitopes.

In Chapter 5, I contributed to the clinical relevant problem of distinguishing bacterial infection from viral infection. Using the Immunosignature technology I am able to develop a classifier that is >10% more accurate than current diagnostics. I further identified the peptides that are most important in the diagnosis and identified the function sequences of those peptides.

Chapter 6 presents the finding of using single measurement (entropy) to represent the complex Immunosignature readings. Various factors that can influence entropy values are first investigated. The distribution of entropy is different between patients with infectious diseases or cancer from the healthy group.

To summarize, this thesis represents my work from optimizing the platform to using Immunosignature to answer various questions that are either clinically relevant or of theoretical research interest. All these results show that Immunosignature is a powerful tool that can be used in diagnosis of various diseases and perform fundamental biological research.

# CHAPTER 2

## OPTIMIZATION OF IMMUNOSIGNATURE PLATFORM WITH MASK DESIGN

### Abstract

Immunosignaturing is a method by which random-sequence peptides in microarray format are used to assess antibody properties from persons suffering from chronic or infectious disease. With 10,000 random-sequence peptides, antibodies against diseases exhibit concerted behaviors allowing disease prediction through deconvolution of antibody-peptide interactions. Early efforts proved feasibility with only 4000-10,000 peptides per array. With more peptides, the precision with which antibody behavior can be determined increases far more than might be predicted. Physically printing peptides, even with high precision non-contact printers, will not enable the density necessary for high content peptide microarrays. However, lithography systems and in-situ synthesis will. Shadow mask technology is very robust, enabling millions of peptides to be created on a standard microscope slide. A downside of this technology is the upfront cost of masks. For creation of a 17mer peptide with 20 different amino acids, one needs 340 masks, and 340 synthesis steps. High numbers of masks are expensive and impose a risk of failure. By reducing the number of masks, one decreases the number of protection/deprotection/synthesis steps. We evaluated 2 Mask generation methods with different parameters using various bioinformatics scoring metrics. Results indicate that a more sophisticated filtering system for peptide selection coupled with mask reduction can enable a very diverse peptide library with a minimum of repetition.

## Introduction

The identification of biomarkers for classification of existing diseases could provide a rapid and inexpensive adjunct to standard diagnosis. Immunosignature technology has provided researchers with a tool for diagnosing disease with a single drop of blood, and leverages the interaction of serum antibodies with random-sequence peptides. The initial product was a 10,000 peptide microarray on which was spotted pre-synthesized 17mer peptides with a constant 3mer linker. This array is responsible for numerous successful disease classifications and analytical techniques specific to immunosignaturing (Legutki, Magee et al. 2010, Brown, Stafford et al. 2011, Restrepo, Stafford et al. 2011, Chase, Johnston et al. 2012, Restrepo, Stafford et al. 2012, Stafford, Halperin et al. 2012, Navalkar, Magee et al. 2014, Stafford, Cichacz et al. 2014). Advantages of this system are purity and ease of mass spectrometry from HPLC-purified peptides, and long shelf-life of lyophilized peptides. The production of the microarray is rapid and simple - by diluting a master mix of peptide into 384-well plates and printing onto commercially produced aminosilane-coated glass slides using commercial non-contact piezo printing (AMI, Tempe, AZ), the cost per slide is fixed and predictable, and the quality is high. However, this manufacturing paradigm does not scale well, with the costs remaining fixed rather than scaling with volume. Also, one can only print ~30,000 spots easily on a microscope slide because the solubility of random peptides differs based on sequence, and thus their printing performance is quite variable. Ironically, random peptides suffer from this much less than life-space peptides due to random distribution of pI and hydrophobicity (Bigelow 1967, Parks 1967, Rose, Geselowitz et al. 1985). Life-space peptides are often quite hydrophobic due to the way nature evolved proteins to interact with water and with cellular membranes.

Computer manufacturers have been able to leverage optical lithography to continually reduce the size of electronic features that can be etched, enabling greater computation speed, reduced energy usage, and reduced cost as feature sizes shrink. There are optical and electronic barriers to this process, but so far Moore's Law has been upheld (Schaller 1997). Manufacturing of peptide microarrays must be made scalable if immunosignaturing is to be useful as a method to continuously monitor health (Stafford, Wrapp et al. 2016). We have developed a method that uses semiconductor-grade equipment to generate peptides on a silicon surface, and have increased the number of peptide features from 10,000 per slide to over 8M peptides per slide. To create the same random library found on the 10,000 peptide glass slides, one needs 340 different masks.

Photolithography of peptides can use either light-activated amino acids that couple upon exposure to light (PepperPrint) or use photoacid or photobase (PAG or PAB) generators that enable BOC or F-MOC synthesis (LC Sciences) (Levenson, Viswanathan et al. 1982, Nuwaysir, Huang et al. 2002). We chose the more conservative approach of using photoacids and BOC synthesis with features of 10um in width spaced at 15um center-to-center distance. This method can use either mask-based illumination or digital light projection (DLP) to produce the acid. We chose the more precise mask-based system because of the number of features that can be created and the precision of near-contact mask-based lithography. This would yield 330,000 peptides per 7mm square area, with 24 different arrays per 1x3" microscope slide, or 342 replicate arrays per 8 inch wafer. To reduce mask complexity, we removed 4 amino acids from the selection pool due to their redundancy and lack of importance in previous studies: C, I, Q and M were left out yielding a 17mer with 16 different amino acids, or 16*17=272 masks for a fully unbiased random set. We developed an algorithm that used the number of masks to restrict the amino acids in a growing peptide. Thus, for any number of masks less than 272, we create peptides that are non-random. Because of the first selection process, the peptides were highly biased at the C and N-terminus due to the selection of amino acids during virtual mask generation. We synthesized a number of wafers using the peptides thus generated, then revisited the peptide generation software to create a version 2.

In this chapter, I performed extensive bioinformatics analyses on the peptides produced and present several attributes that should be considered when designing masks for immunosignaturing microarrays. The performance of old Mask generation method and new generation method are compared to understand the improvements in various metrics scores and stability.

## Method

### Peptide Generation

Peptides are generated using the script written in Matlab by Dr. Neal Woodbury. Variables can be changed including number of Masks and which amino acids to use. For both Mask generation methods, we used 18 amino acids excluding C and M. When 16 amino acids were used, that eliminates I and T, and the 14 amino acid set excludes E and

S. For the “new generation method”, peptides are generated with initial parameters of maximum length of 16 aa, minimum length of 10 aa, minimum new pentamer of 5, maximum percent of each amino acid at the N-Terminus of 10% and number of N-terminus amino acids to constrain of 1. Output from the analysis is sequences of numbers, which are then assigned amino acids in alphabetical order, and then sequence reversed to get the Nterm to Cterm standard nomenclature sequence peptides. The pure random peptide set is generated using a random number generator in R software with length of 17 aa for all peptides. The generated peptide libraries are listed in Table 2.1.

| old design | | | new design | | |
|---|---|---|---|---|---|
| Mask # | aa # | Success or not | Mask # | aa # | Success or not |
| 340 | 20 | Yes | 340 | 20 | Yes |
| 272 | 20 | Yes | 272 | 20 | Yes |
| 140 | 20 | Yes | 140 | 20 | Yes |
| 140 | 18 | Yes | 90 | 20 | Yes |
| 140 | 16 | Yes | 90 | 18 | Yes |
| 140 | 14 | Yes | 90 | 16 | Yes |
| 70 | 20 | Yes | 90 | 14 | No |
| 35 | 20 | Yes | 70 | 20 | Yes |

|  |  |  |  |  |  |
|---|---|---|---|---|---|
| Pure random | 20 | Yes | 35 | 20 | No |

**Table 2.1. Mask design settings used in this study.**

*Both old and new Mask generation methods are used. The number of Masks ranges from 35 to 340. The number of amino acids ranges from 14 to 20aa. 18 separate sets were used in this paper, including 17 sets using the Mask design algorithm and 1 set of pure random sequence peptides generated using a random number generator, which is used as the gold standard to compare the randomness of different Mask settings. Some Mask settings in the new design are tested but are not able to generate 330,000 peptides. They are not used in the subsequent studies.*

### PI distribution analysis

PI value of each peptide is calculated using the ProtParam tools in Biopython (Gasteiger, Hoogland et al. 2005). Distribution is obtained by normalizing the values in each setting with mean of 7 and standard deviation of 1. A distribution figure is generated using SAS software. The difference index is calculated with the formula above and figure generated in Excel.

### Pentamer coverage calculation

Each peptide is dissembled into continuous pentamers. All pentamers from the same Mask setting are analyzed using R to retrieve the percentage. Note that the total number of all possible combination changes with the total number of amino acids used.

### Amino acid position bias calculation

Sequences are imported and percentage calculated in SAS. Peptides are aligned at the N-Terminus. Missing values are introduced at position more than 10 amino acids away from N-Terminus and are disregarded during analysis. Data are then imported into Excel to make the graph.

### Blast experiment procedure

Peptides are blasted against Nr database using the blastp program offered by NCBI (Johnson, Zaretskaya et al. 2008, Madden 2013). The command used to blast is attached below:

```
“blastp -db nr -query input -out output  -outfmt "6 qseqid sgi sacc evalue length nident" -task blastp-short -gapopen 10 -max_target_seqs 100 -num_threads 12 -evalue 10000”
```

Sequences are required to give at most 100 output under e-value of 10000. All output from the same Mask setting are imported into SAS. The frequency of each protein is calculated and then matched with their lengths. The results from different Mask settings are compared to retrieve the one million proteins that have the largest standard deviation of frequency. Data are at last imported into JMP Pro 10 to make the graph.

### Mask deletion experiment

The mask file is imported into R to perform the mask deletion and random selection. The subsequent experiments are carried out using the same methods as above.

## Result

### Testing in-silico produced peptide characteristics

Physical characteristics of the peptides were tested to see whether we can mimic the performance of pure random peptide set. Pure random set contains peptides generated using random number generator. We examined molecular weight, isoelectric point (pI) and hydrophobicity all of which are generated using the ProtParam tools with Python. Distribution of molecular weight and hydrophobicity are the same for all Mask settings. PI distributions are able to illustrate the difference between different Mask settings for both old mask generation method (Figure 2.1) and current new mask generation method (Figure 2.2). Basically, with the decrease of total Mask number, the pI distribution deviates more from the pure random set, which is considered the standard of best performance. But overall, the old Mask generation method yield distribution that are less like random set while the new generation method is able to keep similarity to the random set. This means the information loss accompanying reduction of mask number is significant in old mask generation method but is kept at minimal level for new mask generation method.

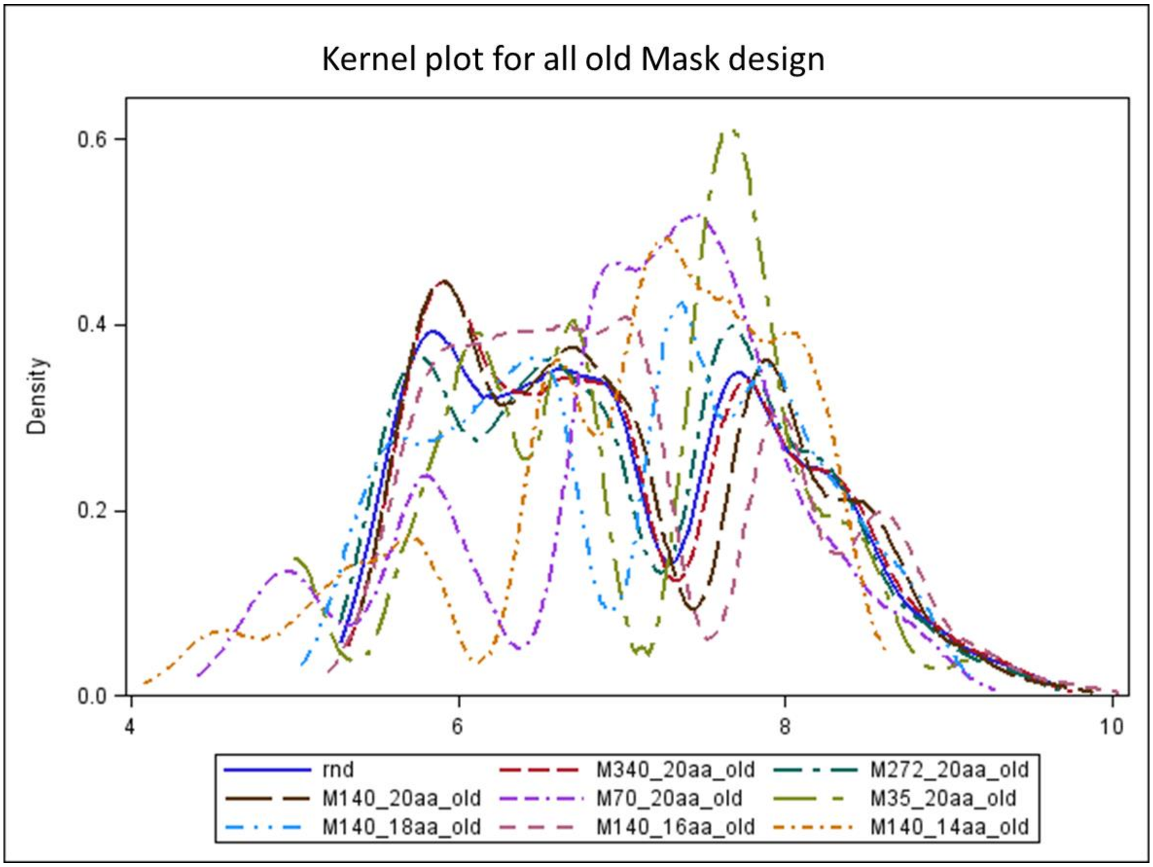


**Figure 2.1. pI distributions for designs using old Mask generation method show large variation.**

*Designs using old Mask generation method with different parameters are generated and pI for each design is calculate. The figure shows various designs have distinctive pI distributions. As the mask number is reduced, the distribution becomes more and more different compared with the random set distribution.*

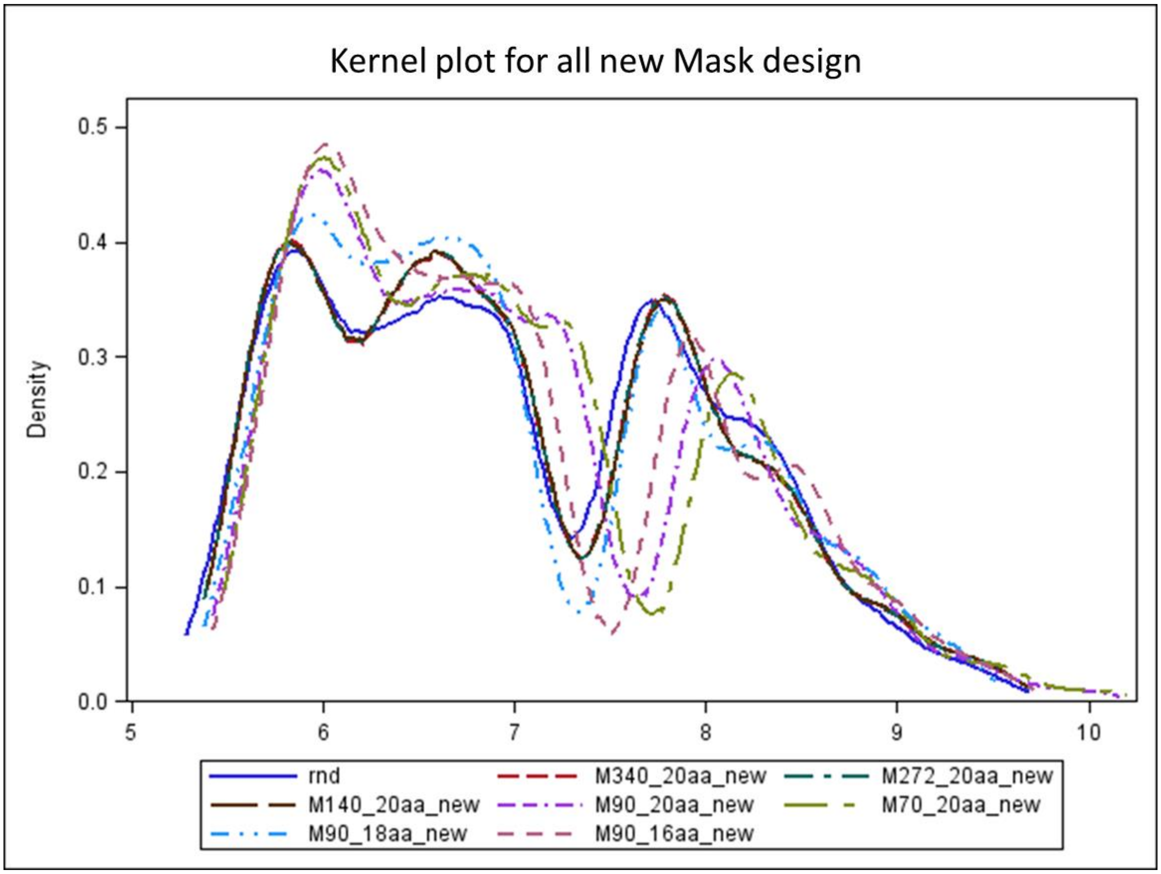


**Figure 2.2. pI distributions for designs using new Mask generation method show large variation.**

*Designs using new Mask generation method with different parameters are generated and pI for each design is calculate. The figure shows various designs have similar pI distributions. As the mask number is deduced, the distribution has minimal changes compared with old Mask generation method.*

Figure 2.3 shows the difference index change with total Mask number and amino acid choices. Difference index is measured using the equation below:

$$\text{Difference index} = \sum abs(\lg(pI(set1)\,/pI(set2)))$$

Abs is absolute value. PI is the isoelectric point.

The pI of each peptide in one set is ranked from low to high and normalized with mean of 7 and standard deviation of 1. Each pI with the same rank in two sets are compared to get the difference index. The more similar the two distributions, the smaller the difference index will be. We compared all Mask settings to the pure random set, each with three replicates. The quantified result shows the same trend. When decreasing the total number of Masks or amino acid choices, the difference index increases. However, the index increases much more in the old Mask algorithm than in the new Mask algorithm.

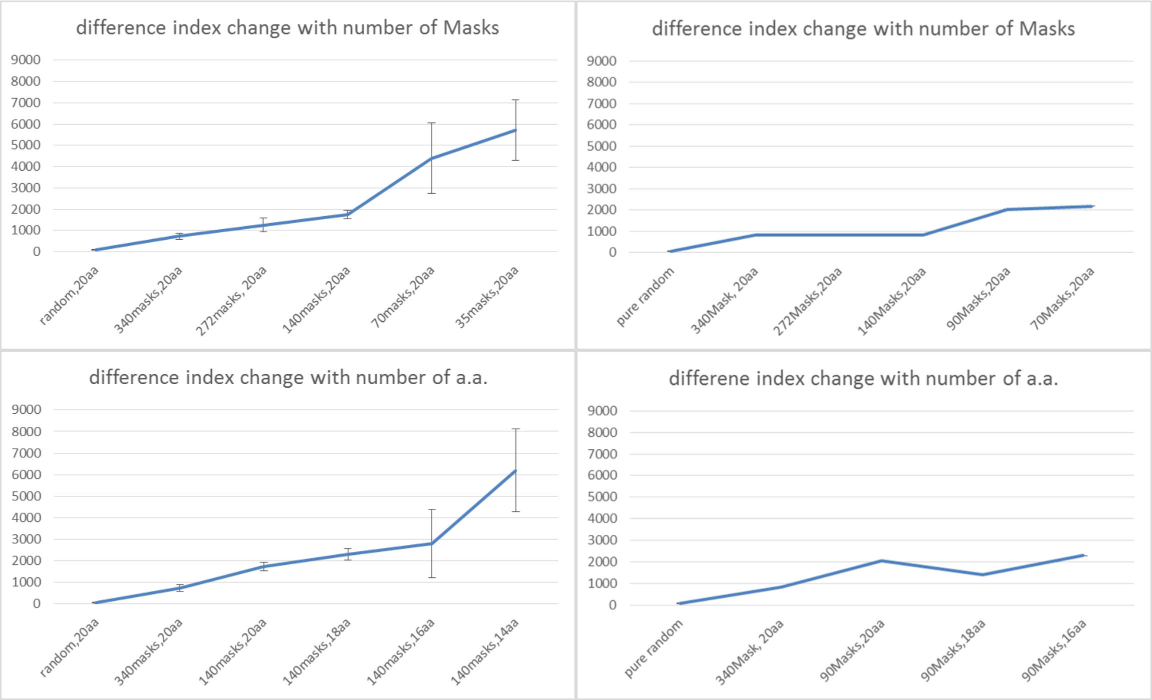


**Figure 2.3. Difference index of pI distributions for designs using different Mask generation method.**

*Each Mask design is compared with the pure random set to calculate the difference index. As the number of Mask or amino acids is reduced, the different becomes more significant for both old and new generation method. However, the new Mask generation method remains more similar to random set compared with old Mask generation method. And the stability is also increased for new Mask generation method, as it is shown in the figure, old designs have large standard error.*

## Testing Random Space Coverage

Immunosignature is used to capture the antibody activity in human. Since there is a large pool of antibodies, which can bind to almost any possible sequence, we want to make sure on our immunosignature the epitopes for every antibody exists, along with their mimotopes, so that we can capture all possible antibody composition in that specific sample. By covering random space, we mean to cover all possible combination of amino acids, which will require infinite number of peptides to accomplish. And since we are limited by manufacture consideration, only certain number of peptides will be used. Through calculation, all tetramers can be covered multiple times. And all pentamers can be covered once if no pentamer is highly repeated. Hexamers can be only covered for a small portion. Effort is made to optimize the pentamer composition for the new Mask design. No optimization of this kind is performed for the old Mask design. Figure 2.4 shows the pentamer coverage plot for all the Mask designs. When decreasing the total Mask number, the system becomes less complex, more pentamers are missed because not many choices are offered. When decreasing the total number of amino acid choices, since the total number of possible pentamers are fewer, which is n^5, n is the total number of amino acid choices, the total coverage percent increases significantly. By comparing the old Mask design and the new Mask design, with the same Mask number, the new Mask design always perform better than the old one.

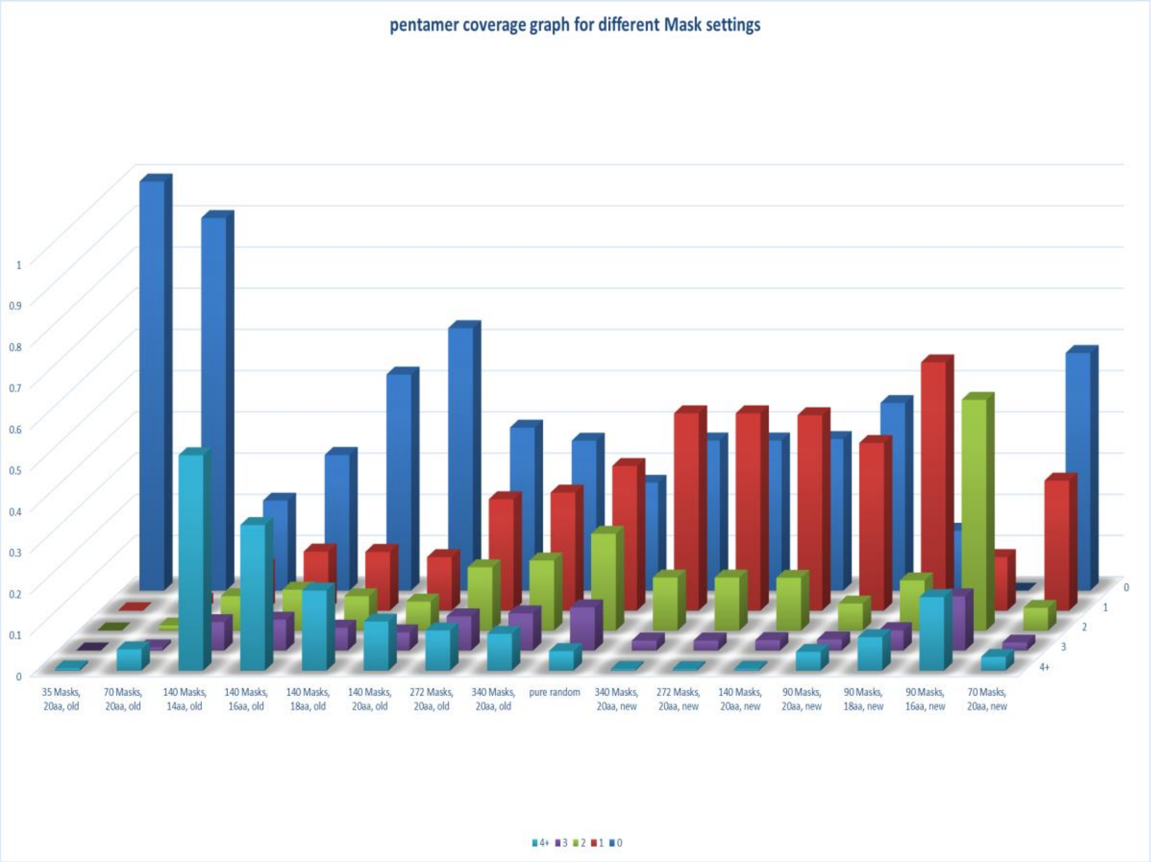


**Figure 2.4. Pentamer coverage graph for different Mask designs.**

*Sequences from each Mask design library are cut into pentamers and counted distribution. New Mask generation method generally can represent more pentamers compared with old Mask method with same parameters.*

### Testing Amino Acid Position Bias

In order to get an unbiased result, each amino acid should be represented evenly at each position of the peptide from the N-Terminus to the C-Terminus. When using less Masks in the old Mask design (shown in Figure 2.5), bias becomes more obvious. It can be generated at any position and at any amino acid, adding instability to the system. However, for the new Mask design, bias is minimal compared to old design. Bias begins to appear at 90 Masks and only appear at the C-Terminates of the first few amino acids.

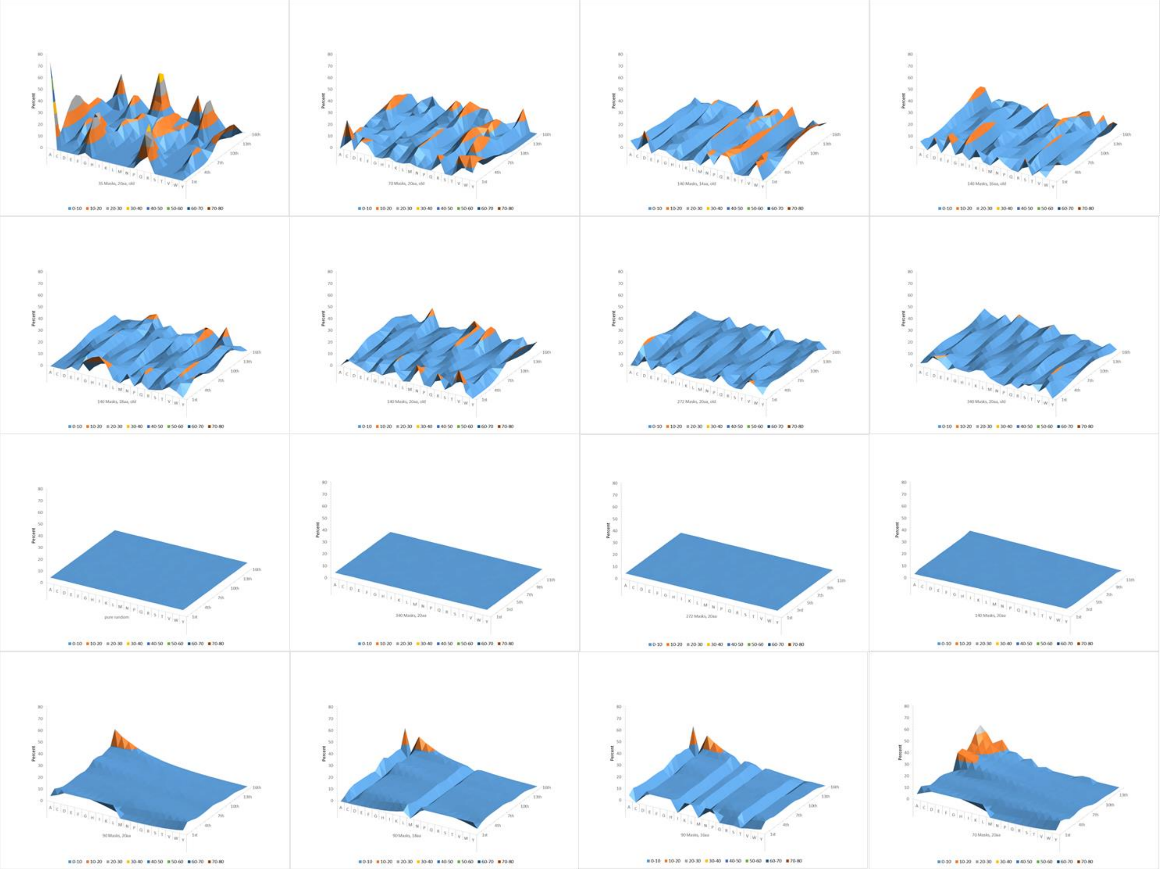


**Figure 2.5. Amino acid position bias for different Mask designs.**

*X axis are the different Mask settings, with 20 amino acids within each Mask setting. Y axis are the position of the amino acid on the peptide from the N terminates to C terminates. Z axis is the percentage of occurrence of the amino acid at specific position. Some amino acids within certain Mask settings are 0% at all position because they are not utilized in that setting. First two rows are designs with old Mask generation method with increasing number of Masks. Lower two rows are designs with new Mask generation method with decreasing number of Masks. New Mask generation method overall has lower amino acid position bias compared with old design.*

## Testing Natural Space Coverage

Since antibodies are mostly targeting proteins in the nature, we also tested the random space coverage of each Mask setting using the blast program. Each peptide in a specific Mask design is blasted against the NR (Non-Redundant) database of NCBI and 100 matches retrieved or all matches below e-value of 10,000, whichever is smaller. All the output from a Mask design should represent the natural space, or biological space that set of peptides can cover. Because we want to capture all possible antibody activity, better natural space coverage should be optimal. And because to make sure the peptides are random, which means they should be not be biased towards certain patterns or sequences, we also measured the correlation of the number of times a protein was hit during the blast search with its length. If the peptides are random, the proteins should be hit with a length-dependent manner: the longer the protein, the more times it gets hit.

In Figure 2.6 we can see most Mask settings are almost the same and behave in a length-dependent manner. The 35Masks setting in the old Mask design is the only exception. Many short proteins are hit much more times than longer ones, indicating that in this setting, the peptides are looking for specific pattern of sequences and the sequences are probably far more similar to each other than to potential proteins.  If a given protein has that particular overrepresented pattern, it will get hit that protein many times. If a protein does not have the pattern, even if it is very long, it is unlikely to be hit.

Notice the transition zone from 140Masks, 20aa setting in the old Mask design to 35Masks setting in the old Mask design. The light blue region becomes broader as the total number of Masks or amino acid choices become smaller. This suggests that the blast program is less sensitive to length. And less sensitivity to length indicates the peptides are becoming less random.

For the coverage of Non-Redundant (NR) database, the complexity of the setting positively correlates with the coverage percentage well. The more Mask used and more amino acid choices, the better the coverage. Notice that both in the old and new Mask setting, the setting using 18aa is always performing better than the setting using 20aa with the same number of Masks.

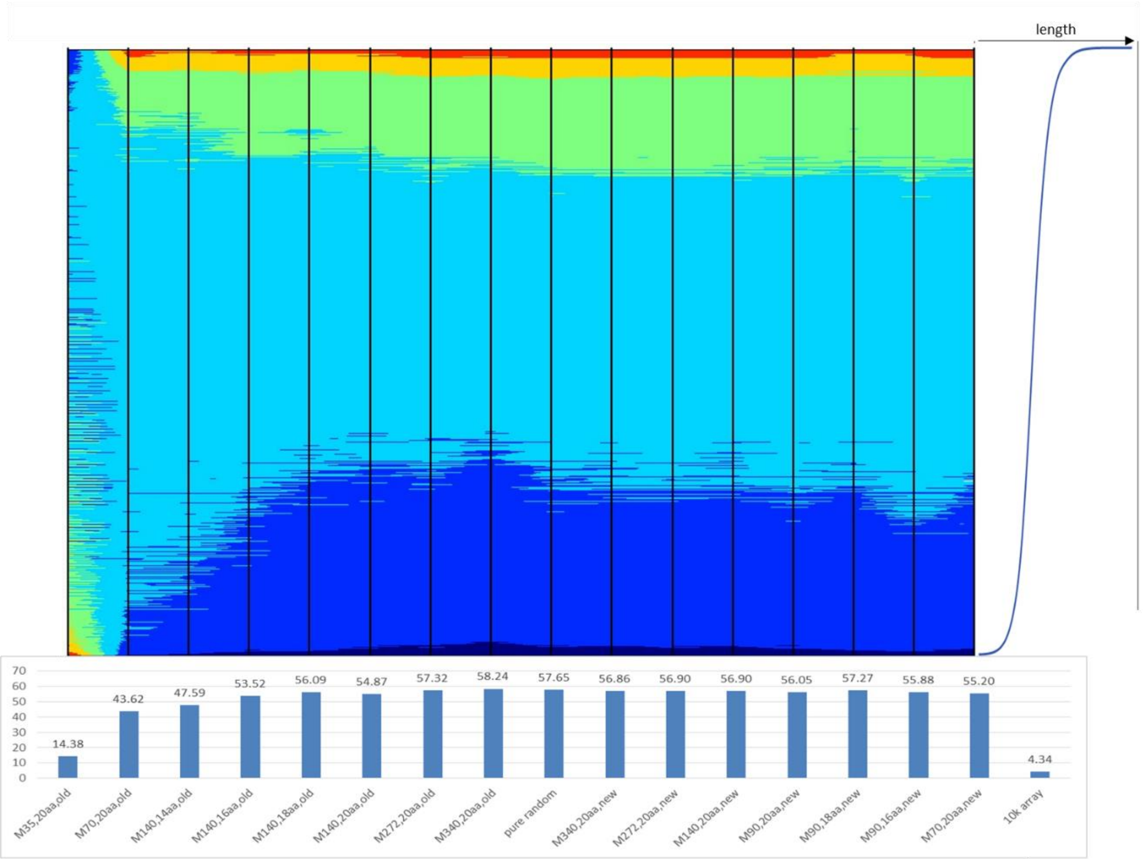


**Figure 2.6. Blast experiments against NR database.**

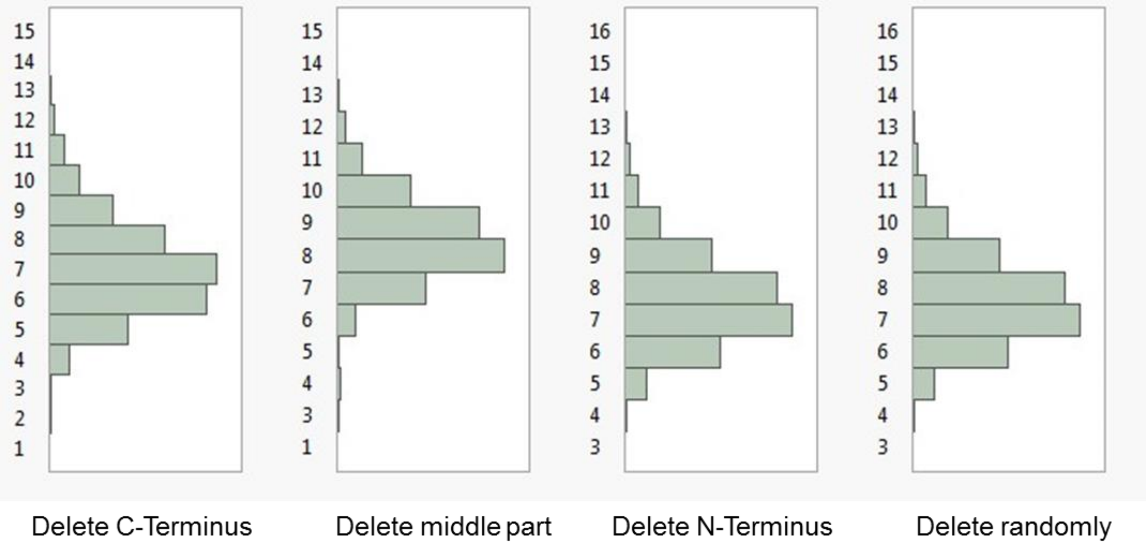
*This graph shows the analysis of the output from blasting all peptides within each Mask setting against the NR database. X axis are the designs. Order the same as figure 2.6. Y axis are one million proteins selected from NR database ranked by their length from short to long. Color in the graph represents the number of times the specific protein is hit during the blast search within each Mask setting (more hits from colder to warmer colors). The histogram below is the percentage of hit proteins within the NR database. Rules for selecting the one million proteins is they must have the largest standard deviation for the value, which is represented by color.*

## Defining roles of Masks at different position

The 90 Masks are carefully designed with sequential order. As a result, Masks at different position should have different functions and we could expect random sampling of the Masks in a different order will distort the design and result in huge performance decrease in all measurements.

