Using Antibodies To Characterize Healthy, Disease, And Age States

by

Kurt Whittemore

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

Approved April 2014 by the
Graduate Supervisory Committee:

Kathryn Sykes, Chair
Stephen Albert Johnston
Bertram Jacobs
Phillip Stafford
Valerie Stout

ARIZONA STATE UNIVERSITY

May 2014

ABSTRACT

The advent of new high throughput technology allows for increasingly detailed characterization of the immune system in healthy, disease, and age states. The immune system is composed of two main branches: the innate and adaptive immune system, though the border between these two states is appearing less distinct. The adaptive immune system is further split into two main categories: humoral and cellular immunity. The humoral immune response produces antibodies against specific targets, and these antibodies can be used to learn about disease and normal states. In this document, I use antibodies to characterize the immune system in two ways: 1. I determine the Antibody Status (AbStat) from the data collected from applying sera to an array of non-natural sequence peptides, and demonstrate that this AbStat measure can distinguish between disease, normal, and aged samples as well as produce a single AbStat number for each sample; 2. I search for antigens for use in a cancer vaccine, and this search results in several candidates as well as a new hypothesis. Antibodies provide us with a powerful tool for characterizing the immune system, and this natural tool combined with emerging technologies allows us to learn more about healthy and disease states.

DEDICATION

I would like to dedicate this to my family and friends and anyone interested in this content.

ACKNOWLEDGEMENTS

I would not have been able to finish my thesis work without the help of many people. I am very grateful to my advisor Dr. Kathryn Sykes for providing so much guidance for many projects over the years. I am also grateful to Dr. Stephen Johnston who created the lab environment I worked in, and also provided guidance and support for the AbStat section of this dissertation. The Center for Innovations in Medicine (CIM) was an exciting place to work since Dr. Johnston and Dr. Sykes pursued many very ambitious projects, and they also encouraged innovation and free thinking by creating projects such as the Gemini project which provided funding for selected proposals with new ideas. I took advantage of this opportunity in the Age Associated Stem Cell Autoimmunity section of my dissertation. I am also very grateful to my other committee members: Phillip Stafford for providing valuable advice about statistics, bioinformatics, and experiments, as well as Bert Jacobs and Valerie Stout for guidance and advice. I am also grateful to the Academic Rewards for College Scientists (ARCS) program and my donors Dr. and Mrs. Nick Theodore for providing me with the ARCS scholarship. I am thankful to Pattie Madjidi and Penny Gwynne who provided managerial support over the years, and Kevin Brown and Preston Hunter for technical support. I would also like to thank Lauren Dempsey, Tony Garcia, Rolf Halden, Maria Hanlin, Laura Hawes, Sudhir Kumar, Brian Smith, and Joann Williams who provided guidance and managerial support for my PhD program.

There are also many other people who have helped me along the way. As I continued to learn and work on projects in CIM there were several people who provided me with extensive training and advice: Valiery Domenyuk trained me to work with peptide arrays when I did my first rotation in CIM; Tien Olson taught me many molecular biology techniques, and helped me do some work with phage antibody libraries; Andrey Loskutov gave me much great advice about PCR and many other molecular biology techniques. Luhui Shen taught me many molecular biology techniques and we also worked together on the "Random sequence mimotopes section". During the course of the cDNA library construction and screening I enjoyed working with and receiving assistance from the students Vanessa Breguez and Maran Montgomery, as well as the high school teacher Rebecca Mestek.

I also received help from the following coworkers and acquaintances: Dr. Yung Chang provided immunology advice and was on my comprehensive exam committee; Dr. Julian Chen provided information about aging research and was on my comprehensive exam committee; Dr. Zbigniew Cichacz helped with peptide array experiments as director of the Peptide Array Core; Dr. Chris Diehnelt helped me understand some of the interactions between antibodies and peptides; Dr. Valentin Dinu helped with the automatic array aligning program; Hu (Tiger) Duan provided me with a mouse cancer time course dataset, and he was also the first person to try out my AbStat program on his own data; Dr. Rebecca Halperin provided bioinformatics help; Dr. Muskan Kukreja provided some advice for data analysis; John Lainson helped me construct and print the tumor cDNA library, as well as provided some interesting discussion on the AbStat section; Heidi Larsen helped me obtain some data about monoclonal antibodies; Dr. Hojoon Lee provided good discussion and bioinformatics help; Dr. Bart Legutki helped me work with mice, work with peptide arrays, and obtain data; Fatjon Leti helped purify peptides; Krupa Navalkar helped me with bioinformatics and obtaining array data; Dr. Lucas Restrepo provided me with the Alzheimer's dataset; Josh Richer helped me with bioinformatics and obtaining data about monoclonal antibodies; Dr. George Runger helped with data mining; Dr. Anshuman Sahu helped with data mining; Dr. John Schloendorn provided some useful discussion and information, particularly in regard to the age associated stem cell autoimmunity hypothesis; Donnie Shepard synthesized some peptides that I used to purify antibodies; Nate Sutton made some improvements to my automatic array aligning program; Lu Wang helped me analyze some of the data in the AbStat section; and Dr. Zhan-Gong Zhao synthesized some peptides for me and helped me with chemistry questions. The following individuals assisted me with molecular biology: Dr. Tricia Carrigan, Felicia Craciunescu, Dr. Debbie Hansen, Kari Kotlarczyk, Dr. Mark Robida, Dr. John-Charles Rodenberry, and Kristen Seifert. The following individuals provided some useful discussion: Diego Chowell-Puente, Lalaine Cordovez, Danielle Lussier, Dr. Clinton (Cosmo) Mielke, Minyao (Steve) Sun, Xiao Wang, Dr. Neal Woodbury, and Liang (Alan) Xiao. I am also very grateful to all of the scientists who came before me and have produced so much rich and fascinating literature. I would also like to thank all of my family and friends.

TABLE OF CONTENTS

Page

CHAPTER Page

APPENDIX Page

APPENDIX Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

Figure Page

Figure Page

Figure Page

LIST OF EQUATIONS

PREFACE

The chapters and appendix sections in this dissertation cover a wide variety of topics. However, all of these topics are connected by the themes of health and the immune system.

The first chapter about an AbStat from an antibody peptide fluorescence intensity distribution became an increasingly larger part of my dissertation work as time went on. The work from this section was originally something I did in my spare time. Around this time, I was listening to an audiobook titled “The Information: A History, a Theory, a Flood” by James Gleick which discussed Claude Shannon and entropy among many other things. This book along with many others increased my general curiosity in the concept of entropy. I, like others, had the general notion that living things are not random, and entropy provided a method for quantitating how random a system is. I thought that perhaps a living system that is more random than another might be sicker or less optimal in some way. Since our lab was quantitating the humoral immune response with a new type of peptide array, I figured that I would try out an entropy calculation to compare a group of young mice with a group of aged mice. I then started writing a Java program to handle the large amounts of data that our lab produces. I was delighted to find that there certainly was a difference between the groups. I would like to express my appreciation to Anshuman Sahu and John Lainson who helped me, through various conversations, to outline my ideas more precisely in order to reach this point. From that time on, as more and more datasets were analyzed the concept seemed to gain more validity. After several initial results, the director of the lab, Dr. Johnston, strongly supported investigating the concept in further detail. Eventually we began using other measures in addition to entropy, and the lab decided by vote to call the concept of measuring the immune response with these broad measures the AbStat for Antibody Status. This exciting development may provide new methods for assessing the state of the immune system.

Chapters 2 and 3 cover my main project with Dr. Sykes. I spent the most time on this project over the years to develop a high-throughput method for screening a tumor cDNA library for immunogens. The immunogens from this type of screening could potentially be used in a cancer vaccine, and several antigens were identified which could be investigated further.

Chapter 4 covers a project which I worked closely with Luhui Shen on to use non-natural sequence peptides to identify epitopes and isolate specific antibodies. The first appendix section covers a project in which I carried out an experiment to test a hypothesis I started to develop in my comprehensive exam document before starting my work in the CIM lab. The project was to use a cytotoxic lymphocyte assay with stem cells to test a hypothesis that part of the aging process is due to an acquired autoimmune disease against stem cells. This work was made possible by an innovative competition which Dr. Johnston and our lab created called the Gemini project. The idea was that anyone in the lab could write a proposal and the proposals that were selected would be funded by our lab. The proposal of Alexander Carpenter who wanted to test synbodies against acne and my proposal was selected, and I began to perform the experiments to test the hypothesis. The last three appendix sections (phage antibody library, purification of antibodies with non-natural sequence peptides, and automatic array alignment) are projects in which some initial work was performed.

1: ABSTAT FROM ANTIBODY FLUORESCENCE INTENSITY DISTRIBUTION ON A PEPTIDE MICROARRAY

1.1 **Introduction**

Methods for determining whether an organism is healthy or sick are very useful for diagnosis as well as for aiding in the prevention of illness. One system of the body that plays some role in virtually every illness is the immune system. The adaptive immune system will often respond in a manner that is specific and unique for the illness, and the B cells of the adaptive immune system will produce antibodies in the sera which bind to specific protein, carbohydrate, and lipid targets. A technology for acquiring information from these antibodies by applying them to a non-natural sequence peptide array has recently been developed [1]. When sera antibodies react with an array of non-natural sequence peptides, each peptide will exhibit a different level of binding of antibodies which are quantified using a fluorescent secondary antibody. The distribution of fluorescent intensities will depend on the number, affinities, and avidities of the antibodies for various peptide sequences. The presumption of the technology is that the number of antibodies, the affinities of the antibodies, and the avidities of the antibodies will depend on the health state of the organism: healthy or disease as well as young or aged. Can the information acquired from the fluorescence intensity distribution resulting from sera antibodies reacting with an array of random sequence peptides be used to distinguish healthy or disease states? Can this information be compressed into a single quantified variable? This research shows that several measures from the fluorescence intensity distribution can be combined to produce an Antibody Status (AbStat) capable of distinguishing healthy and disease states.

1.1.1 Measure humoral immune responses

Methods for interrogating the details of the humoral antibody response have evolved over the years. One of the first methods, ELISA, allows antibodies to react with antigen coated wells in a 96 well microtiter plate [2]. This method is still very popular and useful, but the amount of information that can be acquired is limited since only one antigen is coated per well. In a western

blot, all of the proteins in pathogen lysate are electrophoretically separated in a gel which is then probed with the antibody containing sera [3]. This method surveys many proteins at one time, but there are limits to the amount of separation that can be achieved between proteins with similar sizes and isoelectric points. Note that the proteins on the membrane are largely denatured. High throughput methods have also been developed in which many different proteins are placed in an array and queried with sera [4, 5]. The ELISA, western blot, and protein microarrays provide information about whether an antibody or antibody containing sera binds to a whole protein, but these methods do not provide information about the specific epitope within the protein to which an antibody binds. Some epitopes are conformational, but there are also many epitopes which are linear and would correspond to a short linear sequence within the protein. In order to obtain this epitope information, peptide microarrays consisting of tiled peptides from protein sequences have been used [6]. One disadvantage with this method is that the researcher must know which proteins are present in the pathogen being studied so that peptides from this protein can be tiled on the array. However, there are many situations in which the target antigen is unknown: a patient may be infected with an unknown but previously identified pathogen, the patient may be infected with a new foreign pathogen, or a researcher may be searching for new unidentified mutated proteins in cancer. One method for addressing these situations is to use a peptide microarray consisting of non-natural peptides with sequences randomly selected from all peptide sequence space [1, 7].

These non-natural sequence peptide microarrays have been used for a variety of purposes. The specific epitope of an antibody can be discovered. A few selected peptides from the total array can also be identified which are unique to certain types of diseases. These peptide features are part of the immunosignature, which is a medical diagnostic test obtained by using arrays of non-natural sequence peptides to associate antibodies in a blood sample with disease [1, 8-10]. The peptide features selected from the immunosignature can be used for diagnosis as well as for further characterization of the disease. These peptides are unique for the disease because they have sequences which are mimotopes of epitopes which antibodies for that disease bind to. Note that within one feature on the microarray there are many peptides, but they all have the same sequence. The sequence of the peptide used for each different feature is chosen from

random sequence space as opposed to natural protein sequence space, but the peptide sequence and location of each feature on the array is known. Therefore, if a given feature has a very high intensity whenever breast cancer sera is applied to the microarray, the user will know the sequence of the peptides located at this feature and can study how this may relate to breast cancer for example.

Is there more information that can be obtained from the non-natural sequence peptide microarray in addition to specific peptides that are high or low for a given sera type? When antibodies are applied to the peptide microarray there is a whole feature intensity distribution that results since each feature will have some intensity value which will reflect how much binding with that peptide occurred. This peptide distribution may exhibit behavior characteristic of disease or normal states irrespective of the few specific peptides that are mimotopes of an epitope relevant to the disease. In total, there are approximately 10^{11} B cells in an individual's lymphocyte population [11], and therefore there are many different antibodies in sera. The nature of all of the different affinities and avidities present in the antibody repertoire in a healthy state may be different than the affinities and avidities present in a disease state, and these differences may be detectable in the fluorescence intensity distribution resulting from interaction of these antibodies with a non-natural sequence peptide microarray.

There may be detectable differences in the affinities and avidities of the antibody repertoire in young and aged sera as well as disease. These differences may manifest themselves in the fluorescence intensity distribution that results from applying these antibodies to an array of non-natural sequence peptides. As organisms age, the antibodies they produce against new targets have lower affinities for those targets [12-14]. This would explain why the immune system responds less effectively to vaccines with age [15]. Not only is there a decrease of antibody specificity, but the prevalence of autoimmune disorders tends to increase with age as well [16]. This indicates that the aged immune system is more broadly and non-specifically reactive, and there is evidence of general increased inflammation known as "inflamm-aging" [17, 18]. An abundance of low affinity non-specific antibodies may bind to an array of non-natural sequence peptides in a different manner than antibodies which have very high affinities for their target sequences.

1.1.2 Measures that compose the AbStat

The antibody repertoire will bind to each peptide on the non-natural sequence peptide array with a different level of binding. This level of binding is detected from a scanner which detects the intensity of certain wavelengths of light emitted by a dye conjugated to a secondary antibody that binds to all antibodies of a certain isotype. The final output from this scan is a distribution of intensity numbers corresponding to each peptide in the array. There are a variety of techniques for characterizing a distribution of numbers and compressing this information into a single quantitative value. This quantitative value can then be used for statistics and direct comparisons to other samples that have been applied to the array. The following quantitative measures were used in this research: entropy, minimum, maximum, coefficient of variation, standard deviation, mean, median, 5th percentile, 95th percentile, kurtosis, skew, and dynamic range [19-22]. These measures combined define the AbStat which can be used to distinguish healthy and disease states.

All of these measures characterize the fluorescence intensity distribution in slightly different ways. Data from the peptide microarrays is often median normalized by dividing the fluorescence intensity of each feature by the median fluorescence intensity of all of the feature intensities. This median normalization allows data from different experiments and times to be compared. Median normalizing the data changes the values of some of the measures but not others. The following AbStat measures remain unchanged when analyzing raw data or data that has been median normalized: entropy, coefficient of variation, kurtosis, skew, and dynamic range. All of the other measures will have different values depending on whether the data has been median normalized or not. The minimum and maximum values are simply the lowest and highest values of the fluorescence intensity distribution. The mean aids in defining the “middle” of the distribution, and the median does the same but is less affected by outliers. The standard deviation indicates the level of dispersion from the mean. The coefficient of variation is a normalized measure of dispersion and is defined as the ratio of the standard deviation to the mean. The 95th percentile defines the intensity value which is greater than 95% of all of the other intensity values, the 5th percentile defines the intensity value which is greater than 5% of all of the other intensity values,

and the dynamic range is defined as the ratio of the 95th percentile to the 5th percentile. A high dynamic range indicates that there is a wide range of intensities in the distribution. The kurtosis is the measure of "peakedness" of a distribution, and higher values indicate tighter peaks. The skewness measures the extent to which a distribution "leans" to one side of the mean. The skewness can be positive or negative with zero skew indicating that the distribution is symmetric about the mean without leaning to one side. A large positive value would indicate that there is a long tail on the right side of the distribution with most of the mass on the left. Larger absolute values of skew would indicate a greater degree of lean.

1.1.3 Nature of entropy measure

One measure which distinguishes between healthy and disease samples the best among these measures is entropy, and evidence for this fact is provided in Section 1.3.8 Rank of Measures in this dissertation chapter (crosslink here: 1.3.8 Rank of Measures). Entropy takes on a value of zero when all of the numbers in a distribution have the same value, and the value increases as the distribution becomes more heterogeneous. The maximum value for entropy occurs when each value is represented one time. This is the type of distribution that is most similar to the distribution produced by a random number generator since this algorithm will produce each value with an equal likelihood. The maximum possible value of entropy depends on the number of elements in the distribution. If there are more elements in the distribution, then the maximum possible value of entropy is higher. Entropy can also be conveniently normalized to a value in-between zero and one by dividing the value of entropy by the logarithm (with the same base used in the entropy calculation) of the number of possible values in the distribution. Note that median normalizing a dataset does not affect the value of entropy. In order to calculate the entropy, one must count the number of times that each unique value occurs, and the number of unique values does not change if all of the numbers are multiplied or divided by the same number.

How is the entropy of the fluorescence intensity distribution calculated, and how is this entropy related to statistical thermodynamic entropy and information entropy? The mathematical

concept of entropy was first developed during the 1800s as scientists found that some energy was always lost and not transformed into useful work in combustion engines. The first mathematical formulation of entropy was put forth by Rudolf Clausius in English in 1856, and this representation consisted of bulk quantities that do not consider the state of individual molecules. The formulation took on the form of the ratio of the transfer of heat (Q) to temperature (T) [23]. Ludwig Boltzmann later viewed entropy from a statistical thermodynamic perspective and defined entropy in terms of the distribution of microstates of the system in 1877 [24], and this was later represented in the Gibbs entropy equation (Equation 1). In 1948, Claude Shannon defined the statistical nature of "lost information" in phone-line signals at Bell Telephone Laboratories [22] (Equation 2). Shannon was originally unaware that his equation for “uncertainty” was virtually identical to the thermodynamic entropy developed by Boltzmann and Gibbs [25]. As a brief aside, note that the base of the log only affects the units of the result. In information theory, the base of the log used is often base 2 which results in bits. Using base 10 results in units of dits, and using a base of *e* (the base for the natural logarithm) results in units of nats. Specifically, Shannon information entropy is a measure of the uncertainty of a system represented by discrete outcomes or “information”, and this uncertainty is defined by the number of bits necessary to define the information. An alternative to this definition is that entropy is the average minimum number of yes-no questions necessary to identify an item randomly drawn from a known and discrete probability distribution [26]. High entropy indicates there is a high degree of uncertainty and more information is necessary to define the system.

$$S = -k_b \sum p(x) \ln\left(p(x)\right)$$

Equation 1 Gibbs entropy

Gibbs entropy where kb is the Boltzmann constant (usually expressed in units of Joules/Kelvin), and p(x) is the probability of a microstate.

$$H = -\sum p(x)\log\left(p(x)\right)$$

Equation 2 Shannon information entropy

Shannon information entropy where p(x) is the probability of outcome x

Although both the statistical thermodynamic entropy and the Shannon information entropy equation operate on different input (microstates or energy vs information), they both measure how homogenous or heterogeneous a distribution is. For example, in the statistical thermodynamic form of entropy, a distribution in which many molecules all have the same energy level would have a lower entropy than a distribution in which the energy levels of all of the molecules are more evenly distributed. A bucket of liquid water has a lower entropy than the same water molecules in the form of steam since the molecules in the liquid form will occupy fewer energy levels. Also, note that steam at a higher temperature would occupy more possible energy levels than steam at a lower temperature. In this example, the energy level histogram with a more homogenous distribution with a peaked shape (liquid water) would have a lower entropy than the energy level histogram with a more heterogeneous distribution with a broader peak (gas) since the gas will have molecules at very low energy levels occupied in the water state as well as very high energy levels only occupied in the gas state.

Note that in the high entropy configuration of water molecules there are many more possible ways to set the energy level for each molecule to achieve the same frequency of energy levels, i.e. more possible microstates, whereas a low entropy configuration allows for fewer variations in which molecule occupies which energy level. Many of the molecules need to occupy the same energy level. When transitioning from a liquid to a gas there is an increase in entropy because there are more possible microstates or configurations in the gas than in the liquid. A low entropy energy distribution in a gas is very unlikely because it is very unlikely that many of the gas molecules will suddenly occupy the same energy level. Instead the energy levels will be more evenly distributed, i.e. more randomly distributed. More molecules will occupy the same

energy level in a liquid. Therefore, high entropy configurations are more likely to occur by chance since there are more possible ways to randomly arrive at such configurations.

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 be lower than the entropy of the message “abcd”. The first message would have a more homogenous frequency distribution since all of the values are “a”, but the second message would have a more heterogeneous distribution with a higher entropy since “a”, “b”, “c”, and “d” occur with an equal frequency/probability. Distributions with higher entropy in which the letters occur with equal probability have more possible states than lower entropy distributions, and they are therefore more likely to occur by chance. This is not just true in this particular example, but with any example of information. An object of information with high entropy in which all of the elements of information occur with equal probability have more possible states than low entropy versions of the same type of information. Therefore, high entropy objects are easier to generate with a random information generator than low entropy objects because there are more possible ways to end up in a high entropy state. In order to clarify this point, let’s revisit the “a”, ”b”, ”c”, ”d” example. Note that there are only four ways to write a four character message in which all four positions have the same letter (“aaaa”, “bbbb”, “cccc”, “dddd”), but there are 24 ways (4!=24) to write a message with equal frequency for all four letters (“abcd”, “abdc”, “acbd”, etc.). Therefore, a random letter generator has 24 ways that it could randomly generate one of the highest entropy four character messages, but only four ways of generating the lowest entropy four character message (“aaaa”, “bbbb”, “cccc”, “dddd”).

In summary, entropy is essentially the disorder of a distribution. The most disordered state is the state that is most likely to occur by chance, and this occurs when all of the different elements of the distribution have an equal probability of having any value. The lowest entropy state occurs when all of the different elements of the distribution have exactly the same value, which is less likely to occur by chance. In other words, a high entropy number distribution is most easily reproduced by a random number generator, whereas it is very unlikely for a random number generator to produce a low entropy number distribution in which most of the numbers

have the same value. Therefore, entropy can measure how random a system is. Extending this concept further, many people have a general notion that living systems are ordered and non-random [27]. Therefore, perhaps as a living system becomes more or less ordered from the norm, this deviation will indicate a progression toward a disease state, and the living system will become less lifelike. However, more rigorous scientific data is needed to support this notion.

Since entropy can provide a quantitative value for the homogeneity of a frequency distribution, this concept could also be applied to other types of distributions that do not derive directly from thermodynamic microstates or energy or information. Actually, any system can ultimately be represented as information. In order to calculate the entropy of the fluorescence intensity distribution resulting from the application of antibodies to a non-natural sequence peptide array, one can simply determine the frequency that each intensity value occurs. With the data analyzed in this work, the intensity is acquired from a 16 bit TIFF image, and therefore the intensity can have a value ranging from 0-(2^16-1) or 0-65,535. The entropy can be calculated from the histogram of the fluorescence intensities. A peptide distribution in which there are three peptides with intensity values of 5, 5, and 768 respectively would have an entropy value of $-(2/3*\ln(2/3) + 1/3*\ln(1/3))=0.637$. A homogenous distribution in which all of the intensities are very near 0 would have a very low entropy since all of the peptides have nearly the same value. A heterogeneous distribution in which all of the different peptides have different intensities resulting in a large dynamic range would have a higher entropy.

Characteristics of the antibody repertoire such as the number of antibodies, the affinity of the antibodies for their target sequences, and the avidity of the antibodies will affect the shape of the resulting fluorescence intensity distribution which will be reflected in the quantitative value of the entropy of the distribution. As one quick example, if a monoclonal antibody were applied to a non-natural sequence peptide array, and this monoclonal antibody has a very low affinity for its target sequence but binds randomly with a variety of intensities to other peptides, then this monoclonal antibody would have a higher entropy than an antibody which binds tightly to all peptides with its target sequence with similar intensity. Antibodies with low specificity that bind randomly will result in a higher entropy than antibodies with very high specificity and non-random

binding. Additionally, an antibody mixture with many different specificities with different affinities will result in a higher entropy than a mixture in which more of the antibodies are against a single target with similar affinities. Therefore, entropy is one of the key measures in the list of measures used to define the AbStat for an organism.

1.1.4 Entropy with previous biological data

Using the concept of entropy to provide quantitative correlations with health and disease states has been explored previously. The concept has been used with data associated with cancer, the heart, the brain, the immune system, and infectious diseases. The entropy of these various biological data has also been investigated with regard to age, and changes during the aging process have been observed. These previous research projects provide a foundation and framework of concepts which the AbStat can integrate into. As far as I know, the antibody profile has not been used for this type of assessment. Ultimately, the entropy measurement can be used to assess the health of many different biological systems, and the immune system is an ideal candidate system for health assessment.

There are several papers relating the entropy of biological data with cancer. Many datasets have been analyzed, and the result demonstrates that genomic entropy calculated from aberrations in DNA copy number is higher in a variety of cancer types [28]. Additionally, the increase in genomic entropy is correlated with an increase in gene expression entropy. A separate mathematical study claimed that the increased mortality rates and cancer rates with age observed in models such as the Gompertz equation, Weibull function, and Strehler-Mildvan modification of the Gompertz equation can be associated with increased informational entropy of the genome with time [29]. The author also suggests that increased genomic informational entropy can even be caused by mutations induced by thermal noise over time in the absence of chemical mutagens and radiation.

In addition to DNA focused studies, the entropy of RNA has also been investigated. The entropy of alternative splicing in cancer cells was explored, and the results demonstrate that the alternative splicing entropy is significantly higher in 13 of 27 cancers investigated when compared

to normal tissues of the same anatomical site [30]. The distribution of splicing isoforms present in cancer cells was much broader and flatter, resulting in a higher entropy. Interestingly, the genes that presented the highest entropy in their splicing isoform distribution are splicing factor genes themselves. The study also found that there was a positive linear correlation (correlation coefficient of 0.81) of proliferation level and entropy value for tumors.

If we zoom out from the DNA and RNA level, we find that the entropy of the cell at higher levels has also been determined. In one study, the entropy of structural and numerical chromosomal aberrations were compared among 14 solid tumor types in 1,232 karyotypes [31]. Some cancer types such as lung cancer typically had higher entropy values than other tumor types, and high entropy values were associated with a shorter mean survival time for the patients. In another study, researchers discovered that the entropy of a random walk on the protein interaction network graph was higher in cancer cells than normal cells in all of the 6 different cancer tissue types investigated [32].

Cancer data at the tissue level has also been examined. Histological analysis of photographs of tumor tissue at various magnifications also benefits from the use of entropy as a measure to distinguish normal from tumor tissue [33]. In the histological study, three measures were suggested as particularly useful for diagnosing a sample as a tumor sample using tissues from 34 prostate tumor samples, 34 benign hyperplastic samples, and 34 normal prostate samples: fractal dimension, cell nuclei number, and entropy. All three of these measures were determined from an image file of tumor tissue. Researchers have also calculated the entropy of data from thermal images of women with or without breast cancer. They were able to use these results and other attributes of the thermal image along with classification algorithms to classify samples as cancer better than would be expected by chance [34].

Researchers have correlated health conditions with the entropy of biological data related to the heart. In one study, researchers discovered that the “band limited transfer entropy” *decreases* with age as heart rate complexity also decreases [35]. In the study, the heart rate, respiration rate, and blood pressure were monitored for 20 young subjects aged 21-34 years old and 20 older subjects aged 68-85 years old as they watched the movie Fantasia from Disney.

The researchers then used sophisticated non-linear dynamics techniques to calculate the joint entropy and conditional entropy to determine the contribution of respiration and blood pressure data to correlate with heart rate complexity. They found that the lag between respiration and heart rate was longer in older subjects and the entropy calculated from the data was decreased in older subjects. This age associated effect was more pronounced in the males than the females. Another group found that the spectral entropy of electrocardiogram (ECG) data recorded during sleep was negatively correlated with age [36]. In another study, the researchers calculated the entropy of magnetocardiography data, input these entropy values into a multilayer perceptron neural network as training data, and then classified whether heartbeats were from patients with coronary artery disease or normal individuals with 98% accuracy [37]. The entropy of one of the regions of interest in the magnetocardiography data was considerably lower in the patients diagnosed with coronary artery disease.

Several researchers have used entropy associated with biological data for the brain. One group correlated an increase in entropy in fMRI data with age in a very large dataset with 1,248 samples [38]. The entropy measured the dispersion of the functional connectivities that exist within the brain. In addition to a correlation with age, they found that males exhibit a higher increase of entropy with age then females. They also discovered that schizophrenic patients had a lower functional entropy than normal people. This result illustrates that normal individuals fall within an entropy range, and a value too high or low from this range can indicate a deviation from optimal health, and this also turns out to be the case with AbStat as well. Another MRI study investigated the entropy of the cortical structure complexity determined from images constructed from MRI brain data [39]. They determined that this entropy increases with age, and the value of the entropy was also higher for Alzheimer's patients compared to age-matched controls.

In other studies, researchers have measured the ability of the brain to process information. For example, researchers measured the ability of the brain to process linguistic information at different ages [40]. The researchers asked old and young adults to identify and repeat words which had high or low response entropy. For example, a word with a very high response entropy would occur in a neutral context such as in the sentence: "The word is __.". The final word in this

sentence could be many different words resulting in a very large degree of uncertainty. On the other hand, a word with a very low response entropy would occur in a sentence such as “He wondered if the storm had done much __”. Based on the responses from 100 previous volunteers, the final word in this sentence has a very high probability of being “damage” so there is less uncertainty (and lower response entropy) in this “high context” situation. Older individuals failed to identify the last word of the sentence more often than younger individuals even in low entropy response situations, and the researchers make the argument that this is not due to the loss of hearing acuity that occurs with age. In short, the concept of entropy can also be correlated with the ability of the aged brain to process linguistic information. In a more visually oriented study, older individual’s display a higher variability in an attempt to remember the location of an object, and the researchers claim this is due to a higher entropy in neuronal processing [41]. They create some new terms for their molar entropy model for spatial memory and suggest that the slower processing speed and less accurate outcomes of older individuals is caused by increased neuronal noise or “computational temperature”.

In one study, the researchers did more than just monitor the entropy of the brain, they intervened and changed the entropy. In the study, they found that rhesus monkeys with induced Parkinson’s disease exhibited an increased level of neuronal firing entropy in the subthalamic nucleus area of the brain compared to the entropy in the control group of rhesus monkeys [42]. They then implanted an electrode into the brain and induced high frequency stimulation or low frequency stimulation. High frequency stimulation resulted in lower entropy and a reduction of Parkinson’s disease symptoms, whereas low frequency stimulation increased entropy and exacerbated symptoms.

Researchers have also investigated the entropy of data associated with the immune system, but no one has investigated the entropy resulting from antibodies reacting with an array of thousands of peptides as outlined in this dissertation since this technology is relatively new. Several of these immune research projects are related to the entropy of the sequences of viral, antibody, or MHC elements. One group explored the sequence positional entropy of immunoglobulin C-class and V-class sequences to show that domains can be identified without

any structural information [43]. Entropy has also been used to determine whether certain immunoglobulin domains in aquatic Antarctic species were highly variable or conserved [44]. Other investigations of CDR entropy without the use of structural information have also been performed [45]. Inside a cell, the chromatin textural entropy of erythroid precursor cells in the mouse spleen also increases with age [46]. Note that entropy calculations have not been restricted to immunoglobulins as the entropy of MHC class II promoters has also been investigated [47].

Several studies explore the entropy of infectious diseases. One study found that the rate of HIV virus escape depended on the sequence entropy of the epitope targeted by the immune system of individual patients [48]. In another study, they found that the most effective T cell responses against HIV occurred when the entropy of the overlapping peptides targeted was low [49]. A different study found that the entropy of viral types present in the blood decreased when immunosuppression drugs were administered after an organ transplant [50]. Therefore, one viral type became predominant. Entropy has also been used to measure the diversity of sequences in the influenza virus associated with antibody binding, and the investigators used this entropy metric to determine which regions of the virus were evolving most rapidly [51]. The entropy of epidemiological data has also been explored as researchers used a maximum entropy model to reliably predict the distribution of malaria in Africa [52]. Using this model, they found that the most important factor associated with malaria risk was human population density.

One group developed a method for determining the conformational entropy of protein-protein interactions [53]. They then found that the conformational entropy of the immunoglobulin CDR3 region is significantly correlated with kinetic and affinity constants. They then go one step further and propose an algorithm to replace amino acids with prolines to restrict conformational flexibility and increase antibody affinity.

Immune disorder from the viewpoint of more traditional thermodynamic entropy associated with antibody binding rather than information entropy has also been explored. For example, in one study researchers used isothermal titration calorimetry to identify the thermodynamic entropy change that occurs when the broadly neutralizing 2F5 antibody binds to the HIV gp41 protein

epitope [54]. Isothermal titration calorimetry has also been used to determine the change in entropy when antibodies bind to high-molecular-weight capsular polysaccharides [55].

The entropy of immune structures has also been investigated. The germinal center texture entropy from photographs was negatively correlated with the number of plaque-forming cells [56]. Another study found an increase in information entropy and structural disorganization during the course of proinflammatory diseases [57]. Therefore, just as with cancer, the entropy of the immune system has been calculated from many different types of data at many different levels of biological organization.

In addition to some of the studies mentioned previously with the heart, brain, and immune system which associated entropy measures with age, there are also other measures. One biological characteristic that increases with age is the entropy of DNA methylation. Researchers demonstrated that the methylation changes of the DNMT1 gene, which itself codes for a DNA methyltransferase, exhibited aging-driven entropy characteristics in a White Leghorn chicken model [58]. A larger study in humans with two different cohorts with an n of 482 and 174 with an age range from 19 to 101 has also been performed, and the researchers also found a correlation between the entropy of methylation and age [59]. Another author has discussed the general concept of an increase in entropy among many biological systems with age [27].

There are a few trends in these research articles which address the entropy of biological data. In all of the studies, the value of entropy increases with age, with the exception of the heart studies. The complexity of heart rate data actually decreases with age. Additionally, higher than normal entropy values are correlated with disease states such as cancer. Normal healthy states are demonstrated or implied to fall within a certain entropy range, and values that are too far above or below this range indicate a problem with the system. As an example of an entropy value that was too low, schizophrenic patients exhibited lower than normal entropy values associated with fMRI data as mentioned previously [38]. All of this collective information supports the ideas that entropy is an important measure correlated with health states, healthy systems exhibit entropy within a certain range, and that entropy tends to shift away from this normal range with age.

One could argue against the significance of these ideas. All of these studies start with two diagnosed conditions, and then look for entropy differences. We already know that there are two distinct groups so it is almost circular to show that the entropy between the two groups is different – not necessarily in a predictable direction. Therefore, what good is this entropy measurement for this biological data? The first point to counter this argument is that there is no guarantee that the entropy measure would have been different between the two groups. Many other metrics would fail to exhibit a statistical difference. For example, in my own AbStat analysis, many of the metrics were not significantly different between the groups. One of the simplest examples of this is with the maximum value metric. The maximum value was often exactly the same between normal and disease groups. Other metrics were often not exactly the same, but failed to be statistically different. Therefore, the fact that entropy is significantly different between the two groups is interesting in itself. The second point that makes the entropy measure noteworthy is that this measure may reveal some interesting attributes about the mechanism of the diseases in question. Entropy is a measure of how chaotic a system is, or to think of the metric another way, how likely it would be for a random information generator to produce the system. Therefore, since cancer systems tend to have a higher entropy than normal systems, as suggested via numerous types of biological data, then this may imply that a progression towards cancer is also a progression of the cell towards more chaotic behavior. This observation provides a perspective for viewing cancer which could lead to new discoveries and treatments.

The entropy measure has been applied to many different types of biological data in numerous studies, but the AbStat measure is unique for a few reasons. The first reason the AbStat measure is unique is that it combines entropy with many other global measures to boost the power of classification and diagnosis even further. This type of combination with this many global attributes was not performed in any of the articles previously reviewed. The second reason that the AbStat is unique is that it determines the entropy calculated from the interactions of antibodies with a large number of peptides rather than from the original source of the health problem. The nature of the response of the immune system can often determine the outcome of the patient. Since the immune system plays a very large role in the health of the patient, and

since the immune system responds to many different problems in the body, using the immune system to provide information about the body is a logical strategy.

1.1.5 AbStat applications

The AbStat measures were used to analyze many different types of samples and datasets. The behavior of these measures during controlled antibody experiments was investigated. The behavior of these measures before and after mouse vaccines and infections, human vaccines, lymphomas which decrease antibody repertoire complexity, mouse cancer progression, human disease, and changes with age was analyzed. Throughout the analysis of this data, trends can be identified which can be used to predict the outcome of other datasets, and the interpretation of the results from one experiment to another is fairly compatible. In general, antibody solutions which are more similar to a monoclonal antibody will have a fluorescence intensity distribution with tight "ordered" peaks, whereas antibody solutions which have many different antibodies with many different affinities, particularly low affinities for their targets, result in broad "chaotic" fluorescence intensity distributions. The broadness and degree of chaotic behavior of the fluorescence intensity distribution is quantitated quite well with the entropy measure. A thorough analysis of the ranking of all of the measures for their ability to distinguish healthy from disease and aged states reveals that entropy is indeed one of the best measures to capture differences in these states. However, all of the metrics measure slightly different characteristics of the fluorescence intensity distribution, and there is power in using them all together as input for standard classification algorithms such as SVM and classification trees.

The AbStat could prove useful for a variety of different situations, including situations which cannot be imagined at the present time. For example, individuals could obtain their AbStat on a regular basis. They would be able to observe changes in this measure when they have a vaccine or develop an infection such as the common cold. With a large population undergoing constant monitoring, trends in lifestyle patterns could be correlated with certain AbStat values and health states. This information could be used to improve the health of the population by motivating individuals to make changes that can be observed in the AbStat feedback. The AbStat could also

be used to help diagnose whether someone has a serious illness such as cancer. If the AbStat measures of an individual are far from normal, then this could prompt further investigation into the exact nature of the problem. The behavior of the fluorescence intensity distribution when antibodies bind to a non-natural sequence peptide array can provide information about health, and the patterns identified from constant data collection could ultimately be used to improve health.

1.2 **Materials and Methods**

1.2.1 Array platforms

Two different non-natural sequence peptide array platforms were used: one platform consisted of an array of 10k peptides (CIM10K) and the other platform consisted of an array of 330k peptides (HT330K). There were actually two different CIM10K platforms. The 10k platforms consist of 10k 20 amino acid peptides with sequences of 17 amino acids that were randomly generated, and these peptides covalently bind to a glass slide [60]. The three amino acids of the peptide sequence were constant as glycine, serine, cysteine [61]. These three amino acids formed the linker for attachment to the aminosilane coated glass surface. This linker is on the carboxyl terminus for CIM10Kv1 and on the amino terminus for CIM10Kv2. The CIM10Kv1 arrays were produced by spotting peptides synthesized by Alta Biosciences using a NanoPrint LM60 microarray printer (Arrayit, Sunnyvale, CA). For CIM10Kv2, the 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 was based on the fabrication of 330k peptide microarrays on a silicon wafer. This platform also makes use of peptides selected from random space to maximally distribute the peptides in that space. On this platform, not all of the peptides have exactly the same length, but the average length is 12 amino acids. The manufacturing process used borrows many techniques from the electronics industry during the production of computer processor chips. Once the peptides were synthesized onto the silicon wafer, the arrays were deprotected and soaked overnight in DMF. The arrays were then stepwise transitioned to an aqueous solution.

The residual DMF was removed by two 5 min washes in distilled water, the arrays were soaked in PBS for 30 min, blocked with an incubation buffer (3% BSA in Phosphate Buffered Saline, 0.05% Tween 20 (PBST)), washed, and then spun dry. At this point the, the arrays were ready for the application of sera.

1.2.2 Array procedures with samples

The data I analyzed throughout this chapter was obtained by applying antibodies to peptide microarrays. These experiments were performed by other researchers in the Center for Innovations in Medicine. The key researchers who performed the different experiments will be briefly listed. The "1.3.1.1 Monoclonal affinities and AbStat Measures" experiment was performed by Rebecca Halperin; the "1.3.1.2 Spiking antibody into sera" experiment with antibodies against the GFOD1 protein was performed by Josh Richer; the "1.3.1.2 Spiking antibody into sera" experiment with antibodies against the gp120 HIV protein was performed by Heidi Larsen and Bart Legutki; the "1.3.2 Mouse vaccines and infections" ("1.3.2.1 246 day time course", "1.3.2.2 Multiple mouse immunizations", and "1.3.2.3 6 day time course"), the "1.3.4 Reduction in antibody repertoire complexity with lymphoma", and the "1.3.7.1 Changes with age in mice" experiments were performed by Bart Legutki; the "1.3.5 Mouse cancer progression" experiment was performed by Hu (Tiger) Duan; the "1.3.6.4 Alzheimer's disease" experiment was performed by Lucas Restrepo; the "1.3.3 Human vaccines", "1.3.6.1 HT330K first chip disease dataset", "1.3.6.2 HT330K wafer 46", "1.3.6.3 CIM10K", and "1.3.7.2 Changes with age in humans" experiments were performed by the Peptide Array Core under the direction of Zbigniew Cichacz. The description of the procedure for the experiments with sera samples and peptide microarrays performed by these researchers follows.

The general assay conditions have been published previously [10, 62-64], but they will briefly be described here as well. 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 ^{0}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). I then used these intensity values as input to obtain the AbStat.

1.2.3 Mathematical measures

The AbStat is composed of several measures which are calculated from all of the intensity values from the fluorescence intensity distribution obtained from reacting antibodies with an array of non-natural sequence peptides. The final output for each measure of the fluorescence intensity distribution is a single number. The AbStat consists of the following measures: entropy, minimum, maximum, coefficient of variation, standard deviation, mean, median, 5th percentile, 95th percentile, kurtosis, skew, and dynamic range. The following

measures do not change when working with median normalized data or raw data: entropy, coefficient of variation, kurtosis, skew, and dynamic range. For the numbers that do change, the value from raw data and the value from normalized data were both included in the list of AbStat measures. The values are calculated by a custom Java program, but the resulting values match the values that would be obtained by the corresponding function in Excel, with the exception of entropy since entropy is not a default function in excel. The minimum, maximum, mean, standard deviation, and median are all calculated using very common formulas. The coefficient of variation is the ratio of the standard deviation to mean. The dynamic range is defined as the ratio of the 95th percentile to the 5th percentile. The equations for Shannon information entropy (Equation 2), kurtosis (Equation 3) [65], and skew (Equation 4) [65] are presented. The procedure and equations for calculating the percentile is presented in Equation 5 [65].

$$\left(\frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum\left(\frac{x_j-\bar{x}}{s}\right)^4\right)-\frac{3(n-1)^2}{(n-2)(n-3)}$$

Equation 3 Kurtosis

Kurtosis where n is the number of elements, x_j is element j, $\bar{x}$ is the mean, and s is the standard deviation.

$$\frac{n}{(n-1)(n-2)}\sum\left(\frac{x_j-\bar{x}}{s}\right)^3$$

Equation 4 Skew

Skew where n is the number of elements, x_j is element j, $\bar{x}$ is the mean, and s is the standard deviation.

$$
\begin{aligned}
&N = (n+1)P \\
&If(n = 1) \\
&\{ \\
&\qquad x_0 \\
&\} \\
&If(n = N) \\
&\{ \\
&\qquad x_{\lfloor (n+1)P \rfloor - 1} \\
&\} \\
&Else \\
&\{ \\
&\qquad x_{\lfloor (n+1)P \rfloor - 1} + (n - \lfloor (n+1)P \rfloor)(x_{\lfloor (n+1)P \rfloor} - x_{\lfloor (n+1)P \rfloor - 1}) \\
&\}
\end{aligned}
$$

Equation 5 Procedure for calculating the value at a percentile

Procedure for calculating percentile where n is the number of elements, P is the percentile, and $\lfloor \rfloor$ *is the mathematical floor function to obtain the integer part of a number.*

1.2.4 Java AbStat Program

A custom Java program was written to determine the numerical values of all of the AbStat measures. Calculating the results can be performed in Excel, but this is not practical for a number of reasons. The data from one 330k array file is contained within a GenePix result (gpr) file that is about 40 Mb, and processing this data in excel takes a long time. In fact, using Excel with this data causes many standard desktop computers to freeze for a time before becoming operational. Even if performing the calculations in Excel could be very quick, manually doing this for hundreds of different gpr files in each dataset would be impractical. The custom Java program written is much quicker than excel, and it also automatically process hundreds of gpr files in a matter of minutes. Note that the accuracy of the Java program was validated by manually obtaining all of the AbStat results of a few gpr files in Excel.

The program is run by making a call to a Java jar file from the command line, and the program has six main capabilities: 1. Calculate measures from one gpr file; 2. Calculate measures from a folder of gpr files; 3. Calculate measures from raw data from many samples in one tab delimited text file; 4 °Calculate measures from normalized data from many samples in one tab delimited text file; 5. Collect many different output files into one table; and 6. Place all of

the gpr files in a folder into a new separate folder. The 5th and 6th capabilities can come in handy when there are an extremely large number of files and the user of the program wants to have many different instances of the program running, perhaps even on several different computers. When the operator is using a gpr file as the input, the name of the column with the intensity data is specified, and the program can automatically find this column even if there is a header in the file which takes up several rows before the table of data begins. When the operator is using a tab delimited text file as input, the tab delimited text file can be arranged in any possible manner, and the row and column at which the data starts along with the column at which the data ends is specified. There is no need to format the table of data in a specific way just for the program.

The output of the program is a tab delimited text file with columns for the following values: name of the gpr file, entropy, normalized entropy, max, min, cv, stdev, mean, median, 5th percentile, 95th percentile, max from normalized data, min from normalized data, mean from normalized data, 5th percentile from normalized data, 95th percentile from normalized data, kurtosis, skew, and dynamic range. Each row is then filled for each gpr file which came from one application of sera to a non-natural sequence peptide array. The program takes just a few seconds to complete per 330k gpr file on a standard pc, which is not possible with Microsoft Excel. Using this program one can very quickly determine the AbStat of a sample.

This AbStat program is stored at the following location as of this writing: "\\biofs.biodesign.asu.edu\CIM\Administration\Biostatistics\Immunosignature Entropy\Code".

1.2.4.1 Optimization of Java AbStat Program

The program underwent many revisions before it could quickly calculate all of the measures from a large 330k 40 MB data file. Originally, the program required about 7 min on a standard pc to calculate the entropy value from a 10k peptide file, but the program also was much more customizable and output much more information at the end. For example, one could specify the size of the bin to use when determining the frequency for the entropy calculation, and the program would output all kinds of details such as which peptides were in a certain bin. Later, however, our lab started using 330k gpr files from the new peptide arrays, and the program

required about two days on a standard pc to compute the results. In order to speed up the program dramatically, many changes were made such as removing the customization, removing the detailed output, and replacing many of the modular Java classes with a single class in which primarily primitive types were used such as int and double arrays rather than storing values in the more flexible ArrayLists which would often need to be converted to Strings and then double values. In this lean and fast version of the program, many options are not customizable. Options which seemed useful initially were not necessary for accomplishing the ultimate goal of acquiring information from the peptide arrays which can distinguish between healthy and disease states. The end result of these various optimizations was a very lean and fast program, even on large datasets.

1.2.5 Methods of analysis

The custom AbStat Java program was used to obtain the measures for each sample. This Java program was written using the Eclipse IDE for Java developers (Kepler Service Release 1) [66]. Then several methods and programs were used to analyze the AbStat measures in the different datasets. Grouped box and dot plots of the measures for many samples were often used to get an overview of overall differences between groups. These box and dot plots were created using Deducer (version 1.7-9) [67], which is a graphical user interface for the R statistics software (version 2.15.0) [68]. Student's t-test was performed in Microsoft Excel 2013 (version 15.0.4551.1003) [69] to obtain a p-value and determine if there was a significant difference between groups. Hierarchical clustering of the measures for many samples was performed using the JMP software (JMP Pro 11), and heatmaps were also made from these hierarchical clusters in JMP [70].

Classification machine learning algorithms were also employed using the Weka software (version 3.6.10) [71]. The SMO algorithm (support vector machine using sequential minimal optimization) support vector machine (SVM) was used for input in which there were samples with a given classification (a classification with one of two values) [72], and the attributes for each sample were all of the measures. All of the default values of the SVM were applied, and 10-fold cross validation was used to train and test the machine learning algorithm. The absolute values

of the SVM weights of the measures were often presented in a bar graph to visually determine which measures were the most critical in order for the SVM to distinguish between two classes. Although the SVM classifier can often perform very well, this algorithm is somewhat like a black box and provides the user with little information about how it is making the decision to assign a certain class to a sample. Therefore, a J48graft tree algorithm was also employed. This algorithm outputs a tree displaying which ranges of values for certain measures are associated with certain class assignments. All of the default values for the tree were used, and 10-fold cross validation was set for the training and testing of the machine learning algorithm. The tree output from the algorithm is often presented as a figure, and the tree is made up of lines such as "cv > 1.466364: D (46.0/12.0)". This line indicates that at this level of the tree, if the sample has a cv greater than the indicated value, then it is assigned the class of disease. In this case the algorithm made this assignment 46 times and was incorrect 12 times. A Naïve Bayes algorithm was also employed since this is one of the most commonly used classification algorithms in the Center for Innovations in Medicine.

The classification performance with the AbStat measures is also often compared to the classification performance using a specific peptide analysis. A specific peptide analysis was performed by choosing the top 100 peptides which distinguish samples from two different groups. These peptides were determined by choosing the peptides with the most significant p-value from a t-test with the peptide fluorescence intensity values for the samples in the two groups. However, performing the t-test to select the peptides with the whole dataset can lead to overtraining. Therefore, the samples were split so that half of the samples were randomly assigned to the training set, and the other half was assigned to the test set. Once the peptides were selected, the values for the training set were input into a classification algorithm. This classification algorithm was then used to predict the class of the samples in the test set.

A table for the machine learning results is also presented for many datasets. In this table, the following values are given: percent of instances that were correctly classified, the kappa statistic, and the receiver operating characteristic (ROC) area. The kappa statistic indicates the level at which the classifier performed better than chance. The larger the value of the kappa

statistic from zero, the better the algorithm performed. Zero or lower indicates poor performance. The ROC area is the area under the curve resulting from plotting the true positive rate vs the false positive rate. An area of 1 indicates the best performance, and an area as low as 0.5 indicates poor performance. The tables used also provide data for the performance of the classifier when the class of each sample was randomly assigned in a manner so that the total count of normal and disease or young and aged samples remains the same as the count before the random assignment. With random class assignment, one would expect the machine learning algorithm to acquire little knowledge from the training data, and to perform very poorly when attempting to make classifications. Good performance with the actual class assignments and poor performance with random classifications would indicate that the values of the sample attributes are capable of distinguishing the classes, and that the classifier is not just finding patterns in noise or extraneous data. Therefore, these machine learning algorithms can help determine whether the AbStat measures are useful for diagnosis.

All of these different methods provide information from different angles about the association between certain measures, values, and the state of the sample.

1.3 **Results**

1.3.1 AbStat changes with artificial antibody experiments

Are there changes in the AbStat measurement with artificial antibody experiments? For example, are there relationships between antibody affinities and changes in the AbStat measures? What happens when increasing concentrations of monoclonal antibody is added to sera?

What prompted these questions? After examining some preliminary datasets, it was clear that there was a difference in entropy between normal and disease groups. Therefore, I had to think of a possible explanation for this difference. The hypothesis I proposed was that the differences in entropy were due to a difference in the number of high and low affinity antibodies. This hypothesis is based on the assumption that antibody specificity is correlated with affinity. More precisely, the assumption is that antibodies with low affinity for their cognate epitope would

have less specificity and bind more chaotically to the peptide microarray. Antibodies with high affinity for their cognate epitope would have more specificity and bind less chaotically to the peptide microarray. This assumption of a link between specificity and affinity is not always true. However, the link between the two concepts is hard to investigate quantitatively because specificity is rarely quantified by researchers and companies that produce antibodies, whereas affinity is often quantified. Despite these caveats, perhaps there is a relationship between antibody affinity and specificity, which would lead to a relationship to the entropy calculated from the data from the peptide microarray. A non-specific antibody would bind to many different peptides at a different level for each peptide which would result in varying fluorescence intensities and ultimately a high calculated entropy. A correspondence between entropy and specificity could even serve as a new metric for quantitating the specificity of monoclonal antibodies. The following sections address these questions and display changes in AbStat measures with varying conditions.

1.3.1.1 Monoclonal affinities and AbStat Measures

Understanding the way that antibody affinities affect the AbStat measures is an important concept that provides information about the behavior of normal, disease, and age sera. Antibody affinities are different in normal, disease, and aged states. In order to test the effects of antibody affinity, three different antibodies with different affinities for their cognate epitope were applied to the peptide array. Note that the affinity of an antibody will affect how the antibody will bind to its cognate epitope, and this will also affect how the antibody binds to an array of non-natural sequence peptides. For example, an antibody with a very low Kd, and thus a very high affinity for its cognate sequence, should bind to a few peptides on the array which contain a sequence that closely resembles the original cognate epitope. Therefore, the fluorescence intensity for these peptides would be very high, while the fluorescence intensity for all of the other peptides would be much lower. A fluorescence intensity distribution with a few very high fluorescence intensities would have a high and tight peak near a fluorescence intensity of zero since most of the features would have a very low fluorescence intensity.

A low affinity antibody should exhibit a different behavior. A low affinity antibody would not bind tightly to its cognate epitope, and therefore a low affinity antibody would also not bind tightly to non-natural sequence peptides containing a sequence similar to the cognate epitope. The binding pattern of a low affinity antibody would therefore have less specificity and exhibit more chaotic and random binding. The antibody would bind loosely to a wide variety of different features, but none of these features would be expected to have a very high fluorescence intensity. The resulting fluorescence intensity distribution would be a much broader peak near zero fluorescence intensity when compared to the tight peak produced from an antibody with high specificity and affinity. The relative values of the AbStat measures can be approximated from the shape of the fluorescence intensity distribution.

The researcher Rebecca Halperin applied three monoclonal antibodies with different affinity Kd constants to the CIM10Kv1 non-natural sequence peptide array: mouse IgG1 anti-h-cMyc antibody from AbD Serotec (Cat No MCA2200G) with a Kd of 80 nM, mouse IgG1 anti-h-proenkephalin-B with a Kd of 1.95 nM from AbD Serotec (Cat No 4140-0159), and mouse IgG1 anti-h-p53 with a Kd <0.1 nM from Millipore (Cat No CBL404). Note that one of the key findings from Rebecca Halperin's research with monoclonal antibodies is that different antibodies exhibit vastly different binding distributions since some can bind as much as 70% of the peptides and other antibodies only bind 0.1% of the peptides 73. These very different binding distributions should result in different entropy values. A line graph of the entropy for the three different Kd values is displayed ("Figure 1 Entropy of antibodies versus affinity for cognate sites"). Note that the change in entropy is not linear with the Kd which may hint that the relationship between the affinity of an antibody for its cognate epitope and the degree of randomness with which this antibody binds to a microarray of non-natural sequence peptides is a complex relationship. More data points could help define the relationship more precisely. A heatmap displaying the general trends (increasing or decreasing) for all of the AbStat measures across the Kd range is displayed in "Figure 2 Heatmap of AbStat measures for antibodies versus affinity for cognate sites".

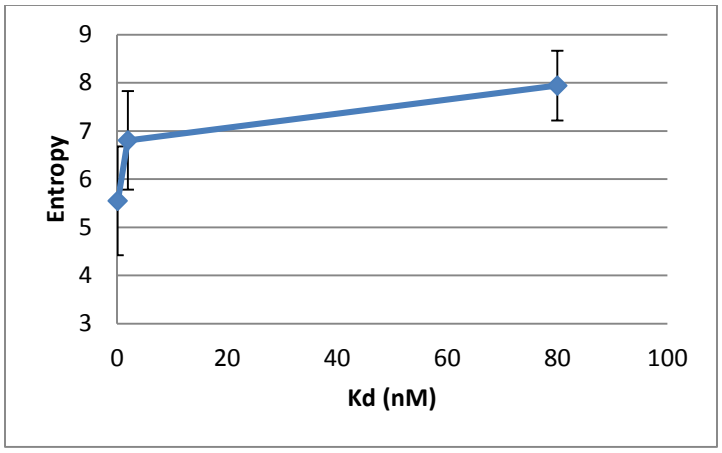


Figure 1 Entropy of antibodies versus affinity for cognate sites

An anti-h-cMyc antibody with a Kd of 80 nM, an anti-h-proenkephalin-B antibody with a Kd of 1.95 nM, and an anti-h-p53 antibody with a Kd<0.1 nM was applied to a peptide microarray separately. The resulting entropy of each sample was calculated and plotted.

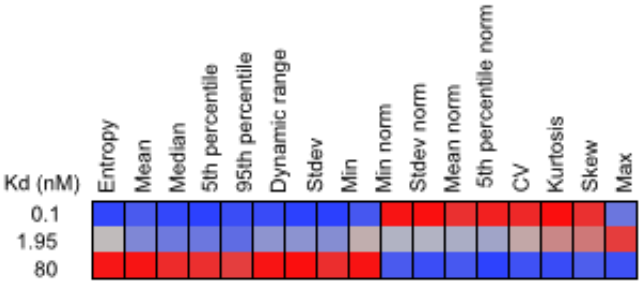


Figure 2 Heatmap of AbStat measures for antibodies versus affinity for cognate sites

Each column corresponds to an AbStat measure and each row corresponds to an antibody. The relative average value of each AbStat measure for three different antibodies is represented by a color with blue indicating the lowest relative value and red indicating the highest relative value. This indicates for example that the 80 nM antibody exhibits the highest entropy (red) while the 0.1 nM antibody exhibits the lowest entropy (blue). Each of the three antibodies represented has a different Kd value for its cognate epitope. The three rows correspond to antibodies with a Kd value of <0.1, 1.95, or 80 nM.

A histogram and box plot of the raw intensities from one of the gpr files for the <0.1 nM antibody is displayed in Figure 3, and a histogram of the raw intensities from one of the gpr files

for the 80 nM antibody is displayed in Figure 4. The high tight peak for the antibody with higher affinity (lower Kd) for its cognate epitope may be the result of tight high intensity binding to a few peptides with mimotopes to the cognate epitope on the non-natural sequence peptide microarray. Therefore, a few peptides bind with high intensity, and the majority of the distribution of peptides does not bind well and exhibits a fluorescence intensity near zero. The antibody with low affinity (high Kd) on the other hand, may not bind to mimotopes of its cognate epitope very well and bind loosely to many different peptides on the array. This would result in a broader flatter fluorescence intensity distribution as many peptides exhibit a fluorescence intensity above zero, but still at a low level.

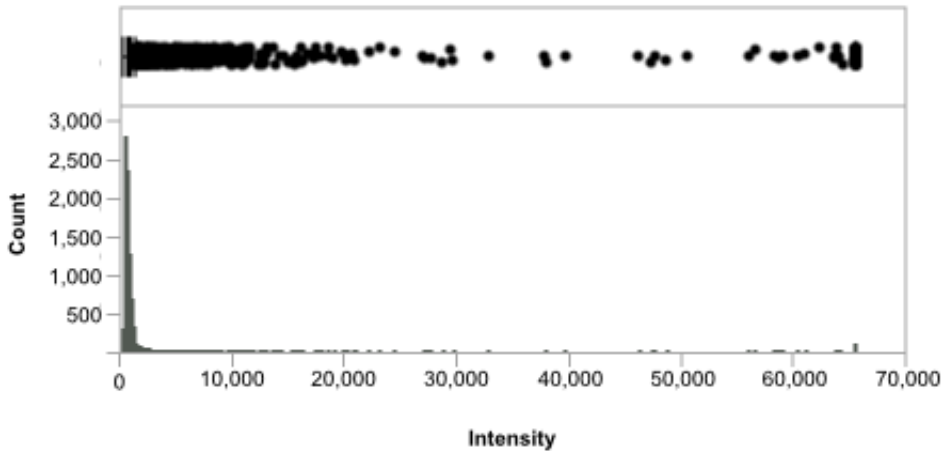


Figure 3 Histogram and box plot for <0.1 nM Kd antibody

Histogram of fluorescence intensities of all 10,000 peptides after an antibody with <0.1 nM Kd for its target is applied. The bin width for this histogram is 140. A box and dot plot is displayed above the histogram with outliers displayed as dots. In this graph, almost all of the fluorescence intensities are less than 1,000, and the "box" part of the box and dot plot is not visible.

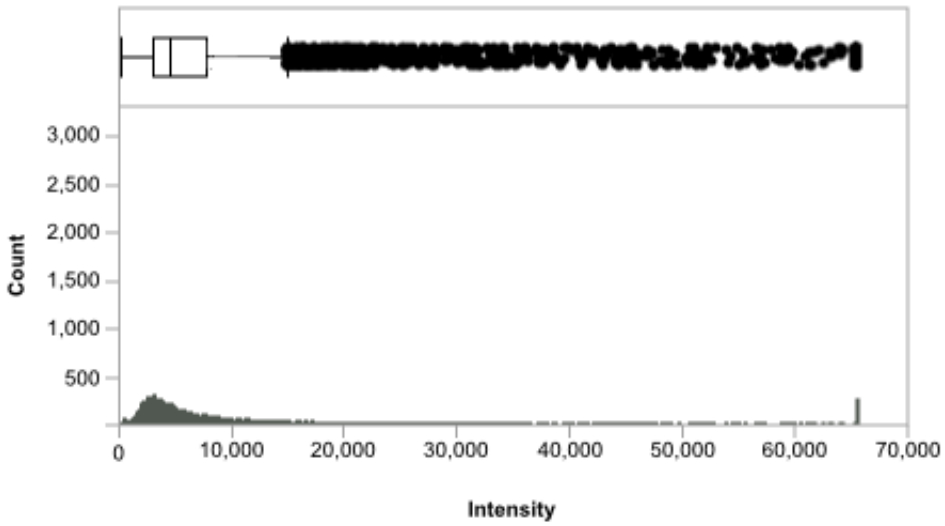


Figure 4 Histogram and box plot for 80 nM Kd antibody

Histogram of fluorescence intensities of all 10,000 peptides after an antibody with 80 nM Kd for its target is applied. The bin width for this histogram is 140. A box and dot plot is displayed above the histogram with outliers displayed as dots.

These results match the predictions. A high affinity antibody produced a fluorescence intensity distribution with a high tight peak near zero. Figure 3 illustrates that the highest count for a given fluorescence intensity bin exceeds 2,500 with a high affinity (low Kd) antibody. Figure 4 illustrates that the highest count for a given fluorescence intensity bin is less than 500 for a lower affinity antibody. Therefore, the low affinity antibody produced a fluorescence intensity distribution with a much broader flat peak near zero. This is the result one could expect from more random binding. If we were to use a random number generator to pick 10,000 intensities, the result would certainly look more similar to Figure 4 than Figure 3 since a perfectly random distribution would appear most like a flat square. Ultimately, the shape of the fluorescence intensity distribution indicates the specificity of the antibody, and these monoclonal antibody results support these claims.

1.3.1.2 Spiking antibody into sera

In the previous example, monoclonal antibodies were applied to the array, but this is far from a practical case. In a real situation, the blood will contain a high diversity of many different antibodies since there are a total of about 10^{11} cells in the B-cell population of an individual [11]. Additionally, the complexity of a whole antibody repertoire can change as a few single antibodies against a few targets may come to dominate the mixture. This scenario can occur when there is a strong immune response against a single virus or bacteria type. Alternatively this scenario could also occur when a lymphoma causes a single B cell to proliferate out of control and produce many antibodies against a single target. In these scenarios, the complexity of the antibody repertoire is reduced as copies of antibodies against a single target make up an increasingly greater portion of the antibody repertoire. This type of situation was simulated by spiking increasing concentrations of monoclonal antibody into sera. The expectation is that a reduction in the antibody repertoire complexity that interacts with the peptide microarray will result in a reduction in complexity in the fluorescence intensity distribution. A lower complexity fluorescence intensity distribution will result in changes in the AbStat measures. For example, the entropy will decrease as the concentration of monoclonal antibody in the sera is increased.

Two experiments were performed in which antibody against a particular target was added to normal sera in increasing concentrations. In the first experiment, polyclonal mouse sera against the human GFOD1 protein was added to normal mouse sera. In the second experiment, human monoclonal antibody against the human gp120 HIV protein was added to normal human sera. Rebecca Halperin also wrote about a similar spiking experiment in her dissertation, and found that there was no significant change in binding pattern if a monoclonal antibody was applied to the array alone or mixed into sera [73]. She was not investigating entropy at that time, and was more interested in peptides that uniquely bind to the antibody.

Polyclonal sera against the human GFOD1 (glucose fructose oxidase 1) protein was added to normal mouse sera at increasing concentrations by Josh Richer. The polyclonal IgG mouse antibody was added to a 1:500 dilution of normal mouse sera in order to obtain the following final antibody concentrations: 0.1 nM, 1 nM, 2.5 nM, 5 nM, 10 nM, and 40 nM. There

were two technical replicates for each condition applied to the CIM10Kv2 arrays. A line graph of the entropy vs antibody concentration ("Figure 5 Entropy for increasing concentrations of α-GFOD1 antibody into normal mouse sera") reveals that the entropy decreases as the concentration increases up to 5 nM and then the entropy starts to increase. The order of the measures with the most significant p-values between normal mouse sera and 5 nM antibody is as follows: 95th percentile, 5th percentile, median, mean, entropy, and cv. The concentration of 5 nM was chosen as the comparison point because this is the point at which the trend in the curves reverses. Note that in Figure 5, the entropy at 2.5, 5, and 10 nM is about the same when error bars are taken into account, but the entropy at lower antibody concentrations (0, 0.1, and 1 nM) is quite different. Note also that the error bars at an antibody concentration of 0, 0.1, and 1 nM do not overlap. These observations suggest that entropy is more sensitive to the addition of antibody at lower concentrations, and resolution of the entropy measure decreases at high antibody concentrations.

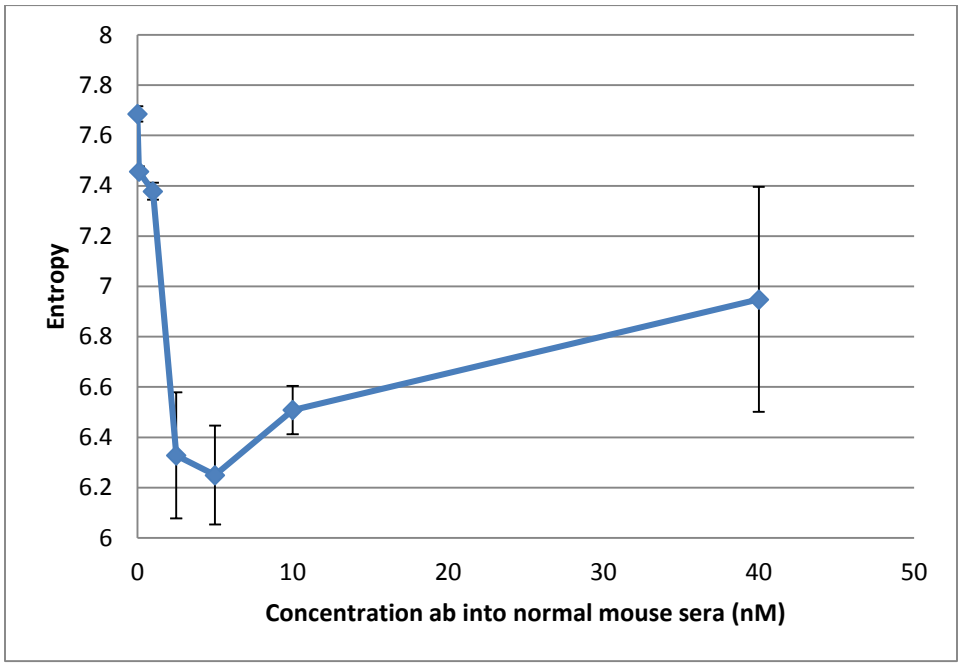
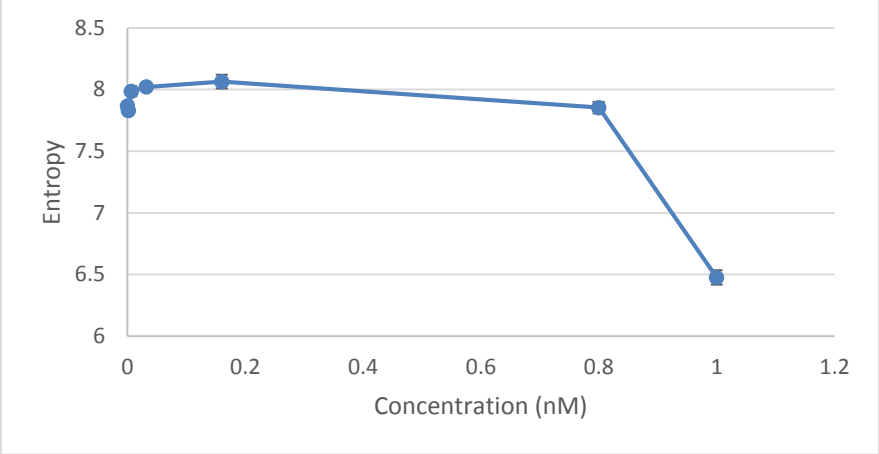


Figure 5 Entropy for increasing concentrations of α-GFOD1 antibody into normal mouse sera

Increasing concentrations of polyclonal sera against the GFOD1 protein were added into normal mouse sera and the entropy of each technical replicate was calculated. The entropy values at each concentration are plotted on a line plot.

A second experiment was performed by Heidi Larsen and Bart Legutki in which the human B12 monoclonal antibody against the gp120 HIV protein was spiked into a 1:500 dilution of normal human sera at increasing concentrations and applied to the CIM10Kv2 arrays. A very wide range of antibody concentrations was tested: 0.00128 nM, 0.0064 nM, 0.032 nM, 0.16 nM, 0.8 nM, 1 nM, 4 nM, 10 nM, 20 nM, and 100 nM. Two technical replicates were performed for each concentration. Since the range of dilutions tested is so wide, the data was graphed across two separate ranges for the entropy (Figure 6 A and B).

A)



B)

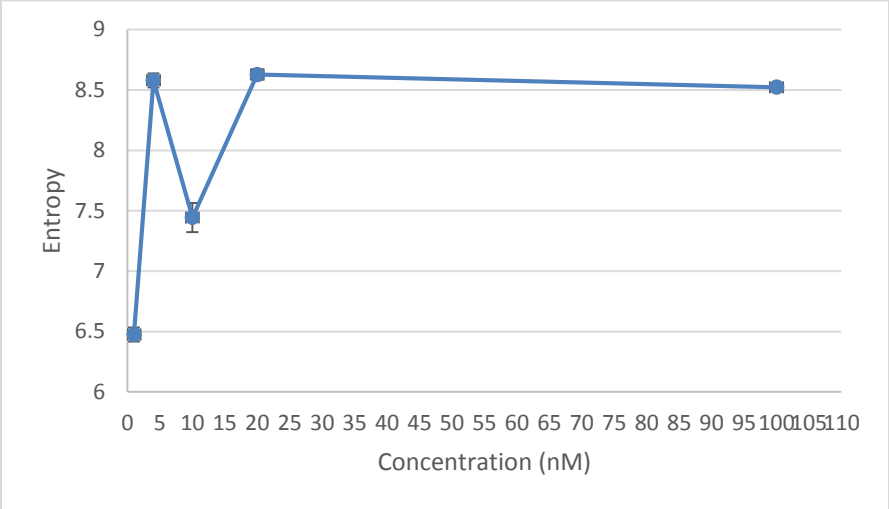


Figure 6 Entropy for increasing concentrations of α–gp120 antibody into normal human sera

Increasing concentrations of antibody against the gp120 HIV protein was spiked into normal human sera and applied to the HT330K peptide microarray. The entropy for each replicate at each concentration was determined and a line plot was created covering α-gp120 antibody concentrations from 0 to 1 nM in A) and 0 to 100 nM in B).

The results from increasing concentrations of monoclonal antibody into sera support the expectations up to a point, but there were also some surprises. The value of the entropy measure did in fact decrease as the concentration of monoclonal antibody in the sera increased. However, after a certain point in both experiments, the entropy began to increase again. This result may have been the result of saturation of target peptides on the array by the monoclonal antibody. After a certain concentration was reached, the monoclonal antibody then began to "spill over" and bind to a variety of different features on the array in a more chaotic manner. This hypothesis is explored more fully in the discussion section ("1.4.1.2 Spiking antibody into sera"). Ultimately, this controlled experiment provides insight into how the AbStat measures can provide information about health, disease, and aged sera states as antibody mixtures become more complex or alternatively are dominated by antibodies against a few targets. For example, in a

lymphoma, one antibody with one target begins to increasingly dominate the repertoire, and this change will result in decreases in entropy up to a point.

1.3.2 Mouse vaccines and infections

Are there changes in the AbStat measurement with artificial mouse vaccines and infections? During an infection or a vaccination, the affinity, avidity and abundance of certain antibodies in the bloodstream should change, and this change should be detectable in changes in the fluorescence intensity distribution acquired from applying sera with antibodies to a non-natural sequence peptide microarray. The expectation is that the antibody repertoire complexity will decrease during a vaccine or infection. During these events a few antibodies against a few specific epitopes will make up an increasing portion of the antibody repertoire, thus reducing the total complexity of the repertoire. This repertoire will produce higher tighter peaks in the fluorescence intensity distributions and therefore a lower entropy value.

There are three experiments to help investigate the issue of mouse vaccines and infections: 246 day time course, a multiple vaccine experiment, and a 6 day time course. The 246 day time course after a vaccine spans a very long time (more than half a year). During this time the entropy should decrease after a vaccine. However, as the mice ages, the specificity of new antibodies will decrease, and therefore the entropy should increase with time. This behavior is also observed in the "Changes with age" section ("1.3.7 Changes with age"). In the multiple vaccine experiment, different groups of mice are immunized with different vaccines. The expectation is that the entropy of the fluorescence intensity distribution from vaccinated mice should be lower than the entropy of the mock vaccinated group. In the 6 day time course after infection the expectation is that entropy would decrease after the mice are vaccinated, unless this time interval is too short to observe any difference. The results from these mouse experiments can provide some insight into what might happen during the course of human vaccines and infections.

1.3.2.1 246 day time course

Bart Legutki performed an experiment with ten 4-5 week old female BALB/c mice, and these mice were vaccinated ("day 0") and challenged with H1N1 A/PR/8/34 influenza 35 days later [9]. Serum was collected 0, 14, 28, 51, and 246 days after vaccinations. The serum was applied to CIM10Kv1 non-natural sequence peptide arrays with three technical replicates for each timepoint with the exception of day 28 which had two timepoints. In this experiment, one of the key findings was that the immune system of the mice responded to 283 influenza-specific peptides after vaccination, and the response to these peptides was present 211 days post-challenge. Therefore, the immune response was present over a long period of time [9]. A heatmap of the AbStat measures for all of the replicates ranging from day 0 to day 246 is displayed in Figure 7. A line graph of the entropy measure over time is displayed in Figure 8. The change in entropy from day 0 to day 246 is statistically significant with a p-value of 6.79E-3. During the experiment, weights of the mice were collected, and these weights indicated that the mice recovered after about 14 to 21 days after infection.

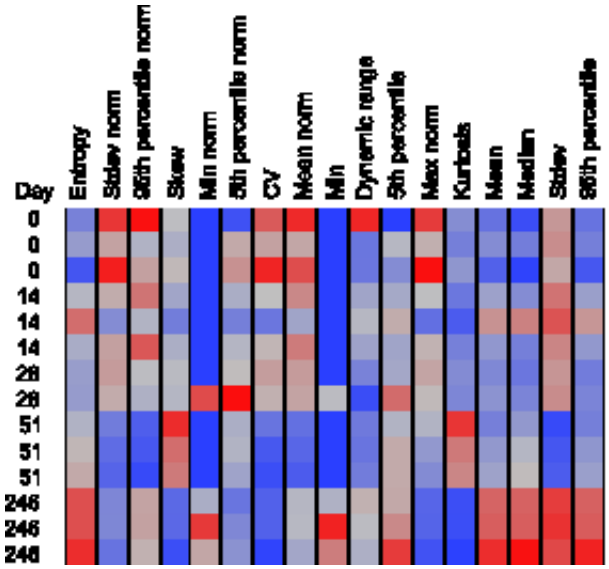


Figure 7 Heatmap of AbStat measures for 246 day mouse time course with 2-3 technical replicates for each timepoint.

Each column corresponds to an AbStat measure and each row corresponds to a replicate mouse sera sample after a certain number of days from vaccination with H1N1 A/PR/8/34 influenza. The

mouse samples were applied to the CIM10kv1 peptide microarray. The relative average value of each AbStat measure is represented by a color with blue indicating the lowest relative value and red indicating the highest relative value. This indicates for example that three 0 day samples exhibit a lower entropy (blue) than the three 246 day samples (red).

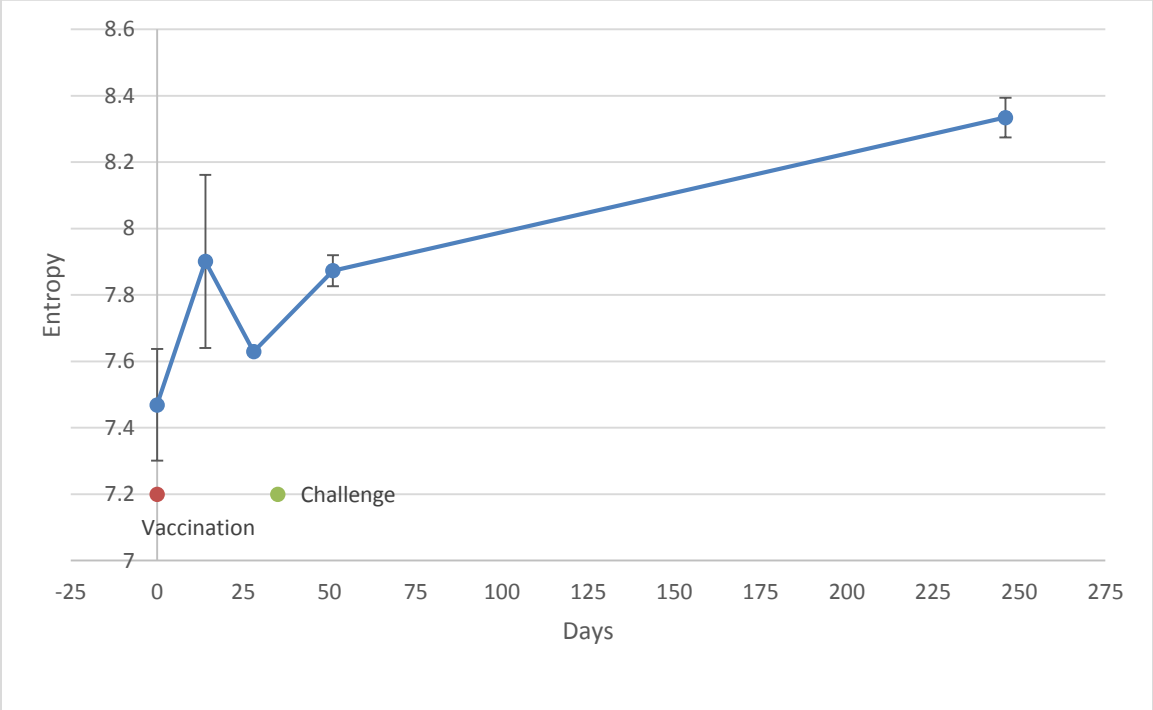


Figure 8 Line graph of entropy over time for 246 day mouse time course

Sera samples from mice after vaccination with H1N1 A/PR/8/34 influenza on day 0 were applied the CIM10kv1 peptide microarray. The mice were challenged with the virus on day 35. The entropy of each technical replicate at each time point was calculated and displayed on a line graph.

The results of the 246 day challenge were interesting. The entropy did decrease after vaccination after enough days had passed. Over the long time course the entropy also increased as expected. Therefore, this experiment provided important information about the behavior of the antibody repertoire and AbStat measures during the course of a vaccine as well as during the normal aging process of a mouse. Using the AbStat measures it was also possible to distinguish between the young and aged timepoints of the mice.

1.3.2.2 Multiple mouse immunizations

Bart Legutki performed an experiment in which different groups of mice received different vaccines. In this experiment, there were a total of four groups: mice infected with the 2006-2007 influenza vaccine (06_INF_vacc), mice infected with the 2007-2008 influenza vaccine (07_INF_vacc), mice infected with influenza (infected), mice injected with formalin fixed PR8 (killed PR8), and a mock group of mice that received PBS intranasally (mock) [8]. There were 10 samples in each group with a replicate for a total of 20 peptide array results (gpr files) per group. The 2006-2007 and the 2007-2008 vaccine had the same H2N3 strain, same B strain, but a different H1N1 strain. Specifically, the 2006-2007 Fluzone vaccine had the following composition: H1N1 A/New Caledonia/20/99, H2N3 A/ Wisconsin/67/2005 and B/Malaysia/2506/20004. The 2007-2008 Fluvarin vaccine had the following composition: H1N1 A/Solomon Islands/3/2006, H2N3 A/ Wisconsin/67/2005, and B/Malaysia/2506/20004. The sera was taken from these mice on day 38 after treatment, and the sera was later applied to CIM10Kv2 arrays. The results from the experiment revealed that the 2006-2007 vaccine resulted in 60% survival after challenge and the 2007-2008 vaccine resulted in 80% survival after challenge [8]. The researchers also found that they could identify reactive peptides on the array which distinguished the groups of mice. Their analysis revealed that some of the reactive peptides contained epitopes found in the PR8 virus.

