Determination of class II peptide-MHC repertoires and recognition via large yeast-displayed libraries

by

Charles Garrett Rappazzo

B.S. Bioengineering
Cornell University, 2014

M.Eng. Biomedical Engineering
Cornell University, 2015

SUBMITTED TO THE DEPARTMENT OF BIOLOGICAL ENGINEERING IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY IN BIOLOGICAL ENGINEERING
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

JUNE 2020

Signature of Author: ____
Department of Biological Engineering
May 15, 2020

Certified by: ____
Michael E. Birnbaum, PhD
Assistant Professor of Biological Engineering
Thesis supervisor

Accepted by: ____
Paul Blainey, PhD
Associate Professor of Biological Engineering
Co-chair of Graduate Program, Department of Biological Engineering

THESIS COMMITTEE

Michael E. Birnbaum, PhD: Assistant Professor of Biological Engineering (Thesis supervisor)

K. Dane Wittrup, PhD: Carbon P. Dubbs Professor of Chemical Engineering and Biological Engineering (Thesis committee chair)

Tyler Jacks, PhD: David H. Koch Professor of Biology

Stefani Spranger, PhD: Howard S. and Linda B. Stern Career Development Assistant Professor of Biology

Determination of class II peptide-MHC repertoires and recognition via large yeast-displayed libraries

by

Charles Garrett Rappazzo

Submitted to the Department of Biological Engineering
on May 15, 2020 in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy in
Biological Engineering

ABSTRACT

T cells occupy essential roles throughout the immune system to prevent and limit disease. As such, breakdowns in their function and recognition underlie poor clinical outcomes across diverse maladies including pathogen infection, cancer, autoimmunity, allergies, and transplant rejection. Yet, when properly directed, T cells drive potent protective and therapeutic responses in prophylactic vaccinations and novel immunotherapies. Therefore, understanding and harnessing T cell function and recognition is of great importance to improving patient care and addressing currently unmet clinical needs.

The function and recognition of T cells are driven through their T cell receptors (TCRs), which bind with great specificity to peptide-MHCs (pMHCs), Major Histocompatibility Complex proteins displaying tissue- and disease-specific peptide antigens derived from their host cell or its surroundings. However, to specifically and comprehensively present and surveil antigens across highly divergent maladies, extreme diversity is required of both the population-level TCR and pMHC repertoires. However, this same diversity which drives T cell function also confounds generalized understanding of these repertoires, as well as their recognition. Therefore, there has been considerable recent interest in the development and application of tools to comprehensively define, predict, and screen these repertoires and their recognition at high throughput.

In this thesis, I both utilize and build upon these tools to define TCR and pMHC repertoires and explore their recognition, particularly with yeast-displayed pMHC libraries for CD4^{+} T cell recognition of class II pMHCs, and especially in the context of cancer. Using these technologies, I empirically define pMHC repertoires, explore the antigenic basis of TCR repertoire convergence in a preclinical tumor model, and explore the antigen reactivity of human T cells with clinical relevance. While these results provide detailed insights into the specific TCRs and pMHCs studied, they also provide guidance for future avenues in the exploration of TCR and pMHC repertoires and their recognition.

Thesis Supervisor: Michael E. Birnbaum, PhD
Title: Assistant Professor of Biological Engineering

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

Writing a thesis acknowledgement section is a rare chance to editorialize and reflect on the state of science, as well as your own position within the scientific community at the outset of your career. So, here we go.

This thesis was written in its entirety from self-isolation amidst the global pandemic of SARS-Cov-2, the causative agent of Covid-19. At the time of writing, March 19, 2020, this novel coronavirus has infected 230,000 people world-wide, caused 9,300 deaths, and halted life-as-usual throughout most of the world. At this time, the only certainties are that this pandemic will get worse, and eventually, it will get better. However, in all likelihood, this pandemic will not fully resolve without the intervention of scientists.

This intervention may be in the form of an antiviral medication, a therapeutic antibody, a prophylactic vaccine, or some yet-to-be-discovered approach. Yet each of these interventions highlights the critical importance of biological research, both in the clinic and in the laboratory, in confronting the world's greatest maladies. In light of this, I am proud to have contributed – in very small part – to the ever-growing fields of immunology and biological engineering during my years at MIT. But more importantly, I am proud to join a community of scientists who continuously strive to improve our understanding of the world, and to find new ways to treat and prevent diseases. Long after Covid-19 is a distant memory and public interest in immunology research wanes, I am confident that this community will continue to endeavor to improve patient care with the same vigor.

My own journey into science has been a somewhat winding path. When I graduated high school, I never would have envisioned that I would be writing a doctoral thesis 10 years later. However, I have been fortunate to have had the inspiration of many talented teachers, the encouragement of steadfast mentors, the loving support of my friends, family, and wife, as well as a whole lot of dumb luck along the way. The remainder of this section is dedicated to the people who made this all possible.

From my time at Cornell University, two professors, Larry Bonassar and Dave Putnam, deserve special acknowledgment. The fall of my junior year, I took a course taught by Larry and his organized, clear, and well-motivated teaching gave me an increased appreciation of biomedical engineering. Two years later, I TA'ed this same course and during this time I came to see him as a mentor. I am forever grateful for his insights into biological engineering, science, and academia, and for his encouragement, advice, and support in my applications to graduate school.

However, my applications to graduate school would have been wholly unsuccessful without my first research advisor, Dave Putnam. I met Dave when he presented on reprogramming the immune system using a bacterially derived vaccine during a senior-year rotating lecturer course. At the time of this presentation, I knew next to nothing about immunology and was still planning on applying to medical schools, not graduate programs. I'm not sure if it was the subject matter or Dave's uniquely energetic presentation, but I was hooked. Even though I was already a senior, Dave agreed to meet with me and let me join his group to work on a universal Influenza A virus vaccine, spurring my interest in immunology and starting research career. I went on to work in the group for a year and a half, staying for an additional year during a Master of Engineering degree, and have so many fond memories of this time. I am thankful to Dave for his continued mentorship, enthusiastic encouragement, his overly generous recommendation letters. I am also so thankful to the many members of the Putnam lab who taught me how to be a researcher and who inspired me to pursue a career in science, despite repeatedly advising me otherwise. In particular, I am thankful to my two mentors in the lab, Cassie and Hannah, and all the other members of team OMV, past and present.

I have also been very fortunate to have very supportive mentors and lab mates at MIT, and I could not have asked for a better thesis advisor than Michael Birnbaum. I first heard of Michael as the new faculty member just starting a lab. Although his research was well over my head at the time, he was smart, approachable, and down-to-earth, making it an easy decision to join his lab (a generous description for the time considering its limited size) that fall. The ensuing four and a half years have had their ups and downs and seen the Birnbaum lab grow from 4 members to 13 (not including the army of undergrads), but through it all, Michael has been a constant source of mentorship, encouragement, and much needed levity. Sorry for all the guff (kinda). To the members of the Birnbaum lab, past and present, thank you for many fruitful conversations and lab meetings, and for pushing me to be a better researcher. I look forward to what you will all accomplish. To my undergraduate student Ben, thank you for your trust and enthusiasm – it was a pleasure to see you grow in your knowledge and abilities. A special thank you to Christine and Anna, who keep the lab running, and to Brooke, who collaborated on nearly all my projects; none of this would have gotten done without you three.

To my committee members, Dane, Tyler, and Stefani, thank you for your consistent support, mentorship, and insights. I am very thankful to have had you as committee members and to have had the chance to collaborate with your lab members on my research throughout the years.

To my friends at MIT and beyond, thank you for so many fond memories throughout these years. Once this pandemic ends, I can’t wait to have you all over for more dinners, drinks, and Lillydog time. To my family, I have no way to fully express how grateful I am for all the ways you have loved and supported me not only during my time at MIT, but every day before and hopefully every day after. A special thank you to my Mom and Dad, who have sacrificed so much to give me a good home, a good education, and a happy and healthy life - all that I accomplish is because of you. To my dog Lilly, you can’t read, but you’re a very good girl. Lastly, to my wife, Rory - thank you for being my constant companion, cheerleader, sounding board, copy editor, comfort, and joy. I can’t wait to start the next chapter with you.

CHAPTER I. Introduction

1.1 What are T cells?

The principal function of the immune system is to prevent and limit disease throughout the body. This function is achieved through the collective action of trillions of individual immune cells interacting with our organs, our environment, commensal microbes, and each other. When they sense danger, such as pathogens, cancers, or injury, these cells are activated to mobilize massive and coordinated protective responses. Yet these cells also have crucial roles in limiting inflammation, promoting wound healing, and maintaining homeostasis. Therefore, the immune system must be able to differentiate self from foreign, healthy from diseased, and commensal from pathogenic [1]. As such, breakdowns in the function or recognition of the immune system underlie diverse ailments from pathogen infection and cancer [1, 2] to autoimmune diseases and allergies [3, 4], as well as chronic and degenerative diseases [5, 6].

The immune system is broadly partitioned by its function and recognition into the innate and adaptive immune systems; the innate immune system generates rapid responses against pathogens, toxins, and cell damage through evolutionarily conserved motifs and receptor, whereas the adaptive immune system generates slower but more specific and specialized responses through highly variable immune receptors [1]. In addition, the adaptive immune system uniquely possesses memory, allowing it to generate memory cell populations that can produce larger, more rapid, more specialized, and even more specific responses upon re-exposure to a given pathogen or disease [1, 7]. As such, the innate immune system serves as a first line of defense, containing potential dangers and recruiting other immune cells, while the adaptive immune system performs higher-level tasks of distinguishing between potential threats and tailoring its response accordingly.

The adaptive immune system is primarily comprised of two cell types: B and T lymphocytes [1, 7]. While both cell types display memory and specificity in their activation and functions, these functions are highly distinct, and are driven by their divergent recognition. In particular, B cell receptors (BCRs) are comprised of two identical recognition domains that bind directly to a diverse array of extracellular proteins. In contrast T cell receptors (TCRs) encode a single recognition site that binds to specialized cell-surface immune proteins called Major Histocompatibility Complexes (MHCs). Furthermore, upon activation, B cells can edit their receptors for higher affinity to their target through somatic hypermutation, and can secrete their receptors as antibodies, which can be oligomerized for additional binding valency [7]. These antibodies can then directly bind to target far from their parent B cell to enact aggregation and direct neutralization, as well as facilitate phagocytosis and cytotoxicity. In contrast, T cells cannot edit or secrete their TCR, and their activation, localization, and function are all driven through their TCR [8]. Yet T cells are uniquely capable of recognizing and responding to threats hidden within the cell, and have roles in coordinating and modulating the immune response [7].

Although T cell populations can be even further partitioned based upon their function and recognition into αβ and γδ subsets [9], as well as innate-like natural killer T (NKT) cells [10], we will focus on their most prevalent subset, αβ T cells, which recognize MHCs displaying short, linear peptides derived from within their host cell or its environment, known as peptide-MHCs (pMHCs).

1.2 The role and recognition of T cells

Owing to their diverse composition and recognition, T cells can fulfill many functions within the immune response across highly divergent diseases. These functions can range from directing and modulating innate and B cell responses, to enacting direct cell killing, and even suppressing other T cell responses [11]. Yet these diverse functions can be largely partitioned among T cells based on their expression of two cell-surface co-receptors into CD4$^+$ 'helper' and CD8$^+$ 'killer' T cells, each with their own recognition [1,7].

As their name suggests, CD8$^+$ 'killer' T cells – also known as cytotoxic T lymphocytes (CTLs) – primarily function to directly kill infected or diseased cells. CD8$^+$ T cells recognize class I peptide-MHC (pMHC) proteins, which are ubiquitously expressed by nucleated cells in the body and principally display peptides derived from within their parent cell [1, 12]. When activated through their TCR, these cells are empowered to directly kill target cells expressing their cognate pMHC. This recognition pathway allows CD8$^+$ T cells to monitor for intracellular infection, impaired function, and even dysregulated signaling pathways, and are therefore crucial for pathogen and tumor surveillance and control [13].

In contrast, CD4$^+$ 'helper' T cells serve in broader but less direct roles, coordinating and aiding the responses of other immune cells. These cells recognize class II pMHC proteins, which are expressed by specialized antigen-presenting cells (APCs) and principally display peptides derived from their environment [1, 12]. These APCs are often phagocytic members of the innate immune system – such as macrophages, dendritic cells, and monocytes – that specialize in rapidly internalizing and processing pathogens and cell debris, but also includes B cells and specialized endothelial cells [12]. This antigen presentation pathway allows CD4$^+$ T cells to respond indirectly to both extracellular and intracellular threats. Yet, in order to respond effectively to this vast array of threats, CD4$^+$ T cells must be able to tailor their response to a threat.

As such, activated CD4$^+$ T cells can adopt a diverse array of distinct T-helper (Th) phenotypes, each with a distinct gene expression and cytokine secretion signature, to tailor the local immune response to a broad variety of threats. A detailed description of these phenotypes – such as Th1, Th2, Th17, Tfh, and Th9 – and their functions are reviewed elsewhere [14]. However, CD4$^+$ T cells can also serve in non-inflammatory, regulatory roles. In contrast to conventional CD4$^+$ T cell (also known as Tconv) populations, regulatory T cell (Treg) populations dampen the immune response and suppress T cell responses through a variety of indirect and direct modalities [15]. While activated Tconv cells can adopt a Treg phenotype [16], most Tregs are a distinct CD4$^+$ T cell lineage with differential recognition [17, 18].

Combined, these diverse functions allow T cells to occupy many crucial roles in the immune system simultaneously. But as such, breakdowns in T cell recognition and function can have disastrous consequences throughout the body, underlying poor clinical outcomes during viral and bacterial infections, facilitating continued outgrowth of tumors, and driving autoimmune diseases, allergies, and transplant rejection [19-23]. Therefore, a detailed understanding of the diverse functions of T cells – and the recognition that drives them – is crucial to our understanding of many diseases, as well as to our ability to develop novel therapies.

In fact, directing, modulating, and coopting T cell function and recognition already underlies many therapies in the clinic: T cells are crucial to forming the protective memory responses required for prophylactic vaccinations [24], can be stimulated and directed by therapeutic vaccinations [25], and can be activated to fight established tumors by novel cancer immunotherapeutics [26]. Furthermore, T cell function can be coopted to potent new purposes by redirecting their recognition through engineered TCRs and chimeric antigen receptors (CARs) [27, 28]. But while each of these applications have already had significant impacts in the clinic, further investigation of T cell function and recognition is still needed to guide further improvements of these therapies, as well as enable future modalities to address unmet clinical needs. Yet in order to fully understand T cell function and recognition, we must first understand their common driver, the T cell receptor.

1.3 Drivers of TCR and pMHC repertoire diversity

As previously discussed, the TCR drives both the recognition and function of T cells. Accordingly, the diverse functions of T cell populations rely on diverse and distinct population-scale TCR repertoires [29]. Therefore, a comprehensive understanding of T cell function and recognition requires detailed knowledge of the drivers of both the diversity and divergence observed in TCR repertoires across T cell populations, as well as how these drivers interact with diverse pMHC repertoires.

Diversity in pMHC repertoires stems from the need to comprehensively present antigenic peptides from a vast array of potential threats, both endogenous and foreign, to antigen-specific T cells, as any holes in these repertoires can blind T cells to ongoing infection or disease. To be properly sampled by both CD4$^+$ and CD8$^+$ T cell subsets, these threats also need to be presented across both class I and class II MHC proteins [12]. In addition, these pMHC complexes must be stable and long-lived in order to be properly surveilled by potentially rare clonal T cell populations [30], which are comprised of as few as 10-100 T cells throughout the entire body [31], and therefore MHCs must be specific to their displayed peptides. Yet the number of peptides required to comprehensively represent all potential immune threats is too vast to be specifically presented by any one MHC protein.

Therefore, to facilitate broad yet specific peptide presentation, MHC proteins themselves are highly diverse [32, 33]. In humans, this diversity is accomplished on an individual basis by the expression of six distinct MHC genes, three class I (HLA-A, -B, and -C) and three class II (HLA-DR, -DP, and –DQ), each with their own peptide specificities [34]. In addition, these MHC genes are extremely polymorphic, providing evolutionarily driven peptide-binding diversity on a population scale [32, 33]. In fact, the MHC locus in humans, also known as the human leukocyte antigen (HLA) locus, is the most polymorphic region in the genome [35], with over 26,000 currently known unique HLA alleles [36]. Importantly, these polymorphisms are clustered in the peptide-binding groove of MHC proteins [37], thereby imparting unique peptide-specificities. Combined, this diversity provides for the expression of up to 12 unique HLA proteins in humans (as the MHC genes are expressed co-dominantly by each chromosome) [37], and therefore up to 12 unique pMHC repertories to comprehensively present potential threats to the TCR repertoires.

To enable specific recognition of these diverse pMHC repertoires, TCR repertoires must also be highly diverse [38]. This is accomplished in the TCR repertoire by VDJ recombination, a process unique to B and T cells that allows receptor formation by recombination of a variable (V), diversity (D), and joining (J) region selected quasi-randomly from diverse collections, with untemplated nucleotide additions, deletions, and substitutions permitted at their junctions to provide additional diversity [1, 7]. These segments are then joined to a constant (C) receptor domain, and the pairing of two independently recombined receptor chains provides additional diversity. A detailed review of this process is available elsewhere [39]. This process allows each of the estimated 10^{12} human T cells [40] to theoretically choose between 10^{15}-10^{20} unique TCR combinations [41]. However, due to biases in VDJ recombination and external factors (discussed below), the true diversity of these repertoires has been recently estimated at approximately 10^{10} unique clones in humans [42].

As a direct result of VDJ recombination, the diversity of the TCR repertoire is greatest at the junction between the V, D, and J regions, and comprises the complementary-determining region 3 (CDR3) of each TCR chain [39]. Importantly, these CDR3 regions are positioned in direct contact with the MHC-displayed peptide in TCR / pMHC complexes, and are the primary drivers of TCR binding and specificity [43, 44]. In addition, these CDR3 regions contribute equally to peptide binding, and therefore the TCR alpha and beta chains are equally important to pMHC binding. The V region-encoded CDR1 and CDR2 regions of each chain contribute additional interactions with the displayed peptide – as well as the MHC peptide-binding groove – providing additional specificity to these interactions [43, 44]. Therefore, the intrinsic diversity of the TCR receptor repertoire drives diverse T cell recognition, yet is essential to provide specific recognition of the even more diverse pMHC repertoire [38].

However, the TCR repertoire not only samples the pMHC repertoire, but is shaped by it. This process primarily occurs during T cell development in the thymus, but occurs to a lesser degree in the periphery following development [8]. Within the thymus, T cell progenitors express both CD4 and CD8 once they successfully formed a TCR capable of signaling through VDJ recombination [1, 8]. These so-called double-positive thymocytes then sample pMHC molecules displayed by specialized antigen presenting cells as well as endothelial cells that express diverse self-derived peptides [45]. The strength of TCR signaling during this this developmental stage determines the T cell's fate; excessive stimulation from strong or frequent recognition of these self-antigens drives clonal deletion to eliminate potentially autoreactive T cells, but a lack of stimulation causes the T cell to die of neglect [8, 45]. Therefore, the productive TCR repertoire is derived from TCRs within this self-specificity sweet spot [46]. However, the nature of this stimulation further determines the lineage of this T cell. In particular, preferential interaction with class I or class II pMHCs drives $CD8^+$ and $CD4^+$ differentiation, respectively, due to stabilizing co-receptor interactions, and $CD4^+$ T cells which receive greater or more frequent TCR stimulation during this stage are preferentially driven towards Treg differentiation [1, 8].

Combined, these developmental influences underlie the ability of T cell populations to distinguish between self and foreign peptides [1], and are responsible for the distinct TCR repertoire of each T cell lineage [29]. These distinct repertoires are in turn responsible for the distinct recognition of these subsets, and underlie their diverse functions throughout the immune system. However, due to their immense diversity and person-to-person variability, individual TCR and pMHC repertoires and their interactions are highly unique. Therefore, tools which help define the composition of these repertoires and their recognition are of great importance for improved understanding of T cell function – as well as our ability to coopt and redirect it – across many diseases.

1.4 Defining pMHC and TCR repertoires and recognition

In line with their importance to understanding and utilizing T cell function, there is great interest in developing tools to study TCR and pMHC repertoires [47, 48], as well as TCR and pMHC recognition [49]. These tools can be used to define individual TCR and pMHC repertoires, screen interactions in high-throughput, and used to train computational algorithms that predict these repertoires and interactions.

As discussed, diverse TCR repertoires drive T cell function and recognition. Therefore, monitoring and defining the TCR repertoire can uncover temporal shifts or convergence in these repertoires that reveal underlying disease biology and T cell targeting. For example, temporal shifts within the TCR repertoire can reveal the temporal dynamics of viral infection and convergence in the TCR repertoire can point to shared targeting of immuno-dominant epitopes [50]. These insights can then be used to guide future studies of the immune response or to design therapeutic modalities.

Although early methods to define and monitor TCR repertoires through their v-region usage (via flow cytometry), and later CDR3 length (via immune spectratyping), provided partial insights into their composition, dynamics, and convergence [51], these techniques did little to define or monitor diversity in the CDR3 junctions that drive recognition. Therefore, the most powerful tool for defining these repertoires is TCR sequencing, which provides detailed coverage of entire TCR, including the CDR3. These sequencing techniques most frequently utilize reverse transcription polymerase chain reactions (RT-PCR) to amplify expressed TCR transcripts, and with the advent of next generation sequencing (NGS) can be used on a repertoire scale [47, 51]. Yet while bulk TCR sequencing of T cell populations (often focused on the TCR beta chain [52]) can be used to identify broad trends in repertoire dynamics and convergence, these approaches lose the linkage between TCR alpha and beta chains that is essential for defining antigen reactivity [43, 44]. Therefore, paired-chain TCR sequencing techniques represent the state-of-the-art for defining TCR repertoires [47]. Although these techniques were originally low throughput [53, 54], recent advances now enable thousands of T cells to be fully defined simultaneously [55-58], and may soon enable full repertoire definition in tissue- and disease-specific contexts.

In contrast, pMHC repertoires cannot be defined with sequencing, as they rely on both MHC allele and peptide diversity. However, sequencing of MHC gene usage, known as HLA typing in humans, can establish MHC allele usage in a given individual, and provide partial insights into these repertoires [59]. This is because while over 19,000 unique class I and 7,000 unique class II HLA alleles have been observed in human populations [36], a small subset of these alleles are frequently observed or dominant within given ethnic groups [60]. This allows researchers to partially generalize allele-specific pMHC repertoires between individuals through the expression of these over-represented alleles. However, these repertoires definitions are only partial because they are also dependent on peptide expression and antigen processing [12, 61]. Therefore, pMHC repertoire definition and utilization is largely dependent on computational algorithms, called antigen prediction algorithms, that allow allele-specific prediction of individual peptide / MHC interactions and pMHC repertoires by extrapolating insights from previously curated datasets [48].

The allele-specific peptide datasets used to train these antigen prediction algorithms are derived from many divergent sources. However, these methods can be broadly partitioned into MHC ligand binding and ligand elution methods [62]. MHC ligand binding assays often use recombinantly-expressed MHC to measure the binding of pre-selected peptides to measure quantitative metrics of binding, such as peptide-binding mode, affinity, or kinetics [48]. Yet while these datasets provide detailed binding information, they are low throughput and do not incorporate antigen processing pathways. In contrast, MHC ligand elution methods utilize natively expressed and loaded pMHC molecules, and therefore capture antigen presentation biases. These methods most frequently utilize mass spectrometry (MS) to determine the sequence of the bound peptide and can be used to define pMHC repertoires in both homeostatic and disease-specific contexts [63]. Furthermore, the recent development of mono-allelic engineered cell lines that express only a single MHC allele has enabled unambiguous linkage between observed peptides and their displaying MHC in MS datasets [64-66]. Yet while these MS-based technologies are high throughput and can broadly define pMHC repertoires, they have notable biases [66] and provide only qualitative binding data [67].

However, neither TCR or pMHC repertoire information can define or predict recognition at their interface. For this, researchers require an entirely separate set of tools. The earliest such tools use biochemical methods for T cell antigen discovery, such as the use fluorescently-labeled tetramers of recombinant pMHC proteins to identify antigen reactive T cells or mass spectrometry to identify pMHC proteins bound to a given TCR. However, these tools are low throughput and require knowledge of at least one component of the interaction [49], and therefore are incompatible with comprehensive or broad T cell antigen discovery [68]. In the absence of a known antigen or eluted ligands, antigen discovery requires time- and resource-intensive ‘epitope mapping’ approaches to determine the specificity of a given TCR [69]. In addition, these methods provide little-to-no information on potential antigen cross-recognition, which is essential to T cell function and comprehensive epitope coverage [38, 68]. While this cross-recognition can be assessed through site-directed mutagenesis [70], or predicted *in silico* [71, 72] these methods are largely restricted to closely related sequences [49].

Therefore, in order to achieve both comprehensive and broad T cell antigen discovery, researchers use pMHC library technologies. Broadly, these technologies consist of a pool of unique cells expressing many copies of a single pMHC protein that are then probed with endogenously or recombinantly expressed TCR [49]. To achieve single pMHC expression, the peptide and MHC are typically expressed as a single construct. The platforms used to express these pMHC library approaches can be mammalian, insect, yeast, or phage, each with unique advantages and short-comings [49]. While mammalian libraries can accurately recapitulate native protein expression and antigen presentation, their potential library size is limited to 10^5-10^6 unique peptides which can limit comprehensive epitope coverage, whereas phage-displayed libraries can express up to 10^{12} unique variants but express very few copies of their pMHC protein [73], and can fail to express many pMHC proteins [68] While insect- and yeast-displayed pMHC libraries exist within the middle of these spectrums, achieving both improved protein expression relative to phage-displayed libraries and achieving larger library sizes than mammalian platforms, yeast-displayed libraries have larger library sizes (up to 10^9 unique variants [49]) and have improved growth rates and modularity [74]. Collectively, these advantages underlie the recent increase in use of yeast-displayed libraries for T cell antigen discovery [75-78], as well as the frequent use of yeast-display for assays and optimizations of many diverse immune proteins [74].

1.5 Thesis overview and motivation

In this thesis, we both utilize and build upon these tools to define TCR and pMHC repertoires and recognition, particularly with yeast-displayed pMHC libraries for CD4$^+$ T cell recognition of class II pMHC repertoires, especially in the context of cancer.

As our lab is located within the Koch Institute for Integrative Cancer Research at MIT, understanding and treating cancer shapes our perspective of immunology, and drives many of our projects and collaborations. In particular, with the advent of potent and efficacious new cancer immunotherapies such as immune checkpoint inhibitors [26], there is great interest in the recognition of tumor-infiltrating T cells [79], which requires T cell repertoire sequencing to identify T cell clones of interest. Although early immunotherapy research focused largely on the antigen reactivity of CD8$^+$ 'killer' T cells, as they directly affect tumor clearance, there is increased appreciation of the role – and therefore the recognition – of CD4$^+$ Tconvs in the anti-tumor response [80], as well as CD4$^+$ Tregs in suppressing tumor clearance [81]. Therefore, in chapter 3 of this thesis, we use single-cell TCR sequencing to investigate the TCR repertoire of tumor-infiltrating Tregs in a preclinical model of lung adenocarcinoma and identify clones of interest. We then screen these TCRs with yeast-displayed pMHC libraries to probe for their antigen recognition.

Furthermore, the clinical successes of personalized cancer vaccines have driven great interest in improved methods for pMHC repertoire definition to facilitate improved tumor-specific antigen prediction for these vaccines [82]. This is especially true for class II pMHC antigen prediction algorithms which underperform their class I counterparts [83]. Therefore, in chapter 2 of this thesis, we develop a new method to define class II pMHC repertoires by modifying an existing yeast-displayed class II pMHC library platform. We then use this data to retrain existing class II antigen prediction algorithms and find that they significantly improve their performance, including in the context of candidate antigen identification for personalized cancer vaccines.

The insights we gained from TCR and pMHC repertoire definition are then applied in chapter 4 to define the native antigen reactivity of both clinically-relevant CD4$^+$ and CD8$^+$ T cells. Finally, through each of these projects we identified pitfalls, bottlenecks, and shortcomings in the application of our yeast-displayed pMHC libraries. Therefore, in chapter 5 of this thesis, we highlight these issues, discuss possible solutions, and discuss future applications of this powerful technology in defining TCR and pMHC repertoires and recognition.

References

1. Chaplin, D.D. Overview of the immune response. *J. Allergy. Clin. Immunol.* **125**, S3-23 (2010).

2. Gonzalez, H., Hagerling, C., & Werb, Z. Roles of the immune system in cancer: From tumor initiation to metastatic progression. *Genes Dev.* **32,** 1267-1284 (2018).

3. Rosenblum, M.D., Remedios, K.A., & Abbas, A.K. Mechanisms of human autoimmunity. *J. Clin. Invest*. **125**, 2228-33 (2015).

4. Galli, S.J., Tsai, M. & Piliponsky, A.M. The development of allergic inflammation. *Nature* **454**, 445-54 (2008).

5. Bennett, J.M., Reeves, G., Billman, G.E, & Sturmberg, J.P. Inflammation-nature's way to efficiently respond to all types of challenges: Implications for understanding and managing "the epidemic" of chronic diseases. *Front Med* **5**, 316 (2018).

6. Labzin, L.I., Heneka, M.T., & Latz, E. Innate immunity and neurodegeneration. *Annu. Rev. Med.* **69**, 437-449 (2018).

7. Bonilla, F.A. & Oettgen, H.C. Adaptive immunity. *J. Allergy. Clin. Immunol.* **125**, S33-40 (2010).

8. Kumar, B.V., Connors, T.J., & Farber, D.L. Human T cell development, localization, and function throughout life. *Immunity* **48**, 202-213 (2018).

9. Chien, Y., Meyer, C., & Bonneville, M. γδ T cells: First line of defense and beyond. *Annu. Rev. Immunol*. **32**, 121-55 (2014).

10. Wu, L. &Van Kaer, L. Natural killer T cells in health and disease. *Front. Biosci.* **3**, 236-51 (2011).

11. Pennock, N.D., White, J.T., Cross, E.W., Cheney, E.E., Tamburini, B.A., & Kedl, R.M. T cell responses: naïve to memory and everything in between. Adv Physiol Educ. **37,** 273–283 (2013).

12. Neefjes, J., Jongsma, M.L.M., Paul, P., & Bakke, O. Towards a Systems Understanding of MHC Class I and MHC Class II Antigen Presentation. *Nat. Rev. Immunol.* **11**, 823-36 (2011).

13. Zhang, N. & Bevan, M.J. CD8+ T cells: Foot soldiers of the immune system. *Immunity* **35,** 161–168 (2011).

14. Geginat, J. *et al*. Plasticity of human CD4 T cell subsets. *Front Immunol* **5**, 630 (2014).

15. Vignali, D.A.A., Collison, L.W., & Workman, C.J. How regulatory T cells work. *Nat. Rev. Immunol*. **8,** 523–532 (2008).

16. Schmitt, E.G. & Williams, C.B. Generation and function of induced regulatory T cells. *Front. Immunol.* **4**, 152 (2013).

17. Pacholczyk, R. & Kern, J. The T-cell receptor repertoire of regulatory T cells. *Immunology* **125,** 450–458 (2008).

18. Golding, A., Darko, S., Wylie, W.H., Douek, D.C., & Shevach, E.M. Deep sequencing of the TCR-β repertoire of human forkhead box protein 3 (FoxP3)+ and FoxP3– T cells suggests that they are completely distinct and non-overlapping. *Clin. Exp. Immunol*. **188**, 12–21 (2017).

19. Blackwell, J. M., Jamieson, S. E., & Burgner, D. HLA and infectious diseases. *Clin. Microbiol. Rev*. **22**, 370-85 (2009).

20. Hadrup, S., Donia, M., Thor-Straten, P. Effector CD4 and CD8 T cells and their role in the tumor microenvironment. *Cancer Microenviron*. **6**, 123-133 (2013).

21. Bluestone J.A., Bour-Jordan, H., Cheng, M., & Anderson, M. T cells in the control of organ-specific autoimmunity. *J. Clin. Invest.* **125**, 2250-2260 (2015).

22. Woodfolk, J. A. T-cell responses to allergens. *J. Allergy Clin Immunol*. **119**, 280-294 (2007).

23. Issa, F., Schiopu, A., Wood, K.J. Role of T cells in graft rejection and transplantation tolerance. *Expert Rev. Clin. Immunol*. **6**, 155-169 (2010).

24. Sallusto, F., Lanzavecchia, A., Araki, K., & Ahmed, R. From vaccines to memory and back. *Immunity* **33**, 451–463 (2010).

25. Gilbert, S.C. T-cell-inducing vaccines – what's the future. *Immunology* **135**, 19–26 (2012).

26. Pardoll, D.M. The blockade of immune checkpoints in cancer immunotherapy. *Nat. Rev. Cancer* **12**, 252-64 (2012).

27. Zhao, L. & Cao, Y.J. Engineered T cell therapy for cancer in the clinic. *Front. Immunol.* **10**, 2250 (2019).

28. Miliotou, A.N. & Papadopoulou, L.C. CAR T-cell therapy: A new era in cancer immunotherapy. *Curr. Pharm. Biotechnol*. **19**, 5-18 (2018).

29. Izraelson, M. *et al*. Comparative analysis of murine T-cell receptor repertoires. *Immunology* **153,** 133-144 (2018).

30. Baumgartner, C.K., Ferrante, A., Nagaoka, M., Gorski, J., & Malherbe, L.P. Peptide-MHC class II complex stability governs CD4 T cell clonal selection. *J. Immunol*. **184**, 573-81 (2010).

31. Jenkins, M.K. & Moon, J.J. The role of naive T cell precursor frequency and recruitment in dictating immune response magnitude. *J. Immunol.* **188**, 4135-40 (2012).

32. Sommer, S. The importance of immune gene variability (MHC) in evolutionary ecology and conservation. *Front. Zool*. **2**, 16 (2005).

33. Piertney, S.B. & Oliver, M.K. The evolutionary ecology of the major histocompatibility complex. *Heredity* **96**, 7-21 (2006).

34. Paul, S., Weiskopf, D., Angelo, M.A., Sidney, J., Peters, B., & Sette, A. HLA class I alleles are associated with peptide-binding repertoires of different size, affinity, and immunogenicity. *J. Immunol.* **191,** 5831-9 (2013).

35. Hughes, A. & Hughes, M. Natural selection on the peptide-binding regions of major histocompatibility complex molecules. *Immunogenetics* **42**, 233–243 (1995).

36. Robinson, J., Halliwell, J.A., Hayhurst, J.H., Flicek, P., Parham, P., & Marsh, S.G.E. The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Res*. **43**, D423-431 (2015).

37. Janeway C.A. *et al.* The major histocompatibility complex and its functions. *Immunobiology: The Immune System in Health and Disease.* (Garland Science, 2001).

38. Sewell, A.K. Why must T cells be cross-reactive? *Nat. Rev. Immunol.* **12**, 669-77 (2012).

39. Jung, D. & Alt, F.W. Unraveling V(D)J recombination; Insights into gene regulation. *Cell* **116**, 299-311 (2004).

40. Arstila, T.P. Casrouge, A., Baron, V., Even, J., Kanellopoulos, J., & Kourilsky, P. A direct estimate of the human alphabeta T cell receptor diversity. Science 286, 958-61 (1999).

41. Laydon, D.J., Bangham, C.R.M., & Asquith, B. Estimating T-cell repertoire diversity: Limitations of Classical Estimators and a New Approach. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **370**, 1675 (2015).

42. Lythe, G., Callard, R.E., Hoare, R.L., & Molina-París, C. How many TCR clonotypes does a body maintain? *J. Theor. Biol.* **389**, 214-24 (2016).

43. Zoete, V., Irving, M., Ferber, M., Cuendet, M.A. & Michielin, O. Structure-based, rational design of T cell ceceptors. *Front. Immunol*. **4**, 268 (2013).

44. Garcia, K.C. & Adams, E.J. How the T cell receptor sees antigen--A structural view. *Cell* **122**, 333-6 (2005).

45. Klein, L., Kyewski, B., Allen, P.M., & Hogquist, K.A. Positive and negative selection of the T cell repertoire: What thymocytes see (and don't see). *Nat. Rev. Immunol.* **14**, 377-91 (2014).

46. Vrisekoop, N., Monteiro, J.P., Mandl, J.N., & Germain, R.N. Thymic positive selection and the mature T cell repertoire for antigen revisited. *Immunity* **41**, 181–190 (2014).

47. Rosati, E., Dowds, C.M., Liaskou, E., Henriksen, E.K.K., Karlsen, T.H., & Franke, A. Overview of methodologies for T-cell receptor repertoire analysis. *BMC Biotechnol.* **17**, 61 (2017).

48. Gfeller, D. & Bassani-Sternberg, M. Predicting antigen presentation-what could we learn from a million peptides? Front. Immunol. 9, 1716 (2018).

49. Gerber, H., Sibener, L.V., Lee, L.J., & Gee, M.H. Identification of antigenic targets. *Trends Cancer* **6**, 299-318 (2020).

50. Ruggiero, E. *et al*. High-resolution analysis of the human T-cell receptor repertoire. *Nat. Commun.* **6**, 8081 (2015).

51. Six, A. *et al.* The past, present, and future of immune repertoire biology - the rise of next-generation repertoire analysis. *Front Immunol* **4**, 413 (2013).

52. Freeman, J.D., Warren, R.L., Webb, J.R., Nelson, B.H., & Holt, R.A. Profiling the T-cell receptor beta-chain repertoire by massively parallel sequencing. *Genome Res.* **19**, 1817–1824 (2009).

53. Han, A., Glanville, J., Hansmann, L., & Davis, M.M. Linking T-cell receptor sequence to functional phenotype at the single-cell level. *Nat. Biotechnol.* **32**, 684-92 (2014).

54. Dash, P., Wang, G.C, & Thomas, P.G. Single-Cell Analysis of T-Cell Receptor αβ Repertoire. *Methods Mol. Biol.* **1343**, 181-97 (2015).

55. Zheng, G.X.Y. *et al*. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).

56. Gierahn, T.M. *et al*. Seq-Well: Portable, low-Cost RNA sequencing of single cells at high throughput. *Nat. Methods* **14**, 395-398 (2017).

57. Lee, E.S. *et al*. Identifying T cell receptors from high-throughput sequencing: dealing with promiscuity in TCRα and TCRβ pairing. *PLoS Comput. Biol.* 13, e1005313 (2017).

58. Holec, P.V., Berleant, J., Bathe, M., & Birnbaum, M.E. A Bayesian framework for high-throughput T cell receptor pairing. Bioinformatics **35**, 1318-1325 (2019).

59. Hosomichi, K. Shiina, T., Tajima, A., & Inoue, I. The impact of next-generation sequencing technologies on HLA research. *J. Hum. Genet.* **60**, 665-73 (2015).

60. Gonzalez-Galarza FF, *et al*. Allele frequency net 2015 update: new features for HLA epitopes, KIR and disease and HLA adverse drug reaction associations. *Nucleic Acid Res.* **39**, D784-8 (2015).

61. Fortier, M. *et al*. The MHC class I peptide repertoire is molded by the transcriptome. *J. Exp. Med.* **205**, 595–610 (2008).

62. Vita, R. *et al.* The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Res.* **47**, D339-D343 (2019).

63. Purcell, A.W., Sri H Ramarathinam, S.H., & Ternette, N. Mass spectrometry-based identification of MHC-bound peptides for immunopeptidomics. *Nat. Protoc.* **14**, 1687-1707 (2019).

64. Abelin, J. G. *et al*. Mass spectrometry profiling of HLA-associated peptidomes in mono-allelic cells enables more accurate epitope prediction. *Immunity*. **46**, 315-326. (2017).

65. Sarkizova, S. *et al.* A large peptidome dataset improves HLA class I epitope prediction across most of the human population. *Nat. Biotechnol*. **38**, 199–209 (2020).

66. Abelin, J. G. *et al.* Defining HLA-II Ligand Processing and Binding Rules with Mass Spectrometry Enhances Cancer Epitope Prediction. *Immunity* **51**, 766-779 (2019).

67. Garde, C. *et al*. Improved peptide-MHC class II interaction prediction through integration of eluted ligand and peptide affinity data. *Immunogenetics* **71**, 445-454 (2019).

68. Birnbaum, M.E., Dong, S., & Garcia, K.C. Diversity-oriented approaches for interrogating T-cell receptor repertoire, ligand Recognition, and function. *Immunol. Rev.* **250,** 82-101 (2012).

69. Provenzano, M. *et al.* MHC-peptide specificity and T-cell epitope mapping: Where immunotherapy starts. *Trends Mol. Med.* **12**, 465-72 (2006).

70. Border, E.C., Sanderson, J.P., Weissensteiner, T., Gerry, A.B., & Pumphrey, N.J. Affinity-enhanced T-cell receptors for adoptive T-cell therapy targeting MAGE-A10: Strategy for selection of an optimal candidate. *Oncoimmunology* **8**, e1532759 (2018).

71. Dash, P. *et al.* Quantifiable predictive features define epitope specific T cell receptor repertoires. *Nature* **547**, 89-93 (2017).

72. Glanville, J. *et al*. Identifying specificity groups in the T cell receptor repertoire. *Nature* **547**, 94-98 (2017).