The Mask setting that will be used in this experiment is the set that will be used for our next generation immunosignature platform, which includes 90 Masks and uses 18 amino acids (excluding C and M).  To test the function of Mask at the C-Terminus, N-Terminus and the middle, we delete the corresponding Masks to test the effect, leaving 60 Masks in each subset. Also, to test the effect of random sampling, 60 Masks are randomly selected and placed in random order.

The peptide sets are then used to retrieve their length distribution and pentamer coverage as before. Results of length distribution is shown in figure 2.7. The pentamer coverage result is listed in table 2.2. From the result, it is easy to see that each part of the masks has distinctive roles. The N-Terminus Masks are used to balance the amino acid position bias at the N-Terminus as they were designed to be. The middle Masks are used to extend the pentamer coverage. And the C-Terminus Masks are used to offer pentamer coverage to some level and ensure peptides meet the minimum length requirement. As can be expected, random sampling of the masks cause the design to fail dramatically.



Delete C-Terminus

Delete middle part

Delete N-Terminus

Delete randomly

**Figure 2.7. Length distribution after deleting specific parts of Masks.**

*The same Mask setting is used to generate the 4 (6) subset. The length distribution is shown in the table. Deleting Masks near the N-Terminus results in significant reduction in length of peptides.*

| Peptide set | Pentamer coverage |
|---|---|
| Delete C-Ter | 39% |
| Delete mid part | 33% |
| Delete N-Ter | 45% |
| Random set | 17% |

**Table 2.2. Pentamer coverage after deleting specific parts of Masks.**

*The same Mask setting is used to generate the 4 (6) subset. The pentamer coverage is shown in the table. Deleting Masks near the mid part results in significant reduction in pentamer coverage.*

## Discussion

*In silico* experiments can be extremely useful in determining how constraints imposed upon a random-sequence generator affect the peptides. In order to reduce mask cost and manufacturing time for creating an in situ-based peptide microarray, we examined methods to reduce these parameters while still producing a 'random' sequence peptide. These methods are not necessary when creating an epitope array, since the sequences are predetermined and must be created in the original order. For random sequences however, the number of masks can be reduced yet the sequence of resulting peptides can be pseudo-random. For reduced masks, the random number generator suggests a particular amino acid for a particular position in a particular peptide. Two Mask designs are currently available. Older design was used to generate CIM 330k version 1 chip. And newer design will be used to generate CIM 330k version 2 arrays. In the old Mask design, total Mask numbers are pre-assigned and then 17 Masks among them will be random selected. Each Mask will be randomly assigned an amino acid. For the new Mask design, pure random sequences are first generated. Total Mask number will be assigned and sequences will be tested to fit into the Masks. Some peptides can fully fit into the Pre-assigned Masks and some can only fit partially. Sequences with fitted length of less than 10 will be discarded. Other requirements the peptides need to meet to be incorporated into the candidate list are: each amino acid cannot be over ten percent at the N- terminus and each new peptide must present a certain number of new pentamers. When decreasing the total Mask number, the effect would be the complexity of the peptides will be reduced since there are less choices for which amino acid can appear at which location, but how much and what is the effect? We examined several Mask settings in both the old and new Mask design along with a pure random peptide set.

Before we created any real peptides, we examined the physical characteristics of resulting peptides from our algorithms, including isoelectric point (pI), molecular weight and hydropathicity. Molecular weight and hydropathicity do not illustrate too much difference between different Mask settings. The result of pI distribution shows profound difference between different settings and is given much investigation. Typically, when using the pI distribution for the pure random set as a baseline, more deviation occurs when less Masks or amino acids are used because there is a bias imposed by the lack of choices in amino acids and positions. The final peptide design is increasingly biased as masks are reduced and generate pseudorandom sequences. Some information in the immunosignature assay is lost because of the bias, as demonstrated by binding and analyzing nearly 300 different monoclonals. Within each Mask setting, the replicates can vary a lot, far more than when more Masks are used, indicating the overall system is less stable and bias can be generated in disparate directions. What is desirable is the set that shares the same distribution as the random set and yet use the minimum number of Masks. We found that using more amino acids does not necessarily guarantee a better distribution. In the new Mask design, the set using 90 Masks and 18 amino acids has more similarity to the set using 90 Masks and 20 amino acids. This should be the result of bias correction introduced by limiting amino acid choices. When limiting the number of Masks, certain biases can be introduced. However, when limiting amino acid choices, another kind of bias is introduced, and it seems in this case, the second kind of bias serves as correction for the first bias, making the total distribution more similar to random. There is a biological impact of restricting amino acids, though.  Fewer amino acids means fewer perfect matches to existing proteins.  Most proteins take advantage of the full set of naturally occurring amino acids.  By restricting the number of amino acids and reducing the mask number, we reduce the amino acid/position bias but impose a less ‘total-variability’ peptide, which is fine if the universe used only those amino acids.  However, in the natural world, that restriction must have some impact which at this point is unknown.

Up to this point, we have worked on the first generation 330k array. The setting we chose is the one using 140 Masks and 16 amino acids, excluding C, Q, M and I using previous immunosignature data as a guide.  Although the average performance for this setting is not perfect, since there is a very large error bar across three replicates, we can still get one set that performs well. And that is what we did: generating many sets and choosing the best among them. Also, from the antibody experiments, a big improvement of performance happens at changing from 70 Masks to 140 Masks. Based on this evidence, we made our first generation 330k array, which is been replaced by the second generation of 330k array using the new mask algorithm.

When constructing the new algorithm, more requirements are considered in order to achieve better performance. In the old algorithm, there is actually no upper limit and no lower limit on the number Masks, except that the Mask number need to be larger or equal to the peptide length. There is no selection of sequences generated, replicates can occur in the same set of peptides. And since all the sequences are generated randomly within the given Mask setting, what we will get are sequences with normal distribution. What we actually want is uniform distribution, where all sequences are represented equally. So sequences in the new design are not purely random by definition, but are biased in a way like uniform distribution. New selection criteria are used to meet the need.

Since we want the sequences to spread out and we only have limited number of sequences, it is not possible to represent all peptides. What would be a logical idea is to optimize in order to represent certain n-mers. If n is too big, we cannot get good coverage. We decided to optimize the representation of pentamer space, because with 330k peptides, we can cover almost 100%, theoretically. These studies show pentamers are most important for antibody binding (Rubinstein, Mayrose et al. 2008, Sun, Xu et al. 2010, Kringelum, Nielsen et al. 2013).

In the pentamer coverage graph, the distribution for the pure random set clearly shows a normal distribution with large deviation. Note that the percent of zero occurrence pentamer is very low. So the overall performance of this setting is not bad. However, we can change the distribution to make it more like uniform distribution, where the distribution has a much smaller deviation. In that way, there will be fewer unrepresented peptides and there will be fewer over-represented peptides. And as can be expected, the percent of unrepresented pentamers goes higher when limiting the number of Masks. And with the same Mask number, the percentage goes lower when limiting amino acid choices, because the total possible combinations are fewer. However, for the settings in the old Mask design, the distributions are skewed towards the unrepresented and over-represented pentamers, which is opposite to what we planned. Too many over-represented pentamers also restrict other pentamers to be represented. However, this is what we can do with the old Mask design and we have to choose one to build the first generation 330k array. For the new mask design, since we are requiring new peptides to present new pentamers, what can be expected is under-represented pentamers are always dominating throughout the Mask settings. Over-represented pentamers are never significant among any setting. The distribution from 340 Masks to 140 masks are the same, skewed towards under-represented pentamers with a small deviation, which is exactly what we want. Information begins to change at 90 Masks. We are able to represent all possible pentamers when limiting amino acid choices to 16 and represent ~80% when limiting to 18. Using 16 amino acids seems to be the best choice in this experiment. However, as it is shown in the pI distribution and the following experiments, deleting too many amino acids is not the optimal choice. Using 18 amino acids should allow a high level of coverage yet without losing too much information.

For the amino acid position bias, as stated in the result part, bias can be generated at any position and at any amino acid in the old Mask algorithm and can only be generated at the starting position of C Terminus and only for the first few amino acids. And the bias can be extreme in the old Mask design, while the bias is only minimal in the new design. So overall, the new Mask design has almost no preference to any amino acids at any position, which makes the new algorithm far more superior in eliminating bias compared to the old design. The overall trend for both designs is still more bias with fewer Mask number.

For the new Mask design, there is no obvious bias at the N terminus because we are restricting each amino acid to be less than 10% arbitrarily. Without that restriction, similar bias can be expected at the N terminus like the C terminus. Notice that biggest bias also occurs at the termini. When investigating what might be the reason for this, some innate shortcoming of the Mask design were discovered. When designing the algorithm, we thought using 340 Masks allows all amino acids to appear at each position, which should mean it is purely random. However, as shown in the Figure 2.5, the old design using 340 Masks still contains bias, albeit small. This makes it different from the purely random set, where there is no bias at all. This is because in the old Mask design, sequences are generated within the Masks.  Although all possible peptides can be generated, they are not of equal probability. This doesn't happen in the new Mask design because the sequences are generated *a priori* using a random number generator without the influence of Masks. When trying to fit the peptides into the Masks, there will be no problem with higher Mask numbers, as shown in the Figure 2.6.  No bias is generated from 340 Masks down to 140 Masks. However, when with low Mask numbers, bias still exists because we wish to fit in at least 10 amino acids into the given Masks, and peptides starting with the first few amino acids have higher chance to pass these criteria, for the same reason as above.

The reason for the bias from the old Mask design is because we are randomly assigning amino acids to each mask instead of assigning amino acids in a specific order. We want everything to be as random as possible in the old Mask design. When the mask number is high, there is not much problem. When the Mask number goes down to lower ones, the sample size is too small that the outcome is usually unpredictable. This should be the main reason for in the old Mask design, large error bar and deviation is shown among replicates.

Although we designed the program, we do not know exactly what the roles are of the masks at different position. By deleting the corresponding masks, we can see the effect of losing those masks and know their functions. From the result, deleting the middle masks yields the longest peptides, indicating there are the fewest amino acids generated using those masks. While deleting the C-Terminus masks yields the shortest peptides, indicating most amino acids are added in those masks. When looking at the pentamer coverage, deleting the middle masks yields the worst pentamer coverage, indicating those masks are crucial to supply the diverse pentamer coverage.

Overall, this chapter represents a method that can be used to generate peptide sequences for Immunosignature. Performance comparisons are made between old and new Mask generation method. The new Mask generation method is superior in all scoring metrics. The improvements in peptide library sequences will enable Immunosignature platform to perform better on distinguishing diseases, since the possibility to catch more antibody binding. And in the following chapters, Immunosignature arrays generated from new Mask algorithm are both used.

# CHAPTER 3

# A COMMON ANTIBODY RESPONSE IS INDUCED BY A WIDE VARIETY OF HUMAN PATHOGENS

## Abstract

An infection is managed by both an innate and an adaptive immune response to the pathogen. It is thought that native antibodies present at the time of infection are a component of the innate response and may play a role by retarding the pathogen (Ochsenbein, Fehr et al. 1999). This delay allows the second arm, the adaptive response, to be activated and evolve to contain the infection (Medzhitov 2007). We have discovered a third arm of the antibody response to infection. We find that 12 different pathogens, including viruses, bacteria and eukaryotes, induce a common set of IgG reactivity. This response was discernible using the immunosignature technology which entails profiling sera antibodies on high-density (125-330k features) peptide arrays (Stafford, Halperin et al. 2012, Sykes, Legutki et al. 2012). The peptides are chosen from random sequence space to maximize chemical diversity. Using sera from 405 infected and non-infected people we find that almost all the infected samples can be sorted by pattern from non-infected people. A signature that separates a single infection type from non-infected consists of both the common signatures and the specific adapted signature. The common signature peptides can be used to separate any other infection from controls. A common signature is not evident in comparison of 4 cancer types to non-cancer subjects. A comparison of the peptides in the common signature to the Immune Epitope Database (IEDB) identified 44 amino acid sequences that are shared between many pathogens in the IEDB and are in the

common signature we identified (Vita, Overton et al. 2014). We propose that viruses, bacteria and eukaryotes that have evolved to become a human pathogen elicit a common IgG antibody response to a limited number of shared epitopes. This common response may, like the native antibodies, serve to modulate the infection in the early stages until the specific adaptive response matures.

## Introduction

Antibodies play a key role in the adaptive immune system. Each time the host is infected with a pathogen and the innate immune system fails to clear the invader, stimulated progenitor B cells followed by short-lived and long-lived plasma cells will produce antibodies that bind to a pathogen and offer partial or in some cases, neutralizing protection (Medzhitov 2007). It is logical then that with each exposure, antibodies will be produced specifically for that pathogen. Subsequent cross-reactivity are usually regarded as imprecision of the immune system. However, there has been no systematical study to test what the general limit is of antibody cross-reactivity or if there is any biological relevance of such phenomenon, mainly because there is no appropriate platform with which to study general cross reactivity .

Immunosignatures are patterns of reactivity between serum antibodies and random-sequence peptides. An immunosignature can detect differences between people based on their history of vaccines and cumulative environmental exposures, as well as differences based on HLA and other genetics of the humoral immune system. It can also detect common reactivity in people exposed to the same pathogen (Chase, Johnston et al. 2012, Hughes, Cichacz et al. 2012, Malin, Kovvali et al. 2012, Restrepo, Stafford et al. 2013, Sykes, Legutki et al. 2013, Stafford, Cichacz et al. 2014). Immunosignatures are inherently multiplexed: they contain enough signals that cross-talk and signature overlap is rare. In one study, 14 different diseases were distinguished simultaneously (Stafford, Cichacz et al. 2014). Thus, this unbiased platform seems ideal to look for sequences that may be represented in many different pathogen exposures.

Here we present data that reveals the extent of cross-reactivity among many individuals' humoral immune response to 7 different pathogens. We included viral, bacterial, and eukaryotic parasite pathogens to ensure representation. We followed an analytical approach where no assumptions were made concerning the infected cohorts, no accommodation made for virus, bacteria or fungus even though the proteome sizes differ considerably, and no compensation was made for number of diagnostic peptides per disease. We asked whether there is a unique and common peptide motif that appears in patients exposed to human pathogens, and did not appear in healthy volunteers. We further asked whether any common signature appeared in cancer patients, and whether a common signature would appear in various pathogen proteomes, even those which were not tested in this experiment. Negative controls for human pathogens include plant pathogens, which would not be expected to share motifs with human pathogens if co-evolution was occurring. This study examines, for the first time, signals in the human antibody repertoire that may suggest that there are common antigenic signatures in human pathogens that may have co-evolved with humans. This new finding suggests new methods for developing broadly protective vaccines against multiple infections at the same time.

## Methods

### Materials

Human sera samples exposed to various pathogens were used. Table S1 shows the total cohort used in this study. Immunosignature arrays are manufactured in batches of 312. Each array is in situ synthesized, and consists of 125,000 or 330,000 random-sequence peptides with average length of 12 amino acids. Among these controls are single and double amino acid missense sequences, designed to identify improper sequence synthesis. Also, 250 blank spots are used to estimate local background and spatial variations in global background signals.

### Immunosignature assay

Sample buffer contains 3% BSA in 1x PBST, pH 7.3. Secondary incubation buffer contains 0.75% Casein in 1x PBST with 0.05% Tween20. Serum samples in 50:50 glycerol were diluted into sample buffer at ratio of 1:1500, then incubated on Immunosignature array with volume of 150ul for a final concentration of 1:750. Incubation was 1h at 37 $^{o}$C with rotation. Arrays were washed 3 times with 1x PBST and rinsed 3 times with ddH$_2$O. 4nM secondary anti-IgG antibodies conjugated with Alexa-Fluor 555 (Life Technologies, St. Louis, MO) was added to the secondary incubation buffer and then added onto entire Immunosignature microarray for a final volume of 2.5 ml to detect the primary antibody binding in the serum. The incubation is 1h at 37 $^{o}$C with gentle agitation, then slides were rinsed with blocking buffer, then washed 3x with 1x PBST and 3x ddH$_2$O then dried. Slides were then scanned at 555nm with Innoscan 910 scanner at 1.0um resolution to acquire the image. Feature intensities were extracted using the GenePix Pro 6.0 software (Molecular Devices, Santa Clara, CA).

### Statistics and Analysis

Analysis was performed using the JMP software (SAS Institute Inc.), R (CRAN repository) and python. Raw data is fetched from each GPR file output by GenePix and normalized to the median before analysis. Whole Immunosignature clustering is performed using all data points for all samples using the hierarchical clustering method. Ward is the distance measure between the samples (columns in heatmaps) and the peptides (rows in heatmaps). Two-tail Student's T-Test is used for feature selection, cutoff is set at either the top 50 or 100 peptides with the best p-value from T-Test. For each set of t-test, the p-value is controlled to be <1/330,000, allowing at most one false positive in 330,000 parallel comparison.

### Epitope identification

The algorithm used to identify significant epitopes is described in detail in (Richer, Johnston et al. 2015). The top 1000 peptides from T-Test obtained by comparing normal samples (control) versus all infected (case) samples are used to identify the epitopes. Epitopes are restricted to 5-mer sequences, ungapped. Once significant epitopes are identified, GLAM2 (http://meme-suite.org/tools/glam2) from the MEME suite software is used to identify the consensus (Frith, Saunders et al. 2008, Bailey, Boden et al. 2009).

### BLAST searches

BLAST (Basic Local Alignment Search) was used to identify matches in the pathogen proteomes. BLASTP by NCBI via web interference is used with default parameters other than not adjusted for short input sequences (the automatic adjustment for short input sequences yields search parameters that are still too relaxed for sequences as short as 5 amino acids), hitlist size = 100, gapcosts = 15 for existence and =2 for extension. Matrix is set to be PAM30 and word size is at 2. Expect threshold is set at 10^10 to ensure we will have desired number of output. Entrez Query is set with "all [filter] NOT predicted [title] NOT hypothetical [title]" to remove predicted and hypothetical proteins. Note that here the E-value is not important, because the input sequence is short, so we will always hit sequences by chance, which is the definition of E-value. RefSeq database is used as the target database for BLASTP because of better annotation and less redundancy (Pruitt, Tatusova et al. 2005). The sequences from the 7 pathogens in the RefSeq database are used in this experiment. Query search against IEDB is performed by finding the exact match of putative conserved sequences (obtained empirically) in the database. BLAST search to identify enrichment of the sequences in the RefSeq database is performed using the BLASTP suite as described above, against all RedSeq proteins. The enrichment is measured by counting the number of unique hits in bacteria and eukaryote and obtaining the percentage of output from bacteria and virus. This information is generated from the BLAST results page from the taxonomy report. Blast search against IEDB and plant pathogens in Figure 6 is performed by using the blast command line program. For each input peptide, the number of matched sequences is recorded. Then a group-wise comparison is performed between the 500 peptides from the disease common signature and

500 randomly selected peptides by T-Test and non-parametric tests. The plant pathogen database is retrieved from Comprehensive Phytopathogen Genomics Resource (Hamilton, Neeno-Eckwall et al. 2011), containing 82 pathogens.

**Result**

The immunosignature diagnostic platform has been shown to separate the immune responses of a variety of infections from non-infected sera samples, as well as different infections from each other (Legutki, Magee et al. 2010, Restrepo, Stafford et al. 2011, Restrepo, Stafford et al. 2012, Johnston, Thamm et al. 2014, Navalkar, Magee et al. 2014, Stafford, Cichacz et al. 2014, Donnell, Maurer et al. 2015). We first demonstrated that the samples we used (Table 3.1) were also distinguishable on this platform. In Figure 3.1, the samples from 5 different infections (BPE, HBV, Dengue, Malaria and Syphilis) are readily distinguished from each other using 500 peptides from the array as a classifier. These peptides are chosen based on their ability to distinguish each infection from the others.

| Group | Count |
|---|---|
| Borrelia | 8 |
| BPE | 12 |
| Dengue | 9 |
| HBV | 15 |
| Malaria | 13 |
| ND | 32 |
| Syphllis | 8 |
| WNV | 21 |
| Total | 118 |

**Table 3.1. Samples cohort used in this study.**

*Seven types of infections along with the normal donor control group are used in this study, with a total sample size of 118.*

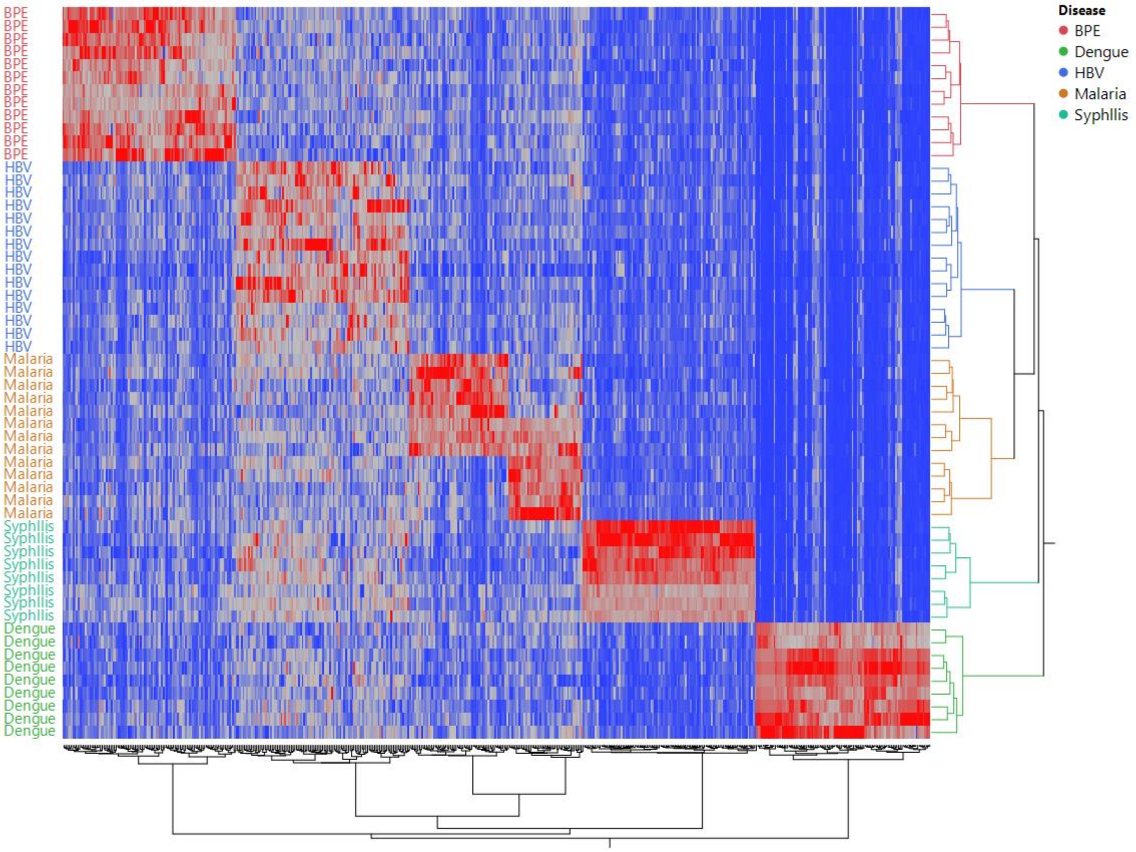


**Figure 3.1. Hierarchical clustering of 5 infections shows separation of each disease.**

*100 peptides are selected for each disease by One-versus-all T-Test comparison. 500 peptides are then combined for use in the clustering. Each disease has its own signature and is different from other diseases.*

The same array data was reanalyzed without separation based on infection type. All 8 sample sets in Table 3.1 were included. Two-way hierarchical clustering of the whole immunosignature with 330,000 features was performed. The result of this clustering (Figure 3.2) shows that most of the non-infection donors (blue label ND) can be differentiated from the 7 pathogens (red label DI) while the infection samples did not fall into obvious groupings by type of infection. To test the robustness of this observation, we performed the same type of analysis including different samples of the 8 groups in Table 3.1, adding 5 different infection types (Flu, HIV, Tuberculosis, Chagas, Valley Fever (a fungus)) and using a different array format containing 125,000 different peptides. As evident in Figure 3.3, most of the 12 different types of infection samples clustered separately from the non-infection samples.

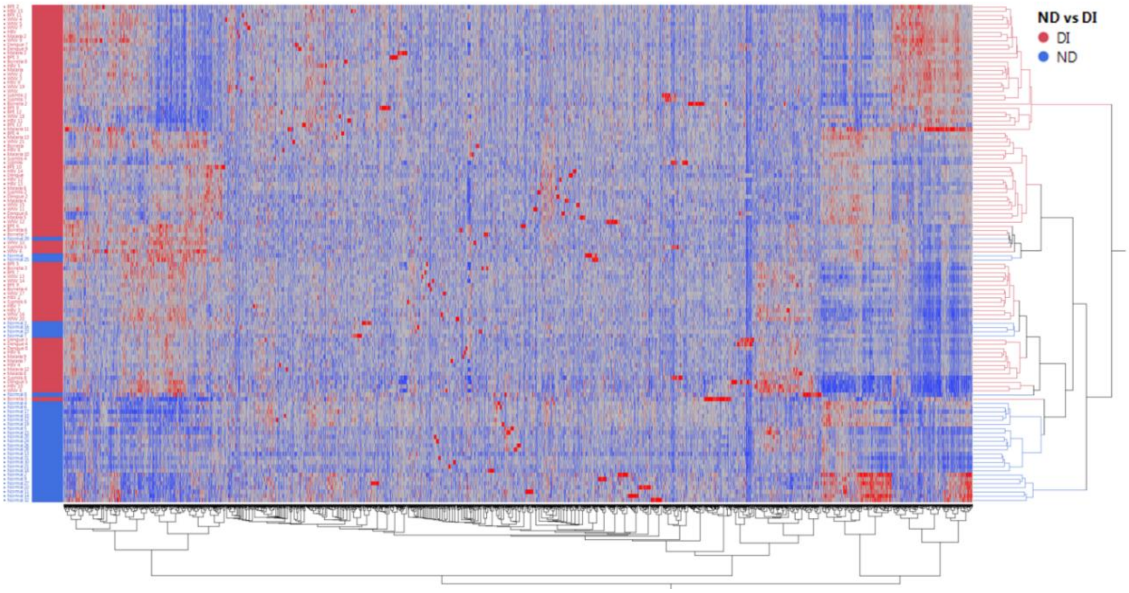


**Figure 3.2. Whole immunosignature clustering of 7 pathogens versus healthy donor.**

*Pathogens share red label indicated using DI. Healthy donors are blue indicated by ND. Samples are placed row-wise. All 330,000 peptides are shown in column-wise direction. Pathogens taking together can be clustered apart from healthy donor, while the pathogens cannot be differentiated with each other. All pathogens share large group of common signature responsible for this hierarchical clustering result*

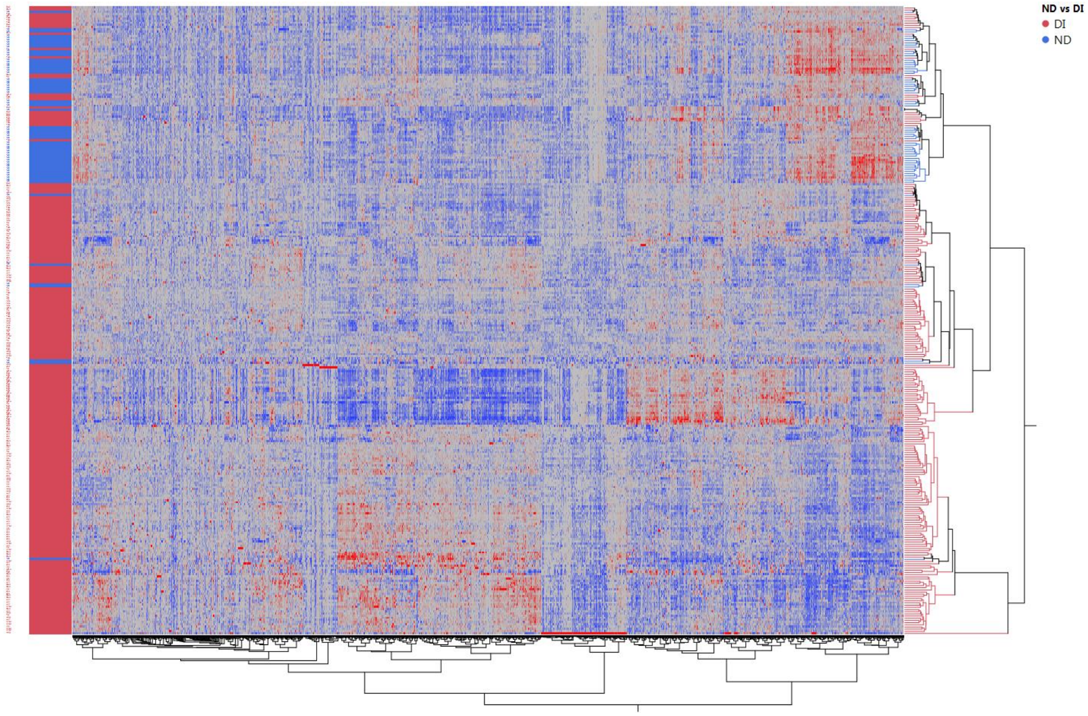


**Figure 3.3. Whole Immunosignature clustering of 12 pathogens versus healthy donor.**

*This analysis used a totally different samples from that in Figure 1, adding Flu, HIV, Tuberculosis, Chagas, VF infections and on a different Immunosignature array with 125,000 peptides to replicate the result as in Figure 1. The same clustering pattern is produced: the infections can be distinguished from the non-infected, while the pathogens are mixed together with each other.*

This analysis implies that very different infections elicit antibodies that bind the same peptides on the array. To test this concept from another angle we individually compared each infection sample set to the non-infection group and selected the top 100 peptides (by p-value) for each comparison.  Of the 700 peptides selected in this manner, 200 peptides appeared in at least two pathogens. These sequences were pooled and two-way hierarchical clustering was performed for the 7 infections and the non-infection samples. The results are presented in Fig 3.4a, showing that these peptides can also be used to separate all infections from non-infection samples.  Principle component analysis (Fig 3.4b) of this data shows that the first component accounts for over 50% of the variance and using only one component can repeat the same separation result as the clustering.

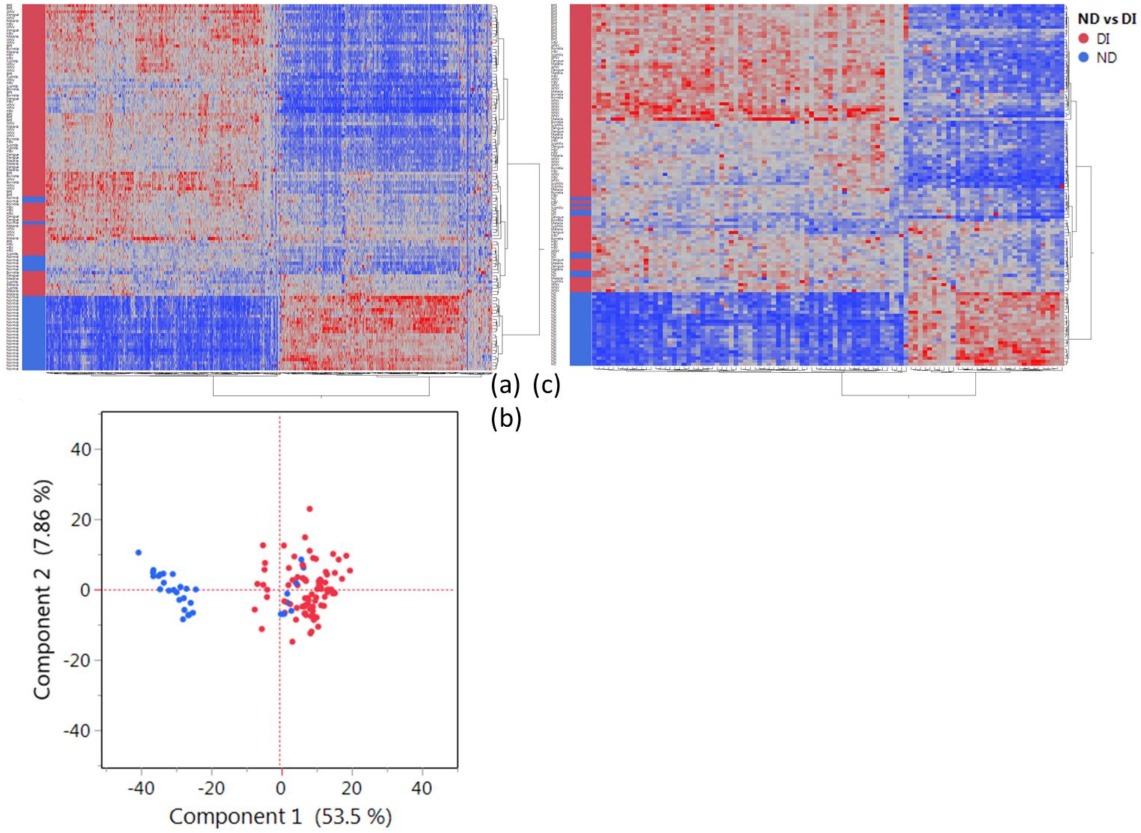


**Figure 3.4. Using selected peptides can repeat the separation of pathogens as a group to healthy donor.**

*(a)Peptides selected from pair-wise T-Test between each pathogen vs Healthy combined together shows separation between the 2 groups. (b) PCA analysis shows same separation and Component 1 accounts for over 50% of the variance. (c) Using peptides from T-Test between healthy donors with only one pathogen (BPE) can also separate all the pathogens from healthy together*

The implication from the results in Figure 3.4a, b is that a signature distinguishing any infection from non-infection will be composed of a common and a specific signature. To test this prediction, we used the 100 peptides chosen that distinguished BPE from non-infection as the basis to cluster the other 6 infection groups from non-infection. As shown in Figure 3.4c, even though these peptides were not chosen against the other six infections, they were very efficient in making the separation between them and the non-infection group. These data support the concept that there is a common set of IgG antibodies elicited by infections.