A graph of the entropy of each group is presented in Figure 9. The entropy value between the mock group and the group which received the 2007-2008 vaccine was statistically significant with a p-value of 0.0251, but the entropy value between mock and infected was not statistically significant (p-value: 0.530). Note that the 2007-2008 vaccine was the more effective vaccine of the two. The AbStat measure with the most significant p-value between mock and infected was the dynamic range (p-value: 0.0453).

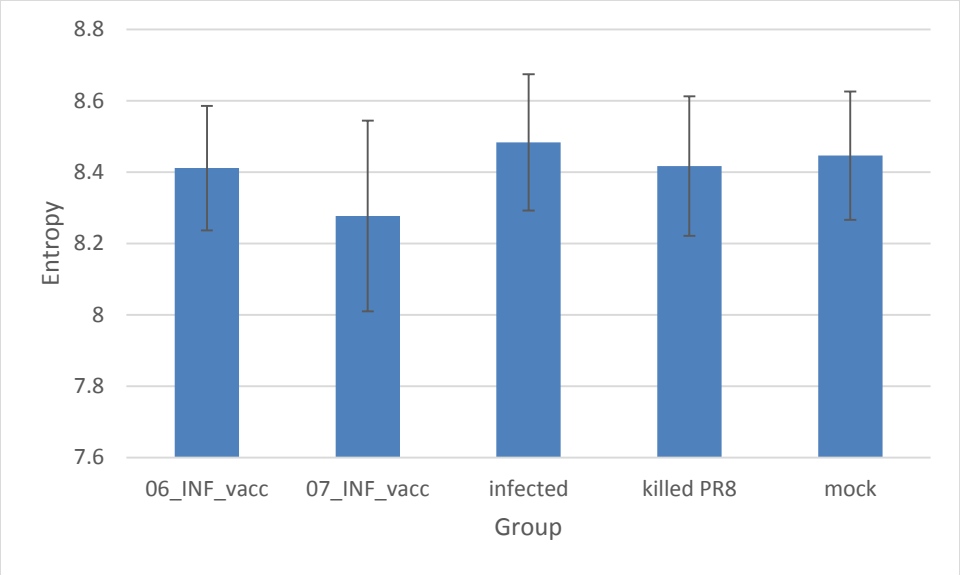


Figure 9 Entropy for multiple mouse immunization experiment

Sera from 5 groups of mice were applied to the CIM10Kv2 array 38 days after an injection with the 2006-2007 influenza vaccine, the 2007-2008 influenza vaccine, the influenza PR8 virus, killed formalin fixed PR8 virus, or a mock PBS treatment. There were 10 mice in each group and there was an array replicate for each mouse for a total of 20 array results per group. The entropy of each array result was calculated and presented in a bar graph.

The AbStat measures for the mock and infected group samples were input into a J48graft tree and SVM classification algorithm. A J48graft tree classification algorithm can only correctly classify 45% of the instances with a kappa statistic of -0.1 and an ROC area of 0.413. The SVM on the other hand can correctly classify 62.5% of the instances with a kappa statistic of 0.25 and an ROC area of 0.625. The instances were also assigned classes (infected or mock) randomly to test whether the SVM performance would remain the same or decrease. With randomly assigned classes, the SVM correctly classified 55% of the instances with a kappa statistic of -0.0496 and an ROC area of 0.478.

One prediction about this experiment was correct. There was a vaccine group that did have lower entropy than the mock immunized group (Figure 9). However, the 06_INF_vacc did not exhibit lower binding than the mock group. Perhaps the resolution provided by the array platform was not high enough to distinguish between the groups even though a difference was

present. Alternatively perhaps the difference in entropy reflects the fact that the 2006 vaccine did not produce an immune response that was as effective as the 2007 vaccine since there was 60% survival vs 80% survival. The more effective 2007 vaccine may have resulted in higher affinity antibodies which resulted in a "tighter" fluorescence intensity distribution from the non-natural sequence peptide array. Another interesting result is that the dynamic range of the fluorescence intensity of the infected group was considerably lower than the other groups. Perhaps the infected group produced a fairly high tight immune response which resulted in a low dynamic range, but the binding within this range may still have been fairly chaotic in a manner which did not significantly decrease the entropy value. Note that there could be competing reactions with divergent effects detected in the immune system through the microarray as well. Overall, these results provide support for the idea that vaccines and infections change the overall binding observed from an antibody repertoire interacting with a peptide array, and this change is reflected in the AbStat measures.

1.3.2.3 6 day time course

One group of five female BALB/c mice was infected with the H1N1 A/PR/8/34 influenza strain on day 0, and another matching mock group of mice was treated with PBS. Sera was collected each day for six days, and the last sera samples were collected by terminal bleed. Samples were applied to the CIM10Kv1 non-natural sequence peptide array with two to three technical replicates for each day.

From this experiment, Bart Legutki noted several key findings in an unpublished manuscript. One finding is that an immune response to the influenza infection in mice can be detected very soon after the initial infection. Twelve peptide features increased in signal within the first three days after infection. He also found that the limit of detection of the non-natural sequence peptide microarray is more sensitive than the limit of detection of the standard ELISA assay. A peptide feature could be detected 7 fold above background at a dilution as high as 1:1,638,400, whereas the limit of detection of the ELISA assay was reached at a dilution of

1:409,600. Once the immune response of mice or humans was raised, this immune response could be detected in the immunosignature as far as one month after the infection.

I used the fluorescence intensity distribution from this mouse influenza experiment to obtain values for the AbStat measures. None of the AbStat measures comparing the mock and infected on day 6 were statistically significant. The best p-value was 0.195 for the mean, and the normalized minimum, median, and entropy followed. A line graph of the entropy for the 6 days is presented in the Figure 10.

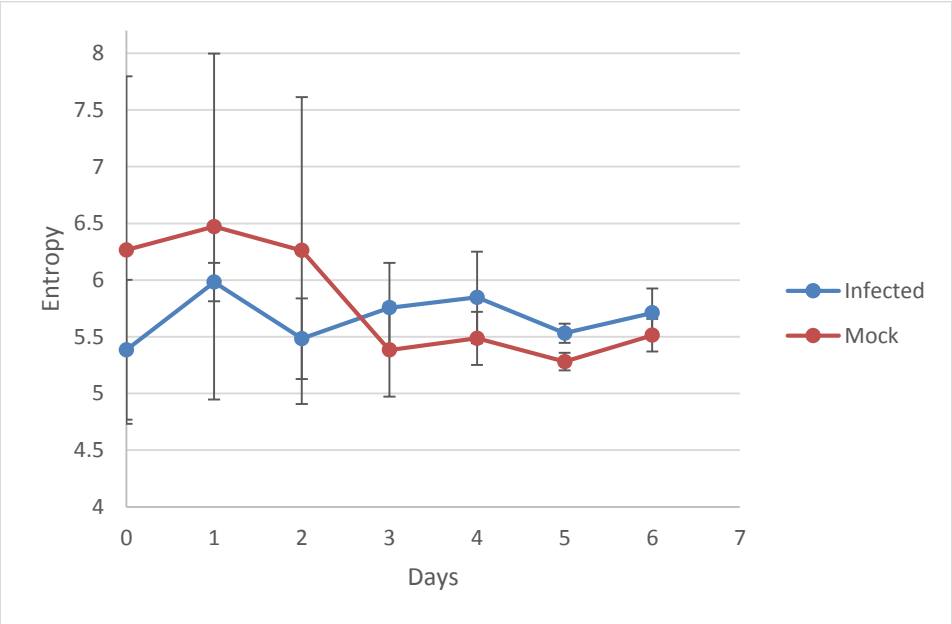


Figure 10 Entropy for 6 day mouse time course

Sera from mice were applied to the CIM10Kv1 array each day after injection with the H1N1 A/PR/8/34 influenza virus or PBS for 6 days. There were five mice in both groups. Each sample was applied to the array with two to three replicates, and the entropy of each replicate was calculated and graphed in a line graph.

The results from this experiment demonstrate that it is not possible to distinguish between an infected and a mock group of mice using the AbStat measures in only 6 days. This short time-frame may have been too demanding to measure an antibody response that can be measured with these broad and global AbStat measures. If more timepoints had been taken, perhaps a difference could have been observed. Alternatively, if the superior peptide arrays

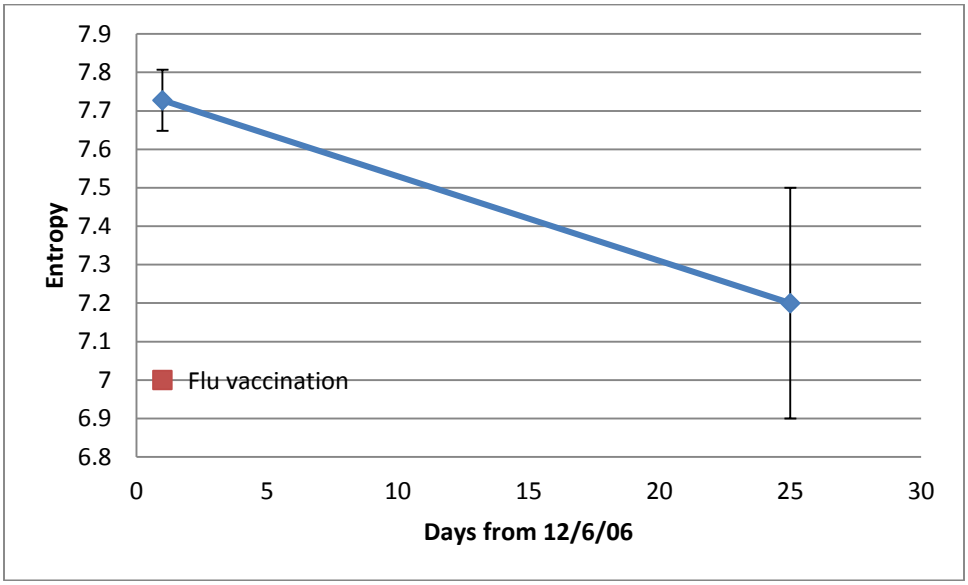
which were developed at a later time had been used, then these arrays may have allowed the samples to be distinguished. These results demonstrate that there are clearly limits on the abilities of the AbStat measures to distinguish infected and mock mouse samples if the time after infection is too short.

1.3.3 Human vaccines

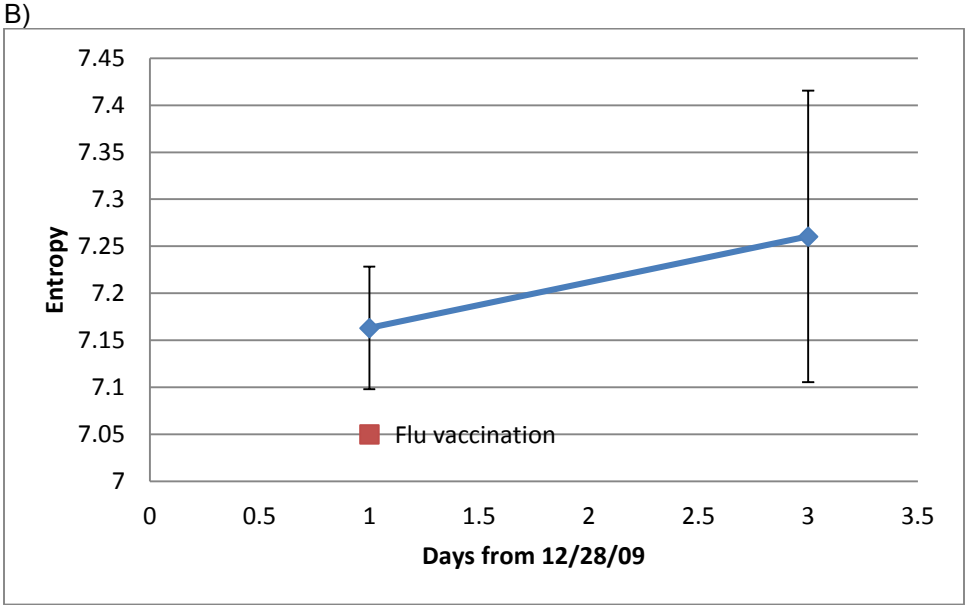
The previous section demonstrated that changes in the AbStat measures can be observed when a mouse is vaccinated, but do the same results hold when a human is vaccinated? If the results do hold true, then this means that the AbStat measures can be used to detect when there is a sudden and strong immune response in humans. This issue was explored by applying human sera pre and post vaccination to the peptide microarray and analyzing the AbStat results. The expectation is that the entropy of the fluorescence intensity distribution will decrease post vaccination after enough days have passed.

Penny Gwynne collected sera from two individuals identified as individual 84 and individual 43 around the time when they received the seasonal flu vaccine. The Peptide Array core under the direction of Zbigniew Cichacz then applied the sera to the CIM10Kv2 non-natural sequence peptide arrays. In the original analysis, the researchers were able to identify specific peptides which significantly increased after the vaccination [9]. The fluorescence intensity distribution was then characterized. Two line graphs of the entropy over time are displayed for individual 43 for the flu vaccine they received in 2006 and 2009 in Figure 11 A) and B). A line graph of the entropy over time for individual 84 for the flu vaccine they received in 2009 is displayed in Figure 11 C.

A)


7.9
7.8
7.7
7.6
7.5
7.4
7.3
7.2
7.1
7
6.9
6.8
Entropy
Flu vaccination
0
5
10
15
20
25
30
Days from 12/6/06

B)


7.45
7.4
7.35
7.3
7.25
7.2
7.15
7.1
7.05
7
Entropy
Flu vaccination
0
0.5
1
1.5
2
2.5
3
3.5
Days from 12/28/09

C)

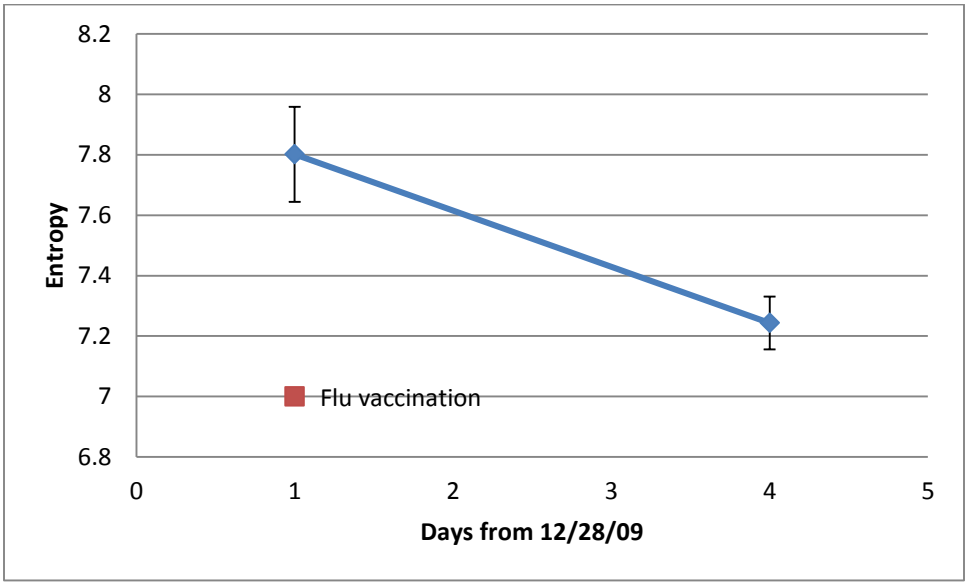
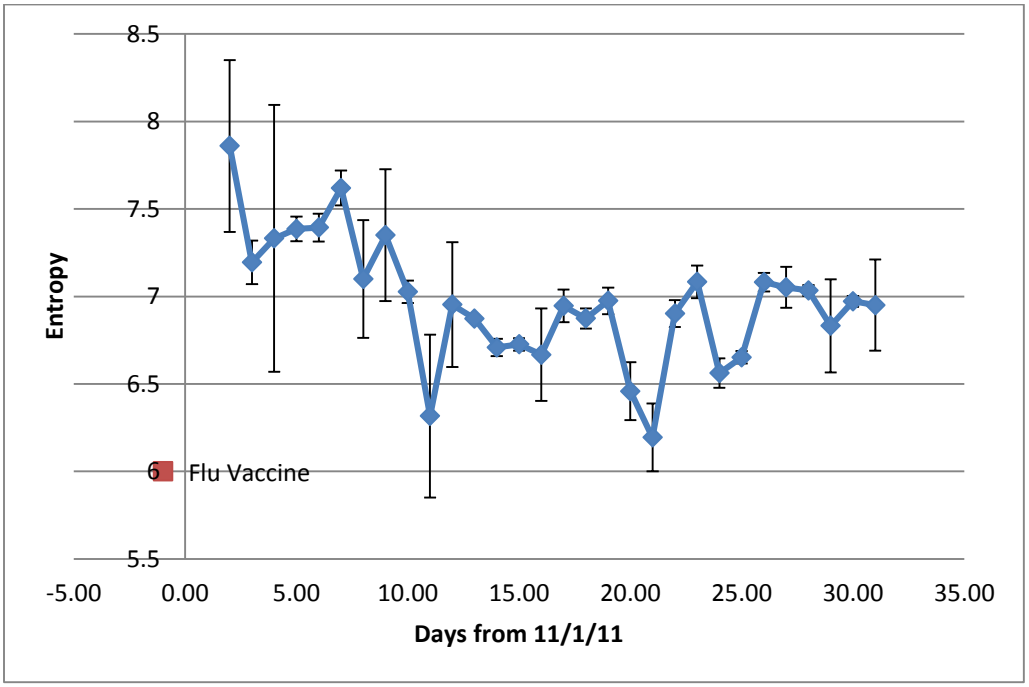


Figure 11 Change in entropy after vaccination

A human individual was immunized with the influenza vaccine, and sera pre and post immunization was applied to the CIM10Kv2 microarray. The entropy of each replicate was calculated and plotted vs time in days in a line graph. In A) individual 43 was vaccinated in December 2006. In B) individual 43 was immunized in December 2009. In C) individual 84 was immunized in December 2009. A red square indicates the time of the vaccination.

During the month of November in 2011 both individuals had their blood drawn on a near daily basis. Just before this month, both individuals received the 2011/2012 seasonal trivalent influenza vaccine. This sera was applied to the CIM10Kv2 arrays. The time course for individual 43 and 84 is presented in Figure 12. Note that individual 84 reported catching a cold on day 17.

A)



B)

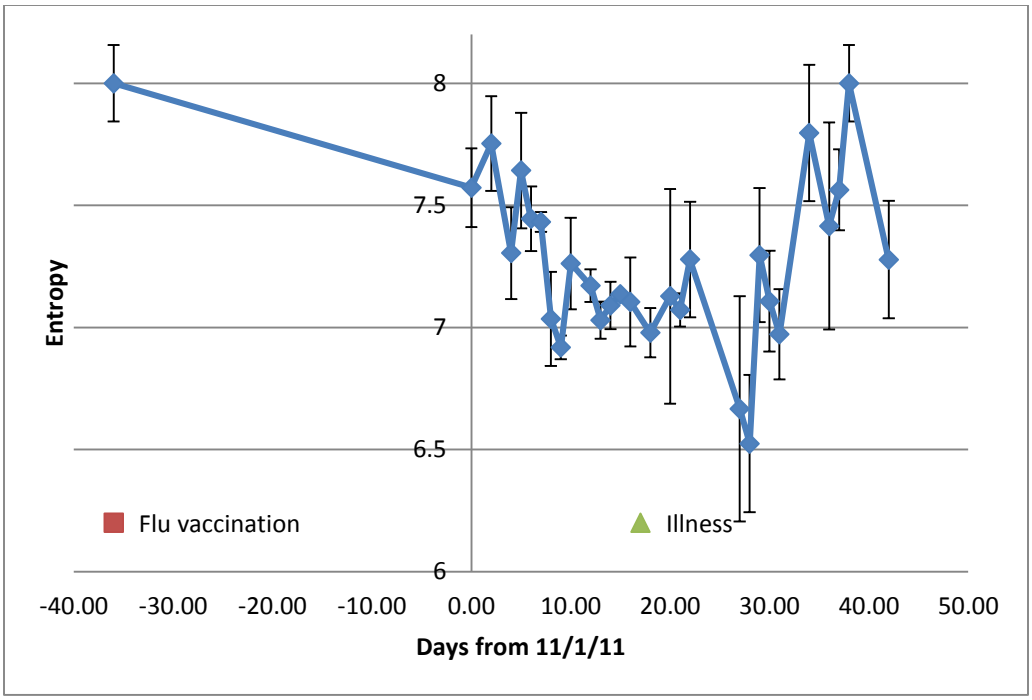


Figure 12 Daily one month entropy change

Sera from two individuals was applied to the CIM10Kv2 during November 2011 on a near daily basis. The entropy of each replicate was calculated and plotted vs time in days in a line graph. The plot in A) represents the entropy results for individual 43, and the plot in B) represents the entropy results for individual 84. The red square indicates the time that the influenza vaccine was received.

In every circumstance in which the entropy was calculated for a sample more than three days after vaccination the entropy decreased. In the one circumstance when the sample was only three days after vaccination, a reduction in entropy could not be observed. These results indicate that a global measure of fluorescence intensities is not capable of detecting a difference between pre and post vaccination in only three days. After three days, the reduction in entropy indicates that the antibody repertoire changes significantly enough for changes in the fluorescence intensity distribution to be observed. These results support the concept that the AbStat measures could reflect changes in the human immune system.

1.3.4 Reduction in antibody repertoire complexity with lymphoma

Are there changes in the AbStat measurement during the course of a lymphoma which reduces the complexity of the antibody repertoire? During the course of a B cell lymphoma one particular antibody in the repertoire becomes dominant as one B cell proliferates uncontrollably and produces large amounts of the antibody. Can this change be detected in the fluorescence intensity distribution produced from applying the sera to an array of peptides? The anticipated result is that the entropy of the fluorescence intensity distribution will decrease when sera associated with a B cell lymphosarcoma (LSA) is applied to the array relative to normal sera. The antibody repertoire for LSA sera will have a higher prevalence of an antibody against one specific target. Therefore, this sera will resemble a monoclonal antibody more than normal sera, and this sera will bind to a few features with high intensity which contain peptides with a mimotope similar to the cognate epitope of the antibody. Many of the other features on the array will have very low intensity ultimately resulting in a less complex fluorescence intensity distribution and a lower entropy value.

The experiment was performed by Bart Legutki using sera from dogs with and without a B cell lymphosarcoma (LSA). These sera samples were obtained from The Flint Animal Cancer Center at Colorado State University. There were 24 different male healthy dogs and 15 female healthy dogs with an age range from 2 to 15 (median of 6) of the following breeds: Mix Breed (19), Golden Retriever (6), Labrador (3), Staffordshire Terrier (2), Australian Cattle Dog (2), Australian Shepherd, Dalmatian, Doberman, German Wire Haired Pointer, Std. Poodle, St. Bernard, and Rottweiler. There were 22 male and 16 female B cell lymphosarcoma dogs with an age range from 2 to 13 (median of 7.9) of the following breeds: Mix Breed (10), Golden Retriever (5), Border Collie (4), German Shepard (2), Rottweiler (2), Scottish Terrier (2), Vizsla (2), Bassett Hound, Belgian Melinois, Boxer, Chesapeake Bay Retriever, Collie, Doberman, Labrador, Miniature Schnauzer, Sheltie, Staffordshire Terrier, and one non-classified breed. The sera from these dogs was applied to the CIM10Kv2 array.

The researcher Bart Legutki and his collaborators made several key findings from this experiment. They found that although LSA is due to the clonal expansion of a single B cell unique to each dog, that there was also a general immunosignature defined by specific reactive peptide features in common among the dogs. Selected peptide features from a training set of dogs was able to predict the health status of a test set of dogs with 97% accuracy (unpublished data). An individual set of peptides unique to each dog was identified as well which consisted of approximately 6 to 8 peptides. The immunosignature at diagnosis was also able to predict which dogs would go into remission and relapse within 120 days. They also found that the median raw feature intensity of the B cell LSA dogs was lower than the healthy dogs (p-value of 1E-4). I calculated the entropy of each sample of the two class types and a dot plot is presented in Figure 13. The p-value between the two groups is 5.47E-10.

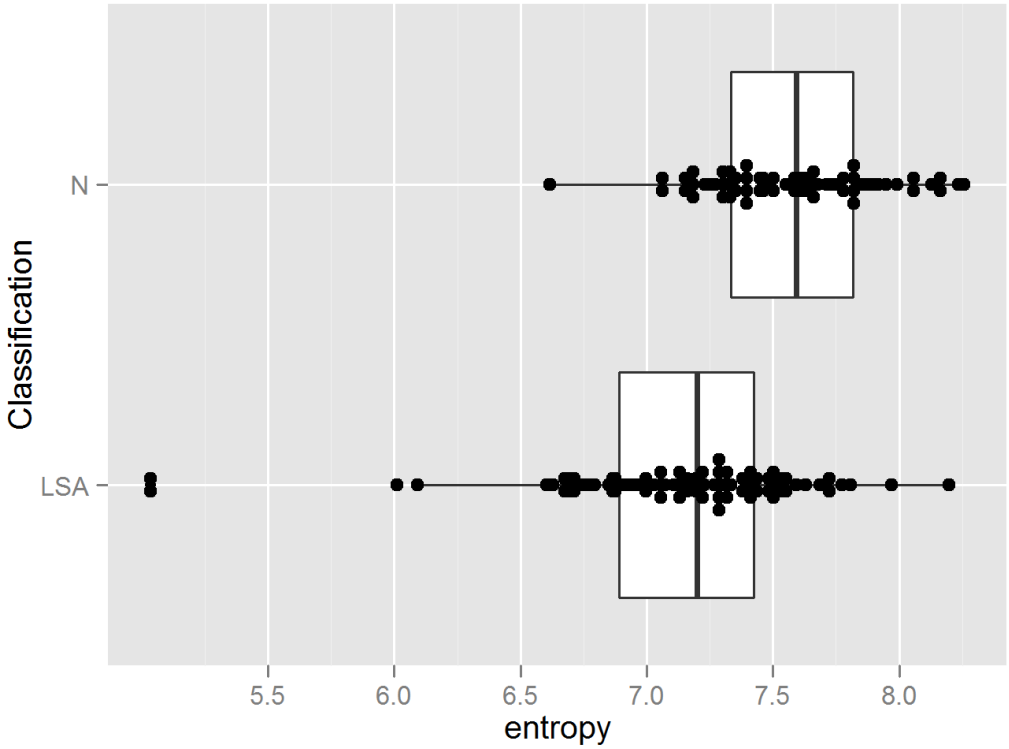


Figure 13 Box and dot plot of entropy for normal (N) and lymphosarcoma (LSA) dogs

Sera from dogs with and without B cell lymphosarcoma were applied to the CIM10Kv2 array. The entropy of each replicate was calculated and the data were plotted in a box and dot plot.

The data for all of the measures for the two groups was input into a SVM algorithm, and the algorithm was able to correctly classify 78.8% of the instances with a Kappa statistic of 0.566 and ROC area of 0.779. When the normal or LSA class was randomly assigned to each sample, the SVM could only correctly classify 52.3% of the instances with a Kappa statistic of 0.0348 and a ROC area of 0.517.

These results demonstrated that the entropy of the LSA group of dogs was indeed lower than the entropy of the normal dogs. The reason for this is that in a B cell lymphoma one antibody comes to predominate the repertoire. The solution comes to resemble more of a monoclonal antibody rather than a complex antibody mixture, and the entropy value is closer to that of a monoclonal antibody. This result was also observed in the section in which a monoclonal antibody was artificially spiked into monoclonal sera (“1.3.1.2 Spiking antibody into

sera"). A similar phenomena is also observed with vaccines as was demonstrated with mouse vaccines and human vaccines ("1.3.2 Mouse vaccines and infections" and "1.3.3 Human vaccines"). The reason that this is the case is due to the fact that vaccines are also characterized by an increase of the number of antibodies against one specific epitope or a few epitopes. In conclusion, these results show that the AbStat measures could be used to monitor dogs for the occurrence of B cell lymphomas. Presumably, the same techniques would work for human sera as well.

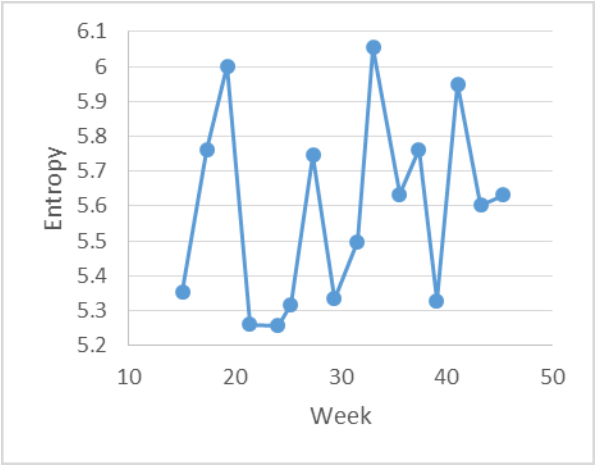
1.3.5 Mouse cancer progression

Are there changes in the AbStat measurements as mice develop cancer? This question was addressed using a wild type mouse and a transgenic mouse which develops mammary tumors. The expectation is that the entropy of the wild type control mouse will increase over time with age. However, the entropy of the cancer mouse should increase more as the antibody repertoire becomes more chaotic and complex during the course of the chronic battle against cancer. If changes can be detected, then the AbStat measure could be used for the early detection of cancer.

Hu (Tiger) Duan performed the experiments to address this issue with a transgenic mouse strain. The strain of the transgenic mouse was FVB/N neuT which begins to develop spontaneous mammary adenocarcinomas around 33 weeks old [74, 75]. One transgenic mouse and one wild type mouse were used. The breeding pairs of transgenic mice were obtained from Dr. Joseph Lustgarten from the Mayo Clinic, Phoenix, AZ. The transgenic mice used for the experiment were bred at Arizona State University. The wild type mice were obtained from Jackson Laboratories. The first tumor in the transgenic mouse was detected after 231 days (33.1 weeks) with a size of 0.5 mm^3. Sera samples from these mice were applied to CIM10Kv2. By analyzing specific peptides, Tiger Duan was able to identify array features that began to exhibit increased fluorescence intensity over time eight weeks before the appearance of the first palpable tumor (33.1 weeks). The broad AbStat measures view the data at a global level without

analyzing specific peptides. A line graph of the entropy over time for the transgenic mouse and the wild type mouse is presented in Figure 14.

A)



B)

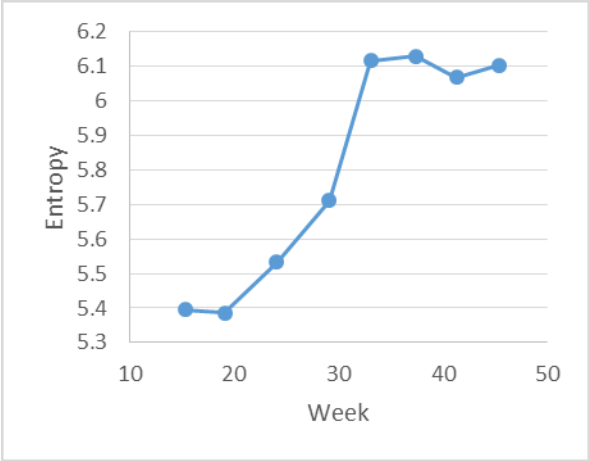


Figure 14 Entropy time course for transgenic and wild type mouse

Sera from a A) transgenic FVB/N neuT mouse and a B) wild type mouse were applied to the CIM10Kv2 microarray. The entropy of each sample was calculated and plotted vs time in weeks in a line graph.

The chaotic nature of the entropy over time for the transgenic mouse was unexpected. The wild type mouse exhibited the steady increase in entropy over time that was expected and also observed in the "246 day time course" section after vaccination ("1.3.2.1 246 day time course"). This result provides further evidence that the entropy obtained from the antibody

repertoire increases with age. The transgenic mouse entropy was expected to increase to a higher level than the normal mouse, but instead the entropy was rather chaotic with constant sharp increases and decreases. This chaotic entropy timeline may be the result of the immune system undergoing constant surges against the tumor. For one period of time, perhaps many antibodies against a single target are produced, and during another period of time perhaps the immune system may subside or produce many different antibodies against many different targets. These effects may or may not be observable in the standard immunosignature which examines specific highly reactive peptides. If the phenomenon is not observable, than the change in entropy may mostly occur in the low intensity range. If the phenomenon is observable, then the fluorescence intensity of the disease specific features may increase and decrease over time. Note that the situation of a transgenic mouse may also be different than a spontaneous tumor. The Center for Innovations in Medicine plans to investigate this in the future.

1.3.6 Human disease

Can the AbStat measurements distinguish between humans with and without disease on different non-natural sequence peptide array platforms? This question is addressed with data from several different experiments across two different platforms: the HT330K silicon wafer platform, and the CIM10K glass slide platform. For each experiment, several normal and disease samples were applied to the array and the quantitative values of all of the AbStat measures were obtained. These measures were then analyzed to determine whether they could distinguish the different groups. The expectation is that the AbStat measures should provide the ability to distinguish between healthy and disease sera because disease sera will contain a different number of antibodies, antibody affinities, and antibody avidities. Therefore, disease sera will bind to an array of non-natural sequence peptides in a different manner than normal sera. More resolution between normal and disease sera should be provided by the HT330K chip over the CIM10K chip since more peptides are present on the HT330K chip. If these expectations are met, then the AbStat measures will prove useful for monitoring and diagnosing the presence or

absence of disease. At this point, further investigations would be required to determine the exact nature of the disease.

1.3.6.1 HT330K first chip disease dataset

In order to address the issue of whether or not normal sera can be distinguished from disease sera using the broad AbStat measures, sera from patients with infectious diseases and sera from normal patients were applied to the HT330K platform type. The infectious disease samples could result in a higher or lower entropy fluorescence intensity distribution than normal samples. If the entropy is lower, this could indicate there is a high focused immune response against few targets for a certain type of infection. If the entropy is higher, then this infectious disease may result in more inflammation and a non-specific immune response with antibodies against many different targets. There may also be more antibody binding to the non-natural sequence peptide microarray in general during the course of an infection.

The following diseases were applied to the array: West Nile virus (WNV), syphilis (SYPH), malaria (MAL), hepatitis B virus (HBV), dengue (DEN), *Bordetella pertussis* (BPE), and *Borrelia* (BORR). The samples were applied to the array and the resulting image files were aligned by the Peptide Array Core under the direction of Zbigniew Cichacz. The samples were obtained from SeraCare Life Sciences, Milford, MA.

Before presenting the results of the AbStat analysis, an analysis based on the traditional approach of selecting specific peptides will be presented. Half of the normal and half of the disease samples were randomly selected as training samples, and the other half of the samples was used as test data. The intensity values of the top 100 most significant peptides by p-value from a t-test from the training data samples were used as input for one of three machine learning algorithms: Naïve Bayes, SVM, or J48graft tree. After the algorithm was trained, the algorithm was used to predict the classification of the test data samples. This same process was repeated with random classification assignments for each sample as disease or normal. Randomly assigned classes should result in a decrease in classification performance. The results of this analysis are presented in Table 1. All three classification algorithms can correctly classify the

samples as disease or normal with about 90% accuracy. Note that for this specific peptide approach the Center for Innovations in Medicine lab often uses Naïve Bayes as the classifier of choice, and Muskan Kukreja has demonstrated that this is the fastest algorithm which still results in acceptable performance [64]. With this particular dataset consisting of 118 samples with 86 disease and 32 normal, all of the algorithms completed in less than 1 second with a standard desktop computer.

Table 1 Machine learning statistics for first chip disease dataset using selection of specific peptides

Algorithm	Correctly Classified Instances	Kappa Statistic	ROC Area
Naïve Bayes	87.9	0.679	0.836
SVM	91.4	0.771	0.863
J48graft	89.7	0.741	0.871
Naïve Bayes Random	70.7	0.0199	0.528
SVM Random	60.3	-0.0537	0.475
J48graft Random	67.2	0.0072	0.503

Three different algorithms were used to predict the class (disease or normal) of samples applied to the HT330K microarrays. Half of the samples were randomly selected as training samples and half were randomly selected as test samples. The top 100 most significant features by p-value from a t-test with the training samples only were used to train the algorithms. The algorithms were then used to predict the class of the samples in the test set. Several machine learning statistics are presented in columns. In a separate analysis, samples were randomly assigned a class as disease or normal and the analysis was repeated.

This specific peptide approach is more susceptible to overtraining than the AbStat approach. For example, if the dataset is not split in half so that features are selected on the training data half and used against the test data half, undesirable results occur. If the top 100

most significant peptides by p-value from a t-test are selected from the whole dataset, and then these peptides are input into the Naïve Bayes classifier with 10-fold cross-validation, then the result is 93.2% correctly classified instances, 0.822 kappa statistic, and 0.934 ROC area. If the process is repeated with samples with randomly assigned class as disease or normal, then the result is 98.3% correctly classified instances, 0.957 kappa statistic, and 0.996 ROC area. The random dataset did not result in worse performance as it should have. The reason for this overtraining is likely that with so many features (330,000) and the presence of some degree of noise in the experimental data as well as some degree of similarity among any randomly chosen samples for some portion of the 330k peptides, there is bound to be at least a small percentage of features which can distinguish between any random grouping of the samples. Another reason for the overtraining with the whole sample set is that the p-value for each peptide is determined by comparing all the samples in both groups, whereas the features of the AbStat are not selected based on comparing both groups. Instead, all of the AbStat measures are calculated independently for each sample, and then these calculated values are input into the machine learning algorithm. The AbStat method is not as susceptible to overtraining, and performance that is no better than random chance results when the classes are randomly assigned to each sample, even when the whole dataset is used.

The AbStat analysis uses only broad global measures, and no specific peptides are selected. A box and dot plot of the entropy of all of the samples is presented in Figure 15. A heatmap of all of the measures for all of the samples is presented in Figure 16. The significance of the measures from a t-test comparing all disease against normal (Figure 17), the SVM weight of the measures (Figure 18), the J48graft tree (Figure 19), and machine learning statistics are also presented (Table 2). All of the diseases had a higher entropy average than the entropy average of the normal group with the exception of BPE (Figure 15). The heatmap of all of the measures reveals that many of the normal samples cluster together with a low entropy and a higher cv, normalized standard deviation, normalized mean, normalized 95th percentile, dynamic range, and normalized maximum than the infectious disease samples. Many of the West Nile virus samples also cluster together with values characteristic for that group for the same

measures and a very low kurtosis and skew (Figure 16). The entropy and kurtosis could distinguish between disease and normal the best by a t-test, and the max normalized and entropy were weighted the most by SVM (Figure 17 and Figure 18). The J48graft tree reveals that 12 samples were correctly classified as normal based on the criteria that their entropy was less than or equal to 7.63 (Figure 19). A total of 46 samples were classified as disease if they had a high entropy and high cv (entropy>7.63 and cv>1.47), and only 12 of these 46 were misclassified. The machine learning statistics illustrate that the machine learning algorithms can classify better than chance with information from the measures (about 80% accuracy for J48graft), but these algorithms can classify no better than chance when the class assignments are randomly assigned (Table 2).

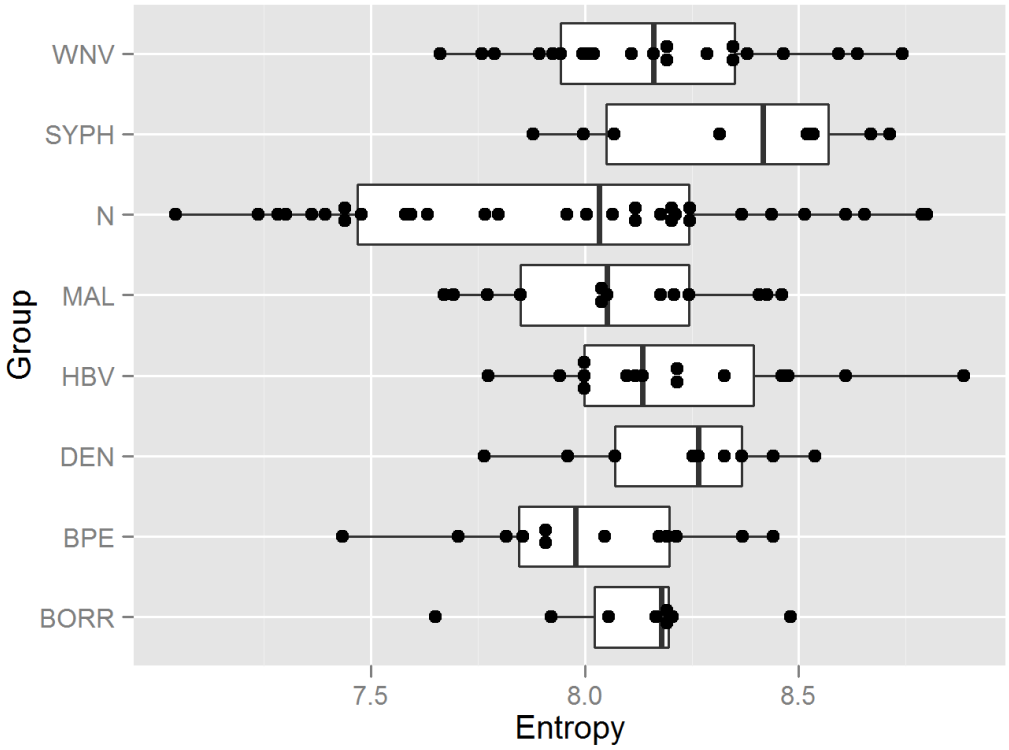


Figure 15 Box and dot plot of entropy for groups in first chip disease dataset

Sera samples were applied to HT330K microarrays. The entropy of each replicate was calculated and presented in a box and dot plot. The class of the sample is designated as follows:

N = normal, WNV = West Nile virus, SYPH = syphilis, MAL = malaria, HBV = hepatitis B virus, DEN = dengue, BPE = Bordetella pertussis, and BORR = Borrelia.

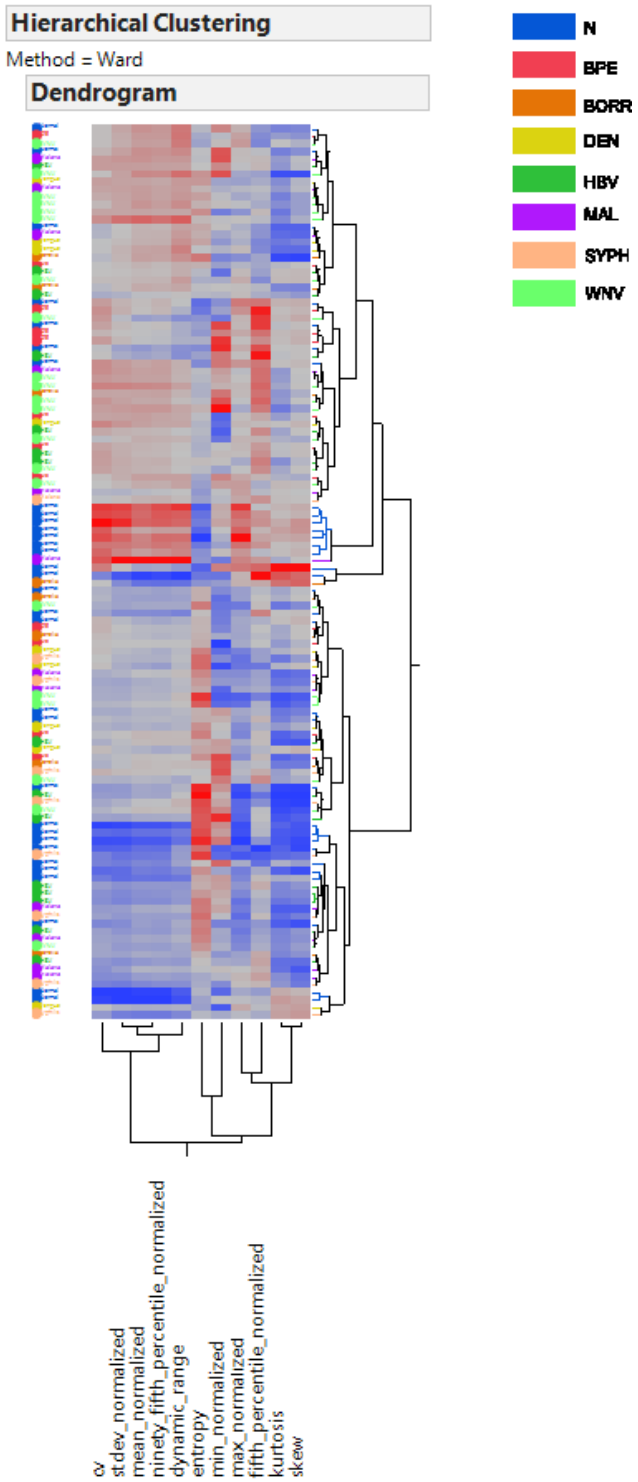


Figure 16 Heatmap of Measures for samples in first chip disease dataset

Each column corresponds to an AbStat measure and each row corresponds to a sera sample. The relative average value of each AbStat measure for the samples is represented by a color with blue indicating the lowest relative value and red indicating the highest relative value. The class of each sample is designated as follows: N = normal, WNV = West Nile virus, SYPH = syphilis, MAL = malaria, HBV = hepatitis B virus, DEN = dengue, BPE = Bordetella pertussis, and BORR = Borrelia.

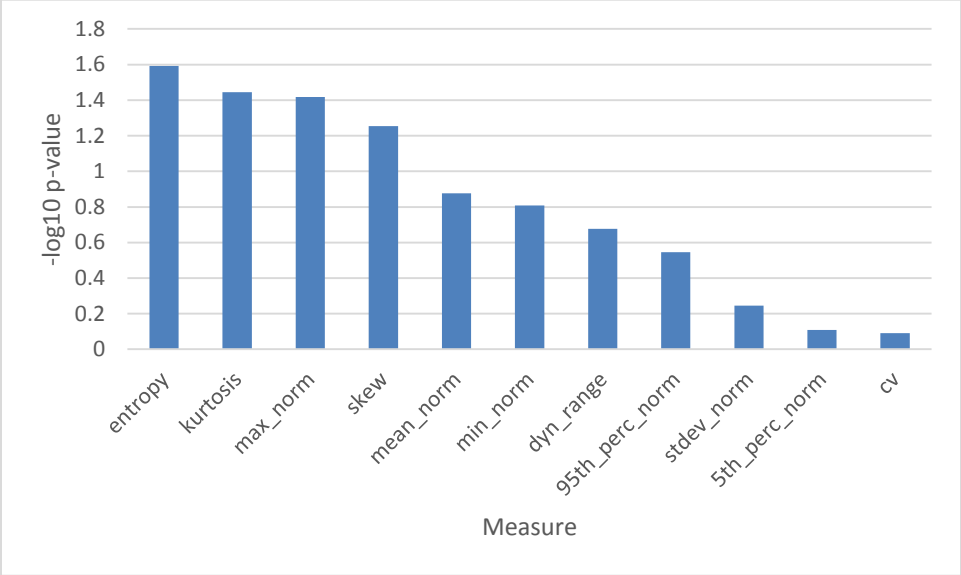


Figure 17 Statistical significance of measures comparing normal with disease in HT330K first chip disease dataset

Sera samples were applied to the HT330K array and the AbStat measures were calculated. A p-value from a t-test was then determined for each measure for disease vs normal. The negative logarithm in base 10 of the p-value was then plotted in a bar graph.

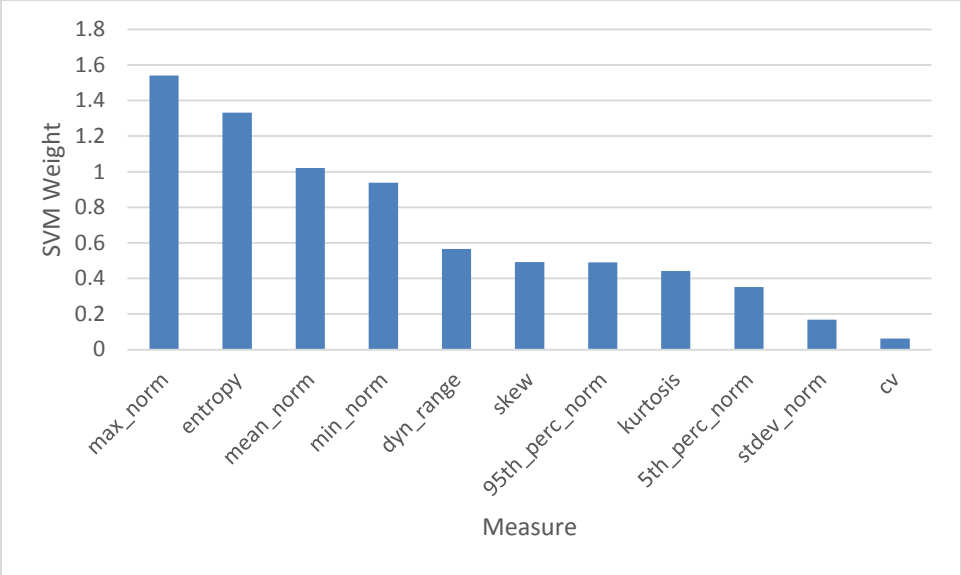


Figure 18 SVM weight of measures comparing normal with disease in first chip disease dataset

Sera samples were applied to the HT330K array and the AbStat measures were calculated. An SVM algorithm with 10-fold cross-validation was then used to predict the class of the samples as disease or normal. The absolute value of the weight of each measure assigned by the SVM was then plotted in a bar graph.

```
entropy <= 7.632743: N (12.0)
entropy > 7.632743
|   cv <= 1.466364: N (8.0)
|   cv > 1.466364: D (46.0/12.0)
```

Figure 19 J48graft tree for first chip disease dataset

Sera samples were applied to the HT330K array and the AbStat measures were calculated. A J48graft tree algorithm with 10-fold cross-validation was then used to predict the class of the samples as disease or normal. The algorithm then selected certain measures with specified cutoff points to construct a classification tree to assign a sample to the normal or disease group.

Table 2 Machine learning statistics for first chip disease dataset

Algorithm	Correctly Classified Instances	Kappa Statistic	ROC Area
SVM	62.1	0.234	0.616
J48graft	78.8	0.5706	0.744
SVM Random	48.8	-0.0351	0.483
J48graft Random	54.5	0.066	0.504

Two different algorithms with 10-fold cross-validation were used to predict the class (disease or normal) of samples applied to the HT330K microarrays. The attributes of each sample were the AbStat measures. Several machine learning statistics are presented in columns. In a separate analysis, samples were randomly assigned a class as disease or normal and the analysis was repeated.

These results demonstrate that it is possible to distinguish disease samples from normal samples better than chance using the AbStat measures. A classification accuracy of approximately 80% can be obtained with the J48graft algorithm. Although much better classification performance can be obtained by other methods, the value of the AbStat measures is that they do not rely on specific features. Instead, this classification performance is achieved from very broad global measures. Additionally, these measures display similar behavior for both disease and age. This is a fact which may be related to some aspect that connects these two different concepts.

Note that the AbStat metric can perform just as well using half of the dataset. Previously, I demonstrated that the specific peptide analysis method is more susceptible to overtraining. Therefore, instead of performing a t-test with all of the samples, and then using the values of the top 100 peptides as input into a classification algorithm with 10-fold cross-validation, a different approach had to be taken. The top 100 peptides from a t-test with half of the samples had to be selected, and then these peptides were input into a classification algorithm for training to then predict the test samples. This approach resulted in the decrease in performance expected with random class assignments for each sample. The AbStat did not suffer from this overtraining,

primarily because all of the AbStat measures are used, and the measures are not selected by comparing them with other samples as with a t-test. However, one can ask the question, does the AbStat perform just as well if data from half of the samples are input into the algorithms and then these algorithms are used to classify the other half, instead of using 10-fold cross-validation? To answer this question half of the samples were randomly assigned to the training set, input into a classification algorithm, and then used to predict the class of the test samples. The same process was repeated with random class assignment samples to see if the expected decrease in classification performance was observed. The results in Table 3 demonstrate that the AbStat method does indeed perform just as well with half the sample set as it does with 10-fold cross-validation. Note that in this particular dataset there were 86 disease samples and 33 normal samples so simply predicting every sample as disease results in 86/(33+86)*100= 72.3% correctly classified instances. Therefore, it is important to look at the kappa statistic and ROC area when evaluating the outcomes in Table 3.

Table 3 Machine learning statistics for first chip disease dataset with half samples as training rather than 10-fold cross-validation

Algorithm	Correctly Classified Instances	Kappa Statistic	ROC Area
Naïve Bayes	82.8	0.551	0.832
SVM	74.1	0.135	0.551
J48Graft	86.2	0.626	0.778
Naïve Bayes Random	51.7	0.077	0.534
SVM Random	72.4	0	0.5
J48Graft Random	65.5	-0.0221	0.388

Three different algorithms were used to predict the class (disease or normal) of samples applied to the HT330K microarrays. The samples were randomly split into training samples or test samples. The attributes of each sample were the AbStat measures. Each algorithm was trained

with the training set and then tested on the other samples. Several machine learning statistics are presented in columns. In a separate analysis, samples were randomly assigned a class as disease or normal and the analysis was repeated.

1.3.6.2 HT330K wafer 46

In the last section, a variety of infectious disease sera samples were applied to the peptide array. In this section two different monoclonal antibodies, pure buffer, and chronic diseases were applied to the peptide array. The chronic diseases were breast cancer and multiple myeloma. The fundamental question addressed was still the same: can AbStat be used to distinguish between disease and normal sera? This dataset may be particularly good for addressing this issue since several researchers with hands on experience with the different HT330K wafers had the opinion that wafer 46 was one of the highest quality wafers that yielded quality data. The expected results are that the chronic disease samples would have a higher entropy than normal, and normal would have an even higher entropy than monoclonal antibodies. Normal sera should have a higher entropy than monoclonal antibodies because normal sera is composed of a more complicated mixture of antibodies than a single monoclonal antibody even though normal sera does not generally have high binding. The differences in the AbStat measures observed should allow for classification between normal and disease sera.

The experiments were performed by the Peptide Array Core under the direction of Zbigniew Cichacz. Two different monoclonal antibodies (Ab1 and Ab2), buffer alone (B), sera from normal donors (N), sera from patients with breast cancer (BC), and sera from patients with multiple myeloma (MM) were applied to 330k peptide arrays. The breast cancer samples were obtained from Dr. Adi Gazday from UT Southwestern, Dallas, TX. The myeloma samples were obtained from Dr. Robert Penny from the Multiple Myeloma Research Foundation, Norwalk, CT. A specific peptide analysis was performed by splitting the dataset in half, choosing the top 100 peptides by p-value from a t-test from the training samples, training a naïve Bayes algorithm, and testing with the testing samples. The results were 85% correctly classified instances, 0.634 kappa statistic, and 0.956 ROC Area. Higher performance would likely be achieved with a larger

dataset since this dataset only had 15 disease samples (5 breast cancer, and 10 multiple myeloma).

The dataset was then subjected to an AbStat analysis. A box and dot plot of the entropy for the groups is presented in Figure 20, and a heatmap of all of the measures for each sample is presented in Figure 21. The statistical significance of each measure when comparing disease and normal is presented in Figure 22. Machine learning was also used to classify samples as disease or normal. The SVM weight of each measure is presented in Figure 23, the J48graft tree is presented in Figure 24, and machine learning statistics for actual and random class assignments are presented in Table 4. In Table 4, the random class assignment test was performed as follows: the random class assignments were made, the values of the attributes and fake class of each sample was input into a machine learning algorithm for training and testing with 10-fold cross-validation, and this process was repeated four more times. The average value plus or minus one standard deviation for each machine learning statistic for the random class assignment test is presented in the table.

The box and dot plot of the entropy shows that the chronic disease groups have a significantly higher entropy than the normal group (p-value of 0.0167 for breast cancer and p-value of 0.00186 for multiple myeloma). The buffer and monoclonal antibody groups on the other hand have a lower entropy (Figure 20). The heatmap of all of the measures with each sample shows that samples of the same group generally cluster together (Figure 21). The multiple myeloma samples have a very high relative entropy and a low normalized 5th percentile. The breast cancer samples have a relatively high entropy, mean, median, 5th percentile, min, standard deviation, and 95th percentile. The monoclonal antibodies and the normal samples are relatively low for all of these measures, but higher for measures to the right of these in the heatmap. The statistical significance from a t-test reveals that entropy and the normalized maximum are the best measures (Figure 22), and the normalized 5th percentile and entropy have the greatest SVM weights (Figure 23). The J48graft tree reveals that 24 samples were classified as normal if the entropy was less than or equal to 9.22, and only 2 of these samples were misclassified. The machine learning algorithms demonstrate that these measures allow about

80% of the samples to be correctly classified with an SVM, and the algorithms perform no better than chance when random class assignments are made (Table 4).

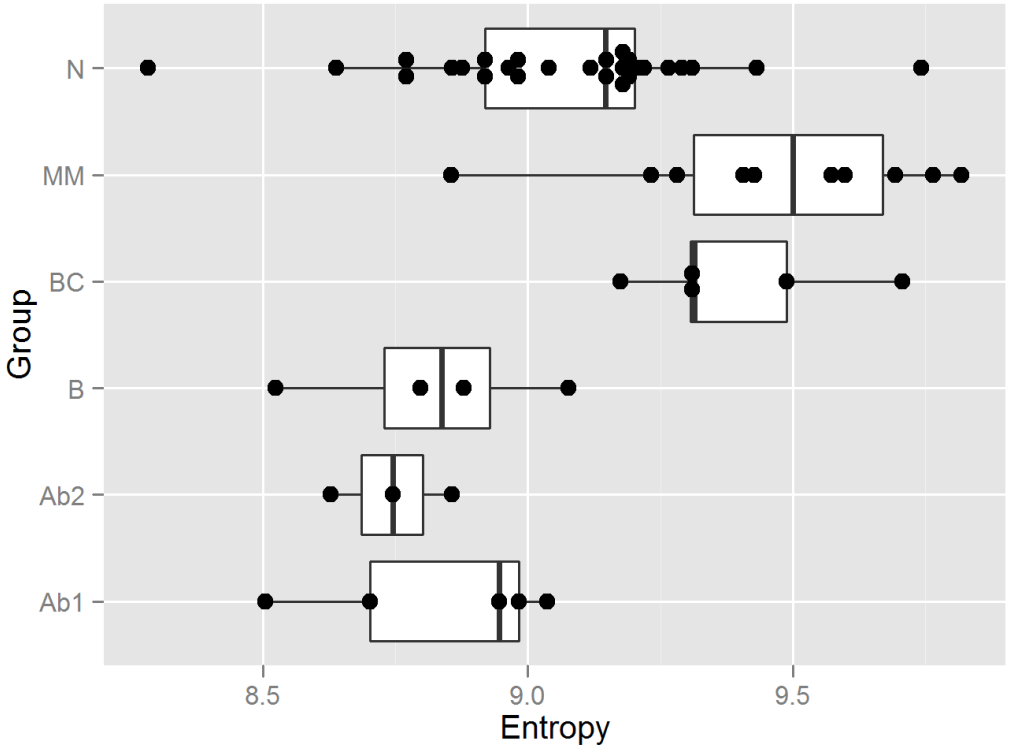


Figure 20 Box and dot plot of entropy for groups on wafer 46

Sera samples were applied to HT330K microarrays. The entropy of each replicate was calculated and presented in a box and dot plot. The class of the sample is designated as follows: N = normal, MM = multiple myeloma, BC = breast cancer, B = buffer, Ab1 = antibody 1 against p53 epitope RHSVV, Ab2 = Ab2 against p53 epitope SDLWKL

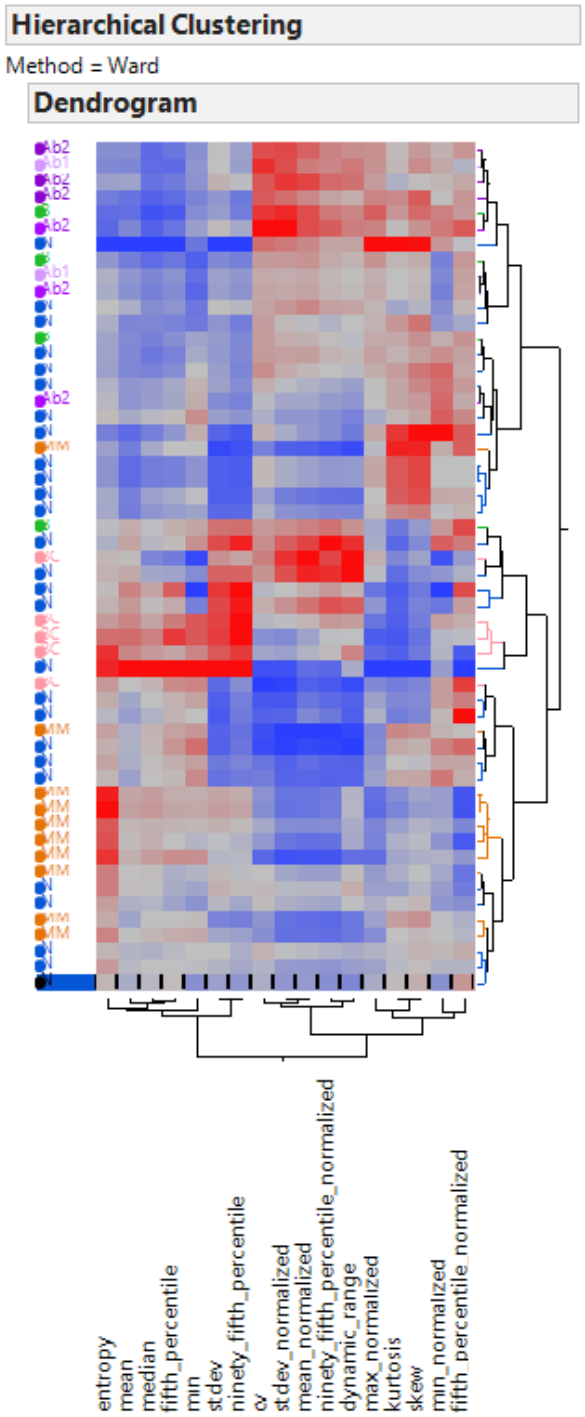


Figure 21 Heatmap of Measures for samples on wafer 46

Each column corresponds to an AbStat measure and each row corresponds to a sera sample. The relative average value of each AbStat measure for the samples is represented by a color with blue indicating the lowest relative value and red indicating the highest relative value. The class of

each sample is designated as follows: N = normal, MM = multiple myeloma, BC = breast cancer, B = buffer, Ab1 = antibody 1 against p53 epitope RHSVV, Ab2 = antibody 2 against p53 epitope SDLWKL.

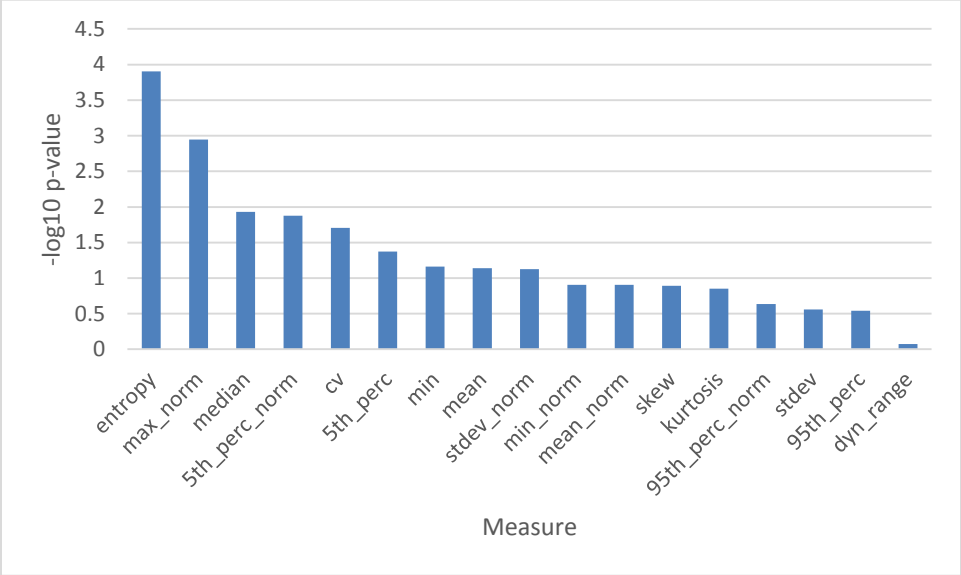


Figure 22 Statistical significance of measures comparing normal with disease for wafer 46

Sera samples were applied to the HT330K array and the AbStat measures were calculated. A p-value from a t-test was then determined for each measure for disease vs normal. The negative logarithm in base 10 of the p-value was then plotted in a bar graph.

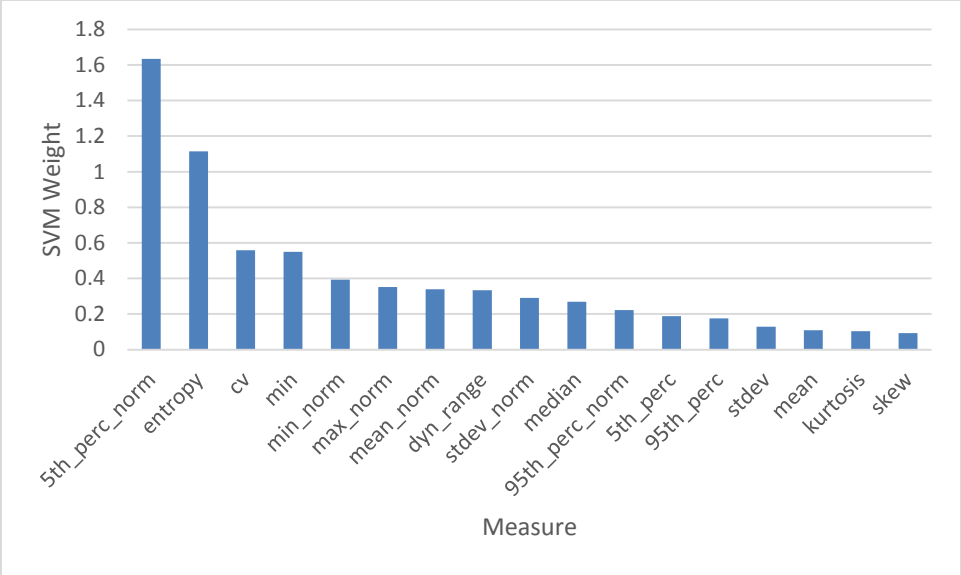


Figure 23 SVM weight of measures comparing normal with disease for wafer 46

Sera samples were applied to the HT330K array and the AbStat measures were calculated. An SVM algorithm with 10-fold cross-validation was then used to predict the class of the samples as disease or normal. The absolute value of the weight of each measure assigned by the SVM was then plotted in a bar graph.

```
entropy <= 9.21943: N (24.0/2.0)
entropy > 9.21943
|  median <= 3951
|  |  min_normalized <= 0.210849
|  |  |  cv <= 1.417799
|  |  |  |  mean <= 5668.328085: N (0.0|13.0/1.0)
|  |  |  |  mean > 5668.328085
|  |  |  |  |  max_normalized <= 15.899668: D (0.0|12.0/1.0)
|  |  |  |  |  max_normalized > 15.899668: N (6.0/2.0)
|  |  |  cv > 1.417799: N (0.0|14.0/1.0)
|  |  min_normalized > 0.210849: N (0.0|9.0)
|  median > 3951: D (12.0/1.0)
```

Figure 24 J48graft tree for wafer 46

Sera samples were applied to the HT330K array and the AbStat measures were calculated. A J48graft tree algorithm with 10-fold cross-validation was then used to predict the class of the samples as disease or normal. The algorithm then selected certain measures with specified cutoff points to construct a classification tree to assign a sample to the normal or disease group.

Table 4 Machine learning statistics for wafer 46

Algorithm	Correctly Classified Instances	Kappa Statistic	ROC Area
SVM	78.6	0.496	0.73
J48graft	66.7	0.274	0.559
SVM Random	62.9 +/-2.74	-0.00742 +/-0.0928	0.498 +/-0.0388
J48graft Random	57.6 +/-4.57	-0.109 +/-0.102	0.354 +/-0.0432

Two different algorithms with 10-fold cross-validation were used to predict the class (disease or normal) of samples applied to the HT330K microarrays. The attributes of each sample were the AbStat measures. Several machine learning statistics are presented in columns. In a separate analysis, samples were randomly assigned a class as disease or normal five different times and the analysis was repeated. The average plus or minus one standard deviation of each machine learning statistic is presented.

The results of the chronic disease samples applied to the HT330K platform matched the expected outcome even though this outcome was opposite that of some infectious diseases and vaccines. The chronic disease samples had a higher entropy than normal samples. Additionally, the monoclonal antibody samples exhibited an even lower entropy than the normal sera samples. Although some of the infectious diseases and vaccines exhibited lower entropy in different experiments, I expected that chronic disease samples would have a higher entropy. Infectious diseases and vaccines may produce high concentrations of antibodies that bind to a few targets resulting in a large and sudden "focused" immune response. A chronic disease on the other hand develops over a long period of time, and the immune response often fails to overcome the challenge. In this situation, the immune system may produce a wide variety of many different antibodies against various targets with lower affinity than might be seen with an infectious disease. An infectious disease may mimic the situation of adding monoclonal antibody to sera ("1.3.1.2 Spiking antibody into sera"), and a chronic disease may mimic the situation of mixing many different antibodies together. Regardless of whether or not these speculations are correct,

the differences among all of the AbStat measures allowed for the correct classification of about 80% of the normal and disease samples. These results provide further evidence that the AbStat measures provide useful information about human disease states.

1.3.6.3 CIM10K

In the last two sections, disease samples were applied to the HT330K platform, but this section aims to address the issue of whether or not disease samples can be distinguished with the AbStat measure on the CIM10K platform with fewer peptides. The disease samples would still be expected to have a higher entropy than the normal samples, unless the disease typically causes a very strong specific immune response. The CIM10K platform should also still provide the resolution necessary to distinguish samples since 10,000 peptides is still a fairly large number of peptides as compared to a few hundred, but the CIM10K platform would be expected to provide less resolution than the HT330K platform. Note that not only are there a different number of peptides on the two platforms, but a slide of each platform type is manufactured using drastically different techniques. Therefore, positive results would indicate that the AbStat measures are robust and can work with peptide arrays produced by a variety of techniques.

The experiments were performed by the Peptide Array Core under the direction of Zbigniew Cichacz. Sera obtained from Lawrence Livermore National Laboratory from patients with West Nile virus (WNV), syphilis (SYPH), hepatitis B virus (HBV), dengue (DEN), *Bordetella pertussis* (BPE), and *Borrelia* (BORR) were applied to the CIM10Kv2 non-natural sequence peptide array. An analysis using specific peptides was performed by splitting the dataset into a training set and a test set. The top 100 peptides by p-value from a t-test with the normal and disease samples in the training set were used as input to a naïve Bayes classifier, and this classifier was used to classify the test set samples. The results were 89.7% correctly classified instances, 0.608 kappa statistic, and 0.899 ROC Area.

I also performed an AbStat analysis. A box and dot plot of the entropy of the different groups (Figure 25), a heatmap of the measures for all of the samples (Figure 26), the statistical significance of the measures from a t-test (Figure 27), the SVM weight of the measures (Figure

28), the J48graft tree of the measures (Figure 29), and a table of the machine learning statistics (Table 5) is presented below. Every disease group had a higher entropy than the normal group with the exception of DEN and BPE (Figure 25). The heatmap of the measures for all of the samples revealed that many of the groups clustered together indicating that some samples within a group have similar values for the different measures (Figure 26). The statistical significance from a t-test showed that the 5th percentile, mean, 95th percentile, median, and skew measures were the best at distinguishing the groups (Figure 27), and the SVM weights of the measures show that the coefficient of variation, standard deviation, 95th percentile, entropy, and skew had the most weight (Figure 28). The J48 graft tree indicates that thirty samples were classified as normal if the dynamic range was less than or equal to 12.6 and the 5^{th} percentile was less than or equal to 678, and eight of these samples were misclassified. Additionally, five samples were classified as disease if the entropy was greater than 7.70 along with the following criteria: $574<5^{th}$ percentile<=678, and dynamic range>12.6. One of these samples was misclassified. The machine learning statistics demonstrate that an SVM can classify about 60% of the instances correctly, and the classification algorithms perform no better than chance when the class of each sample is assigned randomly (Table 5).

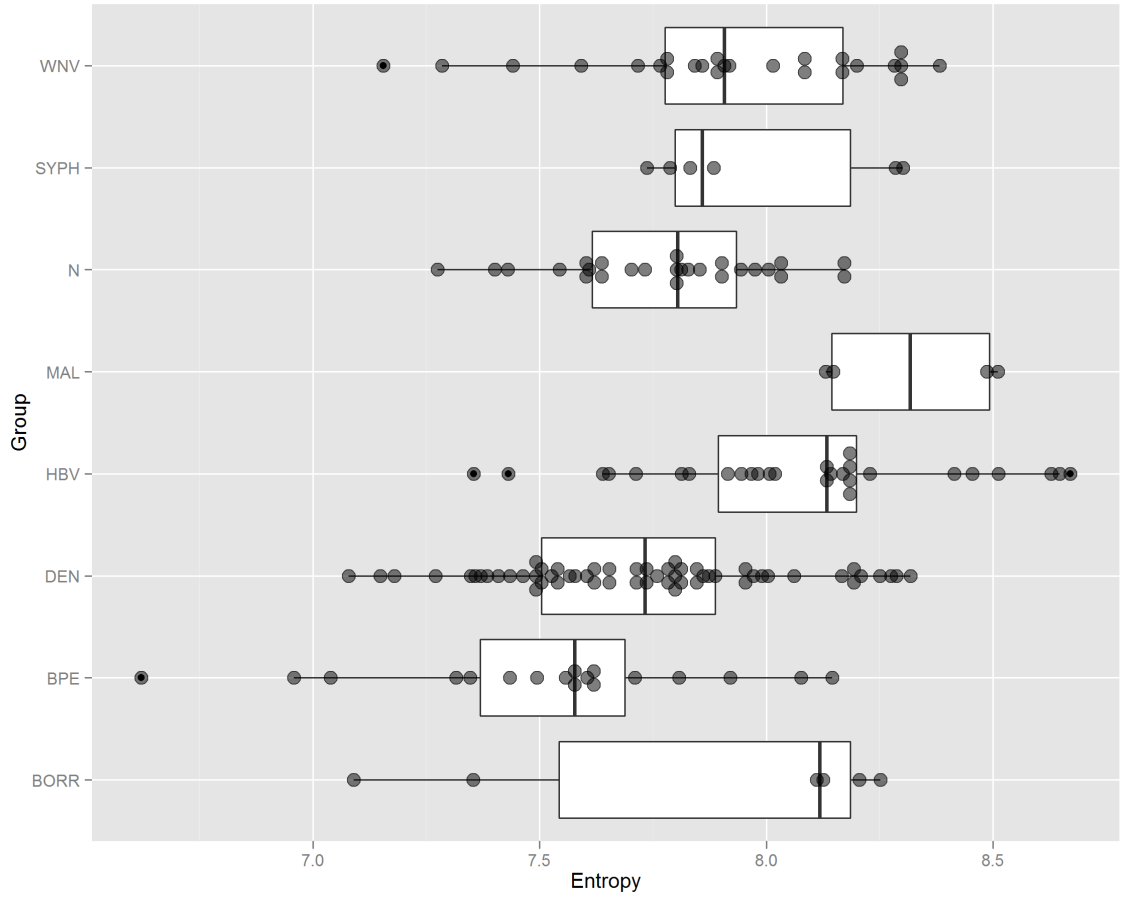


Figure 25 Box and dot plot of entropy for groups on CIM10K

Sera samples were applied to CIM10K microarrays. The entropy of each replicate was calculated and presented in a box and dot plot. The class of the sample is designated as follows: N = normal, WNV = West Nile virus, SYPH = syphilis, MAL = malaria, HBV = hepatitis B virus, DEN = dengue, BPE = Bordetella pertussis, BORR = Borrelia.

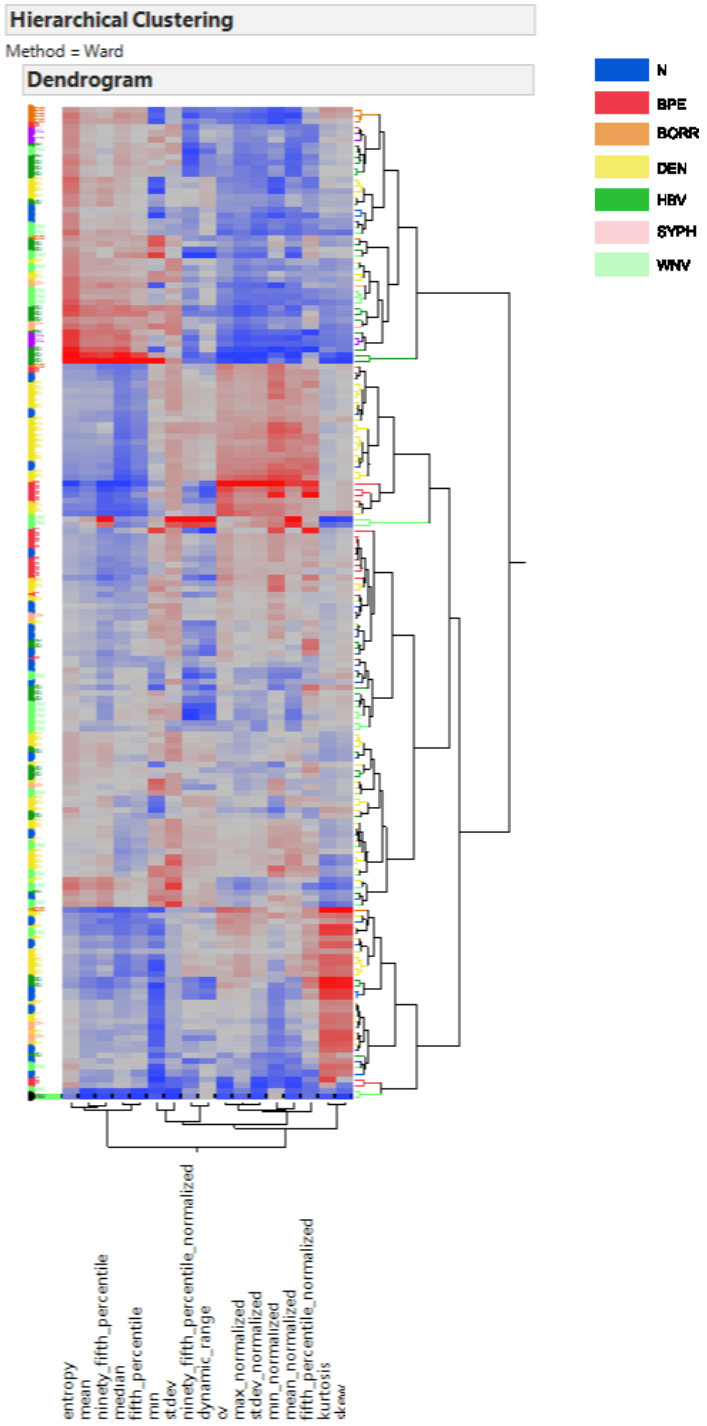


Figure 26 Heatmap of Measures for samples on CIM10K

Each column corresponds to an AbStat measure and each row corresponds to a sera sample. The relative average value of each AbStat measure for the samples is represented by a color with blue indicating the lowest relative value and red indicating the highest relative value. The class of

each sample is designated as follows: N = normal, WNV = West Nile virus, SYPH = syphilis, MAL = malaria, HBV = hepatitis B virus, DEN = dengue, BPE = Bordetella pertussis, BORR = Borrelia.

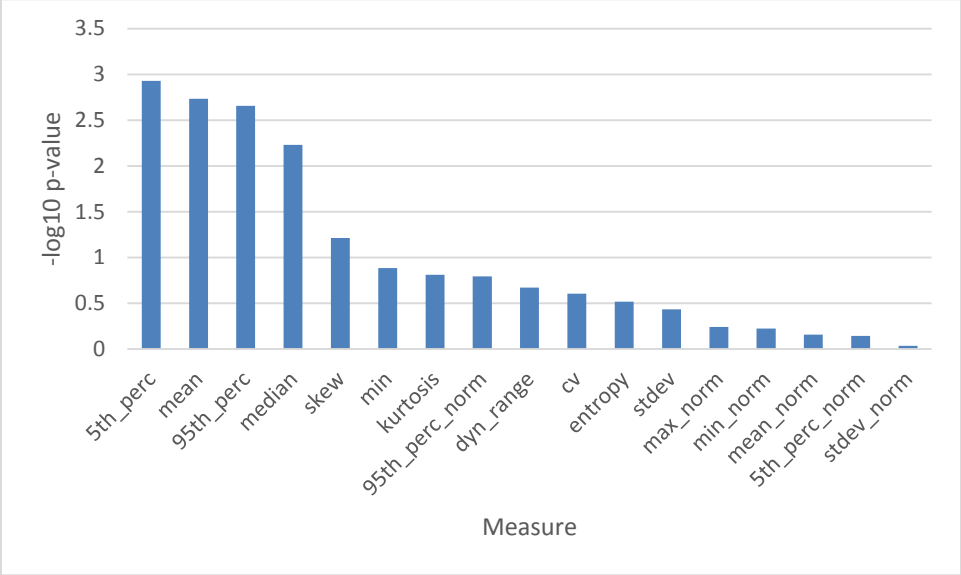


Figure 27 Statistical significance of measures comparing normal with disease for CIM10K

Sera samples were applied to the CIM10K array and the AbStat measures were calculated. A p-value from a t-test was then determined for each measure for disease vs normal. The negative logarithm in base 10 of the p-value was then plotted in a bar graph.

