UNIVERSITY OF CALIFORNIA

Santa Barbara

Serum antibody repertoire analysis for high-throughput epitope discovery and characterization

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Chemical Engineering

by

Joel David Bozekowski

Committee in charge:

Professor Patrick S. Daugherty, Chair

Professor Michelle O'Malley

Professor Siddharth Dey

Professor Kevin Plaxco

June 2018

The dissertation of Joel David Bozekowski is approved.

Michelle O’Malley

Siddharth Dey

Kevin Plaxco

Patrick S. Daugherty, Committee Chair

June 2018

Serum antibody repertoire analysis for

high-throughput epitope discovery and characterization

by

Joel David Bozekowski

ACKNOWLEDGEMENTS

This dissertation would not have been possible without the help and support of so many people throughout the years. I would like to thank my advisor, Patrick Daugherty, for his mentorship and support throughout my doctoral studies. You believed in me from the beginning, allowed me to hit the ground running, and gave me the independence I needed to succeed. I would like to thank my dissertation committee for the continued advice and support over the years. And thank you to all our collaborators over the years who have helped with this work, including Linc Johnson for his helpful insights on AMD.

To the Daugherty Lab members of the past, thank you all for everything. Thanks to Jack for training me when I first arrived in the lab and for all your help over the years. Thanks to Bob, Jen, Serra, Kelly, Tim, and the rest of the group for all of the help and insightful conversations. And a special thanks to Michael for all of the great times in the lab and office. I’ve really enjoyed growing and learning as a scientist with you, but more importantly growing as a person. I’ll always miss our (many) conversations and I couldn’t have asked for a better friend and colleague during this journey. I would also like to thank Kevin Plaxco and the Plaxco Lab for welcoming Michael and I to the lab and for all their support. You all took us in immediately and made us feel at home. It was great getting to know everyone and being in the office and lab was always a good time.

A huge thanks to all of my friends in Santa Barbara who made this an incredible experience from the very beginning. Thanks to Rich, Dan, Thomas, Sean, John, Zach and all the others for the great memories that I’ll never forget. I can’t think of a better group to experience these last five years with. I’m already excited for the reunions.

Ultimately, none of this would have been possible without my parents, Kim and Cindy. I’m only in this position because of your endless sacrifice and support. You taught me how to value learning and to always push myself, but to have fun while doing it. I can’t thank you enough for all of the opportunities you’ve provided me.

And of course, I would like to thank Shelby. These words won’t do it justice but your endless support, encouragement, and patience have meant so much to me. You always make me strive to be my best. We’ve already been through so many great experiences and there’s nothing I look forward to more than what the future holds for us.

VITA OF JOEL DAVID BOZEKOWSKI
June 2018

EDUCATION

University of California, Santa Barbara 2013–2018
Ph.D. Chemical Engineering

University of Colorado, Boulder 2009–2013
B.S. Chemical and Biological Engineering
summa cum laude

RESEARCH EXPERIENCE

Ph.D. Candidate Jan 2014–Present
Advisor: Patrick S. Daugherty
Department of Chemical Engineering
University of California, Santa Barbara

Undergraduate Research Assistant – Senior Thesis Aug 2012–May 2013
Advisor: Hang (Hubert) Yin
Department of Chemical and Biological Engineering
University of Colorado, Boulder

Undergraduate Research Assistant – REU May 2012–Jul 2012
Advisor: Ashutosh Chilkoti
Biomedical Engineering
Duke University

PUBLICATIONS

- **Bozekowski JD**, Plaxco KW, Daugherty PS. Epitope discovery and mapping from herpes simplex virus using random peptide libraries. (*In Preparation*).
- **Bozekowski JD**, Johnson LV, Daugherty PS. Serum antibody repertoire analysis reveals unique binding signatures associated with risk of age-related macular degeneration. (*In Preparation*).
- **Bozekowski JD***, Paull ML*, Daugherty PS. Mapping antibody binding to epitopes using multiplexed epitope substitution analysis. (*In Preparation*).
  *contributed equally
- **Bozekowski JD**, Graham AJ, Daugherty PS. High-titer antibody depletion enhances discovery of diverse serum antibody specificities. J Immunol Methods. 2018;455:1–9.
- Pantazes RJ, Reifert J, **Bozekowski JD**, Ibsen KN, Murray JA, Daugherty PS. Identification of disease-specific motifs in the antibody specificity repertoire via next-generation sequencing. Sci Rep. 2016;6:30312.

CONFERENCE PRESENTATIONS

- 255th ACS National Meeting, New Orleans, LA. 2018 (*Poster*).
- 17th PepTalk, San Diego, CA. 2018 (*Oral*).
    - Featured Poster Presentation
    - Student Fellowship Award
- 10th UCSB Chemical Engineering Graduate Student Symposium, Santa Barbara, CA. 2017 (*Oral*).
- 7th International Conference on Biomolecular Engineering, San Diego, CA. 2017 (*Poster*).
    - 2nd Place Poster Award
- 9th UCSB Chemical Engineering Graduate Student Symposium, Santa Barbara, CA. 2016 (*Poster*).
- 251st ACS National Meeting, San Diego, CA. 2016 (*Poster*).
- 8th UCSB Chemical Engineering Graduate Student Symposium, Santa Barbara, CA. 2015 (*Poster*).

OUTREACH & MENTORING

- Gorman Scholar Undergraduate Research Mentor	Summer 2016
- Visiting Scholar Research Mentor	Spring 2014
- Undergraduate Research Mentor	2014–2015
- Family Ultimate Science Exploration (FUSE)	Fall 2014
- CU Chemical Engineering Alumni Student Mentoring Program	2014–2015

TECHNIQUES & SKILLS

- Biotechnology – high-throughput screening, bacterial display, directed evolution, flow cytometry, FACS, next-generation sequencing design and application, ELISA
- Molecular Biology – PCR, enzymatic digestions, molecular cloning, library design and construction
- Cell Culture – mammalian cell culture (Jurkat T cell, MCF-7, U937, HEK-293), lentiviral transfection
- Bioinformatics – MATLAB, statistical testing, nonparametric testing, machine learning (SVM, PCA), next-generation sequencing data analysis
- Chemical Engineering – unit operations, process design, core ChE curriculum
- Facility Management – UCSB flow cytometry shared facility manager, Daugherty Lab safety and equipment manager

ABSTRACT

Serum antibody repertoire analysis for

high-throughput epitope discovery and characterization

by

Joel David Bozekowski

The serum antibody repertoire is a unique repository of information regarding past immune encounters. Antibodies are produced in response to exposures and bind specifically to their target antigens. Antibodies associated with infection and disease can serve as diagnostic biomarkers upon detection. However, many disease-specific antibodies and their targets remain undiscovered. Therefore, to discover antibody biomarkers, reagents must be developed to specifically bind the disease-specific antibodies. Peptides are suitable reagents as they can often mimic native antibody epitopes with high affinity and specificity. Moreover, the sequences of antibody-binding peptides can be determined and used to identify the original antibody targets, revealing previously unknown antigens that contribute to disease progression and creating new opportunities for therapeutic development. Here, we developed and applied high-throughput methods to discover and characterize peptide epitopes from immune-related diseases.

A large bacterial display peptide library composed of randomized peptides was constructed to screen against human serum specimens and discover peptides that bind antibodies. For each specimen, millions of antibody-binding peptide sequences were determined using next-generation sequencing and analyzed computationally to reveal disease-specific binding motifs. This screening methodology was first employed to discover and characterize epitopes from two highly similar viruses, herpes simplex virus type 1 (HSV-1) and type 2 (HSV-2). Following antibody repertoire analysis, we discovered HSV type-specific motifs that could be used as diagnostic classifiers to achieve 100% diagnostic accuracy when distinguishing HSV-1 from HSV-2 and vice versa. Furthermore, numerous type-specific motifs were mapped to HSV antigens using protein sequence database alignments, including known antigens such as glycoproteins G and D as well as previously unreported antigens. We then applied this screening methodology to age-related macular degeneration (AMD). We identified a large panel of antibody-binding motifs associated with the onset of advanced AMD and developed a classifier with 84% accuracy when distinguishing specimens at high risk of advanced AMD from those at low risk. However, as the risk of developing advanced AMD increased, the classifier performance worsened suggesting a unique antibody signature associated with AMD progression.

Additional methods were developed to improve the discovery and characterization of epitopes from the antibody repertoire. We developed a method to selectively deplete highly abundant antibodies from a specimen to reduce repertoire complexity, improve the limits of detection, and enable the discovery of rare antibodies and their targets. Additionally, we developed a novel computational algorithm to rapidly determine binding motifs for epitopes by utilizing the full extent of next-generation sequencing datasets. Ultimately, these tools for analyzing the antibody repertoire at great depth can be applied to a broad range of infectious, autoimmune, and allergic diseases. The discovery of disease-associated epitopes will create new opportunities for the development of diagnostics, vaccines, and therapeutics.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 Introduction

1.1 Motivation & background

1.1.1 Serum antibody repertoire

The serum antibody repertoire is a collection of immunoglobulin (Ig) protein molecules, known as antibodies, circulating the human body surveilling for foreign target molecules. These target molecules, termed antigens, are often proteins from viruses or bacteria but can also be molecules from the body mistaken as foreign, leading to autoimmunity. Antibodies bind antigens to neutralize and signal for destruction and removal. B cells, a type of white blood cell, constantly produce unique antibody variants in anticipation of possible encounters with diverse antigens. Antibodies are key aspects of adaptive immunity, responsible for mounting an immune response to foreign encounters and establishing immunological memory. When antibodies recognize target antigens for the first time, the strength of binding (affinity) is relatively low but increases through a process of antibody mutation and selection, known as affinity maturation [1]. The antibodies are then produced for extended periods of time, possibly lifelong [2], to circulate the body and function as a component of immunological memory. Upon additional encounters, these antibodies can facilitate an enhanced, targeted response to quickly thwart infection and disease, the mechanism harnessed for vaccines [3]. Over time, the body amasses a set of serum antibodies with high affinity and high specificity towards previously encountered antigens.

Due to immunological memory, the antibody repertoire serves as a unique repository of information regarding immune responses such as infections, inflammation, allergies, and autoimmunity. Because highly specific serum antibodies are produced long after initial exposure, these antibodies are biomarkers that can be detected to inform of previous and current immune encounters. However, the complexity and diversity of the antibody repertoire presents unique challenges for discovering and detecting antibodies. It is estimated that the antibody repertoire contains upwards of 10^6 unique antibody species [4]. While the total concentration of antibodies in serum is homeostatically controlled and is thus relatively stable over time [5], the concentrations of individual antibody species can span orders of magnitudes and are subject to fluctuations due to acute immune responses such as infections and vaccines [6,7]. Additional complexities arise due to the production of different antibody classes (isotypes) that are specialized for various immune responses. For instance, the isotype IgG makes up ~75% of serum antibodies, IgA is primarily located in mucosal tissues, and IgM is expressed during the initial phase of an immune response [8]. The vast diversity of the antibody repertoire therefore presents countless opportunities to detect antibody biomarkers associated with diseases, but significant challenges exist for biomarker discovery and reliable detection.

1.1.2 Antibody biomarkers for diagnostics

Substantial progress has been made for discovering antibody biomarkers and developing antibody-based diagnostic tests. The serum antibody repertoire is easily accessed from blood samples by relatively inexpensive and non-invasive procedures in comparison to other diagnostic tests such as positron emission tomography (PET), magnetic resonance imaging (MRI), lumbar punctures, and tissue biopsies. Additionally, blood/serum is fairly stable and suitable for transport and storage in various forms such as dried blood spots, expanding the options for antibody diagnostics in nonclinical settings [9,10]. Antibody-based diagnostics rely on determining the presence of specific antibodies. This is often accomplished by detecting and measuring antibody binding to known antigens using assays such as immunoblotting, immunofluorescence, and enzyme-linked immunosorbent assays (ELISA). Every diagnostic test has a diagnostic accuracy, which is often reported as the sensitivity and specificity. Sensitivity measures the proportion of positive (disease) classifications correctly identified while specificity measures the proportion of negative (control) classifications correctly identified. Many factors impact the sensitivity and specificity including the type of assay, the antibody biomarkers, and the reagents used for detection such as the antigens and fluorescent probes. While the objective is to have high sensitivity and specificity, there is often a trade-off between the two statistics. To avoid false positive diagnoses, sensitivity is often sacrificed for higher specificity.

Many antibody biomarkers have been identified and utilized for diagnostics in a variety of infectious and autoimmune diseases (**Table 1.1**). Many diseases have multiple antibody biomarkers that can be used for diagnosis, such as rheumatoid arthritis [11], type 1 diabetes [12,13], and celiac disease [14]. In rheumatoid arthritis, antibodies that bind to human IgG-Fc (rheumatoid factor) or cyclic citrullinated peptides can be used as biomarkers. However, rheumatoid factor antibodies may also be present in other rheumatic diseases such as systemic lupus erythematosus and even healthy subjects, leading to a lower diagnostic specificity [15]. The accuracy of antibody-based diagnostics can often be improved by targeting specific antibody isotypes. As celiac disease primarily affects the intestinal tissues where IgA is predominant, an IgA anti-tissue transglutaminase biomarker has been identified with high sensitivity (98%) and specificity (98%) [14]. For infectious diseases, antibody biomarkers can indicate past and present infections. However, early diagnosis is crucial for facilitating treatments and therapies and thus IgM antibodies can be targeted at the early infection stages [16–18]. While many diagnostic tests focus on individual antigens for detecting antibody biomarkers, advances in multiplexed diagnostics for detecting numerous biomarkers will enable higher diagnostic accuracy and personalized testing [19]. Despite the success of many antibody biomarkers, many antigens for diseases remain unknown. As these diagnostics require knowledge of antigens associated with disease, there is an increasing need to develop technologies capable of efficiently mining the serum antibody repertoire for unbiased biomarker discovery.

Table 1.1. Serum antibody biomarker diagnostics.

Disease	Antigen	Isotype	Sensitivity (%)	Specificity (%)	
Infectious					
Herpes Simplex Virus 1	Glycoprotein G-1	IgG	99	95	[20]
Herpes Simplex Virus 2	Glycoprotein G-2	IgG	97	89	[21]
Lyme disease	VlsE C6				[22]
Early localized		IgG	69	99	
Early disseminated		IgG	89	99	
Late disseminated		IgG	98	96	
Autoimmune					
Graves' disease	Thyrotropin receptor	IgG	98	99	[23]
Celiac disease					[14]
	Deamidated gliadin	IgA	88	95	
		IgG	80	98	
	Tissue transglutaminase	IgA	98	98	
		IgG	70	95	
Scleroderma	Topoisomerase 1	IgG	43	100	[24]
Rheumatoid arthritis	IgG-Fc	IgM	69	85	[11]
	Cyclic citrullinated peptide	IgG	67	95	[11]
Type 1 diabetes	Insulin	IgG	54	98	[12]
	Glutamic acid decarboxylase	IgG	82	96	[13]
	Islet antigen-2	IgG	70	99	[13]

Sensitivity = true positive rate; specificity = true negative rate

1.1.3 Peptide epitopes

To enable antibody biomarker discovery and detection, it is useful to identify and characterize the specific region of the antigen that an antibody binds, the epitope. For protein antigens, epitopes are classified as continuous (linear) or discontinuous (structural). Linear epitopes are typically peptides with six to nine amino acids [25], while discontinuous epitopes consist of amino acid residues that are distant in the protein sequence but brought together near the surface and thus depend on the three-dimensional protein structure [26]. While the majority of epitopes are technically discontinuous [27], these epitopes often contain stretches of short linear epitopes [25]. An analysis of antibody-antigen structures from the Protein Data Bank found that more than 85% of the epitopes contained a contiguous stretch of at least five residues [28]. It has also been shown that for many structural epitopes, only three to five residues contribute significantly to antibody binding [29]. Thus, antibody epitopes can often be mimicked by short linear peptides.

The utilization of antibody-binding peptides as diagnostic reagents has numerous advantages. Antigens are often large proteins that contain numerous epitopes and regions susceptible to cross-reactive antibody binding [26], leading to diagnostic false positives. Peptides that mimic epitopes approximate the smallest region responsible for antibody binding, reducing cross-reactivity while maintaining high affinity and specificity. Peptides are also simple, easily produced, and cost-effective, while antigens often require complex production, folding, and purification. Diverse peptides can thus be efficiently screened for antibody binding using high-throughput technologies to discover and characterize epitopes.

In addition to identifying diagnostic reagents for detecting antibody biomarkers, peptide reagents can be used to identify unknown antigens associated with diseases. By determining the epitope sequence, numerous methods can be used to identify the antigen containing the epitope. Antibody-binding peptides can be mapped to antigens using computational epitope mapping algorithms such as EpiSearch [30] and Pepsurf [31], functional binding and inhibition assays [32], or proteome alignments using protein sequence databases [33]. The identification of antigens responsible for antibody production and reactivity can provide important insights into disease pathology and treatments. The identification of gluten antigens responsible for celiac disease allows patients to adopt a gluten-free diet to alleviate symptoms [14]. Autoantibodies in Graves' disease bind and stimulate the thyrotropin receptor leading to hyperthyroidism, which can be treated with antithyroid drugs [34]. Epitopes associated with diseases can also lead to the development of novel therapeutics. The discovery of a complement factor H epitope has led to the development of a therapeutic antibody for early-stage non-small cell lung cancer (NSCLC) entering clinical trials [35]. Additionally, the discovery and characterization of immunodominant epitopes in infectious diseases can aid in designing improved vaccines [36]. As antigens involved in a variety of diseases are still unknown, it is necessary to develop broad, high-throughput technologies to identify disease-specific epitopes and antigens.

1.2 Screening peptide libraries for antibody binding

To effectively probe the antibody repertoire for biomarkers, a vast set of binding targets needs to be utilized. Peptides are suitable targets because diverse ensembles of peptides (libraries) can be readily produced and analyzed. By mimicking epitopes, these peptide libraries can be screened for antibody binding in a high-throughput manner. To accomplish this, numerous platforms have been developed to identify antibody-binding peptides including microarrays [37], phage display [38], and bacterial display [39]. Often, these platforms have been utilized for focused discoveries based on known or suspected antigens. Targeted peptide libraries are composed of overlapping peptides (tiling) that span the entire sequence of a single antigen or a set of antigens. This peptide tiling has been completed for whole proteomes including the human proteome with T7-Pep [40], a phage display library composed of over 400,000 peptides representing all open reading frames in the human genome. T7-Pep screening has been applied to autoimmune diseases such as multiple sclerosis, narcolepsy, rheumatoid arthritis, and type 1 diabetes to discover previously reported and potentially novel autoantigens [41]. A large library (>10^8 peptides) was also constructed with overlapping peptides spanning the proteomes of nearly all viruses that infect humans [42]. This analysis led to the discovery of many novel epitopes and additional insights into viral prevalence.

While targeted peptide screening approaches have led to many interesting discoveries for epitopes and antigens, the focus on specific antigens and proteomes limits the discovery process. Targeted approaches require hypotheses about the candidate antigens and overlook unsuspecting antigens. Moreover, requiring the peptides to represent exact sequences derived from target proteomes is restrictive, as peptide binders with high affinity and specificity can deviate from the cognate epitope sequence. It has also been shown that peptides can bind to antibodies with little to no sequence resemblance to the cognate epitope [43–45]. Therefore, to maximize the scope of peptide discovery, a vast set of peptides with randomized sequences can be screened, termed a random peptide library. Because of the combinatorics of random peptide sequences, an immense number of epitopes can be mimicked providing a relatively unbiased platform for epitope discovery. These random peptide libraries can then be interrogated for antibody binding using high-throughput analyses such as next-generation sequencing (NGS) [46]. Here, a review of random peptide library screening using microarray, phage display, and bacterial display platforms is presented (**Table 1.2**). While additional platforms such as ribosome display [44] and yeast display [47] have been used for random peptide screening, these methods are more often used for single-chain antibody screening and discovery applications [48,49] and are not discussed here.

Table 1.2. Summary of platforms used for random peptide library screening.

Peptide Screening Platform	Peptide Diversity	Advantages	Disadvantages
Microarray	10^4–10^6	Reproducibility Quality control	Low diversity Cost
Phage display	10^9–10^{11}	High diversity Cost Widely used	Low avidity Technical experiments Time consuming
Bacterial display	10^9–10^{11}	High diversity High avidity Cost Flow cytometry	Technical experiments Time consuming

1.2.1 Peptide microarrays

Peptide libraries can be spotted on a solid substrate, often a glass slide, in a two-dimensional array, termed peptide microarrays [37]. The peptides are then interrogated for antibody binding from biological samples such as serum. Antibody binding is detected by utilizing fluorescent secondary probes followed by image analysis to quantify binding. Random peptide microarrays have been used to identify unique antibody-binding signatures associated with various immunological diseases, a method termed immunosignaturing [50]. Immunosignaturing has been applied to vaccines [51], infectious diseases [37,50,52], and cancers [37,53]. For cancer, immunosignatures were identified for six different cancer types utilizing microarrays composed of 330,000 unique peptides [37]. The immunosignatures that resulted from antibody binding to 50 unique peptides for each cancer type achieved 100% diagnostic accuracy using a cross-validation classification approach. For infectious diseases, immunosignatures have been identified for pathogens including *Trypanosoma cruzi*, hepatitis B, hepatitis C, and West Nile virus, demonstrating high diagnostic accuracies and the ability to elucidate disease subclasses [52].

A main advantage for using peptide microarrays to interrogate the antibody repertoire is the ability to control fabrication. Microarray production and synthesis is controlled by highly refined fabrication equipment and thus has high reproducibility and quality control. When spotting and printing peptides, these methods also have the ability to finely control the peptide density, which can be adjusted to alter the avidity and improve antibody detection capabilities [54]. Another advantage of peptide microarrays is that quantitative binding information is obtained for all peptides [55]. Methods such as phage and bacterial display select for peptides based on a certain binding stringency and no quantitative information pertaining to weak binders is obtained. This has ramifications for large libraries when it is unknown if all peptides are being sampled for antibody binding. However, peptide microarrays are limited to diversities in the range of 10^4–10^6 which restricts peptide discovery and leads to challenges for antigen identification. Microarrays with 330,000 unique peptide sequences only sample 83% of all possible tetramers and 27% of pentamers [56]. Thus, these libraries are missing many peptides with sequence information related to cognate antigens, making it very difficult to connect the antibody-binding peptides to antigens. Microarrays also have higher costs relative to other screening methods such as phage and bacterial display that depend on biological growth which ultimately reduces experimental costs.

1.2.2 Phage display

For the screening method phage display, bacteriophage viruses are genetically engineered to display peptides on their surface [38]. Phage display peptide libraries are produced in *Escherichia coli* (*E. coli)* cultures, purified, and screened for antibody binding using a selection process known as panning. Panning requires incubation of the phage display library with antibodies bound to a plate well or magnetic beads, followed by removal of unbound phages via washing. Successive panning rounds can be completed to further enrich peptide libraries. To complete additional panning rounds, the phage library must be infected into a bacterial culture to amplify the phage particles. After the final panning round, the DNA encoding peptide binders from the phage library is extracted and sequenced to reveal peptide sequences. Phage panning has been utilized extensively over the years for identifying and characterizing epitopes [38,43]. Recently, NGS has been integrated with random phage display libraries and used for massively parallel DNA sequencing to deeply characterize antibody-binding peptides, a process referred to as deep panning [57]. Deep panning has been used to discover epitopes from HIV [57], dengue virus [58], and cancers [59]. A similar phage display panning procedure was utilized to identify common epitopes from an individual antibody repertoire [60], but this also required numerous follow-up experiments such as peptide microarray analysis and exhaustive peptide mutagenesis.

Phage display is the most commonly used surface display technology. Phage display libraries can be readily purchased from companies such as New England BioLabs, enabling easy integration into laboratories and companies. Random phage display libraries can have large library diversities (10^9–10^{11}) due to high transformation efficiencies during library construction in *E. coli* [38,57,61], leading to improved epitope discovery and enabling antigen identification. Numerous phage display systems have been developed utilizing different types of bacteriophages and displaying peptides on various phage coat proteins. Most phage display systems utilize filamentous phages (M13) [38]. These are non-lytic phage, simplifying phage amplification as new phage particles are secreted during *E. coli* growth. Peptides are often engineered to be displayed on the N-terminus of the minor coat protein 3 (pIII) at the tip of the filamentous phage. However, only five copies of pIII are displayed on the phage surface and therefore the peptides have a low surface density and low avidity. Avidity is a measure of the net affinity for interactions between an antibody and multiple epitopes. A high concentration of epitopes on a surface increases the net binding strength for antibody interactions and increases the likelihood of selecting the corresponding peptides during screening. For complex systems such as serum antibody repertoires and random peptide libraries, higher avidity is beneficial for binding dilute antibodies or antibodies with low affinity. Therefore, systems with high avidity are more likely to select for peptides that bind rare antibodies associated with diseases. Other phage coat proteins with higher surface density have been engineered for peptide display such as the major coat protein pVIII with 2,700 copies on the surface [38]. However, inserting random peptides into pVIII can cause problems with capsid assembly. Strategies to increase assembly stability have been attempted by interspersing wild-type pVIII proteins with pVIII proteins containing peptide inserts, resulting in 10-100 pVIII stable surface copies [57]. However, the peptide density remains relatively low and difficult to adjust. Another disadvantage of phage display is the panning procedure requires time-consuming experimental screening rounds and amplification stages. Furthermore, the amplification stages of phage libraries in *E. coli* have been demonstrated to decrease library diversity over multiple rounds due to competition between phages [62]. Thus, peptide binders may be lost during the panning process due to the requirement of consecutive infection and amplification stages. However, fewer panning and amplification rounds are required when utilizing NGS and thus these difficulties may be reduced in the future.

1.2.3 Bacterial display

Bacterial display is a screening platform that utilizes bacteria engineered to display peptides on the cell surface. While bacterial species from the *Staphylococcus* genus have been utilized for display systems [63], *E. coli* is an ideal species to use due to its ubiquity in biotechnology and the accompanying wealth of knowledge and opportunities for bioengineering. To facilitate peptide screening using *E. coli* bacterial display, a display scaffold was engineered by circularly permuting an outer membrane protein, allowing both termini to be accessible on the cell surface [64]. Additionally, this display scaffold was evolutionarily optimized for improved surface localization and screening efficiency [65]. The enhanced circularly permuted outer membrane protein X (eCPX) can then be engineered to contain a peptide at the N-terminus or C-terminus for screening applications (**Figure 1.1**).

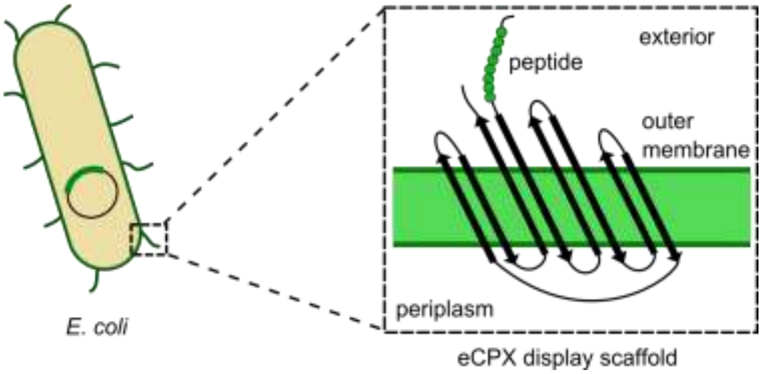


Figure 1.1 Bacterial display eCPX scaffold for peptide screening. *E. coli* cells are genetically engineered with plasmid DNA encoding an outer membrane protein engineered for efficient peptide surface display (eCPX). For bacterial display peptide libraries, each cell displays a unique peptide that can be sorted for antibody binding using a variety of techniques such as magnetic selection or fluorescence-activated cell sorting (FACS). The plasmid DNA can then be extracted and sequenced for peptide sequence analysis and epitope discovery.

Bacterial display is well-suited for peptide screening applications for its ease of genetic manipulation, fast growth rates, and high transformation efficiencies [65,66]. Bacterial display peptide libraries have high diversities (10^9–10^{11}) enabling efficient sampling of potential epitopes from the antibody repertoire. These large library diversities are more stable compared to phage display because of the tightly regulated growth of *E. coli* [67,68]. Another advantage of bacterial display is the ability to utilize flow cytometry for cell population analysis and fluorescence-activated cell sorting (FACS) [39]. The ability to sort cells based on fluorescent binding enables precise selection of cell populations with desired peptide binders and directed evolution of peptides [69–71]. Bacterial library cells can also be efficiently sorted using magnetic selection, a process that uses magnetic beads functionalized with antibody capture reagents to perform a bulk selection of all cells bound by serum antibodies [39].

The eCPX bacterial display system also allows for high avidity antibody binding [39,66]. The eCPX protein is displayed as ~10,000 copies on the bacterial cell surface. Compared to many phage display systems, this increased surface concentration can greatly impact antibody repertoire analysis due to the increased peptide concentration and decreased off-rates for antibody binding, leading to retention and selection of rare antibody-peptide interactions. Other *E. coli* display systems have been developed such as the FliTrx system [66,72], however this system has low avidity and requires a panning procedure similar to phage display rather than magnetic selection or FACS [66]. Additionally, the FliTrx peptides are presented as constrained insertions in exposed loops in the scaffold protein, restricting epitope discovery and leading to poor affinities in solution [64]. Therefore, the eCPX system is an optimal bacterial display system for antibody repertoire analysis.

By screening bacterial display random peptide libraries with FACS, epitopes have been discovered and characterized for various diseases. Peptides with high diagnostic accuracy were identified for celiac disease [70,73,74]. Moreover, preliminary epitopes were further expanded using iterative FACS screening rounds to identify additional epitope sequence information, ultimately enabling the identification of gluten antigens associated with celiac disease [70]. A similar procedure was applied to discover epitopes associated with pre-eclampsia, identifying an epitope with increased binding in pre-eclampsia patients and elucidating a potential mechanism of molecular mimicry between the Epstein–Barr nuclear antigen 1 (EBNA-1) and a G protein-coupled receptor [75].

While bacterial display and FACS has been successful for discovery applications, these methods required numerous and time-consuming experimental screening rounds to reduce the peptide library diversity to a manageable set of peptides capable of low-throughput Sanger sequencing. Additionally, serum specimens were often pooled before analysis to reduce screening rounds needed to identify consensus epitopes. With the advent of NGS, a high-throughput approach is now possible for in-depth characterization of peptide binding to the serum antibody repertoire of individual subjects. To accomplish this, magnetic selection was applied using magnetic beads functionalized with antibody capture reagents to perform a bulk selection of all cells bound by serum antibodies. Analyzing serum specimens from subjects with celiac disease and healthy controls with this next-generation approach revealed gluten associated epitopes with high diagnostic accuracy after only two rounds of magnetic selection, NGS, and computational sequence analysis [76]. The ability to massively catalog antibody-binding peptides for individual specimens coupled with advances in bioinformatics will greatly improve antibody repertoire analysis. A detailed overview of the next-generation methods using bacterial display to identify antibody-binding peptides and characterize epitopes is discussed in **Section 1.4**.

1.3 Infections & diseases

A main advantage of random peptide library screening is the arbitrary applicability to any disease eliciting unique antibody production. Here, we investigated the serum antibody repertoires for diverse systems including the infectious disease herpes simplex and the inflammatory disease age-related macular degeneration (AMD).

1.3.1 Herpes simplex virus

The herpesvirus family contains eight viral species that commonly infect humans [77]. Two of these species are the herpes simplex virus type 1 (HSV-1) and type 2 (HSV-2). The World Health Organization (WHO) estimated that in 2012, the worldwide prevalence of HSV-1 in people under the age of 50 was 67% [20]. The worldwide prevalence of HSV-2 in people aged 15-49 was estimated to be 11% [21]. HSV-1 infection can lead to orofacial herpes resulting in lesions on the face or mouth (cold sores). Infection of HSV-2 can lead to genital herpes resulting in lesions in the genital area, a sexually transmitted disease. Following primary infection of HSV, the viruses infect ganglion neurons and become latent where they remain as lifelong infections. Recurrent infections (reactivation) can occur due to provocative stimuli such as injury, stress, or immunosuppression. However, most HSV infections are asymptomatic and the host remains unaware [78]. HSV is primarily transmitted by physical contact with bodily fluids or lesions of an infected individual but transmission can occur during asymptomatic recurrences.

HSV-1 and HSV-2 are closely related viruses and thus have many physical similarities. HSVs are enveloped, double-stranded DNA viruses with fast replication cycles. The HSV genome contains over 70 genes, all of which are present in both HSV-1 and HSV-2 [79,80]. The HSV structure consists of a protein capsid enclosing the viral DNA, a tegument protein layer surrounding the capsid, all wrapped in a lipid membrane envelope with protruding glycoproteins [81]. The HSVs have at least a dozen different glycoproteins, five of which are involved in viral entry and infection of host cells [82]. Due to the surface exposure and critical role in viral replication, the glycoproteins are convenient antigenic targets for antibodies to neutralize and protect against infection. Detection of these antibodies in serum can therefore inform of HSV infection.

To develop diagnostics for HSV-1 and HSV-2, highly sensitive and specific epitopes for each HSV species need to be identified to enable detection of serum antibodies in infected patients. However, due to the high sequence similarity between the HSV types, few antigenic targets exist that differentiate the two viruses. Numerous studies have therefore been completed to identify epitopes specific to each HSV species, including peptide tiling and epitope mapping of glycoproteins B [83,84], D [85,86], and H/L [87]. Recent studies have utilized microarrays for epitope mapping seven glycoproteins from each HSV species [82]. While type-specific epitopes have been identified using these methods, the most effective epitope for diagnosis has been glycoprotein G. Glycoprotein G has the lowest sequence homology of the glycoproteins due to a large truncation in HSV-1 [80,88] and is therefore an ideal candidate for distinguishing the HSVs. Various targeted peptide screening methods were utilized to identify distinct, type-specific epitopes [89–91]. Glycoprotein G epitope discoveries have enabled the development of antibody-based diagnostics including immunoblots and ELISAs with high diagnostic utility [92,93].

Despite the advances in antibody-based HSV diagnostics, significant limitations exist with these methods. In particular, antibody-based methods are often dependent on the stage of infection and careful assessments need be completed when classifying tests as positive or negative [94]. Therefore, older methods are still widely used for diagnosing active infections, including viral culturing and PCR of viral DNA from a blood/tissue sample [95,96]. These methods are low-cost and fairly reliable, but are still limited by sensitivity and specificity issues and require clinicians and laboratory facilities [92,94]. Given the increasing prevalence of HSV worldwide and the availability of effective antiviral therapy, there is an increasing need to discover and develop highly sensitive and specific reagents to enable point-of-care HSV diagnosis.

1.3.2 Age-related macular degeneration

Age-related macular degeneration (AMD) is a leading cause of blindness in the elderly worldwide [97]. AMD results from the deterioration of the macula, a small region of the retina responsible for sharp, central vision. The degeneration of the macula can lead to a blurry, dark, and distorted central field of view which severely impacts performing simple tasks and quality of life. A major risk factor for AMD is simply age, as the disease is most likely to occur after age 60 [98]. Other risk factors include race, as AMD is more prevalent in Caucasians than other races [99], environmental factors such as smoking and nutrition [100], and numerous genetic factors including polymorphisms in immune-related proteins like various complement factors [101,102]. According to the AMD Alliance International, 40 million individuals worldwide will suffer visual impairment due to AMD by 2020, resulting in $300 billion annual direct costs [103]. As the population of older individuals grows rapidly and life expectancy increases, AMD will become increasingly prevalent in the coming decades [104].

AMD progresses slowly over time in stages designated as early, intermediate, and late. A hallmark of AMD is the formation of extracellular deposits of cellular debris in the retina, known as drusen, that accumulate with age [105,106]. Early AMD is first diagnosed by the presence of medium-sized drusen seen at the back of the eye during a dilated eye exam. Intermediate AMD is characterized by large-sized drusen and/or retinal pigment abnormalities, but vision loss is not common at this stage. Advanced, or late, AMD involves extensive drusen and/or deterioration of the macula and often some degree of vision loss. It is unknown if drusen are a symptom or cause of the disease but an increase in size and number of drusen increases the risk of late AMD [107]. Specifically, the presence of large drusen increases the likelihood of developing late AMD over a five year span by a factor of five. It is postulated that the formation of drusen can stimulate a chronic immune response, exacerbating macular degeneration [105,108], signifying the role drusen could play in AMD progression. Importantly, the presence of drusen does not ensure the development of late

AMD, as only 20% of individuals with early AMD progressed to late AMD over a five year span [97].

There are two different types of late AMD: geographic atrophy (GA) and neovascular (NV). GA is typically characterized by degeneration of the retinal pigment epithelium (RPE) and outer retina while NV is caused by choroidal vascularization under the RPE and is typically associated with the more severe cases of vision loss [98]. GA often develops gradually, while the progression of NV, and therefore the damage, can be rapid. Much work has been devoted to determining ways of slowing or impeding disease progression, including the identification of nutritional supplements that reduce the risk of developing late AMD by 25% over five years [109]. With millions at risk for AMD, this convenient and low cost solution for decreasing risk has enormous potential. Various therapies have been developed to impede vision loss in NV AMD by targeting and inhibiting angiogenesis. NV therapies include photodynamic therapy [110], laser surgery [111], and anti-vascular endothelial growth factor (VEGF) therapies using antibodies (Avastin, Lucentin) [112,113] and aptamers (Macugen) [114,115]. There are currently no therapies for GA AMD although there are many in clinical development to address this unmet need [116]. Additionally, there are no therapies to reverse damage from AMD but there is promising work being done in phase 1 clinical trials using human embryonic stem cell–derived RPE monolayers [117,118].

Fundamental questions remain regarding the complex progression from early and intermediate stages to late AMD, along with the bifurcation of GA and NV AMD. It is known that immune system activation plays a significant role in the development and progression of AMD, but the exact origins and targets of the immune response are still being investigated [98,119]. Compelling evidence has implicated the complement system as a key factor in AMD pathology. Specifically, a single nucleotide polymorphism in complement factor H can account for up to 50% of AMD cases [97,101]. Additionally, numerous complement proteins have been identified as elevated components in drusen using proteomics [106,120]. Targeting complement activation to suppress inflammation has recently been investigated with numerous therapies in clinical development, further demonstrating the role of the complement pathways in AMD pathogenesis [121,122].

Antibodies associated with AMD have also been observed, suggesting a link between the adaptive immune response and AMD. Antibodies targeting carboxyethylpyrrole (CEP) protein adducts have been identified, often generated due to oxidative stress involved with aging [123]. Anti-retinal antibodies have also been identified in AMD [124,125], and unique antibody profiles have been observed at different stages of progression [126]. Despite these discoveries, the roles of antibodies in AMD pathology are still unclear and the exact targets remain undiscovered.

While progress has been made on determining the role of immune activation in AMD, limitations in diagnostics and therapies are largely due to the complexities of the pathology and a lack of understanding the etiology. The detection of antibody biomarkers associated with AMD provides an opportunity for AMD prognostics and diagnostics. However, based on the diversity of antibody targets observed in AMD, it is unlikely that a single antibody species can effectively diagnose the disease [127]. Therefore, a technology such as random peptide library screening could be utilized to discover numerous antibodies associated with AMD and reveal their unique binding targets. These discoveries could lead to the development of antibody-based tests for early diagnosis, the identification of novel therapeutic targets, and insights into disease etiology.

1.4 Techniques

With the recent advances in NGS, the techniques used here focus on screening random bacterial display peptide libraries for antibody binding to individual serum specimens. Enrichment is completed using magnetic selection followed by DNA extraction and sequencing using NGS platforms. A suite of computational algorithms is then applied to mine large peptide sequence datasets for significant binding patterns and motifs. This process enables a deep characterization of binding interactions from the serum antibody repertoire.

1.4.1 Magnetic cell selection

Magnetic cell selection utilizes superparamagnetic microbeads functionalized with capture reagents for separating cells of interest. When screening bacterial display peptide libraries for antibody binding, the magnetic beads are coated with protein A/G, a recombinant fusion protein that binds to antibodies. Protein A/G combines six immunoglobulin binding domains derived from protein A, found in *Staphylococcus aureus*, and protein G, found in *Streptococcus* species [128,129]. Protein A/G binds to the fragment crystallizable (Fc) region of antibodies and therefore does not interfere with epitope binding. In human serum, protein A/G preferentially binds to IgG although weak binding towards IgA and IgM occurs. Protein A/G also has strong binding for mouse and rabbit IgG and can therefore be used in assays with many commercially available monoclonal antibodies [129].