One possibility is that any disease would elicit a common set of antibodies. For example, there are many different types of cancer and they might also elicit a common signature, possibly the same as by infections. To test this, we analyzed the immunosignatures of 4 different cancers (breast, brain, multiple myeloma and pancreatic) in the same manner as we had for the infection samples. As shown in Figure 3.5, there was no clear clustering of cancer versus non-cancer samples.

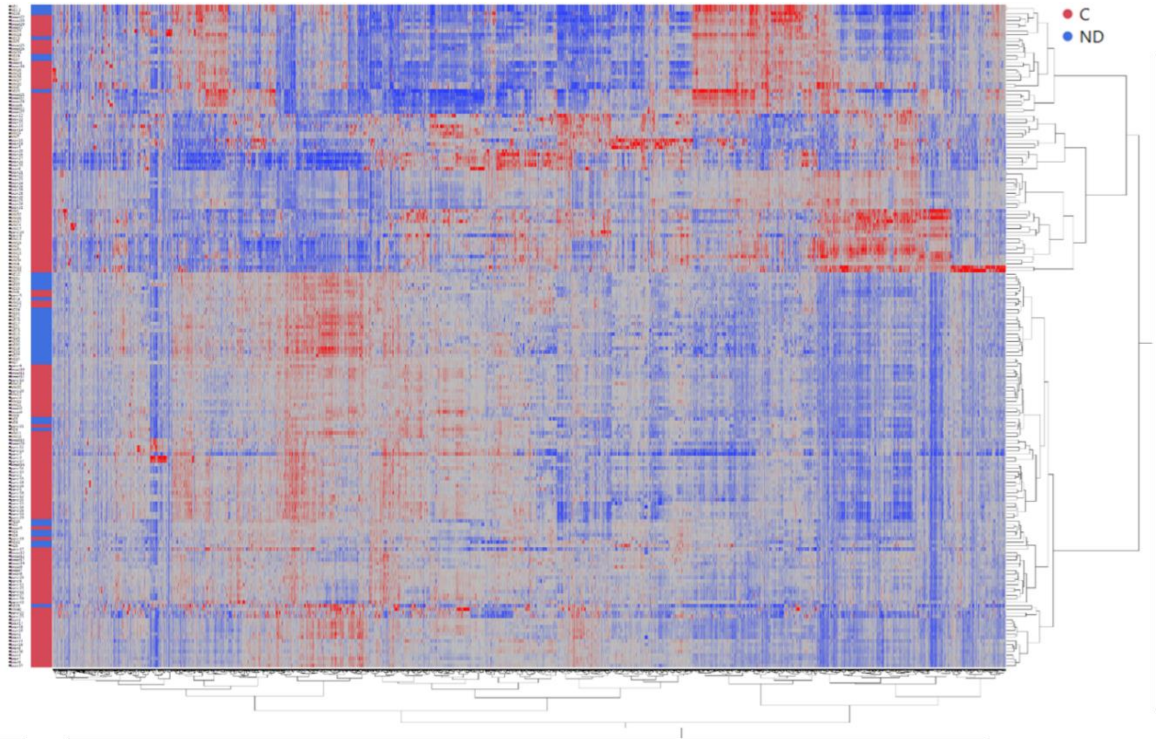
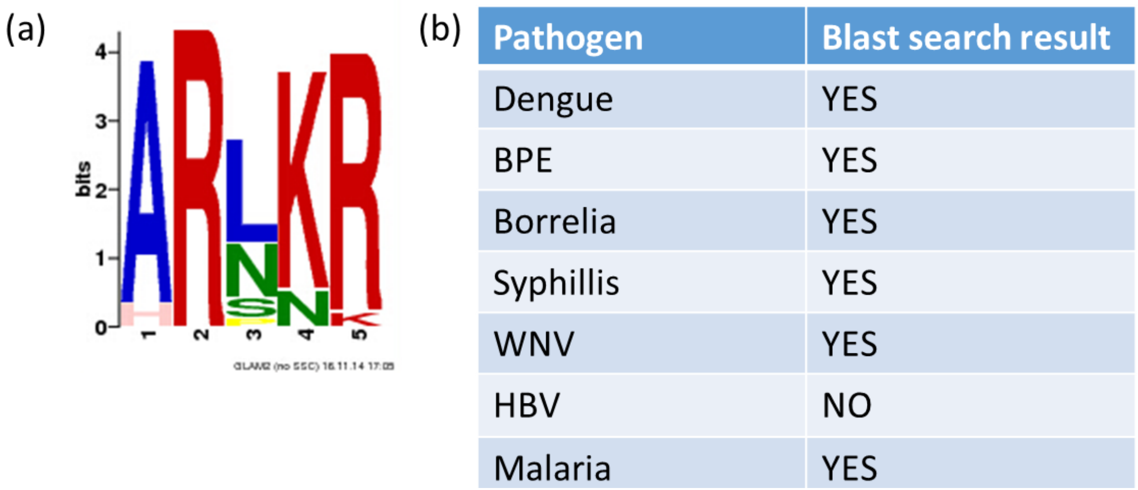


**Figure 3.5. Cancers cannot be differentiated from healthy using the same method.**

*The cancer antibody repertoire will either appear to be normal or different with equal probability. This suggest the immune system of 50% of the cancer patients are suppressed*

A common signature would imply that there are common epitopes in diverse pathogens that elicit an antibody response. The 330,000 peptides on the array used are on average 12aa long and represent approximately 50% of 5mer peptide space. The implication from the common signature is that these peptides would be related to actual pathogen protein sequences. We took two approaches to test this. First, we searched the common signature to identify series of enriched pentamers using methods described in Richer et. Al (Richer, Johnston et al. 2015). The enriched pentamers were then analyzed in GLAM2 to identify consensus epitopes (Bailey, Boden et al. 2009). One dominant epitope, ARLKR, was found (Figure 3.6a). This linear epitope was present in 6 of the 7 pathogens used, with hepatitis B virus the exception (Fig 3.6b). A second approach was to divide all the peptide sequences in the IEBD into pentamers. The IEDB is a database of verified epitopes in infections. A list of the top 2000 recurrent pentamers from the IEDB was compared to the peptides in the common signature. Fourty four pentamers were identified (Table 3.2). These peptides are presumably at least part of the link between the immune response to infection and the common signature.



| Pathogen | Blast search result |
|---|---|
| Dengue | YES |
| BPE | YES |
| Borrelia | YES |
| Syphillis | YES |
| WNV | YES |
| HBV | NO |
| Malaria | YES |

**Figure 3.6. Analysis of the common signature reveals dominant epitope that is enriched in pathogen space.**

*(a) ARLKR epitope was identified as the top consensus epitope after analyzing peptides from the common signature. (b) Blast the epitope against the 7 pathogens found the epitope in most proteomes.*

| | | | |
|---|---|---|---|
| AAGPP | KARRP | PAGDR | RPEGR |
| AGFKG | KGFKG | PDKEV | RPGFG |
| ANPNA | KRGSG | PGAKG | RPSQR |
| APKRG | KRPSQ | PKARR | RPSWG |
| ARHGF | LAGPK | PKRGS | RRPEG |
| FASRG | LGPKG | PPSQG | RSQPR |
| GKWLG | LNPSV | PSQGK | SNKGA |
| GPKGA | LPLGS | PSWGP | SQGKG |
| GPQGA | LSGKP | QRHGS | VHFFK |

| | | | |
|---|---|---|---|
| GSNKG | LSPRG | RGLFG | VYLLP |
| HFDLS | NKPSK | RGSGK | AGPKG |

**Table 3.2. List of the identified enriched epitopes from IEDB.**

*The top 2000 occurring epitopes from IEDB are extracted and tested on immunosignature. 44 epitopes are identified to be enriched.*

We propose that the common signature is the product of the proteomes of diverse pathogens being constrained by the human immune system. If so, one would predict that plant pathogens would not exhibit the same constraints (Jones and Dangl 2006, Király, Künstler et al. 2013). To test this, we first analyzed 500 sequences from the common signature with the highest p-values and 500 randomly picked peptides from the array not in the common signature. Each set was blasted against the IEDB peptides. As shown in Figure 3.7a, the common signature peptides had significantly more hits than the random peptides. This implies that the common signature peptides resemble the IEBD epitopes more than other peptides on the array. We then did the same type of analysis but blasting against a plant pathogen database (Hamilton, Neeno-Eckwall et al. 2011). Interestingly, the common signature peptides were significantly less similar to the plant proteins than random peptides. This may reflect that the plant proteome is also under sequence constraints, but different than from antibodies, due to interactions with plant hosts.

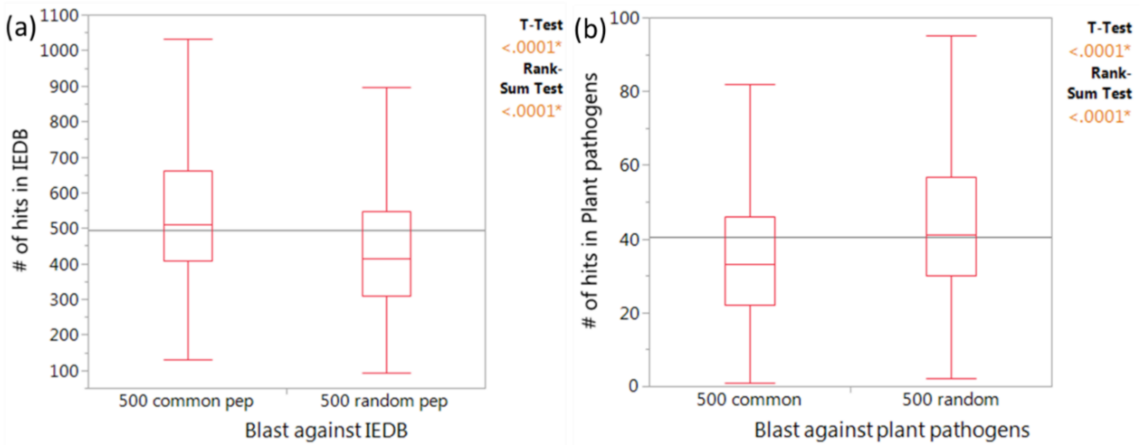


**Figure 3.7. Two sets of sequences blasted against IEDB and plant pathogens.**

*500 peptides from the common signature is compared with 500 randomly selected peptides. Peptides from the common signature shows more similarity to sequences in IEDB. When compared with plant pathogens, 500 common peptides are less similar to them than randomly selected peptides from the immunosignature.*

## Discussion

Other researchers have noted cross reactive antibodies. Natural antibodies, defined as having germline or near germline variable sequences, bind a wide variety of proteins (Notkins 2004), but are not induced on infection. Usually they are IgM class. In contrast, the common signature antibodies are IgG and are only in infected people. Others have noted cross reactive IgG antibodies (Warter, Appanna et al. 2012, Cywes-Bentley, Skurnik et al. 2013). For example, using protein arrays of *Yersina pestis*, Urlich and co-workers found significant cross reactivity with sera from other gram-negative infections (Keasey, Schmid et al. 2009). In at least one example, it was proposed to be caused by reaction to conserved proteins across the gram-negative bacteria. While it is possible there is overlap between previous array based cross reactivity and the common signature we think this is unlikely. The common signature is only approximately 2-fold above the signal in non-infected people, where the adaptive, pathogen specific signal is usually 10-100 fold higher. The immunosignature assay is 10-100x more sensitive than ELISA-type assays (Sykes, Legutki et al. 2012). This level of sensitivity is probably necessary to recognize the common signature.

The B-cells that produce the common signature could be germline cells, as for native antibodies (Ochsenbein, Fehr et al. 1999, Zhou, Zhang et al. 2007). There are native B cells in higher vertebrates (Ochsenbein, Fehr et al. 1999). However, they would need to be induced on infection. On the other hand, these B-cells could have been induced by previous infections and are reactivated on a subsequent infection. Isolation and sequencing of these B-cells should resolve this issue.

The existence of the common signature, and the common epitopes across most human pathogens that may induce them, has interesting evolutionary implications. One idea is that any persistent human pathogen must have these common epitopes. The antibodies comprising the common signature would constrain the infection enough to allow the host to mount a protective response. It would be beneficial for the pathogen so as to not kill the host (Cressler, McLEOD et al. 2016). In the simplest terms, to evolve to be a human pathogen the organism would have to produce the common signature epitopes. If not, it would kill the host too quickly. The implication is that new, highly lethal pathogens from other hosts may not have the common signature epitopes.

Finally, would this common signature have any clinical value?  We note that the level of these antibodies is low relative to the adaptive response.  The samples used in this study were from infected people with clinical symptoms so the common signature was not fully protective, though it may have moderated the infection.  However, it may be possible to augment the low response, by vaccination, to a level that is more protective. Such a vaccine could have broad value.

# CHAPTER 4

# THE IMMUNE PROFILE OF STAGES OF HEMANGIOSARCOMA CANCER IN DOGS CHANGES DRAMATICALLY

## Abstract

It has been amply demonstrated that different stages of a type of cancer can have very different transcriptional profiles. Even one type of cancer at the same stage can present variations in gene expression profiles. The conclusion, at least at the gene expression level, is that tumors are quite variable and the variation extends over time in the evolution of a tumor. We are interested how the immune profile of a cancer changes. The immunosignature technology permits this type of analysis. It involves reacting serum antibodies with arrays of 125K peptides chosen from random sequence space. We have investigated the immunosignatures of Stage 1, 2 and 3 of hemangiosarcoma (HSA) cancer in dogs. HSA is a leading form of cancer in dogs that is usually fatal. It arises in the blood vessels and the spleen and liver forms are highly metastatic. We find that it is possible to define an immunosignature that is diagnostic all three stages. However, we find that all three stages also have a distinctive signature with essentially no overlap of highly significant features between Stage 1 and 3. Further, the signature peptides at each stage present very different patterns over the other stages. Remarkably, the peptides at Stage 1 have much higher similarity to pathogen epitopes than those from Stages 2 and 3. Though these profiles are of antibodies as opposed to T-cells, they may reflect the evolution of the immune system with the tumors.

## Introduction

The oncogenic process evokes considerable and variable cellular changes relative to the tissue of origin. These features evolve in the lineage of a particular cancer, are evident in cancers of the same tissue source and can vary widely between different cancers (Ford, Easton et al. 1998, Reya, Morrison et al. 2001, Marusyk, Almendro et al. 2012, Lawrence, Stojanov et al. 2013, Meacham and Morrison 2013). Considerable effort has been devoted to relate these difference to diagnosis, prognosis and identifying therapeutic targets. The most useful form of characterization has relied on gene expression profiling, using microarrays or RNAseq (Nguyen and Rocke 2002, Wang, Gerstein et al. 2009, Young, Wakefield et al. 2010, Ren, Peng et al. 2012, Patel, Tirosh et al. 2014, Best, Sol et al. 2015). We are interested in expanding cancer profiles to the immune responses to the evolution of tumors.

Gene expression analysis of normal and cancer cells by microarrays, and more recently by RNAseq, has been the most informative aspect of characterization. The analysis of 1000s of tumors has shown that they can differ widely in their variance from the cells of origin (Weinstein, Collisson et al. 2013, Aran, Sirota et al. 2015, Andor, Graham et al. 2016). Hierarchal analysis of expression patterns has revealed subtypes, for example with breast cancer, that were not evident by classical histology (Ivshina, George et al. 2006). In some cases, the gene expression pattern can strongly correlate with prognosis or indicate a specific treatment. However, the gene expression is not useful in analysis of the immune response. While the specific expression of immune regulatory genes can be seen to vary in some tumors, this does not provide antigen specificity. Gene expression measures native genes while it is increasingly clear that it is the immune response to neo-antigens that is important in tumor evolution (Rizvi, Hellmann et al. 2015, Schumacher and Schreiber 2015). With the increasing importance of immunotherapeutics and vaccines in treating cancer it would be helpful to be able to measure the immune profile as broadly as has been done for gene expression patterns (Snyder, Makarov et al. 2014, Erkes, Mohgbeli et al. 2015, Rizvi, Hellmann et al. 2015, Vétizou, Pitt et al. 2015, Riaz, Morris et al. 2016). In this vein, we here explore whether the immunosignature technology could be used to profile the immune response to different stages of cancer.

Immunosignatures (IMS) broadly and unbiasedly profile the antibodies in an individual (Stafford, Halperin et al. 2012). IMS uses arrays of 125,000 peptides chosen from random sequence space to maximize chemical diversity. Diluted blood is applied and the pattern of antibodies binding is detected with a secondary antibody. The same array can be used to profile any condition in any species. IMS has been used as a diagnostic for Alzheimer’s disease, diabetes, chronic fatigue syndrome, various infections and cancers (Legutki, Magee et al. 2010, Brown, Stafford et al. 2011, Restrepo, Stafford et al. 2011, Chase, Johnston et al. 2012, Restrepo, Stafford et al. 2012, Stafford, Halperin et al. 2012, Legutki and Johnston 2013, Johnston, Thamm et al. 2014, Navalkar, Magee et al. 2014, Navalkar, Johnston et al. 2015). In the case of cancer, 14 different types of cancer, mostly late stage, were distinguished simultaneously (Stafford, Cichacz et al. 2014). Here we apply the IMS to different stages of the same cancer, hemangiosarcoma (HSA), in dogs.

Although very infrequent in humans, HSA is one of the most prevalent cancer in dogs. It is estimated to account for 7% of malignant tumors in canines (Vail and Macewen 2000). HSA is more prevalent in breeds like Golden Retrievers and German Shepherd (Ettinger and Feldman 2009). The cancer originates in the endothelium of blood vessels. The patient usually does not show clinical signs until late stage. A common cause of death for this disease is tumor rupture (Simansky, Schiby et al. 1986). There is no diagnostic for the early detection of HSA, thus most dogs are diagnosed at late stage of the disease. There is interest in developing a biomarker, particularly a blood biomarker, for early diagnosis.

In applying the IMS technology to HSA we find that each of the three stages has a distinct set of features characteristic of that stage relative to dogs without HSA. For example, peptides that are highly reactive in Stage 1 are not reactive in Stage 3 cancer. Remarkably, Stage 1 reactive peptides have more similarity to know pathogen epitopes than Stage 2 or 3. The IMS appears capable of staging HSA through its reaction with the humoral immune system.

## Material and methods

### Array Platforms

Immunosignature platform consisting arrays of 125k peptides are used in this study. The 125k platform is in-situ synthesized on silicon wafer(Richer, Johnston et al. 2015, Stafford, Wrapp et al. 2016). Arrays were deprotected after synthesis, soaked in DMF overnight and then transitioned to aqueous solution. The residual DMF was removed by washing 5 min twice in distilled water and arrays were soaked with PBS30 min, followed by blocking with incubation buffer (consisting of 3% BSA in Phosphate Buffered Saline, 0.05% Tween 20 (PBST)). Arrays were then washed, spun dry and ready for the experimenting with sera.

### Array procedures with samples

The assay conditions have been published before (Halperin, Stafford et al. 2010, Brown, Stafford et al. 2011, Kukreja, Johnston et al. 2012, Kukreja, Johnston et al. 2012), but they will briefly be described here as well. Arrays were incubated for 1 hour at room temperature with incubation buffer and diluted sera at final concentration of 1:5000. Arrays were then washed and incubated with IgG secondary antibody with conjugated dye. After washing, the arrays were scanned to determine the signal intensity for each peptide feature at specified wavelength.

After the TIFF image of the array was captured, the intensity values for each feature were extracted using GenePix (Molecular Devices, Santa Clara, CA). The intensity values were used to calculate the analysis described in this paper.

### Software and statistics used for analysis

R programming language and JMP were used for data analysis and to create the graphs. Feature selection is based on Two-Tail Student’s T-Test and sorted by p-value. Clustering and PCA analysis is generated with JMP. Confusion matrix and classification report is generated using R with 10-fold-cross validation with SVM classifier. Venn diagram is drawn using package “VennDiagram” in R. Time series plot and pathogen similarity boxplot are generated using JMP.

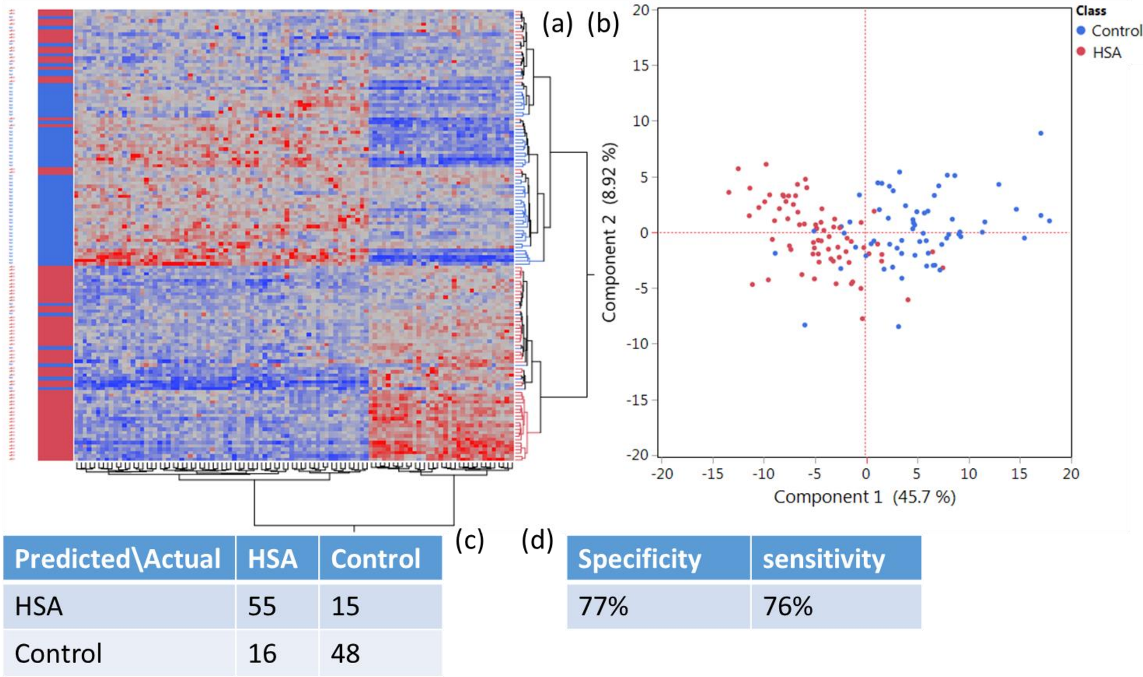
## Results

### Pooled stages of HSA samples can be distinguished from non-cancer samples

The samples used in this study are given in Table 1. The blood was collected as part of a prospective study with Canine Comparative Oncology and Genomics Consortium and from historical samples collected at the Flint Cancer Center at Colorado State University.

We first determined if HSA cancer as one group, that is regardless of stage, can be distinguished from non-cancer samples. A two-tail Student's t-test was performed on all the features' normalized florescence between HSA versus non-cancer samples and 100 peptides were selected based a p-value. This requires peptides that are commonly differentially reactive from samples across all stages relative to non-cancer samples. These features had p-values $<1.97*10^{-8}$. The Bonferroni correction of 0.05 with 125K features gives a p-value cutoff at $\sim4*10^{-7}$.

The selected peptides were used to produce the hierarchical clustering heatmap and principle component analysis (PCA) in Figures 4.1a&b. These figures demonstrate a separation between HSA from the non-cancer group. Note in Figure 1a that peptides with more and less antibody binding contribute to the signature difference. To quantify the difference, we performed classification using a training and test set. Feature selection was based on two tail t-test. Support vector machine (SVM) was used as the classifier. 10-fold cross validation is performed to prevent overtraining, where in each set 90% of the samples were used as training set and an independent 10% as the test set. Feature selection and SVM were performed on the training set data then predicted the test set reiteratively. Only the test sets performance was used for Figures 4.1c&d. At an accuracy of 77% the sensitivity was 76% and the specificity 77%.



| Predicted\Actual | HSA | Control |
|---|---|---|
| HSA | 55 | 15 |
| Control | 16 | 48 |

| Specificity | sensitivity |
|---|---|
| 77% | 76% |

**Figure 4.1. HSA samples can be distinguished from non-cancer controls.**

*Top 100 peptides are selected by T-Test between HSA versus controls samples and sorted by p-value. Hierarchical clustering (a) shows separation of HSA from SE and PCA (b) also shows similar separation. (c) Confusion matrix of SVM classifier with 10-fold cross-validation. (d) Specificity and sensitivity of the classifier. Accuracy at 77%. These HSA samples contain stage 1, 2 and 3 samples. Various stages still have common peptides to distinguish them from control*

### Different stages of HSA have different peptide signatures

We next asked whether each stage of HSA had its own IMS. Pair-wise t-Tests were performed for each stage versus non-cancer dog donors. When a common p-value cutoff of $3.03*10^{-6}$ was used in peptide selection, there are 298 peptides for stage 2 and 169 peptides for stage 3 meeting this cut-off. For stage 1, only 1 peptide met this criterion. Therefore, in the following experiments the top 50 peptides for stage 1 were used. A maximum of 11 peptides out of the 50 are expected to be false positive based on the p-value of these 50 peptides.

The peptides selected are significant in each stage against non-cancer donors. The Venn diagram in Fig. 4.2 shows that most significant peptides for each stage are unique to that stage. Stage 1 and 2 share 1 peptide, while stages 2 and 3 shares 39 peptides. Stages 1 and 3 have no peptides in common. We conclude that the peptides that are significant in each stage against non-cancer dogs are largely unique.

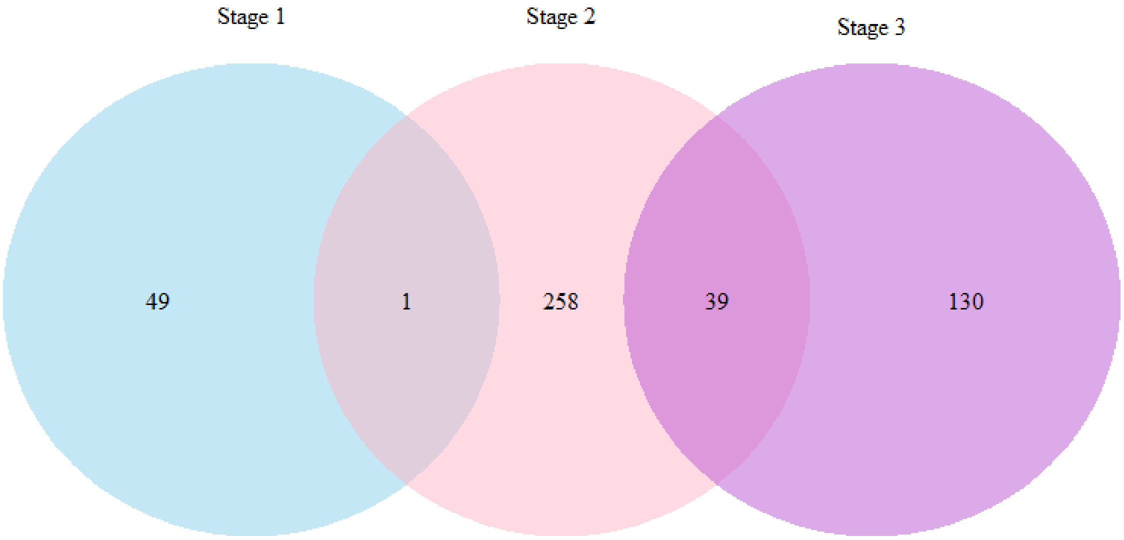


**Figure 4.2. Venn diagram for peptide overlap between stages.**

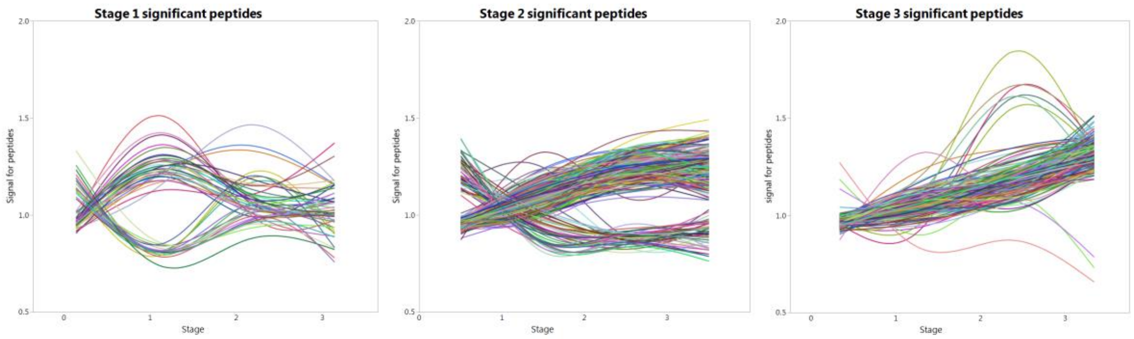
*Peptides are selected by T-Test between specified stages versus non-cancer control. Most peptides belong to only 1 stage, with some peptides being shared between stages. Stage 1 has 50 peptides (49 unique), Stage 2 has 298 peptides (258 unique), and stage 3 has 169 peptides (130 unique). Stage 1 and 2 shares 1 peptides. Stage 2 and 3 shares 39 peptides. Stages 1 and 3 have no peptides in common.*

### IMS peptides from each stage have distinctive stage-series profiles

Ideally, we would want to analyze the IMS profile in each dog over time as it progressed through the stages of cancer – a time series analysis. The samples we have collected were on different dogs at each stage. This does allow us to construct a stage-specific profile for each set of peptides. For example, the relative florescence of each of the 50 stage 1 IMS peptides can be displayed at stage 0 (non-cancer samples), stage 1, 2 and 3. As can be seen in Figure 3, each set of peptides has a unique pattern.

The stage 1 peptides split in approximately half. One half have less binding than in the stage 0 and the other half have higher binding than in stage 0. The lower and higher binding peptides return to the stage 0 levels in stage 2 and 3. If the amount of antibody produced is driven by the antigen, it would imply that there is less of antigen driving the high binders in stages 2 and 3. However, since the low binding returns to the stage 0 level it implies the antibody binding was suppressed in some fashion but not irretrievably.

Stage 2 also presents peptides binding more or less than stage 0. This difference is evident in stage 1 and the difference increases at stage 2. However, in contrast to the stage 1 peptide pattern, this difference is retained in stage 3. This implies that the antigen driving the high binding is continually present through the evolution of the stages. The suppression of antibody for the low binders remains through all stages, potentially from eliminating the B-cells producing the antibodies. In contrast to stages 1 and 2 peptides, the distinguishing peptides for stage 3 display a constant increase over the evolution of the tumor.


Stage 1 significant peptides

Stage 2 significant peptides

Stage 3 significant peptides

**Figure 4.3. Stage significant peptides change during cancer development.**

*Time series analysis on how peptides signal change in different stages. Stage 0 is non-cancer, while stage 1-3 corresponds to the real stage. Stage 1 significant peptides have different signals in stage 1, but return to similar level at stage 0, indicating the elimination of epitopes appeared at stage 1, while stage 2 and 3 peptides keeps increasing with stage, indicating the immune response against these epitopes failed to clear the epitopes, thus is ineffective.*

### Epitope similarity with pathogen is associated with immune response's ability to eliminate the epitope

Studies have found the antibody repertoire is highly skewed by preferential VDJ recombination (Arnaout, Lee et al. 2011, Aoki-Ota, Torkamani et al. 2012). Memory responses for viral antigens are common even for unexposed adults probably due to cross-reactivity with environmental antigens (Su, Kidd et al. 2013). It is highly possible that our immune system has been fine-tuned against environmental pathogens throughout the evolutionary process so that it is no longer effective against antigens not similar to pathogens. Some controversial studies suggest a relationship between the checkpoint inhibitor treatment benefit and the similarity of their tumor neo-antigens and those of pathogens (Snyder, Makarov et al. 2014). Other studies also show the gut microbiome might play an important role in eliminating cancer (Sivan, Corrales et al. 2015, Vétizou, Pitt et al. 2015). Based on these ideas we tested the similarity of the defining IMS peptides at each stage to the Immune Epitope Database (IEDB).  This database is a compilation of experimentally defined immune epitopes of pathogens.

All peptides sequences identified as significant between each stage and non-cancer samples were blasted against the IEDB. Using the same cutoff, the number of matches in IEDB is recorded for each peptide. The higher the number of matches the more presumed similarity to pathogen sequences. The blast hit number was log transformed to ensure a normal distribution.

In Figure 4.4 the peptides in each group are presented in a boxplot. As evident in Figure 4a, when all the significant peptides in each stage are compared, the stage 1 peptides are significantly more similar to the IEDB epitopes than stage 2 and stage 3. Recall that each set of signature peptides for a stage consisted of peptides that were more reactive than non-cancer and ones that were less reactive than non-cancer. Figure 3.4b shows that the similarity of stage 1 peptides to the IEDB is driven by the “up” peptides. As presented in Figure 4.5 and 4.6, the “down” peptides in the stage 1 signature are not significantly different from the stage 2 or 3 “up” or “down” peptides. We conclude that the peptides with higher reactivity in stage 1 signature have significantly more similarity that the stage 2 or 3.