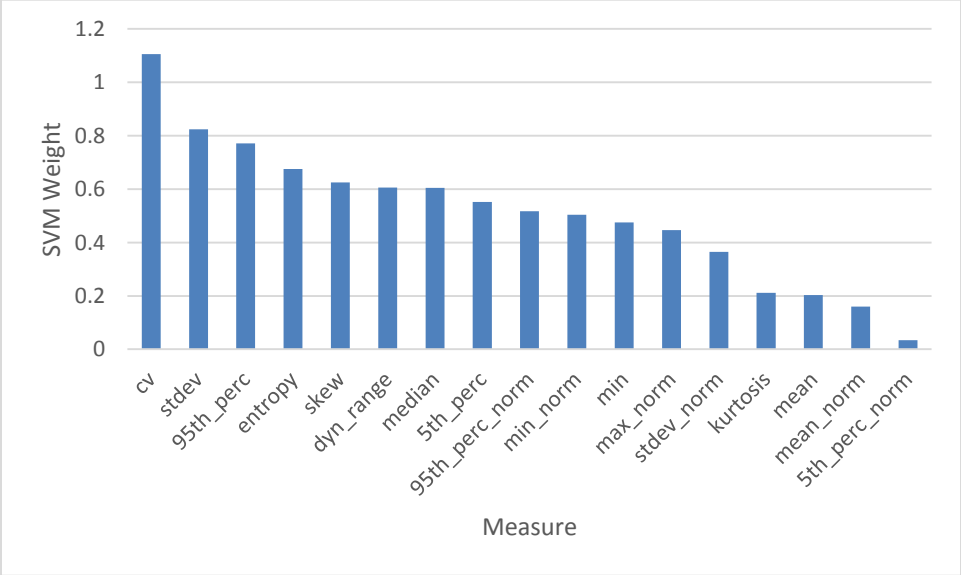


Figure 28 SVM weight of measures comparing normal with disease for CIM10K

Sera samples were applied to the CIM10K array and the AbStat measures were calculated. An SVM algorithm with 10-fold cross-validation was then used to predict the class of the samples as disease or normal. The absolute value of the weight of each measure assigned by the SVM was then plotted in a bar graph.

```
fifth_percentile <= 678
|   dynamic_range <= 12.636133: N (30.0/8.0)
|   dynamic_range > 12.636133
|   |   entropy <= 7.702289: D (6.0)
|   |   entropy > 7.702289
|   |   |   fifth_percentile <= 574: N (3.0)
|   |   |   fifth_percentile > 574: D (5.0/1.0)
fifth_percentile > 678: D (8.0)
```

Figure 29 J48graft tree for CIM10K

Sera samples were applied to the CIM10K array and the AbStat measures were calculated. A J48graft tree algorithm with 10-fold cross-validation was then used to predict the class of the samples as disease or normal. The algorithm then selected certain measures with specified cutoff points to construct a classification tree to assign a sample to the normal or disease group.

Table 5 Machine learning statistics for CIM10K

Algorithm	Correctly Classified Instances	Kappa Statistic	ROC Area
SVM	59.6	0.192	0.596
J48graft	57.7	0.154	0.592
SVM Random	51.9	0.0299	0.515
J48graft Random	48.1	-0.0385	0.469

Two different algorithms with 10-fold cross-validation were used to predict the class (disease or normal) of samples applied to the CIM10K microarrays. The attributes of each sample were the AbStat measures. Several machine learning statistics are presented in columns. In a separate analysis, samples were randomly assigned a class as disease or normal and the analysis was repeated.

The AbStat measures were able to distinguish normal and disease samples better than chance with a classification accuracy of about 60%. This classification was less accurate than the classification accuracy that was obtained from both HT330K datasets which was about 80%. The increased number of peptides, and superior technology used to produce the HT330K slides likely allowed for the increased classification performance. Nevertheless, the AbStat measures were robust enough to perform better than chance on a variety of platforms. The performance of the AbStat measures could possibly be increased further with more peptides and enhanced manufacturing techniques that produce higher quality slides.

1.3.6.4 Alzheimer's disease

Can the AbStat measurement distinguish between aged humans with Alzheimer's and aged human controls? The characteristic of the Alzheimer's disease is quite different from the previous diseases analyzed up to this point. Alzheimer's affects the brain and there is a blood brain barrier which separates the brain from the rest of the bloodstream. Many researchers also do not associate Alzheimer's with the immune system as they do for many cancers and infectious diseases. Nevertheless, there is evidence that the immune system interacts with Alzheimer's and even produce antibodies that target antigens associated with the disease [76]. If the AbStat measures could distinguish between normal and Alzheimer's, then this could provide another tool for monitoring and diagnosing the disease. The expectation is that the AbStat measures should be able to distinguish normal and disease samples if there are enough antibodies with varying affinities and avidities that are different from normal samples which can be detected on a global level with the peptide array.

The Alzheimer's experiments were performed by Lucas Restrepo. Sera from patients with Alzheimer's disease (AD) and sera from patients classified as "non-cognitive impairment" (NC) were applied to the CIM10Kv2 arrays. All of the samples were either obtained from a brain-bank program at Banner's Sun Health Research Institute (Phoenix, AZ) or UT Southwest Medical Center (Dallas, TX) [76]. The NC patients are individuals of approximately the same age as the AD patients, but the NC patients have not been diagnosed with Alzheimer's disease. Lucas Restrepo

analyzed this data and found that he was able to correctly classify 75% of 8 samples in one experiment, and 100% of 24 samples in another experiment using a technique based on the selection of specific array features [76]. I performed another peptide specific analysis as well with 98 Alzheimer's samples and 98 normal samples. The samples were randomly split in half into a training set and a test set. The top 100 peptides from a t-test with the training set were used as input into classification algorithms. These classification algorithms were used to predict the class of the test set. The same process was repeated using random class assignments. The results in Table 6 show that using a specific peptide analysis does result in a better classification performance with the real sample IDs than can be achieved with random class IDs.

Table 6 Machine learning statistics for Alzheimer's disease with specific peptide analysis

Algorithm	Correctly Classified Instances	Kappa Statistic	ROC Area
Naïve Bayes	62.2	0.245	0.646
SVM	61.2	0.225	0.612
J48graft	58.2	0.163	0.588
Naïve Bayes Random	59.2	0.157	0.574
SVM Random	55.1	0.125	0.564
J48Graft Random	47.9	-0.0382	0.482

Three different algorithms were used to predict the class (disease or normal) of samples applied to the CIM10Kv2 microarrays. Half of the samples were randomly selected as training samples and half were randomly selected as test samples. The top 100 most significant features by p-value from a t-test between disease and normal with the training samples only were used to train the algorithms. The algorithms were then used to predict the class of the test set. Several machine learning statistics are presented in columns. In a separate analysis, samples were randomly assigned a class as disease or normal and the analysis was repeated.

Using the AbStat measures, which do not select specific array features, I analyzed this Alzheimer's dataset in the same manner as the previous human disease datasets to answer the question of whether the classes could be distinguished. A box and dot plot of the entropy for the NC and AD group (Figure 30), the statistical significance of the measures from a t-test (Figure 31), and the machine learning statistics are presented (Table 7). The p-values obtained for this dataset (Figure 31) were not as significant as those obtained for the other human disease datasets (Figure 17, Figure 22, and Figure 27). The machine learning algorithms were also incapable of classifying the samples any better than expected by chance (Table 7).

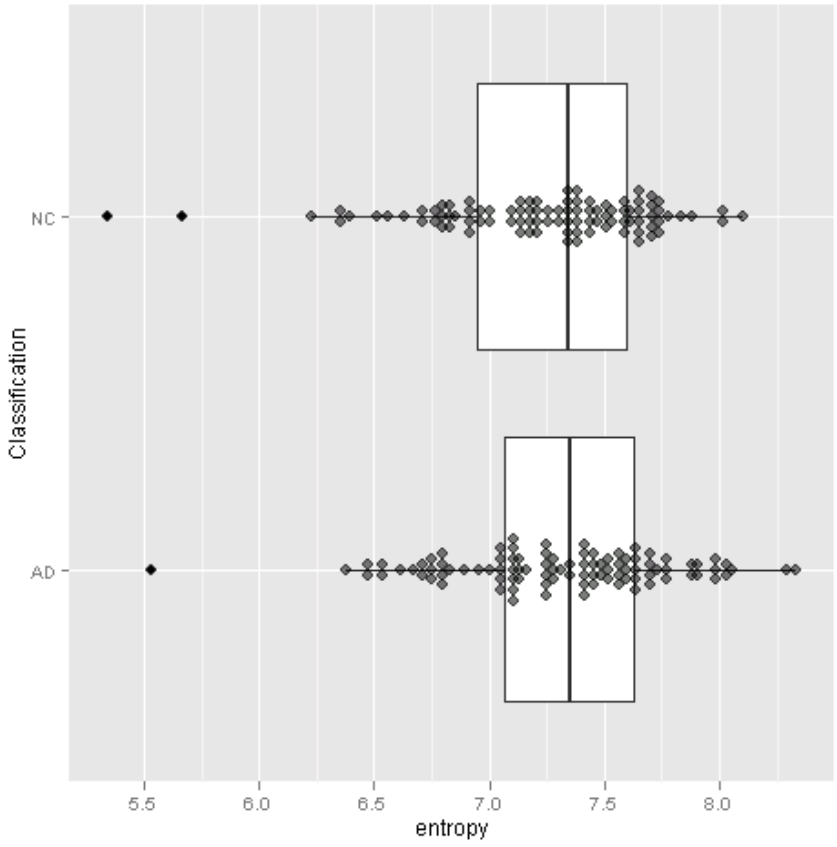


Figure 30 Box and dot plot of entropy for groups for Alzheimer's disease

Sera samples were applied to CIM10K microarrays. The entropy of each replicate was calculated and presented in a box and dot plot. The class of the sample is designated as follows: AD = Alzheimer's disease, NC = non-cognitive impairment control

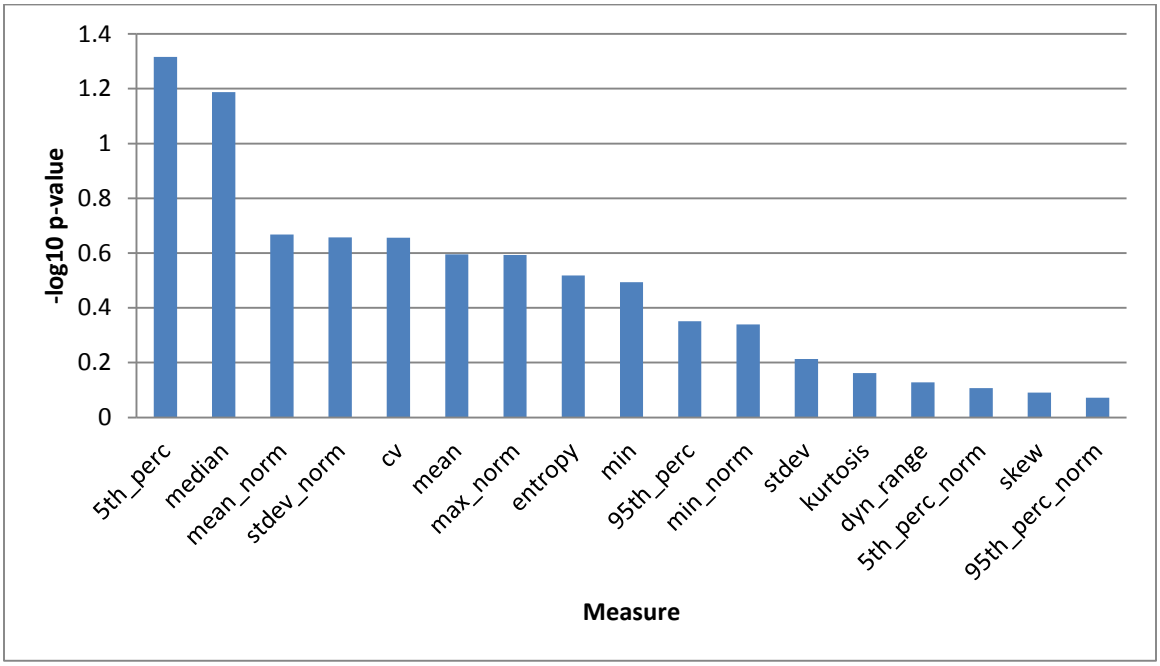


Figure 31 Statistical significance of measures comparing normal with Alzheimer's disease

Sera samples from Alzheimer's patients or controls were applied to the CIM10K array and the AbStat measures were calculated. A p-value from a t-test was then determined for each measure for Alzheimer's vs non-cognitive impairment control. The negative logarithm in base 10 of the p-value was then plotted in a bar graph.

Table 7 Machine learning statistics for Alzheimer's disease

Algorithm	Correctly Classified Instances	Kappa Statistic	ROC Area
SVM	55.2	0.104	0.552
J48graft	47.4	-0.0521	0.472

Two different algorithms with 10-fold cross-validation were used to predict the class (Alzheimer's disease or non-cognitive impairment control) of samples applied to the CIM10K microarrays. The attributes of each sample were the AbStat measures. Several machine learning statistics are presented in columns.

The AbStat measures were not able to distinguish between the controls and the Alzheimer's samples any better than chance. The entropy values were also not significantly

higher with the Alzheimer's samples as compared to the controls. This result certainly does not imply that the immune system is not involved in Alzheimer's disease. There can still be antibodies against selected peptides on the array without providing enough of a global shift in fluorescence intensity for the AbStat measures to detect. Nevertheless, this dataset illustrates some of the limitations of the AbStat measures since normal immunosignaturing analysis techniques are capable of distinguishing the samples. Note that the fact that both the Alzheimer's samples and the normal samples were from elderly individuals may have complicated the AbStat analysis since the AbStat is affected by age.

1.3.7 Changes with age

Can the AbStat measurement be used to distinguish between young and aged organisms? Since there are changes in the antibody repertoire with age, the AbStat measures may be able to capture this change. Additionally, there may be some similarities in many different individual aged samples vs young samples. For example, there is a general trend of loss of antibody specificity and affinity with age, and this change in the antibody repertoire should be detectable with the AbStat measures. One would expect the entropy of the antibody repertoire to increase with age. A validation of this prediction would connect the aging process to an increase in non-specific binding of the antibody repertoire, and provide a means to quantitatively monitor increases and decreases in health.

1.3.7.1 Changes with age in mice

Before investigating human sera, young and aged mice were compared. Although mice only live a little over 2 years, they still undergo a gradual decline of health and die of old age. The anticipated result is that aged mice would have higher entropy fluorescence intensity distributions than young mice.

Bart Legutki applied sera from seven young mice and ten aged mice to the 10k arrays. The aged mice were one year and two months old. The young mice were 6-8 weeks old. All of the mice were infected with 1*10^4 pfu (plaque forming units) of live attenuated PR8 influenza

virus, and blood was collected 40 days after infection. All of the young samples were pooled together, and all of the aged samples were pooled together. These two samples were applied to the 10k arrays with four replicates each. The average entropy was higher for the aged group than the young group (Figure 32). The young and aged samples also cluster together in a heatmap, and the aged samples have generally higher values for most of the measures (Figure 33). The most significant measures that distinguish the two groups in a t-test were the normalized mean and normalized 95th percentile (Figure 34). These trends reflect the broader fluorescence intensity distribution in the aged samples compared to the young samples, as illustrated with one aged and one young example sample (Figure 35). The sample number is too small to apply any machine learning techniques.

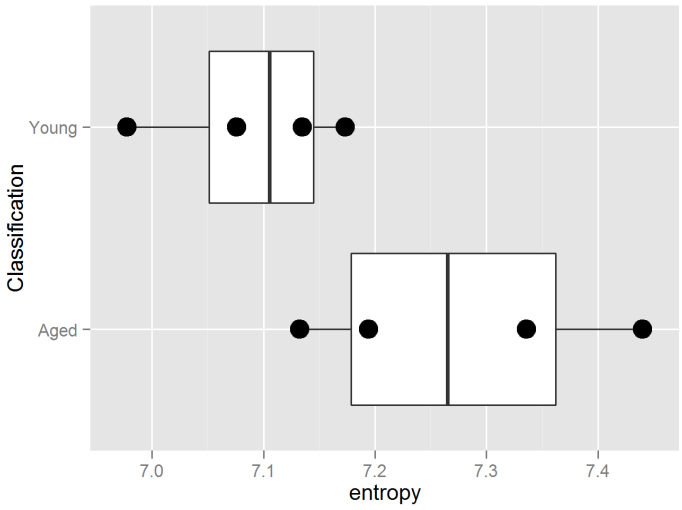


Figure 32 Box and dot plot of entropy for young and aged mice

Four replicates of a pool of seven young mice (6-8 weeks old) and four replicates of a pool of ten aged mice (one year and two months old) were applied to the array and the entropy from each slide was determined.

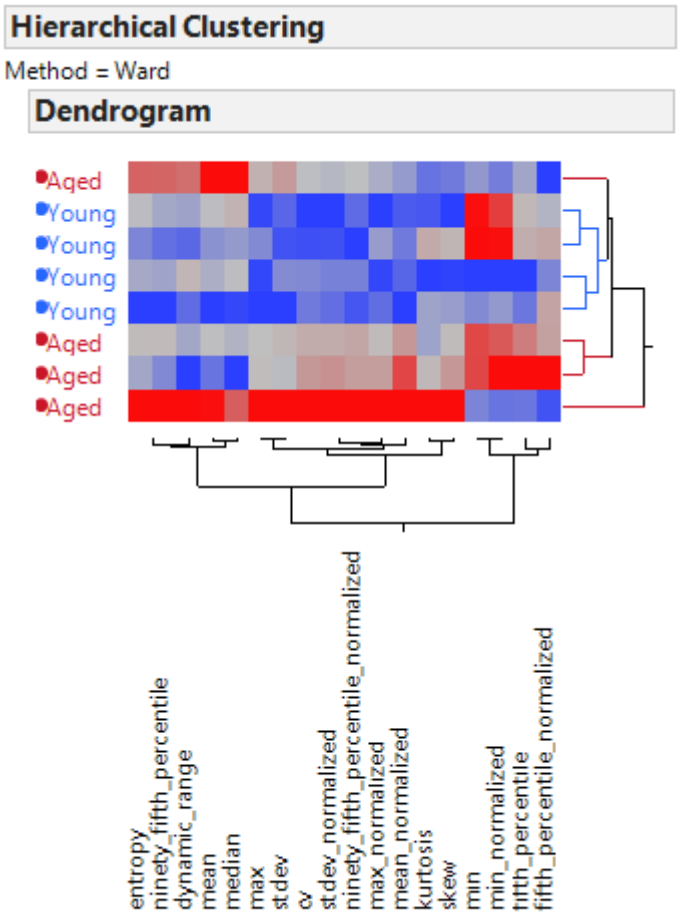


Figure 33 Heatmap of measures for young and aged mice

Four replicates of a pool of seven young mice (6-8 weeks old) and four replicates of a pool of ten aged mice (one year and two months old) were applied to the array and the AbStat values from each slide was determined. The relative average value of each AbStat measure for the samples is represented by a color with blue indicating the lowest relative value and red indicating the highest relative value.

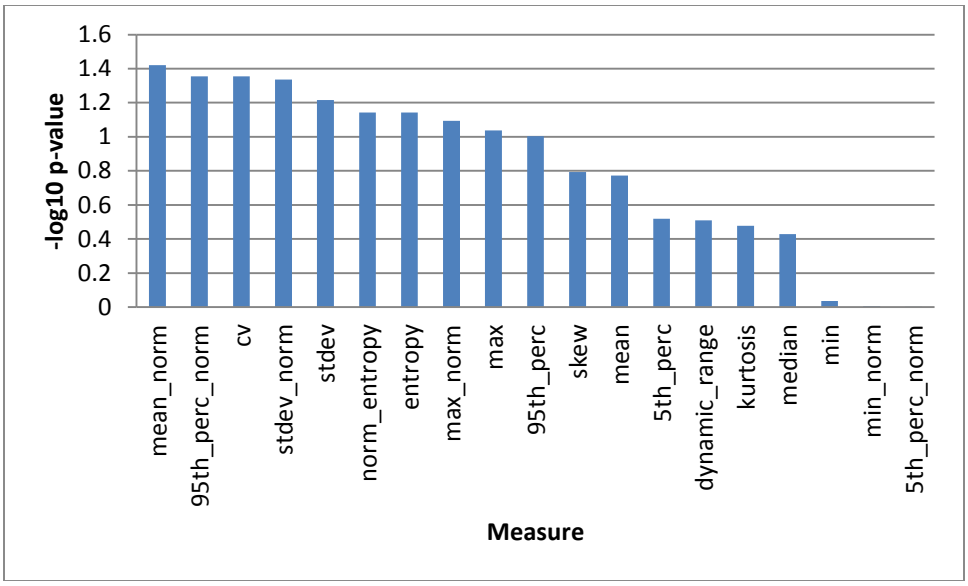
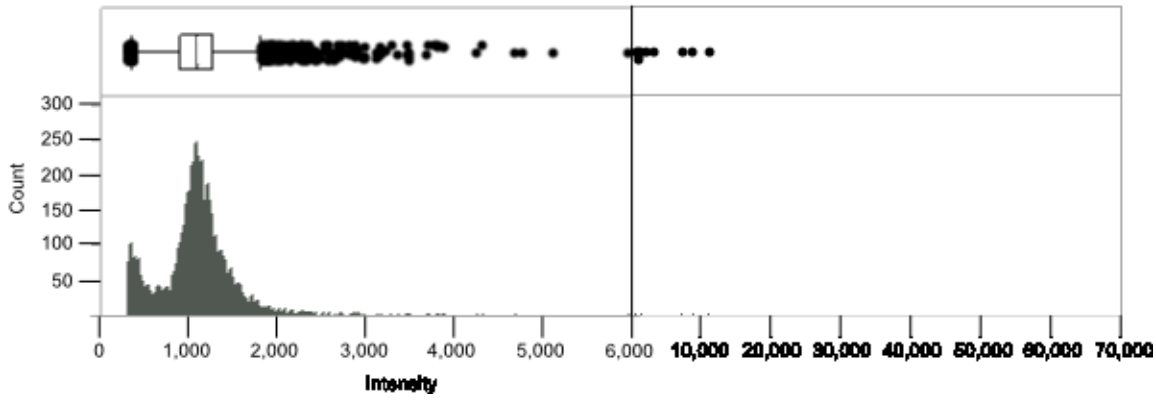


Figure 34 Statistical significance of measures comparing young and aged mice

Sera samples were applied to the CIM10K array and the AbStat measures were calculated. A p-value from a t-test was then determined for each measure for young vs aged. The negative logarithm in base 10 of the p-value was then plotted in a bar graph.

A)



B)

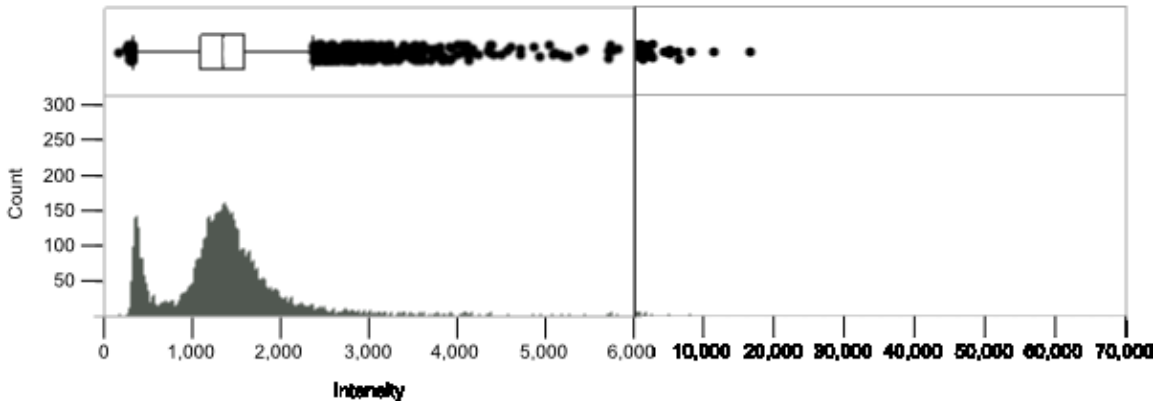


Figure 35 Young and aged mouse fluorescence intensity histograms

Sera from a A) young and B) aged mouse was applied to a CIM10K microarray separately. The fluorescence intensity values of all 10,000 features was then plotted in a histogram. The interval on the x-axis scale is 1,000 from an intensity of 0 to 6,000, and the interval changes (indicated by the vertical line separating the two parts of the graph) to 10,000 from an intensity of 6,000 to 70,000.

The results matched the expectations since aged mice did have a higher entropy fluorescence intensity distribution than young mice. The shape of the distribution is high and tight for the young mice, and the distribution is broader and flatter for aged mice with more outliers spread throughout the intensity range. This result demonstrates that the AbStat can be used to distinguish young and aged mouse sera samples.

1.3.7.2 Changes with age in humans

In the previous section, young and aged mouse samples were compared, and in this section young and aged human samples are compared. The aged samples could have a higher entropy than the young samples as was observed with the mice samples.

All of the normal samples with recorded ages in the 10k experiment (“1.3.6.3 CIM10K”) and from HT330K wafers 20, 22, 25, and 46 were used to analyze the effects of age on the fluorescence intensity distribution in humans. The experiments were performed by the Peptide Array Core under the direction of Zbigniew Cichacz. Many normal samples were collected from volunteers of many ages and nationalities throughout the Biodesign Institute. Across the datasets from the four wafers and the CIM10K experiment previously mentioned there were 99 samples. A sample was placed into the "Young” category if the age was less than or equal to 25 years old, and a sample was placed into the "Aged" category if the age was greater than or equal to 50 years old. Note that an analysis based on the selection of specific peptides for this entire dataset cannot be performed because there are samples from the CIM10K and the HT330K platform which have different peptides. This same issue does not exclude an analysis with the AbStat measure since this measure does not depend on the selection of specific peptides.

The average entropy of the aged group was higher than the average entropy of the young group (p-value of 2.44E-5) as presented in Figure 36. If linear regression is performed to fit a line to the collection of samples with age and entropy values, the resulting line has a slight upward angle of entropy with age with a slope of 0.0278, a y-intercept of 6.728, and an r^2 of 0.0958. Note that the total range of entropy in this dataset is larger than some of the other datasets, and this may be a characteristic unique to the individuals who donated blood. Most of this sera came from individuals who work at the Biodesign Institute, and there are many students and professors who have not been diagnosed with serious illnesses in this dataset. There can also be differences in the non-natural sequence peptide arrays that were produced when compared with other experiments, and this variation could result in more or less overall binding. Median normalizing the data from different batches would not assist one to make multi-batch comparisons with the entropy measure since the value of the entropy measure would not change.

Therefore, a different form of normalization or using the raw entropy values would be the necessary strategy to implement when making comparisons with the entropy measure.

A heatmap reveals that most of the aged samples cluster together with high entropy values (Figure 37). Subsets of normals also group together since there is a fairly large group of normals with relatively low values for all of the measures with the exception of the normalized maximum, normalized 5th percentile, kurtosis, and skew. The most statistically significant measures from a t-test were skew, entropy, and kurtosis (Figure 38).

Machine learning was performed on a subset of the dataset discussed previously and presented in Figure 36. This subset consisted of an equal number of young and aged samples. The young samples to include were chosen randomly. The measures with the greatest SVM weight were the entropy and normalized entropy (Figure 39). Note that the normalized entropy was included for this dataset since some of the samples were run on 10k arrays and some other samples were run on 330k arrays. The J48graft tree algorithm created a tree without even using the entropy measure (Figure 40), but the J48graft tree method was only able to correctly classify 58.7% of the instances; whereas the SVM algorithm could correctly classify 82.6% of the instances (Table 8). The SVM algorithm could not classify as well when samples were randomly assigned to the young or aged groups. A classification accuracy of 65.2% was achieved in this situation.

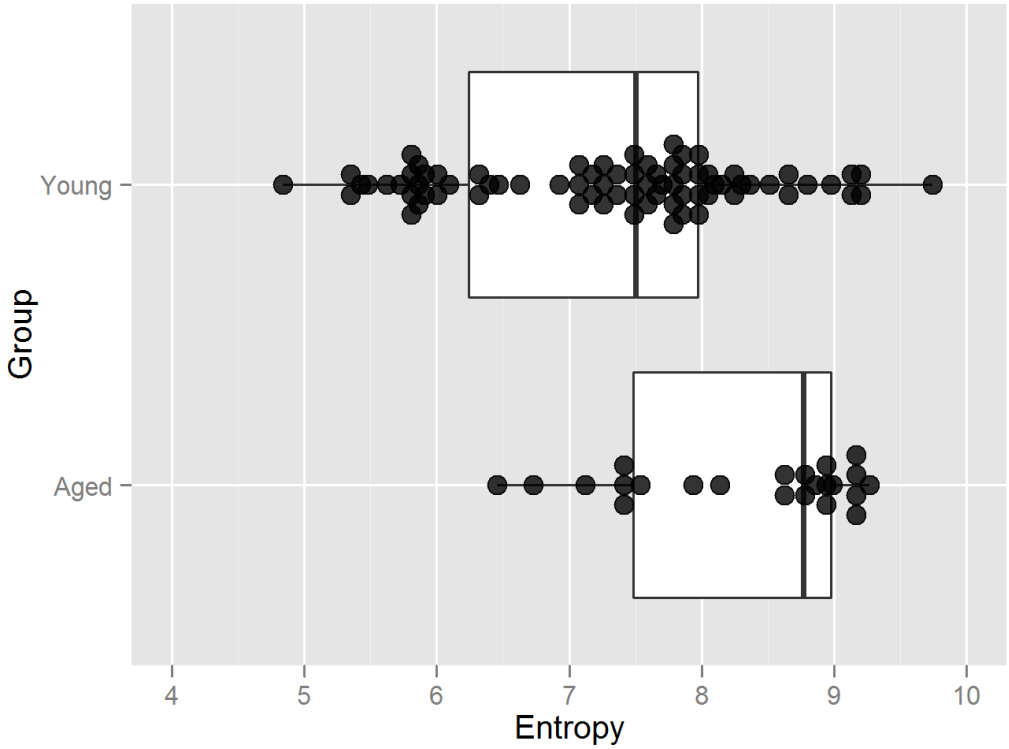


Figure 36 Box and dot plot of entropy for young and aged humans

Sera samples were applied to HT330K microarrays. The entropy of each replicate was calculated and presented in a box and dot plot. Young was classified as less than or equal to 25 years old, and aged was classified as greater than or equal to 50 years old.

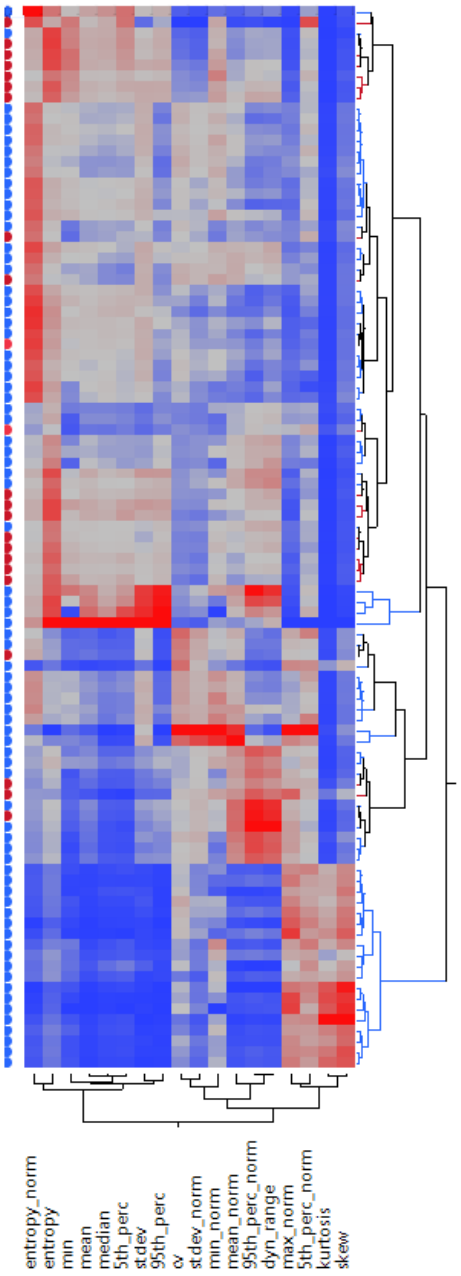


Figure 37 Heatmap of measures for young and aged humans

Each column corresponds to an AbStat measure and each row corresponds to a sera sample. The relative average value of each AbStat measure for the samples is represented by a color with blue indicating the lowest relative value and red indicating the highest relative value. The legend on the left indicates blue for young and red for aged. Each column corresponds to one of the

AbStat measures. Young was classified as less than or equal to 25 years old, and aged was classified as greater than or equal to 50 years old.

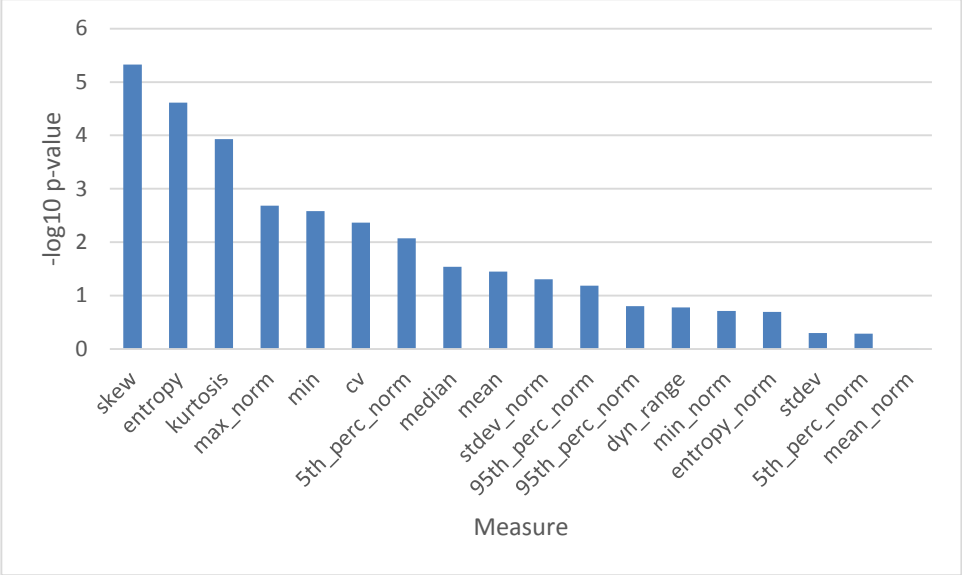


Figure 38 Statistical significance of measures comparing young and aged humans

Sera samples were applied to the HT330K array and the AbStat measures were calculated. A p-value from a t-test was then determined for each measure for young vs aged. The negative logarithm in base 10 of the p-value was then plotted in a bar graph.

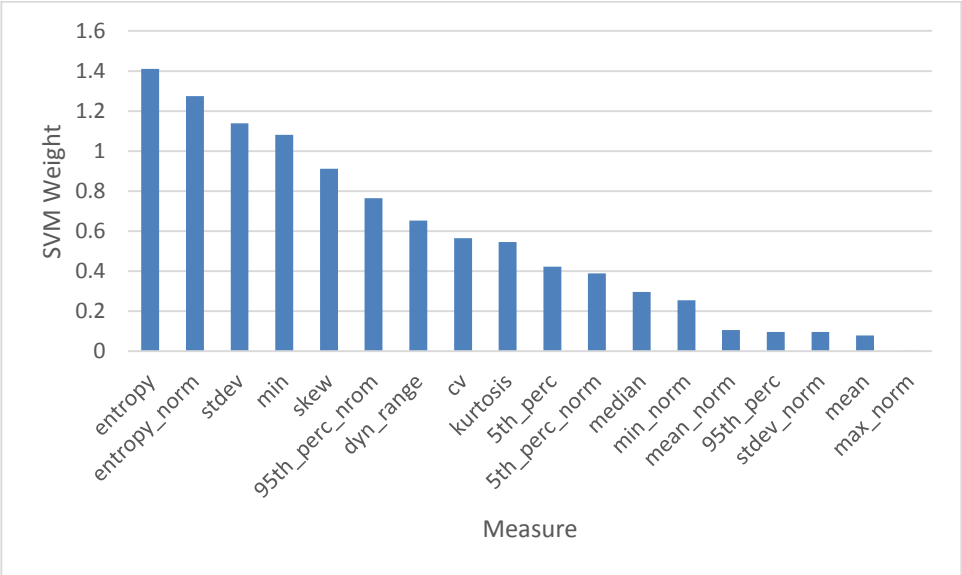


Figure 39 SVM weight of measures comparing young and aged humans

Sera samples were applied to the HT330K array and the AbStat measures were calculated. An SVM algorithm with 10-fold cross-validation was then used to predict the class of the samples as young or aged. The absolute value of the weight of each measure assigned by the SVM was then plotted in a bar graph.

```
kurtosis <= 38.192423: Aged (16.0/1.0)
kurtosis > 38.192423
|   mean_normalized <= 1.60487: Young (9.0)
|   mean_normalized > 1.60487
|   |   min <= 66: Aged (8.0/2.0)
|   |   min > 66: Young (13.0/2.0)
```

Figure 40 J48graft tree of young and aged humans

Sera samples were applied to the HT330K array and the AbStat measures were calculated. A J48graft tree algorithm with 10-fold cross-validation was then used to predict the class of the samples as young or aged. The algorithm then selected certain measures with specified cutoff points to construct a classification tree to assign a sample to the normal or disease group.

Table 8 Machine learning statistics for young and aged humans

Algorithm	Correctly Classified Instances	Kappa Statistic	ROC Area
SVM	82.6	0.652	0.826
J48graft	58.7	0.174	0.56
SVM Random	65.2	0.304	0.652
J48graft Random	63	0.261	0.597

Two different algorithms with 10-fold cross-validation were used to predict the class (young or aged) of samples applied to the HT330K microarrays. The attributes of each sample were the AbStat measures. Several machine learning statistics are presented in columns. In a separate analysis, samples were randomly assigned a class as young or aged and the analysis was repeated.

In order to provide further evidence that age was the attribute of these individuals which allowed the measures to distinguish between young and aged, two different nationalities were compared. One would not expect healthy individuals around the same age from different countries to have dramatically different antibody numbers, affinities, or avidities. Eight Chinese

samples and twelve Indian samples from donors under the age of 40 were compared. The machine learning algorithms were unable to distinguish between Chinese and Indian (Table 9).

Table 9 Machine learning statistics for Chinese and Indian nationality

Algorithm	Correctly Classified Instances	Kappa Statistic	ROC Area
SVM	52.4	-0.117	0.447
J48graft	52.4	-0.117	0.519

Two different algorithms with 10-fold cross-validation were used to predict the class (Chinese or Indian) of samples applied to the HT330K microarrays. The attributes of each sample were the AbStat measures. Several machine learning statistics are presented in columns.

All of these results support the idea that the global behavior of the antibody repertoire interacting with an array of peptide features changes with age, and this change can be detected with the AbStat measures. Approximately 80% of the young and aged samples could be correctly classified using the AbStat, and classification accuracy may improve with more advanced peptide arrays manufactured in the future. Note that it could also be interesting to apply pools of aged sera and pools of young sera to the array and test the classification performance of the AbStat with this data. In general, the aged samples result in higher entropy fluorescence intensity distributions. These findings could provide insight into the nature of the aging process as well as provide a method for monitoring health with age.

1.3.7.3 Specific peptide analysis with aged humans

A specific peptide analysis with the previous dataset could not be performed because this dataset combined human normal samples from the HT330K platform as well as the CIM10K platform in order to obtain a larger sample size for the classification algorithms. The reason a specific peptide analysis cannot be performed with this data is because the identity and number of peptides on the two different platforms are different and a t-test among the samples for a given peptide cannot be performed. Therefore, another analysis is presented here using only samples from the HT330K platform on wafers 20, 22, and 25. Half of the samples were assigned to the

training set, and half of the samples were assigned to the test set. The top 100 peptides by p-value from a t-test with the training samples were input into a classification algorithm. Once trained, the algorithm was used to predict the class of the test samples. The process was then repeated using samples with random class assignments. Additionally, in a separate test, the values of the AbStat measures for the training samples were input into a classification algorithm, and then the algorithm was used to predict the class of the test samples with the AbStat measures. The results of this analysis are presented in Table 10 for the real class assignments and in Table 11 for the random class assignments for each sample which should result in a much lower classification performance. The classification performance of the specific peptide method vs the AbStat method was about the same, with the specific peptide method exhibiting slightly better performance.

Table 10 Machine learning statistics for young and aged humans for specific peptide and AbStat method comparison

Algorithm	Correctly Classified Instances	Kappa Statistic	ROC Area
Naïve Bayes Specific Peptide	76.7	0.493	0.844
SVM Specific Peptide	86.7	0.683	0.841
J48Graft Specific Peptide	54.5	0.085	0.544
Naïve Bayes AbStat	73.3	0.366	0.683
SVM AbStat	66.7	0.333	0.824
J48Graft AbStat	76.7	0.493	0.77

Three different algorithms were used to predict the class (young or aged) of samples applied to the HT330K microarrays. Half of the samples were randomly selected as training samples and the other half was designated as the test set. The top 100 most significant features by p-value from a t-test with the training samples only were used to train the algorithms. The algorithms were then used to predict the class of the test set. Several machine learning statistics are

presented in columns. In a separate analysis, the process was repeated with the AbStat measures rather than from the features with the best p-value.

Table 11 Machine learning statistics for young and aged humans for specific peptide and AbStat method comparison with random class assignments

Algorithm	Correctly Classified Instances	Kappa Statistic	ROC Area
Naïve Bayes Specific Peptide	63.3	-0.122	0.44
SVM Specific Peptide	66.7	-0.0638	0.476
J48Graft Specific Peptide	57.6	0.087	0.542
Naïve Bayes AbStat	50	-0.154	0.421
SVM AbStat	70	0.211	0.55
J48Graft AbStat	70	0	0.5

Samples were randomly assigned to the young or aged class. Then three different algorithms were used to predict the class (young or aged) of samples applied to the HT330K microarrays. Half of the samples were randomly selected as training samples and the other half was designated as the test set. The top 100 most significant features by p-value from a t-test with the training samples only were used to train the algorithms. The algorithms were then used to predict the class of the test set. Several machine learning statistics are presented in columns. In a separate analysis, the process was repeated with the AbStat measures rather than from the features with the best p-value.

1.3.8 Rank of Measures

Which metrics of the AbStat measures provide the most information about healthy and disease states? The measures were evaluated in four different ways: 1. The rank of the measure by p-value acquired from comparing disease with normal (Figure 41); 2. The rank of the measure by p-value multiplied by the p-value acquired from comparing disease with normal (Figure 42); 3. The rank of the absolute value of the SVM weight for the measure acquired from comparing

disease with normal (Figure 43); 4. The rank of the absolute value of the SVM weight multiplied by the reciprocal SVM weight acquired from comparing disease with normal (Figure 44). The reciprocal was used so that smaller values would indicate the best rank as in the other three methods of ranking. The log of this value was taken since the values spanned a very large range. Four datasets were used to calculate the rank values: the first 330k dataset ("1.3.6.1 HT330K first chip disease dataset"); the 330k dataset from wafer 46 ("1.3.6.2 HT330K wafer 46"); the 10k dataset ("1.3.6.3 CIM10K"); and the human age dataset ("1.3.7.2 Changes with age in humans"). These results indicate that entropy is one of the best measures for distinguishing disease and normal samples followed by measures such as skew, kurtosis, cv, and dynamic range.

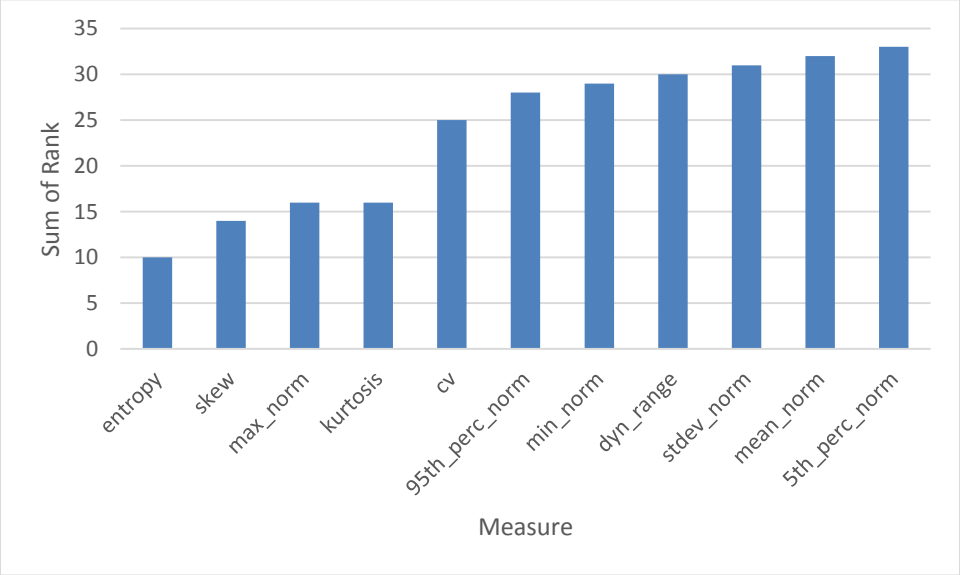


Figure 41 Sum of rank of measures by p-value

The sum of the rank of the AbStat measures by p-value acquired from comparing disease with normal from four different datasets was plotted in a bar graph.

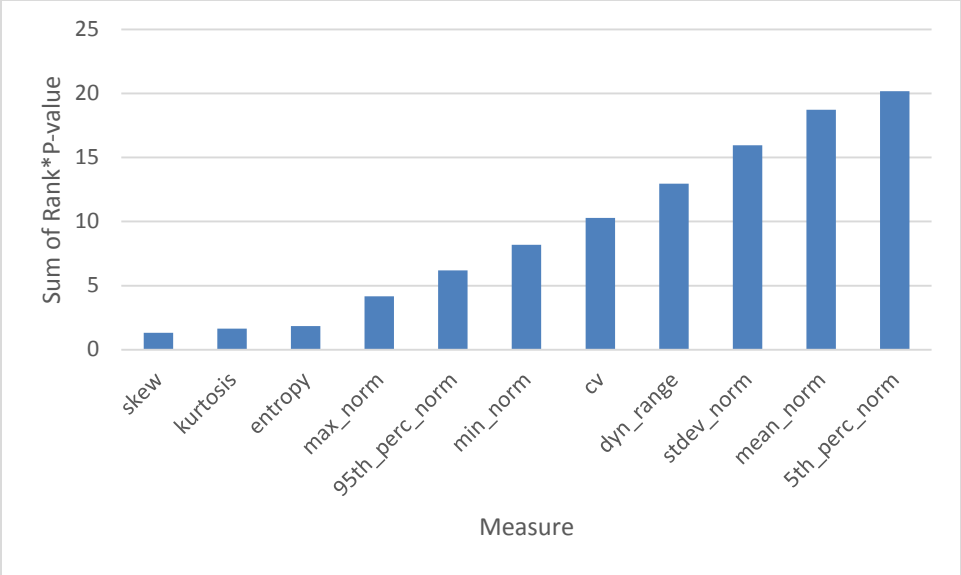


Figure 42 Sum of rank of measures multiplied by the p-value

The sum of the rank of the AbStat measures by the p-value acquired from comparing disease with normal multiplied by the actual value of the p-value from four different datasets was plotted in a bar graph.

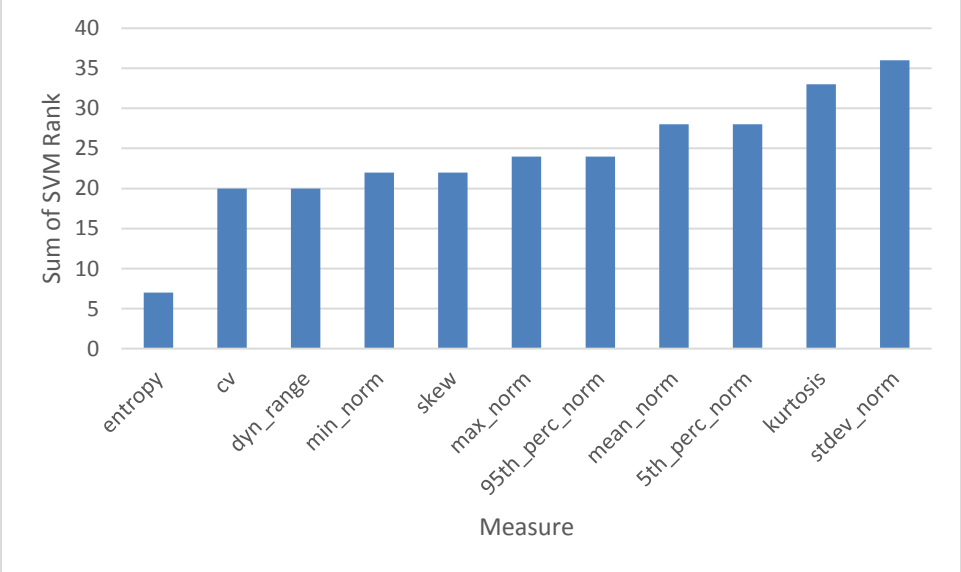


Figure 43 Sum of SVM Rank of measures determined by greatest absolute value SVM weight

The sum of the rank of the AbStat measures by absolute value SVM weight obtained from classifying disease and normal samples from four different datasets was plotted in a bar graph.

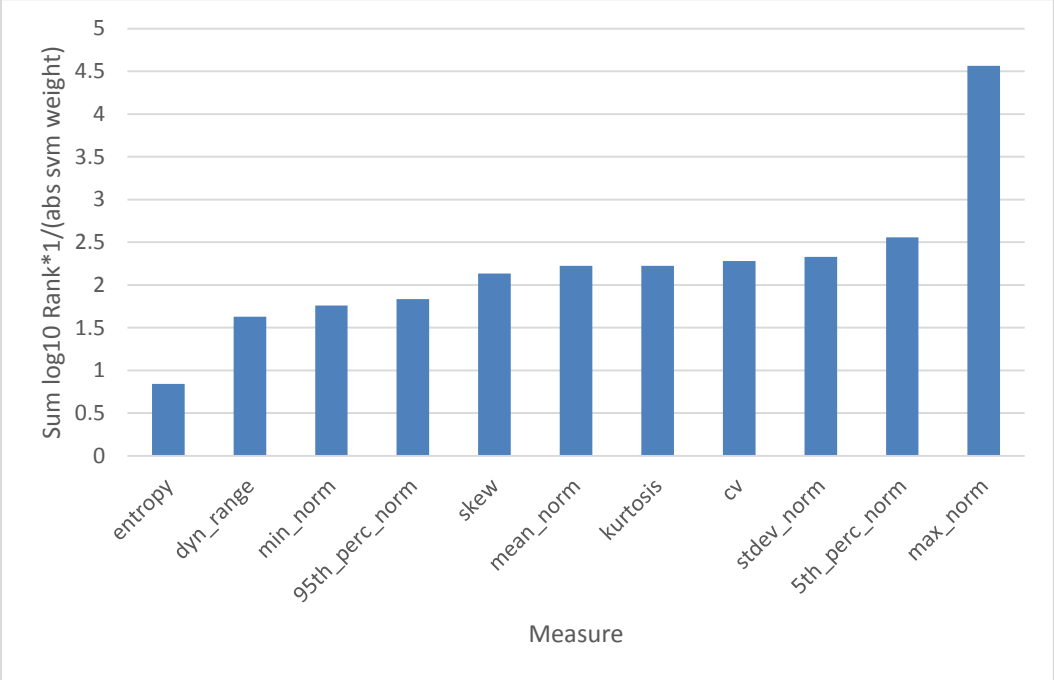


Figure 44 Sum of the log of the SVM rank multiplied by the reciprocal of the absolute value of the SVM weight

The sum of the log of the rank of the AbStat measures by absolute value SVM weight obtained from classifying disease and normal samples multiplied by the actual SVM weight from four different datasets was plotted in a bar graph.

1.3.9 Range of entropy

What is the typical range of entropy in healthy and disease states? A box and dot plot with the mean plus or minus one or two standard deviations for normal and disease samples is presented in Figure 45. This plot was constructed using data from four different datasets: the 1st 330k ("1.3.6.1 HT330K first chip disease dataset"); the wafer 46 330k ("1.3.6.2 HT330K wafer 46"); CIM10K ("1.3.6.3 CIM10K"); and the young/aged dataset ("1.3.7.2 Changes with age in humans"). The plot shows that the average entropy of disease samples is higher than the average entropy of normal samples. The plot also shows that there are many more data points at

the highest range that are disease samples, and there are many more data points at the lowest range that are normal samples. With this dataset, the total range of the normal samples is so large that it includes the range for the disease samples.

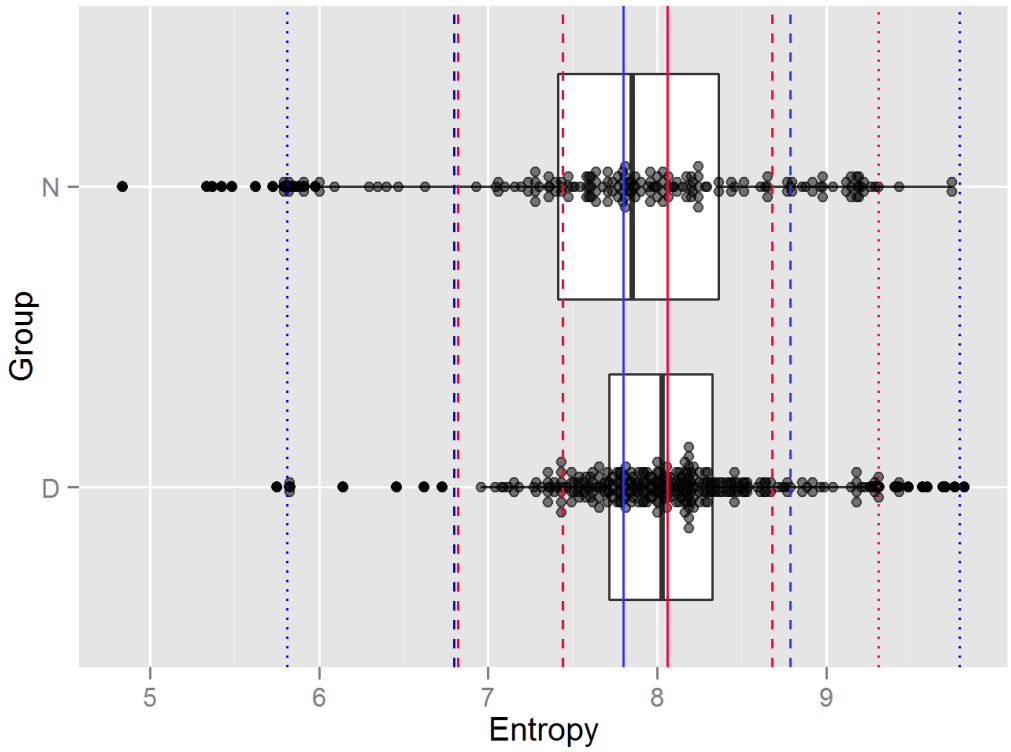


Figure 45 Box and dot plot of entropy values for normal and disease or aged states across four different datasets.

The blue solid line is the mean entropy for normal, the blue dashed line is the mean +/- 1 standard deviation, and the blue dotted line is the mean +/- 2 standard deviations. The red solid line is the mean for disease/age, the red dashed line is the mean+/-1 standard deviation, and the red dotted line is the mean +/- 2 standard deviations.

1.3.10 Quantitative analysis of the entropy measure

This section aims to identify how changes in the numerical analysis of the entropy measure affect the final outcome. From such an analysis, one may identify which parts of the fluorescence intensity distribution are most important for the entropy measure. One may also

identify possible ways to perform the entropy calculation to enhance the power of the entropy metric to distinguish between healthy and disease groups.

1.3.10.1 Changes in entropy measure with removal of peptides

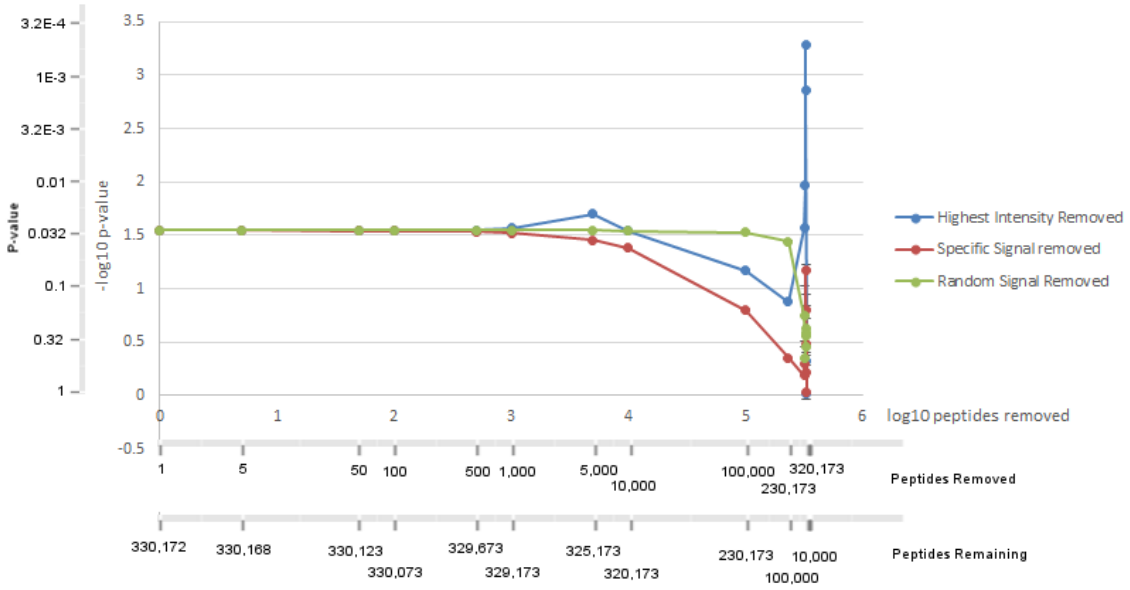
What is the characteristic of the peptides which contribute the most to the entropy measurement of the AbStat? In order to investigate this question, the 5 breast cancer samples and 24 normal samples from wafer 46 ("1.3.6.2 HT330K wafer 46") were compared. Different types of peptides were removed, the entropy was calculated for each sample with these removed peptides, and then a t-test between the two groups was performed. The results of this analysis are presented in ("Figure 46 P-value vs peptides removed"). In this figure, random peptides were removed five different times, and the standard deviation of the p-value for these five different times is plotted on the graph as error bars. However, the error bars are so small that they are not visible until about only 10,000 peptides are remaining. The graph has multiple axes so that the data can be viewed from different perspectives. On the y-axis the –log10 p-value is plotted so that more significant p-values are higher on the graph. The actual p-values at these positions are also displayed on the second y-axis so that one can quickly determine if a value is less than a p-value of 0.05 or another value. The log10 of the number of peptides removed is plotted on the x-axis, and the log10 is used since there is such a wide range of peptides removed. The second x-axis displays the actual number of peptides that were removed at each evaluation, and the third x-axis displays the number of peptides remaining, rather than removed, from the original 330k distribution. Note that as the random peptides were removed, the average entropy of the breast cancer groups (groups is plural because there were 5 different trials of random peptide selection) at each step was higher than the average entropy of the normal groups until there are only 1,000 peptides, at which point the inequality flips.

In addition to the random removal of peptides, peptides that were the most significant based on a t-test for the feature intensities for the two groups were also removed. These are the peptides that constituted the immunosignature of the disease of interest. Note that the average entropy of the breast cancer group is always higher than the average entropy of the normal group

at each step until there are only 50 peptides left at which point the inequality flips. In another analysis, the highest signal peptides for each sample were removed. Each sample was sorted by its own highest intensity. Therefore, in the highest intensity analysis, the same peptides were not removed from each sample at each step. The average entropy of the breast cancer group is higher than the average entropy of the normal group at each step until there are only 50 peptides left, at which point the inequality flips.

Some additional graphs provide more information about Figure 46 A. In this graph, there are many data points that lie in the region of the graph where there are few peptides remaining. Due to the scale, these data points are not well separated. A graph for 5,000 peptides remaining or less is also presented in Figure 46 B. The value of the most significant p-value in the peptide distribution at every point as the peptides were removed, rather than the p-value from a t-test between two groups with entropy values, is presented in Figure 47.

A)



B)

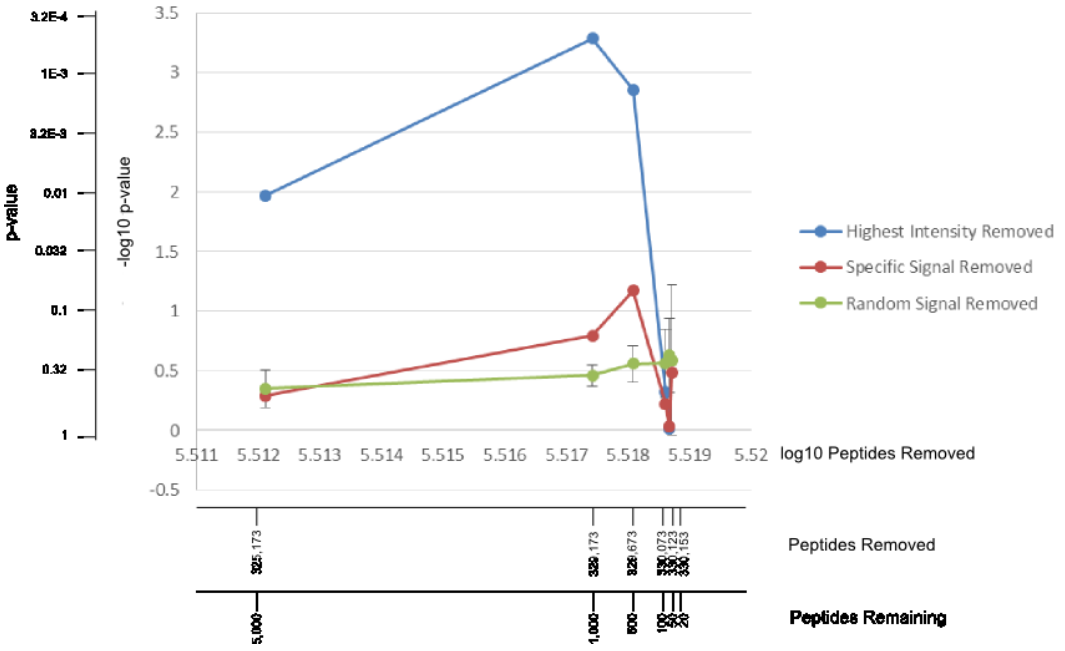


Figure 46 P-value vs peptides removed

A) Highest intensity, disease specific (based on p-value), or random peptide features were removed, the entropy was calculated for each sample with these removed peptides, and then a t-test between the two groups of breast cancer vs normal was performed. The y-axis displays the p-value from the test as well as the negative logarithm of the p-value. The x-axis displays the logarithm of the number of peptides removed, the number of peptides removed, and the number of peptides remaining. B) The same data is plotted from 5,000 to 20 peptides remaining only.

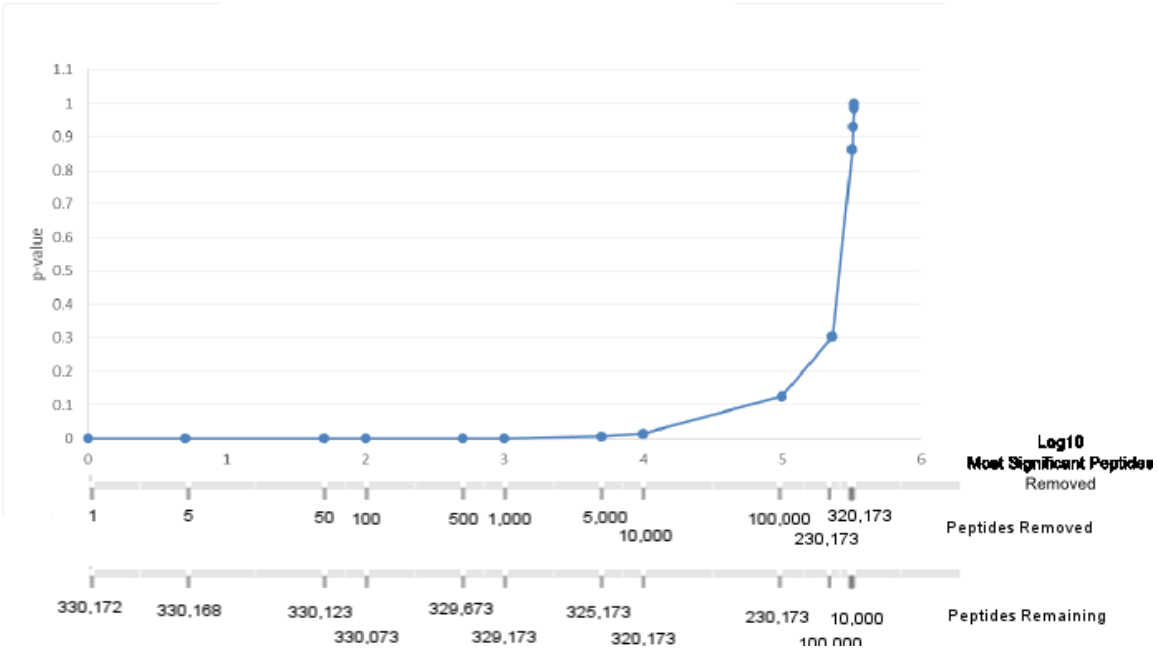


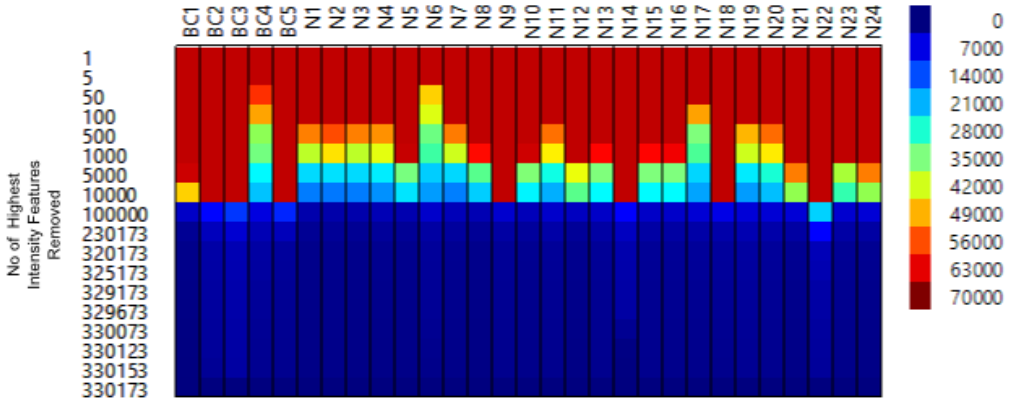
Figure 47 P-value of most significant peptide vs significant peptides removed

The value of the most significant p-value peptide from a t-test between normal and disease (no entropy values were a part of this calculation) was plotted in a line graph as the most significant peptides were removed.

More details about the highest fluorescence intensity features can be investigated. One might be curious about the value of the highest fluorescence intensity at each step as the highest fluorescence intensity peptides are removed. In Figure 46 the p-value becomes much more significant between the normal and breast cancer groups when only 1,000 peptides remain (329,173 peptides removed) after the highest intensity peptides have been removed. What is the fluorescence intensity of the highest intensity peptide for each sample at this point? The heatmap in Figure 48 A shows the value of the highest fluorescence intensity peptide for each sample as the highest fluorescence intensity peptides are removed. The heatmap in Figure 48 B shows the same data with a color bar scale from 0-2000 only instead of 0-70,000. In this figure, the highest fluorescence intensity value for each sample lies around 1,000 when there are only 1,000 features remaining. This information might suggest that calculating the entropy with fluorescence intensities less than about 2,000 could be beneficial in certain circumstances. The “noise” or distribution of fluorescence intensities in this low fluorescence intensity range appears to be

different in normal and disease samples, and the evidence for this is provided in the p-value graph in Figure 46.

A)



B)

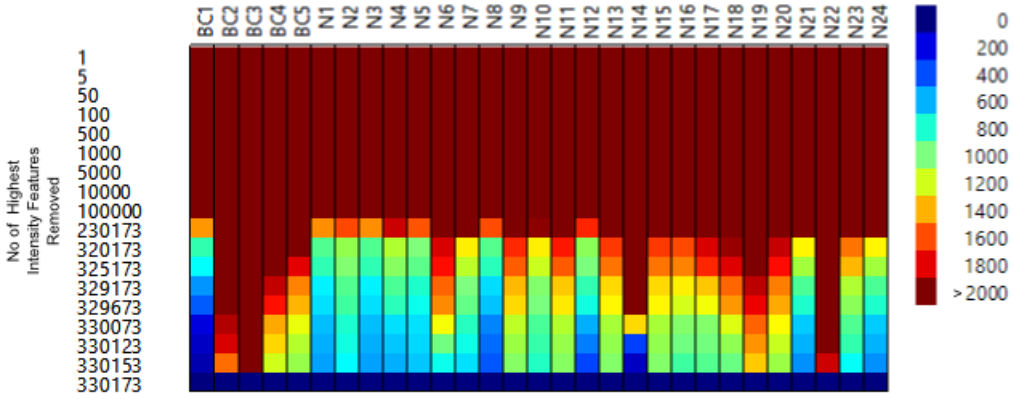


Figure 48 Heatmap of highest intensities as highest intensity features are removed

The columns contain the samples with BC for breast cancer and N for normal. The rows indicate the number of highest intensity features removed. The color of each square corresponds to the intensity value of the highest intensity feature remaining for the given sample. In A) the color bar ranges from 0-70,000 and in B) the color bar ranges from 0-2,000 for the same data.

Another analysis was performed in which the least significant peptides were removed. The entropy of each sample was obtained, and a t-test between the breast cancer and normal group was performed. A graph of the p-value between the two groups as the least significant peptides were removed is displayed in Figure 49. The entropy of the breast cancer group is higher at each step until there are 100,000 peptides remaining, at which point the inequality flips. From there, the average entropy of the breast cancer group is lower than the average entropy of the normal group at each step until there are only 1,000 peptides remaining, at which point the inequality flips. The actual p-value of the least significant peptide as least significant peptides were removed is plotted in Figure 50.

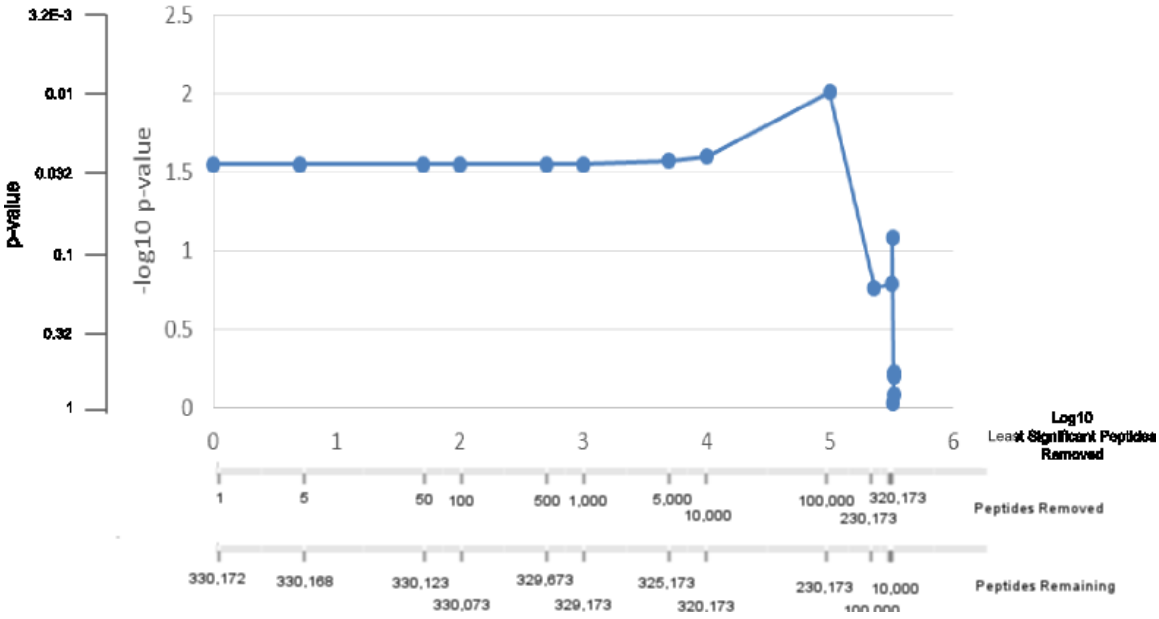


Figure 49 P-value vs least significant peptides removed

The least disease specific (based on p-value) peptide features were removed, the entropy was calculated for each sample with these removed peptides, and then a t-test between the two groups of breast cancer vs normal was performed. The y-axis displays the p-value from the test as well as the negative logarithm of the p-value. The x-axis displays the logarithm of the number of peptides removed, the number of peptides removed, and the number of peptides remaining.

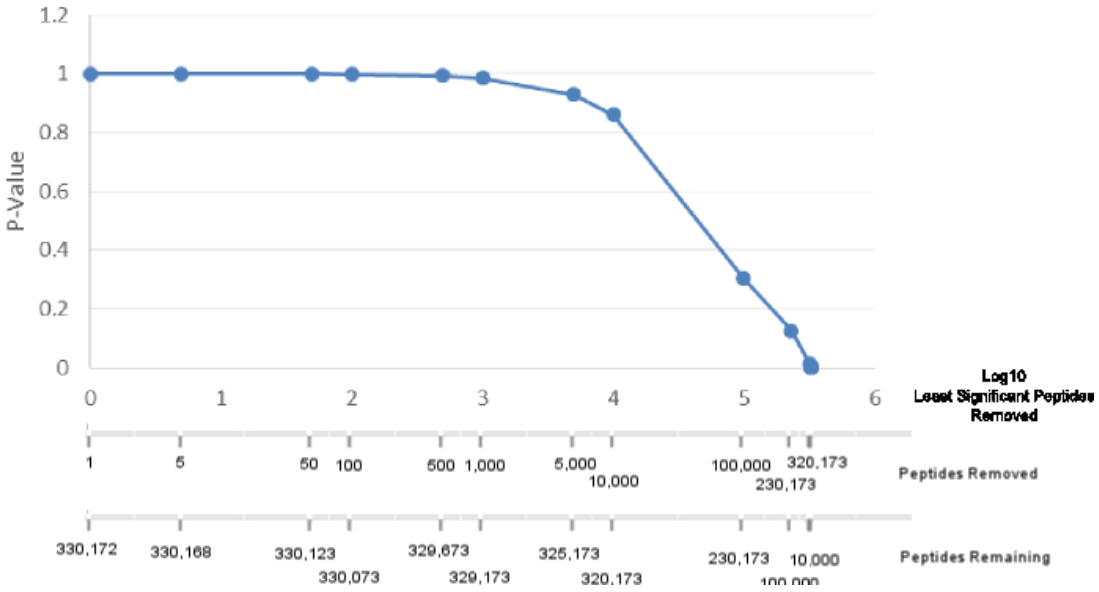


Figure 50 P-value of least significant peptide vs peptides removed

The value of the least significant p-value peptide from a t-test between normal and disease (no entropy values were a part of this calculation) was plotted in a line graph as the least significant peptides were removed.

How did the distinguishing power of the entropy measure change as the lowest intensity peptides were removed? The lowest intensity peptides were removed from each sample individually. Therefore, the identity of the peptides removed in each sample could be slightly different. The entropy of each sample was obtained, and a t-test between the breast cancer and normal group was performed. A graph of the p-value between the two groups as the lowest intensity peptides were removed is displayed in Figure 51. The entropy of the breast cancer group is higher at each step until there are 100,000 peptides remaining, at which point the inequality flips. The inequality is somewhat sporadic after this point. The entropy of the normal group is then higher when there are 10,000 peptides remaining, higher when there are 5,000 peptides remaining, lower when there are 1,000 peptides remaining, higher when there are 500 peptides remaining, and lower when there are 100 peptides remaining. The p-value does not

change much until about only 100,000 peptides remain after removing the lowest intensity peptides. Nevertheless this indicates that the 230,173 lowest intensity peptides are important for the entropy measure. In fact, low intensity peptides appear to be more important than high intensity peptides since removing these peptides decreases the significance of the p-value and taking away high intensity peptides can increase the significance of the p-value as shown in Figure 46.

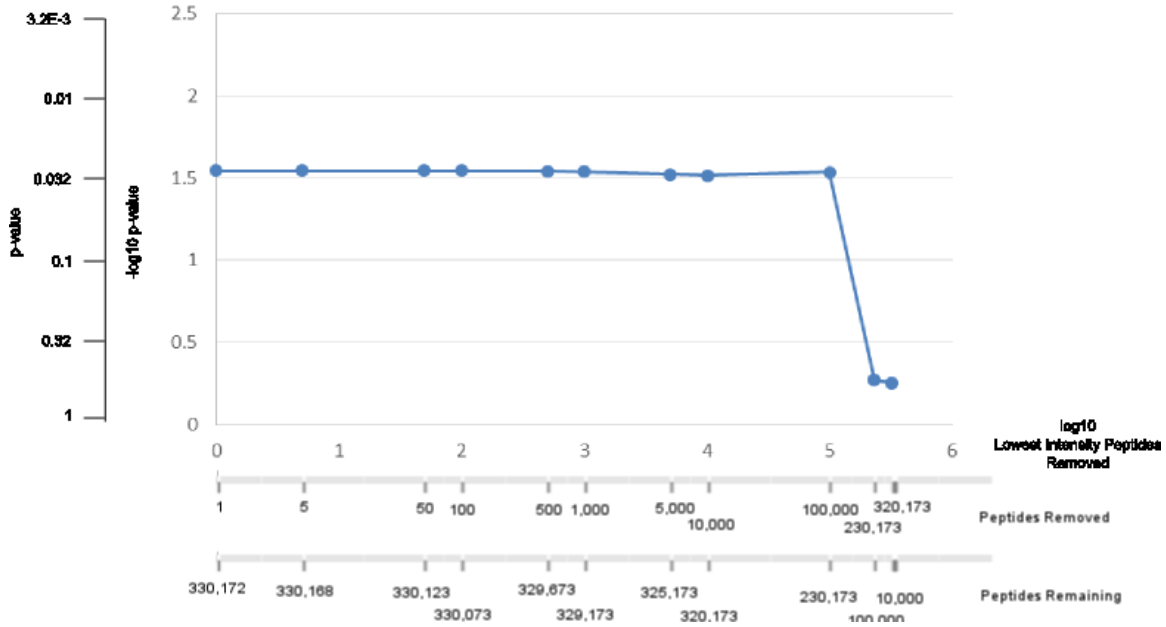


Figure 51 P-value vs lowest intensity peptides removed

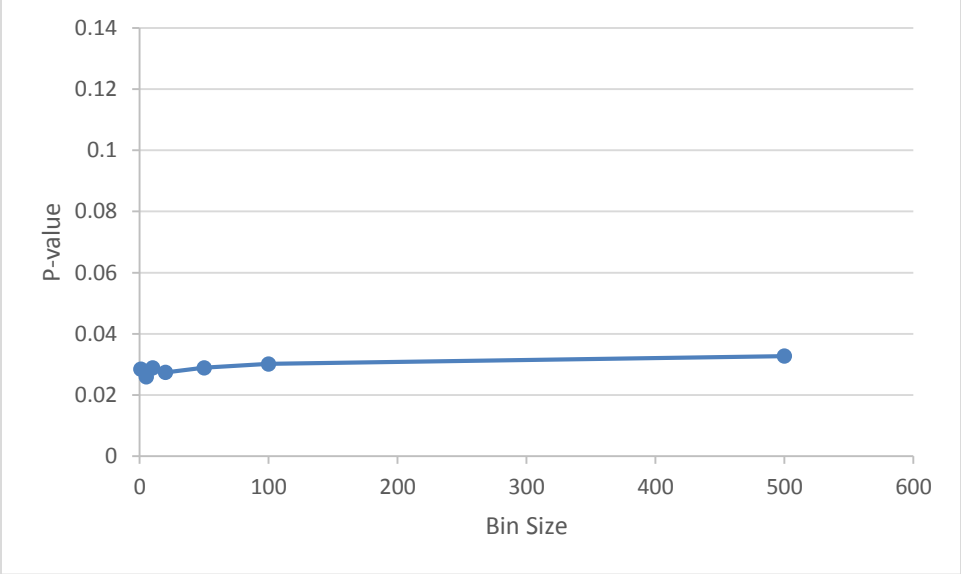
The lowest intensity peptide features were removed, the entropy was calculated for each sample with these removed peptides, and then a t-test between the two groups of breast cancer vs normal was performed. The y-axis displays the p-value from the test as well as the negative logarithm of the p-value. The x-axis displays the logarithm of the number of peptides removed, the number of peptides removed, and the number of peptides remaining.