When screening bacterial display peptide libraries for antibody binding, magnetic selection enables the separation of library cells bound by human serum antibodies (**Figure 1.2**). This screening methodology utilizes a library of *E. coli* cells, each genetically encoded to display a randomized peptide on the N-terminus of the outer membrane protein eCPX. Prior to screening, serum antibodies that bind to *E. coli* displaying eCPX without an Nterminal peptide must be depleted by incubating the serum specimen with *E. coli* and separating cells bound by antibodies from unbound antibodies using centrifugation. Depleted serum can then be incubated with the random bacterial display peptide library to facilitate antibody-peptide binding. Antibodies that do not bind to displayed peptides are washed away and protein A/G magnetic beads are added to facilitate the formation of a cell-antibody-bead ternary complex. A magnet can then be used to physically separate the cells bound by serum antibodies from unbound cells. This process can be repeated for further enrichment of the peptide library towards patient-specific serum antibodies. The resulting library cells contain the genetic information encoding millions of peptides that bound to patient-specific antibodies, which can be accessed and analyzed using NGS and bioinformatics.

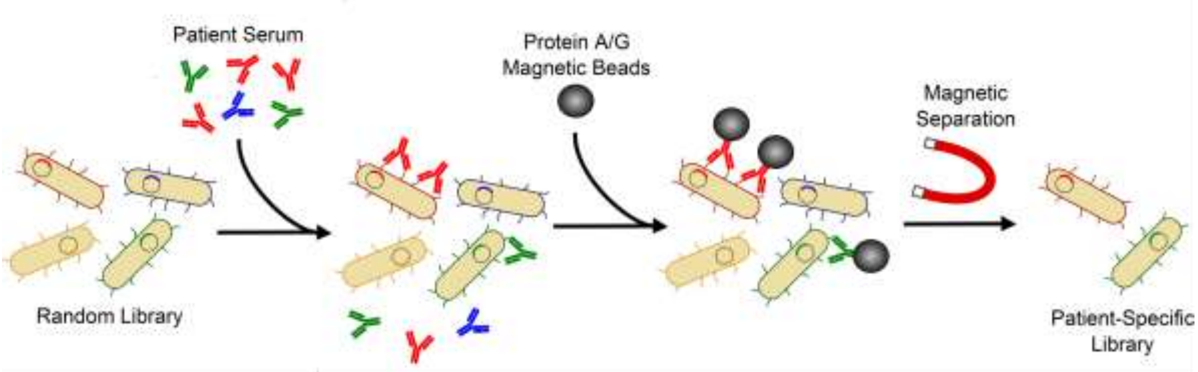


Figure 1.2. Bacterial display peptide library sorting using magnetic selection. A random bacterial display peptide library is incubated with a patient serum specimen to facilitate antibody-peptide binding. Magnetic selection using magnetic particles coated with protein A/G is then applied to physically separate cells bound by antibodies, resulting in a patient-specific library of cells encoding antibody-binding peptides.

1.4.2 Flow cytometry

Flow cytometry can be utilized for a quantitative, real-time analysis of bacterial display peptide libraries. Using a flow cytometer, a suspension of bacterial display cells is hydrodynamically focused into a stream of single cells that are individually interrogated by one or more lasers. Upon interrogation, the light scatter and fluorescence of a single cell are measured and converted to an electronic signal for immediate analysis. Light scattered in a forward direction is used to measure particle size while light scattered at an angle is related to the granularity and internal complexity of the cell. After a random bacterial display peptide library has been screened for antibody binding, the library enrichment can be quantified using flow cytometry. The enriched library is incubated with the corresponding serum specimen to facilitate antibody binding followed by addition of a secondary fluorescent probe that binds to human antibodies, such as phycoerythrin (PE)-conjugated anti-human IgG antibodies. These antibodies bind specifically to the Fc region of human IgG antibodies and have a PE fluorescent protein tag to indicate the presence of antibodies bound to a cell. Flow cytometry can then be used to quantify the proportion of library cells bound by antibodies. Flow cytometry can also be used to precisely sort cells based on fluorescent signal, a method termed fluorescence-activated cell sorting (FACS). FACS is not used here but could be used in future studies to further elucidate epitope sequence information for various disease applications.

1.4.3 Next-generation sequencing

Next-generation sequencing (NGS) can be utilized with bacterial display peptide screening to determine the sequences for millions of antibody-binding peptides [76]. Here, the Illumina NextSeq platform was exclusively used to perform massively parallelized sequencing [130]. To determine the peptide sequences following bacterial display screening, the DNA plasmids encoding display peptides are extracted from the library cells. Upon plasmid recovery, the peptide sequence regions are amplified using PCR. During PCR, various adapter sequences are added to the DNA construct (amplicon) flanking the peptide sequence, which are required for the Illumina platform. These adapter sequences are used to hybridize the amplicons to the flow cell used for sequencing. Additionally, a unique DNA "barcode" sequence is added to the library amplicons as an identifier for patient-specific libraries, enabling pooling of multiple sample libraries for multiplexed sequencing.

A high-output sequencing run on the Illumina NextSeq can achieve over 400 million sequences per run. A control genome (PhiX) must be spiked into each sequencing run to increase sequence diversity for imaging and to serve as a quality control. PhiX is often spiked into these NextSeq runs at 40–50% to provide sufficient sequence diversity, leaving over 200 million sequence reads available for library sequences. Therefore, libraries from 40 specimens can be sequenced to obtain ~5 million sequence reads per sample. This provides sufficient coverage of the peptide library sequences to characterize antibody repertoire binding.

1.4.4 Antibody-binding motif discovery and analysis

The size and complexity of the peptide sequence datasets obtained from random peptide library screening requires the use of computational algorithms to identify and characterize epitopes. The antibody-binding peptide sequences contain short patterns of amino acids that are critical for antibody binding, known as motifs. These motifs can be contiguous amino acid sequences or a series of amino acids interspersed with undefined residues indicating which positions were conserved for antibody binding and which positions were allowed variation. Numerous computational tools have been developed to analyze peptide sequence datasets for binding motifs. MEME is a motif discovery tool that utilizes an expectation maximization algorithm to identify motifs within a set of peptide sequences [131,132]. MEME is a convenient motif discovery platform with an online platform. However, analysis of 5,000 sequences for MEME motif discovery requires ~10 hours. Because NGS datasets often contain millions of peptide sequences, MEME can only be used on small subsets of sequences, thus narrowing the discovery space.

To efficiently analyze large NGS datasets, the algorithm IMUNE was developed [76]. IMUNE includes a suite of computational tools including tools for processing NGS data files and translating the DNA sequences encoding peptides. More importantly, IMUNE performs comprehensive motif searches within the peptide datasets to identify motifs that were statistically enriched in libraries from disease specimens and not enriched in control specimens. This enables the unbiased discovery of disease-specific motifs.

The vast diversity and complexity of the peptide datasets following bacterial display screening and NGS lead to many redundant binding motifs. To reduce redundancy and identify concise motifs, various clustering algorithms can be used to cluster similar motifs and form consensus motifs. The similarity between two motifs can be quantified by aligning the motif sequences and generating a similarity score based on the PAM30 substitution matrix [76,133]. Given a set of motifs, all motif similarity alignments can be calculated and hierarchical clustering can be applied to group motifs based on sequence similarity. The peptide sequences containing the clustered motifs can then be input into MEME to determine the consensus motif.

1.4.5 Disease classification algorithms

The motifs and peptides discovered through random peptide screening can be used as diagnostic reagents for detecting antibody biomarkers in diseases. It has been demonstrated that individual peptides can yield high diagnostic accuracy for autoimmune diseases such as celiac disease [70]. Peptide reactivity to patient sera can be quantified using various methods such as ELISAs, microarrays, and flow cytometry with bacterial display. Patient samples can then be classified as disease or control based on the reactivity towards the diagnostic peptide reagent. Additionally, the enrichment of a motif in a patient-specific peptide library screened for antibody binding can be used as a quantitative metric for classification [76]. However, an individual motif is often insufficient for accurate diagnosis and more sophisticated classification schemes must be utilized to classify a specimen as disease or control. In these cases, reactivity towards numerous peptides or motifs can be used a diagnostic reagents. Algorithms such as support vector machines (SVM) [134] are then applied to classify specimens based on reactivity to the various peptides [73]. An SVM projects the multidimensional data of motif reactivities from groups of specimens into higher dimensions and constructs a hyperplane to optimally separate the disease specimens from the control specimens. The SVM is trained on a discovery cohort of specimens to construct the classifier hyperplane and tested on an independent cohort using the same parameters for classification.

1.4.6 Protein database searches

Following antibody-binding motif discovery, online protein databases can be searched for protein antigens containing motifs of interest. Various protein sequence databases exist including the NCBI protein Basic Local Alignment Search Tool (BLASTp) [135] and ExPASy ScanProsite [136]. Search tools align binding motifs to proteins based on sequence similarity to identify candidate antigens containing the motifs. Sequence alignment similarity is often scored using substitution matrices such as PAM and BLOSUM [133]. Candidate searches can remain unbiased by searching all non-redundant proteins or searches can be targeted by only searching specific proteomes or antigens. These protein alignments can be conducted rapidly for mapping motifs to protein sequences. Other epitope mapping strategies require detailed protein structure data which are unavailable for the overwhelming majority of possible antigens. However, protein sequence database alignments will be unsuccessful for discontinuous epitopes.

A consequence of relying solely on sequence alignments is that sufficient motif or epitope sequence information needs to be determined. For short motifs and epitopes, protein alignments result in too many irrelevant, false positive matches. As a heuristic, the motifs and epitopes must contain at least seven amino acids with >70% sequence resemblance to the native antigen for successful alignments [33]. To determine statistical significance of protein sequence searches in BLASTp, a significance factor (E-value) is reported for each alignment. The E-value is an expected value that estimates the number of alignment matches one can expect to observe by random chance when searching a database of a certain size. The E-value decreases exponentially as the alignment score increases, dependent on factors such as the database size, motif length, and motif accuracy.

2 Bacterial display peptide library construction, characterization, and application for antibody repertoire analysis

Biological systems engineered to display ligands such as phage display and bacterial display have been utilized for applications ranging from discovering protease substrates to developing diagnostic reagents. Recently, random peptide libraries have been utilized to profile the antibody repertoire using these display systems and identify antibody-binding motifs. However, there are limitations to utilizing these biological systems including amino acid frequency biases and stop codon usage which can alter peptide discovery. Here, we constructed a high-quality random bacterial display peptide library. The library (trimer-X_{12}) contains ~10 billion unique peptides with restricted codon usage optimized for random sequences and *Escherichia coli* growth. Next-generation sequencing was utilized to sequence a portion of the naive trimer-X_{12} library and generate statistics to evaluate the quality. We determined that less than 4% of the trimer-X_{12} sequences contained a stop codon, a greater than 10-fold improvement compared to a previously constructed random display library. Additionally, the range of amino acid frequencies was only 5.5% in trimer-X_{12} compared to an 11% range in the previously constructed library. The trimer-X_{12} library was then screened for antibody-binding peptides against serum specimens. We determined that the amino acid frequencies in libraries following selection were highly dependent on the naive library and therefore biases in the naive library propagated forward through the screening process. With the improved trimer-X_{12} constructed and evaluated, we performed various experiments to benchmark the high-throughput screening method for analyzing the antibody repertoire. Additionally, by analyzing fundamental characteristics of antibody-peptide binding, we have gained important insights into how to efficiently and accurately discover motifs, exemplified by discovering monoclonal antibody binding motifs. These results demonstrate that the high-quality trimer-X_{12} library can be readily applied to diverse systems such as infectious, inflammatory, and autoimmune diseases to develop diagnostics, reveal antigens associated with disease, and lead to antibody-based therapeutics.

2.1 Introduction

Random peptide library screening has been successfully utilized for a wide range of applications such as determining protease substrates [69,71], discovering peptide reagents for diagnosing diseases [52,73], and identifying antigens associated with diseases [56,137]. Due to the combinatorial nature of random peptide libraries, the peptides mimic a vast set of naturally occurring epitopes. This unbiased scope can be useful for probing the antibody repertoire for binding motifs due to the large diversity of antibody species and their corresponding targets [4,5]. While there are applications that are useful for targeted approaches such as screening peptides tiled from the proteomes of viruses [42], parasites [138], and humans [40], random peptide libraries avoid the need to bias experiments toward suspected targets. For instance, a random peptide library can be used to identify diverse targets such as dietary antigens in celiac disease [70], viral epitopes associated with pre-eclampsia [75], and antibody signatures in various types of cancers [139].

Random peptide libraries can be utilized with various platforms including synthetic systems such as microarrays [37] as well as biological systems such as phage display [57] and bacterial display [66]. While microarrays have precise control over the peptide synthesis and therefore control over the initial library composition, they are limited by peptide diversities ranging from 10^4–10^6 as well as fabrication and assembly costs [55,140]. Biological systems such as phage display and bacterial display are capable of having much higher diversities of ~10^{11} [141] but must rely on biological processes which have inherent limitations and biases [142]. One such limitation is with the use of codons, sequences of three nucleotides (guanine (G), cytosine (C), adenosine (A), or thymine (T)) that encode for a specific amino acid. The genetic code is redundant due to the existence of 64 codons but only 20 naturally occurring amino acids. A consequence of this redundancy is that the amino acids have varying frequencies which leads to a usage bias. Another limitation of using biological systems for random library construction is the use of stop codons. Stop codons signal termination during protein translation and therefore sequences containing stop codons correspond to non-functional sequences in the library. There are three stop codons in the genetic code. A high frequency of stop codons in the initial, naive library leads to inefficiencies in the screening process as the library diversity is effectively reduced and functional sequences are diluted.

To maximize the scope of random peptide libraries and optimize the discovery of diverse targets, the quality of the naive library is crucial [143]. A classic strategy to circumvent the biases and deficiencies of codon usage is to utilize NNS codons, where N refers to any of the four nucleotides and S refers to the nucleotides C and G [61,144]. In this approach, oligonucleotides are synthesized by sequential addition of single nucleotides. The use of NNS codons reduces the number of possible codons to 32 in a way that reduces the bias of amino acid frequencies while maintaining representation of all 20 amino acids. Additionally, it reduces the number of stop codons from three to one. However, significant bias still exists with several amino acid frequencies three times higher than others, as well as the inclusion of one stop codon. Therefore, to improve screening applications for serum antibody profiling, we sought to construct a new random peptide library with reduced frequency bias and stop codon usage.

To construct an improved random library using the bacterial display system, we utilized a random library of oligonucleotides synthesized using trimer phosphoramidites [145–147]. The oligonucleotides were synthesized by the sequential addition of trimer nucleotides, three nucleotides pre-synthesized in the form of NNN, where N can be any of the four nucleotides but the combination of possible codons is limited to 20 codons corresponding to the naturally occurring amino acids. This is a recently developed alternative approach to synthesize randomized oligonucleotides while avoiding the redundancy, frameshift mutations, and stop codons accompanied by using sequential nucleotide addition. Using trimer phosphoramidites, we constructed a 12-mer random bacterial display library (trimer-X_{12}). With next-generation sequencing (NGS) affordable and readily available, we sequenced the naive library to determine baseline statistics and evaluate stop codon usage and amino acid frequency bias. For a comparison to the trimer-X_{12} library, we analyzed a 15-mer display library previously constructed in the Daugherty Group at UCSB using NNS codons, which has been utilized extensively for various peptide screening applications [70,73–75,148]. We demonstrated that the trimer-X_{12} library had reduced amino acid frequency bias and significantly lower stop codon usage. We utilized this new library to evaluate the screening platform using magnetic selection and NGS for identifying antibody-binding motifs. Following peptide selection, the amino acid frequencies were highly dependent on the naive library frequencies. This indicates that codon bias is propagated through the screening process which could skew motif discovery, demonstrating the importance of a minimally biased naive library. Finally, we used the trimer-X_{12} library to characterize monoclonal antibody (mAbs) motifs and determined general motif characteristics for serum antibody binding. These results demonstrate that the trimer-X_{12} library is high quality and can be utilized for a wide range of applications.

2.2 Results

2.2.1 Random bacterial display peptide library construction

To construct a random bacterial display peptide library with reduced codon bias and minimal stop codons, we obtained oligonucleotides for 12-mer library construction that were synthesized using 20 trimer phosphoramidites optimized for *Escherichia coli* (*E. coli)* codon usage (**Table 2.1**). The 12-mer library (trimer-X_{12}) was designed and constructed to display on the N-terminus of the enhanced circularly permuted outer membrane protein X (eCPX). The trimer-X_{12} oligonucleotide library was ligated into the eCPX scaffold plasmid vector and transformed into *E. coli*. The diversity of the library was estimated to be 8×10^9 unique transformants based on the colony-forming units (cfu) measured following serial dilutions of transformed cultures. A 15-mer eCPX display library previously constructed in the Daugherty Group at UCSB using NNS codons (NNS-X_{15}) was analyzed for comparison. The NNS-X_{15} library had a similar diversity estimated to also be 8×10^9.

Table 2.1. Trimer nucleotides for random libraries optimized for *E. coli.*

Trimer Nucleotide	Amino Acid Encoded
TAC	Tyr (Y)
TCT	Ser (S)
GGT	Gly (G)
CGT	Arg (R)
GCG	Ala (A)
GAT	Asp (D)
AAC	Asn (N)
CAG	Gln (Q)
GAA	Glu (E)
CAT	His (H)
ATC	Ile (I)
CTG	Leu (L)
AAA	Lys (K)
ATG	Met (M)
TTC	Phe (F)
CCA	Pro (P)
ACC	Thr (T)
TGG	Trp (W)
GTT	Val (V)
TGC	Cys (C)

2.2.2 Next-generation sequencing of naive peptide libraries

To generate statistics for each random display library and evaluate the compositions, NGS was performed on the random (naive) libraries. Sequencing was completed using the Illumina NextSeq 500 platform and sequencing and data processing statistics for both libraries were generated (**Table 2.2**). NGS identified 1.5×10^8 and 8.8×10^7 sequences for NNS-X_{15} and trimer-X_{12}, respectively, containing the correct barcode indices and primer annealing regions. The variation in the number of sequences obtained was solely a result of variation in the sequencing process such as the amount of library DNA loaded onto the flow cell and does not reflect the library quality. While these sequences amount to only 1-2% of the estimated library diversities, these proportions are sufficient to garner useful statistics about the naive libraries.

Table 2.2. Next-generation sequencing statistics for naive peptide libraries.

Naive Library	Sequences Obtained	Correct Length (%)	Unique Sequences (%)	Stop Codon Sequences (%)	Functional Unique Sequences
NNS-X_{15}	1.5×10^8	92	82	43	6.2×10^7
Trimer-X_{12}	8.8×10^7	95	94	3.5	7.3×10^7

For both libraries, >90% of the sequences were the correct base pair length. Additionally, >80% of the sequences for both libraries were observed only once in the datasets (unique), suggesting there was minimal overrepresentation of individual sequences which is expected if the estimated diversity is sufficiently large. The discrepancy in the percent unique sequences observed between the libraries is likely due to the discrepancy in sequences obtained and therefore we cannot conclude anything regarding differences in the actual diversities. The most drastic difference between the two libraries was the presence of stop codons, as 43% of sequences in NNS-X_{15} contained at least one stop codon compared to only 3.5% in trimer-X_{12}, consistent with theoretical usages (**Figure 2.1**). While the trimer-X_{12} library should theoretically have zero stop codons, we did observe a small yet finite percentage of stop codon sequences likely due to errors during oligonucleotide synthesis, PCR, and sequencing. However, the construction of the random peptide library using trimer nucleotides ultimately resulted in ~10-fold reduction in stop codon usage compared to NNS.

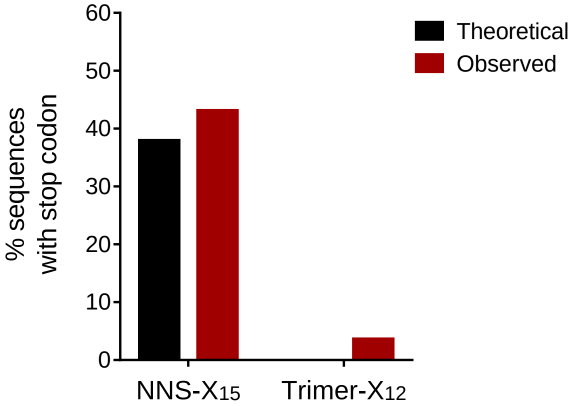


Figure 2.1. Stop codon usage was significantly reduced in the trimer-X_{12} library. Next-generation sequencing was completed for the naive NNS-X_{15} and trimer-X_{12} libraries. The stop codon frequency was identified and compared to theoretical frequencies. For NNS-X_{15}, the theoretical usage is based on the expected frequency of at least one stop codon appearing in a 15-mer peptide using NNS codons. For trimer-X_{12}, the theoretical frequency is zero due to the exclusion of stop codons using trimer nucleotide synthesis. As expected, trimer-X_{12} had significantly reduced stop codon frequencies.

After removing all sequences with stop codons, there were 6.2×10^7 and 7.3×10^7 functional sequences identified for NNS-X_{15} and trimer-X_{12}, respectively, from which the amino acid frequencies were evaluated for bias and skew. The NNS-X_{15} frequencies were first compared to the theoretical frequencies based on NNS expected usages (**Figure 2.2A**). For NNS codons, the theoretical frequencies have a standard deviation of 2.4% and a range of 6.3%. For the NNS-X_{15} library we observed a similar distribution with a standard deviation of 3.2% but a larger range of 11%. The trimer-X_{12} exhibited less frequency bias with a standard deviation of 1.4% and a range of only 5.5%, in agreement with the uniform theoretical usage of 5% for all amino acids (**Figure 2.2B**). Although alanine (A) and threonine (T) frequencies were notably lower than expected at 2.1% and 2.5%, respectively, the trimer-X_{12} library demonstrated reduced frequency bias as only 30% of amino acids were outside of a 4–6% usage range, while 80% of NNS-X_{15} amino acids had frequencies outside of this range. The utilization of trimer nucleotides therefore produced a more uniform frequency distribution in the naive library, providing a less biased platform for screening applications.

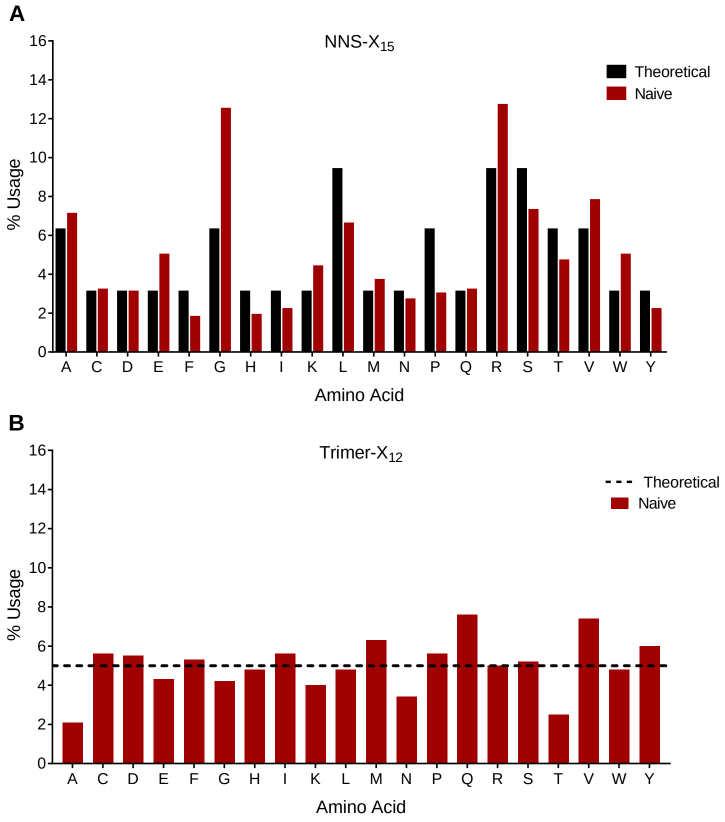


Figure 2.2. Amino acid frequencies in naive libraries were identified using NGS. Next-generation sequencing was completed for the naive NNS-X_{15} and trimer-X_{12} libraries to identify amino acid usages. **(A)** For NNS-X_{15}, the observed frequencies were calculated and compared to the theoretical frequencies based on the expected NNS codon usages. **(B)** For trimer-X_{12}, the observed frequencies were compared to the theoretical frequency of 5% for all amino acids due to the equimolar synthesis using trimer nucleotides. Observed frequencies generally followed expected trends with greater variance in NNS-X_{15} and a more uniform distribution in trimer-X_{12}.

2.2.3 Screening random peptide libraries for antibody-binding peptides

After the construction and analysis of the trimer-X_{12} random library, antibody repertoires were characterized by screening the library against human serum specimens. The random library was screened against individual serum specimens resulting in subject-specific peptide libraries enriched for antibody-binding. Library enrichment was completed using two rounds of magnetic selection, reducing the peptide diversity to ~10^6 antibody-binding peptides. Library enrichment was analyzed using flow cytometry with fluorescent probes to quantify the fraction of *E. coli* cells displaying peptides bound by serum antibodies (**Figure 2.3**). A typical magnetic selection using our standardized protocol, described in **Section 2.4.2**, often achieves >50% enrichment in the first round of selection followed by near-complete enrichment (>95%) after the second selection round. This screening method can then be completed in parallel for many serum specimens and the resulting enriched libraries can be sequenced using multiplexed NGS [76].

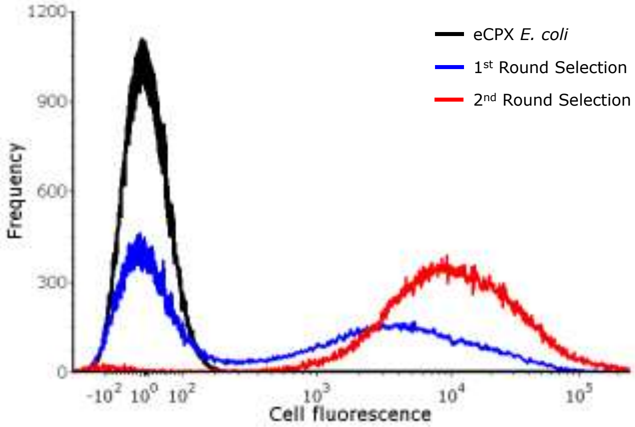


Figure 2.3. Peptide library enrichment was quantified via flow cytometry. The trimer-X_{12} library was enriched for peptides that bound antibodies from a serum specimen using a first round of magnetic selection (blue) followed by a final round of magnetic selection (red). An eCPX *E. coli* population was analyzed as a negative control of unbound cells (black) and used as a background threshold to calculate library enrichment. Enrichment was quantified using flow cytometry with fluorescent anti-human IgG probes to identify cells bound by antibodies. After two rounds of magnetic selection, the peptide library was completely enriched (>95%) for antibody binding.

2.2.4 Amino acid frequencies in libraries following peptide selection

To evaluate the impact that the naive library composition has on peptide selection, we analyzed the peptide sequences obtained from screening random libraries against human serum specimens. The NNS-X_{15} and trimer-X_{12} libraries were each screened against 30 serum specimens to isolate antibody-binding peptides and evaluate the amino acid frequencies in the enriched peptide libraries. For libraries enriched from NNS-X_{15}, the average amino acid frequencies did not deviate greatly from the naive library, with an average absolute difference of 0.6% (**Figure 2.4A**). The exception was arginine (R) which had an average decrease in frequency of 5.2% following selection. This could be due to the abnormally high frequency of 13% in the naive library and possible epitope sequence bias. For libraries enriched using trimer-X_{12}, the average amino acid frequencies again did not deviate from the naive library with an average absolute difference of 0.7% (**Figure 2.4B**). However, the largest absolute difference observed was only a 1.8% decrease for cysteine (C). Therefore, the naive library frequencies were largely maintained following selection for antibody-binding peptides, indicating that selected peptide compositions were skewed based on the initial library screened. These results demonstrate the importance of having an unbiased and a uniform library for random library screening applications.

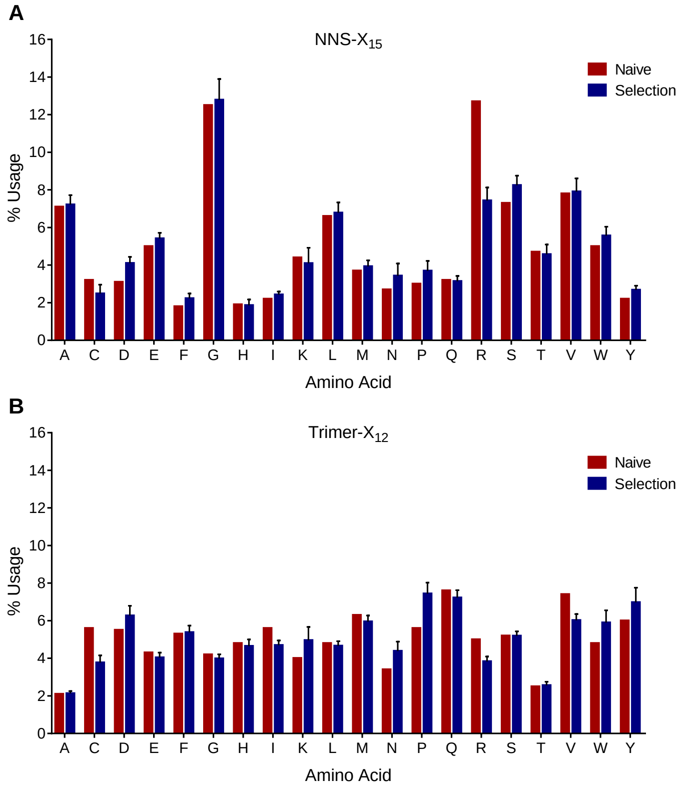


Figure 2.4. Amino acid frequencies in peptide libraries enriched for antibody binding were highly dependent on naive library. Antibody-binding peptide selection was completed for 30 human serum specimens using **(A)** NNS-X_{15} and **(B)** trimer-X_{12} libraries. Next-generation sequencing was completed to identify amino acid frequencies after selection compared to naive library frequencies. Amino acid usages did not largely deviate from the naive frequencies, demonstrating the importance of a uniform naive library to avoid bias during selection.

2.2.5 Antibody-binding motif analysis

After screening numerous serum specimens, we analyzed the resulting peptide libraries to obtain general motif characteristics for antibodies that bound linear peptides. Specifically, we were interested in determining the average motif width as well as the average number of conserved positions present in each motif. We analyzed libraries from 20 serum specimens using both NNS-X_{15} and trimer-X_{12} and identified the 10 highest enriched motifs from each library using the motif search algorithm MEME. Of these motifs, 90% were five to eight amino acids wide with four to six conserved positions (**Figure 2.5**). Because the majority of motifs were at most eight amino acids wide, there were no discernable differences in motifs identified using random libraries with either 12-mers or 15-mers. These results indicate that many serum antibodies have similar motif preferences when binding linear peptides. We can now target our discovery efforts based on this description of common motif profiles for a more efficient and precise discovery platform.

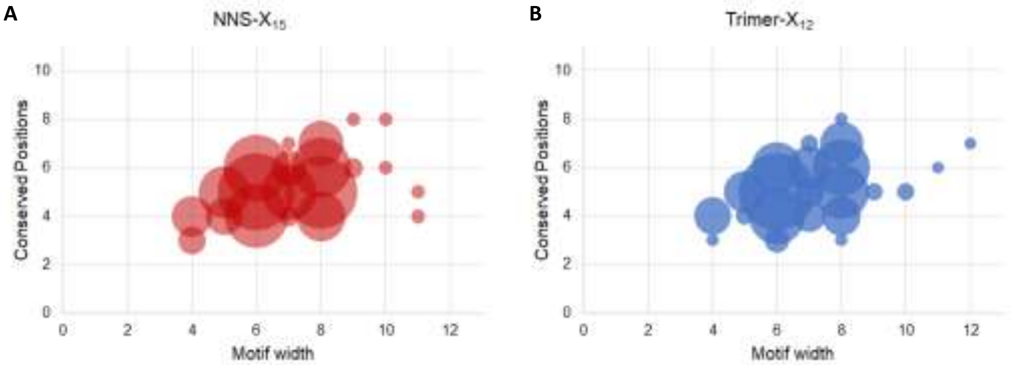


Figure 2.5. Antibody-binding motifs discovered using random peptide libraries have a distinct distribution for widths and numbers of conserved residues. Antibody-binding motifs were discovered from peptide libraries enriched for antibody binding, using both NNS-X_{15} and trimer-X_{12} libraries. The 1,000 highest enriched sequences from 20 serum specimens were used to discover motifs using MEME. The 10 most significant motifs for each library were analyzed for motif widths and numbers of conserved positions. Therefore, a motif characteristic profile was generated for 200 independent motifs from both NNS-X_{15} and trimer-X_{12} derived libraries. Almost all motifs were five to eight amino acids wide with four to six conserved positions. The distribution of widths and numbers of conserved residues using the 12-mer library did not differ from the 15-mer library.

2.2.6 Monoclonal antibody epitope analysis

To validate antibody detection and discovery with random peptide library screening, we utilized mAbs with known linear epitopes as model systems. Three mAbs were analyzed for motif discovery: anti-myc with epitope EQKLISEEDL, anti-V5 with epitope GKPIPNPLLGLDST, and anti-HA with epitope YPYDVPDYA. The trimer-X_{12} library was screened against each mAb and the enriched libraries were sequenced with NGS. To identify consensus binding motifs, 5,000 of the highest enriched peptides from each library were analyzed for motifs using MEME. Of the 5,000 sequences analyzed from each mAb library, 80% contributed to the myc motif LxSEE[DE], 79% contributed to the V5 motif [KR]PxPNxxL, and 87% contributed to the HA motif [YQ]PYD[VI]xD. The mAb binding motifs were displayed graphically as sequence logo plots (**Figure 2.6A-C**). The total height of letters (representing amino acids) at a position indicates the degree of conservation and therefore the importance of that position for binding. The height of individual letters reflects the frequency of specific amino acids at each position. While the exact linear epitopes for each mAb were previously known, we were able to precisely determine which residues were important for binding and the selectivity at each position. The high degrees of sequence convergence and consistency of motifs with the known linear epitopes indicate that these motifs are representative of the conserved mAb binding residues. Furthermore, we determined the enrichment of each motif in the corresponding library and found each motif to be enriched >1,000-fold (**Figure 2.6D**). By studying mAbs as model systems, we successfully integrated random library screening with NGS to discover precise binding motifs and quantify the corresponding enrichment.

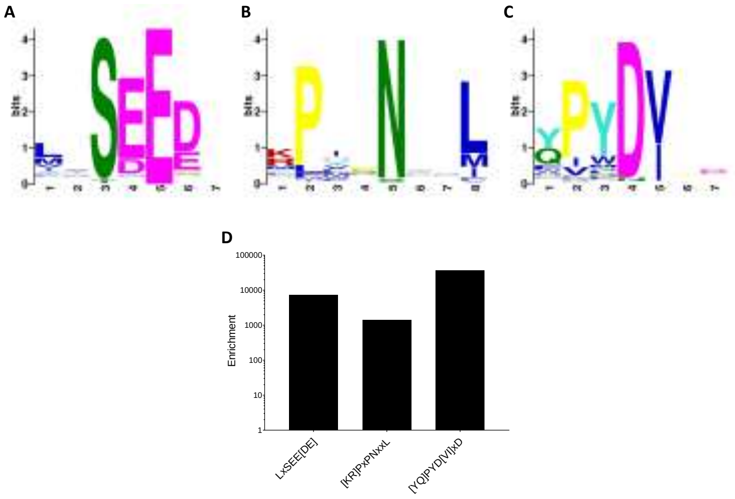


Figure 2.6. Monoclonal antibody motifs were identified through peptide library screening and NGS. Binding motifs for three mAbs (anti-myc, anti-V5, and anti-HA) were discovered by screening trimer-X_{12} libraries followed by NGS. MEME was utilized to discover consensus binding motifs for each mAb and sequence logos were identified: **(A)** myc motif LxSEE[DE] from epitope EQKLISEEDL, **(B)** V5 motif [KR]PxPNxxL from epitope GKPIPNPLLGLDST, and **(C)** HA motif [YQ]PYD[VI]xD from epitope YPYDVPDYA. Enrichments of mAb motifs following selection were calculated and determined to be greater than 1,000 for each mAb motif demonstrating the success of high-throughput screening for binding motifs.

2.3 Discussion

While utilizing biological systems for random peptide display libraries has many advantages, inherent biases and limitations exist that have been difficult to circumvent, such as codon bias and stop codon usage [143,149]. However, due to advances in oligonucleotide synthesis, it is now possible to design and construct improved libraries that circumvent these limitations. Moreover, the advent of high-throughput sequencing has enabled an in-depth characterization of libraries not previously possible. Here, we constructed a random bacterial display peptide library (trimer-X_{12}) using oligonucleotides synthesized with trimer phosphoramidites. This enabled the library to be restricted to codons representing the 20 natural amino acids with minimal overrepresentation and optimized for *E. coli* growth. Utilizing NGS, we obtained ~70 million sequences from the naive library and assessed the quality by generating statistics such as the sequence lengths, stop codon usage, and amino acid frequencies. More than 95% of the sequences identified were the correct length and >89% were unique with only ~4% stop codons, indicating a highly diverse and functional library. By comparison, we sequenced a 15-mer library constructed using NNS codons (NNS-X_{15}) and found that the library contained 43% stop codon sequences. Experimentally, this severely limits the sampling of random peptides because almost every other cell is not displaying a peptide for interrogation. A 10-fold reduction in stop codon usage is therefore a significant improvement to expand the functionality and true diversity of the naive library. In addition, we determined that the trimer-X_{12} library had a more uniform distribution of amino acid frequencies with a range of only 5.5% compared to a range of 11% in NNS-X_{15}, demonstrating the decreased variance. These results signify a high-quality naive library that minimizes biases that were once inherent to biological systems such as bacterial display.

With the trimer-X_{12} library constructed, we utilized a high-throughput protocol to screen for antibody-binding peptides using magnetic selection and NGS. After screening 30 serum specimens for both trimer-X_{12} and NNS-X_{15}, we determined that the amino acid frequencies following selection were highly dependent on the naive library frequencies. The average frequency of glycine (G) following selection was 4.0% using the trimer-X_{12} library but it was 13% using NNS-X_{15}. This large variation in amino acid frequencies due simply to the naive library will negatively impact peptide screening and motif discovery by biasing the motif sequences. Specifically, when quantifying motif enrichment (**Equation (**2.1)), the frequencies of amino acids in the selected library are used to calculate the expected observations. Frequencies that are skewed by the naive library will therefore lead to skewed enrichment calculations. Additionally, when searching protein databases such as BLASTp for proteins that contain the motifs identified during screening, it is critical to have accurately determined as many motif positions as possible to reduce false positive matches [33]. Thus, the trimer-X_{12} library provides an improved platform for discovering antibody-binding motifs and epitopes with minimal amino acid bias.

After determining the quality of the trimer-X_{12} library, we performed high-throughput experiments aimed at evaluating the fundamentals of antibody repertoire profiling. While linear peptides cannot mimic all antibody-antigen interactions [29], an analysis of antibody-antigen structures in the protein data bank found that more than 85% of the epitopes contained a contiguous stretch of at least five residues [28]. By analyzing hundreds of motifs from 20 serum specimens, we determined that the large majority of motifs had a distinct range of widths and numbers of conserved residues, consistent with previous findings. Knowing these ranges enables a targeted approach to motif discovery by focusing discovery on short patterns of amino acids enriched in the peptide sequences, ranging from five to eight amino acids wide. Computationally, this greatly improves the efficiency of searching through the millions of peptide sequences. While we only utilized the 10 most enriched motifs from each specimen library, we can readily identify hundreds of motifs suggesting a broad coverage of many antibody interactions. The precision of discovery was exemplified by analyzing model systems using three mAbs, discovering the distinct binding specificities, and quantifying the enrichment for each motif.

In the past, identifying distinct antibody-binding motifs required laborious rounds of experimental screening using flow cytometry [70,75], constraining throughput and discovery. With the availability and affordability of NGS, the focus has shifted towards fewer experimental stages and more reliance on computational analysis. However, this leaves less room for error in the few experimental stages used, emphasizing the importance of using high quality libraries for screening applications. With an improved random peptide display library and better understanding of the fundamentals of antibody repertoire profiling, we are better suited to apply this technology to more complicated and diverse immunological systems such as infectious and autoimmune diseases for diagnostics, antigen discovery, and therapeutic development.