73. Pedersen, L.O. *et al*. Efficient assembly of recombinant major histocompatibility complex class I molecules with preformed disulfide bonds. *Eur. J. Immunol*. **31**, 2986-96 (2001).

74. Pepper, L.R., Cho, Y.K., Boder, E.T., & Shusta, E.V. A decade of yeast surface display technology: Where are we now? *Comb. Chem. High Throughput Screen.* **11**, 127–134 (2008).

75. Birnbaum, M.E. *et al.* Deconstructing the peptide-MHC specificity of T cell recognition. *Cell* **157**, 1073-87 (2014).

76. Gee, M. H. *et al.* Antigen identification for orphan T cell receptors expressed on tumor-infiltrating lymphocytes. *Cell* **172**, 549-563 (2018).

77. Sibener, L.V. *et al.* Isolation of a structural mechanism for uncoupling T cell receptor signaling from peptide-MHC binding. *Cell* **174**, 672-687 (2018).

78. Saligrama, N. *et al.* Opposing T cell responses in experimental autoimmune encephalomyelitis. *Nature* **572**, 481-487 (2019).

79. Savage, P.A., Leventhal, D.S. & Malchow, S. Shaping the repertoire of tumor-infiltrating effector and regulatory T cells. Immunol. Rev. 259, 245-258 (2015).

80. Borst, J., Ahrends, T., Bąbała, N., Melief, C.J.M., & Kastenmüller, W. CD4 + T cell help in cancer immunology and immunotherapy. *Nat. Rev. Immunol.* **18**, 635-647 (2018).

81. Togashi, Y., Shitara, K., & Nishikawa, H. Regulatory T cells in cancer immunosuppression — implications for anticancer therapy. *Nat. Rev. Clin. Oncol*. **16**, 356-371 (2019).

82. Hu, Z., Ott, P.A., & Wu, C.J. Towards personalized, tumour-specific, therapeutic vaccines for cancer. *Nat. Rev. Immunol.***18**, 168-182 (2018).

83. Jensen, K. K. *et al*. (2018) Improved methods for predicting peptide binding affinity to MHC class II molecules. *Immunology*. **154**, 394-406 (2018).

CHAPTER 2: Empirical determination of class II MHC peptide repertoires for improved antigen prediction.

Abstract

$CD4^+$ 'helper' T cells play a central role in the immune system, coordinating the immune response in pathogen infection and cancer, but also in autoimmune diseases and allergies. Accordingly, accurate prediction of which antigens can be displayed on class II MHCs for recognition by $CD4^+$ T cells is an important consideration for the study of immune disorders, and for the design of novel antigen-targeted vaccines and cancer immunotherapies. While many algorithms have been developed to predict peptide binding to class II MHCs, they under-perform their class I counterparts – even for well-characterized alleles – due in part to gaps and inaccuracies within their underlying training sets, and deficiencies arising from limited peptide diversity.

In this chapter, we describe a yeast-display-based platform to screen libraries of 10^8 peptides for binding to a co-expressed class II MHC, identifying over an order of magnitude more unique binders than comparable approaches. The enriched peptide data contains strong motifs that reflect previous reports, but also highlight gaps and inaccuracies in current data collection techniques and frequently used prediction algorithms, which are validated by *in vitro* binding assays. We further validated that these gaps and inaccuracies are rectified when existing prediction algorithms are trained upon our yeast-display library data, providing improved prediction of peptide-binding affinity and improved antigen prediction for pathogen and tumor-associated peptides. Together, these findings demonstrate that this platform yields large, high-quality peptide-binding datasets that can be used to improve the accuracy of class II MHC prediction algorithms for improved understanding and application of $CD4^+$ T cell recognition.

2.1 Introduction

T cells recognize short, linear peptides displayed by Major Histocompatibility Complexes (MHCs), or Human Leukocyte Antigens (HLAs) in humans, through their T cell receptors (TCRs). Upon recognition of a cognate peptide-MHC (pMHC) complex, the T cell is activated, initiating an immune response that can protect against infectious diseases and cancer [1, 2], but that can also potentiate autoimmunity, allergy, and transplant rejection [3-5].

As T cells occupy a central role in many diseases and underlie the successes of novel antigen-targeted vaccinations and immunotherapies [6-8], there is considerable interest in determining which peptides can be presented by MHCs for T cell surveillance. However, the highly polymorphic peptide-binding groove of MHCs and the immense diversity of potential binding peptides necessitates the use of allele-specific antigen prediction algorithms. Recent advances have described improvements of these computational algorithms [9-11], their underlying training data [12, 13], or both [15-18]. But while these advances have benefited antigen prediction for both classes of MHC proteins – class I and II, which are canonically recognized by 'killer' CD8$^+$ and helper CD4$^+$ T cells, respectively – there is sustained interest in improving the performance of class II MHC prediction algorithms [19], which frequently under-perform their class I counterparts [11, 20-24].

Although this under-performance is at least partially due to a paucity of curated peptide-binding data for class II MHC alleles [25] – as under-performance is particularly pronounced for alleles with few reported binders [20, 21] – these predictions under-perform for even well-characterized alleles [20, 24]. This is likely due to challenges inherent to class II MHCs, which have more degenerate peptide-binding motifs than class I MHCs [26], and have an open peptide-binding groove that requires an added algorithmic step of peptide-register determination [21, 27-29]. Additionally, publically available class II MHC peptide-binding datasets contain many redundant nested peptide sets and single amino-acid variants of well-characterized peptides, limiting their effective depth and generalizability [25, 30]. Therefore, we hypothesize that the under-performance of class II MHC prediction algorithms is driven primarily by deficiencies in their underlying training data that can be rectified with large and diverse high-quality peptide datasets.

In this chapter, we describe a yeast-display-based platform to screen 10^8 peptides for their ability to bind a co-expressed class II MHC to highlight and rectify these deficiencies. This platform generates over an order of magnitude more unique peptide data than comparable approaches for two human class II alleles, and enriches peptide motifs that mirror previous reports. However, our datasets contain additional peptide-binding information that results in consequential differences from existing prediction algorithms and other state-of-the-art data collection techniques. We demonstrate that these differences represent systemic gaps and inaccuracies in current class II MHC peptide-binding data that are rectified by training existing algorithms on yeast-display data. Finally, we show that an algorithm trained on our data improves prediction of peptide-binding affinity and improves antigen prediction for pathogen- and tumor-associated peptides. These data show the importance of large, unbiased pMHC repertoires to improve existing antigen prediction training datasets, and suggest our approach can facilitate improved understanding of CD4$^+$ T cell recognition and improved patient benefit from antigen-targeted therapeutics.

2.2 Results

2.2.1 Yeast-displayed class II MHC platform identifies peptide binding

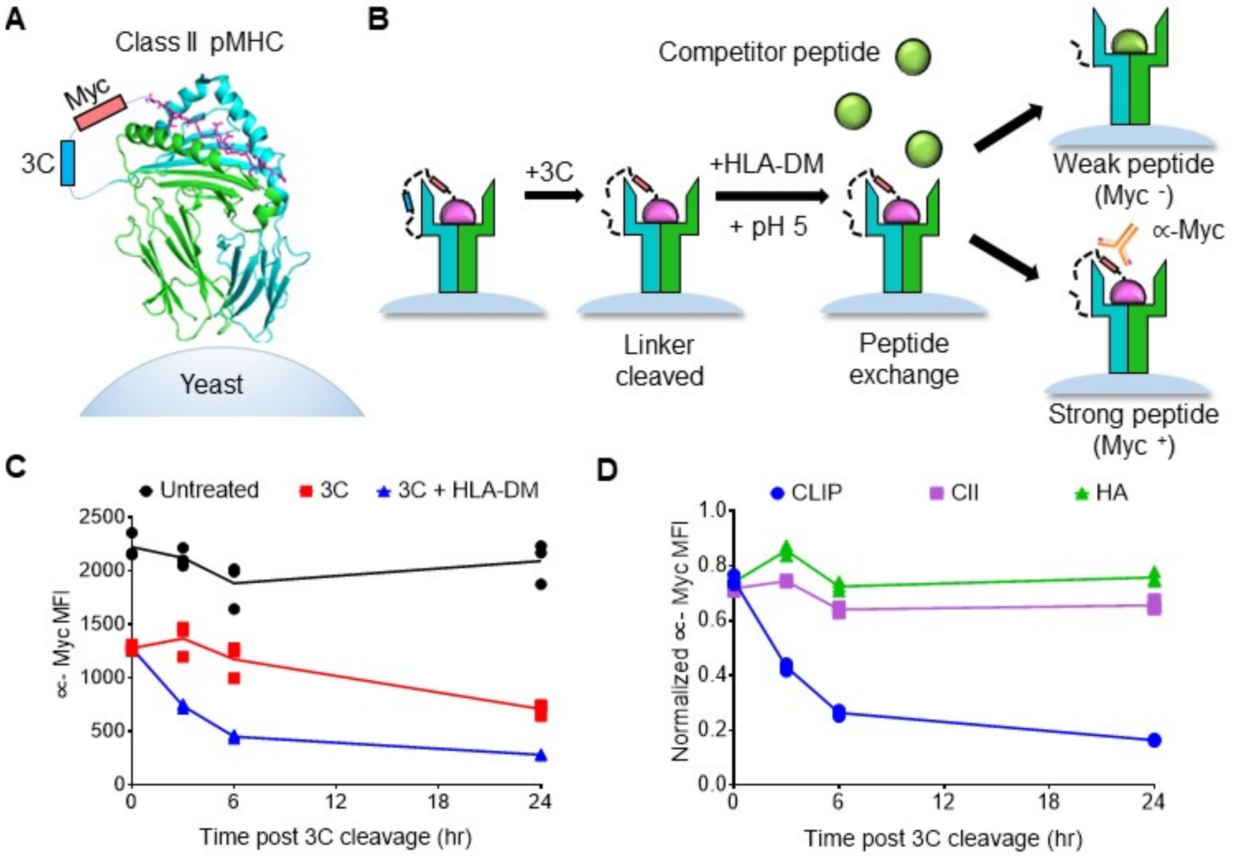


Figure 2.1. *Design and validation of a yeast-display platform to identify peptide binding to a co-expressed class II MHC. A) Structural representation (adapted from PDB 1J8H) of platform highlighting 3C protease cleavage site and Myc epitope tag within the linker connecting the peptide and MHC β1 domain. B) Schematic of validation protocol, including linker cleavage with 3C, peptide exchange at low pH in the presence of HLA-DM and high-affinity competitor peptide, and quantification of remaining bound peptide with an anti-Myc antibody. C) Time course of mean fluorescence intensity (MFI) of a fluorescently labeled anti-Myc antibody for HLA-DR401-CLIP$_{81\text{-}101}$-encoding yeast without treatment (Untreated), with linker cleavage (3C), or with linker cleavage and peptide exchange (3C + HLA-DM), as determined by flow cytometry. D) Comparison of peptide retention for HLA-DR401-CLIP$_{81\text{-}101}$, -CII$_{261\text{-}273}$, or -HA$_{306\text{-}318}$-encoding yeast with linker cleavage and peptide exchange, as determined by flow cytometry and normalized to MFI before treatment.*

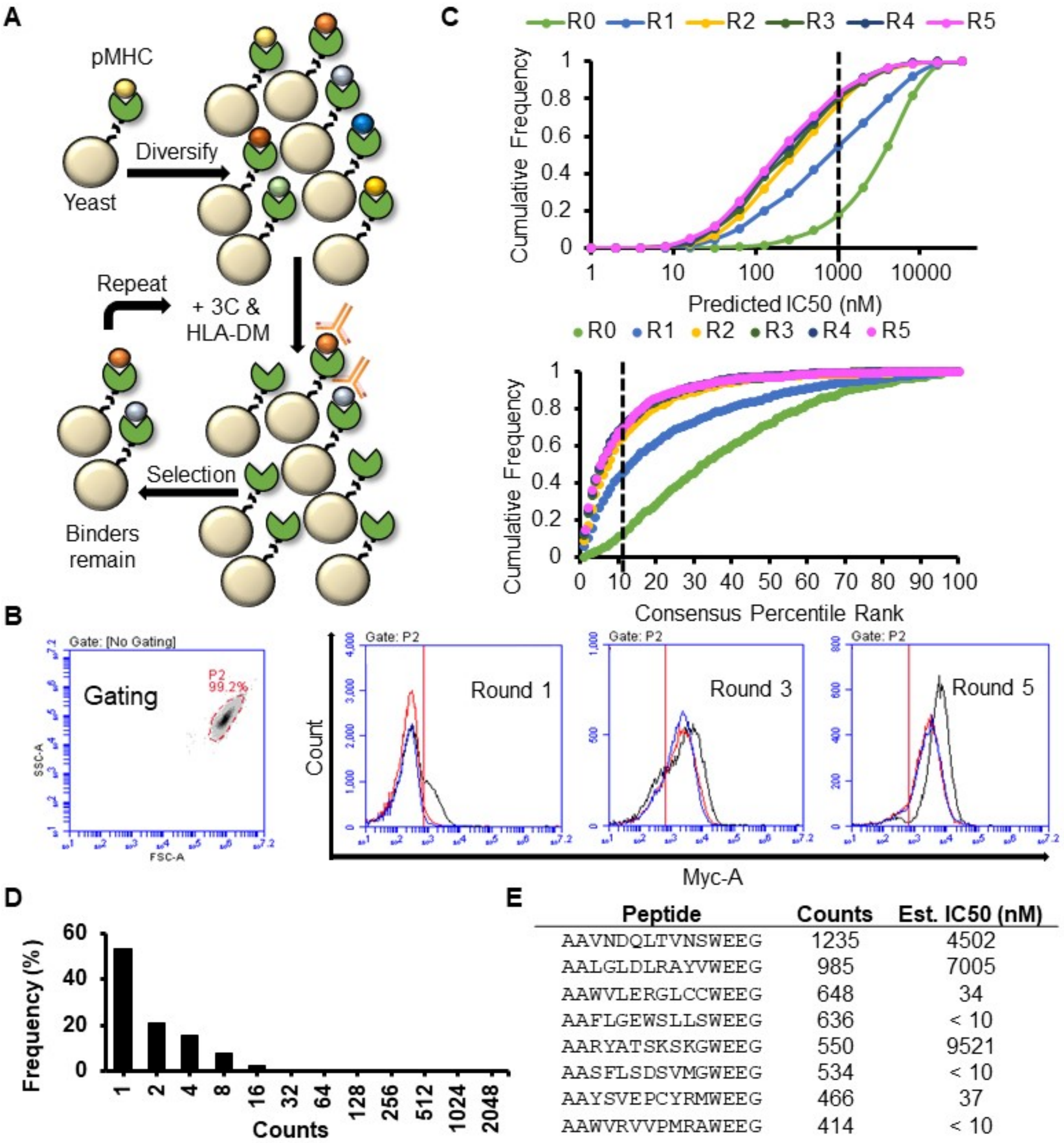
Yeast-displayed class II pMHC platforms have previously been used to identify peptides that facilitate TCR binding [31, 32]. To enable this platform to determine peptide binding to the MHC, we modified the previously-described design of HLA-DR401 (HLA DRA1*01:01, HLA-DRB1*04:01) [32] to express a 3C protease site and an antibody-trackable Myc epitope tag within the flexible linker connecting the peptide to the N-terminus of the HLA β chain (Figure 2.1A). Protease treatment cleaves the linker, allowing unbound peptides to freely disassociate. Peptide exchange is then initiated at low pH in the presence of a high-affinity competitor peptide and the catalyst HLA-DM (Figure 2.1B), emulating the native endosomal environment of peptide loading and exchange [33]. If the original peptide is displaced by the competitor peptide, the peptide-proximal epitope tag is lost, enabling us to differentiate yeast encoding binding and non-binding peptides by flow cytometry with a fluorescently-labeled antibody directed against the peptide-proximal Myc tag.

This platform was validated through its specificity in peptide retention. Yeast expressing HLA-DR401 and the class II-associated invariable chain peptide (CLIP$_{81-101}$), the peptide displaced during endogenous antigen presentation [33], exhibited significant loss of peptide-proximal epitope tag signal following linker cleavage that increased with incubation at low pH with a competitor peptide (Figure 2.1C). Consistent with its role as a peptide-exchange catalyst, the addition of HLA-DM significantly accelerated signal loss. However, yeast expressing known binders of HLA-DR401, HA$_{306-318}$ [34, 35] and CII$_{261-273}$ [35-37], exhibited retention of their peptides when treated with 3C and HLA-DM (Figure 2.1D).

2.2.2 Selection, analysis, and validation of a class II MHC peptide library

To enable large-scale identification of HLA-DR401-binding peptides, we generated a yeast library encoding 1x10^8 random MHC-linked peptides. To simplify downstream analysis, peptides were designed as a randomized 9mer flanked by constant residues that favor binding to the MHC in a single register, as the class II MHC peptide-binding groove is open at either end and allows binding in many possible registers [22,23]. The library was subjected to iterative rounds of linker cleavage, peptide exchange, and selection for epitope tag retention (Figure 2.2A), resulting in a pool of strong binders after five rounds (Figure 2.2B). Upon deep sequencing, we observed enrichment of predicted binders (Figure 2.2C). The enriched library was highly diverse, consisting of 81,422 unique peptides in the correct register, of 85,756 total peptides. To visualize positional amino acid enrichment and depletion, positional residue frequencies were benchmarked against the unselected library to generate positional log2-fold-change enrichment values. The resulting data are presented as heatmaps representing unweighted averages, as the distribution of peptide frequency in the enriched library was largely flat, with no observed correlation between individual peptide frequency and affinity (Figures 2.2D, E).

We observed the strongest enrichments at peptide positions P1, P4, P6, and P9 (Figure 2.3A), which are considered ‘anchor’ positions where the peptide backbone orients the amino acid side chain directly into pockets of the MHC surface (Figure 2.3B) [28]. These enrichments largely match previous reports for HLA-DR401 [37, 39-43]: the deep P1 pocket favors large hydrophobic residues; the basic P4 pocket favors acidic residues; P6 favors polar residues Ser, Thr, and Asn; and the shallow P9 pocket favors Ala, Gly, and Ser. However, the observed enrichment of P9 Cys and P6 Asp do not match this consensus, and only the latter has been previously reported [43, 44]. We also observed a less stringent preference for Pro and Asn at P7, which is considered to be an auxiliary anchor position [45]. While the remaining positions are considered to be determinants of TCR binding [46], each displayed marked preferences, such as the uniform depletion of Trp, the enrichment of Pro and Asp at P5, the strong depletion of P2 Pro, and the previously described preference for P2 Arg [37, 40]. Each described enrichment or depletion was highly statistically significant ($p < 0.001$).



Peptide	Counts	Est. IC50 (nM)
AAVNDQLTVNSWEEG	1235	4502
AALGLDLRAYVWEEG	985	7005
AAWVLERGLCCWEEG	648	34
AAFLGEWSLLSWEEG	636	< 10
AARYATSKSKGWEEG	550	9521
AASFLSDSVMGWEEG	534	< 10
AAYSVEPCYRMWEEG	466	37
AAWVRVVPMRAWEEG	414	< 10

Figure 2.2. *Selection of a yeast-displayed HLA-DR401 randomized peptide library displays rapid convergence. A) Schematic of sequential rounds of library selection to eliminate non-binding peptides and enrich binders. B) Histogram of the fluorescence intensity of a labeled anti-Myc antibody for 10,000 yeast in each round of selection either before linker cleavage (Black), following cleavage (Red), or after 24h peptide exchange (Blue), with gating strategy. C) Cumulative distribution function of predicted peptide IC$_{50}$ (top) and percentile rank (bottom) of 1000 peptides from each round of selection, as determined by NetMHCII and the IEDB consensus tool, respectively. Dashed lines represent previously established cut-offs for peptide binding. D) Histogram of occurrences of each unique peptide found in round 5 of selection. E) Table of the most enriched peptides found within round 5 of selection with occurrences and estimated IC$_{50}$ value from two-point fluorescence polarization competition assays.*

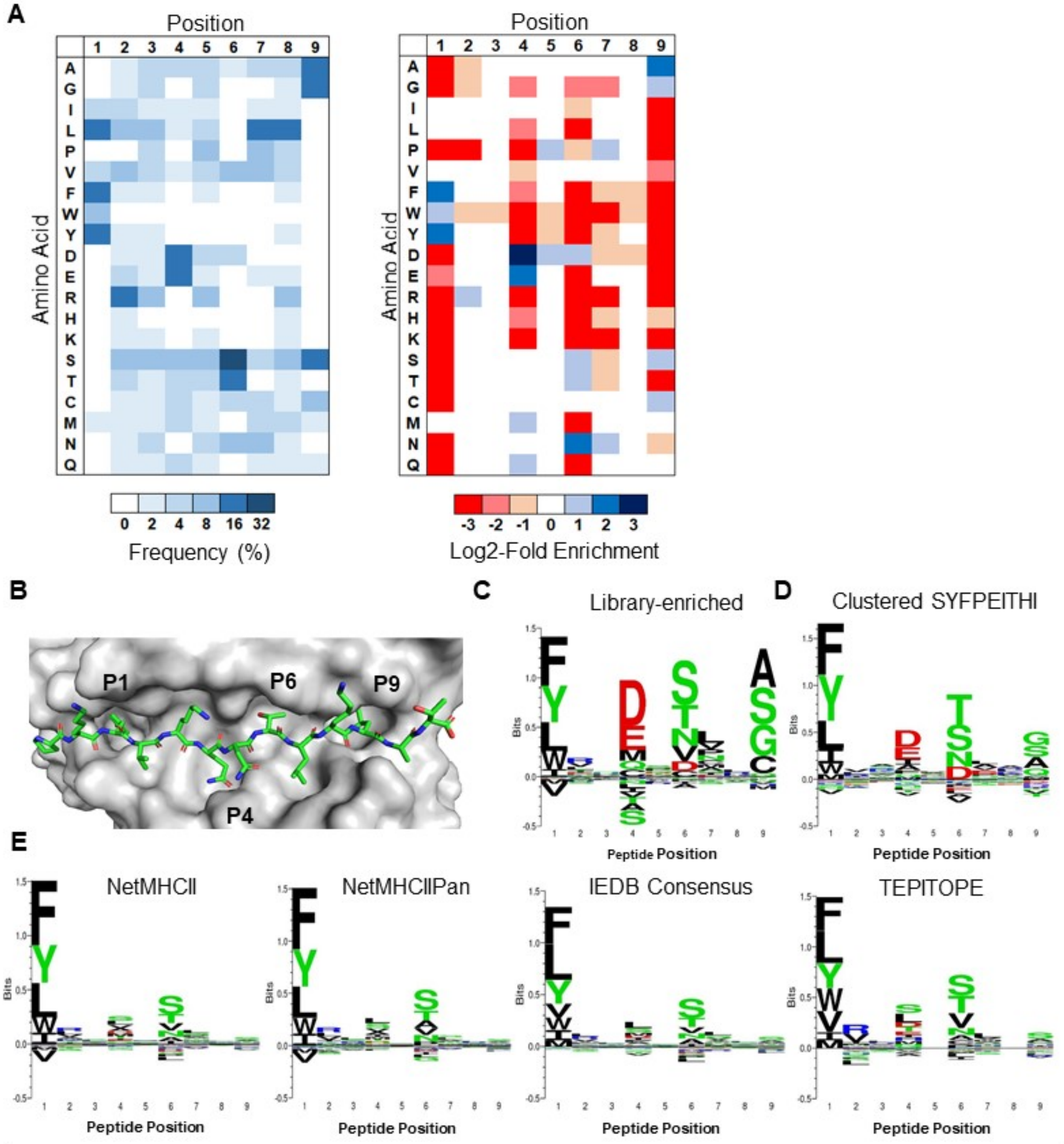


Figure 2.3. *Selection of a yeast-displayed HLA-DR401 randomized peptide library reveals a strongly enriched binding motif that differs from existing prediction algorithms. A) Unweighted heat maps of positional percent frequency and log2-fold enrichment of each amino acid in round 5 of selection (N = 81,422 unique peptides). B) Structure of HA$_{306\text{-}318}$ peptide in the HLA-DR401 peptide-binding groove (PDB 1J8H), with primary peptide 'anchor' positions denoted in bold. (C-E) Kullback-Leibler relative entropy motifs of the core nine amino acids of HLA-DR401-binding peptides, as determined (C) empirically from our yeast-display library, (D) by clustering of binders curated on the SYFPEITHI database, or (E) by application of existing class II MHC prediction algorithms to computationally-generated peptides.*

Notably, our overall library-enriched motif (Figure 2.3C) closely resembles that of known HLA-DR401 binders (Figure 2.2D), generated by clustering previously reported HLA-DR401-binding peptides curated on the SYFPEITHI database [30]. In particular, we found stringent preferences at each anchor position that matched our own, with the exception of P9 Cys. However, these motifs were highly dissimilar to those generated by existing class II MHC prediction algorithms NetMHCII [11], NetMHCIIPan [11], TEPITOPE [48], or IEDB consensus [49] (Figure 2.2E). Comparison suggests that while these algorithms mirror the importance and nature of preferences at P1 and P6 (excepting P6 Asp), they have increased uncertainty or miscall preferences at the remaining anchor positions, P4 and P9.

In addition, we performed library selection without the addition of the endosomal class II peptide-exchange catalyst HLA-DM in order to quantify its impact on the peptide repertoire. With the exception of differences in their magnitudes, the observed enrichments and depletions were consistent with HLA-DM addition (Figure 2.4), suggesting that HLA-DM selects for the retention of high-affinity peptides uniformly across each position, but does not impart unique positional preferences, consistent with previous reports [18, 47].

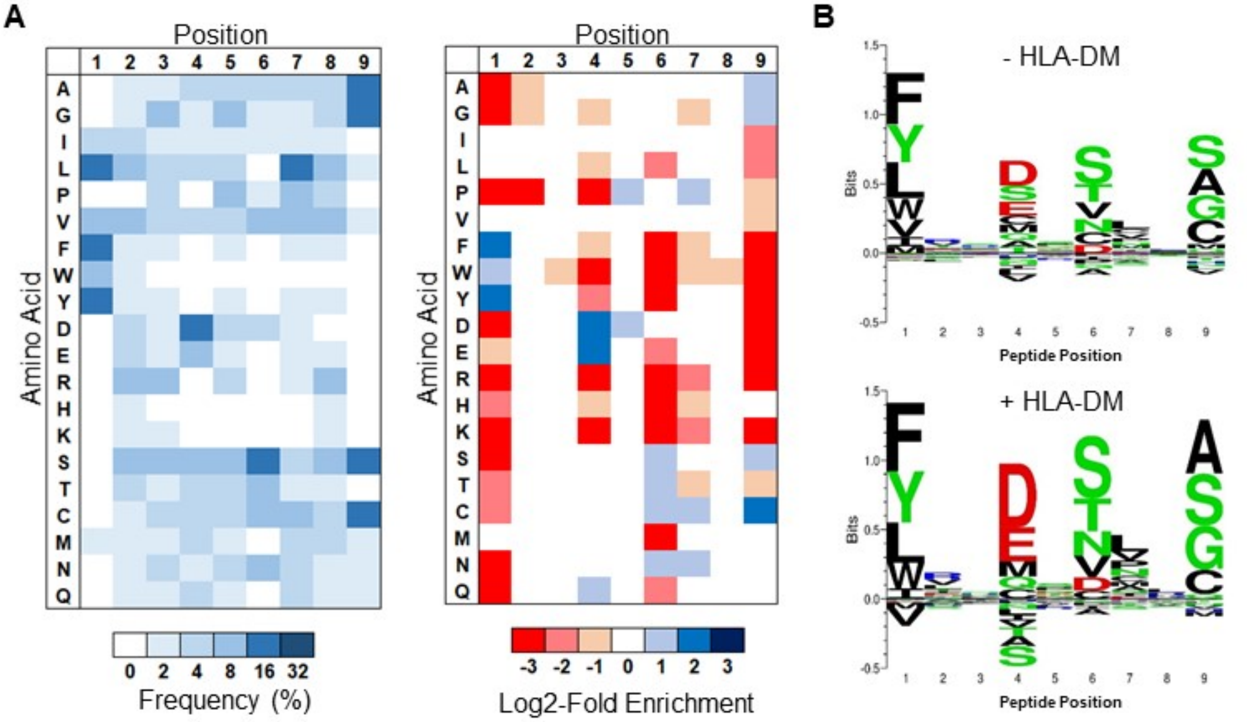
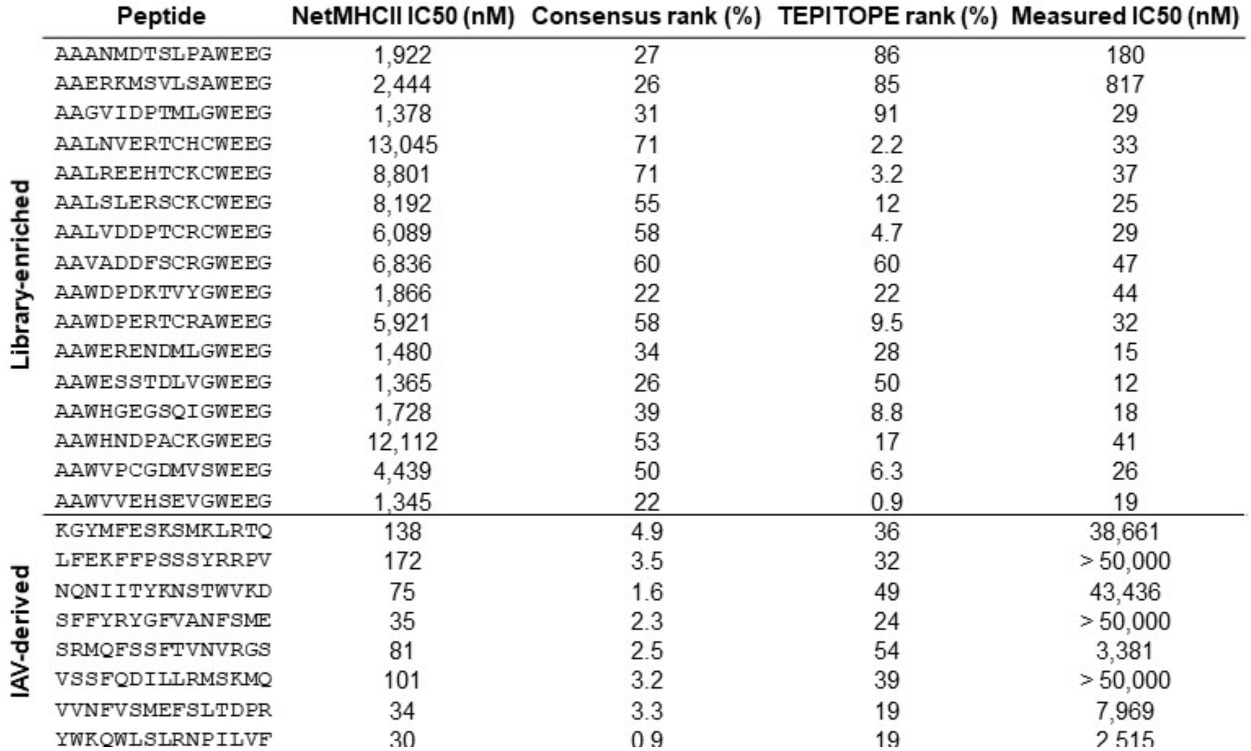


Figure 2.4. *HLA-DM addition increases the stringency of library selection. A) Unweighted heat maps of the positional percent frequency and log2-fold enrichment of each amino acid in round 5 of selection without the addition of HLA-DM (N = 105,717 unique peptides). B) Kullback-Leibler relative entropy motifs of the core nine amino acids of HLA-DR401-binding peptides, determined empirically from round 5 of library selection with or without HLA-DM addition.*

To quantify the consequence of these differences, we performed fluorescence polarization competition assays on selected peptides to determine their IC_{50} values for recombinant HLA-DR401, which correlate with affinity [50]. We selected 16 peptides that were enriched by our library but deemed non-binders by both NetMHCII and the IEDB consensus tool (predicted $IC_{50} > 1$ μM, consensus rank > 10 [45, 49]); each contained either cysteine at P4, P7, or P9, aspartic acid at P6, or an unfavorable P1 residue (Table 2.1). Upon measurement, each of these peptides had an IC_{50} less than 1 μM, and 14/16 bound stronger than $HA_{306-318}$ (76 nM), a known strong binder [34, 35] (Figure 2.5A). Importantly, the binding of the cysteine-containing peptides was specific, as two allele-mismatched cysteine-containing peptides did not exhibit binding (Figure 2.5B). We further identified 8 peptides from Influenza A virus [A/Victoria/3/75 (H3N2)] that both NetMHCII and IEDB consensus predicted as binders (IC_{50} < 200 nM, consensus rank < 5) but that did not match our enriched motif, largely due to departures at P4 and P9 (Table 2.1). Each had a measured IC_{50} > 2 μM, and 6/8 bound weaker than $CLIP_{89-101}$. Overall, there was minimal concordance between the measured IC_{50} of these peptides and the predictions of NetMHCII, IEDB consensus, or TEPITOPE, and the predictions of both NetMHCII and IEDB consensus were negatively correlated with measured IC_{50} (Figure 2.5C).

Table 2.1. *Peptides either enriched by our randomized 9mer HLA-DR401 library selections but not predicted to bind HLA-DR401 by NetMHCII or IEDB Consensus (Library-enriched), or derived from Influenza A virus and predicted to bind HLA-DR401 but not matching our enriched motif (IAV-derived), with prediction values and IC_{50} measured via from fluorescence polarization competition assays.*



	Peptide	NetMHCII IC50 (nM)	Consensus rank (%)	TEPITOPE rank (%)	Measured IC50 (nM)
Library-enriched	AAANMDTSLPAWEEG	1,922	27	86	180
	AAERKMSVLSAWEEG	2,444	26	85	817
	AAGVIDPTMLGWEEG	1,378	31	91	29
	AALNVERTCHCWEEG	13,045	71	2.2	33
	AALREEHTCKCWEEG	8,801	71	3.2	37
	AALSLERSCKCWEEG	8,192	55	12	25
	AALVDDPTCRCWEEG	6,089	58	4.7	29
	AAVADDFSCRGWEEG	6,836	60	60	47
	AAWDPDKTVYGWEEG	1,866	22	22	44
	AAWDPERTCRAWEEG	5,921	58	9.5	32
	AAWERENDMLGWEEG	1,480	34	28	15
	AAWESSTDLVGWEEG	1,365	26	50	12
	AAWHGEGSQIGWEEG	1,728	39	8.8	18
	AAWHNDPACKGWEEG	12,112	53	17	41
	AAWVPCGDMVSWEEG	4,439	50	6.3	26
	AAWVVEHSEVGWEEG	1,345	22	0.9	19
IAV-derived	KGYMFESKSMKLRTQ	138	4.9	36	38,661
	LFEKFFPSSSYRRPV	172	3.5	32	> 50,000
	NQNIITYKNSTWVKD	75	1.6	49	43,436
	SFFYRYGFVANFSME	35	2.3	24	> 50,000
	SRMQFSSFTVNVRGS	81	2.5	54	3,381
	VSSFQDILLRMSKMQ	101	3.2	39	> 50,000
	VVNFVSMEFSLTDPR	34	3.3	19	7,969
	YWKQWLSLRNPILVF	30	0.9	19	2,515

Together, these data highlight consequential gaps and inaccuracies in current class II prediction algorithms, and suggest that our yeast-display platform enriches high-quality peptides that may rectify these deficiencies.

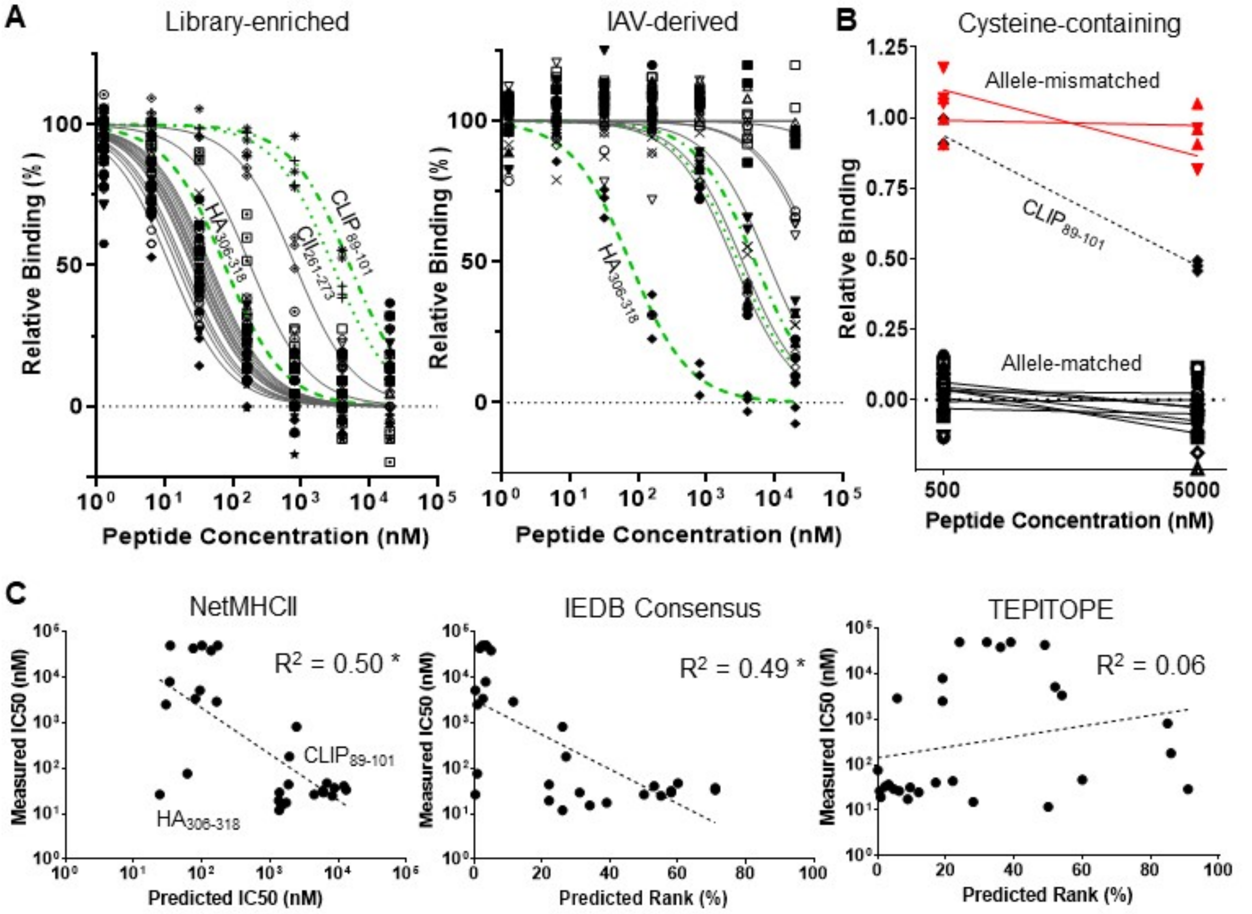


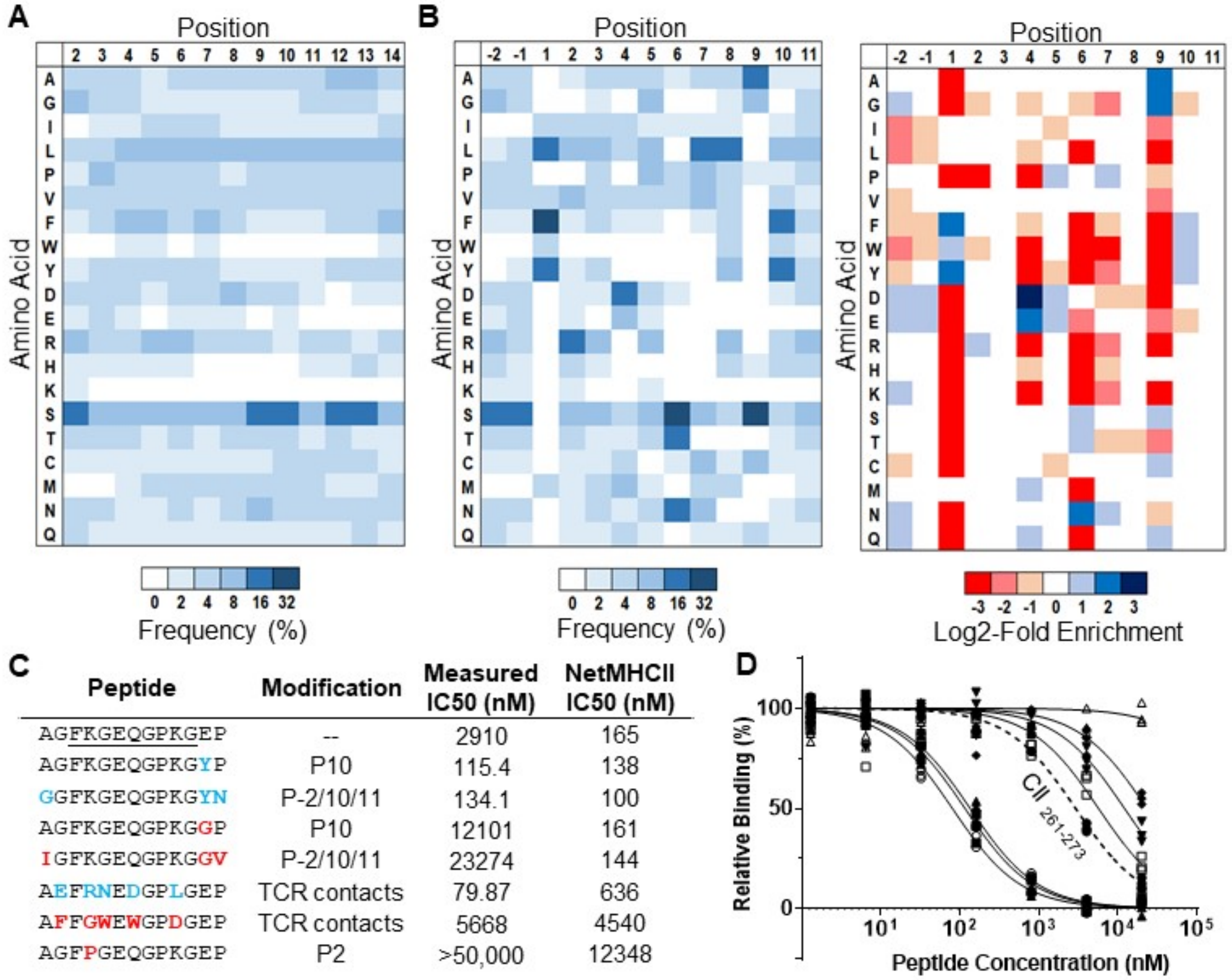
Figure 2.5. *Validation of library-enriched HLA-DR401-binding motif reveals consequential gaps and inaccuracies in class II prediction algorithms. A) Relative binding curves for HLA-DR401 in fluorescence polarization competition assays for peptides either enriched by selection of a 9mer HLA-DR401 library but not predicted to bind HLA-DR401 (Library-enriched) or derived from Influenza A virus and predicted to bind HLA-DR401 but not matching our enriched motif (IAV-derived), with selected control peptides (green). B) Relative binding of cysteine-containing peptides found in round 5 of selection of either the randomized 9mer HLA-DR401 (allele-matched) or HLA-DR402 (allele-mismatched) libraries, tested at two concentrations with HLA-DR401 in a fluorescence polarization competition assay. Curves are fit to N = 3 replicates. C) Scatterplots of algorithmic predictions versus measured IC_{50} with lines of best fit and their associated coefficients of determination (R^2). Asterisk denotes R^2 values of negative correlations.*

2.2.3 Preferences outside the peptide 'core' greatly affect binding

Canonically, peptide positions P1 through P9 are considered to form the 'core' of the interface with the class II MHC peptide-binding groove [27, 28]. However, positions outside of the MHC groove, also known as peptide flanking residues (PFRs), can reportedly affect peptide binding [51-53]. Most notably, modifications at position P10 can reportedly alter peptide IC_{50} up to two orders of magnitude [52], without altering the peptide 'core' or TCR interactions [46]. We therefore sought to investigate peptide preferences outside the groove using a randomized peptide library.

