Molecular Display of Synthetic Oligonucleotide Libraries and their Analysis with High Throughput DNA Sequencing

by Harry Benjamin Larman

B.S., University of California, Berkeley (2002)

Submitted to the Department of Materials Science & Engineering
and the Harvard-MIT Division of Health Sciences and Technology

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2012

Signature of Author..Harry Benjamin Larman
May 5, 2012

Certified by...Stephen J. Elledge, Ph.D.
Gregor Mendel Professor of Genetics, Harvard Medical School
Thesis Supervisor

Certified by...K. Dane Wittrup, Ph.D.
C.P. Dubbs Professor of Chemical Engineering and Biological Engineering, MIT
Thesis Committee Chair

Accepted by...Ram Sasisekharan, Ph.D.
Edward Hood Taplin Professor of Health Sciences & Technology and Biological Engineering, MIT
Director, Harvard-MIT Division of Health Sciences and Technology

Molecular Display of Synthetic Oligonucleotide Libraries and their Analysis with High Throughput DNA Sequencing

by

Harry Benjamin Larman

Submitted to the Harvard-MIT Division of Health Sciences and Technology
on May 5, 2012
in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in
Medical Engineering & Medical Physics

Abstract

High throughput methods in molecular biology have changed the landscape of biomedical research. In particular, advances in massively parallel DNA sequencing and synthesis technologies are defining our genomes and the products they encode. In the first part of this thesis, we have constructed a rationally designed antibody library and analysis platform optimized for use with deep sequencing technologies. Libraries of fully defined oligonucleotides encode three complementarity determining regions (CDRs; L3 from the light chain, H2 and H3 from the heavy chain), and were combinatorially cloned into a synthetic single chain variable fragment (scFv) framework for molecular display. Our novel CDR sequence design utilized a hidden Markov model (HMM) that was trained on all antibody-antigen co-crystal complexes present in the Protein Data Bank. The resultant ~10^{12} member library has been produced in ribosome display format, and was comprehensively analyzed over four rounds of antigen selections by multiplex paired-end Illumina sequencing. The HMM library generated multiple antibodies against an emerging cancer antigen and is the basis of a next generation antibody production platform.

In a second application of these technologies, we have created a synthetic representation of the complete human proteome, which has been engineered for display on bacteriophage. We use this library together with deep DNA sequencing methods to profile the autoantibody repertoires of individuals with autoimmune disease in a procedure called phage immunoprecipitation sequencing (PhIP-Seq). In a proof-of-concept study, this method identified both known and novel autoantibodies contained in the spinal fluid of a control patient with paraneoplastic neurological syndrome. The study was then expanded to include a large scale automated screen of 289 independent antibody repertoires, including those from a large number of healthy donors, multiple sclerosis patients, rheumatoid arthritis patients, and type 1 diabetics. Our data describes each individual's unique "autoantibodyome", and defines a small set of recurrently targeted peptides in health and disease.

Thesis Supervisor: Stephen J. Elledge, Ph.D.

Title: Gregor Mendel Professor of Genetics, Harvard Medical School

Credits

The work presented here is the result of a collaborative effort among a group of research clinicians, academic scientists, and industrial researchers. The individuals associated with each project and their contributions are documented at the outset of each chapter. Four people, however, deserve special recognition, as their contributions were pivotal to the success of this dissertation. In order of appearance: Stephen J. Elledge Ph.D. has directed this research, served as my professional mentor, and has provided unlimited wisdom, scientific insight, innovation, and friendship. Nicole Solimini Ph.D. was instrumental in building the T7-Pep library and acquiring the clinical samples that eventually proved the viability of PhIP-Seq. Uri Laserson played a lead role in the development of the statistical frameworks required to make sense of the colossal datasets presented herein. George Xu co-starred in the HMM scFv library project.

Dedication

This Thesis is dedicated to my loving wife Tasha Larman.

Quotes to set the mood

"Believe those who are seeking the truth. Doubt those who find it."
- André Gide

"The cure for boredom is curiosity. There is no cure for curiosity."
- Ellen Parr

"An inventor is simply a fellow who doesn't take his education too seriously."
- Charles F. Kettering