2.4 Materials and methods

2.4.1 Construction of bacterial display peptide libraries

To construct a random 12-mer peptide library on the N-terminus of the outer membrane protein scaffold eCPX, *E. coli* strain MC1061 (F^- araD139 Δ(ara-leu)7696 galE15 galK16 Δ(lac)X74 rpsL (Str^R) hsdR2 (r_K^- m_K^+) mcrA mcrB1) [67] was used with surface display vector pB33eCPX [65]. Library construction was adapted from a protocol previously described [39]. Forward oligonucleotide primers were designed to contain the trimer-X_{12} random library and anneal upstream of eCPX and extend to a SfiI site:

ACTTCCGTAGCTGGCCAGTCTGGCCAGGGTGGA-NNN-NNN-NNN-NNN-NNN-NNN-NNN-NNN-NNN-NNN-NNN-NNN-GGAGGGCAGTCTGGGCAGTCTG

where -NNN- represents one of 20 codons optimized for *E. coli* expression (**Table 2.1**) (Metkinen Chemistry). The reverse primer was designed to anneal downstream of a SfiI site on the C-terminal end of eCPX: GGCTGAAAATCTTCTCTC (Eurofins MWG Operon).

To prepare the insert DNA for library construction, PCR amplification was accomplished using a KAPA HiFi HotStart ReadyMix PCR Kit (Kapa Biosystems). Touchdown PCR was used to increase specificity (20 cycles with annealing at 67 °C and reduced 0.5 °C/cycle followed by 20 cycles with annealing at 57 °C). The PCR products were purified and concentrated with a PCR purification kit (GeneJET) and double digested with SfiI in CutSmart buffer (10 U/μg insert DNA) (NEB) for 4 hrs at 50 °C. The digested insert product (~550 bp) was extracted using a gel DNA recovery kit (Zymo Research). To prepare vector DNA, pB33eCPX vector was isolated using a plasmid maxiprep kit (Qiagen) and double digested with SfiI in CutSmart buffer (10 U/μg vector DNA) (NEB) for 4 hrs at 50 °C. The large product band (~5,400 bp) was extracted using a gel DNA recovery kit (Zymo Research). The extracted vector product was dephosphorylated with Antarctic Phosphatase (5 U/μg vector DNA) (NEB) for 1 hr at 37 °C followed by heat inactivation for 5 min at 70 °C. Insert and vector products were ligated using T4 DNA Ligase (1 U/50 ng vector) (Invitrogen) overnight at 14 °C for 16 hrs followed by heat inactivation for 10 min at 70 °C. The ligation product was concentrated and purified by DNA precipitation and desalted via drop dialysis. Transformation of the ligated product was completed by electroporation (1.8 kV, 50 uF, and 100 Ω) (Bio-Rad Gene Pulser II) in batches of 1 μg vector added to 70 μL electrocompetent MC1061 cells in 1 mm electroporation cuvettes (Fisher). Transformation batches were recovered in 1.75 mL of SOB media (BD Difco) supplemented with 15% (v/v) glycerol for 1 hr at 37 °C with 250 rpm shaking. If batch electroporation time constants were consistent, the transformations were pooled and expanded overnight in LB (10 g tryptone, 5 g yeast extract, 10 g/L NaCl) (BD Difco & Fisher) supplemented with 34 μg/mL chloramphenicol (CM) and 0.2% (w/v) glucose. Serial dilutions were performed and plated on LB-agar/CM plates to estimate the library diversity under the assumption that each colony was a unique transformant. Glycerol stocks of ~10^{11} cells were aliquoted in cryovials (Nalgene) and stored at -80 °C.

2.4.2 Random bacterial display peptide library screening

To screen bacterial display peptide libraries for serum antibody-binding peptides, a previous protocol was adapted for high-throughput magnetic selection and NGS [150]. Prior to screening, antibodies that bind to *E. coli* must be removed from serum specimens of interest to ensure that cells will be selected only if displayed peptides are bound by antibodies. To accomplish this, *E. coli* cultures expressing the eCPX scaffold with no peptide insert are used as depletion reagents. First, this eCPX *E. coli* strain is grown overnight at 37 °C with shaking (250 rpm) in LB supplemented with 34 μg/mL CM (LB + CM) and 0.2% glucose. The next day, LB + CM is inoculated at 1:50 with the overnight culture, grown to an OD_{600} of 0.4–0.6, and induced for eCPX expression for 1 hr at 37 °C with 0.02% (w/v) L(+)-arabinose. After induction, cells are centrifuged at 3,000 rcf for 5 min, washed once with PBST (1x PBS with 0.05% Tween 20), and resuspended in serum diluted 1:50 in PBST. For every μL of undiluted sera, ~10^9 cells (~1 mL of induced culture) are used for depletion. The samples are incubated overnight at 4 °C with gentle mixing on an orbital shaker (20 rpm). Antibodies that bound to *E. coli* or the eCPX scaffold are then removed by centrifugation at 5,000 rcf for 5 min twice, with depleted serum recovered as the supernatant after each centrifugation step. The depleted serum is stored at 4 °C for at most two weeks. All serum specimens utilized for library screening were depleted in this process unless stated otherwise.

Following serum depletion, a random bacterial display peptide library is screened for antibody binding towards an individual serum specimen using magnetic selection. A trimer-X_{12} or NNS-X_{15} library glycerol stock containing ~10^{11} cells (>10x the estimated diversity) is thawed and inoculated into at least 500 mL of LB + CM. The culture is grown to an OD_{600} of 0.4–0.6 at 37 °C with 250 rpm shaking and induced with 0.02% arabinose for 1 hr. For each serum specimen being analyzed, 5×10^{10} cells (>5x oversampling) are collected by centrifugation at 3,000 rcf for 10 min and resuspended in 1 mL PBST. The library cells are then cleared of peptides that bind to the selection reagents, protein A/G magnetic beads (Pierce), by incubating cells with beads at a ratio of 1 bead per 200 cells for 45 min at 4 °C with gentle mixing. Protein A/G beads are washed with PBST (x3) using magnetic separation prior to use. Two rounds of magnetic separation are then used to recover the unbound cells. The recovered cells are centrifuged, resuspended in 1 mL serum diluted 1:100, and incubated for 45 min at 4 °C with gentle mixing. Following serum incubation, cells are washed with PBST by centrifugation (x3) and resuspended in 1.5 mL PBST. After washing and resuspension, protein A/G beads are added at a ratio of 1 bead per 50 cells and incubated for 45 min at 4 °C with gentle mixing. Magnetic separation is utilized to separate cells displaying peptides that are bound by antibodies. The unbound cells in the supernatant are discarded and the separated cells bound by beads are washed with 1 mL PBST. This process is repeated five times to enrich the library population for antibody binding. After the final wash, the cells are resuspended in 1 mL of LB, inoculated into 10 mL LB + CM and 0.2% glucose, and grown overnight at 37 °C with shaking at 250 rpm.

The first magnetic selection enriches the library for antibody binding and also reduces the library size to facilitate a more efficient selection process in subsequent rounds. A second magnetic selection round is then completed to further enrich the peptide library. Here, the protocol for the first magnetic selection is modified to accommodate smaller working volumes. Following the overnight library growth of the first magnetic selection, cells are inoculated at 1:50 into 10 mL LB + CM, grown to an OD_{600} of 0.4–0.6, and induced with

0.02% arabinose for 1 hr. A volume of cells containing >20x the library diversity is collected, centrifuged, and resuspended in 200 μL PBST. Library cells are again cleared of cells that bind the magnetic separation reagents using the same protocol as the first selection but with a higher ratio of 1 bead per 2 cells. After magnetic separation, the recovered unbound library cells are centrifuged, resuspended in 200 μL serum diluted 1:100, and incubated for 45 min at 4 °C with gentle mixing. After the serum incubation, cells are washed with PBST by centrifugation (x3) and resuspended in 200 μL PBST. Protein A/G beads are added at a ratio of 1 bead per cell and incubated for 45 min at 4 °C with gentle mixing. Magnetic separation is utilized in the same manner as the first selection round to separate cells displaying peptides that are bound by antibodies. After the final wash, the separated cells are resuspended in 1 mL of LB, inoculated into 10 mL LB + CM and 0.2% glucose, and grown overnight at 37 °C with shaking at 250 rpm. Glycerol culture stocks are prepared the following day after each selection round for storage at -80 °C. All serums specimens were de-identified and obtained with consent according to institutional guidelines.

2.4.3 Peptide library enrichment analysis using flow cytometry

To quantify the enrichment of peptide libraries toward serum antibody binding, the libraries cells can be analyzed using flow cytometry. Enriched library cultures are grown and induced as described for magnetic selection (**Section 2.4.2**). After the libraries are induced, ~10^7 cells are centrifuged and resuspended in 50 μL serum diluted 1:100 for 45 min at 4 °C with gentle mixing. The antibody bound cells are then washed with PBST and resuspended in 50 μL of anti-human goat IgG conjugated to phycoerythrin (Jackson ImmunoResearch) diluted 1:100 in PBST. The library is incubated at 4 °C for 45 min and the cells are washed and resuspended in 500 μL PBS for flow cytometry on a FACSAria (BD Bioscience). A culture of eCPX scaffold cells is also included in this protocol and analyzed to quantify background serum reactivity (negative control) for each serum specimen used for library enrichment. Library cells are analyzed with flow cytometry using a blue excitation laser (488 nm) and a 576 nm PMT to quantify the number of library cells with red fluorescence above the eCPX culture background reactivity. A gate is created as a background threshold containing >99% of negative control events and any library cell with fluorescent signal greater than this threshold is considered enriched. Libraries with >90% enrichment are processed for NGS.

2.4.4 Amplicon library DNA preparation for next-generation sequencing

To isolate the DNA encoding bacterial displayed peptides, library cultures grown overnight following the second magnetic selection round were harvested and the plasmid DNA was extracted using a plasmid miniprep kit (Qiagen). To sequence the peptide region, a DNA amplicon must be constructed. The Nextera XT (Illumina) protocol was adapted for this system [151]. The peptide region was amplified using a two-step touchdown PCR. For the first PCR, the primers contain adaptors specific to the Illumina sequencing platform with annealing regions that flank the peptide region on the eCPX scaffold. The forward primer sequence is TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGVBHDV**GGCCAGTCTGGCCAG**, with bolded sequences indicating eCPX scaffold annealing regions. Five semi-random base pairs, VBHDV (IUPAC codes), were inserted at the beginning of the annealing region to provide sequence diversity for cluster identification during sequencing with the Illumina platform. The reverse primer is GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG**GTGATGCCGTAGTACTGG**. The first PCR uses 12.5 ng of library plasmid DNA, 5 μL of forward and reverse primers (1 μM), and 12.5 μL 2x KAPA HiFi HotStart ReadyMix

(Kapa Biosystems) in a 25 μL reaction volume. Touchdown PCR was completed using 15 cycles with the annealing temperature at 72 °C and reduced 0.5 °C/cycle followed by 15 additional cycles with annealing at 65 °C. The PCR products were purified using Agencourt AMPure XP magnetic beads (Beckman Coulter). The resulting product undergoes a second PCR to index the amplicon for specific samples and provide flanking adaptor regions for Illumina NGS according the Nextera XT protocol. These adaptors provide unique base pair barcodes enabling the identification of DNA sequences corresponding to individual samples. The second PCR is completed according to the Nextera XT protocol using 8 cycles with annealing at 70 °C. The PCR products are again purified using Agencourt AMPure XP magnetic beads, quantified, and pooled for sequencing. All sequencing was completed using a NextSeq500 (Illumina).

2.4.5 Peptide sequence processing

Following NGS, the Illumina platform automatically de-multiplexes the sequencing data and separates sequences based on the individual barcode indices. The algorithm IMUNE [76] is used to convert the FASTQ formatted files containing DNA sequences into library-specific files containing peptide sequences. Each DNA sequence contains primer annealing regions that flank the peptide sequence region. IMUNE identifies the upstream and downstream annealing sequences and if either of the annealing regions contains greater than 25% sequence errors such as insertions, deletions and/or mutations, the sequence is discarded. Once both annealing regions are successfully identified and located, the DNA sequence encoding the peptide is determined and translated. Only sequences that correspond to the correct peptide length are retained. To be conservative not to propagate any sequence errors, peptide sequences with nine or more identical positions were assumed to be the result of PCR or sequencing errors and were combined into a single peptide sequence.

2.4.6 Motif discovery using MEME

To determine motifs from peptide sequence datasets, the discovery algorithm MEME [131] was utilized. For this application, MEME takes peptide sequences as input and outputs motifs that were significantly present in the dataset. MEME determines the sequence logo plot as well as the motif regular expression. MEME is not suitable to analyze large NGS datasets and therefore the MEME algorithm is run locally with a maximum of 5,000 sequences.

To determine general motif characteristics from serum antibodies, motifs were analyzed from 20 libraries enriched against human serum specimens. Following library sequencing, the 1,000 highest enriched peptide sequences from each sample library were input into MEME for motif discovery. The 10 most significant motifs discovered via MEME for each library were examined for motif widths and numbers of conserved positions. Regular expressions were determined using default MEME output, with amino acids included in the expression if the frequency was greater than 0.2 at each position. This analysis was completed using both the NNS-X_{15} and trimer-X_{12} libraries for a total of 40 independent libraries.

2.4.7 Monoclonal antibody analysis

For mAb analysis, three rabbit mAbs were used to screen peptide libraries for binding and identify specificities: myc-Tag 71D10 (EQKLISEEDL), V5-Tag D3H8Q (GKPIPNPLLGLDST), and HA-Tag C29F4 (YPYDVPDYA) (Cell Signaling Technology). The trimer-X_{12} library was screened against each mAb at 20 nM in PBST and sequenced with NGS to reveal consensus binding motifs. To resolve the consensus binding motif for each mAb, the 5,000 most enriched peptides from each library were used as input for MEME motif discovery with minimum widths of seven for myc and HA and eight for V5 due to a longer epitope. The mAb motifs were defined using the default regular expressions reported via MEME.

2.4.8 Motif enrichment calculations

Motif enrichment is defined as the ratio of actual motif observations in the library to the expected observations, where expected observations were calculated using the amino acid frequencies observed in the library and assuming all positions were independent of one another. To calculate the enrichment of a motif in a peptide library, enrichment is calculated as:

$$enrichment = \frac{N}{E} \tag{2.1}$$

where N is the number of motif observations in the library and E is the expected observations calculated as:

$$E = N_{total} * (L_{sequence} - L_{motif} + 1) * P_{motif} \tag{2.2}$$

where N_{total} is the total number of sequences in the library, $L_{sequence}$ and L_{motif} are the lengths of sequences and motifs, respectively, and P_{motif} is the probability of observing the motif in the library given the amino acid frequencies in the library and assuming each position occurs independently. For example, the probability of the motif ACxDx[EF]G, where “x” is any of the 20 amino acids and bracketed positions indicate one of the enclosed amino acids is present at that position, would be calculated as:

$$P_{motif} = P_A * P_C * P_x * P_D * P_x * (P_E + P_F) * P_G \tag{2.3}$$

where P_i represents the frequency of amino acid in the enriched library.

We generally use frequencies from selected libraries as opposed to naive library frequencies for the enrichment calculations because the naive library frequencies do not account for bias in the expression of peptides containing certain amino acids, which could impact the peptide screening. Therefore it is often more accurate to use the library frequencies following selection as it is known that those peptides were actually expressed, displayed, and bound. For the mAb screens in PBST, the amino acid frequencies in the enriched library were highly biased due to the presence of a single antibody species. In this case, the naive library frequencies were used to calculate enrichments.

To determine statistical significance of motif enrichment, Poisson distribution statistics are used. The probability P of observing N motifs when E are expected is calculated with:

$$P(N,E) = \frac{E^N e^{-E}}{N!} \tag{2.4}$$

and the probability of at least N observations is:

$$P(\geq N,E) = 1 - \sum_{i=0}^{N-1} P(i,E) \tag{2.5}$$

3 Discovery and mapping of epitopes from herpes simplex viruses using random peptide library screening

Herpes simplex viruses (HSV) are highly prevalent and endemic throughout the world. There are two HSV species, HSV-1 and HSV-2, which cause numerous diseases including oral and genital herpes. While diagnostics for HSV exist, including viral culturing and PCR, difficulties in accurate diagnosis arise due to the high degree of similarity between the two species as well as the asymptomatic nature of HSV infections. Additionally, diagnosis often requires a laboratory facility and time-consuming experiments. There is an increasing need to develop highly specific reagents that can be used for rapid and accurate detection of HSV. Here, high-throughput random peptide library screening was applied to identify antibody-binding motifs associated with HSV infections. Specifically, we identified 14 HSV-2 specific motifs that each demonstrated 100% sensitivity and 100% specificity when distinguishing HSV-2 from HSV-1. We also identified nine HSV-1 specific motifs with high specificities and moderate sensitivities. The diagnostic sensitivity was improved by developing a classifier utilizing multiple HSV-1 specific motifs, achieving 100% sensitivity and 100% specificity. Further validation was completed by analyzing HSV-specific motif enrichments in peptide libraries from general populations including youth and adult cohorts to assess diagnostic utility and estimate HSV prevalence. Finally, online protein databases were used to align motifs to HSV proteomes and determine candidate antigens. We discovered motifs that mapped to known HSV antigens including glycoproteins G and D as well as additional antigens such as viral tegument proteins not previously investigated for epitopes. These discoveries could lead to improved diagnostics for HSV and help prevent the spread of disease. Additionally, antigen identification could lead to a better understanding of HSV pathogenesis and the development of novel therapeutics and vaccines. Moreover, the successful application of this high-throughput peptide screening to the highly similar HSV species demonstrates the versatility of this platform for epitope discovery and characterization.

3.1 Introduction

Herpes simplex virus type 1 (HSV-1) and type 2 (HSV-2) are highly infectious viruses that are widespread in humans and cause numerous diseases. HSV-1 is the primary cause of oral herpes (cold sores) while HSV-2 is responsible for most cases of genital herpes. Following primary infections, HSVs establish lifelong latent infections in the nervous system [77]. While many infections are asymptomatic, viral reactivation can occur leading to recurrent lesions. HSVs are usually transmitted by direct contact with active lesions or bodily fluids of an infected individual but transmission can also occur during asymptomatic periods [78]. Active infections are often diagnosed by viral culturing or PCR, but these methods are limited by requiring a laboratory facility and long experimental procedures, while often lacking sensitivity (true positive rate) and specificity (true negative rate) [94–96]. Recently, various serological antibody diagnostics have been developed and utilized to detect type-specific HSV infections during asymptomatic periods [89,92–94]. However, these tests can be highly dependent on the stage of infection and therefore need to be conducted and interpreted correctly. Despite these limitations, antibody-based diagnostics have a promising future for rapid and accurate HSV diagnosis [94,152].

To enable the development of novel antibody-based diagnostics, there is an increasing need to identify and characterize epitopes specific to each HSV types. Difficulties in HSV type-specific diagnosis arise due to the high degree of antigen similarity between HSV-1 and HSV-2, as over 80% of the protein-coding regions from each virus share sequence homology [94]. There have been many efforts to identify epitopes, largely focused on HSV surface glycoproteins, using peptide tiling methods [82,85,90,153] and predictive epitope mapping [154]. However, the majority of the studies utilizing peptide tiling methods identified epitopes with low precision by identifying large epitope regions that bound to HSV antibodies. Additionally, there have been many efforts to identify epitopes and antigens for vaccine development [83,155,156], however no vaccine has been proven safe and efficacious [157,158]. To enable the development of highly sensitive and specific reagents for use in diagnostics and vaccines, detailed epitope mapping at the amino acid resolution may be required.

Here, we applied an unbiased, high-throughput peptide screening technology to identify and map HSV epitopes. By screening a random bacterial display peptide library against HSV-1 and HSV-2 patient sera, millions of peptides were sequenced and analyzed to identify HSV type-specific binding motifs. These binding motifs were utilized as diagnostic reagents to classify patient serum specimens with high sensitivity and specificity. Additionally, binding motifs were associated with native HSV antigens by sequence alignment using protein databases. These discoveries could lead to the development of novel diagnostic reagents and vaccines. Moreover, these results demonstrate the success of high-throughput peptide screening which can be readily applied to discover and characterize epitopes from diverse pathogens.

3.2 Results

3.2.1 Identification of HSV-specific antibody-binding peptides

To identify epitopes associated with HSV-1 and HSV-2, we analyzed the antibody repertoires from serum specimens from 11 subjects positive for HSV-1, 12 subjects positive for HSV-2, and six subjects negative for both HSV types. A random 12-mer bacterial display peptide library was screened for antibody-binding peptides against individual serum specimens. Peptides were selected for antibody-binding using two rounds of magnetic cell selection. Following peptide screening, the DNA plasmids encoding the library peptides were extracted and sequenced using next-generation sequencing (NGS). An average of two million unique 12-mer sequences was obtained for each specimen analysis. Peptide sequence libraries were then analyzed for HSV type-specific binding motifs. Specimen cohorts were divided into discovery cohorts used for identifying type-specific motifs and validation cohorts used for independently evaluating the diagnostic accuracy of motifs. Additional specimens from a general population were utilized for analysis including a youth cohort composed of 22 specimens under the age of 16 years and an adult cohort composed of 128 specimens over the age of 50.

3.2.2 Discovery of HSV-2 specific motifs

To identify antibody-binding motifs specific to HSV-2, the algorithm IMUNE was utilized to identify highly enriched motifs in HSV-2 libraries but absent in control libraries. These motifs were clustered based on sequence similarity to identify distinct, consensus binding motifs. Motif enrichment was then evaluated in the validation cohort libraries to determine highly sensitive and specific motifs for HSV-2. We identified 14 motifs that were highly enriched (average enrichment >10) in HSV-2 libraries and exhibited minimal enrichment in negative controls and HSV-1 libraries (**Figure 3.1**). The validation HSV-2 specimens exhibited high enrichment for all motifs, comparable to the discovery cohort. The motifs were non-reactive in HSV-1 specimens, indicating the motifs were HSV-2 type-specific. Additionally, the motifs displayed 100% specificity for the youth cohort, which we expect to be HSV-2 negative based on demographics. Therefore, we identified 14 distinct antibody-binding motifs that demonstrated high diagnostic accuracy for HSV-2.

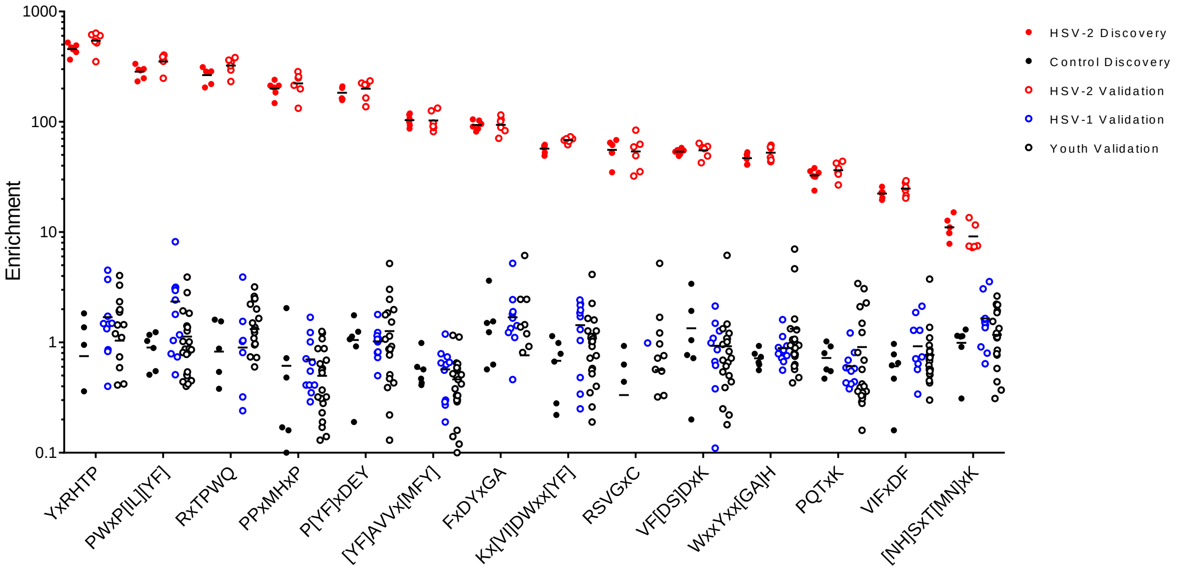


Figure 3.1. HSV-2 binding motifs were discovered with high sensitivities and specificities. Antibody-binding motifs highly enriched in HSV-2 discovery libraries (N=6, red closed circles) and absent in negative control libraries (N=6, black closed circles) were identified and motif enrichments were evaluated. Enrichment was then calculated in the validation cohorts of HSV-2 libraries (N=6, red open circles), HSV-1 libraries (N=11, blue open circles), and youth specimens (N=22, black open circles). Enrichment is calculated as the ratio of the number of motifs observed to the number of expected motifs given random chance. The plot uses a logarithmic scale and any specimens with motif enrichment of zero were omitted from the plot but still included in the average calculation (black line) for each cohort. All motifs demonstrated 100% sensitivity and 100% specificity.

To further assess the diagnostic accuracy of the HSV-2 binding motifs, adult specimens from a general population were evaluated for motif enrichment. The four HSV-2 motifs with the highest enrichments were analyzed in the adult cohort (**Figure 3.2**). By utilizing an enrichment threshold of >10 for classifying a specimen as HSV-2 positive, each motif demonstrated reactivity in 2–4% of the adult cohort with 10% of specimens positive for at least one motif. Based on the estimate of 14% of the adult population being positive for HSV-2 in America [21], and assuming a similar prevalence in the adult cohort utilized, these motifs displayed high diagnostic specificity in a large population.

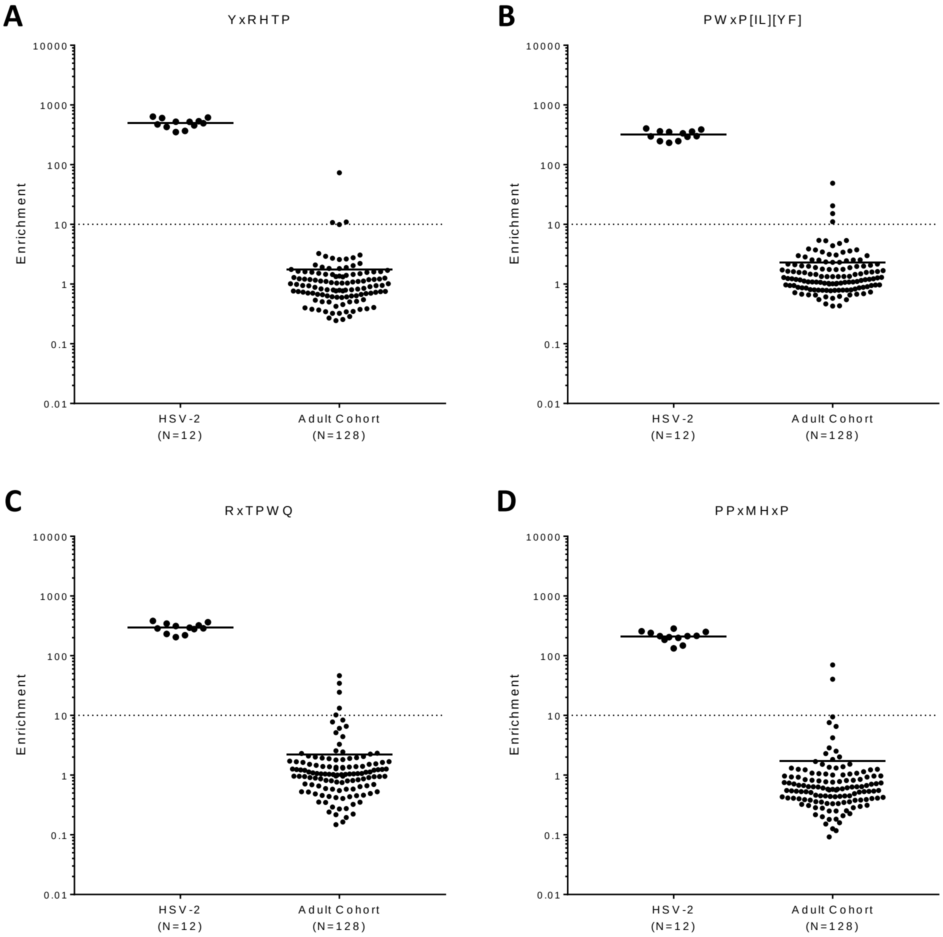
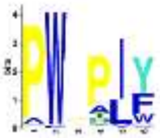
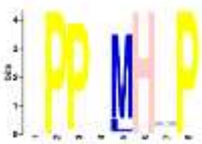
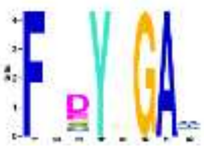
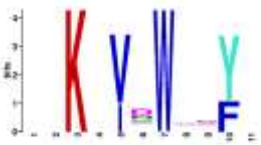
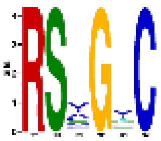


Figure 3.2. HSV-2 motifs were highly sensitive and specific in a large general population. Four HSV-2 specific motifs were evaluated for enrichment in an adult cohort (N=128), representing a general population, and compared to the discovery and validation HSV-2 cohort enrichments (N=12). An enrichment threshold of >10 (dashed line) was used to estimate the proportion of adult cohort specimens positive for HSV-2 for each motif: **(A)** YxRHTP (2% positive), **(B)** PWxP[IL][YF] (3% positive), **(C)** RxTPWQ (4% positive), and **(D)** PPxMHxP (2% positive). In total, 10% of adult cohort specimens were positive for at least one HSV-2 specific motif.

3.2.3 Determining candidate antigens from the HSV-2 proteome

After discovery of HSV-2 specific motifs, we sought to map the binding motif sequences to the HSV-2 proteome to identify candidate antigens containing the motifs. The protein database BLASTp was utilized for proteome alignments with the HSV-2 motifs. To improve the accuracy of antigen identification, the motifs were expanded to include additional positions and amino acids based on the clustering results. The expanded motifs were aligned and five significant antigen matches (E-value <1) were reported (**Table 3.1**). Among the candidate antigens identified, the motifs PWxP[IL][YF] and RSVGxC mapped to envelope glycoproteins G2 and D2, respectively. Numerous candidate antigens were therefore readily identified by proteome alignment using highly sensitive and specific HSV-2 motifs.

Table 3.1. HSV-2 candidate antigens.

Motif **Sequence Logo**	**Expanded Motif**	**Candidate** **Epitope**	**Candidate** **Antigen**	**BLASTp** **E-Value**
PWxP[IL][YF] 	PWPP[IL][YFW]	PWPPIW	Envelope glycoprotein G2	0.70
PPxMHxP 	PPR[ML]H[QM]P	PPRLHQP	Transcriptional regulator RS1	0.16
FxDYxGA 	F[DQ][DE]YPGA[VMI]	FDDYPGAV	Tegument protein UL47	0.20
Kx[VI]DWxx[YF] 	KY[VI][DN]W[DQ]D[YF]	KYIDWDDY	Minor capsid UL6	0.14
RSVGxC 	RS[VAIL]G[VIY]C	RSLGVC	Envelope glycoprotein D2	0.17

3.2.4 Discovery of HSV-1 specific motifs

To identify antibody-binding motifs specific to HSV-1, we utilized a procedure similar to that applied to HSV-2 motif discovery. IMUNE was used to discover highly enriched motifs in HSV-1 libraries and absent from negative control libraries. Motifs were again clustered to reveal consensus HSV-1 binding motifs. Motif enrichments were quantified in the discovery cohorts and displayed in a heatmap (**Figure 3.3A**). Nine motifs were identified but individual motifs demonstrated lower sensitivities, as each motif was reactive in at least two of five HSV-1 specimens. The motifs were then evaluated for enrichment in the validation cohorts of HSV-1 and HSV-2 specimens to assess motif capabilities for type-specific diagnostic utility (**Figure 3.3B**). The motifs exhibited minimal enrichment in HSV-2 libraries and similar sensitivities to those observed in the discovery cohort, suggesting the motifs discovered were HSV-1 specific motifs. However, the lower sensitivity of individual motifs indicates poor diagnostic utility. To circumvent this, we combined the nine motifs into a diagnostic panel to improve the accuracy for classifying HSV-1 specimens apart from HSV-2 specimens. A support vector machine (SVM) classifier was trained using the motif enrichments from the discovery cohorts to classify HSV-1 specimens and then tested on the validation cohorts. A receiver operating characteristic (ROC) curve was generated for the classification of HSV-1 from HSV-2 validation specimens (**Figure 3.4**) which illustrates the number of true positive and false positive classifications based on varying discrimination thresholds. The area under the curve (AUC) was determined to be 1, indicating correct classification for all validation HSV-1 (100% sensitivity) and HSV-2 specimens (100% specificity). Thus, utilizing multiple HSV-1 specific motifs enabled accurate type-specific diagnosis of HSV-1 from HSV-2.

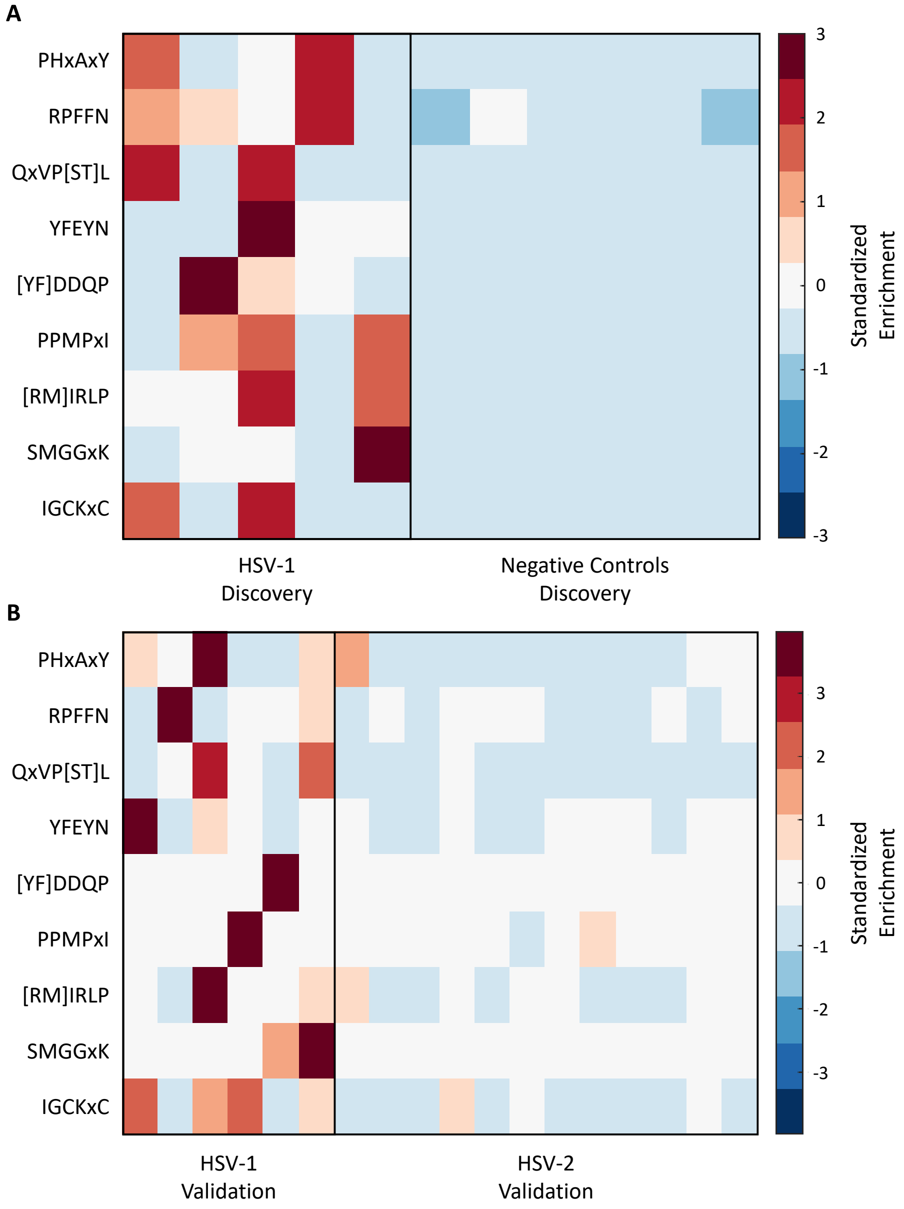


Figure 3.3. HSV-1 motifs were discovered with moderate sensitivity and high specificity. (A) Nine motifs were identified as significantly enriched in multiple HSV-1 specimen libraries (N=5) and not enriched in negative control specimen libraries (N=6). **(B)** Enrichments were evaluated in independent validation cohorts including HSV-1 specimen libraries (N=6) and HSV-2 specimen libraries (N=12). Enrichments were standardized for each motif (row) across specimens (columns).

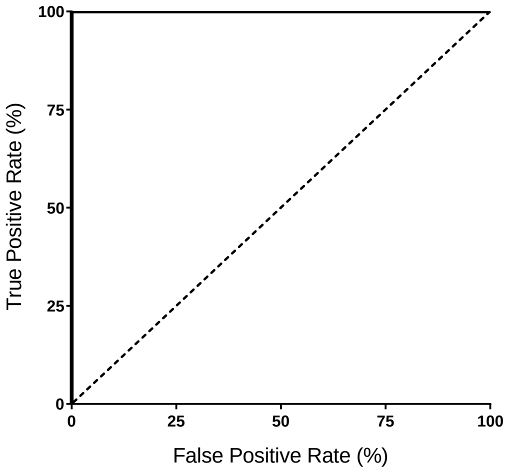


Figure 3.4. Motif panel classifier demonstrated high sensitivity and specificity for classifying HSV-1 from HSV-2. Enrichments of nine HSV-1 specific motifs were utilized to classify specimens. A support vector machine (SVM) was trained using the discovery cohort of HSV-1 specimens (N=5) and negative controls (N=6). The classifier was then tested on the validation cohort of HSV-1 specimens (N=6) and HSV-2 specimens (N=12). A receiver operating characteristic (ROC) curve was generated to illustrate the true positive rate (sensitivity) and false positive rate (100% - specificity) for the validation cohort classification based on varying discrimination thresholds. The area under the curve (AUC) was 1 indicating perfect classification. The identity line (dashed) represents a classification due to random chance (AUC=0.5).

The enrichments of four HSV-1 motifs were then evaluated in the adult cohort libraries to estimate the prevalence of HSV-1 in a general population (**Figure 3.5**). Utilizing an enrichment threshold of >10 for classifying a specimen as HSV-1 positive, the motifs demonstrated reactivity ranging from 18–55% in HSV-1 specimens. In the adult cohort, the motifs exhibited reactivity in 10–14% of specimens with 41% of specimens positive for at least one motif, suggesting a sufficient proportion of the adult population had been exposed to HSV-1. The prevalence of HSV-1 in American adults is estimated to be 40–50% [20]. These results further indicate that the motifs discovered are specific to HSV-1 and could be utilized as diagnostic reagents.