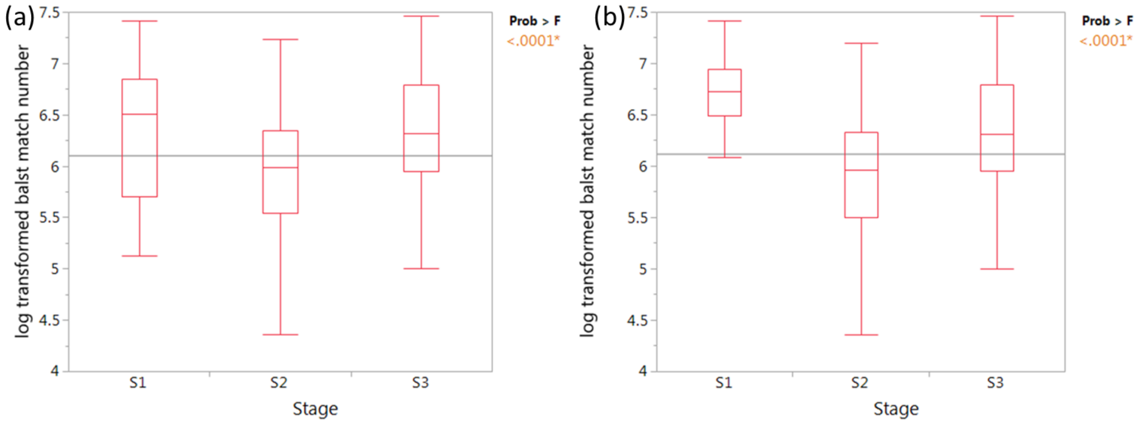


**Figure 4.4. Blast against IEDB shows differential similarity with pathogen of peptides from different stages.**

*Each peptide is blasted against IEDB and recorded the number of matches under a common cutoff. Peptides from same group are put into boxplot and represent the pathogen similarity of the group. (a) Stage 1 peptides are more similar to pathogen than stage 2 peptides, (b) Stage 1 peptides are more similar to stage 2 peptide when using only the up peptides.*

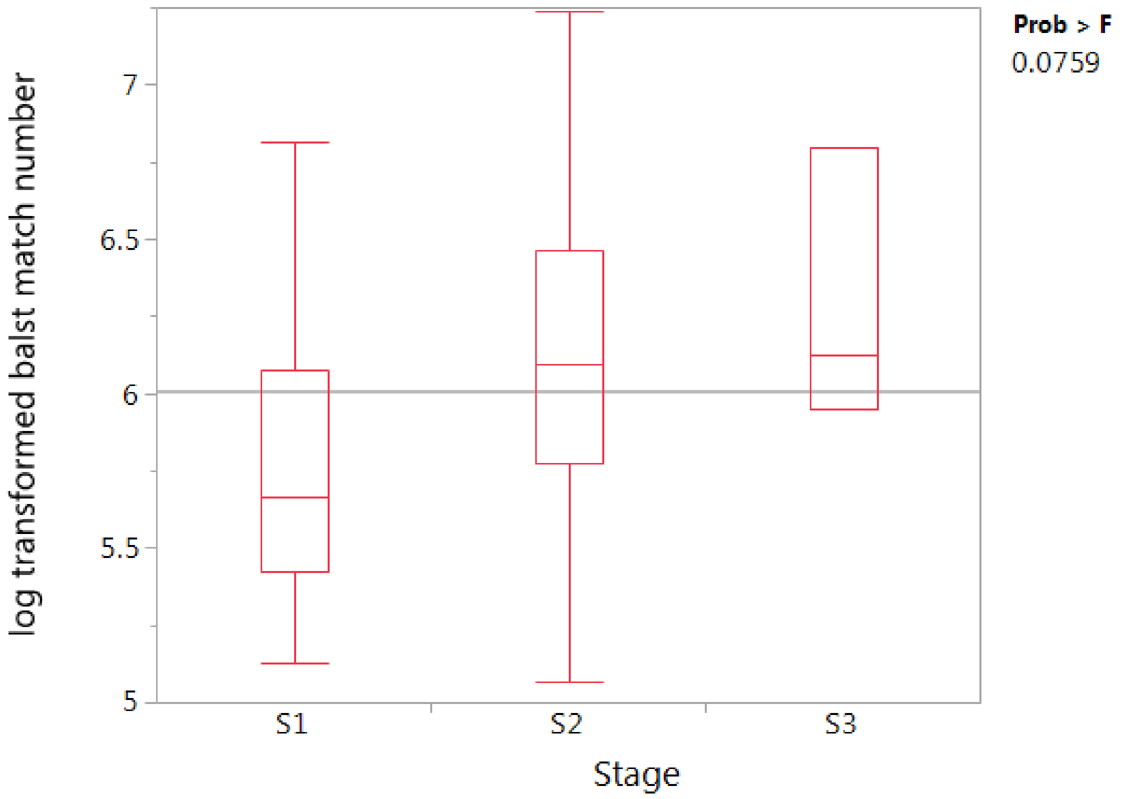


**Figure 4.5. Blast against IEDB shows differential similarity with pathogen of peptides from different stages, down peptides only.**

*The overall P-value for this comparison is not significant, but the trend is very clear: the down peptides become more and more like pathogen from stage 1 to stage 3.*

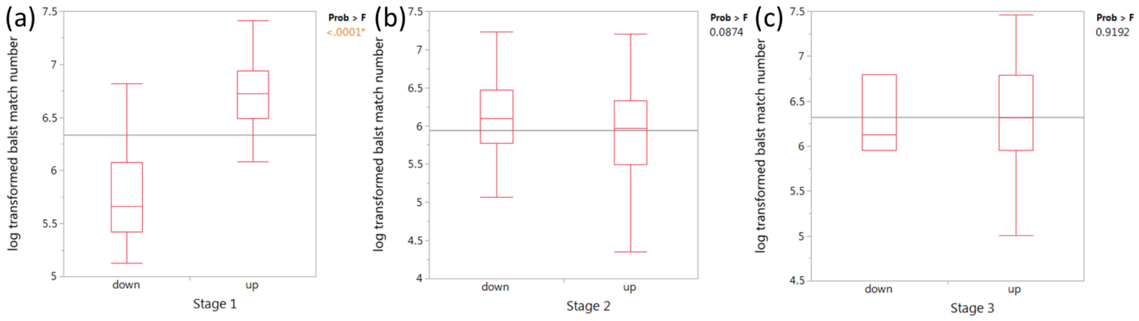


**Figure 4.6. Comparison of blast results for peptides within same stages.**

*Peptides are grouped first by stage and then by whether they have signal that is higher than non-cancer samples or lower than non-cancer samples. In stage 1(a), the up peptides are a lot more like pathogen than the down peptides. In stage 2 and Stage 3, there is no statistical difference between the two groups. While the function of the down peptides are still unknown, this result shows there is extensive selection in stage 1 but not as much in stage 2 and stage 3*

## Discussion

In this chapter, we performed characterization of different cancer stages using the immunosignature platform. We first show that dog HSA samples from stage 1-3 combined share a common signature that distinguishes them from non-cancer samples, with a classification accuracy of 77%. We then focused on understanding the differences between stages. We found stage 1, 2 and 3 have different signature peptides defining them from non-cancer. Most of the identified peptides are stage-specific, with a small proportion of the epitopes overlapping between stages. Analysis of the florescence intensities of each signature peptide over the 3 tumor stages revealed three distinct patterns for each set of peptides. While the peptides for the stage 3 signature increased from non-cancer to stage 3, both stage 1 and 2 peptides included ones that declined in reactivity from non-cancer. Finally, based on earlier reports of a link between cancer and infection epitopes, we compared the signature peptides to the IEDB data base of infection epitopes. The stage 1 "up" peptides were significantly more similar to the IEDB epitopes than the other signature types.

The IMS diagnostic has been applied to a number of diseases including cancers. We reported earlier on defining a signature for the reoccurrence of lymphoma in dogs. Generally, it has been more difficult to find pan-cancer signature peptides than stage specific peptides. Here we report a signature for HSA that includes all three stages examined, relative to dogs without HSA. The accuracy was 77%. The mis-calls were not biased by stage. Though low this accuracy may be clinically useful, considering that there is no current screen for HSA and that dogs have very poor survivability, often measured in months.

Ideally, it would be best to detect HSA as early as possible (Ogilvie, Powers et al. 1996). Toward this end we analyzed samples from early stages separately for a distinctive signature. The numbers in stage 1 was small to result in any meaningful classification. However, with 10-fold cross-validation the accuracy for stage 1 and 2 combined was at 77%. Again, given the current lack of a diagnostic this may be useful. It will require obtaining more samples, possibly through a prospective study, to verify this usefulness.

As we have found earlier, there was little to no overlap between the diagnostic peptides for each stage. Stage 2 and 3 had approximately 8% peptides in common while there were none between Stage 1 and 3. Given the large histological differences in the stages it is not too surprising that it would be reflected in the immune response. Stage 1 involves small, local tumors, while Stage 2 tumors are larger, may have ruptured, invaded nearby tissues and spread to a regional lymph node. Stage 3 is classified by further invasion of adjacent structures and metastasis (Thamm 2012)

The large differences in profiles between stages highlights the unique aspect of developing IMS as a diagnostic. Late stage samples will not be useful in identifying the classifying peptides for early stage cancer. We will need to use samples from early stage. This makes immunological sense in that a tumor would evolve the proteins it is producing over time and therefore the immunological response would change.

Arguably the most interesting observation was the three distinct patterns of immune reactivity of the sets of stage specific peptides. Stage 3 peptides had the simplest and expected pattern. These peptides increased in reactivity in stage 1 versus non-cancer and further in stage 2 versus stage 1, with stage 3 having the highest level of reactivity. The simplest interpretation is that the antigen eliciting this response was made early in the development of the tumor and continued to do so as the tumor grew. Increasing the amount of the tumor would present more antigen to the B-cells and stimulate more antibody production. An implication is that this antigen was not selected against as the tumor evolved so it would probably not be a good therapeutic target. And there is research showing strong cancer antigens are selected against in early cancer (Marty, Kaabinejadian et al. 2017, McGranahan, Rosenthal et al. 2017).

Stage 1 and 2 signature peptides had more complex but distinct patterns for reactivity. There were two types of reactivities in stage 1. There are approximately an equal number of peptides that displayed more reactivity or less reactivity than in non-cancer samples. Interestingly, by stage 2 both sets of reactivities returned to non-cancer levels and remained so in stage 3. The “up” set of peptides may represent a new tumor antigen that elicits antibodies. Their decline at stage 2 and 3 may be because of a loss of antigen or the suppression or elimination of the B-cells producing the antibody. If it is the loss of antigen these may represent therapeutic targets. The origin of the “down” set of peptides is puzzling. One possibility is that the antibodies that mature to the “up” set cause the “down” signature. This seems unlikely since the level increases in stage 2, 3. Another possibility is that they are caused by the suppression of specific B-cells that is relieved in stages 2 and 3. A third, at least theoretically possible, is that the “down” antibodies are sequestered by the stage 1 tumor but not by stage 2, 3 tumors. For stage 2 the “up” peptides could be explained in the same fashion as for the stage 3. However, the “down” peptides, unlike for stage 1, remain so in stage 3. If this represents selection against the antigen responsible these could also be therapeutic targets.

Clearly it would be useful to find the antigens responsible for the antibodies in these signatures, particularly ones that may offer therapeutic targets. While using peptides from random sequence space offers higher resolution of antibody differences and a non-disease specific platform, it is difficult to translate from this random space to identify a specific protein in the human or dog proteome. While this has been possible to a limited degree, this this continues to be an area for progress.

In this regard, we did try to simplify the comparison by limiting the search space to the IEDB. There is some, though controversial, basis for similarity between infectious disease reactive peptides and those produced by tumors (Snyder, Makarov et al. 2014). It does seem clear that the sequence space occupied by antibodies is somewhat constrained and that this constraint may have evolved in interaction with infectious agents (Chapter 3). By implication these constraints may be reflected in the immune response to tumor antigens.

Interestingly in this regard, Schreiber's group observed that a tumor lacking mutations in spectrin-β2 is likely to survive (Matsushita, Vesely et al. 2012). The mutant spectrin-β2 sequence (QIAL) has 8 matches in the B cell epitopes and 24 matches in T cell epitopes, while the wild type sequence (QIAR) is matched only twice in B cell epitopes and twice in T cell epitopes in the IEDB database. At least in this case, a mutation that is protective is more similar to defined IEDB epitope.

In our comparisons we do find that the "up" set of peptides from stage 1 has significantly more similarity to the IEDB than the other peptides sets. Presumably this set of peptides are the early responses of the immune system to the cancer. A limitation of this analysis is that we compare the dog cancer immune response to the IEDB data base which is largely composed of reactive epitopes in human infections.

In conclusion, we have shown that it may be possible to use the IMS diagnostic for the detection of at least stage 2 HSA. It remains to validate this with larger sample sets and to determine if the diagnostic can be effective for stage 1 detection. We demonstrate that each stage of disease has a distinctive set of diagnostic peptides and that these peptides have different patterns of reactivity over the stages, implying a complex interaction of the immune system and the tumor over time. The relatedness of the stage 1 diagnostic peptides to pathogen epitopes is highly speculative but bears further exploration in other cancer types.

# CHAPTER 5

# DISTINCTION OF BACTERIAL FROM VIRAL INFECTIONS BY IMMUNOSIGNATURES

## Abstract

A blood-based diagnostic that could readily distinguish a bacterial from a viral infection could have a major impact on antibiotic resistance and over-prescription. Ideally, the diagnostic would be a serological test rather than a nucleic acid test, and would work upon presentation of symptoms. Here we explore whether antibody signatures could meet these requirements. We started by looking for common immunosignatures between 4 different bacteria and 5 different viruses, in 157 samples. Immunosignatures (IMS) are patterns of antibody binding on 125,000 peptide feature chips. The peptides are chosen from random peptide sequence space to maximize chemical diversity. Immunosignatures have been demonstrated to readily distinguish different types of infections and chronic diseases. Here we wished to determine if IMS could distinguish the class of bacteria from viral infections. A training set of 95 samples and validation set of 31 samples composed of bacterial and viral infections were used to establish the signature. The training set was used to train the model and parameters were fine-tuned on the validation set. Then the model was tested on another completely independent test set of 31 samples to evaluate performance. We discovered 1000 peptides could make the distinction with 0.84 specificity and 0.83 sensitivity in the test set. Misclassified samples are spread out in all infections. This assay would be more practical if fewer peptides were required for distinction. To examine this issue, we tested each peptide for performance. We determined

that 2 peptides performed as well as the 1000 in making the bacteria versus virus call. To further explore the limits of IMS we included samples from 3 eukaryotic pathogens. Given the aim that decision is needed for whether antibiotics should be prescribed, we found that the accuracy of distinguishing bacterial from non-bacterial pathogen increased. We believe these results suggest IMS could be used to develop a simple, serological assay to distinguish bacterial from viral infections.

## Introduction

Antibiotic resistance is a global problem (Spellberg, Guidos et al. 2008, Davies and Davies 2010, Shallcross and Davies 2014). It is mainly due to the overuse of antibiotics in clinical settings. Overuse is mainly due to the lack of accurate diagnosis that can distinguish bacterial infections. This is especially true for respiratory tract infections and pediatric sepsis (Sweeney, Wong et al. 2016, Tsalik, Henao et al. 2016). More accurate diagnosis at the time of first clinical visit that can distinguish bacterial from other infections would greatly curb the antibiotic overuse problem (2014, OBAMA 2014).

Current research on distinguishing bacterial from viral infections has mostly been focusing on genome-wide expressions (GWAS) (Sweeney, Wong et al. 2016, Tsalik, Henao et al. 2016). The notion is that gene expression will change upon infections of different pathogens. However, a serological test detection method for pathogens is antibody response. There are many complicating factors that make analysis of antibodies between viral and bacterial infections complex – one of the most important is the study platform. Immunosignature offers the best chance for solving this difficulty. Immunosignature is a peptide microarray that derives peptide sequences from random space rather than biological sequence space. The analysis of semi-random sequences allows for a mostly unbiased search for antibodies that may display a common binding motif.  We would not focus on sequences for any given pathogen; this allows us to look more broadly for antibodies that may fall into a pattern that overlaps bacteria and virus. Immunosignature has shown its potential at distinguishing various infections, along with chronic diseases and cancer (Restrepo, Stafford et al. 2011, Chase, Johnston et al. 2012, Restrepo, Stafford et al. 2012, Stafford, Halperin et al. 2012, Stafford, Halperin et al. 2012, Legutki and Johnston 2013, Navalkar, Magee et al. 2014, Stafford, Cichacz et al. 2014, Donnell, Maurer et al. 2015, Navalkar, Johnston et al. 2015) and should be a plausible approach to distinguish bacterial infections from viral infections.

In this chapter, we asked whether we could diagnose samples with various types of infection using the Immunosignature platform at the level of bacteria and viral. We will show that Immunosignature by measuring the antibody response against pathogens, can distinguish bacterial from viral infections. We identified 2 peptides that can distinguish the two classes, which would yield a biomarker with more clinical utility. Finally, we tested the idea that Immunosignature can distinguish bacterial from generally a non-bacterial infection, which is of more clinical relevance, since there are always non-bacterial and non-viral infections present in clinical settings. Our study would provide the first diagnosis measuring antibody response to distinguish bacterial infections and would provide better clinical guidance for whether antibiotics should be prescribed.

## Material and methods

### Study Design

Serum samples were collected at various sources described in detail below and received at Arizona State University (ASU). All samples have informed consent and were anonymized. Every disease sample was tested positive for the specified disease before rendering to ASU. Bordetella pertussis samples were provided by Seracare Life Sciences (Seracare). Tuberculosis from University of Texas at El Paso (UTEP). Malaria from Seracare. HIV from Creative Testing Solutions (CTS). Flu from BioreclamationIVT. Dengue from UTEP. WNV from CTS. VF from Sonora Lab. Chagas from CTS. Lyme from Seracare. Hepatitis B from CTS, Syphilis from Seracare.

Bordetella pertussis, Lyme, Syphillis, Tuberculosis, Dengue, Flu, Hepatitis B, HIV and WNV samples are used in the bacterial versus viral experiment. Chagas, Malaria and Valley Fever were added in the bacterial versus non-bacterial experiment. All samples are randomly assigned into training, validation and test set with equal probability.

### Immunosignature assay

Serum samples were diluted 1:1500 into the sample buffer (3% BSA in 1x PBST) before incubated on Immunosignature microarrays at a final volume of 150ul for 1h at 37 $^{o}$C with rotating. Primary antibodies from the serum were then washed with 1x PBST for 3 times and rinsed with $ddH_2O$ for 3 times. 4nM Secondary anti-human IgG antibodies with Alexa-Fluor 555 conjugation from Life Technologies are added in secondary incubation buffer (0.75% Casein in 1x PBST with 0.05% Tween20) to detect primary antibody binding. Secondary antibodies were incubated on the array for 1h at 37 $^{o}$C before washed off with blocking buffer. Slides were then washed with 1x PBST and $ddH_2O$ before drying. Images were obtained from scanning arrays at 555nm using Innoscan 910 scanner. Signal intensity for features were extracted using GenePix Pro 6.0.

**Statistical Analysis**

Analysis is performed using scripts written in R or the JMP software (SAS Institute Inc.). Raw intensity reads for all samples are normalized to the median per sample. Quality Control (QC) for the samples is performed by checking each sample's average correlation against all other samples. Samples with correlation<0.2 are deleted. 226 samples are run on Immunosignature and 212 samples passed QC and were analyzed.

Feature selection is done by using samples in the training and validation set. Two-tail Student's T-Test is performed for each peptide by comparing bacterial infection samples versus viral infection samples (non-bacterial infection samples). Cutoff is controlled at allowing 1 false positive for all test, which is 1/124,000 or 1000 peptides, whichever is smaller.

PCA is performed using selected peptides with all samples, with the test set samples highlighted in right PCA plot. Hierarchical clustering is performed using the selected peptides with all samples. Ward method is used in calculating the distance between the samples. The same method is used in calculating distance for the features in two-way clustering.

Random Forest is carried out with maximum 100 trees in the forest. Minimum split per tree is set at 10 and maximum at 2000. Early stopping rule is applied on validation set. And performance of the classifier is evaluated and output as confusion matrix for the training, validation and test set. Neural Network is built with one hidden layer and 3 nodes, with TanH as the activation function.

Stepwise regression for reducing number of features is used with stopping rule of p-value cutoff at 0.1 for both entering and leaving the model. The model starts empty with no feature. Features become included in the model if below cutoff p-value and will be removed from the model once p-value larger than the cutoff. This process is done recursively until the model stabilize, with no feature entering and leaving the model. Then the selected features are tuned to maximum RSquare for the validation set. Then Logistic regression is used in building model with the 2 selected peptides.

Blast search of the 2 peptides was done using the NCBI blast server. Protein Blast (blastp) suite is used. Database is Reference proteins and organism is limited to Bacteria (taxid:2). Algorithm parameters is set to adjust for short sequences, and max target sequences at 100. Then the matched sequences are processed to contain only linear matched part. The 100 matched sequences are imported into MEME suite to identify epitopes, with configurations of 10 minimum sites per epitope and 3 maximum epitopes.

## Result

### Correlation of the infections shows possible distinction between bacterial and viral infection

Immunosignatures can classify between infections (Restrepo, Stafford et al. 2011, Restrepo, Stafford et al. 2012, Legutki and Johnston 2013, Navalkar, Magee et al. 2014, Stafford, Cichacz et al. 2014, Donnell, Maurer et al. 2015, Navalkar, Johnston et al. 2015).

However, until now no one has published a successful serological test that can distinguish bacterial from viral infections.

Here, we have used 4 types of bacterial infections, 5 types of viral infections and 3 types of non-bacterial and non-viral infections with a sample size ~280 including non-infected controls to test whether distinguishing bacterial and viral infection is feasible on Immunosignature platform. Samples are listed in table 5.1. They represent a wide range of bacterial and virus species. There were between 9-22 sera samples from each type of pathogen. Each sample was run on the standard CIMV7 arrays containing 125K peptides. The process has been described (Stafford, Cichacz et al. 2014, Stafford, Wrapp et al. 2016). In the assays reported here, IgG was detected.

| Class of infection | type of infection | sample number | Count per class |
|---|---|---|---|
| Bacteria | Bordetella pertussis | 9 | 64 |
| Bacteria | Lyme | 13 | |
| Bacteria | Syphillis | 22 | |
| Bacteria | Tuberculosis | 20 | |
| Virus | Dengue | 22 | 105 |
| Virus | Flu | 22 | |
| Virus | Hepatitis B | 20 | |
| Virus | HIV | 21 | |
| Virus | WNV | 20 | |

| | | | |
|---|---|---|---|
| Other | Chagas | 19 | 57 |
| Other | Malaria | 17 | |
| Other | Valley Fever | 21 | |

**Table 5.1. Sample information used in this study.**

*12 classes of infections are included in addition to a group of non-infected individuals coded as normal.*

If the immune system responds to bacterial and viral infections differently, then we can expect to high correlation for the immune responses within each group and low correlation between them. As a result, we are using correlation of the Immunosignature as the first predictive method to test the idea of distinguishing the 2 groups. Correlation is calculated for each pair of samples using all 125K features from the Immunosignature array. Then the samples belong to the same comparison combination and are averaged to a single correlation value (Fig. 5.1). For example, correlations for all comparisons between any Dengue samples versus any WNV samples are averaged into a single value, representing the average correlation between the two groups. Hierarchical clustering was used to distinguish bacteria from virus. Figure 1 demonstrates the initial unsupervised division showing that influenza virus is the sole misclassified group, classified with bacteria. Non-infected samples and non-bacterial non-viral pathogens are mixed when included in the correlation table (fig. 5.2).

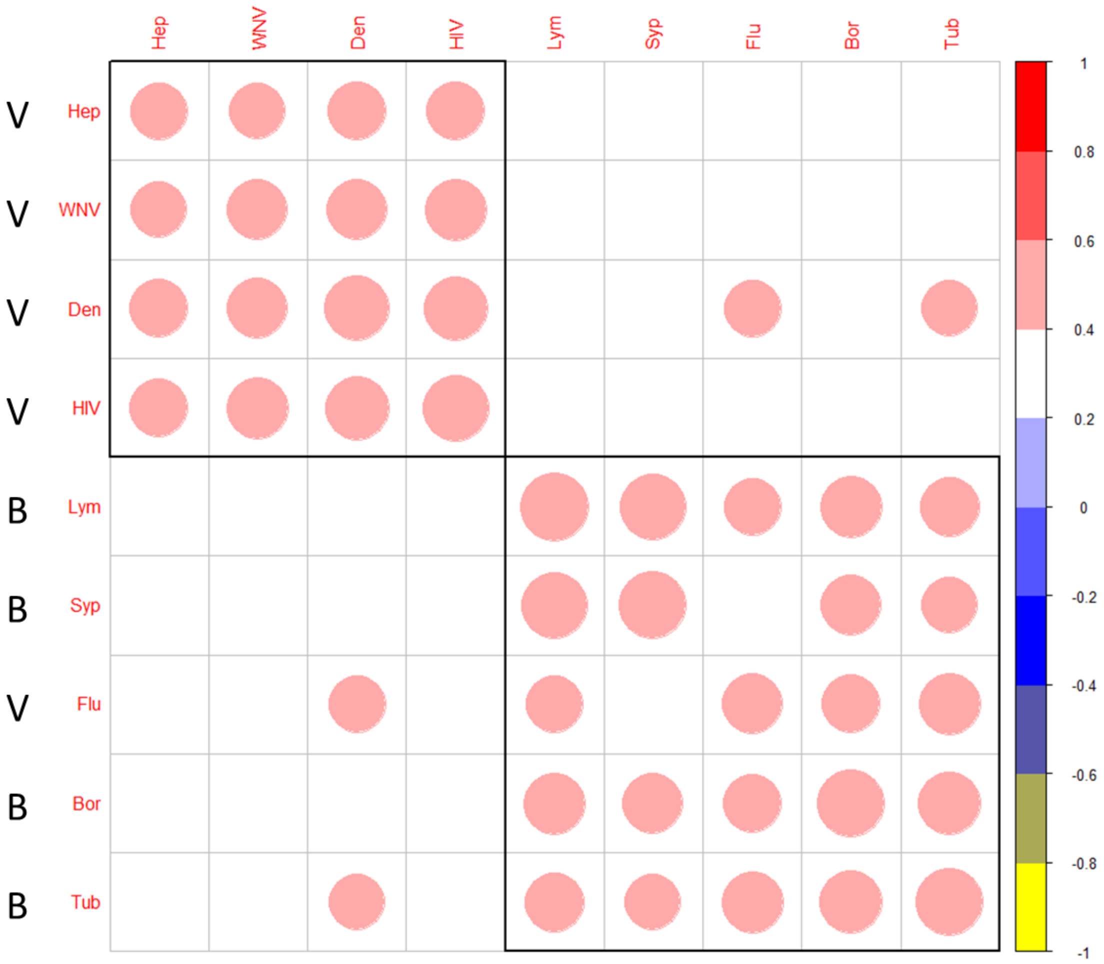


**Figure 5.1. Hierarchical clustering for the correlation of the whole Immunosignature by type of infection shows potential classification of bacterial versus viral infection.**

*Correlation is calculated for each pair-wise sample comparison, then the samples belong to the same class are averaged to a single correlation value. The clustering table shows most viruses can be distinguished from the bacteria, with the exception of flu.*

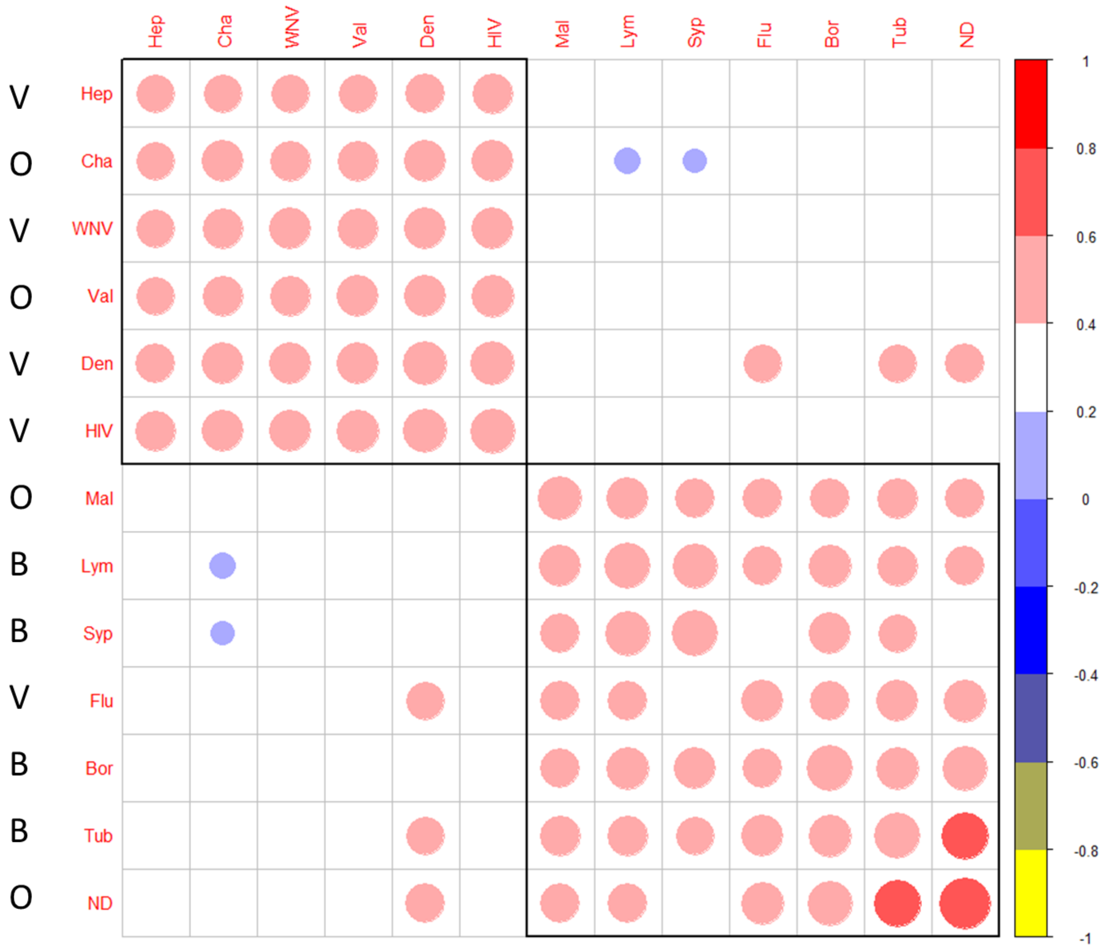


**Figure 5.2. Hierarchical clustering for the correlation of the whole Immunosignature by type of infection including all classes.**

*Non-infected class is more similar to bacterial infection, while the non-bacterial and non-viral infections are spread out in groups.*

A further breakdown per samples is shown in Figure 5.3. Hierarchical clustering using the correlations for every sample (no sample is averaged) is shown in Figure 5.3. The specificity for viral infections is near 100%, with some viruses being classified as bacteria, mostly influenza. This result is consistent with the class level clustering result.

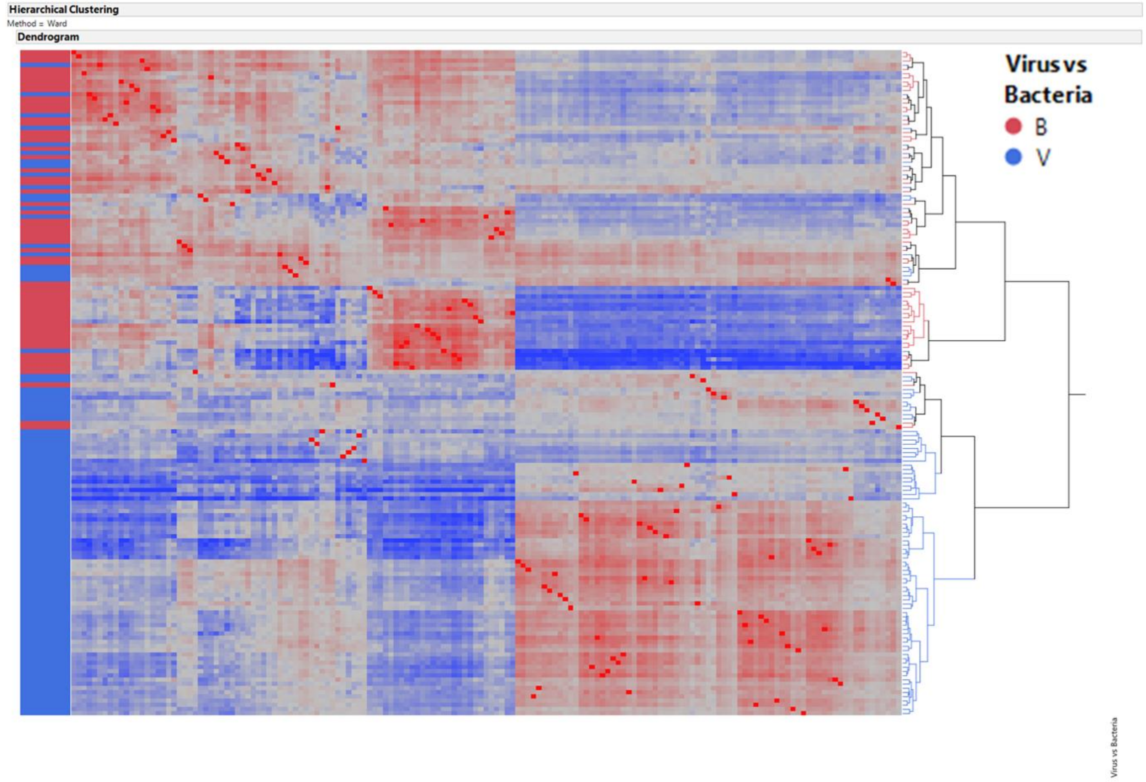


**Figure 5.3. Hierarchical clustering for the correlation of the whole Immunosignature of each sample within bacterial and viral infections.**

*More virus samples are misclassified as bacteria and mostly are influenza samples. Specificity for virus is near to 100%*

### Build bacterial versus viral infection classifier shows robust distinction

Once we confirmed the viability of distinguishing the two types of infections, we utilized machine learning techniques to classify the samples. In this experiment, only bacterial and viral infection samples are used, with a total of 157 samples. Experimental workflow is outlined in Figure 5.4. All samples are randomly divided into training, validation and held-out test set, with a ratio of 60%, 20%, 20%. Training and validation sets are used to build the classifier. Test sets remains untouched until the final model is constructed and used only for evaluation.