"I'm Winston Wolfe. I solve problems."
- The Wolf

Table of Contents

Figures and Tables

1. Introduction

1.1 Background and motivation

Technologies that enable the high throughput analysis of biomolecules have changed the landscape of biomedical research. Certainly, one of humankind's most consequential triumphs has been the determination of our own genetic code, a feat made possible by technical innovations in the field of DNA sequencing. Continuing advances are fueling efforts to sequence the genomes of thousands more human individuals in health and disease,[1, 2] as well as a plethora of nonhuman organisms[3] and the microbial communities that populate our bodies.[4] "Next generation" and "third generation" (or "single molecule") DNA sequencing technologies work by separating and analyzing single molecules of fragmented DNA. Next generation sequencing achieves analysis by distributing single DNA fragments onto a surface or beads, where it can be locally and clonally amplified as a polymerase colony ("polony").[5] These micron-sized DNA clusters can then be sequenced by cycles of base addition or ligation. Third generation technologies bypass the polony formation, and directly sequence single molecules of DNA.[6] Both next and third generation DNA sequencing strategies are often referred to as "massively parallel" or "deep" sequencing technologies, since they are capable of generating a very large number of short sequencing "reads" at the same time. One can imagine that as single molecule technologies mature, we will witness another sea change in biomolecular research: the emergence of whole genome DNA sequencing as routine and inexpensive as a PCR reaction (Figure 1.1).

While these advances in DNA sequencing technologies are incredibly important, alone they do little to further our functional understanding of the genetically encoded proteins. Knowledge of a protein's function comes from observing phenotypic responses to a change in its abundance or localization, or by determining its set of interaction partners and/or enzymatic activities. Standard techniques can now be used to label, knock out, knock down, mutate, or overexpress a protein under investigation, either in tissue culture or in an animal. Recent advances in throughput have been made possible by construction of RNA interference (RNAi) libraries[7] and the cloning of a large fraction of our genome's expressed sequences (the "ORFeome").[8] With regard to the determination of protein-protein interactions, major advances have come with the development of mass spectrometry and high throughput expression cloning technologies. The latter, which utilize methods of linking libraries of genetic fragments to the polypeptides they encode (also known as "molecular display"), have enabled, for example, the rapid discovery of therapeutic drug targets, viral epitopes important for immunity, and the elucidation of highly interconnected protein-protein interaction networks. Importantly, these approaches can be used to test otherwise unattainable numbers of interactions, and so are inherently less biased than candidate, hypothesis-driven experiments. This thesis seeks to improve upon current methods in molecular display, by incorporating recent advances in both high throughput oligonucleotide synthesis and deep sequencing of DNA libraries.

The availability of complex libraries of high quality, relatively long oligonucleotides has only recently become a reality (Figure 1.1). Currently, the most important methods for producing sequence defined, custom oligonucleotide libraries include inkjet printing (Agilent Technologies), digital micromirror device (DMD) based photolithography (Roche Nimblegen, LC Sciences, Mycroarray.com), and microelectrochemical array synthesis (Combimatrix). These techniques share a common strategy in that the oligos are first synthesized in the form of a DNA microarray, before being chemically released into solution and shipped to the customer. The range of ways in which these synthetic oligonucleotide libraries can be exploited is only beginning to emerge. For example, libraries of DNA oligos derived from microarrays have recently been utilized in the context of pooled RNAi screening,[7] and for the parallel assembly of synthetic genes.[9]