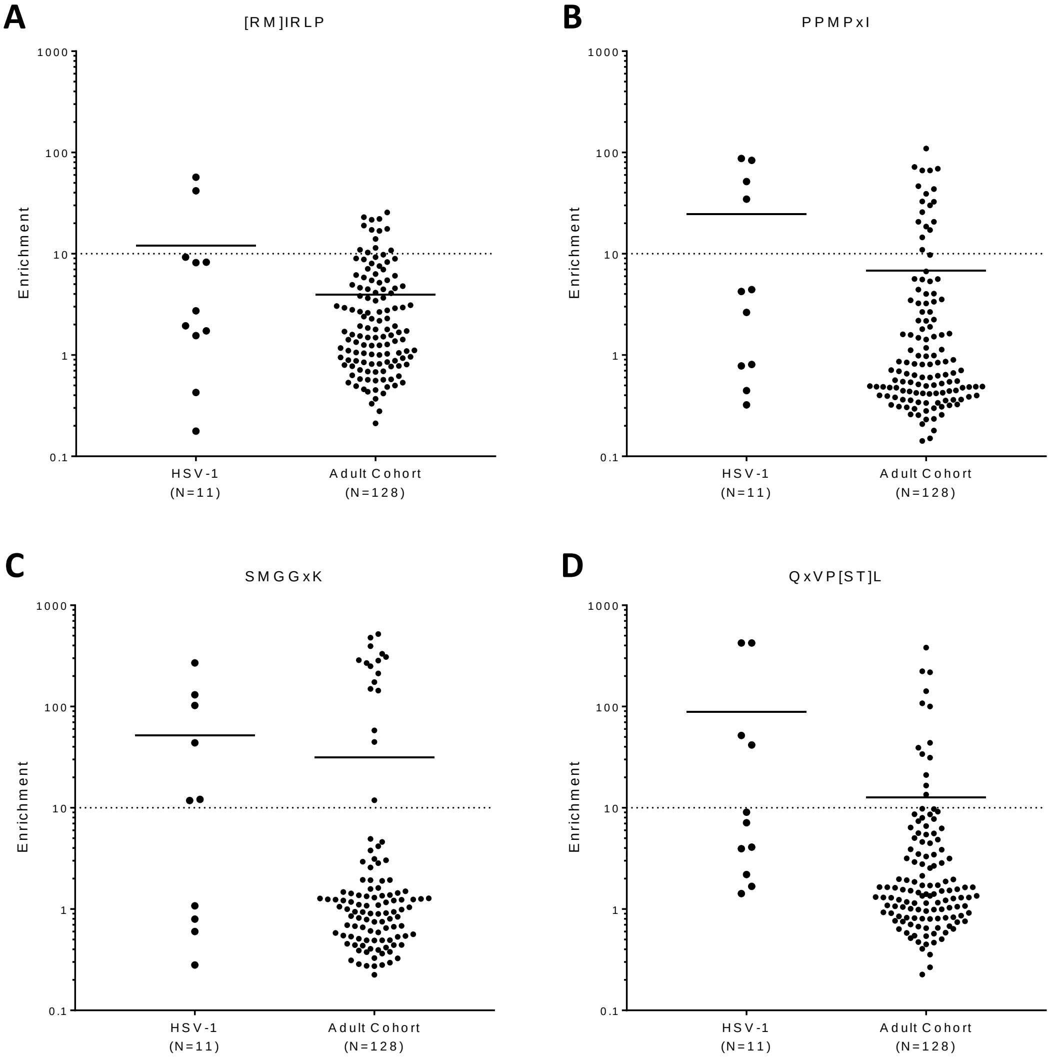
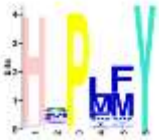
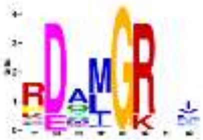
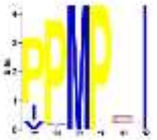
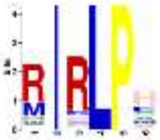


Figure 3.5. HSV-1 motifs exhibited reactivity in a large proportion of an adult population consistent with the expected HSV-1 prevalence. Four HSV-1 specific motifs were evaluated for enrichment in an adult cohort (N=128) and compared to the HSV-1 discovery and validation cohort enrichments (N=11). An enrichment threshold of >10 (dashed line) was used to estimate the proportion of adult cohort specimens positive for HSV-1. **(A)** The motif [RM]IRLP had 18% sensitivity for HSV-1 specimens and was 10% positive in adults. **(B)** The motif PPMPxI had 36% sensitivity for HSV-1 specimens and was 14% positive in adults. **(C)** The motif SMGGxK had 55% sensitivity for HSV-1 specimens and was 13% positive in adults. **(D)** The motif QxVP[ST]L had 36% sensitivity for HSV-1 specimens and was 10% positive in adults. In total, 41% of adult cohort specimens were positive for at least one HSV-1 specific motif.

3.2.5 Determining candidate antigens from the HSV-1 proteome

To identify HSV-1 candidate antigens, the protein database BLASTp was utilized to align the HSV-1 specific motifs to the HSV-1 proteome. Due to the small sample size, we applied a cross-validation procedure to discover additional HSV-1 specific motifs. The discovery and validation cohorts were exchanged, and motifs were discovered in the new discovery cohort. All motifs were again expanded based on the clustering results to include additional sequence information and improve the accuracy of proteome database alignments. Following BLASTp alignments, four significant candidate antigens were identified containing HSV-1 specific motifs (**Table 3.2**). Three of the four antigens mapped to either envelope glycoproteins D1 or G1. The high-throughput peptide library screening therefore revealed motifs specific to HSV-1 and enabled the mapping of numerous candidate antigens associated with HSV-1 infection.

Table 3.2. HSV-1 candidate antigens.

Motif Sequence Logo	Expanded Motif	Candidate Epitope	Candidate Antigen	BLASTp E-Value
HxP[LM][FM]Y 	H[QM]P[LMF][FM]Y	HQPLFY	Envelope glycoprotein D1	0.03
DA[ML]GR 	[RP][DE][AQY][MLI]G[RK][YV][IVLM]	PDYMGRYL	Tegument protein UL7	0.48
PPMPxI 	[PIV]PMP[SDP]I	PPMPSI	Envelope glycoprotein G1	0.05
[RM]IRLP 	[RM]IRLP[HF]	RIRLPH	Envelope glycoprotein D1	0.07

3.3 Discussion

In this study, we utilized a high-throughput peptide screening technology to identify antibody-binding motifs specific to HSV-1 or HSV-2. This non-targeted approach utilizing a large random peptide display library enabled the identification of highly sensitive and specific motifs for both types of HSVs, despite the high degree of antigen similarity between the viruses. For HSV-2, 14 motifs were identified demonstrating 100% sensitivity and 100% specificity. Additionally, many motifs displayed a large dynamic range between HSV-2 and HSV-1 specimens, with a 250-fold difference in enrichment on average for the top four motifs. The large dynamic range and homogeneity of response observed in the HSV-2 cohort is likely a consequence of HSV-2 infection status, as the specimens were IgM positive and therefore likely had active or recent infections [159,160]. For HSV-1, nine motifs were identified as moderately sensitive and highly specific. Sensitivities were likely lower as these specimens were IgG positive, likely corresponding to previous HSV-1 exposure but not active infections. A classifier was then developed utilizing multiple motifs to improve diagnostic sensitivity. By utilizing a random peptide library for high-throughput peptide screening, we readily identified diverse antibody-binding motifs that exhibited type-specific diagnostic utility for both HSV-1 and HSV-2.

While motif discovery was successful and diagnostic utility was demonstrated, the discovery was limited due to relatively small sample sizes. Diagnostic utility is ultimately established by analyzing hundreds of clinical samples and therefore additional work will be required to further validate the motifs identified here. However, one advantage of utilizing random peptide library screening and NGS is the ability to assemble large databases of peptide libraries screened for antibody binding from diverse specimens. Here, a youth cohort composed of specimens with ages ranging from 2–16 years was utilized as a cohort likely negative for HSV-2, a primarily sexually transmitted virus. By evaluating HSV-2 specific motifs in the youth cohort, we were able to further confirm that the motifs were highly specific for HSV-2. Similarly, a cohort consisting of adult specimens was analyzed for enrichment of HSV type-specific motifs to assess reactivity in a general population. Based on the epidemiological data for HSV, we observed reactivity trends consistent with HSV-1 and HSV-2 prevalence. Therefore, discovery of disease-specific motifs can be improved by analyzing motif enrichments in these large databases where false positives can be readily identified and discarded if reactivity is significantly different than expected. While caution needs to be taken when generating hypotheses regarding expected prevalences in uncontrolled cohorts, these large databases can be useful tools for additional validation of disease-specific motifs.

Along with motif discovery, another capability of random peptide library screening is the identification of candidate antigens. The HSV proteomes were searched for antigens containing the motifs discovered through screening to determine possible viral proteins responsible for antibody binding. Numerous candidate antigens were determined for both HSV-1 and HSV-2, including envelope glycoproteins D and G. These are known antigenic proteins that have been studied previously for antigen mapping and discovery [82,85,90]. Moreover, multiple epitopes discovered previously contain motifs identified in this study, such as the HSV-1 epitopes DDQPSSHQPLFY and HRRTRKAPKRIRLPHIR from glycoprotein D1 [85] containing the motifs HxP[LM][FM]Y and [RM]IRLP, respectively. Additionally, the HSV-1 epitope AISLTTPDHTPPMPSIGLEE has been identified from glycoprotein G1 [82] which contains the motif PPMPxI identified here. While the HSV glycoproteins are useful antigens to investigate for epitopes, less-studied but relevant antigens may exist and go overlooked. Notably, we identified multiple tegument proteins as candidate antigens for both HSV-1 and HSV-2, which could be the source of valuable antibody biomarkers previously undiscovered.

While numerous candidate antigens were identified, many motifs were not connected to candidates. One main reason for this is that short motifs (four to six amino acids) generally do not contain enough sequence information to reliably connect to candidate antigens through database alignments [33]. Follow-up experiments are likely required to elucidate additional sequence information to enable antigen identification [70]. Additionally, some motifs are likely structural mimics (mimotopes) with low sequence resemblance to the cognate epitopes and thus protein database alignments will not be successful [161]. Random peptide library screening therefore facilitated the identification of binding motifs belonging to a diverse set of antigens as well as possible mimotopes with diagnostic utility.

Ultimately, the identification of HSV type-specific motifs with high diagnostic accuracy could facilitate the development of new gold standard HSV diagnostics. Random peptide library screening enabled the discovery of numerous motifs that could be used in a multiplexed manner to detect antibodies associated with various stages of HSV infection. Multiple antigens were also identified as potential antibody targets that could reveal insights into viral pathogenesis and lead to the development of novel therapies and vaccines. Moreover, this versatile screening and discovery platform can be applied to any infectious disease for developing diagnostic reagents and determining antigens. As point-of-care technology continues to advance [162,163], these binding motifs could be integrated for rapid and accurate detection of antibody biomarkers for various diseases. In addition to improving diagnostics and enabling therapeutic intervention, these types of reagents could be used to monitor changes in the prevalence and spread of infectious diseases worldwide.

3.4 Materials and methods

3.4.1 Study population

HSV serum specimens were obtained from subjects positive for HSV-1 IgG (N=11) (Discovery Life Sciences) and subjects positive for HSV-2 IgM (N=12) (NOVA Biologics). Serum specimens negative for HSV-1 and HSV-2 (N=6) (Discovery Life Sciences) were utilized as a negative control cohort. Additionally, serum specimens were analyzed from a youth cohort (N=22) composed of specimens under the age of 16 and an adult cohort (N=128) composed primarily of Caucasian American specimens over the age of 50.

3.4.2 Bacterial display peptide library screening

To determine antibody-binding peptides from individual serum specimens, the screening procedure described in **Section 2.4.2** was applied. Briefly, a 12-mer random bacterial display peptide library (trimer-X_{12}) was screened for antibody binding using serum specimens diluted 1:100. Cells displaying peptides bound by antibodies were sorted and enriched using two rounds of magnetic selection using protein A/G magnetic beads. From each enriched peptide display library, plasmid DNA was extracted and DNA amplicons containing the peptide sequence region were amplified and sequenced using NGS from Illumina.

3.4.3 Identifying antibody-binding motifs

The motif identification algorithm IMUNE [76] was used to identify antibody-binding motifs specific for HSV-1 or HSV-2. For HSV-2 discovery, the HSV-2 cohort was divided into a discovery cohort (N=6) and a validation cohort (N=6). IMUNE was then used to identify motifs significantly enriched (-$\log_{10}$ p >8) in all HSV-2 validation libraries and not significant enriched (-$\log_{10}$ p >2) in any discovery cohort controls. Motif enrichment was calculated as the ratio of motifs observed to motifs expected, and Poisson distribution statistics were used to determine statistical significance (**Section 2.4.8**). Motifs were restricted to four defined amino acids in a frame of four to eight positions possibly interspaced by undefined positions. Following HSV-2 specific motif identification, motifs with an average enrichment >15 in HSV-2 discovery libraries (N=548) were selected for clustering to determine consensus motifs. A similarity score between two motifs was generated by aligning the motifs and evaluating the sequence similarity based on PAM30 alignment scoring [76,164]. Hierarchical motif clustering was then applied based on the similarity scores using average distance clustering (MATLAB, clustergram). Motifs within a cluster were traced back to the peptide sequences containing the motif that were observed at least once in all discovery HSV-2 libraries. A maximum of 1,000 cluster sequences were input into the motif discovery algorithm MEME to determine the consensus binding motif. Using the consensus motifs, a second round of hierarchical clustering and MEME discovery was applied to reduce motif redundancy and generate the final set of consensus motifs. For each cluster, MEME generated a sequence logo profile and a motif regular expression, restricted to the five most significant positions in the motif and including amino acids with frequencies greater than 20% at a given position. Bracketed positions in a consensus motif indicate one of the amino acids enclosed was present at that position while an “x” represents any amino acid. Following motif discovery, motif enrichments were calculated in the remaining validation cohorts.

HSV-1 motif discovery was accomplished using the same procedure as used for HSV-2, but with parameters adjusted to optimize discovery. The discovery cohorts included HSV-1 specimens (N=5) and negative control specimens (N=6). IMUNE was used to discover motifs significantly enriched ($-\log_{10}$ p >8) in at least three HSV-1 specimens and not significantly enriched ($-\log_{10}$ p >4) in any negative control specimens. Motifs were restricted to five defined amino acids in a frame of five to eight positions possibly interspaced by undefined positions. Motifs with an average enrichment >20 in HSV-1 specimens were utilized for hierarchical clustering and motif identification. Following hierarchical clustering, peptide sequences containing motifs were identified and used for MEME motif discovery if the sequence was observed in at least two of the five HSV-1 discovery libraries. Consensus motif discovery was then accomplished using the procedure outlined for HSV-2.

Fewer motifs were identified for HSV-1 than HSV-2. To enable the discovery of additional candidate antigens, a second discovery process was completed by exchanging the HSV-1 discovery and validation cohorts and repeating the procedure. Motifs discovered in this cross-validation were only utilized for candidate antigen identification, not diagnostic validation.

3.4.4 Support vector machine classification

Due to the moderate sensitivities of HSV-1 specific motifs, a binary support vector machine (SVM) classifier was developed to classify HSV-1 positive specimens from HSV-1 negative specimens. A support vector machine (MATLAB, fitcsvm) with a linear kernel was trained using nine HSV-1 motif enrichments from the original discovery cohort of HSV-1 specimens (N=5) and negative control specimens (N=6). The SVM classifier was then tested on HSV-1 (N=6) and HSV-2 (N=12) validation specimens to evaluate the diagnostic accuracy.

3.4.5 Motif alignments using protein sequence databases

The protein basic local alignment search tool (BLASTp) was utilized to align HSV type-specific motifs to HSV proteomes for candidate antigen identification. Motifs were first expanded to include additional sequence information from the MEME motif discovery to improve antigen identification and reduce false positive alignments. Expanded motifs were defined as MEME regular expressions including all amino acids with frequency greater than 10% at a given position. Each unique variant of the expanded motif was then aligned to the corresponding HSV proteome using the BLASTp non-redundant database search and default parameters for short input sequences. Organism queries were either herpes simplex virus 1 (taxid 10298) or herpes simplex virus 2 (taxid 10310). Candidate motifs were reported if the E-value was less than one.

4 Serum antibody repertoire analysis reveals unique binding signatures associated with risk of age-related macular degeneration

Age-related macular degeneration (AMD) is the leading cause of blindness in the developed world. AMD progresses over time, predominantly after the age of 60, and is classified into early, intermediate, and late stages of the disease. As the prevalence of AMD continues to increase, there is an increasing need to identify novel biomarkers associated with the onset of late AMD to enable earlier intervention. However, current AMD classification schemes are largely limited to clinicians and researchers analyzing and grading fundus photographs of the retina based on abnormalities. Here, we analyzed the antibody repertoire for serological biomarkers at early stages of AMD. Specifically, we utilized high-throughput peptide screening to identify antibody-binding motifs associated with the onset of late AMD progression. An SVM binary classifier was developed using the reactivities of 1,125 binding motifs to distinguish onset late AMD serum specimens from specimens at varying risk of developing late AMD. When testing classification of onset late AMD specimens versus specimens in a youth cohort, the diagnostic accuracy was 89%. Similarly, classification of onset late AMD versus an age-matched cohort with low risk of developing AMD exhibited a diagnostic accuracy of 84%. However, the classifier performance worsened as the risk of developing late AMD increased, suggesting a unique antibody response associated with AMD progression. From the set of motifs used for classification, we identified three consensus binding motifs that demonstrated increased reactivity in onset late AMD specimens. Furthermore, these motifs exhibited reactivity over years of disease progression, indicating that the corresponding antibodies were sustained over time. These findings suggest that antibody-binding motifs could be utilized to assess risk of late AMD at early stages and assist in monitoring disease progression. Additionally, the binding motifs could lead to the identification of novel targets for therapies and improve our understanding of AMD development and progression.

4.1 Introduction

Age-related macular degeneration (AMD) is the most common cause of blindness in developed countries [165,166] and the number of individuals with AMD worldwide is projected to increase to 288 million by 2040 [167]. AMD progresses slowly over time, predominantly after the age of 60, and is classified into early, intermediate, and late stages of disease [168]. The early and intermediate stages of AMD are typically asymptomatic but can be detected during a dilated eye exam by the presence of yellowish extracellular deposits under the retina, known as drusen, as well as possible pigmentary irregularities [169]. The relationship between drusen formation and AMD progression is correlative, with an increased risk of developing late AMD as the number and size of drusen increase [107]. Late AMD is further classified as either geographic atrophy (GA) or neovascular (NV) AMD which are both often accompanied by vision impairment. GA AMD is characterized by atrophy of the retinal pigment epithelium (RPE) and photoreceptor cells whereas NV AMD is characterized by the abnormal growth of leaky choroidal blood vessels which leads to photoreceptor degeneration [98]. While therapies are limited, there are interventions that can reduce the risk of developing late AMD, such as nutritional supplements [109] and antiangiogenic therapeutics for NV AMD to impede additional retinal damage [112]. Because damage occurs at the late stages of AMD, it is critical to detect AMD early to enable risk assessment and appropriate intervention.

An ideal strategy for early diagnosis would be to detect serological biomarkers associated with AMD progression as these tests would be convenient and readily available. It is known that immune activation plays a critical role in the development and progression of AMD [98]. Accumulating evidence has implicated the complement system as a key factor in AMD, including a single nucleotide polymorphism in complement factor H that can account for up to 50% of AMD cases [97,101]. Additionally, numerous complement proteins have been identified as elevated components in drusen using proteomics [106]. The formation of drusen can also stimulate a chronic immune response exacerbating AMD progression [105,108,170]. Thus, there have been many efforts to identify serological biomarkers associated with inflammation in AMD. Elevated levels of various serum/plasma biomarkers have been associated with AMD [169] including C-reactive protein [171,172] and numerous complement proteins [173]. Additionally, various antibodies associated with AMD have been identified, including antibodies targeting carboxyethylpyrrole (CEP) protein adducts which are generated due to oxidative stresses involved with aging [123]. Anti-retinal antibodies have also been associated with AMD [124,125] and observed in patients with early AMD [174]. Moreover, unique antibody profiles have been observed at different stages of AMD progression [126]. While these findings suggest a link between the adaptive response and AMD, it is unclear whether antibodies associated with AMD are causal or epiphenomenal. Despite these uncertainties, serological antibodies provide a unique opportunity to serve as biomarkers for AMD prognostics and diagnostics and even provide insights into disease etiology and/or pathology. However, due to the complexity and progressive nature of AMD, it is unlikely that a single antibody species will perform as an effective prognostic or diagnostic [127]. Therefore, a promising strategy is to identify an antibody profile that encompasses a set of diverse antibody species associated with AMD.

Here, we applied an unbiased, high-throughput peptide screening platform to identify antibody-binding peptides associated with AMD. We utilized a large random bacterial display peptide library [76] to screen against serum specimens from the Age-Related Eye Disease Study (AREDS). Specifically, we focused on serum specimens collected years prior to the development of GA and/or NV AMD to identify antibody biomarkers associated with the onset of late stage progression. By discovering peptides reactive in onset late AMD specimens, we identified unique antibody profiles that could be used to classify the onset of late AMD from healthy controls. Additionally, consensus binding motifs associated with onset late AMD were identified and determined to be reactive to antibodies throughout the progression of early to late AMD. These results suggest that unique antibodies exist at the early stages of AMD progression and could be used to inform of the risk of late AMD and enable earlier diagnosis and therapy. Furthermore, the binding motifs could reveal targets associated with late AMD to provide insights into the disease pathology and be used as therapeutic targets.

4.2 Results

4.2.1 Discovery of antibody-binding motifs associated with onset late AMD

Five cohorts of serum specimens from subjects with various stages of AMD were analyzed in this study to identify antibody profiles associated with onset late AMD (**Table**

4.1). Four distinct cohorts from AREDS were analyzed. Three cohorts were composed of subjects that had either small, medium, or large drusen and did not progress beyond that stage throughout the trial. As the focus of this study was the onset of late AMD, these three cohorts were treated as various controls. The fourth cohort, deemed onset late AMD, developed GA and/or NV AMD during the trial. Serum specimens obtained at the beginning of the trial were analyzed, before the development of late AMD, to focus on antibody biomarkers associated with the early stages of progression. The final cohort was composed of non-AREDS youth subjects with an average age of 6 years (range of 2–16 years).

Table 4.1. Clinical characteristics of the study population.

	Youth	Small Drusen	Medium Drusen	Large Drusen	Onset Late AMD
Population, no.	22	19	30	30	49
Age (years), mean (SD)	6 (5)*	69 (5)	68 (4)	69 (5)	70 (5)
Gender, no. (%)					
Male	12 (55)	9 (47)	8 (27)	14 (47)	15 (31)
BMI (kg/m^2), mean (SD)	NA	26 (4)	27 (4)	26 (4)	27 (4)
Ethnicity, no. (%)					
White	NA	17 (89)	29 (97)	30 (100)	49 (100)
Black/Hispanic/Asian/Other	NA	2 (11)	1 (3)	0 (0)	0 (0)
Smoking, no. (%)					
Never	NA	9 (47)	13 (43)	16 (53)	25 (51)
Former smoker	NA	8 (42)	17 (57)	13 (43)	20 (41)
Current smoker	NA	2 (11)	0 (0)	1 (3)	4 (8)
AREDS treatment*, no. (%)					
Placebo	n/a	11 (58)	8 (27)	10 (33)	11 (22)
Antioxidant	n/a	8 (42)	3 (10)	4 (13)	4 (8)
Zinc	n/a	0 (0)	14 (47)	10 (33)	22 (45)
Antioxidant + Zinc	n/a	0 (0)	5 (17)	6 (20)	12 (24)
AMD subtype developed, no. (%)					
Geographic atrophy (GA)	n/a	n/a	n/a	n/a	18 (37)
Neovascular (NV)	n/a	n/a	n/a	n/a	18 (37)
Both	n/a	n/a	n/a	n/a	13 (27)

BMI = body mass index; NA = not available; n/a = not applicable; no. = number; SD = standard deviation. Significant differences (*p <0.05): analysis of variance for continuous variables followed by Dunnett's multiple comparisons test and chi-square test for nominal variables.

To identify antibody-binding profiles associated with the onset of late AMD, we screened a random bacterial display peptide library for antibody-binding peptides with each serum specimen. Peptides were enriched for antibody-binding using two rounds of magnetic cell selection. Following library screening, the DNA plasmids encoding the library peptides were extracted and sequenced using NGS. Therefore, for every serum specimen, we obtained a peptide sequence library enriched for antibody binding.

To focus discovery on late AMD onset, the peptide libraries were analyzed for binding motifs highly enriched in AMD specimens but not enriched in control cohorts. The study population was divided into a discovery cohort used to identify motifs specific to onset late AMD and a validation cohort to independently validate the motif enrichments (reactivities). For discovery, motifs were identified that were significantly enriched in at least 4 of 26 libraries (>15%) from onset late AMD specimens. Additionally, the motifs identified were not significantly enriched in all small drusen specimens (N=10) and all medium drusen specimens (N=20). In total, 1,470 motifs were discovered that met these criteria. We then evaluated each cohort for reactivity towards each motif. Pie charts were generated to display the number of motifs that were reactive in different proportions of each cohort (**Figure 4.1**). Small and medium drusen controls in the discovery cohort were not reactive to any motifs as per the discovery criteria. For onset late AMD, the motifs demonstrated reactivity in only a subset of specimens, as 87% of motifs (1,282) were reactive in only 15% of the cohort (**Figure 4.1A**), corresponding to the minimum criteria used for motif discovery. There were only three motifs that exhibited reactivity in 27% of the cohort, the maximum proportion observed. This demonstrates the diversity of antibody reactivity in AMD, as no single binding motif was reactive in a majority of specimens.

Each motif was then evaluated for reactivity in the validation cohorts (**Figure 4.1B**). We observed increased motif reactivity across the cohorts as drusen size, and therefore risk of late AMD, increased. For the youth and small drusen cohorts, 81% (1,188) and 85% (1,249) of motifs, respectively, were non-reactive in all specimens. In the medium drusen cohort, 72% of motifs (1,064) were non-reactive in all specimens. Conversely, 56% of motifs (830) were reactive in at least one large drusen specimen and 63% of motifs (924) were reactive in at least one onset late AMD specimen. The majority of motifs were again reactive in a subset of the onset late AMD specimens, exhibiting a similar reactivity profile to that observed in the discovery cohort. Thus, we observed increased antibody reactivity correlated with increased risk of developing late AMD.

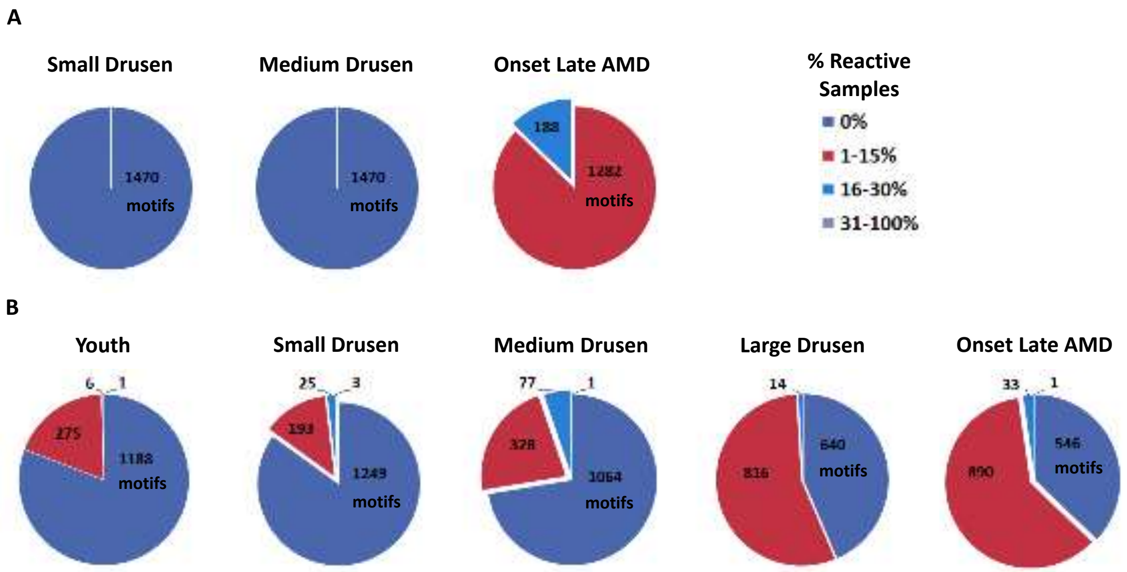


Figure 4.1. Antibody-binding motifs demonstrated reactivity correlated with increased risk of late AMD. Antibody-binding motifs were identified by screening random peptide libraries against serum specimens. A total of 1,470 motifs were discovered in onset late AMD specimen libraries and evaluated for reactivity in validation cohorts. Pie charts depict the number of motifs that were reactive in various proportions of specimens from the **(A)** discovery cohort with 10 small drusen specimens, 20 medium drusen specimens, and 26 onset late AMD specimens and **(B)** the validation cohort with 22 youth specimens, 9 small drusen specimens, 10 medium drusen specimens, 30 large drusen specimens, and 23 onset late AMD specimens. Statistically significant motifs were determined using Poisson distribution statistics for motif observations given an expected value. Motifs were considered reactive if $-\log_{10}$ p >8 for all cohorts except small and medium drusen discovery cohorts where a higher stringency was used ($-\log_{10}$ p >2).

4.2.2 Classification of onset late AMD using a motif panel

Although antibody-binding motifs were identified in onset late AMD specimens, reactivity was limited to subsets of the cohort suggesting individual motifs were not sufficient to serve as biomarkers and therefore a multiplexed approach must be taken. We observed that numerous motifs could be utilized in parallel to increase the coverage of reactivity for all specimens, as displayed in a heatmap of enrichments for motifs identified in the discovery cohort (**Figure 4.2**). Therefore, we utilized a support vector machine (SVM) to classify onset late AMD using a set of motifs. To train the classifier, the 1,470 motif enrichments were used to classify specimens from the discovery cohort as either onset late AMD or controls (small or medium drusen). A cross-validation procedure was performed to determine a subset of motifs (N=1,125) with minimal misclassification rate in the discovery cohort.

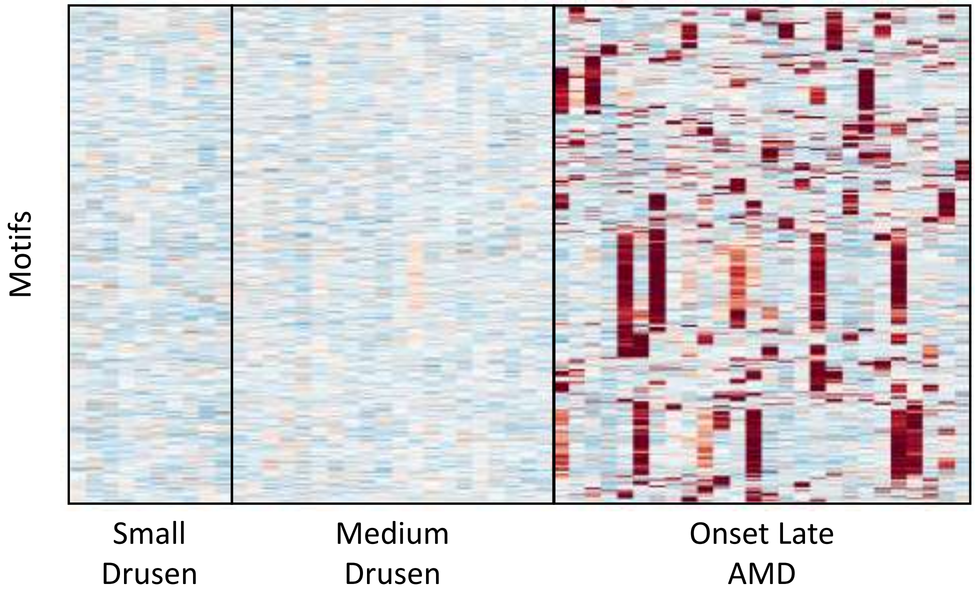


Figure 4.2. Utilization of a panel of motifs enabled coverage of entire onset late AMD cohort. Motifs (N=1,470) were discovered in onset late AMD specimens and motif enrichments were displayed in a heatmap, with each row representing enrichment values for a motif in small drusen (N=10), medium drusen (N=20), and onset late AMD (N=26) discovery specimens (columns). Enrichments were transformed using $\log_{10}$(enrichment+1) and standardized for each motif. Hierarchical motif clustering based on enrichment was utilized to aid visualization of distinct motif clusters and reactivity patterns across the cohorts.

The motif panel was then tested on the validation cohorts to assess independent classification. Specifically, the binary classifier was utilized to classify onset late AMD against each remaining cohort (youth and small, medium, and large drusen). For each test, a receiver operating characteristic (ROC) curve was generated (**Figure 4.3**) which determines the number of true positive and false positive classifications based on varying discrimination thresholds. The area under the curve (AUC) was calculated for each ROC curve to estimate the probability of correctly classifying a validation sample chosen randomly. We observed an increased AUC as the risk of late AMD decreased. Classifications with the youth (AUC=0.95) and small drusen (AUC=0.84) controls were significantly different than classification by random chance (identity line, AUC=0.5). Classifications with the medium (AUC=0.66) and large (AUC=0.53) drusen controls were not statistically significant. For additional validation, we randomized all AREDS samples, identified discovery and validation cohorts, and repeated the motif discovery and classifier development to determine the AUC obtained with randomized samples. By completing this procedure 10 times, we generated an average AUC of 0.52 with a standard error (SE) of 0.03 (data not shown). Using this set of random classifiers as a null distribution, we again determined that onset late AMD classifications using our original SVM classifier with the youth cohort and small drusen cohort were statistically significant. A summary of ROC curve classification results is presented (**Table 4.2**). These results demonstrate that motif enrichments could be utilized as classifiers to discriminate onset late AMD from cohorts with varying risks of developing AMD.

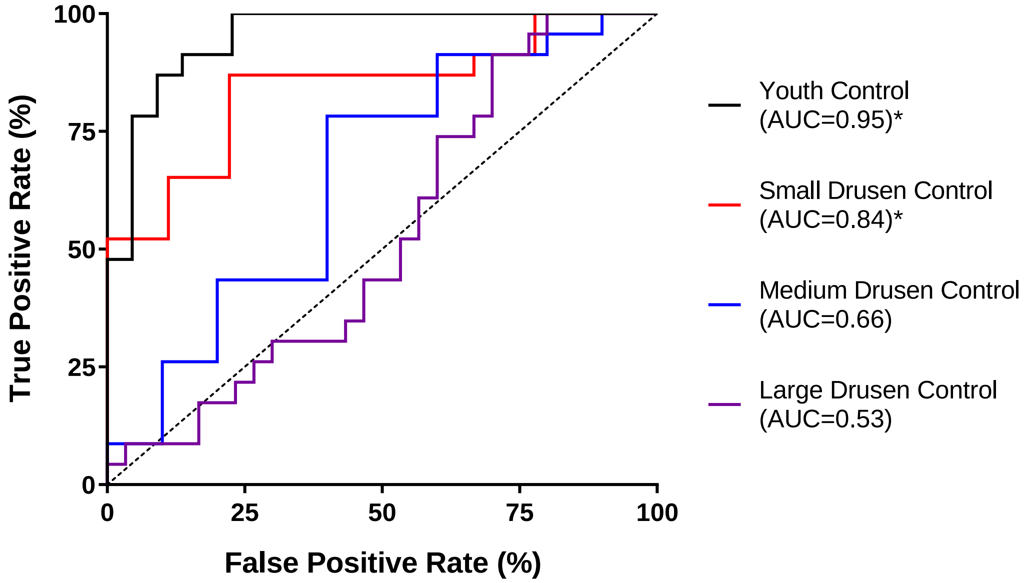


Figure 4.3. Motif panel classifier exhibited improved performance as risk of developing late AMD decreased. A support vector machine (SVM) algorithm for binary classification of onset late AMD was trained using a panel of motif enrichments (N=1,125). Training and optimization were performed on the discovery cohort. The SVM classifier was tested on an independent validation cohort of onset late AMD specimens (N=23) and various control cohorts including a youth cohort (N=22) and small (N=9), medium (N=10), and large (N=30) drusen cohorts. For each binary classification, the classifier performance was displayed using a receiver operating characteristic (ROC) curve which determines the number of true positive and false positive classifications based on varying discrimination thresholds. The area under the curve (AUC) represents an estimate of the overall accuracy of correctly classifying a validation sample chosen randomly. The identity line (dashed) represents a classification due to random chance (AUC=0.5). The AUCs for the youth and small drusen cohorts were determined significantly different (*p <0.01) than the identity AUC, as well as the average AUC of 10 randomly generated classifiers (data not shown).

Table 4.2. Motif panel classification statistics.

Disease		Control		AUC	SE	95% CI	Accuracy	p-value (Identity)	p-value (Randomized Trials)
Onset Late AMD	N=23	Young	N=22	0.95	0.03	0.89–1	0.89	<0.0001	<0.0001
		Small Drusen	N=9	0.84	0.07	0.70–0.98	0.84	0.0031	<0.0001
		Medium Drusen	N=10	0.66	0.11	0.44–0.87	0.73	0.1585	0.2373
		Large Drusen	N=30	0.53	0.08	0.37–0.68	0.57	0.7331	0.9256

AUC = area under the curve; CI = confidence interval; SE = standard error.

4.2.3 Identification of consensus binding motifs associated with onset late AMD

From the set of motifs discovered and utilized for classification of onset late AMD, we sought to identify distinct, consensus binding motifs. Of the motifs discovered, some have similar sequence identity and likely bind to the same antibody species. Motifs with similar sequence identity therefore exhibit similar reactivities, as observed in the enrichment heatmap (**Figure 4.2**) where motifs were clustered based on enrichments revealing distinct clusters of motifs. To further investigate and identify these consensus motifs, the motifs were grouped using hierarchical clustering based on the PAM30 similarity score between each motif. Three of the largest clusters were selected for analysis. For each cluster, the motifs were traced back to the peptide sequences containing the motifs from the enriched peptide libraries. These sequences were input into the motif discovery algorithm MEME [131] to determine a single consensus binding motif for each cluster. A sequence logo plot was determined for each cluster to visualize the consensus binding motif and enrichments were evaluated in the discovery and validation cohorts (**Figure 4.4**). The first consensus motif was identified as [ASN]LVN and exhibited reactivity (enrichment >3) in a subset of the onset late AMD validation specimens. Similarly, the second consensus motif, AGxx[VI]N, demonstrated reactivity in multiple onset late AMD specimens in addition to moderate enrichment in multiple medium and large drusen specimen. The third motif, [KR]DLxYP, exhibited high reactivity in a subset of large drusen and onset late AMD validation specimens but also showed moderate reactivity in two specimens from the youth cohort, suggesting it may not be highly specific. All motifs were non-reactive in the validation small drusen control specimens. The distributions of enrichments for each cohort were tested for normality, as non-reactive cohorts should have enrichments normally distributed around a mean enrichment of one. Onset late AMD cohorts had non-normal distributions for all three motifs due to reactivity in a subset of specimens. Conversely, the small drusen cohorts for all three motifs were normally distributed.

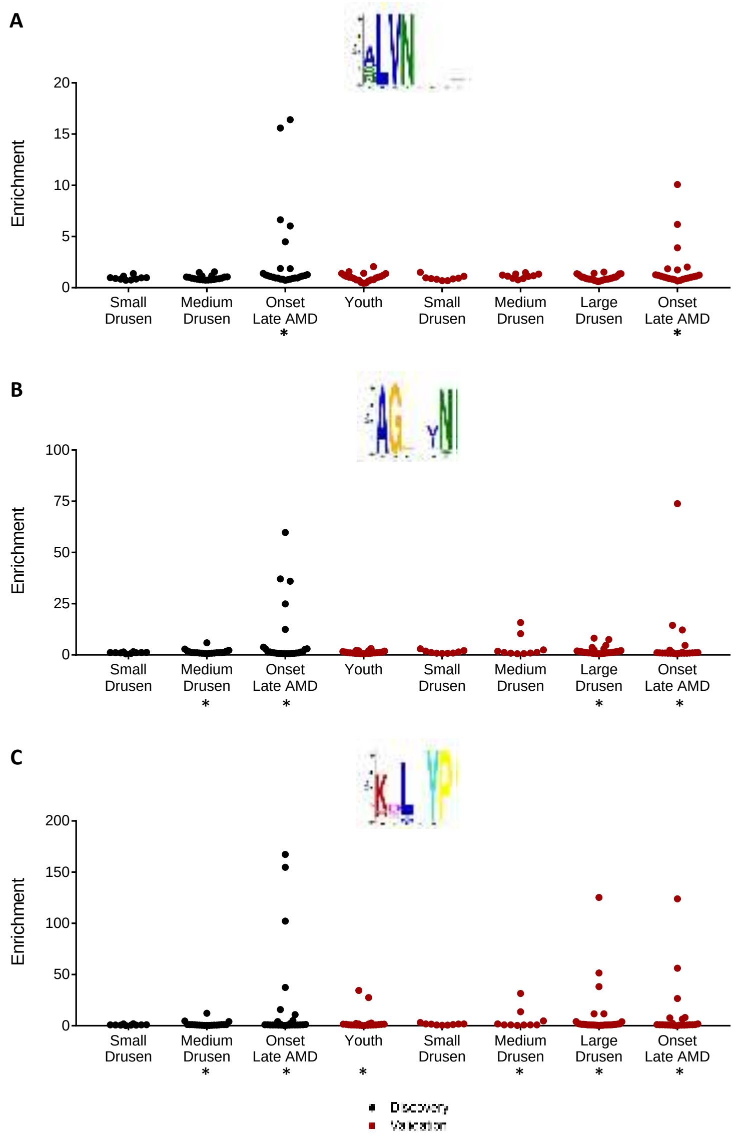


Figure 4.4. Consensus binding motifs demonstrated reactivity in subsets of onset late AMD specimens and cohorts with increased risk of late AMD. Motifs were clustered using hierarchical clustering to determine consensus binding motifs. Peptide sequences containing motifs from within a cluster were input into MEME to determine the consensus motif. Three consensus motifs **(A)** [ASN]LVN, **(B)** AGxx[VI]N, and **(C)** [KR]DLxYP were identified and the enrichments were evaluated and plotted for the discovery (black) and validation (red) cohorts. The discovery cohort includes small (N=10) and medium drusen (N=20) specimens and onset late AMD specimens (N=26). The validation cohort includes youth specimens (N=22), small (N=9), medium (N=10), and large (N=30) drusen specimens, and onset late AMD specimens (N=23). The distributions of enrichments for each cohort were tested for normality using the D'Agostino-Pearson normality test (*p <0.0001).