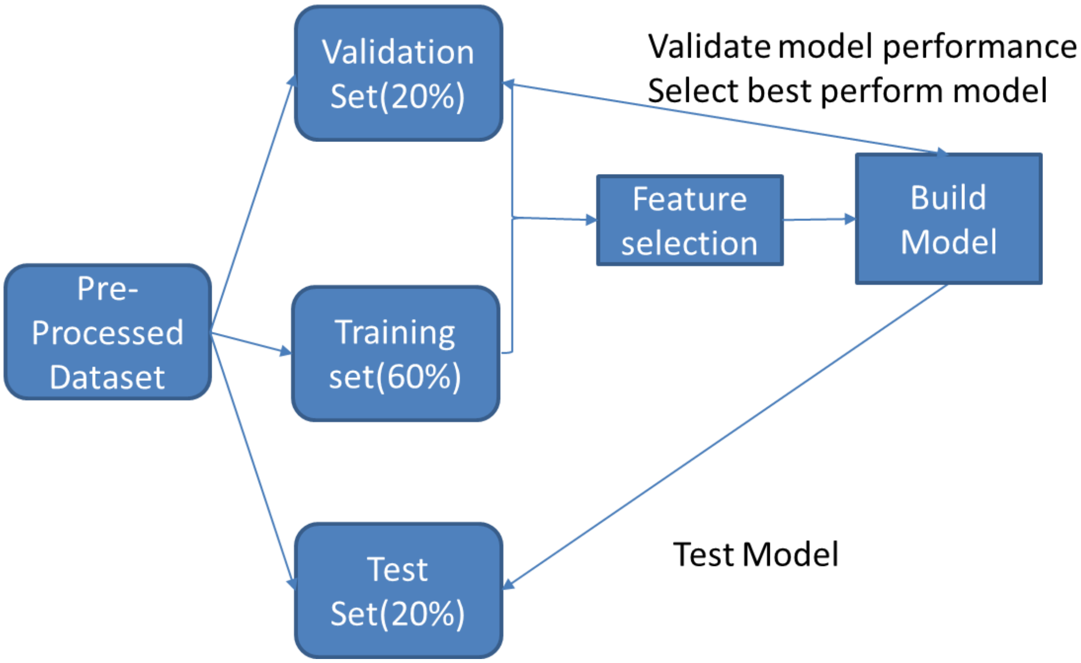


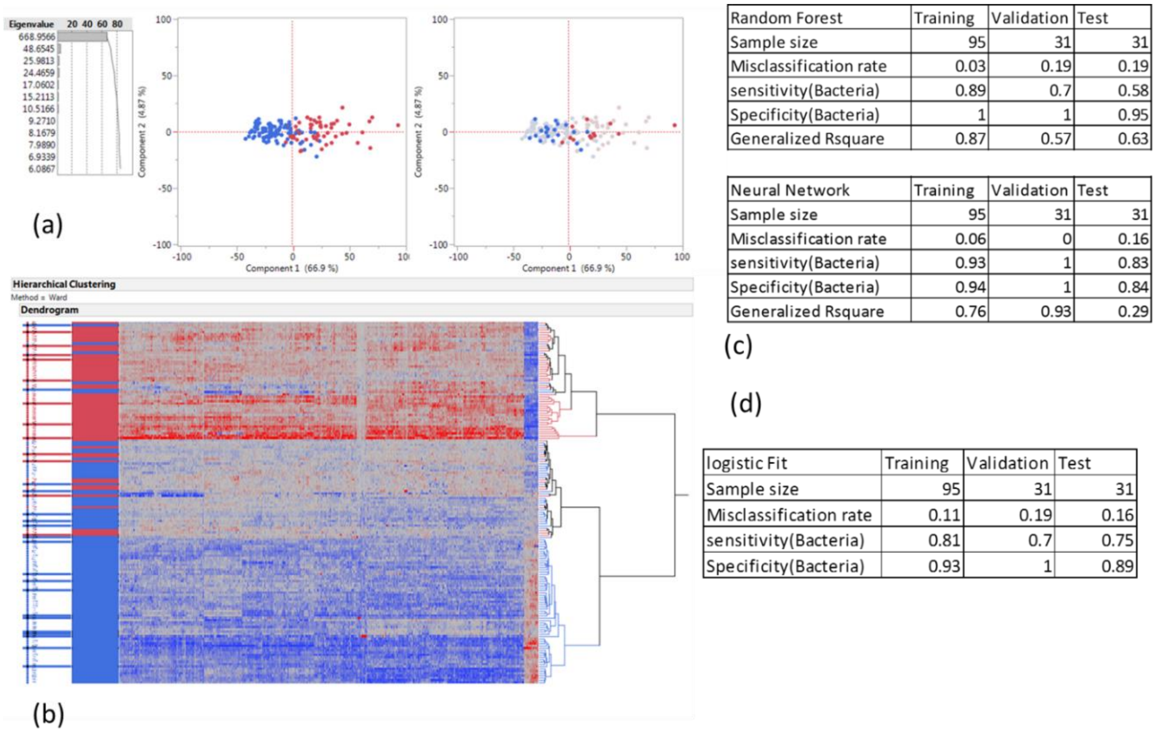
**Figure 5.4. Experiment workflow.**

*Samples are divided into training, validation and test set. Feature selection and model is constructed using training and validation set. Performance is evaluated using test set.*

Since we have 125,000 features on the Immunosignature platform, it is plausible to first do feature selection to find the most useful peptides and remove noise. Feature selection is performed using training and validation set data via two-tail t-test for every peptide and top 1000 significant peptides are used. Note that the general cutoff is either selecting top 1000 peptides or p-value<1/125,000, controlling overall false positive sample to be less than 1. Whichever cutoff has smaller peptide numbers is used in real experiment.

For tests we performed, the p-values are much lower than 1/125,000. As a result, a common cutoff of top 1000 peptides is used throughout the paper.

Using the selected features, Principle Component Analysis (PCA) is performed to determine how many components are responsible for the majority of the variability (Fig. 5.5a). We found component 1 alone explains over 60% of the variability, indicating at least one factor is strongly driving the variance across groups, at least for the selected features. The test samples are not used in feature selection, however, when analyzed with PCA (highlighted in Fig. 5.5a) the test set samples are well separated as the validation set would suggest, suggesting overfitting is negligible. Hierarchical clustering is performed using the selected features to visualize the data (Fig 5.5b). As we can see most peptides are relatively higher in intensity in bacterial than viral infections. This suggests the one component from the PCA analysis may be highly bacterial-specific, suggesting that the peptides that are being selected are from antibody response raised to the bacterial infection. The test set samples are also highlight in the clustering heatmap to show their clustering group location compared with the training and validation set. No obvious overfitting is noted as test set samples are generally clustered in the right class.



| Random Forest | Training | Validation | Test |
|---|---|---|---|
| Sample size | 95 | 31 | 31 |
| Misclassification rate | 0.03 | 0.19 | 0.19 |
| sensitivity(Bacteria) | 0.89 | 0.7 | 0.58 |
| Specificity(Bacteria) | 1 | 1 | 0.95 |
| Generalized Rsquare | 0.87 | 0.57 | 0.63 |

| Neural Network | Training | Validation | Test |
|---|---|---|---|
| Sample size | 95 | 31 | 31 |
| Misclassification rate | 0.06 | 0 | 0.16 |
| sensitivity(Bacteria) | 0.93 | 1 | 0.83 |
| Specificity(Bacteria) | 0.94 | 1 | 0.84 |
| Generalized Rsquare | 0.76 | 0.93 | 0.29 |

| logistic Fit | Training | Validation | Test |
|---|---|---|---|
| Sample size | 95 | 31 | 31 |
| Misclassification rate | 0.11 | 0.19 | 0.16 |
| sensitivity(Bacteria) | 0.81 | 0.7 | 0.75 |
| Specificity(Bacteria) | 0.93 | 1 | 0.89 |

**Figure 5.5. Performance of distinguishing bacterial versus viral infection.**

*(a). PCA analysis on the selected peptides shows one factor is responsible for most variability, test set samples are highlight in the right figure. (b). Clustering of the selected peptides shows most peptides are bacteria specific peptides. (c) Performance of the classification algorithms. (d) 2 selected peptides can achieve similar performance of classification*

Machine learning classifiers like Random Forest and Neural Networks are used to build the model of classification between the two groups. For each classifier, model is trained using training data and validation set is used to fine-tune the model and gain an initial performance evaluation to limit overfitting. After the established model is used on the test set, we perform a final performance evaluation on this independent dataset.

Experiments with training group only usually results in overfitting because the classifier might adjust to the random variations in the training group to gain best fit scores. Validation set only also pose the same issue because the model is generated with information from the validation dataset. In microarray studies, there are inevitably more variables than observations, overfitting becomes more pronounced. Independent datasets are needed to test the performance of the classifier, like what we are using in this paper, the test set data are not used in feature selection to model generating and is only used for the final evaluation of the model.

As it is shown in Figure 5.5c, Random Forest and Neural Networks both have minimal misclassification rate on both training and validation. The final performance on the test set is also similar for both classifiers. Random Forest tends to exhibit less sensitivity to the bacterial infections (sensitivity at 0.58) but is extremely specific (0.95). This is a bias toward true negatives as the cost of lower true positives. Neural Network models yielded more balance for TP and FP between the two groups, with sensitivity and specificity at 0.83 and 0.84 respectively (Fig. 5.6). Both models yield misclassification rates of less than 20%. 60% of human infections are from viruses (Boone and Gerba 2007). Consequently, if doctors follow the immunosignature result, we would reduce the use of antibiotics by over 50% in conditions where doctors prescribe antibiotics to all patient.

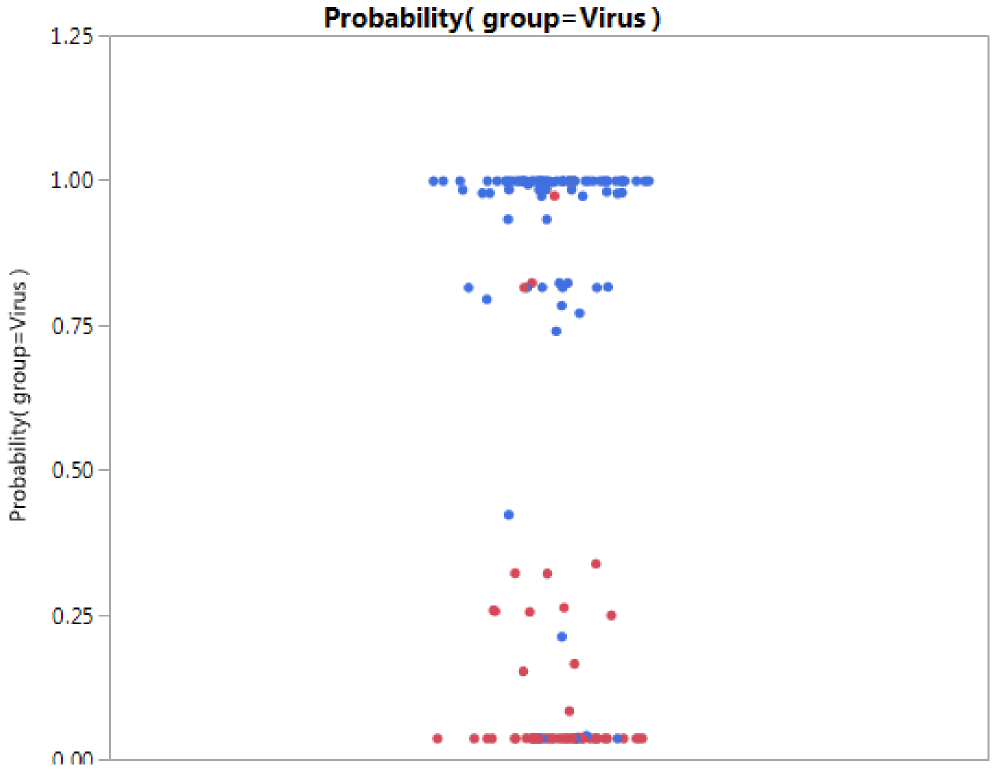


**Figure 5.6. Probability graph for being virus using Neural Network method in bacteria vs viral infection experiment.**

*Color is true label. All samples are included in this figure. Graph shows good separation between the two groups.*

The model was created using 1000 features (peptides). This is difficult to apply in clinical settings as a biomarker test. It will be interesting to see what the minimum number of peptides is that can still achieve similar classification results.

Stepwise regression is used to find the optimal, non-redundant peptides that can be used to fit the model. Each peptide has to meet a p-value cutoff of 0.1 to enter the model and will exit the model upon the exceeding the cutoff p-value of 0.1. Regression is started, with no peptides. The regression process is iterated until the model stabilized, meaning no peptides leave or enter the model. Then the model is fine-tuned to maximize RSquare for the Validation set (Fig. 5.5D). The final regression model only includes two peptides, GLSNGASSFGKASGVAL and GALSRSFANVSFPGVAG (Fig. 5.7). Specificity and sensitivity for the test set comes to 0.75 and 0.89, only marginally worse than the complete models using all 1000 peptides. And the misclassification rate is at 0.16, no worse than the complete models.

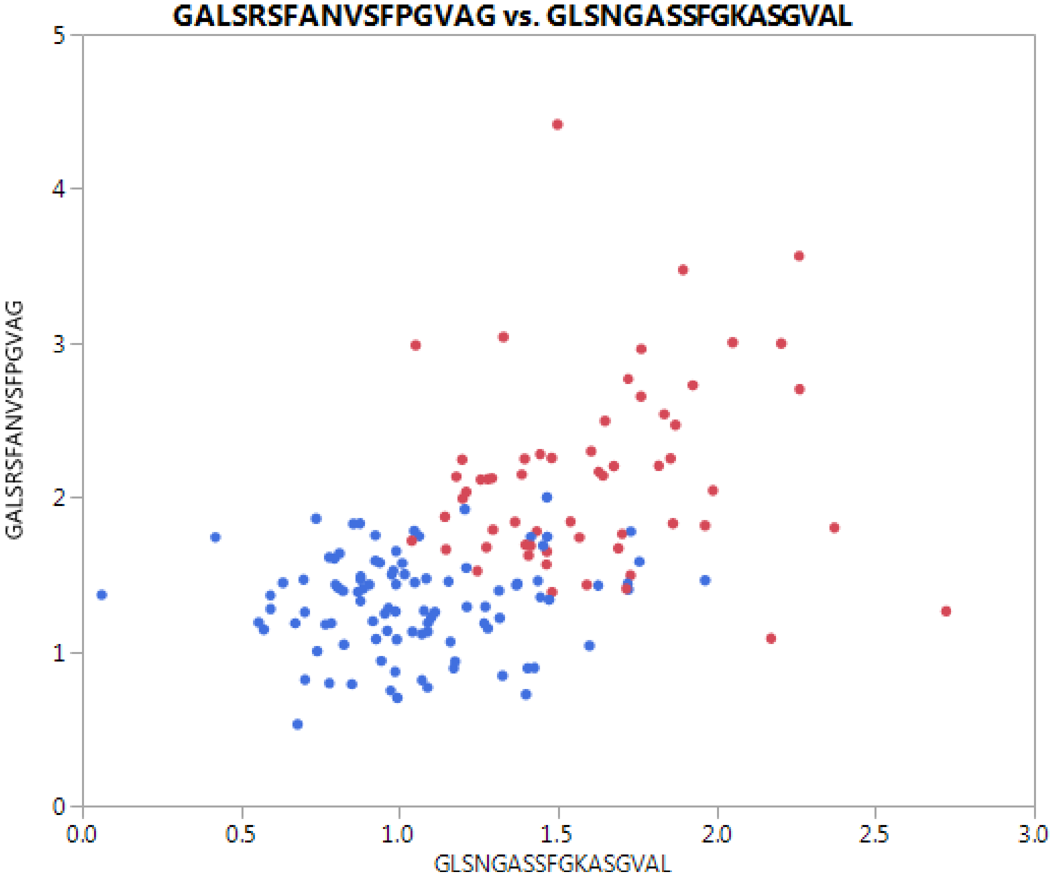


**Figure 5.7. Scatterplot of the 2 selected peptides.**

*Color is true class. All samples are included in this figure. Both peptides are bacteria specific peptides.*

This reduction to 2 features may allow development of a more clinic-friendly serology test. BLAST search on these 2 peptides against the RefSeq database excluding Homo sapiens, Models (XM/XP) and Uncultured/environmental sample sequences, found them highly enriched in bacteria but not in viruses. Furthermore, they are prevalent in all types of bacteria and all types of proteins, suggesting they are indeed good bacterial infection biomarkers.

### Epitopes of bacteria are identified via blast search of the 2 peptides followed by ungapped motif mapping

Once we identified the 2 peptides that are distinguishing bacterial from viral infection, we performed further experiments to identify the epitopes within the sequence. The 2 peptides must contain bacterial epitopes or mimotopes that enhance bacteria-specific antibody binding. We then did a protein blastp search of the 2 peptides against the Bacteria (taxid:2)(Altschul, Madden et al. 1997), with no E-value cutoff. We identified 100 matched sequences in bacteria proteomes which were then submitted to the MEME tool in the MEME suite.  This method identifies consensus motifs (Bailey and Elkan 1994, Bailey, Boden et al. 2009). The identified motif(s) will be the epitope(s) from bacteria that the 2 peptides represent. Results are shown in table 5.2. 1 epitope is identified for peptide 1 while 2 epitopes were identified for peptide 2. It is interesting to note that for peptide 1, only 6 amino acids seem to be the target of bacterial specific antibodies, while for peptide 2, the full length of the peptide is used. Each epitope is matched with at least 20 sequences from the bacterial proteome, so the epitopes are broadly represented in the bacterial world.

| | epitope 1 | epitope 2 |
|---|---|---|
| GALSRSFANVSFPGVAG | RSFANV | |
| GLSNGASSFGKASGVAL | SFGKASGV | LSNGAS |

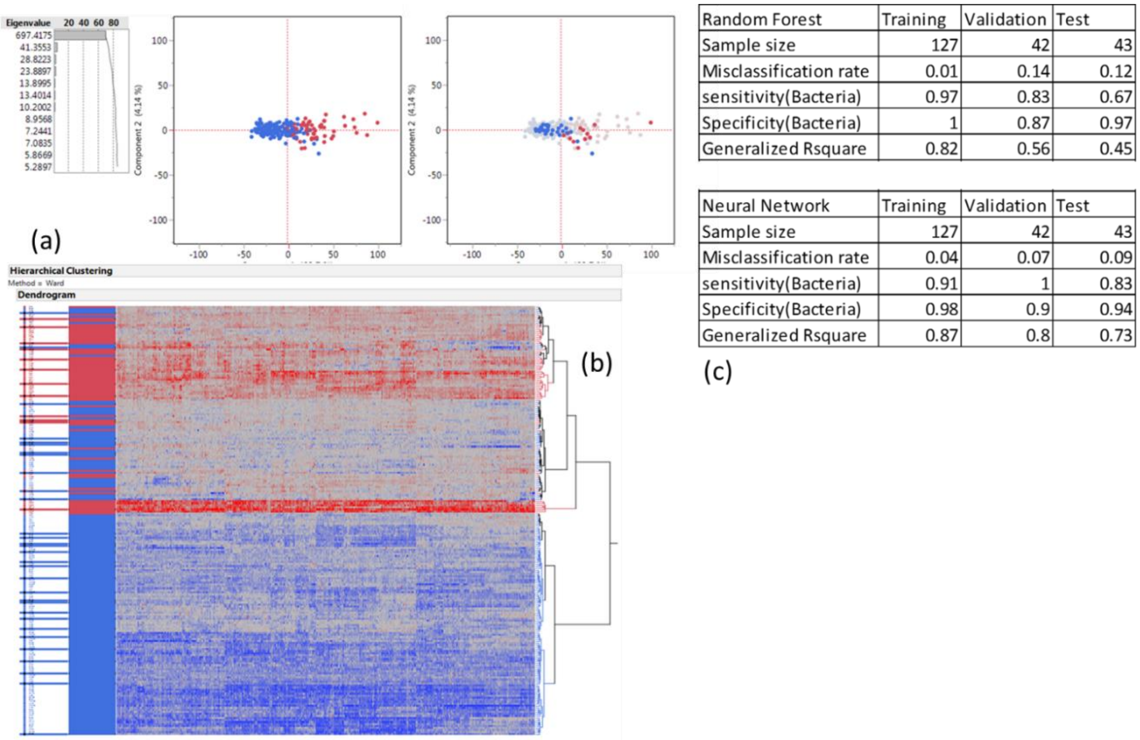
**Table 5.2. Identified epitopes of bacteria with the 2 bacterial-viral distinguishing peptides.**

*Peptide 1 has 1 epitope with length of 6 a.a. While peptide 2 has 2 matched epitopes with length of 8 a.a. and 6 a.a. correspondingly. Matched part is highlighted with color in peptides. This implies only part of peptide 1 is identified by bacterial specific antibody while the whole sequence of peptide 2 is the target for bacterial antibodies.*

**Broad bacterial versus non-bacterial infection classifier shows robust distinction and better performance**

Once we finished constructing a model that is able to distinguish bacterial vs viral infections, we want to test whether we can still distinguish bacterial infection from non-bacterial infections if other types of infections are added as noise. In clinical settings, it is likely that non-bacterial or non-viral infection may be present. Here we ask whether Chagas, malaria and Valley Fever disrupt the original bacterial vs. viral classification performance.

Experiments are performed as described above. Samples are divided into training, validation and test set. Training and validation sets are used to do feature selection and construct model, then test the performance on the independent test set. Results are summarized in Fig 5.8. PCA analysis (Fig. 5.8a) and hierarchical clustering (Fig. 5.8b) show similar separation of the two group as in Figure 5.5, suggesting performance does not deteriorate when noise is added. Random Forest model and Neural Network models misclassify at 0.12 and 0.09 for the test set, which is an improvement compared with the bacterial vs viral only model. The better performing Neural Network model is at 0.83 sensitivity and 0.94 specificity for bacteria with Generalized RSquare at 0.73, all improve vs. the original bacterial vs viral model. This improvement might be the result of more samples being used for model construction, or by including more types of infections as the non-bacteria comparison, the bacterial specific signature becomes more specific.



| Random Forest | Training | Validation | Test |
|---|---|---|---|
| Sample size | 127 | 42 | 43 |
| Misclassification rate | 0.01 | 0.14 | 0.12 |
| sensitivity(Bacteria) | 0.97 | 0.83 | 0.67 |
| Specificity(Bacteria) | 1 | 0.87 | 0.97 |
| Generalized Rsquare | 0.82 | 0.56 | 0.45 |

| Neural Network | Training | Validation | Test |
|---|---|---|---|
| Sample size | 127 | 42 | 43 |
| Misclassification rate | 0.04 | 0.07 | 0.09 |
| sensitivity(Bacteria) | 0.91 | 1 | 0.83 |
| Specificity(Bacteria) | 0.98 | 0.9 | 0.94 |
| Generalized Rsquare | 0.87 | 0.8 | 0.73 |

**Figure 5.8. Performance of distinguishing bacterial versus viral and other types of infection.**

*(a). PCA analysis on the selected peptides shows one factor is responsible for most variability, test set samples are highlight in the right figure. (b). Clustering of the selected peptides shows most peptides are bacteria specific peptides. (c) Performance of the classification algorithms*

In this experiment, we also attempted to find minimal number of peptides that can achieve similar performance compared with using all selected peptides. However, after the same stepwise regression process, the best performance we can get is using 5 peptides to gain a misclassification rate of 0.23, significantly worse than the complete model using all 1000 peptides (Table. 5.3). Also, the sensitivity for bacteria only coms at 0.44, also significantly worse than the Neural Network model. In this case, we cannot find minimal number of peptides to achieve good classification result.

| logistic Fit | Training | Validation | Test |
|---|---|---|---|
| Sample size | 127 | 42 | 43 |
| Misclassification rate | 0.06 | 0.14 | 0.23 |
| sensitivity(Bacteria) | 0.89 | 0.58 | 0.45 |
| Specificity(Bacteria) | 0.96 | 0.97 | 0.875 |

**Table 5.3.Performance of bacterial vs non-bacterial infection classification using 5 selected peptides**

*Peptides are selected from stepwise regression using mixed p-value model at cutoff of 0.1. Logistic fit is then performed using the selected peptides. Test set performance is much lower compared with the complete model using all selected peptides from T-Test.*

**Discussion**

In this chapter we attempted to discriminate viral from bacterial infections using immunosignatures, a microarray-based serological test that uses semi-random peptides to splay out the antibody repertoire from infected individuals. Previously, it has been demonstrated that IMS can distinguish specific infections with high accuracy. This suggests that Immunosignatures are detecting antibodies specific to the infection. However, we asked a broader question: can we identify peptides that generally separate bacterial from viral infections? We built machine learning models to identify the predictive performance of a given set of peptides across 169 patients, 105 with bacterial infections and 64 with viral infections. We achieved over 84% accuracy, 84% specificity, and 83% sensitivity, and could achieve this performance with as few as two peptides. These two peptides are overrepresented in bacterial proteomes, and underrepresented in viral proteomes. Even when adding fungal and protozoan infections, we maintained high specificity, an important goal when attempting to reduce improperly prescribed antibiotics.

Accurate diagnose of bacterial and viral infections is needed in clinical settings. The current imprecise diagnosis results in either over use of antibiotics or delayed treatment for patients. Here we present a novel diagnosis based on Immunosignature technology that is able to reliably diagnose bacterial infection from viral infections. By measuring the antibody response of patients with different infections, we showed that correlation of infections is already able to distinguish the majority of the bacterial and viral infections. We further constructed models based on selected features and applying machine learning algorithms to the selected features. This model is able to classify the two types of infections with a misclassification rate of less than 20%, exceeding current biomarkers either used in research or clinical settings (Oved, Cohen et al. 2015). We further reduced the number of peptides to 2 both to test the limit of distinction and for easy application in clinical settings. The reduced model is performs as well as the full model and we identified the epitope from both sequences. Since in clinical settings, non-bacterial, non-viral infections will be expected, we also construct a model aimed at distinguishing bacterial versus all other non-bacterial infections, consisting of viral infection and noise infections including Chagas, Malaria and Valley Fever. This model shows even better performance with misclassification rate at about 10%. These results suggest using antibody response measured from the Immunosignature platform is a viable approach to develop clinically usable bacterial versus viral infection diagnosis.

Several studies using gene expression profile has shown potential to diagnose bacterial vs viral infections (Oved, Cohen et al. 2015, Sweeney, Wong et al. 2016, Tsalik, Henao et al. 2016). The logic behind those studies is genes will be differentially regulated when encountering different infections. So is it the case for antibody response. Antibody response is the most direct reaction for an infection. Given the fact that genes as indirect reaction can still work to distinguish infections types, antibody response should be an even better approach because of it directly targeting the pathogens. Bacteria and viruses have totally different structures while within the class they share commonality. This gives the foundation for why Immunosignature platform that measures antibody response might be viable in the classification. One thing worth noting is that compared with gene microarrays, where it is usually one-to-one binding, antibodies will usually bind to multiple peptides on Immunosignature as long as the peptides are mimotopes of the true epitope (Stafford, Halperin et al. 2012). As a result, more peptides are used in analysis for the Immunosignature experiments.

Correlation of the infections was used to first test the possibility of distinction at the antibody system level. The logic behind using correlation of infections is that the immune system might systematically see the difference between bacterial and viral infection by activating different pathways (Begitt, Droescher et al. 2014). Immunosignature platform is measuring antibody repertoire in the blood. If you use all the data from the platform, then you are measuring the immune system. Correlation of the immune system can then be tested by calculating the correlation of the Immunosignature for different pathogens. The results from the correlation offer insights into understanding both diagnosis and how the immune system works. It seems the immune system is able to distinguish most bacterial and viral infections and mount totally different immune response, since only one infection is misclassified. This confirms the notion that our immune system probably knows the source of the infection and responds accordingly. Or we can propose that maybe the immune system does not know the source of infection but because all infections within the same class are so similar, the immune system always produce similar antibodies against various bacterial infections. The same might be the case for viral infections. As later on described in the chapter, most of the signatures that can distinguish bacterial and viral infection are bacterial specific signatures, implying the immune system is producing various antibodies against bacterial infection in ways like broad-spectrum anti-biotics.

The result that influenza virus is misclassified into bacteria is interesting because it suggests somehow influenza virus successfully tricked the immune system into thinking it as bacteria and produce antibodies against bacteria. This is consistent with the fact that the virus is highly contagious worldwide, implying the immune system cannot quickly mount an effective immune response because influenza virus is regarded as bacteria. This complication adds to the existing problem for the virus including ever-evolving and easy transmission (Cox and Subbarao 2000). This misclassification by the immune system might also explain why there are already pre-existing neutralizing antibodies within the immune system, but they were not usually elicited during flu infection (Xu, Kula et al. 2015). Even though by correlation influenza virus is a problem for diagnosis, they do not appear any different compared with other viruses in methods described later in the paper. As a result, they were not highlighted in the experiments following the correlation study.

Once we found the notion of using immune system to classify types of infections held up, we continued to build a model using feature selection following machine learning classifiers and validated it using independent samples. We envision the major question we can answer in this chapter is whether an antibiotic should be prescribed for an incoming patient. Without accurate diagnosis, a doctor can choose to offer antibiotic, which will results in over-use of the drug, followed by antibiotic resistance (Spellberg, Guidos et al. 2008, Davies and Davies 2010, Shallcross and Davies 2014). Instead, a doctor can also choose to not offer an antibiotic, which will result in delayed treatment of the patient, maybe followed by higher mortality rate and more suffering (Kollef 2008). To solve this problem, all we need is the ability to do classification of bacterial versus non-bacterial infection. We suggest the peptides could be used to develop a binary classification of bacterial infection versus viral or non-bacterial infections. We first tested the model using bacterial versus viral infection and then expand the datasets by including other type of infections as the non-bacterial class to mimic real clinical settings, where there is no assurance the patient only has bacterial or viral infection.

Overfitting has been a major problem in microarray studies (Smialowski, Frishman et al. 2010). Here we approach the experiment with a pre-isolated test set data to avoid the problem. The whole model construction process is without information from the test set. After the model is stabilized, its performance is tested with the test set data. Our results from shows there is little overfitting when migrating the model from training, validation set to the test set.

In the bacterial versus viral infection model, we are achieving accuracy of over 80% in both classifiers tested, which is better than clinical or lab used biomarkers. Clinicians can choose which classifier to use based on experience, since following the random forest classifier will minimize the diagnosis of viral infection into bacterial infection, hence lower the usage of antibiotics, while the neural network classifier tends to balance the error rate in each class, resulting in more usage of antibiotics but less suffering of patients. Features being selected from this study are almost exclusively from bacterial infection, indicating there is more commonality in the immune response. The ability to classify the two classes may just be because our immune system recognizes bacterial infection better than viral infection. And the 2 selected peptides that we found that achieved similar classification accuracy compared with the larger set of peptides suggests there is conservation of antibody response against all bacterial infections.

We further queried the 2 peptides by asking for matched sequences from real bacteria proteomes and then used the matched sequences to identify the consensus motifs. These consensus motifs should be the real target within the 2 peptides in the Immunosignature. We found that only 6 a.a. is the target in one of the peptides while the full length in the other peptide is being matched by the bacterial antibodies. This indicates these two peptides are recognizing different antibodies.

Surprisingly, when non-bacterial and non-viral infections are added to the non-bacterial class, the performance of the model actually increased. Accuracy was ~90% in both classifiers. Specificity for Bacteria is ~95% in both classifiers, indicating this model is good at distinguishing non-bacterial infections. When coupled with the result of the clustering heatmap, we are relatively comfortable to suggest that our immune system sees the commonality for bacterial infections but not other types. This is interpreted from the classifier result that all features are bacteria specific features and as long as you don’t have those features, you are classified into the non-bacterial class. Interestingly when applying stepwise regression to reduce the number of peptides used in the model, we are not able to maintain similar accuracy with it.

Our study is limited by sample size and disease cohort. This will result in instability in the classifiers and is reflected in not being able to minimize peptide number in the largest set. The overall performance of the model is also influenced by the sample size, since all models work better when you have more observations.  In this study, we are using 13 infections, which is relatively small compared with all possible pathogens. However, we are approaching the problem by only doing binary classification. And the fact that all the signature is bacterial specific strengthens the model because for classification of infection, as long as it is non-bacteria, then it will not share the bacteria specific signature and should be classified correctly.

In summary, we are able to construct classifiers that are better performing for bacterial versus viral infection. We validated each model using independent datasets to confirm the robustness of the model. We are able to confirm the source of the selected features, which in turn offers a logic for the success of the model. We believe Immunosignature can be beneficial when used in clinical settings to both combat the antibiotic overdose problem and reduce suffering of the patients.

# CHAPTER 6

# ENTROPY IS A SIMPLE MEASURE OF THE ANTIBODY PROFILE AND IS AN INDICATOR OF HEALTH STATUS

## Abstract

We have previously shown that the diversity of antibodies in an individual can be displayed on chips on which 125,000 peptides chosen from random sequence space have been synthesized. This immunosignature technology is unbiased in displaying antibody diversity, and has been shown to have diagnostic and prognostic potential for a wide variety of diseases and vaccines. Here we show that a global measure such as Shannon's entropy can be calculated for each immunosignature. The immune entropy was measured across a diverse set of 800 people and in 5 individuals over 3 months. The immune entropy is affected by some population characteristics and varies widely across individuals. We find that people with infections or breast cancer, generally have higher entropy values than non-diseased individuals. We propose that the immune entropy as measured from immunosignatures may be a simple method to monitor health in individuals and populations.

This chapter contains significant input from Dr. Kurt Whittemore. He originally came up with the idea of using Entropy as measurement of Immunosignature and performed the early studies. His Java script is used for calculation of entropy in this chapter. I analyzed new, larger datasets containing more diseases and asked new questions about using the entropy measurement. I am responsible for majority of the results and figures presented in this chapter.

## Introduction

The antibodies in an individual's blood offer a tremendously valuable source of information. The $10^9$ types in an individual and $10^{12}$ total variants exist in widely different concentrations and affinities for their original targets (Legutki, Magee et al. 2010, Legutki, Zhao et al. 2014, Stafford, Cichacz et al. 2014). There are also 5 major isotypes adding to the richness of this information (Rajewsky 1996). Many strategies have been employed to decipher this complexity. Arrays of proteins representing some or all of the proteome of a species are produced commercially (MacBeath 2002, Templin, Stoll et al. 2002, Michaud, Salcius et al. 2003, Miller, Zhou et al. 2003). These can be used to discover antibodies against pathogen proteins or autoantibodies. Peptide arrays representing the proteomes provide higher resolution for the antibody binding to known proteins. Alternatively, high throughput sequencing can be used to read the total variable regions of B and T cells (Briney, Willis et al. 2012, Georgiou, Ippolito et al. 2014). The composite of all of the sequences represents the profile of the antibody coding regions for a particular sample. We have developed an approach, immunosignatures (IMS), that also uses peptide arrays, but the peptides are chosen from random sequence space to maximize chemical diversity and to allow for the presence of mimotopes to epitopes which may be novel, such as a mutation in a cancer cell (Halperin, Stafford et al. 2010, Sykes, Legutki et al. 2012). These peptide arrays can be used to discover biomarkers or vaccine candidates. IMS can be used as a diagnostic tool (Legutki, Magee et al. 2010, Restrepo, Stafford et al. 2011, Restrepo, Stafford et al. 2012, Navalkar, Magee et al. 2014, Stafford, Cichacz et al. 2014). In contrast, here we explore the application of IMS to measure the immune entropy of individuals across time, populations and health status.

The IMS technology is based on creating arrays of $10^4$ to $3x10^5$ peptides, 9-20 amino acids long, in an area of ~$0.5cm^2$ (Stafford, Halperin et al. 2012, Sykes, Legutki et al. 2012, Stafford, Cichacz et al. 2014, Stafford, Wrapp et al. 2016). They are chosen from random peptide sequence space to optimize chemical diversity and therefore, presumably, binding distinctions between antibodies. Given that most epitopes of antibodies are 5-20aa long, it is unlikely that the exact cognate epitope for any antibody is present in the arrays. However, because of the avidity effect each antibody will bind many peptides in a characteristic signature (Halperin, Stafford et al. 2010, Stafford, Halperin et al. 2012). Therefore, when blood from an individual is applied, a complex pattern of antibody binding (IMS) is produced unique for each sample. The binding varies in which features are bound and the amount of antibody on each feature. An attractive feature of IMS is its simplicity. A drop of blood can be sent on a filter paper thru the mail, diluted and applied to the array to make the measurement, greatly facilitating monitoring individuals (Chase, Johnston et al. 2012).