1.3.10.2 Change in entropy with change in bin size

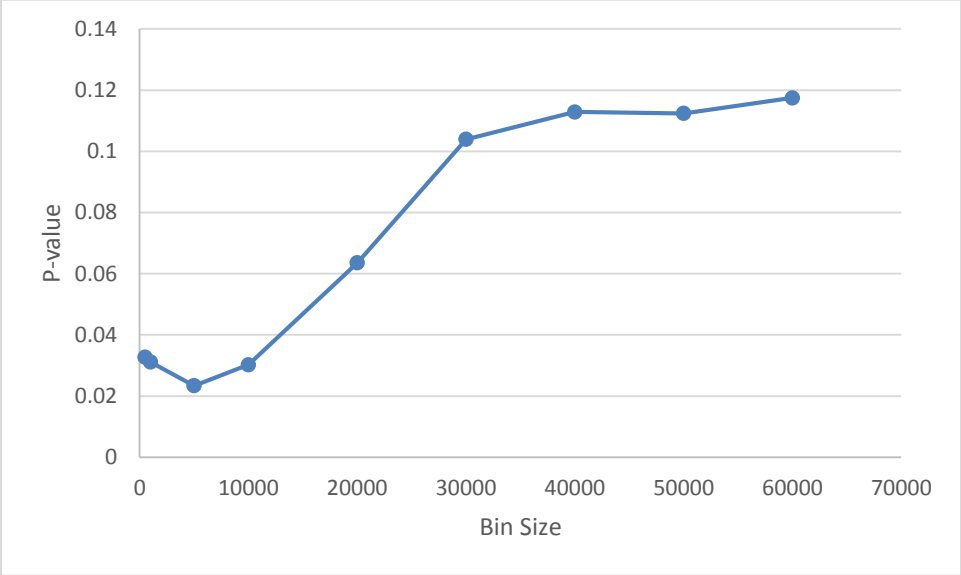
When calculating entropy, it is necessary to set the size of the bins. How does the ability of the entropy metric to distinguish between disease and healthy groups change with changes in

bin size? In order to investigate this question, the 5 breast cancer samples and 24 normal samples from wafer 46 (1.3.6.2 HT330K wafer 46) were compared. The bin size was adjusted from 1 to 60,000 (the max fluorescence intensity is 65,535), the entropy was calculated for each sample at each bin size, and a p-value from a t-test with the normal donor and breast cancer groups was acquired. The results are presented in Figure 52. Note that throughout this dissertation, if the bin size was not specified, then the smallest bin size of 1 was used. These results demonstrate that the p-value generally increases (becomes less significant) with increasing bin size with some exceptions like the exception at a bin size of 5,000 (Figure 52 B). Note that at every bin size, the average entropy of the breast cancer group was higher than the average entropy of the normal group.

A)



B)



C)

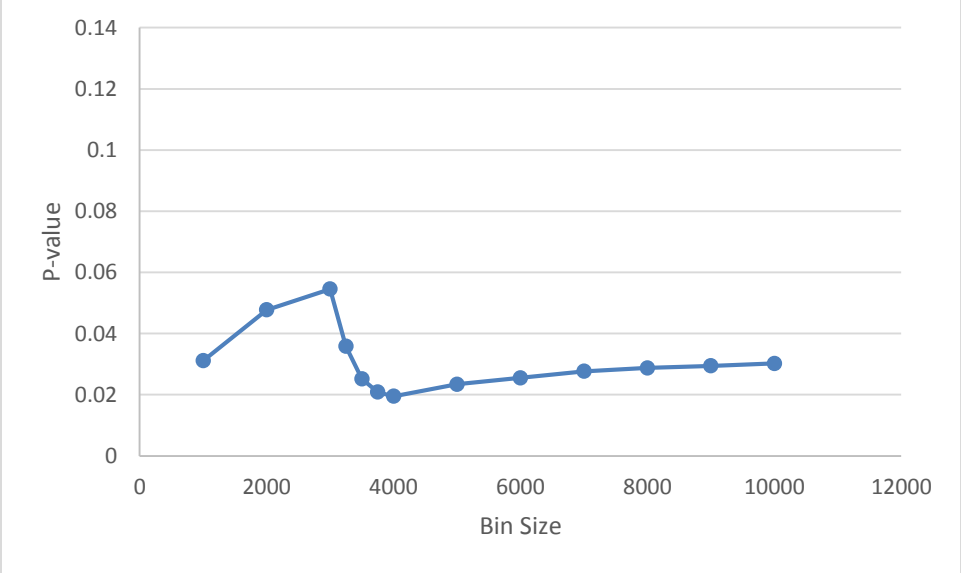


Figure 52 P-value vs bin size

The bin size was adjusted and the entropy of each sample was calculated. A p-value from a t-test between breast cancer and normal with the entropy values was then performed. The p-value vs bin size was plotted on a line graph. The bin size was adjusted from 1 to 500 in A), 500 to 60,000 in B), and 1,000 to 10,000 in C).

The slight drop in p-value at a bin size of 5,000 was somewhat surprising (Figure 52 B). The p-values at more bin sizes near a bin size of 5,000 were acquired to increase the resolution of the curve near this area. This data is presented in Figure 52 C. This curve shows that there is a drop in p-value after a bin size of 3,000, and the lowest p-value is near a bin size of 4,000. Note that the p-value at a bin size of 4,000 is not dramatically different than the p-value at a bin size of 5,000.

1.4 **Discussion**

The AbStat measures provide indicators for a wide variety of different immunological situations. I first demonstrated that the entropy measure of three different antibodies was different and correlated with the affinity of these antibodies for their cognate epitopes. Next I investigated the behavior of the AbStat measures as the concentration of a monoclonal antibody spiked into sera was increased, and I found that the entropy decreased up to a certain point in two different experiments. Next I investigated the behavior of the AbStat with mouse vaccines and infections and found that entropy decreased after vaccination, while also increasing in general over long periods of time. Human vaccines also showed that the entropy decreased after vaccination if more than three days had passed. Dog lymphoma samples revealed that the entropy decreases when one single antibody begins to comprise a larger portion of the antibody repertoire, and these findings supported the results from the controlled antibody spiking experiments. A cancer mouse was also compared to a normal mouse. The normal mouse exhibited a steady increase in entropy over time, but the entropy of the cancer mouse over time was much more sporadic. Human infectious disease, chronic disease, and normal samples were investigated and I demonstrated that the AbStat measures can be used to correctly classify the samples with about 80% accuracy with the HT330K platform, but only 60% accuracy with the CIM10K platform. The entropy was generally higher for the disease samples. Alzheimer's samples could not be distinguished from the controls. Finally, I demonstrated that the AbStat measures were capable of distinguishing young mice from aged mice as well as young humans from aged humans with a higher entropy in the aged state.

I also used the data from these experiments to rank the most important measures of the AbStat, and the entropy measure turns out to be the most distinguishing measure. A more detailed analysis of this entropy measure reveals that it is more affected by peptides with certain characteristics such as high intensity or statistical significance than randomly selected peptides. Ultimately, the concept of using entropy and other broad measures of a distribution to assess the vitality of a complex adaptive system is not restricted to antibodies and non-natural sequence peptides. The potential of the AbStat for the immune system though is that it can aid with diagnosis of health and disease states, and the concept may also lead to strategies to mitigate the development of disease states such as old age.

1.4.1 AbStat changes with artificial antibody experiments

Three different controlled antibody experiments were performed in order to understand the relationship of the AbStat measures with certain types of antibody mixtures: 1. Antibodies with a range of recorded affinities for their targets were applied to the array; and 2. Increasing concentrations of antibody against a single target were spiked into sera. The knowledge gained from these experiments allowed for better understanding and prediction of the AbStat behavior with different health states.

1.4.1.1 Monoclonal affinities and AbStat measures

The different types of measures exhibit different behaviors across the range of affinities tested. An affinity is quantified by the Kd value which measures the strength of the interaction between an antibody and its cognate epitope. Although the cognate epitope may not be present in any of the non-natural peptides on the microarray, an antibody with a high affinity for its cognate epitope will also have a higher affinity for mimotopes present in some of the peptides on the array. The following measures increase as the Kd increases and the monoclonal antibody binds less tightly to its specific cognate epitope: entropy, mean, median, 5th percentile, 95th percentile, 95th percentile normalized, dynamic range, standard deviation, and min. The following measures decrease as the Kd increases: minimum normalized, standard deviation

normalized, mean normalized, 5th percentile normalized, cv, kurtosis, and skew. The maximum did not follow a trend in one direction.

What is a possible explanation for the trends observed? In general an antibody with a low Kd will bind tightly to peptides with sequences similar to the target, and the intensities of the interactions may be in a very similar range. This will produce a "spike" in the fluorescence intensity distribution. The entropy of such a distribution will be lower than a broad distribution, and the mean and median may remain on the lower end if there are only a few specific high signals. This logic would also apply to the 95th percentile, 5th percentile, and dynamic range which is actually the ratio of the 95th percentile to the 5th percentile. An antibody with a high Kd for the target, on the other hand, will not bind tightly to peptides with sequences similar to the target. These antibodies may bind more broadly to many different peptides on the array in a more chaotic manner producing many different intensity values thus leading to a higher entropy and an overall higher mean, median, and dynamic range. Note that the kurtosis and skew often exhibit the opposite behavior of entropy, and many of the other measures which decrease as the Kd increases are measures calculated after the fluorescence intensity distribution has been median normalized.

The observations can be concisely summarized by stating that high affinity antibodies (low Kd) will result in fluorescence intensity distributions with tight peaks ("Figure 3 Histogram and box plot for <0.1 nM Kd antibody"), and low affinity antibodies (high Kd) will result in fluorescence intensity distributions with broad peaks ("Figure 4 Histogram and box plot for 80 nM Kd antibody"). A high affinity antibody will bind tightly to a few peptides and ultimately result in a high fluorescence intensity for these array features. The rest of the fluorescence intensity distribution will form a tight high peak in the range with almost no intensity value (values close to zero). A low affinity antibody on the other hand will bind chaotically with mostly weak binding to many different peptides forming a shorter broader peak in the lower intensity range with moderate binding in the mid and high intensity ranges. Note that the reason high affinity antibodies can bind tightly to a few specific sequences is due to the fact that high affinity antibodies are affinity matured antibodies representing the immune system's mature, late response to an antigen. The summary

measures used can distinguish between narrow and broad fluorescence intensity distribution shapes and summarize the observation in a single quantitative number. Although all of the different measures are calculated in a different manner, most of them will have general trends for narrow or broad distributions.

1.4.1.2 Spiking antibody into sera

Two separate experiments in which antibody was spiked into normal sera at increasing concentrations resulted in similar trends. In both experiments, both the entropy and mean from raw intensity values decreased as increasing concentrations of antibody against a single target were added ("1.3.1.2 Spiking antibody into sera"). Then once a certain point was reached the trend reversed and both the mean and entropy increased. An explanation for the decrease in entropy as more target specific antibody is added is that the antibody will bind to a few peptides with sequences similar to the target epitope with roughly the same intensity. Therefore, a "spike" in the fluorescence intensity distribution may appear which would lead to a lower calculated entropy. After a certain concentration is reached, however, the antibody will have already bound to all of the peptides on the array with sequences similar to the target epitope. Once this saturation point is reached, the antibody may "spill over" and bind to many other peptides with a variety of intensities which would lead to a broader fluorescence intensity distribution profile and therefore a higher calculated entropy value. These observations suggest that a solution which behaves more like a monoclonal antibody will have a lower entropy unless the monoclonal antibodies have completely saturated the array. A corollary is that the entropy may show a concentration curve dependent on the ratio of peptides to antibodies.

Also note that although the entropy and the mean followed similar trends, they are different measures and one or the other may provide more or less resolution between samples depending on the situation. For example, when comparing the 1 nM concentration to the 10 nM concentration in "Figure 6 Entropy for increasing concentrations of α–gp120 antibody into normal human sera", the entropy provides a more significant difference between the two groups than the mean (p-value of 0.0253 compared to 0.140). Also note that while the mean will change

depending on whether or not the data has been median normalized, the value of the entropy will remain unchanged. This is a particularly important feature when trying to compare data across different array types.

How reasonable is the explanation that the trend of the entropy reversed due to saturation of the target peptides? A few very approximate calculations can be used to address this question. For the sake of simplicity, the system of the peptide microarray is modeled by pretending that the target peptides of the antibody are present at a homogenous concentration in the chamber rather than just at spots on the slide surface. We can use the Kd equation to determine the Kd of the antibody if the system has reached saturation so that all of the target peptides are bound by antibody. The entropy trend in "Figure 5 Entropy for increasing concentrations of α-GFOD1 antibody into normal mouse sera" for the GFOD antibody may indicate that saturation has been reached at about 5 nM since the trend reverses at this point. Note that the Kd values of antibodies typically range from 1 pM to 10 μM. If the system is saturated and 95% of the target peptides have been bound by antibody when the concentration of free antibody is approximately 5 nM, then the Kd value would be 0.3 nM which lies within the normal range of Kd values for antibodies (Equation 6). If 95% of the target peptides have been bound at an antibody concentration of approximately 1 nM, as Figure 6 may suggest for the gp120 antibody, then the Kd value would be 50 pM. These approximate calculations support the idea that the entropy may start to increase once the target peptides have been saturated. Beyond this antibody concentration, the additional antibodies may bind with less specificity and even bind somewhat randomly to the remaining peptides on the array.

$$Kd = \frac{[A] * [B]}{[AB]} \rightarrow Kd = \frac{5nM * 0.05 * [B]}{0.95 * [B]} = 0.3nM$$

Equation 6 Dissociation constant calculation for saturation at 5 nM

In this calculation [A] can be considered to be the concentration of free antibody, [B] is the concentration of free peptide, and [AB] is the concentration of peptide bound to antibody. When 95% of the free peptide is bound by antibody, the concentration of free peptide is 0.05 multiplied

by the original concentration of peptide, and the concentration of antibody bound by peptide is 0.95 multiplied by the original concentration of peptide.

These approximate Kd calculations can be compared with the Kd values determined from a one to one binding model. The software Graphpad Prism can implement a one to one binding model to perform non-linear regression to fit the equation $y = B_{max}*X/(K_d+X)$ where y is the specific binding, B_{max} is the maximum specific binding, X is the concentration of ligand, and the K_d is the equilibrium binding constant defined as the ligand concentration needed to achieve half-maximum binding. The input in this case is the fluorescence intensity value of a peptide at various concentrations of antibody. In order to use this model the top 10 peptides by intensity were chosen. Note that a peptide was not chosen if the intensity was at the maximum value for all concentrations. The Kd value was then calculated for each peptide, and the calculated Kd values from peptides with an $r^2>0.90$ for the non-linear regression were averaged. The resulting Kd for the GFOD antibody is 61.9 nM, and the resulting Kd for the gp120 antibody is 15.9 nM. The approximated Kd from the entropy described previously was 0.3 nM for the GFOD antibody and 0.05 nM for the gp120 antibody. Although the Kd values from these two different methods of approximation differ, the order of Kd values for the two different antibodies remains the same, and both methods yield Kd values within the normal range observed for antibodies.

In summary, entropy appears to decrease as increasing concentrations of antibody against a single target are added. However, once the concentration is high enough, the entropy reverses and begins to increase. The point at which the trend in entropy reverses may be near the concentration of antibody at which the targets have been bound and the array is saturated. At this point, antibody may “spill over” and bind to many other peptides somewhat non-specifically. A few approximate methods for determining the Kd from this data support this hypothesis since the Kd values obtained are within a normal antibody Kd range. Note that the relationship of the entropy to the antibody concentration may be non-linear since the “spill over” antibody binding may be driven by avidity, which may be a more diffuse process than a process driven by affinity.

1.4.1.3 Summary of AbStat changes with artificial antibody experiments

These controlled antibody experiments reveal trends in the AbStat measures for certain types of solutions. In these examples, solutions that behave more like a monoclonal antibody have low values (with the exception of kurtosis and skew which will be high), and solutions that are complex with low affinity antibodies will have high values. This conclusion is derived from two types of experiments. In the first type of experiment antibodies with high or low affinities were applied to the array, and the data revealed that measures which often followed entropy increase for low affinity antibodies. A second type of experiment involved antibody spiking experiments. Two examples of antibody spiked into normal sera revealed that measures that follow the entropy decreased as monoclonal antibody was added up until a certain point, and then at concentrations greater than this point the trend reversed indicating that the array may be fully saturated. The lessons learned from these experiments indicate that if there is a disease that raises a few high affinity antibodies or alternatively results in a complex mixture of many different antibodies, then the AbStat measures such as entropy will be significantly different enough from normal sera to detect a change.

Note that it is extremely important to point out that not all of the measures are always positively or negatively correlated with entropy. In these tightly controlled artificial antibody experiments, this was mostly the case. However, there are certainly situations in which this would not be the case. For example, one could have a very low entropy, but a very high mean, which was not observed in these experiments. This situation would occur when there is a very high tight narrow peak at a very high intensity value. Each measure characterizes the peptide distribution in a unique way and they do not all positively follow or negatively follow one another in all circumstances. The results from the human disease data demonstrates this (“1.3.6 Human disease”). Three measures which do appear to be quite connected are entropy, kurtosis, and skew. When the entropy is low, kurtosis and skew are usually high; when the entropy is high, kurtosis and skew are usually low.

The observations from these controlled artificial antibody experiments aid in understanding the state of human sera. Diseases which make the sera appear more like a

monoclonal antibody should decrease the entropy measure. For example, in a B cell lymphoma a great portion of the antibody repertoire consists of one single antibody from a B cell that is proliferating out of control and producing large amounts of one antibody. Vaccines may also result in a similar observation since a vaccine will induce the immune system to produce large amounts of antibody against a single well defined target in a short amount of time. Diseases in which there is a large amount of inflammation and many non-specific and specific antibodies against many different targets would have complex antibody mixtures with many low affinity antibodies. These types of diseases may result in a high entropy value. Diseases that could fall into this category are chronic diseases such as cancer, and autoimmune diseases. The aging process may also fall into this category as overall inflammation (non-specific immune responses) and autoimmunity increases. Even without an increase in inflammation and autoimmunity, the aging process might lead to an increase in the complexity of the antibody repertoire as more antibodies are produced against more targets over time.

There are detectable and quantifiable changes in the overall antibody repertoire when transitioning from different types of health and disease states. These controlled antibody experiments provide insight into the correlation between different types of antibody mixtures and several quantitative measures. This insight can be further applied to understand actual sera in various health states.

1.4.2 Mouse vaccines and infection

There were three separate experiments in which mice were vaccinated against or infected with an infectious pathogen: 246 day time course experiment, a multiple mouse immunization experiment, and a 6 day time course. The 6 day time course may have been too short to detect many changes in the AbStat. However, the 246 day time course spanned a considerable portion of the mouse lifespan and the AbStat exhibited some clear trends over this long period of time. The multiple mouse immunization experiment revealed differences in measures of the AbStat for the different groups. These experiments provided some initial insight before examining human vaccine data.

1.4.2.1 246 day time course

In the 246 day time course many of the AbStat measures change in a consistent way over several time points. For example, the entropy measure increases as the mouse age increases (Figure 8). The normalized standard deviation and normalized 95th percentile start high and decrease with age, while the mean, median, and ninety fifth percentile start low, but then increase with age (Figure 7). Therefore, there are clear changes that occur in the overall binding of the array, and these changes can be detected without focusing on specific peptides which may have a mimotope with a known antigen. For this particular experiment, it is unclear whether these changes are due to the vaccine and challenge, the aging process alone, or a combination of both since there is not a mock group to compare the experimental group with.

1.4.2.2 Multiple mouse immunizations

The multiple mouse immunization experiment had data for one time point 37 days after treatment which consisted of one of two immunizations, a killed PR8 injection, an infection, or a mock treatment. The group with the greatest entropy was the infected group, and the group with the lowest entropy was the 2007-2008 vaccinated group. For this experiment, the dynamic range was the measure that was able to distinguish between some of the groups the very best since the dynamic range had the best p-value for mock vs infected, and the SVM algorithm assigned the greatest weight to the 5^{th} percentile normalized and dynamic range measures. The dynamic range was lower in the infected group than the normal group which indicates that there was a narrower range of intensities produced by antibodies binding to non-natural sequence peptides for the infected group. Overall, a SVM algorithm was able to make more accurate predictions of mock vs infected with the real data rather than data with random assignments. This increase in prediction accuracy indicates that it is possible to detect changes in the immune system from broad measures which summarize all of the data from antibodies interacting with an array of non-natural sequence peptides.

1.4.2.3 6 day time course

Significant differences in the AbStat measures for the mock vs infected groups could not be detected after 6 days of infection with this mouse experiment. The timeframe may have been too short. Another possibility may be that differences would be able to be detected on newer arrays with 330k peptides instead of 10k peptides. Our lab has also performed experiments on several versions of 10k slides over the years, and this experiment was performed on the earliest version of 10k arrays.

1.4.3 Human vaccines

The results from the human vaccines indicate that in general the entropy decreases shortly following vaccination. A decrease in entropy after vaccination was observed in individual 43 for the 2006 vaccination and in individual 84 for the 2009 vaccination ("Figure 11 Change in entropy after vaccination"). A decrease in entropy was not observed for individual 43 in 2009 3 days after the vaccination. This was the shortest time point observed after vaccination, and the data may have been collected too soon to observe a difference in entropy. In the one month trial, both individuals generally exhibit a decrease in entropy after vaccination, which then later increases back again ("Figure 12 Daily one month entropy change"). Individual 84 reported catching a cold on day 17, and there was a very dramatic drop in the entropy on day 27, which was just 10 days later. The entropy stayed low for one more day before returning back to the previous level. Note that no blood was drawn on days 23-26 due to the individual's illness. These observations indicate that changes in the overall fluorescence intensity distribution can capture the event of an individual catching a cold.

The observed decrease in entropy following vaccination is likely due to an increasing predominance of antibodies in the antibody repertoire against a single target present in the vaccine. The notion that the global AbStat measures could detect a change in the antibody repertoire makes some sense in relation to the amount of antibody a human body can produce to protect itself from an infection. The exact amount of antibody required for protection varies from disease to disease, but the amount of antibody against a target can often be as much as 0.1 mg

[77], and the Center for Innovations of Medicine has always identified a unique set of peptides for each disease [1, 9]. Note that after a vaccine there will be a decrease in entropy, but a higher mean. This higher mean is likely due to more peptides binding to the antibodies produced against the vaccine target.

1.4.4 Reduction in antibody repertoire complexity with lymphoma

During the course of a lymphoma, a single B cell proliferates uncontrollably and can produce a large amount of the single type of antibody which it produces. The rate at which a single B cell can produce antibodies is astounding, with 5,000-20,000 antibodies secreted per minute [1]. The B cell itself divides about every 70 hr, which would ultimately lead to an amplification of 10^{11} for a specific antibody in just one week. Altogether there are about $4*10^{16}$ IgG antibody molecules in one milliliter of blood [78]. Therefore, the complexity of the antibody repertoire is reduced as an antibody that binds to one target makes up an increasingly larger proportion of the repertoire. Note also that researchers in the Center for Innovations in Medicine have also previously demonstrated that an individual lymphoma signature can be detected for each dog as well as a general immunosignature in common among the dogs. The unique individual immunosignature consists of approximately 6 to 8 peptides (unpublished data). Therefore, the antibodies from the cancer B cell clone can be detected on the non-natural sequence peptide array. These facts provide an explanation for the ability of the AbStat metrics to detect a change in the dogs with a lymphosarcoma. The reduction in antibody repertoire complexity manifests itself as a reduction in entropy in the fluorescence intensity distribution acquired from application of the sera to a non-natural sequence peptide array (Figure 13). This reduction in entropy is caused by a narrower fluorescence intensity distribution, and this result is consistent with the results from spiking increasing concentrations of a single monoclonal into normal sera (“1.3.1.2 Spiking antibody into sera”).

The results from applying the SVM algorithm to the data indicates that there truly are unique characteristics associated with the fluorescence intensity distribution that help classify a sample as either normal or LSA. The SVM could correctly classify with an accuracy of 78.8%, but when

the classes were randomly assigned, the SVM could only correctly classify 52.3% of the samples which is the performance one would expect if one were to randomly guess the class of each sample. Note that the measure with the best p-value is not necessarily the measure with the highest SVM weight. For example, with this dataset entropy had the 2nd best p-value but the 8th highest absolute value SVM weight. Overall these results demonstrate that these measures can be used to classify a dog sample as normal or LSA with approximately 80% accuracy.

1.4.5 Mouse cancer progression

The behavior of the entropy of the fluorescence intensity distribution over time for the transgenic mouse and the normal mouse was dramatically different. The entropy of the transgenic mouse was somewhat sporadic as it continually increased for a short time, decreased for a short time, and then repeated ("Figure 14 Entropy time course for transgenic and wild type mouse"). For example the entropy increased from about week 15 to week 19, decreased from week 19 to 21, increased from week 24 to 27, decreased from 27 to 29, increased from 29 to 33, decreased from 33 to 39, increased from 39 to 41, and decreased from week 41 to 43. The entropy of the wild type mouse on the other hand constantly increases from week 15 to week 33 after which the value of the entropy basically reaches a plateau relative to the other points (Figure 14). One possible explanation for this behavior is that the immune system of the transgenic mouse is constantly battling the cancer and changing. At one time the antibody repertoire might consist of a large amount of antibody against a single target, and at another time there may be more general inflammation and many different antibodies with varying affinities against different targets. The wild type mouse on the other hand may just have a slow increase of general inflammation or simply just an increase in the number of antibodies with time resulting in a steady increase in entropy. The constant change in entropy observed in the transgenic mice as opposed to the steady increase in entropy of the normal mice may provide a method for detecting cancer early. Constant changes in the entropy value may indicate that a cancer is developing.

1.4.6 Human disease data

Samples from patients diagnosed with diseases could be classified with an accuracy better than chance based on the AbStat measures as sample attributes. The HT330K human disease samples could be classified based on the values of the measures with an accuracy of approximately 80%, the CIM10K samples could be classified with an accuracy of approximately 60%, and the Alzheimer's dataset could be classified no better than chance. This performance was not achieved when the class of each sample (disease or normal) was randomly assigned. Therefore, there are differences in the measures that are unique to disease which distinguish disease from normal samples. The poor performance obtained with the Alzheimer's dataset indicates that there is not enough resolution provided by these measures to distinguish between Alzheimer's patients and non-cognitive impairment controls of approximately the same age. As stated in the Alzheimer's results section ("1.3.6.4 Alzheimer's disease"), a classification accuracy of about 75% was achieved by using methods which select specific peptides instead of the global measure methods of the AbStat. Therefore, AbStat is not appropriate for all scenarios.

On the other hand, one of the interesting aspects of the classification of human diseases with the AbStat is that specific peptides with a high or low intensity for each disease were not identified. In certain situations, this could be an advantage since the classification is not dependent on the presence of peptides with specific sequences. Instead, the overall characteristics of the fluorescence intensity distribution summarized into single quantitative values were able to provide enough information for the classification performance observed. These characteristics were able to classify a sample as disease over a very wide range of diseases from a variety of different infectious diseases to a variety of different chronic diseases. Therefore, there appears to be a fundamental pattern in the antibody repertoire that applies to a very wide spectrum of disease states that is not present in normal states.

1.4.7 Changes with age

Does the antibody repertoire contain information about an organism's age? Can this information be captured and quantified? Differences in the fluorescence intensity distribution

resulting from the application of young or aged sera to a non-natural sequence peptide array are present. These differences were detected in both mice and humans. Using quantitative measures of this fluorescence intensity distribution from humans, it is possible to classify a sample as young (≤25 yo) or aged (≥50 yo) with about 80% accuracy. The algorithms can classify no better than chance when samples are assigned a class (young or aged) randomly. Note that the changes that occur with age are similar to the changes that occur with disease since both groups exhibit an increase in entropy.

What are other metrics of aging, and how is the entropy calculated from the data of antibodies reacting with a non-natural sequence peptide microarray a useful metric for aging? Everyone has a general notion of what "aging" is, but there can be disagreement about precise definitions. Chronological age can be precisely quantitated, but chronological age and biological age are two very different things. Aging can be generally viewed as the decline of health with increasing chronological age, as well as increased vulnerability to death [79]. Some definitions of health include "the capacity to adapt to changing external and internal circumstances", "wellbeing and not merely the absence of disease" from the World Health Organization, the opposite of "disorder of structure or function in an organism" from the Oxford dictionary for the definition of disease, and "the capacity to love and work" from Sigmund Freud [80]. As people age and health declines, many physiological process go awry: wound healing rates decrease [81], lean muscle body mass decreases, muscle strength decreases, progressive demineralization leads to decline of bone strength, neurodegeneration occurs, balance decreases, reaction times decrease, memory declines, sensory (vision, hearing, taste) abilities decrease, resting metabolic rate decreases, and homeostasis pathways consisting of hormones and inflammatory molecules provide less resistance to stress [80]. Leonard Hayflick stated "The fundamental aging process is not a disease but it increases vulnerability to disease" [82]. In recent decades, researchers have gained the ability to associate quantitative levels of molecular markers with age such as telomere length [83, 84], DNA methylation [59], damaged proteins [82, 85-87], hormones [88, 89], and markers associated with the following: genomic instability, epigenetic alterations, deregulated nutrient sensing, mitochondrial dysfunction, cellular senescence, and stem cell exhaustion [79]. The role of

the AbStat and the entropy measure among data from all of these other biological features and processes is to provide a metric of aging and health that is specifically related to the humoral immune system. Information from these metrics can be used not only to monitor "aging", but to monitor health, and make changes to slow down the decline of health or even maintain or enhance health.

One interesting characteristic of the machine learning results with the age data was that the SVM weighted the entropy measure the most (Figure 39), but the J48graft tree algorithm did not (Figure 40). The SVM algorithm achieved higher performance: 82.6% correctly classified by SVM, and 58.7% correctly classified by J48graft (Table 8). Also note that the SVM algorithm provides further evidence that age is the attribute that distinguishes fluorescence intensity distributions of the individuals since the algorithm was unable to correctly classify samples of two different nationalities: Chinese or Indian (Table 9). Another interesting aspect of this analysis was that the normalized entropy was not weighted by the SVM more than the entropy. Therefore, even though the samples were applied to arrays with a very different number of features, the SVM still weighted the entropy without normalization as the most important feature.

1.4.8 Rank and range of measures

The AbStat measures were ranked in four different ways. The entropy measure was ranked as the best measure with three of these ranking methods. With the fourth ranking method, entropy was ranked as the 3rd best measure. The following ranks after the 1st position are different depending on whether the rank is based on the p-value or the SVM weight. The entropy, normalized maximum, skew, and kurtosis perform well in p-value based ranks, and the entropy, dynamic range, and normalized minimum perform well by SVM rank. Therefore, the entropy, dynamic range, normalized minimum, normalized maximum, skew, and kurtosis are the measures which provide the most information when distinguishing normal, disease, and aged states.

The range of normal entropy values is around 7.80 +/-1. The disease entropy values have less variation and the values range from around 8.10 +/-0.60. Therefore, the disease/aged

range is enclosed within the normal range. The knowledge acquired from these datasets indicates that the higher the entropy, the more likely it is that a sample is actually a disease/aged sample or progressing towards this state. A normal sample with a very low entropy below the mean minus one standard deviation could exhibit this value due to a variety of different situations: 1. the individual is healthy and normal; 2. the individual has recently been vaccinated; 3. the individual has an acute infection such as a common cold; 4. the individual has a serious illness for which they are producing large amounts of high affinity antibodies against a single target; or 5. the individual has a lymphoma which is causing a few B cells to produce a large amount of antibody against a single target. In the cases 2-5, the expectation would be that the entropy would be low while the mean, on the other hand, would be higher than normal since antibodies would bind to a few peptides with high intensity. Therefore, the mean is another important measure to monitor with time.

1.4.9 Quantitative analysis of the entropy measure

Several conclusions can be made from the results of the quantitative analysis of the entropy measure presented in "1.3.10 Quantitative analysis of the entropy measure". A quick summary is that disease specific peptides are more important to the entropy measure than randomly selected peptides. The "maxed out" fluorescence intensity peptides with the highest possible intensity dilute the entropy calculation. The distribution of low intensity peptides is very important for the entropy measure, particularly since the majority of the distribution consists of low intensity peptides. Another conclusion is that the sensitivity of the entropy metric generally decreases with increasing bin size, but there are exceptions. A more thorough discussion is presented in the subsections below.

1.4.9.1 Analysis of important peptides

Analysis was performed to identify the types of peptides in the distribution which contributed the most to the ability of the entropy measure to distinguish between healthy and disease groups ("1.3.10.1 Changes in entropy measure with removal of peptides"). The data

used for the analysis consisted of 5 breast cancer samples and 24 normal samples. The entropy of each sample was calculated, and a t-test was performed with the two groups. Before the entropy was calculated, however, various numbers of peptides of a given type were removed. If non-natural peptides which bound specifically to disease sera were more important than randomly selected peptides, then fewer specific peptides could be removed before the significance of the p-value started to decrease. Here specific peptides referred to peptides that had the most significant p-value in a t-test with the intensities of that peptide in breast cancer and normal samples. The results showed that specific peptides are indeed more important than random peptides since the p-value with entropy between the two groups started to become less significant after about 5,000 peptides have been removed ("Figure 46 P-value vs peptides removed"). With random peptides, on the other hand, about 230,173 peptides must be removed before the p-value starts to become less significant. A targeted removal of peptides that are most different from breast cancer and normal caused the entropy measure to lose the ability to distinguish between the groups more than removal of random peptides.

What does this result mean relative to standard immunosignaturing techniques? In the Center for Innovations of Medicine, the standard immunosignaturing technique involves performing a t-test for each peptide in all of the samples, and selecting the peptides which are specific to the disease with a p-value more significant than a certain cutoff value. These peptides are then used with a classification algorithm such as naïve Bayes to classify samples as normal or disease samples. The number of peptides selected to be input into a classification algorithm is typically in the range of about fifty to a few hundred peptides. However, this entropy analysis demonstrated that the distinguishing power of the entropy measure for normal and disease samples did not drop off until about 5,000 specific peptides had been removed as stated previously. Perhaps it would be far too extreme to state that these 5,000 peptides are "disease specific". However, this result may mean that these 5,000 peptides are more similar to mimotopes of the disease than other randomly chosen peptides, but not that these peptides actually are mimotopes. Let's assign some hypothetical numbers to illustrate the point. The top 20 peptides may have a similarity score of 90% to a cognate epitope of the disease, whereas a

peptide that ranks as the 4,500th most specific peptide may have a similarity score of 15% to the same cognate epitope. Alternatively, these 5,000 peptides may not have any similarity to disease specific epitopes at all, but rather they may exhibit non-specific avidity interactions with the antibody repertoire of a disease sample. These avidity effects may not be manifest by other randomly selected peptides on the array due to their level of non-specific binding or general characteristics such as hydrophobicity or hydrophilicity. In conclusion, there appears to be more information present than can be captured from selecting highly specific array features alone.

What is the actual p-value of the most significant peptide when the p-value with the entropy between the groups begins to become less significant? The p-value with the entropy becomes less significant than a p-value of 0.1 after 100,000 peptides have been removed. At this point the p-value of the most significant peptide is about 0.15 (Figure 47). Therefore, in this dataset, peptides with a p-value more significant than 0.15 seem to contribute the most to the ability of entropy to distinguish between groups.

Removal of the highest intensity peptides yielded some interesting results as well ("Figure 46 P-value vs peptides removed"). Here different peptides were removed from each sample since the peptides were removed in order of highest intensity unique to each sample. Naturally, there will be some similarities. Many of the top 1,000 peptides in one sample will be in the top 1,000 peptides of another sample, but there will not be a 100% overlap due to natural variation. In the analysis, when the top 5,000 peptides were removed, the p-value actually becomes more significant than when all of the peptides were present. This result makes sense because many of the top several thousand peptides were maxed out at 65,535 in this dataset. Therefore, these peptides which had exactly the same value in both groups was diluting the calculations so that there was less of a difference in entropy between the two groups. Once these maxed out peptides are removed, then the p-value becomes more significant than it was with the full 330k distribution. Next the p-value between groups with entropy becomes extremely significant with the remaining very low intensity peptides. This makes sense since a fairly large group of peptides will have intensity values near zero, and the distribution of these near-zero

peptides will have a different shape in normal and breast cancer samples. Figure 46 shows the graph zoomed into the portion with 5,000 or less peptides remaining.

In another analysis, the least significant peptides, rather than the most significant, were removed from the distribution. In this experiment, the p-value of the entropy between the groups becomes more significant after the 100,000 least significant peptides have been removed, and then the p-value drops off to become less significant (Figure 49). The increase in significance observed was not as great as the significance obtained when removing the highest intensity peptides. What is the actual value of the p-value of the least significant peptide when there is an increase in significance with entropy? The p-value of the least significant peptide when 100,000 least significant peptides have been removed is about 0.25 (Figure 50). Therefore, removing peptides with a p-value greater than 0.25 improved the ability of entropy to distinguish between the two groups.

In summary, peptides that are different by a t-test with fluorescence intensity contribute the most to the ability of the entropy to distinguish between the two groups. The highest maxed out peptides do not contribute much to the ability of entropy to distinguish. Also, the shape of the group of peptides in the very low range is very important for the ability of entropy to distinguish between groups.

Many questions can be asked from these results. Would the set of low peptides that allow entropy to distinguish well be mostly the same or different for different diseases such as breast cancer and lung cancer or syphilis? If another breast cancer dataset was analyzed would the low peptides be mostly the same or different than the original breast cancer dataset? In other words, is this distribution of low peptides caused by a type of specific low binding unique to a disease or set of diseases, or is this distribution of low peptides mostly random. If it is random, then non-specific antibodies may be binding randomly to peptides on the array based on their location at the time and whether they happen to interact with a certain peptide somewhat due to chance. Technical replicates are usually performed in the lab and high correlation coefficients above 0.95 are frequently obtained. This means that a peptide with a given intensity in one replicate will have a very similar value in another replicate. Therefore, it would not be appropriate

to describe the binding to the array as “random”. However, there may be more or less variation in certain intensity ranges. For example, array features with a high fluorescence intensity may often be due to binding with high affinity antibodies, and the standard deviation of the fluorescence intensity of these array features may be small e.g. 20 intensity units. Binding to array features in the lower intensity range (<1,000) may be due to lower affinity antibodies and the standard deviation of the fluorescence intensity of an array feature in this range may be higher e.g. 150 since the binding could be more random and chaotic in this range. In this situation, technical replicates could still have a very high correlation coefficient, but the degree of variability in the signals could be different in different intensity ranges.

1.4.9.2 Change in entropy with change in bin size

One of the main parameters of an entropy calculation is the bin size. The analysis in section “1.3.10.2 Change in entropy with change in bin size” demonstrates that in general the best bin size is the smallest bin size. This makes sense since smaller bin sizes will allow for increased resolution. As the bin size was increased, the ability of the entropy metric to distinguish between disease and normal samples decreased. However, there were some exceptions since the p-value did become more significant near a bin size of 5,000. The distributions between disease and normal exhibited the largest difference when a bin size of 5,000 was used with this particular dataset, but different datasets may have a different optimal bin size. A bin size of 5,000 is likely optimal because the majority of the distribution at the low intensity range would fit into or not fit into the first bin depending on whether the sample was normal or disease. A quantitative examination of this dataset supports this conjecture. In the five breast cancer samples present in this dataset, 46.7 +/-18.1% of the features have a fluorescence intensity less than 5,000. However, an analysis of five randomly selected normal samples reveals that 69.7 +/-4.87% of the features have a fluorescence intensity less than 5,000 due to a distribution with a high tight peak less than 5,000. Therefore, choosing a bin size of 5,000 places most of the feature intensities in the normal samples in the first bin, and splits the feature intensities in the breast cancer samples to a greater degree. This leads to a difference in

calculated entropy. Interestingly, the total range of p-values did not change dramatically with the wide range of different bin sizes tested. The p-values ranged from 0.0195 to 0.117, which is a smaller range than one might have anticipated.

1.4.10 Limitations of AbStat

The AbStat measures clearly have limitations since some samples could not be distinguished from normal. For example, the Alzheimer's samples could not be distinguished from the controls, but these samples could be distinguished when using other classification techniques ("1.3.6.4 Alzheimer's disease"). For many other diseases there was also very poor separation from normal. For example, most of the *Bordetella pertussis* (BPE) samples fall within a normal range (Figure 15). The classification accuracy of the AbStat measures is also typically low even though it is better than chance, in the range of 60-80%. Higher accuracy can be obtained with analysis methods that focus on specific peptides. Note, however, that the classification accuracy of the broad global measures that compose the AbStat improved when analyzing samples on the HT330K platform as opposed to the CIM10K platform. This result suggests that with more peptides and higher quality slides made possible with more advanced technology, the power of the AbStat measures may improve with time.

Another clear limitation of the AbStat is that it cannot aid a researcher in identifying specific antibody targets. The measures used in the AbStat are broad global measures that take all of the fluorescence intensities and compress them into one single numerical value. All amino acid sequence information associated with an array feature is discarded. Therefore, it would not be possible to use mimotopes on the array to trace back and identify the cognate epitopes of the antibodies in the sample. These mimotopes and epitopes are critical for biological discoveries, diagnosis, and vaccine development. Although the AbStat measures cannot be used to study biological sequence information, they should be applied to scenarios that they do provide information for: monitoring overall health states associated with disease and age, studying the overall characteristics of an antibody mixture, and making enlightening connections between diseases. More detail on the value of AbStat is provided in the next section.

1.4.11 Value of AbStat compared to other classification methods

What is the value of the AbStat measures when compared to other classification methods? In the Center for Innovations in Medicine the typical procedure for classifying samples is to perform a t-test for all of the features comparing the intensity of each feature to the intensity of that same feature obtained with other samples. Features which yield a p-value that is more significant than the cutoff value are then used as the features to input into a classification algorithm such as naïve Bayes to determine whether a sample can be classified as normal or as a certain disease. The p-value cutoff point used to select the significant disease specific features is often the reciprocal of the number of features. This process usually results in significantly better classification performance than can be obtained with the AbStat. Note, however, that the specific peptide analysis approach is also more susceptible to overtraining as presented in the "1.3.6.1 HT330K first chip disease dataset" section. Therefore, it is important to use caution with this approach, and analyze the data with appropriate training and test data. Nevertheless, since it is possible to obtain better classification performance with specific peptide analysis methods, one can ask the question: What is the use of AbStat when more accurate classification methods exists?

The benefits of AbStat are found not with its classification ability alone, but rather with many other features which other methods do not provide. First, the AbStat measures are broad global measures which do not rely on selecting specific features from the rest of the features. Using other methods, if certain disease specific peptides were removed or not originally included on an array, then classification performance may decrease significantly. However, the classification performance of the AbStat measures would likely remain about the same as long as enough non-natural sequence peptides selected from random space were included. Additionally, many of the AbStat measures are unaffected by normalization, whereas other classification methods would not work as well without normalization. Since there can be disagreement about which normalization techniques are most suitable for various situations, the fact that normalization does not even affect some of the AbStat measures could be perceived as an advantage in certain situations.

The AbStat methods can also detect differences between young and aged sera. The overall specificity of the antibody repertoire seems to change with age and these changes are reflected in the AbStat measurement. These AbStat measurements may ultimately allow individuals to mitigate the effects of age as they learn which factors exacerbate or ameliorate changes in the AbStat measurements such as entropy.

Another benefit of the AbStat measures is that they connect several different health states together, and some of these connections may not have been previously apparent. For example, the entropy measure of the AbStat increases to higher than normal levels for both chronic diseases and old age. The general public would probably often associate chronic disease and old age with poor health, but the AbStat provides another quantitative means of directly comparing these two states of health. This association also suggests that in both cases the antibody repertoire may become more complex and "chaotic", and this can reveal something fundamental about both states. As another example, the entropy measure of the AbStat decreases for both vaccines and lymphomas which produce large quantities of a single antibody. Before this finding, one might not see much of a similarity between these two different states. However, the AbStat measure can find some clear commonalities between these two seemingly unrelated states. The vaccine and lymphoma states are conceptually connected by the fact that they both undergo a decrease in repertoire complexity by increasing the proportion of a particular antibody against a specific epitope or a few epitopes. Therefore, the AbStat measures can help reveal truths and connections which were not previously obvious.

Not only can the AbStat make connections between different health states, but the AbStat can also make connections between different fields of study. One of the critical measures of the AbStat is entropy which is used in the field of physics and thermodynamics as well as information theory. Therefore, now entropy can be used in biology, physics, and information theory, and the knowledge gained from the other fields may be useful in biology as well. For example, in physics and information theory entropy can reflect the degree of randomness in a system, and this could be true in biology as well. Additionally, in thermodynamics, entropy must always increase as the universe moves towards disorder, and scientists and philosophers have

claimed that this provides the universe the arrow of time in which we operate [90]. The entropy measure of the AbStat also demonstrates that the entropy of the antibody repertoire interacting with an array of peptides also increases with time/age. Many individuals often have the general notion that the entropy of biological systems increases with age [27], but the AbStat entropy provides some quantitative evidence for this impression. Therefore, the AbStat is important because it ties several fields together, and this measure may lay the groundwork for future discoveries and insights.

1.4.12 Use and possible future applications

1.4.12.1 Overview of use and possible future applications

Since these AbStat measures have proved useful for identifying interesting immune states associated with chronic disease, infectious disease, vaccine, and age, individuals may be interested in monitoring their AbStat constantly over time. By constantly monitoring this information, one might be able to catch the progression of disease early and take action to mitigate the effects of disease. Researchers with access to large databases of this data can also identify new trends which can be used to inform the public to make healthy decisions. Basically, the AbStat can provide people with another metric that allows them to evaluate their fitness more quickly and make adjustments as they deem fit for optimum health. This metric could accompany many others in the emerging Quantified Self revolution [91].

In the scenario that people are constantly monitoring their AbStat measures it would be useful to have a simple method for conveying the AbStat information to the individual who may not be an expert on these topics. Two different approaches could be suitable for this task. One approach is to simply plot one of the most informative measures, entropy vs time with lines for the population mean +/- one standard deviation so that the individual can compare their state with the population norm. The individual may also just prefer to view their entropy in relation to other people of their age group or even only their own past data. The key here is to keep it simple, and to allow the individual to obtain more information if they desire. The second approach is basically the same as the first. However, instead of plotting entropy on the y axis, the distance to the SVM

boundary would be plotted. As the data throughout this document was analyzed, all of the measures were input into an SVM, and the SVM used some training data in order to draw a linear boundary in a high dimensional space which can separate as many samples of two different classes as possible. Therefore, the distance to the SVM boundary would incorporate information from all of the measures rather than entropy alone, and the information acquired from the training data would also inherently be included. The larger a positive value was the more likely it would be that an individual is in a disease state, and the larger a negative value was the more likely it would be that an individual is in a normal state.

AbStat could also be used to quantitatively characterize monoclonal antibodies. The results in "1.3.1.1 Monoclonal affinities and AbStat Measures" indicate that there is a relationship between monoclonal antibody affinity and entropy. Lower affinity antibodies appear to bind more non-specifically to the array. Companies which produce monoclonal antibodies often do not describe the specificity of an antibody, and often don't even determine the dissociation constant for the affinity. However, Dr. Johnston had the idea that the entropy measure obtained from reacting a monoclonal antibody with a microarray of non-natural sequence peptides could be used to quantify the specificity of an antibody. This could result in higher quality antibodies as companies produce antibodies to optimize this metric. Note that there may be a relationship between affinity and specificity because of the rigidity of the structure of an antibody. Antibodies with little specificity for one unique target (termed polyreactive antibodies) often have more flexibility in their structure [92]. On the other hand, an antibody with high specificity and high affinity may have a more rigid structure which does not allow the antibody to bind many other sequences that are not similar to the cognate epitope [53, 93].

The concepts, metrics, and data analysis presented in this document could have much wider implications than a new metric for the immune system. In particular the entropy measure may prove interesting in other situations, but may be aided by other measures such as the mean just as it was in this document. Anytime one can obtain a distribution of numbers, one could test the entropy which would essentially be the same as asking how easily a random number generator could reproduce the frequency of values observed. This would test how random a

distribution is. Entropy has already been used to demonstrate that the entropy of the alternative splicing distribution is higher in cancer cells than normal cells [30, 94]. Entropy has also been used in many other situations as reviewed in the "1.1.4 Entropy with previous biological data" section. Using quantitative values to characterize the shape of a distribution, and then using those values to assess the vitality or stability of a system, is not limited to biology either. For example, the statistician George Zipf made the argument that certain levels of heterogeneity in the income distribution of a country can precede revolutions [95].

There are many other datasets which one could apply this type of analysis to. For example, the AbStat of mass spectrometry data of all of the proteins in a cell may be different between healthy and disease cells since diseased cells may have different numbers and types of proteins, and this difference may be quantifiable in the entropy measure. An analysis of brain scan data in which the number of connections of each neuron was determined could allow one to calculate the entropy of the neuron degree distribution, where degree refers to the number of connections of each neuron. Individuals diagnosed with Alzheimer's disease or insanity may have distinctly different entropy values than individuals who have not been diagnosed with Alzheimer's disease or insanity, just as schizophrenic patients exhibited abnormally low entropy fMRI data compared to normal patients [38].

Once one begins to apply the entropy calculation to assess the vitality of systems represented as networks, such as the brain, the situation could become extremely fascinating for two reasons: 1. Many systems can be represented as networks from genetic networks, protein networks, metabolic networks, social networks, airplane networks, software since any software can represented as a network, etc.; and 2. There is a great deal of existing network science that can be applied to understand such networks. Theoretically, one could apply the entropy calculation to all of these networks and associate certain entropy values with well-functioning versions of these networks and certain entropy values with poor functioning versions of these networks. So is the node degree entropy of a real-world network lower than a random version of the same network? The answer to this question is yes. A quick calculation of the entropy of a real world network, which turns out to be a small-world scale-free network [96], such as the network

of airport connections in the United States [97] revealed that this network had an entropy value of 5.03 whereas a random version of the same network had a value of 5.76 (personal calculation). These type of real world networks are “small” because the average distance between any two nodes is small. These networks are “scale-free” because the basic shape of the network looks about the same at small or large scales so there is no scale. Note, however, that quite different results are obtained with the node degree entropy for networks that are not scale-free such as road and power-grid networks which are exponential networks rather than scale-free networks. Scale-free networks have a node degree distribution that obeys a power law where very few nodes have a very high number of connections, and exponential networks have a node degree distribution that obeys a Poisson distribution in which most of the nodes have the average number of connections.

The lowest node degree entropy for a network occurs when the network forms a ring so that every single node has exactly the same number of connections (two connections), but this network certainly does not have the form that actual real world networks exhibit. These real world networks are also ordered and non-random. A ring network is also extremely inefficient at passing messages to different nodes since the average shortest path is very high (one must pass through every single node to reach a node half-way around the ring). Networks that represent living systems fall into the category of small-world scale-free networks as mentioned previously, and these networks have a definite complexity and organization to them that is not random. Perhaps the best metric or version of an entropy metric for a network would be one that is low for very simple “boring” networks such as a ring, low for random networks, and high for a small-world scale-free network. Other methods for calculating the entropy of a network as opposed to node degree entropy have already been investigated such as search information entropy, target entropy, and road entropy [98, 99]. Other network measures such as the average shortest path length and the clustering coefficient may also be very important. A metric which is the highest for a very well organized efficient network, may send an alarm when the network changes and the value for this metric drops off indicating that the network has become more chaotic or less optimal. This event may indicate that the complex adaptive system which the network

represents, whether it be a human brain or a system of airport connections or even software, is beginning to lose its proper function. The essential idea here is that one can obtain a distribution of numbers that measure something about a complex system, calculate a value from that distribution, and then have knowledge about the vitality of that system.

Once one knows how healthy or sick a system is, what good is this information? The primary use of this information is to then take action to mitigate damage to the system and to get it functioning optimally. For example, one could predict from the results in this document, that if one were to immunize an aged person and a young person against the same target that neither of their immune systems had ever seen, then the resulting entropy of the fluorescence intensity distribution would be higher for the aged individual. The explanation would be that the aged individual has an immune system that does not function as optimally, and the aged individual produces antibodies with a lower affinity to the target. These less specific antibodies then bind more randomly (not completely randomly by any means) to a variety of different peptides with varying intensities which results in a higher entropy.

However, what if there were actions one could take to mitigate the increase in entropy in older individuals? How could this be done? The immune system is a complex adaptive system with similarities to many other complex adaptive systems such as a brain or a collection of muscle cells. These systems stagnate when there is nothing to adapt to. However, when they are constantly challenged, a brain or a muscle will respond to that challenge and maintain better performance for future challenges. Perhaps the immune system could benefit from constant training or challenge as well. Perhaps an aged individual who is vaccinated against targets throughout their lifetime will be able to produce more specific antibodies than an aged individual who has an immune system that is rarely used. Whether or not these speculations are true, the entropy of the peptide distribution provides a quantitative way to test these predictions.

If training or exercising the immune system is beneficial, another fascinating topic to investigate is what is the nature of the optimal training? For any complex adaptive system, the challenge needs to be at an appropriate level for the ability of the system to currently handle that challenge. Additionally, the challenges need to vary with an appropriate frequency and level.

Someone does not lift the heaviest weight in a weight room if they have never lifted a weight before. Additionally, people training with weights do not lift the heaviest weight they can possibly lift at all times every day and do nothing else; they often perform some type of warm-up and/or cool-down, and often don't lift the heaviest weight they can every day. Challenging a system too harshly will damage it, and challenging a system too little will lead to stagnation so that the system will be unprepared for future challenges. What are the mathematical truths that underlie the optimal training and maintenance of a complex adaptive immune system, and how can these concepts be applied to the immune system? Although the answers to these questions are unknown, the entropy of the fluorescence intensity distribution provides a method for quantitating the results of any experiments to address such questions.

1.4.12.2 Immune system training hypothesis with mouse experiment

Let's discuss a simple experiment one could perform to test the hypothesis that exercising and stimulating the immune system will slow the progression of AbStat entropy. One could immunize one group of mice frequently over the lifetime of a mouse, and then the sera of these mice could be applied to the peptide microarrays to determine the increase in entropy overtime. This entropy time course would then be compared to a control group of mice which only received PBS (phosphate buffered saline) injections. At the end of the time course, both groups would be immunized with an antigen neither group had ever been exposed to. If stimulating and exercising the immune system truly does result in a more "fit" immune system capable of producing higher affinity antibodies to new targets, then the expectation is that the frequently immunized group would exhibit a lower entropy than the control group after both groups were immunized with a new antigen. Such a result would imply that frequent immunization would be a method for slowing down the decline of immune system health with age. This result would support other studies which demonstrate that exercising muscles or memory can slow the deterioration of these systems with age. Note that exploring the frequency of immunization, quantity of antigen, and diversity of antigens immunized could also be interesting

as well. How much is too much and how much is too little in regard to obtaining the goal of improved immune function?

1.4.12.3 Immune system training hypothesis with rural and non-rural individuals

Another strategy for testing the immune system training hypothesis would be to compare the AbStat from immune systems of individuals from rural areas and individuals from large industrialized cities. The phenomena that individuals from industrialized countries develop more allergies than those from developing countries has led to the hygiene hypothesis. The hygiene hypothesis states that the industrialized country individuals were not exposed to infectious agents, parasites, and microorganisms during childhood, and therefore their immune systems did not develop properly [100-102]. In addition to an increase in allergies often associated with a Th2 immune response, there is also an increase of Th1 immune response autoimmune diseases in industrialized nations [103]. The exact cause of this phenomena is still debated, but the fact that there are fewer of these types of immunological disorders in developing nations is a statistical fact. Therefore, perhaps the AbStat approach can detect a difference in the immune systems of rural and non-rural individuals. This strategy for testing the immune training hypothesis has advantages over the mouse immunization approach since it is not known whether frequently immunizing a mouse can enhance the immune system, but it is already known that individuals from rural areas have fewer allergies and autoimmune disorders on average. In fact, with the mouse approach, frequent immunizations may actually be harmful, since this non-natural intervention may cause the immune system to become imbalanced in some way.

The AbStat measures may be able to classify individuals from rural or non-rural areas. One might predict that the rural individuals typically have a lower AbStat entropy than the non-rural individuals since lower AbStat entropy is typically correlated with absence of disease, with several exceptions (shortly after a vaccine, the occurrence of a cold, and a lymphoma). To be more precise, one might expect rural individuals to have a low AbStat entropy as well as a low AbStat mean compared to non-rural individuals. AbStat numbers that are more associated with normal healthy states may be found more often in rural individuals than non-rural individuals, and

these better AbStat numbers may result from the increased immune system training that the rural individuals had been exposed to.

1.4.12.4 AbStat and overall health hypothesis

One hypothesis could be that the AbStat entropy of individuals that are extremely healthy may be lower than the AbStat entropy of individuals who have just not been diagnosed with a disease. If this hypothesis were true, then the public would have a reason to monitor their AbStat because they would know that it would be one more metric of their general health. Not only would it be one more metric of general health, but it would also be one specifically associated with the immune system which people often do not monitor currently. Throughout this dissertation, "normal" individuals were simply not diagnosed with a disease, but it is natural to assume that there is a whole spectrum of health and disease states. Some "normal" people may be significantly healthier than other "normal" people. Additionally, general metrics of health may also often be correlated with the health of various specific systems since the whole body is connected together, and the health of one system will often impact the health of another. Therefore, it could be very interesting to correlate the AbStat measures with various other metrics of general health such as body mass index, resting heart rate, blood pressure, VO_2 max, etc. Another measure could be a score of health from a panel based on photographs of individuals. Although this method may seem unscientific and not quantitative at first, at least one study has already found that the perceived age of twins from photographs correlated significantly with time of death, physical functioning, cognitive functioning, and leukocyte telomere length [104]. The prediction with AbStat entropy is that the entropy would be lower for healthy individuals compared to less healthy individuals of the same age. Health would be correlated with various metrics of general health. Such a finding would provide a motivation to monitor and adjust one's AbStat.

Note that it may be necessary to immunize individuals in the study against a foreign antigen to create a strong immune response. Individuals with better overall health may have healthier immune systems, and they may produce higher affinity antibodies against new targets than less healthy individuals. Although there could be a difference between the two groups

without immunization, the difference may be too small to detect with statistical significance. I believe that this sub-hypothesis may hold true for the immune system training hypothesis with rural and non-rural individuals as well.

1.4.13 Conclusion of use and possible future applications

This discussion can be concluded by revisiting the practical applications of AbStat. The immediate use of the AbStat could be for frequent monitoring of health and for classification/diagnosis purposes. Individuals could obtain their AbStat on a regular basis, and choose how they want to interpret the information themselves given the normal range of entropy for all other individuals tested. A simple line graph could display their entropy over time, and they could correlate this with personal events in their own lives such as “This timepoint is when I started this new exercise routine”, “This timepoint is when I received an immunization”, and “This timepoint is when I caught a cold”. Alternatively, they might observe a strange sudden drop in their entropy which is not correlated with any life event that they are aware of. This information could prompt them to see a doctor, and further investigation might lead to the diagnosis of an illness which can be treated more effectively since it was noticed early. Another practical application of the AbStat is to use it as a metric in tests of future hypotheses such as the immune training hypothesis mentioned previously. Therefore, the practical value of AbStat lies in monitoring, diagnosis/classification, and as a metric used for testing new hypothesis about the immune system.

1.5 **Conclusion**

A group of measures of a fluorescence intensity distribution produced from reacting antibody containing sera with a non-natural sequence peptide array result in an AbStat that can distinguish healthy and disease as well as young and aged states. One of the key measures of the AbStat is entropy, which has been used to determine how random a system is since the 1800s. Scientists used this calculation in the field of thermodynamics to measure the chaotic state of particles, they used it to measure the chaotic state of information, and now it is being used to measure the

chaotic state of the humoral immune system. The quantitative results from measuring the chaotic state of the immune system has now been associated with the level of health of both mice and humans. Chronic disease, infectious disease, the common cold, and vaccines are all events that the AbStat can detect. The results obtained with AbStat can be used for practical purposes such as diagnosis and monitoring for many different diseases, and the ultimate results may be actions people take to maintain health. The AbStat and the entropy measure have yielded interesting philosophical questions and practical applications.

2: PLATFORM FOR SCREENING CDNA LIBRARIES

2.1 Introduction

2.1.1 The need for better approaches against cancer

Cancer is the second leading cause of death worldwide [105]. Conventional treatments for cancer such as surgery, chemotherapy, and radiation have harsh side effects and usually slow progression but do not cure disease. There is now a newer class of treatments comprised of monoclonal antibodies such as Herceptin, which recognizes the breast tumor-associated antigen HER2/neu [106], and Rituximab which binds the lymphoma associated antigen CD20 [107]. These immunotherapies have the significant benefit of being far more specific than conventional approaches, but the disadvantage of their high cost and requirement for continuous re-administration. Additionally, only HER2+ or CD20+ tumors are responsive to these single surface-marker targeted treatments. Only 20% of breast cancers are HER2+ [108], and the most aggressive cancers are actually HER2- [109]. Furthermore, even receptor positive tumors are responsive initially, but can develop escape mechanisms, and patient tumors recur.

An even less conventional and recently emerging approach to treating cancer is vaccination. Vaccines are typically considered for battling infectious pathogens, not one's own cells. However, at least in theory the immune system could be capable of learning to recognize the patterns that set tumor cells apart from normal self-cells. Not only does a cancer vaccine theoretically make sense, but cancer vaccines have been demonstrated to work in animal models. For example, vaccination of rats with a tumor-associated antigen formulation significantly delayed growth of chemically induced tumors [110]. Injecting rats with lysates from whole tumor cells was also able to provide some level of tumor growth protection [111]. Human self-tumor mixtures in which tumor lysates are prepared from patients' own tumor are also being tested in clinical trials [112]. While self-tumor mixtures hold potential, preparing these tumor mixtures from tumor-bearing patients is often not feasible relative to both disease timeframes and costs. In addition, effectiveness is compromised by the very fact that disease must be sufficiently

progressed such that a tumor can be detected and extracted. A more attractive cancer vaccine that is economically and technically more practical might consist of a subunit vaccine containing tumor-specific antigens to be administered prior to tumor development. Note that the feasibility of component vaccines, instead of whole pathogen vaccines, has been demonstrated [113-116]. In order to develop an effective subunit cancer vaccine, prior knowledge of tumor antigens is needed.

2.1.2 Methods for discovering tumor antigens

Identifying tumor specific antigens can be tackled using a variety of methods. One could sequence the DNA of cancer and normal cells to search for genomic differences [117]. Alternatively, a researcher could examine the RNA transcripts to look for gene expression differences [118]. Another method is to actually use the immune system of the patient to discern antigenic differences. Researchers have demonstrated that both T and B cell based immune responses are generated against tumors [119, 120]. The presence of lymphocytes in a tumor is associated with a better patient outcome [121, 122, 123, 124]. While many of these immune responses are cell based, there are also a great diversity of humoral responses to tumors. Antibodies produced against tumor proteins make robust biomarkers since antibodies can persist in the body longer than a tumor protein biomarker, and antibodies are continuously amplified in the presence of continued immunogen. Significantly, tumor specific antibodies can even be elicited before disease symptoms are manifest.

Many experimental methods to probe tumor proteins with antibodies have been developed. Before the advent of high throughput methods, researchers could only probe a few antigens at a time using assays such as 1D SDS/PAGE and ELISA. Later on in 1995, the serological analysis of recombinant cDNA expression (SEREX) libraries was developed [125]. Briefly, the method involved producing a tumor cDNA expression library, lysing clones on a nitrocellulose membrane filter, and then probing this membrane with autologous sera. Using this approach a large number of antibody and recombinant tumor protein interactions can be assayed at one time as compared with the single antigen assays. Nevertheless the SEREX method does have some disadvantages such as the fact the proteins must be capable of being expressed in

bacteria, proteins with the most transcripts in the tumor cell may be detected most frequently while tumor associated antigens of low abundance may be missed, and there will be antibodies to non-tumor associated antigens. Most vexing has been the poor reproducibility of the process and the time consuming and labor-intensive protocol (Figure 53).

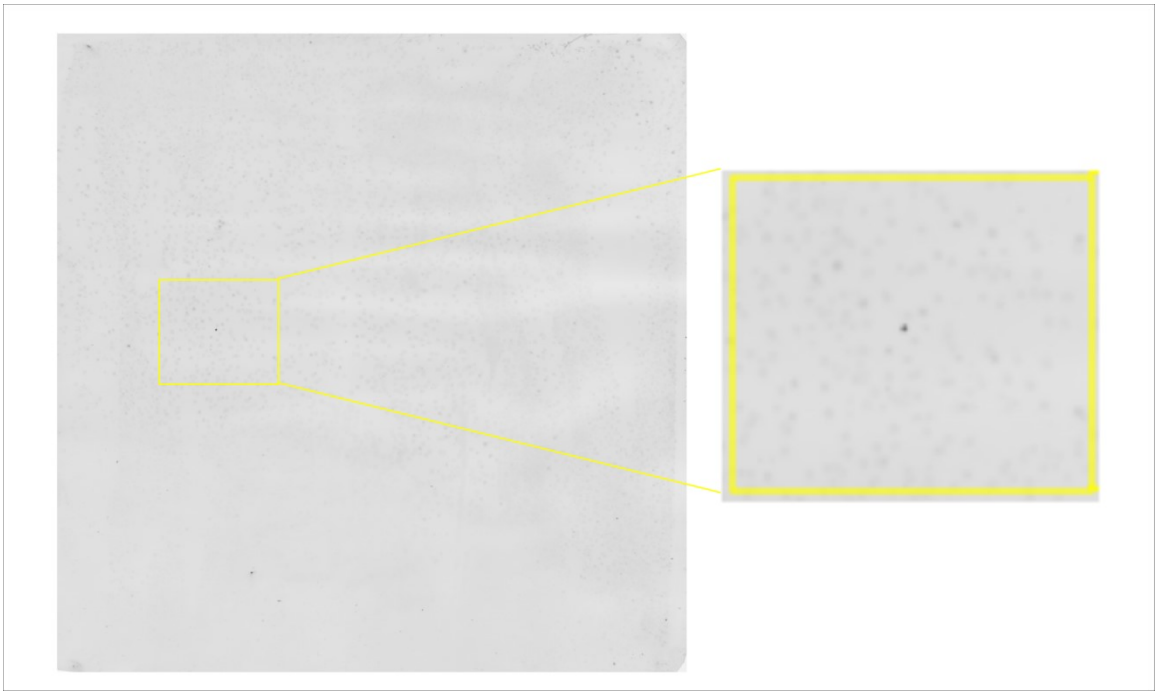


Figure 53 Example SEREX nitrocellulose membrane

This 245X245 mm nitrocellulose membrane was pressed against a plate of colonies, the colonies were lysed, and then the membrane was probed with sera by myself. One can imagine that tracing back from an antibody-lysate spot signal to the original colony on an agar plate would be difficult since the colonies are not arrayed out in a set pattern.

Microarrays overcome some of the limitations of SEREX. For example, recombinant proteins are systematically organized onto a small slide at addressable locations, rather than spread out randomly across a large membrane. The slide printing arrangement allows for a small sample amount to be sufficient. The microarray may be printed with known or unknown proteins. The biased approach contains known proteins. For example, Invitrogen manufactures a ProtoArray containing approximately 9,000 proteins which researchers can probe with sera [4]. The disadvantage of this method is that new immunogens, such as undiscovered mutated

proteins that may not be present in current databases, cannot be discovered. Alternatively, recombinant phage or cDNA library cell lysate can be spotted onto a slide. This method allows for the discovery of previously unknown proteins since any protein expressed by the original tumor can be in the library. The benefits of this microarray approach include its compatibility with high-throughput protocols, reproducibility, small reaction surface, even sera distribution across the surface, high dynamic range of signal intensities, and the quantity of data that one experiment can produce [126]. The disadvantages currently involve the cost, and the fact that so much data are produced that more sophisticated data analysis software must be used to interpret the results.

Serological proteome analysis (SERPA) is another method for probing tumor proteins with antibody containing sera [127]. In this method, 2D electrophoresis is first performed to fractionate the lysate's tumor proteins. A western blot is then performed to detect binding events between sera from the tumor-bearing host and the tumor proteins. Tumor proteins of interest are then extracted from the gel and the identity of the protein is determined by mass spectrometry. This method does not require the construction of a cDNA library. Post-translational modifications can potentially be detected and heterologous expression is not needed since the tumor proteins are directly assessed. Unfortunately, this method suffers from limiting amounts of material, poor reproducibility, and the low resolution of 2D electrophoresis.

2.1.3 Known cancer immunogens

The experimental methods described above have discovered >2,500 tumor proteins, some of which may be immunogenic. Many of these proteins are stored in the Cancer Immunome Database [128]. An analysis of these proteins reveals that there are many different categories of non-normal proteins recognized by the immune system of a tumor-bearing host. Well-studied tumor antigens include cancer testis (CT) antigens, heat shock proteins [129], and oncoproteins such as HER-2/Neu and p53. These have been identified as normal cell antigens that are overexpressed in tumor cells [130]. Viral antigens such as antigens from the human papillomavirus [131], and cell differentiation antigens such as RAB38, and NY-BR-1 have also been identified [94].

Other categories of tumor antigens are the result of frameshift mutations due to microsatellite regions in the genome such as the CDX2 antigen in colon cancer [119].

2.1.4 Ideal cancer vaccine

Although many cancer immunogens have been discovered, other criteria need to be satisfied if they are to be useful components of a cancer vaccine. Even if tumor specific antigens do exist in each individual tumor, these antigens would not prove very useful targets for a vaccine unless they occurred with a relatively high prevalence within the human population. If every tumor-specific antigen was unique to each tumor, then a separate vaccine would need to be discovered and developed for each patient. Not only is this impractical, but this is also not possible if the goal is to make a prophylactic vaccine. If antigens could be identified that were commonly found among cancer patients, then vaccines could be economically and effectively developed to treat them. Supporting this possibility, common mutations in oncogenes p53 and Ras have been found [132, 133]. In addition, the ideal cancer vaccine should be comprised of multiple antigens so as to prevent immune escape that can occur as tumors continually evolve to evade the detection of the immune system. Note that the challenges of immune escape are mitigated in earlier stages of cancer and for prophylactic vaccines since there will be fewer tumor cells at these stages. Even though many tumor antigens have been discovered, the requirements for using these to treat cancer are high, and the current list may not be sufficient for the development of an effective vaccine. There may still be many critical proteins left to discover which will aid in the understanding of cancer as well as the development of a cancer vaccine.

2.1.5 Developing platform for screening pooled tumor cDNA library lysates

In the presented research, a platform was developed for high throughput screening of a tumor cDNA library with sera from a tumor-bearing mouse. The platform consists of screening 3,000 lysate-pools, isolated from a tumor-cDNA library, and arrayed onto a nitrocellulose slide. During the development of this platform many parameters had to be tested and optimized such as the protein production process in 96 well plates, the slide surface, and the incubation time with

sera. The platform was also tested using sera to a known protein along with lysate containing that protein. As an additional test, the efficacy of the method was verified by determining that sera to a known protein bound to that protein as it naturally occurred within the tumor cDNA library. In the presented research, a microarray of pooled tumor cDNA library lysates was used to screen tumor antigens. This high-throughput approach would not have been practical at the time when some tumor antigen screening methods such as SEREX were first developed since this new approach makes use of automated equipment for handling large numbers of PCR plates, small nitrocellulose slides, high resolution scanners, and accurate printing machinery.

2.2 Materials and Methods

2.2.1 Procedures with BALB/c mice

Fifteen four-week old female BALB/c mice were purchased from Jackson laboratories and divided into three groups: a non-treated group, a tumor challenged group, and a Structural Maintenance Chromosome 1A frameshift (SMC1Afs) protein immunized group. Sera was collected weekly from each group starting at five weeks of age. Seven thousand 4T1 cells in 100 µL PBS were injected subcutaneously (s.c.) into the flank of the mice when they were six weeks old. The tumor volume was measured daily after the tumor was palpable, and mice were euthanized when the tumor volume exceeded 2,000 mm^3 as calculated by the formula $L^2*W/2$. Splenocytes and tumor tissue were collected after euthanasia in accordance with the protocols outlined in Current Protocols in Immunology [134, 135] (Figure 54). Splenocytes were used to construct a phage library which is work not presented in this chapter. Tumor tissues were used to construct a tumor cDNA library. Murine experiments were conducted under a protocol approved by the Arizona State University Institutional Animal Care and Use Committee.

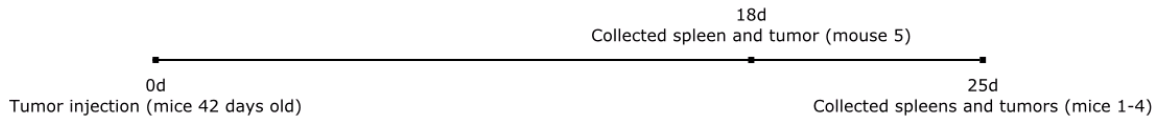


Figure 54 Timeline for tumor group

Timeline for the five mice injected with tumor cells. "0d" refers to 0 days. The spleen and tumor was collected from mouse 5 on day 18 since the tumor of this mouse exceeded the allowed volume. The tumors of the remaining mice were collected on day 25 post injection.

Mice in the SMC1Afs designated group were immunized with a SMC1Afs DNA expression construct followed by the corresponding protein, in a prime-boost regimen. The subsequently harvested serum was used as a positive control for reactivity against the SMC1Afs protein in the next set of experiments. Five female six week old mice were genetically immunized with 1 µg SMC1-pCMVi plasmid and 1 µg CpG2395 adjuvant using a gene gun for delivery at day zero, day one, and day 15. Two protein boosts were performed: i) with 5 µg 17 amino acid SMC1Afs fused N terminally to GST at day 34, and ii) with Alum and 5 µg of a SMC1Afs 17 amino acid peptide at day 57 (Figure 55). All mice were euthanized at day 73; sera and splenocytes were collected.

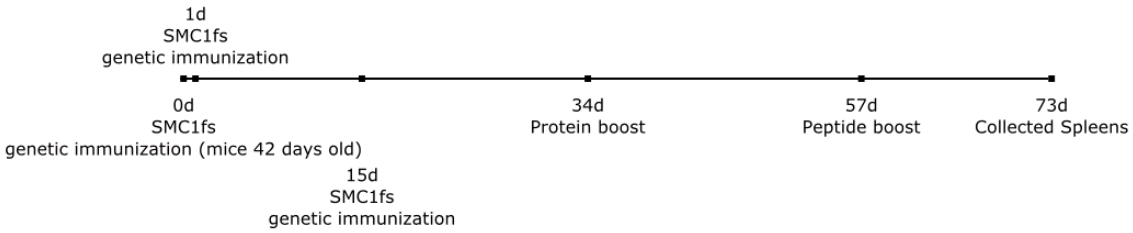


Figure 55 Timeline for SMC1fs group

Five female BALB/c mice were immunized with SMC1-pCMVi plasmid on day 0 (mice were 42 days old), day one, and day 15. Two protein boosts were performed on day 34 and 57. Mice were euthanized and sera and splenocytes were collected on day 73.

2.2.2 Construction of tumor cDNA library

A tumor cDNA library was constructed from the tumor tissue from the tumor bearing BALB/c mice. RNA was extracted from 100–1,000 mg of tumor tissue from mice using TRIzol (Cat No 15596-018) from Invitrogen according to the manufacturer's instructions.

The In-Fusion SMARTer cDNA Library Construction Kit (Cat No of 634929) from Clontech was used to construct the cDNA library from the obtained RNA. The first strand cDNA synthesis was performed according to the manual instructions. Briefly the initial reaction

consisted of 4 µg of RNA mixed with the poly T 3' SMART CDS Primer IIA oligo and ddH2O, incubated at 72°C 3 min, and then incubated at 42°C 2 min. The following components were added to the mixture and incubated 42°C 90 min: 5X First-Strand buffer, DTT, dNTP, SMARTer IIA Oligonucleotide, RNase Inhibitor, and SMARTscribe Reverse Transcriptase. For the second strand synthesis several different numbers of cycles were used in different reactions. The reaction consisted of first-strand cDNA, ddH2O, 10X Advantage 2 PCR buffer, 50X dNTP mix, 5 PCR Primer IIA, and 50X Advantage 2 Polymerase Mix. The PCR reaction was performed with the following conditions: 95°C 1 min, (95°C 15 sec, 65°C 30 sec, 68°C 6 min) Xcycles. The sample which yielded a moderately bright smear spanning from about 0.1 to 4 kb with a few prominent bands for abundant transcripts was selected for further reactions, and this sample corresponded with a cycle number of 21. The ds cDNA was purified to remove many small side reaction products. The purification was performed with CHROMA SPIN DEPC-1,000 Columns from Clontech which exclude molecules larger than the pore size from the resin so that they pass through the column quickly. An ethanol precipitation was performed to purify the DNA further from primers and small products.

An In-Fusion reaction was performed to clone the cDNA products into the pSMART2IF vector. An In-Fusion reaction fuses DNA fragments that contain the same 15 base pairs at their ends using the In-Fusion enzyme. In this library the sequence of overlapping 15 bp was identical at both sides of the insertion site into the plasmid. Therefore, sequences from the library could insert into the plasmid in the forward or reverse orientations. Additionally, translation of the sequence to protein in the bacterial clone starts from the Lac promoter in the plasmid, and therefore the frame of the tumor library sequence expressed may be different than the wild type frame. The In-Fusion reaction was performed by combining 5X In-Fusion reaction buffer, pSMART2IF linearized vector, cDNA, In-Fusion enzyme, and ddH2O and then incubating at 15 min 37°C followed by 15 min 50°C. This In-Fusion product was ethanol precipitated and electroporated into DH010B T1 electrocompetent cells from Invitrogen (Cat No 12033-015). These transformed cells were then induced to produce protein corresponding to the cDNA

sequence inserted into their plasmid. Note that the final protein lysate that was used for screening was produced in several different batches of transformation and expression.

2.2.3 Automated colony counting

Transformed colonies on numerous plates were counted using ImageJ software from the NIH [136]. Images of the plates were obtained using the Universal Hood II from BIO-RAD Laboratories. The colony counting was performed by adjusting the brightness and contrast, setting the image to an 8 bit image type, applying a threshold so that only the colonies remained on the image, and using the analyze particles function to obtain a cell count (Figure 56).

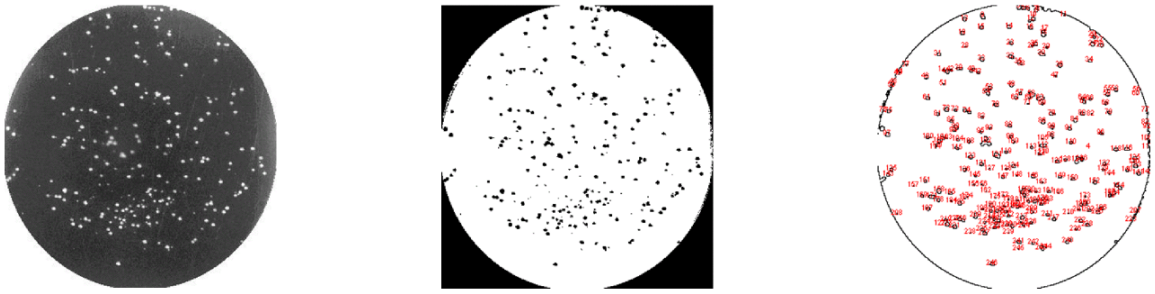


Figure 56 Automated colony counting with ImageJ

Colony counting was performed by converting the image to an 8 bit image type (left image), applying an intensity threshold (middle image), and counting the number of particles (right image) with ImageJ software.