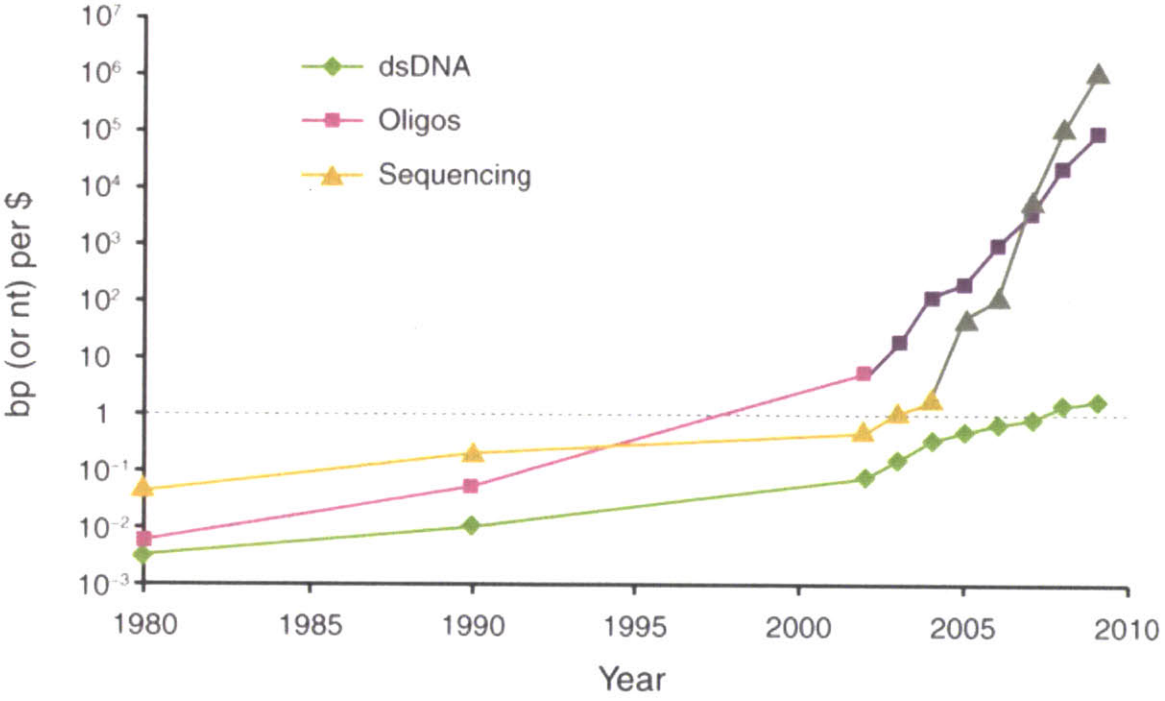


Figure 1.1: Efficiency trends in synthesis and sequencing over the past 30 years (base pairs per dollar)

Transition of yellow to gray illustrates improvement due to massively parallel sequencing technologies, and transition from pink to purple denotes improvements due to microarray oligo synthesis. This figure is from Carr & Church, *Nat Biotech*, 2009.[10]

The application of massively parallel DNA sequencing and oligonucleotide synthesis to molecular display technologies is particularly interesting in the context of immunology. The adaptive immune systems of vertebrate animals are essentially an extraordinarily elaborate, yet elegant, molecular display platform. Lymphocytes undergo complex genetic processes (e.g. recombination, untemplated nucleotide insertions and hypermutation) to create libraries of unique cell surface receptors which are then selected for their ability to recognize potentially important molecular shapes. Analyses of lymphocyte receptor repertoires had been almost impossible prior to the development of deep sequencing technologies, which are now becoming an increasingly popular method for characterizing them.[11-13] A population-scale characterization of lymphocyte receptor recognition specificities, however, is only now becoming feasible with sophisticated proteomic methods (Chapter 1.1.1).

A second way in which molecular display technology harmonizes with the immune system is by emulating it. One of the most extensively utilized type of display libraries is based on the antigen binding domain of antibodies.[14] In addition to their role as important laboratory reagents, antibodies have become an extremely successful pharmaceutical molecule, and so it is not surprising that protein engineers have sought to harness the power of molecular display for the production of antibodies with well-defined properties. One property of particular interest is that the antibody be as close to a human polypeptide sequence as possible, so as to minimize the likelihood of a patient developing inhibitor antibodies that inactivate the therapeutic. Antibodies against self proteins are not typically retrievable directly from human donors, thus a number of techniques have been developed to "humanize" animal immune systems,[15] or to display and evolve naive human antibody repertoires *in vitro*. Indeed, efforts of the latter sort led to the development of the first fully human antibody to be approved by the FDA, adalimumab (Humira, targets TNF-α), which in 2009 produced over $5 billion in annual sales. Finally, fully synthetic human repertoires have also been constructed and used to generate high affinity antibodies (Chapter 1.1.2).[16, 17] In the context of synthetic antibody libraries, massively parallel DNA sequencing and synthesis technologies promise to dramatically expand the potential of *in vitro* antibody display techniques, and is the focus of Chapter 2.

1.1.1 Screening immune receptor repertoires

The adaptive immune system is charged with the task of generating enough receptor diversity to recognize virtually any molecular shape, while at the same time removing or inactivating those receptors that happen to recognize molecular shapes present in the healthy host organism. Failure in the first case permits establishment of infection, while failure in the second case can result in chronic inflammation and/or autoimmune disease. Immune receptors that recognize self molecules and trigger immune responses can do harm to the host organism in several ways. Systemic lupus erythematosis is thought to develop largely due to a failure of programmed B cell death, resulting in populations of autoreactive B cells that damage the body's tissues via multiple distinct mechanisms (e.g. lupus nephritis due to immune complex deposition and complement activation in the renal glomeruli, secondary antiphospholipid syndrome due to antibodies that bind components of the cell membrane and result in hypercoagulability, etc.). At the other end of the spectrum is development of pathogenic receptors that target a very specific self epitope, such as the activating antibodies against the receptor for thyroid-stimulating hormone that occur in Grave's disease.[18] Between these extremes lie a large number of autoimmune processes with more complex, and often mysterious etiologies. For example, Wegener's granulomatosis (part of a larger group of vasculitic syndromes) is characterized by granuloma formation on top of a nonspecific inflammatory background and is associated with characteristic anti-neutrophil cytoplasmic antibodies (ANCAs) that are now believed to be pathogenic.[19]