To this end, we constructed a randomized 13mer HLA-DR401 library, which was selected analogously to the original library. While peptides from round 5 showed no initially obvious motif

(Figure 2.6A), register deconvolution by Gibbs Cluster [54] identified 7 distinct registers among the 15,147 unique peptides, of which 3,374 were found to occupy the central register where positions P-2 through P11 are diversified (Figure 2.6B). Analysis at P10 showed a preference for aromatic residues, consistent with previous findings [52], and depletion of both Gly and Glu. We also observed depletion of hydrophobic residues and enrichment of acidic residues at positions P-2 and P-1. Positional preferences between positions P1 and P9 were consistent with the original library, suggesting our motif was not influenced by the fixed peptide flanking residues in our original design.



Peptide	Modification	Measured IC50 (nM)	NetMHCII IC50 (nM)
AGFKGEQGPKGEP	--	2910	165
AGFKGEQGPKG**Y**P	P10	115.4	138
GGFKGEQGPKG**YN**	P-2/10/11	134.1	100
AGFKGEQGPKG**G**P	P10	12101	161
IGFKGEQGPKG**GV**	P-2/10/11	23274	144
A**E**F**RN**E**D**GP**L**GEP	TCR contacts	79.87	636
A**F**F**GW**E**W**GP**D**GEP	TCR contacts	5668	4540
AGF**P**GEQGPKGEP	P2	>50,000	12348

Figure 2.6. *Preferences at TCR contacts and positions outside the peptide core affect peptide binding. (A-B) Unweighted heat maps of log2-fold enrichment and/or positional percent frequency of each amino acid for either (A) all peptides in round five of selection of a randomized 13mer HLA-DR401 library (N = 15,147 unique peptides) or (B) only those determined to bind in the third peptide register (N = 3,374 unique peptides). C) Table of modified CII$_{261\text{-}273}$ peptides with associated IC$_{50}$ values and the predictions of NetMHCII. Peptide positions 1-9 of are underlined, and detrimental (red) or beneficial (blue) modifications are denoted in bold. D) Relative binding curves for HLA-DR401 fluorescence polarization competition assays of CII$_{261\text{-}273}$ variants. Wild-type peptide is shown by dashed line and curves are fit to N=3 replicates.*

To validate these observations, we performed competition assays with variants of CII$_{261\text{-}273}$. Notably, modifying P10 to its most enriched residue, tyrosine, resulted in a 30-fold decrease in measured IC_{50}, transforming CII into a strong binder (Figure 2.6C, D) Furthermore, modification to its most depleted residue, glycine, resulted in a 4-fold increase in IC_{50}. Interestingly, added modification of P-2 and P11, which sit outside the groove but are not considered TCR contacts [46], did not further benefit peptide binding for favorable residues, but furthered loss of binding for unfavorable residues. We observed comparable effects from modifying each TCR contact (P-1, P2, P3, P5, and P8) to favorable or unfavorable residues, and the singular modification of P2 Pro resulted in the loss of any detectable binding, consistent with its strong depletion. Although NetMHCII reportedly considers PFRs [11, 29], we did not observe substantial changes in predicted IC_{50} when positions -2, 10, or 11 were modified (Figure 2.6C).

These data demonstrate that peptide binding is greatly affected by positions outside the MHC groove, especially at P10, highlighting an additional deficiency in existing class II prediction algorithms that is highlighted by our yeast-displayed libraries.

2.2.4 Application to a poorly characterized HLA-DR allele

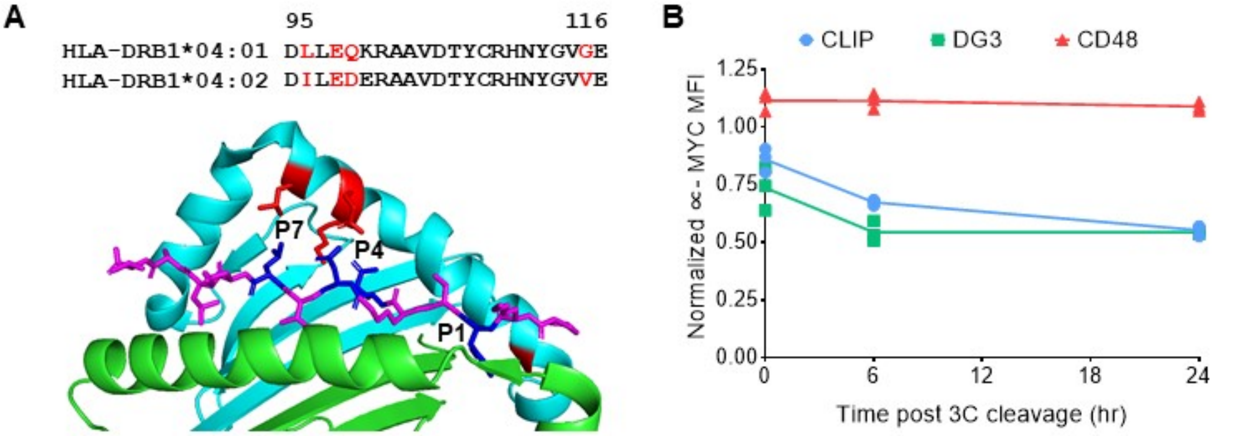


Figure 2.7. *Design and validation of a yeast-displayed HLA-DR402 construct. A) Structure of HLA-DR401 complexed with $HA_{306-308}$ (PDB 1J8H) highlighting HLA-DR402 polymorphisms (red) and polymorphism-proximal peptide positions (blue), with associated sequence alignment. B) Comparison of peptide retention for HLA-DR402-$CLIP_{81-101}$, -$DG3_{190-204}$, or -$CD48_{36-43}$-encoding yeast, with linker cleavage and peptide exchange, as determined by flow cytometry.*

Among human class II MHC alleles, HLA-DR401 is well-studied, with over 5,000 peptides curated in the Immune Epitope Database (IEDB) [25]. However, many alleles have few, or no, reported binders. We therefore sought to apply our platform to one such allele, where the need for high-quality peptide data is greatest. We chose HLA-DR402 (HLA-DRA1*01:01, HLA-DRB1*04:02), which differs from HLA-DR401 by only four amino acids (Figure 2.7A) yet has only 256 peptides curated on the IEDB, many of which are non-unique nested sets and single amino-acid variants of a parental sequence [25, 30].

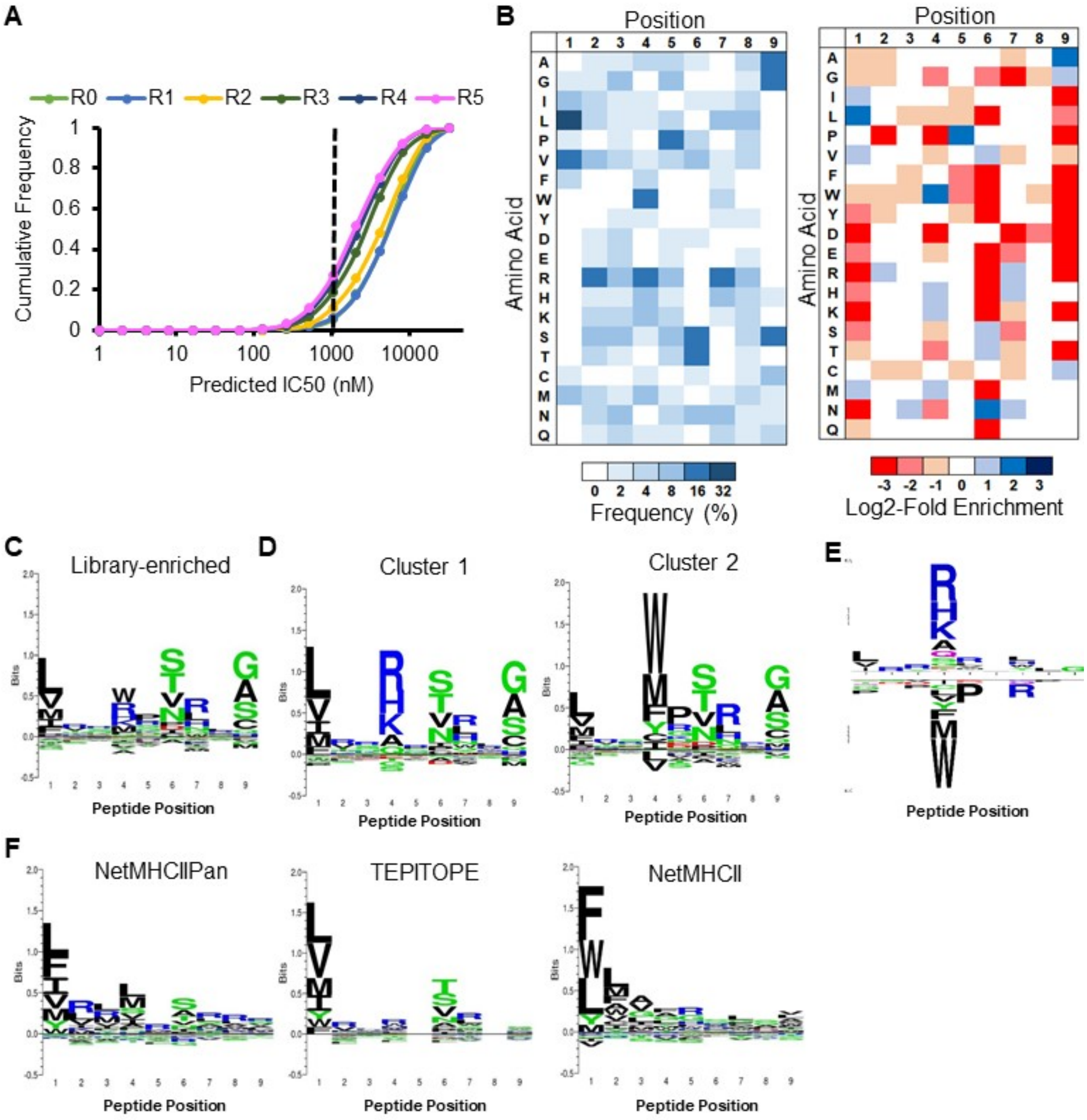
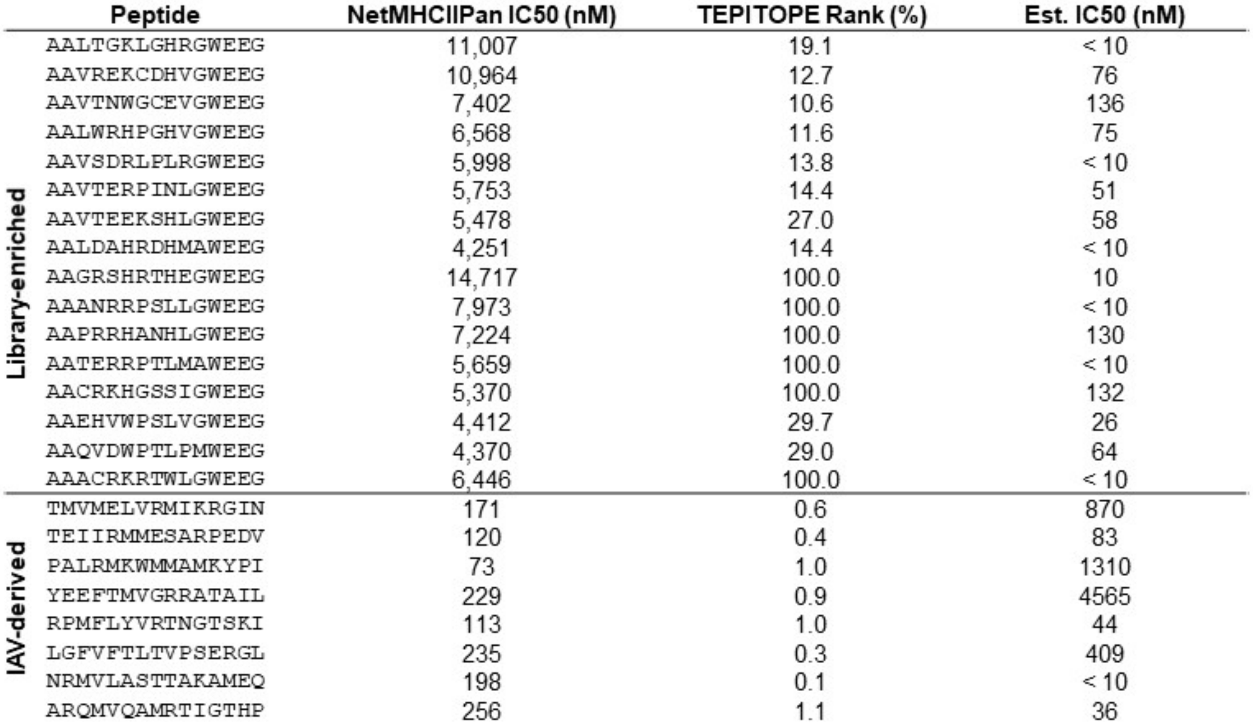


Figure 2.8. *HLA-DR402 library selection enriches two distinct peptide-binding motifs that differ from prediction algorithms. A) Cumulative distribution function of the predicted IC$_{50}$ of 1000 peptides from each round of selection of a randomized 9mer HLA-DR402 library, as determined by NetMHCIIPan, with a previously reported cut-off for peptide binding (dashed line). B) Unweighted heat maps of positional percent frequency and log2-fold enrichment of each amino acid in round 5 of selection (N = 7,692 peptides). C-D) Kullback-Leibler relative entropy motifs of the core nine amino acids of HLA-DR402-binding peptides, determined empirically from (C) all peptides in round 5 of selection or (D) in each distinct cluster. E) Amino acids at each position within the core of HLA-DR402-binding peptides significantly ($p < 0.05$) differentially distributed between clusters. Residue size correlates with statistical significance. F) Kullback-Leibler relative entropy motifs of the core nine amino acids of HLA-DR402-binding peptides, determined by applying existing class II MHC prediction algorithms to computationally-generated peptides.*

Yeast-displayed HLA-DR402 was validated through its ability to specifically retain previously reported peptide binders [43, 55-59] (Figure 2.7B). A randomized 9mer HLA-DR402 library was constructed, selected, and analyzed as above. While the predicted affinity of observed sequences increased throughout selection, the final proportion of predicted binders was low (27%), suggesting a large divergence between our enriched library and prediction algorithms (Figure 2.8A). Sequences from round 5 of selection again revealed a strongly enriched motif (Figure 2.8B), with 7,692 unique peptides within the correct register, of 10,189 total peptides.

Consistent with the location and nature of its polymorphisms, residue preferences at positions P2, P3, P6, P8, and P9 mirror those of HLA-DR401, yet differ notably at positions P1, P4, P5, and P7 (Figure 2.8C). Specifically, the truncated P1 pocket favors small hydrophobic residues; P4 favors basic residues and large hydrophobic residues Trp and Met; P5 increasingly favors Pro as well as basic residues; and P7 favors basic residues, consistent with the consensus of previous reports [37, 43, 44, 56, 60-62]. However, the decreased information content at positions P1 and P4 was notable. Further analysis revealed that the enriched sequences represented two unique motifs (Figure 2.8D): The first, a conventional HLA-DR motif with strong preferences at anchor positions P1, P4, P6, and P9; the second, an unconventional motif dominated by hydrophobic residues, especially Trp, at P4, and significantly ($p < 0.05$) less dependent on hydrophobic residues at P1, but more dependent on P5 Pro (Figure 2.8E).

Table 2.2 *Peptides either found enriched by in round 5 of selection of a randomized 9mer HLA-DR402 library but not predicted to bind HLA-DR402 (Library-enriched), or derived from Influenza A virus and predicted to bind HLA-DR402 but not matching our enriched motif (IAV-derived), with associated predictions of NetMHCIIPan and TEPITOPE, and estimated IC*$_{50}$ *values from by two-point fluorescence polarization competition assays*



	Peptide	NetMHCIIPan IC50 (nM)	TEPITOPE Rank (%)	Est. IC50 (nM)
Library-enriched	AALTGKLGHRGWEEG	11,007	19.1	< 10
	AAVREKCDHVGWEEG	10,964	12.7	76
	AAVTNWGCEVGWEEG	7,402	10.6	136
	AALWRHPGHVGWEEG	6,568	11.6	75
	AAVSDRLPLRGWEEG	5,998	13.8	< 10
	AAVTERPINLGWEEG	5,753	14.4	51
	AAVTEEKSHLGWEEG	5,478	27.0	58
	AALDAHRDHMAWEEG	4,251	14.4	< 10
	AAGRSHRTHEGWEEG	14,717	100.0	10
	AAANRRPSLLGWEEG	7,973	100.0	< 10
	AAPRRHANHLGWEEG	7,224	100.0	130
	AATERRPTLMAWEEG	5,659	100.0	< 10
	AACRKHGSSIGWEEG	5,370	100.0	132
	AAEHVWPSLVGWEEG	4,412	29.7	26
	AAQVDWPTLPMWEEG	4,370	29.0	64
	AAACRKRTWLGWEEG	6,446	100.0	< 10
IAV-derived	TMVMELVRMIKRGIN	171	0.6	870
	TEIIRMMESARPEDV	120	0.4	83
	PALRMKWMMAMKYPI	73	1.0	1310
	YEEFTMVGRRATAIL	229	0.9	4565
	RPMFLYVRTNGTSKI	113	1.0	44
	LGFVFTLTVPSERGL	235	0.3	409
	NRMVLASTTAKAMEQ	198	0.1	< 10
	ARQMVQAMRTIGTHP	256	1.1	36

Our enriched motif again differed substantially from those generated by existing prediction algorithms (Figure 2.8F). For this allele, the dearth of curated peptides necessitates the use of algorithms that incorporate structural data, such as TEPITOPE [48], or nearest-neighbor algorithms, such as NetMHCIIpan [11]. This is demonstrated by the inconclusive motif of NetMHCII, which is trained on only allele-specific curated peptides. While TEPITOPE and NetMHCIIPan reflect the truncation of the P1 pocket and consistent preferences at P6 – again excepting P6 Asp – they continue to have increased uncertainty at P9. While TEPITOPE mirrored our observed preferences at P4 and P7, it does not consider P5 [48], which was essential to our observed motif.

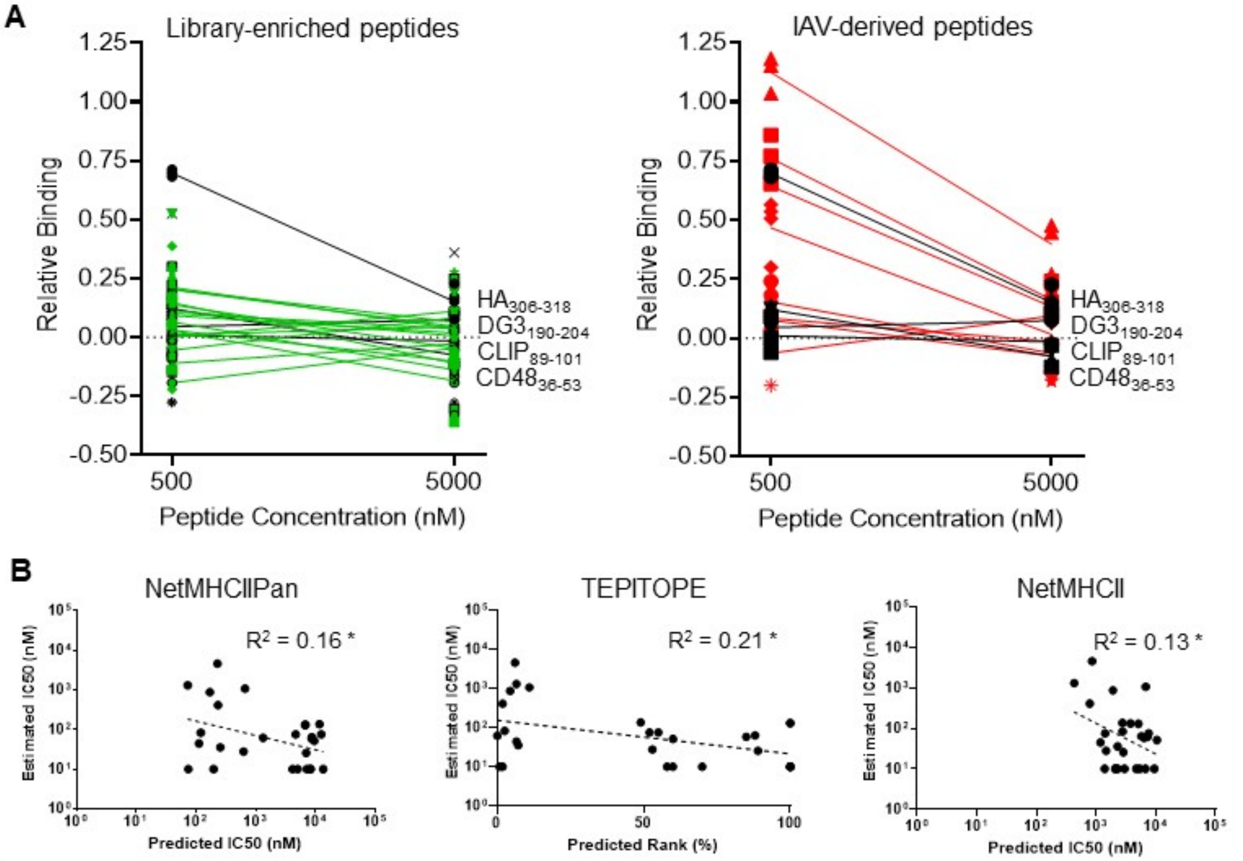


Figure 2.9. *Validation of library-enriched HLA-DR402-binding motif reveals further gaps and inaccuracies in existing class II MHC prediction algorithms. A) Two-point fluorescence polarization competition assay relative binding curves for peptides either found enriched by our randomized 9mer HLA-DR402 library but not predicted to bind HLA-DR402 (Library-enriched) or derived from influenza A virus and predicted to bind HLA-DR402 but not matching our enriched motif (IAV-derived). Selected control peptides are shown in black and curves are fit to N = 3 replicates. B) Scatterplots of algorithmic predictions versus measured IC_{50} with lines of best fit and their associated coefficients of determination (R^2). Asterisk denotes R^2 values of negative correlations*

Our enriched motif was supported by competition assays that validated 16/16 library-enriched peptides (measured IC_{50} < 150 nM) that were not predicted to bind HLA-DR402 by both NetMHCIIpan and TEPITOPE (Table 2.2, Figure 2.9A). These peptides were derived from both clusters within our data, supporting each motif. We further identified 8 peptides from Influenza A virus predicted to be strong binders by both NetMHCIIPan and TEPITOPE, but not matching our overall enriched motif. Interestingly, only 3/8 were found to be weak or non-binders (IC_{50} > 500 nM). This discrepancy may have been caused by treating the two overlapping motifs as one averaged motif. However, the predictions of existing algorithms were again negatively correlated with measured IC_{50} (Figure 2.9B).

These results demonstrate that our platform can generate large quantities of high-quality training data even for alleles for which there are no allele-specific reagents to validate fold and function. It further revealed that HLA-DR alleles can bind peptides in multiple distinct peptide motifs, including non-conventional motifs, introducing inaccuracies in algorithms that overweight hydrophobic preferences at position P1.

2.2.5 Benchmarking performance of yeast-display trained algorithms

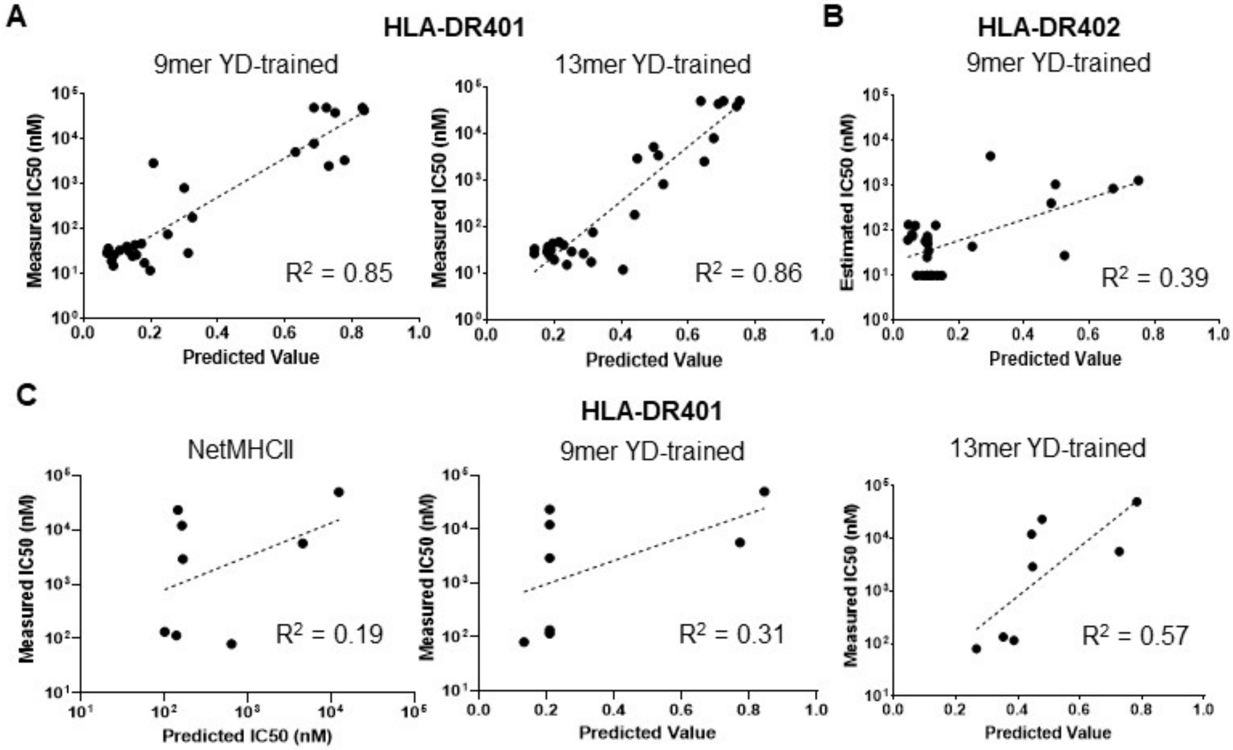


Figure 2.10. *Training on yeast-display library data rectifies deficiencies in existing class II MHC prediction algorithms. Scatter plots of predicted value and measured IC_{50}, with associated lines of best fit and coefficients of determination (R^2) for the following peptides: (A) Enriched by selection of a 9mer HLA-DR401 library but not predicted to bind HLA-DR401, or derived from influenza A virus and predicted to bind HLA-DR401 but not matching our enriched motif; (B) Enriched by selection of a 9mer HLA-DR402 library but not predicted to bind HLA-DR402, or derived from influenza A virus and predicted to bind HLA-DR402 but not matching our enriched motif; C) or variants of wild-type $CII_{261-273}$ peptide. The allele of predicted and measured binding is displayed in bold above each panel.*

We hypothesize that the deficiencies observed in existing class II prediction algorithms in each of the above examples is primarily driven by deficiencies in their underlying training data, rather than the training architectures. To address this hypothesis, we trained prediction algorithms with our yeast-displayed library data using NN-Align, the training architecture underlying NetMHCII and

NetMHCIIPan [63], facilitating direct comparison of the training data. Importantly, training on our yeast-display data improved prediction performance for both alleles when applied to the peptide sets that differentiated our enriched motif from existing algorithms above. Specifically, they displayed greatly improved correlation with measured IC_{50} (Figure 2.10A, B), and correctly classified 24/24 of the HLA-DR401 peptides and 21/24 of the HLA-DR402 peptides as binders or non-binders (rank < 10%, measured IC_{50} < 1 μM, or rank > 10 %, measured IC_{50} > 1 μM, respectively). Furthermore, consistent with the effect of peptide flanking residues on binding, training on the 13mer HLA-DR401 yeast-display data resulted in improved correlation with measured IC_{50} for the $CII_{261-273}$ variant peptides, relative to training on the 9mer library, or to NetMHCII (Figure 2.10C).

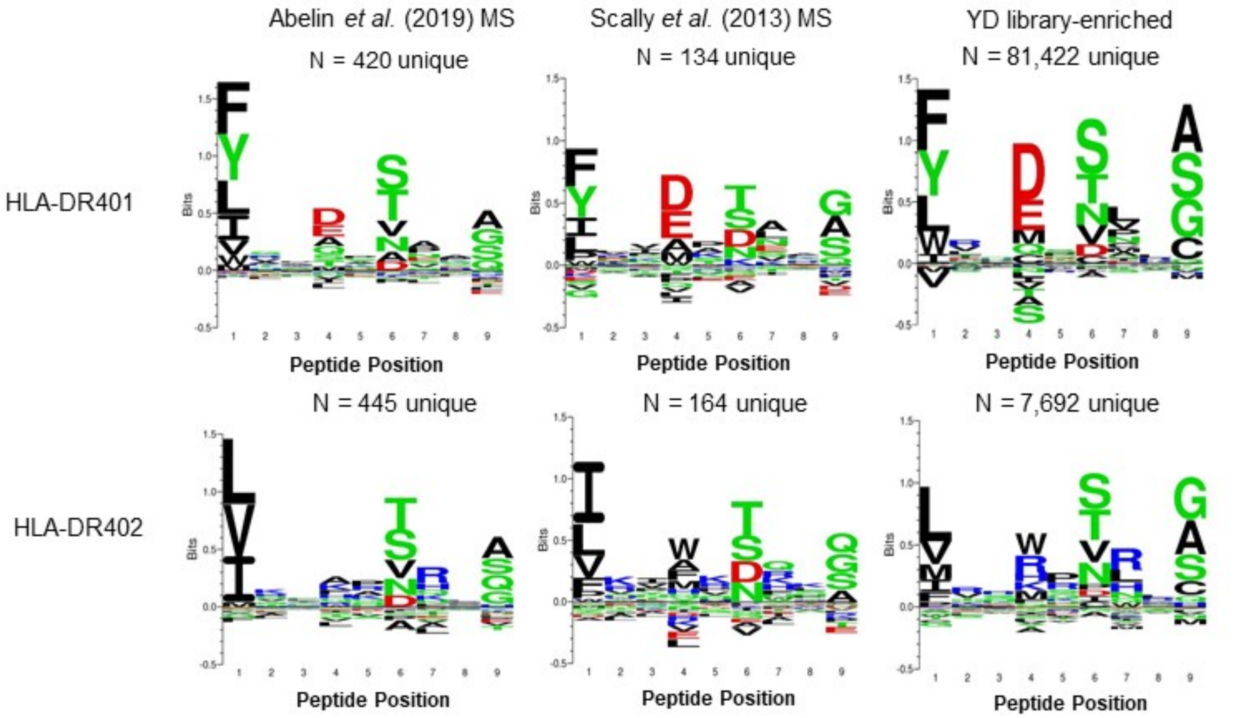
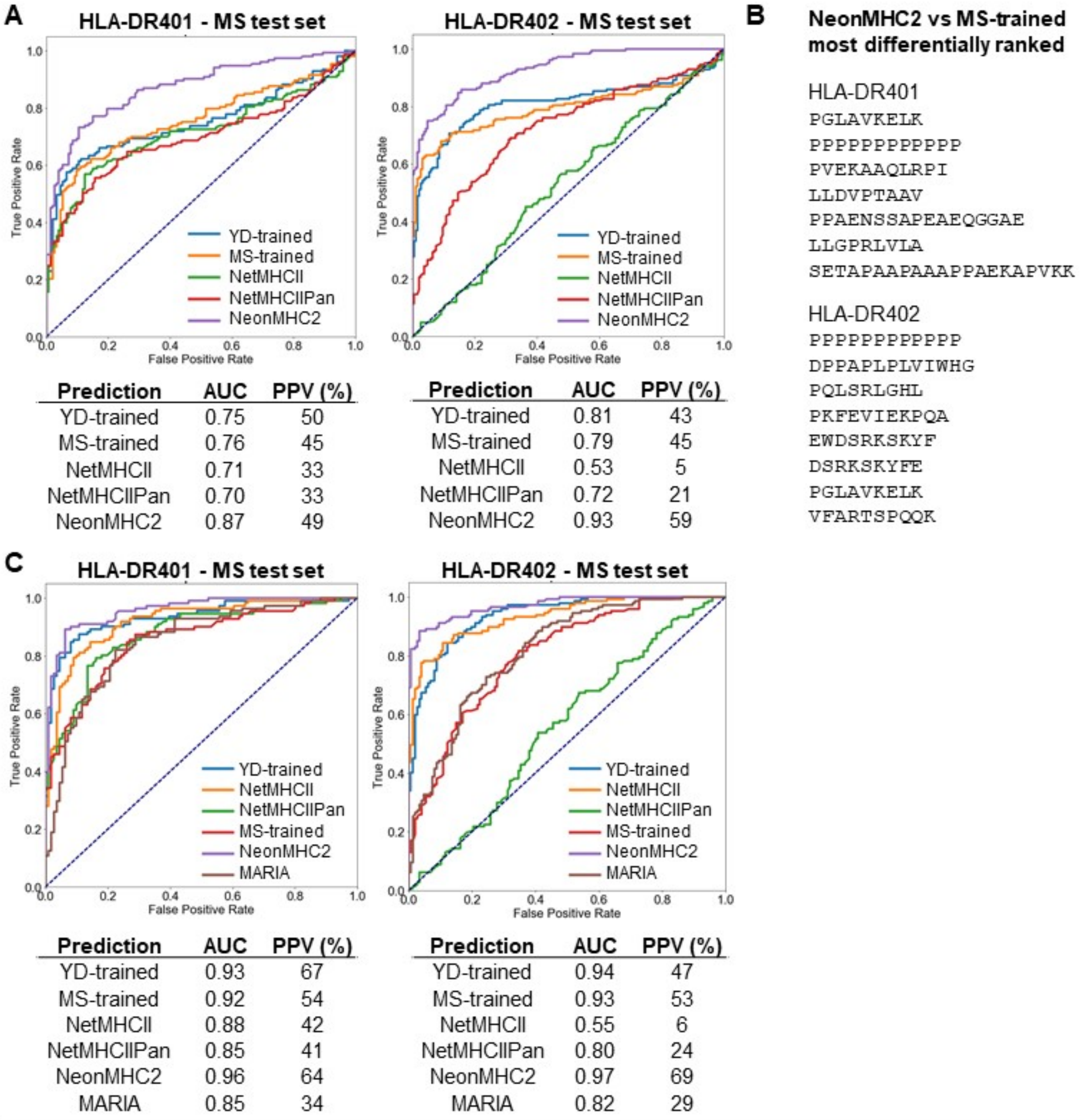


Figure 2.11. *Eluted ligand mono-allelic mass spectrometry and yeast-display library datasets display similar motifs, but vary in size and result in divergent preferences in trained algorithms. A) Kullback-Leibler relative entropy motifs of the core 9 amino acids in HLA-DR401 or -DR402-binding peptides, determined from clustering of the filtered minimum core epitopes of nested sets in eluted ligand mono-allelic mass-spectrometry (MS) datasets, or empirically from round 5 of selection of randomized 9mer yeast-display libraries (YD library-enriched). Number of unique cores comprising each motif are shown.*

To further benchmark the predictive performance of each algorithm, we identified two peptide-binding datasets for each allele that are not represented in current prediction training data [18, 43]. These datasets were generated from eluted ligand mono-allelic mass spectrometry (MS), which utilizes antigen-presenting cells that express only a single class II MHC allele, eliminating the ambiguity in allelic assignment encountered in conventional poly-allelic MS [14, 26]. This method has recently been used to generate high-quality data for many class I and II MHC alleles [14, 15, 18 43,]. While these datasets are over an order of magnitude smaller than those generated by yeast-display in terms of unique peptide cores (Figure 2.11), their peptide motifs are largely consistent with yeast-display, with the exception of P9 Cys, and the absence of two distinct motifs for HLA-

DR402. As the latter dataset [18] underlies the recently published prediction algorithm NeonMHC2, we chose to generate an additional prediction algorithm on this data – again using NN-Align – to provide an additional comparison for training data versus training algorithm.