2.2.4 Protein production and lysate printing

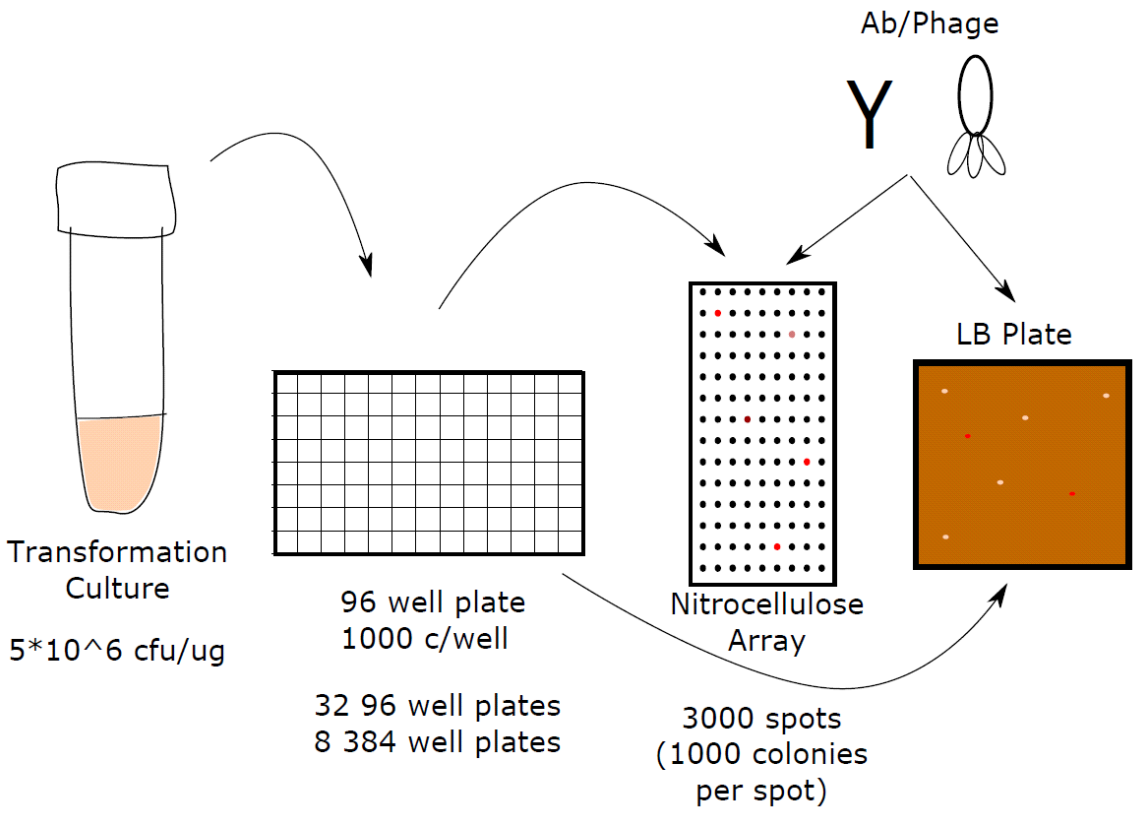


Figure 57 Logistics of printing

*A typical transformation efficiency of bacteria with plasmid was 5*10^6 cfu/ug. This culture was diluted and separated into wells within 96 well plates at a concentration of 1,000 colonies per well. Lysate from the bacteria could be transferred to 384 well plates to be spotted onto nitrocellulose slides for a total of 3,000 spots corresponding to approximately 1,000 unique bacterial clones per spot. Alternatively, bacteria could be spread onto LB plates and grown. These bacteria could be lysed and the lysate could be transferred to a nitrocellulose membrane. Proteins from lysate could be probed with sera containing antibodies or probed with phage antibody libraries.*

The proteins of the tumor library sequences were produced in pooled cultures in 96 well plates. For the library with about 3,000 lysate features, 32 96-well plates containing bacteria cultures were used (Figure 57). The transfer of liquid cultures and reagents for these plates was performed with the Biomek FX Laboratory Automation Workstation from Beckman Coulter (Figure

58). Briefly, the transformed cultures were incubated overnight at 37°C at 250 rpm. These plates were incubated in stacks in a HiGro shaker from DIGILAB to make handling the numerous plates more convenient (Figure 59). After the overnight incubation, the culture was diluted 1 to 100 in fresh LB with Carbenicillin at a final concentration of 50 ug/mL and grown to an OD600 of 0.5. Then, IPTG was added to a final concentration of 1 mM to each well in the 96 well plates to induce the production of protein. The cultures were incubated at 37°C for 4 h. Cultures were harvested by centrifuging at about 4,000 rcf for 10 min, discarding the supernatant, and resuspending in 100 µL of lysis buffer. The lysis buffer consisted of PBS, 1% Triton X-100 detergent, 1 mM PMSF protease inhibitor, and Complete Protease Inhibitor Cocktail from Roche (1 tablet per 10 mL solution). Lysozyme was added to the culture at a final concentration of 0.05 ug/µl. Plates were incubated for 15 min at room temperature with moderate shaking. The plates were then subjected to three freeze/thaw cycles to break up the cell membrane. DNase and MgCl2 were added to the solutions for a final concentration of 0.02 ug/mL and 1.1 mM MgCl2 respectively. Plates were incubated for 1 h at 4°C with moderate shaking. Soluble protein was isolated by centrifuging at 4,000 rcf for 20 min, and the supernatant was transferred to a new plate. EDTA was added at a final concentration of 1 mM to prevent contaminating growth, and glycerol was added at a final concentration of 2% to help maintain the integrity of the protein throughout future freeze/thaw cycles. A portion of the solution in each well was transferred from a 96 well plate to a 384 well plate to allow for printing.

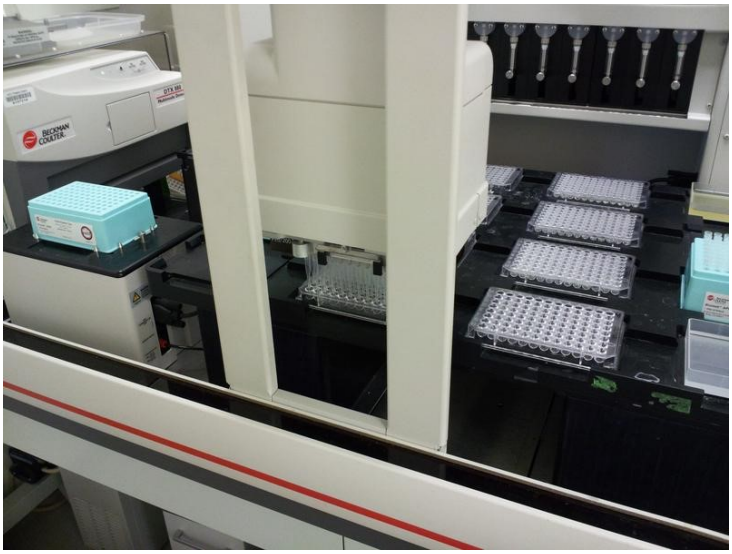


Figure 58 Biomek FX Laboratory Automation Workstation to automate liquid handling

The Biomek FX is a workstation capable of automating the transfer of liquid cultures and reagents. In this image the "arm" has obtained pipette tips from the light blue case on the left, collected liquid from the container on the far bottom right, and then dispensed this liquid into the 96 well plate in the middle.

Figure 59 HiGro shaker for shaking multiple plates

Four towers of the HiGro shaker are displayed. Each tower can contain a stack of 96 well plates for shaking and incubating at the set level.

Positive control lysate pools were included on the first library array containing 3E6 clones. These positive control lysates were from clones containing the pGEX-SMC1fs-27mer plasmid, and these clones produced 17 amino acids of frameshift SMC1Afs sequence and 10 amino acids of wild-type SMC1A sequence fused to GST. Antibodies from mice immunized with this protein were later tested against this protein. Negative control lysates from transformants containing the pUC19 plasmid were also included in the library.

The soluble protein lysate was printed onto slides with the NanoPrint 60 microarray printer from Arrayit Technologies. The final protocol made use of ONCYTE SuperNova Nitrocellulose Film-Slides slides from Grace Biolabs (Cat No GBL705177) (Figure 60) with 500 µm spacing between spots. Any printed slides were used for an experiment with sera one day after the print run to allow for enough time for the lysates to completely dry, but not allow for enough time for the proteins to degrade significantly. The lysate plates were stored at −80°C when not in use.

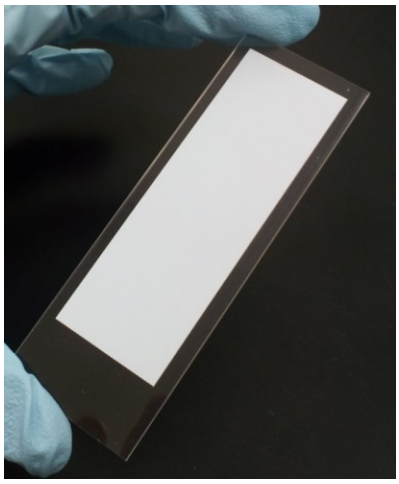


Figure 60 ONCYTE SuperNova Nitrocellulose Film-Slide

Proteins within lysate were spotted onto ONCYTE supernova nitrocellulose slides with the NanoPrint 60 microarray printer. Each slide can contain 3,000 spots with 500 µM spacing between spots.

2.2.5 Test prints with SMC1Afs dilution series

Many lysate test print runs were performed to develop a method which would allow for non-smeared spot morphology on a slide surface and also allow for the antibody detection of dilute samples of the cognate protein. This section will very briefly present the various slide surfaces and conditions that were tested during the evolution of an effective protocol. The ultimate goal was to demonstrate that a protein can be detected at low concentrations in a lysate pool. Demonstrating this capability was important since the library would contain pooled lysate from many different cDNA-library clones.

2.2.5.1 Initial Aminosilane and CodeLink Slide Test

For other projects, researchers in our lab print non-natural sequence peptides on aminosilane-coated glass slides. The slide is coated with sulfo-SMCC linker. This molecule binds to the free amines on the slide surface at one of its ends and makes available an NHS ester to bind a cysteine on a protein or non-natural sequence peptide on the other end through its maleimide group (Thermo Fisher Scientific Cat No 22122). Since these slides were abundant in our lab and compatible with all of our instruments, it was natural for us to attempt to print the cell lysate onto these slides.

In addition to printing lysate onto the aminosilane slides, I also wanted to test another slide surface and selected CodeLink Amine-binding slides (Surmodics Cat No DN01-0025). These slides have a polymer coating along with a reactive group that binds to amines. Ideally the polymer coating would help trap cell lysate proteins, and prevent smearing of lysate spots on the slide.

A lysate dilution series was printed onto both of these slide types. Lysate from bacteria induced to produce the SMC1fs peptide fused to GST (GST_SMC1fs) from the pGEX-SMC1fs-27mer plasmid was diluted into negative control lysate from non-induced bacteria containing the pUC19 plasmid. The following lysate spot dilutions were printed onto the array in triplicate in the same order that one reads: 1X, 0.5X, 0.1X, 0.001X, 0X, Blank, GST_SMC1fs purified, and GST

purified where 1X indicates undiluted GST_SMC1fs lysate, 0.1X indicates a 1/10 dilution of GST_SMC1fs lysate, and 0X indicates pure negative control lysate with no SMC1fs.

Sera were applied to both of these slide types in a similar manner. First the slides were non-specifically blocked. The aminosilane slides were blocked with BSA for 1 hr at 37 °C. The CodeLink slides were blocked for 30 min with CodeLink ethanolamine blocking buffer which was composed of 0.1 M Tris and 50 mM ethanolamine at pH 9, and then 30 min with BSA containing buffer. The slides were then washed and 5 nM purified rabbit α-hSMC1fs 17mer antibody was applied to 4 slides while a 1:500 dilution of rabbit α-hSMC1fs 27mer sera was applied to 6 slides for 1 hr 37 °C. The slides were washed and 5 nM biotinylated goat α-Rabbit IgG (Bethyl) was applied as the secondary reagent for 1 hr 37 °C. The slides were then washed and 5 nM streptavidin AF647 was applied as the tertiary reagent for 1 hr 37 °C. This method resulted in smeared lysate spots as illustrated in Figure 64a in the results section.

2.2.5.2 Denatured lysate

Next I tried to denature the proteins in the lysate to determine whether this would cause less smearing on the slide. SDS was added to the lysate at a final concentration of 1%, and this solution was incubated at 50 °C for 1 hr. The procedure for applying the sera was very similar to the procedure described in the previous section. The denatured lysate still smeared across the slide as presented in Figure 64b in the results section.

2.2.5.3 HEPES and glycerol buffer

Since glycerol is a viscous substance, I hypothesized that adding glycerol to the lysate would prevent smearing across the slide. I also tested HEPES (4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid) buffer instead of glycerol as well. Two aminosilane slides and two CodeLink slides were tested. Each slide contained a block of denatured HEPES lysate, non-denatured HEPES, denatured glycerol, and non-denatured glycerol. The final concentration of HEPEs in the lysate was 7 mM with 1 mM EDTA, and the final concentration of glycerol was 5% glycerol. The primary reagent was 1:500 rabbit α-SMC1fs-17mer; the secondary reagent was 5

nM biotinylated goat α-rabbit IgG; and the tertiary reagent was 5 nM streptavidin AF647. Some of the most extreme smearing observed was found with these buffers as seen in Figure 64c in the results section.

2.2.5.4 PEI polymer slides

Another researcher by the name of Valeriy Domenyuk in our lab was using custom PEI (polyethylenimine) polymer slides with whole cells [137], and I decided to test these slides with cell lysate. These slides are glass slides coated with epoxysilane connected to a hyper-branched PEI polymer. Other than the change of the slide surface, the rest of the procedure remained essentially the same as in the previous section. This slide surface still resulted in smearing as demonstrated in Figure 64d in the results section.

2.2.5.5 Nitrocellulose slides

Nitrocellulose slides were the next slide type tested. Nitrocellulose slides have a 3D porous structure that proteins can become trapped in, and this is the same type of material that other protein assays such as dot blots and Western blots are performed with. Lysate was printed and sera was applied to this material the same way as described in the previous sections. With this slide type very good morphology with no smearing on the slide was obtained as illustrated in Figure 64e in the results section.

2.2.5.6 Fresh SMC1fs lysate on nitrocellulose slides

The nitrocellulose slides resulted in great morphology, but only the positive control protein was detected above background signal. I hypothesized that the diluted SMC1fs may have been too old at this point, and I might obtain better results with the nitrocellulose slides and fresh SMC1fs lysate. All of the other conditions remained the same. This did result in better detection of the diluted SMC1fs lysate by the α-hSMC1fs-17mer antibody, but the fresh lysate did smear more and the spots were too close as visualized in Figure 64f in the results section.

2.2.5.7 Nitrocellulose with 800 µM spacing

The fresh lysate had larger spots, and adjacent spots ran into each other. Therefore, another print run using farther spacing between each spot at 800 µM was performed. This did result in well-spaced lysate features (Figure 64g).

2.2.5.8 Concentrated primary with nitrocellulose

In an attempt to detect lower dilutions of SMC1fs lysate, the concentration of primary serum was increased. Instead of using a 1:500 dilution of serum, a 1:20 dilution was used. The incubation time for the primary sera was also increased from 1 hr to 2 hr. The features were printed with 800 µM spacing as before. The primary sera used in this experiment was BALB Nat α-mSMC1-27mer, the secondary was a 1:100 dilution of goat α-mouse IgG (Bethyl), and the tertiary was 5 nM AF647.

2.2.5.9 Nitrocellulose with overnight primary

The concentrated sera resulted in better detection of the SMC1fs dilutions, but there was too much high uneven signal across the slide. This condition would not be good for a real library print run with many features across the entire slide. Therefore, I tried using the standard dilution of primary serum (1:500) with an overnight incubation (16 hr) instead of 1 hr or 2 hr at a lower temperature of 23 °C instead of 37 °C (Figure 64i).

2.2.5.10 Nitrocellulose with Super G Blocking Buffer and 2hr incubation

Next I tried using an alternative to the BSA blocking buffer: Super G Blocking Buffer. These slides were also incubated for 2 hr at 37 °C with the primary sera.

2.2.5.11 Nitrocellulose blocking buffer and incubation time test

One last test print was performed to directly compare the different blocking buffers and incubation times. Both the Super G Blocking Buffer and the BSA based blocking buffer were used. An incubation time of 1 hr at 37 °C and an incubation time at 16 hr 23 °C were also tested. The primary serum dilution was 1:500.

2.2.6 PCR Screen

Pools were screened for the presence of particular frameshift transcripts using a touchdown PCR protocol. All of the transcripts screened were chimeric transcripts in mice that have been previously identified by our lab [138]. The reaction was setup as follows: 5 µl 10X DyNAzyme buffer, 1 µl 25 mM dNTP each, 1 µl 10 µM forward primer, 1 µl 10 µM reverse primer, 1 µl template, 40.5 µl PCR grade H20, and 0.5 µl DyNAzyme II DNA polymerase (Thermo Scientific Cat no F-503S). Ten cycles of touchdown PCR followed by 40 normal cycles were performed as follows: (95 °C 30s, (61-cy) °C 30s, 72 °C 30s)X10cy, (95 °C 30s, 50 °C 30s, 72 °C 30s)X40, 72 °C 5m, 4 °C hold. Therefore, with this protocol the first cycle would use an annealing temperature of 60 °C, the second would use an annealing temperature of 59 °C, etc. The primers used for the different transcripts are listed in Table 12. For some of the transcripts investigated, a primer which bound directly to base pairs across the fusion site was used (designated with the letters “fus”) in addition to primers on either end of the chimeric fusion site (“F” and “R”).

Table 12 Primers used for PCR screen

Primer	Sequence	Length (bp)	Annealing Temperature (°C)
TR_E20151F	TGGTGAGCTACCCTAAGCTG	20	53.7
TR_E20151R	ACAGCCGTCTTCTGAGTTTG	20	51.7
TR_E20131F	ACCTTTCAACAGGCTCACAG	20	51.7
TR_E20131R	TTCATGCAATACCACCAATG	20	47.6
TR_E20026F	GTGTTCTTTCATGTCCCTGC	20	51.7
TR_E20026R	CCTTCCTTCTTTTCCCTTTG	20	49.6
TR_E10176F	AGCACATCTCTCTGCTGGTC	20	53.7
TR_E10176R	TGGACTGACCTTTCATCCTC	20	51.7
TR_E10142F	GACCATTCTGGAGTGTCGAG	20	53.7
TR_E10142R	CTCTTCTGCAGAGAACCAGC	20	53.7
TR_E10028F	CTGGACCATGAGGTGAAGAC	20	53.7
TR_E10028R	TCTCCAGACTCTCCAGGTTG	20	53.7
Thap2F	CATCAGCTTCCACAGGTTTC	20	51.7
Thap2R	CGATGGTTAAGATGAATCCG	20	49.6
Thap_fus	CAGGCCGAAAGTTACCAGAGAC	22	56.6
SMC1-mou-R	GAGCTGTCCTCTCCTTG	17	49.3
SMC1-mou-F	CTGTCATGGGTTTCCTG	17	46.9
Smc_fus	CACTGCACCCAGAGCCATTG	20	55.8
Slain2F	CCTGCCCTTAATAGGTTTTCTCC	23	55.2
Slain2R	CTTTCCAACTGTCATCTTTCATGC	24	53.9
Slain2_fus	TGGCAATGATTGGCTGAAGCC	21	54.2
Rbm14F	CAATGTCGATGGGGCGGATA	20	53.7
Rbm14R	CCACACCGATAGCAACCACT	20	53.7
Rbm14_fus	GACAAGAGCCTACCACGTCAC	21	56.2

List of primers used to screen for the presence of chimeric transcripts in the tumor cDNA library. These transcripts have been previously identified in mice in our lab.

2.2.7 Application of sera to tumor library lysates

The nitrocellulose slides were first locked into the Tecan HS 4800 Pro Microarray Hybridization Station from Tecan. Two slides were used for each condition: tumor sera, naïve sera, and sera from SMC1Afs immunized mice. The slides were first washed for 30 sec with Tris-Buffered Saline and Tween 20 (TBST), blocked with Super G blocking buffer, incubated at 23 °C for 1 h, and washed for 30 sec with TBST. Primary sera at a 500 fold dilution in a 200 µL volume

of incubation buffer consisting of BSA, Tween 20, and PBS was then applied to the nitrocellulose slides. This sera was from the tumor bearing, naïve, or SMC1Afs immunized mice described in the "Procedures with BALB/c mice" section. The slides were incubated at 23 °C for 16 h, washed for 30 sec with TBST, and then 5 nM of AF647 goat α-mouse IgG H+L antibody from Bethyl was added. The slides were incubated at 23 °C 1 h, washed 30 sec with TBST, washed 30 sec with ddH2O, and then dried with nitrogen for 5 min. The slides were scanned in the Tecan Power Scanner at 10 μm resolution to produce 16 bit images with intensity values for each pixel ranging from 0–65,535. The analysis of the data is described in the "Data Analysis" section.

2.2.8 Data analysis

High intensity spots in the images of the scanned slides were correlated with the identity of specific lysate pools or lysates containing single clones by aligning the spots with labeled circles in a gal file. This process was performed with GenePix software (Molecular Devices, Santa Clara, CA). The resulting intensity values for each lysate were then analyzed in Microsoft Excel (version 15.0.4551.1003) [69] to perform T-tests to determine the p-value for a lysate with tumor sera or naïve sera. The fold change between naïve and tumor was also determined. The QVALUE program for the R statistics software (version 2.15.0) [68] was also used to determine the false discovery rate (fdr) distribution of the intensities from a tumor library lysate screen [139]. All of the default settings were used to calculate the false discovery rates. Graphs and figures were created using GraphPad Prism 4 for Windows (GraphPad Software, www.graphpad.com) and Inkscape 0.48 for Windows (www.inkscape.org).

2.3 Results

2.3.1 Complexity of cDNA library

From the 15 colonies that were sequenced after the construction of the tumor cDNA library, an estimate of the complexity of the transcripts represented in the library was calculated. Several of the sequences were very short products (three sequences), and nine of the remaining 12 insert sequences were oriented in the forward orientation relative to the Lac promoter of the bacterial

expression plasmid. One of these nine sequences was a duplicate, such that eight from this sampling of 15 library components were found to be unique probes for screening the binding reactivities of test and control sera.

2.3.2 Troubleshooting protein production in 96 well plate

The protocol I followed in order to produce protein in 96 well plates was adapted from the protocols referred to in Current Protocols in Molecular Biology Unit 16.5 [140]. The final protocol I consistently used is described in the "2.2.4 Protein production and lysate printing" section. However, before arriving at this protocol, I had some trouble producing proteins in the 96 well plate format. I observed very low amounts of the induced protein and high amounts of a mysterious 14 kD protein. The 14 kD band was detected in pGEX-SMC1fs-27mer plasmid type solution (induced and non-induced) and pET32b plasmid type solution (both induced and non-induced) when probed with antibody against SMC1fs or thioredoxin (Figure 61). The 14 kD protein was also observed in a Coomassie stain during which no antibodies are used (Figure 62).

John-Charles Rodenberry, Dr. Sykes, and I later discovered that this band was caused by the addition of too much lysozyme during the protein production protocol to degrade the bacterial cell wall. During the protocol to produce protein, bacteria are grown overnight, diluted, induced to produce protein with IPTG, allowed to produce protein for several hours, centrifuged, resuspended in lysis buffer, and then lysozyme is added. The solution is then subjected to freeze/thaw cycles, DNA is degraded with DNase, the solution is centrifuged, and the soluble protein in the supernatant is stored. When developing the protocol I had to modify the procedure we used in the lab for producing protein in large flasks rather than 96 well plate wells. For the protocol with a large flask, 2 mg of lysozyme is added per 1 mL resuspension solution. For the plate, I was adding 0.4 mg per 1 mL resuspension solution. However, the original protocol with a large flask was actually adding 0.05 mg lysozyme per 1 mL of original culture before the bacteria solution was diluted or centrifuged. Therefore, I was adding too much lysozyme. Once the correct amount of lysozyme (0.05 mg lysozyme per 1 mL of original culture) was used, a strong

SMC1fs band could be detected in a western blot with only a faint 14 kD lysozyme band in the background (Figure 63).

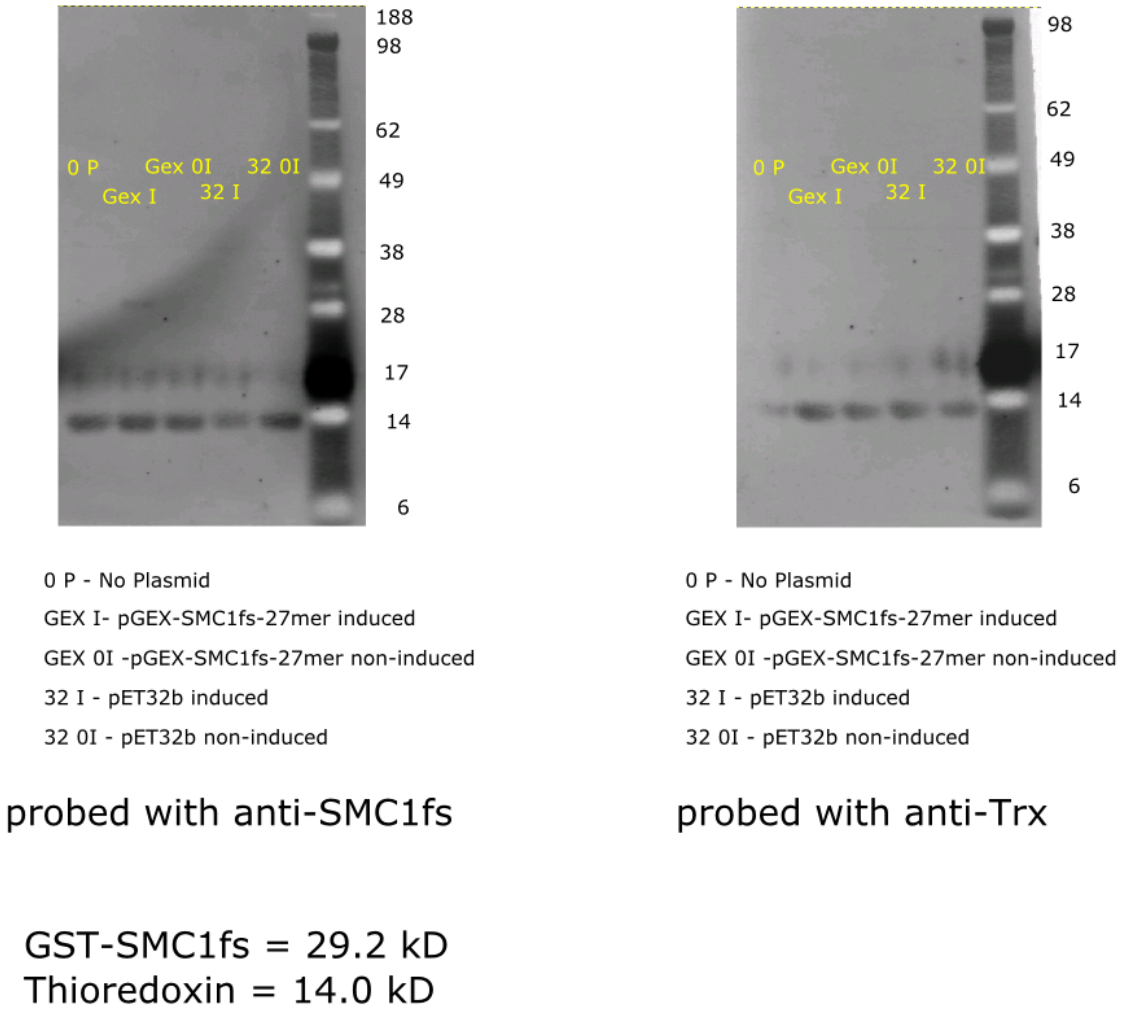


Figure 61 Western blot for production of protein in plate format with 14.3 kD Lysozyme band

E. coli containing different plasmids were lysed, and the lysate was probed with antibody against SMC1fs or thioredoxin. The "0 P" plasmid refers to the negative control E. coli with no plasmid, "GEX I" refers to the pGEX-SMC1fs-27mer plasmid with E. coli that were induced to produce protein with IPTG, "GEX 0I" refers to the pGEX-SMC1fs-27mer plasmid without induction, "32 I" refers to the pET32b plasmid with induction, and "32 0I" refers to pET32b with no induction.

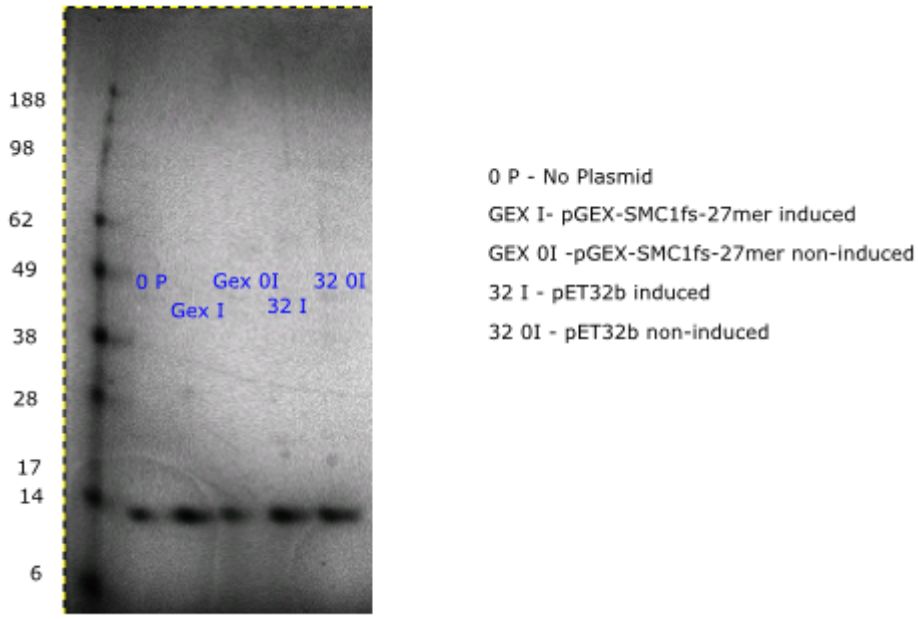


Figure 62 Coomassie stain for protein production in plate format with 14.3 kD band

E. coli containing different plasmids were lysed, and the lysate was subjected to a Coomassie stain to visualize protein bands. The "0 P" plasmid refers to the negative control E. coli with no plasmid, "GEX I" refers to the pGEX-SMC1fs-27mer plasmid with E. coli that were induced to produce protein with IPTG, "GEX 0I" refers to the pGEX-SMC1fs-27mer plasmid without induction, "32 I" refers to the pET32b plasmid with induction, and "32 0I" refers to pET32b with no induction.

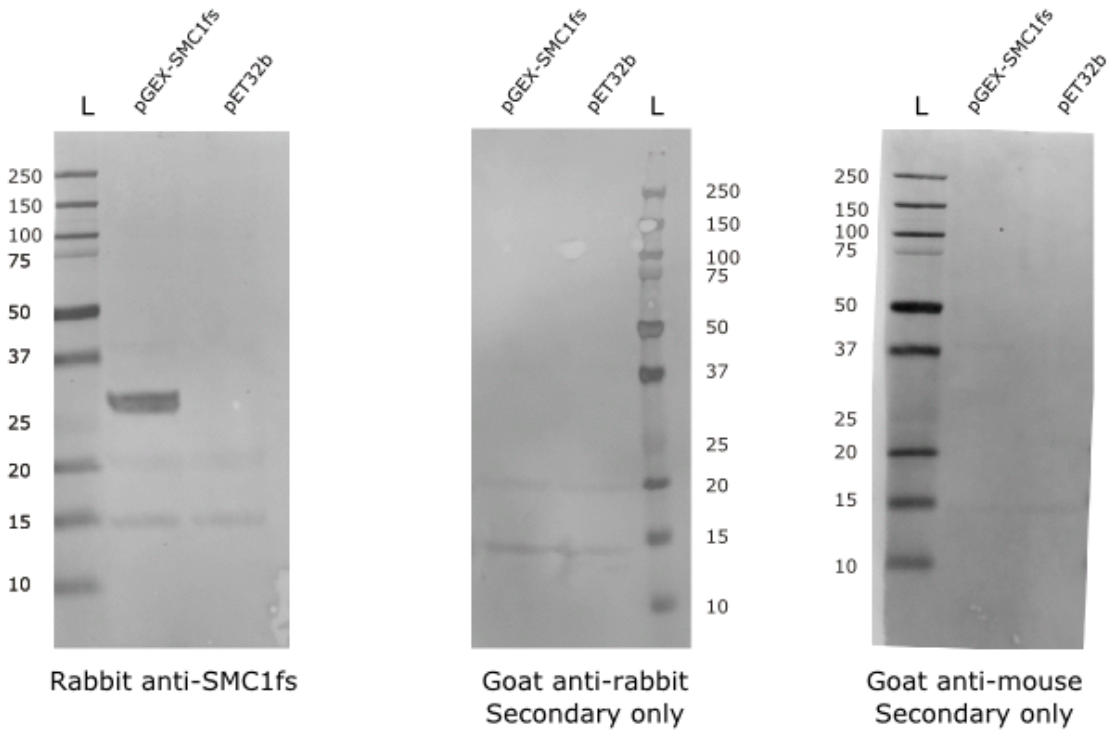


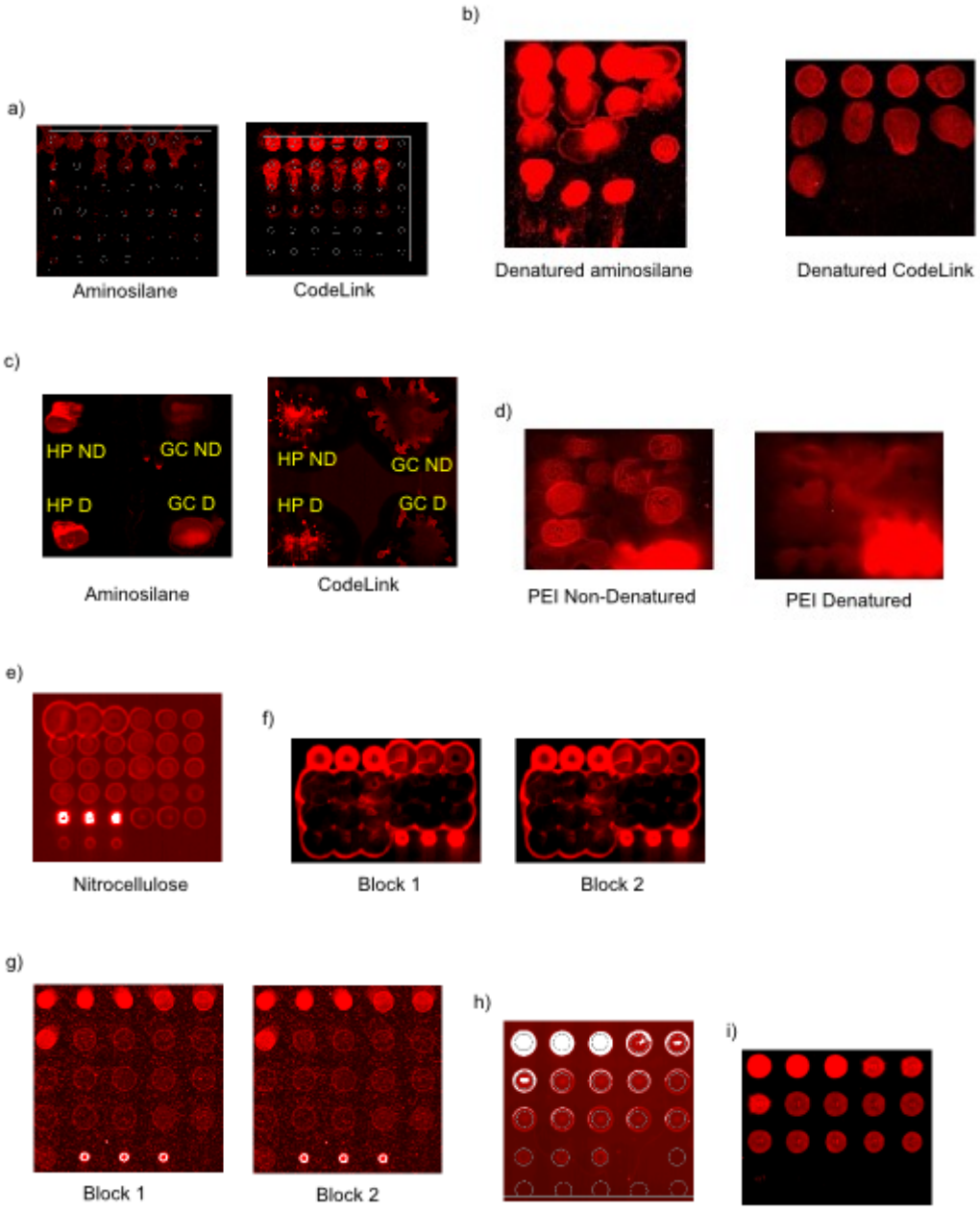
Figure 63 Western blot of protein production in plate format with appropriate lysozyme concentration

E. coli containing either the pGEX-SMC1fs or the pET32b plasmid were induced to produce protein with IPTG. Lysate from this E. coli was then used in a western blot probed with antibody against SMC1fs (left), rabbit IgG (middle), or mouse IgG (right). The molecular weight of GST-SMC1fs is 29.2 kD, and the expected size of lysozyme is 14 kD. During the production of the protein 0.05 mg of lysozyme per 1 mL of original E. coli culture was used.

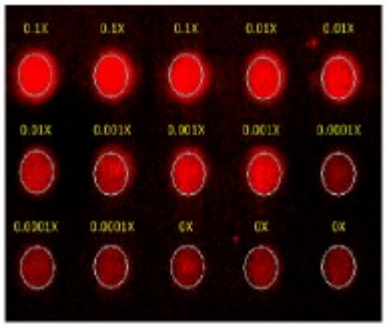
2.3.3 Test prints with SMC1Afs dilution series

Before arriving at the final protocol used for the tumor cDNA library lysate slides, many different slide surfaces, incubation conditions, and reagent concentrations were tested (Figure 64 and Figure 65). The final protocol optimized tumor lysate spot morphology without smearing one lysate pool into another (Figure 66). I determined that a test antibody can recognize its cognate antigen when diluted up to 1,000 fold into control lysate with a fold change of 1.66 and a p-value of 0.01. An experiment in which α-SMC1Afs antibody was applied to the three million component

tumor library array demonstrated that the SMC1Afs containing lysate pools were selectively recognized (Figure 67 and Figure 68).



j)



k)

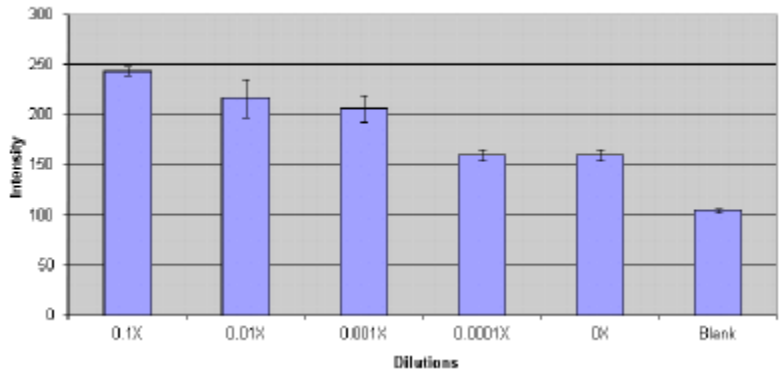
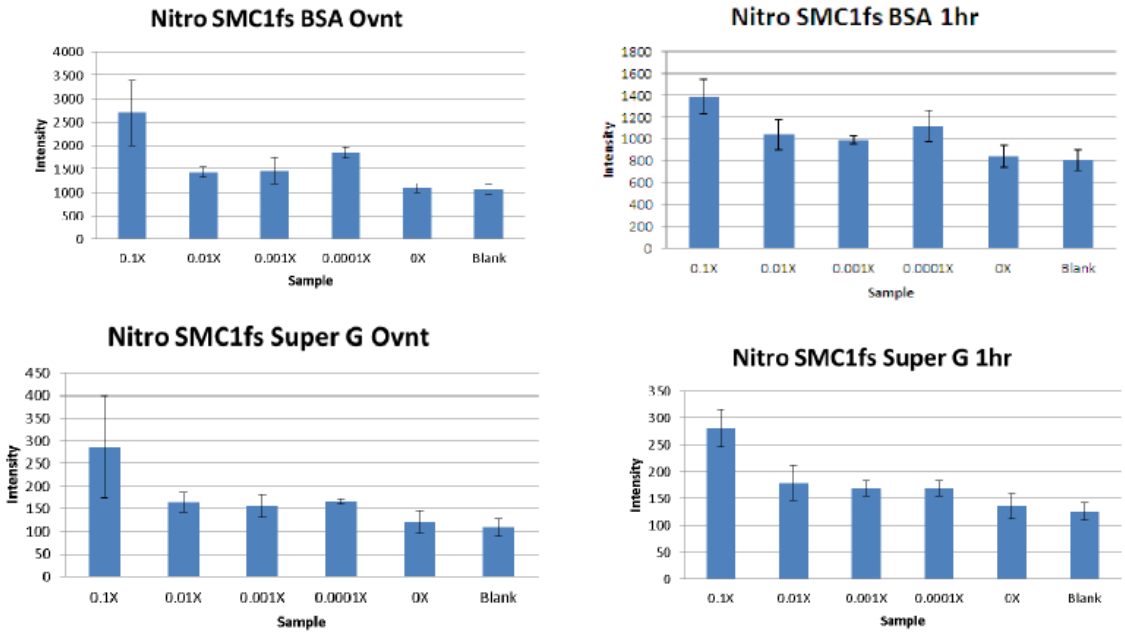


Figure 64 Test prints

a) Lysate morphology on CodeLink and Aminosilane slides. b) Denatured lysate on CodeLink and Aminosilane slides. c) Denatured and non-denatured lysate with HEPES and glycerol buffer on CodeLink and Aminosilane slides (HP = HEPES, GC = Glycerol, D = denatured, ND = non-denatured). d) Denatured and non-denatured lysate on PEI polymer slides. e) Lysate printed onto nitrocellulose slide. f) Fresh SMC1fs lysate printed onto nitrocellulose slides. g) Lysate printed onto nitrocellulose slides with farther 800 µM spacing. h) Concentrated 1:20 primary sera applied to nitrocellulose slide. i) Overnight primary incubation with nitrocellulose slides. j) Super G blocking buffer used instead of BSA blocking buffer. k) Bar graph of intensity values detected in each dilution of SMC1fs lysate with Super G blocking buffer.

a)



b)

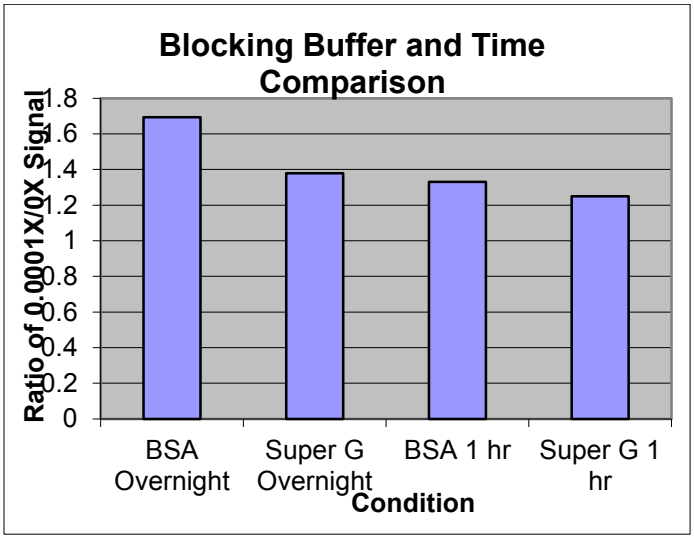


Figure 65 Test of blocking buffer and incubation condition

a) Bar graphs of detected intensity of features in an SMC1fs lysate dilution series on nitrocellulose (nitro) slides with Super G Blocking Buffer or BSA blocking buffer, and an overnight (Ovnt) 16 hr 23 °C incubation or 1 hr 37 °C incubation. b) Ratio of 0.001X/0X signal for the different conditions.

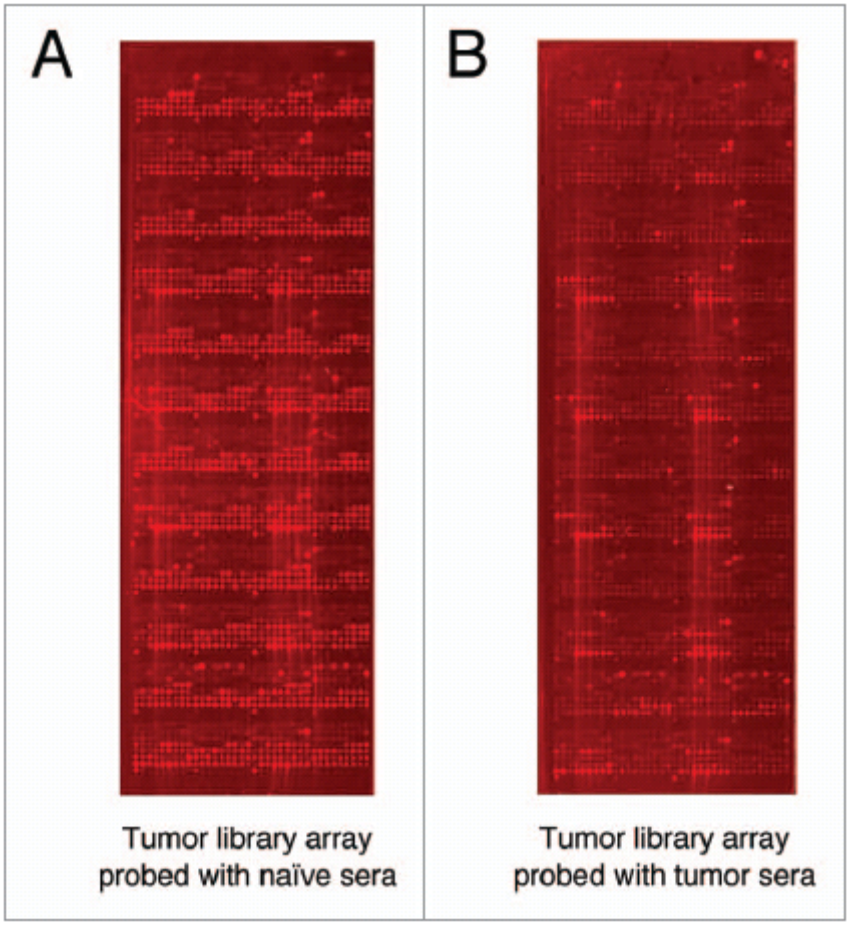


Figure 66 Array scan of slide surface.

Lysates from a 4T1 tumor cDNA expression library printed onto nitrocellulose slides. (A) Library probed with naïve mouse serum. (B) Library probed with serum from 4T1 tumor-bearing mouse. Nitrocellulose slides were spotted with 3,000 pools of bacterially expressed tumor cDNA library clones, and probed with the indicated sera. Specific lysate binding was detected with an α-mouse IgG secondary antibody conjugated to AF647 and then scanned (Tecan Power Scanner) at 647 nm. An image of one slide with naive sera and one slide with tumor sera is presented. Each one of the 3,000 pools is comprised of 1,000 original 4T1 tumor cDNA library transformants.

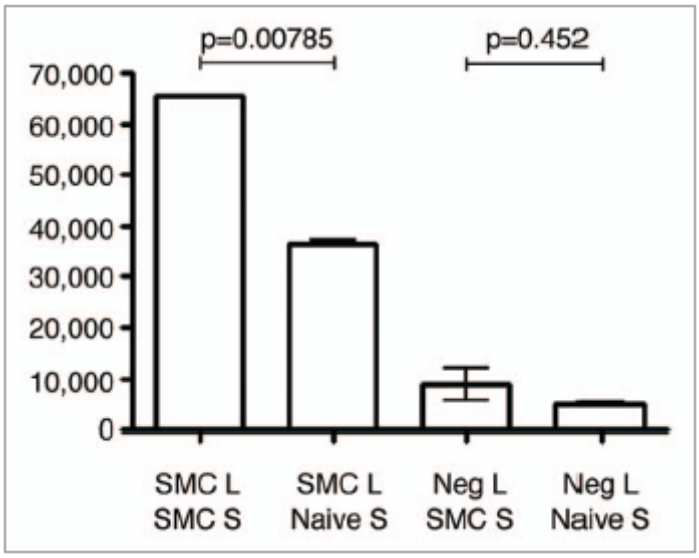


Figure 67 Bar graph of detected intensity of controls

Intensity of lysate containing either SMC1Afs peptide or negative control β-galactosidase protein. The intensities were detected after either sera against SMC1Afs or sera from a naïve mouse were applied to the 3 million component tumor lysate library. SMC, SMC1Afs; Neg, negative control β-galactosidase; S, sera; and L, lysate.

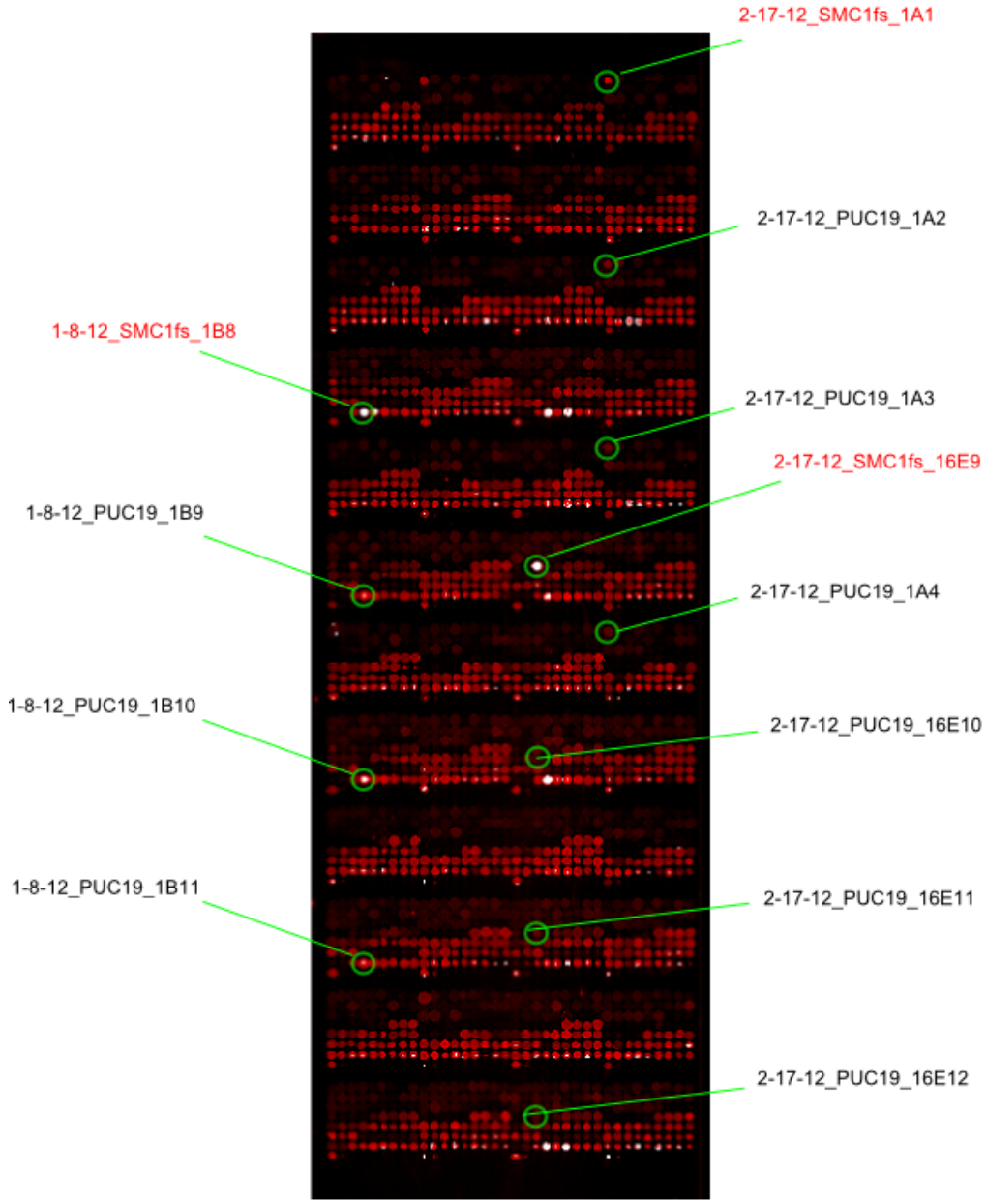
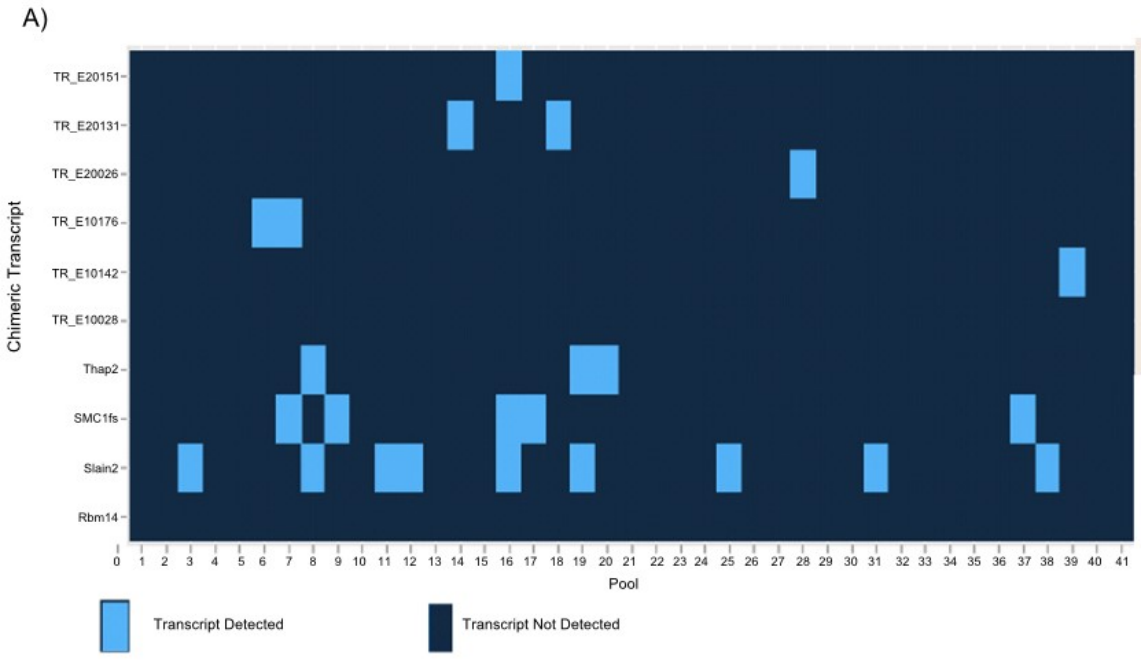


Figure 68 Tumor library array probed with SMC1fs sera

The positive and negative control lysate spots on the array are labeled with a date that the lysate was produced, a label indicating whether the lysate is a pUC19 or SMC1fs spot, and information about the PCR plate that the lysate came from (the plate number, the row number, and the column number).

2.3.4 PCR Screen

A PCR was performed to detect whether certain pools in the library contained certain frameshift transcripts and chimeric transcripts which our lab has investigated over time. Our lab has sera against the bottom four transcripts in Figure 69A. The pools which the selected frameshift or chimeric transcripts were detected in are displayed in Figure 69A. Each pool in this screen of 41 pools consists of approximately 96 subpools. Pool 16 of the 41 pools was selected for further screening for the SMC1fs and Slain2 transcripts (Figure 69B). This pool was chosen because it had both transcripts and the bands on a PCR gel looked distinct. Note that only one pool from this screen contains the SMC1fs transcript in the forward orientation relative to the *lac* promoter in the bacterial expression plasmid, and this pool was pool number 87 in Figure 69B.



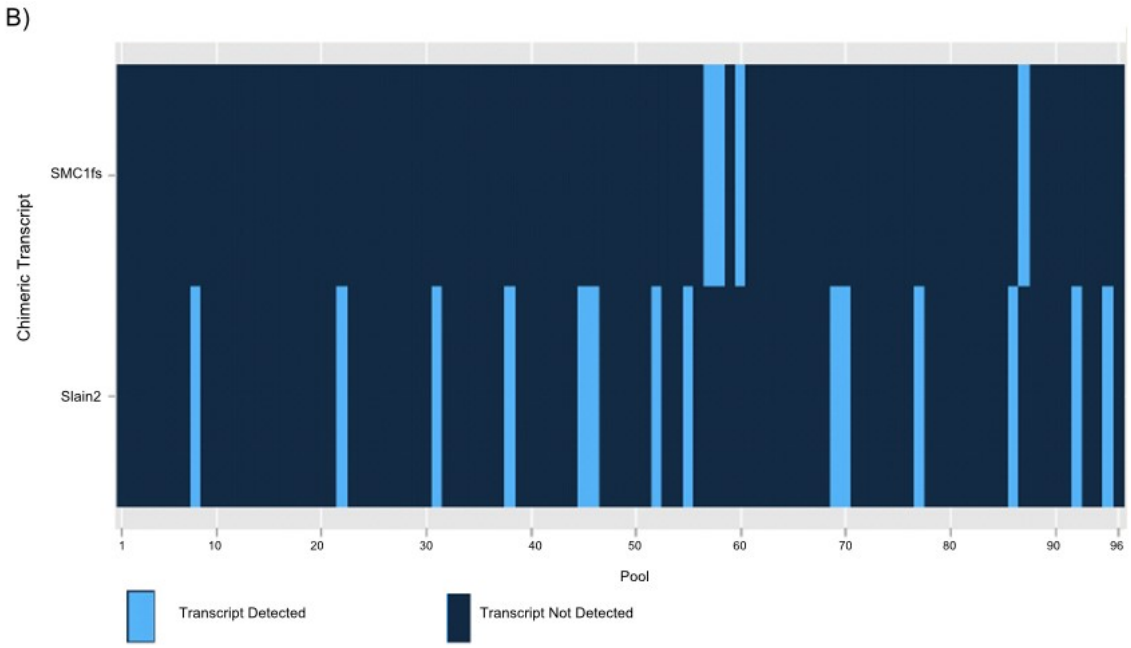


Figure 69 PCR Screens of Pools in Tumor cDNA Library

Pools in the tumor cDNA library were screened for the presence of specific transcripts using PCR. A) 41 pools each consisting of approximately 96 subpools of the 3,000 pool library representing 3 million clones were screened for the presence of 10 chimeric transcripts. B) Pool 16 from the screen in A was screened further to determine whether the 96 subpools contained the SMC1fs transcript or the Slain2 transcript.

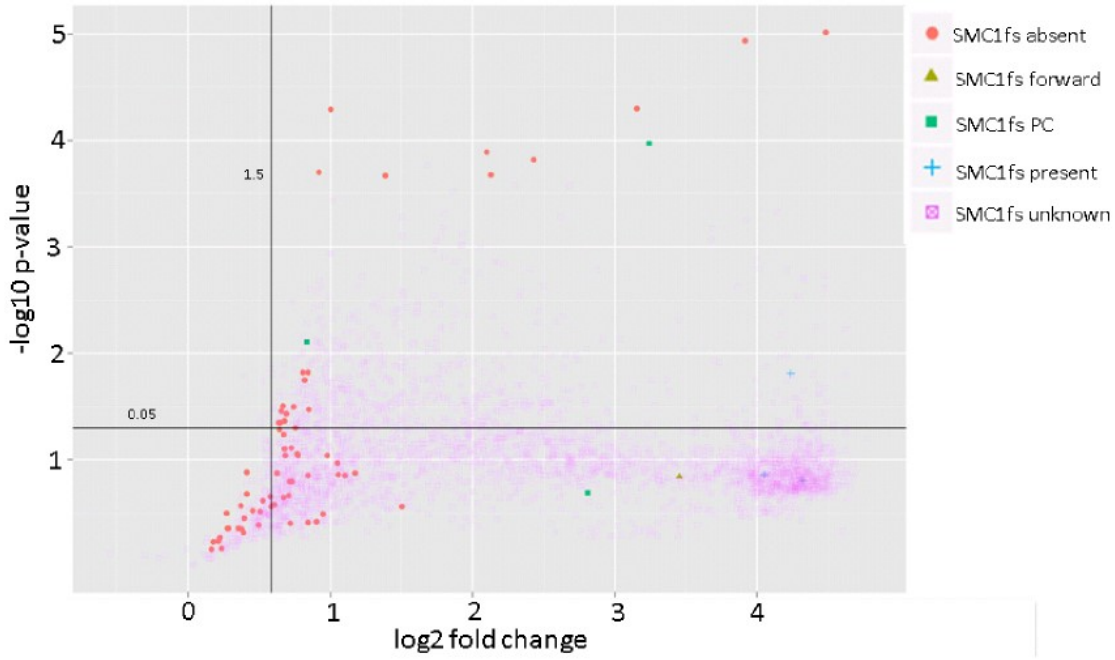
2.3.5 Application of sera to tumor library

Sera from mice immunized with SMC1fs and sera from the tumor-bearing mice from which the tumor cDNA library was constructed were applied to the 3,000 pool tumor cDNA library lysate array. From the PCR screen described in the "2.2.6 PCR Screen" section, I knew which pools in the library contained the SMC1fs transcript. Three pools on the array corresponded to the SMC1fs positive control pools containing lysate from bacteria with the pGEX-SMC1fs-27mer plasmid. One pool is known to contain the SMC1fs transcript in the forward orientation relative to the *lac* promoter in the plasmid.

After applying sera to the tumor cDNA library lysate array, an intensity value for each pool was obtained. These intensity values were compared to the intensity values for each pool

obtained after applying naïve sera, and the p-value and fold change of each pool probed with naïve vs SMC1fs sera, or naïve vs tumor sera was obtained. A figure showing where each of these pools lie on a p-value and fold change scatterplot is presented with SMC1fs sera (Figure 70A) and tumor sera (Figure 70B).

A)



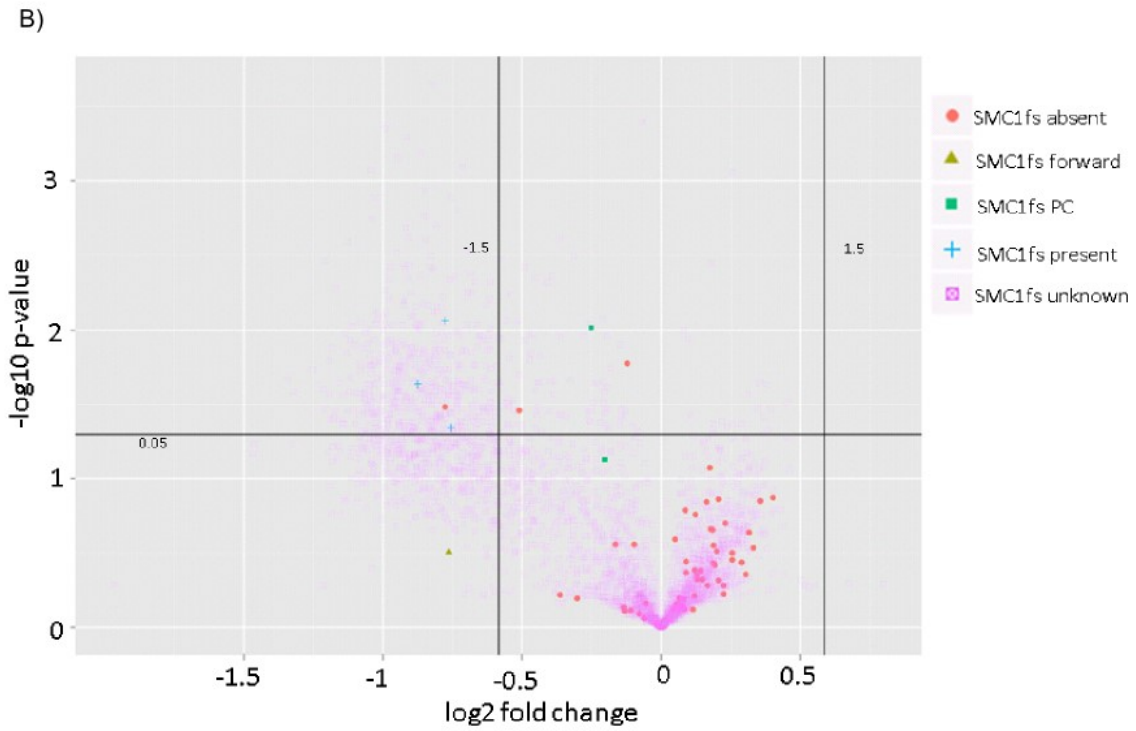


Figure 70 Reactivity of SMC1fs lysate with sera on 3K array

Sera were applied to the 3K lysate array and compared with naïve sera. A scatterplot of the p-value obtained in comparison with naïve sera and the fold change in comparison with naïve sera is displayed. A) Results obtained from SMC1fs applied to the array. B) Results obtained from tumor sera applied to the array.

2.4 Discussion

2.4.1 Library transcript representation

The complexity of the sequences contained in the constructed cDNA library indicated that 53.3% of the sequences were not truncated products and were correctly oriented for expression. While informative, this number is likely an overestimate caused by the small sampling size. Namely, more than half of the clones in the sampled set were in the correct orientation (11/15 = 73% in correct orientation). Unless there is some fortuitous bias that I am unaware, 50% should be forward oriented and 50% should be reverse oriented). If more clones were sequenced, the calculated expressed-clone complexity would presumably approach the expected. This can be

approximated as (1/2)*11/15 = 37% since 11/15 of the transcripts were unique non-truncated sequences, and approximately 50% of these sequences would be expected to insert into the plasmid in the correct orientation. Of these 37% only 33% would be anticipated to be in the same frame as the original transcript and therefore approximately 12% of the clones in the library would be expected to express products faithful to the original tumor RNA products.

The nitrocellulose slides are printed with E. coli cell lysate from this tumor cDNA library. There are 3,000 features total. Each feature represents proteins translated from approximately 1,000 original transformants. Therefore, the total number of components present on a single microarray is ~3 million. This number includes transcripts that were inserted backward or out-of-frame. After taking these artifacts into account, I estimate a total of 3.6E5 unique transcripts are properly translated and presented on the microarray. There are between 20,000 and 50,000 transcripts in a mammalian cell, depending on how one defines a unique transcript, and not taking into account post-translational modifications or rare splice variants. Most of the mammalian transcripts (94%) occur between one and five times [141], and about 25% are present in one or fewer copies per cell [142]. Given 3,000,000 clones, each mammalian transcript would be represented in our library 3E6/50,000 = 60 times on average. However, since only 12% of the library would correspond to the original RNA transcript sequence, each transcript would be expected to be represented approximately 60*0.12 = 7.2 times on average. I expect that there are certain transcripts for which this would not be true, given that some transcripts are extremely high copy and would appear more than 7.2 times presenting an opportunity to create a noise threshold. However, given a 7.2-fold minimum representation for low-copy transcripts, I expect sufficient sensitivity to detect single-copy events. A spike-in dilution experiment indicated our sensitivity is greater than one copy per 1,000 clone pool (feature).

2.4.2 Experiment conditions

Different slide surfaces, incubation conditions, and reagent concentrations were tested during the development of the protocol. A brief description of these tests will be provided. Many slide surfaces were used before consistently using the nitrocellulose slide surfaces used for the

screens presented in the Materials and Methods and Results section. I first tried to use CodeLink and Aminosilane slides, but the printed cell lysate features would smear across the slide. I then tried to denature the lysate by incubating with a final concentration of 1% SDS detergent at 50 °C for 1 hr, but this strategy was not effective at preventing smearing. A 1/10 dilution of SMC1fs into pUC19 lysate could not be detected above background. Next I tried using HEPES and glycerol buffer, but this actually resulted in the most smearing of any of the print runs. Note that there was high humidity around this time, and this may have adversely affected the print run as well. I then tried printing onto a different slide type: PEI polymer slides. No improvement was obtained. The final slide type tested was composed of nitrocellulose, and this material yielded very good cell lysate spot morphology. However, the SMC1fs dilution series did not show any signal. Perhaps the SMC1fs lysate which had been stored at -80 °C and frozen and thawed several times was too old at this point in time.

Once I identified that the nitrocellulose material worked very well to prevent smearing, I tried to increase our ability to detect the diluted protein. Fresh SMC1fs lysate was prepared and another experiment was performed. Only the 1X and 0.5X lysate spots were detected above background, and this fresh lysate also exhibited more smearing than the older lysate. Another print run with 800 µM spacing instead of 500 µM was performed to prevent adjacent spots from running into one another. In order to detect the lower dilutions of SMC1fs protein, the concentration of the primary sera was increased from 1:500 to 1:20. Lower dilutions were detected, but there was too much bright uneven intensity across the slide which would have caused problems for screening an entire library with many features. I then tried using the standard primary sera dilution of 1:500, but I increased the incubation time from 1 hr to 16 hr and lowered the temperature to 23 °C. This did allow for the detection of the 1/100 lysate spot, but the 1/1,000 dilution was still not detectable.

Next I tried using the Super G Blocking Buffer instead of the BSA blocking buffer. The first experiment with this buffer was performed with a primary sera incubation at 37 °C for 2 hr. This Super G Blocking Buffer decreased all detected intensities considerably. However, the ability to distinguish between different lysate dilutions and the ability to detect lower dilutions

increased. The error bars of replicate lysate spots were also very small with this blocking buffer. The final test print directly compared the Super G and BSA blocking buffer as well as the 1 hr 37 °C primary incubation condition and the 16 hr 23 °C incubation condition. The Super G blocking buffer results in less smearing, and the overnight condition results in higher signals. In this experiment I was even able to detect 1/10,000 SMC1fs dilutions above background.

The results from these test prints led us to use nitrocellulose slides, Super G blocking buffer, 1:500 sera dilution, and 16 hr 23 °C primary sera incubation conditions for future experiments. A secondary antibody titration experiment was also performed to determine the optimal concentration of secondary to use, and a 1 nM concentration was then chosen for many experiments. These conditions were then used to demonstrate that α-SMC1Afs antibody could detect the SMC1Afs protein in lysate from clones producing this protein as presented in the Results section.

Note that there are groups of features showing similar intensities. These were obtained in the same bacterial production batch, and this effect can be seen from the patterns of bright and faint spots on the nitrocellulose slide (Figure 66). This bias in intensity which is introduced by technical experiment variation rather than a biological phenomenon does not affect the data analysis because each lysate feature intensity obtained with naïve sera is compared with the same lysate feature intensity obtained with tumor sera. In other words, the lysates compared are identical and are of the same batch, and lysates from different batches were not compared against one another.

2.4.3 PCR Screen

There were 3,000 tumor cDNA library pools and some of these pools were screened for the presence of specific frameshift and chimeric transcripts. The positive rate of the transcripts (displayed in Figure 69) seems reasonable considering that approximately 7 copies of any single copy transcripts are expected to be present in the library as indicated in the “2.4.1 Library transcript representation” section. Also note that most of the transcripts searched for were

detected with the exception of 2 transcripts. These may be very low copy transcripts which did not make it into the cDNA library.

2.4.4 Application of sera to tumor cDNA library

Sera samples from naïve mice, tumor mice, and SMC1fs-immunized mice were applied to the cDNA-library. Note that the sera from the mice with the tumor were the same mice from which the tumor cDNA library was constructed (see sections "2.2.1 Procedures with BALB/c mice" and "2.2.2 Construction of tumor cDNA library" for more details). The intensity for each pool with SMC1fs sera was compared to the intensity obtained with naïve sera and the p-value and fold change were determined (Figure 70A). The same values were determined with tumor sera. These results show that pools which are known to contain the SMC1fs transcript as determined from a PCR screen (section "2.3.4 PCR Screen") had either significantly different p-values or fold changes relative to naïve sera. Almost all of the SMC1fs positive control pools fall into this same category as well. These findings indicate that the platform developed works as planned. Antibodies against a specific target (SMC1fs) detected this target when produced artificially in bacteria containing the pGEX-SMC1fs-27mer plasmid. These antibodies also detected the target as it naturally occurred in the tumor cDNA library. These results suggest that this array platform can be used to detect new immunogens against which a host has produced antibodies. These proteins could be studied further to better understand cancer and the relationship of these proteins with the immune system, and these proteins are also good candidates for a cancer vaccine.

2.5 Conclusion

This research outlines a platform for screening tumor cDNA library lysate for tumor-specific antigens. The platform involves using automated equipment to handle large numbers of PCR plates, small nitrocellulose slides, high resolution scanners, and the screening of several rounds of pools. As researchers use high throughput methods such as the method demonstrated in this paper to discover new tumor immunogens, the reactivity of antibodies for specific proteins will

reveal which proteins the immune system responds to during the course of cancer. Using this knowledge scientists will be able to select optimal components to include in a preventative cancer vaccine.

3: DISCOVERING IMMUNOGENS

3.1 Introduction

3.1.1 Screening pooled tumor cDNA library lysates

A tumor cDNA library was constructed and 3,000 lysate pools each containing approximately 1,000 original transformants were contact printed onto a nitrocellulose slide. This approach allowed three million library components to be quickly and simultaneously screened on one platform. One high binding sublibrary lysate-pool was then selected and re-arrayed onto another slide, as 3,000 individual features. Thus ~1,000 unique constituents would be represented at approximately one per spotted position since the colonies were diluted to contain an estimated one colony per well volume. High binding lysate spots from this array were then used to construct an array containing exactly one clone for a final round of screening. This multi-round pooled lysate approach allowed for the screening of many clones at once. The high-throughput tumor lysate screening for tumor antigens approach described here may be the type of method that could lead to the discovery of important tumor immunogens that aid in the understanding and treatment of cancer, and several potential immunogens from this screen are presented.

3.2 Materials and Methods

3.2.1 Protein production and lysate printing

The proteins of the tumor library sequences were produced in pooled cultures in 96 well plates. There were three libraries handled: a library containing 3,000 lysate pools representing three million clones, a library containing 3,000 lysate pools representing 3,000 clones, and a library containing 12 lysates representing 12 clones.

The soluble protein lysate was printed onto ONCYTE SuperNova Nitrocellulose Film-Slides slides from Grace Biolabs (Cat No GBL705177) with the NanoPrint 60 microarray printer from Arrayit Technologies with 500 µm spacing between the center of one spot and the center of the

next spot. For the data presented in this chapter, four print runs were performed: (1) a three million clone library containing ≈3,000 lysate pools (actually 3,072) each containing approximately 1,000 unique clones each, (2) a validation of performance print of 15 lysate pools of ≈3,000 each selected based on p-values and intensity values from the three million clone library, (3) a pool reduction screen print consisting of a 3,000 feature sublibrary (specifically 3,072) with approximately one unique clone per feature, and (4) a single clone array screen print consisting of 12 lysates deconvoluted from features in the 3,000 feature sublibrary.

3.2.2 Application of sera to tumor library lysates

The data from four experiments is presented (Figure 71). The same protocol was followed for all four of these experiments with slight variations as knowledge was gained throughout the screening process.

3.2.2.1 Tumor library screen

The first tumor library screen was performed as described previously during the development of the tumor cDNA library screening platform ("2.2.7 Application of sera to tumor library lysates"). After the p-value was determined by comparing the tumor and naïve samples, three sets of spots were selected to be used in a validation screen. Namely, the lysate pools with the top five most significant increases in intensity of tumor relative to control sera, the most significant lysate pools displaying decreases in intensity in tumor sera, and the five least significant p-value lysate pools. The five least significant p-value lysate pools were chosen as negative controls. The top five most significant lysate pools displaying decreases in intensity in tumor sera could be important proteins for which the tumor has somehow caused an immune suppression against. The lysate pools with increases in intensity with tumor sera could be important because there could be a strong immune response against these proteins in the mice with a tumor.

3.2.2.2 Validation of performance of selected lysate pools

Selected lysate pools (15 total) were printed onto a set of new nitrocellulose slides. Each lysate pool was present on the array with ten replicates. Sera was applied as before at a 500 fold dilution to two slides for tumor sera and two slides for naïve sera. This was followed by 1 nM secondary antibody. The lysate pool with the most significant p-value calculated by comparing the tumor and naïve sample was used to construct a new library containing 3,000 lysate pools with ≈1,000 unique clones per pool.

3.2.2.3 Pool reduction screen

Sera was applied at a 500 fold dilution to four nitrocellulose slide replicates per group containing lysate representing ≈3,000 unique clones. This was followed by 5 nM secondary antibody. Several lysate pools were chosen for further screening. These lysates corresponded to the three lysate pools with the most significant p-value, the greatest fold change, and the highest detected intensity with tumor sera.

3.2.2.4 Single clone array screen

Individual unique clones derived from the selected pools in the ≈3,000 clone screen were printed onto new nitrocellulose slides with three replicates per group and sera was applied at a 1,000 fold dilution. This was followed by 1 nM secondary antibody. The p-value and fold change of the lysates between tumor and naïve samples was then analyzed.

3.3 Results

3.3.1 Tumor lysate screening

3.3.1.1 Tumor library screen

Pooled tumor lysates were subjected to several rounds of screening as described in the Materials and Methods section. **Table 13** lists the array feature selection criteria used through the screening protocol, and the results of the screen are presented in **Figure 71** A–D. In the first

round of screening, the nitrocellulose slide contained 3,000 pools each consisting of approximately 1,000 unique bacteria clones. Several pools from this round of experiments were selected for reduction and further characterization. The five tumor library lysate pools displaying the most significant increases in reactivity between normal and tumor sera samples were selected (p-values ranged from 2.04E-5 to 4.27E-3; fdr q-values ranged from 3.78E-3 to 1.86E-2). The five pools with the most significant decreases in reactivity were also selected (p-values ranged from 2.40E-6 to 2.04E-5; q-values ranged from 1.35E-3 to 3.78E-3). The five pools with the least significant p-values were selected as negative controls (p-value of 1; fdr q-value of 0.5). These fifteen selected pools were printed onto a new array for further characterization.

A scatterplot of the 3,000 lysate pools is presented in Figure 71A. The base two log of the fold change is represented on the x-axis so that values with a decreasing fold change from naïve to tumor are on the left side of the y-axis and values with an increasing fold change from naïve to tumor are on the right side of the y-axis. The negative base ten log of the p-value is represented on the y-axis so that values with the most significant p-values are present at the top of the graph. Note that false discovery rates (FDR) are useful for making assessments about many p-values from an experiment. An FDR adjusted p-value, or q-value, of 3.78E−3 implies that 0.378% of all tests with a q-value less than or equal to 3.78E−3 are false positives. This provides a metric to use across the entire experiment rather than to use only for each individual lysate pool. In support of the validity of the array-based library screening method, the pool with the most significant p-value within the set of increased reactivities in one screening experiment was also in the top three of the same category in a second screening experiment (data not shown).

Table 13 Categories of features

Symbol	Description	Stage of Screen
△	Significant p-value low tumor	3E6 component library
◇	Significant p-value high tumor	3E6 component library
◆	Pool selected from 3e6 component library for further partitioning	3E6 component library
+	Greatest fold change	3,000 component sublibrary
▽	Greatest tumor intensity	3,000 component sublibrary
✳	Significant p-value	3,000 component sublibrary

Categories of specific features displayed in the scatterplots in Figure 71.

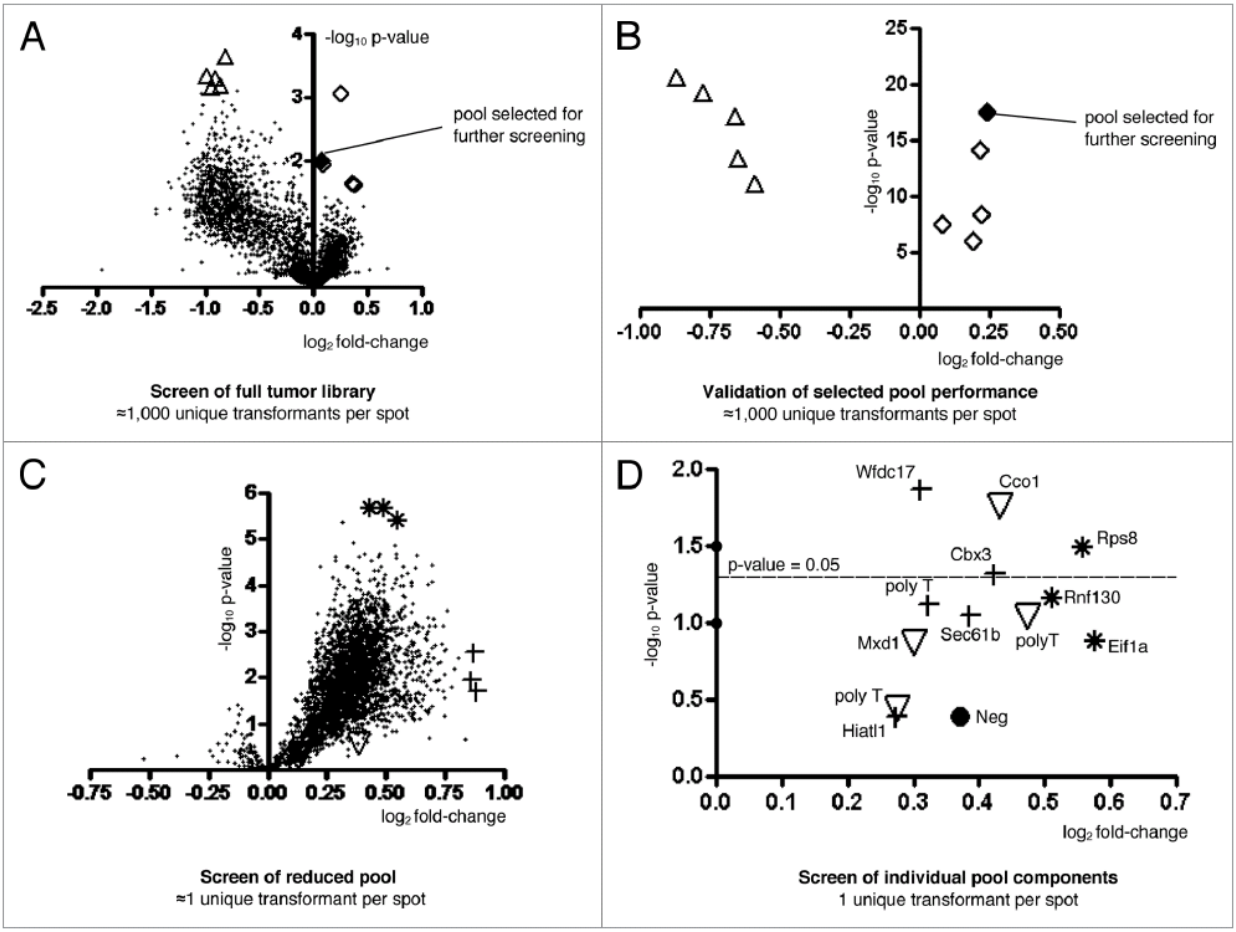


Figure 71 Scatterplots of tumor library lysate screens

Naive and tumor sera was applied to the tumor library and sublibrary features. Values of features from tumor to control are displayed for the screen of the tumor library (A). At this stage of the

screen, there are ≈1,000 unique transformants per spot. In (B) the validation of performance of selected lysate pools from the tumor library screen is presented. In this validation experiment to confirm that selected pools exhibited reproducible behavior in a new sera screen, there are still ≈1,000 unique transformants per spot. The most significant pool was partitioned onto a new array in the pool reduction screen presented in (C) with approximately one unique transformant per spot. A total of 9 spots were selected from this screen and partitioned onto a new array for the single clone array screen presented in (D). In this screen each spot was comprised of exactly one single unique transformant. The categories of features are described in Table 13*. In (D), each plotted lysate result is labeled with a wild type transcript which the cloned sequence aligned with when using BLAST.*

3.3.1.2 Validation of performance of selected lysate pools from tumor library screen

A validation array containing only the ten lysate pools from the three million clone library selected as reacting significantly different to the tumor vs. control sera, and the five negative control pools was printed and tested. The results support the outcomes obtained from the original full screen. The least significant p-value lysates in the tumor library screen had the least significant p-value in the validation screen with p-values and fold changes that would place them near the origin of the p-value vs. fold change plots. The lysate pool within the increased intensity set with the most significant p-value (2.93E-18) was selected to partition and re-array on a nitrocellulose slide as distinct features for further testing and reduction of pool complexity (Table 13 and Figure 71B).