As autoantigen discovery technologies continue to mature, we are sure to elucidate a myriad of disease pathogeneses that have eluded our understanding for decades. With the high resolution definition of eliciting epitopes will come an insight into disease triggering events, such as those involving viral infection and molecular mimicry that can lead to viral-self epitope spreading. One can imagine that in the future, individuals with risk-conferring HLA haplotypes will be immunized against those viral pathogens with a propensity for triggering loss of tolerance. In the few cases where autoantigens are well defined, there is now ongoing effort to develop antigen-targeted therapies to alleviate the immune attack. For example, pathogenic antibodies against the acetylcholine receptor (AChR) have been temporarily removed from a patient with myasthenia gravis by immunoadsorption on immobilized recombinant AChR domains.[20] Antigen immunization strategies seeking to skew the Th1/Th2 balance have been explored in type 1 diabetes, but have so far met mostly with disappointing results.[21, 22] In a number of animal studies, introducing autoantigens within a tolerogenic context can reverse or alleviate disease, thus prompting the initiation of clinical trials in humans.[23, 24] Finally, efforts to increase the abundance and/or activity of tolerogenic, CD4+Foxp3+ regulatory T cells are also ongoing.[25] In addition to these highly anticipated antigen-specific therapies of the future, disease-associated autoantibodies are today invaluable biomarkers that are routinely used for clinical diagnosis.

A variety of methods have been successfully utilized to identify important autoantigens. For example, Lüdemann et al. used autoantibody affinity purification followed by protein sequencing to identify proteinase 3 as the target autoantigen in Wegener's granulomatosis.[26] Szabo et al. and Buckanovich et al. both used λ ZAP phage libraries of neuronal cDNAs to identify the Hu and NOVA paraneoplastic neurological disorder autoantigens, respectively.[27, 28] Using tissue expression patterns and candidate-based approaches, Wenzlau et al. identified the zinc transporter ZnT8 as a major autoantigen in diabetes,[29] and Lennon et al. found aquaporin 4 to be targeted by autoantibodies in neuromyelitis optica.[30] The method we describe in Chapter 3 is thus an additional tool that can be used to identify targeted self antigens in autoimmune disease by profiling secreted antibody specificities.

1.1.2 Synthetic immune systems

The converse of using molecular display libraries for the discovery of disease-associated autoantigens is their use in the production of desirable autoantibodies targeting a disease-associated antigen. Human(ized) monoclonal antibodies against endogenous proteins are now the fastest growing segment of the pharmaceutical industry.[31] TNF agents, B cell depletion therapies, and inactivating growth factor receptor antibodies (HER2, EGFR) are among the most successful examples, and we are sure to witness the rate of antibody FDA approvals to continue accelerating for the foreseeable future. In addition to self antigens, a great deal of effort has been directed at the production of monoclonal antibodies against dangerous toxins[32] or viral epitopes. Of particular interest, broadly neutralizing influenza and HIV antibodies are in development,[33-35] and will be useful in providing passive immunity to high-risk individuals.

Protein engineers have devised molecular display technologies based on minimal antibody fragments from the antigen combining site of the molecule (Figure 1.2). The multiplicity of display formats and their inherent strengths and weaknesses have been reviewed extensively.[36-38] A significant barrier to the success of antibody display involves the analysis of enriched libraries. Experiments typically involve many rounds of enrichment, which are followed by the sequencing of "representative" clones. In addition to experimentally confounding variables such as clonal growth advantage, valuable information about population dynamics is never captured with these methods. Deep sequencing of the libraries is the obvious solution to this problem, but challenges associated with sample preparation, as well as candidate antibody rescue have largely prohibited its adoption. These problems are further discussed and addressed in Chapter 2.