HLA-DR401 - MS test set (A)

Prediction	AUC	PPV (%)
YD-trained	0.75	50
MS-trained	0.76	45
NetMHCII	0.71	33
NetMHCIIPan	0.70	33
NeonMHC2	0.87	49

HLA-DR402 - MS test set (A)

Prediction	AUC	PPV (%)
YD-trained	0.81	43
MS-trained	0.79	45
NetMHCII	0.53	5
NetMHCIIPan	0.72	21
NeonMHC2	0.93	59

B NeonMHC2 vs MS-trained most differentially ranked

HLA-DR401
PGLAVKELK
PPPPPPPPPPPP
PVEKAAQLRPI
LLDVPTAAV
PPAENSSAPEAEQGGAE
LLGPRLVLA
SETAPAAPAAAPPAEKAPVKK

HLA-DR402
PPPPPPPPPPPP
DPPAPLPLVIWHG
PQLSRLGHL
PKFEVIEKPQA
EWDSRKSKYF
DSRKSKYFE
PGLAVKELK
VFARTSPQQK

HLA-DR401 - MS test set (C)

Prediction	AUC	PPV (%)
YD-trained	0.93	67
MS-trained	0.92	54
NetMHCII	0.88	42
NetMHCIIPan	0.85	41
NeonMHC2	0.96	64
MARIA	0.85	34

HLA-DR402 - MS test set (C)

Prediction	AUC	PPV (%)
YD-trained	0.94	47
MS-trained	0.93	53
NetMHCII	0.55	6
NetMHCIIPan	0.80	24
NeonMHC2	0.97	69
MARIA	0.82	29

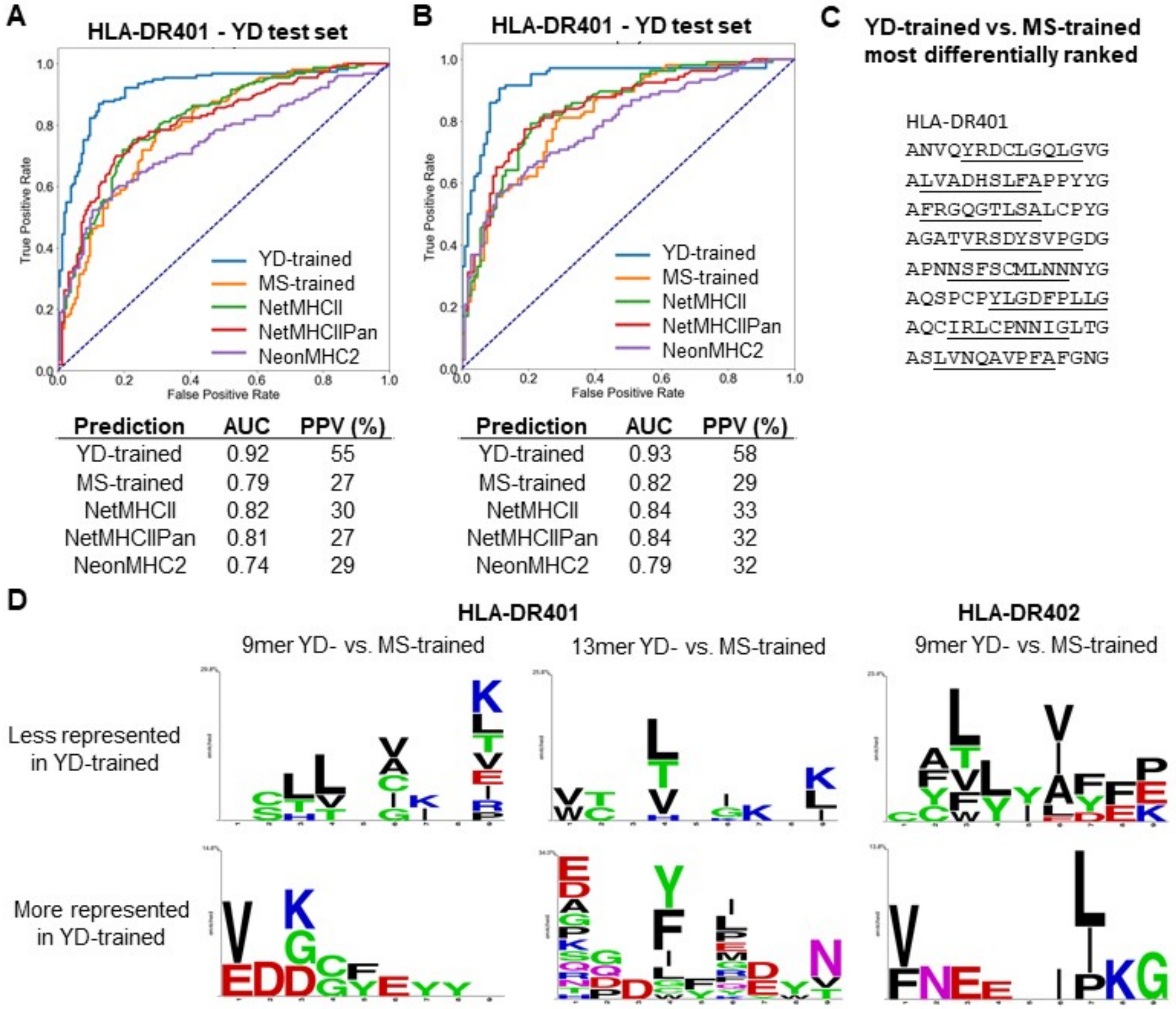
Figure 2.12. *Benchmarking predictive performance on eluted ligand MS data reveals source of algorithmic false positives. A,C) Receiver operating characteristic (ROC) curves for prediction of eluted ligand MS data for HLA-DR401 and -DR402, with expression-matched decoy peptides, before (A) or after (C) unsupervised outlier removal, for existing algorithms or algorithms trained on our 9mer yeast-display library (YD-trained) or independent mono-allelic MS (MS-trained) data. The area under the ROC curve (AUC) and positive predictive value (PPV) of each prediction is shown. B) List of the 8 most differentially ranked peptides within these benchmarking datasets for NeonMHC2, relative to the MS-trained algorithms.*

Each algorithm was applied to the remaining allele-matched dataset [43], with length- and expression-matched decoy peptides, to determine two metrics of predictive performance: the area under the receiver operating characteristic curve (AUC), and the positive predictive value (PPV). For both alleles, the MS- and 9mer yeast-display (YD)-trained models performed comparably to one another, and outperformed NetMHCII and NetMHCIIPan, yet were substantially outperformed by NeonMHC2 (Figure 2.12A). However, analysis of the peptides most differentially valued by NeonMHC2 relative to the MS-trained algorithms (Figure 2.12B) revealed that NeonMHC2 valued proline-rich peptides (two of which were shared between alleles), and highly basic and hydrophobic peptides, which are prone to non-specificity. This finding suggests that eluted ligand MS datasets may contain substantial amounts of non-specific peptides that drive an erroneous outperformance of NeonMHC2, which may be over-fit to these non-specific peptides. In fact, unsupervised clustering of each evaluation set via Gibbs Cluster [54] revealed that each contains a substantial portion (26% for HLA-DR401, 19% for HLA-DR402) of outliers, including peptides with long stretches of Gly or Pro – previously reported to non-specifically populate eluted ligand datasets [64] – and 15/16 of the peptides most differentially valued by NeonMHC2.

In further support of this notion, removal of outlier peptides from the evaluation sets greatly diminished the outperformance of NeonMHC2, but universally improved prediction performance (Figure 2.12C). For both alleles, the MS- and YD-trained algorithms performed comparably (AUC 0.92-0.94), and outperformed NetMHCII and NetMHCIIPan, especially for HLA-DR402. This outperformance was more pronounced for PPV, with the YD-trained algorithm reaching 67% for HLA-DR401. NeonMHC2 also demonstrated impressive AUC (0.96-0.97) and PPV (64-69%) performance for both alleles. As NeonMHC2 is built upon the same underlying data as the MS-trained algorithms, its improved performance here may be due to the incorporation of peptide processing information, such as peptide cleavage preferences [18]. In addition, NeonMHC2 outperformed the recently published class II MHC prediction algorithm, MARIA [16], which is also trained on eluted ligand MS data and considers peptide processing, but utilizes conventional poly-allelic MS data, displaying the importance of non-ambiguous allelic assignment.

While our data suggests that benchmarking on MS-derived data may underestimate false positives for MS-trained algorithms due to over-fitting, benchmarking on these datasets may also underestimate algorithmic false negatives due to gaps in MS-derived data, such as systemic under-sampling of cysteine [18, 43]. Therefore, we further evaluated predictive performance of each algorithm on the 13mer HLA-DR401 YD library data. Here, we observed comparable performance between the MS-trained algorithm, NetMHCII, and NetMHCIIPan (AUC 0.79-0.82, PPV 27-30%) (Figure 2.13A). Additionally, NeonMHC2 slightly underperformed its NN-Align-based counterpart, even though it was used in 'tiling mode' which ignores peptide cleavage preferences, suggesting that the previously noted outperformance was due to factors inherent to MS-derived data, such as peptide processing. Notably, the 9mer YD-trained model clearly outperformed each of the four alternatives, with an AUC of 0.92 and a positive predictive value of 55%, and prediction performance was only minimally improved by removal of outlier peptides (Figure 2.13B).

Considered together, the comparable predictive performance of the yeast-display-trained algorithm on MS-derived data and its over-performance on non-overlapping yeast-display data suggests that there may be peptide motifs in yeast-display data that are not adequately sampled by MS. Direct comparison of the MS- and YD-trained algorithms at a positional level revealed significantly ($p < 0.05$) more stringent preferences at P9 for the YD-trained algorithms (Figure 2.13D), but was less notable for the 13mer-trained algorithm, suggesting it is at least partially driven by register uncertainty in the MS-derived data. Consistent with its under-representation in MS-derived data, Cys was significantly over- or under-represented at multiple positions. Additionally, MS-trained algorithms had a greater preference for small hydrophobic residues at multiple positions. Consistent with these findings, the most differentially ranked peptides in the 13mer YD evaluation set displayed these motifs (Figure 2.13C), and did not appear non-specific.



Prediction	AUC	PPV (%)
YD-trained	0.92	55
MS-trained	0.79	27
NetMHCII	0.82	30
NetMHCIIPan	0.81	27
NeonMHC2	0.74	29

Prediction	AUC	PPV (%)
YD-trained	0.93	58
MS-trained	0.82	29
NetMHCII	0.84	33
NetMHCIIPan	0.84	32
NeonMHC2	0.79	32

HLA-DR401
ANVQYRDCLGQLGVG
ALVADHSLFAPPYYG
AFRGQGTLSALCPYG
AGATVRSDYSVPGDG
APNNSFSCMLNNNYG
AQSPCPYLGDFPLLG
AQCIRLCPNNIGLTG
ASLVNQAVPFAFGNG

Figure 2.13. *Benchmarking on yeast-display data suggests the presence of unique motifs. Receiver operating characteristic (ROC) curves for prediction of round five 13mer yeast-display library data, with decoys from the naïve library, for existing algorithms or algorithms trained on 9mer yeast-display library (YD-trained) or mono-allelic MS (MS-trained) data, before (A) or following (B) unsupervised outlier removal, with associated area under the ROC curve (AUC) and positive predictive values (PPV). C) List of peptides most differentially ranked between the YD- and MS-trained algorithms, with 9mer core underlined. D) Amino acids significantly (p < 0.05) differentially represented at each position within the 9mer core of HLA-DR401 or -DR402-binding peptides, as determined by the YD-trained algorithms relative to the MS-trained algorithms. Displayed size of residues corresponds with statistical significance.*

A

	Name	Peptide	WT NeonMHC2 Rank (%)	WT YD-trained Rank (%)	Mut. NeonMHC2 Rank (%)	Mut. YD-trained Rank (%)
Neoantigens	HEATR1 G170V	LLPVKQS**V**VPLAKGTLITHC	7.65	5.4	13.7	1.8
	PHF20L1 R499M	ASDSSYRNECP**M**AEKEDTQM	30.2	21.5	21.6	1.8
	DNAH9 L2180I	MKRRPVWTD**I**NPKAVTNDEL	13.6	4.3	11.1	0.8
	PARP8 R94S	PIPVIFHRIATEL**S**KTNDIN	0.8	28.4	< 0.1	0.4
	NUP210L K1682N	VYGWATLVSERS**N**NGMQRIL	0.4	26.8	0.3	0.5
	COG6 A318S	YVGDMLAWLHQ**S**TASEKEHL	1.0	46.2	0.5	33.5
	C2orf63 P161R	EVKYCTFSKD**R**SKPIPGMTL	1.3	32.2	1.5	17.5
	PXDNL L1117I	WRAPSY**I**LSPELTQRLFSAA	2.1	37.9	1.8	23.6
	POTEF R162M	RKDLIVML**M**DTDVNKQDKQK	1.8	29.8	1.0	18.6
	OR3A1 Q189R	YCDLPQLF**R**LSCSSTQLNEL	2.1	30.6	2.0	16.0
	PCDH10 G42C	FVGNIAEDL**C**LDITKLSARG	2.4	43.2	1.6	18.5
	MYO1H D248Y	ESSISDKN**Y**WKTVSNAFSVI	5.8	25.4	2.0	11.5
	ADAMTS14 V267L	YSIEVLLV**L**DDSVVRFHGKE	1.4	30.3	1.2	29.9
IAV-derived	HA 486-505	FYHKCDNECMESVRNGTYDY	51.1	1.7	-	-
	PB2 447-466	LFQNWGVEPIDNVMGMIGIL	12.9	1.6	-	-
	PA 300-319	EGIPLYDAIKCMRTFFGWKE	0.1	16.6	-	-
	M1 2-21	ARQMVQAMRTIGTHPSSSAG	1.7	31.5	-	-
	PA 562-581	MFLYVRTNGTSKIKMKWGME	0.9	43.1	-	-

B

Name	NeonMHC2 Rank (%)	YD-trained Model Rank (%)	Measured IC50 (nM)
HEATR1 G170V	13.7	1.8	307
PHF20L1 R499M	21.6	1.8	47
DNAH9 L2180I	11.1	0.8	21
PARP8 81-100	0.8	28.4	698
NUP210L 1670-89	0.4	26.8	245
COG6 A318S	0.5	33.5	406
C2orf63 P161R	1.5	17.5	> 50,000
PXDNL L1117I	1.8	23.6	248
POTEF R162M	1.0	18.6	3173
OR3A1 Q189R	2.0	16.0	39
PCDH10 G42C	1.6	18.5	86
MYO1H D248Y	2.0	11.5	25
ADAMTS14 V267L	1.2	29.9	> 50,000
HA 486-505	51.1	1.7	7273
PB2 447-466	12.9	1.6	2524
PA 300-319	0.1	16.6	61
M1 2-21	1.7	31.5	3994
PA 562-581	0.9	43.1	6102

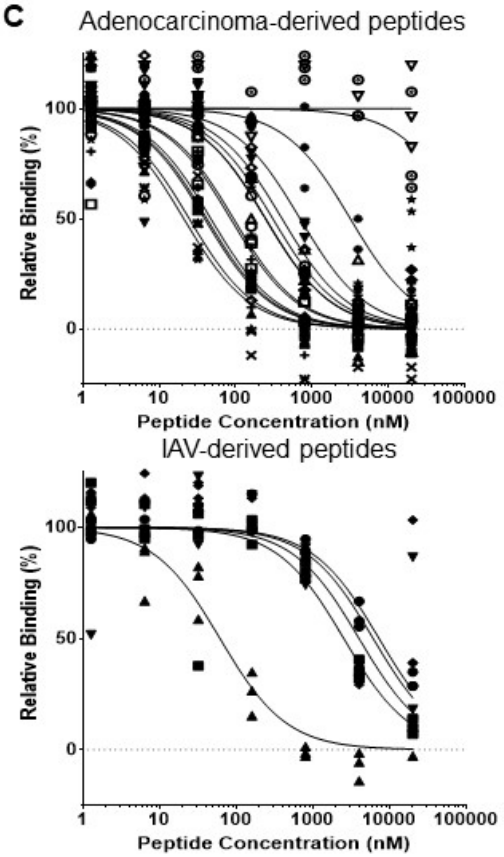


Figure 2.14. *Prediction of pathogen and tumor-associated peptides shows unique improvements for training existing algorithms on yeast-display library data. A) Table of mutant peptides from human lung adenocarcinomas differentially predicted as neoantigens for HLA-DR401, or peptides derived from influenza A virus differentially predicted as strong- versus non-binders, by NeonMHC2 or a model trained on our 9mer yeast-display library data. Mutations in adenocarcinoma-derived peptides are noted in bold. B) Table of predicted percentile ranks and measured IC_{50} values of peptides on which NeonMHC2 or a yeast-display trained algorithm disagreed, with associated binding curves for HLA-DR401 from fluorescence polarization competition assays for these peptides (C). Curves are fit to N = 3 replicates.*

To investigate the effect these differences may have on the prediction of clinically relevant peptides, we performed antigen prediction for HLA-DR401 with NeonMHC2 and the 9mer yeast-display-trained algorithm on the proteome of Influenza A virus (IAV), and expression-validated mutations from human lung adenocarcinoma patients [65]. From these datasets, the 9mer yeast-display-trained model differentially classified (relative to NeonMHC2) 5 IAV-derived peptides as strong or non-binders, and differentially classified 13 adenocarcinoma-derived peptides as potential neoantigens (Figure 2.14A). Interestingly, these algorithms displayed non-overlapping algorithmic misses (Figure 2.14B), suggesting that there are peptide motifs unique to each training set that contribute to improved peptide prediction performance. Importantly, however, when all peptides assayed for binding to HLA-DR401 in this study were considered, yeast-display-trained models displayed substantially improved correlation with measured IC_{50} relative to all alternative algorithms, which performed poorly (Figure 2.15A). Consistent with our findings on peptide flanking residues, the predictions of the 13mer yeast-display-trained model displayed the greatest correlation with measured IC_{50} ($R^2 = 0.62$). These findings also held true when each prediction was converted to percent rank to match the output of NeonMHC2 (Figure 2.15B).

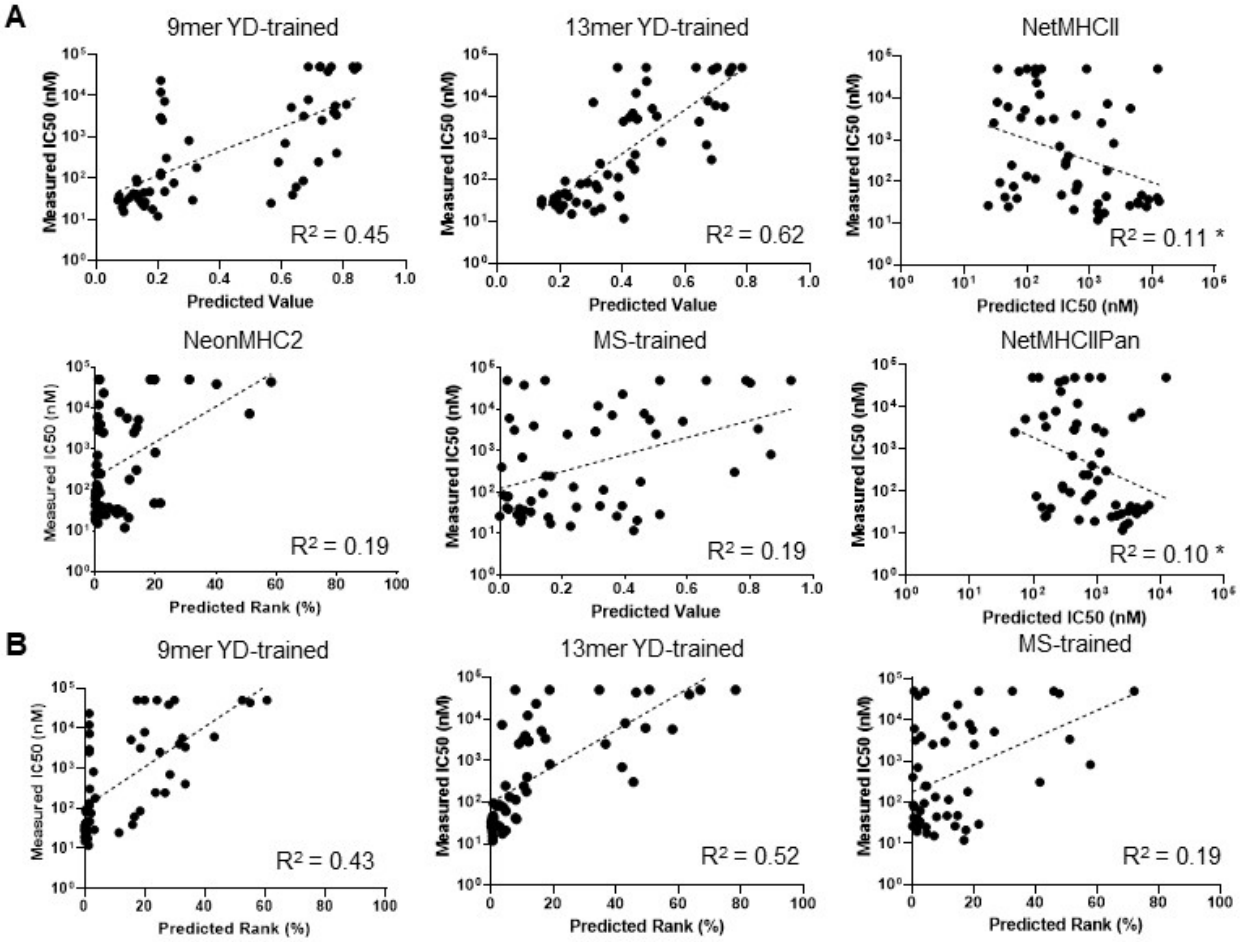


Figure 2.15. *Training existing algorithms on yeast-display library data improves estimation of peptide-binding affinity. Scatterplots of algorithmic predictions versus measured IC_{50} values for 56 peptides assayed for binding to HLA-DR401 in fluorescence polarization competition assays for native output (A) or converted to percentile rank (B), with associated lines of best fit and coefficients of determination (R^2).*

Overall, our results demonstrate that both mono-allelic-MS- and yeast-display-generated peptide datasets greatly improve the performance of class II prediction algorithms, and can identify unique peptide motifs that contribute to improved prediction performance. However, we find that yeast-display provides much larger datasets than mono-allelic MS, and provides improved performance in predicting peptide affinity.

2.3 Discussion

The central role of CD4$^+$ T cells across infection, cancer, autoimmunity, and allergy motivates a need to better predict which peptide antigens can be presented by class II MHCs. However, class II prediction algorithms suffer from consequential gaps and inaccuracies in coverage, especially for poorly characterized alleles [11, 20-24].

Here, we present a platform for large-scale unbiased identification of class II MHC-binding peptides to identify and rectify these gaps and inaccuracies. We showed that this platform does not require extensive prior knowledge of a peptide-binding motif or allele-specific reagents, allowing application to poorly characterized class II MHC alleles. We demonstrated that our platform generates over an order of magnitude more unique data than comparable approaches for two human class II MHC alleles, and identifies motifs that are missed by current data collection techniques and frequently used prediction algorithms. We further validate that these deficiencies and inaccuracies are rectified when existing algorithms are trained upon our yeast-display library data, and use these algorithms to discover *bona fide* peptide binders that are not predicted by other prediction algorithms.

Analysis of the training data underlying existing prediction algorithms revealed multiple sources of underperformance. For both alleles studied, we found large numbers of nested sets and single amino acid variant peptides within curated training sets. While training algorithms account for redundant information from nested sets [29], their presence diminishes the functional size of the training set. However, single amino acid variants are considered unique peptides, and can therefore impart biases. Furthermore, a systemic absence of cysteine in training sets resulted in substantial algorithmic false negatives for both alleles. While this is likely due in part to an aversion to working with cysteine-containing peptides, it is also driven by the difficulties inherent to sampling them in mass-spectrometry (MS) datasets [66]. A systemic underrepresentation of acidic residues in the IEDB has also been reported [18]. In comparison, no systematic absences were observed within our yeast-display data.

In addition, we found that yeast-displayed libraries uniquely benefit from their large size and engineered composition. By engineering randomized peptide libraries with defined flanking residues, we reduced register uncertainty and increased anchor preference resolution. Meanwhile, the large size of our libraries enabled identification of consequential preferences at non-anchor residues, including those outside the peptide-binding groove. Our libraries also enabled us to identify two distinct motifs for HLA-DR402 that were not adequately captured by curated peptides or eluted ligand MS. The coexistence of two unique binding motifs, including one of which defies the conventional notion of a hydrophobic P1 residue-driven HLA-DR motif in favor of hydrophobic residue at P4, is unique relative to recent reports of HLA-DR alleles [17-18]. The smaller size of the mono-allelic MS-derived dataset and its under-representation of Trp [18] – which dominated this newly-described motif – may account for its absence.

By using our data to train prediction algorithms and benchmark their performance against existing algorithms and those generated by comparable approaches, we identified key considerations for antigen prediction moving forward. First, our results demonstrate that high-quality training data improves the performance of class II prediction algorithms without alteration of underlying training algorithm architectures, especially for poorly characterized alleles. However, there are important opportunities for algorithmic improvement, such as improved binding register determination and increased focus on peptide flanking residues. Second, we find that each source of data has non-overlapping strengths and weaknesses for improving prediction performance. Therefore, we believe that an ideal class II MHC prediction algorithm may be trained on high-quality datasets that reflect native processing [66], such as mono-allelic MS datasets, as well as large and diverse peptide datasets, such as those generated by our yeast-display platform. Third, we highlight the importance of the choice of validation sets for benchmarking prediction algorithms, as frequently used metrics of prediction performance underestimate false negatives due to gaps in test sets, allowing entire classes of peptides to be missed without impacting performance metrics. Additionally, we found that false positives may be under-represented when benchmarking on eluted MS data when using MS-trained algorithms, such as NeonMHC2, which appears to be over-fit to this data source. Finally, we find that yeast-display-trained algorithms are markedly superior at predicting peptide affinity, which is a crucial consideration in identifying peptides suitable for antigen-targeted therapeutics [6-8]. The non-binary nature of yeast-display data, which is trained on peptides from five rounds of selection, possibly accounts for this key disparity.

Lastly, as this platform does not require allele-specific reagents, we believe it can generate high-quality repertoire-scale data for many additional class II alleles, even those with few curated binders, greatly increasing its applicability. As such, we believe this technology will greatly benefit the field of class II MHC antigen prediction, and therefore the study and application of CD4$^+$ T cell recognition across pathogen infection, cancer, and immune disorders.

2.4 Methods

2.4.1 Yeast-displayed pMHC design and peptide exchange

Full-length yeast-displayed HLA-DR401 (HLA-DRA1*01:01, HLA-DRB1*04:01) with a cleavable peptide linker was based upon a previously described HLA-DR401 construct optimized for yeast display with the mutations Mα36L, Vα132M, Hβ62N, and Dβ72E to enable proper folding without perturbing either TCR- or peptide-contacting residues [32]. The alpha and beta chain ectodomains are expressed as a single transcript connected by a self-cleaving P2A sequence. The peptide is joined through a flexible linker to N-terminus of MHC β1 domain. This construct was further modified to express a 3C protease site (LEVLFQ/GP) and MYC epitope tag (EQKLISEEDL) within the flexible linker, for a total of 32 amino acids between the peptide and β1 domain. HLA-DR402 (HLA-DRA1*01:01, HLA-DRB1*04:02) was generated by modification of this construct with each native HLA-DRβ polymorphism of HLA-DR402. All yeast-display constructs were produced on the pYAL vector as N-terminal fusions to AGA2. All yeast strains were grown to confluence at 30°C in pH 5 SDCAA yeast media then subcultured into pH 5 SGCAA media at OD_{600} = 1.0 for 48h induction at 20°C [67].

For peptide retention experiments, the linker between peptide and MHC was cleaved with 1 μM 3C protease in PBS pH 7.4 at a concentration of $2x10^8$ yeast/mL for 45 minutes at room temperature. After linker cleavage, yeast expressing the pMHC were washed into pH 5 citric acid saline buffer (20mM citric acid, 150 mM NaCl) at $1x10^8$ yeast/mL with 1 μM HLA-DM and a high-affinity competitor peptide at 4°C to catalyze peptide exchange. HLA-DR401-expressing yeast were incubated with 1 μM $HA_{306-318}$ (PKYVKQNTLKLAT) and HLA-DR402-expressing yeast were incubated with 5 μM $CD48_{36-53}$ (FDQKIVEWDSRKSKYFES) (Genscript, Piscataway NJ). Peptide dissociation was tracked through a AlexaFluor647-labeled ∝-Myc antibody (Cell Signaling Technologies, Danvers MA) on an Accuri C6 flow cytometer (Becton Dickinson, Franklin Lakes NJ). For each construct, N = 3 aliquots were treated independently and measured for each time point and condition. Statistical evaluation of dissociation experiments was performed by repeated measures two-way ANOVA with Dunnett's test for multiple comparison within treatment conditions, or Tukey's test for multiple comparisons across treatment conditions, in Prism 8.0 (GraphPad Software Inc, San Diego CA).

2.4.2 Library design and selection

Randomized peptide yeast libraries were generated by polymerase chain reaction (PCR) of the pMHC construct with primers encoding NNK degenerate codons. To ensure only randomized peptides expressed within the library, the template peptide-encoding region encodes multiple stop codons. Randomized 9mer libraries were designed as [AAXXXXXXXXXWEEG…] to constrain peptide register and randomized 13mer libraries were designed as [AXXXXXXXXXXXXXG…]. Randomized pMHC PCR product and linearized pYAL vector backbone were mixed at a 5:1 mass ratio and electroporated into electrically competent RJY100 yeast [68] to generate libraries of at least $1x10^8$ transformants. Libraries were subjected to 3C cleavage and peptide exchange for 16-18 h, as described above, and were selected for peptide-retention via binding of ∝-Myc-AlexaFluor647 antibody and magnetic ∝-AlexaFluor647 magnetic beads (Miltenyi Biotech, Bergisch Gladbach, Germany). Selected yeast were re-cultured, induced, and selected for an additional four rounds, for five total rounds of selection.

2.4.3 Library deep sequencing and analysis

Libraries were deep sequenced to determine the peptide repertoire at each round of selection. Plasmid DNA was extracted from $5x10^7$ yeast from each round of selection with the Zymoprep Yeast Miniprep Kit (Zymo Research, Irvine CA), according to manufacturer's instructions. Amplicons were generated by PCR with primers designed to capture the peptide encoding region through the polymorphic region that differentiates HLA-DR401 from HLA-DR402. An additional PCR round was then performed to add P5 and P7 paired-end handles with inline sequencing barcodes unique to each library and round of selection. Amplicons were sequenced on an Illumina MiSeq (Illumina Incorporated, San Diego CA) with the paired-end MiSeq v2 500bp kit at the MIT BioMicroCenter.

Paired-end reads from were assembled via FLASH [69] and processed with an in-house pipeline which filters for assembled reads with exact matches for the expected length, polymorphic sequences, and 3C protease cleavage site, then sorts each read based on its inline barcode and extracts the peptide-encoding region. To ensure only high-quality peptides were analyzed, reads were discarded if any peptide-encoding base pair was assigned a Phred33 score less than 20, or did not match the expected codon pattern at NNK sites (N = any nucleotide, K = G or T). To account for PCR and read errors of high-prevalence peptides, reads were discarded if their peptide-encoding regions were Hamming distance 1 from any more prevalent sequence, Hamming distance 2 from a sequence 100 times more prevalent, or Hamming distance 3 from a sequence 10,000 times more prevalent within the same round, in line with previously published analysis methods [70]. Unique DNA sequences were translated by Virtual Ribosome [71] and filtered for peptides not encoding a stop codon.

2.4.4 Heat map visualization of library peptide preferences and determination of peptide register

Heat maps were generated from filtered sequences from each round to visually represent positional preferences. For each round, the unweighted prevalence of each amino acid at each position was calculated as a percentage. This positional percent prevalence was compared to its matched value in the unselected library to generate log2-fold enrichment values. The significance of deviations from unselected library positional amino frequencies was evaluated using an unweighted binomial test using 10,000 peptides to establish each distribution in kpLogo [72], with a Bonferroni correction for multiple hypothesis testing.

For randomized 9mer libraries, these log2-fold enrichment values were used to generate 20x9 position-specific scoring matrices (PSSMs) that were used to identify out-of-register peptides in round 5 of selection. Each 15mer peptide was scored in each of its seven possible 9mer registers by the PSSM, without positional weighting. Peptides which scored highest in a shifted register, regardless of score, were deemed out-of-register. For the randomized 13mer library, peptide register was determined by Gibbs Cluster 2.0 [54], with settings imported from 'MHC class I ligands of the same length', a motif of 13 amino acids, no discarding of outlier peptides, and background amino acid frequencies derived from the data. This allowed visualization of each peptide register independently, without collapsing to a common 9mer motif. The number of unique clusters was determined by maximum Kullback-Leibler distance. Results were comparable between both methods of register determination for the 9mer peptide data.

2.4.5 Analysis of peptide data from external data sources

External MHC-binding peptide data was curated either from the SYFPEITHI database [30] or from two previously-published eluted ligand mono-allelic mass-spectrometry (MS) datasets [18], [43]. Eluted ligand mono-allelic MS peptide data was analyzed as previously recommended [43], identifying the minimum epitope of nested peptide sets and filtering for those which do not map to immunoglobulin or HLA proteins. Each dataset was clustered by Gibbs Cluster 2.0 [54] with default settings for 'MHC class II ligands', excepting the default removal of outlier peptides, and amino acid frequencies 'from data', to identify the core 9mer of each peptide. In each case, Kullback-Leibler distance was maximized for one cluster. For identification of outlier peptides, the default removal of outlier peptides was enabled.

2.4.6 Generation and comparison of peptide motifs

Kullback-Leibler relative entropy motifs were generated with Seq2Logo 2.0 [73]. For yeast-display data, the core 9mers of round 5 sequences were input with background amino acid frequencies derived from their average in their matched unselected library. For externally sourced peptide data, unique core 9mers were input with background frequencies from the UNIPROT [74] average of each amino acid. Motifs for prediction algorithms were generated by application of each prediction to a computationally-generated set of 50,000 unique 15mer peptides with the UNIPROT average frequency of each amino acid. Prediction with each of NetMHCII 2.3 [11], TEPITOPE [48], NetMHCIIPan 3.2 [11], the IEDB consensus tool [49] produces a predicted value and core 9mer. Predicted core 9mers of peptides that met published recommendations for binding (NetMHCII and NetMHCIIPan: $IC_{50} < 500$ nM, TEPITOPE: rank < 6, IEDB Consensus: rank < 10) were input into Seq2Logo with UNIPROT average background frequencies.

Comparison of the two clusters found within round 5 of our randomized 9mer HLA-DR402 library selection data was performed with Two Sample Logo [75]. Significance was determined by two-sided unweighted binomial test for $p < 0.05$, with a Bonferroni correction for multiple hypothesis testing.

2.4.7 Training of peptide prediction algorithms

Allele-specific class II MHC prediction models were generated from yeast-display library data or from external mono-allelic MS data [18, 43] using NN-Align 2.0 [63]. For yeast-display library data, the randomized residues of up to 80,000 sequenced peptides were assigned a target value commensurate with the final round of selection in which they were observed between 0 and 1, with increasing target value for observation in later rounds. The 9mer library data was used for training with default settings for 'MHC class II ligands', excepting expected peptide length set to 9 amino acids and expected PFR (peptide flanking residue) length set to 0 amino acids. The 13mer library data was used for trained with default settings, excepting expected peptide length set to 13 amino acids.

For the mono-allelic MS data, curated filtered minimum epitopes were assigned a target value of 1. In order to prevent the algorithm from conflating altered amino acid frequencies arising from MS data collection with peptide-binding preferences, each peptide was scrambled to generate negative instances and assigned a target value of 0, in line with previously published recommendations [18]. These algorithms were trained with default 'MHC class II ligands' settings.

Reported prediction values are the inverse of model output prediction values (1-value) to match other prediction models for ease of comparison. Percentile ranks were established by comparison of prediction values to the distribution of prediction values generated by applying each prediction to 50,000 computationally generated random 15mer peptides (see above).

2.4.8 Benchmarking and comparison of prediction algorithms

Prediction algorithms were benchmarked against independently generated allele-specific eluted ligand mono-allelic MS or yeast-display library data, with matched decoy peptides. For the MS datasets, the filtered minimum core epitopes (see above) were classified as positive instances, and length- and expression-matched decoy peptides were randomly selected from a pool of computationally generated peptides as previously described [18]. For each protein observed within the dataset, we tiled across its sequence with peptide lengths randomly selected from the length distribution of the observed peptides, starting at the first amino acid in the protein and allowing an eight amino acid overlap between subsequent proteins. If the length of the last peptide extended beyond the end of the protein, we randomly shifted the starting amino acid such that the starting amino acid of the first peptide and last amino acid of the final peptide were all within the protein. We randomly selected decoy peptides from this set such that the length distribution of decoy peptides matched that of the positive instances, and that there was no 9mer sequence match with the other decoys or positive instances. For the yeast-display dataset, a randomly selected size-matched set of peptides found enriched in round 5 of selection were classified as positive instances, and decoy peptides were randomly selected from peptides only observed in their respective unselected library. A 1:1 ratio of positive instances and decoy peptides was used to generate receiver operating characteristic (ROC) curves. A 1:19 ratio of positive instances and decoy peptides was used for calculation of positive predictive value (PPV), which is calculated as the fraction of true instances observed in the top 5% of predicted value for each algorithm [18].

Prediction algorithms were compared at a positional level by Two Sample Logo [17]. For each comparison, the two algorithms were applied to a common set of 50,000 computationally-generated 15mer peptides (see above). The predicted core 9mer of peptides which rank within the 90th percentile or higher of predicted value for only one algorithm are evaluated against the cores of peptides which rank within the 90th percentile or higher of predicted value for both algorithms. Significance was determined by two-sided unweighted binomial test for $p < 0.05$, with a Bonferroni correction for multiple hypothesis testing.

2.4.9 Recombinant protein production

Recombinant soluble HLA-DM, HLA-DR401, and HLA-DR402 were produced in High Five (Hi5) insect cells (Thermo Fisher) via a baculovirus expression system, as previously described for other class II MHC proteins [31]. Briefly, ectodomain sequences of each chain followed by a poly-histidine purification site were cloned into pAcGP67a vectors. For each construct, 2 µg of plasmid DNA was transfected into SF9 insect cells with BestBac 2.0 linearized baculovirus DNA (Expression Systems, Davis CA) using Cellfectin II reagent (Thermo Fisher, Waltham MA. Viruses were propagated to high titer, co-titrated to maximize expression and ensure 1:1 MHC heterodimer formation, and co-transduced into Hi5 cells, which were then grown at 27°C for 48-72h. Proteins were purified from the pre-conditioned media supernatant with Ni-NTA resin and size purified via size exclusion chromatography using a S200 increase column on an AKTAPURE FPLC (GE Healthcare, Chicago IL). HLA-DRB1*04:01 and HLA-DRB1*04:02 chains were expressed with $\text{CLIP}_{81\text{-}101}$ peptide connected by a 3C-protease-cleavable flexible linker to the MHC N-terminus, which improved protein yields.

2.4.10 Fluorescence polarization competition assays and peptide IC50 determination

The IC_{50} of characterized peptides was quantified with a protocol adapted from Yin, L. and Stern, L.J. (2014) [50]. Relative binding values were generated at each concentration according to the equation (FP_{sample} – FP_{free})/($FP_{no\ comp}$ – FP_{free}), where FP_{free} is the polarization value of the fluorescent peptide before addition of MHC, FP_{no_comp} is the polarization value with added MHC but no competitor peptide, and FP_{sample} is the polarization value with added MHC and competitor peptide. Relative binding curves were generated and fit by Prism 8.0 (GraphPad Software Inc, San Diego CA) to the equation $y = 1/(1+[pep]/IC_{50})$, where [pep] is the concentration of competitor peptide, to determine the IC_{50} of each peptide, its concentration of half-maximal inhibition.

For each 200 µL assay, 100 nM soluble MHC was combined with 25 nM of fluorescently-modified peptide in pH 5 binding buffer and incubated at 37°C for 72h in black 96-well flat bottom plates (Greiner Biotech, Kremsmünster, Austria). Modified $HA_{306-308}$ peptide [APRFV{Lys(5,6 FAM)}QNTLRLATG] was used for HLA-DR401 and modified $CD48_{36-53}$ peptide [AQRIVEWDSR{Lys(5,6) FAM)}SRYG] was used for HLA-DR402. N=3 replicates were performed for each unlabeled peptide (Genscript, Piscataway NJ) concentration, ranging in five-fold dilutions from 20 µM to 1.28 nM. Plates were read on a Tecan M1000 (Tecan Group Ltd., Morrisville NC) with 470nm excitation, 520 nm emission, optimal gain, and a G-factor of 1.10. An important modification of our protocol is the presence of the MHC-linked CLIP peptide, which was released by incubation with 3C protease at a 1:100 molar ratio at room temperature for 1h prior to dilution into plates. Residual cleaved CLIP peptide at 100 nM is not expected to alter peptide binding.

Due to poor soluble expression of HLA-DR402, the assay for HLA-DR402-binding peptides was limited to two concentrations of unlabeled competitor peptide for this allele. However, we found high correlation between two-point estimated IC_{50} values and those obtained from full titration curve fitting for HLA-DR401.

Lines of best fit between predicted and measured affinity for characterized peptide, and associated determinants of determination (R^2), were generated in Prism 8.0 (GraphPad Software Inc, San Diego CA).

2.5 Acknowledgements

This contents of this chapter has been submitted for peer review as Rappazzo, C.G, Huisman, B.D., & Birnbaum, M.E. "Repertoire-scale determination of class II MHC peptide binding via yeast display improves antigen prediction". Brooke D. Huisman contributed greatly to the described work, especially to the development, benchmarking, and comparative analysis of the yeast-display algorithms. Michael E. Birnbaum contributed to the initial conceptualization of this work, guided construct and experimental design, and contributed to analysis of the data. Brooke and Michael additionally contributed to the writing of this chapter. We would like to thank K. Christopher Garcia (Stanford University) for providing the plasmids encoding 3C protease and yeast-displayed HLA-DR401, Patrick Holec and Robert Wilson for their assistance in analyzing computational data, and Lauren Stopfer and Forest White for their insight and feedback. We would also like to thank Anthony W. Purcell (Monash University) and Michael S. Rooney (Neon Therapeutics) for generously providing the mono-allelic mass spectrometry data for HLA-DR401 and HLA-DR402. This work was supported by the staff of the Koch Institute Swanson Biotechnology Center, especially the staff of the flow cytometry, biopolymers and proteomics, high-throughput sciences, and genomics cores. This work was also supported by the National Cancer Institute (P30-CA14051), the Packard Foundation, the V Foundation and the AACR-TESARO Career Development Award for Immuno-oncology Research [17-20-47-BIRN] for M.E.B.

References

1. Blackwell, J. M., Jamieson, S. E., & Burgner, D. HLA and infectious diseases. *Clin. Microbiol. Rev*. **22**, 370-85 (2009).

2. Hadrup, S., Donia, M., Thor-Straten, P. Effector CD4 and CD8 T cells and their role in the tumor microenvironment. *Cancer Microenviron*. **6**, 123-133 (2013).

3. Bluestone J.A., Bour-Jordan, H., Cheng, M., & Anderson, M. T cells in the control of organ-specific autoimmunity. *J. Clin. Invest.* **125**, 2250-2260 (2015).