With consensus binding motifs determined, we then sought to evaluate motif reactivity over time as AMD progressed. For 19 specimens from both the onset late AMD cohort (discovery) and small drusen cohort (discovery and validation), two additional serum specimens obtained at later time points in the trial were analyzed for motif reactivity. The average time span between the three longitudinal specimens was 7 years (SD=1). All but six onset late AMD specimens developed GA and/or NV by the third time point. For the remaining six specimens, GA and/or NV developed within three years of the third time point. For the three consensus motifs, enrichment was calculated at each time point for both cohorts and plotted to observe changes in motif reactivity over time (**Figure 4.5**). For the onset late AMD cohort, the general trend was that motif enrichment did not fluctuate greatly over time. Therefore, if a specimen was positive for reactivity at the beginning of the trial, the reactivity was sustained over time as AMD increased in severity. Additionally, if a specimen was non-reactive at the beginning of the trial, there were no substantial increases in enrichment over time. Specifically, all but one small drusen specimen were non-reactive throughout the longitudinal span. The specimen with increased reactivity was for the motif [KR]DLxYP (**Figure 4.5C**) with an enrichment of eight at the final time point, which was still ~10-fold lower than the average enrichment for reactive onset late AMD specimens. We therefore identified distinct antibody-binding motifs that were reactive in subsets of specimens with increased risk of developing late AMD. Furthermore, these motifs were reactive for years with minimal reactivity in control specimens that did not develop AMD during the trial. These results suggest that the corresponding antibodies persisted over time as AMD progressed from early/intermediate stages towards late AMD.

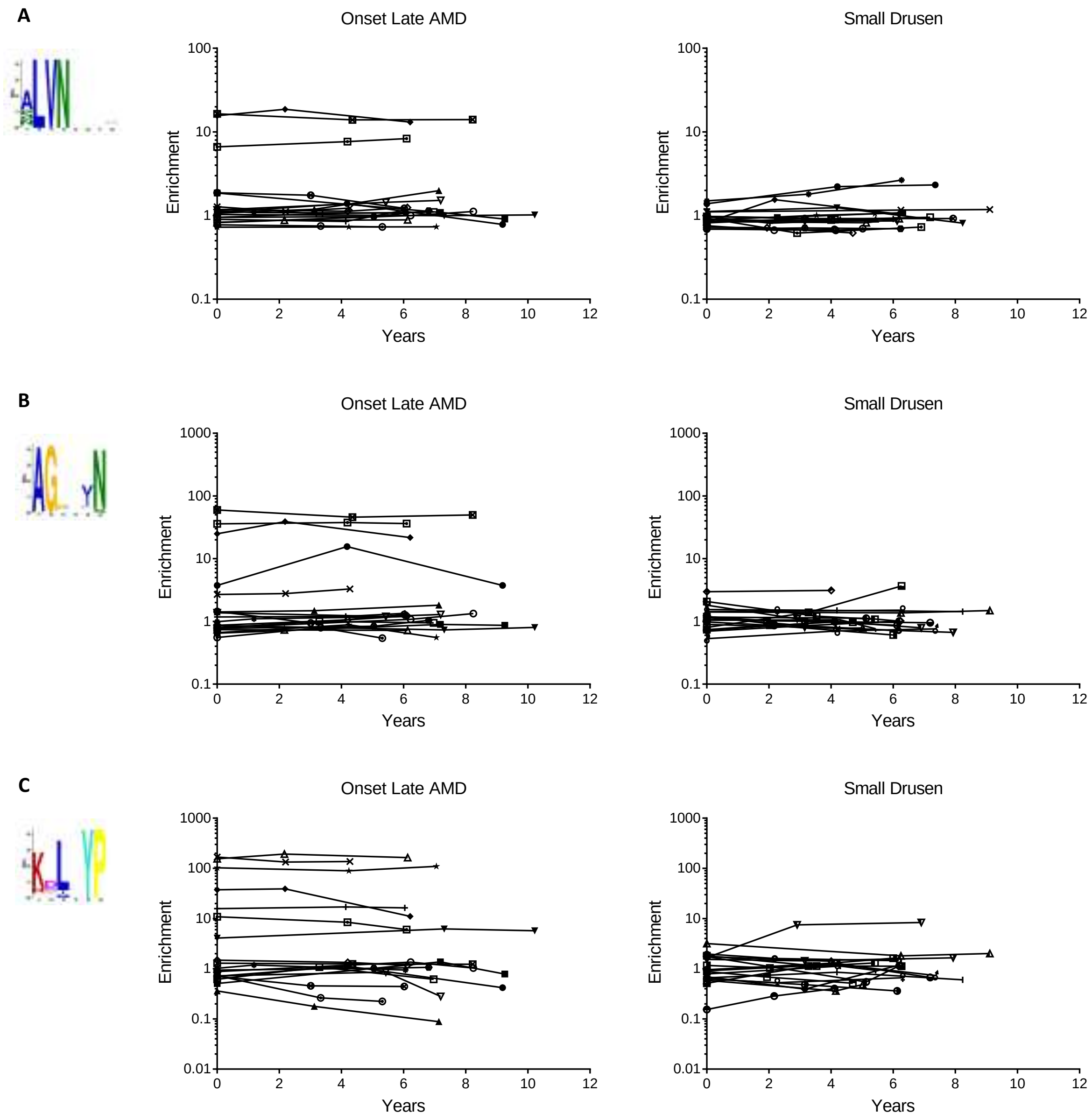


Figure 4.5. Consensus binding motifs showed sustained reactivity over years of disease progression. Enrichments of three consensus motifs **(A)** [ASN]LVN, **(B)** AGxx[VI]N, and **(C)** [KR]DLxYP were evaluated over time from serum specimens obtained at three visits throughout the AREDS. Each line represents a different subject. For onset late AMD (N=19), the specimens at the first time point were specimens utilized for motif discovery and correspond to time points prior to late AMD development. Each subject developed late AMD during the AREDS. Small drusen specimens (N=19) at the first time point were used for motif discovery (N=10) and validation (N=9). Each subject had small or no drusen for the entire trial. The years correspond to the time since the first serum specimen collection (within three years of baseline).

4.3 Discussion

As the prevalence of AMD increases, millions of lives will be impacted worldwide and an estimated $300 billion will be spent on direct costs annually in 2020 [103]. Currently, assessment of risk and diagnosis is largely limited to standardized classification schemes conducted by clinicians and researchers based on drusen size and pigment abnormalities in the retina [168]. There is an increasing need for more sophisticated, analytical methods utilizing various biomarkers for diagnosing AMD at the earliest stages and monitoring progression. In this study, we utilized high-throughput peptide screening to identify antibody-binding motifs associated with the onset of late AMD progression. Specimens from the onset late AMD cohort correspond to subjects with medium or large drusen at the time of serum collection, but who eventually developed late AMD. These specimens therefore provide a unique opportunity to study the antibody repertoire at the early stages of late AMD.

After identifying motifs significantly enriched in a discovery cohort of onset late AMD specimens, we evaluated the motif enrichments in independent validation cohorts. We observed that cohorts with increased risk of developing late AMD demonstrated increased reactivity towards the motifs. However, the motifs were only reactive in subsets of onset late AMD specimens. Individual motifs therefore demonstrated poor diagnostic utility, likely due to the complexity of AMD and the diversity of antibody targets. Similar trends have been observed in previous studies where fractions of late AMD specimens were highly reactive to diverse autoantigens [124]. To overcome these diagnostic limitations, we developed a classifier using a combination of motifs. The classifier exhibited improved diagnostic capabilities for correctly classifying specimens as onset late AMD when compared to cohorts with lower risk of developing late AMD, such as the youth cohort and the small drusen cohort. The small drusen cohort serves as a useful age-matched control due to the low risk (<2%) of subjects with small drusen developing late AMD over the course of 10 years [168]. Classifier performance decreased when distinguishing between onset late AMD and medium drusen specimens. The 10-year risk of developing late AMD when medium drusen is present is ~4% for unilateral drusen and ~13% for bilateral drusen [168]. Over 60% of the medium drusen specimens tested had bilateral medium drusen and were therefore at higher risk of developing late AMD. The classifier performed the poorest when classifying onset late AMD from the large drusen cohort, which was only slightly better than binary classification by random chance (AUC=0.5). The risk of late AMD in subjects with bilateral large drusen varies based on additional risk factors such as pigment abnormalities, but the five-year risk factor can be upwards of 50% [168,175]. All large drusen specimens in this study had large drusen in both eyes at baseline and were therefore at higher risk of late AMD. Therefore, the panel of motifs demonstrated increased reactivity as the risk of developing late AMD increased, suggesting a unique antibody response associated with the progression of AMD.

By clustering motifs based on sequence similarity, we identified distinct consensus binding motifs enriched in subsets of onset late AMD specimens. By analyzing the enrichments of consensus motifs in longitudinal specimens, obtained over years during the AREDS, we observed that motif enrichment did not fluctuate in onset late AMD specimens positive for reactivity. These findings indicate that the antibodies present at the early stages of AMD progression were sustained over time as AMD progressed to the late stages. Moreover, specimens with small drusen remained non-reactive across all time points. These consensus motifs could be used as specific biomarkers associated with late AMD progression and lead to the identification of novel targets for AMD therapy. For example, elevated levels of CEP protein adducts and autoantibodies targeting these adducts have been observed in AMD [123]. It has been hypothesized that CEP adducts stimulate angiogenesis and a monoclonal antibody targeting CEP adducts demonstrated potential for neutralizing angiogenesis [176]. The identification of antibody targets such as CEP could therefore provide novel therapeutic opportunities and inform about AMD development and progression.

In this study, we demonstrated the potential of serum antibody repertoire analysis for identifying biomarkers associated with the onset of late AMD. Due to the complexities of AMD, a multifaceted approach is necessary to appropriately assess risk and monitor the disease. By identifying antibody-binding motifs associated with onset late AMD, these peptides could be utilized as reagents for detecting antibody biomarkers. The low sensitivity (true positive rate) of motifs observed in this study exemplifies the need to utilize multiplexed antibody biomarkers. Follow-up studies with larger sample sizes will be critical to validating and advancing the diagnostic performance of these motifs. Due to limitations of motif sensitivities and sample sizes, we could not significantly evaluate differences between the late AMD categories of GA and NV. Future analyses of these cohorts could identify additional biomarkers and improve our understanding of late AMD. Recent studies have taken a different approach, focusing on plasma biomarkers at various stages of AMD using metabolomics [177]. These diverse analytical approaches to AMD biomarker discovery could be used in tandem with conventional protocols to strengthen diagnosis and risk assessment. As more advanced therapies emerge for treating AMD [117,118], these biomarkers could provide the opportunity to intervene earlier, limit AMD progression, and prevent vision impairment.

4.4 Materials and methods

4.4.1 Study population

Serum specimens were obtained from the AREDS conducted in collaboration with the National Eye Institute (NEI), part of the National Institutes of Health (NIH). This study consisted of four cohorts from the AREDS. The specimens were categorized based on the AREDS categories [104]. Briefly, subjects in the small drusen cohort (N=19) were essentially free of AMD abnormalities, with no drusen or small drusen (<63 μm diameter), for the duration of the trial. Subjects in the medium drusen cohort (N=30) had unilateral or bilateral medium drusen (63–124 μm) at the baseline examination and had medium drusen observed at least 80% of the time during follow-up while never developing large drusen or GA or NV AMD. The large drusen cohort (N=30) had large drusen (≥125 μm) at baseline and had large drusen observed at least 80% of the time during follow-up and never developed GA or NV AMD. Lastly, the onset late AMD cohort (N=49) had small, medium, or large drusen at baseline and developed GA and/or NV during follow-up. For motif discovery, serum specimens obtained within the first three years of the trial were analyzed. Serum specimens from two additional follow-up visits were obtained for longitudinal analysis from 19 subjects in both the small drusen and onset late AMD cohorts. For one small drusen subject, only a single follow-up specimen was obtained for analysis. For additional validation, specimens from an independent youth cohort (N=22) were analyzed, with an average age of 6 years ranging from 2 years to 16 years. All specimens were de-identified and obtained with consent according to institutional guidelines.

To describe the clinical and demographic characteristics of the study population, the mean and standard deviation for continuous variables were reported and percentages of nominal variables were reported. The different cohorts were analyzed for statistical differences ($p < 0.05$) using chi-squared tests or analysis of variance, followed by Dunnett's multiple comparison tests for comparing the young cohort to the AREDS cohorts.

4.4.2 Bacterial display peptide library screening

To identify antibody-binding peptides, the screening procedure described in **Section 2.4.2** was utilized. Briefly, a 12-mer random bacterial display peptide library (trimer-X_{12}) was screened for antibody binding using individual serum specimens diluted 1:100. Cells displaying peptides bound by antibodies were enriched with two rounds of magnetic selection using protein A/G magnetic beads. Following magnetic selection, plasmid DNA was extracted from the enriched cellular libraries and DNA amplicons containing the peptide sequence region were constructed and sequenced using Illumina NGS.

4.4.3 Antibody-binding motif discovery

To discover antibody-binding motifs associated with onset late AMD, we utilized the motif identification algorithm IMUNE [76]. The cohorts were divided into a discovery cohort used for motif identification and a validation cohort to independently validate the motif reactivity. The discovery cohort was composed of 26 onset late AMD specimens as disease positive and 10 small drusen and 20 medium drusen specimens as disease negative (controls). The validation cohort was composed of 23 onset late AMD specimens as disease positive with various disease negative control cohorts including 9 small, 10 medium, and 30 large drusen specimens, as well 22 youth specimens.

The peptide sequence libraries from each specimen in the discovery cohort were searched for motifs that were statistically enriched in onset late AMD while absent from designated controls. Motifs are stretches of amino acid sequences possibly interspersed with undefined residues. Here, we restricted motifs to five defined amino acids in a frame of five to eight positions, denoted by a regular expression such as ACDEF, ACxDxEF, or AxCxDExF, where the capital letters represent the corresponding amino acids and an “x” represents any of the 20 amino acids. Motif enrichment was calculated as the ratio of motif observations to expected observations, where expected observations were calculated using the amino acid frequencies observed in the enriched library while assuming all positions were independent of one another (**Section 2.4.8**). Additionally, statistical significance was determined using Poisson distribution statistics to evaluate the probability of observing a motif in a sequence dataset given the expected value. Motifs were identified if they were statistically significant ($-\log_{10} p > 8$) in at least 15% of the onset late AMD discovery cohort and not significant in any discovery cohort controls ($-\log_{10} p > 2$). Different p-values were utilized for discovery to provide additional stringency for what was considered significantly enriched in control specimens. Motifs were then evaluated in the validation cohorts for statistical significance ($-\log_{10} p > 8$).

4.4.4 Motif panel construction and classification

Motifs that met the discovery criteria were used in a diagnostic panel for binary classification. Motif enrichments were transformed using a $\log_{10}$(enrichment+1) transformation and standardized for each motif. A support vector machine with a linear kernel (MATLAB, fitcsvm) was trained using the motif enrichments from the discovery cohort. To optimize the SVM classifier, the number of motifs included in the panel was varied and a 10-fold cross-validation was performed to evaluate misclassification rate in the discovery cohort. Motifs were added to the classifier based on rank of sensitivity (true positive rate), specificity (true negative rate), and average enrichment, starting with the highest ranked motif and ending with the full set. This procedure was repeated 100 times and the set of motifs with the lowest average misclassification rate was used in the classifier. The classifier was then tested on the validation cohort for its ability to correctly classify onset late AMD specimens from various controls, including small, medium and large drusen cohorts as well as the young cohort. A ROC curve was generated for each classification and the AUC was calculated. Statistical significance ($p < 0.01$) was determined based on the null hypothesis that the classification was by random chance (AUC=0.5) assuming normal distributions. Diagnostic accuracy was calculated using the maximum Youden's index [178] as a threshold for classification.

An additional statistical test was completed by generating the average ROC curve AUC for motif panels constructed based on randomized specimen cohorts. All AREDS samples were randomized and divided into cohorts sized equivalently to the actual study cohorts for discovery and validation. Motifs were identified that met the defined criteria and used to build an SVM classifier, trained on the randomized discovery cohort. The AUC from the ROC curve was generated for classification of 23 random specimens designated "positive" from 9 random specimens designated "negative", a test equivalent to the classification of onset late AMD specimens from small drusen specimens. This procedure was repeated 10 times and the average AUC and SE were determined. This randomized trial was then used as the null distribution to determine statistical significance ($p < 0.01$) for determining an AUC from a given trial compared to random chance.

4.4.5 Hierarchical motif clustering to determine consensus motifs

To identify consensus motifs, the motifs from the diagnostic panel were clustered using hierarchical clustering based on motif similarity. A similarity score between each motif was generated by aligning the motifs and evaluating the sequence similarity based on PAM30 alignment scoring [76,164]. Hierarchical motif clustering based on similarity was completed using single-linkage clustering (MATLAB, clustergram). Three of the four largest clusters were selected for consensus analysis, as the fourth cluster did not show reactivity in the validation cohort. Each motif within a cluster was traced back to the peptide sequences containing the motif (observed in at least one onset late AMD library), and a maximum of 4,000 sequences were input into the motif discovery algorithm MEME to determine the consensus binding motif. MEME generated a sequence logo profile and a motif regular expression. Regular expressions were restricted to positions with >1 bit of information and amino acids with frequencies >10% at a given position. Bracketed positions in the consensus motif indicate one of the amino acids enclosed was present at that position.

After determining a consensus motif, the motif enrichments were calculated for each cohort. Here, specimen reactivity was evaluated based on an enrichment value >3. Additionally, the distributions of enrichments for each cohort were tested for normality using the D'Agostino-Pearson normality test (p <0.0001), as only subsets of specimens were reactive and thus the distributions were highly skewed. Non-reactive cohorts should therefore have enrichments normally distributed around a mean of one. The consensus motif enrichments were also calculated from AREDS specimens collected at various time points. For 19 specimens from both the onset late AMD and small drusen cohorts, a total of three serum specimens spanning an average of seven years were analyzed from the AREDS (one specimen only had two longitudinal time points). Exact dates for specimen collection were used when available, however if a date was unavailable, the date was interpolated from other visit dates.

5 High-titer antibody depletion enhances discovery of diverse serum antibody specificities

The human antibody repertoire is a unique repository of information regarding infection, inflammation, and autoimmunity of the past, present, and future. However, antibodies can span vast ranges of concentrations with varying affinities and the repertoire is often heavily polarized by a few species. These complexities lead to difficulties detecting and characterizing low abundance antibody species that may be relevant to disease. We therefore developed a method to selectively remove antibodies from a sample in proportion to the titer of the species prior to analysis, referred to as high-titer depletion (HTD). Peptides from a large random peptide display library were enriched for binding towards high-titer antibody species and utilized as binding reagents to deplete the corresponding species from the specimen. HTD enabled the discovery of antibody-binding specificities using random peptide library screening with reduced cross-reactivity and background signal and improved coverage of low abundance species. With HTD, three monoclonal antibody species were detected at concentrations at least an order of magnitude lower than without HTD. Additionally, 92 serum antibody specificities were readily discovered from an individual specimen using HTD compared to only 25 specificities without HTD. Parameters affecting the extent of depletion such as the concentration of depleted serum were also adjusted to reproducibly improve the coverage of antibody specificities. These results demonstrate that HTD could be employed for the discovery of rare specificities related to disease and enable extensive characterization of the antibody repertoire. Moreover, the strategy of depletion in proportion to titer could be extended to other applications with complex biological samples to improve discovery.

5.1 Introduction

Technological advancements in biomedical research have facilitated the analysis of complex biological samples for the identification and development of prognostics, diagnostics, therapeutic targets, and patient-specific therapies [179,180]. The serum proteome can be mined for protein biomarkers [181], the transcriptome sequenced to identify splice variants [182], and circulating tumor cells quantified for predictions of survival [183]. Recently, due largely to innovations in sequencing capabilities and other high-throughput platforms, the antibody repertoire has garnered significant attention [184]. This collection of antibodies produced by the adaptive immune system, acting as an immunological history record, can be analyzed by sequencing B cell populations and/or by elucidating antibody functions and binding specificities [5]. In-depth analysis of the antibody repertoire has revealed antibody biomarkers for diagnosing disease [37,185,186], disease-specific antibody epitopes [40,76], and fundamental insights into dynamics of the adaptive immune system with implications for pathology [187,188], vaccine design [6,189,190], and therapeutic antibody development [191].

When analyzing biological samples, a significant challenge involves sufficiently detecting targets of interest due to the complexity and diversity of samples. For example, the serological antibody repertoire is estimated to have 10^3–10^6 distinct binding specificities [4] and can be highly polarized by dominant antibody species [6,7]. This complicates the use of technologies designed to probe antibody binding such as peptide/protein arrays and libraries by diluting low abundance species while increasing background activity and cross-reactivity [192]. Similarly, the serum proteome has an extremely large dynamic concentration range spanning greater than 10 orders of magnitude and is dominated by a small number of proteins such as albumin and immunoglobulins [193]. The asymmetries in the serum proteome severely impede protein biomarker discovery by masking lower abundance proteins related to disease [194]. However, researchers have addressed this issue by selectively depleting known, highly abundant serum protein species prior to profiling, thereby reducing the dynamic range and enhancing coverage of low abundance species [195,196]. This type of targeted depletion strategy has also been employed for analyzing other complex biological samples such as the bulk removal of ribosomal RNA for improved RNA-seq capabilities [197] and the depletion of blood cells to enable quantification of circulating tumor cells [198,199]. Unfortunately, when analyzing the antibody specificity repertoire, targeted depletion strategies to improve detection do not exist largely because there are no ubiquitously abundant species, as repertoires vary greatly across individuals and time.

To improve discovery of antibody specificities for low abundance species, we therefore developed a method to remove high-titer serum antibody species from a sample prior to analysis, referred to as high-titer depletion (HTD). Peptides from a large random peptide display library were preferentially enriched towards binding high-titer antibody species and then utilized as binding reagents to deplete the corresponding antibody species from the sample. This approach was designed to deplete antibody species in proportion to their titer (i.e. concentration and affinity), enabling the depleted serum sample to be analyzed with reduced cross-reactivity and background signal and improved coverage of low abundance species. We compared antibody specificity repertoire analysis with and without HTD using bacterial display and NGS and found that HTD increased detection limits and coverage of the specificity repertoire. Importantly, HTD does not require information about the specimen prior to depletion and can therefore be robustly applied. Our results suggest that HTD could be used for discovery of rare specificities related to disease or infection and aid characterization of the antibody repertoire at a depth not previously possible. Furthermore, this general strategy of depletion based on concentration/titer could be extended to other areas of biological research to enhance discovery capabilities.

5.2 Results

5.2.1 Depleting highly abundant antibody species to improve antibody specificity discovery

To enhance the discovery of low-titer antibody specificities in human sera, a method was developed for high-titer depletion (HTD) of antibody species prior to library enrichment by magnetic selection. We reasoned that enrichment of a peptide display library against antibody-containing serum would generate a pool of peptide binders whose prevalence would be roughly proportional to the serum antibody titers and that the resulting binder pool could be used as a reagent to effectively deplete antibodies from a sample in proportion to their titer (**Figure 5.1**). Briefly, a bacterial display peptide library of 10 billion members [76], composed of *Escherichia coli* (*E. coli)* cells engineered to display a unique 12-mer peptide on the surface of each cell, is enriched for peptides binding to antibodies from an individual serum sample. Enrichment is completed via two rounds of magnetic separation using protein A/G coated magnetic beads. The enriched peptide library (the depletion reagent) is then incubated with the original serum sample to allow antibody binding, followed by centrifugation to separate unbound antibodies from antibodies bound to peptides displayed on cell surfaces. The unbound fraction, or depleted serum, is collected and used for a new peptide library enrichment (**Figure 5.2**). The library is then sequenced using NGS to enable peptide enrichment analysis.

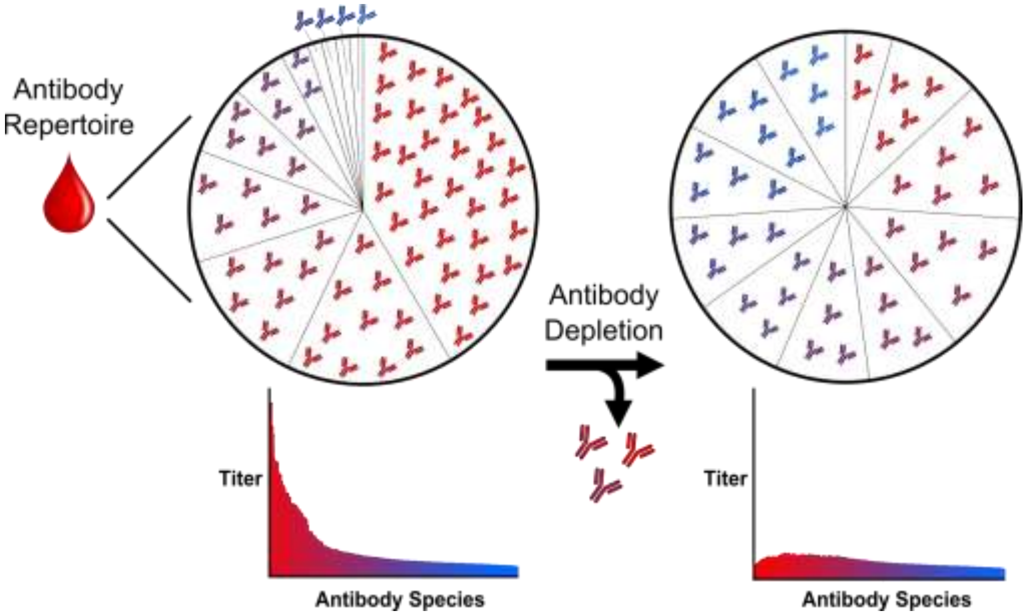


Figure 5.1. Conceptual approach for high-titer antibody depletion. The pie charts illustrate the abundance of various antibody species in a serum sample. The histograms demonstrate the titer (combination of concentration and affinity) of individual antibody species. On the left, high-titer antibodies (red) dominate the antibody repertoire and obscure lower-titer species (blue). On the right, by removing high-titer species through antibody depletion, a depleted serum sample will contain a higher fraction of low-titer species enabling improved detection.

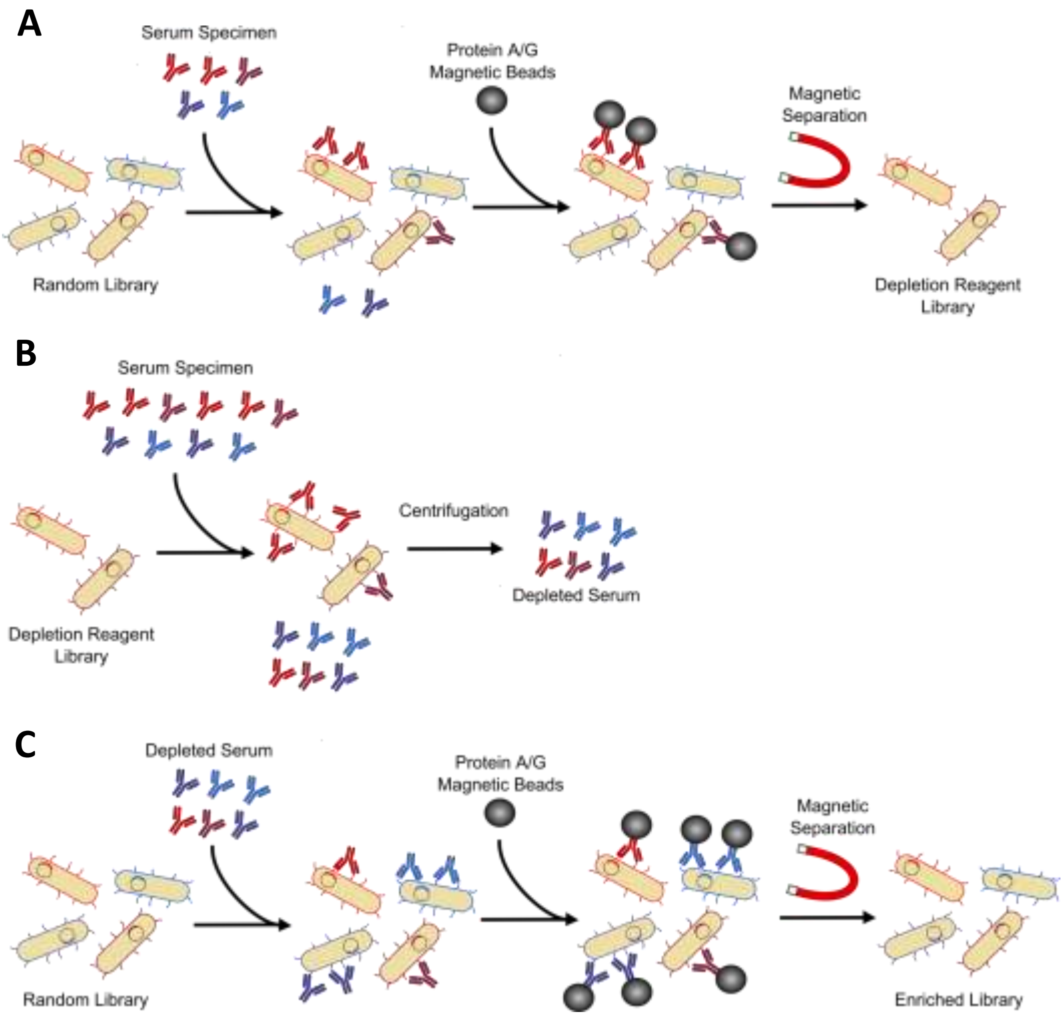


Figure 5.2. High-titer depletion method using bacterial display peptide libraries. (A) A random bacterial display peptide library is screened for antibody binding from serum using magnetic selection. Peptide enrichment will be skewed toward binding high-titer species (red) that dominate the repertoire. **(B)** The enriched library, functioning as a depletion reagent, is incubated with the serum and the unbound fraction, the depleted serum, is recovered. **(C)** The depleted serum is then used to enrich a random bacterial display peptide library for antibody binding, enabling enhanced detection of low-titer antibodies (blue).

5.2.2 Effects of antibody depletion on discovering individual antibody specificities

To determine whether depleting serum of high-titer antibodies improves antibody specificity detection capabilities, a serum specimen (Specimen 1) was spiked with three monoclonal antibodies (mAbs) that have known peptide epitopes (anti-V5, anti-myc, and anti-HA), enabling corresponding peptide enrichments to be monitored. The mAbs were spiked into serum at equal concentrations for a range of concentrations (200, 20, 2, 0.2, and 0.02 nM) to reflect the large dynamic range of titers within the antibody repertoire. A bacterial display peptide library was enriched with each spiked serum sample, as well as serum without spiked mAbs, diluted 1:100. Each enriched library, functioning as a depletion reagent, was then incubated with the corresponding spiked serum sample and the unbound serum fraction was recovered. The depleted sera were then used to enrich new peptide libraries using 1:50 serum dilutions. To allow comparisons with the depletion method, peptide libraries were also enriched using non-depleted spiked sera diluted 1:20. A higher concentration of non-depleted sera was used to demonstrate the effects of high antibody titers on the peptide enrichment process. The compositions of the twelve enriched libraries (six with HTD and six without HTD) were determined by preparing amplicon libraries from isolated plasmid DNA and performing NGS to obtain 1–9×10^6 total sequences and 5–9×10^5 unique sequences per enriched library.

To assess the improvements of discovering antibody specificities with depleted serum, the enrichments of the mAb binding motifs were quantified for each library enriched with spiked serum. The consensus binding motifs for each mAb were previously determined (**Section 2.2.6**) to be [KR]PxPNxxL (anti-V5), LxSEE[DE] (anti-myc), and [YQ]PYD[VI]xD (anti-HA). For V5 and myc, motif enrichment values were increased or maintained with HTD at 200 and 20 nM mAbs (**Figure 5.3A,B**). This observation indicated that the mAbs at these concentrations were not completely removed during the depletion, either because they were not sufficiently high-titer or they were in such abundance that depletion did not sufficiently reduce their concentration. Moreover, motif enrichment was improved due to the extensive polyclonal depletion of numerous other serum antibody species. At 2 nM spiked mAb, V5 and myc motif enrichments with HTD were statistically significant (p <0.001), with 5-fold and 18-fold increases from enrichments without HTD, respectively. Without HTD, HA motif enrichment was observed at as low as 2 nM (**Figure 5.3C**). The increased detection range suggests that the anti-HA mAb has a higher affinity than the other two mAbs. After depletion, HA motif enrichment increased at 200 nM (for the same reasons as V5 and myc) but decreased at 20 and 2 nM. This result suggests that anti-HA mAbs present at concentrations between 2-20 nM were depleted by the depletion reagent because they were effectively high-titer antibodies. Correspondingly, HTD increased the sensitivity of anti-HA detection at the lowest concentrations of 0.2 and 0.02 nM (i.e. increased sensitivity for low-titer antibodies) demonstrating that the removal of abundant serum antibody species enabled the detection of this antibody species at very low concentrations.

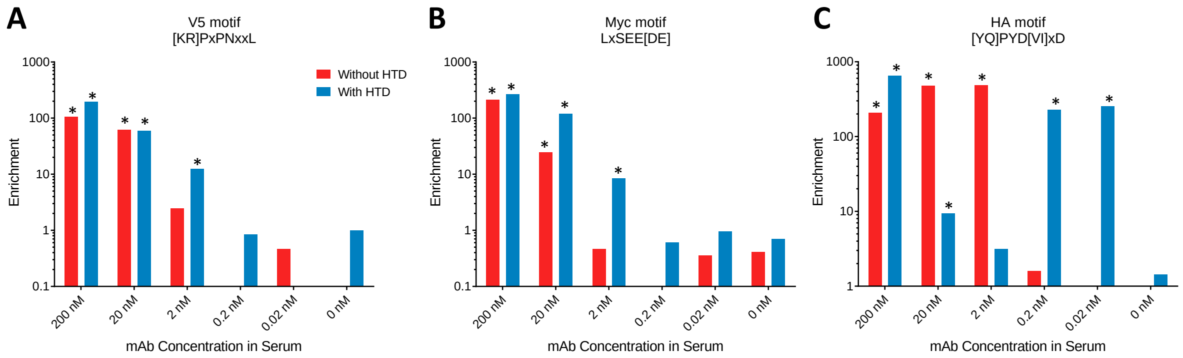


Figure 5.3. High-titer depletion improves mAb motif enrichment at low concentrations. Three mAbs were spiked into a serum sample at various concentrations and used to screen bacterial peptide libraries with and without HTD. Histograms demonstrate enrichment of the mAb motifs at a range of concentrations after screening spiked serum with (blue) or without (red) HTD. **(A)** V5-mAb (GKPIPNPLLGLDST) with motif [KR]PxPNxxL, **(B)** myc-mAb (EQKLISEEDL) with motif LxSEE[DE], and **(C)** HA-mAb (YPYDVPDYA) with motif [YQ]PYD[VI]xD. Enrichment was calculated as the ratio of actual motif observations to expected observations. Statistically significant enrichment (*p <0.001) at each concentration was determined using Poisson distribution statistics based on motif observations and expected observations.

5.2.3 Quantifying the extent of antibody specificity discovery with antibody depletion

To fully investigate the potential benefits of HTD, the richness of the resulting peptide sequence datasets from screening Specimen 1 serum was assessed by determining the number of distinct epitope motif clusters identified using the motif discovery algorithm MEME [131]. Because the MEME algorithm is poorly suited for clustering the large peptide sequence sets of ~millions that result from each serum-specific library enrichment, we reduced the dataset size by only considering peptides that were reproducibly enriched. Without HTD, 146,783 unique peptides were observed at least once in all six replicate libraries (i.e., reproducibly enriched peptides). With this reduced dataset, MEME was used for motif discovery within the top 4,000 most enriched peptide sequences. In total, MEME identified 25 unique motifs (E-value <0.05) representing the serum antibody-binding motifs enriched without HTD. For libraries enriched with HTD, 114,341 unique peptides were observed at least once in each of the six replicate libraries. After MEME analysis of the top 4,000 sequences, 92 unique motifs were identified. By comparing the motifs between the two data sets using a modified PAM30 similarity matrix, 10 motifs were common to both data sets and represented prominent motifs discovered with both methods. Thus, depletion increased the capacity of unique motif discovery by 5- to 6-fold (**Figure 5.4A**). Further analysis of all motifs revealed that 69% of the top 4,000 sequences enriched without HTD belonged to one of the top five motif clusters (**Figure 5.4B**). In contrast, the top five motifs discovered with HTD comprised only 17% of the sequences analyzed, signifying a more uniform distribution of motif enrichment. Taken together these results demonstrate that depletion of high-titer antibodies increased data richness and enabled discovery of a greater number of distinct antibody-binding specificities.

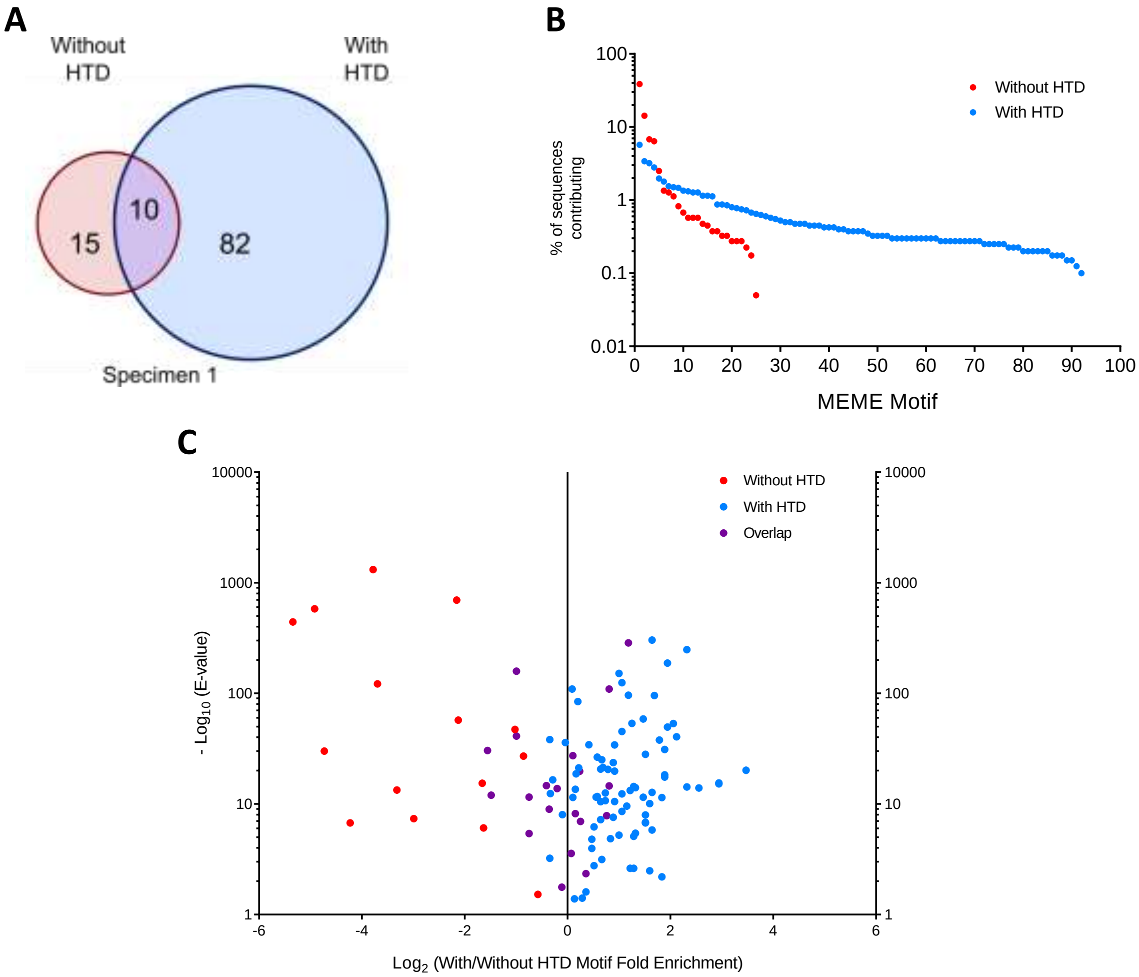


Figure 5.4. High-titer depletion removes dominant species and increases coverage of serum antibody specificities. (A) Motifs were discovered using MEME from peptide libraries screened with (blue) and without (red) HTD. The Venn diagram shows the number of unique motifs (E-value <0.05) discovered by analyzing 4,000 enriched peptides from each library, with shared motifs common to both discovery sets (purple) at the intersection. **(B)** The MEME motifs discovered with and without HTD were ranked by significance (E-value) within each dataset and the percentage of the sequences that contributed to each motif was reported. **(C)** The enrichment of each motif discovered with and without HTD was calculated from each peptide dataset. The log$_2$ ratio of enrichment with HTD to enrichment without HTD was calculated for each motif and plotted against the motif's MEME significance (E-value).