3.3.1.3 Pool reduction screen

Since the number of original transformants being expressed within each pool was estimated to be 1,000, the identification of an individual antigen requires further testing. The sublibrary pool with the most significant p-value and increasing tumor intensity was partitioned and regrown in 3,000 microtiter wells. The 3-fold redundancy is to facilitate capturing a complete sublibrary representation, and each new sample is estimated to be comprised of approximately one to three unique clones. Slides were printed and tumor test and naïve control sera were

applied. Analysis of fluorescent reactivity readouts was used to select several samples for further characterization. Three lysate samples were selected based on the best p-value for increase-reactivities (2.05E-6, 2.05E-6, and 3.80E-6; q-values of 5.55E-4, 5.55E-4, and 5.84E-4), three lysate samples were selected based on the highest fold change in reactivity relative to control sera (1.81, 1.82, and 1.84), and three lysate samples were selected based on the highest measured reactivity levels (10,052; 9,205; and 9,598) (Figure 71C).

3.3.1.4 Single clone array screen

A final set of lysate slides were constructed that contained single clones derived from these nine lysate pools (Figure 72), which were selected based on their p-value, fold change, or high intensity. This was done by sequencing ten single colonies from each of the nine complexity-reduced pools. These were determined to contain from one to three unique clones. Therefore, the final lysate array consisted of three lysates derived from the high p-value lysate pools, five lysates derived from high fold change lysate pools, and four lysates from high intensity lysate pools. Lysates from a bacterium producing no library-derived clone was included as a negative control. A scatterplot of the fold changes in reactivity vs. p-values between test and control demonstrates that one of the clone-lysates from each of the three characteristic-category displayed a significant increase in reactivity to tumor sera relative to naïve sera (Figure 71D). Note that in Figure 71D, the category of each single clone feature represents the category of the 3,000 component sublibrary feature from which the single clone feature was derived from. In other words, the Eif1a feature did not have one of the most significant p-values in the single clone array screen, but the Eif1a feature was derived from a feature pool which did have the most significant p-value in the 3,000 component sublibrary screen.

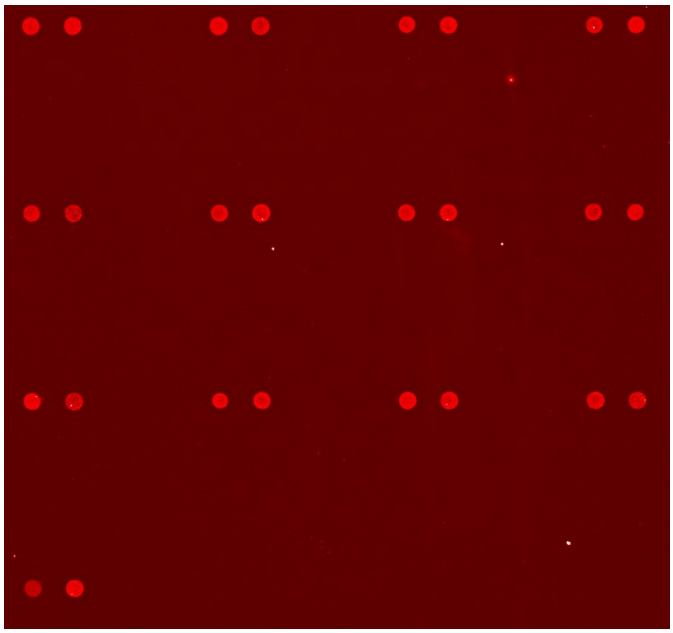


Figure 72 Single clone lysate array

In this image there are replicate spots, but they were printed in reverse order so that replicate spots are not adjacent.

3.3.2 Sequence information

Sequence information was obtained for the single clones tested as lysates on the final nitrocellulose arrays. Most of these sequences matched the 3' end of wild-type mouse RNA transcripts but were truncated at the 5' end. Some of the cDNAs inserted into the plasmid out of frame relative to the wild type protein (5/9) or in reverse orientation (1/9). Several of the sequences also contained point mutations relative to the murine database. Table 14 summarizes some of the characteristics of these sequences. Note that the CcoI sequence is an unusual sequence since it contains the 3' end of the CcoI mouse mitochondrial gene transcript upstream of the tRNA-Ser sequence and a poly A tail. For the reverse-oriented HiatI1, a ribosome translating this would slip along the poly T region present in the sequence and all three reading frames would be expressed which is why multiple lengths to the stop codon are listed in the table for this transcript. The Wfdc17 sequence almost spans the full wild-type sequence; there are only 12 bp missing from the 5' end.

Table 14 Information for sequences in clones

Sequence[a]	Length (bp)[b]	Length to SC (bp)[c]	Forward Orientation?[d]	5' truncated?[e]	Frame-shift?[f]	No. of PM[g]	Extra Seq (bp)[h]
Rps8	443	9	Yes	Yes	Yes	0	0
Eif1a	288	51	Yes	Yes	No	0	0
Rnf130	467	276	Yes	Yes	No	2	17
Cbx3	483	27	Yes	Yes	Yes	0	0
Hiatl1	676	336 or 42 or 12	No	Yes	Yes	2	0
Sec61b	278	177	Yes	Yes	Yes	0	0
Wfdc17	458	111	Yes	Yes; only 12 bp missing	Yes	0	0
Mxd1	101	20	Yes	Yes	No	0	0
Cco1	447	18	Yes	Yes	No	3	0

Information for the sequences present in the tumor cDNA library clones in the final round of screening with tumor and naïve mouse sera. [a]The wildtype mouse transcript which the cloned sequence aligned with. [b]The length of the inserted sequence. [c]The length to the stop codon from the insertion site. [d]Indicates whether the sequence inserted into the plasmid in the forward orientation relative to the Lac promoter of the bacterial expression plasmid. [e]Indicates whether the sequence in the clone is missing the normal 5' end of the corresponding wild type transcript. [f]Indicates whether the sequence expressed from the Lac promoter in the bacterial plasmid is frame-shifted relative to the frame of the wild-type transcript. [g]Indicates the number of point mutations in the sequence. [h]Indicates how many additional base pairs are present in the sequence which are not present in the wildtype transcript sequence.

3.4 Discussion

3.4.1 Summary

The goal of this research was to explore methods for screening tumor immunogens efficiently, and much knowledge has been acquired throughout this process. The basic process utilizes several technologies which were not widely available when the SEREX method was first developed. Such technologies include automated equipment to handle large numbers of 96-well plates, accurate printing machinery, and high resolution scanners. The process of screening

multiple rounds of pooled lysates also allows the researcher to hone in on a particular clone of interest beginning from clones that cover the entire transcriptome. Several potential tumor-specific antigens have also been identified.

3.4.2 Experiment conditions

These rounds of cDNA library screening have produced some candidate antigens worthy of further investigation. There are several lysates which bound to tumor sera more than naïve sera at the last stage of screening. Note that the last round of screening resulted in less significant p-values than was obtained in the previous rounds of screening. This last round of screening was repeated several times, and p-values comparable to the previous p-values were never obtained (data not shown). A likely reason for these less significant p-values is that the protein and lysate production for this batch of slides was not of the same quality as the protein and lysate production of the previous experiments. Nevertheless, p-values less than 0.05 were still often obtained. Note that several primary sera dilutions were also tried for this single clone screen and some counterintuitive results were obtained. The greatest fold change values between naïve and tumor with higher intensity for the tumor sera were obtained at the lower dilutions used. Perhaps this reflects greater affinity demanded for successful binding at the lower antibody concentrations, thereby reducing off target binding. Dilutions of 100 fold, 500 fold, 1,000 fold, and 2,000 fold were tested, and the results from the 1,000 fold dilution are presented in Figure 71D. A general observation from the experiments performed is that the Rps8, Wfdc17, Cbx3, Rnf130, and Sec61b clones often bound to sera better than other clones such as Mxd1 and the poly T clones. Some of the clones at this final stage of screening came from the same pool in a previous round of screening, but this does not imply that only one of the sequences is tumor specific. On the contrary, a lysate pool may have bound to antibodies in the tumor sera more than naïve because it contained more than one positive clone (tumor specific sequence). With lysate from a higher quality lysate preparation, there is the possibility that all of the positive clones might demonstrate higher binding reactivity to tumor than normal sera.

3.4.3 Transcript sequences and cancer

Do any of these sequences encode for proteins that have functions which relate to cancer in any way? Wfdc17, CcoI, Rps8, and Cbx3 produced the best p-values in a screen with tumor vs. control sera. The protein encoded by Wfdc17 acts as a counter-regulator of proinflammatory responses [143]. No papers associating this sequence with cancer were found. CcoI is a part of a complex in the mitochondria electron transport chain [144]. This protein has been found to have altered expression levels in prostate cancer [145]. The Rps8 gene encodes a ribosomal protein that is a component of the 40S subunit [146]. Increased expression of this gene and other ribosomal proteins has been observed in some cancers such as colorectal cancer [147]. The protein encoded by Cbx3 can act as a transcriptional regulator and is a component of heterochromatin [148]. This sequence has previously been associated with neoplastic transformation and progression [149]. Rps and Cbx sequences are currently present in the Cancer Immunome Database, but Wfdc and Cco proteins are not [128].

The other sequences had a p-value less than 0.05 in the screen displayed in Figure 71 D. Rnf130 encodes a protein that is a zinc finger protein structural domain, and these proteins are often involved in the ubiquitin degradation pathway [150]. RNF130 interacts with miRNAs which are overexpressed in tumors [151]. The protein encoded by Sec61b is a component of the machinery necessary to translocate proteins across the endoplasmic reticulum [152]. Sec61b has been found to be one of the top 20 genes down-regulated by miR-133a which is a tumor-suppressive microRNA [153]. Mxd1 is a dimerization protein that competes with MYC for binding to MAX to form a DNA binding complex [154], and a mutated version of Myc is found in many cancers [155, 156]. Eif1a is an RNA binding protein essential for translation initiation [157]. The EIF1A locus has been determined to be hypermethylated in human ovarian carcinoma CP70 cells [158]. Hiatl1 belongs to the major facilitator superfamily and is involved in facilitating transport across membranes [159]. The expression level of Hiatl1 was found to aid in the classification of gastric cancer [160]. Sec, Rnf, and Eif sequences are currently present in the Cancer Immunome Database, but Hiatl, and Mxd sequences are not [129].

Analysis of the sequences in the clones reveals that most of the sequences were 5' truncated. The library was constructed using poly dT primers which would bind to the 3' poly A tail of RNA transcripts. Although the reverse transcriptase used, polymerase used, and the SMART (Switching mechanism at 5' end of RNA template) mechanism should have resulted in mostly full-length transcripts, many 5' truncated transcripts were present. There could be important information upstream of the 3' region in the original RNA transcript which was not detected in the library. Such information could include certain mutations as well as possible gene fusions which could cause expression of a frameshifted string of amino acids. A number of the sequences of the clones in the library were frameshifted relative to wild type murine databases. Some of these frameshifts may correspond to genomic events; others may be at the transcript level such as alternative splicing or trans-splicing events. Frameshifts would be good candidate proteins or genes to include in a vaccine because frameshifts would be more immunogenic since they would be unique or more common in the tumor cells.

3.4.4 Speculation about translation errors and cancer

The ultimate goal of this research was to identify specific transcripts which produced proteins recognized by the immune system of animals with tumors. Ideally these antigens would be frameshift or chimeric transcripts unique to the tumor cells and not found in normal cells. These transcripts could then be used to study cancer further or to develop a cancer vaccine. However, after analyzing the results of these immune screens, there may be another approach to a cancer vaccine. Instead of searching for specific frameshift or chimeric transcripts, one could develop a cancer vaccine with the frame-shifted versions of abundant proteins.

More than half of the transcripts that made it into the final single clone screen (“3.3.1.4 Single clone array screen”) would produce frame-shifted proteins. This frame-shift is not the result of a known frameshift in the RNA transcript in the original cell, but instead is actually an artifact of the cDNA construction process since transcripts could be truncated randomly at any position along their length. These truncated transcripts would insert into the plasmid and be expressed in a certain frame relative to the bacterial *lac* promoter, and this frame may or may not

be true to the original frame in the tumor cell. If one could not find an actual frame-shifted transcript in the tumor cell, one might conclude that antibodies don't actually bind to this frame-shifted transcript. Instead, the results of any antibody detection are just an anomaly. If the protein were an abundant protein such as a ribosomal protein like Rps8 detected in this screen (Table 14), there would be an even greater reason to suspect that antibody detection does not reflect actual biological events because these abundant transcripts are more likely to be present in clones in the tumor cDNA library. These abundant clones would have a higher chance of being detected.

However, what if these antibody detection events are not an anomaly, and there also is no frame-shifted transcript in the tumor cell? Perhaps, tumor cells produce more frame-shifted proteins from wild-type transcripts during translation even though there are no frame-shifted transcripts for many proteins originally in the cell. This increased error rate during translation in tumor cells may be due to an increased rate of translation as the tumor cells rapidly proliferate, or these errors may be due to a decrease in quality control. Whatever the cause may be, if it is true that tumor cells produce more frame-shifted proteins during translation than normal cells, then one could create a cancer vaccine simply by including many frame-shifted proteins. Which proteins would be the best to choose? Perhaps frame-shifted versions of house-keeping proteins would be the best to choose since there is a high abundance of house-keeping proteins and there would be the most frame-shifted versions of these proteins for the tumors to present. Therefore, an effective cancer vaccine might simply consist of frame-shifted versions of abundant proteins since there may be more errors during translation in tumor cells.

3.5 Conclusion

A high-throughput platform for screening tumor cDNA libraries was used to discover several immune reactive proteins. The platform takes advantage of automated equipment to handle large numbers of PCR plates, small nitrocellulose slides, high resolution scanners, and the screening of several rounds of pools. In this demonstration of the technology, nine sequences were discovered which are potentially immunoreactive, and at least one of the proteins is a very well-

known protein involved in cancer pathways (Mxd1). Several of the sequences would produce frame-shifted proteins, but chimeric transcripts which would cause these frame-shifts were never discovered in an investigation with PCR. These results inspired the hypothesis that the immune system may be producing antibodies against frame-shifted proteins which resulted from defective translation of protein in tumor cells, instead of defective DNA or RNA. If defective translation is more abundant in tumor cells than normal cells, then a vaccine composed of a cocktail of frame-shifted proteins may prove effective. Further investigation will be required to support or disprove this hypothesis, as well as to identify protein sequences which would be effective in a subunit cancer vaccine.

4: RANDOM SEQUENCE MIMOTOPES

4.1 Introduction

Characterizing the interactions between disease-specific antibodies and their cognate antigens has proven highly informative in the study of host-pathogen relationships and critical in the development of effective biomedical products. Similarly in chronic diseases, the discovery of modified antigens or autoantigens that are specifically recognized by patient antibodies is of growing importance in disease research and target development for diagnostics, vaccines, and therapeutics. These complexes are typically found by querying immune sera against possible ligands in lysates, or in libraries of proteins or peptides made recombinantly or synthetically. A myriad of binding assays such as immunoblots [161], ELISAs [162], phage display [163], ribosome display [164], beads [165], and microarrays [166] has been employed to identify the antigen recognized by an antibody and the epitope sequence it binds. While the pros and cons of each assay vary with respect to cost, simplicity, bias, breadth, and sensitivity, these methods have collectively contributed significantly to our knowledgebase of disease-relevant antigen/antibody complexes. However, a rate limiting step toward more comprehensively identifying immune targets of biologically relevant antibodies has been developing a robust method for screening ligands against complex antibody mixtures.

In its original description, phage display was used to survey a library of peptides for binding to a given antibody [167]. It has now been used extensively to display libraries of peptides or antibody fragments, expressed as coat protein fusions on the phages surface, for “panning” against a particular molecule of interest. Phages are washed across an isolated, immobilized target; bound recombinant phage are collected and amplified in bacteria for additional rounds of panning [168]. One of the major drawbacks of phage display is the technique’s reliance on multiple *in vivo* steps that cannot be well-controlled and incur biases to the output. For example, any peptide-coat protein fusions that reduce the fitness of the phage or reduce secretion to the phage surface will not be well represented if at all. In the initial panning round each phage recombinant is present in such limited numbers that the probability of a ligand finding a target can be

stochastic. Yet only those recombinants that survive this first round are subsequently propagated reiteratively. As an alternative, *in vitro* translation systems such as ribosome display have been developed for studying protein-protein interactions, including antigen/antibody binding. Like phage, very large libraries can be constructed at minimal cost but the diversities of these recombinant mixtures are difficult to maintain and are not reproducible. In addition, apparent diversities can be misleading since the redundancy of the genetic code, incidental stop codons, and peptide-dependent effects on translation efficiency will limit the ultimately displayed diversity.

In vitro combinatorial synthesis of peptides on beads and microarrays of either proteins or peptides have been explored as library formats for surveying target binding [165, 166]. Both of these methods are performed entirely *in vitro*, and thereby resolve the vagaries of *in vivo* propagation and biological compatibility. Since peptides are used directly, the issues of translating DNA are avoided. However for libraries in bead format, the binding steps must be followed by decoding what is bound through peptide-sequencing, chemical-tracking, or other reading methods. The synthesis, binding, and decoding steps tend to be laborious, time consuming, and often lack reproducibility [169]. For libraries in either bead- or array- format, peptide synthesis costs limit the size and number of libraries that can be built, relative to the recombinant libraries. Furthermore, the antigens or epitopes within the antigen may not be in the known proteome. Studies can be confounded by the fact that immune sera often carry antibodies to mutant, unknown, or even exogenously-derived antigens of a host or pathogen proteome. In these cases, lysates or libraries of proteins or epitopes that are made recombinantly or synthetically will be incomplete.

Non-natural sequence peptide libraries, whether in vivo or in vitro, provide a means for identifying mimotopes of unknown antigens and are far more economical since one library can be used for all screens. For example, an antibody panned against phages displaying random sequence peptide fusions might select recombinants that mimic or hold similarity to a previously unknown ligand [170]. Microarrays consisting of random sequence peptides or peptoids have been explored for identifying ligands [171]. Probing of a SPOT synthesis array carrying 5,520 random 15-mers with three different monoclonal antibodies followed by substitutional analysis was able to identify mimotopes of the known wild type epitopes [7].

Each of the described methods, using either natural or random sequences, requires knowledge and preparation of either the antibody or ligand, in order to identify the other. If antigens instead of epitope-size peptides are screened, then epitope identification requires additional knowledge. Namely, antigen sequence information and a sufficient budget would be required to perform "peptide tiling" assays. Broader identification and characterization of disease-relevant antibody targets will require an inexpensive method for performing screens for these antigen/antibody interactions without specific knowledge of any of the components. Recently, an array platform has been developed in which 10,000 non-natural sequence peptides are printed onto a functionalized glass slide and complex sera mixtures are surveyed in parallel [9]. Binding analyses have been able to consistently define patterns of immune responses against diseases [172]. This non-natural sequence peptide array platform is applied to the design of a method for capturing disease-relevant antibodies and identifying their biological targets.

As proof of principle for demonstrating the ability of non-natural sequence peptide arrays to identify disease specific antigen/antibody complexes, a tumor-associated mutant of the structural maintenance chromosome 1A protein (SMC1A) and immune serum were used. The SMC1A protein was first described in yeast as an essential protein necessary for nuclear division [173, 174]. It is also involved in human cell division, serving a number of functions including roles in the stabilization of sister chromatids during their replication and separation, DNA repair, activation of S phase check point, and regulation of gene transcription [175]. Given the critical activity of SMC1A in regulating cell division, the possible effect of its mutation in progressing cancer has been suggested [175]. In fact increased levels of a particular aberrant transcript encoding a frame-shifted SMC1A (SMC1Afs) has been correlated with tumor cells relative to normal cells (Luhui Shen, manuscript in prep). This exon 4 reading frame shift produces a truncated protein ending with 17 unique amino acids. Here I demonstrate that this non-natural sequence peptide microarray screening process can be used to isolate immune-specific antibodies and to identify important epitopes to which α-SMC1Afs antibodies specifically bind. The same methods could be used to identify epitopes in previously unknown antigens and aid in the discovery of important targets for any infectious or chronic disease.

4.2 Materials and Methods

4.2.1 Peptides and Beads

A number of free peptides as well as Tentagel bead conjugated peptides were used in experiments. The sequences of these peptides and their assigned names are presented in Table 15.

Table 15 Amino acid sequences of free peptides and peptide-bead conjugates

Assigned Name	Sequence	Description
SMCfs	CCGIYCHEEPQREDSSI	human SMC1A frameshift 17mer peptide
SMCfs-27	TAIIGPNGSGCCGIYCHEEPQREDSSI	human SMC1A frameshift 27mer peptide
RP1	TISKYVMVEPMRQHEEWGSC	SMCfs mimotope
RP2	AVSHQEMNEGEQGPMREGSC	SMCfs mimotope
RP3	RVGEMPMREYDISGGSGGSC	SMCfs mimotope
RP4	TAFYRTLTKHEVDPGIAGSC	SMCfs mimotope
CP1	AVLLMCQLYQPWMCKEYYRLL	negative control peptide which is not a mimotope of SMCfs
SMCfs-B	CCGIYCHEEPQREDSSI	human SMC1A frameshift 17mer peptide conjugated to Tentagel beads
RP1-B	TISKYVMVEPMRQHEEWGSC	SMCfs mimotope conjugated to Tentagel beads
RP2-B	AVSHQEMNEGEQGPMREGSC	SMCfs mimotope conjugated to Tentagel beads
RP3-B	RVGEMPMREYDISGGSGGSC	SMCfs mimotope conjugated to Tentagel beads
RP4-B	TAFYRTLTKHEVDPGIAGSC	SMCfs mimotope conjugated to Tentagel beads
CP2-B	ATKAAIPGPNTVPRAP	negative control peptide which is not a mimotope of SMCfs conjugated to Tentagel beads

The amino acid sequence and name of free peptides and peptide-bead conjugates used is listed. The "B" at the end of the assigned name indicates that the peptide is conjugated to a bead, and the absence of a "B" indicates that the component is a free peptide.

4.2.2 Rabbit α-SMCfs Sera

Serum against the human SMC1Afs protein was generated by Global Peptide Service LLC (Fort Collins, CO). The 17 amino acid (a.a.) SMC1A frameshift mutant sequence (SMCfs),

identified in human tumor cDNA (CCGIYCHEEPQREDSSI), was synthesized by the peptide synthesis lab at Arizona State University and then conjugated to keyhole limpet hemocyanin (KLH) by Global Peptide Service LLC. A New Zealand white rabbit was immunized with the SMCfs-KLH conjugate. Blood for the experiment was collected at exsanguinations after two immunizations.

4.2.3 ELISA

ELISA plates were coated with 50 µL of 10 µg/mL of peptide or protein in carbonate coating buffer and incubated at 4 °C overnight. The coated plates were washed 3X with PBST and blocked with 200 µL of 3% BSA in PBST at 37 °C for 30 minutes. The blocked plate was washed 3X with PBST and 50 µL of primary anti-serum or purified antibody diluted in 3% BSA in PBST was applied. The plate was then incubated at 37 °C for 1 hr. After the incubation, the plate was washed 3X with PBST. The antibody was detected with 50 µL HRP-goat α-rabbit IgG diluted 1:2,000 in 3% BSA in PBST. After the plate was incubated at 37 °C for 1 hr, the plate was washed 3X and developed with 50 µL TMB for 10 minutes at room temperature. The development was stopped by adding 50 µL of 0.5 N HCl, and the plate was read with a SpectraMax 190 Molecular Devices instrument at OD 450 nm.

4.2.4 Antibody absorption

Specific antibodies were absorbed from the rabbit α-SMCfs sera by applying the sera to the SMCfs coated plate. The rabbit sera was diluted 1:250 with 3% BSA in PBST and incubated with the SMCfs peptide coated plate at 37 °C for 1 hr. The unbound antibody in the supernatant was then removed and applied to another peptide coated well to remove more antibody specific for the peptide. This process was repeated up to 20 times, and this serum was then later applied to the peptide microarray at a dilution of 1:500. This same method was used to produce negative control antibody absorbed sera using the negative control CP1 peptide.

4.2.5 Non-natural sequence peptide array printing

The 10,000 randomly generated peptide sequences were custom synthesized by Sigma, Inc. These 20-mers were designed with 17 random a.a. residues (excluding cysteine) and a 3 amino acid (GSC) linker sequence on the C terminus. The C terminal cysteine binds to a sulfo-SMCC coated aminosilane glass slide. The solutions of different non-natural sequence peptides were printed onto the glass slide using a NanoPrint 60 instrument (Arrayit Technologies).

4.2.6 Application of sera to non-natural sequence peptide array

Rabbit sera samples were applied to the non-natural sequence peptide microarray using a Tecan HS 4800 Pro microarray hybridization station. Slides were first washed for 30 s with TBST, and then blocked with a blocking buffer consisting of BSA, mercaptohexanol, Tween 20, and PBS for 1 hr at 23 °C. Duplicate samples of sera were diluted 1:500 in an incubation buffer consisting of BSA, Tween 20, and PBS and incubated with the slide for 1 hr at 37 °C. The slide was then washed, and 5 nM of goat α-rabbit IgG conjugated with AlexaFluor 647 dye was applied for 1 hr at 37 °C. The slide was then washed and dried for 5 min.

4.2.7 Scanning and analysis of array

The slides were scanned with an Agilent Technologies DNA Microarray Scanner with Surescan High-Resolution Technology instrument and analyzed with GenePix Pro 6.0 software (Molecular Devices, Santa Clara, CA) to determine the fluorescence intensity of each spot. GeneSpring GX 7.3 (Agilent Technologies, Santa Clara, CA), Microsoft Excel (version 15.0.4551.1003) [69], simple custom Java code, and GraphPad Prism 4 (GraphPad Software, La Jolla California USA, www.graphpad.com) were then used to perform further analysis of this data.

4.2.8 Antibody purification

Specific antibodies were purified from sera by flowing sera through a column filled with TentaGel beads with synthesized peptides on their surface such as the SMCfs peptide, selected non-natural sequence peptides, or the irrelevant CP2-B peptide. The total IgG of the rabbit α-SMCfs sera was purified using Pierce Protein A/G Agarose beads with the protocol of the

manufacturer. A volume of 1 mL of the TentaGel beads were mixed with 3 mL of the purified total IgG, and this solution was incubated for 1 hr at room temperature. The column was then washed with 10 mL PBST. The specific IgG was eluted with 4 fractions of 1 mL IgG Elution Buffer (Pierce, Inc) and each fraction was neutralized with 100 μl 1M TRIS. All of the eluted fractions were measured at an absorbance of 280nm. The two fractions with the highest absorbance were combined and used for further analysis at a 1 to 40 dilution in 3% BSA in PBST.

4.2.9 Analysis of Motifs

GLAM2 software [176] was used to identify and score common motifs among the high binding peptides. Bepipred [177] was used to find and score B cell epitopes.

4.3 Results and Discussion

4.3.1 Peptide microarray screening of α-SMCfs serum

Antibodies against the α-SMCfs peptide were generated by immunization with the synthetically produced 17-mer conjugated to KLH. Harvested rabbit serum was applied to a functionalized microarray slide that had been printed with a NanoPrint instrument with a defined set of 10,000 (10k) non-natural sequence 20-mers, as detailed in the section ("4.2.5 Non-natural sequence peptide array printing"). Serum-antibodies bound to the peptides after washes were measured using fluorescently-labeled α-IgG secondary antibody reagents. Genepix software (Molecular Devices) was used to align the feature-reactivities on the array, and GeneSpring (Agilent Technologies) was used to analyze the results by quantifying fluorescence units at each feature. Both specific antibody levels and their binding avidities are components of these fluorescence level readouts. From this queried peptide library, 108 peptides exhibited highly specific binding to the immune serum relative to naïve rabbit control serum (p <0.0001). In Figure 1, fluorescence of each of the 108 peptides, as a measure of antibody binding activity, are compared between SMCfs-immune and naïve sera. A color range from green to red is used to display these fluorescence intensities and is represented as a heatmap. This analysis allows

visualization of the differential profiles of antibody-reactivity of the immune and naïve sera to the selected set of non-natural sequence peptides.

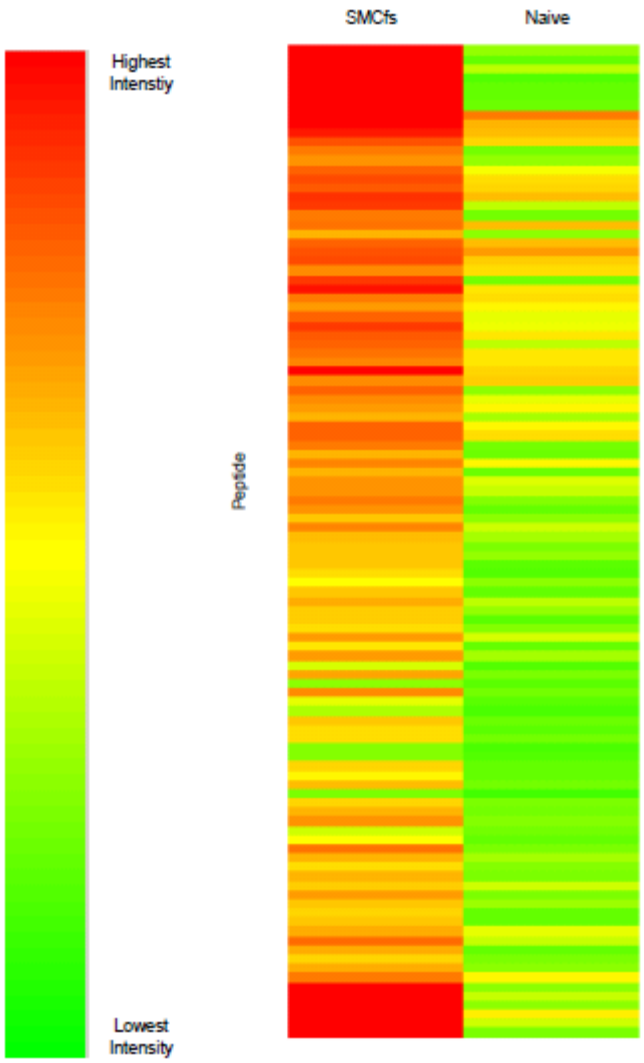


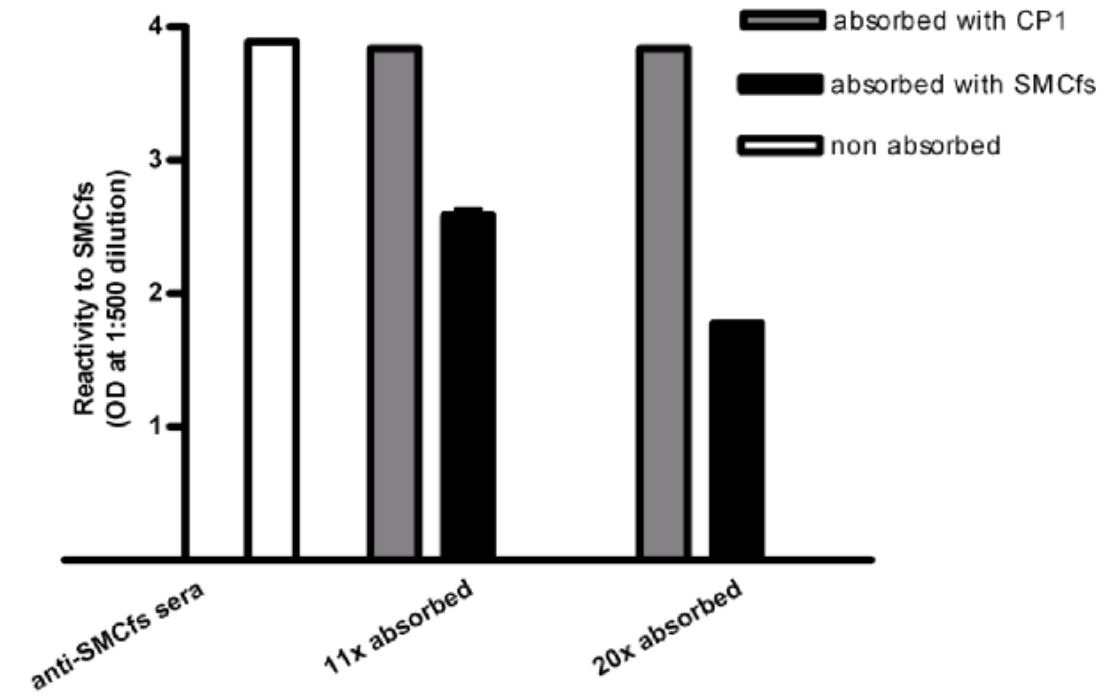
Figure 73 Selective binding of α-SMCfs serum to a set of random sequence peptides displayed on a microarray

Polyclonal rabbit serum generated against the SMCfs peptide conjugated to KLH and naive rabbit serum were applied to an array of 10,000 synthetic peptides of randomly generated sequence. Student's t-test analyses comparing the probing results identified 108 random 20-mers with differential binding to the immune versus naïve sera (p-values <0.0001). These differences in peptide-binding intensities are visually represented as a heatmap.

4.3.2 SMCfs immune serum absorption and array analysis

Total α-SMCfs immune serum contains many different antibody species, including antibodies to the co-administered KLH adjuvant carrier molecule. To differentially measure the SMCfs-peptide antigen reactivity from that of KLH, an absorption experiment was conducted so as to remove the α-SMCfs specific antibodies. Analysis of the depleted serum would indicate SMCfs specific peptide reactivities, by their specific loss. The immune serum was bound to wells of a plate coated with SMCfs, and absorption was performed in a step-wise manner. ELISA-based confirmation of antibody depletion was assessed after 11 and after 20 re-iterative rounds of absorption. The results of an ELISA experiment performed with the peptide-depleted serum samples are shown in Figure 74A. Binding to the SMCfs peptide decreased as serum samples subjected to additional rounds of re-absorption and re-testing were used; binding of the absorbed samples to an irrelevant peptide (CP1) peptide was unchanged.

A)



B)

Figure 74 Analyses of α-SMCfs serum pre-absorbed against its cognate SMCfs peptide.

A) Validation of SMCfs antibody depletion from immune serum. The serum-absorption steps (11 or 20) refer to the number of iterative rounds of SMCfs-peptide or CP1-peptide absorption experiments conducted before application of the depleted serum to ELISA plates coated with SMCfs peptide. B) Changes in non-natural sequence peptide binding intensities following α-SMCfs antibody depletion of serum. The depleted immune sera samples were applied to the peptide microarray. Binding intensities to the 108 peptides, which were shown to be selectively recognized by the original α-SMCfs serum, are displayed as a heatmap for their visual comparison. Four α-SMCfs sera samples were applied to the peptide microarray, 1) non-

absorbed α-SMCfs serum, 2) α-SMCfs serum absorbed 11x against SMCfs peptide, 3) α-SMCfs serum absorbed 20x against SMCfs peptide, 4) α-SMCfs serum absorbed 20x against the CP1 negative control peptide Colors toward red indicate highest relative fluorescence intensity, and colors toward green indicate lowest relative fluorescence intensity.

The α-SMCfs antibody-depleted samples, as well as the undepleted serum, were applied to the 10k non-natural sequence peptide array. Reactivity profiles of these samples for the 108 peptides selected in the original screen were compared, and are presented in Figure 74B. The naïve sera reactivity for these 108 peptides is found in Figure 73. A range of profiles were observed. For example, some peptides displayed binding intensities higher than that of naïve serum when probed with serum samples absorbed against the negative control irrelevant peptide, but also showed these intensities when probed with the α-SMCfs-absorbed serum samples. Other peptides displayed a decrease in signal when probed with the α-SMCfs antibody depleted samples, but also showed decreased signal with the control absorbed serum. Lastly, a set of peptides showed decreasing fluorescence intensities when serum from increasing rounds of α-SMCfs depletion were applied. These peptides showed no loss of signal when probed with the control-absorption serum sample. Based on these evaluations of specificity, each of the 108 peptides was assigned a score. The peptides were scored such that large scores were assigned to peptides that displayed low binding to the SMCfs-absorbed serum samples and high binding to undepleted sera. The score was assigned to each peptide as follows:

$$peptide\ score = \frac{\mathbf{max\ signal} - absorbed\ signal}{max\ signal} + \frac{negative\ control\ absorbed\ signal}{\mathbf{max}\ signal}$$

Equation 7 Peptide score

Peptide score assigned to each peptide where the variables are as follows: max signal refers to the signal of the peptide with the greatest intensity; absorbed signal refers to the signal of a peptide with α-SMCfs serum absorbed 20x against SMCfs peptide; and negative control peptide refers to the signal of a peptide with α-SMCfs serum absorbed 20x against the CP1 negative control peptide.

Four of the top 5 scoring peptides were selected for re-synthesis and further testing against the α-SMCfs serum.

4.3.3 ELISA-based validation of α-SMCfs serum binding to screen-selected peptides

To validate that the peptides which were identified in the peptide-array screening were specifically bound by the α-SMCfs immune serum, standard ELISA assays were performed. The ELISA determinations confirmed the results of the peptide array: the 4 array-selected peptides could be specifically recognized by the rabbit α-SMCfs sera whereas the negative control CP1 peptide was not (Figure 75A).

Since the α-SMCfs immune serum was generated by immunization with a SMCfs peptide attached to KLH, the specificities of the immune serum-selected peptides for SMCfs were compared to KLH. The IgG-binding activity of each selected peptide against mouse α-KLH sera was measured. The results demonstrated that all four of these selected peptides as well as the SMCfs peptide were not recognized by α-KLH antibodies; while the cognate KLH peptide bound strongly (Figure 75B). Therefore, the array-selected peptides behaved as mimotopes of the SMCfs peptide and could be specifically recognized by antibodies against SMCfs peptide. The reactivity of the rabbit α-SMCfs serum to the four selected peptides was lower than that to the SMCfs peptide, as measured by α-IgG-fluorescence readout. This quantitative difference indicates that the selected mimotope peptides are lower affinity ligands relative to the original SMCfs immunogen. This may be due to less optimal or partial epitope sequences.

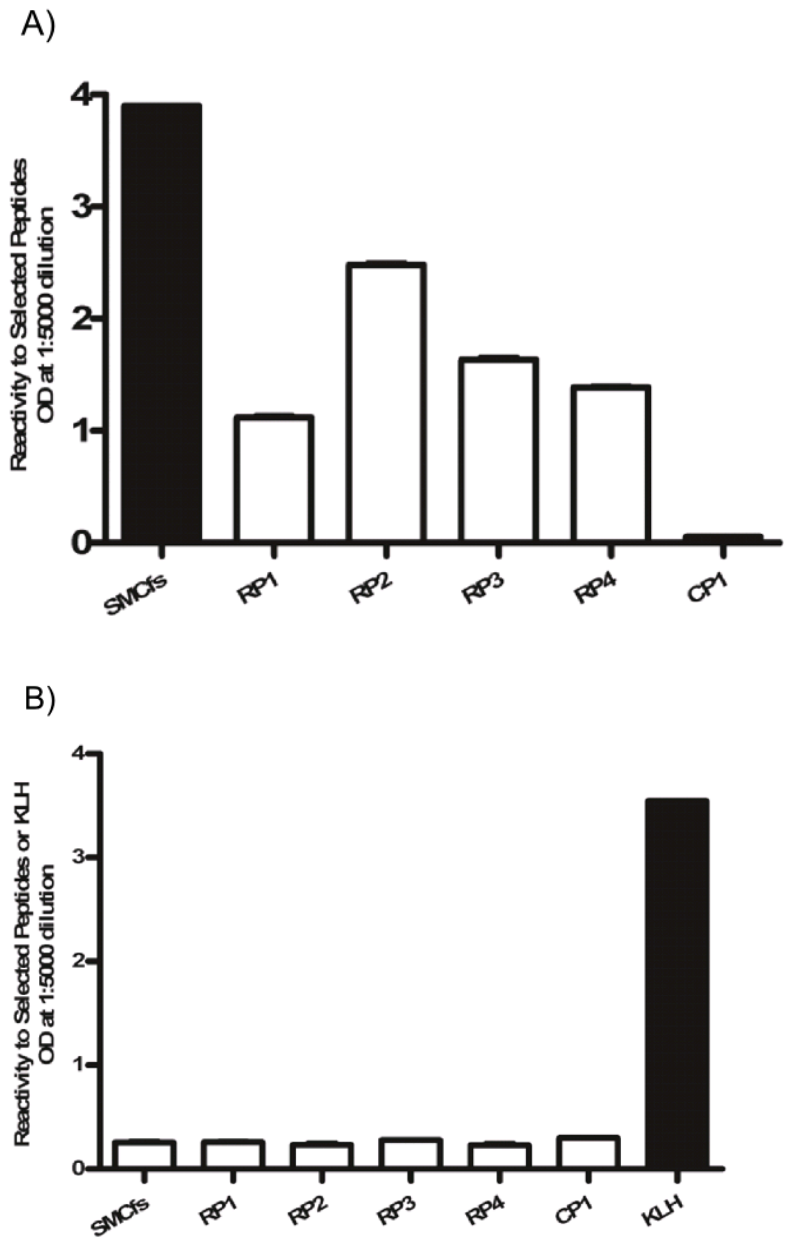


Figure 75 ELISA determinations of α-SMCfs sera binding to array-selected peptides

A) Immune serum: rabbit polyclonal α-SMCfs serum was applied to the cognate peptide-antigen, four of the array-selected peptides, and a control peptide. B) Control serum: mouse polyclonal α-KLH serum was applied to the same set of peptides described in A.

4.3.4 Affinity purification of α-SMCfs antibodies with mimotope peptides

To determine whether the non-natural sequence peptide mimotopes could be used to specifically isolate antibodies against the original immunogen, Tentagel-attached peptides were prepared. Columns containing beads conjugated to one of the non-natural sequence peptides, the SMCfs peptide, or an irrelevant peptide (CP2-B) were used to capture antibodies from the α-SMCfs immune serum. The binding specificities of the peptide-purified antibodies were measured by determining the activity of the eluted antibodies against the SMCfs-27 or KLH antigen by ELISA (Figure 76). The SMCfs-27 is comprised of the 17 unique, frame-shift created residues plus the 10 amino acids found in the wild type protein immediately upstream of the frame shift. This 27mer was used so as to accommodate any possible antibody binding that spanned the normal/mutant sequence junction. However when the 17-mer SMCfs peptide was replaced with SMCfs-27 in an ELISA, the results were indistinguishable from those shown in Figure 3a. Therefore, the frameshift junction site does not appear to be critical for the α-SMCfs antibody binding activity within this immune serum.

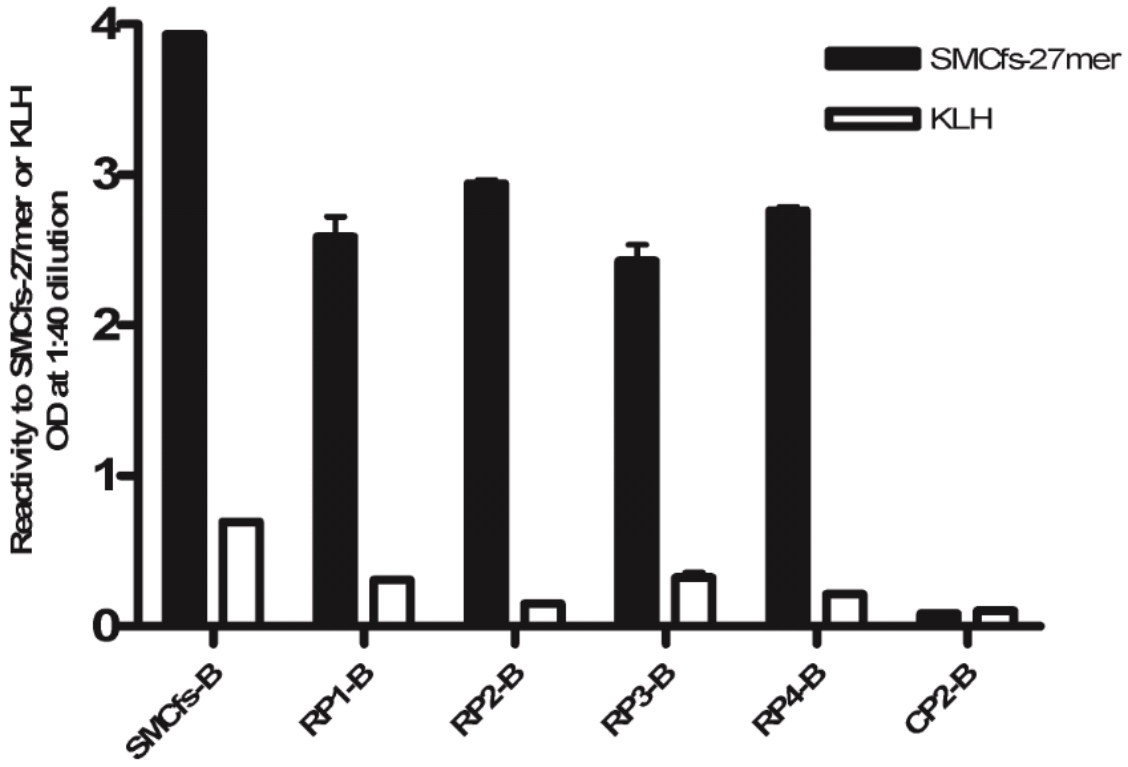


Figure 76 ELISA analysis of affinity-purified antibodies

The cognate SMCfs peptide, four highly array-reactive peptides, and a control peptide (CP2-B) were synthesized on Tenta-gel beads. These were used to prepare individual affinity columns. Purified total IgG of the α-SMCfs serum was applied; bound antibodies were eluted. These peptide-purified antibody samples were analyzed by ELISA against SMCfs-27 coated or KLH coated plates.

Analysis of mimotope-binding antibodies isolated from the total serum demonstrates their specific binding to SMCfs-27 and not to KLH. Antibodies eluted from the negative control CP2-B bead column did not show any reactivity against the SMCfs-27 peptide or KLH. Thus, this non-natural sequence peptide library-screen identified mimotopes of an immunogen, and these mimotopes successfully captured immunogen-specific antibodies.

4.3.5 Measuring cross-reactivity of α-SMCfs antibodies to mimotopes

The antibody samples that were affinity-purified from the SMCfs immune serum with the SMCfs peptide or with the array-selected mimotopes display a range of cross-reactivities for one another (Figure 77). For example, antibodies purified with the RP1, RP2, or RP3 peptides cross-react strongly with each of the other two mimotopes, as measured by ELISA. Antibodies purified with the RP4 bound beads recognize the RP4 peptide, but these antibodies do not cross react well with the other three peptides. One possible explanation for this behavior is that RP1, RP2, and RP3 possess a similar sequence or structural motif whereas RP4 carries a unique one. Notably, all four mimotope-purified antibody samples display the highest reactivity levels against the cognate SMCfs immunogen, rather than the non-natural sequence peptide ligand used to purify them. This indicates that the mimotope affinities for the SMCfs-elicited antibodies are weaker than the SMCfs antibody affinities to the cognate sequence. This may be a consequence of the sparse sampling of random space provided by the library of only 10e4 random sequences; whereas, greater than 1e22 unique 17-mers with 20 amino acids are possible.

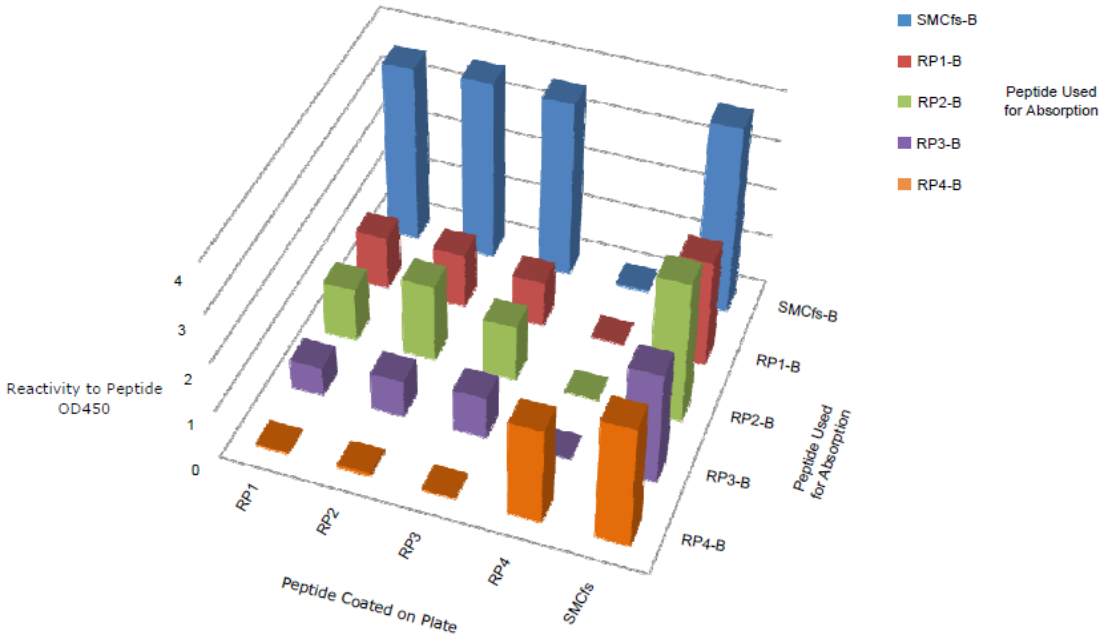


Figure 77 Differential binding of affinity-purified α-SMCfs antibodies to cognate peptide and mimotopes, displayed in a three-dimensional bar graph

The antibodies affinity purified by SMCfs-B and four array-selected non-natural sequence peptide beads (RP1-B through RP4-B) were measured for reactivity against the cognate and mimotope peptides by ELISA.

4.3.6 Sequence Analyses

Bioinformatic analysis was used to identify the epitopes in the non-natural sequence peptides which match with the SMCfs immunogen. Using the epitope search software Bepipred, the strongest B cell epitope within the SMC1A frameshift mutation (CCGIYCHEEPQREDSSI) was predicted to be at the C-terminal end of the 17-mer (SMCfs), with the highest score assigned to the proline within HEEPQRE. The GLAM2 software identifies sequences in the peptide mimotopes that are similar to sequence stretches within SMCfs. The RP1 peptide contains two motifs: HEE and YXXXXPMRQ, although they are in reverse order relative to those sequences of SMCfs. RP2 contains PMREGS and RP3 contains EMPMRE. RP4 contains only the simple dimer HE (Table 16). This analysis shows that three of the four peptides contain a PQRE-like motif found in the cognate SMCfs peptide. Although RP1 contains reversed YXXXXPMRQ and HEE motifs, the mimotope is specifically recognized by α-SMCfs antibodies. This may indicate that i) the HEE motif acts independently of the PQRE-motif, ii) they work in concert but order is not important, iii) or that HEE is not relevant to binding and our analysis failed to identify those a.a. that are relevant in RP4.

Table 16 Motifs in common between RP1-4 and SMCfs as identified by the GLAM2 software

Sequence ID	Sequence	Motif in Peptide Matching with SMCfs	Matching SMCfs sequence
SMCfs	CCGIYCHEEPQREDSSI	-	-
RP1	TISKYVMVEPMRQHEEWGSC	YXXXXPMRQ	YXXXXPQRE
		PMRQ	PQRE
		HEE	HEE
RP2	AVSHQEMNEGEQGPMREGSC	PMREGS	PQREDS
		PMRE	PQRE
RP3	RVGEMPMREYDISGGSGGSC	EMPMRE	EEPQRE
		PMRE	PQRE
RP4	TAFYRTLTKHEVDPGIAGSC	HE	HE

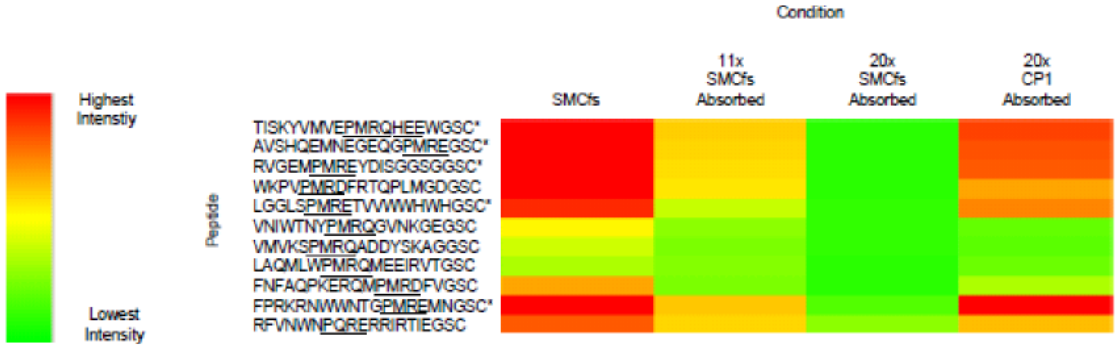
The entire amino acid sequence of each peptide is displayed. The amino acid sequence of only the motif within the whole sequence that matches with the SMCfs sequence as well as the corresponding sequence of the motif in the SMCfs sequence itself is also presented.

The GLAM2-assigned similarity scores of the mimotopes to SMCfs led to a ranking of the peptides in the same order that they were ranked by the array absorption scoring described in section 3.2. For example, the RP1 peptide holds two motifs in common with the original antigen, exhibits the highest intensity on the peptide microarray, and scored the highest by the GLAM2 software. By contrast, the absorbance intensities obtained in the ELISA experiment did not yield rankings identical with the other readouts. This suggests that the array format may enable higher resolution of antibody-peptide interactions.

Broader analysis of the array results shows that other peptides on the peptide array, which were not selected as α-SMCfs antibody specific binders in the original screen, contain the motifs found in the four peptide mimotopes. The observation that not all peptides containing the same motifs exhibit the same binding reactivity empirically suggests that the context in which the motif occurs may be important. Alternatively, the residue positions within the consensus motifs may differentially influence binding strength. Informatic analysis of all 10k random sequences shows: 11 peptides with the PQRE-like motifs (4 peptides with PMRQ, 4 peptides with PMRE, 1 peptide with PQRE, and 2 peptides with PMRD), 920 peptides with HE, 19 peptides with HEE,

and no peptides with a Y preceding the PQRE-like motif with more than one amino acid of separation other than the RP1 peptide. Of these, only 5 of the 11 PQRE-like motif containing peptides and 1 of the 19 HEE motif containing peptides are in the list of 108 screen-selected peptides defined by immune serum-specific reactivity (Figure 73 and Figure 78). However, the rankings of the peptides differ based on these assays that measured specificity by either total serum binding or loss of peptide-absorbed serum binding. This may be a consequence of the significant difference in what the two assay formats will detect as a specific response. Namely, non-SMCfs reactivities to KLH will be included in the analysis of total immune to naïve serum, while only SMCfs-specific reactivities will be measured as specific interactions by the peptide-absorbed serum samples. Thus, this two-step screening approach permitted identification of the most informative mimotopes.

A)



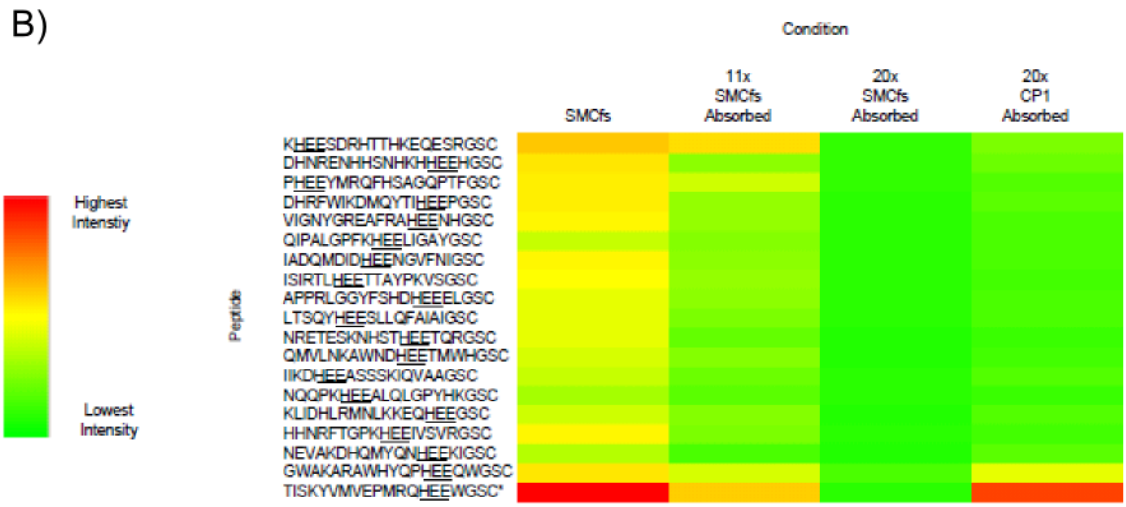


Figure 78 Motif analysis of peptides bound by the SMCfs antibody depleted

A heatmap representation of the changes in peptide-binding intensities of SMCfs-depleted immune sera samples versus the control sera for PQRE-like and HEE motifs. A) Heatmap displaying differential intensities of binding by the α-SMCfs sera samples to the 11 peptides on the peptide microarray with PQRE-like motifs. B) Heatmap of differential intensities of sera sample reactivity to the 19 peptides on the peptide microarray with HEE motifs.

4.3.7 Conclusion

The techniques demonstrated here could be more generally employed for epitope definition and antigen discovery. Non-natural sequence peptide arrays can be used to identify specific antibodies since the antibodies will separate out across the array due to their varying affinities for the different peptides. Non-natural sequence peptides which are unique and specific for a mixture of antibodies found in sera can be used to absorb the desired unique and specific antibodies from the rest of the antibodies in the sera. Once purified, these antibodies can be used to identify the particular epitopes in an antigen to which antibodies are raised. This technique of using non-natural sequence peptides is particularly useful for defining the epitopes in mutant proteins such as the human SMC1Afs protein used as an example in this paper. Non-natural sequence peptides make no assumptions about the epitopes that will be present in a biological sample, and therefore new mutant immunogens which are absent from current

databases can be discovered. Furthermore, the information acquired from probing non-natural sequence peptides with antibodies could be useful in the development of subunit vaccines. In addition to defining epitopes, there is also potential for discovering new antigens since the purified antibodies can be used to probe biological material. Non-natural sequence peptide microarrays hold promise for efficiently and inexpensively dissecting the complexity of vast antibody repertoires.

REFERENCES

1. Stafford P, Halperin R, Legutki JB, Magee DM, Galgiani J, Johnston SA. Physical characterization of the "immunosignaturing effect". Mol Cell Proteomics 2012; 11:M111.011593.

2. Yalow RS, Berson SA. Immunoassay of endogenous plasma insulin in man. J Clin Invest 1960; 39:1157-75.

3. Towbin H, Staehelin T, Gordon J. Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets: procedure and some applications. Proc Natl Acad Sci U S A 1979; 76:4350-4.

4. Vermeulen N, de Beeck KO, Vermeire S, Van Steen K, Michiels G, Ballet V, et al. Identification of a novel autoantigen in inflammatory bowel disease by protein microarray. Inflamm Bowel Dis 2011; 17:1291-300.

5. Stoevesandt O, Taussig MJ, He M. Protein microarrays: high-throughput tools for proteomics. Expert Rev Proteomics 2009; 6:145-57.

6. Ngo Y, Advani R, Valentini D, Gaseitsiwe S, Mahdavifar S, Maeurer M, et al. Identification and testing of control peptides for antigen microarrays. J Immunol Methods 2009; 343:68-78.

7. Reineke U, Ivascu C, Schlief M, Landgraf C, Gericke S, Zahn G, et al. Identification of distinct antibody epitopes and mimotopes from a peptide array of 5520 randomly generated sequences. J Immunol Methods. Netherlands, 2002:37-51.

8. Legutki JB, Johnston SA. Immunosignatures can predict vaccine efficacy. 2013; 110:18614-9.

9. Legutki JB, Magee DM, Stafford P, Johnston SA. A general method for characterization of humoral immunity induced by a vaccine or infection. Vaccine. Netherlands: 2010 Elsevier Ltd, 2010:4529-37.

10. Brown JR, Stafford P, Johnston SA, Dinu V. Statistical methods for analyzing immunosignatures. BMC Bioinformatics 2011; 12:349.

11. Glanville J, Zhai W, Berka J, Telman D, Huerta G, Mehta GR, et al. Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. Proc Natl Acad Sci U S A 2009; 106:20216-21.

12. Dunn-Walters DK, Banerjee M, Mehr R. Effects of age on antibody affinity maturation. Biochem Soc Trans 2003; 31:447-8.

13. Howard WA, Gibson KL, Dunn-Walters DK. Antibody quality in old age. Rejuvenation Res 2006; 9:117-25.

14. LeMaoult J, Szabo P, Weksler ME. Effect of age on humoral immunity, selection of the B-cell repertoire and B-cell development. Immunol Rev 1997; 160:115-26.

15. Furman D, Jojic V, Kidd B, Shen-Orr S, Price J, Jarrell J, et al. Apoptosis and other immune biomarkers predict influenza vaccine responsiveness. Mol Syst Biol 2013; 9:659.

16. Wick G, Grubeck-Loebenstein B. The aging immune system: primary and secondary alterations of immune reactivity in the elderly. Exp Gerontol 1997; 32:401-13.

17. Cevenini E, Caruso C, Candore G, Capri M, Nuzzo D, Duro G, et al. Age-related inflammation: the contribution of different organs, tissues and systems. How to face it for therapeutic approaches. Curr Pharm Des 2010; 16:609-18.

18. Salminen A, Ojala J, Kaarniranta K, Kauppinen A. Mitochondrial dysfunction and oxidative stress activate inflammasomes: impact on the aging process and age-related diseases. Cell Mol Life Sci 2012; 69:2999-3013.

19. Holmes D, Moody P, Dine D. Research Methods for the Biosciences. Oxford University Press, 2011.

20. Agarwal BL. Programmed Statistics (Question-Answers). New Age International, 2007.

21. Black K. Business Statistics: For Contemporary Decision Making, 8th Edition. Wiley Global Education, 2013.

22. Shannon CE. A mathematical theory of communication. ACM SIGMOBILE Mobile Computing and Communications Review 2001; 5:3-55.

23. Clausius R. On a modified form of the second fundamental theorem in the mechanical theory of heat. Philosophical Magazine Series 4 1856; 12:81-98.

24. Boltzmann L. On the Relation of a General Mechanical Theorem to the Second Law of Thermodynamics. The Kinetic Theory Of Gases Series: History of Modern Physical Sciences 2003; 1:362-7.

25. Tribus MEC, McIrvine. Energy and information. Scientific American 1971:224.

26. Brillouin L. Science and Information Theory. Courier Corporation, 2013.

27. Kenner T. Age, time, entropy, and biological optimality--some remarks on the general problem of biological optimization. Wien Med Wochenschr 1996; 146:104-7.

28. van Wieringen WN, van der Vaart AW. Statistical analysis of the cancer cell's molecular entropy using high-throughput data. Bioinformatics 2011; 27:556-63.

29. Riggs JE. Carcinogenesis, genetic instability and genomic entropy: insight derived from malignant brain tumor age specific mortality rate dynamics. J Theor Biol 1994; 170:331-8.

30. Ritchie W, Granjeaud S, Gautheret D. Entropy Measures Quantify Global Splicing Disorders in Cancer. PLOS Computational Biology 2008; 4:e1000011.

31. Castro MA, Onsten TT, de Almeida RM, Moreira JC. Profiling cytogenetic diversity with entropy-based karyotypic analysis. J Theor Biol 2005; 234:487–95.

32. West J, Bianconi G, Severini S, Teschendorff AE. Differential network entropy reveals cancer system hallmarks. Sci Rep 2012; 2:802.

33. de Arruda PF, Gatti M, Facio FN, Jr., de Arruda JG, Moreira RD, Murta LO, Jr., et al. Quantification of fractal dimension and Shannon's entropy in histological diagnosis of prostate cancer. BMC Clin Pathol 2013; 13:6.

34. Lee MY, Yang CS. Entropy-based feature extraction and decision tree induction for breast cancer diagnosis with standardized thermograph images. Comput Methods Programs Biomed 2010; 100:269-82.

35. Nemati S, Edwards BA, Lee J, Pittman-Polletta B, Butler JP, Malhotra A. Respiration and heart rate complexity: effects of age and gender assessed by band-limited transfer entropy. Respir Physiol Neurobiol 2013; 189:27-33.

36. Ohisa N, Ogawa H, Irokawa T, Kurosawa H, Yoshida K. Effect of age on electrocardiogram entropy value of parameters by sleep respiratory disturbance. Rinsho Byori 2010; 58:1073-7.

37. Steinisch M, Torke PR, Haueisen J, Hailer B, Gronemeyer D, Van Leeuwen P, et al. Early detection of coronary artery disease in patients studied with magnetocardiography: an automatic classification system based on signal entropy. Comput Biol Med 2013; 43:144-53.

38. Yao Y, Lu WL, Xu B, Li CB, Lin CP, Waxman D, et al. The increase of the functional entropy of the human brain with age. Sci Rep 2013; 3:2853.

39. Chen Y, Pham TD. Sample entropy and regularity dimension in complexity analysis of cortical surface structure in early Alzheimer's disease and aging. J Neurosci Methods 2013; 215:210-7.

40. Lash A, Rogers CS, Zoller A, Wingfield A. Expectation and entropy in spoken word recognition: effects of age and hearing acuity. Exp Aging Res 2013; 39:235-53.

41. Allen PA, Kaufman M, Smith AF, Propper RE. A molar entropy model of age differences in spatial memory. Psychol Aging 1998; 13:501-18.

42. Dorval AD, Russo GS, Hashimoto T, Xu W, Grill WM, Vitek JL. Deep brain stimulation reduces neuronal entropy in the MPTP-primate model of Parkinson's disease. J Neurophysiol 2008; 100:2807-18.

43. Proctor EA, Kota P, Demarest SJ, Caravella JA, Dokholyan NV. Metric to distinguish closely related domain families using sequence information. J Mol Biol 2013; 425:475-8.

44. Coscia MR, Cocca E, Giacomelli S, Cuccaro F, Oreste U. Immunoglobulin from Antarctic fish species of Rajidae family. Mar Genomics 2012; 5:35-41.

45. Culler S, Hsiao TR, Glassy M, Chau PC. Cluster and information entropy patterns in immunoglobulin complementarity determining regions. Biosystems 2004; 77:195-212.

46. Pantic I, Pantic S, Paunovic J. Aging increases nuclear chromatin entropy of erythroid precursor cells in mice spleen hematopoietic tissue. Microsc Microanal 2012; 18:1054-9.

47. Cowell LG, Kepler TB, Janitz M, Lauster R, Mitchison NA. The distribution of variation in regulatory gene segments, as present in MHC class II promoters. Genome Res 1998; 8:124-34.

48. Liu KPM, Hawkins N, Ritchie J. A. Vertical T cell immunodominance and epitope entropy determine HIV-1 escape. Journal of Clinical Investigation 2014; 123:380-93.

49. Mothe B, Llano A, Ibarrondo J, Daniels M, Miranda C, Zamarreno J, et al. Definition of the viral targets of protective HIV-1-specific T cell responses. J Transl Med 2011; 9:208.

50. De Vlaminck I, Khush KK, Strehl C, Kohli B, Luikart H, Neff NF, et al. Temporal response of the human virome to immunosuppression and antiviral therapy. Cell 2013; 155:1178-87.

51. Pan K, Deem MW. Quantifying selection and diversity in viruses by entropy methods, with application to the haemagglutinin of H3N2 influenza. J R Soc Interface 2011; 8:1644-53.

52. Moffett A, Shackelford N, Sarkar S. Malaria in Africa: vector species' niche models and relative risk maps. PLoS One 2007; 2:e824.

53. Haidar JN, Zhu W, Lypowy J, Pierce BG, Bari A, Persaud K, et al. Backbone Flexibility of CDR3 and Immune Recognition of Antigens. J Mol Biol 2013.

54. Crespillo S, Casares S, Mateo PL, Conejero-Lara F. Thermodynamic Analysis of the Binding of 2F5 (Fab and Immunoglobulin G Forms) to Its gp41 Epitope Reveals a Strong Influence of the Immunoglobulin Fc Region on Affinity. J Biol Chem 2014; 289:594-9.

55. Harris SL, Fernsten P. Thermodynamics and density of binding of a panel of antibodies to high-molecular-weight capsular polysaccharides. Clin Vaccine Immunol 2009; 16:37-42.

56. Pantic I, Pantic S. Germinal center texture entropy as possible indicator of humoral immune response: immunophysiology viewpoint. Mol Imaging Biol 2012; 14:534-40.

57. Avtandilov GG, Barsukov VS. Information analysis of immune and endocrine organs. Morphological changes in the course of infection. Zentralbl Pathol 1992; 138:345-9.

58. Jiuzhou S. Complexity and Entropy Analysis of DNA Methyltransferase. Journal of Data Mining in Genomics & Proteomics 2013.

59. Hannum G, Guinney J, Zhao L, Zhang L, Hughes G, Sadda S, et al. Genome-wide methylation profiles reveal quantitative views of human aging rates. Mol Cell 2013; 49:359-67.

60. Sykes FK, Legutki, B. Joseph, Stafford, Phillip. Immunosignaturing: a critical review. Trends in Biotechnology 2013; 31:45-51.

61. Restrepo L, Stafford P, Johnston S. Feasibility of an early Alzheimer's disease immunosignature diagnostic test. J Neuroimmunol 2013; 254:154–60.

62. Halperin RF, Stafford P, Johnston SA. Exploring antibody recognition of sequence space through random-sequence peptide microarrays. Mol Cell Proteomics 2011; 10:M110.000786.

63. Phillip S. Immunosignaturing Microarrays Distinguish Antibody Profiles of Related Pancreatic Diseases. Journal of Proteomics & Bioinformatics 2013.

64. Kukreja M, Johnston SA, Stafford P. Comparative study of classification algorithms for immunosignaturing data. BMC Bioinformatics 2012; 13:139.

65. Jeschke E, Reinke H, Unverhau S, Pfeifer E, Fienitz B, Bock J. Microsoft® Excel® 2010 Formulas and Functions Inside Out. O'Reily Media, Inc, 2012.

66. Hou D. Studying the evolution of the Eclipse Java editor. Proceedings of the 2007 OOPSLA workshop on eclipse technology eXchange: ACM, 2007:65-9.

67. Fellows I. Deducer: A Data Analysis GUI for R. Journal of Statistical Software 2012; 49:1-15.

68. Team RDC. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 2008.

69. Albright SC, Winston, L. Wayne, Zappe, J. Christopher. Data Analysis and Decision Making with Microsoft Excel Revised,. Cengage Learning, 2008.

70. Inc. SI. JMP 11 Documentation Library. Cary, NC: SAS Institute Inc., 2013.

71. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. ACM SIGKDD Explorations Newsletter 2009; 11:10-8.

72. Keerthi SS, Shevade SK, Bhattacharyya C, Murthy KRK. Improvements to Platt's SMO Algorithm for SVM Classifier Design. Neural Comput 2006; 13:637-49.

73. Halperin R. Characterization and Analysis of a Novel Platform for Profiling the Antibody Response. ARIZONA STATE UNIVERSITY, 2011:290.

74. Boggio K, Nicoletti G, Di Carlo E, Cavallo F, Landuzzi L, Melani C, et al. Interleukin 12-mediated prevention of spontaneous mammary adenocarcinomas in two lines of Her-2/neu transgenic mice. J Exp Med 1998; 188:589-96.

75. Abe F, Dafferner AJ, Donkor M, Westphal SN, Scholar EM, Solheim JC, et al. Myeloid-derived suppressor cells in mammary tumor progression in FVB Neu transgenic mice. Cancer Immunol Immunother 2010; 59:47-62.

76. Restrepo L, Stafford P, Magee DM, Johnston SA. Application of immunosignatures to the assessment of Alzheimer's disease. Annals of Neurology; 70:286-95.

77. Plotkin SA. Immunologic correlates of protection induced by vaccination. Pediatr Infect Dis J 2001; 20:63-75.

78. Gonzalez-Quintela A, Alende R, Gude F, Campos J, Rey J, Meijide LM, et al. Serum levels of immunoglobulins (IgG, IgA, IgM) in a general adult population and their relationship with alcohol consumption, smoking and common metabolic abnormalities. Clin Exp Immunol 2008; 151:42-50.

79. Lopez-Otin C, Blasco MA, Partridge L, Serrano M, Kroemer G. The hallmarks of aging. Cell 2013; 153:1194-217.

80. Brüssow H. What is health? Microbial Biotechnology 2013; 6:341-8.

81. Gosain A, DiPietro LA. Aging and Wound Healing. World Journal of Surgery 2004; 28:321-6.

82. Hayflick L. Entropy explains aging, genetic determinism explains longevity, and undefined terminology explains misunderstanding both. PLoS Genet 2007; 3:e220.

83. Vera E, Bernardes de Jesus B, Foronda M, Flores JM, Blasco MA. The rate of increase of short telomeres predicts longevity in mammals. Cell Rep 2012; 2:732-7.

84. Zhu H, Belcher M, van der Harst P. Healthy aging and disease: role for telomere biology? Clin Sci (Lond) 2011; 120:427-40.

85. Viner RI, Ferrington DA, Williams TD, Bigelow DJ, Schoneich C. Protein modification during biological aging: selective tyrosine nitration of the SERCA2a isoform of the sarcoplasmic reticulum Ca2+-ATPase in skeletal muscle. Biochem J 1999; 340 (Pt 3):657-69.

86. Sharov VS, Dremina ES, Galeva NA, Williams TD, Schoneich C. Quantitative mapping of oxidation-sensitive cysteine residues in SERCA in vivo and in vitro by HPLC-electrospray-tandem MS: selective protein oxidation during biological aging. Biochem J 2006; 394:605-15.

87. Cloos PA, Fledelius C. Collagen fragments in urine derived from bone resorption are highly racemized and isomerized: a biological clock of protein aging with clinical potential. Biochem J 2000; 345:473-80.

88. Haimov I, Laudon M, Zisapel N, Souroujon M, Nof D, Shlitner A, et al. Sleep disorders and melatonin rhythms in elderly people. BMJ 1994; 309:167.

89. Ferrini RL, Barrett-Connor E. Sex hormones and age: a cross-sectional study of testosterone and estradiol and their bioavailable fractions in community-dwelling men. Am J Epidemiol 1998; 147:750-4.

90. Lineweaver CH, Davies PCW, Ruse M. Complexity and the Arrow of Time. Cambridge University Press, 2014.

91. Swan M. Emerging patient-driven health care models: an examination of health social networks, consumer personalized medicine and quantified self-tracking. Int J Environ Res Public Health 2009; 6:492-525.

92. Notkins AL. Polyreactivity of antibody molecules. Trends in Immunology 2004; 25:174-9.

93. Acierno JP, Braden BC, Klinke S, Goldbaum FA, Cauerhff A. Affinity maturation increases the stability and plasticity of the Fv domain of anti-protein antibodies. J Mol Biol 2007; 374:130-46.

94. Jager D, Stockert E, Gure AO, Scanlan MJ, Karbach J, Jager E, et al. Identification of a tissue-specific putative transcription factor in breast tissue by serological screening of a breast cancer library. Cancer Res 2001; 61:2055-61.

95. Zipf KG. National Unity and Disunity: The Nation as a Bio-social Organism. Principia Press Inc., 1941.

96. Barabasi A-L, Bonabeau, Eric. Scale-Free Networks. Scientific American 2003:60-9.

97. Batagelj V, Mrvar, Andrej. Pajek datasets. 2006.

98. Matossian JV. Analyzing the Impact of Local Perturbations of Network Topologies at the Application-level. ProQuest, 2007.

99. Sneppen K, Trusina A, Rosvall M. Hide-and-seek on complex networks. Europhys Lett 2005; 69:853.

100. Strachan DP. Family size, infection and atopy: the first decade of the "hygiene hypothesis". Thorax 2000; 55 Suppl 1:S2-10.

101. Martinez FD. Heterogeneity of the association between lower respiratory illness in infancy and subsequent asthma. Proc Am Thorac Soc 2005; 2:157-61.

102. Martinez FD. Gene-environment interaction in complex diseases: asthma as an illustrative case. Novartis Found Symp 2008; 293:184-92; discussion 92-7.

103. Yazdanbakhsh M, Kremsner PG, van Ree R. Allergy, parasites, and the hygiene hypothesis. Science 2002; 296:490-4.

104. Christensen K, Thinggaard M, McGue M, Rexbye H, Hjelmborg JV, Aviv A, et al. Perceived age as clinically useful biomarker of ageing: cohort study. Bmj 2009; 339:b5262.

105. Yach D, Hawkes C, Gould CL, Hofman KJ. The global burden of chronic diseases: overcoming impediments to prevention and control. JAMA. United States, 2004:2616-22.

106. Shak S. Overview of the trastuzumab (Herceptin) anti-HER2 monoclonal antibody clinical program in HER2-overexpressing metastatic breast cancer. Herceptin Multinational Investigator Study Group. Semin Oncol 1999; 26:71-7.

107. Plosker GL, Figgitt DP. Rituximab: a review of its use in non-Hodgkin's lymphoma and chronic lymphocytic leukaemia. Drugs 2003; 63:803-43.

108. Slamon DJ, Clark GM, Wong SG, Levin WJ, Ullrich A, McGuire WL. Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. Science 1987; 235:177-82.

109. Dent R, Trudeau M, Pritchard KI, Hanna WM, Kahn HK, Sawka CA, et al. Triple-Negative Breast Cancer: Clinical Features and Patterns of Recurrence. 2007.

110. Ben-Hur H, Kossoy G, Sandler B, Zusman I. Vaccination with soluble low-molecular weight tumor-associated proteins suppresses chemically-induced mammary tumorigenesis in rats. In Vivo 2000; 14:551-4.

111. Ward S, Casey D, Labarthe MC, Whelan M, Dalgleish A, Pandha H, et al. Immunotherapeutic potential of whole tumour cells. Cancer Immunol Immunother 2002; 51:351-7.

112. Biragyn A, Tani K, Grimm MC, Weeks S, Kwak LW. Genetic fusion of chemokines to a self tumor antigen induces protective, T-cell dependent antitumor immunity. Nat Biotechnol 1999; 17:253-8.

113. Sykes KF, Lewis MG, Squires B, Johnston SA. Evaluation of SIV library vaccines with genetic cytokines in a macaque challenge. Vaccine 2002; 20:2382–95.

114. Li D, Borovkov A, Vaglenov A, Wang C, Kim T, Gao D, et al. Mouse model of respiratory Chlamydia pneumoniae infection for a genomic screen of subunit vaccine candidates. Vaccine 2006; 24:2917–27.

115. Sykes K. Progress in the development of genetic immunization. Expert Rev of Vaccines 2008; 7:1395-404.

116. Pissani F, Malherbe DC, Schuman JT, Robins H, Park BS, Krebs SJ, et al. Improvement of antibody responses by HIV envelope DNA and protein co-immunization. Vaccine 2013; 32:507-13.

117. Kallioniemi OP, Kallioniemi A, Piper J, Isola J, Waldman FM, Gray JW, et al. Optimizing comparative genomic hybridization for analysis of DNA sequence copy number changes in solid tumors. Genes Chromosomes Cancer 1994; 10:231-43.

118. Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci U S A 1999; 96:6745-50.

119. Casiano CA, Mediavilla-Varela M, Tan EM. Tumor-associated antigen arrays for the serological diagnosis of cancer. Mol Cell Proteomics. United States, 2006:1745-59.

120. Anderson KS, LaBaer J. The sentinel within: exploiting the immune system for cancer biomarkers. J Proteome Res 2005; 4:1123-33.

121. Naito Y, Saito K, Shiiba K, Ohuchi A, Saigenji K, Nagura H, et al. CD8+ T cells infiltrated within cancer cell nests as a prognostic factor in human colorectal cancer. Cancer Res 1998; 58:3491-4.

122. Nakano O, Sato M, Naito Y, Suzuki K, Orikasa S, Aizawa M, et al. Proliferative activity of intratumoral CD8(+) T-lymphocytes as a prognostic factor in human renal cell carcinoma: clinicopathologic demonstration of antitumor immunity. Cancer Res 2001; 61:5132-6.

123. Schumacher K, Haensch W, Roefzaad C, Schlag PM. Prognostic significance of activated CD8(+) T cell infiltrations within esophageal carcinomas. Cancer Res 2001; 61:3932-6.

124. Eerola AK, Soini Y, Paakko P. A high number of tumor-infiltrating lymphocytes are associated with a small tumor size, low tumor stage, and a favorable prognosis in operated small cell lung carcinoma. Clin Cancer Res 2000; 6:1875-81.

125. Sahin U, Tureci O, Schmitt H, Cochlovius B, Johannes T, Schmits R, et al. Human neoplasms elicit multiple specific immune responses in the autologous host. Proc Natl Acad Sci U S A 1995; 92:11810-3.

126. Stempfer R, Syed P, Vierlinger K, Pichler R, Meese E, Leidinger P, et al. Tumour auto-antibody screening: performance of protein microarrays using SEREX derived antigens. BMC Cancer. England, 2010:627.

127. Suzuki A, Iizuka A, Komiyama M, Takikawa M, Kume A, Tai S, et al. Identification of melanoma antigens using a Serological Proteome Approach (SERPA). Cancer Genomics Proteomics. Greece, 2010:17-23.