4. Woodfolk, J. A. T-cell responses to allergens. *J. Allergy Clin Immunol*. **119**, 280-294 (2007).

5. Issa, F., Schiopu, A., Wood, K.J. Role of T cells in graft rejection and transplantation tolerance. *Expert Rev. Clin. Immunol*. **6**, 155-169 (2010).

6. Backert, L. & Kohlbacher, O. Immunoinformatics and epitope prediction in the age of genomic medicine. *Genome Med*. **7**, 119 (2015).

7. Hu, Z., Ott, P. A., & Wu, C. J. Towards personalized, tumour-specific, therapeutic vaccines for cancer. *Nat. Rev. Immunol*. **18**, 168-182 (2018).

8. Patronov, A. & Doytchinova, I. T-cell epitope vaccine design by immunoinformatics. *Open Biol*. **3**, 120139 (2013).

9. Jurtz, V. *et al*. NetMHCpan-4.0: Improved Peptide-MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. *J. Immunol*. **199**, 3360-3368 (2017).

10. O'Donnell, T. J. *et al*. MHCflurry: Open-Source Class I MHC Binding Affinity Prediction. *Cell Syst*. **7**, 129-132. (2018).

11. Jensen, K. K. *et al*. (2018) Improved methods for predicting peptide binding affinity to MHC class II molecules. *Immunology*. **154**, 394-406 (2018).

12. Bassani-Sternbern, M. *et al*. Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry. *Nat. Commun*. **7**, 13404 (2016).

13. Graham, D. B. *et al*. Antigen discovery and specification of immunodominance hierarchies for MHCII-restricted epitopes. *Nat. Med*. **24**, 1762-1772 (2018).

14. Abelin, J. G. *et al*. Mass spectrometry profiling of HLA-associated peptidomes in mono-allelic cells enables more accurate epitope prediction. *Immunity*. **46**, 315-326. (2017).

15. Sarkizova, S. *et al.* A large peptidome dataset improves HLA class I epitope prediction across most of the human population. *Nat. Biotechnol*. **38**, 199–209 (2020).

16. Chen, B. *et al.* Predicting HLA class II antigen presentation through integrated deep learning. *Nat. Biotechnol.* **37**, 1332–1343 (2019).

17. Racle, J. *et al.* Robust prediction of HLA class II epitopes by deep motif deconvolution of immunopeptidomes. *Nat. Biotechnol*. **37**, 1283–1286 (2019).

18. Abelin, J. G. *et al.* Defining HLA-II Ligand Processing and Binding Rules with Mass Spectrometry Enhances Cancer Epitope Prediction. *Immunity* **51**, 766-779 (2019).

19. Editorial. The problem with neoantigen prediction. *Nat. Biotechnol*. **35**, 2, (2017)

20. Zhao, W. & Sher, X. Systematically benchmarking peptide-MHC binding predictors: From synthetic to naturally processed epitopes. *PLoS Comput Biol*. **14**, (2018).

21. Nielsen, M., Lund, O., Buus, S., & Lundegaard, C. MHC Class II epitope predictive algorithms. *Immunology*. **130**, 319-328 (2010).

22. Lin, H. H., Zhang, G. L., Tongchusak, S., Reinherz, E. L., & Brusic, V. Evaluation of MHC-II peptide binding prediction servers: applications for vaccine research. *BMC Bioinformatics*. **9**, S12:S22 (2008).

23. Andreatta, M. *et al*. An automated benchmarking platform for MHC class II binding prediction methods. *Bioinformatics* **34**, 1522-1528 (2018).

24. Wang, P. *et al.* A systematic assessment of MHC class II peptide binding predictions and evaluation of a consensus approach. *PLoS Comput. Biol.* **4**, e1000048 (2008).

25. Vita, R. *et al.* The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Res.* **47**, D339-D343 (2019).

26. Alvarez, B., Barra, C., Nielsen, M., & Andreatta, M. Computational tools for the identification and interpretation of sequence motifs in immunopeptidomes. *Proteomics* **18**, (2018).

27. Stern, L. J. *et al*. Crystal structure of the human class II MHC protein HLA-DR1 complexed with an influenza virus peptide. *Nature*. **368**, 215-221 (1994).

28. Jones, E. Y., Fugger, L., Strominger, J. L., Siebold, C. MHC class II proteins and disease: a structural perspective. *Nat. Rev. Immunol*. **6**, 271-282 (2006).

29. Nielson, M. & Lund, O. NN-align: An artificial neural network-based alignment algorithm for MHC class II peptide binding prediction. *BMC Bioinformatics*. **10**, 296 (2009).

30. Rammensee, H., Bachmann, J., Emmerich, N. P., Bachor, O. A., & Stevanović, S. SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics.* **50**, 213-219 (1999).

31. Birnbaum, M. E. *et al.* Deconstructing the peptide-MHC specificity of T cell recognition. *Cell* **157**, 1073-87 (2014).

32. Birnbaum, M. E., Mendoza, J., Bethune, M., Baltimore, D. and Garcia, K. C. Ligand discovery for T cell receptors. US20170192011A1. (2017).

33. Roche, P. A. & Furuta, K. The ins and outs of MHC class II-mediated antigen processing and presentation. *Nat. Rev. Immunol.* **15**, 203-216 (2015).

34. Hennecke, J. & Wiley, D. C. Structure of a complex of the human alpha/beta T cell receptor (TCR) HA1.7, influenza hemagglutinin peptide, and major histocompatibility complex class II molecule, HLA-DR4 (DRA*0101 and DRB1*0401): insight into TCR cross-restriction and alloreactivity. *J. Exp. Med.* **195**, 571-581 (2002).

35. Fridkis-Hareli, M. & Strominger, J. L. Promiscuous binding of synthetic copolymer 1 to purified HLA-DR molecules. *J. Immunol.* **190**, 4386-4397 (1998).

36. Rosloniec, E. F., Whittington, K. B., Zaller, D. M., & Kang, A. H. HLA-DR1 (DRB1*0101) and DR4 (DRB1*0401) use the same anchor residues for binding an immunodominant peptide derived from human type II collagen. **168**, 253-259 (2002).

37. Dessen, A., Lawrence, C. M., Cupo, S., Zaller, D. M., & Wiley, D. C. X-ray crystal structure of HLA-DR4 (DRA*0101, DRB1*0401) complexed with a peptide from human collagen II. *Immunity* **7**, 473-81 (1997).

38. Fugger, K., Rothbard, J. B., & Sonderstrup-McDevitt, G. Specificity of an HLA-DRB1*0401-restricted T cell response to type II collagen. *J. Immunol.* **26**, 928-933 (1996).

39. Bolin, D. R. *et al*. Peptide and peptide mimetic inhibitors of antigen presentation by HLA-DR class II MHC molecules. Design, structure−activity relationships, and x-ray crystal structures. *J. Med. Chem.* **43**, 2135-2148 (2000).

40. Hammer, J. *et al*. High-affinity binding of short peptides to major histocompatibility complex class II molecules by anchor combinations. *Proc. Natl. Acad. Sci.* **91**, 4456-4460 (1994).

41. Sette, A. *et al.* HLA DR4w4-binding motifs illustrate the biochemical basis of degeneracy and specificity in peptide-DR interactions. *J. Immunol.* **151**, 3163-3170 (1993).

42. Hammer, J. *et al*. Promiscuous and allele-specific anchors in HLA-DR-binding peptides. *Cell.* **74**, 197-203 (1993).

43. Scally, S. W. *et al.* A molecular basis for the association of the HLA-DRB1 locus, citrullination, and rheumatoid arthritis. *J. Exp. Med.* **210**, 2569-82 (2013).

44. Hammer, J. *et al.* Peptide binding specificity of HLA-DR4 molecules: correlation with rheumatoid arthritis association. *J. Exp. Med.* **181**, 1847-1855 (1995).

45. Southwood, S. *et al*. Several common HLA-DR types share largely overlapping peptide binding repertoires. J. Immunol. **160**, 3363-3370 (1998).

46. Reinherz, E. L. *et al*. The crystal structure of a T cell receptor in complex with peptide and MHC class II. *Science*. **286**, 1913-1921 (1999).

47. Yin, L. *et al*. Susceptibility to HLA-DM protein is determined by a dynamic conformation of Major Histocompatibility Complex class II molecule bound with peptide. *J. Bio. Chem.* **289**, 23449-23464 (2014).

48. Sturniolo, T. *et al*. Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices. *Nat. Biotechol*. **17**, 555-561 (1999).

49. Fleri, W. *et al*. The Immune Epitope Database and analysis resource in epitope discovery and synthetic vaccine design. *Front. Immunol*. **8**, 278. (2017).

50. Yin, L. & Stern, L. J. Measurement of peptide binding to MHC class II molecules by fluorescence polarization. *Curr. Protoc. Immunol.* **106**, 5.10.1–5.10.12. (2014).

51. O'Brien, C., Flower, D. R., & Feighery, C. Peptide length significantly influences in vitro affinity for MHC class II molecules. *Immunome Res.* **4**, 6 (2008).

52. Zavala-Ruiz, Z., Strug, I., Anderson, M. W., Gorski, J., & Stern, L. J. A polymorphic pocket at the P10 position contributes to peptide binding specificity in class II MHC proteins. *Chem. Biol.* **11**, 1395-1402 (2004).

53. Lovitch, S. B., Pu, Z., & Unanue, E. R. Amino-terminal flanking residues determine the conformation of a peptide-class II MHC complex. *J. Immunol*. **176**, 2958-2968 (2006).

54. Andreatta, M., Alvarez, B., & Nielsen, M. GibbsCluster: unsupervised clustering and alignment of peptide sequences. *Nucleic Acids Res.* **45**, W458-463 (2017).

55. Veldman, C. M. *et al*. T Cell Recognition of Desmoglein 3 Peptides in Patients with Pemphigus Vulgaris and Healthy Individuals. *J. Immunol.* **172**, 3883-3892 (2004).

56. Wucherpfennig, K. W. *et al.* Structural basis for major histocompatibility complex (MHC)-linked susceptibility to autoimmunity: charged residues of a single MHC binding pocket confer selective presentation of self-peptides in pemphigus vulgaris. *Proc. Natl. Acad. Sci.* **92** 11935-11939 (1995).

57. Kirschmann, D. A. *et al*. Naturally processed peptides from rheumatoid arthritis associated and non-associated HLA-DR alleles. *J. Immunol.* **155**, 5655-5682 (1995).

58. Freide, T. *et al*. Natural ligand motifs of closely related HLA-DR4 molecules predict features of rheumatoid arthritis associated peptides. *Biochim. Biophys. Acta*. **1316**, 85-101 (1996).

59. Patil, N. S. *et al*. Rheumatoid arthritis (RA)-associated HLA-DR alleles form less stable complexes with class II-associated invariant chain peptide than non-RA-associated HLA-DR alleles. *J. Immunol*. (2001).

60. Woulfe, S. L. *et al*. Negatively charged residues interacting with the p4 pocket confer binding specificity to DRB1*0401. *Arthritis Rheum.* **38**, 1744-1753 (1995).

61. Fu, X. T. *et al*. Pocket 4 of the HLA-DR(α,β 1*0401) molecule is a major determinant of T cells recognition of peptide. *J. Exp. Med*. **181**, 915-926 (1995).

62. Busch, R., Hill, C. M., Hayball, J. D., Lamb, J. R., & Rothbard, J. B. Effect of natural polymorphism at residue 86 of the HLA-DR beta chain on peptide binding *J. Immunol*. **147**, 1292-1298 (1991).

63. Nielsen, M. & Andreatta, M. NNAlign: a platform to construct and evaluate artificial neural network models of receptor-ligand interactions. *Nucleic Acids Res*. **45**, W344-349 (2017).

64. Heyder, T. *et al.* Approach for identifying human leukocyte antigen (HLA)-DR bound peptides from scarce clinical samples. *Mol. Cell. Proteomics*. **15**, 3017-29 (2016).

65. Cai, W. *et al*. MHC class II restricted neoantigen peptides predicted by clonal mutation analysis in lung adenocarcinoma patients: implications on prognostic immunological biomarker and vaccine design. *BMC Genomics* **19**, 582 (2018).

66. Barra, C. *et al*. Footprints of antigen processing boost MHC class II natural ligand predictions. *Genome Med.* **10**, 84 (2018).

67. Chao, G. *et al.* Isolating and engineering human antibodies using yeast surface display. *Nature Protocols* **1**, 755–768 (2006).

68. Van Deventer, J.A., Kelly, R.L., Rajan, S., Wittrup, K.D., & Sidhu, S.S. A switchable yeast display/secretion system. *Protein Eng Des Sel*. **28**, 317-325 (2015).

69. Magoč, T. & Salzberg, S. L. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*. **27**, 2957-2963 (2011).

70. Christiansen, A. *et al*. High-throughput sequencing enhanced phage display enables the identification of patient-specific epitope motifs in serum. *Sci. Rep.* **5**, 12913 (2015).

71. Wernersson, R. Virtual ribosome - a comprehensive DNA translation tool with support for integration of sequence feature annotation. *Nucleic Acids Res.* **34**, W385-W385 (2006).

72. Wu, X., & Bartel, D.P. kpLogo: positional k-mer analysis reveals hidden specificity in biological sequences. *Nucleic Acids Res*. **45**, W534-538 (2017).

73. Thomsen, M.C.F. & 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.* **40**, W281-W287 (2012).

74. Apweiler, R. *et al*. UniProt: the universal protein knowledgebase. *Nucleic Acids Res*. **32**, D115-D119 (2004).

75. Vacic, V., Iakoucheva, L.M., & Radivojac, P. Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments. *Bioinformatics* **22**, 1536-7 (2006).

CHAPTER 3 - Decoding the origin and antigen discovery of a conserved regulatory T cell response to highly immunogenic murine lung adenocarcinomas

ABSTRACT

Tumor-infiltrating regulatory T (Treg) cells suppress anti-tumor immune responses – facilitating continued outgrowth and treatment resistance – and are associated with poor clinical prognoses across many tumor types, including non-small cell lung cancer (NSCLC). Although the success of checkpoint-blockade inhibitors in a subset of NSCLC patients demonstrates the potential for immune-mediated tumor clearance, T cell responses to these tumors are highly inhibited and successful therapy is largely dependent on elevated tumor immunogenicity. Previous studies in the 'KP' genetically engineering mouse model of NSCLC have identified tumor-infiltrating Tregs as a key driver of this immunosuppression, and demonstrated that depletion or modulation of these Tregs improves anti-tumor immunity, even in the absence of potent immunogens. However, therapies that systemically deplete or modulate Tregs greatly increase the risk for immune-mediated adverse events, prompting increased focus on specifically targeting tumor-infiltrating Treg populations.

In this chapter, we investigate the identity, origin, and antigenic basis of Tregs infiltrating highly immunogenic KP lung adenocarcinomas to facilitate improved targeting of this population. We find that these populations are highly diverse and functionally unique, yet contain prevalent clones with nearly identical T cell receptor (TCR) pairings across mice. These 'public' Tregs appear to be lung-resident, expand synchronously with anti-tumor immune responses, and appear in independent studies of acute lung inflammation. However, their antigenic basis remains uncertain. Nevertheless, these data provide novel insights into the composition of tumor-infiltrating Tregs, and may provide a model system for the study of Treg immunosuppression in a highly treatment-resistant murine model of NSCLC.

3.1 Introduction

The infiltration of regulatory T (Treg) cells into many tumor types is correlated with poor clinical prognoses [1, 2]. Within the tumor microenvironment, Tregs – which limit immune-mediated damage to healthy tissues during inflammation and autoimmunity [3] – suppress anti-tumor responses, facilitating continued outgrowth and treatment resistance [1, 2]. Although the systemic depletion or modulation of Tregs can alleviate immunosuppression, these treatments increase the risk of immune-related adverse events, particularly autoimmune toxicities [1]. As such, specific targeting of tumor-infiltrating Tregs is of great interest for the development of more effective immunotherapies. However, the composition of these populations is still poorly understood, and in particular, their clonal identity, origin, and antigen-specificity are rarely known [4]. Therefore, in order to improve targeting of tumor-infiltrating Tregs for improved cancer immunotherapies, it is essential to understand their clonal identity, origin, and antigen specificity.

One tumor type which exhibits robust immunosuppression and treatment-resistance is non-small cell lung cancer (NSCLC), which despite treatment advances still accounts for ~25% of all cancer-related deaths in the U.S., and has an 80% 5-year mortality rate [5]. Although the success of immune-checkpoint inhibitors in a subset of NSCLC patients demonstrates the presence of T cells responses against these tumors, immunosuppression greatly inhibits their function [6, 7]. Furthermore, successful therapy is largely limited to patients with high densities of non-synonymous coding mutations – which give rise to potently immunogenic cancer 'neoantigens' [8, 9] – leaving an unmet clinical need for NSCLC patients with lesser mutational burdens and tumor immunogenicity, such as never smokers [8].

The use of genetically engineered mouse models (GEMMs) has greatly informed our understanding of the tumor biology and treatment-resistance of NSCLC [10]. In particular, the 'KP' (*Kras*$^{\text{LSL-G12D/+}}$, *Trp53*$^{\text{fl/fl}}$) model of Kras-driven lung adenocarcinomas (which account for ~25% of NSCLCs [11]) – wherein intranasal administration of a Cre recombinase-encoding virus initiates multi-focal autochthonous tumorigenesis in the presence of a fully functional immune system – mirrors the progression, poor immunogenicity, and robust immunosuppression of the human disease [10, 12-15]. Furthermore, 'LucOS' (firefly luciferase fused to the potent OT-I, OT-II, and 2C T cell epitopes)-encoding variants of these tumors display potent immunogenicity, yet remain treatment-resistance due to robust immunosuppression [12, 13]. Previously, it has been demonstrated that Treg infiltrates drive this immunosuppression, and that their systemic depletion releases potent anti-tumor responses, even in the absence of defined antigens [13]. Furthermore, these tumor-infiltrating Tregs are functionally diverse, but differentiate towards an effector phenotype, and modulation of their signaling bolsters anti-tumor immunity [16]. Yet while these findings suggest that tumor-infiltrating Tregs are a key deterrent of immune-mediated tumor clearance in these lung adenocarcinomas, little is known about their clonal composition or recognition.

Here, we study the identity, origin, and antigenic basis of Tregs infiltrating 'KP' lung adenocarcinomas by investigating their T cell receptor (TCR) repertoires. We discover that KP.LucOS-infiltrating Tregs are highly diverse and distinct from matched conventional CD4^{+} (Tconv) populations or healthy lung-resident Tregs, yet converge on two nearly identical TCRs that are highly conserved between mice. Our findings suggest that these 'public' Tregs are lung-resident and expand to suppress the anti-tumor response directed against highly immunogenic

KP.LucOS tumors, but their antigen specificity remains unknown. We believe that the description of these highly conserved public Tregs provides novel insights into the composition of this key immunosuppressive population, and may provide a defined model system of Treg immunosuppression in highly treatment-resistant lung adenocarcinomas.

3.2 Results

3.2.1 Highly conserved public Tregs emerge in highly immunogenic lung adenocarcinomas

The defined-antigen autochthonous KP.LucOS murine lung adenocarcinoma model has previously been shown to elicit robust anti-tumor immune responses [12, 13]. Following intra-tracheal delivery of a lentivirus encoding Cre recombinase and the LucOS fusion protein [12, 17], Cre recombinase causes spontaneous loss of p53 and expression of oncogenic Kras^{G12D} mutant protein, resulting in multi-focal tumorigenesis and progression to stark adenocarcinoma, mimicking the progression of Kras-driven NSCLC in humans [17, 10]. Although these tumors have a very low burden of non-synonymous protein coding mutations [14] – and therefore very few cancer 'neoantigens' – they are highly immunogenic owing to the expression of defined T cell antigens, with large infiltrates of T and B cells within 4 weeks of tumor initiation [12]. However, these immune responses become highly suppressed and these tumors display high frequencies of infiltrating Tregs [13]. Previous studies have shown that these tumor-infiltrating Tregs are critical to tumor immunosuppression, as their systemic depletion or modulation bolsters anti-tumor immunity and delays tumor progression [13, 15, 16], yet little is known about their composition. We therefore sought to investigate the identity, origin, and antigenic basis of Tregs within the highly immunogenic KP.LucOS lung adenocarcinoma model in order to better understand their role in tumor immunosuppression.

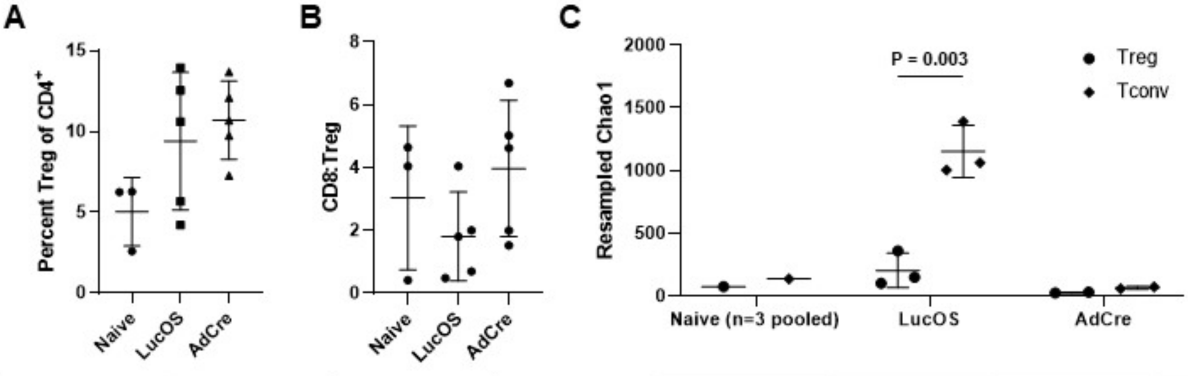


Figure 3.1. *Highly immunogenic LucOS lung adenocarcinoma induce large and diverse Treg response. A, B) Percent of regulatory T cells (Tregs) among all lung-resident CD4+ T cells (A) or ratio of lung-resident CD8+ T cells to Tregs (B) from naïve mice or from mice bearing late-stage LucOS- or AdCre lung adenocarcinomas. C) Mean resampled Chao1 diversity of pooled lung-resident regulatory (Treg) or conventional (Tconv) CD4+ T cells from T cell receptor (TCR) beta chain amplicon sequencing. Asterisk denotes value from n=3 samples pooled prior to sequencing.*

To facilitate this analysis, we also analyzed Tregs and conventional CD4$^+$ T cells (Tconvs), from LucOS tumor-bearing mice, as well as from naïve mice and mice bearing AdCre lung adenocarcinomas. In contrast to LucOS tumors, AdCre tumors are established by transient infection with an adenovirus encoding Cre recombinase but no defined antigens [17], and therefore express no foreign antigens following tumorigenesis and are poorly immunogenic [15]. In line with previous reports, we found that the lungs of mice bearing late-stage (week 16-20 post-initiation) LucOS and AdCre tumors had elevated fractions of regulatory T cells among lung-resident CD4$^+$ T cells relative to naïve mice (Figure 3.1A). However, only LucOS tumor-bearing mice displayed decreased ratios of CD8$^+$ T cell to Treg infiltration (Figure 3.1B) indicative of strong immunosuppression.

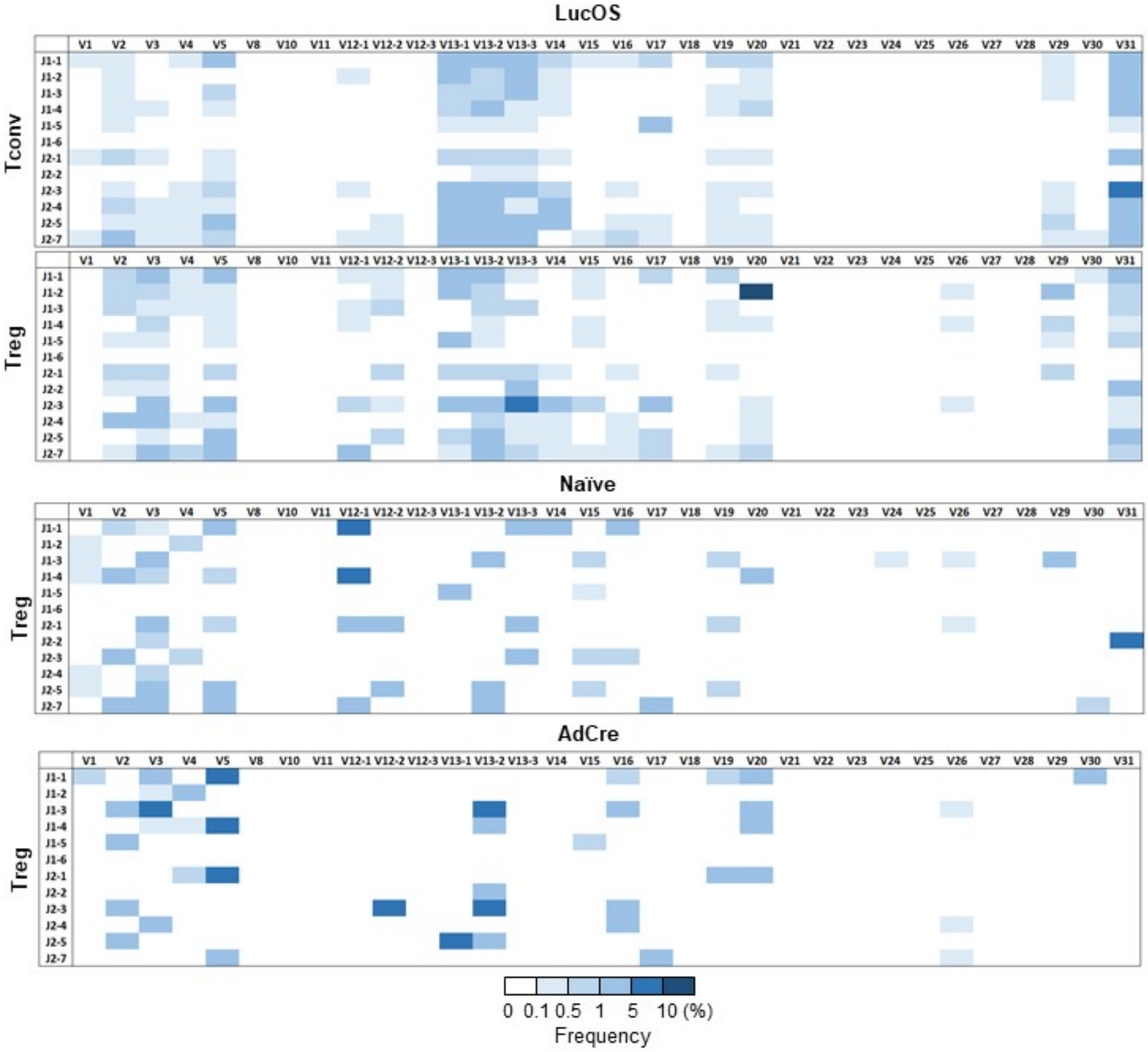


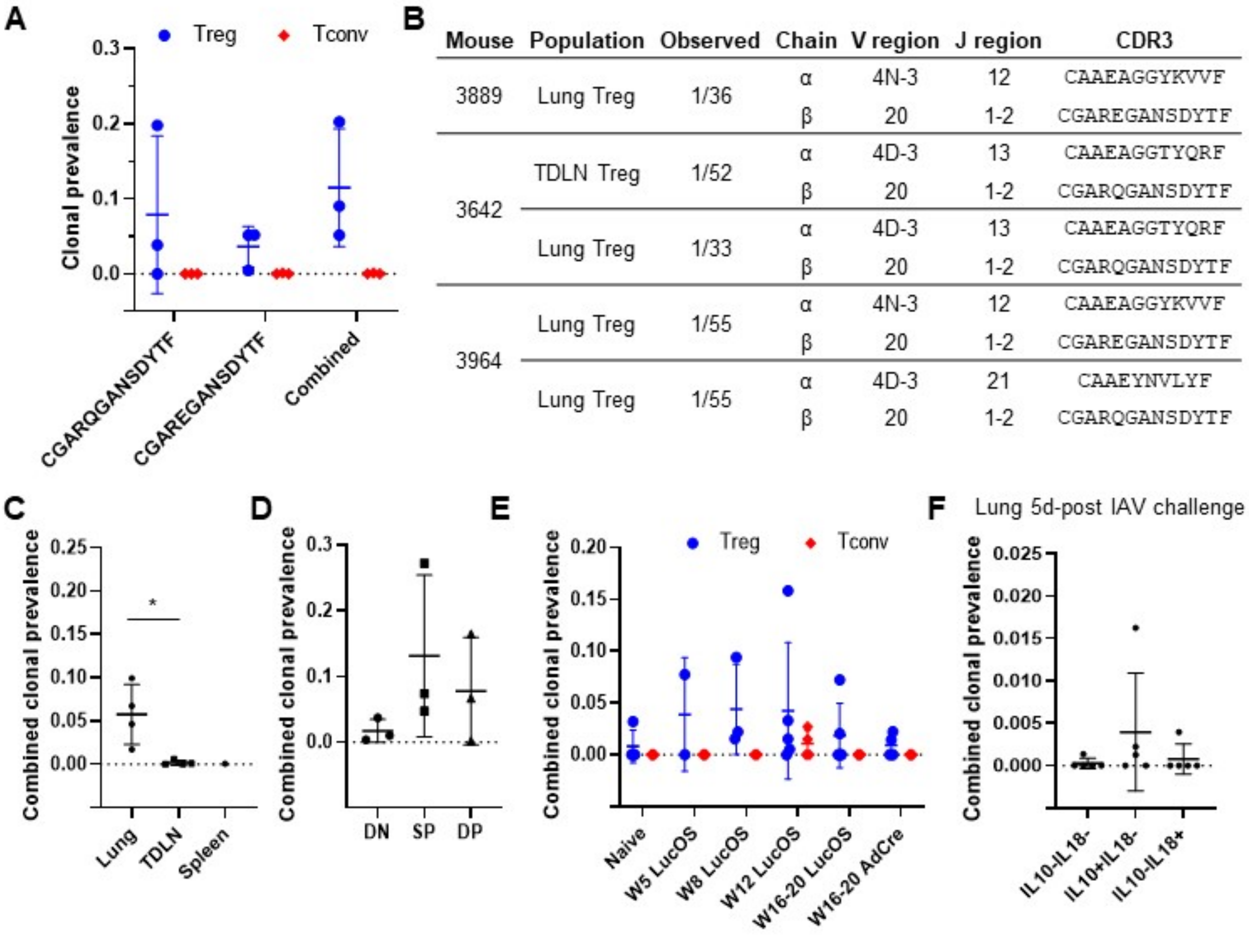
Figure 3.2. *Regulatory T cell repertoire converges on common motif during the immune response to a highly immunogenic lung adenocarcinoma. Percent frequency of V and J region pairings from TCR beta chain amplicon sequencing of pooled lung-resident regulatory T cells from naïve or late-stage LucOS or AdCre tumor-bearing mice, averaged within conditions.*

To investigate their identity, lung-resident Tconvs and Tregs from a subset of these mice were bulk sorted and submitted for targeted T cell receptor (TCR) beta-chain amplicon sequencing. TCR beta chain sequencing is frequently used to probe T cell populations for clonal expansions and can reveal convergence in receptor composition that is indicative of shared antigen recognition.

Analysis revealed that both lung-resident Tconvs and Tregs from LucOS tumor-bearing mice were more diverse than matched populations from AdCre tumor-bearing or naïve mice (Figure 3.1C), consistent with increased immune infiltration. Furthermore, the Tconv response in LucOS tumor-bearing mice was significantly more diverse than the matched Treg response, and overlap between these two repertoires was minimal (average F2 = 0.033 ± 0.007), in line with previous reports [4]. Notably, despite the overall diversity of the LucOS-infiltrating Treg response, we observed overlap in Treg populations between mice (average F2 = 0.066 ± 0.025) and strong convergence (> 10% of reads) upon a single TRBV and TRBJ pairing (TRBV20/TRBJ1-2) across mice (Figure 1D). This convergence was not observed in the matched Tconv population (Figure 3.2), or in the Treg populations of naïve or AdCre tumor-bearing mice, suggesting a unique antigenic basis.

Analysis of this convergent TRBV/TRBJ pairing revealed that it was largely comprised of two closely related TCR beta (TRB) chain sequences that were highly conserved across mice. The two beta chains share near identical CDR3β sequences (CGARQGANSDYTF and CGAREGANSDYTF) and at least one was observed in each LucOS-infiltrating Treg population (Figure 3.3A). While these sequences were also observed in the matched Tconv populations, their prevalence was diminished by orders of magnitude, suggesting that this Treg population is not derived from Tconvs (*i.e.* induced Tregs). Neither sequence was observed in the lung-resident Treg or Tconv populations of the n = 3 naïve or n = 2 AdCre tumor-bearing populations submitted for beta-chain sequencing. Single-cell sequencing further revealed that these T cells share highly similar TCR alpha (TRA) chains that utilize closely related V regions TRAV 4D-3 or 4N-3, and share common CDR3α motifs (Figure 3.3B). These T cells therefore represent a highly conserved 'public' regulatory T cell response to KP.LucOS lung adenocarcinomas, and are likely to have a common origin and antigenic basis [18].

To investigate this origin and antigenic basis, we analyzed previously published bulk RNA-seq datasets of Tregs from the lungs, tumor-draining lymph nodes (TDLN), and spleen of late-stage LucOS tumor-bearing mice [16]. We observed substantially increased prevalence of the conserved TRB sequences in the lung-infiltrating populations, relative to the TDLN or spleen (Figure 3.3C), supporting a lung-specific antigenic basis, as opposed to a systemic response to inflammation. Furthermore, these conserved sequences were found more frequency in previously described [16] tissue-resident CD103 single-positive (SP), and tissue-resident effector CD103 KLRG1 double-positive (DP), Treg subpopulations within the lung (Figure 3.3D), further supporting lung-specific function. In addition, analysis of Treg and Tconv RNA-seq datasets collected at multiple time points during LucOS tumor progression revealed that this Treg population expands as early as 5 weeks post tumor-initiation and appears to peak in prevalence around week 8 before collapsing in late-stage tumors (Figure 3.3E), matching the dynamics of the anti-tumor immune response, which peaks between weeks 4-8 [12]. Importantly, these sequences were only observed in Tconv samples at 12 weeks post tumor initiation, following the peak immune response, further supporting a Tconv-independent origin for this conserved Treg response.



Mouse	Population	Observed	Chain	V region	J region	CDR3
3889	Lung Treg	1/36	α	4N-3	12	CAAEAGGYKVVF
			β	20	1-2	CGAREGANSDYTF
3642	TDLN Treg	1/52	α	4D-3	13	CAAEAGGTYQRF
			β	20	1-2	CGARQGANSDYTF
	Lung Treg	1/33	α	4D-3	13	CAAEAGGTYQRF
			β	20	1-2	CGARQGANSDYTF
3964	Lung Treg	1/55	α	4N-3	12	CAAEAGGYKVVF
			β	20	1-2	CGAREGANSDYTF
	Lung Treg	1/55	α	4D-3	21	CAAEYNVLYF
			β	20	1-2	CGARQGANSDYTF

Figure 3.3. *Highly conserved public regulatory T cells arise specifically in lungs during peak immune response. A) Prevalence of two highly related public regulatory T cell beta chains, and combined prevalence, as a fraction of all TRB-associated reads in pooled TCR beta-chain amplicon sequencing from tumor-infiltrating Treg and Tconv populations in late-stage LucOS tumor-bearing mice. B) Paired-chain TCR sequences identified from single cell RNA-seq of lung- or tumor-draining lymph node (TDLN)-resident CD4$^+$ Tregs from KP.LucOS tumor-bearing mice. C) Combined prevalence of the conserved TRB chains, as a fraction of all TCR-associated reads in RNA-seq of pooled Tregs from the lung, TDLN, or spleen of late-state LucOS tumor-bearing mice. Asterisk denotes statistical significance from two-sided paired t test at $p < 0.05$. D) Combined prevalence of the conserved TRB chains, as a fraction of all TCR-associated reads in RNA-seq of CD103$^-$ KLRG1$^-$ (DN), CD103$^+$ KLRG1$^-$ (SP), or CD103$^+$ KLRG1$^+$ (DP) Treg subpopulations from late-stage LucOS tumor-bearing mice. E) Combined prevalence of the conserved TRB and paired TRA sequences in RNA-seq, as a fraction of all TCR associated reads, of pooled lung-resident Treg or Tconv cells from various time points in LucOS or AdCre tumor development. E) Combined prevalence of conserved TRB chains, as a fraction of all TRB-associated reads from three subpopulations of lung-resident Tregs isolated 5d post Influenza A virus (IAV) challenge.*

Lastly, these conserved TRB sequences were present at low levels in 1/4 naïve and 2/4 late-stage AdCre tumor-bearing mice from this dataset in lung-resident Treg populations. This finding suggests that these public Tregs may be present as low frequencies in lung-resident Treg populations (possibly explaining their absence in beta chain sequencing datasets), but may preferentially expand in response to conditions unique to LucOS lung adenocarcinomas. These unique conditions may be the expression of foreign antigens encoded by the LucOS tumors, or conditions secondary to the vigorous anti-tumor immune response. In support of the latter origin, we observe these same conserved TRB sequences in an independent RNA-seq dataset [19] of lung-resident Tregs collected 5 days post-infection with Influenza A virus at the onset of the anti-viral immune response (Figure 3.3F). Within this dataset, these sequences were observed preferentially in an IL-10 single-positive Treg subpopulation that was differentiated by TCR-dependent IL-10 secretion and a tissue-resident effector phenotype.

Together, these data suggest that this conserved public Tregs population are derived from a lung-resident population that expand to suppress lung-specific inflammation secondary to the vigorous immune response directed against KP.LucOS lung adenocarcinomas. However, we sought to understand the antigenic basis of their localization and expansion to better understand their origin and function within the tumor microenvironment.

3.4.2 Attempts to define antigenic basis of conserved public regulatory T cell response in vitro

To discover antigens for these TCRs, we turned to yeast-displayed pMHC libraries, which we have previously used to discover *bona fide* stimulatory antigens for class II MHC-restricted murine $CD4^+$ T cells [20]. To this end, we created a single-chain yeast-displayed pMHC platform design of I-A^b (H2-Ab), the sole class II molecule expressed in C57BL/6 mice. Analogous to the previously described murine class II molecule I-E^k [20], this design expresses a peptide, the β1 and α1 domains of I-A^b (which encode the MHC peptide-binding groove), and a FLAG epitope tag, connected by flexible Gly-Ser linkers, as an N-terminal fusion to Aga2, facilitating surface display through its linkage to the yeast-surface protein Aga1 (Figure 3.4A). To validate the expression and fold of this construct, we expressed a p1Y variant of the 3K peptide and stained with fluorescently-labeled tetramers of B3K506 TCR, which is known to bind I-A^b/3K [21]. The p1Y variant of 3K was used to improve predicted peptide binding without alteration of TCR recognition.

As is typical of this system, this original construct did not display binding to its cognate TCR (data not shown). Therefore, to rescue the fold and function of this construct, we subjected its MHC-encoding domains to error-prone PCR to construct a yeast-displayed library of 10^8 unique I-A^b/3K variants (Figure 3.4A). This library was then subjected to sequential rounds of affinity-based selection B3K506-coated magnetic beads to enrich a population of constructs with gain-of-function mutations. Single-cell analysis of this enriched population yielded a tetramer-positive clone with 3 alpha chain mutations, α63D, α67A, α75I (Figure 3.4B). However, α63 and α67 lie within the peptide-binding groove of I-A^b, and are therefore undesirable for faithful reproduction of its native peptide-binding preferences. These mutations were therefore reverted to wild-type and the α75I mutation was combined with two frequently observed beta chain mutations, β28F and β31, which reduce the hydrophobicity in a previously solvent-inaccessible region of the native MHC. Although this construct failed to display tetramer binding, further mutation of three additional small hydrophobic residues within this region rescued tetramer binding (Figure 3.4B). This final construct design also bound strongly to another known I-A^b/3K-restricted TCR, YAe62 [22] (Figure 3.4C), demonstrating faithful and generalizable reproduction of I-A^b recognition.

This platform was then used to construct a yeast-displayed library of randomized MHC-linked peptides to identify antigenic peptides for the conserved Treg TCRs. Using degenerate primers, we constructed a yeast-displayed library of 10^8 peptide variants linked to I-A^b (Figure 3.5A). This library was then subjected to sequential rounds of selection with magnetic beads coated with recombinantly expressed CGR7 and CGR8 TCRs, which were reconstructed from single-cell sequencing data (Figure 3.5B). However, after three rounds of affinity-based selection we did not observe notable enrichment of a TCR-binding yeast population for either TCR, and sequencing of selected clones revealed peptides that did not display convergence but displayed signs of non-specific binding, such as up-facing cysteine and tryptophan residues (Figure 3.5C).