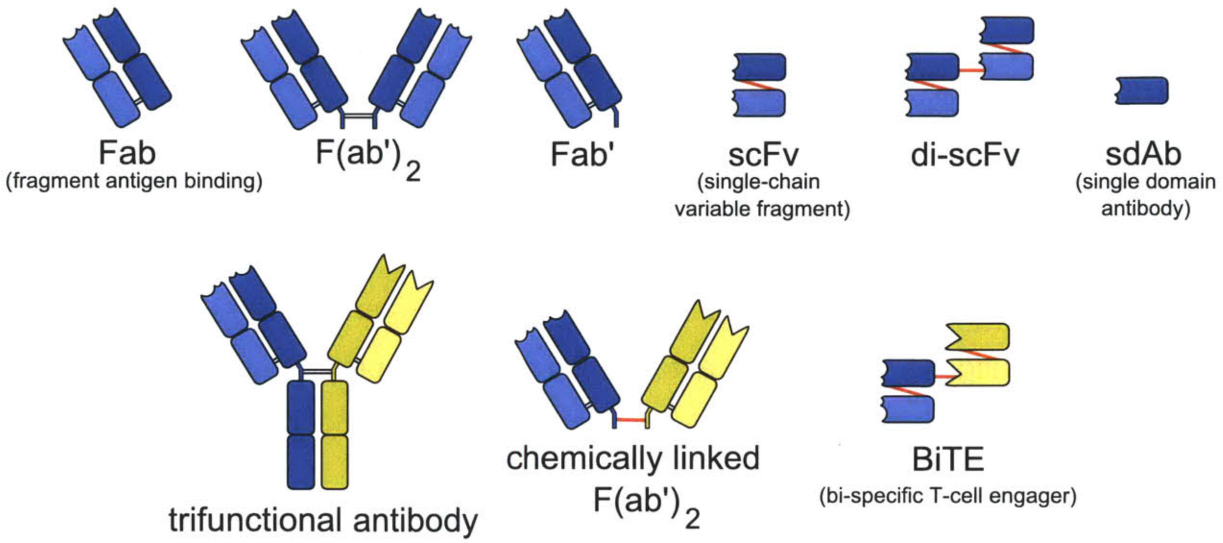

Fab
(fragment antigen binding)
F(ab')$_2$
Fab'
scFv
(single-chain variable fragment)
di-scFv
sdAb
(single domain antibody)
trifunctional antibody
chemically linked F(ab')$_2$
BiTE
(bi-specific T-cell engager)

Figure 1.2: Engineered monoclonal antibodies

Types of monoclonal antibodies with other structures than naturally occurring antibodies. Top row: monospecific antibodies (fragment antigen-binding, F(ab')2 fragment, Fab' fragment, single-chain variable fragment, di-scFv, single domain antibody). Bottom row: bispecific antibodies (trifunctional antibody, chemically linked F(ab')2, bi-specific T-cell engager). Heavy chains have a darker shade, light chains a lighter one. Parts of antibodies with different targets are colored differently. Constant regions are shown as regular round-edged boxes, variable regions as boxes with an irregularly shaped end. Artificial links between fragments are colored red. This figure is from Wikimedia Commons.

1.2 Objectives and findings of this dissertation

This dissertation presents novel approaches in the use of synthetic, defined oligonucleotide libraries in molecular display platforms, as well as their analysis by massively parallel DNA sequencing. The work described herein has been divided into two parts reflecting the above discussion: 1) The development of a rationally designed antibody library compatible with deep sequencing analysis, and 2) The identification of candidate self antigens by profiling autoantibody binding specificities with a synthetic human peptidome.

1.2.1 Antibody library engineering

Chapter 2 of this thesis approaches a set of important problems faced by the antibody engineering community. Adapting deep sequencing technologies to naive human repertoires is problematic for two main reasons. The first challenge has to do with analysis of enriched repertoires. Illumina sequencing routinely obtains ~10^8 "short" (currently up to 100 nt, soon to be up to 200 nt) reads and this level of complexity is well suited to the analysis of highly diverse antibody libraries. However, sample preparation and sequencing analysis is made complicated by the diversity of naturally occurring variable domain genes, since amplification and sequencing requires a complex set of primers. The second challenge to be overcome is the recovery of desirable clones identified by sequencing. Recovering a unique clone from a mixture of complex variable domains can only be accomplished by complete synthesis (expensive), hybridization capture (technically challenging), or by PCR rescue and variable domain re-assembly strategies (complicated). Both of these challenges can be circumvented by using a single variable domain ("single framework") antibody library, which is the focus of Chapter 2. The limitation of single framework libraries is their generally inconsistent performance due to lack of framework diversity. We therefore set out to design a set of optimized CDR sequences to maximize the functionality of a single framework antibody library.

We developed the concept that sequence biases inherent to high affinity, antigen binding hypervariable loops can be captured by a contact/noncontact two state hidden Markov model (HMM). Indeed, this model not only recapitulated much of what was already known about the structure-function of CDRs, but is also a source of novel predictions. In addition, we report a method that can be utilized to mimic junctional diversity using type IIS restriction enzymes and shuffling ligation. Such combinatorial techniques are important in the context of fully defined oligo libraries, as their complexities tend to be on the order of 10^4 - 10^5 sequences.

The HMM scFv library was subjected to four rounds of enrichment on the PVRL4 cancer antigen, and the library population dynamics were monitored by deep sequencing using a highly streamlined protocol. Analysis of the sequencing data allowed us to identify candidate scFvs that could be PCR-rescued for clonal expression in a scalable, two step procedure. The HMM scFv platform described in Chapter 2 can serve as a template for further improvements in the integration of designed, synthetic oligo libraries and deep sequencing-assisted antibody production.

1.2.2 Antibody profiling

In a separate set of studies, we applied massively parallel DNA synthesis and sequencing to the interrogation of individual human antibody repertoire binding specificities. Chapter 3 describes our proof-of-principle work using programmable microarray-derived oligonucleotides to encode an unbiased collection of phage-displayed 36 amino acid peptides that together span the entire open reading frame of the human genome ("T7-Pep"). This is a valuable, and renewable resource that we have shared with the community. We optimized protocols to enable single-round autoantigen enrichments that were detectable via deep sequencing analysis of the phage library ("PhIP-Seq").