Here we calculate the information entropy of each IMS. Shannon information entropy (defined as $H= -\sum p(x)*\log(p(x))$ where p(x) is the probability of outcome x) can be applied to any type of information to quantify how predictable the information is. In information theory, the entropy can be determined from the frequency of values for all of the elements contained in an object of information. For example, the entropy of the message “aaaa” would have a lower entropy value than the message “abcd”. The entropy value of the first message is –(4/4*log(4/4))=0, and the entropy of the second message is –(1/4*log(1/4)+ 1/4*log(1/4)+ 1/4*log(1/4)+ 1/4*log(1/4))=1.39. Therefore, high entropy information is most similar to the information that would be output by a random information generator.

Global measures, and the entropy measure in particular, have been applied to a variety of biological data previously. Global measures such as the mean and median of a sample are used extensively in scientific research. Application of information entropy is less common, but it has been used to characterize a wide range of different biological data. In cancer, the entropy calculated from aberrations in DNA copy number is higher in a variety of cancer types (van Wieringen and van der Vaart 2011), alternative splicing entropy is higher in some cancers (Ritchie, Granjeaud et al. 2008), the entropy of structural and numerical chromosomal aberrations is higher in cancers (Castro, Onsten et al. 2005), the entropy of a random walk on the protein interaction network graph was higher in cancer cells (West, Bianconi et al. 2012), and the entropy of photographs of tissues was higher in cancer tissues (de Arruda, Gatti et al. 2013). In the brain, the entropy of fMRI data increases with age and Alzheimer's disease in a dataset of 1,248 samples (Chen and Pham 2013, Yao, Lu et al. 2013). Schizophrenic patients had a lower entropy value than normal subjects, which indicates that entropy values that are too low or too high may indicate that something is altered from normal in the system being investigated (Yao, Lu et al. 2013). Rhesus monkeys with induced Parkinson's disease had higher levels of neuronal firing entropy compared to controls (Dorval, Russo et al. 2008). Entropy has also been used for data related to the immune system. For example, Vilar *et al.* assessed entropy from data sets on immune cells (Vilar 2014). Merilli *et al.* applied entropy values to the putative idiotypic network of antibodies (Rucco, Castiglione et al. 2016). Asti et al used maximum-entropy models based on antibody gene sequence data to predict antibody binding from complex mixtures (Asti, Uguzzoni et al. 2016).

Here we calculate the Shannon information entropy of the peptide fluorescence intensity distribution that results from applying sera to a complex peptide microarray surface. The immune entropy (IE) was measured in a wide array of people, the same people over time and the people with diseases.

## Material and methods

### Array Platforms

Two different immunosignature peptide array platforms were used: two different libraries of 10,000 peptide microarrays, the CIM10Kv1(NCBI GEO accession number pending), the CIM10Kv2 (GPL17600) and HT330K (GPL17679). The 10K random peptide platforms consists of 10K 20 residue peptides linked to glass slides through a maleimide conjugation to a linker coupled to an aminosilane-coated glass surface. This linker is on the carboxyl terminus for CIM10Kv1 and on the amino terminus for CIM10Kv2(Stafford, Halperin et al. 2012). The CIM10Kv1 arrays were produced by spotting peptides synthesized by Alta Biosciences using a NanoPrint LM60 microarray printer (Arrayit, Sunnyvale, CA). The CIM10Kv2, peptides were synthesized by Sigma Genosys (St. Louis, MO), and they were printed by Applied Microarrays (Tempe, AZ) using a piezo non-contact printer.

The 330K platform (GPL17679) uses an *in situ* synthesis method to create 330,000 peptides on a silicon wafer (Legutki, Zhao et al. 2014). This platform uses peptides selected from random space to maximize chemical diversity. On this platform, not all of the peptides have exactly the same length, but average 12 amino acids plus or minus 6 amino acids at the 95th percentile. Arrays are deprotected following synthesis, soaked overnight in dimethyl formamide. The residual DMF was removed by two 5 min washes in distilled water, then arrays are soaked in PBS pH 7.3 for 30 min, blocked with an incubation buffer (3% BSA in Phosphate Buffered Saline, 0.05% Tween 20 (PBST)), washed, and spun dry, 1500RPM x 5'. At this point the, the arrays were ready for the application of sera.

### Array procedures with samples

The general assay conditions have been published previously (Halperin, Stafford et al. 2010, Brown, Stafford et al. 2011, Kukreja, Johnston et al. 2012, Kukreja, Johnston et al. 2012), and briefly described here. The procedure for applying sample to the arrays of the two different types of platforms is nearly identical, and less than 1 µl of sample is required. For the CIM10K platform, the microarrays are pre-washed in 10% acetonitrile, 1% BSA to remove unbound peptides. Then the slides are blocked with 1XPBS pH 7.3, 3% BSA, 0.05% Tween 20, 0.014% β-mercaptohexanol for 1 hr RT. Without drying, slides are immersed in sample buffer consisting of 3% BSA, 1X PBS, and 0.05% Tween 20 pH 7.2. Serum is diluted 1:500 and applied to the peptide array for 1 hr at 37 °C. The slides are washed in 1X Tris-buffered saline with 0.05% Tween 20 (TBST) pH 7.2. Then a mouse anti-human secondary antibody conjugated to a dye is applied to the array. The slides are washed again as before and dried by centrifugation. The slides are then scanned in an

Agilent ‘C’ scanner to determine the intensity of each peptide. For the 330k platform, the arrays were loaded into a multi-well Array-It gasket. Then a volume of 100 µl of incubation buffer was added to each well, and then 100 µl of 1:2,500 diluted sera was added for a final concentration of 1:5,000. Arrays were incubated for 1 hr at room temperature (RT) with rocking, and then washed with PBST using a BioTek 405TS plate washer. An anti-human IgG-DyLight 549 secondary antibody with a conjugated dye (KPL, Gaithersburg, MD) was added to the sera at a final concentration of 5 nM. This solution was incubated 1 hr at RT with rocking, and unbound secondary was then removed with PBST followed by distilled water. The arrays were removed from the gasket while submerged, dunked in isopropanol, and centrifuged dry at 800Xg for 5 min. These arrays were then scanned with a commercially available scanner to determine the intensity of a certain wavelength at each peptide feature position.

Once the 16 bit TIFF image file from either type of array was obtained, the intensity values from each feature were obtained using GenePix 8.0 (Molecular Devices, Santa Clara, CA). These fluorescence intensity values were then used to calculate the value of global measures such as the mean and Shannon information entropy.

### Java Entropy program

A custom Java program was written to calculate Shannon’s entropy from the fluorescence intensity files (.gpr, or “Gene Pix Array Format”) from the peptide microarray. Most image alignment software allows output as a gpr file, and that is how the program recognizes data columns. However, any datatype could be used with minor modifications. There are two programs listed in the Appendix, an algorithm class and a test class. The algorithm class provides values entropy given an immunosignature data file, but for comparison sake it also provides CV (coefficient of variance), mean, median, kurtosis, skew, 95th percentile, 5th percentile, and dynamic range.  Tests have shown that entropy is the most sensitive and robust to health changes, but the other calculations provide comparisons.  The test class allows the user to input their data directories and filenames, and serves as the Java main class.

### Software and statistics for general analysis

Microsoft Excel and JMP were used for data analysis and to create the graphs. Linear fit of entropy on age is by ordinary least squares. P-value is the probability of aging is actually influencing entropy. Either ANOVA test or t-Test is used in testing if entropy is being influenced by specific factors.

## Results

### Entropy can differentiate a monoclonal antibody solution from a mixed antibody solution

Entropy can generally measure the difference in the distribution of two datasets as illustrated by example in Figure 6. 1. As applied to an IMS, the expectation is that more antibody types would produce more randomness, which should result in a higher entropy number. This hypothesis was tested by measuring the entropy of binding of two different monoclonal antibodies individually and then in an equal mixture. The results are shown in Figure 6.2. The two monoclonals target different sites (RHSVV and SDLWKL) on the p53 protein.  When each was applied separately to the array, they bound a different set of peptides but the distribution was approximately the same, so the IEs were similar.

However, when the two antibodies were mixed, the distribution of the IMS signal expanded, which in turn caused the entropy to be higher than a single antibody.  This result confirms that entropy can in principle be used as a measure of the disorder in an IMS.

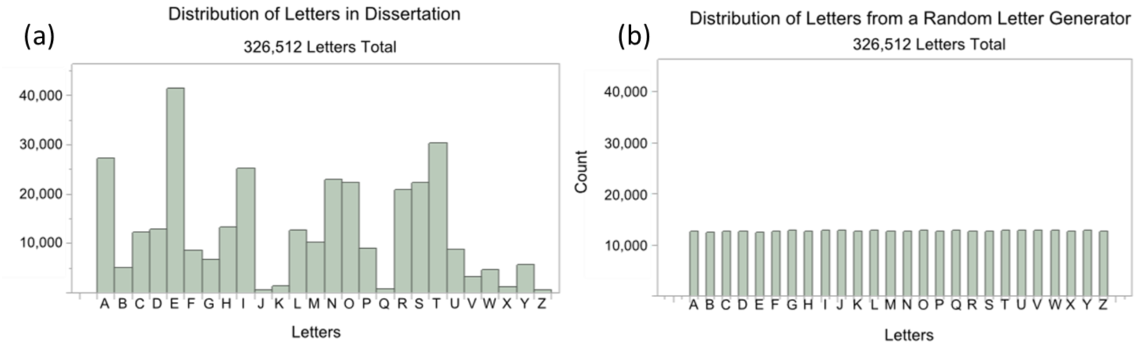


**Figure 6.1. Example of entropy measuring the difference in an information distribution.**

*(a) is the letter distribution from a real dissertation(Whittemore 2014). (b) is the letter distribution of randomly generated thesis with the same total number of letters. The selective use of words results in order for the distribution. The outcome is that the normalized entropy is lower in the real dissertation than the randomly generated one, 0.887 compared with 1.*

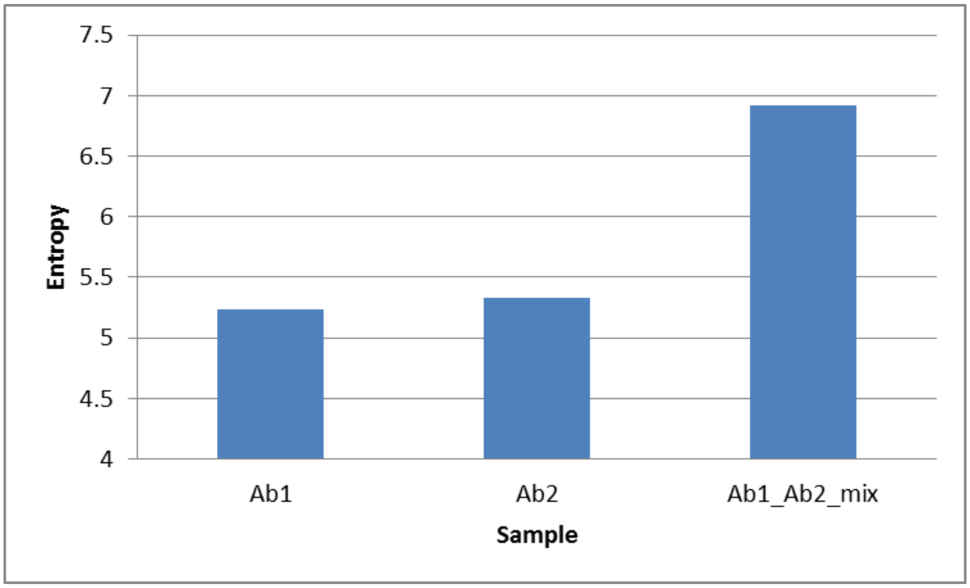


**Figure 6.2. Entropy measurement is able to distinguish a single monoclonal antibody profile from a mixed monoclonal profile.**

*Antibody1 and antibody 2 are individually applied to the Immunosignature platform and then mixed together to apply for the Immunosignature platform. The entropy value is calculated for each distribution. The two monoclonal antibody entropies cannot be differentiated, while both of them are obviously lower than mixing the two antibodies together.*

**IE varies with gender, blood type, and ethnicity but not age or location**

In order to identify factors associated with IE, we examined the sera of 800 healthy individuals using the IMS platform. These samples were obtained from Clinical Testing Solutions (CTS Inc., Tempe, AZ) and were chosen to equally represent the proportion of genders, ethnicity, blood types, and ages in the Southwest US population. They were collected from centers in California, Arizona and Texas.

In Figure 6.3 the distribution of entropy values across the whole set of 800 samples is presented.  The entropy values ranged from 6.6 to 8.8 with a median of 8.1.  The values are approximately normally distributed.

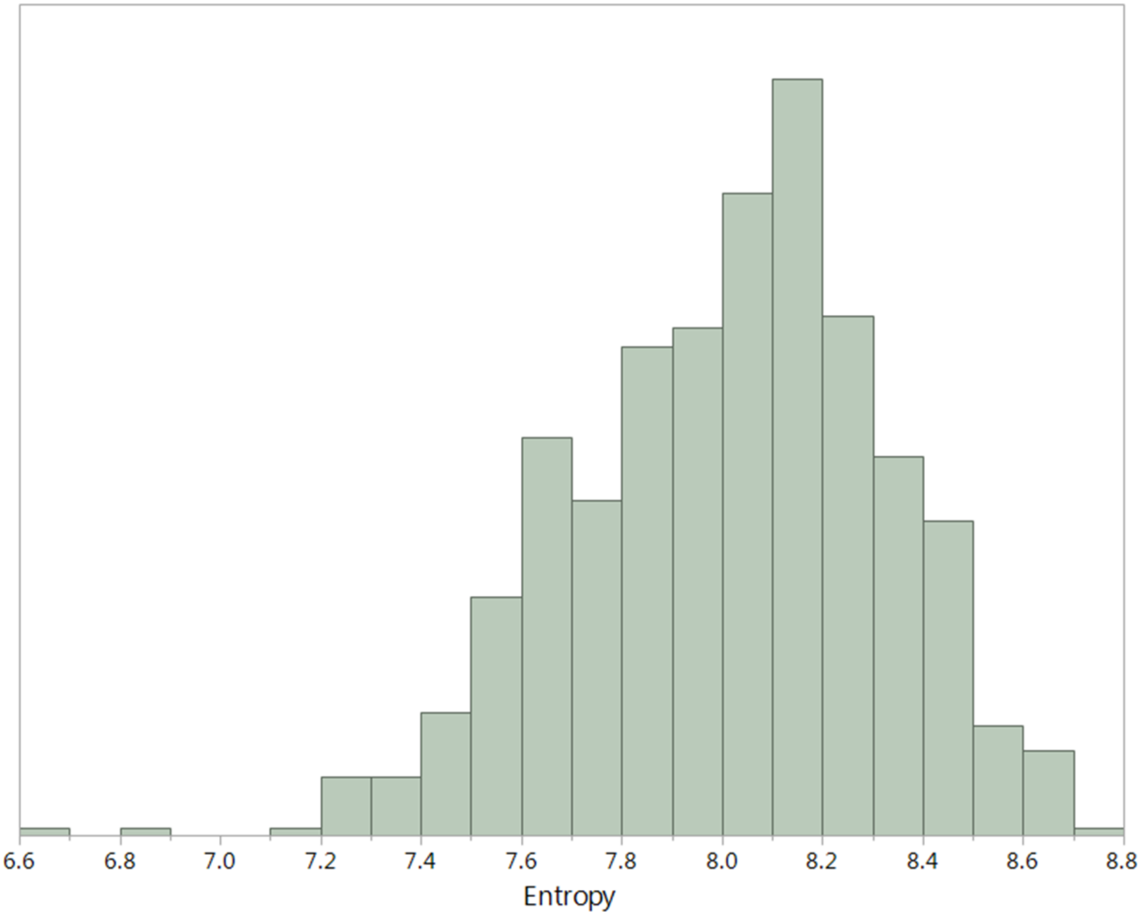


**Figure 6.3. Distribution of entropy values for 800 healthy individuals.**

*The entropy value ranges from 6.6 to 8.8 with a median of 8.1. The distribution is approximately normal.*

Figure 6.4 shows the IE distribution with various factors including age, location, gender, blood type, and ethnicity. The distribution in every group follows a near normal distribution. We asked if there were any significant differences in pairwise comparisons of the entropy with regard to these factors. We found none with respect to age and location. However, we did find that the entropy values are influenced by gender, blood type, and ethnicity.

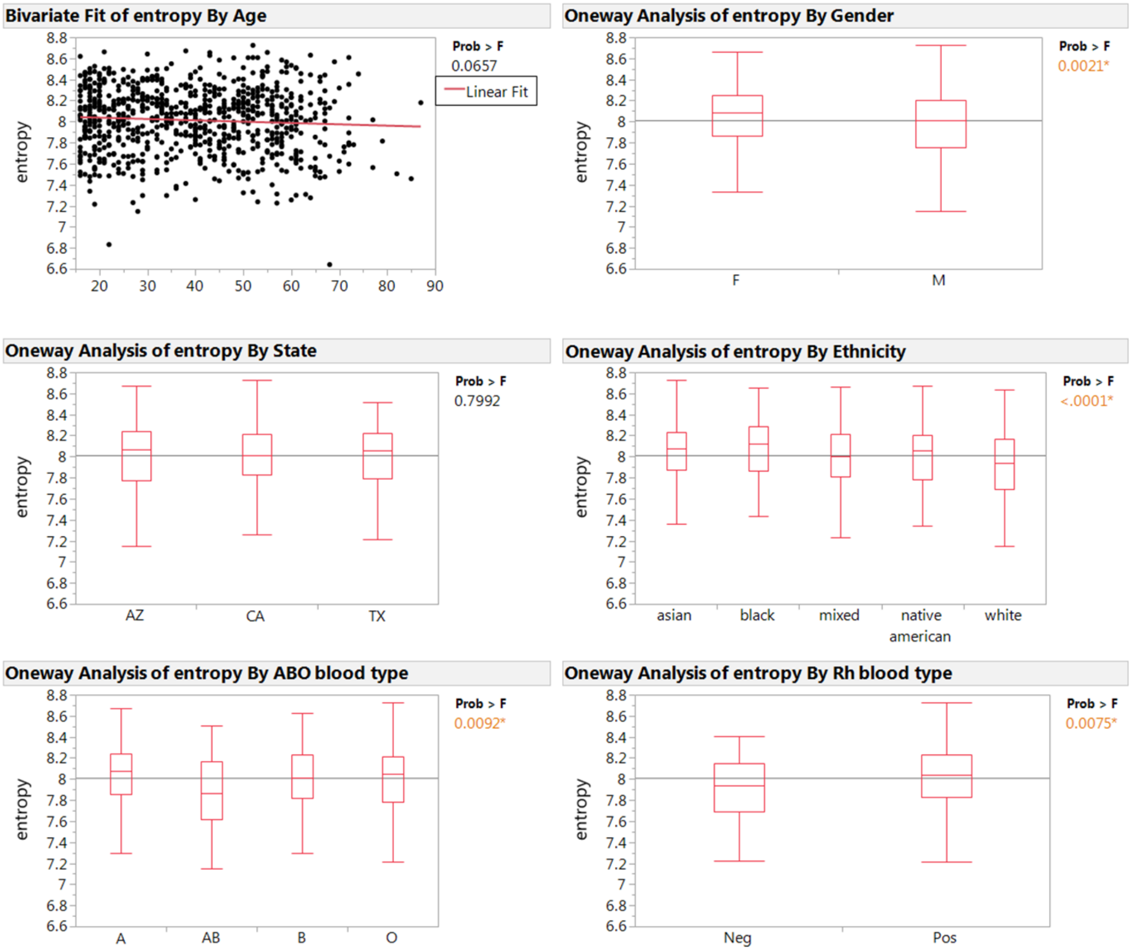


**Figure 6.4. Entropy measurement variance by different factors.**

*Entropy value was tested with factors of age, gender, location (state), ethnicity and blood type. Age, gender, and location are found to not influence the entropy value, while ethnicity and blood type has significant influence on the entropy value. The p-value is obtained from an ANOVA test for each comparison.*

Generally, females have slightly higher entropy than males. Caucasians had a lower entropy level than Asian or African-Americans. The difference of these two sets of comparisons were at a significance level of <0.005 by a t-Test and <0.0001 by an ANOVA test.

We found differences in IE both in the ABO blood group system and the Rh blood group system. People with AB blood type have on average the lowest entropy value, whereas the other blood types are similar to each other. The Rh blood system also shows that Rh- blood type has lower entropy compared with Rh+ blood type.

As noted the Caucasian and Asian populations had different average entropy levels and Rh+ and Rh- have different average values. Caucasians have a frequency of 17% for Rh- while Asians have a frequency of <2%(Garratty, Glynn et al. 2004). Given these differences we inquired whether the differences in ethnic backgrounds could be accounted for by Rh differences. The Rh- samples were subtracted from the Asian and Caucasian derived samples and reanalyzed. The difference in entropy averages was not affected. Therefore, it appears the differences at least between the Asian and Caucasian groups is not due to differences in Rh factor.

### The entropy value varies between individuals, in the same individual over time, and can reflect health status

One would assume that the entropy value between individuals would be different even if just due to random fluctuations in the immune system. However, it is not known what the range of the variation is and how it differs from person to person. In this experiment, we obtained the IMS of 5 individuals over a period of time. Blood was drawn daily for 1 month and every week for 2 subsequent months, the IMS determined for each sample and the entropy calculated. The variance for each individual is summarized in a box plot in Figure 6.5a. An ANOVA test shows a p-value<0.0001, indicating there is significant difference from the grand mean in the mean entropy for the five individuals. This suggests that random fluctuations alone are not sufficient to explain the difference between individuals. It is of interest to note that people with lower average entropy tend to have lower variation overall. The standard error correlates well with the average entropy value. This is especially the case for volunteers 4 and 5, both of whom had the lowest average entropy and variance.

We were also interested in how entropy changes over time within an individual and between them. Instead of plotting the entropy values in a boxplot graph, we illustrated the entropy change with time in each of the individuals in Figure 6.5b. Five volunteers are monitored during the same time period. As it shown, the entropy for all individuals varies during this period and does not show a time correlation between individuals. It appears that the variance in entropy is quite different between individuals.

To determine whether entropy can truly reflect the health status of an individual, we recorded the volunteers' health and vaccine history during the monitored time period. An example of one individual is graphed in Figure 6.5c. Volunteer 4 received 3 vaccines, and was self-reported sick during the monitoring period. Aside from the missing data points from July 25th to early August, we found that there was a trend for the entropy value to increase on health intervention. This gives us a first indication that entropy can be used to monitor health status as it changes with exposure to infections or vaccines.

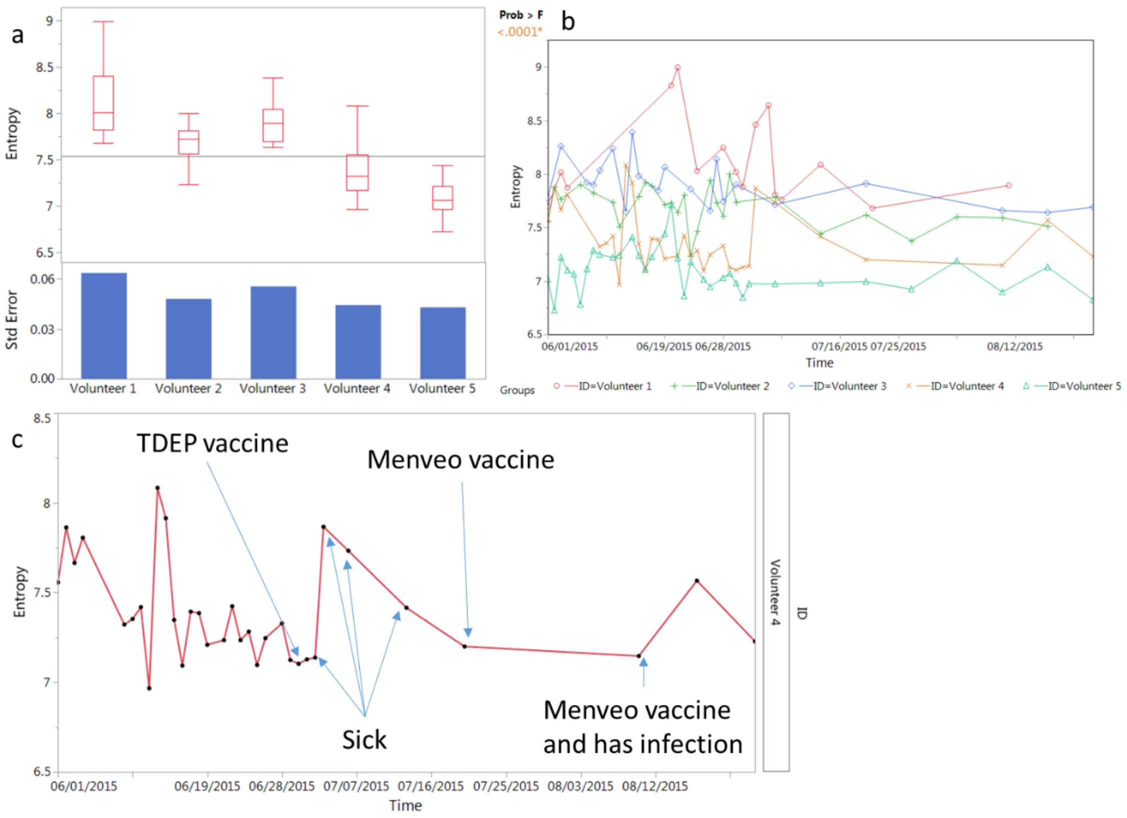


**Figure 6.5. Entropy measurement variance between individuals over time and with changes in health states.**

*(a), boxplot of 5 individual's entropy recorded over a period of time shows difference from person to person. (b), plotting entropy against time for the volunteers shows variation of entropy that is independent between individuals. (c), recorded volunteer's activity shows entropy changes with vaccine administration and sickness. Black dots are blood draw points and the red line connects the dots.*

### Entropy is higher for people infected with pathogens

Once we established how entropy changes in healthy individuals, we asked whether entropy value changes with different forms of health disturbance. We first tested this with infectious diseases. Sera from 7 types of infections were assayed, including Borrelia (8), Bordetella pertussis (12), dengue (9), Hepatitis B virus (15), malaria (13), syphilis (8) and West Nile Virus (21). All samples were from convalescent people. These pathogens, including bacterial, viral and parasite infections, were chosen to broadly reflect the infectious population.

When comparing them with non-infected samples, the infection group shows significantly higher entropy level (Figure 6.6). This result implies that entropy can indeed distinguish people with different health status. Result of the un-mixed 7pathogens' entropy comparison is attached in Figure 6.7.

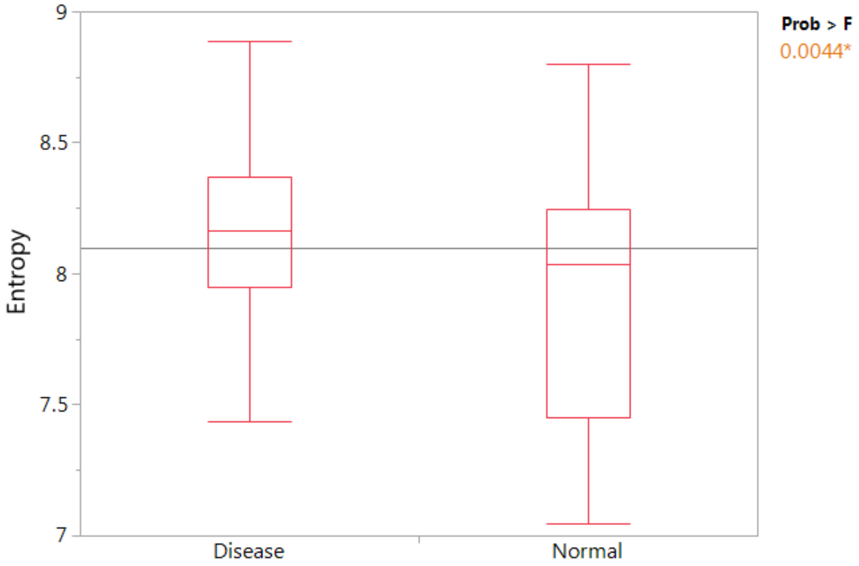


**Figure 6.6. People recovering from infectious diseases have a higher entropy values compared with normal donors.**

*Samples from 7 types of infections are mixed together to represent the disease group. A t-test shows that the entropy from the disease group is significantly higher compared with the normal donors. P-value <0.0044.*

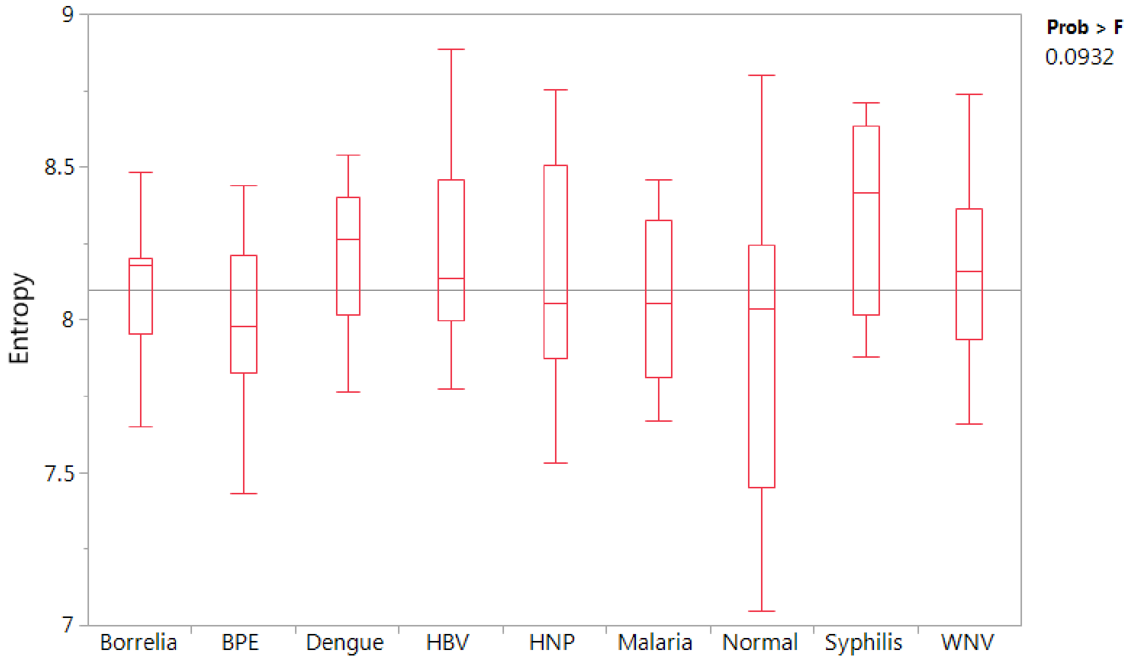


**Figure 6.7. Infections listed individually and in comparison with normal donors.**

*The overall p-value from AVONA test is not significant from this comparison. 6 of the 7 infections have higher mean entropy than normal donors.*

### Sera from people with cancer exhibited a higher level of entropy

We also tested if people with cancer have differences in average entropy. Cancer signatures are distinct by type and from infections (Hanahan and Weinberg 2000, Hanahan and Weinberg 2011). A tumor presumably presents more antigens, including neo-antigens, to the immune system and is often subject to immune suppression (Kawakami, Fujita et al. 2004, Whiteside 2006, Reiman, Kmieciak et al. 2007, Andersen, Thrue et al. 2012).

Here we used datasets from normal donors and from people with several types of cancer to represent general cancer patients, including breast cancer (5), esophageal cancer (2), *Glioblastoma multiforme* (1), lung cancer (1), meningioma (1) and multiple myeloma (1). Analysis is performed with sample sizes of 11 cancer and 21 healthy donors. As shown in Figure 6.8a, cancer samples have significantly higher entropy value compared with healthy donors. The P-value from T-Tests is <0.0096.

In some B-cell lymphomas, a large amount of the same antibody is produced, which changes the antibody composition in the blood (Kuppers 2005, Shaffer, Young et al. 2012). We predict that this may lead to lower entropy value compared with healthy donors. To test this prediction we determined the IMS for dogs with a B-cell lymphosarcoma (LSA) to healthy dogs. IMS uses the same chip for all diseases and species, just requiring the appropriate, in this case dog, secondary, labeled antibody. 68 normal dogs were compared to 83 LSA samples. As evident the entropy is significantly lower in the LSA compared with healthy dogs. This is consistent with the prediction.

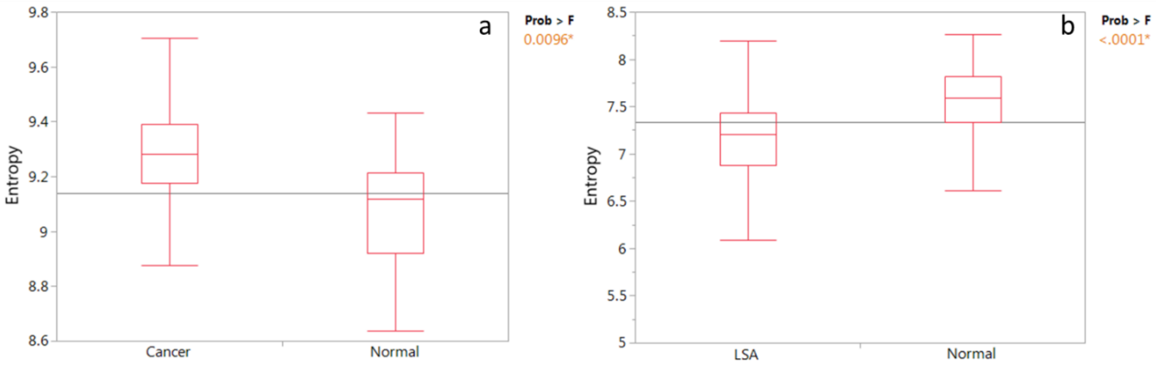


**Figure 6.8. Comparision of cancer patients with normal donors.**

*(a)Various cancer samples are used to represent the general cancer group. The boxplot shows that cancer samples have a higher entropy value compared with normal donors by T-Test with p-value<0.0096. (b) Dog LSA samples are compared with non-cancer normal samples shows lowered entropy for LSA samples with p-value<0.0001 by T-Test.*

## Discussion

We have explored the application of Shannon information entropy to immunosignatures. We first showed that two different monoclonal antibodies that bind to a different set of peptides and have comparable entropy measures, produce an increase in entropy when mixed and added to the arrays, as predicted. We then used a collection of sera from 800 people who equally represent gender, age, ethnic background and three geographic locations to measure the entropy of IMS for each. We found that the entropy values ranged from ~6.6 to 8.8 and were approximately normally distributed over the 800 samples. In pairwise comparison of various sets of signatures we found that there were no significant differences in average entropy values between age or geographic location. We did find the average values females were slightly higher than males, and Asian and African-American donors were significantly higher than that of Caucasian donors. While there were no differences in averages between A, B and O blood types, AB blood types were significantly lower on average. Rh- samples were on average lower than Rh+. We found that the difference between Asian and Caucasian donor samples could not be explained by differences on Rh- frequency between the two groups. We extended the analysis to samples from people infected with 7 different pathogens and found that as a pool these samples had on average significantly higher entropy values than uninfected controls.  The same was true for samples from people with three different cancers compared to people without cancer.  However, we found that dogs with a B-cell lymphoma, as might be predicted for a clonal production of a particular antibody, actually had lower average entropy levels.