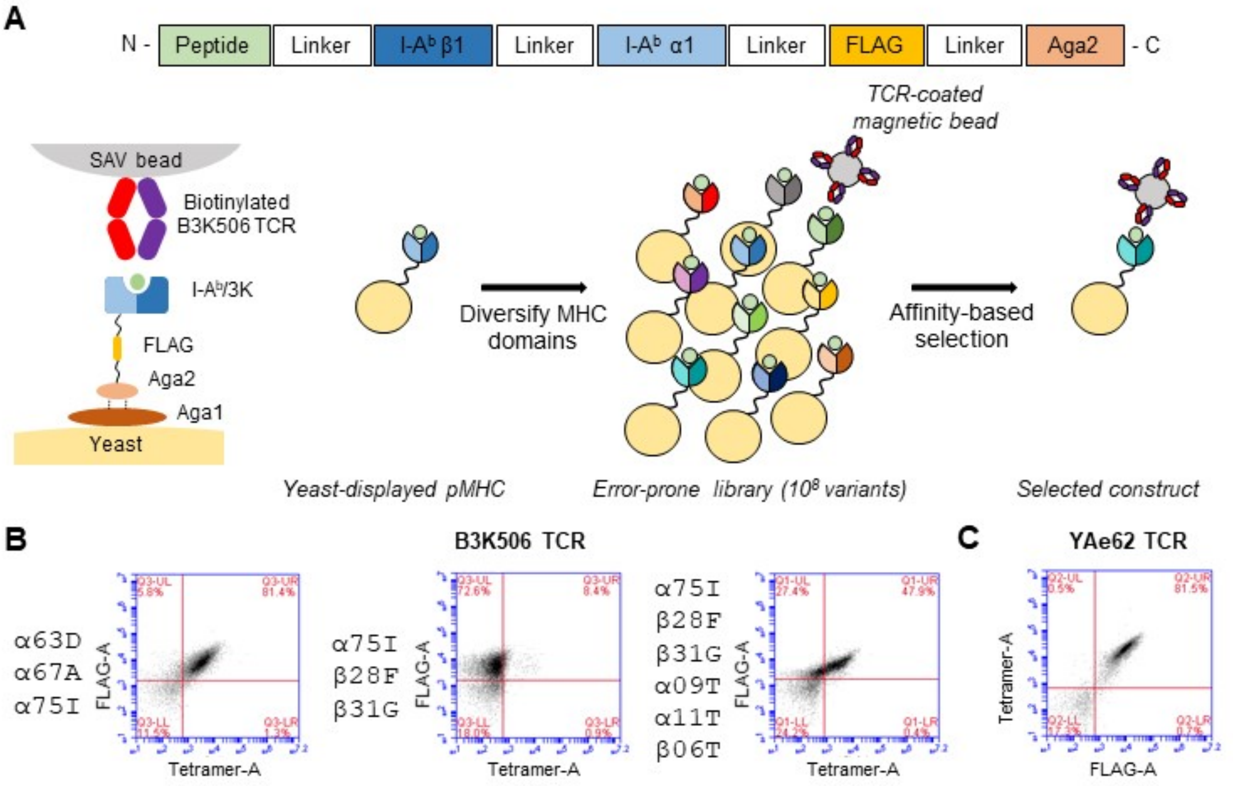
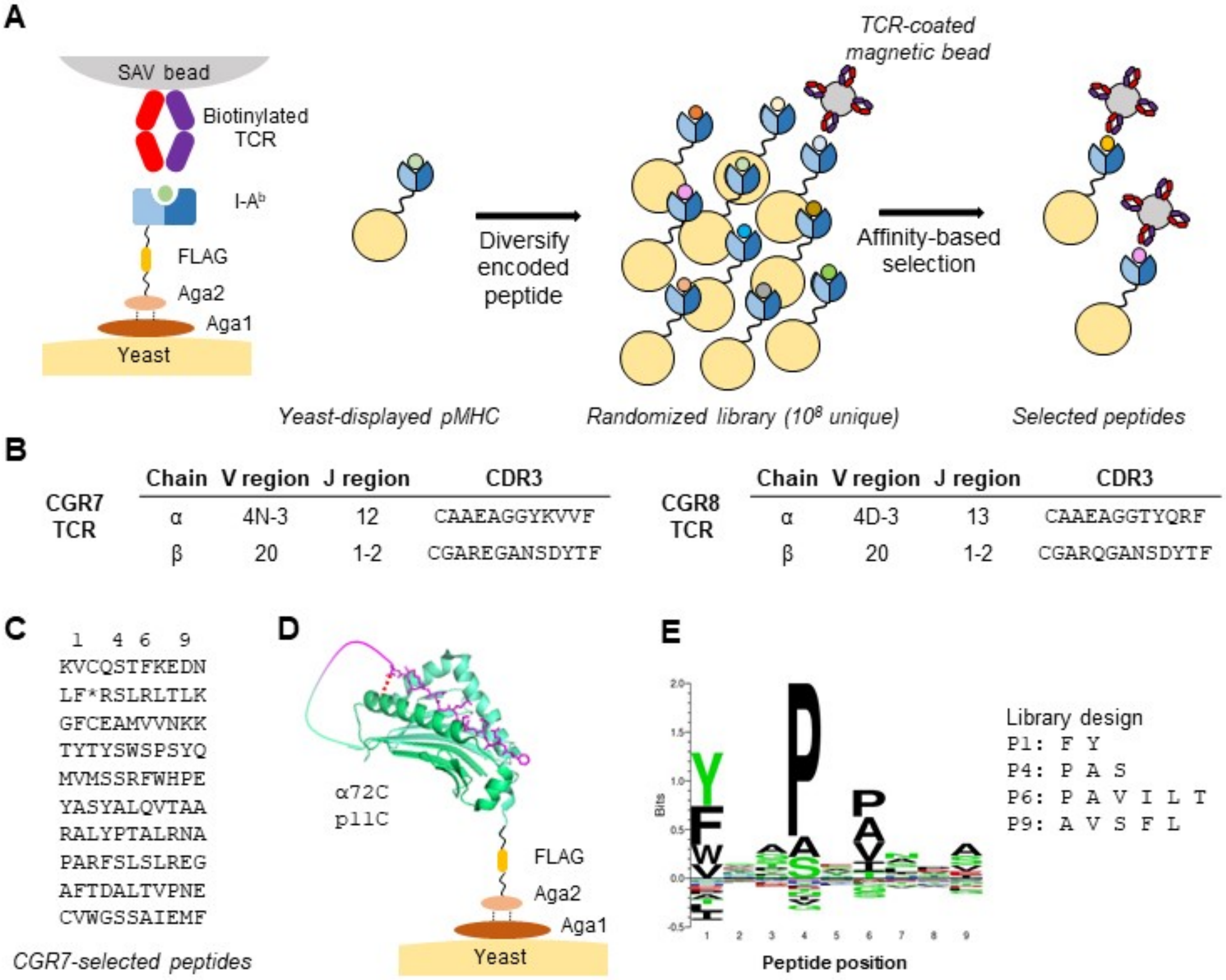


Figure 3.4. *Design and validation of a yeast-displayed I-A^b peptide-MHC (pMHC) platform. A) Schematics of the design of a single-chain yeast-displayed murine I-A^b/3K pMHC construct, and library-based identification of MHC mutations facilitating successful expression and fold. B,C) Flow cytometry analysis of yeast expressing I-A^b/3K pMHC variant constructs with fluorescently-labeled tetramers of B3K506 (B) or YAe62 (C) cognate TCRs and anti-FLAG epitope tag antibody, with associated MHC mutations.*

As murine H2-A alleles have previously shown weak peptide binding [23], we attempted to rescue binding by increasing the apparent MHC-binding affinity of library encoded peptides or increasing the fraction of MHC-binding peptides within the library. First, we engineered a peptide 'disulfide trap' by modifying peptide position p11 and neighboring MHC position α72 to cysteines (Figure 3.5C). This technique has previously been used to study low-affinity peptides on other murine class II MHC molecules [23] and moderately improved B3K506 tetramer staining of I-A^b/3K construct (data not shown). However, yeast-displayed libraries built upon this design again failed to show enrichment. In addition, to increase the proportion of library-encoded peptides which bind to I-A^b (and therefore increase the likelihood of a TCR-binding peptides), we limited diversity at MHC-facing anchor positions (P1, P4, P6, and P9) to preferred residues derived from clustering of eluted ligand mass-spectrometry [24] (Figure 3.5E). However, these this library also failed to display enrichment when selected with CGR7 and CGR8 TCRs.



CGR7 TCR	Chain	V region	J region	CDR3
	α	4N-3	12	CAAEAGGYKVVF
	β	20	1-2	CGAREGANSDYTF

CGR8 TCR	Chain	V region	J region	CDR3
	α	4D-3	13	CAAEAGGTYQRF
	β	20	1-2	CGARQGANSDYTF

```
 1  4 6  9
KVCQSTFKEDN
LF*RSLRLTLK
GFCEAMVVNKK
TYTYSWSPSYQ
MVMSSRFWHPE
YASYALQVTAA
RALYPTALRNA
PARFSLSLREG
AFTDALTVPNE
CVWGSSAIEMF
```

CGR7-selected peptides

Library design
P1: F Y
P4: P A S
P6: P A V I L T
P9: A V S F L

Figure 3.5. *Investigating the antigenic basis of a conserved Treg response with yeast-displayed peptide-MHC (pMHC) libraries. A) Schematic of the design and sequential affinity-based selection of a randomized yeast-displayed peptide-MHC (pMHC) library with TCR-coated magnetic beads. B) Composition of CGR7/8 conserved Treg TCRs derived from single-cell sequencing and used for library screening. C) Sequential rounds of selection with CGR7 TCR-coated yielded no convergence in the linked peptide among selected clones from round 3 of selection. Select peptide positions are labeled. D) Structural representation (adapted from PDB ID: 3C5Z) of 'disulfide-trap' construct design, showing engineered cysteine linkage between peptide position p11 and MHC alpha chain position α72. D) Core 9mer of I-A^b-binding peptides, as determined by clustering of mono-allelic eluted ligand mass spectrometry data, with associated library design to enforce observed peptide 'anchor' positions.*

3.4.3 Attempts to define antigenic basis of conserved public regulatory T cell response in vivo

The absence of observed antigen reactivity from our large yeast-displayed libraries suggests the possibility that these conserved Treg-derived TCRs may not be restricted to the class II molecule I-A^b, or may have very low affinity for their cognate antigen, thereby confounding our affinity-based selections. Therefore, to study the reactivity of these TCRs in the context of native and complete antigen presentation, we studied the localization and expansion of TCR-transduced primary T cells following adoptive transfer. To this end, a mixed CD4$^+$ and CD8$^+$ population of primary splenic T cells from wild-type hosts were transduced with retroviruses encoding CGR7 or B3K506 TCR. Successfully transduced T cells were sorted through their co-expression of EGFP fluorescent protein and adoptively transferred into wild-type or *Rag2*$^{-/-}$ hosts (which lack B cells and T cells), and monitored their organ-specific accumulation after 10 days. Although we observed little to no accumulation of the transduced T cells in wild-type hosts, we observed sizeable accumulation in each *Rag2*$^{-/-}$ mouse (Figure 3.6A), likely due to the lack of niche competition from other T cells. For each mouse, the transduced T cells displayed greatest accumulation in the spleen, but accumulation was indistinguishable between the two TCRs in the spleen, lung, and mediastinal and inguinal lymph nodes. This result suggests that the cognate antigen of the conserved Tregs is either not expressed or not presented within the lungs under normal healthy conditions.

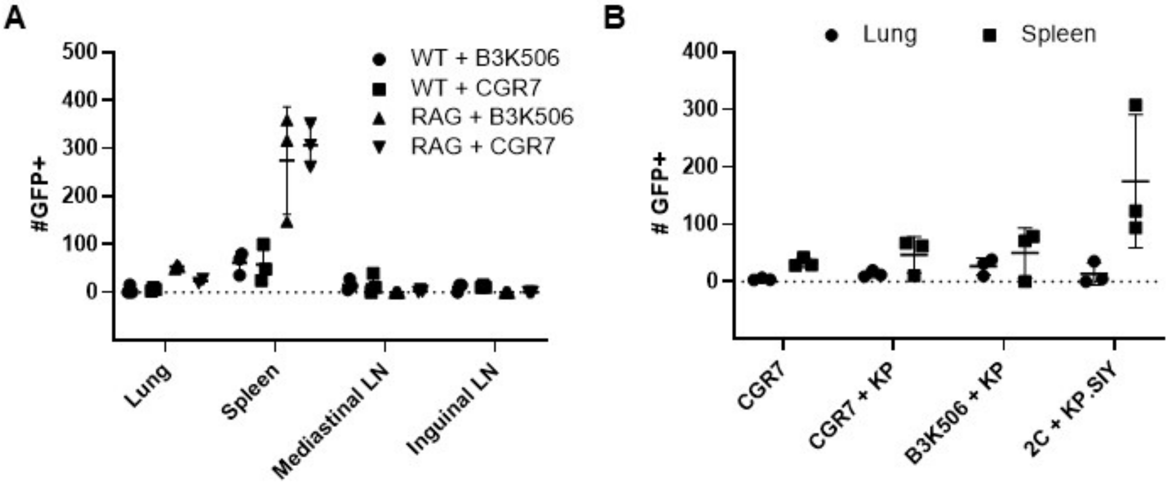


Figure 3.6*. Investigating the antigenic basis of a conserved public Treg response in vivo. Organ-specific accumulation of GFP$^+$ TCR-transduced primary T cells following adoptive transfer into (A) healthy wild-type (WT) or Rag2$^{-/-}$ (RAG) mice or (B) healthy or orthotopic 'KP' tumor-bearing Rag2$^{-/-}$ mice, as assessed by flow cytometry for N = 3 mice per group.*

To explore whether this cognate antigen is expressed or presented in the context of KP lung adenocarcinomas, we then adoptively transferred TCR-transduced primary T cells into *Rag2*$^{-/-}$ hosts bearing orthotopic KP lung tumors. Unlike their autochthonous variants, transplanted orthotopic KP tumors are rapidly established, and are more susceptible to anti-tumor immunity, possibly due to increased T cell priming [12, 15]. As a positive control for T cell tumor reactivity, 2C TCR-transduced T cells – which are specific to the class I MHC-restricted antigen SIY [25] – were adoptively transferred into mice bearing established KP.SIY orthotopic lung tumors. Although we did not observe accumulation of 2C-transduced T cells in the lungs of these mice, they were present in large numbers within the spleen, and the lung tumors were no longer visible at sacrifice (data not shown), consistent with rapid tumor rapid rejection. However, neither CGR7 or B3K506 TCR-transduced T cells displayed comparable accumulation in either the lungs or spleens of mice bearing orthotopic KP lung tumors lacking defined antigens, and accumulation was again indistinguishable between TCRs (Figure 3.6B). This result suggests that the conserved Treg response is not directed directly against KP lung tumors in the absence of defined antigens, nor against a self-antigen that is mis- or over-expressed within the lungs during KP tumor outgrowth. However, this result does not rule out direct reactivity to the foreign antigens encoded by LucOS tumors or to self-antigens uniquely presented in the highly inflammatory environment created by the anti-tumor response against these foreign antigens. Therefore, further experimentation is needed to determine the antigenic basis of this highly conserved Treg response.

3.3 Discussion

Specific targeting of tumor-infiltrating Tregs for deletion or modulation is an attractive therapeutic modality to alleviate tumor immunosuppression and drive anti-tumor immunity without inducing systemic autoimmunity [1]. This modality is especially attractive in poorly immunogenic tumors – such as those with low burdens of non-synonymous coding mutations [8, 9] – as greater alleviation of immunosuppression is required to overcome the barrier for anti-tumor immunity [2]. However, the identity, origin, and antigenic basis of tumor-infiltrating Treg populations are rarely known [4], complicating efforts to specifically target these antigen-specific populations. Here, we investigate the composition of tumor-infiltrating Tregs in the KP murine model of Kras-driven NSCLC that displays robust immunosuppression and resistance to chemo- and immune-therapeutics [12, 13, 15], even when engineered to be highly immunogenic. Within this population, we identify and explore the origin and antigen reactivity of a prevalent and highly conserved public Treg clone that may provide a defined model system for Treg-mediated tumor immunosuppression.

Through population-level TCR beta chain sequencing, we found that the Treg response to the highly immunogenic yet immunosuppressed KP.LucOS lung adenocarcinoma model is both highly diverse and distinct from matched Tconv populations. Although peripherally induced Treg (iTreg) populations have been shown to form from Tconv populations in immunosuppressive environments [26], Treg and Tconv TCR repertoires are largely distinct [27, 28], and previous studies have failed to observe Treg induction in a tumor context [29, 30], suggesting thymically-derived Tregs (tTregs) dominate the tumor-infiltrating Treg response. Importantly, the TCR repertoire of these tTregs reportedly skews towards self-reactivity [1, 4] and previous studies have identified self-reactivity in tumor-infiltrating Tregs [30, 31].

Although the Treg response to KP.LucOS lung adenocarcinomas was highly diverse, there was a convergence on two nearly-identical TCRβ chains, and T cells bearing these beta chains share highly similar paired TCRα chains, suggesting shared antigen recognition. Similar 'public' T cell responses have been noted previously in many human and murine studies [18], and Treg responses to murine tumors have been previously reported to skew towards public responses in a tumor- and antigen-specific manner [32]. Consistent with this notion, these conserved Tregs were observed almost exclusively within the lungs, and were preferentially found in tissue-resident effector populations, both within this model and in the Treg response to Influenza A virus infection. In addition, this conserved Treg population appears to expand and contract with the anti-tumor immune response, and was largely restricted to Treg populations, suggesting a thymic origin for this population. These observations and the purported self-reactivity of tumor-infiltrating Tregs lead us to postulate that these Tregs are reactive to a lung-specific self-antigen and their expansion is driven by the highly inflammatory immune response to the defined antigens in LucOS tumors. However, we were unable to identify antigen reactivity in the context of large yeast-displayed class

II peptide-MHC libraries, or in the lungs of healthy mice or mice bearing KP tumors without defined antigens using adoptive transfer of TCR-transduced T cells.

These results suggest either a weak affinity for their cognate antigen or an unconventional mode of antigen recognition, and further suggest that this antigen is not presented during healthy lung function or during outgrowth of KP tumors without defined antigens. However, as KP tumors without defined antigens are poorly immunogenic [12, 15], these findings do not rule out self-reactivity during acute lung inflammation. Therefore, these results must be supplemented with adoptive transfer in a model of lung inflammation in the absence of tumor, such as Influenza A virus challenge [19], to confirm this hypothesis. In the absence of reactivity in this model, adoptive transfer into mice bearing orthotopic or autochthonous KP.LucOS tumors is warranted, as these Tregs could be cross reactive to the foreign antigens encoded within the LucOS fusion protein. Although we were unable to define the antigen reactivity of this conserved Treg response, our findings provide insight into the identity and origin of tumor-infiltrating Tregs in a highly immunosuppressive and treatment-resistant model of Kras-driven NSCLC, and may provide a model system for future modalities to specific targeting or modulation of tumor-infiltrating Tregs.

3.4 Methods

3.4.1 Animal studies

C57BL/6 mice were used for all experiments. *Kras*$^{\mathrm{LSL\text{-}G12D/+}}$ *Trp53*$^{\mathrm{fl/fl}}$ (KP) *FoxP3*$^{\mathrm{GFP}}$ C57BL/6 mice have previously been described [13], and were used for sorting and sequencing of T cells. Wild-type C57BL/6 mice were used for isolation of primary T cells for peptide restimulation and adoptive transfer experiments. C57BL/6 wild-type or *Rag2*$^{-/-}$ mice were used as hosts for adoptive transfer. Experimental and control mice were co-housed whenever appropriate. All studies were performed under an animal protocol approved by the Massachusetts Institute of Technology (MIT) Committee on Animal Care. Mice were assessed for morbidity according to MIT Division of Comparative Medicine guidelines and humanely sacrificed prior to natural expiration.

For *in vivo* labelling of circulating immune cells, anti-CD4-PE (1:400, eBioscience) and anti-CD8β-PE (1:400, eBioscience), or anti-CD45-PE-CF594 (1:200, BD Biosciences) antibodies were diluted in PBS and administered by intravenous injection 5 minutes before harvest [33]. Labeled cells constitute the IV^{+} population in flow cytometry experiments.

3.4.2 Tumor induction

Lentiviral Lenti-LucOS and adenoviral AdCre vectors have been described previously [12, 17]. Lentiviral plasmids and packaging vectors were prepared using endo-free maxiprep kits (Qiagen). Lentiviruses were produced by co-transfection of 293FS* cells with Lenti-LucOS, psPAX2 (gag/pol), and VSV-G vectors at a 4:2:1 ratio with Mirus TransIT LT1 (Mirus Bio, LLC). Supernatant was collected 48 and 72h after transfection and filtered through 0.45mm filters before concentration by ultracentrifugation (25,000 RPM for 2 hours with low decel). Virus was then resuspended in 1:1 Opti-MEM (Gibco) - HBSS. Aliquots of virus were stored at -80°C and titered using the GreenGo 3TZ cell line. AdCre virus was provided by the University of Iowa Viral Vector core at a titer of 1x10^{12} particles/mL (1 x 10^{10} PFU/mL). For tumor induction, mice between 8-15 weeks of age received 2.5 × 10^4 PFU of Lenti-LucOS or 2.5 x 10^7 PFU of AdCre intratracheally as described previously [17].

For orthotopic KP tumor challenges, tumor lines were previously generated [12] and propagated in DMEM supplemented with 10% FBS, 1% penicillin-streptomycin, 1% non-essential amino acids, and 25 mM HEPES. Cells were trypsinized, washed, counted, and 250,000 were injected intravenously via the tail vein into WT or R*ag2*-/- female mice.

3.4.3 Tissue isolation and preparation of single cell suspensions

After sacrifice, lungs were placed in 2.5mL collagenase/DNAse buffer in gentleMACS C tubes (Miltenyi) and processed using program m_impTumor_01.01. Lungs were then incubated at 37°C for 30 minutes with gentle agitation. The tissue suspension was filtered through a 100 µm cell strainer and centrifuged at 1700 RPM for 10 minutes. Red blood cell lysis was performed by incubation with ACK Lysis Buffer (Life Technologies) for 3 minutes. Samples were filtered and centrifuged again, followed by resuspension in RPMI 1640 (VWR) supplemented with 1% heat-inactivated FBS and 1X penicillin-streptomycin (Gibco), and 1X L-glutamine (Gibco).

Bone marrow-derived dendritic cells were collected from the femurs and tibias of mice by perfusion with PBS. BMDCs were cultured with GM-CSF (40 ng/ml, Biolegend) for 8 days post-isolation in RPMI supplemented with 10% heat-inactivated FBS, 1X penicillin-streptomycin, 25 mM HEPES, and 1% non-essential amino acids (Gibco). Cells were frozen and stored in liquid nitrogen until future use.

Spleens were mechanically disrupted in culture dishes in PBS supplemented with 2% FBS. Cells were passed through a 70 µm cell strainer (Fisher), centrifuged at 300 x g for 10 minutes and resuspended at 1 x 10^8 cells/mL. Primary T cells were isolated from mixed splenic populations using EasySep mouse T cell or CD4 T cell isolation kits (Stem cell technologies), according to the manufacturer's instructions. T cells were expanded in RPMI media supplemented with 10% FBS, 1X penicillin-streptomycin, 50 µM beta-2-mercaptoethanol, and 100 U/mL recombinant human IL-2 (RnD systems) with Dynabeads mouse T cell activator beads (Gibco) at a 1:1 ratio.

3.4.4 Retroviral transduction of T cells and adoptive transfer

Retrovirus MP-71 was used to deliver TCR genes into primary mouse T cells. These plasmids were designed essentially as previously described [34], encoding full length TCR alpha and beta chains, and the fluorescent protein EGFP, connected by self-cleaving P2A sequences. B3K506 and CGR7 TCR constructs were encoded with synthetic gene blocks (Integrated DNA Technologies) and OT-II TCR was generated from the murine TCR OTII-2A.pMIG II plasmid (Plasmid #52112, AddGene). For each virus, TCR plasmid DNA was combined with pCL-Eco retroviral vector and polyethylenimine (PEI) at a 5:3:24 mass ratio, diluted in Opti-MEM, and applied to a confluent layer of HEK 293 cells (ATCC) cultured in DMEM supplemented with 10% FBS, 1X penicillin-streptomycin, and 25 mM HEPES. Media was exchanged after 4 hours and virus-containing supernatant was collected at 48-72h, passed through 0.45 µM filters, and stored in aliquots at -80°C until use.

Each virus aliquot was applied to 2 x 10^6 activated primary murine T cells at 48h post-isolation in 6 well cell culture plates (Corning). Plates were centrifuged for 90 minutes at 1000 x g (with low acceleration and deceleration) to aid transduction. T cells were collected 48-72h post-transduction and bulk sorted for EGFP expression on a MA900 Cell Sorter (Sony) and returned to culture for expansion.

Following expansion, 1 x 10^5 TCR transgenic T cells were injected into the tail vein of the mouse in 100 µL PBS. Mice were sacrificed for analysis 7-14 days later. Following harvest (below), samples were gated on $\mathrm{IV}^{\mathrm{neg}}$ live (Fixable Live/Dead discriminator, 1:500, eBioscience), $\mathrm{CD45}^{+}$ (1:200, Biolegend) singlets, and further gated on $\mathrm{TCR}\beta^{+}$ (1:200, Biolegend) GFP^{+} cells. Cells were also analyzed for CD4 (1:200, Biolegend) and CD8 (1:200, Biolegend) expression. Before analysis, 5 x 10^3 count bright beads (Life Tech) were added to the sample for normalization of cell counts, and the fraction of beads recovered was assumed to be the fraction of sample analyzed.

3.4.5 T cell sequencing and analysis

For bulk TCR beta chain sequencing, 1,000-10,000 Tconv ($\mathrm{IV}^{\mathrm{neg}}$ $\mathrm{CD4}^{+}$ $\mathrm{GFP}^{\mathrm{neg}}$) or Treg ($\mathrm{IV}^{\mathrm{neg}}$ $\mathrm{CD4}^{+}$ GFP^{+}) cells were sorted directly into 250µl RNAprotect buffer (Qiagen), spun down for 1 minute at 2000 RPM, and immediately frozen at −80°C. Naïve samples were pooled to reach minimum sample size requirements. Samples were sent to iRepertoire (Huntsville, AL) for library preparation and sequencing. TCR sequences were analyzed and compared with MiXCR and VDJtools software [35, 36].

All analyzed bulk and single-cell RNA-seq datasets have been previously reported [16, 19]. TCRs were reconstructed and compared from bulk RNA-seq datasets with MiXCR and VDJtools software [35, 36] using recommended RNA-seq settings. TCRs were reconstructed from previously published single-cell RNA-seq datasets [4] using TraCeR [37], run in short read mode with the following settings '--inchworm_only=T --trinity_kmer_length=17'.

3.4.6 Soluble protein production

Recombinant soluble TCR for yeast selections were produced in High Five (Hi5) insect cells (Thermo Fisher) via a baculovirus expression system, as previously described [20]. Briefly, ectodomain sequences of each chain followed either an acidic or basic lysine zipper domain and a poly-histidine purification site were cloned into pAcGP67a vectors. An AviTag peptide (GLNDIFEAQKIEWHE) was expressed between the acidic leucine zipper and poly-histidine site for single chain biotinylation. For each construct, 2 µg of plasmid DNA was transfected into SF9 insect cells with BestBac 2.0 linearized baculovirus DNA (Expression Systems) using Cellfectin II reagent (Thermo Fisher). Viruses were propagated to high titer, co-titrated to maximize expression and ensure 1:1 heterodimer formation, and co-transduced into Hi5 cells, which were then grown at 27°C for 48-72h. Proteins were purified from the pre-conditioned media supernatant with Ni-NTA resin and biotinylated overnight through addition of BirA ligase, ATP, and biotin. Protein were size purified via size exclusion chromatography using a S200 increase column on an AKTAPURE FPLC (GE Healthcare) and stored in 20% glycerol aliquots at -80°C until use.

3.4.7 Yeast-display construct design and validation

Yeast-displayed I-A^b was designed and validated for correct fold essentially as previously described for I-E^k [20]. Briefly, a single construct encoding a peptide, the β1 domain of H2-Ab1, the α1 domain of H2-Aa, and a Flag (DYKDDDDK) epitope tag connected by flexible Gly-Ser linkers was expressed as an N-terminal fusion to Aga2 in a pYAL vector. All constructs were transformed into electrically competent RJY-100 yeast. Yeast were grown to confluence at 30°C in pH 5 SDCAA yeast media then subcultured into pH 5 SGCAA media at OD600 = 1.0 for 48h induction at 20°C.

Proper fold was assessed by expression of p1Y-modifed 3K (FEYQKAKANKAVD) peptide and probing with fluorescently-labeled tetramers of known binders B3K506 [21] and YAe62 [22]. Following initial absence of tetramer staining, the MHC-encoding domains were mutagenized using a GeneMorph II kit (Agilent), according to manufacturer's instructions, and yeast libraries were created via homologous recombination of linearized pYAL vector and mutagenized pMHC construct. The MHC mutations Vβ6T, Iβ28F, Rβ31G, Iα9T, Vα11T, and Tα75I, none of which are located within the peptide- or TCR-binding interfaces, rescued fold and function, largely through breaking hydrophobic patches within previous protein interfaces.

Peptide libraries were created through use of mutagenic primers allowing all 20 amino acids via NNK codons. The libraries allowed limited diversity at the known MHC anchor residues to maximize the number of correctly folded and displayed pMHC clones in the library. Final libraries contained approximately 2 x 10^8 yeast transformants. Yeast libraries were selected for binding to the TCR of interest coupled to streptavidin-coated magnetic beads (Miltenyi) through magnetic-activated cell sorting. Plasmid DNA was extracted from 5x10^7 yeast from the final round of selection with the Zymoprep Yeast Miniprep Kit (Zymo Research), according to manufacturer's instructions, and transformed into chemically competent DH5α *E. coli* for single colony selection and sanger sequencing.

3.4.8 Statistical analyses

Statistical analyses were performed with paired or unpaired two-sided t tests, or two-sided one-way ANOVA tests with Tukey's correction for multiple comparisons, where appropriate. Size of test groups and statistical tests used are indicated in figure legends.

3.5 Acknowledgements

The work presented in this chapter was the result of a multi-year collaboration with many members of the Birnbaum, Jacks, Regev, and Spranger groups. Amy Li and David Canner of the Jacks lab provided the T cells for targeted TCR sequencing. Rebecca Herbst of the Regev lab at the Broad Institute of Harvard and MIT performed the single-cell sequencing analysis with TraCeR. Amy Li, Rebecca Herbst, and AJ Bhutkar of the Jacks lab provided access to bulk RNA-seq data. Brendan Horton of the Spranger lab performed adoptive transfer, collection, and analysis of transduced T cells. Khloe Gordon, Connor Dobson, and Taeyoon Kyung of the Birnbaum lab provided reagents and expertise for viral production and transduction of primary T cells. Khloe Gordon additionally provided spleens for primary T cell isolation. Jason Schenkel of the Jacks lab provided input into experimental design and contributed to related work that is not described here.

Amy Li, David Canner, Rebecca Herbst, and Brendan Horton also made intellectual contributions to the interpretation of results.

References

1. Togashi, Y., Shitara, K., & Nishikawa, H. Regulatory T cells in cancer immunosuppression — implications for anticancer therapy. *Nat. Rev. Clin. Oncol*. **16**, 356-371 (2019).

2. Roychoudhuri, R., Eil, R. L. & Restifo, N. P. The interplay of effector and regulatory T cells in cancer. *Curr. Opin. Immunol.* **33,** 101–111 (2015).

3. Sakaguchi, S., Yamaguchi, T., Nomura, T., Ono, M. Regulatory T cells and immune tolerance. *Cell* **133**, 775-787 (2008).

4. Savage, P.A., Leventhal, D.S. & Malchow, S. Shaping the repertoire of tumor-infiltrating effector and regulatory T cells. *Immunol. Rev.* **259**, 245-258 (2015).

5. Seer.cancer.gov. National Cancer Institute (NCI) Surveillance, Epidemiology, and End Results (SEER) program. (2019). at <https://seer.cancer.gov/statfacts/html/lungb.html>

6. Pardoll, D.M. The blockade of immune checkpoints in cancer immunotherapy. *Nat. Rev. Cancer* **12**, 252–264. (2012).

7. Carbone, D. P. *et al.* First-line Nivolumab in stage IV or recurrent non–small-cell lung cancer. *N. Engl. J. Med.* **376,** 2415–2426 (2017).

8. Yarchoan, M., Johnson III, B. A., Lutz, E. R., Laheru, D. A. & Jaffee, E. M. Targeting neoantigens to augment antitumour immunity. *Nat. Rev. Cancer* **17,** 209–222 (2017).

9. Schumacher, T. N. & Schreiber, R. D. Neoantigens in cancer immunotherapy. *Science* **348,** 69-74 (2015).

10. Singh, M, Murriel, C.L., & Johnson, L. Genetically Engineered Mouse Models: Closing the Gap Between Preclinical Data and Trial Outcomes. *Cancer Res.* **72**, 2695-700 (2012).

11. Mao, C. *et al*. KRAS mutations and resistance to EGFR-TKIs treatment in patients with non-small cell lung cancer: A meta-analysis of 22 studies. *Lung Cancer* **69**, 272-278 (2010).

12. DuPage, M. *et al*. Endogenous T cell responses to antigens expressed in lung adenocarcinomas delay malignant tumor progression. *Cancer Cell* **19**, 72-85 (2011).

13. Joshi, N.S. *et al.* Regulatory T cells in tumor-associated tertiary lymphoid structures suppress anti-tumor T cell responses. *Immunity* **43,** 579-90 (2015).

14. McFadden, D.G. *et al.* Mutational landscape of EGFR-, MYC-, and Kras-driven genetically engineered mouse models of lung adenocarcinoma. *Proc. Natl. Acad. Sci.* **113**, E6409-17 (2016).

15. Pfirschke, C. *et al.* Immunogenic chemotherapy sensitizes tumors to checkpoint blockade therapy. *Immunity* **44**, 343–354 (2016).

16. Li, A. *et al.* IL-33 signaling alters regulatory T cell diversity in support of tumor development. *Cell Rep.* **29**, 2998-3008 (2019).

17. DuPage, M., Dooley, A.L., Jacks, T. Conditional mouse lung cancer models using adenoviral or lentiviral delivery of Cre recombinase. *Nat. Protoc.* **4**, 1064-72 (2009).

18. Venturi, V., Price, D.A., Douek, D.C. & Davenport, M.P. The molecular basis for public T-cell responses? *Nat. Rev Immunol.* **8**, 231–238 (2008).

9. Arpaia, N. *et al.* A distinct function of regulatory T cells in tissue protection. *Cell* **162**, 1078–1089 (2015).

20. Birnbaum, M.E. *et al*. Deconstructing the peptide-MHC specificity of T cell recognition. *Cell* **157**, 1073-87 (2014).

21. Rees, W. *et al*. An inverse relationship between T cell receptor affinity and antigen dose during CD4(+) T cell responses in vivo and in vitro. *Proc. Natl. Acad. Sci*. **96**, 9781-9786 (1999).

22. Huseby, E.S. *et al.* How the T cell repertoire becomes peptide and MHC specific. *Cell* **122**, 247-60 (2005).

23. Stadinski, B.D. *et al.* Diabetogenic T cells recognize insulin bound to IAg7 in an unexpected, weakly binding register. *Proc. Natl. Acad. Sci.* **107**, 10978-83 (2010).

24. Graham, D.B. *et al*. Antigen discovery and specification of immunodominance hierarchies for MHCII-restricted epitopes. *Nat. Med.* **24**, 1762-1772 (2018).

25. Eisen, H.N., Sykulev, Y., & Tsomides, T.J. Antigen-specific T-cell receptors and their reactions with complexes formed by peptides with major histocompatibility complex proteins. *Adv. Protein Chem.***49**, 1-56 (1996).

26. Schmitt, E.G. & Williams, C.B. Generation and function of induced regulatory T cells. *Front. Immunol.* **4**, 152 (2013).

27. Pacholczyk, R. & Kern, J. The T-cell receptor repertoire of regulatory T cells. *Immunology* **125,** 450–458 (2008).

28. Golding, A., Darko, S., Wylie, W.H., Douek, D.C., & Shevach, E.M. Deep sequencing of the TCR-β repertoire of human forkhead box protein 3 (FoxP3)+ and FoxP3– T cells suggests that they are completely distinct and non-overlapping. *Clin. Exp. Immunol*. **188,** 12–21 (2017).

29. Hindley, J.P. *et al*. Analysis of the T-cell Receptor Repertoires of Tumor-Infiltrating Conventional and Regulatory T Cells Reveals No Evidence for Conversion in Carcinogen-Induced Tumors. *Cancer Res.* **71**, 736-46 (2011).

30. Malchow, S. *et al*. Aire-dependent thymic development of tumor-associated regulatory T cells. *Science* **339**, 1219 (2013).

31. Leonard, J.D. *et al*. Identification of natural regulatory T cell epitopes reveals convergence on a dominant autoantigen. *Immunity* **47**, 107-117 (2017).

32. Sainz-Perez, A., Lim, A., Lemercier, B., & Leclerc, C. The T-cell receptor repertoire of tumor-infiltrating regulatory T lymphocytes is skewed toward public sequences. *Cancer Res.* **72**, 3557-3569 (2012).

33. Anderson, K.G. *et al.* Cutting edge: Intravascular staining redefines lung CD8 T cell responses. *J. Immunol.* **189**, 2702-2706 (2012).

34. Sommermeyer, D. *et al.* Designer T cells by T cell receptor replacement. *Eur. J. Immunol.* **36**, 3052-3059 (2006).

35. Bolotin, D.A. *et al.* MiXCR: Software for comprehensive adaptive immunity profiling. *Nat. Methods* **12**, 380-381 (2015).

36. Shugay, M. *et al*. VDJtools: Unifying post-analysis of T cell receptor repertoires. *PLoS. Comput. Biol.* **11**, e1004503 (2015).

37. Stubbington, M.J.T. *et al.* T cell fate and clonality inference from single-cell transcriptomes. *Nat. Methods* **13**, 329-332 (2016).

CHAPTER 4. Design and application of yeast-displayed peptide-MHC libraries for cognate antigen discovery

ABSTRACT

T cells achieve specificity in their activation, localization, and function through highly specific interactions between their T cell receptor (TCR) and specialized immune MHC proteins, which display potential antigenic peptides on their surface for TCR surveillance. When TCRs interact with their cognate peptide-MHC (pMHC) proteins, T cell responses are initiated or perpetuated. However, due to the complexity of these interactions, the cognate antigens underlying T cell responses are rarely known, and efforts to uncover them are time and resource intensive. However, recent advances have enabled high-throughput screening and computational prediction of TCR / pMHC interactions to facilitate faster and easier cognate antigen determination. One such advance is the use of yeast-displayed pMHC libraries, which express distinctively large (10^8) collections of unique pMHC complexes on the surface of engineered yeast to screen for TCR recognition. However, likely due to their large size, these libraries are prone to discovering so-called 'mimotopes', peptides that mimic the function of – but bear little-to-no sequence or structural homology with – the true cognate antigen. Therefore, improvements to this technology are needed to facilitate rapid and unambiguous cognate antigen discovery

In this chapter, we present two case studies of the design and application of yeast-display pMHC libraries for cognate antigen discovery for clinically relevant T cell populations. While successful, these case studies highlight shortcomings in previously established methods, and demonstrate how deeper understanding of the underlying system can be utilized during the design and analysis of library selections to overcome these shortcomings. However, as these improvements rely on an in-depth understanding of the system, broad application of these methods still remains infeasible. Therefore, the methods described herein should serve as a template for future cognate antigen discovery efforts, but must be supplemented by further improvements in library design to facilitate broad and rapid application of this powerful technology.

4.1 Introduction

T cells achieve specificity in both their activation and function through interactions between their T cell receptor (TCR) and short, linear peptides displayed on the surface of Major Histocompatibility Complexes (MHCs). These peptide-MHC (pMHC)-TCR interactions are highly specific and can initiate a robust T cell response with as little as a single short-lived interaction [1]. However, the specific cognate pMHC which initiates the activation of a given T cell is rarely known due to the complexity of the system, which arises from the diversity of both the pMHC and TCR repertoires. In particular, the MHC locus encodes multiple MHC genes, each of which are highly polymorphic, yielding expression of many distinct MHC proteins with divergent repertoires of presentable peptides (see Chapter 2). In humans, this results in the expression of up to 12 unique MHC proteins (known as HLA proteins in humans) presenting unique peptide repertoires, convoluting even the MHC-restriction of a given TCR. In addition, TCRs are inherently cross-reactive, and can recognize peptides with little sequence or structural homology [2-4]. While this increases the probability of identifying a peptide agonist for a T cell of interest, it complicates identification of the original peptide agonist of T cell responses as the mimetic epitopes, or mimotopes, can bear little-to-no resemblance to the original agonist [5].

Early methods to identify the cognate antigen of a T cell response relied upon labor intensive techniques to identify the MHC restriction and source protein reactivity of the T cell response [6]. This process culminates in assays with pools of overlapping peptides to identify the agonist peptide, a process known as epitope mapping [6]. While advances in the field have enabled high-throughput screening and computational prediction of pMHC-TCR interactions (recently reviewed extensively [7]), researchers still utilize early methods when to determine the original peptide agonist of a T cell response [8]. This is due to limitations inherent to these computational and high-throughput screening techniques, which can require extensive *a priori* knowledge of the system, specialized techniques, reagents, and equipment, and can fail to identify candidate antigens or identify only peptide mimotopes. Therefore, improvement of these techniques can greatly simplify and hasten the determination of the antigenic bases of T cell responses.