To further quantify the impact HTD had on the discovery of serum antibody-binding motifs, enrichment of each motif within the full sequence datasets was determined with and without HTD. In particular, the ratio of motif enrichment obtained with HTD to that obtained without HTD was calculated for each motif, to assess the relative enrichment achieved by each approach. For the 15 motifs discovered exclusively without HTD, enrichment was decreased ~7-fold on average in libraries with HTD, with four of the five most significant motifs decreasing ~13- to 40-fold, demonstrating substantial depletion of the antibodies that bind these peptides (**Figure 5.4C**). In contrast, the 82 additional motifs discovered using HTD had on average 2-fold greater enrichment than obtained without HTD, verifying that enrichment was broadly improved and not skewed towards any particular antibody species. Lastly, motifs common to both data sets exhibited no average change in enrichment (average fold change ~1) with a maximum fold increase or decrease of 3. These results demonstrated that HTD effectively removed antibodies that otherwise dominate peptide enrichment and that screening depleted serum allowed for the discovery of many more antibody specificities with greater sensitivity.

5.2.4 Optimizing antibody specificity discovery and coverage

While one of the major benefits of HTD is that information regarding which species are abundant is not required prior to depletion, the possibility of exhaustively depleting unbeknownst antibody species may be undesirable for certain applications. Here, "exhaustive depletion" is the removal of an antibody species prior to peptide screening resulting in enrichment indistinguishable from statistical chance. We therefore adjusted the screening conditions to minimize exhaustive depletion of highly abundant species while maintaining increased sensitivity and coverage of low-titer antibody species. To accomplish this, we increased the depleted serum concentration during screening, hypothesizing that the concentrations of highly depleted species in previous experiments were below detection level but not zero. Specifically, a bacterial display peptide library was first enriched toward serum antibodies diluted 1:100. This depletion reagent was then used to deplete the serum sample of antibodies, resulting in depleted serum. A new bacterial display library was then enriched against the depleted serum diluted 1:25, a 2-fold increase in concentration compared to the previous screens of Specimen 1 samples. To allow comparisons with and without HTD, both the depletion reagent (enriched at 1:100) and the library enriched via depleted serum (enriched at 1:25) were sequenced using NGS. Although this library was screened without HTD using a 1:100 dilution instead of the previous 1:20 dilution, we found that enrichments for high-titer samples with 1:100 or 1:20 dilution only vary slightly (data not shown). This process was completed for three additional serum specimens. Libraries screened without HTD had 2–3×10^6 total sequences and 5–6×10^5 unique sequences and libraries screened with HTD had 6–10×10^6 total sequences and 2×10^6 unique sequences per library.

Antibody specificities were discovered by analyzing the 4,000 highest enriched peptide sequences with MEME for each sample with and without HTD. Depletion again resulted in improved discovery capacity with 2- to 3-fold more unique motifs discovered (**Figure 5.5A)**. Without HTD, enrichment profiles were skewed towards few motifs for each specimen with ~40–60% of the sequences analyzed belonging to the top five motifs (**Figure 5.5B**). In contrast, the top five motifs for each depletion sample contained only 8–14% of the sequences analyzed, suggesting depletion allowed a more uniform enrichment. To determine the extent to which high-titer antibody species were depleted, the fold change enrichment was again determined for each motif (**Figure 5.5C**). The motif enrichment analysis for all three specimens revealed that 79% of motifs discovered without HTD had $\log_2$ ratios between -2 and 0, independent of motif significance, with a ~3-fold decrease on average and a maximum decrease of 7-fold. These more modest decreases suggest that antibodies with the highest abundance were not completely removed and were still available for enrichment. Motifs discovered with HTD had larger variations in fold enrichment but over 60% of motifs had increased enrichment with a maximum of 25-fold increase. Although some motifs had moderate decreases in enrichment, the net effect of HTD enabled discovery of these motifs. All shared motifs decreased in enrichment after HTD with an average decrease of 2.5-fold. These results demonstrate that HTD reproducibly improved motif discovery and enrichment while reducing inadvertent, exhaustive depletion of antibody species. However, there exist trade-offs between minimizing exhaustive depletion and maximizing discovery because minimizing depletion inherently reduces the advantageous effects depletion can have on enrichment. Furthermore, these findings indicate that HTD parameters, such as concentration of depleted serum, can be tuned to optimize antibody specificity discovery.

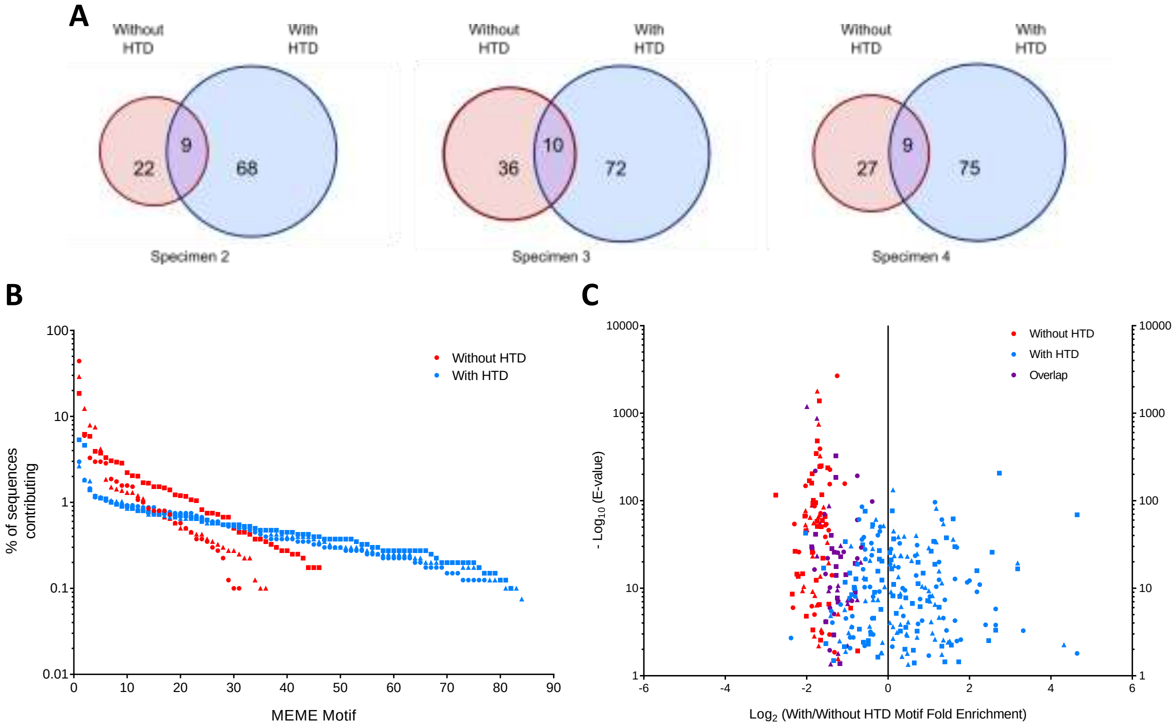


Figure 5.5. Modified HTD conditions reduce exhaustive depletion while maintaining increased coverage of antibody specificities. Three additional serum specimens, Specimen 2 (circles), Specimen 3 (squares), and Specimen 4 (triangles), were analyzed with (blue) and without (red) HTD. The concentration of serum used for screening was increased with HTD to reduce the exhaustive depletion of antibody species. **(A)** The Venn diagrams show the number of unique motifs (E-value <0.05) discovered through MEME analysis of 4,000 highly enriched peptides with and without HTD for each specimen with shared motifs common to both discovery sets (purple) at the intersection. **(B)** The MEME motifs discovered with and without HTD were ranked by significance (E-value) within each dataset and the percentage of the sequences that contributed to each motif was reported. **(C)** For each specimen, the enrichment of each motif discovered with and without HTD was calculated from the two datasets. The $\log_2$ ratio of enrichment with HTD to enrichment without HTD was calculated for each motif and plotted against the motif's MEME significance (E-value) to determine the fold change enrichment.

5.2.5 Identification of high-titer antibody specificities present in multiple specimens

After analyzing individual repertoires for high-titer antibody specificities, which were significantly depleted with HTD, we found that some high-titer specificities were common among multiple specimens. Specifically, the most significant motif from Specimen 1, which was subsequently depleted 13-fold following HTD, was also the most significant motif from Specimens 2 and 4. Of the 4,000 highest enriched sequences analyzed for motif discovery from these three samples, 29–44% of the sequences contained this motif. This suggests that there may exist high-titer specificities common among human specimens that could be utilized for a targeted HTD approach.

To identify high-titer specificities present in multiple specimens, libraries from Specimens 2, 3, and 4 were analyzed, as these three samples were analyzed using the same HTD conditions. From libraries analyzed without HTD, there were 31,631 peptides observed in at least two of the three specimens. Of these shared peptides, the 4,000 highest enriched sequences were analyzed for motif discovery using MEME. In total, 30 unique motifs (E-value <0.05) were discovered, representing specificities common to at least two of the three specimens. To determine if these motifs corresponded to high-titer antibodies and were subsequently depleted after HTD, the enrichments for the 10 most significant motifs were evaluated with and without HTD (**Figure 5.6**). Motif enrichment decreased at least 2-fold for five or six motifs in each subject library after applying HTD, while enrichment increased over 2-fold after HTD in only one motif (Motif 4, Specimen 4). These findings indicate that the serum antibodies that bound these motifs were high-titer and present in multiple human specimens and were effectively depleted following HTD. Therefore, if enough common, high-titer specificities like these exist, it is possible that a targeted depletion approach could be taken based on catalogued high-titer specificities to improve the efficiency of the depletion process and reduce inadvertent depletion.

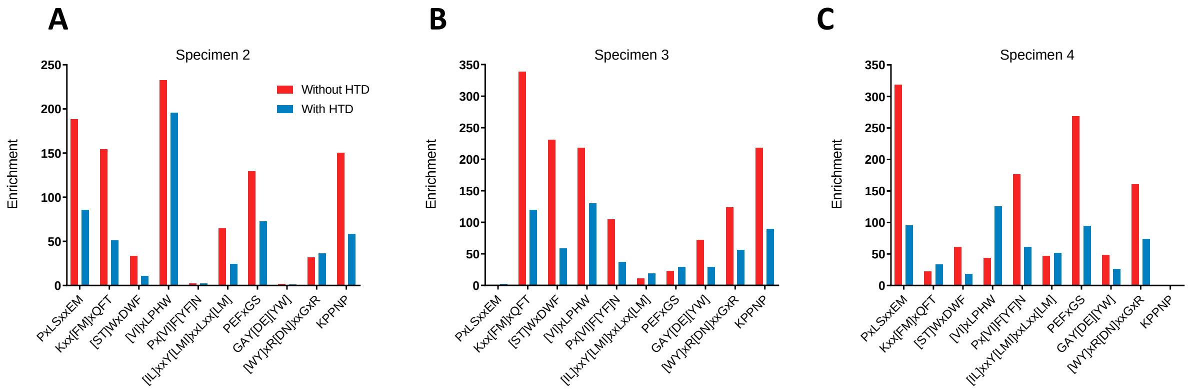


Figure 5.6. High-titer specificities were identified to be present in multiple specimens. Peptides observed in at least two of three libraries from **(A)** Specimen 2, (**B)** Specimen 3, and (**C)** Specimen 4 obtained without HTD were analyzed for common motifs. The 4,000 most enriched peptides were input for MEME motif discovery, and the 10 most significant motifs ranked by E-value were selected for analysis. Enrichments of the 10 motifs in specimen libraries obtained with and without HTD were determined to observe the effects of HTD on the shared motifs and to classify specificities as corresponding to common, high-titer antibodies.

5.3 Discussion

While advances have been made in technologies capable of interrogating the antibody repertoire with protein and peptide libraries, fundamental hurdles remain for identification of biomarkers and antigens associated with disease. Specifically for peptide libraries, the focus for improving antibody screening has largely been on design and construction [37,200], avidity-effects [54,66], library composition and complexity [42,201,202], and bioinformatic analysis [33,56,76]. However, the complexity and diversity of the antibody repertoire itself presents a significant challenge for discovering specificities, as rare or infrequent species need to be reliably detected and distinguished from abundant species. Other areas of proteomics research have overcome similar difficulties by employing methods to reduce sample complexity prior to analysis, such as bulk removal of abundant serum proteins [196,203]. Therefore, we developed a general method (HTD) to deplete sera of antibody species in proportion to titer prior to screening to reduce antibody repertoire complexity.

When analyzing the antibody repertoire with peptide libraries, removing antibody species with high abundance and affinity reduces competition for binding peptides that contain multiple motifs and decreases the proportion of sequences from NGS that correspond to the most dominant species. Collectively, these effects enable increased coverage of target binding and quantification. For individual species, three spiked mAbs were each detected at lower concentrations with HTD through significant increases in mAb motif enrichment, demonstrating improved detection limits. The mAb analysis also confirmed that HTD has varying effects on individual antibody species dependent upon concentration, affinity, and specificity. For instance, enrichment of the anti-HA motif was observed at two orders of magnitude lower concentration with HTD, suggesting that high affinity species could be detected at very low concentration. Focusing on the polyclonal serum antibodies from an individual specimen, the unaltered repertoire was dominated by only a few specificities. This is consistent with antibody repertoire sequencing where relatively few clonotypes dominate the repertoire, especially following an immune challenge [6,7]. After depletion, enrichments of these motifs were substantially reduced suggesting the corresponding antibody species were sufficiently depleted which enabled the discovery of over 5-fold more unique motifs. This increased depth and breadth of discovery could reveal previously inaccessible antibody specificities relevant for diagnostics and provide insights into disease etiology.

Implementing HTD inevitably brings about trade-offs between gaining access to rare interactions while potentially losing some common interactions. For applications of biomarker or antigen discovery, the loss of information regarding abundant species in return for low-titer specificities may be advantageous. However, this loss of information may be undesirable for other applications and a strategy for minimizing this possibility would be beneficial. The extent of depletion for any antibody species is dependent on several factors. Some factors are species-specific such as antibody affinity and specificity (and therefore prevalence of motifs in the peptide library) demonstrated by the varied effects that depletion had on the three mAbs tested at equal concentrations. Other factors are adjustable like serum concentration as well as concentration of pre-enriched peptides (the depletion reagent) used to physically deplete the serum. Therefore, we sought to mitigate inadvertent depletion by tuning the depletion screening parameters. Specifically, we increased the concentration of depleted serum during peptide screening, thereby increasing enrichment of all antibody species including those that were significantly reduced via depletion. This depletion strategy, applied to three specimens, exhibited only moderate depletion of the highest-titer species yet still increased the discovery of unique specificities. Collectively, these results confirm that the extent of depletion can be adjusted and illustrate the trade-offs of depletion: the presence of high-titer species affects the detection of low-titer species.

While certain parameters can be adjusted to optimize the depletion of a given sample, the complexity and diversity of the antibody repertoire may make it challenging to devise a universal depletion scheme. One strategy could be to compile a set of peptides specific to antibody species that are relatively common throughout the population for use as depletion reagents. This would minimize off-target depletion. For example, peptides could be used to remove antibody species that target common viruses such as the Epstein–Barr virus which infects over 90% of adults [204]. Interestingly, it was found that Specimens 1, 2, and 4 had the same motif ranked as most significant when analyzing the repertoires without HTD. This consensus motif, P[SA][LF]SxxE[MTS], occupied a dominant share of each specimen’s enriched peptide library. This motif is likely from the enterovirus family which has been studied and characterized for decades and is known to be highly immunogenic [205–207]. Further investigation of common specificities readily revealed additional motifs that were determined to be highly enriched in multiple specimens and subsequently depleted after applying HTD. Therefore, it is conceivable that a panel of peptides targeting various high-titer antibody species like these could be catalogued, constructed, and employed for sera depletion. However, it is unclear whether this strategy would be robust enough to sufficiently deplete samples throughout a diverse population. This highlights the importance of a depletion method that is independent of the specimen and simply depletes species relative to titer. If inadvertent depletion is a serious concern, a solution is to simply combine the peptide data sets with and without HTD, maximizing coverage of the antibody specificity repertoire. This concept has also been suggested in other areas of proteomic biomarker discovery [208]. This eliminates the concern for unintentional depletion and allows for maximizing depletion leverage. For instance, combining both anti-HA mAb enrichment data sets (with and without HTD) would lead to detection over at least five orders of magnitude (0.02-200 nM). Because HTD in this study already requires the enrichment of a peptide library toward the abundant fraction of serum antibodies, and NGS is becoming more accessible and affordable, sequencing and combining both library data sets could be a reasonable, and perhaps desirable, option.

Although HTD was applied to the analysis of the antibody specificity repertoire using bacterial display, the general concept of depleting species in relation to abundance and affinity could be extended to other applications. Simplifying a biological sample by depletion prior to analysis is not new to biotechnology but has been limited to depleting species known to be present *a priori* and has yet to be extended to more heterogeneous mixtures. The HTD method presented here accompanies the shift toward personalized medicine with a robust depletion method specific to individual subject specimens and provides a framework for applying novel depletion strategies to biological samples. As high-dimensional data technologies and immunology continue to merge [209], this technique could greatly expand our ability to probe the complex and dynamic interactions of our immune systems and environments.

5.4 Materials and methods

5.4.1 Bacterial display peptide library screening for antibody binding

Bacterial display peptide library screening for antibody binding with NGS was carried out as described in **Section 2.4.2**. Briefly, a 12-mer random *E. coli* display peptide library (trimer-X_{12}) was screened for antibody binding using a human serum specimen previously depleted of *E. coli* binders [70,76]. Peptides bound by antibodies were enriched with two rounds of magnetic selection using protein A/G magnetic beads (Thermo Scientific Pierce) at a bead to cell ratio of 1:100 for the first round and 1:1 for the second round. Following magnetic selection, plasmid DNA from the enriched cellular libraries was extracted and amplicons containing the peptide region were constructed using a two-step PCR based on the Nextera XT (Illumina) protocol and sequenced as previously described [76]. For libraries screened without HTD, serum specimen 1 was screened at a 1:20 dilution in PBST while serum specimens 2, 3, and 4 were screened at a 1:100 dilution. All subjects provided informed consent and were adults with no known immunological conditions at sample collection. Specimens were de-identified and acquired according to institutional guidelines. Serum specimens were stored at -80 °C in long-term storage cryogenic vials (Nalgene).

5.4.2 High-titer serum antibody depletion

To deplete a serum sample of high-titer antibodies, a bacterial display peptide library was first enriched towards the specimen as described in **Section 2.4.2**. A frozen aliquot of the enriched library cells was inoculated at 1:40 in 10 mL of LB supplemented with 34 µg/mL chloramphenicol (CM), grown to an OD_{600} of 0.4–0.6 at 37 °C with orbital shaking (250 rpm), and induced with 0.02% wt/vol L(+)-arabinose. After 1 hr of induction, cells were centrifuged at 3,000 rcf for 5 mins, washed with PBST, and resuspended with *E. coli* depleted serum diluted 1:10 with PBST. For all serum depletions used in this study, 2×10^9 induced library cells (depletion reagent) were used per 300 µL of 1:10 *E. coli* depleted serum. Samples were incubated overnight at 4 °C with gentle mixing on an orbital shaker (20 rpm). Unbound antibodies were separated from antibodies bound to peptides by centrifugation of the sample at 5,000 rcf for 5 min and extraction of the supernatant, the depleted serum. This separation process was completed three times to ensure no library cells remained in the depleted serum. The depleted serum was then used for a new peptide library screen. For libraries screened with HTD, serum was depleted with a corresponding library screened with 1:100 diluted serum. Depleted serum was then used at 1:50 for Specimen 1 library screens and 1:25 for Specimen 2, 3, and 4 library screens.

5.4.3 Monoclonal antibody spike-in

Three rabbit mAbs were used to screen peptide libraries for mAb binding and identify specificities: V5-Tag D3H8Q (GKPIPNPLLGLDST), myc-Tag 71D10 (EQKLISEEDL), and HA-Tag C29F4 (YPYDVPDYA) (Cell Signaling Technology). For Specimen 1 serum samples, each mAb was added at equal concentrations to mimic antibodies found in serum at various concentrations (200, 20, 2, 0.2, 0.02, and 0 nM). Specifically, the three mAbs were added to *E. coli* depleted Specimen 1 sera and then further diluted for peptide library screening without HTD or depleted against a corresponding enriched peptide library and used for HTD screening.

5.4.4 Determining antibody-binding specificities

The motif discovery algorithm MEME [131] was used to identify antibody-binding specificities (motifs) from peptide sequence datasets. Because MEME is not suitable to analyze large NGS datasets, the MEME algorithm was run locally with a maximum of 5,000 sequences. The consensus binding motifs for each mAb were determined previously (**Section 2.4.7**). With the mAb motifs defined as the regular expressions discovered via MEME, peptide datasets with and without HTD from Specimen 1 spiked screens were searched for these motifs to determine mAb motif enrichment in spiked sera. Motif enrichment was defined as the ratio of actual motif observations in the library to the expected observations, where expected observations were calculated using the amino acid frequencies observed in the library and assuming all positions were independent of one another. Poisson distribution statistics were used to calculate the p-value of observing a motif given the expected observation value (**Section 2.4.8**).

For serum antibody-binding specificities, the 4,000 most enriched peptides for each sample were used as input for MEME motif discovery (minimum width of four). Regular expressions for serum antibody-binding motifs were defined from MEME but were restricted to the five most significant positions. Although sequences exhibiting the highest number of observations (counts) were used for motif discovery, the motif enrichments were calculated based on observations of the motifs in the full peptide dataset. However, for serum antibody-binding motif discovery from Specimen 1, only peptides enriched in each library (screened with serum spiked with mAbs at various concentrations) were used to reduce the peptide dataset size and avoid analyzing mAb-binding peptides. Enrichment values for motifs of interest were then averaged among the sample libraries, with or without HTD.

5.4.5 Aligning and scoring motifs for similarity

To compare two motifs and quantify the similarity, the two motifs were aligned and a similarity score was calculated using a modified PAM30 similarity matrix [76]. A similarity score greater than 12 was used to classify two motifs as similar. For each library, motifs discovered with MEME were scored for similarity to reduce redundancy within the discovery list due to subtle variations. If two motifs within a motif set were similar, the motif with lower significance (higher E-value) was removed. The resulting motifs were designated as unique. To identify motifs common to both sets, with and without HTD, motifs were scored for similarity and enrichments of motifs designated as similar were averaged for any calculations.

6 Mapping antibody binding to epitopes using multiplexed epitope substitution analysis

A more complete understanding of antibody-binding epitopes would aid the development of diagnostics, therapeutic antibodies, and vaccines. However, current methods for mapping the binding to epitopes require targeted experimental approaches, which limit throughput. To address these limitations, we developed Multiplexed Epitope Substitution Analysis (MESA) which can rapidly characterize target epitopes from millions of antibody-binding peptides. We selected peptides from a random 12-mer library that bound to human serum antibody repertoires and determined their sequences using next-generation sequencing (NGS). Next, we evaluated the enrichment of all 5- and 6-mers in the peptide dataset. Computationally, we divided target epitope sequences into overlapping k-mers. Then, the positions in each k-mer were substituted with all 20 amino acids and the enrichments of the substituted k-mers were quantified in the peptide dataset. The substitution data for overlapping k-mers spanning the target epitope were compiled to identify substitutions favored for binding at each position in the target epitope, revealing the precise binding motif. This analysis can be completed rapidly for numerous target epitopes from a single peptide dataset, multiplexing epitope binding analysis. To validate MESA, we generated binding motifs for monoclonal antibodies spiked into a serum specimen, recovering the expected binding positions and amino acid preferences. To characterize epitopes bound by a population, we analyzed 50 serum specimens to determine the consensus binding motifs within various target epitopes including known pathogen epitopes. Binding motifs identified by MESA agreed with those discovered using alternative computational approaches. However, MESA's ability to utilize the full depth of NGS datasets enabled the identification of an Epstein–Barr virus binding motif that was not discovered with alternative approaches. These results demonstrate that MESA can rapidly identify binding motifs for multiple epitopes in parallel to enhance our understanding of antibody interactions and characterize epitopes.

*This chapter was co-authored with Michael L. Paull (UCSB, ChE) with equal contributions.

6.1 Introduction

The capability to map the binding of antibodies to epitopes has become essential in applications ranging from basic research to therapeutic and diagnostic development. For example, therapeutic antibody development requires precise determination of epitopes to achieve desired biological activity [210] and to avoid undesired cross reactivity [211]. Similarly, epitope characterization can aid in the development of vaccines that generate neutralizing antibody responses by identifying epitopes that are immunodominant, variable, or hypervariable in infections such as HIV [212]. Additionally, the performance of many antibody-based diagnostic tests is limited by knowledge of the most sensitive and specific epitopes [139]. Finally, the identification of epitopes can yield more effective affinity-capture reagents for research applications [213].

While the gold standard for characterizing antigen epitopes is X-ray crystallography [214], epitope mapping methods using substitution analysis and peptide libraries have become commonplace [141,215]. To determine the extent that each position in an epitope contributes to binding, alanine scanning mutagenesis has commonly been employed [42]. Extending this approach, exhaustive mutagenesis can be used wherein each position in an epitope is mutated to all amino acids [216]. Given the labor-intensive nature of these methods, there remains a need for more efficient, high-throughput methods to characterize and map antibody binding to epitopes.

Peptide microarrays have been used extensively to determine antibody specificities [217]. Here, the antigen is tiled into overlapping peptides and exhaustively mutated. The importance of each amino acid for antibody binding in the entire protein sequence can then be inferred from extent of antibody binding to each peptide. Unfortunately, since this method uses a targeted library, additional microarrays must be prepared for each protein of interest. For the analysis of many antigens, or antibodies within unknown antigens, this process becomes impractical.

To address the limitations of targeted epitope characterization approaches, we developed a method termed Multiplexed Epitope Substitution Analysis (MESA), which utilizes random peptide libraries, NGS, and bioinformatics to simulate exhaustive mutagenesis of arbitrary epitopes. First, we selected antibody-binding peptides from a large surface-displayed peptide library. Next, we evaluated the enrichment of all 5-mers or 6-mers in the antibody-binding peptide dataset. We then divided target epitope sequences into overlapping k-mers and evaluated the enrichments of the k-mers and all possible single-amino acid substitutions. By compiling the segment analyses that span the epitope, we determined the effect of amino acid substitution at each position in the epitope and revealed the binding positions and amino acid preferences within the epitope. Because MESA utilizes millions of peptides selected from a random library, many protein epitopes can be characterized simultaneously.

6.2 Results

6.2.1 MESA maps antibody binding to epitopes using random peptide libraries

MESA characterizes antibody binding to linear targets through random peptide library screening. MESA determines binding motifs of target epitopes that indicate which positions are conserved or variable and which amino acids are preferred at each position. We adapted the algorithm ArrayPitope [217] to use peptide sequences derived from random peptide libraries, rather than microarray binding data. By using random libraries, MESA can examine numerous target sequences via substitution analysis to determine binding motifs (**Figure 6.1**). In this context, a binding motif can be visually represented as an "epitope logo" that reveals amino acid preferences at each position in a target sequence. Large peptide sequence datasets are obtained by enriching a random peptide library for serum antibody binding and identifying the enriched peptides using next-generation sequencing (NGS) [76]. The epitope region of interest (the target sequence) is first transformed into k-mer sequences (5- or 6-mers) with one amino acid overlap spanning the entire sequence. Then, the enrichment of each k-mer in the antibody-binding peptide dataset is calculated. Each position of the overlapping k-mers is substituted with each amino acid and the enrichments of substituted k-mers are calculated to determine the effect of amino acid substitution. Finally, the effects of substitution on each k-mer are compiled to determine the effects of substitution in the whole target sequence.

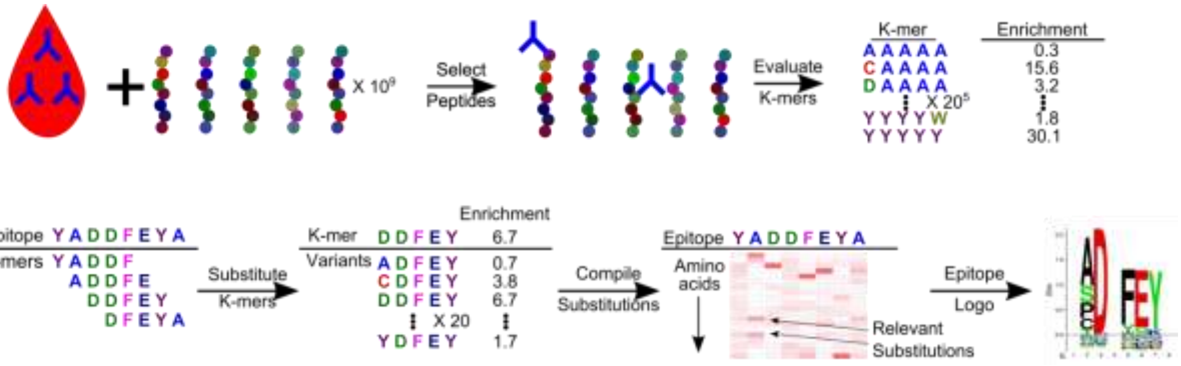


Figure 6.1. An overview of Multiplexed Epitope Substitution Analysis (MESA). A random peptide library is screened for antibody binding to serum specimens. Peptide sequences are determined and the enrichments for all k-mers of a set length are calculated. A target epitope is divided into k-mers with a single amino acid overlap. For each k-mer, every position is substituted with all amino acids to generate 20 variants. Enrichments for all k-mers are then compiled to determine statistically significant positions and valid substitutions in the epitope. The amino acid preferences for each position in the epitope are displayed in an epitope logo. Positions which were not important for binding are blank.

6.2.2 Determining binding motifs for monoclonal antibodies with known epitopes

To validate MESA, we analyzed antibody species with known linear epitopes to determine binding motifs. First, two monoclonal antibodies (mAbs), anti-myc and anti-HA, were spiked into a human serum specimen at 200 nM each. The linear epitopes for anti-myc and anti-HA are EQKLISEEDL and YPYDVPDYA, respectively. We identified 619,527 12-mer antibody-binding sequences by screening this specimen. MESA was applied to each linear mAb epitope by dividing the epitopes into 5-mers and evaluating enrichments in the peptide dataset. The results generated by MESA are visually displayed with alignment heatmaps (**Figure 6.2A,C**) and epitope logos (**Figure 6.2B,D**). The relative frequencies of amino acids at each position in the target sequence are displayed in an epitope logo, wherein the total height of letters at a position represents the importance to binding of that position and the height of individual letters reflects the amino acid preference. Blank positions indicate the position was not statistically significant. MESA also generates a regular expression that represents the epitope logo. For each epitope, alignment heatmaps were generated to visualize the importance of positions within each k-mer. In an alignment heatmap, the substituted and target k-mer enrichments are summed for each position in the k-mers. Lower values represent a greater effect of substitution at a position because at an insignificant position, all substituted k-mers will have the same enrichment, leading to a 20 times larger enrichment total than for a significant position. Additionally, the contribution of each k-mer to the epitope logo is proportional to its total enrichment, which corresponds to the total of each row.

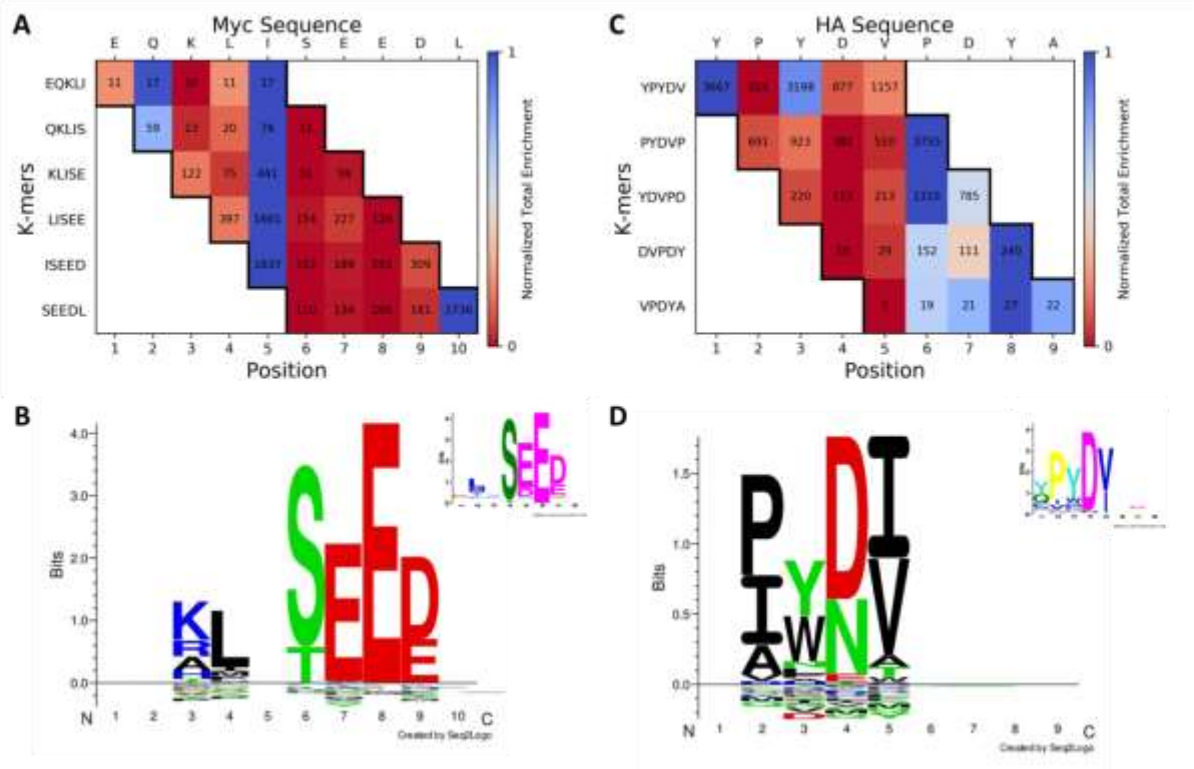


Figure 6.2. MESA determined binding motifs for monoclonal antibody epitopes. A random bacterial display peptide library was screened against human serum spiked with anti-myc and anti-HA mAbs, each at 200 nM, resulting in 619,527 12-mer sequences. MESA was applied to this dataset to generate an **(A)** alignment heatmap and **(B)** epitope logo for the myc target EQKLISEEDL, and an **(C)** alignment heatmap and **(D)** epitope logo for the HA target YPYDVPDYA. MESA used 5-mers, a 10% score threshold, and a 25% minimum enrichment threshold. In alignment heatmaps, the substituted and target k-mer enrichments are summed for each position in the k-mers. Lower values represent a greater effect of substitution at a position. For epitope logos, the absolute height of letters at a position represents the relative effect on binding due to amino acid substitution and the height of individual letters reflects the binding preference at that position relative to the initial target sequence. MEME sequence logos (insets) were obtained via MEME analysis of 5,000 sequences obtained from libraries screened with each mAb in PBST.

For the two mAbs tested, MESA identified distinct amino acid preferences at each position that were significant for binding. For example, position 8 of the myc epitope logo (**Figure 6.2B**) was highly conserved for binding with almost exclusive preference for glutamic acid (E), while position 9 was half as conserved with roughly equal preferences for aspartic acid (D) and glutamic acid (E). The regular expressions corresponding to the binding motifs were KLxSEE[DE] and [PI]Y[DN][IV] for myc and HA, respectively. To validate these epitope logos, we determined mAb binding motifs by screening each mAb spiked into buffer at 20 nM. After NGS, 5,000 peptides with the highest observations from each library were analyzed using MEME to identify sequence logos for each mAb (**Figure 6.2B,D insets**). While MEME and MESA identified similar binding motifs for the mAbs, only MESA could identify binding motifs for mAbs spiked into serum due to MEME's sequence input limitations. Thus, MESA precisely determined the mAb binding motifs despite the presence of background serum antibodies.

6.2.3 Applying MESA to identify binding motifs from a single specimen

When exact antibody targets are unknown, MESA can instead be used with single peptides that were enriched for antibody binding. MESA can then reveal epitopes within the enriched peptides. We analyzed a single serum specimen library, which had 364,411 antibody-binding peptides. Epitope logos and alignment heatmaps were generated for two of the most observed peptide sequences in the library, YADVFEYQYDWP (P1) and TWRDWWSKQPFQ (P2) with 1,429 and 611 observations, respectively (**Figure 6.3**). For P1, the regular expression was ADxFEY which, along with the alignment heatmap and epitope logo, indicated strong amino acid preferences at all positions except position 3 (**Figure 6.3A,B**). MEME analysis of the 5,000 highest enriched antibody-binding peptides revealed a motif with a highly similar logo (**Figure 6.3B inset**), suggesting that using MESA with the target peptide P1 converged upon the true binding motif. Similar success was obtained for P2, which had a regular expression of S[WF][KR]xW[FYW] and an epitope logo that was nearly identical to the MEME sequence logo (**Figure 6.3C,D**). Notably, even though the P2 target peptide TWRDWWSKQPFQ contained a tryptophan (W) at position 6, MESA accurately identified the preference for phenylalanine (F) and tyrosine (Y) as well.

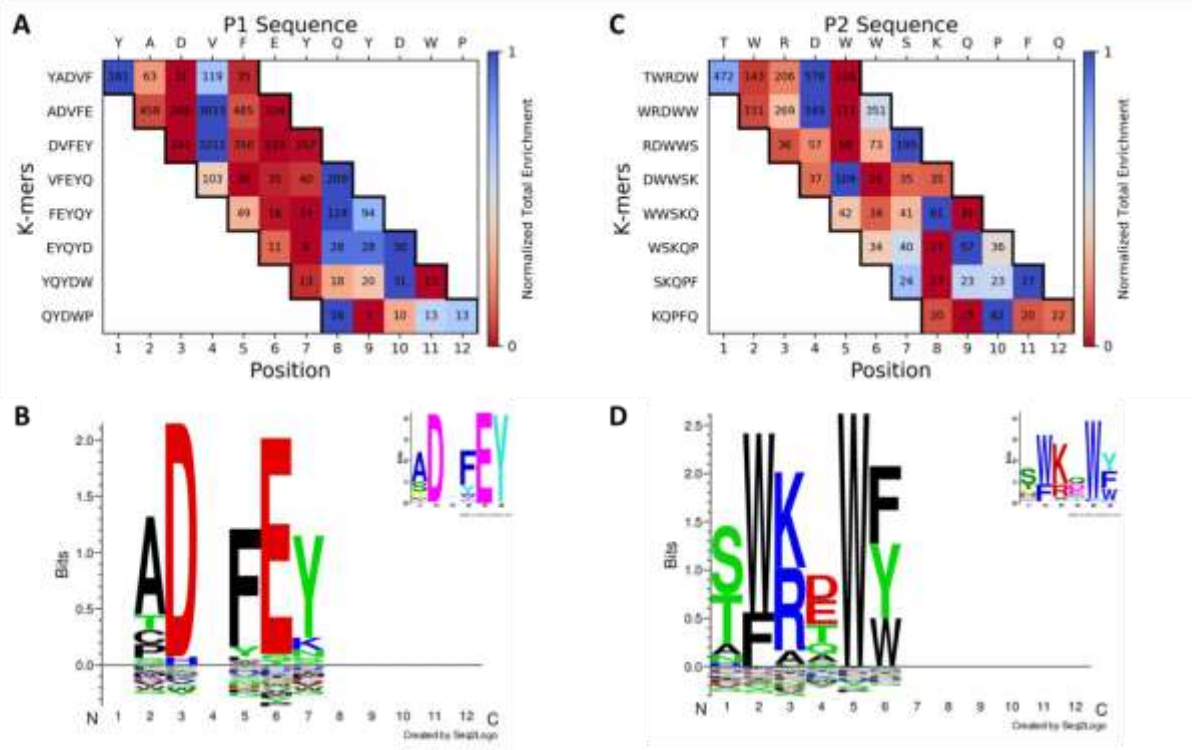


Figure 6.3. MESA identified binding motifs for antibodies in an individual serum specimen. We analyzed two highly enriched antibody-binding peptides using MESA with a peptide library from a serum specimen containing 364,411 sequences. MESA generated an **(A)** alignment heatmap and **(B)** epitope logo for the P1 target YADVFEYQYDWP, and an **(C)** alignment heatmap and **(D)** epitope logo for the P2 target TWRDWWSKQPFQ. MESA used 5-mers, a 10% score threshold, and a 25% minimum enrichment threshold. MEME sequence logos (insets) were obtained via MEME analysis of 5,000 sequences obtained from the specimen library.