In the proof of principle experiment we used two different high affinity monoclonal antibodies to two different sites on P53 (Figure 6.2).  We have shown that monoclonal antibodies can vary greatly in the number of peptides they bind in the array (Halperin, Stafford et al. 2010).  We suggest that the entropy assessment of an antibody may be a good predictor of off-target binding.  It would have the value of being a simple, single number standard that could be applied to all antibodies.

While there was a wide range of entropy values in each of the groups in the 800 samples (Figure 6.3), there were significant differences in the average for gender, ethnicity, and blood groups (Figure 6.4).  The underlying causes of these differences is unknown. Given that the immune system is highly sensitive to both intrinsic and extrinsic factors it would take more studies to associated a cause(s) of the differences.  Where there are no significant differences, for example geographic location, we can exclude differences in flora, for example, as inducing different average entropy levels.

Five people were monitored daily for one month and then weekly for an addition two months (Figure 6.5).  This allowed us to determine the differences in averages overtime and the variance for each person over time.  The entropy averages of the 5 people happened to represent approximately the range we observed in the 800 samples.  Each person generally maintained the differences between each other over the three months. The person with the highest average entropy also had the highest variance and the one with the lowest the lowest variance. It will be interesting to see in a larger set of individuals whether this generally holds true. In order to see if a health event changed the entropy value of an individual, one person received a vaccine. There was subsequently a sharp increase in the entropy number for this individual (Figure 6.5c), although the increase was within the range they previously presented. Additionally, one individual later had an undiagnosed illness and this was accompanied by an increase in entropy (Figure 6.9). These are single events so the association between entropy increase and illness could be coincidental.

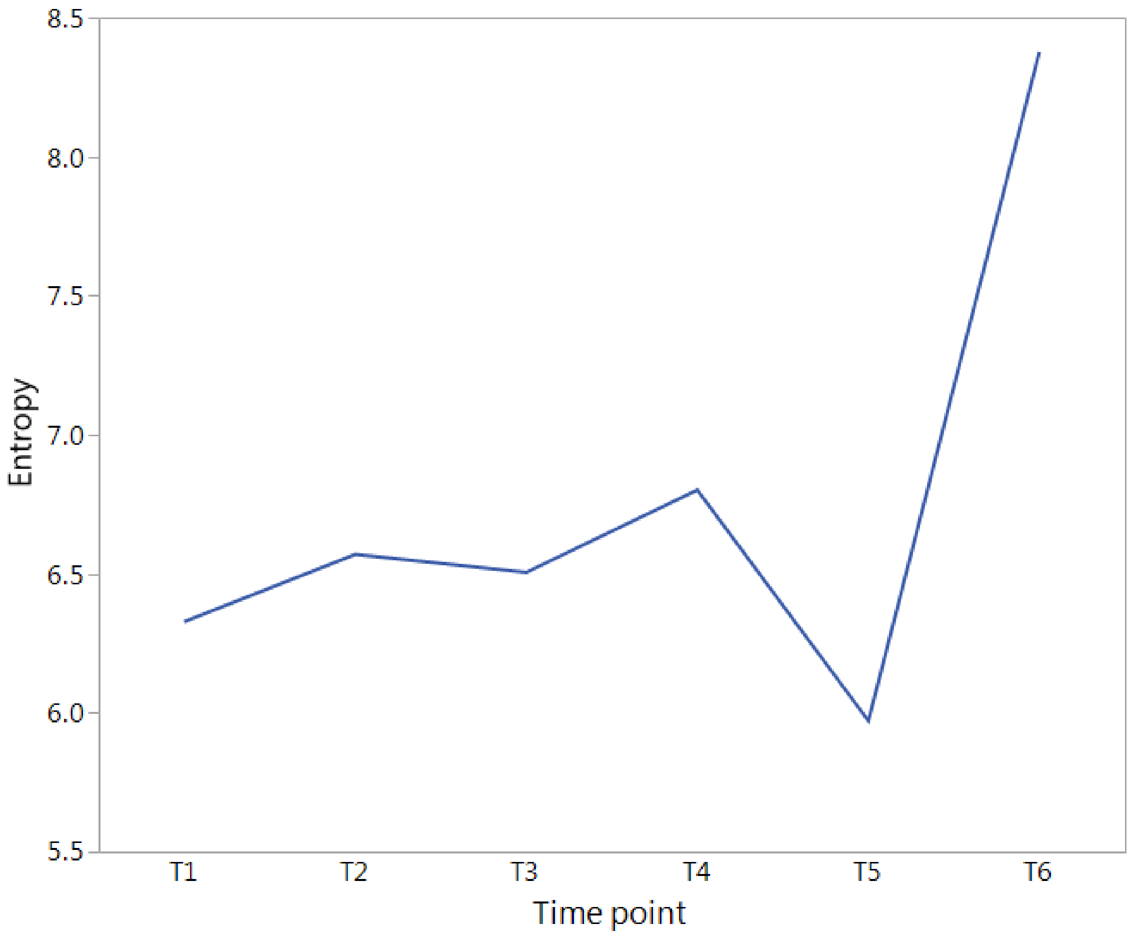


**Figure 6.9. Entropy record of one individual at different time points.**

*The volunteer is healthy at the first 5 data points but report unknown illness at T6. Dramatic increase is observed at T6.*

The results of the monitoring of individuals suggests two potential applications for entropy monitoring. On an individual level if a person monitors their entropy over time on a regular basis, one could detect a significant change from baseline or normal variance. To be useful this would change would need to be present before symptoms occurred. Whether entropy changes are present before symptoms is another area of future investigation.

Another potential application would be for population monitoring for a disease outbreak or an intentional biological attack. If a population was monitoring their IMS on a regular basis, presumably in order to detect early signs of a chronic disease, a disturbance in the entropy levels of a large number of people could be an indicator of an event. As evident from the data in Figure 6.5 on monitoring individuals, this would need to be based on multiple measures of time of each individual. It may be possible to identify the peptides that were responsible for the change in entropy in each person and determine if there was a common signature. In the case of a natural outbreak or attack, this signature would represent the immune response to the infectious agent.

In the data presented in Figures 6.5c and 6.9, the disturbance health event was accompanied by an increase in entropy. We investigated whether this is generally the case. We found that for both infections (Figure 6.6) and cancers (Figure 6.8a) the people with the health problem had on average higher entropy levels. However, within both diseases there was a wide range in entropy values for different people. Therefore, even for a health disturbance that causes and increase in entropy, it would need to be measured against the personal baseline. As an example of entropy decreasing we presented analysis of dogs diagnosed with a B-cell lymphosarcoma (LSA). In contrast to the data in Figure 6.8a, the average entropy was lower in the disease state. B-cell cancers may be a special case as they are characterized by overproduction of one antibody species.

Infections induce a set of high affinity antibodies to the pathogen. In order for this to register as an increase in entropy the induced antibodies would need to expand the number of sites bound relative to the peptides bound by the non-infected samples. The implication is that there would need to be unoccupied features that the induced antibodies could bind to expand the diversity. Presumably, this would also be the case for the cancer samples. In the case of the LSA samples the preponderance of the antibody produced by the cancerous B-cell would decrease the total diversity of antibodies in the sample to lead to a decrease in average entropy.

As discussed in the Introduction, the concept of entropy has been applied to various measures of the immune system. The approach of sequencing B-cell variable regions in depth most closely resembles our concept. For example, Asti et al(Asti, Uguzzoni et al. 2016) used deep sequencing data on HIV patients as applied to predict binding to HIV antigens. Using IMS to measure entropy of the antibody repertoire has several advantages. The blood spots for the IMS analysis can be sent through regular mail and only requires a small amount of blood, making large population surveys feasible (Chase, Johnston et al. 2012). The assay itself is simple and inexpensive. We hope that the simplicity of this approach to measuring the humoral immune component will encourage further investigations and applications.

# CHAPTER 7

# CONCLUDING REMARKS

Immunosignature technology is a powerful tool to perform diagnosis on various diseases. The platform itself requires advanced skills both in manufacturing and data analysis. In this thesis I presented my contribution in improving the Immunosignature technology and using Immunosignature to perform diagnosis on various diseases as well as uncovering fundamental biological phenomena.

I first contributed to the optimization of the immunosignature platform by introducing scoring metrics to select optimal parameters considering performance as well as practicality. Next, I primarily worked on identifying a signature shared across various pathogens that can distinguish them from the healthy population. I further retrieved consensus epitopes from the disease common signature and proposed that most pathogens could share the signature by studying the enrichment of the common signature in the pathogen proteomes. Following this, I worked on studying cancer samples from different stages and correlated the immune response with whether the epitope presented by tumor is similar to pathogen. An effective immune response is defined as an antibody titer increasing followed by decrease, suggesting elimination of the epitope. I found that an effective immune response usually correlates with epitopes that are more similar to pathogens. This suggests that the immune system might have a limit and can be effective against only certain epitopes that have similarity with pathogens. I then participated in the attempt to solve the antibiotic resistance problem by developing a classification algorithm that can distinguish bacterial versus viral infection. This algorithm outperforms other currently available classification methods. Finally, I worked on the concept of deriving a single number to represent all the data on the immunosignature platform. This resembles the concept of temperature, which is an indirect measurement of whether an individual is healthy. The measure of Immune Entropy was found to work best as a single measurement to describe the immune system information derived from the immunosignature. Entropy is relatively invariant in a healthy population, but shows significant differences when comparing healthy donors with patients either infected with a pathogen or have cancer.

The future of healthcare relies on early diagnosis of diseases. Immunosignature is a good choice to fulfill this task because of its ability to diagnosis various diseases simultaneously with high accuracy in single assay and its low cost. No other technology has the same capacity like Immunosignature. My work during my Ph.D. study presents some unique usages of Immunosignature and moves one step closer for Immunosignature to become a single test for diagnosing all diseases.

# REFERENCES

(2014). Longitude Prize.

Abraham, E. P., E. Chain, C. M. Fletcher, A. D. Gardner, N. G. Heatley, M. A. Jennings and H. W. Florey (1941). "Further observations on penicillin." The Lancet **238**(6155): 177-189.

Aebersold, R. and M. Mann (2003). "Mass spectrometry-based proteomics." Nature **422**(6928): 198-207.

Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic acids research **25**(17): 3389-3402.

Anderl, J. N., M. J. Franklin and P. S. Stewart (2000). "Role of antibiotic penetration limitation in Klebsiella pneumoniae biofilm resistance to ampicillin and ciprofloxacin." Antimicrobial agents and chemotherapy **44**(7): 1818-1824.

Andersen, R. S., C. A. Thrue, N. Junker, R. Lyngaa, M. Donia, E. Ellebaek, I. M. Svane, T. N. Schumacher, P. Thor Straten and S. R. Hadrup (2012). "Dissection of T-cell antigen specificity in human melanoma." Cancer Res **72**(7): 1642-1650.

Andor, N., T. A. Graham, M. Jansen, L. C. Xia, C. A. Aktipis, C. Petritsch, H. P. Ji and C. C. Maley (2016). "Pan-cancer analysis of the extent and consequences of intra-tumor heterogeneity." Nature medicine **22**(1): 105.

Andresen, H. and C. Grotzinger (2009). "Deciphering the antibodyome-peptide arrays for serum antibody biomarker diagnostics." Current Proteomics **6**(1): 1-12.

Angenendt, P. (2005). "Progress in protein and antibody microarray technology." Drug Discovery Today **10**(7): 503-511.

Angenendt, P., J. Glökler, J. Sobek, H. Lehrach and D. J. Cahill (2003). "Next generation of protein microarray support materials:: Evaluation for protein and antibody microarray applications." Journal of chromatography A **1009**(1): 97-104.

Aoki-Ota, M., A. Torkamani, T. Ota, N. Schork and D. Nemazee (2012). "Skewed primary Igkappa repertoire and V-J joining in C57BL/6 mice: implications for recombination accessibility and receptor editing." J Immunol **188**(5): 2305-2315.

Aran, D., M. Sirota and A. J. Butte (2015). "Systematic pan-cancer analysis of tumour purity." Nature communications **6**.

Arnaout, R., W. Lee, P. Cahill, T. Honan, T. Sparrow, M. Weiand, C. Nusbaum, K. Rajewsky and S. B. Koralov (2011). "High-resolution description of antibody heavy-chain repertoires in humans." PLoS One **6**(8): e22365.

Asti, L., G. Uguzzoni, P. Marcatili and A. Pagnani (2016). "Maximum-Entropy Models of Sequenced Immune Repertoires Predict Antigen-Antibody Affinity." PLoS Comput Biol **12**(4): e1004870.

Auer, H., S. Lyianarachchi, D. Newsom, M. I. Klisovic and K. Kornacker (2003). "Chipping away at the chip bias: RNA degradation in microarray analysis." Nature genetics **35**(4): 292-293.

Baidya, B., O. S. Dandekar and V. K. Singh (2016). Photolithography mask synthesis for spacer patterning, Google Patents.

Bailey, T. L., M. Boden, F. A. Buske, M. Frith, C. E. Grant, L. Clementi, J. Ren, W. W. Li and W. S. Noble (2009). "MEME SUITE: tools for motif discovery and searching." Nucleic Acids Res **37**(Web Server issue): W202-208.

Bailey, T. L. and C. Elkan (1994). "Fitting a mixture model by expectation maximization to discover motifs in bipolymers."

Ballew, J. T., J. A. Murray, P. Collin, M. Mäki, M. F. Kagnoff, K. Kaukinen and P. S. Daugherty (2013). "Antibody biomarker discovery through in vitro directed evolution of consensus recognition epitopes." Proceedings of the National Academy of Sciences **110**(48): 19330-19335.

Beck, I. L., M. G. McKeown and E. S. McCaslin (1983). "Vocabulary development: All contexts are not created equal." The Elementary School Journal **83**(3): 177-181.

Begitt, A., M. Droescher, T. Meyer, C. D. Schmid, M. Baker, F. Antunes, K. P. Knobeloch, M. R. Owen, R. Naumann, T. Decker and U. Vinkemeier (2014). "STAT1-cooperative DNA binding distinguishes type 1 from type 2 interferon signaling." Nat Immunol **15**(2): 168-176.

Begovich, A. B., V. E. Carlton, L. A. Honigberg, S. J. Schrodi, A. P. Chokkalingam, H. C. Alexander, K. G. Ardlie, Q. Huang, A. M. Smith, J. M. Spoerke, M. T. Conn, M. Chang, S. Y. Chang, R. K. Saiki, J. J. Catanese, D. U. Leong, V. E. Garcia, L. B. McAllister, D.

A. Jeffery, A. T. Lee, F. Batliwalla, E. Remmers, L. A. Criswell, M. F. Seldin, D. L. Kastner, C. I. Amos, J. J. Sninsky and P. K. Gregersen (2004). "A missense single-nucleotide polymorphism in a gene encoding a protein tyrosine phosphatase (PTPN22) is associated with rheumatoid arthritis." Am J Hum Genet **75**(2): 330-337.

Bell, B. G., F. Schellevis, E. Stobberingh, H. Goossens and M. Pringle (2014). "A systematic review and meta-analysis of the effects of antibiotic consumption on antibiotic resistance." BMC infectious diseases **14**(1): 13.

Bertone, P. and M. Snyder (2005). "Advances in functional protein microarray technology." The FEBS journal **272**(21): 5400-5411.

Best, M. G., N. Sol, I. Kooi, J. Tannous, B. A. Westerman, F. Rustenburg, P. Schellen, H. Verschueren, E. Post and J. Koster (2015). "RNA-Seq of tumor-educated platelets enables blood-based pan-cancer, multiclass, and molecular pathway cancer diagnostics." Cancer cell **28**(5): 666-676.

Bigelow, C. C. (1967). "On the average hydrophobicity of proteins and the relation between it and protein structure." Journal of Theoretical Biology **16**(2): 187-211.

Bilek, M. M. (2014). "Biofunctionalization of surfaces by energetic ion implantation: review of progress on applications in implantable biomedical devices and antibody microarrays." Applied Surface Science **310**: 3-10.

Boone, S. A. and C. P. Gerba (2007). "Significance of fomites in the spread of respiratory and enteric viral disease." Appl Environ Microbiol **73**(6): 1687-1696.

Borrebaeck, C. A. (2017). "Precision diagnostics: moving towards protein biomarker signatures of clinical utility in cancer." Nature Reviews Cancer **17**(3): 199-204.

Briney, B. S., J. R. Willis, B. A. McKinney and J. E. Crowe, Jr. (2012). "High-throughput antibody sequencing reveals genetic evidence of global regulation of the naive and memory repertoires that extends across individuals." Genes Immun **13**(6): 469-473.

Brown, J., P. Stafford, S. Johnston and V. Dinu (2011). "Statistical methods for analyzing immunosignatures." BMC Bioinformatics **12**(1): 349.

Butte, A. J. and I. S. Kohane (2000). Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. Pac Symp Biocomput.

Camargo, M. C., A. García, A. Riquelme, W. Otero, C. A. Camargo, T. Hernandez-García, R. Candia, M. G. Bruce and C. S. Rabkin (2014). "The problem of Helicobacter pylori resistance to antibiotics: a systematic review in Latin America." The American journal of gastroenterology **109**(4): 485-495.

Castro, M. A., T. T. Onsten, R. M. de Almeida and J. C. Moreira (2005). "Profiling cytogenetic diversity with entropy-based karyotypic analysis." J Theor Biol **234**(4): 487-495.

Chandler, D. (1987). "Introduction to modern statistical mechanics." Introduction to Modern Statistical Mechanics, by David Chandler, pp. 288. Foreword by David Chandler. Oxford University Press, Sep 1987. ISBN-10: 0195042778. ISBN-13: 9780195042771: 288.

Chase, B. A., S. A. Johnston and J. B. Legutki (2012). "Evaluation of biological sample preparation for immunosignature-based diagnostics." Clinical and Vaccine Immunology **19**(3): 352-358.

Chen, H., J. Liu, B. A. Merrick and M. P. Waalkes (2001). "Genetic events associated with arsenic-induced malignant transformation: applications of cDNA microarray technology." Molecular carcinogenesis **30**(2): 79-87.

Chen, Y. and T. D. Pham (2013). "Sample entropy and regularity dimension in complexity analysis of cortical surface structure in early Alzheimer's disease and aging." J Neurosci Methods **215**(2): 210-217.

Chizhikov, V., A. Rasooly, K. Chumakov and D. D. Levy (2001). "Microarray analysis of microbial virulence factors." Applied and environmental microbiology **67**(7): 3258-3263.

Choi, J.-W., K. W. Oh, J. H. Thomas, W. R. Heineman, H. B. Halsall, J. H. Nevin, A. J. Helmicki, H. T. Henderson and C. H. Ahn (2002). "An integrated microfluidic biochemical detection system for protein analysis with magnetic bead-based sampling capabilities." Lab on a Chip **2**(1): 27-30.

Cleven, B. E., M. Palka-Santini, J. Gielen, S. Meembor, M. Krönke and O. Krut (2006). "Identification and characterization of bacterial pathogens causing bloodstream infections by DNA microarray." Journal of clinical microbiology **44**(7): 2389-2397.

Cole, W. C., S. J. DeNardo, C. F. Meares, M. J. McCall, G. L. DeNardo, A. L. Epstein, H. A. O'Brien and M. K. Moi (1987). "Comparative serum stability of radiochelates for antibody radiopharmaceuticals." Journal of nuclear medicine: official publication, Society of Nuclear Medicine **28**(1): 83-90.

Cover, T. M. and J. A. Thomas (2012). Elements of information theory, John Wiley & Sons.

Cox, N. and K. Subbarao (2000). "Global epidemiology of influenza: past and present." Annual review of medicine **51**(1): 407-421.

Cressler, C. E., D. V. McLEOD, C. Rozins, J. Van Den Hoogen and T. Day (2016). "The adaptive evolution of virulence: a review of theoretical predictions and empirical tests." Parasitology **143**(7): 915-930.

Cretich, M., F. Damin, G. Pirri and M. Chiari (2006). "Protein and peptide arrays: recent trends and new directions." Biomolecular engineering **23**(2): 77-88.

Cummings, C. A. and D. A. Relman (2000). "Using DNA microarrays to study host-microbe interactions." Emerging infectious diseases **6**(5): 513.

Cunha, B. A. (2010). Antibiotic Essentials 2009, Jones & Bartlett Publishers.

Cywes-Bentley, C., D. Skurnik, T. Zaidi, D. Roux, R. B. DeOliveira, W. S. Garrett, X. Lu, J. O’Malley, K. Kinzel and T. Zaidi (2013). "Antibody to a conserved antigenic target is protective against diverse prokaryotic and eukaryotic pathogens." Proceedings of the National Academy of Sciences **110**(24): E2209-E2218.

D'costa, V. M., C. E. King, L. Kalan, M. Morar, W. W. Sung, C. Schwarz, D. Froese, G. Zazula, F. Calmels and R. Debruyne (2011). "Antibiotic resistance is ancient." Nature **477**(7365): 457.

Davies, D. H., X. Liang, J. E. Hernandez, A. Randall, S. Hirst, Y. Mu, K. M. Romero, T. T. Nguyen, M. Kalantari-Dehaghi, S. Crotty, P. Baldi, L. P. Villarreal and P. L. Felgner (2005). "Profiling the humoral immune response to infection by using proteome microarrays: High-throughput vaccine and diagnostic antigen discovery." Proceedings of the National Academy of Sciences of the United States of America **102**(3): 547-552.

Davies, J. and D. Davies (2010). "Origins and evolution of antibiotic resistance." Microbiol Mol Biol Rev **74**(3): 417-433.

de Arruda, P. F. F., M. Gatti, F. N. F. Junior, J. G. F. de Arruda, R. D. Moreira, L. O. Murta, L. F. de Arruda and M. F. de Godoy (2013). "Quantification of fractal dimension and Shannon’s entropy in histological diagnosis of prostate cancer." BMC clinical pathology **13**(1): 6.

Degliangeli, F., P. Kshirsagar, V. Brunetti, P. P. Pompa and R. Fiammengo (2014). "Absolute and direct microRNA quantification using DNA–gold nanoparticle probes." Journal of the American Chemical Society **136**(6): 2264-2267.

Diehl, F., S. Grahlmann, M. Beier and J. D. Hoheisel (2001). "Manufacturing DNA microarrays of high spot homogeneity and reduced background signal." Nucleic acids research **29**(7): e38-e38.

Dodge, Y. (2006). The Oxford dictionary of statistical terms, Oxford University Press on Demand.

Donnell, B., A. Maurer, A. Papandreou-Suppappola and P. Stafford (2015). "Time-Frequency Analysis of Peptide Microarray Data: Application to Brain Cancer Immunosignatures." Cancer Informatics(4906-CIN-Time-Frequency-Analysis-of-Peptide-Microarray-Data:-Application-to-Bra.pdf): 219-233.

Dorval, A. D., G. S. Russo, T. Hashimoto, W. Xu, W. M. Grill and J. L. Vitek (2008). "Deep brain stimulation reduces neuronal entropy in the MPTP-primate model of Parkinson's disease." J Neurophysiol **100**(5): 2807-2818.

Duburcq, X., C. Olivier, F. Malingue, R. Desmet, A. Bouzidi, F. Zhou, C. Auriault, H. Gras-Masse and O. Melnyk (2004). "Peptide− protein microarrays for the simultaneous detection of pathogen infections." Bioconjugate chemistry **15**(2): 307-316.

Duttagupta, R., R. Jiang, J. Gollub, R. C. Getts and K. W. Jones (2011). "Impact of cellular miRNAs on circulating miRNA biomarker signatures." PloS one **6**(6): e20769.

Eccles, M. P., J. M. Grimshaw, M. Johnston, N. Steen, N. B. Pitts, R. Thomas, E. Glidewell, G. Maclennan, D. Bonetti and A. Walker (2007). "Applying psychological theories to evidence-based clinical practice: Identifying factors predictive of managing upper respiratory tract infections without antibiotics." Implementation Science **2**(1): 26.

Erkes, D. A., T. Mohgbeli and C. M. Snyder (2015). "Virus-specific CD8+ T cells infiltrate melanoma lesions and retain function despite high PD-1 expression." Journal for ImmunoTherapy of Cancer **3**(Suppl 2): O6.

Ettinger, S. J. and E. C. Feldman (2009). Textbook of Veterinary Internal Medicine-eBook, Elsevier health sciences.

Felsenfeld, O., I. F. Volini, S. J. Ishihara, M. C. Bachman and V. M. Young (1950). "A study of the effect of neomycin and other antibiotics on bacteria, viruses, and protozoa." The Journal of Laboratory and Clinical Medicine **35**(3): 428-433.

Finberg, R. W., R. C. Moellering, F. P. Tally, W. A. Craig, G. A. Pankey, E. P. Dellinger, M. A. West, M. Joshi, P. K. Linden and K. V. Rolston (2004). "The importance of bactericidal drugs: future directions in infectious disease." Clinical infectious diseases **39**(9): 1314-1320.

Ford, D., D. F. Easton, M. Stratton, S. Narod, D. Goldgar, P. Devilee, D. T. Bishop, B. Weber, G. Lenoir, J. Chang-Claude, H. Sobol, M. D. Teare, J. Struewing, A. Arason, S. Scherneck, J. Peto, T. R. Rebbeck, P. Tonin, S. Neuhausen, R. Barkardottir, J. Eyfjord, H. Lynch, B. A. J. Ponder, S. A. Gayther, J. M. Birch, A. Lindblom, D. Stoppa-Lyonnet, Y. Bignon, A. Borg, U. Hamann, N. Haites, R. J. Scott, C. M. Maugard, H. Vasen, S. Seitz, L. A. Cannon-Albright, A. Schofield and M. Zelada-Hedman (1998). "Genetic Heterogeneity and Penetrance Analysis of the BRCA1 and BRCA2 Genes in Breast Cancer Families." The American Journal of Human Genetics **62**(3): 676-689.

Frith, M. C., N. F. Saunders, B. Kobe and T. L. Bailey (2008). "Discovering sequence motifs with arbitrary insertions and deletions." PLoS Comput Biol **4**(4): e1000071.

Garratty, G., S. A. Glynn and R. McEntire (2004). "ABO and Rh (D) phenotype frequencies of different racial/ethnic groups in the United States." Transfusion **44**(5): 703-706.

Gaseitsiwe, S., D. Valentini, S. Mahdavifar, I. Magalhaes, D. F. Hoft, J. Zerweck, M. Schutkowski, J. Andersson, M. Reilly and M. J. Maeurer (2008). "Pattern recognition in pulmonary tuberculosis defined by high content peptide microarray chip analysis representing 61 proteins from M. tuberculosis." PloS one **3**(12): e3840.

Gasteiger, E., C. Hoogland, A. Gattiker, S. e. Duvaud, M. R. Wilkins, R. D. Appel and A. Bairoch (2005). Protein identification and analysis tools on the ExPASy server, Springer.

Ge, H. (2000). "UPA, a universal protein array system for quantitative detection of protein–protein, protein–DNA, protein–RNA and protein–ligand interactions." Nucleic Acids Research **28**(2): e3-e3.

Geijersstam, V., M. Kibur, Z. Wang, P. Koskela, E. Pukkala, J. Schiller, M. Lehtinen and J. Dillner (1998). "Stability over time of serum antibody levels to human papillomavirus type 16." The Journal of Infectious Diseases **177**(6): 1710-1714.

Georganopoulou, D. G., L. Chang, J.-M. Nam, C. S. Thaxton, E. J. Mufson, W. L. Klein and C. A. Mirkin (2005). "Nanoparticle-based detection in cerebral spinal fluid of a soluble pathogenic biomarker for Alzheimer's disease." Proceedings of the National Academy of Sciences of the United States of America **102**(7): 2273-2276.

Georgiou, G., G. C. Ippolito, J. Beausang, C. E. Busse, H. Wardemann and S. R. Quake (2014). "The promise and challenge of high-throughput sequencing of the antibody repertoire." Nat Biotechnol **32**(2): 158-168.

Ghotaslou, R., H. E. Leylabadlo and Y. M. Asl (2015). "Prevalence of antibiotic resistance in Helicobacter pylori: A recent literature review." World journal of methodology **5**(3): 164.

Gladyshev, G. P. (1999). "On thermodynamics, entropy and evolution of biological systems: What is life from a physical chemist's viewpoint." Entropy **1**(2): 9-20.

Goossens, H., M. Ferech, R. Vander Stichele, M. Elseviers and E. P. Group (2005). "Outpatient antibiotic use in Europe and association with resistance: a cross-national database study." The Lancet **365**(9459): 579-587.

Gottardo, F., C. G. Liu, M. Ferracin, G. A. Calin, M. Fassan, P. Bassi, C. Sevignani, D. Byrne, M. Negrini and F. Pagano (2007). Micro-RNA profiling in kidney and bladder cancers. Urologic Oncology: Seminars and Original Investigations, Elsevier.

Gunderson, K. L., F. J. Steemers, G. Lee, L. G. Mendoza and M. S. Chee (2005). "A genome-wide scalable SNP genotyping assay using microarray technology." Nat Genet **37**(5): 549-554.

Gupta, N., J. Lainson, V. Domenyuk, Z.-G. Zhao, S. A. Johnston and C. W. Diehnelt (2016). "Whole-Virus Screening to Develop Synbodies for the Influenza Virus." Bioconjugate chemistry **27**(10): 2505-2512.

Gupta, N., J. C. Lainson, P. E. Belcher, L. Shen, H. S. Mason, S. A. Johnston and C. W. Diehnelt (2017). "Cross-reactive synbody affinity ligands for capturing diverse noroviruses." Analytical chemistry **89**(13): 7174-7181.

Haab, B. B. (2005). "Antibody arrays in cancer research." Molecular & cellular proteomics **4**(4): 377-383.

Hall, D. A., J. Ptacek and M. Snyder (2007). "Protein microarray technology." Mechanisms of ageing and development **128**(1): 161-167.

Halperin, R. F., P. Stafford, J. B. Legutki and S. A. Johnston (2010). "Exploring antibody recognition of sequence space through random-sequence peptide microarrays." Molecular and Cellular Proteomics **28**(1): e101230.101236.

Hamilton, J. P., E. C. Neeno-Eckwall, B. N. Adhikari, N. T. Perna, N. Tisserat, J. E. Leach, C. A. Levesque and C. R. Buell (2011). "The Comprehensive Phytopathogen Genomics Resource: a web-based resource for data-mining plant pathogen genomes." Database (Oxford) **2011**: bar053.

Hanahan, D. and R. A. Weinberg (2000). "The hallmarks of cancer." cell **100**(1): 57-70.

Hanahan, D. and R. A. Weinberg (2011). "Hallmarks of cancer: the next generation." Cell **144**(5): 646-674.

Hartmann, M., J. Roeraade, D. Stoll, M. F. Templin and T. O. Joos (2009). "Protein microarrays for diagnostic assays." Analytical and bioanalytical chemistry **393**(5): 1407-1416.

Hawkey, P. M. and A. M. Jones (2009). "The changing epidemiology of resistance." Journal of Antimicrobial Chemotherapy **64**(suppl_1): i3-i10.

Hegde, P., R. Qi, K. Abernathy, C. Gay, S. Dharap, R. Gaspard, J. Hughes, E. Snesrud, N. Lee and J. Quackenbush (2000). "A concise guide to cDNA microarray analysis." Biotechniques **29**(3): 548-563.

Heller, M. J. (2002). "DNA microarray technology: devices, systems, and applications." Annual review of biomedical engineering **4**(1): 129-153.

Honeyfield, J. (1977). "Word frequency and the importance of context in vocabulary learning." RELC journal **8**(2): 35-42.

Hu, B., X. Niu, L. Cheng, L. N. Yang, Q. Li, Y. Wang, S. C. Tao and S. M. Zhou (2015). "Discovering cancer biomarkers from clinical samples by protein microarrays." PROTEOMICS-Clinical Applications **9**(1-2): 98-110.

Hueber, W., P. J. Utz, L. Steinman and W. H. Robinson (2002). "Autoantibody profiling for the study and treatment of autoimmune disease." Arthritis Research & Therapy **4**(5): 290.

Hughes, A. K., Z. Cichacz, A. Scheck, S. W. Coons, S. A. Johnston and P. Stafford (2012). "Immunosignaturing can detect products from molecular markers in brain cancer." PloS one **7**(7): e40201.

Huttner, B., H. Goossens, T. Verheij and S. Harbarth (2010). "Characteristics and outcomes of public campaigns aimed at improving the use of antibiotics in outpatients in high-income countries." The Lancet infectious diseases **10**(1): 17-31.

Ivshina, A. V., J. George, O. Senko, B. Mow, T. C. Putti, J. Smeds, T. Lindahl, Y. Pawitan, P. Hall and H. Nordgren (2006). "Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer." Cancer research **66**(21): 10292-10301.

Joanes, D. and C. Gill (1998). "Comparing measures of sample skewness and kurtosis." Journal of the Royal Statistical Society: Series D (The Statistician) **47**(1): 183-189.

Johnson, M., I. Zaretskaya, Y. Raytselis, Y. Merezhuk, S. McGinnis and T. L. Madden (2008). "NCBI BLAST: a better web interface." Nucleic acids research **36**(suppl_2): W5-W9.

Johnston, S. A., D. H. Thamm and J. B. Legutki (2014). "The immunosignature of canine lymphoma: characterization and diagnostic application." BMC cancer **14**(1): 657.

Jones, J. D. and J. L. Dangl (2006). "The plant immune system." Nature **444**(7117): 323-329.

Jost, L. (2006). "Entropy and diversity." Oikos **113**(2): 363-375.

Kardos, N. and A. L. Demain (2011). "Penicillin: the medicine with the greatest impact on therapeutic outcomes." Applied microbiology and biotechnology **92**(4): 677.

Kawakami, Y., T. Fujita, Y. Matsuzaki, T. Sakurai, M. Tsukamoto, M. Toda and H. Sumimoto (2004). "Identification of human tumor antigens and its implications for diagnosis and treatment of cancer." Cancer science **95**(10): 784-791.

Keasey, S. L., K. E. Schmid, M. S. Lee, J. Meegan, P. Tomas, M. Minto, A. P. Tikhonov, B. Schweitzer and R. G. Ulrich (2009). "Extensive antibody cross-reactivity among infectious gram-negative bacteria revealed by proteome microarray analysis." Mol Cell Proteomics **8**(5): 924-935.

Keating, C. D. (2005). "Nanoscience enables ultrasensitive detection of Alzheimer's biomarker." Proceedings of the National Academy of Sciences of the United States of America **102**(7): 2263-2264.

Kingsmore, S. F. (2006). "Multiplexed protein measurement: technologies and applications of protein and antibody arrays." Nature reviews. Drug discovery **5**(4): 310.

Király, L., A. Künstler, R. Bacsó, Y. Hafez and Z. Király (2013). "Similarities and differences in plant and animal immune systems — what is inhibiting pathogens?" Acta Phytopathologica et Entomologica Hungarica **48**(2): 187-205.

Knudsen, E., G. N. Rolinson and R. Sutherland (1967). "Carbenicillin: a new semisynthetic penicillin active against Pseudomonas pyocyanea." British medical journal **3**(5557): 75.