One such method to identify T cell epitopes is the use of yeast-displayed pMHC libraries. Compared to other techniques, yeast-displayed pMHC libraries, which encode up to 10^8 unique peptide variants [2], allow high-throughput screening for peptides that facilitate TCR-binding yet retain linkage to their parent MHC [7]. However, their engineered nature can fail to fully recapitulate the biology of the native system, resulting in the identification of peptides that bind but fail to stimulate the TCR [9]. Additionally, the randomized nature of their encoded peptides often leads to the identification of peptides that strongly bind and activate T cells – sometimes stronger than the native antigen – but do not map to any known proteome [2]. Although this method has been used to identify natively expressed agonist peptides using computational extrapolation of the TCR recognition motif and proteomic database searches [2, 3], this method can overlook subdominant recognition motifs or fail to identify natively expressed antigens [10].

In this chapter, we describe two case studies for the identification of the natively expressed cognate antigens underlying T cell responses using yeast-displayed pMHC libraries. While these studies were primarily motivated by a desire to elucidate the underlying biology of these T cell responses, they also highlight shortcomings in previously established methods and demonstrate novel methods to improve the identification of *bona fide* cognate antigens. Therefore, the methods highlighted in this chapter may greatly improve our ability to rapidly and unambiguously determine the antigenic bases of T cell responses using yeast-displayed pMHC libraries.

4.2 Results

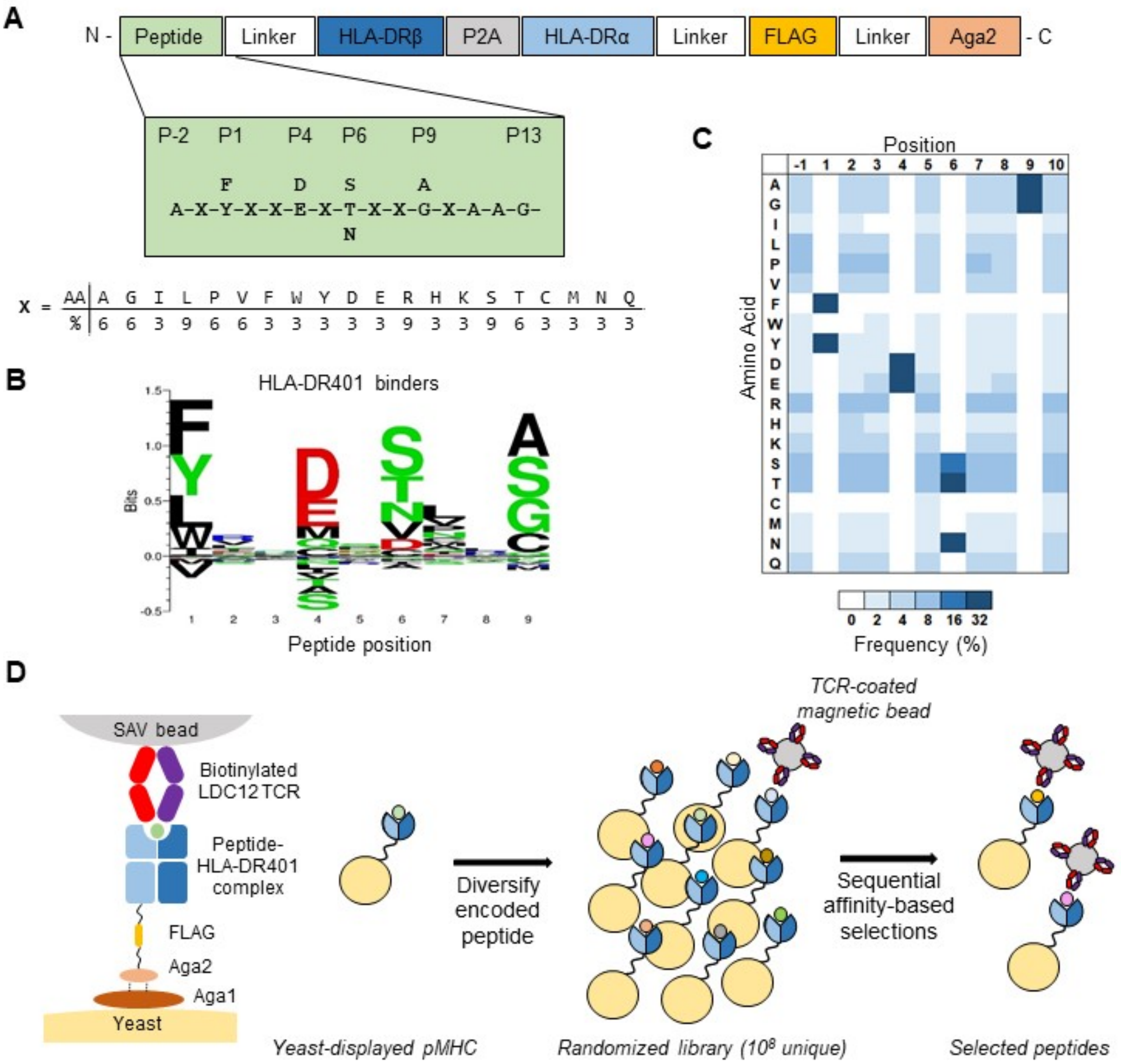
4.2.1 Antigen discovery for an allo-HLA reactive virus-specific TCR

The first case study is of the LDC12 TCR, also known as clone 12 TCR [11]. This TCR was first isolated from a CD8^{+} T cell reactive to the IPSINVHHY (IPS) peptide from human cytomegalovirus (CMV) protein pp65 presented by HLA-B*35:01 [12]. This TCR is a δ/αβ TCR, a rare class of T cells formed by VDJ rearrangement of a TCR-δ variable gene (Vδ1) with α joining (Jα) and constant (Cα) domains, paired with a TCR-β chain, but displays a typical αβ TCR class I pMHC binding mode [11]. Notably, this TCR was derived from a healthy donor but displays reactivity to antigen-presenting Epstein Barr virus (EBV)-immortalized lymphoblastiod cell lines (LCLs) in the absence of exogenous antigen [12]. This allo-HLA reactivity was restricted to LCLs expressing HLA-DRB1*04:01 (HLA-DR401) or HLA-DRB1*14:01, but not observed in antigen presentation-deficient T2 cells engineered to express HLA-DR401 (personal correspondence, M.H.M. Heemskerk). These finding suggests that this allo-HLA reactivity is peptide dependent in the context of HLA-DR401 presentation, and that this antigen is endogenously expressed, in line with previous reports [13].

Despite a known HLA restriction and a candidate protein pool from endogenously expressed genes, the fine peptide specificity of this TCR in the context of HLA-DR401 presentation has not been determined. However, a using a cDNA library, the HLA-DR401-restricted mimetic antigen 'GSEFVSALVRPAASGPQ' (GSE) peptide was discovered to activate LDC12 TCR (personal correspondence, M.H.M. Heemskerk). Yet, as this peptide does not map to any known human protein, the endogenous antigen driving this TCRs reactivity remains unknown.

In order to elucidate this endogenous antigen, we designed a yeast-displayed HLA-DR401 peptide library for selection with recombinant LDC12 TCR. This library was based on the previously described (see Chapter 2) construct of full-length yeast-displayed HLA-DR401 (Figure 4.1A). In contrast to previously described HLA-DR401 libraries, this library was designed to encode peptides that favor MHC binding in a set register but retain diversity at canonical TCR contact positions. In particular, the amino acid composition of MHC anchor residues P1, P4, P6, and P9 was designed to recapitulate the enriched motif for HLA-DR401 binders derived from selection on the basis of peptide retention (Figure 4.1B), and canonical class II pMHC TCR contact positions P-1, P2, P3, P5, and P8, as well as auxiliary MHC anchor positions P7 and P10, were encoded with degenerate 'NNK' primers (Figure 4.1A). This design is an extension of previously described yeast-displayed class II pMHC libraries [2], and deep sequencing of the resulting pre-selection library reveals successful execution of this design (Figure 4.1C).

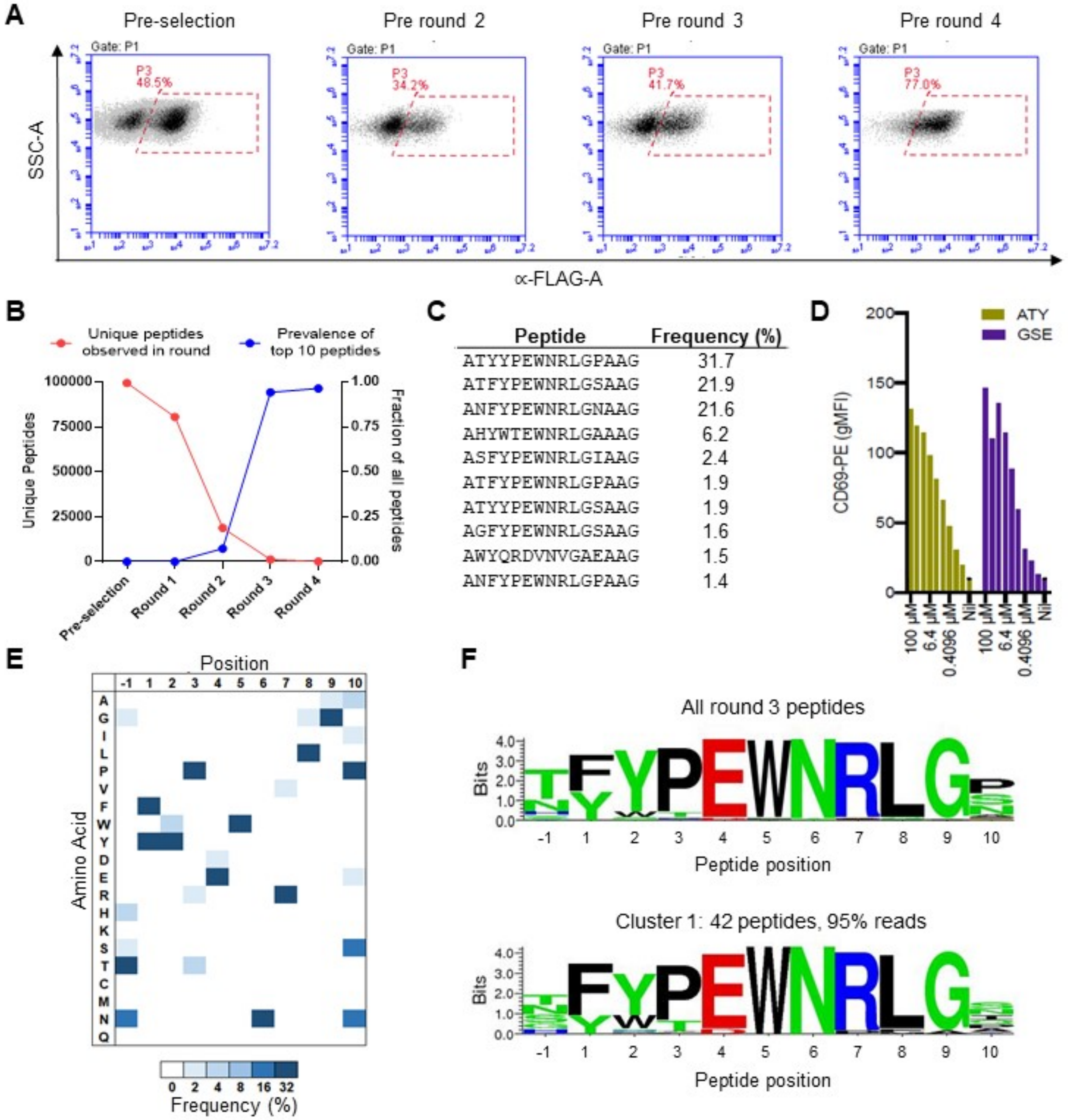
This library was then subjected to sequential rounds of affinity-based selections with magnetic beads coated with recombinant LDC12 TCR (Figure 4.1D), resulting in a population of yeast enriched for strong expression of linked FLAG epitope tag following round 3 of selection (Figure 4.2A), demonstrating enrichment of yeast with successful construct expression. Deep sequencing of this TCR-enriched yeast population revealed strong convergence on a small set of peptides by round 3 of selection (Figure 4.2B) that were highly related (Figure 4.2C). Importantly, the most enriched peptide 'ATYYPEWNRLGPAAG' (ATY) displays comparable stimulatory activity to the GSE mimotope in LDC12-expressing T cells (Figure 4.2D), demonstrating enrichment of a functionally relevant motif.



X =

AA	A	G	I	L	P	V	F	W	Y	D	E	R	H	K	S	T	C	M	N	Q
%	6	6	3	9	6	6	3	3	3	3	3	9	3	3	9	6	3	3	3	3

Figure 4.1. *Design and selection of randomized peptide yeast-displayed pMHC library to identify LDC12 TCR binders. A) Schematic of yeast-displayed HLA-DR401 pMHC construct and design of partially randomized peptide library that favors MHC binding but retains diversity at TCR contact positions. B) Kullback-Leibler relative entropy motif of the core 9mer of HLA-DR401 binders, derived empirically from a yeast-displayed HLA-DR401-linked peptide library selected for peptide retention. C) Weighted heat map of positional percent frequency of each amino acid found in peptides from the pre-selection library. D) Schematic of library selection to enrich LDC12 TCR peptide-MHCs through sequential rounds of affinity-based selections with LDC12 coated magnetic beads*



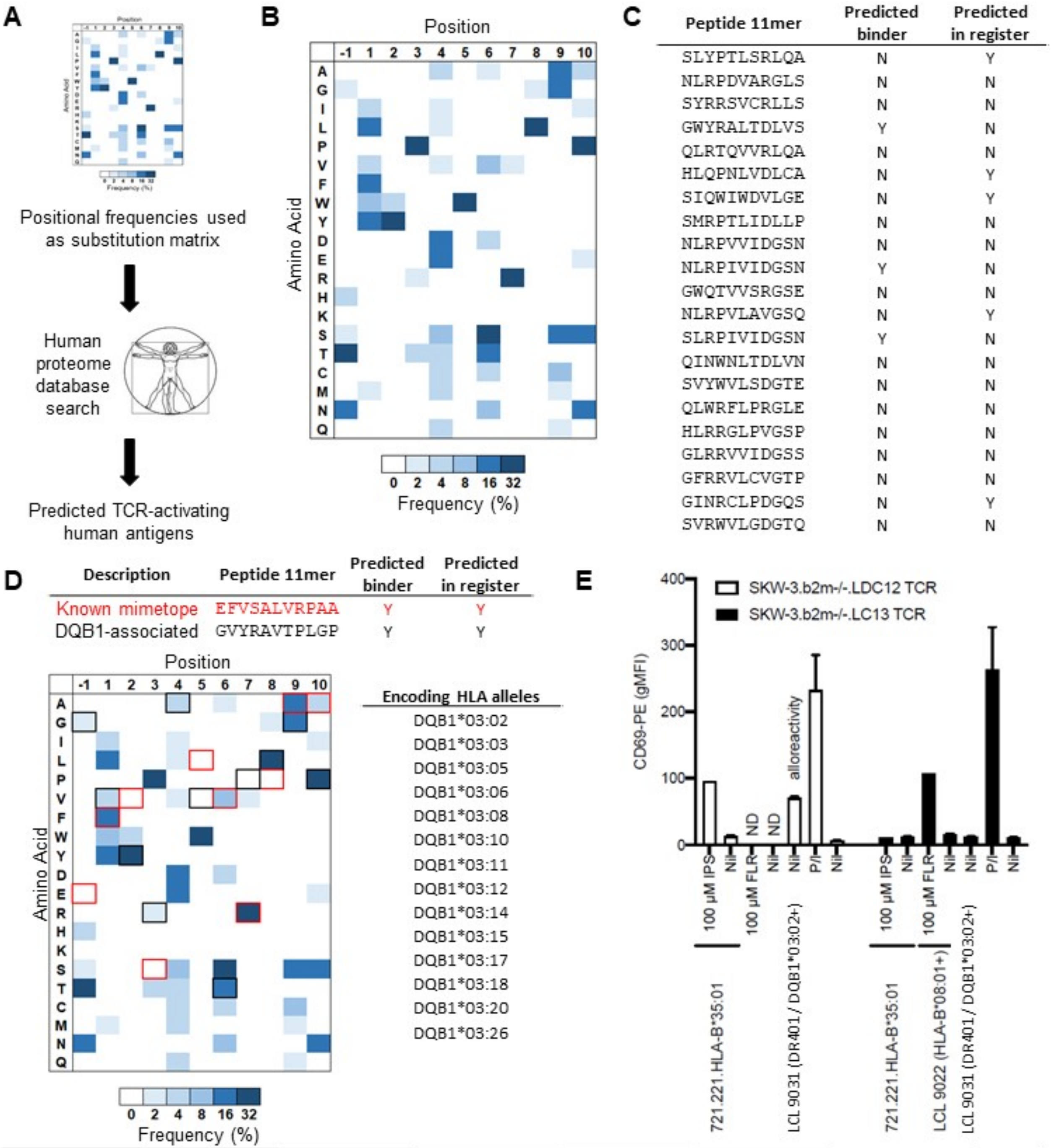
Peptide	Frequency (%)
ATYYPEWNRLGPAAG	31.7
ATFYPEWNRLGSAAG	21.9
ANFYPEWNRLGNAAG	21.6
AHYWTEWNRLGAAAG	6.2
ASFYPEWNRLGIAAG	2.4
ATFYPEWNRLGPAAG	1.9
ATYYPEWNRLGSAAG	1.9
AGFYPEWNRLGSAAG	1.6
AWYQRDVNVGAEAAG	1.5
ANFYPEWNRLGPAAG	1.4

Figure 4.2. *Library selection with LDC12 TCR enriches yeast population with a dominant and stimulatory peptide motif. A) FLAG epitope tag staining prior to each round of selection shows enrichment of yeast expressing well-expressed pMHC protein. B) Multivariate plot of the number of unique peptides observed, and proportion of peptide reads assigned to the 10 most frequent peptides, in each round of yeast library selection. C) Table of the 10 most frequently observed peptides in round 3 of library selection. D) Activation of LDC12 TCR-expressing T cell line following stimulation with various concentrations of the most frequently observed peptide (ATY) and a previously discovered LDC12 agonist (GSE), as assessed by flow cytometry. E) Weighted positional percent frequency of each amino acid in round 3 of library selection. F) Unweighted Shannon entropy sequence logos of all peptides (top) or peptides found within the most dominant cluster (bottom) found in round 3 of library selection*

It has previously been shown in reports of TCR selections of yeast-displayed pMHC libraries that peptides in round 3 of selection display TCR specificity, but retain positional diversity that is indicative of sub-optimal but tolerated substitutions [2]. Consistent with this notion, analysis of the read count-weighted positional frequency heat map of each amino acid in peptides found in round 3 of selection, and analysis of the unweighted Shannon entropy logo of these peptides (Figure 4.2F), reveals strong but not absolute amino acid preferences at each TCR contact position (Figure 4.2E). In particular, we observe strong selection for P2 Tyr (87% usage), P3 Pro (88% usage), P5 Trp (97% usage), and P8 Leu (95% usage), and more tolerant selection at P-1, with Thr and Asn most favored (59% and 24% usage, respectively). However, we also observe subdominant amino acids at each of these positions, such as P2 Trp (8% usage), P3 Thr (7% usage), P5 Val (2% usage), and P8 Gly (3% usage). While these subdominant amino acids could indicate an orthogonal and subdominant peptide motif, these preferences are retained in the dominant peptide motif (95% of reads) following clustering (Figure 4.2F), indicating that they instead represent tolerated but sub-optimal TCR contacts. In addition, we observe convergent residue preferences at primary MHC anchor positions P4, P6, and P9, as well as auxiliary positions P7 and P10, indicating that amino acid usage at these positions may contribute to preferential orientation of TCR contacting residues, or may directly contribute to TCR binding.

Combined, these amino acid preferences provide a footprint for LDC12's peptide recognition in the context of HLA-DR401 presentation. However, as expected for the selection of a randomized peptide library, none of the most enriched peptides (Figure 4.2C) map to any known protein. Therefore, in order to discover endogenously expressed candidate antigens that may underlie the allo-HLA reactivity of LDC12, we used the weighted positional frequency heat map to generate a substitution matrix to search the human proteome (Figure 4.3A), as previously described [2, 3]. However, even with a very lenient amino acid positional percent frequency threshold of 0.1%, we were unable to find a single candidate antigen in the human proteome (data not shown). This was likely due to overly stringent constraints at MHC anchor positions P1, P4, P6, and P9, due to their semi-fixed engineered composition. Therefore, we replaced these amino acid positional percent frequencies with those generated from round 5 of selection of a randomized 9mer yeast-displayed HLA-DR401 pMHC library selected for peptide retention (see Chapter 2). This loose anchor design represents an improvement of previously used substitution matrix-based proteome searches, as it significantly broadened our search space for candidate antigen discovery.

Accordingly, using this loose anchor motif (Figure 4.3B) we discovered 22 candidate antigens at a moderate positional percent frequency threshold of 0.5%. However, possibly due to loosened constraints on these anchor positions, none of these candidate antigens were predicted to bind HLA-DR401 in the correct register for LDC12 recognition, according to a prediction algorithm trained on the randomized 9mer yeast-displayed HLA-DR401 peptide library selection data (see Chapter 2). Consistent with these predictions, 5/5 candidate antigens from this list failed to exhibit LDC12 reactivity *in vitro* (data not shown). While loosening the positional percent frequency threshold increases the number of candidate antigens, it decreases the likelihood that any given antigen will bind HLA-DR401 or the TCR of interest. However, by loosening this threshold slightly at primary and auxiliary MHC anchor positions, we discovered the candidate antigen 'GVRAVTPLGP' (GVR), which maps to HLA-DQB1*03 proteins, the beta chain of a subset of HLA-DQ alleles (Figure 4.3D). Importantly, as HLA-DQ is co-expressed in all HLA-DR expressing cells, this antigen would be consistent with endogenous presentation by HLA-DR401.



Peptide 11mer	Predicted binder	Predicted in register
SLYPTLSRLQA	N	Y
NLRPDVARGLS	N	N
SYRRSVCRLLS	N	N
GWYRALTDLVS	Y	N
QLRTQVVRLQA	N	N
HLQPNLVDLCA	N	Y
SIQWIWDVLGE	N	Y
SMRPTLIDLLP	N	N
NLRPVVIDGSN	N	N
NLRPIVIDGSN	Y	N
GWQTVVSRGSE	N	N
NLRPVLAVGSQ	N	Y
SLRPIVIDGSN	Y	N
QINWNLTDLVN	N	N
SVYWVLSDGTE	N	N
QLWRFLPRGLE	N	N
HLRRGLPVGSP	N	N
GLRRVVIDGSS	N	N
GFRRVLCVGTP	N	N
GINRCLPDGQS	N	Y
SVRWVLGDGTQ	N	N

Description	Peptide 11mer	Predicted binder	Predicted in register
Known mimetope	EFVSALVRPAA	Y	Y
DQB1-associated	GVYRAVTPLGP	Y	Y

Encoding HLA alleles
DQB1*03:02
DQB1*03:03
DQB1*03:05
DQB1*03:06
DQB1*03:08
DQB1*03:10
DQB1*03:11
DQB1*03:12
DQB1*03:14
DQB1*03:15
DQB1*03:17
DQB1*03:18
DQB1*03:20
DQB1*03:26

Figure 4.3. *Library-guided search for human-origin antigens uncovers an endogenous candidate antigen underlying LDC12 allo-HLA reactivity. A) Schematic of library-guided identification of endogenous TCR-activating antigens. B) Weighted positional percent frequency of each amino acid in following round 3 of selection with LDC12, adjusted for native MHC anchor position frequencies. C) Antigens identified from proteome search with 0.5% positional frequency threshold, with associated predictions for HLA-DR401 binding. D) Minimum core epitope of a known LDC12-activating mimotope (red) and DQB1-associated candidate antigen identified from proteome search with loosened search threshold at P7 (black), with residues mapped onto weighted positional frequency heat map. E) List of DQB1 alleles which encode the candidate antigen. F) Activation of LDC12- or LC13-TCR expressing T cells following co-culture with antigen-presenting cell lines, with or without the addition of exogenous peptides.*

Unlike other candidate antigens, this HLA-DQB1*03-associated antigen is predicted to bind HLA-DR401 in the correct register for recognition (Figure 4.3D). In addition, this antigen exceeds the positional percent frequency threshold at 5/5 primary TCR contact (least dominant contact is P5 Val with 2% usage), and contains dominant P2 Tyr and P8 Leu residues (Figure 4.3D). In addition, this antigen contains favorable residues at each primary MHC anchor position and at the auxiliary anchor position P10, yet narrowly evaded our previous search due to P7 Pro, which was only observed in 0.45% of reads. Importantly, this peptide is encoded by a polymorphic region of HLA-DQB1 and therefore only encodes each of these favorable TCR contacts for a subset of HLA-DQB1*03 alleles (Figure 4.3 D). Therefore, we hypothesized that LDC12 recognizes this antigen when presented by HLA-DR401 in cells expressing one of these HLA-DQB1*03 alleles. In support of this hypothesis, HLA-DR401$^+$ LCL 9031 antigen-presenting cells stimulate LDC12-expressing T cells even in the absence of endogenous peptides and express HLA-DQB1*03:02, which encodes the candidate antigen. However, further testing of this epitope in antigen-presentation deficient T2.DR4 cells and investigation of LDC12 allo-HLA reactivity in a diverse panel of LCL antigen-presenting cells is still needed to confirm this hypothesis.

In contrast to the DQB1-associated antigen, the minimum stimulatory epitope (data not shown) of the GSE peptide, 'EFVSALVRPAA' (EFV), has only one TCR contact above 0.5% usage, P5 Leu, but does contain the highly favored P7 Arg residue and is predicted to bind HLA-DR401 in this register (Figure 4.3D). This suggests that LDC12 TCR may bind EFV peptide in an orthogonal mode to the dominant peptide motif found in our library selections. This hypothesis will be investigated by producing crystal structures of LDC12 bound to HLA-DR401-EFV, -ATY, or -GVR complexes, which will provide detailed information on differential TCR binding modes.

4.2.2 Antigen discovery for tumor-infiltrating T cells with defined antigen libraries

The second case study is of antigen discovery for T cells found enriched in tumor-infiltrating populations of patients with glioblastomas following treatment with an immunotherapeutic regimen [14]. These patients received a 'neoantigen' vaccine comprised of an immune-activating adjuvant and peptides that are predicted to be displayed by the patient's HLA molecules and contain a coding mutation unique to the tumor. These vaccinations have shown promise in other cancer types [15], and in treated glioblastoma patients extended survival and increased tumor-infiltrating T cell activity [14]. To understand their clonality and the antigen basis of their localization and function, these tumor-infiltrating T cells were isolated from treated patients and single-cell sequenced to recover their TCR pairings. By re-expressing these TCRs in T cell lines, it was found that many of these expanded T cell clones were specific to the tumor neoantigens or common viral peptides. However, after neoantigen and viral reactivity was ruled out, there remained expanded T cell clones which had no known antigen. Therefore, to discover their antigen reactivity, we expressed these TCR in soluble format and used them to select yeast-displayed pMHC libraries.

Unlike each of the previously described antigen discovery efforts in this thesis, these TCRs were derived from CD8$^+$ 'killer' T cells, which are canonically restricted to class I pMHCs [16]. Although relative to CD4$^+$ T cells, CD8$^+$ T cells have divergent roles in the immune response, TCR repertoires, and recognize different classes of MHC molecules, they are equally amenable to antigen discovery with yeast-displayed pMHC libraries [3, 9, 10]. These yeast-displayed pMHC constructs are expressed as previously described [3], as a single-chain trimer of peptide, beta-2-

microglobulin (β2M), and MHC α1-α3 domains that is linked to Aga2 through flexible Gly-Ser linkers, with a Myc epitope tag (Figure 4.4A). In addition, each yeast-displayed pMHC construct has a Y84A mutation that opens the natively closed class I MHC peptide-binding groove to facilitate single-chain expression. Importantly, this mutation does not alter TCR recognition [3].

In further contrast to the previously described antigen discovery efforts, this library was designed to encoded a defined pool of peptides expressed on multiple MHC alleles, rather than randomized peptides expressed on a single MHC allele. The MHC alleles were selected based on the class I HLA type of the patient from which the TCRs were derived, and the peptides were either directly observed in eluted ligand mass spectrometry (MS) of pMHC complexes expressed by immortalized lines of the patient's tumor cells, or were predicted to bind the patient's HLA alleles and derived from genes found over-expressed (via RNA-seq) by these immortalized cells. These directly observed and predicted tumor-associated antigens (TAA) – as well as predicted neoantigen and viral peptides – were synthesized as pooled oligonucleotides and used to generate a peptide library for each patient HLA allele (Figure 4.1A). Combined, this defined composition library design theoretically allows us to recapitulate the class I pMHC repertoire of the patient's tumor to unambiguously determine the cognate antigen of the patient-derived tumor-infiltrating T cells in the absence of mimetopes that may arise from randomly-encoded libraries. However, unlike randomized libraries, design of these peptide pools required deep knowledge of the system.

Yeast-displayed constructs of each of the class I MHC molecules that comprised this patient's HLA type, HLA-A*24:02 / 68:01, HLA-B*35:02 / 35:03, and HLA-C*04:01 (for which the patient was homozygous) were generated and expressed. Yeast-displayed HLA-B*35:03 was previously designed and validated [9] and HLA-B*35:02 was generated with two (D114N, F116Y) mutations of this template. HLA-A*24:02 and HLA-A*68:01 were validated by expression of known allele-restricted HIV antigens and selection with known cognate TCRs S19.2 [17] and c23 [18], respectively. This HLA-A*24:02 construct displayed binding to fluorescently labeled S19.2 TCR tetramer (Figure 4.4B) and HLA-A*68:01-expressing yeast were enriched when selected with c23-coated magnetic beads, despite the absence of c23 tetramer binding (Figure 4.4C). The remaining allele, HLA-C*04:01, could not be validated for fold because there are no previous reports of HLA-C*04:01-restricted TCRs, but was included in the library as a wild-type construct for completeness. These alleles showed successful expression following library generation (Figure 4.3D) and deep sequencing revealed representation of the majority of candidate antigens for each allele in the pre-selection library (range 92-96%, data not shown).

Based on their enrichment in tumor-infiltrating CD8$^+$ T cells, eight TCRs (Figure 4.5A) of unknown antigen specificity were selected to screen these libraries. Sequential rounds of affinity-based selections with TCR-coated magnetic beads (Figure 4.4A) produced an enriched pool of TCR-bound yeast for 3/8 TCRs of interest, GBM 04, 34, and 35 (Figure 4.5B). Each of these TCRs dominantly enriched a single pMHC construct (Figure 4.5B), in contrast to previously described libraries that enrich many related pMHC constructs, but consistent with our expectation for a library encoding defined antigen pools. Two of these antigens were TAAs bound to HLA-A*68:01 and the remaining was a mutant variant of a library-encoded TAA bound to HLA-B*35:02 (Figure 4.5C). Of note, each of these peptides were predicted to bind their linked MHC by NetMHCPan (data not shown), and each construct displayed specific binding to tetramers of their respective TCR (Figure 4.5D), confirming their specificity.

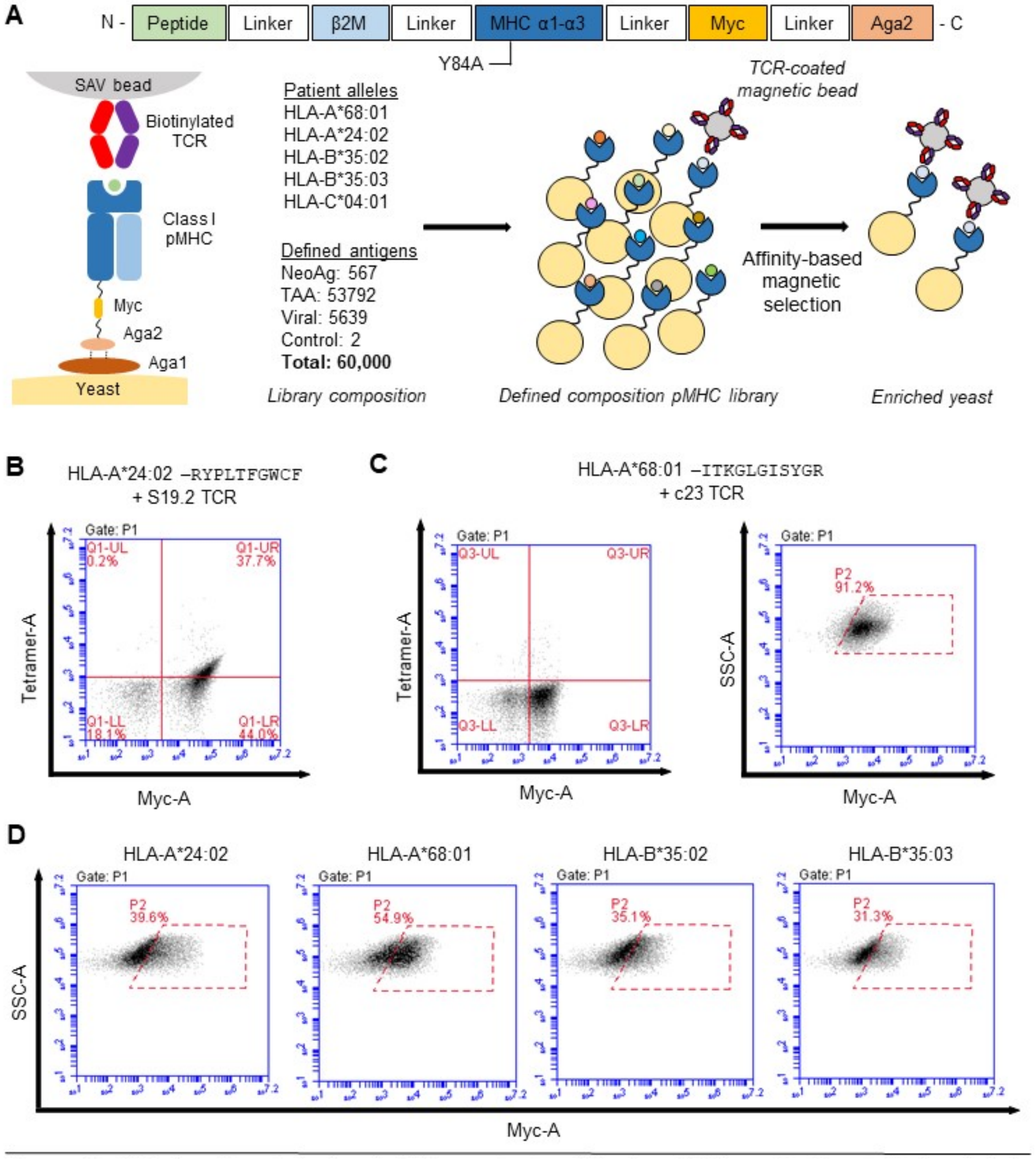
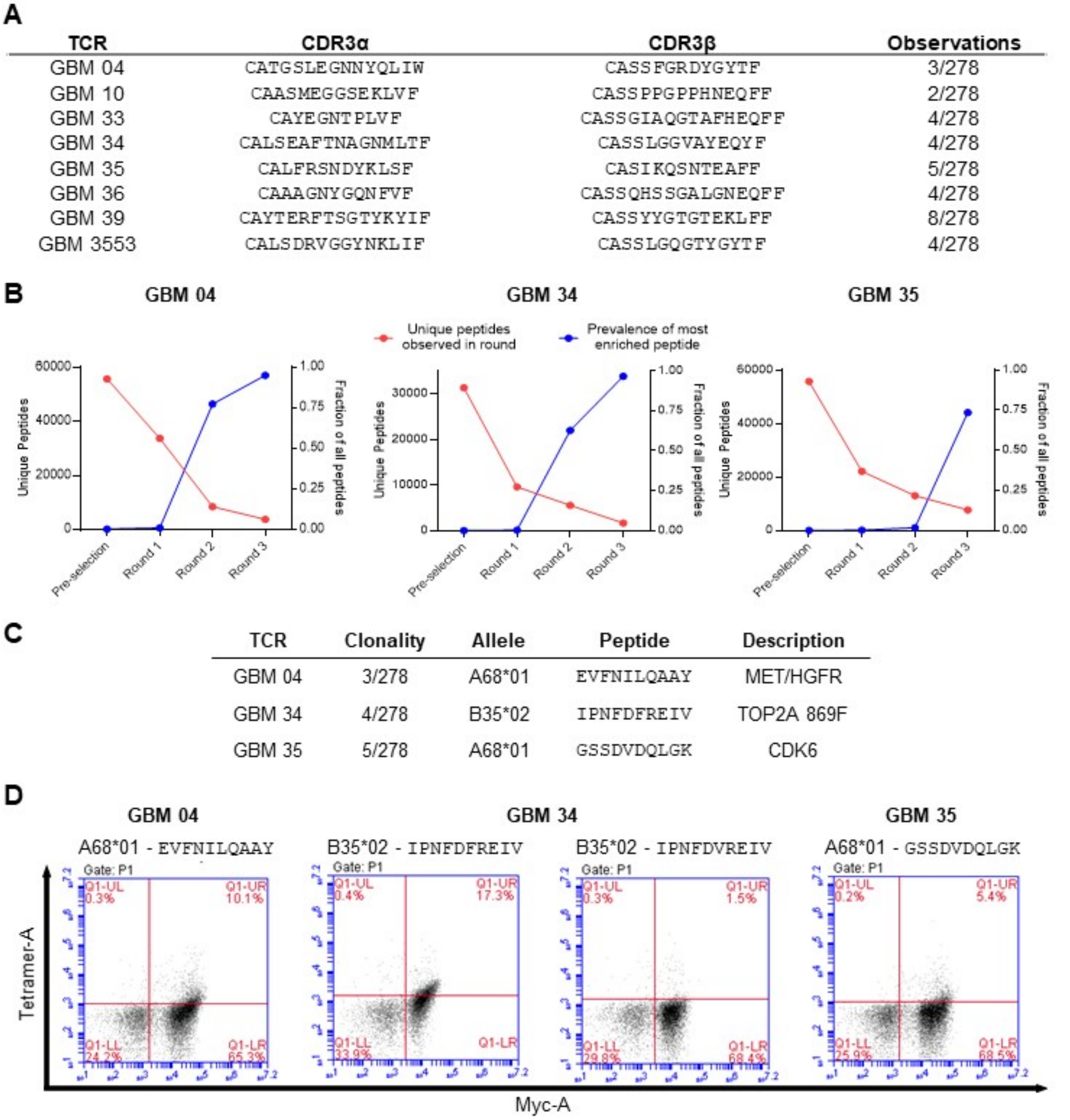


Figure 4.4. *Validation of yeast-displayed pMHC constructs and libraries. A) Schematic of yeast-displayed single chain trimer pMHC expressed as an N-terminal fusion to Aga2, with construct and library design, and TCR affinity-based selection strategy B,C) Validation of fold for yeast-displayed single-chain trimer pMHC constructs for HLA-A*24:02 (B) and HLA-A*6801 (C) alleles, using previously established cognate peptide and TCR combinations, with fluorescently labeled tetrameric TCR. In the absence of tetrameric TCR binding, fold was validated by enrichment of Myc^{+} yeast following selection with TCR-coated magnetic beads. D) Validation of successful pMHC library generation and induction for each allele studied, as assessed by percent Myc^{+} yeast prior to library selection.*



A

TCR	CDR3α	CDR3β	Observations
GBM 04	CATGSLEGNNYQLIW	CASSFGRDYGYTF	3/278
GBM 10	CAASMEGGSEKLVF	CASSPPGPPHNEQFF	2/278
GBM 33	CAYEGNTPLVF	CASSGIAQGTAFHEQFF	4/278
GBM 34	CALSEAFTNAGNMLTF	CASSLGGVAYEQYF	4/278
GBM 35	CALFRSNDYKLSF	CASIKQSNTEAFF	5/278
GBM 36	CAAAGNYGQNFVF	CASSQHSSGALGNEQFF	4/278
GBM 39	CAYTERFTSGTYKYIF	CASSYYGTGTEKLFF	8/278
GBM 3553	CALSDRVGGYNKLIF	CASSLGQGTYGYTF	4/278

C

TCR	Clonality	Allele	Peptide	Description
GBM 04	3/278	A68*01	EVFNILQAAY	MET/HGFR
GBM 34	4/278	B35*02	IPNFDFREIV	TOP2A 869F
GBM 35	5/278	A68*01	GSSDVDQLGK	CDK6

Figure 4.5. *Defined composition yeast-displayed pMHC libraries discover tumor-associated-antigen (TAA)-reactive T cells. A) Table of TCRs found enriched in tumor-infiltrating CD8$^+$ T cells in a glioblastoma patient following administration of an immunotherapeutic regime. B) Representative plot of unique peptide reads and the percent of reads associated with the most enriched peptide, for each round of library selection with TCRs found enriched in tumor-infiltrating CD8$^+$ T cells, for each TCR which enriched a population of yeast during selections. C) Table of TCRs which enriched yeast populations, with respective library-discovered candidate antigen. D) Scatter plot of fluorescent intensity of fluorescently labeled TCR tetramer and anti-Myc antibody, for yeast expressing the candidate antigen (or variants thereof) of selected TCRs, as assessed by flow cytometry.*

In support of their purported cognate antigen reactivity, the two HLA-A*68:01-linked TAAs were derived from genes found upregulated in glioblastoma patients, *MET* and *CDK6*, that are implicated in the pathogenesis or progression of glioblastoma [19, 20]. However, while the HLA-B*35:02-linked mutant antigen was derived from an over-expressed *TOP2A* gene the V869F mutation was not observed within this patient, but was encoded in the pre-selection library due to an error during oligonucleotide synthesis. As the wild-type variant of this peptide fails to show binding to GBM 34 (Figure 4.5D), this peptide may represent a mimotope for the true antigen. Therefore, individual validation of each of these peptides, as well as the wild-type variant of the TOP2A V869F antigen, in TCR-transduced primary T cell lines is needed to further confirm each these observed antigens.