128. Jongeneel V. Towards a cancer immunome database. Cancer Immun. United States, 2001:3.

129. Uemura M, Nouso K, Kobayashi Y, Tanaka H, Nakamura S, Higashi T, et al. Identification of the antigens predominantly reacted with serum from patients with hepatocellular carcinoma. Cancer 2003; 97:2474-9.

130. Scanlan MJ, Gout I, Gordon CM, Williamson B, Stockert E, Gure AO, et al. Humoral immunity to human breast cancer: antigen definition and quantitative analysis of mRNA expression. Cancer Immun. United States, 2001:4.

131. Jochmus I, Osen W, Altmann A, Buck G, Hofmann B, Schneider A, et al. Specificity of human cytotoxic T lymphocytes induced by a human papillomavirus type 16 E7-derived peptide. J Gen Virol 1997; 78 (Pt 7):1689-95.

132. Ang HC, Joerger AC, Mayer S, Fersht AR. Effects of common cancer mutations on stability and DNA binding of full-length p53 compared with isolated core domains. J Biol Chem. United States, 2006:21934-41.

133. Andreyev HJ, Norman AR, Cunningham D, Oates J, Dix BR, Iacopetta BJ, et al. Kirsten ras mutations in patients with colorectal cancer: the 'RASCAL II' study. Br J Cancer. Scotland: 2001 Cancer Research Campaign., 2001:692-6.

134. Pulaski BA, Ostrand-Rosenberg S. Mouse 4T1 breast tumor model. Curr Protoc Immunol 2001; Chapter 20:Unit 20 2.

135. Reeves JP, Reeves PA, Chin LT. Survival surgery: removal of the spleen or thymus. Curr Protoc Immunol 2001; Chapter 1:Unit 1 10.

136. Schneider CA, Rasband WS, Eliceiri KW. NIH Image to ImageJ: 25 years of image analysis. Nat Methods 2012; 9:671-5.

137. Domenyuk V, Loskutov A, Johnston SA, Diehnelt CW. A technology for developing synbodies with antibacterial activity. PLoS One. United States, 2013:e54162.

138. Lee H. Identification of Neo-antigens for a Cancer Vaccine by Transcriptome Analysis. ARIZONA STATE UNIVERSITY, 2012:168.

139. Storey JD, Stanford University U. A direct approach to false discovery rates. Journal of the Royal Statistical Society: Series B (Statistical Methodology); 64:479-98.

140. Hoey T. UNIT 16.5 Expression and Purification of lacZ and trpE Fusion Proteins. Current Protocols in Molecular Biology: John Wiley & Sons, 1994:16.5.1-.5.6.

141. Scott H. What transcripts are found in a human cell? Genome Biology 2000; 1.

142. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, et al. Landscape of transcription in human cells. Nature 2012; 489:101-8.

143. Karlstetter M, Walczak Y, Weigelt K, Ebert S, Van den Brulle J, Schwer H, et al. The novel activated microglia/macrophage WAP domain protein, AMWAP, acts as a counter-regulator of proinflammatory response. J Immunol. United States, 2010:3379-90.

144. Li Y, Park JS, Deng JH, Bai Y. Cytochrome c oxidase subunit IV is essential for assembly and respiratory function of the enzyme complex. J Bioenerg Biomembr 2006; 38:283-91.

145. Herrmann PC, Gillespie JW, Bichsel VE, Paweletz CP, Calvert VS, Kohn EC, et al. Mitochondrial proteome: Altered cytochrome c oxidase subunit levels in prostate cancer. PROTEOMICS; 3:1801-10.

146. Davies B, Fried M. The structure of the human intron-containing S8 ribosomal protein gene and determination of its chromosomal location at 1p32-p34.1. Genomics. United States, 1993:68-75.

147. Pogue-Geile K, Geiser JR, Shu M, Miller C, Wool IG, Meisler AI, et al. Ribosomal protein genes are overexpressed in colorectal cancer: isolation of a cDNA clone encoding the human S3 ribosomal protein. Mol Cell Biol 1991; 11:3842-9.

148. Ye Q, Worman HJ. Interaction between an integral protein of the nuclear envelope inner membrane and human chromodomain proteins homologous to Drosophila HP1. J Biol Chem 1996; 271:14653-6.

149. Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, et al. Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. Proc Natl Acad Sci U S A. United States, 2004:9309-14.

150. Lovering R, Hanson IM, Borden KL, Martin S, O'Reilly NJ, Evan GI, et al. Identification and preliminary characterization of a protein motif related to the zinc finger. Proc Natl Acad Sci U S A 1993; 90:2112-6.

151. Lee KH, Goan YG, Hsiao M, Lee CH, Jian SH, Lin JT, et al. MicroRNA-373 (miR-373) post-transcriptionally regulates large tumor suppressor, homolog 2 (LATS2) and stimulates proliferation in human esophageal cancer. Exp Cell Res. United States, 2009:2529-38.

152. Hartmann E, Sommer T, Prehn S, Gorlich D, Jentsch S, Rapoport TA. Evolutionary conservation of components of the protein translocation complex. Nature 1994; 367:654-7.

153. Uchida Y, Chiyomaru T, Enokida H, Kawakami K, Tatarano S, Kawahara K, et al. MiR-133a induces apoptosis through direct regulation of GSTP1 in bladder cancer cell lines. Urol Oncol 2013; 31:115-23.

154. Ayer DE, Lawrence QA, Eisenman RN. Mad-Max transcriptional repression is mediated by ternary complex formation with mammalian homologs of yeast repressor Sin3. Cell. United States, 1995:767-76.

155. Arends MJ, McGregor AH, Toft NJ, Brown EJ, Wyllie AH. Susceptibility to apoptosis is differentially regulated by c-myc and mutated Ha-ras oncogenes and is associated with endonuclease availability. Br J Cancer 1993; 68:1127-33.

156. Niklinski J, Furman M. Clinical tumour markers in lung cancer. Eur J Cancer Prev 1995; 4:129-38.

157. Battiste JL, Pestova TV, Hellen CU, Wagner G. The eIF1A solution structure reveals a large RNA-binding surface important for scanning function. Mol Cell. United States, 2000:109-19.

158. Shi H, Wei SH, Leu YW, Rahmatpanah F, Liu JC, Yan PS, et al. Triple analysis of the cancer epigenome: an integrated microarray system for assessing gene expression, DNA methylation, and histone acetylation. Cancer Res 2003; 63:2164-71.

159. Law CJ, Maloney PC, Wang DN. Ins and outs of major facilitator superfamily antiporters. Annu Rev Microbiol 2008; 62:289-305.

160. Cui J, Li F, Wang G, Fang X, Puett JD, Xu Y. Gene-expression signatures can distinguish gastric cancer grades and stages. PLoS One 2011; 6:e17819.

161. Billings PB, Hoch SO, White PJ, Carson DA, Vaughan JH. Antibodies to the Epstein-Barr virus nuclear antigen and to rheumatoid arthritis nuclear antigen identify the same polypeptide. Proc Natl Acad Sci U S A 1983; 80:7104-8.

162. Usuda S, Okamoto H, Iwanari H, Baba K, Tsuda F, Miyakawa Y, et al. Serological detection of hepatitis B virus genotypes by ELISA with monoclonal antibodies to type-specific epitopes in the preS2-region product. J Virol Methods 1999; 80:97–112.

163. Wang LF, Yu M. Epitope Identification and Discovery Using Phage Display Libraries: Applications in Vaccine Development and Diagnostics. Curr Drug Targets 2004; 5:1-15.

164. Schimmele B, Pluckthun A. Identification of a functional epitope of the Nogo receptor by a combinatorial approach using ribosome display. J Mol Biol. England, 2005:229-41.

165. Lam KS, Lake D, Salmon SE, Smith J, Chen ML, Wade S, et al. A One-Bead One-Peptide Combinatorial Library Method for B-Cell Epitope Mapping. Methods, 1996:482-93.

166. Hueber W, Kidd BA, Tomooka BH, Lee BJ, Bruce B, Fries JF, et al. Antigen microarray profiling of autoantibodies in rheumatoid arthritis. Arthritis & Rheumatism; 52:2645-55.

167. Scott JK, Smith GP. Searching for peptide ligands with an epitope library. Science 1990; 249:386-90.

168. Coomber DW. Panning of antibody phage-display libraries. Standard protocols. Methods Mol Biol 2002; 178:133-45.

169. Gao X, Pellois JP, Na Y, Kim Y, Gulari E, Zhou X. High density peptide microarrays. In situ synthesis and applications. Mol Divers 2004; 8:177-87.

170. Folgori A, Tafi R, Meola A, Felici F, Galfre G, Cortese R, et al. A general strategy to identify mimotopes of pathological antigens using only random peptide libraries and human sera. EMBO J 1994; 13:2236-43.

171. Lim H-S, Reddy MM, Xiao X, Wilson J, Wilson R, Connell S, et al. Rapid identification of improved protein ligands using peptoid microarrays. Bioorg Med Chem Lett 2009; 19:3866–9.

172. Stafford P, Halperin R, Legutki JB, Magee DM, Galgiani J, Johnston SA. Physical characterization of the "immunosignaturing effect". Mol Cell Proteomics 2012; 11:M111 011593.

173. Strunnikov AV, Larionov VL, Koshland D. SMC1: an essential yeast gene encoding a putative head-rod-tail protein is required for nuclear division and defines a new ubiquitous protein family. J Cell Biol 1993; 123:1635-48.

174. Larionov VL, Karpova TS, Kouprina NY, Jouravleva GA. A mutant of Saccharomyces cerevisiae with impaired maintenance of centromeric plasmids. Curr Genet 1985; 10:15-20.

175. Yazdi PT, Wang Y, Zhao S, Patel N, Lee EY, Qin J. SMC1 is a downstream effector in the ATM/NBS1 branch of the human S-phase checkpoint. Genes Dev 2002; 16:571-82.

176. Frith MC, Saunders NF, Kobe B, Bailey TL. Discovering sequence motifs with arbitrary insertions and deletions. PLoS Comput Biol 2008; 4:e1000071.

177. Larsen JE, Lund O, Nielsen M. Improved method for predicting linear B-cell epitopes. Immunome Res 2006; 2:2.

178. Harrison DE, Astle CM, Doubleday JW. Cell lines from old immunodeficient donors give normal responses in young recipients. J Immunol 1977; 118:1223-7.

179. Carlson BM, Faulkner JA. Muscle transplantation between young and old rats: age of host determines recovery. Am J Physiol 1989; 256:C1262-6.

180. Conboy IM, Conboy MJ, Wagers AJ, Girma ER, Weissman IL, Rando TA. Rejuvenation of aged progenitor cells by exposure to a young systemic environment. Nature 2005; 433:760-4.

181. Biteau B, Karpac J, Supoyo S, Degennaro M, Lehmann R, Jasper H. Lifespan extension by preserving proliferative homeostasis in Drosophila. PLoS Genet 2010; 6:e1001159.

182. Cortez-Gonzalez X, Zanetti M. Telomerase immunity from bench to bedside: round one. J Transl Med 2007; 5:12.

183. Yui J, Chiu CP, Lansdorp PM. Telomerase activity in candidate stem cells from fetal liver and adult bone marrow. Blood 1998; 91:3255-62.

184. Rao MS, Mattson MP. Stem cells and aging: expanding the possibilities. Mech Ageing Dev 2001; 122:713-34.

185. Chiang CL, Kandalaft LE, Coukos G. Adjuvants for enhancing the immunogenicity of whole tumor cell vaccines. Int Rev Immunol 2011; 30:150-82.

186. Kyte JA. Cancer vaccination with telomerase peptide GV1001. Expert Opin Investig Drugs 2009; 18:687-94.

187. Harrison DE, Strong R, Sharp ZD, Nelson JF, Astle CM, Flurkey K, et al. Rapamycin fed late in life extends lifespan in genetically heterogeneous mice. Nature 2009; 460:392-5.

188. Abeliovich H, Dunn WA, Jr., Kim J, Klionsky DJ. Dissection of autophagosome biogenesis into distinct nucleation and expansion steps. J Cell Biol 2000; 151:1025-34.

189. Oldham S, Hafen E. Insulin/IGF and target of rapamycin signaling: a TOR de force in growth control. Trends Cell Biol 2003; 13:79-85.

190. Zhang X, Goncalves R, Mosser DM. The isolation and characterization of murine macrophages. Curr Protoc Immunol 2008; Chapter 14:Unit 14.1.

191. Wonderlich J, Shearer G, Livingstone A, Brooks A. Induction and measurement of cytotoxic T lymphocyte activity. Curr Protoc Immunol 2006; Chapter 3:Unit 3.11.

192. Danneman PJ, Suckrow MA, Brayton C. The Laboratory Mouse, Second Edition (HANDBOOK OF EXPERIMENTAL ANIMALS). CRC Press, 2012.

193. Deursen MHHaJv. Transgenic Mouse Methods and Protocols. Humana Press, 2011.

194. Kim GG, Donnenberg VS, Donnenberg AD, Gooding W, Whiteside TL. A novel multiparametric flow cytometry-based cytotoxicity assay simultaneously immunophenotypes effector cells: comparisons to a 4 h 51Cr-release assay. J Immunol Methods 2007; 325:51-66.

195. Appendix E: Stem Cell Markers. Stem Cell Information [World Wide Web site]. Bethesda, MD: National Institutes of Health, U.S. Department of Health and Human Services, 2009.

196. Ozato K, Mayer N, Sachs DH. Hybridoma cell lines secreting monoclonal antibodies to mouse H-2 and Ia antigens. J Immunol 1980; 124:533-40.

197. Zhou H, Fisher RJ, Papas TS. Optimization of primer sequences for mouse scFv repertoire display library construction. Nucleic Acids Res 1994; 22:888-9.

198. Winter G, Griffiths AD, Hawkins RE, Hoogenboom HR. Making antibodies by phage display technology. Annu Rev Immunol 1994; 12:433-55.

199. Scott J, Smith G. Searching for peptide ligands with an epitope library. Science 1990; 249:386-90.

200. Pitaksajjakul P, Lekcharoensuk P, Upragarin N, Barbas CF, 3rd, Ibrahim MS, Ikuta K, et al. Fab MAbs specific to HA of influenza virus with H5N1 neutralizing activity selected from immunized chicken phage library. Biochem Biophys Res Commun 2010; 395:496–501.

201. Carter JM. Unit 9.3 Using an affinity column with acidic, basic, or chaotropic elution. Current Protocols in Immunology: John Wiley & Sons, 2003.

202. Burgess R. Advances in gentle immunoaffinity chromatography. Current Opinion in Biotechnology 2002; 13:304–8.

203. Rocha R, Nunes C, Rocha G, Oliveira F, Sanches F, Gobbi H. Rabbit monoclonal antibodies show higher sensitivity than mouse monoclonals for estrogen and progesterone receptor evaluation in breast cancer by immunohistochemistry. Pathol Res Pract 2008; 204:655-62.

204. Nunes CB, Rocha RM, Reis-Filho JS, Lambros MB, Rocha GF, Sanches FS, et al. Comparative analysis of six different antibodies against Her2 including the novel rabbit monoclonal antibody (SP3) and chromogenic in situ hybridisation in breast carcinomas. J Clin Pathol 2008; 61:934-8.

205. Rossi S, Laurino L, Furlanetto A, Chinellato S, Orvieto E, Canal F, et al. Rabbit monoclonal antibodies: a comparative study between a novel category of immunoreagents and the corresponding mouse monoclonal antibodies. Am J Clin Pathol 2005; 124:295-302.

206. Stangegaard M, Dufva IH, Dufva M. Reverse transcription using random pentadecamer primers increases yield and quality of resulting cDNA. Biotechniques 2006; 40:649-57.

APPENDIX A

AGE ASSOCIATED STEM CELL AUTOIMMUNITY

A.1 Introduction

Could part of the aging process be due to an autoimmune reaction against stem cells necessary for repairing damaged tissue? The idea that the aging process is partly caused by autoimmunity against stem cells has not been previously proposed or even suggested to the knowledge of the authors, and therefore this hypothesis requires some justification supported with several background facts. One symptom of the aging process is a decrease in wound healing rates [81]. Some of the cells required to heal wounds and damaged tissue in the body are proliferating stem cells. One very interesting and recent finding is that stem cells from an aged organism can repair damage very well in a younger host, even though these very same stem cells are incapable of repairing damage as well in the aged hosts [178,179,180]. This somewhat counterintuitive result suggests that the problem may lie in the environment of the stem cells rather than in the aged stem cells themselves. A number of explanations could account for this result. Perhaps the aged environment is damaged beyond repair or perhaps epigenetic shifts in gene expression may hinder proper cell to stem cell signaling. In reality, probably both the stem cells and the environment become damaged with age. However, the explanation proposed here is that the immune system has learned to recognize and attack these stem cells, particularly proliferating stem cells. The T cell assay experiments that follow in this paper support this hypothesis that the immune system interferes with stem cells, and this finding may have implications for aging research.

The immune system may attack stem cells because it may fail to distinguish between stem cells and cancerous cells since these cell types have many similarities. The immune system undergoes many changes with the aging process. For example, the immune system responds less effectively to vaccines with age [15], and autoimmunity also increases with age [16] so an autoimmune reaction is not unlikely. Cancer rates also increase with age. Some of this increased cancer risk is due to a change in behavior in stem cells. In fact, researchers have discovered that in aging drosophila, intestinal stem cells over-proliferate and mis-differentiate

which leads to dysplasia [181]. Since cancer cells and stem cells have many similarities, the proposition that the immune system would fail to distinguish between the two is not unreasonable. In fact, more than 85% of all cancers overexpress the telomerase enzyme [182], which is also expressed in stem cells [183], particularly when stem cells are proliferating. Stem cells also become cancerous more often than somatic cells [184] since they already meet one of the requirements for becoming cancerous: telomerase expression or an alternative lengthening of telomeres mechanism.

The immune system plays a large role in dealing with cancer, and would therefore also play a role in interacting with stem cells and therefore the aging process. The immune system is capable of recognizing and combating cancer cells as evidenced by methods such as whole cell tumor vaccines [185]. Vaccination with telomerase, which is found in both cancer cells and stem cells, in mice has demonstrated a decrease in cancer occurrence [186] which demonstrates the immune system is capable of targeting telomerase. Note that the lifespan of the mice was not documented. Additionally, one link between the immune system and aging is that immune suppression drugs such as rapamycin have lengthened the lifespan of mice when maintained in pathogen free environments [187]. This increase in lifespan is usually explained by referring to the interaction of rapamycin with the mTOR pathway which results in autophagy [188], and decreased cellular growth and proliferation [189]. However, the immune suppression itself along with the decreased inflammation associated with this suppression may play a large role in the health benefits observed.

All of these facts combined suggest there is a relationship between aging, cancer, stem cells, and the immune system. In this paper, a chromium release T cell cytotoxicity assay is performed to determine whether aged splenocytes kill young cells from the bone marrow more than young splenocytes kill young cells from the bone marrow. Aged and young splenocytes were also combined with aged cells from the bone marrow. During the experiment some of the aged mice became wounded, and these wounds dramatically affected the results. More wounds correlated with more T cell killing, and this may have been the result of increased inflammation overall.

One can speculate about the implications if the age associated stem cell autoimmunity hypothesis presented here is true. Experiments that address the hypothesis that the immune system incorrectly interferes with normal healthy stem cells in old age may lead to new strategies for mitigating the aging process. While stem cell autoimmunity is certainly not the only cause of aging, many improvements in health might be made by inducing tolerance for healthy stem cells. Although no widespread effective method for inducing tolerance exists, there is no reason to believe there will never be such methods, and developing such methods will be required to effectively treat autoimmune diseases. Additionally, the immune system may continue to operate properly if one were to vaccinate against cancer cells so that the immune system does not later confuse stem cells for cancer cells. Future stem cell treatments may improve the health of certain organs, but such treatments would be impaired if the immune system impairs stem cell function in old age. Understanding the relationship between the immune system and stem cells may result in increased understanding of the mechanisms of the aging process as well as result in methods to aid stem cells to more effectively repair the aging body.

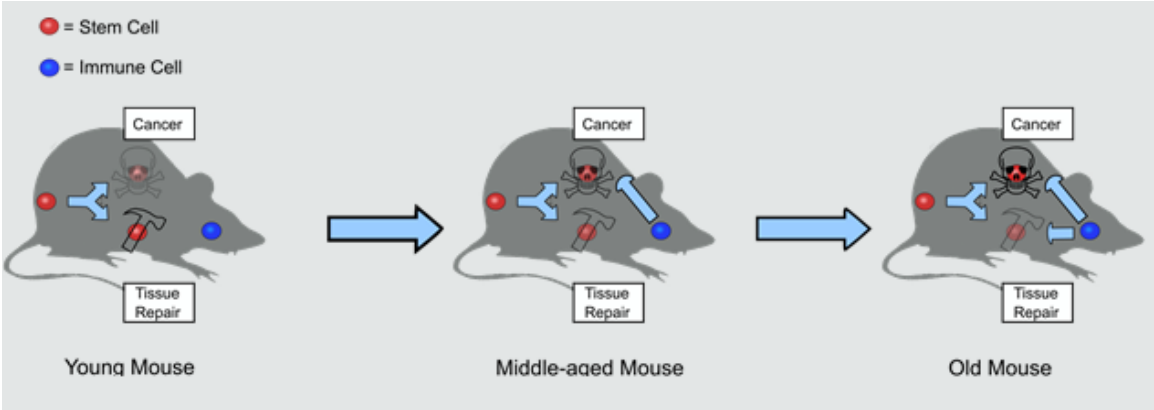


Figure 79 Age Associated Stem Cell Autoimmunity Hypothesis

In a young mouse stem cells repair tissues and few stem cells become cancerous. In a middle aged mouse, fewer stem cells can repair tissues and more stem cells become cancerous as mutations accumulate. At this stage the immune system will start to combat some of the cancers. In an old mouse, increasingly fewer stem cells repair and increasingly more stem cells become

cancerous. At this point, I hypothesize that the immune system fails to discern between the two cell types and combats cancerous stem cells as well as healthy stem cells.

A.2 Materials and Methods

Two experiments were performed to test the reactivity of young and aged splenocytes for young and aged cells derived from the bone marrow. The first experiment used one young and one aged mouse. The second experiment used three young and three aged mice.

A.2.1 One young and one aged mouse

A.2.1.1 Mice

One 3 month 11 day old C57BL/6 mouse was used for this experiment, and this mouse was classified as the "young" mouse. One 1 year 11 month 25 day old C57BL/6 mouse was used, and this mouse was classified as the "aged" mouse.

A.2.1.2 Splenocyte preparation

Splenocytes were obtained from the spleen of mice according to a protocol outlined in Current Protocols in Immunology [135]. Briefly, the mice were anaesthetized with 300 µl of 0.05 mg/mL Avertin injected i.p. A cervical dislocation was then performed. The surgical area was cleaned with ethanol and then an incision was made midway between the last rib and the hip. The spleen was removed and placed in RPMI-10 media on ice. The spleen was homogenized and passed through a 40 µM filter. Splenocytes were collected by centrifugation at 1,200 rpm 5 m 4 °C. 3 mL of Red Blood Cell Lysis Buffer Hybri-Max (Sigma Cat No R7757) solution was added, and the mixture was incubated at RT 5 m. 10 mL complete RPMI-10 media was added to stop the lysis reaction, and the cells were washed twice with 10 mL RPMI-10.

A.2.1.3 Cells from bone marrow

Cells from the bone marrow were obtained in a manner similar to a protocol found in Current Protocols in Immunology [190]. The mice were anaesthetized with 300 µl of 0.05 mg/mL Avertin injected i.p. The hind legs were removed from the mice. Muscle tissue was removed with a

Clorox disinfecting wipe. The thigh bone was then placed into DMEM media. The tip of the bottom and top of the thigh bone was cut, and the marrow was flushed with cold RPMI-10 media with a 3 mL syringe using a 25-G needle. The resulting suspension was filtered through a 40 µM mesh to remove any remaining bone spicules. Hematopoietic progenitor cells were isolated from the bone marrow through a process of negative selection which removed CD5, CD11b, CD19, CD45R, Ly-6G/C, TER119, and 7-4 positive cells using the EasySep Negative Selection Mouse Hematopoietic Progenitor Cell Enrichment Kit (Cat no. 19756) according to the manufacturer's instructions.

A.2.1.4 Chromium release assay protocol

T cell cytotoxicity assays were performed using the chromium release assay method outlined in Current Protocols in Immunology [191]. Cells were first counted using trypan blue. The target cells from the bone marrow were incubated with 20 µl FBS per 100 µl cell solution and 0.1 mL 1 mCi/mL radioactive sodium chromate in a 14 mL round bottom tube for 1 to 3 hr at 37 °C 5% CO2. After the target cells were labeled with chromium, the cells were washed 2 to 3 times with 10 mL complete RPMI-10 media. The cells were then resuspended to a concentration of 1e5 cells/mL. A volume of 100 µl of effector cells from the spleen were dispensed into the wells of 96-well microtiter plates with 4 replicates of each effector cell concentration: 1e7 cells/mL, 3e6 cells/mL, 1e6 cells/mL, and 3e5 cells/mL. This resulted in E:T (effector to target) ratios of 100, 30, 10, and 3. The chromium labeled target cells were then added to wells containing effector cells. The 96 well plate was then centrifuged for 30 s at 200Xg to promote contact between the effector and target cells. Triton X-100 detergent was added to some of the wells as a control to lyse all of the target cells. Some wells also contained target cells alone without any splenocytes. The plate was then incubated 6 hr 37 °C 5% CO_2. The plates were then centrifuged 5 m at 200Xg and 0.1 mL of each supernatant was added into the wells of a LumaPlate (Packard Cat no 4096). The plate was dried at 55 °C overnight in an Affymetrix GeneChip Hybridization Oven 640 (Model no. OVNAA115S) and then sealed with TopSeal-A (PerkinElmer TopSeal-A 605019). The

amount of radioactive chromium in each well was detected using a TopCount HTS (Packard model no. B384V00) instrument to detect cpm (counts per minute) for each well.

A.2.2 Three young mice and three aged mice with zero, one, or two wounds

A second chromium release T cell assay experiment was performed using three young mice and three old mice.

A.2.2.1 Mice

Three 3 month 16 day old C57BL/6 mice were classified as the "young" mice. Three 1 year 11 month 30 day old C57BL/6 mice were classified as the "aged" mice. Two of the aged mice were wounded. One mouse had no wounds (aged mouse no. 1). One had one wound on the side (aged mouse no. 2). Another mouse had a very large wound on the back of the neck with a large wound on the side (aged mouse no. 3). These wounds were not inflicted by the authors. Perhaps one of the mice had wounded the other two since the unwounded mouse was larger than the other two. The weights of the mice and the size of the wounds were not quantitatively measured.

A.2.2.2 Splenocyte preparation

The splenocytes were prepared in the same manner as the first experiment.

A.2.2.3 Cells from bone marrow

The cells were isolated from the bone marrow in the same manner implemented for the first experiment except the hematopoietic progenitor cells were not isolated using the EasySep Negative Selection Mouse Hematopoietic Progenitor Cell Enrichment Kit (Cat no. 19756). Therefore, all of the cells derived from the bone marrow were classified as "stem cells" for this experiment.

A.2.2.4 Chromium release assay protocol

The chromium release assay protocol was performed in the same manner as the first experiment. Stem cells from the six different mice in the two groups (three young mice and three aged mice) were combined with young and aged splenocytes at various E:T ratios.

A.2.2.5 Graphs and calculations

Calculations and t-tests were performed with Microsoft Excel 2007 for Windows (Microsoft, Redmond Washington USA, www.microsoft.com). All t-tests were performed as two-tailed distributions with the two-sample unequal variance type selected. Graphs were created using GraphPad Prism 4 for Windows (GraphPad Software, La Jolla California USA, www.graphpad.com). The corrected percent lysis was calculated using the following formula:

$$\%lysis = 100 * (\frac{test\_^{51}Cr - control\_^{51}Cr}{max\_^{51}Cr - control\_^{51}Cr})$$

In this equation, test_^{51}Cr refers to the amount of chromium released by a particular sample, control_^{51}Cr refers to the amount of chromium released by target cells alone without any splenocytes or Triton X-100 detergent, and max_^{51}Cr refers to the amount of chromium released by target cells with Triton X-100 detergent added to lyse all of the cells.

A.3 Results

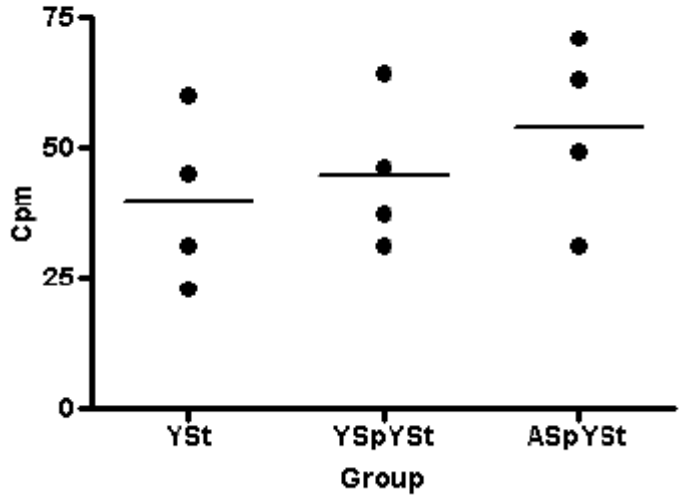
The results can be grouped into two different experiments. The first experiment was performed with one young and one aged mouse, and the second experiment was performed with three young mice and three aged mice. The second experiment produced results for each group of radioactive sodium chromate labeled stem cells from the six different mice used in these experiments. All of the figures in this section have the same format, and the names of groups obey the following convention: Y = young, A = aged, Sp = splenocyte, St = stem cell, and _# indicates an effector to target ratio. The number following Sp or St indicates which mouse the splenocytes or stem cells derived from. For example, the group ASp1YSt2_3 indicates that splenocytes from aged mouse no. 1 were combined with stem cells from young mouse no. 2 at an effector to target ratio of three to one for this group.

The results from these experiments help answer several questions. How do aged and young splenocytes compare in regards to their reactivity for young stem cells? How do aged and young splenocytes compare in regards to their reactivity for aged stem cells? How do aged splenocytes from an unwounded mouse (aged mouse no. 1), a mouse with one wound (aged mouse no. 2), and a mouse with two large wounds (aged mouse no. 3) compare in regards to their reactivity for aged stem cells? How do splenocytes from individuals of the same age compare in regards to their reactivity for aged stem cells? Are aged stem cells killed by young splenocytes and aged splenocytes more than young stem cells are killed?

A.2.3 One young and one aged mouse

The results from the experiment with one aged mouse and one young mouse are presented in Figure 1. The average cpm from the YSpYSt (young splenocytes with young stem cell) group is higher than the average of YSt alone, and the average cpm from the ASpYSt group is higher than YSpYSt (Figure 80A). However, this difference is not statistically significant (p-value of 0.459). In Figure 80B, the young and aged splenocytes are combined with aged stem cells. The average of YSpASt is about 400 cpm higher than the average cpm of ASpASt, but the difference is not statistically significant (p-value of 0.0565). A following experiment with three young mice and three aged mice was then performed.

A)



B)

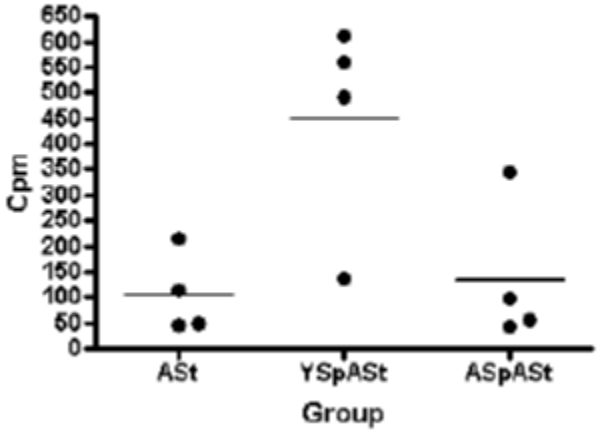


Figure 80 T cell assay results for one young mouse and one aged mouse

A) Young splenocytes and aged splenocytes combined with young stem cells. B) Young splenocytes and aged splenocytes combined with aged stem cells. Cpm refers to counts per minute detected from a scintillation counter in radioactive chromium release assay.

A.2.4 Three young mice and three aged mice with zero, one, or two wounds

A.2.4.1 Young and aged splenocytes with young stem cells

How do aged and young splenocytes compare in regards to their reactivity for young stem cells? Aged and young splenocytes were combined with young stem cells at several different effector to target ratios to answer this question. The results from YSp1, YSp2, ASp1, ASp2, and ASp3 with YSt1 targets at E:T ratios of 3, 10, 30, and 100 are presented in Figure 81A. For any given E:T ratio the average cpm of the aged splenocyte groups two and three is higher than the average of the young splenocyte groups one and two with the exception of the YSp1YSt1_100 group. The same results are presented for YSt2 and YSt3 targets in Figure 81B and Figure 81C respectively. The average cpm of each ASp group is higher than the average cpm of each YSp group within each E:T ratio with the exception of ASp1YSt2_3. These results may indicate that aged splenocytes kill young stem cells more than young splenocytes. Note that results from the combination of every possible group were not obtained due to insufficient cell numbers. These results were obtained with young stem cells and the following results are with aged stem cells.


A)
Cpm
0
100
200
YSt1
YSp1YSt1_3
YSp2YSt1_3
ASp1YSt1_3
ASp2YSt1_3
ASp3YSt1_3
YSp1YSt1_10
YSp2YSt1_10
ASp1YSt1_10
ASp2YSt1_10
ASp3YSt1_10
YSp1YSt1_30
YSp2YSt1_30
ASp1YSt1_30
ASp2YSt1_30
ASp3YSt1_30
YSp1YSt1_100
YSp2YSt1_100
ASp1YSt1_100
ASp2YSt1_100
ASp3YSt1_100
YSt1T

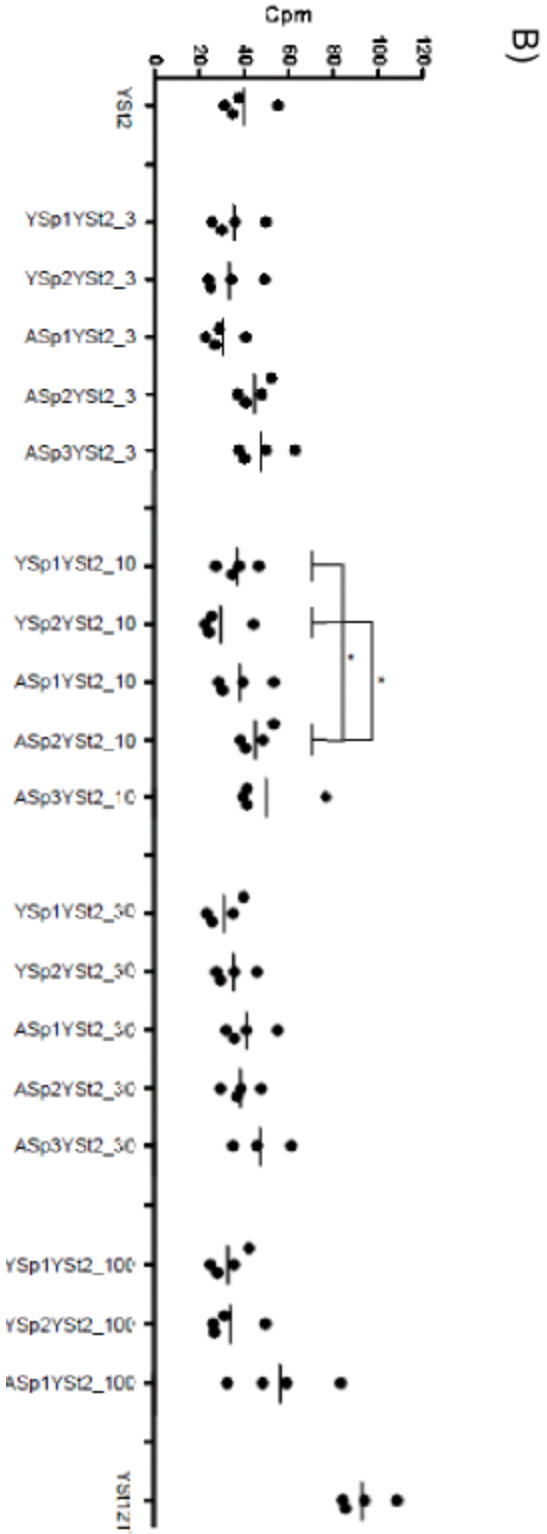

B)
Cpm
0
20
40
60
80
100
120
YSt2
YSp1YSt2_3
YSp2YSt2_3
ASp1YSt2_3
ASp2YSt2_3
ASp3YSt2_3
YSp1YSt2_10
YSp2YSt2_10
ASp1YSt2_10
ASp2YSt2_10
ASp3YSt2_10
YSp1YSt2_30
YSp2YSt2_30
ASp1YSt2_30
ASp2YSt2_30
ASp3YSt2_30
YSp1YSt2_100
YSp2YSt2_100
ASp1YSt2_100
YSt2Z1

Figure 81 Young and aged splenocytes with young stem cells

(A) Cpm counts for YSt1 target cells. (B) Cpm counts for YSt2 targets. (C) Cpm counts for YSt3 targets. P-values from a t-test are indicated as follows: one star (*) for less than 0.05 and greater than 0.01, two star (**) for less than 0.01 and greater than 0.001, and three star (***) for less than 0.001.

A.2.4.2 Young and aged splenocytes with aged stem cells

How do young and aged splenocytes compare in regards to their reactivity for aged stem cells? Young and aged splenocytes were combined with aged stem cells in Figure 82. Specifically YSp1, ASp1, and ASp2 were combined with ASt3. In every comparison of YSp with ASp, the ASp groups have a higher average cpm with a statistically significant p-value with the

exception of the ASp2ASt3_30 group since this group did not have any replicates to allow for a t-test.

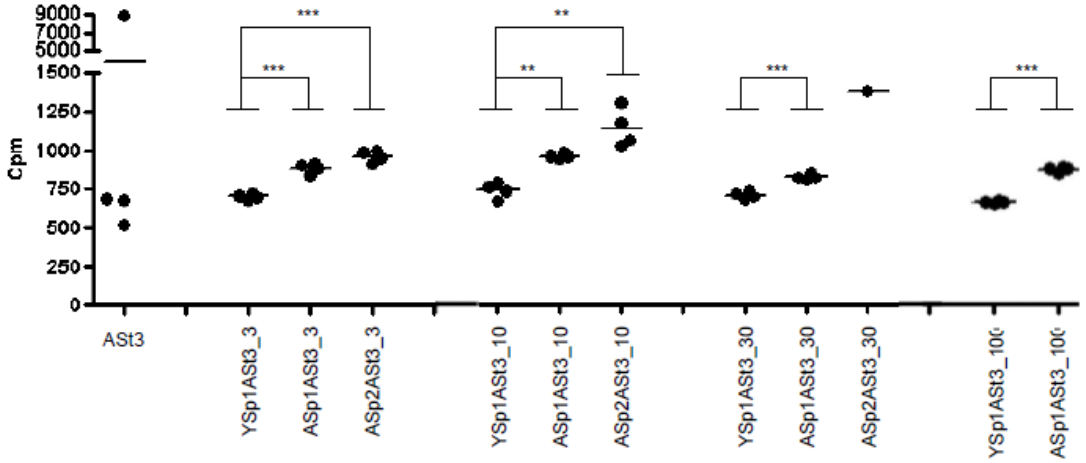


Figure 82 Young and aged splenocytes with aged stem cells

The cpm values from ASt3 are displayed.

A.2.4.3 Aged splenocytes from wounded mice with aged stem cells

How do aged splenocytes from an unwounded mouse (aged mouse no. 1), a mouse with 1 wound (aged mouse no. 2), and a mouse with two large wounds (aged mouse no. 3) compare in regards to their reactivity for aged stem cells? The results obtained from combining ASp1, ASp2, and ASp3 with ASt2 at various E:T ratios are presented in Figure 83. The average cpm of the ASp3 group which contains splenocytes from a mouse with two wounds is higher than the average of the ASp1 group (no wounds) at every E:T ratio.

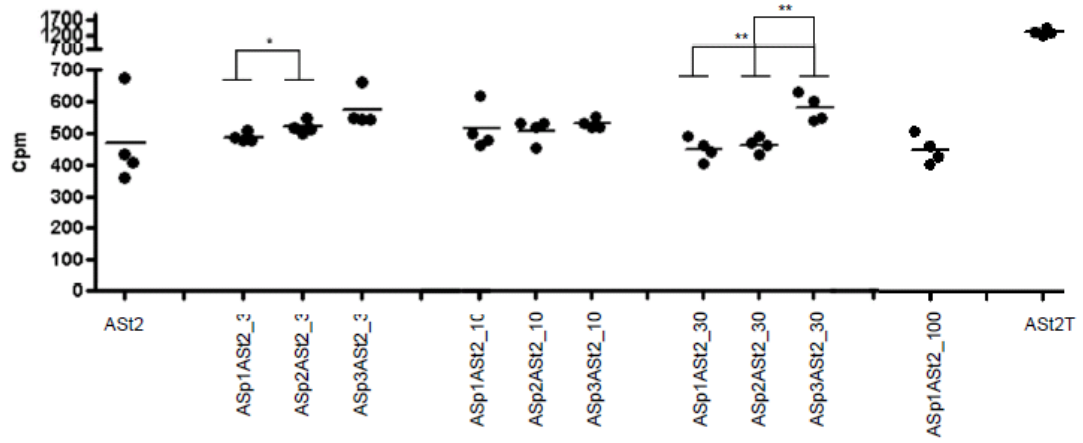


Figure 83 Aged splenocytes from wounded mice with aged stem cells

Aged mouse no. 1 had no wounds, aged mouse no. 2 had one wound, and aged mouse no. 3 had two large wounds. The cpm results for the ASt2 target cells are displayed.

A.2.4.4 Comparison of splenocytes of same age

How do splenocytes from different individuals of the same age compare in regards to their reactivity for aged stem cells? The results from YSp1 and YSp3 combined with ASt1 are presented in Figure 84. There was no statistical difference between YSp1 and YSp3 at any of the E:T ratios.

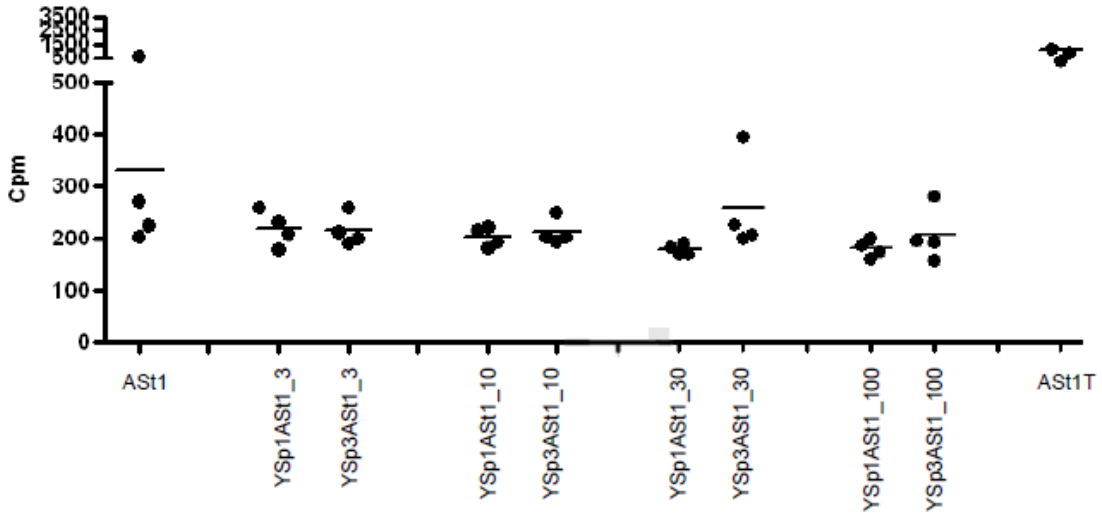


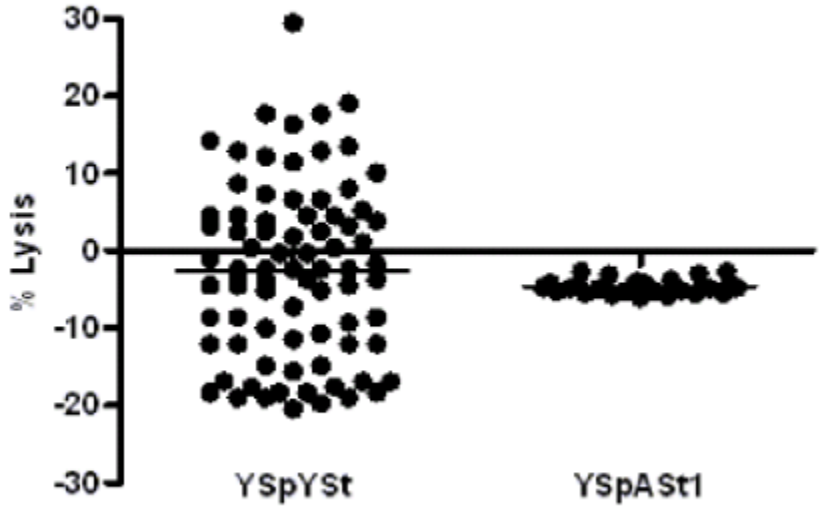
Figure 84 Comparison of splenocytes of same age

The cpm results of ASt1 targets are displayed.

A.2.4.5 Comparison of stem cell age

Are aged stem cells killed more by young splenocytes and aged splenocytes more than young stem cells are killed? One cannot directly compare the cpm values of young stem cells with aged stem cells since the two cell types uptake the radioactive sodium chromate very differently. When the Triton X-100 detergent is used to completely lyse the stem cells to release all of the chromate that entered the cell, much more radioactivity is detected from the aged cells than the young cells in a statistically significant manner. The average YStT cpm was 196 and the average AStT cpm was 2,350 (p-value of 1.10E-4). Therefore, the corrected percent lysis as defined in the Materials and Methods was used to compare YSt groups and ASt groups with YSp (Figure 85A). The corrected percent lysis from the YSt groups and ASt groups with ASp is plotted in Figure 85B. Note that outliers were determined from a box and whisker plot and were excluded from Figure 85A and Figure 85B. All of the raw cpm values for each St group without any splenocytes are presented in Figure 85C.

A)



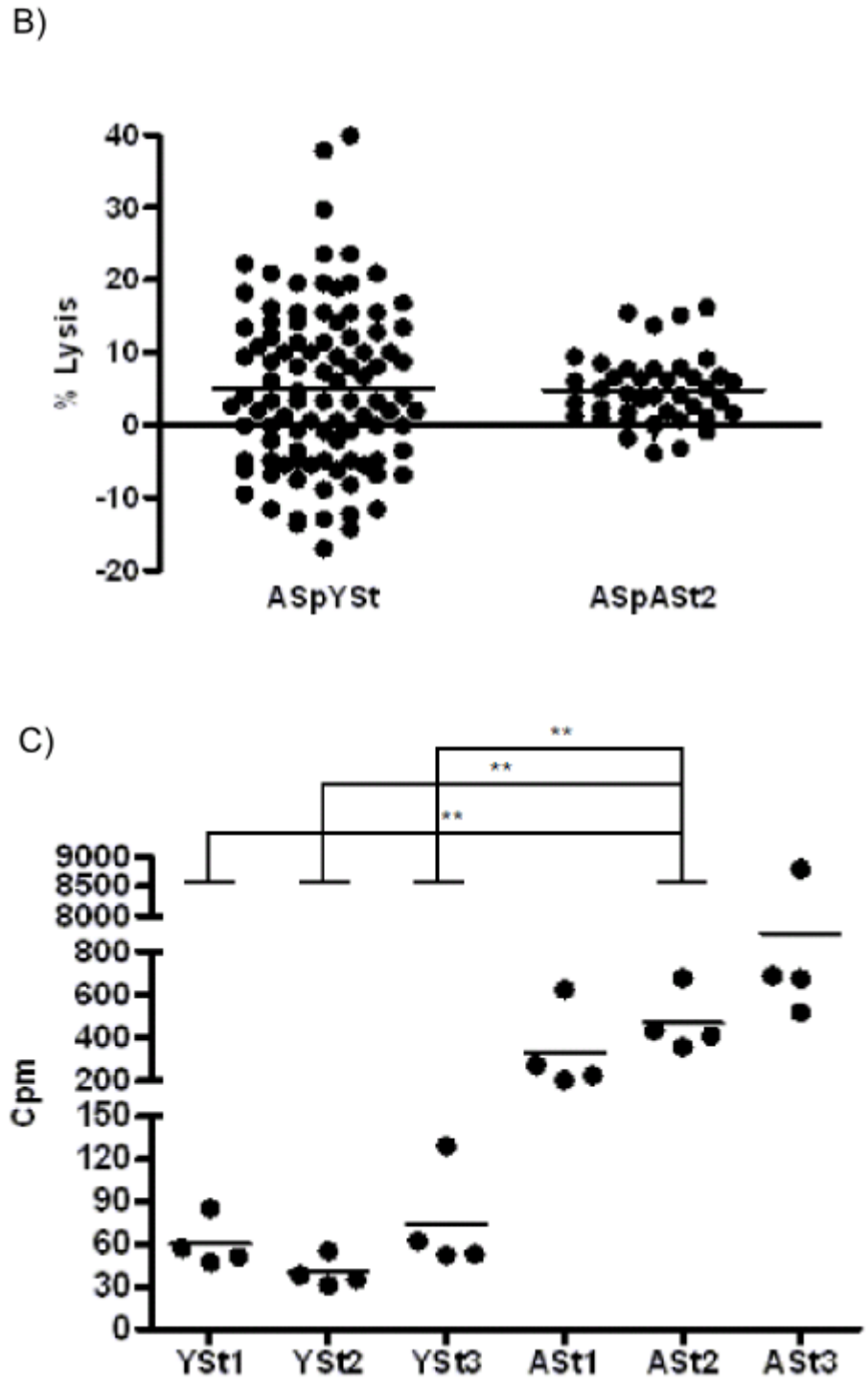


Figure 85 Comparison of young and aged stem cells

A) Corrected percent lysis of YSt and ASt with YSp. B) Corrected percent lysis of YSt and ASt with ASp. C) Raw cpm values of each St group without any Sp.

A.4 Discussion

In general, these experiments support the hypothesis that an aged immune system kills young stem cells more than a young immune system, and unintended discoveries were made as well. Two experiments were performed in which young and aged splenocytes and stem cells were collected from mice and used in T cell killing assays to assess whether the T cells killed the stem cells. Comparisons were then made about whether young or aged splenocytes kill young or aged stem cells more. I found that aged splenocytes do more killing. Serendipitously, I also discovered that splenocytes from wounded aged mice demonstrate an even more pronounced effect of stem cell killing.

The result that aged splenocytes kill young stem cells is not completely intuitive since the argument could be made that no young stem cells would be killed by aged splenocytes or young splenocytes since the immune system should not be attacking self-cells. Note that in these experiments the term “young” refers to mice that were about 3 months old and the term “aged” refers to mice that were about 2 years old. Mice usually finish sexually maturing by 6-8 weeks [192], and they usually die around 2 years of age[193]. Aged splenocytes killed young stem cells more than young splenocytes killed young stem cells in the first experiment presented in Figure 80A as well as in the subsequent experiment in Figure 81A, Figure 81B, and Figure 81C. Although the average cpm of YSp groups vs ASp groups with YSt targets was not always statistically significant, 40 of the 51 (78%) possible YSp vs ASp group comparisons reveal that the ASp groups had a higher average. Additionally, of the three instances in which statistical significance was achieved, the ASp group had the higher average in each case.

The results obtained from aged stem cells are less clear than the results from young stem cells. YSp killed ASt more than ASp killed ASt in the first experiment in Figure 80B, but this result was not statistically significant. In the following experiment, ASp killed ASt more than YSp killed ASt (Figure 82), and statistical significance was achieved in every instance with more than one replicate. The second experiment may have more reliable data since the protocol was better established in the lab at this point, the data points are less sporadic with smaller standard deviations, and the data from the second experiment has statistical significance. Interestingly,

the aged mouse with one wound (aged mouse no. 2) always had a higher average than the aged mouse with no wounds (aged mouse no. 1). Perhaps there was more inflammation, non-specific killing, and an increased immune response in the aged mouse with one wound compared with the unwounded aged mouse.

Wounded mice were not part of the plan of the second experiment, but the results from these mice were quite interesting. The cause of these wounds is still unknown. However, the unwounded aged mouse (aged mouse no. 1) was much larger than the other two aged mice, and I speculate that this mouse wounded the other two: aged mouse no. 2 and aged mouse no. 3. The more wounds a mouse had, the more the splenocytes derived from this mouse killed YSt and ASt as presented in Figure 81A, Figure 81B, Figure 82, and Figure 83. This increased killing may be due to increased inflammation.

Splenocytes from two different mice of the same age did not exhibit any statistical difference in their reactivity for aged stem cells as presented in Figure 84. This result is expected since two mice of the same age, strain, and cage would not be expected to have a different immune response against stem cells.

In addition to the immune system becoming more inflammatory and non-specific with age, the stem cells themselves are liable to accumulate more mutations as well as errors in RNA splicing. This may make these stem cells more prone to become cancer cells as well as to become targets of the immune system. Therefore, the following question can be asked: Are aged stem cells killed more by young splenocytes and aged splenocytes than young stem cells are killed? This may be the case, but this conclusion cannot be reached from the results of this experiment. At first glance, the aged stem cells seem to be killed much more than young stem cells since the cpm values for aged stem cells are much higher. However, the aged stem cells label with sodium chromate much more easily than the young stem cells do as well. The better sodium chromate labeling of aged cells may be due to a more permeable cell membrane. Results from the two different types of stem cells are never placed side by side in any of the figures in this paper until Figure 85. In order to compare the two different types of stem cells in Figure 85A and Figure 85B, the corrected percent lysis was calculated as described in the Materials and

Methods. Figure 85 A and B show that there is no statistical difference in the percent lysis for YSt or ASt with YSp or ASp. A more sensitive assay may detect a difference if there is one.

Many of the comparisons between groups were statistically significant while others were not, and this result may be due to the qualities of the chromium release assay protocol used. Four replicates were included for each group since this is often the standard for chromium release assays. If one of the comparisons which was not statistically significant is selected as an example, how many replicates would need to be performed with the same difference and standard deviation to obtain a significant p-value? For example, with YSp3YSt3_30 and ASp1YSt3_30 in Figure 81C, a few calculations show that 58 replicates instead of 4 would be needed to achieve a p-value less than 0.05 between these two groups which have an average of 58.3 and 65.5 as well as a standard deviation of 14.5 and 18.9 respectively. This many replicates would certainly not be practical and indicates that an alternative more sensitive method for detecting T cell cytotoxicity may be more suitable for investigating the phenomenon in question further. Note that there also did not appear to be a dose response with increased E:T ratios. Perhaps the E:T ratios tested were too similar for the particular phenomenon being tested. The chromium release assay is also a highly variable method. Despite the fact that researchers often perform four technical replicates for every combination of effectors and targets, chromium release interassay cv values are often above 20% [194].

The hypothesis that the immune system interferes with stem cells in old age could be investigated further in many ways. The current results could be repeated and validated using the same chromium release assay method. The experiment could also be performed using a method such as flow cytometry since this method can actually be even more sensitive than the chromium release assay, and more information about the markers on cells and specific types of cells used can be acquired [194]. Valuable information might also be gained by determining exactly which antigens in stem cells that the aged splenocytes might be targeting. Such antigens could be elucidated using assays such as ELISPOT and ELISA. For example, since cancer cells express telomerase, aged splenocytes might target stem cells that express telomerase as well. Other potential antigens to investigate could be stem cell antigen (Sca-1), stem cell factor (SCF), Pax6,

MyoD, as well as many other known stem cell markers [195]. Another very novel method of searching for antigens could be to use young and aged antibody containing sera on a non-natural sequence peptide array and search for patterns that distinguish young and aged sera [1,9]. If the immune system impairs the ability of stem cells to repair damaged tissue, one could test this hypothesis by measuring wound healing rates in SCID mice and normal mice of the same age treated with young stem cells. The SCID mice would be predicted to receive a greater benefit from the treatment. As an alternative to using SCID mice, one could try to implement a bone marrow transplant to replace the immune system of an aged mouse with fresh young cells and compare wound healing rates with non-treated aged mice. Apheresis would be another method that could be performed to remove aged immune cells.

Results from these initial experiments support the hypothesis that aged splenocytes kill stem cells more than young splenocytes. More experiments should be performed to elucidate the exact relationship between splenocytes and stem cells and to verify or disprove the current results. The knowledge gained from the fact that the aged immune system impairs stem cells necessary for repairing damage could lead to treatments to mitigate the aging process. For example training the immune system against true cancer cells or inducing tolerance for stem cell self-proteins may result in an increased healthspan.

APPENDIX B

PHAGE ANTIBODY LIBRARY

B.1 Introduction

One of the main targeting molecules of the immune system is an antibody or immunoglobulin which is a Y-shaped molecule produced by B cells. These antibodies can be used to assess whether a certain immune system targets a cancer protein. An antibody is made up of four chains connected by disulfide bonds with two heavy chains, and two light chains. The antibody is a mirror image of itself so that the two heavy chains are identical and the two light chains are identical. A large portion of the antibody is the constant region which is not involved in binding to the antigen protein target. A small region at the tip of both arms of the "Y" known as the variable region contains the paratope of the antibody which binds to the epitope of the antigen. Both the heavy and light chains have a variable region, but they each have a slightly different composition. The heavy chain variable region is made up of a combination of V, D, and J genes whereas the light chain variable region only consists of a combination of V and J genes. This variable region allows different antibodies to bind to different antigens.

Since antibodies are so useful for targeting molecules for research purposes and therapeutic purposes, researchers have attempted to devise ways to make them quickly and cheaply. One of the strategies employed is to make hybridomas formed by the combination of a B cell of interest and a lymphoma cell [196]. However, this method can be expensive and labor intensive. The resulting molecule is also a large full antibody molecule with the constant region. This constant region is not always desired, and in many applications a small molecule is ideal.

Another strategy developed is to produce single chain variable fragments (scFv). These scFv are small molecules that only consist of a combination of the heavy chain variable region and the light chain variable region. They are created by performing PCR to amplify the heavy and light chain variable regions from the antibody encoding mRNA of B cells [197]. These amplified heavy and light chains are then connected by an artificial flexible linker. This new scFv gene can then be cloned into the genome of a phage which is a virus that infects bacteria. The scFv is then produced as a fusion with the phage coat protein. These phage can replicate indefinitely in bacterial hosts. This system allows for the amplification of antibodies in an abbreviated form by phage [198].

Once a phage antibody library has been produced, it can be used much the same way that antibody containing sera would be used to probe for antigens to which the antibodies bind. One of the first uses of a phage library actually still used antibodies. The phage displayed a library of peptides and the antibodies were screened against this library [199]. Once phage could display scFv, the phage could take the place of the antibody and be "panned" against a molecule of interest. Basically the phage are applied to a surface with an immobilized target, and the unbound phage are washed away. The bound phage are then eluted, and can be used for further rounds of panning [168]. The identity of a particular scFv of interest can also be characterized since the sequence of the scFv is contained within the phage genome.

Although phage libraries have many advantages such as the amplification of antibodies and the production of small molecule versions of antibodies, these libraries also have some disadvantages. For example, although scFv exhibit very similar binding patterns to the antibodies from which they were derived there may still be specific instances in which the binding is not similar enough for the application in question. These scFv are smaller molecules with a slightly different structure that could bind differently in some cases. Phage libraries also require work to produce. This is work that would not be required if the sera were used directly. The diversity of the phage library is also dependent on biological factors such as whether the bacteria and phage are still viable when interacting with particular phage and scFv combinations. Some scFv might display better on the surface of the phage coat than other scFv as well which could introduce some bias into the library. Despite these drawbacks however, researchers have already identified many scFv which bind to specific antigens. The fact that these libraries can identify antigen binders, allow for amplification of the antibody, allow for the sequencing of the antibody, and also utilize small molecules ensures that scFv libraries will continue to be used for some time.

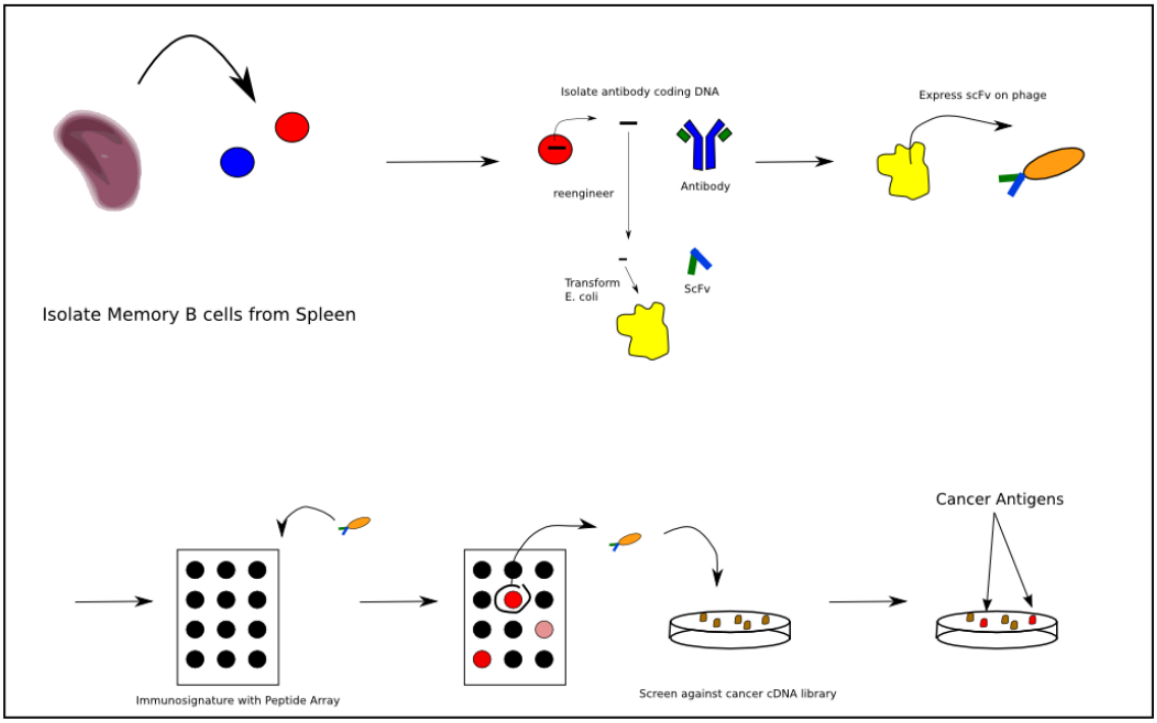


Figure 86 Potential of phage antibody libraries for discovering cancer antigens

Phage antibody libraries could be used to discover cancer antigens. In this figure, memory B cells are isolated from the spleen of a mouse with cancer or a normal mouse as a control, and then antibody coding DNA is isolated from these cells. This DNA is re-engineered to code for scFv, instead of the original large antibody structure, and this scFv DNA is inserted in phage and infected into E. coli. The E. coli are then induced to produce the phage displaying scFv on the surface, and these phage can be used to probe the peptide microarray to determine which peptide features bind significantly different to tumor phage antibody libraries compared to normal phage antibody libraries. The identified peptides could be used to isolate phage for further screening against a cancer cDNA library to identify cancer antigens.

B.2 Materials and Methods

B.2.1 Isolation of RNA from B cells

RNA was isolated from B cells using Trizol from Invitrogen (Cat No 15596-018) according to the instructions from the manufacturer.

B.2.2 PCR Construction of scFv

A total of 73 PCRs were required to construct the scFv fragments. The developed PCR scheme is presented in Figure 87, and the primers designed derived from the primers in a 1994 paper with mouse primers [197]. The sequences of the primers used are listed in Table 17. In the first reaction to amplify the variable domains of the immunoglobulin genes, the heavy chain variable domains and the light chain variable domains were amplified separately. A total of 27 reactions are required to amplify the light chain kappa segments since a PCR reaction has to be performed for every combination of the light chain forward primers (MKV_FOR 1-3) and the light chain back primers (MKV_BACK 1-9). Just one PCR reaction was performed for the lambda light chain with the MLV_FOR and MLV_BACK primers. A total of 40 reactions was performed for every combination of heavy chain forward (MHV_FOR 1-4) and heavy chain back (MHV_BACK 1-10) primers.

Once the light chain and heavy chain variable fragments were amplified, they were linked together. First a linker section (Figure 88) was added to the light chain fragments and heavy chain fragments in a total of 3 PCRs. A linker was added to the light chain kappa fragments with Linker_Light and MKV_FOR_NOTI; a linker was added to the light chain lambda fragments with Linker_Light and MLV_FOR_NOTI; and a linker was added to the heavy chain fragments with Linker_Heavy and MHV_BACK_SFII. This process also added NotI and SfiI restriction enzyme sites to aid in cloning the fragments into a vector at later steps. Next the fragments with the added linkers were linked together in two PCRs: one PCR for the kappa fragments (MKV_FOR_NOTI and MHV_BACK_SFII primers) and one PCR for the lambda fragments (MLV_FOR_NOTI and MHV_BACK_SFII). These scFv fragments were then ready to clone into a phagemid vector with NotI and SfiI sites.

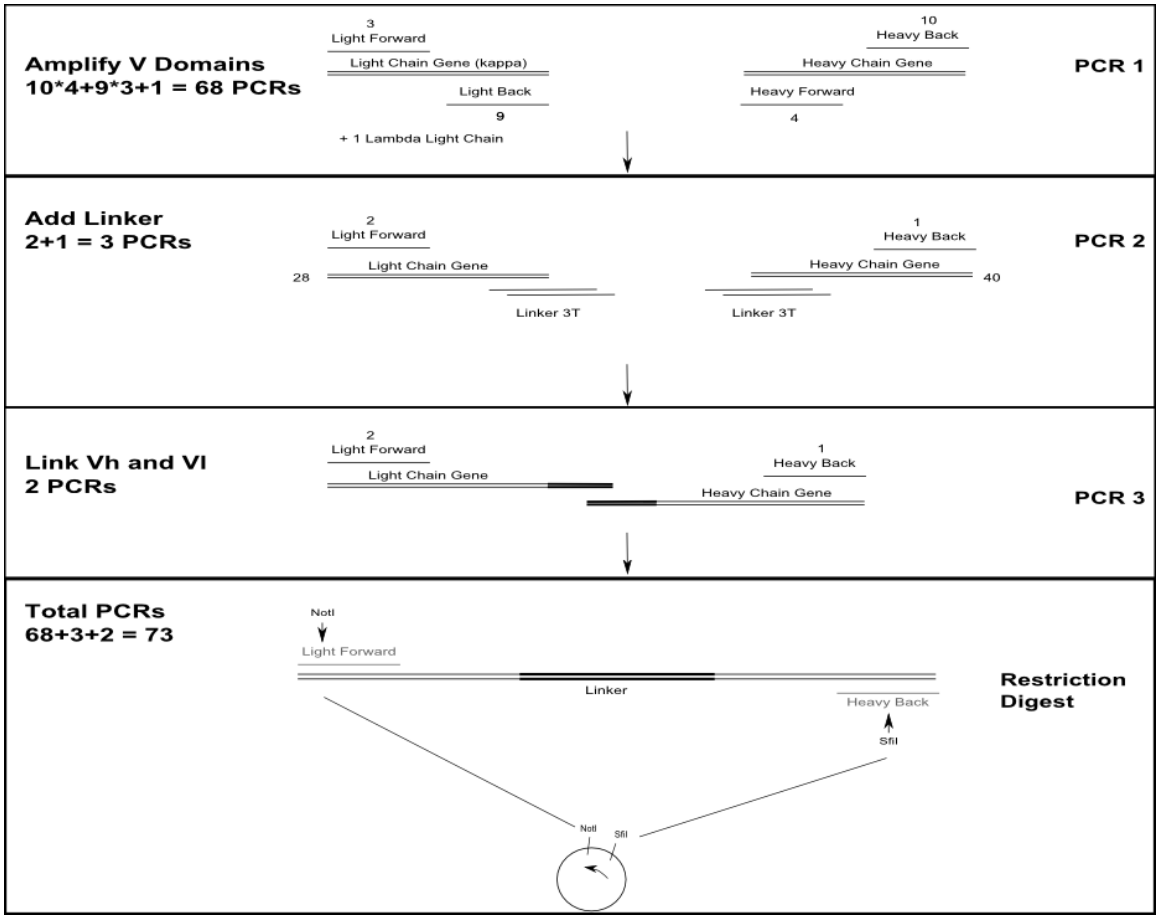


Figure 87 PCR Construction of scFv

The scFv genes can be constructed in several stages of PCR. First the variable domains are amplified by performing PCR with the different variants of light chain and heavy chain variable genes. Next a linker is added to connect the 28 different types of light chains amplified and 40 different types of heavy chain genes amplified. After the linker is added, an additional PCR is performed to amplify the construct for both the lambda and kappa construct types. The scFv can then be cloned into a vector using the SfiI and NotI restriction sites.

Table 17 Primers used for scFv Construction

Primer Type	Primer Name	Primer Sequence
Light Chain Kappa	MKV_FOR1	GTTCTTGCGGCCGCCCGTTTCAGCTCCAGCTTG
	MKV_FOR2	GTTCTTGCGGCCGCCCGTTTTATTTCCAGCTTGGT
	MKV_FOR3	GTTCTTGCGGCCGCCCGTTTTATTTCCAACTTTG
	MKV_BACK1	TCTGGCGGTGGCGGATCGGATGTTTTGATGACCCAAACT
	MKV_BACK2	TCTGGCGGTGGCGGATCGGATATTGTGATGACGCAGGCT
	MKV_BACK3	TCTGGCGGTGGCGGATCGGATATTGTGATAACCCAG
	MKV_BACK4	TCTGGCGGTGGCGGATCGGACATTGTGCTGACCCAATCT
	MKV_BACK5	TCTGGCGGTGGCGGATCGGACATTGTGATGACCCAGTCT
	MKV_BACK6	TCTGGCGGTGGCGGATCGGATATTGTGCTAACTCAGTCT
	MKV_BACK7	TCTGGCGGTGGCGGATCGGATATCCAGATGACACAGACT
	MKV_BACK8	TCTGGCGGTGGCGGATCGGACATCCAGCTGACTCAGTCT
	MKV_BACK9	TCTGGCGGTGGCGGATCGCAAATTGTTCTCACCCAGTCT
Light Chain Lambda	MLV_BACK	TCTGGCGGTGGCGGATCGCAGGCTGTTGTGACTCAGGAA
	MLV_FOR	GTTCTTGCGGCCGCCTTGGGCTGACCTAGGACAGT
Heavy Chain	MHV_BACK1	CATATGGCCCAGCCGGCCATGGCCGATGTGAAGCTTCAGGAGTC
	MHV_BACK2	CATATGGCCCAGCCGGCCATGGCCCAGGTGCAGCTGAAGGAGTC
	MHV_BACK3	CATATGGCCCAGCCGGCCATGGCCCAGGTGCAGCTGAAGCAGTC
	MHV_BACK4	CATATGGCCCAGCCGGCCATGGCCCAGGTTACTCTGAAAGAGTC
	MHV_BACK5	CATATGGCCCAGCCGGCCATGGCCGAGGTCCAGCTGCAACAATCT
	MHV_BACK6	CATATGGCCCAGCCGGCCATGGCCGAGGTCCAGCTGCAGCAGTC
	MHV_BACK7	CATATGGCCCAGCCGGCCATGGCCCAGGTCCAACTGCAGCAGCCT
	MHV_BACK8	CATATGGCCCAGCCGGCCATGGCCGAGGTGAAGCTGGTGGAGTC
	MHV_BACK9	CATATGGCCCAGCCGGCCATGGCCGAGGTGAAGCTGGTGGAATC
	MHV_BACK10	CATATGGCCCAGCCGGCCATGGCCGATGTGAACTTGGAAGTGTC
	MHV_FOR1	GCCTCCACCTGCAGAGACAGTGACCAGAGT
	MHV_FOR2	GCCTCCACCTGAGGAGACTGTGAGAGTGGT
	MHV_FOR3	GCCTCCACCTGAGGAGACGGTGACTGAGGT
	MHV_FOR4	GCCTCCACCTGAGGAGACGGTGACCGTGGT
Primers for Linking	MKV_FOR_NOTI	GTTCTTGCGGCCGCCCGT
	MLV_FOR_NOTI	GTTCTGCGGCCGCCTTGG
	MHV_BACK_SFII	CATATGGCCCAGCCGGCCA
	Linker_Light	TTCAGGCGGAGGTGGCTCTGGCGGTGGCGGATCG
	Linker_Heavy	CAGAGCCACCTCCGCCTGAACCGCCTCCACCTGAGGAGAC

The nucleotide sequences of the primers used in the scheme presented in Figure 87 are presented. There are three light chain kappa forward primers, nine light chain kappa back

primers, one light chain lambda back primer, one light chain lambda forward primer, four heavy chain forward primers, ten heavy chain back primers, a kappa chain primer to add a NotI restriction site, a lambda chain primer to add a NotI site, a heavy chain primer to add a SfiI site, a light chain linker primer, and a heavy chain linker primer.

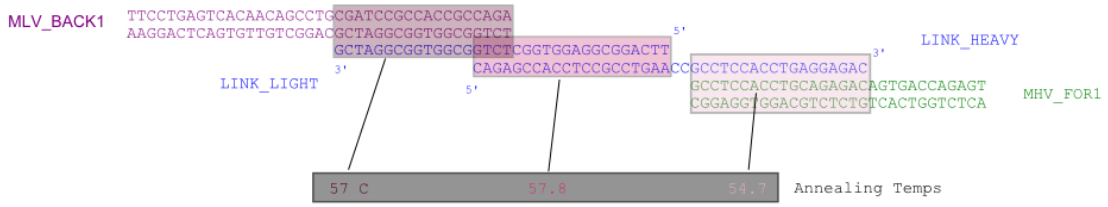


Figure 88 Diagram of linker primers

The light chain and heavy chain genes are linked together using linker primers. In this diagram the MLV_BACK1 light chain sequence and the MHV_FOR1 heavy chain sequence are presented, but the same linker primers would work for any of the light chain or heavy chain combinations. The annealing temperature of the linker primers at each section of the interaction with other DNA strands is presented.

B.2.3 Preparation of vector

Modifications were made to the pComb3XSS vector to prepare to clone in the scFv fragments. The pComb3XSS vector was from the Scripps Research Institute, and this vector has previously been used by other researchers to produce phage libraries [200]. This vector did not originally contain NotI and SfiI sites at the correct location, and therefore these changes had to be made. The NotI site had to be removed, and the SfiI site had to be replaced with a NotI site in order to insert the scFv fragments after the *lacZ* promoter and in front of GeneIII necessary for phage production. The necessary changes are outlined in Figure 89. Plans were made to make these changes by linearizing the vector with primers that would make the appropriate changes, and then performing an In-Fusion reaction (Clontech Cat no 120513).

The plasmid linearization was performed with multiple polymerase enzymes to select the best one: Herculase II, Iproof, PrimeSTAR Max, Pfu Turbo, and Advantage HD. For the

Herculase II polymerase a 0.5% step DMSO gradient was performed so that many reactions were tried with different amounts of DMSO. Two different template amounts with 1 ng and 5 ng were tested with the Iproof, Pfu Turbo, and Advantage HD polymerases. Two different extension times (35 s and 3 min) were also tested for PrimeSTAR Max. The results with PrimeSTAR Max resulted in clear bands so then multiple annealing temperatures were tested with this polymerase from 50 to 68 °C.

Note that after vector preparation, the scFv fragments were never cloned into the vector as planned.

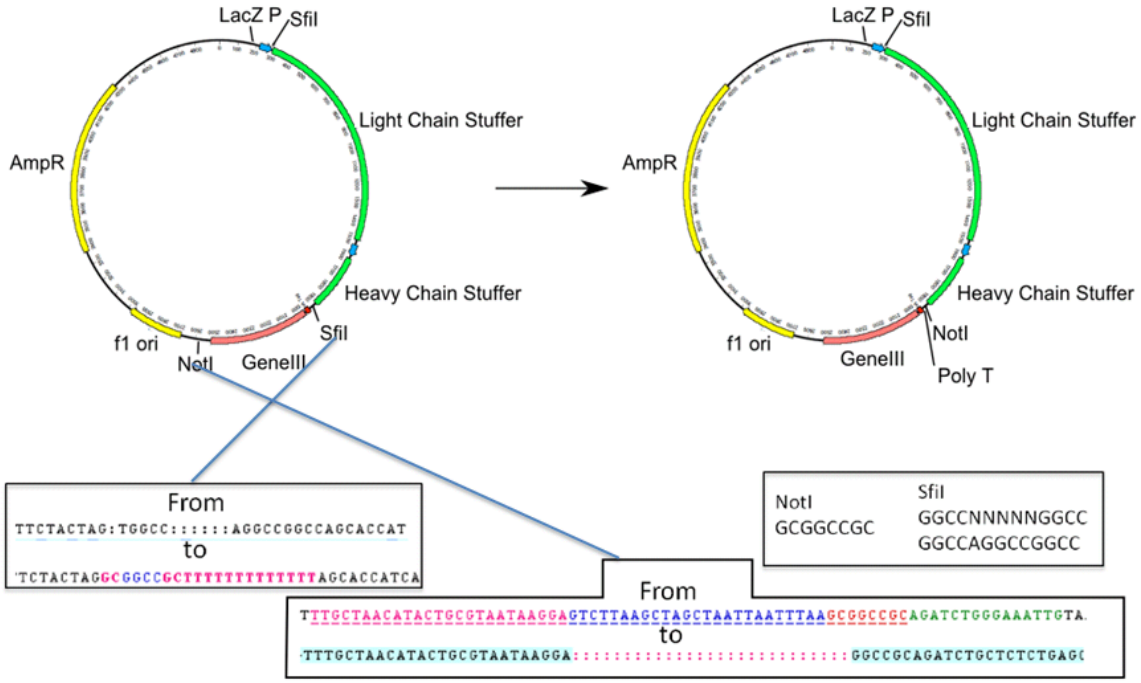


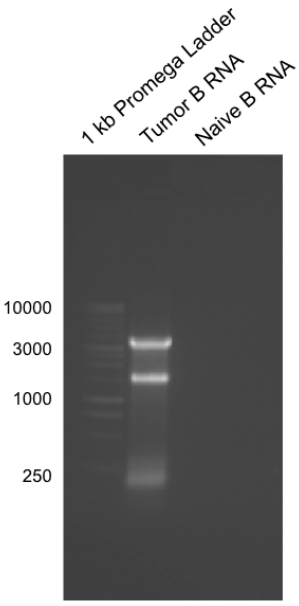
Figure 89 Planned pComb3XSS Modifications

The plan to modify the pComb3XSS primer would remove the NotI site furthest downstream from the lac promoter and change the SfiI site to a NotI site. This would allow scFv flanked by NotI and SfiI sites to be quickly inserted after the lac promoter and before the pIII gene fusion. The exact nucleotide sequence changes to be made are given in boxes with the final sequences given in the box on the right. Note that the SfiI restriction site contains "N" nucleotides so the exact sequence of the particular SfiI site planned is given below these "N" nucleotides.

B.3 Results

B.3.1 Isolation of RNA from B cells

The quantitative results of the isolation of RNA from B cells from tumor or naïve mice are presented in Figure 90. The 260/280 absorbance ratio for the Tumor B cell RNA was 2.08 and the 260/280 ratio of the Naïve B cell RNA was 1.89.



RNA Quality

260/280 around 2.0
DNA 260/280 is around 1.8

Bands:
28S around 5 kb
18S around 2 kb
Low MW RNA (tRNA, 5S) between 0.1-0.3

The 28S (5 kb) rRNA band should be 1.5-2 fold brighter than the 18S (2 kb) band.

Nanodrop Readings

Sample	**Concentration (ng/uL)**	**Abs. 260/280**
Tumor B RNA	695.9	2.08
Naive B RNA	136.4	1.89

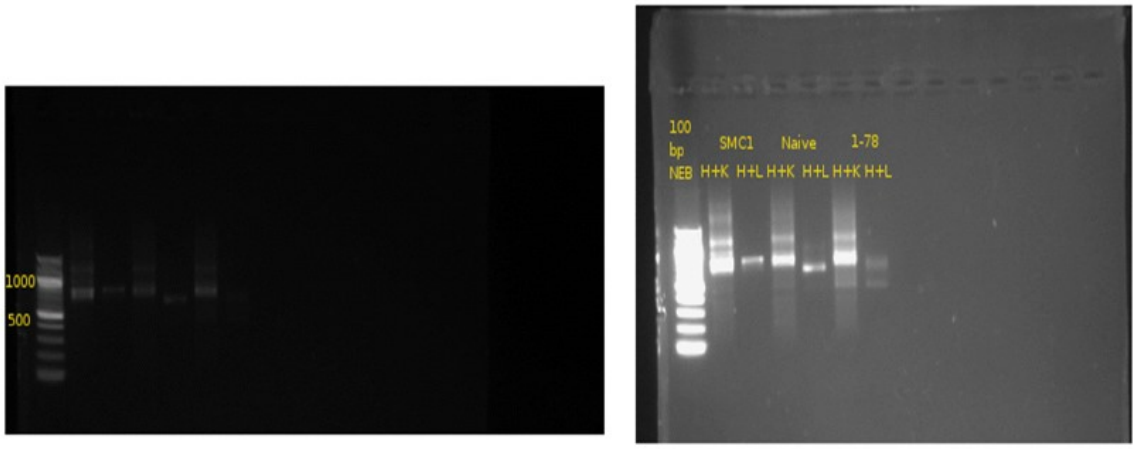
Figure 90 RNA isolated from B cells

RNA isolated from B cells of BALB/c mice. One group of mice was injected with 4T1 cells and had a tumor at the time the RNA was isolated.