Kollef, M. H. (2008). "Broad-spectrum antimicrobials and the treatment of serious bacterial infections: getting it right up front." Clin Infect Dis **47 Suppl 1**: S3-13.

Kringelum, J. V., M. Nielsen, S. B. Padkjær and O. Lund (2013). "Structural analysis of B-cell epitopes in antibody: protein complexes." Molecular immunology **53**(1): 24-34.

Krupp, K. and P. Madhivanan (2015). "Antibiotic resistance in prevalent bacterial and protozoan sexually transmitted infections." Indian journal of sexually transmitted diseases **36**(1): 3.

Kukreja, M., S. A. Johnston and P. Stafford (2012). "Comparative study of classification algorithms for immunosignaturing data." BMC Bioinformatics **13**(139).

Kukreja, M., S. A. Johnston and P. Stafford (2012). "Immunosignaturing microarrays distinguish antibody profiles of related pancreatic diseases." Proteomics and Bioinformatics **S6**.

Kuppers, R. (2005). "Mechanisms of B-cell lymphoma pathogenesis." Nat Rev Cancer **5**(4): 251-262.

Kwon, J.-a., H. Lee, K. N. Lee, K. Chae, S. Lee, D.-k. Lee and S. Kim (2008). "High diagnostic accuracy of antigen microarray for sensitive detection of hepatitis C virus infection." Clinical chemistry **54**(2): 424-428.

Lainson, J. C., M. F. Fuenmayor, S. A. Johnston and C. W. Diehnelt (2015). "Conjugation approach to produce a Staphylococcus aureus synbody with activity in serum." Bioconjugate chemistry **26**(10): 2125-2132.

Lawrence, M. S., P. Stojanov, P. Polak, G. V. Kryukov, K. Cibulskis, A. Sivachenko, S. L. Carter, C. Stewart, C. H. Mermel, S. A. Roberts, A. Kiezun, P. S. Hammerman, A. McKenna, Y. Drier, L. Zou, A. H. Ramos, T. J. Pugh, N. Stransky, E. Helman, J. Kim, C. Sougnez, L. Ambrogio, E. Nickerson, E. Shefler, M. L. Cortes, D. Auclair, G. Saksena, D. Voet, M. Noble, D. DiCara, P. Lin, L. Lichtenstein, D. I. Heiman, T. Fennell, M. Imielinski, B. Hernandez, E. Hodis, S. Baca, A. M. Dulak, J. Lohr, D. A. Landau, C. J. Wu, J. Melendez-Zajgla, A. Hidalgo-Miranda, A. Koren, S. A. McCarroll, J. Mora, R. S. Lee, B. Crompton, R. Onofrio, M. Parkin, W. Winckler, K. Ardlie, S. B. Gabriel, C. W. Roberts, J. A. Biegel, K. Stegmaier, A. J. Bass, L. A. Garraway, M. Meyerson, T. R. Golub, D. A. Gordenin, S. Sunyaev, E. S. Lander and G. Getz (2013). "Mutational heterogeneity in cancer and the search for new cancer-associated genes." Nature **499**(7457): 214-218.

Lawrence, R., J. R. Brown, K. Al-Mafraji, W. C. Lamanna, J. R. Beitel, G.-J. Boons, J. D. Esko and B. E. Crawford (2012). "Disease-specific non–reducing end carbohydrate biomarkers for mucopolysaccharidoses." Nature chemical biology **8**(2): 197-204.

Legutki, J. B. and S. A. Johnston (2013). "Immunosignatures can predict vaccine efficacy." Proceedings of the National Academy of Sciences.

Legutki, J. B., D. M. Magee, P. Stafford and S. A. Johnston (2010). "A general method for characterization of humoral immunity induced by a vaccine or infection." Vaccine **28**(28): 4529-4537.

Legutki, J. B., Z.-G. Zhao, M. Greving, N. Woodbury, S. A. Johnston and P. Stafford (2014). "Scalable high-density peptide arrays for comprehensive health monitoring." Nature communications **5**.

Leinberger, D. M., U. Schumacher, I. B. Autenrieth and T. T. Bachmann (2005). "Development of a DNA microarray for detection and identification of fungal pathogens involved in invasive mycoses." Journal of clinical microbiology **43**(10): 4943-4953.

Lessa-Aquino, C., J. C. Lindow, A. Randall, E. Wunder, J. Pablo, R. Nakajima, A. Jasinskas, J. S. Cruz, A. O. Damião and N. Nery (2017). "Distinct antibody responses of patients with mild and severe leptospirosis determined by whole proteome microarray analysis." PLoS neglected tropical diseases **11**(1): e0005349.

Levenson, M. D., N. Viswanathan and R. A. Simpson (1982). "Improving resolution in photolithography with a phase-shifting mask." IEEE Transactions on electron devices **29**(12): 1828-1836.

Li, H., S. Wetten, L. Li, P. L. S. Jean, R. Upmanyu, L. Surh, D. Hosford, M. R. Barnes, J. D. Briley and M. Borrie (2008). "Candidate single-nucleotide polymorphisms from a genomewide association study of Alzheimer disease." Archives of neurology **65**(1): 45-53.

Li, J., S. Chen and D. H. Evans (2001). "Typing and subtyping influenza virus using DNA microarrays and multiplex reverse transcriptase PCR." Journal of clinical microbiology **39**(2): 696-704.

Li, Z., B. Zhao, D. Wang, Y. Wen, G. Liu, H. Dong, S. Song and C. Fan (2014). "DNA nanostructure-based universal microarray platform for high-efficiency multiplex bioanalysis in biofluids." ACS applied materials & interfaces **6**(20): 17944-17953.

Liang, P.-H., C.-Y. Wu, W. A. Greenberg and C.-H. Wong (2008). "Glycan arrays: biological and medical applications." Current opinion in chemical biology **12**(1): 86-92.

Ligon, B. L. (2004). Penicillin: its discovery and early development. Seminars in pediatric infectious diseases, Elsevier.

Lueking, A., O. Huber, C. Wirths, K. Schulte, K. M. Stieler, U. Blume-Peytavi, A. Kowald, K. Hensel-Wiegel, R. Tauber and H. Lehrach (2005). "Profiling of alopecia areata autoantigens based on protein microarray technology." Molecular & cellular proteomics **4**(9): 1382-1390.

MacBeath, G. (2002). "Protein microarrays and proteomics." Nat Genet **32 Suppl**: 526-532.

Madden, T. (2013). "The BLAST sequence analysis tool."

Maksimov, P., J. Zerweck, A. Maksimov, A. Hotop, U. Groß, U. Pleyer, K. Spekker, W. Däubener, S. Werdermann and O. Niederstrasser (2012). "Peptide microarray analysis of in silico-predicted epitopes for serological diagnosis of Toxoplasma gondii infection in humans." Clinical and Vaccine Immunology **19**(6): 865-874.

Malin, A., N. Kovvali, A. Papandreou-Suppappola, J. J. Zhang, S. Johnston and P. Stafford (2012). Beta process based adaptive learning for immunosignature microarray feature identification. Signals, Systems and Computers (ASILOMAR), 2012 Conference Record of the Forty Sixth Asilomar Conference on, IEEE.

Mardia, K. V. (1970). "Measures of multivariate skewness and kurtosis with applications." Biometrika **57**(3): 519-530.

Martin, A. B., M. Hartman, B. Washington, A. Catlin and N. H. E. A. Team (2016). "National health spending: faster growth in 2015 as coverage expands and utilization increases." Health Affairs: 10.1377/hlthaff. 2016.1330.

Marty, R., S. Kaabinejadian, D. Rossell, M. J. Slifker, J. van de Haar, H. B. Engin, N. de Prisco, T. Ideker, W. H. Hildebrand and J. Font-Burgada (2017). "MHC-I genotype restricts the oncogenic mutational landscape." Cell **171**(6): 1272-1283. e1215.

Marusyk, A., V. Almendro and K. Polyak (2012). "Intra-tumour heterogeneity: a looking glass for cancer?" Nat Rev Cancer **12**(5): 323-334.

Matsushita, H., M. D. Vesely, D. C. Koboldt, C. G. Rickert, R. Uppaluri, V. J. Magrini, C. D. Arthur, J. M. White, Y. S. Chen, L. K. Shea, J. Hundal, M. C. Wendl, R. Demeter, T. Wylie, J. P. Allison, M. J. Smyth, L. J. Old, E. R. Mardis and R. D. Schreiber (2012). "Cancer exome analysis reveals a T-cell-dependent mechanism of cancer immunoediting." Nature **482**(7385): 400-404.

McGranahan, N., R. Rosenthal, C. T. Hiley, A. J. Rowan, T. B. Watkins, G. A. Wilson, N. J. Birkbak, S. Veeriah, P. Van Loo and J. Herrero (2017). "Allele-specific HLA loss and immune escape in lung cancer evolution." Cell **171**(6): 1259-1271. e1211.

McNulty, C. A., P. Boyle, T. Nichols, P. Clappison and P. Davey (2007). "The public's attitudes to and compliance with antibiotics." Journal of Antimicrobial Chemotherapy **60**(suppl_1): i63-i68.

Meacham, C. E. and S. J. Morrison (2013). "Tumour heterogeneity and cancer cell plasticity." Nature **501**(7467): 328-337.

Medzhitov, R. (2007). "Recognition of microorganisms and activation of the immune response." Nature **449**(7164): 819-826.

Michaud, G. A., M. Salcius, F. Zhou, R. Bangham, J. Bonin, H. Guo, M. Snyder, P. F. Predki and B. I. Schweitzer (2003). "Analyzing antibody specificity with whole proteome microarrays." Nat Biotechnol **21**(12): 1509-1512.

Miller, J. C., H. Zhou, J. Kwekel, R. Cavallo, J. Burke, E. B. Butler, B. S. Teh and B. B. Haab (2003). "Antibody microarray profiling of human prostate cancer sera: antibody screening and identification of potential biomarkers." Proteomics **3**(1): 56-63.

Mishra, A. and M. Verma (2010). "Cancer biomarkers: are we ready for the prime time?" Cancers (Basel) **2**(1): 190-208.

Mitchell, P. S., R. K. Parkin, E. M. Kroh, B. R. Fritz, S. K. Wyman, E. L. Pogosova-Agadjanyan, A. Peterson, J. Noteboom, K. C. O'Briant, A. Allen, D. W. Lin, N. Urban, C. W. Drescher, B. S. Knudsen, D. L. Stirewalt, R. Gentleman, R. L. Vessella, P. S. Nelson, D. B. Martin and M. Tewari (2008). "Circulating microRNAs as stable blood-based markers for cancer detection." Proc Natl Acad Sci U S A **105**(30): 10513-10518.

Mitscher, L. A., S. P. Pillai, E. J. Gentry and D. M. Shankel (1999). "Multiple drug resistance." Medicinal research reviews **19**(6): 477-496.

Montecalvo, M. A., H. Horowitz, C. Gedris, C. Carbonaro, F. C. Tenover, A. Issah, P. Cook and G. P. Wormser (1994). "Outbreak of vancomycin-, ampicillin-, and aminoglycoside-resistant Enterococcus faecium bacteremia in an adult oncology unit." Antimicrobial Agents and Chemotherapy **38**(6): 1363-1367.

Moran, S., C. Arribas and M. Esteller (2016). "Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences." Epigenomics **8**(3): 389-399.

Navalkar, K., D. M. Magee, J. Galgiani, Z. Cichacz, S. A. Johnston and P. Stafford (2014). "Application of immunosignatures to diagnosis of Valley Fever." Clinical and Vaccine Immunology **in press**.

Navalkar, K. A. (2014). Antibody based strategies for multiplexed diagnostics, Arizona State University.

Navalkar, K. A., S. A. Johnston and P. Stafford (2015). "Peptide based diagnostics: Are random-sequence peptides more useful than tiling proteome sequences?" Journal of Immunological Methods **417**: 10-21.

Nguyen, D. V. and D. M. Rocke (2002). "Tumor classification by partial least squares using microarray gene expression data." Bioinformatics **18**(1): 39-50.

Notkins, A. L. (2004). "Polyreactivity of antibody molecules." Trends in immunology **25**(4): 174-179.

Nuwaysir, E. F., W. Huang, T. J. Albert, J. Singh, K. Nuwaysir, A. Pitas, T. Richmond, T. Gorski, J. P. Berg and J. Ballin (2002). "Gene expression analysis using oligonucleotide arrays produced by maskless photolithography." Genome research **12**(11): 1749-1755.

OBAMA, B. (2014). Executive Order—Combating Antibiotic-Resistant Bacteria.

Ochsenbein, A. F., T. Fehr, C. Lutz, M. Suter, F. Brombacher, H. Hengartner and R. M. Zinkernagel (1999). "Control of early viral and bacterial distribution and disease by natural antibodies." Science **286**(5447): 2156-2159.

Ogilvie, G. K., B. E. Powers, C. H. Mallinckrodt and S. J. Withrow (1996). "Surgery and doxorubicin in dogs with hemangiosarcoma." Journal of Veterinary Internal Medicine **10**(6): 379-384.

Organization, W. H. (2004). "The World health report: 2004: changing history."

Oved, K., A. Cohen, O. Boico, R. Navon, T. Friedman, L. Etshtein, O. Kriger, E. Bamberger, Y. Fonar, R. Yacobov, R. Wolchinsky, G. Denkberg, Y. Dotan, A. Hochberg, Y. Reiter, M. Grupper, I. Srugo, P. Feigin, M. Gorfine, I. Chistyakov, R. Dagan, A. Klein, I. Potasman and E. Eden (2015). "A novel host-proteome signature for distinguishing between acute bacterial and viral infections." PLoS One **10**(3): e0120012.

Packer, N. H., C. W. von der Lieth, K. F. Aoki-Kinoshita, C. B. Lebrilla, J. C. Paulson, R. Raman, P. Rudd, R. Sasisekharan, N. Taniguchi and W. S. York (2008). "Frontiers in glycomics: Bioinformatics and biomarkers in disease An NIH White Paper prepared from discussions by the focus groups at a workshop on the NIH campus, Bethesda MD (September 11–13, 2006)." Proteomics **8**(1): 8-20.

Pankey, G. and L. Sabath (2004). "Clinical relevance of bacteriostatic versus bactericidal mechanisms of action in the treatment of Gram-positive bacterial infections." Clinical infectious diseases **38**(6): 864-870.

Park, T. (2016). "1488 Toxicity of antibiotics on rumen protozoan and its associated microbes." Journal of Animal Science **94**(supplement5): 722-722.

Parks, G. A. (1967). Aqueous surface chemistry of oxides and complex oxide minerals: Isoelectric point and zero point of charge, ACS Publications.

Patel, A. P., I. Tirosh, J. J. Trombetta, A. K. Shalek, S. M. Gillespie, H. Wakimoto, D. P. Cahill, B. V. Nahed, W. T. Curry and R. L. Martuza (2014). "Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma." Science **344**(6190): 1396-1401.

Periyannan Rajeswari, P. K., L. M. Soderberg, A. Yacoub, M. Leijon, H. Andersson Svahn and H. N. Joensson (2017). "Multiple pathogen biomarker detection using an encoded bead array in droplet PCR." J Microbiol Methods **139**: 22-28.

Petrik, J. (2001). "Microarray technology: the future of blood testing?" Vox sanguinis **80**(1): 1-11.

Pritchard, C. C., E. Kroh, B. Wood, J. D. Arroyo, K. J. Dougherty, M. M. Miyaji, J. F. Tait and M. Tewari (2012). "Blood cell origin of circulating microRNAs: a cautionary note for cancer biomarker studies." Cancer prevention research **5**(3): 492-497.

Pruitt, K. D., T. Tatusova and D. R. Maglott (2005). "NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins." Nucleic Acids Res **33**(Database issue): D501-504.

Qiu, J., J. Madoz-Gurpide, D. E. Misek, R. Kuick, D. E. Brenner, G. Michailidis, B. B. Haab, G. S. Omenn and S. Hanash (2004). "Development of natural protein microarrays for diagnosing cancer based on an antibody response to tumor antigens." Journal of proteome research **3**(2): 261-267.

Raghavan, M., D. M. Lillington, S. Skoulakis, S. Debernardi, T. Chaplin, N. J. Foot, T. A. Lister and B. D. Young (2005). "Genome-wide single nucleotide polymorphism analysis reveals frequent partial uniparental disomy due to somatic recombination in acute myeloid leukemias." Cancer research **65**(2): 375-378.

Rajewsky, K. (1996). "Clonal selection and learning in the antibody system." Nature **381**(6585): 751.

Rea, M. C., A. Dobson, O. O'Sullivan, F. Crispie, F. Fouhy, P. D. Cotter, F. Shanahan, B. Kiely, C. Hill and R. P. Ross (2011). "Effect of broad-and narrow-spectrum antimicrobials on Clostridium difficile and microbial diversity in a model of the distal colon." Proceedings of the National Academy of Sciences **108**(Supplement 1): 4639-4644.

Reiman, J. M., M. Kmieciak, M. H. Manjili and K. L. Knutson (2007). "Tumor immunoediting and immunosculpting pathways to cancer progression." Seminars in Cancer Biology **17**(4): 275-287.

Ren, S., Z. Peng, J.-H. Mao, Y. Yu, C. Yin, X. Gao, Z. Cui, J. Zhang, K. Yi and W. Xu (2012). "RNA-seq analysis of prostate cancer in the Chinese population identifies recurrent gene fusions, cancer-associated long noncoding RNAs and aberrant alternative splicings." Cell research **22**(5): 806.

Rényi, A. (1961). On measures of entropy and information. Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics, The Regents of the University of California.

Restrepo, L., P. Stafford and S. A. Johnston (2012). "Feasibility of an early Alzheimer's disease immunosignature diagnostic test." Journal of Neuroimmunology.

Restrepo, L., P. Stafford and S. A. Johnston (2013). "Feasibility of an early Alzheimer's disease immunosignature diagnostic test." Journal of neuroimmunology **254**(1): 154-160.

Restrepo, L., P. Stafford, D. M. Magee and S. A. Johnston (2011). "Application of immunosignatures to the assessment of Alzheimer's disease." Annals of Neurology: 5-18.

Reya, T., S. J. Morrison, M. F. Clarke and I. L. Weissman (2001). "Stem cells, cancer, and cancer stem cells." nature **414**(6859): 105-111.

Riaz, N., L. Morris, J. J. Havel, V. Makarov, A. Desrichard and T. A. Chan (2016). "The role of neoantigens in response to immune checkpoint blockade." Int Immunol **28**(8): 411-419.

Richer, J., S. A. Johnston and P. Stafford (2015). "Epitope Identification from Fixed-complexity Random-sequence Peptide Microarrays." Molecular & Cellular Proteomics **14**(1): 136-147.

Rifai, N., M. A. Gillette and S. A. Carr (2006). "Protein biomarker discovery and validation: the long and uncertain path to clinical utility." Nat Biotech **24**(8): 971-983.

Ritchie, W., S. Granjeaud, D. Puthier and D. Gautheret (2008). "Entropy measures quantify global splicing disorders in cancer." PLoS Comput Biol **4**(3): e1000011.

Rizvi, N. A., M. D. Hellmann, A. Snyder, P. Kvistborg, V. Makarov, J. J. Havel, W. Lee, J. Yuan, P. Wong and T. S. Ho (2015). "Mutational landscape determines sensitivity to PD-1 blockade in non–small cell lung cancer." Science **348**(6230): 124-128.

Rödiger, S., C. Liebsch, C. Schmidt, W. Lehmann, U. Resch-Genger, U. Schedler and P. Schierack (2014). "Nucleic acid detection based on the use of microbeads: a review." Microchimica Acta **181**(11-12): 1151-1168.

Rose, G. D., A. R. Geselowitz, G. J. Lesser, R. H. Lee and M. H. Zehfus (1985). "Hydrophobicity of amino acid residues in globular proteins." Science **229**: 834-839.

Rossolini, G. M., F. Arena, P. Pecile and S. Pollini (2014). "Update on the antibiotic resistance crisis." Current opinion in pharmacology **18**: 56-60.

Rubinstein, N. D., I. Mayrose, D. Halperin, D. Yekutieli, J. M. Gershoni and T. Pupko (2008). "Computational characterization of B-cell epitopes." Molecular immunology **45**(12): 3477-3489.

Rucco, M., F. Castiglione, E. Merelli and M. Pettini (2016). Characterisation of the Idiotypic Immune Network Through Persistent Entropy. Proceedings of ECCS 2014: European Conference on Complex Systems. S. Battiston, F. De Pellegrini, G. Caldarelli and E. Merelli. Cham, Springer International Publishing**:** 117-128.

Russo, G., C. Zegar and A. Giordano (2003). "Advantages and limitations of microarray technology in human cancer." Oncogene **22**(42): 6497.

Ruuskanen, O., E. Lahti, L. C. Jennings and D. R. Murdoch (2011). "Viral pneumonia." The Lancet **377**(9773): 1264-1275.

Salser, W., J. Janin and C. Levinthal (1968). "Measurement of the unstable RNA in exponentially growing cultures of Bacillus subtilis and Escherichia coli." Journal of molecular biology **31**(2): 237-266.

Schaller, R. R. (1997). "Moore's law: past, present and future." Spectrum, IEEE **34**(6): 52-59.

Scherrer, K., H. Latham and J. E. Darnell (1963). "Demonstration of an unstable RNA and of a precursor to ribosomal RNA in HeLa cells." Proceedings of the National Academy of Sciences **49**(2): 240-248.

Schumacher, T. N. and R. D. Schreiber (2015). "Neoantigens in cancer immunotherapy." Science **348**(6230): 69-74.

Schwalbe, R., L. Steele-Moore and A. C. Goodwin (2007). Antimicrobial susceptibility testing protocols, Crc Press.

Seki, M., J. Ishida, M. Narusaka, M. Fujita, T. Nanjo, T. Umezawa, A. Kamiya, M. Nakajima, A. Enju and T. Sakurai (2002). "Monitoring the expression pattern of around 7,000 Arabidopsis genes under ABA treatments using a full-length cDNA microarray." Functional & integrative genomics **2**(6): 282-291.

Shaffer, A. L., 3rd, R. M. Young and L. M. Staudt (2012). "Pathogenesis of human B cell lymphomas." Annu Rev Immunol **30**: 565-610.

Shallcross, L. J. and D. S. Davies (2014). "Antibiotic overuse: a key driver of antimicrobial resistance." Br J Gen Pract **64**(629): 604-605.

Shannon, C. E. (1948). "A mathematical theory of communication, Part I, Part II." Bell Syst. Tech. J. **27**: 623-656.

Shannon, C. E. (1951). "Prediction and entropy of printed English." Bell Labs Technical Journal **30**(1): 50-64.

Sharma, V. K., N. Johnson, L. Cizmas, T. J. McDonald and H. Kim (2016). "A review of the influence of treatment strategies on antibiotic resistant bacteria and antibiotic resistance genes." Chemosphere **150**: 702-714.

Shaw, C. A., S. Seneff, S. D. Kette, L. Tomljenovic, J. W. Oller and R. M. Davidson (2014). "Aluminum-induced entropy in biological systems: implications for neurological disease." Journal of toxicology **2014**.

Simansky, D. A., G. Schiby, Z. Dreznik and E. T. Jacob (1986). "Rapid progressive dissemination of hemangiosarcoma of the spleen following spontaneous rupture." World journal of surgery **10**(1): 142-144.

Singh, S., P. Stafford, K. A. Schlauch, R. R. Tillett, M. Gollery, S. A. Johnston, S. F. Khaiboullina, K. L. De Meirleir, S. Rawat and T. Mijatovic "Humoral Immunity Profiling of Subjects with Myalgic Encephalomyelitis Using a Random Peptide Microarray Differentiates Cases from Controls with High Specificity and Sensitivity." Molecular neurobiology: 1-9.

Sivan, A., L. Corrales, N. Hubert, J. B. Williams, K. Aquino-Michaels, Z. M. Earley, F. W. Benyamin, Y. M. Lei, B. Jabri and M.-L. Alegre (2015). "Commensal Bifidobacterium promotes antitumor immunity and facilitates anti–PD-L1 efficacy." Science **350**(6264): 1084-1089.

Smialowski, P., D. Frishman and S. Kramer (2010). "Pitfalls of supervised feature selection." Bioinformatics **26**(3): 440-443.

Snyder, A., V. Makarov, T. Merghoub, J. Yuan, J. M. Zaretsky, A. Desrichard, L. A. Walsh, M. A. Postow, P. Wong, T. S. Ho, T. J. Hollmann, C. Bruggeman, K. Kannan, Y. Li, C. Elipenahli, C. Liu, C. T. Harbison, L. Wang, A. Ribas, J. D. Wolchok and T. A. Chan (2014). "Genetic basis for clinical response to CTLA-4 blockade in melanoma." N Engl J Med **371**(23): 2189-2199.

Spellberg, B., R. Guidos, D. Gilbert, J. Bradley, H. W. Boucher, W. M. Scheld, J. G. Bartlett, J. Edwards, Jr. and A. Infectious Diseases Society of (2008). "The epidemic of antibiotic-resistant infections: a call to action for the medical community from the Infectious Diseases Society of America." Clin Infect Dis **46**(2): 155-164.

Stafford, P., Z. Cichacz, N. W. Woodbury and S. A. Johnston (2014). "Immunosignature system for diagnosis of cancer." Proceedings of the National Academy of Sciences **111**(30): E3072-E3080.

Stafford, P., R. Halperin, J. B. Legutki, D. M. Magee, J. Galgiani and S. A. Johnston (2012). "Physical characterization of the 'Immunosignaturing Effect'." Molecular & Cellular Proteomics.

Stafford, P., R. Halperin, J. B. Legutki, D. M. Magee, J. Galgiani and S. A. Johnston (2012). "Physical characterization of the "immunosignaturing effect"." Mol Cell Proteomics **11**(4): M111 011593.

Stafford, P., D. Wrapp and S. A. Johnston (2016). "General assessment of humoral activity in healthy humans." Molecular & Cellular Proteomics **15**(5): 1610-1621.

Su, L. F., B. A. Kidd, A. Han, J. J. Kotzin and M. M. Davis (2013). "Virus-specific CD4(+) memory-phenotype T cells are abundant in unexposed adults." Immunity **38**(2): 373-383.

Sun, J., T. Xu, S. Wang, G. Li, D. Wu and Z. Cao (2010). "Does difference exist between epitope and non-epitope residues? Analysis of the physicochemical and structural properties on conformational epitopes from B-cell protein antigens." Immunome research **7**(3): 1-11.

Sweeney, T. E., H. R. Wong and P. Khatri (2016). "Robust classification of bacterial and viral infections via integrated host gene expression diagnostics." Science Translational Medicine **8**(346): 346ra391-346ra391.

Sykes, K. F., J. B. Legutki and P. Stafford (2012). "Immunosignaturing: a critical review." Trends in Biotechnology.

Sykes, K. F., J. B. Legutki and P. Stafford (2013). "Immunosignaturing: a critical review." Trends in biotechnology **31**(1): 45-51.

Tabar, L., A. Gad, L. Holmberg, U. Ljungquist, C. Fagerberg, L. Baldetorp, O. Gröntoft, B. Lundström, J. Månson and G. Eklund (1985). "Reduction in mortality from breast cancer after mass screening with mammography: randomised trial from the Breast Cancer

Screening Working Group of the Swedish National Board of Health and Welfare." The Lancet **325**(8433): 829-832.

Tanaka, K., H. Waki, Y. Ido, S. Akita, Y. Yoshida, T. Yoshida and T. Matsuo (1988). "Protein and polymer analyses up to m/z 100 000 by laser ionization time-of-flight mass spectrometry." Rapid communications in mass spectrometry **2**(8): 151-153.

Tang, C., A. Leung and K. Lam (2006). "Entropy application to improve construction finance decisions." Journal of construction engineering and management **132**(10): 1099-1113.

Teillant, A., S. Gandra, D. Barter, D. J. Morgan and R. Laxminarayan (2015). "Potential burden of antibiotic resistance on surgery and cancer chemotherapy antibiotic prophylaxis in the USA: a literature review and modelling study." The Lancet infectious diseases **15**(12): 1429-1437.

Templin, M. F., D. Stoll, M. Schrenk, P. C. Traub, C. F. Vöhringer and T. O. Joos (2002). "Protein microarray technology." Drug discovery today **7**(15): 815-822.

Thamm, D. H. (2012). Hemangiosarcoma. Withrow and MacEwen's Small Animal Clinical Oncology, W.B. Saunders**:** pp. 679-687.

Tsalik, E. L., R. Henao, M. Nichols, T. Burke, E. R. Ko, M. T. McClain, L. L. Hudson, A. Mazur, D. H. Freeman and T. Veldman (2016). "Host gene expression classifiers diagnose acute respiratory illness etiology." Science translational medicine **8**(322): 322ra311-322ra311.

Uhlmann, K., A. Brinckmann, M. R. Toliat, H. Ritter and P. Nürnberg (2002). "Evaluation of a potential epigenetic biomarker by quantitative methyl-single nucleotide polymorphism analysis." Electrophoresis **23**(24): 4072-4079.

Vail, D. M. and E. G. Macewen (2000). "Spontaneously occurring tumors of companion animals as models for human cancer." Cancer investigation **18**(8): 781-792.

van Wieringen, W. N. and A. W. van der Vaart (2011). "Statistical analysis of the cancer cell's molecular entropy using high-throughput data." Bioinformatics **27**(4): 556-563.

Vétizou, M., J. M. Pitt, R. Daillère, P. Lepage, N. Waldschmitt, C. Flament, S. Rusakiewicz, B. Routy, M. P. Roberti and C. P. Duong (2015). "Anticancer immunotherapy by CTLA-4 blockade relies on the gut microbiota." Science **350**(6264): 1079-1084.

Vilar, J. M. G. (2014). "Entropy of Leukemia on Multidimensional Morphological and Molecular Landscapes." Physical Review X **4**(2).

Vita, R., J. A. Overton, J. A. Greenbaum, J. Ponomarenko, J. D. Clark, J. R. Cantrell, D. K. Wheeler, J. L. Gabbard, D. Hix and A. Sette (2014). "The immune epitope database (IEDB) 3.0." Nucleic acids research **43**(D1): D405-D412.

Walsh, C. (2003). Antibiotics: actions, origins, resistance, American Society for Microbiology (ASM).

Wang, D. G., J.-B. Fan, C.-J. Siao, A. Berno, P. Young, R. Sapolsky, G. Ghandour, N. Perkins, E. Winchester and J. Spencer (1998). "Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome." Science **280**(5366): 1077-1082.

Wang, Z., M. Gerstein and M. Snyder (2009). "RNA-Seq: a revolutionary tool for transcriptomics." Nature reviews genetics **10**(1): 57-63.

Warter, L., R. Appanna and K. Fink (2012). "Human poly-and cross-reactive anti-viral antibodies and their impact on protection and pathology." Immunologic research **53**(1-3): 148-161.

Weinstein, J. N., E. A. Collisson, G. B. Mills, K. R. M. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, J. M. Stuart and C. G. A. R. Network (2013). "The cancer genome atlas pan-cancer analysis project." Nature genetics **45**(10): 1113-1120.

Werner, S., H. Chen, S. Tao and H. Brenner (2015). "Systematic review: serum autoantibodies in the early detection of gastric cancer." International journal of cancer **136**(10): 2243-2252.

West, J., G. Bianconi, S. Severini and A. E. Teschendorff (2012). "Differential network entropy reveals cancer system hallmarks." Sci Rep **2**: 802.

Whiteside, T. L. (2006). "Immune suppression in cancer: effects on immune cells, mechanisms and future therapeutic intervention." Semin Cancer Biol **16**(1): 3-15.

Whittemore, K. (2014). Using Antibodies to Characterize Healthy, Disease, and Age States. Doctor of Philosophy, Arizona State University.

Xu, G. J., T. Kula, Q. Xu, M. Z. Li, S. D. Vernon, T. Ndung'u, K. Ruxrungtham, J. Sanchez, C. Brander, R. T. Chung, K. C. O'Connor, B. Walker, H. B. Larman and S. J. Elledge

(2015). "Viral immunology. Comprehensive serological profiling of human populations using a synthetic human virome." Science **348**(6239): aaa0698.

Yang, S.-Y., K.-Y. Lien, K.-J. Huang, H.-Y. Lei and G.-B. Lee (2008). "Micro flow cytometry utilizing a magnetic bead-based immunoassay for rapid virus detection." Biosensors and Bioelectronics **24**(4): 855-862.

Yao, Y., W. L. Lu, B. Xu, C. B. Li, C. P. Lin, D. Waxman and J. F. Feng (2013). "The increase of the functional entropy of the human brain with age." Sci Rep **3**: 2853.

Young, M. D., M. J. Wakefield, G. K. Smyth and A. Oshlack (2010). "Gene ontology analysis for RNA-seq: accounting for selection bias." Genome biology **11**(2): R14.

Zhao, X., C. Li, J. G. Paez, K. Chin, P. A. Jänne, T.-H. Chen, L. Girard, J. Minna, D. Christiani and C. Leo (2004). "An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays." Cancer research **64**(9): 3060-3071.

Zhong, L., G. E. Hidalgo, A. J. Stromberg, N. H. Khattar, J. R. Jett and E. A. Hirschowitz (2005). "Using Protein Microarray as a Diagnostic Assay for Non–Small Cell Lung Cancer." American Journal of Respiratory and Critical Care Medicine **172**(10): 1308-1314.

Zhou, J., D. Thompson and L. Wu (2002). Detecting microorganisms using whole genomic DNA or RNA microarray, Google Patents.

Zhou, Z.-H., Y. Zhang, Y.-F. Hu, L. M. Wahl, J. O. Cisar and A. L. Notkins (2007). "The Broad Antibacterial Activity of the Natural Antibody Repertoire Is Due to Polyreactive Antibodies." Cell Host and Microbe **1**(1): 51-61.

Zhu, H., S. Hu, G. Jona, X. Zhu, N. Kreiswirth, B. M. Willey, T. Mazzulli, G. Liu, Q. Song and P. Chen (2006). "Severe acute respiratory syndrome diagnostics using a coronavirus protein microarray." Proceedings of the National Academy of Sciences of the United States of America **103**(11): 4011-4016.

# APPENDIX A
# PUBLICATIONS AND SUBMISSIONS

**CHAPTER 3**

This work is submitted for publication in Nature.

Lu Wang, Phillip Stafford, Stephen Albert Johnston

*A Common Antibody Response Is Induced by a Wide Variety of Human Pathogens*

All co-authors have granted permission for the inclusion of this work in this dissertation.

**CHAPTER 4**

This work is submitted for publication in Oncotarget.

Lu Wang, Luhui Shen, Stephen Albert Johnston

*The Immune Profile of Stages of Hemangiosarcoma Cancer in Dogs Changes Dramatically*

All co-authors have granted permission for the inclusion of this work in this dissertation.

**CHAPTER 6**

This work is published in Scientific Report.

Lu Wang, Kurt Whittemore, Stephen Albert Johnston, Phillip Stafford

*Entropy is a Simple Measure of the Antibody Profile and is an Indicator of Health Status*

All co-authors have granted permission for the inclusion of this work in this dissertation.