4.3 Discussion

Due to the complexity of the system, the antigenic basis of most T cell responses is uncertain. Although technologies such as yeast-displayed pMHC libraries can elucidate the antigen reactivity of T cells, these methods are time and labor intensive, and can fail to uncover viable candidate antigens underlying a T cell response. Here, we described two case-studies that demonstrate both the successes and pitfalls of the application of yeast-displayed pMHC libraries to this purpose, and describe improvements to their design and use.

The first case study for LDC12 shows that while fully randomized peptide MHC libraries can discover many peptide antigens that both bind and stimulate a TCR of interest, these antigens often fail to map to any known protein. Therefore, methods have been designed to computationally search for antigens that match the overall peptide-recognition motif of the TCR of interest [2, 3]. However, as we have shown, these methods are highly sensitive to arbitrary thresholds that hold no physiologic relevance but attempt to balance antigen discovery with the likelihood of recognition. Comparable to another recent study [10], our original attempt at antigen discovery failed to produce a viable candidate antigen. However, by incorporating information on MHC-binding preferences at peptide anchor positions derived from a previous study (see Chapter 2) and setting lower thresholds at non-TCR contact positions, we were able to identify a candidate antigen that is endogenously expressed by the target cell line of interest.

While this modification represents an improvement to yeast-displayed pMHC library-guided candidate antigen discovery, it was accomplished through detailed information about the peptide repertoire of the MHC allele that is often not available. Beyond this limitation, this case study highlights other pitfalls in this method. In particular, the large divergence between our observed motif and a known LDC12 stimulatory peptide suggests that our method is only able to identify antigens that match one of potentially many orthogonal TCR recognition modes. These orthogonal recognition modes may rely on TCR contacts not covered in our library (outside of P-1 to P10), or may rely on non-conventional MHC contact residues not encoded in our semi-fixed design, yet could encompass the true antigenic peptide underlying the allo-HLA reactivity of LDC12 TCR. In addition, as seen in a previous application [2], this candidate antigen search may have yielded many plausible antigens for which further validation and testing would be required. Therefore, while this case-study displays an improvement in the design and application of yeast-displayed pMHC libraries for the discovery of antigens underlying T cell responses, it shows that this improved method is still not yet broadly applicable.

The second case study of antigen discovery for glioblastoma tumor-infiltrating T cells represents the first description and application of a defined antigen yeast-displayed pMHC library. As demonstrated by the TCR-mediated enrichment of two candidate antigens that map to genes over-expressed in the patient tumor from which these TCRs were derived, this library design has a clear advantage for directly enriching antigens of interest. However, as demonstrated by the enrichment of a peptide containing a mutation not found in this patient's tumor, this method may still be susceptible to discovering mimotopes that are irrelevant to the underlying immune response. Furthermore, this design was highly dependent on a detailed understanding of the pMHC repertoire of this particular tumor, relying on patient and tumor-specific information such as HLA type, gene expression, and gene mutations. This information then required individual validation of each MHC allele for yeast display, which was not uniformly possible, as well as antigen prediction, which is subject to its own pitfalls (see Chapter 2). The potential impact of these limitations may be best demonstrated by the absence of antigen reactivity for 5/8 TCRs found enriched in this tumor, as their cognate antigens may not have been included in our defined antigen pool or could have failed to be presented properly by our yeast-displayed libraries. Therefore, while defined antigen library approaches for antigen discovery, such as described here and in the recently described method T-Scan [21], have clear advantages for limiting the confounding effects of mimotopes, they require both detailed patient- and tissue-specific information, as well as extensive validation, in order to function properly, limiting their broad and rapid application.

Combined, these case studies show the unique advantages and pitfalls of using yeast-displayed pMHC libraries to guide TCR antigen discovery, and in particular, the discovery of *bona fide* cognate antigens that underlie T cell responses. Although the distinctly large size of these libraries enables broad searches of peptide space, discovered antigens can be physiologically irrelevant. Furthermore, attempts to limit false positive results can blind us to potential antigens. Therefore, while yeast-displayed pMHCs libraries were successfully applied in these case studies, further advances are needed to guide broad but relevant antigen searches and improve the applicability of this powerful method.

4.4 Methods

4.4.1 Yeast-displayed pMHC designs

Full-length yeast-displayed HLA-DR401 (HLA-DRA1*01:01, HLA-DRB1*04:01) with a cleavable peptide linker was based upon a previously described HLA-DR401 construct optimized for yeast display with the mutations Mα36L, Vα132M, Hβ62N, and Dβ72E to enable proper folding without perturbing either TCR- or peptide-contacting residues [22]. The alpha and beta chain ectodomains are expressed as a single transcript connected by a self-cleaving P2A sequence. The peptide is joined through a flexible linker to N-terminus of MHC β1 domain. This construct was further modified to express a 3C protease site (LEVLFQ/GP) and MYC epitope tag (EQKLISEEDL) within the flexible linker, for a total of 32 amino acids between the peptide and β1 domain

Full length yeast-displayed HLA-A*68:01, -A*24:02, -B*35:02, -B*35:03, and -C*04:01 were designed as previously described [3, 9] as a single-chain trimer of peptide, β2M, and the MHC ectodomains α1-α3 (containing the modification Y84A), connected by flexible Gly-Ser linkers. The HLA-B*35:03 construct was previously generated and optimized for yeast display with the MHC mutation T71I to enable proper folding without perturbing either TCR- or peptide contacting residues [9]. This construct was modified with HLA-B*35:02 polymorphisms D114N and F116Y to generate the HLA-B*35:02 construct. Yeast-displayed HLA-A*24:02 and HLA-A*68:01 constructs were validated without mutations through the previously described interactions of HLA-A*24:02 expressing RYPLTFGWCF peptide with recombinant S19.2 TCR [17], and HLA-A*68:01 expressing ITKGLGISYGR peptide with recombinant c23 TCR [18]. HLA-C*04:01 was generated without mutations but could not be validated for fold due to the absence of previously characterized HLA-C*04:01-restricted TCRs.

All yeast-display constructs were produced on the pYAL vector as N-terminal fusions to Aga2. All yeast strains were grown to confluence at 30°C in pH 5 SDCAA yeast media then sub-cultured into pH 5 SGCAA media at OD_{600} = 1.0 for 48h induction at 20°C [23].

4.4.2 Library design and selection

The randomized HLA-DR401-linked peptide library was generated by polymerase chain reaction (PCR) of the pMHC construct with mutagenic primers allowing all 20 amino acids via NNK codons, as previously described [2]. The libraries allowed limited diversity at the known MHC anchor residues P1, P4, P6, and P9 to maximize the number of correctly folded and displayed pMHC clones in the library. To ensure only randomized peptides expressed within the library, the template peptide-encoding region encodes multiple stop codons. Randomized pMHC PCR product and linearized pYAL vector backbone were mixed at a 5:1 mass ratio and electroporated into electrically competent RJY100 yeast [24] to generate a library of at least 1 x 10^8 transformants.

Defined antigen class I pMHC libraries were generated by PCR of synthesized pooled oligonucleotides encoding the defined antigens (Twist Biosciences) with DNA encoding constant regions of the yeast-displayed construct. Defined antigens were selected based upon their presence in eluted ligand mass spectrometry (MS) datasets from immortalized lines of patient tumor cells with or without addition of recombinant IFN-γ, or upon predicted binding to one or more patient-derived MHC alleles using a previously described antigen prediction algorithm [25], for genes found over-expressed in patient tumor cells or expressed by common human viruses. Defined antigens were individually designed with codons randomly chosen at frequencies that mimic their native usage in yeast to ensure expression and to add nucleotide diversity to highly similar peptides. PCR product and linearized DNA containing MHC allele-encoding regions and pYAL vector backbone were mixed at a 5:1 mass ratio electroporated into electrically competent RJY100 yeast to generate libraries of at least 1x10^7 transformants. Libraries were generated separately for each MHC allele to prevent homologous recombination between alleles, and were pooled prior to selection.

Yeast libraries were selected for binding to the TCR of interest coupled to streptavidin-coated magnetic beads (Miltenyi) through magnetic-activated cell sorting, as previously described [2]. Selected yeast were washed into SDCAA media for regrowth and sequential rounds of selection.

4.4.3 Library deep sequencing and analysis

Libraries were deep sequenced to determine the peptide repertoire at each round of selection. Plasmid DNA was extracted from 5x10^7 yeast from each round of selection with the Zymoprep Yeast Miniprep Kit (Zymo Research), according to manufacturer's instructions. Amplicons were generated by PCR with primers designed to capture the peptide encoding region through regions that differentiate alleles. An additional PCR round was then performed to add P5 and P7 paired-end handles with inline sequencing barcodes unique to each library and round of selection. Amplicons were sequenced on an Illumina MiSeq (Illumina Incorporated) at the MIT BioMicroCenter.

For the randomized HLA-DR401 linked peptide library, paired-end reads from were assembled via FLASH [26] and processed with an in-house pipeline which filters for assembled reads with exact matches for the expected length then sorts each read based on its inline barcode and extracts the peptide-encoding region. To ensure only high-quality peptides were analyzed, reads were discarded if any peptide-encoding base pair was assigned a Phred33 score less than 20, or did not match the expected codon pattern at NNK (N = any nucleotide, K = G or T) or semi-fixed sites. To account for PCR and read errors of high-prevalence peptides, reads were discarded if their peptide-encoding regions were Hamming distance 1 from any more prevalent sequence, Hamming distance 2 from a sequence 100 times more prevalent, or Hamming distance 3 from a sequence 10,000 times more prevalent within the same round, in line with previously published analysis methods [27]. Unique DNA sequences were translated by Virtual Ribosome [28] and filtered for peptides not encoding a stop codon.

For the defined antigen class I pMHC libraries, reads were not assembled due to a designed absence of overlap (due to a long amplicon). Forward reads were filtered for sequences with exact matches for defined flanking regions and the peptide-encoding region was extracted. Reads that did not encode a peptide (original template sequences) were removed from analysis. Reverse reads were filtered for sequences with exact matches to encoded alleles. Extracted peptide sequences were compared with designed DNA sequences and sequences observed from sequencing of pooled oligonucleotide library. Sequences not designed or observed in the supplied library were removed from analysis, as well as sequences that were out of frame or encoded stop codons.

4.4.4 Soluble protein production

Recombinant soluble TCR for yeast selections were produced in High Five (Hi5) insect cells (Thermo Fisher) via a baculovirus expression system, as previously described [2]. Briefly, ectodomain sequences of each chain followed either an acidic or basic lysine zipper domain and a poly-histidine purification site were cloned into pAcGP67a vectors. An AviTag peptide (GLNDIFEAQKIEWHE) was expressed between the acidic leucine zipper and poly-histidine site for single chain biotinylation. For each construct, 2 μg of plasmid DNA was transfected into SF9 insect cells with BestBac 2.0 linearized baculovirus DNA (Expression Systems) using Cellfectin II reagent (Thermo Fisher). Viruses were propagated to high titer, co-titrated to maximize expression and ensure 1:1 heterodimer formation, and co-transduced into Hi5 cells, which were then grown at 27°C for 48-72h. Proteins were purified from the pre-conditioned media supernatant with Ni-NTA resin and biotinylated overnight through addition of BirA ligase, ATP, and biotin. Protein were size purified via size exclusion chromatography using a S200 increase column on an AKTAPURE FPLC (GE Healthcare) and stored in 20% glycerol aliquots at -80°C until use.

4.4.5 Antigen search in human proteome

Human antigen searches were conducted as previously described [2, 3]. Briefly, at each peptide position, amino acids above a set threshold (for example, 0.5%) were considered viable. Peptides were then generated with every combination of viable amino acids. These peptides were then searched across the known human proteome (Uniprot UP000005640) for peptides with a 100% match. Peptides found within the human proteome were then assigned a score corresponding to the sum of its positional amino acid frequencies.

Each peptide was also evaluated using an allele-specific class II MHC prediction model generated from yeast-display library data (see Chapter 2) using NN-Align 2.0 [28]. Briefly, this algorithm was generated using up to 80,000 sequenced peptides assigned a target value commensurate with the final round of selection in which they were observed between 0 and 1, with increasing target value for observation in later rounds. This library data was used for training with default settings for 'MHC class II ligands', excepting expected peptide length set to 9 amino acids and expected PFR (peptide flanking residue) length set to 0 amino acids. Peptides were considered a binder if they had a predicted value higher than 0.47 (the 90th percentile of scores for 50,000 computationally generated peptides) in the correct peptide register.

4.5 Acknowledgements

The work described in this chapter represents two productive inter-institution collaborations. The LDC12 TCR project is a collaboration with the research groups of Jim McCluskey of the University of Melbourne and of Mirjam H.M. Heemskerk of Leiden University Medical Center. Sidonia Eckle of the McCluskey lab initiated the collaboration, provided recombinant LDC12 TCR for analysis, performed all *in vitro* T cell peptide restimulation assays, and has contributed to the direction and analysis of experiments. The Heemskerk group discovered LDC12 TCR and Dr. Heemskerk has provided valuable insight into its origin and recognition, and contributed to the direction and analysis of experiments. Brooke Huisman of the Birnbaum group designed and conducted the human antigen search.

The glioblastoma tumor-infiltrating T cell project is a collaboration with the group of Catherine J. Wu MD of the Dana Farber Cancer Institute and the Broad Institute of Harvard and MIT. Giacomo Oliveira provided the patient-derived TCR sequences and list of defined antigens for study and has contributed to the direction, design, and analysis of experiments. Brooke Huisman of the Birnbaum lab designed and constructed the pooled oligonucleotide libraries, and conducted computational analyses of selected peptides. We would also like to thank Leah Sibener and K. Christopher Garcia of Stanford University for providing the yeast-displayed construct of HLA-B*35:03.

References

1. Huang, J. *et al.* A single peptide-major histocompatibility complex ligand triggers digital cytokine secretion in CD4(+) T cells. *Immunity* **39**, 846-57 (2013).

2. Birnbaum, M.E. *et al.* Deconstructing the peptide-MHC specificity of T cell recognition. *Cell* **157**, 1073-87 (2014).

3. Gee, M. H. *et al*. Antigen identification for orphan T cell receptors expressed on tumor-infiltrating lymphocytes. *Cell* **172**, 549-563 (2018).

4. Riley, T.P. *et al.* T cell receptor cross-reactivity expanded by dramatic peptide-MHC adaptability. *Nat. Chem. Biol.* **14**, 934-942 (2018).

5. Crawford, F., Huseby, E., White, J. Marrack, P., & Kappler, J.W. Mimotopes for alloreactive and conventional T cells in a peptide-MHC display library. *PLoS Biol.* **2**, E90 (2004).

6. Provenzano, M. *et al.* MHC-peptide specificity and T-cell epitope mapping: Where immunotherapy starts. *Trends Mol. Med.* **12**, 465-72 (2006).

7. Gerber, H., Sibener, L.V., Lee, L.J., & Gee, M.H. Identification of antigenic targets. *Trends Cancer* **6**, 299-318 (2020).

8. Leonard, J.D. *et al*. Identification of natural regulatory T cell epitopes reveals convergence on a dominant autoantigen. *Immunity* **47**, 107-117 (2017).

9. Sibener, L.V. *et al.* Isolation of a structural mechanism for uncoupling T cell receptor signaling from peptide-MHC binding. *Cell* **174**, 672-687 (2018).

10. Saligrama, N. *et al.* Opposing T cell responses in experimental autoimmune encephalomyelitis. *Nature* **572**, 481-487 (2019).

11. Pellicci, D.G. *et al.* The molecular bases of δ/αβ T cell-mediated antigen recognition. *J. Exp. Med.* **211**, 2599-615 (2014).

12. Amir, A.L. *et al.* Allo-HLA reactivity of virus-specific memory T cells is common. *Blood* **115**, 3146-57 (2010).

13. D'Orsogna, L.J., Nguyen, T.H.O., Claas, F.H.J., C Witt, C., & Mifsud, N.A. Endogenous-peptide-dependent alloreactivity: New scientific insights and clinical implications. *Tissue Antigens* **81**, 399-407 (2013).

14. Keskin, D.B. *et al*. Neoantigen vaccine generates intratumoral T cell responses in phase Ib glioblastoma trial. *Nature* **565**, 234-239 (2019).

15. Hu, Z., Ott, P.A., & Wu, C.J. Towards personalized, tumour-specific, therapeutic vaccines for cancer. *Nat. Rev. Immunol.***18**, 168-182 (2018).

16. Neefjes, J., Jongsma, M.L.M., Paul, P., & Bakke, O. Towards a systems understanding of MHC class I and MHC Class II antigen presentation. *Nat. Rev. Immunol.* **11**, 823-36 (2011).

17. Shi, Y. *et al*. Conserved Vδ1 binding geometry in a setting of locus-disparate pHLA recognition by δ/αβ T cell receptors (TCRs): Insight into recognition of HIV peptides by TCRs. *J. Virol.* **91**, 17 (2017).

18. Gostick, E. *et al.* Functional and biophysical characterization of an HLA-A*6801-restricted HIV-specific T cell receptor. *Eur. J. Immunol.* **37**, 479–486 (2007).

19. Pearson, J.R.D. & Tarik Regad, T. Targeting cellular pathways in glioblastoma multiforme. *Signal Transduct. Target. Ther.* **2**, 17040 (2017).

20. Lam, P.Y., Di Tomaso, E., Ng, H.K., Pang, J.C., Roussel, M.F., & Hjelm, N.M. Expression of p19INK4d, CDK4, CDK6 in glioblastoma multiforme. *Br. J. Neurosurg.* **14**, 28-32 (2000).

21. Kula, T. *et al.* T-scan: A genome-wide method for the systematic discovery of T cell epitopes. *Cell* **178**, 1016-1028 (2019).

22. Birnbaum, M.E., Mendoza, J., Bethune, M., Baltimore, D. and Garcia, K. C. Ligand discovery for T cell receptors. US20170192011A1. (2017).

23. Chao, G. *et al.* Isolating and engineering human antibodies using yeast surface display. *Nature Protocols* **1**, 755–768 (2006).

24. Van Deventer, J.A., Kelly, R.L., Rajan, S., Wittrup, K.D., & Sidhu, S.S. A switchable yeast display/secretion system. *Protein Eng. Des. Sel*. **28**, 317-325 (2015).

25. Abelin, J.G. *et al.* Mass spectrometry profiling of HLA-associated peptidomes in mono-allelic cells enables more accurate epitope prediction. *Immunity* **46**, 315-326 (2017).

26. Magoč, T. & Salzberg, S. L. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*. **27**, 2957-2963 (2011).

27. Christiansen, A. *et al*. High-throughput sequencing enhanced phage display enables the identification of patient-specific epitope motifs in serum. *Sci. Rep.* **5**, 12913 (2015).

28. Nielsen, M. & Andreatta, M. NNAlign: a platform to construct and evaluate artificial neural network models of receptor-ligand interactions. Nucleic Acids Res. 45, W344-349 (2017).

CHAPTER 5 – Perspectives and future directions

5.1 Summary

T cell responses have a critical role in our immune systems' defense against pathogens and cancer, but can also potentiate and exacerbate autoimmune diseases, allergies, and transplant rejection [1-5]. Importantly, each of these roles is dependent on highly specific interactions between their T cell receptor (TCR) and peptide-MHC (pMHC) molecules. Therefore, to gain insight into the role and function of T cells across many diseases, and to leverage this insight into new therapeutic modalities, it is crucial to understand the TCR / pMHC interactions that drive T cell responses.

However, the simultaneous requirements for specificity and comprehensive epitope recognition within this system necessitates immense diversity across both the TCR and pMHC repertoires [6], confounding generalized understanding. Furthermore, evolutionary pressure for population-level immunity drives immense person-to-person variability in these repertoires [1], contributing additional complexity. Therefore, in order to understand the TCR / pMHC interactions that drive T cell responses, we must first understand the underlying repertoires of each constituent component within this system. To this end, many technologies have been developed to probe the composition of TCR repertoires [7], pMHC repertoires [8], and recognition at their intersection [9].

In this thesis, we employ and build upon these technologies to define pMHC repertoires, explore the antigenic basis of TCR repertoire convergence in a preclinical tumor model, and explore the antigen reactivity of human T cells with clinical relevance. While these results provide detailed insights into the specific TCRs and pMHCs studied, they also provide guidance for future avenues in the exploration of TCR / pMHC repertoires and their recognition, some of which are discussed below. In addition, these studies have revealed critical pitfalls and bottlenecks in current technologies that limit their rapid or broad application. These pitfall and bottlenecks, as well as potential technologies and innovations that may rectify them, are also discussed.

5.2 Future directions

5.2.1 Broadening current applications

In chapter 2 of this thesis, we presented the development and first application of randomized yeast-displayed pMHC libraries for defining pMHC repertoires. While this technology was successfully implemented for the specific MHC alleles for which it was designed, there remain vast opportunities for expanding the breadth and comprehensiveness of its application.

In particular, we developed a randomized yeast-displayed pMHC library to empirically define the peptide-MHC repertoire of the human class II HLA alleles, HLA-DR401 and HLA-DR402 (HLA-DRA1*01 / HLA-DRB1*04:01, HLA-DRA1*01 / HLA-DRB1*04:02, respectively). Most notably, we found that while these repertoires mirrored the structure of the MHC groove and well-curated datasets, they confounded antigen prediction algorithms that attempt to predict class II MHC peptide presentation. Furthermore, it was found that training these same algorithms on our yeast-displayed library data improved their performance and revealed novel peptide motifs that had been overlooked due to systemic amino acid under-representations in their training data. But while this technology improved our understanding of the peptide repertoires of these two HLA-DR alleles, as well as our ability to predict them, there are over 7000 known human class II alleles [10]. Therefore, for this technology to broadly benefit the field of class II MHC antigen predictions, it must be expanded to many more class II MHC alleles.

There are two primary limitations to the broad application of this technology: Allele validation and allele-specific reagents. For the original development of this technology, HLA-DR401 was chosen both because it is well-characterized and because it was previously developed as a full-length yeast-displayed class II pMHC construct. A full-length class II MHC was desired to fully capture the interaction with the class II MHC peptide-exchange catalyst, HLA-DM, that we suspected (and later confirmed) would shape the resulting peptide repertoire. It is uncertain whether HLA-DM would exert the same influence on a 'platform' yeast-displayed class II MHC design, as described in Chapter 3. As HLA-DR401 was previously validated for successful fold and replication of native function through the introduction of stabilizing mutations [11], this construct was immediately portable to this new application. In addition, HLA-DR402 was designed by minor modification of this scaffold, again allowing rapid application.

However, validation of MHC alleles for yeast display is not always straight-forward or simple. As described in Chapter 3 for yeast-displayed I-A^b, these constructs can require many modifications that require individual validation. Furthermore, we have found that the modifications which enable HLA-DR401 and -DR402 yeast display do not stabilize the display of other more distantly-related HLA-DR alleles (data not shown), suggesting that broad application of yeast-displayed pMHC libraries to defining class II MHC peptide repertoires will require many allele-specific validations. In addition, as seen in Chapter 4 for HLA-C*04:01, many alleles do not have previously characterized interactions with TCRs – or other proteins – to validate their fold and function.

However, there are a few possible strategies to avoid this bottleneck. One possible route is to validate fold and function through peptide retention, as was accomplished with HLA-DR402. However, a novel selection strategy to select for successfully folded MHC variants would be required, and some MHC alleles have few-to-no known curated or characterized peptide binders [12]. Another potential work-around would be to use recombinant MHC molecules that can be expressed in insect cells (see Chapter 2) to select for binding and retention of peptides expressed on the yeast surface. However, while this alternative obviates the need for allele-specific validation, it may be prone to non-specific peptide binding often observed in phage-displayed libraries [13], and highly hydrophobic and cysteine-containing peptides (which were essential to our observed peptide-binding motifs) may be prone to aggregation and disulfide-bond formation in the absence of a co-expressed MHC protein.

Therefore, innovations in the design or selection of yeast-displayed pMHC libraries are still needed to more broadly improve our understanding of class II MHC peptide repertoires and transform the field of class II MHC antigen prediction.

5.2.2 Expanding to new applications

Beyond their current applications, such as defining class II pMHC repertoires and antigen discovery for tumor-infiltrating T cells, the technologies developed in this thesis can be expanded to new applications to broaden their benefit.

One such expansion is the application of the cleavable linker peptide exchange platform, developed in Chapter 2 for class II MHC proteins, to class I MHC proteins. As for class II MHC proteins, this expansion would allow empirical determination of class I MHC peptide repertoires and may improve class I MHC antigen predictions. This design would be directly portable to the yeast-displayed class I pMHC designs described in Chapter 4 due to their engineered peptide-binding grooves. In addition, these selections could also be performed in the presence of soluble TAPBPR, which has been reported that to act as a class I MHC peptide-exchange catalyst [14], similar to HLA-DM in the class II MHC antigen-presentation pathway.

However, class I MHC antigen prediction algorithms already perform substantially better than their class II counterparts [15], and this performance has only improved with the advent of eluted ligand mono-allelic mass spectrometry (MS) training datasets [16, 17]. In addition, as the peptide-binding groove in yeast-displayed class I pMHCs is engineered to be open at the C-terminal aspect of the peptide (as opposed to closed in its' native form), not all peptide lengths and binding modes may be accurately captured by this technology. Yet it is possible that the methods and datasets currently used to benchmark the performance of class I MHC antigen prediction algorithm underestimate the true false negative rate of these algorithms, due to the systemic under-sampling of some residues (such as Cys, Trp, and Met [18]), as observed in our class II MHC system. In addition, as we and others [19] have observed, algorithms trained on MS-derived data underperform yeast-displayed library-derived data in predicting peptide affinities. Therefore, class I MHC antigen prediction algorithms may still benefit from our newly described yeast-displayed pMHC platform.

Another beneficial expansion would be the application of our defined antigen yeast-displayed pMHC libraries to antigen discovery following pathogen infection. While conventional antigen-discovery techniques are well suited to pathogens with small genomes, such as using pMHC tetramers libraries to cover the entire HIV proteome [20], which is comprised of only 15 proteins [21]. However, these techniques are not well suited to bacterial pathogens that encode an average of 5000 unique genes across a genome of 5x10^6 base-pairs [22]. However, the entire proteome of a bacterial pathogen could be encoded within a single yeast-displayed library, which can present up to 10^8 unique peptides [9]. In addition, these peptide libraries can be synthesized to exhaustively cover the bacterial proteome regardless of patient HLA type, obviating the need for potentially flawed antigen predictions. This feat requires significantly less peptides for class II pMHC libraries due to the natively open class II MHC peptide-binding groove, which allows a peptide length-independent binding in many possible registers [23], but is also feasible for class I pMHC libraries.

This technology could therefore be immediately useful for the study of T cell targeting in chronic bacterial infections such as tuberculosis, in which CD4$^+$ T cells of uncertain antigen specificity control the infection [24]. Beyond pathogens, this technology could be used to comprehensively encoding peptides from proteins frequently found over-expressed in certain tumor types for tumor-associated antigen discovery for T cells found enriched in those tumors, regardless of patient HLA type. Yeast-displayed pMHC libraries are also much better suited to these tasks than mammalian cell-based defined antigen pMHC libraries, such as the recently described T-Scan [25], due to their superior size. Therefore, the possible applications for defined antigen pMHC libraries for antigen discovery are many, and will only become more accessible with the decreasing cost of oligonucleotide synthesis [26].

5.2.3 Future technologies

Regardless of the application, there are many technological innovations that could facilitate more rapid, broad, and effective application of pMHC libraries for antigen discovery. These include innovations in TCR sequencing and recombinant expression, as well as high-throughput library selection platforms.

Two of the greatest bottlenecks in our antigen discovery pipeline were the identification and expression of TCRs of interest. As T cell receptor function relies on unique pairings between diverse TCR alpha and beta chains, any T cell sequencing strategy must contain paired-chain sequencing. Early strategies for paired-chain T cell sequencing relied on single-cell polymerase chain reactions (PCR) in 96-well plates [27, 28], greatly limiting their throughput. However, recent advances in T cell sequencing such as improved TCR transcript recovery [29, 30], single-cell transcript barcoding [31, 32], and improved computational deconvolution of sequencing data [33, 34] now enable paired-chain TCR information to be recovered from thousands of T cells simultaneously.

Yet while these advances greatly have greatly diminished the bottleneck of identifying TCRs of interest, these TCRs must still be recombinantly expressed individually for use in yeast-displayed pMHC library selections. This process involves cloning individual TCRs into vectors for recombinant expression in bacterial [35], insect [36], or mammalian [37] expression vectors, followed by protein isolation and purification. In total, the time between TCR identification and recombinant expression was approximately 50 days in our hands using an insect expression system. However, two classes of technological innovation may diminish this bottleneck: direct porting of amplified transcripts into expression vectors, and small-scale TCR expression.

While plate-based single-cell sequencing strategies allow direct porting of amplified TCR-encoding regions into expression vectors [38], this is not currently possible in higher throughput sequencing technologies. This is because many of these sequencing technologies do not provide coverage of the entire V region of either TCR chain and fragment the amplified DNA to allow for more rapid and accurate sequencing [7]. Therefore, DNA encoding TCRs of interest must be synthesized either in part or in their entirety for recombinant expression, increasing the time from identification to expression. Therefore, innovations which allow high-throughput recovery of full-length TCR variable (V, D, and J) regions will greatly reduce this bottleneck.

In fact, one such innovation was recently described [39], and uses microfluidics and single-cell emulsion PCR to generate single-transcript paired-chain TCR constructs for millions of T cells. While these transcripts were generated for expression as full-length TCRs in T cell lines, the ectodomains of both TCR chains could be extracted from these paired-chain constructs for use in soluble expression vectors with a single PCR step. Therefore, adaptation of this high-throughput microfluidic strategy for extracting full-length paired-chain TCRs would greatly diminish the time between TCR identification and expression, allowing for more rapid antigen discovery with yeast-displayed pMHC libraries.

Another innovation that could facilitate more rapid antigen discovery by diminishing the time to expression is small-scale TCR expression. In our current insect expression protocol, we recover up to 1 mg of soluble recombinant TCR per liter of culture. However, we typically used less than 30 µg of TCR during library selection. Despite this, the larger format expression is used in part to ensure optimal TCR chain pairing ratios [40]. However, successful soluble expression of the TCR using a single paired-chain construct would obviate the need to balance TCR chain ratios, and may enable small-scale TCR expression without the need for additional viral amplification steps (which currently require an additional 7-9 days). In addition, insect cell expression of BirA ligase [41] (used for biotinylation of the purified protein expressing an AviTag peptide) may further facilitate rapid small-scale expression of biotinylated TCR for use in library selections. These innovations would also greatly diminish the cost of protein expression, which is largely driven by the cost of insect media. Therefore, innovations for small-scale expression of soluble recombinant TCRs will greatly reduce both the time and monetary costs associated with generating TCRs to screen our yeast-displayed pMHC libraries.

Should these innovations be successfully implanted and substantially reduce both the time and effort required for TCR discovery and expression, higher throughput selection methods will be needed to facilitate broader antigen discovery efforts. Fortunately, as current selection methods utilize magnetic-activated cell sorting technologies, commercially available (AutoMACs, Miltenyi) or custom automated magnetic cell sorting platforms can be used to stream-line selections for higher throughput discovery.

Yet even with each of these above described innovations, yeast-displayed pMHC library-guided antigen discovery campaigns will still be limited to tens of unique TCRs at a time. As such, higher throughput antigen discovery will require a complete reformatting of current strategies. These so-called ‘library-versus-library’ selection strategies [42] will therefore require significant innovations of existing TCR and pMHC expression platforms, or the development of entirely new platforms, but represents the ideal format for future TCR antigen discovery efforts.

5.3 Closing thoughts

In conclusion, yeast-displayed pMHC libraries are a uniquely powerful tool for exploring TCR and pMHC repertoires and recognition due to their large size and engineered composition. In this thesis, we described the development and application of these libraries to define class II MHC peptide repertoires, explore CD4$^+$ regulatory T (Treg) cell TCR repertoire convergence in a pre-clinical tumor model, and discover the native antigen reactivity of clinically relevant human T cells. Yet while this thesis focused principally on class II pMHCs and CD4$^+$ T cells, the technologies we developed, as well as the insights we gained through their application, are readily applicable to class I pMHCs and CD8$^+$ T cells. Therefore, we believe that the research described herein will contribute not only to a detailed understanding of the systems studied, but will greatly enable future applications of yeast-displayed pMHC libraries for the study of T cell responses across many distinct maladies. However, as we have discussed, key advancements are still needed to optimize this powerful technology for rapid and broad application, providing a foundation for both continued innovation and discovery in this field.

References

1. Blackwell, J. M., Jamieson, S. E., & Burgner, D. HLA and infectious diseases. *Clin. Microbiol. Rev*. **22**, 370-85 (2009).

2. Hadrup, S., Donia, M., Thor-Straten, P. Effector CD4 and CD8 T cells and their role in the tumor microenvironment. *Cancer Microenviron*. **6**, 123-133 (2013).

3. Bluestone J.A., Bour-Jordan, H., Cheng, M., & Anderson, M. T cells in the control of organ-specific autoimmunity. *J. Clin. Invest.* **125**, 2250-2260 (2015).

4. Woodfolk, J. A. T-cell responses to allergens. *J. Allergy Clin Immunol*. **119**, 280-294 (2007).

5. Issa, F., Schiopu, A., Wood, K.J. Role of T cells in graft rejection and transplantation tolerance. *Expert Rev. Clin. Immunol*. **6**, 155-169 (2010).

6. Turner, S.J., La Gruta, N.L., Kedzierska, K., Thomas, P.G., & Doherty, P.C. Functional implications of T cell receptor diversity. *Curr. Opin. Immunol.* **21**, 286-90 (2009).

7. De Simone, M., Rossetti, G., & Pagani, M. Single cell T cell receptor sequencing: Techniques and future challenges. *Front. Immunol.* **9**, 1638 (2018).

8. Gfeller, D. & Bassani-Sternberg, M. Predicting antigen presentation-what could we learn from a million peptides? *Front. Immunol.* **9**, 1716 (2018).

9. Gerber, H., Sibener, L.V., Lee, L.J., & Gee, M.H. Identification of antigenic targets. Trends Cancer 6, 299-318 (2020).

10. Robinson, J., Halliwell, J.A., Hayhurst, J.H., Flicek, P., Parham, P., & Marsh, S.G.E. The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Res.* **43**, D423-431 (2015).

11. Birnbaum, M. E., Mendoza, J., Bethune, M., Baltimore, D. and Garcia, K. C. Ligand discovery for t cell receptors. US20170192011A1. (2017).

12. Vita, R. *et al.* The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Res.* **47**, D339-D343 (2019).

13. Wu, C.H., Liu, I.J., Lu, R.M., &Wu, H.C. Advancement and applications of peptide phage display technology in biomedical science. *J. Biomed. Sci.* **23**, 8 (2016).

14. Hermann, C. *et al.* TAPBPR alters MHC class I peptide presentation by functioning as a peptide exchange catalyst. *Elife* **4**, (2015).

15. Zhao, W. & Sher, X. Systematically benchmarking peptide-MHC binding predictors: From synthetic to naturally processed epitopes. PLoS Comput Biol. 14, (2018).

16. Abelin, J. G. *et al*. Mass spectrometry profiling of HLA-associated peptidomes in mono-allelic cells enables more accurate epitope prediction. *Immunity*. **46**, 315-326. (2017).

17. Sarkizova, S. *et al.* A large peptidome dataset improves HLA class I epitope prediction across most of the human population. *Nat. Biotechnol*. **38**, 199–209 (2020).

18. Abelin, J. G. *et al.* Defining HLA-II Ligand Processing and Binding Rules with Mass Spectrometry Enhances Cancer Epitope Prediction. *Immunity* **51**, 766-779 (2019).

19. Garde, C. *et al*. Improved peptide-MHC class II interaction prediction through integration of eluted ligand and peptide affinity data. *Immunogenetics* **71**, 445-454 (2019).

20. Campion, S.L. *et al*. Proteome-wide analysis of HIV-specific naive and memory CD4(+) T cells in unexposed blood donors. *J. Exp. Med.* **211**, 1273-80 (2014).

21. Frankel, A.D. & Young, J.A. HIV-1: Fifteen proteins and an RNA. *Annu. Rev. Biochem.* **67**, 1-25 (1998).

22. Land, M. *et al.* Insights from 20 years of bacterial genome sequencing. *Funct. Integr. Genomics* 15, 141-61(2015).

23. Jones, E. Y., Fugger, L., Strominger, J. L., Siebold, C. MHC class II proteins and disease: a structural perspective. *Nat. Rev. Immunol*. **6**, 271-282 (2006).

24. Sakai, S., Mayer-Barber, K.D., & Barber, D.L. Defining features of protective CD4 T cell responses to mycobacterium tuberculosis. *Curr. Opin. Immunol.* **29**, 137-42 (2014).

25. Kula, T. *et al.* T-scan: A genome-wide method for the systematic discovery of T cell epitopes. *Cell* **178**, 1016-1028 (2019).

26. Kosuri, S. & Church, G.M. Large-scale de novo DNA synthesis: Technologies and applications. *Nat. Methods* **11**, 499-507 (2014).

27. Han, A., Glanville, J., Hansmann, L., & Davis, M.M. Linking T-cell receptor sequence to functional phenotype at the single-cell level. *Nat. Biotechnol.* **32**, 684-92 (2014).

28. Dash, P., Wang, G.C, & Thomas, P.G. Single-Cell Analysis of T-Cell Receptor αβ Repertoire. *Methods Mol. Biol.* **1343**, 181-97 (2015).

29. Tu, A.A. *et al*. TCR sequencing paired with massively parallel 3' RNA-seq reveals clonotypic T cell signatures. *Nat. Immunol.* **20**, 1692-1699 (2019).

30. Li, S. *et al*. RNase H-dependent PCR-enabled T-cell receptor sequencing for highly specific and efficient targeted sequencing of T-cell receptor mRNA for single-cell and repertoire analysis. *Nat. Protoc.* **14**, 2571-2594 (2019).

31. Zheng, G.X.Y. *et al*. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).

32. Gierahn, T.M. *et al*. Seq-Well: Portable, low-Cost RNA sequencing of single cells at high throughput. *Nat. Methods* **14**, 395-398 (2017).

33. Lee, E.S. *et al*. Identifying T cell receptors from high-throughput sequencing: dealing with promiscuity in TCRα and TCRβ pairing. *PLoS Comput. Biol.* 13, e1005313 (2017).

34. Holec, P.V., Berleant, J., Bathe, M., & Birnbaum, M.E. A Bayesian framework for high-throughput T cell receptor pairing. Bioinformatics **35**, 1318-1325 (2019).

35. Novotny, J. *et al*. A soluble, single-chain T-cell receptor fragment endowed with antigen-combining properties. *Proc. Natl. Acad. Sci*. **88**, 8646–8650 (1991).

36. Jordan, K.R. *et al*. Baculovirus-infected insect cells expressing peptide-MHC complexes elicit protective antitumor immunity. *J. Immunol.* **180**, 188-97 (2008).

37. Walseng, E. *et al*. Soluble T-cell receptors produced in human cells for targeted delivery. *PLoS One* **10**, e0119559 (2015).

38. Guo, X.J. *et al*. Rapid cloning, expression, and functional characterization of paired αβ and γδ T-cell receptor chains from single-cell analysis. *Mol. Ther. Methods Clin. Dev.* **3**, 15054 (2016).

39. Spindler, M.J. *et al*. Massively parallel interrogation and mining of natively paired human TCRαβ repertoires. *Nat. Biotechnol*. (2020).

40. Birnbaum, M. E. et al. Deconstructing the peptide-MHC specificity of T cell recognition. Cell 157, 1073-87 (2014).

41. Tykvart, J. *et al*. Efficient and versatile one-step affinity purification of in vivo biotinylated proteins: Expression, characterization and structure analysis of recombinant human glutamate carboxypeptidase II. *Protein Expr. Purif.* **82**, 106-15 (2012).

42. Pelletier, J.N., Arndt, K.M., Plückthun, A., & Michnick, S.W. An in vivo library-versus-library selection of optimized protein-protein interactions. *Nat Biotechnol.* **17**, 683-90 (1999).