T7-Pep compared favorably to alternative cDNA-based libraries. Whereas typical cDNA libraries are dominated by sequences not normally expressed, we found 83% of T7-Pep library members to express full-length peptides in the correct reading frame. As a uniform representation of the proteome (78% of the library within 10 fold abundance), the library can be used to screen any autoimmune disease cohort (regardless of target tissue specificity) and inter-experiment data is immediately comparable. Our pilot study analyzed the autoantibody repertoire found in the cerebrospinal fluid of patients suffering from paraneoplastic neurological disorder. PhIP-Seq with T7-Pep rediscovered a control patient's known autoantigen (NOVA1) with very high confidence, in addition to several novel autoantigens, including a putative cancer-testis antigen, TGIF2LX. We also detected GAD65 autoantibodies that were not capable of recognizing the fully denatured antigen, revealing that the library of 36-mer peptides retains an important degree of conformational information.

The technology was further developed in Chapter 4 to enable the first high throughput, low cost screen of a large number of individual patients with different autoimmune diseases for comparison to each other and to their healthy counterparts. By adapting automation and 96-plex DNA barcoding to the PhIP-Seq protocol, we performed a proteomic-scale assessment of autoreactivities found within a collection of 289 individual antibody repertoires. Several important autoimmune diseases were represented in this collection, and we were thus able to search for novel disease-associated autoantigens in type 1 diabetes (T1D), rheumatoid arthritis (RA), and multiple sclerosis (MS). The T1D patients we screened had previously been evaluated for their autoantibody status, which allowed us to determine that PhIP-Seq has a relatively high false negative discovery rate. Despite this finding, however, we were able to rediscover the important islet-specific antigen and biomarker, PTPRN (IA-2), in addition to potentially novel autoantigens, which are now undergoing further evaluation via radioimmunoassay. Our collection of MS patients' CSF and sera also provided rich data, allowing the rediscovery of reported epitope motifs and the generation of new candidate autoantigens. Interestingly, the screen data from the healthy controls revealed the existence of recurrent, likely benign, autoantibodies present in a large fraction of the population. In addition to shared autoantibodies, the dataset demonstrates that each individual possesses a unique autoantibody repertoire and the extent to which it shapes our phenotype will only come to light with continued screening.

2. Development of a Rationally Designed Antibody Platform

Collaborator affiliations and contributions

H. Benjamin Larman[1,2,3,*], George J. Xu[1,3,4,*], Natalya N. Pavlova[3] & Stephen J. Elledge[3]

[*] These authors contributed equally to this work.

[1] Harvard-MIT Division of Health Sciences and Technology, Cambridge, MA, USA

[2] Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA

[3] Department of Genetics, Harvard University Medical School, and Division of Genetics, Howard Hughes Medical Institute, Brigham and Women's Hospital, Boston, MA, USA

[4] Biophysics Program, Graduate School of Arts and Sciences, Harvard University, Cambridge, MA, USA

Special thanks to Uri Laserson for helpful discussions about how to model the contact biases in the antibody-antigen structure database.

S.J.E. conceived and supervised the project. H.B.L. designed and constructed the HMM scFv library and optimized the ribosome display protocol. G.J.X. performed the selections, analyzed the sequencing data, and performed candidate validation studies. N.N.P. provided the PVRL4 antigen and performed the FACS analyses. This chapter was prepared by H.B.L. and edited by G.J.X., N.N.P. and S.J.E.

2.1 Abstract

Antibody discovery platforms have become an important source of both therapeutic biomolecules and general purpose affinity reagents. Massively parallel DNA sequencing can be used to assist antibody selection, thereby greatly expanding the potential of these systems. We have constructed a rationally designed scFv library and analysis platform, which is optimized for use with short read deep sequencing technologies. Libraries of fully defined oligos encoding three complementarity determining regions (CDRs; L3 from the light chain, H2 and H3 from the heavy chain) were combinatorially cloned into a single, synthetic scFv framework for molecular display. Our novel CDR sequence design utilized a hidden Markov model (HMM) that was trained on all antibody-antigen co-crystal complexes present in the Protein Data Bank. The resultant ~10^{12} member library has been produced in ribosome display format, and was comprehensively analyzed over four rounds of antigen selections by multiplex paired-end Illumina sequencing. The HMM library generated multiple antibodies against an emerging cancer antigen and is the basis of a next generation antibody production platform.