6.2.4 Identifying binding motifs using multiple serum specimens

By using large peptide datasets from multiple specimens, MESA can identify binding motifs that represent a population. Peptide libraries screened against eight individual serum specimens were sequenced to obtain a total of 1×10^7 sequences. This dataset was sufficiently large that MESA could utilize 6-mers for increased resolution relative to 5-mers. Due to the large sequence dataset from using multiple samples, MESA parameter stringencies for determining statistical significance were increased. From the dataset, the two peptides with the most observations in at least six specimen libraries, DPYLPHWSTVEV (P3) and KYAFPQRIFVSS (P4), were used as targets for MESA (**Figure 6.4**). For P3, MESA identified the regular expression as LPHW, with highly conserved residues at positions 4–7 (**Figure 6.4A,B**). The epitope logo for P3 agreed with the MEME sequence logo determined using all 1,865 sequences present in at least six of the eight libraries. For P4, the regular expression was KxxFPQx[IV], in strong agreement with the MEME sequence logo (**Figure 6.4C,D**). Despite the diverse dataset of $>1\times10^7$ sequences, MESA accurately characterized the binding of two antibody species present in multiple serum specimens.

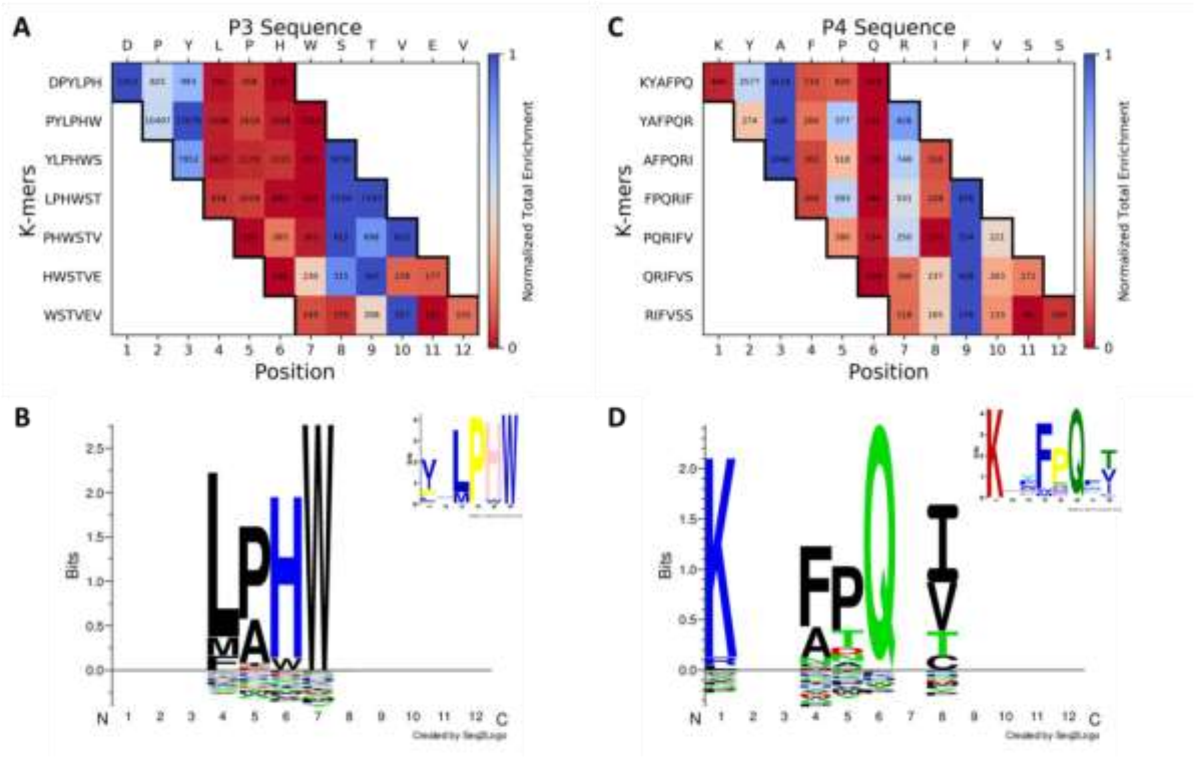


Figure 6.4. Consensus binding motifs from multiple specimens were identified using MESA. Libraries screened against eight individual serum specimens were sequenced to obtain a total of 1×10^7 12-mer sequences. The two peptides with the most observations in at least six specimen libraries, DPYLPHWSTVEV (P3) and KYAFPQRIFVSS (P4), were used as targets for MESA. MESA generated an **(A)** alignment heatmap and **(B)** epitope logo for the P3 target DPYLPHWSTVEV, and an **(C)** alignment heatmap and **(D)** epitope logo for the P4 target KYAFPQRIFVSS. Due to the larger sequence dataset from eight specimens, MESA used 6-mers, a 5% score threshold, and a 50% minimum enrichment threshold. To validate the MESA epitope logos, motifs were identified by MEME analysis of all 1,865 sequences observed in at least six of the eight libraries (insets).

To validate MESA with an even larger population of specimens, we analyzed epitope sequences from common antigens using 50 specimens (>1×10^8 sequences). We utilized MESA to generate the epitope logo of the common epitope SGSPPRRPPPGRRPFFHPVG from Epstein–Barr virus nuclear antigen 1 (EBNA1) [218] (**Figure 6.5A**). The regular expression generated for this epitope was RRP[FW]FHP, which was highly enriched (enrichment >10) in 64% of the specimen libraries. The EBNA1 motif RRPFF has been found in multiple previous analyses [41,219,220]. The binding motif for another EBNA1 epitope, EADYFEYHQEGGPDGEPDVP [219], was also generated using MESA (**Figure**

6.5B). The regular expression for this epitope was ADYxEY, which is the same specificity identified with MESA using target P1 (**Figure 6.3**). This motif was highly enriched in 30% of the specimen libraries. The third epitope analyzed, VPEFKGSLP, was from the *Staphylococcus aureus* antigen extracellular matrix protein-binding protein emp [60]. MESA generated the binding motif for this epitope and identified the regular expression VPEFxG[AS], which was highly enriched in 92% of libraries (**Figure 6.5C**). As additional validation, we conducted a MEME analysis of 5,000 sequences observed in at least 10 of the 50 specimen libraries. This analysis revealed motifs that corresponded to the binding motifs determined with MESA (**Figure 6.5 insets**). MESA therefore identified consensus binding motifs from public epitopes through substitution analysis of a multi-subject library with $>1\times10^8$ sequences. This demonstrates the power of MESA to efficiently mine large datasets for antibody binding characterization.

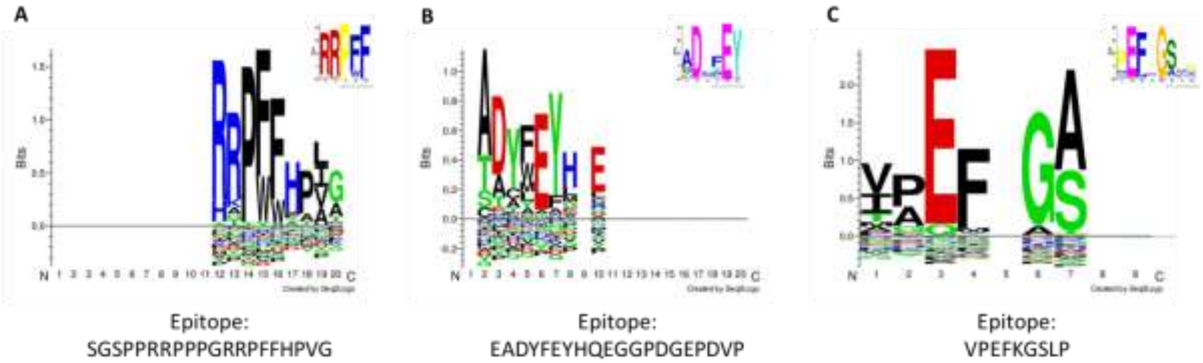
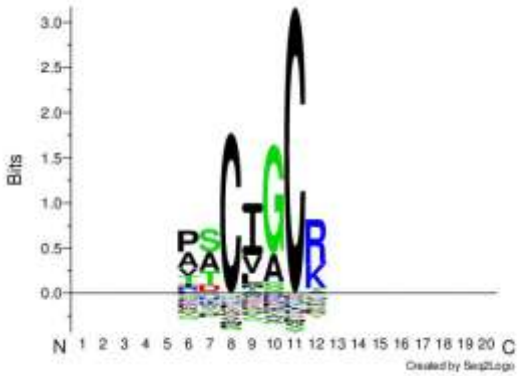


Figure 6.5. MESA determined binding motifs of public epitopes from a large population. Libraries from 50 specimens ($>1\times10^8$ sequences) were analyzed with MESA to determine the binding motifs of three common epitopes: **(A)** the EBNA1 epitope SGSPPRRPPPGRRPFFHPVG, **(B)** the EBNA1 epitope EADYFEYHQEGGPDGEPDVP, and **(C)** the extracellular matrix protein-binding protein emp epitope VPEFKGSLP from *Staphylococcus aureus*. MESA utilized 6-mers, a 2.5% score threshold, and a 50% minimum enrichment threshold for **(A)** and **(B)**, but a 25% enrichment threshold was used for **(C)** due to the shorter epitope length. MEME sequence logos (insets) were obtained via MEME analysis of 5,000 sequences observed in at least 10 of 50 specimen libraries.

Importantly, MESA can generate binding motifs for less common epitopes that are represented in a small fraction of the sequence datasets, and would therefore be difficult to discover using algorithms like MEME. We utilized MESA to determine the binding motif for another common EBNA1 epitope, RPQKRPSCIGCKGTHGGTGA [218] (**Figure 6.6**). We determined the regular expression for the binding motif as CIGCR, but we did not identify a corresponding motif using MEME. However, we determined CIGCR to be highly enriched in 34% of specimen libraries, suggesting this epitope was bound by a significant proportion of the population.



Epitope:
RPQKRPSCIGCKGTHGGTGA

Figure 6.6. A binding motif from a EBNA1 public epitope was discovered exclusively using MESA. Libraries from 50 specimens (>1×10^8 sequences) were analyzed with MESA to discover a binding motif for the EBNA1 epitope RPQKRPSCIGCKGTHGGTGA. MESA utilized 6-mers, a 2.5% score threshold, and a 50% minimum enrichment threshold. MEME was not able to identify a corresponding motif due to the limited number of input sequences.

6.3 Discussion

Here, we introduced an algorithm for determining the most important positions and amino acid preferences in epitopes using random peptide library screening. Notably, this method enables the identification of binding motifs for arbitrary epitopes, avoiding the need to bias experiments. MESA computationally substitutes individual positions in an epitope and calculates the effect of the substitutions. Positions which are the most affected by mutations are plausibly important to antibody binding. Because the large, random bacterial display library has ~100% coverage of 5-mers and ~96% coverage of 6-mers, this approach can utilize k-mers to precisely evaluate the effect of a substitution in a peptide. It would be difficult to correctly identify binding motifs for approaches with smaller libraries such as microarrays with low k-mer coverage [56]. By analyzing known epitopes for mAbs spiked into serum, we demonstrated that only a few positions dominated binding for an antibody species. Additionally, by analyzing antibody-binding peptides and known pathogen epitopes, we could refine epitopes into shorter motifs representing the most important binding positions. We focused on common epitopes from antigens such as EBNA1 because the specimens were from a general population where Epstein–Barr virus infects over 90% [204]. We were able to validate our results by comparison to MEME [221], with the exception of the epitope CIGCR (**Figure 6.6**) which was exclusively identified using MESA and has been previously reported [75]. This is unsurprising since we could only input ~5,000 sequences to MEME, which for the single specimen and 50 specimen analyses was 1% and 0.005% of all sequences, respectively. Thus, MESA can more thoroughly explore the depth of NGS datasets for epitope characterization. Another limitation of using algorithms such as MEME to find significant positions in a target sequence is that it requires the cumbersome process of searching through motifs for similarity to the target. In contrast, MESA simply substitutes a target sequence and returns the binding motif. While the motif discovery algorithm IMUNE efficiently searches full NGS datasets [76], it is primarily designed to analyze groups of specimens for disease-specific motifs, while MESA can be utilized for mapping antibody binding to target epitopes in single specimens or populations.

MESA has significant advantages over other epitope mutagenesis approaches. Targeted mutagenesis schemes like ArrayPitope tile peptides from a protein sequence in a microarray format, and then mutate individual positions to all other amino acids [217]. The main issue with using a non-random library is that experiments must be repeated for each additional protein target. While researchers have used random bacteriophage libraries to identify epitopes [60], targeted libraries were required to refine epitopes. Non-random libraries can also lead to inaccurate results if serum antibodies have a preference towards a variant of a protein sequence. For example, celiac disease antibodies target deamidated gliadin, thus a microarray using the gliadin sequence could miss this binding interaction [70]. MESA represents the first algorithm that can multiplex epitope substitution using random libraries. With a computational mutagenesis scheme using k-mers and random libraries, there is no need to bias the experimental approach towards a specific antigen. Thus, epitopes corresponding to many antigens can be analyzed.

Limitations of MESA are mostly related to its inclusion of noise, a necessary result of avoiding time-consuming motif discovery algorithms like pair-wise sequence comparisons. However, MESA's substitution analysis sometimes leads to the inclusion of sequences that are irrelevant to the true binding motif, leading to a poor signal-to-noise ratio. Since this method evaluates many different 5- or 6-mers, it is crucial that there are $>10^5$ sequences for the analysis of serum antibodies, which for typical NGS datasets should not be difficult to achieve [76]. Also, if a group of subjects does not have homogenous binding to an epitope, MESA may not generate a clear binding motif.

MESA could be used to discover and refine epitopes identified through epitope mapping for the development of more effective diagnostics, vaccines, and reagents. Generating a binding motif for a disease-specific epitope would allow for the optimization of diagnostic peptides [70]. Additionally, associating motifs with diseases could provide insights into etiology [75,222] and lead to the development of antibody therapeutics, such as for cancer [35]. MESA could aid the study of broadly neutralizing antibodies for vaccine design and therapeutic development in infections such as HIV [223,224]. This approach could also be used to improve the development of peptides vaccines [225] enabling a more focused immune response. Moreover, there is an increasing need for robust methods capable of studying and resolving cross-reactivity in vaccine development [26].

MESA can be used to refine epitopes in NGS datasets using antibody-binding peptides or using known protein epitopes. The substitution analysis of the algorithm allows it to rapidly probe the full depth of large NGS datasets. The ability to identify epitopes relating to arbitrary antigens will likely become indispensable to fully interrogating antibody repertoires.

6.4 Materials and methods

6.4.1 Bacterial display library screening and sequencing

The protocol for screening bacterial display peptide libraries was carried out as previously described (**Section 2.4.2**). Briefly, we added 1:100 diluted human serum to an *Escherichia coli* display library of 8×10^9 random 12-mer peptides (trime-X_{12}), and sorted cells with bound antibodies through two rounds of magnetic selection using Protein A/G magnetic beads (Thermo Scientific Pierce). DNA amplicon libraries were prepared from the enriched library cells for sequencing using the Illumina NextSeq. All sera (N=60) were obtained as de-identified specimens from biobanks according to institutional guidelines and handled according to CDC-recommended BSL-2 guidelines.

6.4.2 Monoclonal antibody spike-in

Two rabbit mAbs, myc-Tag 71D10 (EQKLISEEDL) and HA-Tag C29F4 (YPYDVPDYA) (Cell Signaling Technology), were used for MESA analysis of known target sequences. The mAbs were added to a human serum specimen at 200 nM to mimic antibodies in serum (**Section 5.4.3**) [164].

6.4.3 Sequence processing

We generated non-redundant sequences from FASTQ files and calculated 5-mer and 6-mer enrichments using an adapted version of IMUNE [76]. Enrichment is defined as the ratio of observations of a k-mer to the “expected” observations. The number of “expected” observations is calculated by multiplying the number of frames the k-mer could fit in a 12-mer peptide with the total number of sequences and the probability of the k-mer appearing based on amino acid usage. For computations with large specimen datasets using 6-mers, we sum the enrichments from all specimens for each k-mer. Running these programs requires at least 16 GB of free RAM and 100 GB of hard-drive space. Analysis was carried out on a Dell Optiplex 9020 with an Intel® Core™ i7-4790 CPU @ 3.60 GHz, 64-bit operating system, and 32.0 GB of RAM. Processing FASTQ files into subsequences from 12 specimens (~1.5 million unique sequences per specimen) and calculating 5-mer and 6-mer enrichments required 136 mins, 10.1 mins, and 140 mins, respectively. The duration of these calculations scales approximately linearly with the number of specimens and sequences.

6.4.4 MESA algorithm

Multiplexed Epitope Substitution Analysis (MESA) determines which positions in an antibody-binding peptide are important for binding and the amino acid (AA) preferences at each binding position. This binding motif is displayed as an epitope logo in which the effect of AA substitution at a position corresponds to the absolute height at that position. Positions with an insignificant contribution to binding have a height of zero. The height of individual AA letters indicates the relative importance of an AA at a position.

The approach for determining binding motifs is based on the methodology used in ArrayPitope [217]. To start, a sequence with known antibody binding, denoted the target sequence, is tiled into k-mers of 5 or 6 AA with a single AA overlap. For each k-mer, one position at a time is substituted with all AAs. This results in 100 variants for 5-mers and 120 variants for 6-mers. The k-mer that corresponds to the native antigen sequence is termed the original k-mer. When using multiple specimens, the enrichment of a k-mer is the sum of enrichments in all specimens. All the enrichments are normalized by the original k-mer enrichment. The normalized enrichments are used to populate a position-specific scoring matrix (PSSM). In a PSSM, columns represent positions in the target sequence and rows represent AA substitutions.

The average of each column in the PSSM, denoted the score, indicates whether a position in a k-mer is conserved (‘N’) or variable (‘X’). The “conservation string” for a k-mer indicates which positions are conserved (e.g. NXNXN). If a conserved position is mutated, antibody binding will diminish, whereas antibody binding is nearly unaffected by the mutation of a variable position. A score near 1 indicates that a position has little preference for the original AA at that position and it is therefore a variable position. If the score is far below 1, this indicates a strong preference for the target sequence AA and it is considered a conserved position. Therefore, we use a binary cutoff value, termed the score threshold, on the score to determine if a position should be considered conserved or variable. The score threshold is determined by generating PSSMs for 1,000 random k-mers and compiling their scores into a list. The assumption is that these random k-mers should have all variable positions and should thus possess scores approximating random chance. A score threshold is chosen using a low percentile (e.g. 5%) of the list of random scores. Additionally, a position was considered variable if the sum of enrichments for the 20 variants that substitute that position ("the enrichment sum") is less than the minimum enrichment threshold. This threshold excludes spurious results where a position appears conserved because the enrichment is too low. The minimum enrichment threshold is defined as a percentile in the distribution of all enrichment sums generated from a target sequence. To determine a "sequence conservation string", the conservation string for each k-mer is aligned to the target sequence. If there is at least one k-mer that is conserved at a position in the alignment, then that position is conserved in the sequence conservation string.

To determine the binding motif, a substitution matrix was generated for each position in the target sequence. A substitution matrix can then be used to describe the AA preferences for each position (the "substitution position") in the target sequence. PSSM columns from positions that overlap with the substitution position are used to construct a substitution matrix. PSSM columns are only included in the substitution matrix if they are conserved in the k-mer. To determine the frequency of each AA at a substitution position, we determine the average of each column in the substitution matrix and normalize the averages by the sum of all column averages.

From the substitution matrices, each position in the target sequence has AA frequency values. The information that is directly used to generate the epitope logo is contained in the relative entropy matrix. To generate the relative entropy matrix, the following formula is applied to each AA frequency:

$$entropy = f_{aa} \log_2 \frac{f_{aa}}{u_{aa}} \tag{6.1}$$

where f_{aa} is the frequency of the AA and u_{aa} is the usage of the AA in the peptide dataset. In the relative entropy matrix, each row corresponds to a position in the initial sequence and each column corresponds to one of the 20 AAs. If a position is determined to be variable, it is represented by a row of zeros so that it will not appear in the epitope logo. The relative entropy matrix is then input into Seq2Logo to generate a sequence logo [226]. For generating an epitope logo from the single specimen for the peptide YADVFEYQYDWP (P1), it required 0.016 sec to generate the relative entropy matrix and 0.75 sec to generate the sequence logo using Seq2Logo.

The regular expression of the sequence logo allows for a single textual representation of the binding motif. To determine the regular expression, each position in the target peptide is made into 'X' if it is variable or is replaced with one or more AAs if the position is conserved. AAs are included in the regular expression if they meet a frequency cutoff (0.2 in our analysis). Leading and trailing 'X's are then trimmed from the regular expression.

Elements that were adapted from ArrayPitope [217] include dividing a target sequence into shorter overlapping subsequences, generating PSSMs and substitution matrices, and visualizing the results using Seq2Logo. MESA differs from ArrayPitope by using peptides selected from a random library rather than from targeted microarrays, dividing sequences into k-mers, using a nonparametric statistical approach, and generating an entropy matrix for input to Seq2Logo.

6.4.5 Identifying antibody-binding motifs with MEME

To validate the binding motifs discovered using MESA, the motif discovery algorithm MEME [131] was used to determine antibody-binding motifs in a set of peptide sequences. MEME uses pair-wise sequence comparisons in small sequence sets of less than ~5,000 members, while MESA utilizes substitution analyses throughout full NGS datasets. Although variations in the results from MESA and MEME will exist due to differences in algorithms and data input, comparison to MEME is effective for broadly confirming antibody-binding motifs and assessing MESA performance.

MEME is not suitable to analyze large peptide datasets due to run-time constraints. Therefore, all MEME analyses were run with a maximum of 5,000 sequences. To identify mAb motifs, a random bacterial display peptide library was screened against each mAb at 20 nM in PBST [164] and sequenced. A minimum motif width of 8 was utilized for the MEME mAb analyses. For all analyses of serum antibody motifs, a minimum motif width of 4 was used. For 5,000 sequences, MEME analysis required 10.0 ± 1.4 hours.

7 Conclusion

7.1 Perspectives

7.1.1 Development of high-throughput screening method

Random bacterial display peptide library screening has the potential to offer a high-throughput platform for interrogating the serum antibody repertoire to discover and characterize epitopes associated with a wide range of immunological diseases. However, bacterial display can suffer from biological limitations such as codon biases and stop codons that can impact epitope discovery. Therefore, it is critical to construct and characterize a high-quality display library before the application to complex systems for discovery. Here, we constructed a large random bacterial display peptide library (trimer-X_{12}) containing ~10 billion peptides with restricted codon usage optimized for randomized sequences and *E. coli* growth. Next-generation sequencing (NGS) was performed to determine ~70 million sequences from the naive library, enabling in-depth characterization. The trimer-X_{12} library was compared to a previously constructed library using the traditional NNS codon method (NNS-X_{15}) to reveal major improvements. We determined that less than 4% of the trimer-X_{12} sequences contained a stop codon compared to 43% in the NNS-X_{15} library. Additionally, the trimer-X_{12} library had less variability of amino acid frequencies with a standard deviation of 1.4% and a range of only 5.5%, compared to the NNS-X_{15} library frequencies which had a standard deviation of 3.2% and a range of 11%. NGS was also utilized to determine the amino acid frequencies following library screening with serum specimens. We determined that the amino acid frequencies after selection were highly dependent on the naive library, and bias in the naive library was propagated through the screening process. Therefore, the minimization of stop codon usage and codon bias in the trimer-X_{12} library will enable efficient sampling of the highly diverse peptides for unbiased epitope discovery.

With a high-quality peptide library constructed and characterized, we screened libraries for antibody-binding against a set of serum specimens to reveal fundamental characteristics of antibody binding to linear peptides. Through the analysis of hundreds of binding motifs, we determined that 90% of motifs were five to eight amino acids wide with four to six conserved residues. Knowing the structure of the majority of motifs now provides us with a basis for motif identification in large peptide datasets. Finally, we demonstrated unbiased motif discovery using monoclonal antibodies (mAb) with known epitopes by identifying precise binding motifs for each mAb and quantifying motif enrichment. Ultimately, the availability of NGS has enabled an in-depth characterization of random and enriched peptide libraries, providing a better understanding of the fundamentals of peptide library screening for epitope discovery. With a large, high-quality library constructed and characterized, we have developed a reproducible and high-throughput screening methodology that can be readily applied to interrogate the antibody repertoires from immunological diseases.

7.1.2 Application of random peptide screening to diverse immune-related diseases

A major advantage of random peptide library screening is the applicability to diverse immune-related diseases to discover antibody-binding peptides. This facilitates the discovery of diagnostic reagents for detecting antibody biomarkers and the identification of antigens associated with disease. Here, we analyzed the serum antibody repertoires associated with herpes simplex virus (HSV) for epitope discovery and characterization. The two HSV species, HSV-1 and HSV-2, are highly similar with over 80% of the protein-coding regions from each virus sharing sequence homology [94]. Therefore, highly-specific reagents need to be identified for accurate diagnostics. By applying our peptide screening methodology, we identified 14 HSV-2 specific motifs that each demonstrated 100% sensitivity and specificity when distinguishing HSV-2 from HSV-1. Motifs specific for HSV-1 were also identified but with lower sensitivity. A classification algorithm was developed to improve diagnostic accuracy by utilizing reactivity towards nine motifs in parallel, achieving 100% type-specific accuracy. To bolster evidence for the specificity of the HSV motifs, our database of peptide libraries from various sample cohorts was analyzed for motif enrichment. A general population of adults was evaluated for motif enrichment and the estimated HSV prevalence was consistent with epidemiological trends. Finally, the HSV-specific motifs were aligned to the HSV proteomes using online protein sequence databases to identify candidate antigens. Numerous motifs mapped to HSV glycoproteins, such as the HSV-1 glycoproteins G1 and D1 and the HSV-2 glycoproteins G2 and D2, known for being antigenic [82,85,90]. Moreover, we identified previously unreported candidate antigens containing highly specific motifs, including tegument proteins and capsid proteins. While additional studies are needed to further validate these antigens, these results demonstrate the broad capabilities for discovering epitopes in highly similar infectious diseases. The numerous epitopes discovered with high diagnostic utility could be integrated into novel point-of-care devices for rapid, multiplexed diagnosis of viral species. The ability to accurately detect highly similar species could improve diagnostics and enable the monitoring of disease prevalence and the spread of infectious diseases.

We next applied our antibody repertoire analysis platform to age-related macular degeneration (AMD). It is known that AMD progression involves immune dysregulation [98] and unique antibody signatures have been previously identified [124,125]. However, unlike an infectious disease, age-related macular degeneration (AMD) does not have clear antibody targets. By analyzing the serum antibody repertoires from specimens obtained at different stages of AMD, we identified a set of binding motifs specific to the onset of late AMD. An SVM classifier was developed to classify onset late AMD specimens from cohorts with varying risk of developing late AMD. The classifier demonstrated improved performance as the risk of developing late AMD decreased. Importantly, the classifier demonstrated 84% diagnostic accuracy when classifying independent onset late AMD specimens from age-matched controls with low risk of developing late AMD (small drusen). AMD-specific motifs were also found to be reactive over years of disease progression, indicating that the corresponding antibodies were present and sustained over time. These findings suggest unique antibody signatures exist at the beginning stages of late AMD that could be detected to facilitate early treatment and AMD prevention. Due to the severity of late stage AMD and the irreversibility of the accompanied vision loss, there is an increasing need to accurately diagnosis the disease as early as possible. To further validate the AMD motif panel for diagnostic accuracy, cohorts with increased sample sizes will need to be analyzed. Additional work can be completed to evolve the AMD-specific motifs to elucidate additional epitope sequence identity and enable novel antigen identification. This could be completed by constructing bacterial display focused motif libraries and screening the motif libraries against additional AMD specimens. The identification of antigens associated with AMD would greatly improve our understanding of the complex disease pathology and provide new opportunities for therapies.

The discoveries of antibody-binding motifs associated with diverse infections and diseases such as HSV and AMD demonstrate the versatility of antibody repertoire analysis using bacterial display. Peptide reagents can be readily identified and utilized with an assortment of classification schemes to develop diagnostic panels with high accuracy. The unbiased approach of random peptide screening enables the discovery of novel peptides and antigens not possible in targeted approaches.

7.1.3 Advancements in peptide screening and epitope discovery

While we successfully applied the bacterial display screening platform for a variety of antibody repertoire analyses, the complexity of the serum antibody repertoire remains a barrier for biomarker discovery. Specifically, the antibody repertoire is often dominated by high-titer antibody species, limiting the detection of low-titer antibodies that may be relevant to diseases. We therefore developed a novel method to selectively deplete high-titer antibodies from serum specimens, termed high-titer depletion (HTD). With HTD, random peptides were enriched towards binding high-titer antibody species and then utilized as binding reagents to deplete the corresponding species from the specimen. Importantly, the depletion reagents are customized for each unique antibody repertoire, optimizing the depletion process for each specimen. A random peptide library can then be screened using the depleted serum specimen to enhance the discovery of peptides bound by low-titer antibodies. HTD was validated by detecting three monoclonal antibodies at concentrations at least an order of magnitude lower than without HTD. Moreover, 92 serum antibody specificities were readily discovered from an individual specimen using HTD compared to only 25 specificities without HTD. These results indicate that the HTD method could be reproducibly applied for the discovery of rare antibody biomarkers related to disease and characterization of the serum antibody repertoire at depths not previously possible.

Aside from improving the sensitivity and depth of the peptide screening technology, there is an increasing need for the development of advanced bioinformatics to analyze the large NGS datasets. Currently, motif discovery algorithms such as MEME and IMUNE are unbiased discovery algorithms. However, if an antigen or epitope is known or suspected, it is difficult to identify the binding motifs associated with the target epitope using unbiased discovery methods. Moreover, these algorithms are limited when analyzing the peptide libraries screened against a single antibody repertoire. In the past, studies have required a targeted approach of tiling peptide sequences from an antigen and interrogating binding for each antigen of interest. However, this approach is low-throughput and restrictive. Therefore, we developed an approach to rapidly map the antibody binding to epitopes of interest following random peptide library screening, termed Multiplexed Epitope Substitution Analysis (MESA). Given a target epitope sequence, MESA computationally mutates each epitope position and evaluates the enrichment of the peptide variant. By evaluating all amino acid substitutions, a precise map of binding preferences for each position in the epitope can be determined. Because MESA utilizes data from random peptide library screening, any peptide epitope can be characterized for antibody binding. MESA was applied to individual peptide libraries to accurately characterize patient-specific epitopes as well as large cohort libraries to map consensus epitopes from common pathogens such as the Epstein–Barr virus. The ability of MESA to efficiently search entire NGS datasets allows for maximum utilization of the library sequences and optimal epitope discovery and characterization.

7.2 Future directions

The integration of bacterial display peptide library screening and NGS has enabled unprecedented access to the serum antibody repertoire. The immunological memory of the antibody repertoire provides countless opportunities to discover antibody biomarkers and reveal novel antigens from infections, allergens, the microbiome, autoimmune diseases, and cancers. Random peptide library screening provides a broad, high-throughput platform for discovery and characterization. The antibody-binding peptides discovered can be utilized as diagnostic reagents and mapped to antigens to inform of disease pathology and novel therapies.

The methodologies developed and applied in this work are part of the emerging field of immunoinformatics, the integration of immunology and bioinformatics [209]. The complexity and dynamic nature of the immune system requires sophisticated computational analysis to expose the network of interactions. The serum antibody repertoire is one such network connecting the adaptive immune response to interactions with the environment. By cataloging millions of antibody-peptide interactions for patient specimens, we have constructed a database that can be readily searched for antibody-binding motifs. We utilized this database to examine motif reactivity in various populations, including youth cohorts and adult cohorts, to further validate diagnostic specificity. Diagnostics require massive collections of patient specimens to validate accuracy, and thus databases of these types generated from high-throughput analyses could greatly impact the field of diagnostic development. Numerous other databases involving epitopes and antigens have been developed, such as the immune epitope database (IEDB) [227] and the B-cell epitope database Bcipep [228]. As these databases grow and antigen information becomes readily available, these databases will be indispensable tools for antibody biomarker discovery.

While random peptide library screening has many advantages for epitope discovery, all screening platforms currently suffer from fundamental limitations for discovering structural epitopes and epitopes with post-translational modifications (PTMs). While advances have been made for mapping linear antibody-binding peptides onto three-dimensional antigen structures [229], this is only possible for the small number of antigens with elucidated structures. Recently, high-throughput yeast display has been used to display mutated antigens and screen for antibody binding to map conformational epitopes [230]. These methods are still limited to known antigens and are targeted approaches with narrow scopes but nonetheless provide a foundation for complex epitope screening platforms. To further expand the scope of epitope discovery, screening methods incorporating PTMs into epitopes will need to be developed. A prominent PTM epitope with diagnostic relevance is the cyclic citrullinated peptide used to diagnose rheumatoid arthritis [11]. PTMs have successfully been integrated into bacterial display peptide screening by utilizing enzymes to modify the displayed peptides [231].This approach could be used to identify novel linear and cyclic epitopes with PTMs such as citrullinated peptides. Although these methods will inevitably be lower throughput than traditional peptide library screening, they will expand the breadth of epitope discovery and create new opportunities for diagnostics and therapeutics.

While the antibody repertoire analyses completed in this work primarily focused on populations of specimens with disease, experimental and computational advances will enable the characterization of individual repertoires. The ability to analyze an individual antibody repertoire could enable precision health monitoring and personalized diagnostics and therapies. Additional advancements in personalized medicine technologies such as single B-cell sequencing [7,232] could be used in conjunction with functional antibody repertoire analysis to characterize the repertoire at great depth for patient-specific diagnostics and therapeutics and to gain important insights into disease pathology. Moving forward, these antibody repertoire technologies will greatly impact the development of diagnostics and therapeutics and the study of immune system networks.

8 References

1. Foote J, Eisen HN. Kinetic and affinity limits on antibodies produced during immune responses. Proc Natl Acad Sci. 1995;92(5):1254–6.

2. Hammarlund E, Lewis MW, Hansen SG, Strelow LI, Nelson JA, Sexton GJ, et al. Duration of antiviral immunity after smallpox vaccination. Nat Med. 2003;9(9):1131–7.

3. Sallusto F, Lanzavecchia A, Araki K, Ahmed R. From vaccines to memory and back. Immunity. 2010;33(4):451–63.

4. Lavinder JJ, Horton AP, Georgiou G, Ippolito GC. Next-generation sequencing and protein mass spectrometry for the comprehensive analysis of human cellular and serum antibody repertoires. Curr Opin Chem Biol. 2015;24:112–20.

5. Wine Y, Horton AP, Ippolito GC, Georgiou G. Serology in the 21st century: the molecular-level analysis of the serum antibody repertoire. Curr Opin Immunol. 2015;35:89–97.

6. Lavinder JJ, Wine Y, Giesecke C, Ippolito GC, Horton AP, Lungu OI, et al. Identification and characterization of the constituent human serum antibodies elicited by vaccination. Proc Natl Acad Sci U S A. 2014;111(6):2259–64.

7. Lee J, Boutz DR, Chromikova V, Joyce MG, Vollmers C, Leung K, et al. Molecular-level analysis of the serum antibody repertoire in young adults before and after seasonal influenza vaccination. Nat Med. 2016;22(12):1456–64.

8. Parham P. The immune system. 4th ed. Garland Science; 2014.

9. McDade TW, Williams S, Snodgrass JJ. What a drop can do: dried blood spots as a minimally invasive method for integrating biomarkers into population-based research. Demography. 2007;44(4):899–925.

10. Chase BA, Johnston SA, Legutki JB. Evaluation of biological sample preparation for immunosignature-based diagnostics. Clin Vaccine Immunol. 2012;19(3):352–8.

11. Nishimura K, Sugiyama D, Kogata Y, Tsuji G, Nakazawa T, Kawano S, et al. Meta-analysis: diagnostic accuracy of anti–cyclic citrullinated peptide antibody and rheumatoid factor for rheumatoid arthritis. Ann Intern Med. 2007;146(11):797.

12. Schlosser M, Mueller PW, Törn C, Bonifacio E, Bingley PJ. Diabetes antibody standardization program: evaluation of assays for insulin autoantibodies. Diabetologia. 2010;53(12):2611–20.

13. Törn C, Mueller PW, Schlosser M, Bonifacio E, Bingley PJ. Diabetes antibody standardization program: evaluation of assays for autoantibodies to glutamic acid decarboxylase and islet antigen-2. Diabetologia. 2008;51(5):846–52.

14. Leffler DA, Schuppan D. Update on serologic testing in celiac disease. Am J Gastroenterol. 2010;105(12):2520–4.

15. Shmerling RH, Delbanco TL. The rheumatoid factor: an analysis of clinical utility. Am J Med. 1991;91(5):528–34.

16. Ho DWT, Field PR, Sjogren-Jansson E, Jeansson S, Cunningham AL. Indirect ELISA for the detection of HSV-2 Specific IgM and IgM antibodies with glycoprotein G (gG-2). J Virol Methods. 1992;36(3):249–64.

17. Helfand RF, Heath JL, Anderson LJ, Maes EF, Guris D, Bellini WJ. Diagnosis of measles with an IgM capture EIA: the optimal timing of specimen collection after rash onset. J Infect Dis. 1997;175(1):195–9.

18. Langenhuysen MM, The TH, Nieweg HO, Kapsenberg JG. Demonstration of IgM cytomegalovirus-antibodies as an aid to early diagnosis in adults. Clin Exp Immunol. 1970;6(3):387–93.

19. Ellington AA, Kullo IJ, Bailey KR, Klee GG. Antibody-based protein multiplex platforms: technical and operational challenges. Clin Chem. 2010;56(2):186–93.

20. Looker KJ, Magaret AS, May MT, Turner KME, Vickerman P, Gottlieb SL, et al. Global and regional estimates of prevalent and incident herpes simplex virus type 1 infections in 2012. PLoS One. 2015;10(10):1–17.

21. Looker KJ, Magaret AS, Turner KME, Vickerman P, Gottlieb SL, Newman LM. Global estimates of prevalent and incident herpes simplex virus type 2 infections in 2012. PLoS One. 2015;10(1):1–23.

22. Wormser GP, Schriefer M, Aguero-Rosenfeld ME, Levin A, Steere AC, Nadelman RB, et al. Single-tier testing with the C6 peptide ELISA kit compared with two-tier testing for Lyme disease. Diagn Microbiol Infect Dis. 2013;75(1):9–15.

23. Tozzoli R, Bagnasco M, Giavarina D, Bizzaro N. TSH receptor autoantibody immunoassay in patients with Graves' disease: improvement of diagnostic accuracy over different generations of methods. systematic review and meta-analysis. Autoimmun Rev. 2012;12(2):107–13.

24. Basu D, Reveille JD. Anti-scl-70. Autoimmunity. 2005;38(1):65–72.

25. Berglund L, Andrade J, Odeberg J, Uhlen M. The epitope space of the human proteome. Protein Sci. 2008;17(4):606–13.

26. Van Regenmortel MHV. Specificity, polyspecificity, and heterospecificity of antibody-antigen recognition. J Mol Recognit. 2014;27(11):627–39.

27. Haste Andersen P, Nielsen M, Lund O. Prediction of residues in discontinuous B-cell epitopes using protein 3D structures. Protein Sci. 2006;15(11):2558–67.

28. Kringelum JV, Nielsen M, Padkjær SB, Lund O. Structural analysis of B-cell epitopes in antibody:protein complexes. Mol Immunol. 2013;53(1–2):24–34.

29. Van Regenmortel MHV. Mapping epitope structure and activity: from one-

dimensional prediction to four-dimensional description of antigenic specificity. Methods a Companion To Methods Enzymol. 1996;9(3):465–72.

30. Negi SS, Braun W. Automated detection of conformational epitopes using phage display peptide sequences. Bioinform Biol Insights. 2009;(3):71–81.

31. Mayrose I, Shlomi T, Rubinstein ND, Gershoni JM, Ruppin E, Sharan R, et al. Epitope mapping using combinatorial phage-display libraries: a graph-based algorithm. Nucleic Acids Res. 2007;35(1):69–78.

32. Balass M, Heldman Y, Cabilly S, Givol D, Katchalski-Katzir E, Fuchs S. Identification of a hexapeptide that mimics a conformation-dependent binding site of acetylcholine receptor by use of a phage-epitope library. Proc Natl Acad Sci U S A. 1993;90(22):10638–42.

33. Bastas G, Sompuram SR, Pierce B, Vani K, Bogen SA. Bioinformatic requirements for protein database searching using predicted epitopes from disease-associated antibodies. Mol Cell Proteomics. 2007;7(2):247–56.