B.3.2 PCR Construction of scFv

The final fragments from the scFv PCR construction resulted in bands of the correct size for the heavy chain genes connected to the kappa or lambda light chain genes for all four groups (SMC1, Naïve, 1-78, and Tumor). The expected size of the scFv fragment was approximately 750 bp, and the gel of these fragments is displayed in Figure 91.

A)



B)

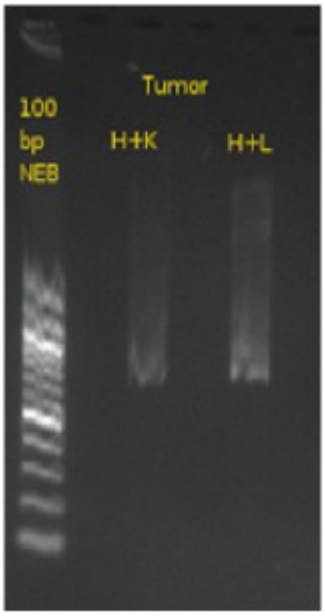
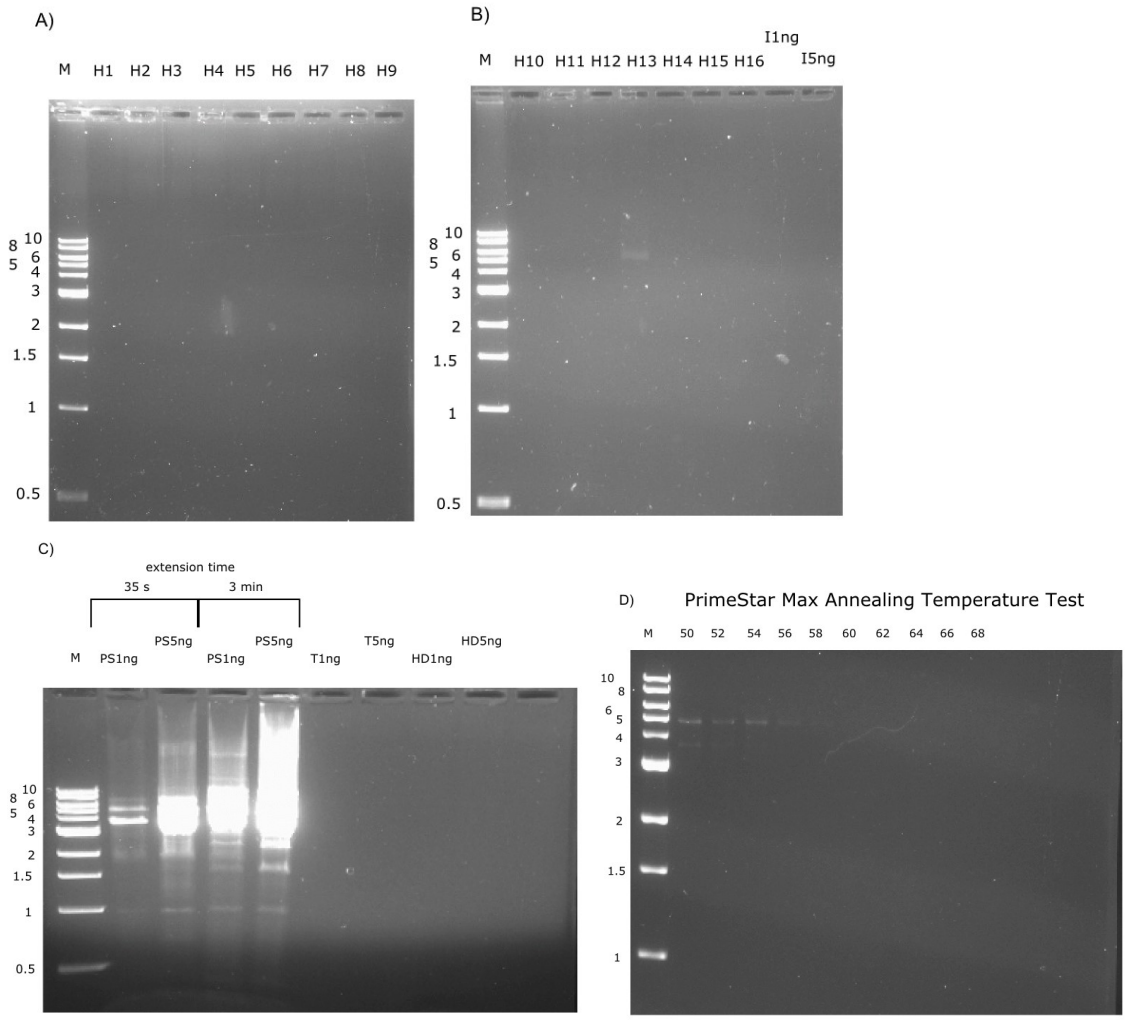


Figure 91 Constructed scFv DNA Fragments

A gel with PCR products from the scFv gene construction is displayed for scFv constructed from naïve mice, mice immunized with SMC1fs, or mice immunized with the 1-78 frameshift transcript in (A) and mice injected with tumor cells in (B). In (A) two different pictures of the same gel with different exposure times to increase or decrease the brightness of the gel are presented. For each mouse type both the "H+L" (heavy chain linked to lambda light chain) and the "H+K" (heavy chain linked to kappa chain) scFv are presented.

B.3.3 Preparation of vector

The results of the pComb3XSS linearization with a variety of different polymerases and conditions are presented in Figure 92. The PrimeSTAR Max polymerase produced the strongest bands, and another test with this polymerase was performed at different annealing temperatures in Figure 92 part D.



H= Herculase II
H1-16 correspond to 0.5% step DMSO gradient
I = Iproof
PS = PrimeSTAR Max
T = Pfu Turbo
HD = Advantage HD

Figure 92 Plasmid Linearization of pComb3XSS

In A) and B) a DMSO gradient with the Herculase II polymerase was performed. In B) the Iproof polymerase was also tested with 1 ng and 5 ng of starting template material. In C) the

PrimeSTAR Max polymerase, the Pfu Turbo polymerase, and the Advantage HD polymerase were tested. In D) the PrimeSTAR Max polymerase was tested with annealing temperatures ranging from 50-68 °C.

B.4 Discussion

Phage antibody libraries provide a tool for amplifying an antibody repertoire, and an attempt was made here to produce a phage antibody library to probe for cancer antigens. First quality RNA was isolated from B cells from the spleens of mice with tumor or no tumor. Next this RNA was used with PCR to construct scFv fragments. An efficient method was developed with Dr. Andrey Loskutov and Dr. Kathryn Sykes to produce the scFv needed to cover the mouse immunoglobulin repertoire in just 73 total PCRs, which is less than the number of PCRs that would be required by many other methods for a fully representative library. Next, initial steps were made to clone the scFv fragments into a vector. We originally tried to clone the fragments into the PCANTAB5E vector, but we later discovered that the PCANTAB5E plasmid in our lab's possession was actually a different plasmid. We then decided to use the pComb3XSS vector. First, however, the pComb3XSS vector had to be modified to contain the appropriate restriction enzymes which required linearization of the plasmid. Several conditions were tested to optimize the PCR of such a long DNA fragment, and the PrimeSTAR Max polymerase proved to be the most suitable polymerase for the challenge. Ultimately, the fragments were never cloned into the vector since other projects took priority. In theory though, the scFv fragments produced from 73 PCR reactions could be cloned into a modified pComb3XSS vector to produce a phage antibody library with an abundant amount of material (as opposed to pure sera) for discovering tumor antigens for a tumor vaccine.

APPENDIX C

PURIFICATION OF ANTIBODIES WITH NON-NATURAL SEQUENCE PEPTIDES

C.1 Introduction

A researcher may want to purify antibodies using the non-natural sequence peptide microarray before screening a cDNA library. Using purified antibodies, it might be possible to obtain much stronger signals, and fewer false positives. The purification of antibodies should be possible since researchers in our lab have discovered that there are some non-natural sequence peptides which specifically bind to tumor sera compared to naïve sera. Therefore, these non-natural sequence peptides can be used to bind and elute just the antibodies of interest.

C.2 Materials/Methods and Results

An attempt was made to purify specific antibodies using selected non-natural sequence peptides. The basic process followed derives from a protocol outlined in Current Protocols in Immunology 9.3 Basic Protocol 3 Using an Affinity Column with Acidic, Basic, or Chaotropic Elution [201]. Six peptide sequences which bound more to tumor sera than naïve sera were synthesized onto Tentagel beads (RAPP-Polymere). A 100 µl aliquot of a 50% DMF slurry of these beads was washed in the organic solvent DMF in a 5 mL polypropylene column (Thermo Cat no 29922), and were gradually transitioned to a wash with approximately 5 mL PBS aqueous solution. The sera was then diluted 1 to 10 in equilibration buffer (TBST) and washed through the column. The unbound sera fraction was collected. The column was then sealed and incubated on a rotisserie shaker for 1 to 2 hr at 37 °C. The column was then washed with approximately 20 mL equilibration buffer. Several Protein LoBind tubes (Eppendorf Cat no 0030 108.116) for collecting the elution fractions were prepared by adding 0.2 mL neutralization buffer (1.5 M Tris · Cl 150 mM NaCl pH 8) to each tube. Elution was performed by adding 5 mL acidic elution buffer (0.1 M glycine 0.15 M NaCl pH 2.8) and collecting 1 mL fractions in the tubes containing neutralization buffer. A Nanodrop instrument (Thermo Scientific) was used to determine which fractions contained

antibody and these fractions were pooled together. Dialysis was performed with these solutions using Slide-A-Lyzer (Thermo Scientific) according to the manufacturer's instructions.

After the antibodies were purified, these antibodies were used in a dot blot to test for reactivity to the selected peptides. Peptides were spotted (2 µl of 30 ng/µl peptide) onto a nitrocellulose membrane. The membrane was then dried for 20 min and blocked in 5% BSA in TBST overnight. The selected sample (original sera, elution fraction, or unbound flow-through) was then diluted into 20 mL 5% BSA TBST and used to probe the membrane 1hr RT. The membrane was washed and the 2 mg/mL goat anti-mouse IgG H+L AF647 secondary at a 1:10,000 dilution in TBST 5% BSA was applied to the membrane 1 hr RT. The membrane was washed, dried, and then scanned with a Typhoon Trio+ instrument (Amersham Biosciences). Specific binding for the selected peptides in the eluted fraction could not be detected.

The selected sample (original sera, elution fraction, or unbound flow-through) was then applied to the non-natural sequence peptide microarray using the standard protocols for the microarrays in our lab (1:500 dilution of primary and 5 nM secondary). The expectation was that the purified and eluted antibodies would bind to the same peptides used to purify them. However, the purification did not work. The flow through which did not bind to the peptides could still bind to the 6 peptides on the array, but the purified fraction did not bind to these 6 peptides Figure 93. Therefore, the method of purification had to be redesigned. The major issue may have been that the peptides on the array are at a much higher density than the peptides on the Tentagel mesh. The orientation of the peptides may also have been an issue.

Peptides were synthesized onto the Tentagel beads at 4X the density, but similar results as before were obtained.

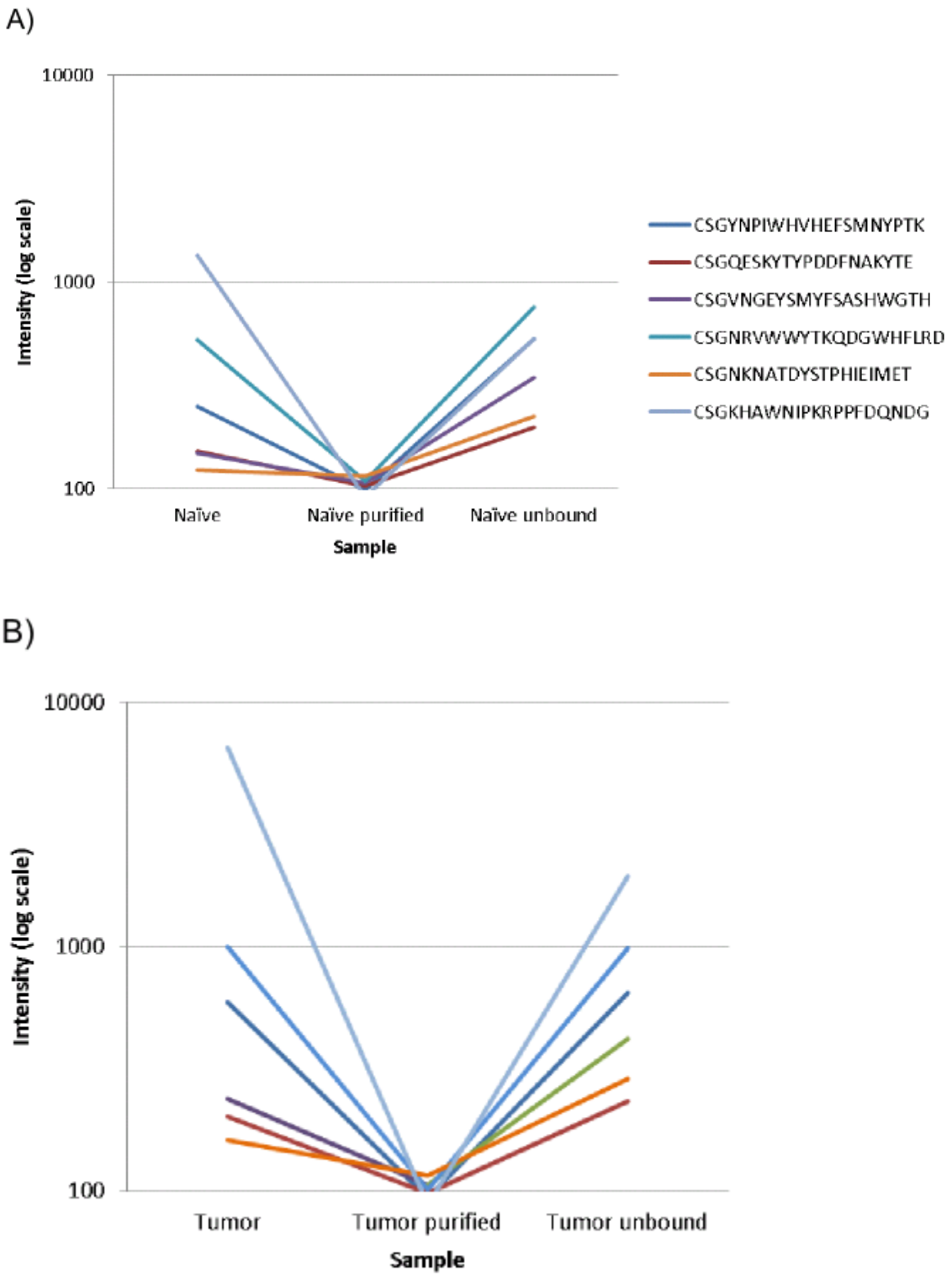


Figure 93 Purification of selected antibodies with peptides

A column containing TentaGel beads with six different peptides was used to purify antibodies from A) naïve sera and B) tumor sera. Unpurified sera, sera eluted from the column, and sera in the unbound fraction was applied to the non-natural sequence peptide microarray and the intensity of each of the six peptides used for the purification is presented on the y-axis.

C.3 Discussion

A method for purifying antibodies using selected peptides from a non-natural sequence peptide microarray would be very useful. The purified antibodies could then be used to probe for antigens in a cDNA library screen, western blot, or pull-down assay. Theoretically, one could apply sera from a sick patient onto a non-natural sequence peptide microarray, identify unique peptides for the patient, use these peptides to isolate the relevant antibodies, and then use these antibodies to probe for important antigens which could be used in a vaccine or to study the disease. An attempt was made here to purify antibodies using peptides synthesized onto Tentagel beads. The results demonstrated that reactivity against the selected peptides could be detected, but only with the unbound fraction and not the eluted fraction which should have contained antibodies which bound to the peptides. Other approaches will need to be pursued in order to identify a method which can successfully purify antibodies using selected peptides. Perhaps a different equilibration buffer (such as ethylene glycol [202]) or peptide synthesis substrate instead of Tentagel beads could work more effectively. Note that non-natural sequence peptides were successfully used to purify antibodies in "4: RANDOM SEQUENCE MIMOTOPES" in this dissertation. However, in that chapter, sera from a hyper-immunized rabbit rather than sera from a mouse with a tumor was used. There may not have been enough antibody to a specific target in the mouse with a tumor. Also note that antibodies from rabbits typically exhibit superior antigen recognition when compared to mouse antibodies [203-205]. Once an effective protocol is developed, this process could be an invaluable tool for quickly identifying important antigens and infectious agents.

APPENDIX D

AUTOMATIC ARRAY ALIGNMENT

D.1 Introduction

In the Center for Innovations in Medicine many students, technicians, graduate students, post-docs, and professors have spent a considerable amount of time aligning lighted features in an image with assigned features, which is a necessary but tedious task. On some of the arrays there are 10,000 features, and on some of the arrays there are 330,000 features. This alignment process is performed using the GenePix 8.0 software (Molecular Devices, Santa Clara, CA). While this software does have some rudimentary automatic aligning features, there are many situations in which this automatic alignment does not perform well and the user still has to manually look through the alignment and make corrections. Additionally, the user still has to open up a file, roughly align all the grids before performing any automatic alignment, obtain the results, and then save the file. Ideally, it would be better to have software that could just automatically go through an entire folder of files and output the results, which would be a process the program could perform all night and day without fatigue. I was ambitious and thought that I could create such a program. I did create a program which successfully reached some benchmarks, but ultimately a program which could automatically align an entire array was never completed.

D.2 Materials and Methods

A Java program was created to perform the automatic alignment of one block on the array. The program takes a red reflect (pre-experiment) 16 bit TIFF image file, a gal file, and a post-experiment 16 bit TIFF image file containing the spots for one block of the array as input and outputs the x-y coordinates of each feature thus performing an alignment. Note that a gal file contains the general pattern of features along with their identity.

The constructed Java program proceeds through several steps. First the intensity values are extracted from the red reflect image. Note that in the red reflect image, every spot has an intensity and this is scanned before any sera is applied to the array. In order to extract intensity values from the 16 bit TIFF image the Java Advanced Imaging (JAI) API was used. Specifically the jai_codec.jar, jai_core.jar,

and mlibwrapper_jai.jar files were included in the project. A threshold was applied to retain pixel intensity values above a certain value, and then the density-based spatial clustering of applications with noise (DBSCAN) algorithm was used to find the coordinates of clusters of high intensity pixels (spots). Once the x,y coordinates of all of the points are identified the program then finds the vertical and horizontal edges formed by rows and columns of spots. These edges are found by identifying ranges in which many points lie within approximately the same x or y value. From the edges, the corner points are also identified. Next the program roughly aligns a gal file which contains the feature names and the basic pattern of the spots with the grid identified in the red reflect. The gal file is rotated and resized appropriately. Next the position of each feature in the roughly aligned gal file points are fine-tuned by moving them to an actual image spot in the red reflect if there is an image spot very nearby.

Once the features have been identified in the red reflect image, this arrangement of features is then matched up with the actual image after the experiment has been performed and sera has been applied to the array. Note that the alignment was performed with the red reflect image first because this is a very clean image in which just about each spot is present, whereas in the image after the experiment there are many false spots and real spots that are missing from the image. Matching with the post-experiment image is performed by first identifying the landing lights, which are like the regular spots, but they have a very high intensity. The pattern of points established with the red reflect image is then rotated and resized to match the landing lights. These points are then fine-tuned to move any point to a new post-experiment image spot. In other words any features are moved to very close spots in the post-experiment image. The last step was to simply extract the median intensity of each spot. The code to perform these functions is stored at "\\biofs.biodesign.asu.edu\CIM\Administration\Biostatistics\Kurt\Automatic Slide Alignment" as of this writing.

D.3 Results

A diagram illustrating the process the program takes as it automatically aligns a slide is presented in Figure 94. Note that the image at each stage was not a constructed diagram, but was actually an image output by the program itself. The end result was a program that has a 0.94 correlation coefficient

with a human alignment, and that is correctly aligned with the bright spots of the image upon manual visual inspection. The program could only work for one block, however, and was never completed enough to automatically align an entire array.

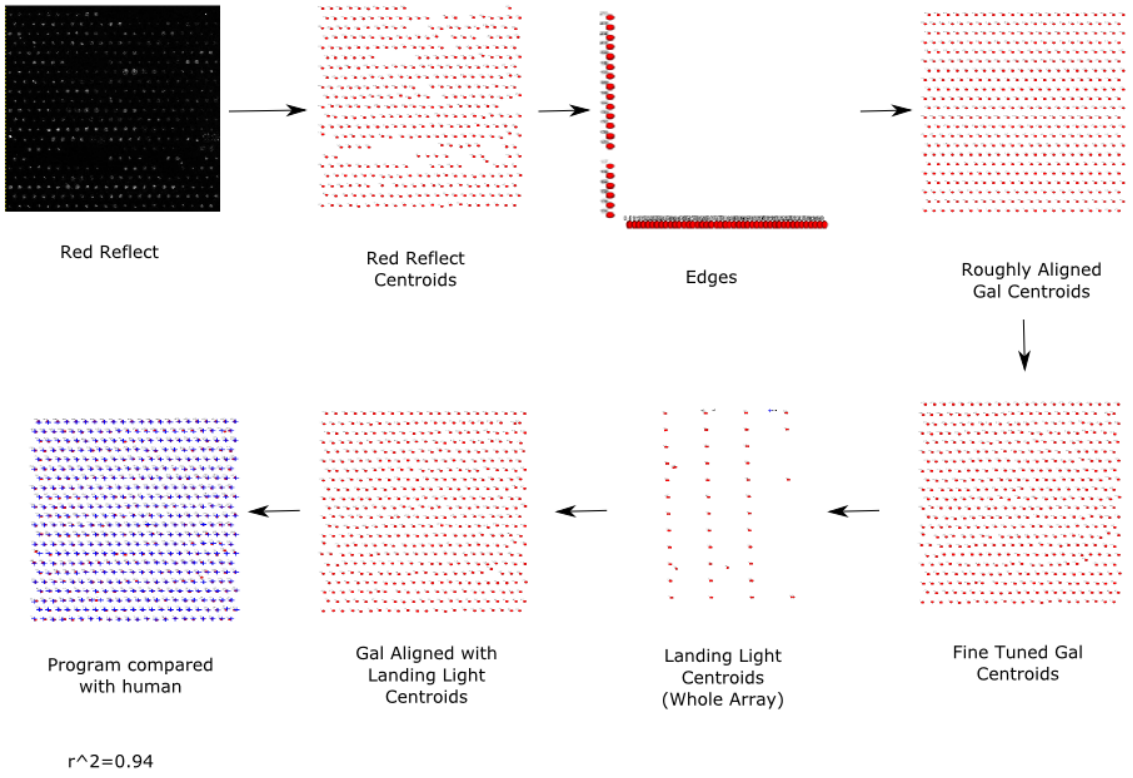


Figure 94 Automatic array alignment figure

The image of the red reflect is first input into the program and the centroids identified from the DBSCAN algorithm are located. The program then determines the locations of the edges of columns and rows to determine where the corners of the grid are located, and roughly aligns the centroids from the gal file with every labeled peptide feature present to the centroids identified from the red reflect. This rough alignment is accomplished by rescaling, rotating, and moving the gal file grid to match up with the corner points of the red reflect centroid grid. The program then fine tunes the rough alignment to move each centroid to the nearest red reflect centroid if there is one nearby. Next the program identifies the high intensity landing light centroids on the experiment array image (rather than the red reflect), and then transforms the gal file aligned with the red reflect to the real image from the experiment using the landing light

information. The correlation coefficient obtained by comparing a human alignment with the program alignment was 0.94.

D.4 Discussion

Aligning features on our non-natural sequence peptide microarrays is a time consuming task which could potentially be automated. I attempted to write a program to automatically align the features on the array, and I achieved some success. My program demonstrated that it could align one block on the array (there are 48 blocks on the CIM10K array) with a correlation of 94% to the alignment of a human

The program operated simply by using an algorithm called DBSCAN (density-based spatial clustering of applications with noise) to identify bright spots in the image. Before an experiment is performed with the peptide microarray slides, the slides are scanned to obtain a red reflect image. These red reflect images are much cleaner than the images obtained after an experiment since all of the peptide feature spots are present, the spots have about equal intensity, and there are not smears or streaks across the array. My program took the alignment from the red reflect and then performed a transformation to match up as closely as it could with the actual array image. Then the spots were moved to the closest DBSCAN identified spots on the experiment image if a DBSCAN spot was within a certain limit range. After I demonstrated that the technique could work for one block, the project was transferred to the programmer Nate Sutton in another lab to automate the alignment of the whole array with all of the blocks. Unfortunately correctly identifying the position of all of the blocks turned out to be a difficult task and the project was not completed. Perhaps a more distinguishing arrangement of the landing lights in the blocks would have made the task easier.

As of the current date several programmers in our lab have attempted to write programs to automate this task: Preston Hunter, Muskan Kukreja, and Kevin Brown. Our lab has also tried to make some arrangements with a Chinese software company to automate this task, but this arrangement did not pan out. Automatically aligning the arrays has proved to be a challenging task, but I still maintain that this is certainly something which should be possible.

If I were to start the project all over again, I think I would try to use a method with SVMs (support vector machines) which would not require any landing lights to be present on the slide at all. Basically, a

SVM can classify an object as belonging to one of two categories based on values of the features of that object. The SVM learns to make this classification from a set of training data. A method for alignment could involve using DBSCAN to identify the bright spots on the array. Then an SVM would determine whether a certain DBSCAN spot was spot “A” by looking at the features of that spot. The features would be assigned to be the position of other features on the array. In other words, the spot would basically be defined by its relative position to other spots. An SVM would be applied to each spot on the array to determine whether this is spot “A”, whether this is spot “B”, etc. Each SVM could make these judgments based on some previously human aligned training data.

This SVM alignment method could be performed by using the x,y coordinates of the 10 nearest spots to the spot in question. I think multiple SVMs would have to be trained with all possible orderings of the 10 nearest spots so that the collective group of SVMs is not restricted to assigning an identity to a spot only if "feature 1" is in an approximate location. If "feature 2" is in that approximate location then the correct assignment would be made by one of the SVMs in the group. In other words, "feature 1" can be at any number of one of the 10 spot locations. This is necessary because there could be an extra stain on the array which makes all the subsequent features have a different assignment. For example, if the 4th point is a stain, then the real 4th point would be the 5th, the real 5th point would be the 6th, etc. Likewise, instead of a stain which adds an extra spot, there could be spots that are missing which would also change the normal order of the real spots. Therefore, it would be necessary to train multiple SVMs for the identity of a feature with each different SVM trained with a different ordering of the closest 10 spots. In order to account for every possible ordering of the 10 closest spots, 10! = 3,628,800 SVMs would need to be trained in order to assign the identity of one spot. Although this is a very large number, this approach may still be computationally feasible since the speed of one single SVM with only 10 features is virtually instantaneous on a standard PC. One of the beautiful aspects of this method is that the assignment of an identity to one feature is completely independent from the assignment of an identity to the other features. Therefore, even if there is very large bad region on the array which makes it impossible to assign an identity to a large group of spots, this would not interfere with the correct assignment of spots in a different good region of the array. This method also does not rely on any

landing lights. For the spots that did not receive an assignment, a human could even go in and inspect these unassigned spots manually if desired.

In summary, a strategy was pursued to automatically align bright spots in an image with a pattern of features. This strategy implemented the DBSCAN algorithm, landing lights, and many spatial transformations. The end result was a program which could align one block of a 10k array of 48 blocks with a correlation coefficient of 0.94 with a human alignment. The program was never completed to automatically align an entire array. If the project was started again, a strategy that implemented the SVM algorithm would have many benefits and may be a superior method. Regardless of the method used, a program capable of automatically aligning the arrays would save researchers a considerable amount of time so that they can pursue other tasks.

APPENDIX E

CONSTRUCTION OF HUMAN TUMOR CDNA

First steps toward the construction of a human tumor cDNA library were taken. Human breast tumor RNA was obtained from Dr. Stephen Lyle at the Cancer Center Tissue and Tumor Bank from the University of Massachusetts. The characteristics of this RNA sample are presented in Table 18.

Table 18 Human breast tumor RNA characteristics

Sample ID	Sample Date	Organ	Cancer/Met	Cancer Type	Grade	ER	PR	HER-2	Age	Gender
2802T	10/25/2010	Breast	Malignant	Ductal	Carcinoma	-	-	-	53	Female

The attributes of the human breast tumor RNA sample are presented. The RNA came from a malignant breast ductal carcinoma from a 53 year old female. The carcinoma was negative for all three of the markers ER, PR, and HER-2.

Next cDNA was synthesized from this RNA using random hexamers as presented in Figure 95. The cDNA made from 18 cycles was selected for cloning into a vector and transforming.

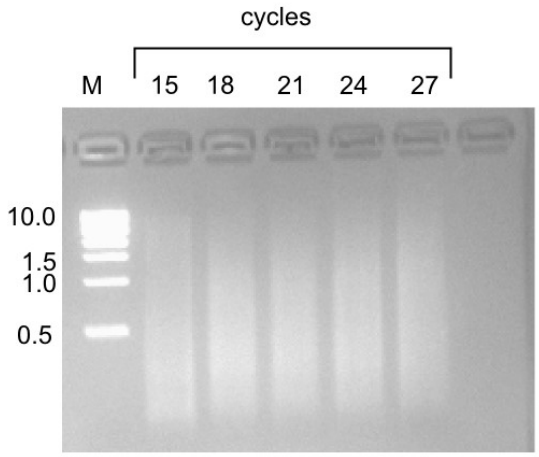


Figure 95 Tumor cDNA synthesized from random hexamer primers

The DNA from a PCR of human breast tumor RNA with random hexamer primers with 15, 18, 21, 24, or 27 PCR cycles.

After this DNA was transformed into *E. coli* and the plasmids inside some of the colonies were sequenced, most of the results were short poly T sequences. Therefore, the library had to be remade. However, when remaking the library, random pentadecamers were used instead of random hexamers since pentadecamers increase the yield and quality of the resulting cDNA [206]. The integrity of the RNA

was checked before constructing the random pentadecamer library (Figure 96). After the library was constructed with random pentadecamers, the cDNA was purified using AMPURE magnetic beads, but then there were no colonies after the transformation into *E. coli*. When the library was constructed again (Figure 97), electroelution was used to purify the cDNA rather than using magnetic beads (Figure 98). An ethanol precipitation was then performed on this electroeluted DNA (Figure 99). This purified DNA was ligated into a vector by performing an in-fusion reaction, and another ethanol precipitation was performed (Figure 100). After all of these steps, the final percent recovery of cDNA was calculated as 6.19% with a final sample concentration of 13.2 ng/µl.

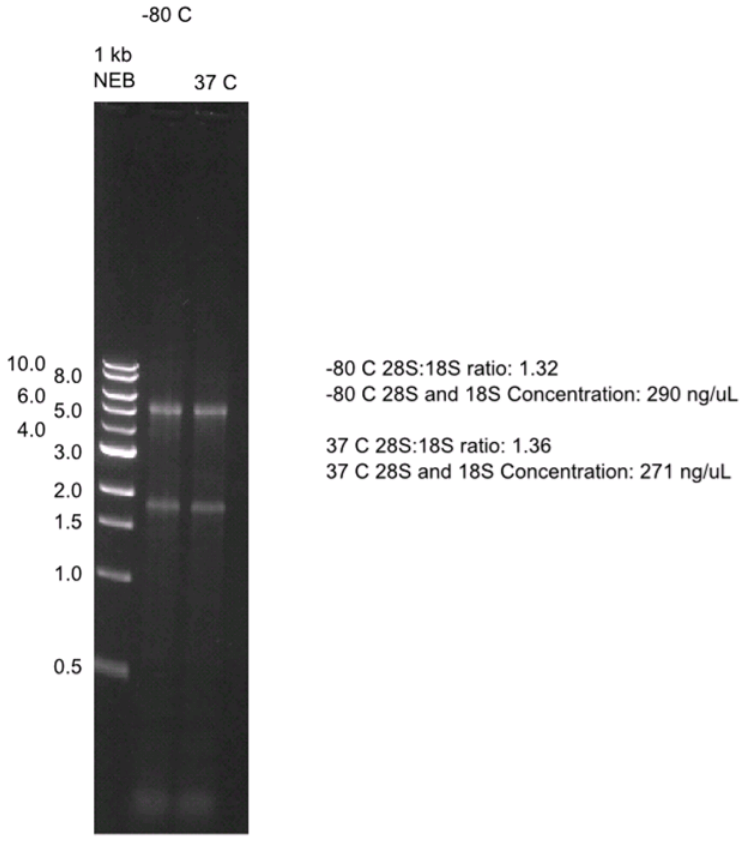


Figure 96 RNA integrity check before random pentadecamer library construction

One aliquot of the RNA was incubated at 37 °C, and one aliquot was incubated at -80 °C for 2 hours. The samples were then run on a gel and the intensity of the 28S to 18S bands were compared.

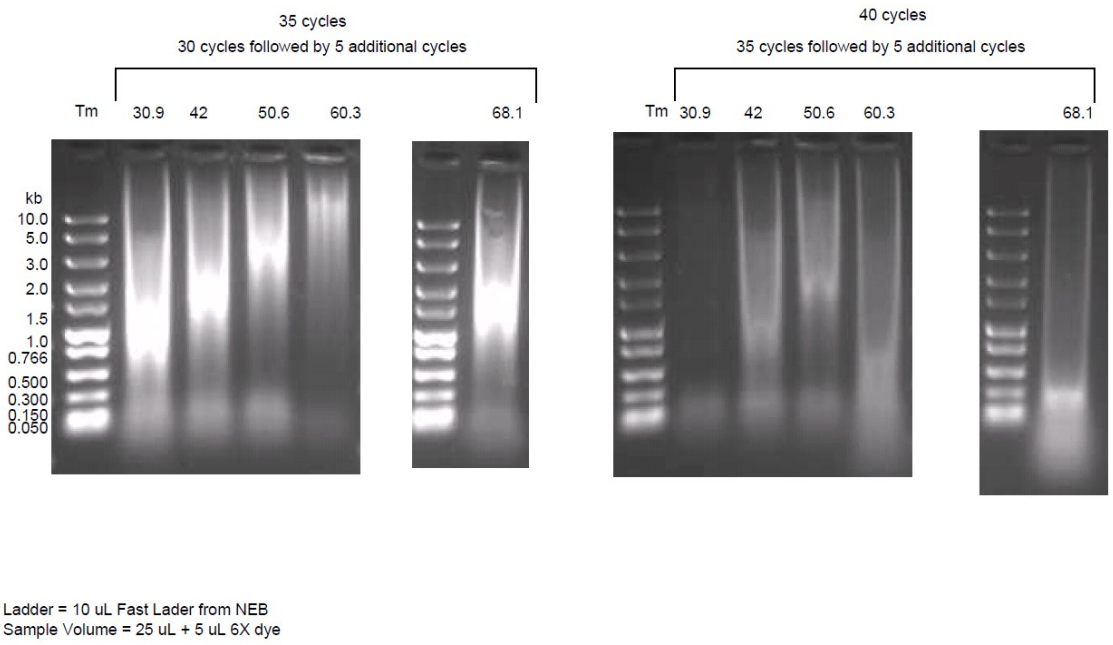


Figure 97 Second strand synthesis with random pentadecamer primers

Second strand synthesis was produced using random pentadecamer primers rather than random hexamer primers. In the first two gel images, the PCR is performed with 35 cycles at a variety of different annealing temperatures (30.9, 42, 50.6, 60.3, and 68.1 °C). In the two gel images on the right, the PCR is performed with 40 cycles at a variety of different annealing temperatures (30.9, 42, 50.6, 60.2, and 68.1 °C).

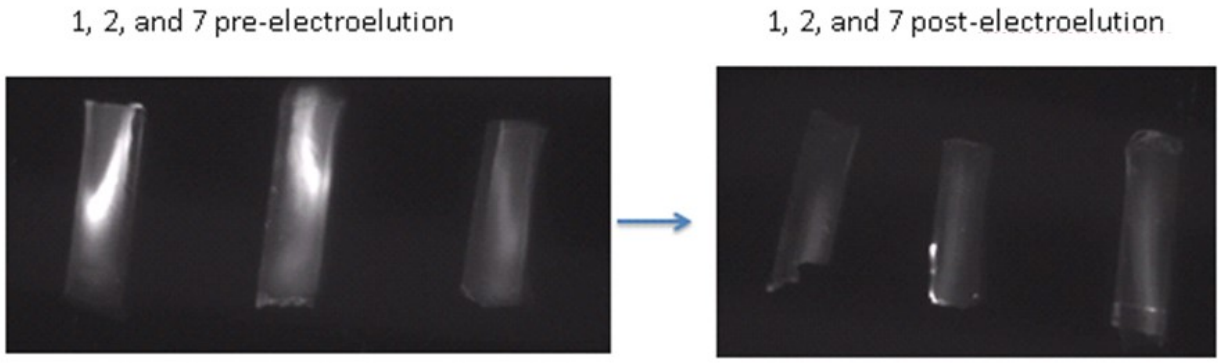
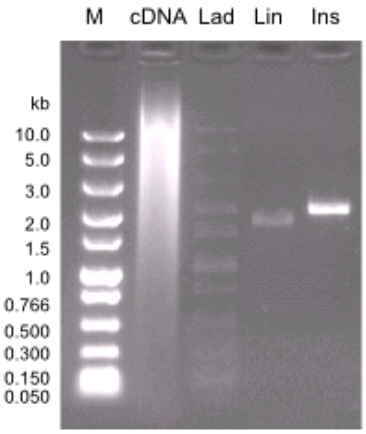


Figure 98 Gel fragments pre and post electroelution

Images of cut out gel fragments pre and post electroelution. The "1", "2", and "7" just refer to the lanes from which the fragments were cut from in Figure 97, and in this case the "1" refers to cDNA made with 35 cycles 30.9 °C Tm, "2" corresponds to 35 cycles 42 °C Tm, and "7" refers to 40 cycles 42 °C Tm.

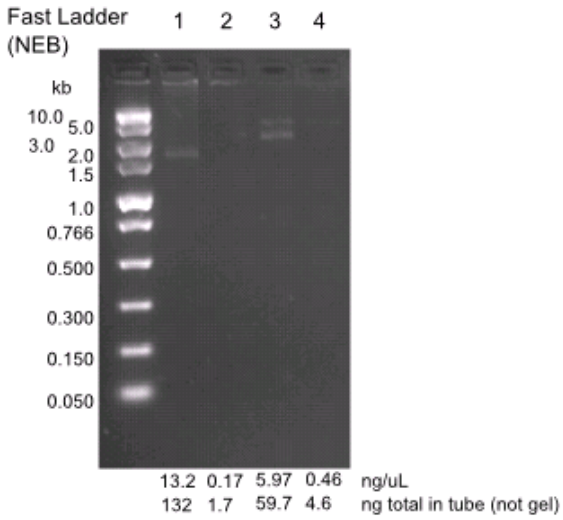


M	cDNA	Lad	Lin	Ins	
	43	74.7	21.2	30.8	Estimated Percent Recovery after electroelution and ethanol precipitation
	248	56	12.3	10.6	Estimated Concentration (ng/uL)

M = 10 uL Fast NEB Ladder
cDNA = cDNA constructed from random pentadecamers
Lad = Fast NEB ladder
Lin = pUC19 Control Vector, linearized
Ins = 2 kb Control Insert

Figure 99 cDNA library after electroelution and precipitation

Human breast cancer cDNA after electroelution and ethanol precipitation purification. The other ladder, linear vector, and control insert were controls that would be subjected to the following in-fusion and precipitation steps along with the cDNA.



1 = 4T1 cDNA created from random pentadecamers (13.2 ng/uL)
2 = Positive control eluted (2.7 kb PUC19 and 2 kb control insert) (0.17 ng/uL)
3 = Positive control (2.7 kb PUC19 and 2 kb control insert) (5.97 ng/uL)
4 = Negative Control (2.7 kb pSMART2IFD) (0.46 ng/uL)

Figure 100 cDNA library after electroelution, ethanol precipitation, in-fusion, and ethanol precipitation

The cDNA and PUC19 control insert were electroeluted, ethanol precipitated, in-fused into a vector, and ethanol precipitated. The amount of DNA in each lane estimated from the brightness of the bands is indicated below the gel image.

Some experiments were also performed to test variables that would affect the transformation efficiency. The conditions tested were as follows: presence or absence of glycogen; use of heat (50 °C 10 m) or not on DNA sample before transformation; and use of 10 mM Tris buffer or water (Figure 101). These results demonstrate that if glycogen is used as a carrier molecule, then a much higher transformation efficiency can be obtained if the sample is heated preceding transformation.

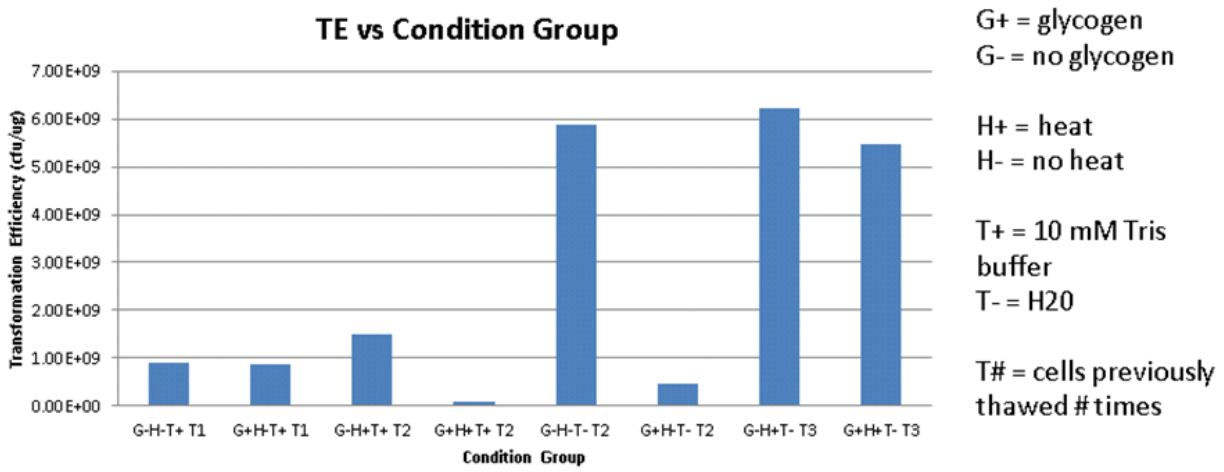


Figure 101 Electroporation condition test

The transformation efficiency of electroporation with many different conditions was tested. A "-" indicates absence of a condition, and a "+" indicates presence of a condition. "G" indicates that glycogen was added to the mixture, "H" indicates that the DNA was heated to 50 °C for 10 m before electroporation, and "T" indicates that 10 mM Tris buffer was used instead of water. A "T1", T2", or "T3" at the end of the group designation indicates that the cells used were previously thawed one, two, or three times respectively.

In conclusion, initial steps were taken to construct a human tumor cDNA library. Unlike many other cDNA libraries, this library was constructed using random pentadecamers instead of random hexamers. The DNA was also purified using electroelution instead of beads. Transformation conditions were also tested, and heating the glycogen DNA sample to 50 °C before transformation resulted in a higher transformation efficiency. Using these procedures may result in a library with good representation of the original transcripts, but this particular library was not completed as other projects took priority.

APPENDIX F

ABSTAT ANALYSIS OF TUMOR CDNA LIBRARY MICE

The first chapter of this dissertation discusses the AbStat metric, and the second and third chapters discuss screening a tumor cDNA library constructed from a mouse tumor. Therefore, one might be curious about the AbStat results of the tumor mice used to construct the tumor cDNA library. Sera samples from these naïve and tumor mice were applied to the non-natural sequence peptide microarray, and therefore the AbStat results can be determined. The tumor sera sample was composed of a pool of sera from the five tumor mice, and the naïve sera sample was composed of a pool of sera from the five naïve mice. The Peptide Array Core then applied these samples to the CIM10Kv2 peptide microarrays with a replicate for each sera sample. The resulting fluorescence intensity distributions from the experiment were used to obtain AbStat results. Graphs of the entropy for both groups (Figure 102), a heatmap displaying the relative intensities of the AbStat values for each sample (Figure 103), and the statistical significance of each measure (Figure 104) are displayed below. These results show that the tumor sample has a higher average entropy than the normal group. The p-value from a t-test with the entropy values is 0.0638. The p-value may have been more significant if the sample size of each group was larger than the minimum of two required for a t-test. The heatmap shows that most of the AbStat measures are relatively higher for the tumor samples compared to the normal samples. The bar graph of the statistical significance for each measure reveals that the max, median, and entropy were the measures which distinguished between the two groups the best. Overall these results support the previous findings that samples from organisms with cancer often have higher entropy values ("1.3.5 Mouse cancer progression" and "1.3.6.2 HT330K wafer 46").

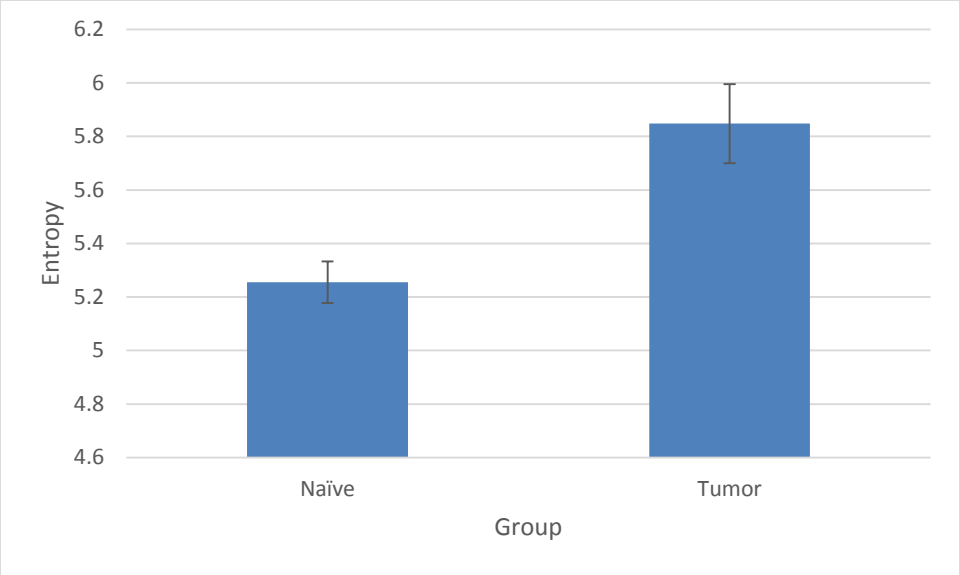


Figure 102 Entropy of naive mice and tumor mice used to construct a cDNA library

The entropy calculated from the peptide fluorescence intensity distribution after the application of naïve mouse sera or tumor sera is plotted in a bar graph. There was a technical replicate for each group for a total of two samples per group. The naïve sera sample consisted of a pool of five naïve mice, and the tumor sample consisted of a pool of five tumor mice.

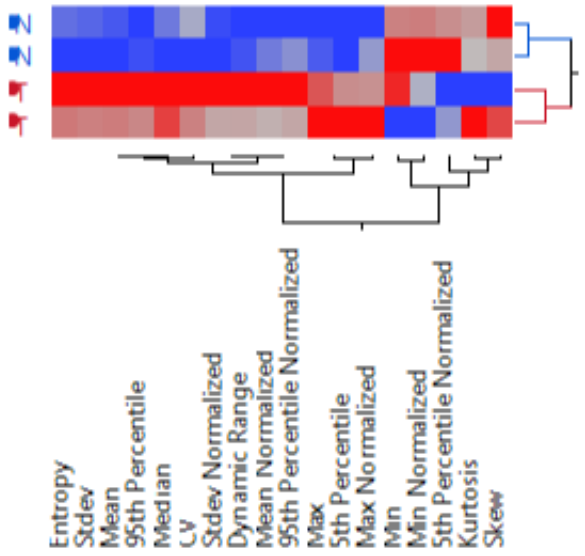


Figure 103 Heatmap of measures for naive mice and tumor mice used to construct cDNA library

Each column corresponds to an AbStat measure and each row corresponds to a sera sample. The relative average value of each AbStat measure for the samples is represented by a color with blue indicating the lowest relative value and red indicating the highest relative value. The class of each sample is designated as follows: N = normal, T = tumor.

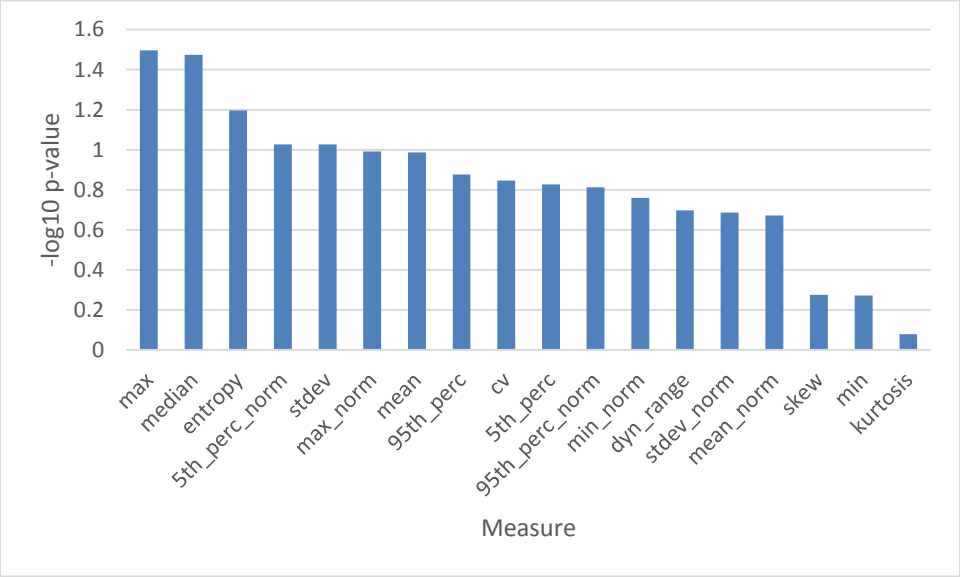


Figure 104 Statistical significance of measures comparing normal and tumor mice used to construct cDNA library

Sera samples were applied to the CIM10Kv2 array and the AbStat measures were calculated. A p-value from a t-test was then determined for each measure for tumor vs normal. The negative logarithm in base 10 of the p-value was then plotted in a bar graph.

APPENDIX G

INSTITUTIONAL REVIEW BOARD (IRB)

ARIZONA STATE UNIVERSITY
Office of Research Integrity and Assurance
660 S. Mill Avenue Suite 315
Arizona State University
Tempe AZ 85287-6111
(Mail Code 6111)
Phone: 480-965-6788
Fax: (480) 965-7772

CONTINUING REVIEW FORM- IRB

- In accordance with Federal Regulations 45CFR46, the IRB must review nonexempt protocols at least annually, or more frequently if warranted.
- Please type your responses in the boxes provided. Use as much space as necessary (the boxes will expand). Please answer each question – if a question is not applicable, please put N/A in the box.
- Studies that are in the data analysis phase are considered open, researchers must complete this form.

1. Principal Investigator	
Principal Investigator: Stephen Albert Johnston	
ASU department address: Center for Innovations in Medicine, Biodesign Institute B230 MC5901	
E-mail address: Stephen.johnston@asu.edu	
Phone number: 480-727-0792	Fax Number: 480-727-0782
Co-Investigator(s) Name(s) and Contact Information: Phillip Stafford, Phillip.stafford@asu.edu; Kathryn Sykes, Kathryn.sykes@asu.edu; Lucas Restrepo, lucas.restrepo@asu.edu; Muskan Kukreja, muskan.kukreja@asu.edu	

2. Protocol Information
2a) Title of protocol: Profiling Human Sera for Unique Antibody signatures
2b) HS #: 0912004625
2c) If project is funded or funding is being sought, provide list of all sponsors and grant numbers: DTRA HDTRA1-11-1-0010; DoD BCRP W81XWHO710549; Please indicate the grant status for each source of funding: ☒ Active ☐ Pending
2d) ASU account number/project number: FQS0052, FQS0030
2e) Location(s) of research activity: Biodesign Center for Innovations in Medicine B225, B229, B233, B237
2f) IRB approval dates from additional institutions: All samples are provided to us from institutions that are current with their IRB approval. We currently do not have that information but can obtain if needed. **Please note that copies of current IRB approvals from additional institutions are required.*

3. Protocol Status
3a) Active: ☒ Yes ☐ No (If no, submit a close out report: http://researchintegrity.asu.edu/humans/forms
3b) Please indicate remaining duration of the study: 9 years
4. Participant Information
4a) Is this study closed to enrollment of new subjects: ☒ Yes ☐ No

4b) Total number of participants approved for the study (to be enrolled): N/A
4c) Number of participants enrolled (e.g. signed a consent form) during the **past approval period**: N/A
4d) Total number of participants enrolled **since study began**: N/A
4e) Total number of individuals screened (e.g. individuals that responded to study advertisements or other recruitment tractices and were questioned by investigators) in the past approval period (if applicable): N/A (this includes the number that was later enrolled)
4f) Of the total number of individuals screened in the past approval period, what percentage has been ineligible to participate in the study (if applicable)? N/A
4g) Number of enrolled participants who withdrew from the study: N/A Please state the reason(s) the participant(s) withdrew.
4h) Number of participants still to be enrolled: N/A (If this brings the sample to greater than what is listed in 4b, submit a request for modification see 7d).
4i) Participant enrollment breakdown by gender, age and ethnicity: (This information is required for all studies that are NIH-sponsored. It is recommended, but not required, that other researchers provide this information). N/A

5. Data Sources Check all categories that apply to your protocol:	
☐	Human subjects intervention with use of informed consent form
☐	Discarded, identified pathological materials, no intervention
☐	Genetic analysis
☐	Interviews or questionnaires
☐	Medical records or other records from human subjects
☒	Other please specify: **We have clinical diagnosis and treatment status, genotype information and maybe smoking history. All information though is provided to us.**

6. Adverse Events or Unexpected Problems

6a) Have there been any complaints from subjects in the past approval period?
☐ Yes If yes, describe ☒ No

6b) Have there been any ***adverse events*** or unexpected problems in the past approval period?
☐ Yes ☒ No
If yes, please explain in detail and indicate when the IRB was notified of the event or problem. If the IRB was not notified, please explain why this was not done.

6c) Does the study have a Data Safety Monitoring Board (DSMB)? ☐ Yes ☒ No
If yes, please indicate the date of the last DSMB review:

Please note that investigators are required to submit DSMB reports to the ASU IRB at the time they are made available to the investigator.

7. Protocol Modifications or Revisions

7a) Have there been any modifications or revisions to the protocol in the past approval period?
☐ Yes ☒ No
If yes, please indicate the date of the approval from the Committee for the modification or revision and provide a brief description.

7b) Have there been any deviations from the approved protocol? ☐ Yes ☒ No
If yes, please describe to self-report the protocol violation.
7c) Do you want to add any new co-investigators to the study? ☐ Yes ☒ No
If yes, submit their names and copies of the human subjects training required by the IRB: http://researchintegrity.asu.edu/training/humans
7d) Do you wish to submit a modification at this time? ☒ Yes ☐ No
If yes, please describe the modification request and rationale for the changes: Please remove Dr. Patricia Carrigan from the protocol. She has left ASU.

8. Current Consent Form

8a) Please attach a copy of your current consent form for renewal if you are enrolling new subjects. N/A

8b) Is this the original consent form or a revised form? ☐ Original ☐ Revised (If revised, please provide date of ASU IRB approval for the revision. Attach a copy of the stamped form and unstamped form)

9. Protocol Progress Report

9) Please submit a ***detailed*** progress report. The progress report must be substantive and complete, and include the goal(s) of the study, findings to-date, how data is being stored, and plans for the next year/review period. If this project is funded, please send a copy of the most recent progress report that was sent to the funding agency:

The last year, our team has optimized the immunosignature microarray and has contracted with Applied Microarrays (AMI) in Tempe, AZ to print our arrays. We obtained a new set of 10,000 different random peptides, as the last set had been depleted. We ensured that the new peptides were carefully diluted in a new buffer/organic mix that is compatible with AMI's printing process. The added precision of commercial printing has allowed us to obtain higher reproducibility across patients, and find much more subtle changes in antibody responses. We have completed the Valley Fever project by printing a set of 100-peptide 'diagnostic arrays' to do the test-training sample sets. We have obtained 65 look-back blinded samples from John Galgiani at U of A in Tucson that were all false negative samples from his clinic. We classified these samples with 0% error (after excluding problematic samples that were inherently high-background or had been subject to degradation effects). We are in the process of writing these data up in a manuscript.

We completed a project on glioblastoma multiformae, using blinded samples from Barrow Neurological Institute (BNI) in which we were able to identify brain cancer grade as well as presence or absence of an important methylation enzyme, MGMT. This enzyme's status has been shown to be an effective predictor of response to Temozolamide. We have submitted this manuscript to NeuroOncology.

We have completed a project on Esophageal Cancer, using blinded samples obtained from Mayo Clinic, in which we were able to distinguish presence or absence of Esophageal cancer in patients. We are currently examining samples from patients with Barrett's Esophagus, to determine whether we can detect early cancer predisposition.

We have built a pathogen microarray, in which 5K peptides from human pathogens were tiled on a standard glass slide. We are currently optimizing this platform to distinguish patients who are convalescent from one or another infectious agent. We have found that printing methods that enhance the immunosignaturing effect are deleterious to the discrimination of our pathogen epitope arrays. Thus we are altering the printing characteristics for these arrays, and are using a slide surface that spaces the peptides out much further than our aminosilane slides allow.

10. Publications, Presentations and Recent Findings

10a) Have there been any presentations or publications resulting from this study during the past approval

period? ☒ Yes ☐ No If yes, please submit a copy of the abstract, or the publication, with this application.

"Immunosignaturing can detect products from molecular markers in brain cancer" – submitted to NeuroOncology
"Physical parameters Affecting Antibody Profiles as Biomarkers of Health Status" – revision resubmitted to Molecular and Cellular Proteomics
"Sample Preparation for Immunosignaturing" – revision resubmitted to Vaccine

Presentations:
BRP FY11 Vision Setting Meeting, November 2010 – Panel Member
3rd Annual Oncology Biomarker Conference, January 2011 – Invited speaker
Leading Innovation and Knowledge Sharing (LINKS) BCMRP meeting, February, 2011 - Panel member
BCRP FY10 Programmatic Review Meeting, March 2011 – Panel member
Canary Foundation, March 2011 – Invited participant
Era of Hope Abstract Placement Meeting, April 2011 – Committee member
NBCC Artemis Project, April 2011 – Workshop participant
Era of Hope Meeting, Orlando, August 2011 – Invited speaker, Organizing committee member

10b) Have there been any recent findings either from this study, or a related study (through a literature review for example), that would have an effect on this study's risk/benefit analysis? ☐ Yes ☒ No
If yes, please describe and cite references:

11.Conflicts of Interest and Commercialization

11. Does any member of the research team have a potential conflict of interest with this study that could affect study participants and/or study outcome? For more information about examples of conflicts of interests, please visit the ASU objectivity website: http://researchintegrity.asu.edu/coi
☐ Yes (If yes, please describe and disclose in the consent form) ☒ No

b) Does the PI or Co-I have a current conflict disclosure form on file at the ASU Office of Research Integrity and Assurance?
☐ Yes ☐ No

c) If there are conflicts of interests, please describe the ways in which you have and will minimize harm to research subjects and/or the objectivity of research. No prospective human trials have been proposed, only blinded retrospective samples are currently being run on the immunosignaturing platform.

12. Training

12.The research team must verify completion of human subjects training within the last 3 years. (http://researchintegrity.asu.edu/training/humans)

CITI training – Provide the date that the PI and Co-I's completed the training:
If you completed **NIH** training prior to 9/15/10 this will be accepted. Provide a copy of the certificate.

13. Required Signatures

Principal Investigator:	Date: 10/19/11

FOR IRB USE Chair or Committee member name: Signature: Date:

Revised 8/11

ASU Knowledge Enterprise Development

Office of Research Integrity and Assurance

To:	Stephen Johnston BDB
From:	Carol Johnston, Chair Biosci IRB
Date:	10/24/2011
Committee Action:	**Renewal**
Renewal Date:	10/24/2011
Review Type:	Expedited F5
IRB Protocol #:	0912004625
Study Title:	Profiling Human Sera for Unique Antibody Signatures
Expiration Date:	11/21/2012

The above-referenced protocol was given renewed approval following Expedited Review by the Institutional Review Board.

It is the Principal Investigator's responsibility to obtain review and continued approval of ongoing research before the expiration noted above. Please allow sufficient time for reapproval. Research activity of any sort may not continue beyond the expiration date without committee approval. Failure to receive approval for continuation before the expiration date will result in the automatic suspension of the approval of this protocol on the expiration date. Information collected following suspension is unapproved research and cannot be reported or published as research data. If you do not wish continued approval, please notify the Committee of the study termination.

This approval by the Biosci IRB does not replace or supersede any departmental or oversight committee review that may be required by institutional policy.

Adverse Reactions: If any untoward incidents or severe reactions should develop as a result of this study, you are required to notify the Biosci IRB immediately. If necessary a member of the IRB will be assigned to look into the matter. If the problem is serious, approval may be withdrawn pending IRB review.

Amendments: If you wish to change any aspect of this study, such as the procedures, the consent forms, or the investigators, please communicate your requested changes to the Biosci IRB. The new procedure is not to be initiated until the IRB approval has been given.

ASU Knowledge Enterprise Development

Office of Research Integrity and Assurance

Biosci IRB

To: Johnston,Stephen Albert **Date:** 09 / 26 / 2011

From: Biosci IRB **Expiration Date:** 11 / 22 / 2011

Re: Protocol # 0912004625: Profiling Human Sera for Unique Antibody Signatures

This letter serves as a IRB notification reminder by the Biosci IRB. It is the primary responsibility of the Principal Investigator to ensure that the re-approval status for lapsed protocols is achieved. All protocols must be re-approved annually by the IRB unless shorter intervals have been specified.

Please note that the level of review given to the continuing review process is the same as that of any new protocol. All requests for re-approval must be reviewed at a convened IRB meeting, except for those protocols that meet the criteria for expedited review.

Please submit the following documents at least three weeks prior to the expiration date to allow for full committee review:

1) A completed Continuing Review Form.

2) Two (2) copies of each consent form(s) used in the study (If data collection is ongoing).

Please note that you can obtain a copy of the Continuing Review Form through our web site: http://researchintegrity.asu.edu/humans.

As of July 1, 2003, all personnel involved in human subjects research must complete the Human Subjects training course. It is the responsibility of the Principal Investigator to make sure all personnel associated with this study have completed the human subjects training course (see the Office of Research Integrity and Assurance website for a link to the NIH training).

It is a violation of Arizona State University policy and federal regulations to continue research activities after the approval period has expired. If the IRB has not reviewed and re-approved this research by its current expiration date, all enrollment, research activities and intervention on previously enrolled subjects must stop. If you believe that the health and welfare of the subjects will be jeopardized if the study treatment is discontinued, you may submit a written request to the IRB to continue treatment activities with currently enrolled subjects.

Your assistance and cooperation in ensuring that the above-mentioned protocol is received for re-approval evaluation at the Office of Research Integrity and Assurance before the lapse date is greatly appreciated.

Sent to Office of Research Integrity
10/18/11

APPENDIX H

INSTITUTIONAL ANIMAL CARE & USE COMMITTEE

utional Animal Care and Use Committee (IACUC)
izona State University

Tempe, Arizona 85287-3503
(480) 965-2179 FAX: (480) 965-8013

Animal Protocol Review

Protocol Number: 05-817R
Protocol Title: Genetic Cancer Vaccines
Principal Investigator: Stephen Johnston
Date of Action: 06/17/2005 Final Action Date: 06/17/2005

The animal protocol review was considered by the Committee and the following decisions were made:

- ☐ The original protocol was APPROVED as presented.
- ☒ The revised protocol was APPROVED as presented.
- ☐ The protocol was APPROVED with RESTRICTIONS or CHANGES as listed below. The project can only be pursued, subject to your acceptance of these restriction or changes. If you are not agreeable, contact the IACUC Chairperson immediately.
- ☐ The Committee requests CLARIFICATIONS or CHANGES in the protocol as described below. Approval is contingent upon review and approval of the required revisions by the IACUC Chair.
- ☐ The protocol was approved, subject to the approval of a WAIVER of provisions of NIH policy as noted below. Waivers require written approval from the granting agencies.
- ☐ The protocol was DISAPPROVED for reasons outlined in the attached memorandum.
- ☐ The Committee requests you to contact _____ to discuss this proposal.
- ☐ A copy of this correspondence has been sent to the Vice President for Research.

RESTRICTIONS, CHANGES OR WAIVER REQUIREMENT:

Approved Number of Animals: 3,000 Mice

Approval Period: 06/17/2005 – 06/16/2008

Signature: _____ Date: 06/17/2005
IACUC Chair or Designee

Investigator cc: IACUC Office, IACUC Chair, ORSPA

VI. DUPLICATION AND ALTERNATIVES

A. Provide the following details for the most recent literature search used to explore for duplicative research, alternatives to painful procedures and most currently relevant teaching use of animals.

Date that search was conducted: 4/5/2005
Database used: Medline
Publication years covered by the search: Last 5 years
Keywords used: Genetic immunization, cancer vaccine, tumor antigens, melanoma, and breast cancer

B. Describe any other procedures (e.g., participation in meetings, review of journals) that are used to evaluate duplication and explore alternatives:

We will continue to monitor commercial sources should one be available. In addition, we will monitor scientific literature for comparable reagents as needed. Journals that we routinely monitor include Science, Nature groups, Vaccine, PNAS, EMBO, Cancer Research, to name a few. In addition, Dr. Johnston attends multiple scientific meetings each year to keep abreast of new developments.

C. Does this research replicate previous work? **NOTE: Teaching protocols need not address Item VI.c.**

☑ No. Proceed to section VII.
☐ Yes. Explain why the replication is necessary:

VII. ASSURANCE:

The information contained herein is accurate to the best of my knowledge. I have carefully compared the proposed work with the current state of knowledge in this field by reviewing the literature and it is my professional opinion that the proposed work meets high standards of scientific merit. If the study involves pain and distress to the animal, whether or not it is relieved by anesthetics or analgesics, I have (1) reviewed the literature related to this work and have found no significant studies which could make this protocol unnecessarily duplicative, and (2) considered alternatives to animal use and found none available, as described above. Procedures involving animals will be carried out humanely and all procedures will be performed by or under the direction of trained or experienced persons. Any revisions to animal care and use in this project will be promptly forwarded to the Animal Care and Use Committee for review. Revised protocols will not be used until Committee clearance is received. The use of alternatives to animal models has been considered and found to be unacceptable at this time.

The principal investigator, by signing below, and the IACUC recognize that other medications may be given to the animals for veterinary care purposes (including humane euthanasia of animals in pain that cannot be controlled, as determined by the University Veterinarian or a euthanasia-certified principal investigator).

Individual listed on P.A. — Date: 6/30/05

Stephen Albert Johnston, Director CTM — Date: 6/30/05
Insert Name and Title Here – copy and paste as necessary

***Department Chair — Date

***College Dean — Date

***ASU East requires these signatures.

NOTE: Principal investigators are requested to attach a two-page biosketch reflecting their most recent pertinent experience.

Institutional Animal Care and Use Committee (IACUC)
Arizona State University

Tempe, Arizona 85287-1103
(480) 965-2179 FAX: (480) 965-7772

Animal Protocol Review

ASU Protocol Number: 08-1000R
Protocol Title: Genetic Cancer Vaccines
Principal Investigator: Stephen Johnston
Date of Action: 07/01/2008

The animal protocol review was considered by the Committee and the following decisions were made:

- ☒ The original protocol was APPROVED as presented.
- ☐ The revised protocol was APPROVED as presented.
- ☐ The protocol was APPROVED with RESTRICTIONS or CHANGES as noted below. The project can only be pursued, subject to your acceptance of these restriction or changes. If you are not agreeable, contact the IACUC Chairperson immediately.
- ☐ The Committee requests CLARIFICATIONS or CHANGES in the protocol as described in the attached memorandum. The protocol will be reconsidered when these issues are clarified and the revised protocol is submitted.
- ☐ The protocol was approved, subject to the approval of a WAIVER of provisions of NIH policy as noted below. Waivers require written approval from the granting agencies.
- ☐ The protocol was DISAPPROVED for reasons outlined in the attached memorandum.
- ☐ The Committee requests you to contact _____ to discuss this proposal.
- ☐ A copy of this correspondence has been sent to the Vice President for Research.
- ☐ Amendment was approved as presented.

Approved # of Animals: 6,336 Mice **Pain Level:** D
Approval Period: 07/01/2008 - 06/29/2011
Funded: Department of Defense
Title: Towards Developing a Prophylactic Breast Cancer Vaccine

Signature: _____ Date: 7/1/08
IACUC Chair or Designee

Original: Principal Investigator
cc: IACUC Office
IACUC Chair
ORSPA/SPS

Institutional Animal Care and Use Committee (IACUC)
Office of Research Integrity and Assurance
Arizona State University
660 South Mill Avenue, Suite 315
Tempe, Arizona 85287-6111
Phone: (480) 965-4387 *FAX:* (480) 965-7772

Animal Protocol Review

ASU Protocol Number: **11-1197R**
Protocol Title: **Genetic Cancer Vaccines**
Principal Investigator: **Stephen Johnston**
Date of Action: **06/24/2011**

The animal protocol review was considered by the Committee and the following decisions were made:

- ☐ The original protocol was APPROVED as presented.
- ☒ The revised protocol was APPROVED as presented.
- ☐ The protocol was APPROVED with RESTRICTIONS or CHANGES as noted below. The project can only be pursued, subject to your acceptance of these restriction or changes. If you are not agreeable, contact the IACUC Chairperson immediately.
- ☐ The Committee requests CLARIFICATIONS or CHANGES in the protocol as described in the attached memorandum. The protocol will be considered when these issues are clarified and the revised protocol is submitted.
- ☐ The protocol was approved, subject to the approval of a WAIVER of provisions of NIH policy as noted below. Waivers require written approval from the granting agencies.
- ☐ The protocol was DISAPPROVED for reasons outlined in the attached memorandum.
- ☐ The Committee requests you to contact _____ to discuss this proposal.
- ☐ A copy of this correspondence has been sent to the Vice President for Research.
- ☐ Amendment was approved as presented.

RESTRICTIONS, CHANGES OR WAIVER REQUIREMENTS:

Total # of Animals: **9,024** **Pain Level:** **B-720; C-3,074; D-5,230** **Species:** **Mice**
Sponsor: **Department of Defense**
Title: **Towards Developing a Prophylactic Breast Cancer Vaccine**
Proposal #: **W81XWH0710549**
Approval Period: **06/24/2011 – 06/23/2014**

Signature: _____ Date: 6/24/11
IACUC Chair or Designee

Original: Principal Investigator
Cc: IACUC Office
IACUC Chair

Date: 6/7/2012

ARIZONA STATE UNIVERSITY
IACUC ANNUAL REVIEW

I. Currently approved protocol

Protocol Number: 11-1197R
Protocol Title: Genetic Cancer Vaccine
Principal Investigator: Stephen Johnston

II. Status of Project

A. Was the research or teaching conducted?

i. ☐ No. If no,
1. Will the protocol be terminated?

a. ☐ Yes. Proceed to item VI.
b. ☐ No. Proceed to item II B.

ii. ☒ Yes. If yes,
1. Were there any significant animal welfare issues (morbidity or mortality, complications, etc.) encountered over the past 12 months?

a. ☒ Yes. Please describe (include the problem, approximate number of animals affected, and resolution). Proceed to item II B when completed.
About 5% of the FVB/N-NeuT females spontaneously died. This is a common characteristic of this transgenic mouse strain. All of the other mice behaved normally.

b. ☐ No. Proceed to item II B.

B. Will the research or teaching continue with no anticipated protocol changes in animal species, animal numbers, or categories listed below for the next 12-month period?

- Procedures
- Criteria to Measure/Monitor Pain or Distress
- Alternatives to Painful Procedures
- Restraint
- Amelioration and Control of Painful Procedures
- Estimation of Potential Postoperative/Intervention Pain
- Postoperative/Chronic Care
- Euthanasia/Disposition of Animals
- Animal Care and/or Use Sites

i. ☒ Yes. Proceed to item III.
ii. ☐ No. If there will be proposed changes, you must complete an Amendment Request form describing all proposed changes as well as the scientific rational for these changes. Proceed to item III.

III. Updated Information

A. Please evaluate the Category of Pain as stated in your currently approved protocol. Do you feel it remains appropriate for the procedures performed?

i. ☒ Yes. Proceed to item III B.
ii. ☐ No. If no, please describe: Proceed to item III B when completed.

B. Have there been any recent findings, either from this study or a related study, that would change the planned use of animals?

i. ☐ Yes. If yes, cite references below or in an attachment and submit an Amendment Request form. Proceed to item IV when completed.

ii. ☒ No. Proceed to item IV.

IV. Progress Report

Provide a statement on progress of your teaching or research under this protocol over the past 12 months. Include any presentations or publications that have resulted from this protocol during the past 12 months.

We optimized the immunization regime for our prophylactic cancer vaccine candidates. The protection by individual and pool FS antigens were confirmed in both FVB/N-NeuT and BALB-NeuT mice models. We are currently working on further optimization of the immunization to achieve the additive protection by pooling more FS antigen candidates. We presented "Frameshift Peptides as Prophylactic Cancer Vaccines Antigens" at 2012 annual meeting of American Association for Cancer Research at April 2012.

V. Personnel

All personnel who work with animals are required to have animal care training within the last three years. ASU IACUC training modules can be completed at the LATA ASU homepage. Training dates can also be verified by users at this site: **http://balsam.forest.net/latanet/records/asut/search3.htm**

A. List the names, titles, affiliations, and roles of **ALL** persons currently involved in the research or teaching activity.

Name	Title	ASURITE name	Role in Protocol (What procedures will each person be doing?)	Species with which individual will have direct contact ("all" or list species)*	IACUC USE ONLY Training (mm/yy)
Stephen Johnston, Ph.D.	PI, Center Director, CIM		Design experiments	None	4/10 HSQ
Kathryn Sykes, Ph.D.	Adjunct Professor		Design Experiments, interpret data, troubleshoot	None	11/11 HSQ
Christopher Diehnelt, Ph.D.	Res. Asst. Professor		Design Experiments	None	Basic 9/09, Mice 4/10 HSQ
Danielle Lussier	Graduate Student		Immunization, /bleed mice /Euthanasia	Mouse	HSQ 11/11
Andrey Loskutov, Ph.D.	Research Scientist		Immunization, /bleed mice /Euthanasia	Mouse	2/11 HSQ

Luhui Shen	Graduate Student		Immunization, /bleed mice /Euthanasia/Breed Tg mice/tumor cell injection/monitoring	Mouse	11/11 HSQ
Felicia Craciunescu	Researcher		Immunization, /bleed mice /Euthanasia/tumor cell injection/monitoring	Mouse	10/09 HSQ
John Charles Rodenberry	Researcher		Immunization, /bleed mice /Euthanasia/Breed Tg mice/tumor cell injection/monitoring	Mouse	2/11 HSQ
Kurt Whittemore	Graduate Student		Immunization, /bleed mice /Euthanasia/Breed Tg mice/tumor cell injection/monitoring	Mouse	6/12 HSQ
Kari Kotlarczyk	Technician		Immunization, /bleed mice /Euthanasia/Breed Tg mice/tumor cell injection/monitoring	Mouse	7/10 HSQ
Hu Duan	Graduate Student		Immunization, /bleed mice /Euthanasia/Breed Tg mice/tumor cell injection/monitoring	Mouse	6/10 HSQ

B. List the names of any individuals no longer involved with the research (these individuals will be removed from the protocol and DACT will be notified):
Mark Robida, Kristen (Day) Seifert

VI. Certification

By signing this report, I certify that, to the best of my knowledge, the information included herein is accurate and complete. I understand that continued animal use past the scheduled termination date of the protocol requires IACUC approval. I also understand that should the animal use under this protocol require any change from that stated in the protocol, prior approval by the IACUC is required.

Principal Investigator's Signature

6/7/12
Date

FOR IACUC USE ONLY
Annual Review Determination

ANNUAL REVIEW APPROVAL SIGNATURES:

Chair, IACUC — Date: 6-28-12

Attending Veterinarian — Date: 6-28-12

Additional IACUC Member — Date: 6-28-12

Curriculum Vitae

Kurt Whittemore

Home: 920 S Terrace Rd Apt 208., Tempe, Arizona 85281 (480) 559-9781

Work: CIM, The Biodesign Institute 1001 S. McAllister Ave, Tempe, AZ 85287

E-mail: kurtwhittemore@gmail.com

CURRENT POSITION:

Research Associate Student **Center for Innovations in Medicine**

The Biodesign Institute

August 2008 – May 2014

Member of the Biological Design Ph.D. program.
Research Projects/Accomplishments
1) I developed a high through-put platform for screening tumor cDNA libraries with antibody containing sera to discover new immunogens. These immunogens could be used to study cancer further and develop cancer vaccines. I am first author on a paper titled "A microarray method for identifying tumor antigens by screening a tumor cDNA expression library against cancer sera" for this work which will be published in the journal for Human Vaccines and Immunotherapeutics in October 2013.
2) I demonstrated that random sequence peptide mimotopes can be used to capture immune-specific antibodies and predict epitope motifs.
3) I constructed a scFv (single chain fragment variable) DNA library from B cell RNA so that a phage antibody library could later be constructed.
4) I developed the hypothesis that part of the aging process is due to an acquired autoimmune disease against stem cells since the stem cells necessary for repairing damage share some similarities with cancer cells such as increased telomerase expression. This proposal was entered into the Gemini Contest, and was selected for funding. Some initial experiments to test this hypothesis were performed using a chromium release CTL (cytotoxic T lymphocyte) assay. These early results lend some support for the hypothesis.
5) I developed a program using Java, Weka, and the DBSCAN algorithm which could automatically align one block of antibody-peptide feature reactivities with the corresponding features within an array of 48 blocks. The alignment from the program had a correlation coefficient of 0.94 with a human alignment.
6) I developed the idea and method for calculating the entropy of the data acquired from reacting antibody containing sera with an array of non-natural sequence peptides. Preliminary results indicate that there is a statistically significant difference between the entropy values of young and aged sera when reacted with the non-natural sequence peptide array.

EDUCATION:

Ph.D. **Arizona State University**

August 2008 – May 2014

Major: Biological Design

Bachelor of Science **Southern Utah University**

August 2005-May 2008

Major: Chemistry

PROFESSIONAL EXPERIENCE:

Water Lab Analyst Southern Utah University
08/2005-05/2008

August 2005 – May 2008

The water lab at Southern Utah University tests the water of Cedar City Utah for the presence of specific chemicals and contaminants. During my employment there, I used knowledge gained from my Chemistry studies to work as a lab technician in the water lab.

SKILLS AND EXPERTISE:

Molecular biology
PCR
Western blotting
Biochemistry
Gel Electrophoresis
Cell culture

TEACHING EXPERIENCE:

Teaching Assistant for Biological Design Class

Lectured and graded papers

AWARDS AND FELLOWSHIPS:

2011	Gemini Proposal regarding an “Age Associated Stem Cell Autoimmunity” hypothesis selected for funding
2010-2013	ARCS (Achievement Rewards for College Scientists) Scholarship
2008	College of Science Outstanding Scholar at Southern Utah University
2008	Graduated Summa Cum Laude from Southern Utah University
2007	Gold medal at iGEM (international Genetically Engineered Machines) competition
2007	Listed on Dean's List

COMPUTATIONAL SKILLS:

Programming languages: Java and C++

Familiar with: BLAST, WEKA, CLUSTALW, Sequencher

Comfortable working in Windows and Linux

Extra-curricular Activities:

Biodesign Graduate Student Organization	2009-present
Honor Society	2005-2008
Alpha Chi Member	2005-2008
Chemistry Club	2005-2008

PUBLICATIONS:

Whittemore K, Sykes K. A microarray method for identifying tumor antigens by screening a tumor cDNA expression library against cancer sera. Hum Vaccin Immunother 2013; 9.

BIOGRAPHICAL SKETCH

There is an interesting quote from Freeman Dyson: "a new generation of artists will be writing genomes with the fluency that Blake and Byron wrote verses". What if it were possible to engineer biological systems as easily and reliably as computer systems are engineered today, and as a result enjoy much longer healthier lives? With the advent of biological engineering and synthetic biology, this is becoming more of a reality. I have taken several steps to be a part of this exciting new area of science. I obtained by undergraduate degree in Chemistry at Southern Utah University where I also worked at the university Water Lab to test for the presence of harmful chemical compounds. During this time, I started an iGEM (international Genetically Engineered Machines) team to make bacteria with a cyanide biosensor. During my PhD degree I screened a tumor cDNA library for antigens which the immune system would recognize. In another project, I tested my own proposal that part of the aging process may be due to an age associated stem cell autoimmune disease. I also developed the idea that the way a repertoire of antibodies from an organism reacts with an array of non-natural sequence peptides will change during the aging process and during poor health. I calculated the Shannon information entropy of the resulting fluorescence intensity distribution from the non-natural sequence peptide array, and found that the entropy was higher in aged organisms and organisms with cancer compared to healthy individuals. I want to continue to participate in research that will lead to the maintenance of good health at a level that was not possible with previous technology.