2.2 Introduction

Antibodies are useful for their ability to bind molecular surfaces with incredible specificity. The genetic basis for their structural diversity is partially found in the germline, and partially the result of stochastic genetic events, including chromosomal rearrangements, non-templated nucleotide insertions, and hypermutation. The vast majority of this diversity is localized to the Complementarity Determining Regions (CDRs), which are the six peptide chains that protrude from the variable domain framework to form the antigen binding surface of the antibody molecule. Three CDR loops are contributed by the heavy chain (H1, H2, and H3) and three by the light chain (L1, L2, and L3), all six of which come together to form the antigen combining surface. CDRs 1 and 2 are encoded in the germline, and thus more constrained in their diversity. L3 is characterized by "junctional diversity," formed during the recombination of two gene segments (V and J). Finally, H3 is formed by two consecutive genetic rearrangements (first between D and J, and then between V and DJ), and is accompanied by the addition of non-templated "N" nucleotides, making this CDR the source of most naturally occurring antibody diversity.

Our goal was to develop an antibody production platform that could be seamlessly integrated with massively parallel DNA sequencing analysis. We reasoned that for this to be the case, neither library amplification nor sequencing reactions should depend upon the complex mixture of primers necessary for amplification and analysis of the naturally occurring heavy and light chain variable domains. Importantly, however, the natural diversity of variable domain framework regions contributes significantly to the "shape space" of a natural antibody repertoire, despite that fact that CDRs are considered the major determinant of antigen combining site topology.[39] Indeed, single framework libraries have generally not performed consistently when tested on a diverse set of antigens.[40] We therefore focused our attention on maximizing the functional diversity in our library's CDR repertoire.

Our first step was to identify a suitable framework into which we could combinatorially insert libraries of rationally designed CDRs. Lloyd et al. screened a very large pre-immune human scFv library against a panel of 28 different antigens, and after sequencing >5,000 post-selection clones, they found a strong enrichment for a small subset of heavy and light chain variable domains.[41] Among them, the most highly enriched were V_H1-69 and the lambda V_L1-44. They attributed these framework enrichments to increased expression and optimal folding within the periplasm of the E. coli host cells. These findings were further corroborated by the work of Glanville et al.[42] who found that of the subset of V_H chains tested, V_H1-69 was the most successful in generating binders against a panel of 16 different antigens. We therefore housed our CDR libraries within a V_H1-69, V_L1-44 framework.

As a source of inspiration for CDR design features, we turned to IMGT's annotated database ("IMGT/3Dstructure-DB") of all antibody-antigen co-crystal structures present within PDB as of May, 2009.[43, 44] A similar database was constructed by Schlessinger et al. in 2006.[45] Amino acid residues within CDRs can contribute to antigen binding via two distinct mechanisms. The first is direct, via contribution of a side group that contacts the antigen (Figure 2.1). The second role is indirect, affecting the conformation of the peptide backbone in a way that permits the direct interaction of neighboring amino acid side groups. This behavior of CDR amino acid sequences can be captured in a two state hidden Markov model (HMM). The “contact” state should be enriched for amino acids capable of sharing/exchanging electrons or of helping to bury hydrophobic surfaces, whereas the “noncontact” state should be enriched for amino acids capable of appropriately constraining or relaxing the CDR polypeptide backbone. One feature of all HMMs is that the state of each position depends upon its nearest neighbor. It is thus important to note that traditional approaches to synthetic scFv library construction utilize degenerate nucleotides or codons, and so can at best control only the composition of one amino acid position at a time. We took a novel approach, which was to encode complete HMM-generated CDR sequences as releasable ssDNA, which were synthesized on a silicon wafer. An additional advantage of specifying complete sequences is the ability to filter out deleterious properties, such as restriction sites and unwanted peptide motifs (e.g. glycosylation, immunogenic, etc). These HMM CDR libraries were subsequently combinatorially cloned into the single V_H1-69, V_L1-44 framework.

The transformation efficiency of bacterial cells with plasmid DNA is a significant barrier to constructing molecular libraries of a complexity greater than $\sim 10^{10}$. Since the utility of an scFv library scales with its diversity, we took advantage of the *in vitro* ribosome display technique which has been used to generate antibodies with picomolar affinities.[46] In this approach, mRNA molecules are tethered to the proteins they encode via noncovalent interactions with a ribosome. The mRNA is made to lack the stop codon necessary for transcript release, and so a population of ternary complexes composed of mRNA, encoded scFvs, and ribosomes are thus formed. The libraries can be subjected to repeated rounds of selection and (optionally mutagenic) re-amplification to enrich for scFvs that bind to a target antigen (Figure 2.2A).

After characterizing the quality of the HMM scFv library, we tested our system by sequencing the library as it evolved over multiple rounds of selection on a protein antigen. We also developed robust methods to specifically rescue desirable clones for expression and analysis in a simple two-step process. Our platform successfully produced antibodies against the emerging cancer antigen PVRL4, and sets the stage for a new paradigm in sequencing-assisted selection of rationally designed human antibodies.