34. Weetman AP. Graves ' disease. N Engl J Med. 2000;343(17):1236–48.

35. Bushey RT, Moody MA, Nicely NL, Haynes BF, Alam SM, Keir ST, et al. A therapeutic antibody for cancer, derived from single human B cells. Cell Rep. 2016;15(7):1505–13.

36. Rappuoli R. Bridging the knowledge gaps in vaccine design. Nat Biotechnol. 2007;25(12):1361–6.

37. Legutki JB, Zhao ZG, Greving M, Woodbury N, Johnston SA, Stafford P. Scalable high-density peptide arrays for comprehensive health monitoring. Nat Commun. 2014;5:4785.

38. Smith GP, Petrenko VA. Phage display. Chem Rev. 1997;97(2):391–410.

39. Getz JA, Schoep TD, Daugherty PS. Peptide discovery using bacterial display and flow cytometry. In: Methods in Enzymology. Academic Press; 2012. p. 75–97.

40. Larman HB, Zhao Z, Laserson U, Li MZ, Ciccia A, Gakidis MAM, et al. Autoantigen discovery with a synthetic human peptidome. Nat Biotechnol. 2011;29(6):535–41.

41. Larman HB, Laserson U, Querol L, Verhaeghen K, Solimini NL, Xu GJ, et al. PhIP-Seq characterization of autoantibodies from patients with multiple sclerosis, type 1 diabetes and rheumatoid arthritis. J Autoimmun. 2013;43:1–9.

42. Xu GJ, Kula T, Xu Q, Li MZ, Vernon SD, Ndung'u T, et al. Comprehensive serological profiling of human populations using a synthetic human virome. Science. 2015;348(6239):aaa0698.

43. Cortese R, Felici F, Galfre G, Luzzago A, Monaci P, Nicosia A. Epitope discovery using peptide libraries displayed on phage. Trends Biotechnol. 1994;12(7):262–7.

44. Heyduk E, Heyduk T. Ribosome display enhanced by next generation sequencing: a tool to identify antibody-specific peptide ligands. Anal Biochem. 2014;464:73–82.

45. Geysen HM, Rodda SJ, Mason TJ. A priori delineation of a peptide which mimics a discontinuous antigenic determinant. Mol Immunol. 1986;23(7):709–15.

46. Metzker ML. Sequencing technologies–the next generation. Nat Rev Genet. 2010;11(1):31–46.

47. Gai SA, Wittrup KD. Yeast surface display for protein engineering and characterization. Curr Opin Struct Biol. 2007;17(4):467–73.

48. Hanes J, Schaffitzel C, Knappik A, Plückthun A. Picomolar affinity antibodies from a fully synthetic naive library selected and evolved by ribosome display. Nat Biotechnol. 2000;18(12):1287.

49. Boder ET, Midelfort KS, Wittrup KD. Directed evolution of antibody fragments with monovalent femtomolar antigen-binding affinity. Proc Natl Acad Sci U S A. 2000;97(20):10701–5.

50. Legutki JB, Magee DM, Stafford P, Johnston SA. A general method for characterization of humoral immunity induced by a vaccine or infection. Vaccine. 2010;28(28):4529–37.

51. Legutki JB, Johnston SA. Immunosignatures can predict vaccine efficacy. Proc Natl Acad Sci. 2013;110(46):18614–9.

52. Rowe M, Melnick J, Gerwien R, Legutki JB, Pfeilsticker J, Tarasow TM, et al. An ImmunoSignature test distinguishes Trypanosoma cruzi , hepatitis B , hepatitis C and West Nile virus seropositivity among asymptomatic blood donors. PLoS Negl Trop Dis. 2017;11(9):1–30.

53. Hughes AK, Cichacz Z, Scheck A, Coons SW, Johnston SA, Stafford P. Immunosignaturing can detect products from molecular markers in brain cancer. PLoS One. 2012;7(7):1–7.

54. Stafford P, Halperin R, Legutki JB, Magee DM, Galgiani J, Johnston SA. Physical characterization of the "immunosignaturing effect." Mol Cell Proteomics. 2012;11(4):M111-11593.

55. Sykes KF, Legutki JB, Stafford P. Immunosignaturing: a critical review. Trends Biotechnol. 2013;31(1):45–51.

56. Richer J, Johnston SA, Stafford P. Epitope identification from fixed-complexity random-sequence peptide microarrays. Mol Cell Proteomics. 2015;14(1):136–47.

57. Ryvkin A, Ashkenazy H, Smelyanski L, Kaplan G, Penn O, Weiss-Ottolenghi Y, et al. Deep panning: steps towards probing the IgOme. PLoS One. 2012;7(8):1–11.

58. Frietze KM, Pascale JM, Moreno B, Chackerian B, Peabody DS. Pathogen-specific deep sequence-coupled biopanning: a method for surveying human antibody responses. PLoS One. 2017;12(2):e0171511.

59. Frietze KM, Roden RBS, Lee JH, Shi Y, Peabody DS, Chackerian B. Identification of anti-CA125 antibody responses in ovarian cancer patients by a novel deep sequence-

coupled biopanning platform. Cancer Immunol Res. 2016;4(2):157–64.

60. Weber LK, Palermo A, Kügler J, Armant O, Isse A, Rentschler S, et al. Single amino acid fingerprinting of the human antibody repertoire with high density peptide arrays. J Immunol Methods. 2017;443:45–54.

61. Clackson T, Wells JA. In vitro selection from protein and peptide libraries. Trends Biotechnol. 1994;12(5):173–84.

62. Derda R, Tang SKY, Li SC, Ng S, Matochko W, Jafari MR. Diversity of phage-displayed libraries of peptides during panning and amplification. Molecules. 2011;16(2):1776–803.

63. Löfblom J. Bacterial display in combinatorial protein engineering. Biotechnol J. 2011;6(9):1115–29.

64. Rice JJ, Schohn A, Bessette PH, Boulware KT, Daugherty PS. Bacterial display using circularly permuted outer membrane protein OmpX yields high affinity peptide ligands. Protein Sci. 2006;15(4):825–36.

65. Rice JJ, Daugherty PS. Directed evolution of a biterminal bacterial display scaffold enhances the display of diverse peptides. Protein Eng Des Sel. 2008;21(7):435–42.

66. Daugherty PS. Protein engineering with bacterial display. Curr Opin Struct Biol. 2007;17(4):474–80.

67. Casadaban MJ, Cohen SN. Analysis of gene control signals by DNA fusion and cloning in Escherichia coli. J Mol Biol. 1980;138(2):179–207.

68. Daugherty PS, Olsen MJ, Iverson BL, Georgiou G. Development of an optimized expression system for the screening of antibody libraries displayed on the Escherichia coli surface. Protein Eng. 1999;12(7):613–21.

69. Boulware KT, Daugherty PS. Protease specificity determination by using cellular libraries of peptide substrates (CLiPS). Proc Natl Acad Sci. 2006;103(20):7583–8.

70. Ballew JT, Murray JA, Collin P, Mäki M, Kagnoff MF, Kaukinen K, et al. Antibody biomarker discovery through in vitro directed evolution of consensus recognition epitopes. Proc Natl Acad Sci U S A. 2013 Nov 26;110(48):19330–5.

71. Boulware KT, Jabaiah A, Daugherty PS. Evolutionary optimization of peptide substrates for proteases that exhibit rapid hydrolysis kinetics. Biotechnol Bioeng. 2010;106(3):339–46.

72. Tamerler C, Sarikaya M. Molecular biomimetics: utilizing nature's molecular ways in practical engineering. Acta Biomater. 2007;3(3):289–99.

73. Spatola BN, Murray JA, Kagnoff MF, Kaukinen K, Daugherty PS. Antibody repertoire profiling using bacterial display identifies reactivity signatures of celiac disease. Anal Chem. 2013;85(2):1215–22.

74. Spatola BN, Kaukinen K, Collin P, Mäki M, Kagnoff MF, Daugherty PS. Persistence of elevated deamidated gliadin peptide antibodies on a gluten-free diet indicates

nonresponsive coeliac disease. Aliment Pharmacol Ther. 2014;39(4):407–17.

75. Elliott SE, Parchim NF, Kellems RE, Xia Y, Soffici AR, Daugherty PS. A pre-eclampsia-associated Epstein-Barr virus antibody cross-reacts with placental GPR50. Clin Immunol. 2016;168:64–71.

76. Pantazes RJ, Reifert J, Bozekowski J, Ibsen KN, Murray JA, Daugherty PS. Identification of disease-specific motifs in the antibody specificity repertoire via next-generation sequencing. Sci Rep. 2016;6:30312.

77. Brooks GF, Carroll KC, Butel JS, Morse SA, Mietzner TA. Jawetz, Melnick, & Adelberg's Medical Microbiology. 26th ed. McGraw-Hill; 2013.

78. Ryan KJ, Ray CG. Sherris Medical Microbiology. 4th ed. McGraw-Hill; 2004.

79. McGeoch DJ, Rixon FJ, Davison AJ. Topics in herpesvirus genomics and evolution. Virus Res. 2006;117(1):90–104.

80. Bowden RJ, McGeoch DJ. Evolution of herpes simplex viruses. In: Herpes simplex viruses. CRC Press; 2017. p. 1–34.

81. Taylor TJ, Brockman MA, McNamee EE, Knipe DM. Herpes simplex virus. Front Biosci. 2002;7:752–64.

82. Risinger C, Sørensen KK, Jensen KJ, Olofsson S, Bergström T, Blixt O. Linear multiepitope (glyco)peptides for type-specific serology of herpes simplex virus (HSV) infections. ACS Infect Dis. 2017;3(5):360–7.

83. Liu K, Jiang D, Zhang L, Yao Z, Chen Z, Yu S, et al. Identification of B- and T-cell epitopes from glycoprotein B of herpes simplex virus 2 and evaluation of their immunogenicity and protection efficacy. Vaccine. 2012;30(19):3034–41.

84. Goade DE, Bell R, Yamada T, Mertz GJ, Jenison S. Locations of herpes simplex virus type 2 glycoprotein B epitopes recognized by human serum immunoglobulin G antibodies. J Virol. 1996;70(5):2950–6.

85. Isola VJ, Eisenberg RJ, Siebert GR, Heilman CJ, Wilcox WC, Cohen GH. Fine mapping of antigenic site II of herpes simplex virus glycoprotein D. J Virol. 1989;63(5):2325–34.

86. Whitbeck JC, Huang Z-Y, Cairns TM, Gallagher JR, Lou H, Ponce-de-Leon M, et al. Repertoire of epitopes recognized by serum IgG from humans vaccinated with herpes simplex virus 2 glycoprotein D. J Virol. 2014;88(14):7786–95.

87. Cairns TM, Shaner MS, Zuo Y, Baribaud I, Eisenberg RJ, Cohen GH, et al. Epitope mapping of herpes simplex virus type 2 gH/gL defines distinct antigenic sites, including some associated with biological function. J Virol. 2006;80(6):2596–608.

88. McGeoch DJ, Moss HWM, McNab D, Frame MC. DNA sequence and genetic content of the HindIII l region in the short unique component of the herpes simplex virus type 2 genome: identification of the gene encoding glycoprotein G, and evolutionary comparisons. J Gen Virol. 1987;68(1):19–38.

89. Levi M, Ruden U, Wahren B. Peptide sequences of glycoprotein G-2 discriminate between herpes simplex virus type 2 (HSV-2) and HSV-1 antibodies. Clin Diagn Lab Immunol. 1996;3(3):265–9.

90. Marsden HS, MacAulay K, Murray J, Smith IW. Identification of an immunodominant sequential epitope in glycoprotein G of herpes simplex virus type 2 that is useful for serotype-specific diagnosis. J Med Virol. 1998;56(1):79–84.

91. Grabowska A, Jameson C, Laing P, Jeansson S, Sjögren-Jansson E, Taylor J, et al. Identification of type-specific domains within glycoprotein G of herpes simplex virus type 2 (HSV-2) recognized by the majority of patients infected with HSV-2, but not by those infected with HSV-1. J Gen Virol. 1999;80(7):1789–98.

92. Ashley RL. Sorting out the new HSV type specific antibody tests. Sex Transm Infect. 2001;77(4):232–7.

93. Oladepo DK, Klapper PE, Marsden HS. Peptide based enzyme-linked immunoassays for detection of anti-HSV-2 IgG in human sera. J Virol Methods. 2000;87(1–2):63–70.

94. Legoff J, Péré H, Bélec L. Diagnosis of genital herpes simplex virus infection in the clinical laboratory. Virol J. 2014;11(1):83.

95. Wald A, Huang M, Carrell D, Selke S, Corey L. Polymerase chain reaction for detection of herpes simplex virus (HSV) DNA on mucosal surfaces: comparison with HSV isolation in cell culture. J Infect Dis. 2003;188(9):1345–51.

96. Van Doornum GJJ, Guldemeester J, Osterhaus ADME, Niesters HGM. Diagnosing herpesvirus infections by real-time amplification and rapid culture. J Clin Microbiol. 2003;41(2):576–80.

97. Gehrs KM, Anderson DH, Johnson LV, Hageman GS. Age-related macular degeneration - emerging pathogenetic and therapeutic concepts. Ann Med. 2006;38(7):450–71.

98. Ambati J, Atkinson JP, Gelfand BD. Immunology of age-related macular degeneration. Nat Rev Immunol. 2013;13(6):438–51.

99. Age-Related Eye Disease Study Research Group. Risk factors for the incidence of advanced age-related macular degeneration in the Age-Related Eye Disease Study (AREDS): AREDS report no. 19. Ophthalmology. 2005;112(4):533–9.

100. Donoso LA, Kim D, Frost A, Callahan A, Hageman G. The role of inflammation in the pathogenesis of age-related macular degeneration. Surv Ophthalmol. 2006;51(2):137–52.

101. Haines JL, Hauser MA, Schmidt S, Scott WK, Olson LM, Gallins P, et al. Complement factor H variant increases the risk of age-related macular degeneration. Science. 2005;308(5720):419.

102. Black JRM, Clark SJ. Age-related macular degeneration: genome-wide association studies to translation. Genet Med. 2016;18(4):283–9.

103. AMD Alliance International. The global economic cost of visual impairment. Access Econ. 2010;26.

104. Age-Related Eye Disease Study Research Group. The age-related eye disease study (AREDS): design implications AREDS report no. 1. Control Clin Trials. 1999;20(6):573–600.

105. Anderson DH, Mullins RF, Hageman GS, Johnson LV. A role for local inflammation in the formation of drusen in the aging eye. Am J Ophthalmol. 2002;134(3):411–31.

106. Crabb JW. The proteomics of drusen. Cold Spring Harb Perspect Med. 2014;4(7):1–14.

107. Wang JJ, Foran S, Smith W, Mitchell P. Risk of age-related macular degeneration in eyes with macular drusen or hyperpigmentation: the Blue Mountains Eye Study cohort. Arch Ophthalmol. 2003;121(5):658–63.

108. Nowak JZ. Age-related macular degeneration (AMD): pathogenesis and therapy. Pharmacol Rep. 2006;58(3):353–63.

109. Coleman H, Chew E. Nutritional supplementation in age-related macular degeneration. In: Medical Retina. Springer Berlin Heidelberg; 2007. p. 105–11.

110. Pandey RK. Recent advances in photodynamic therapy. J Porphyr Phthalocyanines. 2000;4(4):368–73.

111. Ivandic BT, Ivandic T. Low-level laser therapy improves vision in patients with age-related macular degeneration. Photomed Laser Surg. 2008;26(3):241–5.

112. Volz C, Pauly D. Antibody therapies and their challenges in the treatment of age-related macular degeneration. Eur J Pharm Biopharm. 2015;95:158–72.

113. Saharinen P, Eklund L, Alitalo K. Therapeutic targeting of the angiopoietin–TIE pathway. Nat Rev Drug Discov. 2017;16(9):635.

114. Drolet DW, Green LS, Gold L, Janjic N. Fit for the eye: aptamers in ocular disorders. Nucleic Acid Ther. 2016;26(3):127–46.

115. van Lookeren Campagne M, LeCouter J, Yaspan BL, Ye W. Mechanisms of age-related macular degeneration and therapeutic opportunities. J Pathol. 2014;232(2):151–64.

116. Evans JB, Syed BA. New hope for dry AMD? Nat Rev Drug Discov. 2013;12(7):501–2.

117. Kashani AH, Lebkowski JS, Rahhal FM, Avery RL, Salehi-Had H, Dang W, et al. A bioengineered retinal pigment epithelial monolayer for advanced, dry age-related macular degeneration. Sci Transl Med. 2018;10(435):eaao4097.

118. da Cruz L, Fynes K, Georgiadis O, Kerby J, Luo YH, Ahmado A, et al. Phase 1 clinical study of an embryonic stem cell–derived retinal pigment epithelium patch in age-related macular degeneration. Nat Biotechnol. 2018;36(4):328.

119. Ozaki E, Campbell M, Kiang AS, Humphries M, Doyle SL, Humphries P. Inflammation in age-related macular degeneration. In: Retinal Degenerative Diseases. New York, NY: Springer; 2014. p. 229–35.

120. Umeda S, Suzuki MT, Okamoto H, Ono F, Mizota A, Terao K, et al. Molecular composition of drusen and possible involvement of anti-retinal autoimmunity in two different forms of macular degeneration in cynomolgus monkey (Macaca fascicularis). FASEB J. 2005;19(12):1683–5.

121. Khandhadia S, Cipriani V, Yates JRW, Lotery AJ. Age-related macular degeneration and the complement system. Immunobiology. 2012;217(2):127–46.

122. Geerlings MJ, de Jong EK, den Hollander AI. The complement system in age-related macular degeneration: a review of rare genetic variants and implications for personalized treatment. Mol Immunol. 2017;84:65–76.

123. Gu X, Meer SG, Miyagi M, Rayborn ME, Hollyfield JG, Crabb JW, et al. Carboxyethylpyrrole protein adducts and autoantibodies, biomarkers for age-related macular degeneration. J Biol Chem. 2003;278(43):42027–35.

124. Morohoshi K, Patel N, Ohbayashi M, Chong V, Grossniklaus HE, Bird AC, et al. Serum autoantibody biomarkers for age-related macular degeneration and possible regulators of neovascularization. Exp Mol Pathol. 2012;92(1):64–73.

125. Morohoshi K, Ohbayashi M, Patel N, Chong V, Bird AC, Ono SJ. Identification of anti-retinal antibodies in patients with age-related macular degeneration. Exp Mol Pathol. 2012;93(2):193–9.

126. Adamus G, Chew EY, Ferris FL, Klein ML. Prevalence of anti-retinal autoantibodies in different stages of age-related macular degeneration. BMC Ophthalmol. 2014;14(1):154.

127. Adamus G. Can innate and autoimmune reactivity forecast early and advance stages of age-related macular degeneration? Autoimmun Rev. 2017;16(3):231–6.

128. Eliasson M, Olsson A, Palmcrantz E, Wiberg K, Inganas M, Guss B, et al. Chimeric IgG-binding receptors engineered from staphylococcal protein A and streptococcal protein G. J Biol Chem. 1988;263(9):4323–7.

129. Choe W, Durgannavar TA, Chung SJ. Fc-binding ligands of immunoglobulin G: an overview of high affinity proteins and peptides. Materials. 2016;9(12):994.

130. Reuter JA, Spacek D, Snyder MP. High-throughput sequencing technologies. Mol Cell. 2015;58(4):586–97.

131. Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in bipolymers. Proc Second Int Conf Intell Syst Mol Biol. 1994;28–36.

132. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, et al. MEME suite: tools for motif discovery and searching. Nucleic Acids Res. 2009;37:202–8.

133. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc

Natl Acad Sci. 1992;89(22):10915–9.

134. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning. Vol. 112. Springer; 2012.

135. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25(17):3389–402.

136. de Castro E, Sigrist CJA, Gattiker A, Bulliard V, Langendijk-Genevaux PS, Gasteiger E, et al. ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. Nucleic Acids Res. 2006;34:362–5.

137. Sadam H, Pihlak A, Kivil A, Pihelgas S, Jaago M, Adler P, et al. Prostaglandin D2 receptor DP1 antibodies predict vaccine-induced and spontaneous narcolepsy type 1: large-scale study of antibody profiling. EBioMedicine. 2018;29:47–59.

138. Carmona SJ, Nielsen M, Schafer-Nielsen C, Mucci J, Altcheh J, Balouz V, et al. Towards high-throughput immunomics for infectious diseases: use of next-generation peptide microarrays for rapid discovery and mapping of antigenic determinants. Mol Cell Proteomics. 2015;14(7):1871–84.

139. Stafford P, Cichacz Z, Woodbury NW, Johnston SA. Immunosignature system for diagnosis of cancer. Proc Natl Acad Sci U S A. 2014;111(30):E3072-80.

140. Liotta LA, Espina V, Mehta AI, Calvert V, Rosenblatt K, Geho D, et al. Protein microarrays: meeting analytical challenges for clinical applications. Cancer Cell. 2003;3(4):317–25.

141. Paull ML, Daugherty PS. Mapping serum antibody repertoires using peptide libraries. Curr Opin Chem Eng. 2018;19:21–6.

142. Rodi DJ, Makowski L. Phage-display technology–finding a needle in a vast molecular haystack. Curr Opin Biotechnol. 1999;10(1):87–93.

143. Ryvkin A, Ashkenazy H, Weiss-Ottolenghi Y, Piller C, Pupko T, Gershoni JM. Phage display peptide libraries: deviations from randomness and correctives. Nucleic Acids Res. 2018;46(9):e52.

144. Lowman HB, Wells JA. Monovalent phage display: a method for selecting variant proteins from random libraries. Methods. 1991;3(3):205–16.

145. Kayushin AL, Korosteleva MD, Miroshnikov AI, Kosch W, Zubov D, Piel N. A convenient approach to the synthesis of trinucleotide phosphoramidites—synthons for the generation of oligonucleotide/peptide libraries. Nucleic Acids Res. 1996;24(19):3748–55.

146. Mauriala T, Auriola S, Azhayev A, Kayushin A, Korosteleva M, Miroshnikov A. HPLC electrospray mass spectrometric characterization of trimeric building blocks for oligonucleotide synthesis. J Pharm Biomed Anal. 2004;34(1):199–206.

147. Yagodkin A, Azhayev A, Roivainen J, Antopolsky M, Kayushin A, Korosteleva M, et

al. Improved synthesis of trinucleotide phosphoramidites and generation of randomized oligonucleotide libraries. Nucleosides, Nucleotides and Nucleic Acids. 2007;26(5):473–97.

148. Kenrick SA, Daugherty PS. Bacterial display enables efficient and quantitative peptide affinity maturation. Protein Eng Des Sel. 2010;23(1):9–17.

149. Rodi DJ, Soares AS, Makowski L. Quantitative assessment of peptide sequence diversity in M13 combinatorial peptide phage display libraries. J Mol Biol. 2002;322(5):1039–52.

150. Kenrick S, Rice J, Daugherty P. Flow cytometric sorting of bacterial surface-displayed libraries. Curr Protoc Cytom. 2007;42(1):4.6.1-4.6.27.

151. Illumina. 16S metagenomic sequencing library preparation. 2013;1–28.

152. Philip SS, Ahrens K, Shayevich C, de la Roca R, Williams M, Wilson D, et al. Evaluation of a new point-of-care serologic assay for herpes simplex virus type 2 infection. Clin Infect Dis. 2008;47(10):e79–82.

153. Dietzschold B, Eisenberg RJ, De Leon MP, Golub E, Hudecz F, Varrichio A, et al. Fine structure analysis of type-specific and type-common antigenic sites of herpes simplex virus glycoprotein D. J Virol. 1984;52(2):431–5.

154. Pan M, Wang X, Liao J, Yin D, Li S, Pan Y, et al. Prediction and identification of potential immunodominant epitopes in glycoproteins B, C, E, G, and I of herpes simplex virus type 2. Clin Dev Immunol. 2012;2012.

155. Clo E, Kracun SK, Nudelman AS, Jensen KJ, Liljeqvist J-A, Olofsson S, et al. Characterization of the viral O-glycopeptidome: a novel tool of relevance for vaccine design and serodiagnosis. J Virol. 2012;86(11):6268–78.

156. Kalantari-Dehaghi M, Chun S, Chentoufi AA, Pablo J, Liang L, Dasgupta G, et al. Discovery of potential diagnostic and vaccine antigens in herpes simplex virus 1 and 2 by proteome-wide antibody profiling. J Virol. 2012;JVI-05194.

157. Alami Chentoufi A, Kritzer E, Yu DM, Nesburn AB, BenMohamed L. Towards a rational design of an asymptomatic clinical herpes vaccine: the old, the new, and the unknown. Clin Dev Immunol. 2012;2012.

158. Belshe RB, Leone PA, Bernstein DI, Wald A, Levin MJ, Stapleton JT, et al. Efficacy results of a trial of a herpes simplex vaccine. N Engl J Med. 2012;366(1):34–43.

159. Kalimo KO, Marttila RJ, Granfors K, Viljanen MK. Solid-phase radioimmunoassay of human immunoglobulin M and immunoglobulin G antibodies against herpes simplex virus type 1 capsid, envelope, and excreted antigens. Infect Immun. 1977;15(3):883–9.

160. Kurtz JB. Specific IgG and IgM antibody responses in herpes-simplex-virus infections. J Med Microbiol. 1974;7(3):333–41.

161. Halperin RF, Stafford P, Johnston SA. Exploring antibody recognition of sequence space through random-sequence peptide microarrays. Mol Cell Proteomics.

2011;10(3):M110.000786.

162. Ranallo S, Rossetti M, Plaxco KW, Vallée-Bélisle A, Ricci F. A modular, DNA-based beacon for single-step fluorescence detection of antibodies and other proteins. Angew Chemie. 2015;127(45):13214–6.

163. Kang D, Sun S, Kurnik M, Morales D, Dahlquist FW, Plaxco KW. New architecture for reagentless, protein-based electrochemical biosensors. J Am Chem Soc. 2017;139(35):12113–6.

164. Bozekowski JD, Graham AJ, Daugherty PS. High-titer antibody depletion enhances discovery of diverse serum antibody specificities. J Immunol Methods. 2018;455:1–9.

165. The Eye Diseases Prevalence Research Group. Causes and prevalence of visual impairment among adults in the united states. Arch Ophthalmol. 2004;122(4):477–85.

166. Klaver CCW, Wolfs RCW, Vingerling JR, Hofman A, de Jong PTVM. Age-specific prevalence and causes of blindness and visual impairment in an older population: the Rotterdam Study. Arch Ophthalmol. 1998;116(5):653–8.

167. Wong WL, Su X, Li X, Cheung CMG, Klein R, Cheng CY, et al. Global prevalence of age-related macular degeneration and disease burden projection for 2020 and 2040: a systematic review and meta-analysis. Lancet Glob Heal. 2014;2(2):e106–16.

168. Ferris FL, Wilkinson CP, Bird A, Chakravarthy U, Chew E, Csaky K, et al. Clinical classification of age-related macular degeneration. Ophthalmology. 2013;120(4):844–51.

169. Lambert NG, ElShelmani H, Singh MK, Mansergh FC, Wride MA, Padilla M, et al. Risk factors and biomarkers of age-related macular degeneration. Prog Retin Eye Res. 2016;54:64–102.

170. Doyle SL, Campbell M, Ozaki E, Salomon RG, Mori A, Kenna PF, et al. NLRP3 has a protective role in age-related macular degeneration through the induction of IL-18 by drusen components. Nat Med. 2012;18(5):791–8.

171. Mitta VP, Christen WG, Glynn RJ, Semba RD, Ridker PM, Rimm EB, et al. C-reactive protein and the incidence of macular degeneration: pooled analysis of 5 cohorts. JAMA Ophthalmol. 2013;131(4):507–13.

172. Hong T, Tan AG, Mitchell P, Wang JJ. A review and meta-analysis of the association between C-reactive protein and age-related macular degeneration. Surv Ophthalmol. 2011;56(3):184–94.

173. Silva AS, Teixeira AG, Bavia L, Lin F, Velletri R, Belfort R, et al. Plasma levels of complement proteins from the alternative pathway in patients with age-related macular degeneration are independent of Complement Factor H Tyr^{402}His polymorphism. Mol Vis. 2012;18:2288–99.

174. Cherepanoff S, Mitchell P, Wang JJ, Gillies MC. Retinal autoantibody profile in early age-related macular degeneration: preliminary findings from the Blue Mountains Eye Study. Clin Exp Ophthalmol. 2006;34(6):590–5.

175. Age-Related Eye Disease Study Research Group. A simplified severity scale for age-related macular degeneration: AREDS report no. 18. Arch Ophthalmol. 2005;123(11):1570–4.

176. Ebrahem Q, Renganathan K, Sears J, Vasanji A, Gu X, Lu L, et al. Carboxyethylpyrrole oxidative protein modifications stimulate neovascularization: implications for age-related macular degeneration. Proc Natl Acad Sci U S A. 2006;103(36):13480–4.

177. Laíns I, Kelly RS, Miller JB, Silva R, Vavvas DG, Kim IK, et al. Human plasma metabolomics study across all stages of age-related macular degeneration identifies potential lipid biomarkers. Ophthalmology. 2018;125(2):245–54.

178. Youden WJ. Index for rating diagostic tests. Cancer. 1950;3(1):32–5.

179. Subramanyam M, Goyal J. Translational biomarkers: from discovery and development to clinical practice. Drug Discov Today Technol. 2016;21:3–10.

180. McDermott JE, Wang J, Mitchell H, Webb-Robertson BJ, Hafen R, Ramey J, et al. Challenges in biomarker discovery: combining expert insights with statistical analysis of complex omics data. Expert Opin Med Diagn. 2013;7(1):37–51.

181. Mallick P, Kuster B. Proteomics: a pragmatic perspective. Nat Biotechnol. 2010;28(7):695–709.

182. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10(1):57–63.

183. Christofanilli M, Budd T, Ellis M, Stopeck A, Matera J, Miller MC, et al. Circulating tumor cells, disease progression, and survival in metastatic breast cancer. N Engl J Med. 2004;351(8):781–91.

184. Georgiou G, Ippolito GC, Beausang J, Busse CE, Wardemann H, Quake SR. The promise and challenge of high-throughput sequencing of the antibody repertoire. Nat Biotechnol. 2014;32(2):158–68.

185. Burbelo PD, O'Hanlon TP. New autoantibody detection technologies yield novel insights into autoimmune disease. Curr Opin Rheumatol. 2014;26(6):717–23.

186. He J, Wu J, Jiao Y, Wagner-Johnston N, Ambinder RF, Diaz LA, et al. IgH gene rearrangements as plasma biomarkers in Non-Hodgkin's lymphoma patients. Oncotarget. 2011;2(3):178–85.

187. Boyd SD, Marshall EL, Merker JD, Maniar JM, Zhang LN, Sahaf B, et al. Measurement and clinical monitoring of human lymphocyte clonality by massively parallel V-D-J pyrosequencing. Sci Transl Med. 2009;1(12):12ra23.

188. Tschumper RC, Asmann YW, Hossain A, Huddleston PM, Wu X, Dispenzieri A, et al. Comprehensive assessment of potential multiple myeloma immunoglobulin heavy chain V-D-J intraclonal variation using massively parallel pyrosequencing. Oncotarget. 2012;3(4):502–13.

189. Laserson U, Vigneault F, Gadala-Maria D, Yaari G, Uduman M, Heiden JA Vander, et al. High-resolution antibody dynamics of vaccine-induced immune responses. Proc Natl Acad Sci. 2014;111(13):4928–33.

190. Wang C, Liu Y, Cavanagh MM, Le Saux S, Qi Q, Roskin KM, et al. B-cell repertoire responses to varicella-zoster vaccination in human identical twins. Proc Natl Acad Sci U S A. 2015;112(2):500–5.

191. Cheung WC, Beausoleil SA, Zhang X, Sato S, Schieferl SM, Wieler JS, et al. A proteomics approach for the identification and cloning of monoclonal antibodies from serum. Nat Biotechnol. 2012;30(5):447–52.

192. Robinson WH. Sequencing the functional antibody repertoire—diagnostic and therapeutic discovery. Nat Rev Rheumatol. 2015;11(3):171–82.

193. Anderson NL, Anderson NG. The human plasma proteome: history, character, and diagnostic prospects. Mol Cell Proteomics. 2002;1(11):845–67.

194. Tirumalai RS, Chan KC, Prieto DA, Issaq HJ, Conrads TP, Veenstra TD. Characterization of the low molecular weight human serum proteome. Mol Cell Proteomics. 2003;2(10):1096–103.

195. Echan LA, Tang HY, Ali-Khan N, Lee K, Speicher DW. Depletion of multiple high-abundance proteins improves protein profiling capacities of human serum and plasma. Proteomics. 2005;5(13):3292–303.

196. Fonslow BR, Stein BD, Webb KJ, Xu T, Choi J, Park SK, et al. Digestion and depletion of abundant proteins improves proteomic coverage. Nat Methods. 2013;10(1):54–6.

197. O'Neil D, Glowatz H, Schlumpberger M. Ribosomal RNA depletion for efficient use of RNA-seq capacity. Curr Protoc Mol Biol. 2013;4–19.

198. Yang L, Lang JC, Balasubramanian P, Jatana KR, Schuller D, Agrawal A, et al. Optimization of an enrichment process for circulating tumor cells from the blood of head and neck cancer patients through depletion of normal cells. Biotechnol Bioeng. 2009;102(2):521–34.

199. Ozkumur E, Shah AM, Ciciliano JC, Emmink BL, Miyamoto DT, Brachtel E, et al. Inertial focusing for tumor antigen-dependent and -indpendent sorting of rare circulating tumor cells. Sci Transl Med. 2013;5(179):179ra47.

200. Schirwitz C, Loeffler FF, Felgenhauer T, Stadler V, Breitling F, Bischoff FR. Sensing immune responses with customized peptide microarrays. Biointerphases. 2012;7(1–4):47–55.

201. Wang LF, Yu M. Epitope identification and discovery using phage display libraries: applications in vaccine development and diagnostics. Curr Drug Targets. 2004;5(1):1–15.

202. Ayoglu B, Schwenk JM, Nilsson P. Antigen arrays for profiling autoantibody repertoires. Bioanalysis. 2016;8(10):1105–26.

203. Björhall K, Miliotis T, Davidsson P. Comparison of different depletion strategies for improved resolution in proteomic analysis of human serum samples. Proteomics. 2005;5(1):307–17.

204. Cohen JI. Epstein-Barr virus infection. N Engl J Med. 2000;343(7):481–92.

205. Cello J, Samuelson A, Stalhandske P, Svennerholm B, Jeansson S, Forsgren M. Identification of group-common linear epitopes in structural and nonstructural proteins of enteroviruses by using synthetic peptides. J Clin Microbiol. 1993;31(4):911–6.

206. Oberste MS, Maher K, Kilpatrick DR, Flemister MR, Brown BA, Pallansch MA. Typing of human enteroviruses by partial sequencing of VP1. J Clin Microbiol. 1999;37(5):1288–93.

207. Ding Y, Chen X, Qian B, Wu G, He T, Feng J, et al. Characterization of the antibody response against EV71 capsid proteins in Chinese individuals by NEIBM-ELISA. Sci Rep. 2015;5:10636.

208. Yadav AK, Bhardwaj G, Basak T, Kumar D, Ahmad S, Priyadarshini R, et al. A systematic analysis of eluted fraction of plasma post immunoaffinity depletion: implications in biomarker discovery. PLoS One. 2011;6(9):e24442.

209. Kidd BA, Peters LA, Schadt EE, Dudley JT. Unifying immunology with informatics and multiscale biology. Nat Immunol. 2014;15(2):118–27.

210. Nybakken GE, Oliphant T, Johnson S, Burke S, Diamond MS, Fremont DH. Structural basis of West Nile virus neutralization by a therapeutic antibody. Nature. 2005;437(7059):764–9.

211. Michaud GA, Salcius M, Zhou F, Bangham R, Bonin J, Guo H, et al. Analyzing antibody specificity with whole proteome microarrays. Nat Biotechnol. 2003;21(12):1509–12.

212. Soria-Guerra RE, Nieto-Gomez R, Govea-Alonso DO, Rosales-Mendoza S. An overview of bioinformatics tools for epitope prediction: implications on vaccine development. J Biomed Inform. 2015;53:405–14.

213. Forsström B, Axnäs BB, Stengele KP, Bühler J, Albert TJ, Richmond TA, et al. Proteome-wide epitope mapping of antibodies using ultra-dense peptide arrays. Mol Cell Proteomics. 2014;13(6):1585–97.

214. Ahmad TA, Eweida AE, Sheweita SA. B-cell epitope mapping for the design of vaccines and effective diagnostics. Trials Vaccinol. 2016;5:71–83.

215. Pashova S, Schneider C, von Gunten S, Pashov A. Antibody repertoire profiling with mimotope arrays. Hum Vaccines Immunother. 2017;13(2):314–22.

216. Buus S, Rockberg J, Forsström B, Nilsson P, Uhlen M, Schafer-Nielsen C. High-resolution mapping of linear antibody epitopes using ultrahigh-density peptide microarrays. Mol Cell Proteomics. 2012;11(12):1790–800.

217. Hansen CS, Østerbye T, Marcatili P, Lund O, Buus S, Nielsen M. ArrayPitope:

automated analysis of amino acid substitutions for peptide microarray-based antibody epitope mapping. PLoS One. 2017;12(1):e0168453.

218. Hecker M, Fitzner B, Wendt M, Lorenz P, Flechtner K, Steinbeck F, et al. High-density peptide microarray analysis of IgG autoantibody reactivities in serum and cerebrospinal fluid of multiple sclerosis patients. Mol Cell Proteomics. 2016;15(4):1360–80.

219. Sundström P, Nyström M, Ruuth K, Lundgren E. Antibodies to specific EBNA-1 domains and HLA DRB1*1501 interact as risk factors for multiple sclerosis. J Neuroimmunol. 2009;215(1):102–7.

220. Rand KH, Houck H, Denslow ND, Heilman KM. Molecular approach to find target(s) for oligoclonal bands in multiple sclerosis. J Neurol Neurosurg Psychiatry. 1998;65(1):48–55.

221. Bailey T, Elkan C. Unsupervised learning of multiple motifs using expected minimization. Mach Learn. 1995;21:51–80.

222. Amornsiripanitch N, Hong S, Campa MJ, Frank MM, Gottlin EB, Patz EF. Complement factor H autoantibodies are associated with early stage NSCLC. Clin Cancer Res. 2010;16(12):3226–31.

223. Burton DR, Desrosiers RC, Doms RW, Koff WC, Kwong PD, Moore JP, et al. HIV vaccine design and the neutralizing antibody problem. Nat Immunol. 2004;5(3):233–6.

224. Ferrari G, Haynes BF, Koenig S, Nordstrom JL, Margolis DM, Tomaras GD. Envelope-specific antibodies and antibody-derived molecules for treating and curing HIV infection. Nat Rev Drug Discov. 2016;15(12):823–34.

225. Wang CY, Walfield AM. Site-specific peptide vaccines for immunotherapy and immunization against chronic diseases, cancer, infectious diseases, and for veterinary applications. Vaccine. 2005;23(17–18):2049–56.

226. Thomsen MCF, Nielsen M. Seq2Logo: a method for construction and visualization of amino acid binding motifs and sequence profiles including sequence weighting, pseudo counts and two-sided representation of amino acid enrichment and depletion. Nucleic Acids Res. 2012;40(W1):281–7.

227. Fleri W, Paul S, Dhanda SK, Mahajan S, Xu X, Peters B, et al. The immune epitope database and analysis resource in epitope discovery and synthetic vaccine design. Front Immunol. 2017;8:278.

228. Saha S, Bhasin M, Raghava GPS. Bcipep: a database of B-cell epitopes. BMC Genomics. 2005;6(1):79.

229. Ibsen KN, Daugherty PS. Prediction of antibody structural epitopes via random peptide library screening and next generation sequencing. J Immunol Methods. 2017;451:28–36.

230. Kowalsky CA, Faber MS, Nath A, Dann HE, Kelly VW, Liu L, et al. Rapid fine conformational epitope mapping using comprehensive mutagenesis and deep

sequencing. J Biol Chem. 2015;290(44):26457–70.

231. Henriques ST, Thorstholm L, Huang Y-H, Getz JA, Daugherty PS, Craik DJ. A novel quantitative kinase assay using bacterial surface display and flow cytometry. PLoS One. 2013;8(11):e80474.

232. DeKosky BJ, Kojima T, Rodin A, Charab W, Ippolito GC, Ellington AD, et al. In-depth determination and analysis of the human paired heavy- and light-chain antibody repertoire. Nat Med. 2015;21(1):86–91.