2.3 Results

2.3.1 Library design, assembly and characterization

We set out to diversify the three CDR loops most relevant to antigen binding. By examining the IMGT/3Dstructure-DB, we determined the average number of contacts per structure contributed by each CDR. Of contacts reported in this database, 76% were contributed by residues contained in the CDRs. As expected, L3 and H3 contributed the most contacts, with H2 providing the third most. In sum, 71% of all CDR contacts were made by amino acids in these three CDRs (Figure 2.1A).

To estimate the HMM-defining parameters for L3 and H3, we identified 236 unique L3 and 241 unique H3 sequences, each residue of which was classified as either making contact or not with the protein antigen. In the IMGT nomenclature, L3 and H3 sequences from position 105 to 117 were used to train the model. Finally, since position 118 frequently contributes contacts, but is outside the defined hypervariable domain, this residue was randomly assigned according to its frequency of occurrence in the database, regardless of its contact status.

The resulting HMM state transition rates and amino acid emission probabilities for L3 and H3 are illustrated in Figure 2.2B and 2.2C. Notable features of these models are: 1) enrichment for the noncontact state at positions closer to the framework (i.e. probability of S (start) --> N (noncontact) and N --> E (end) transitions are greater than S --> C (contact) and C --> E, respectively); 2) in H3, a tendency for blocks of contact/noncontact states (i.e. probability of staying in the same state is higher than transitioning between states); 3) a strong enrichment in both L3 and H3 for contacts consisting of tyrosine and tryptophan (reported by Ofran et al.[47]), and 4) L3/H3-specific enrichments for certain amino acids in each state (e.g. noncontact proline in L3, and contact glutamic acid in H3).

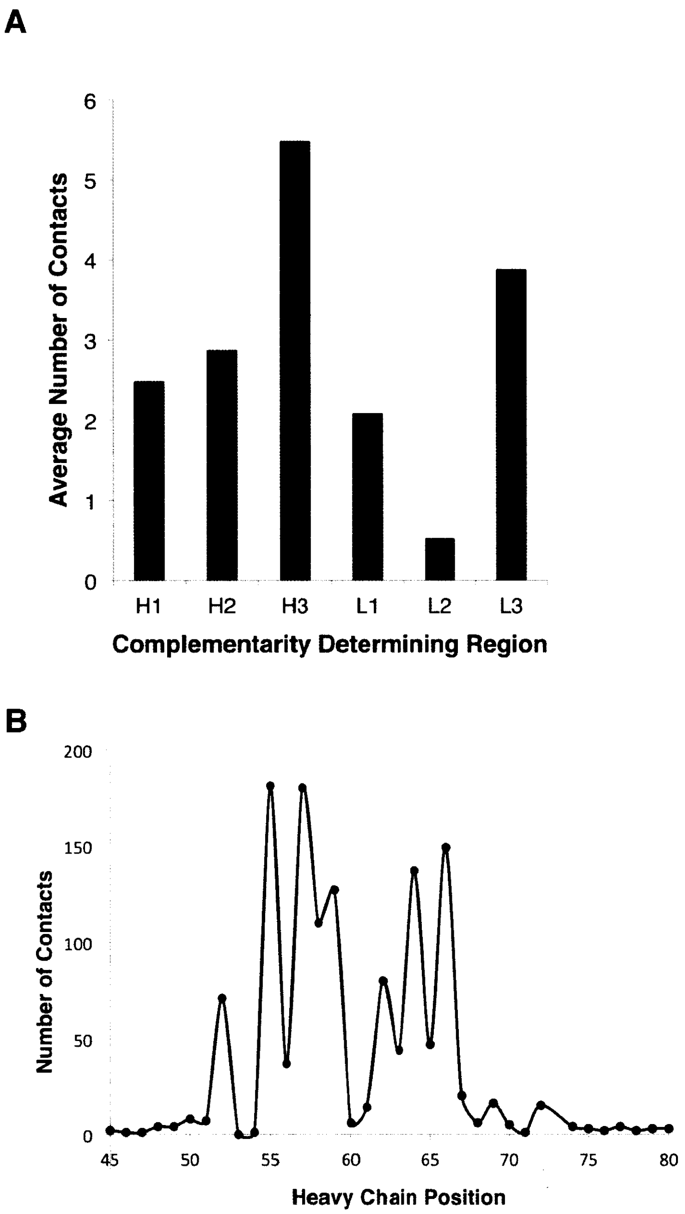


Figure 2.1: CDR contact distribution and H2 contact profile.

A. Contacts reported in the IMGT/3Dstructure-DB database. Contact assignment is based on IMGT definition of CDR positions. Data was obtained from 241 antibody-antigen co-crystal structures.

B. Position-dependent contact distribution in H2. Valleys represent amino acids more likely to play a role in framework stability.

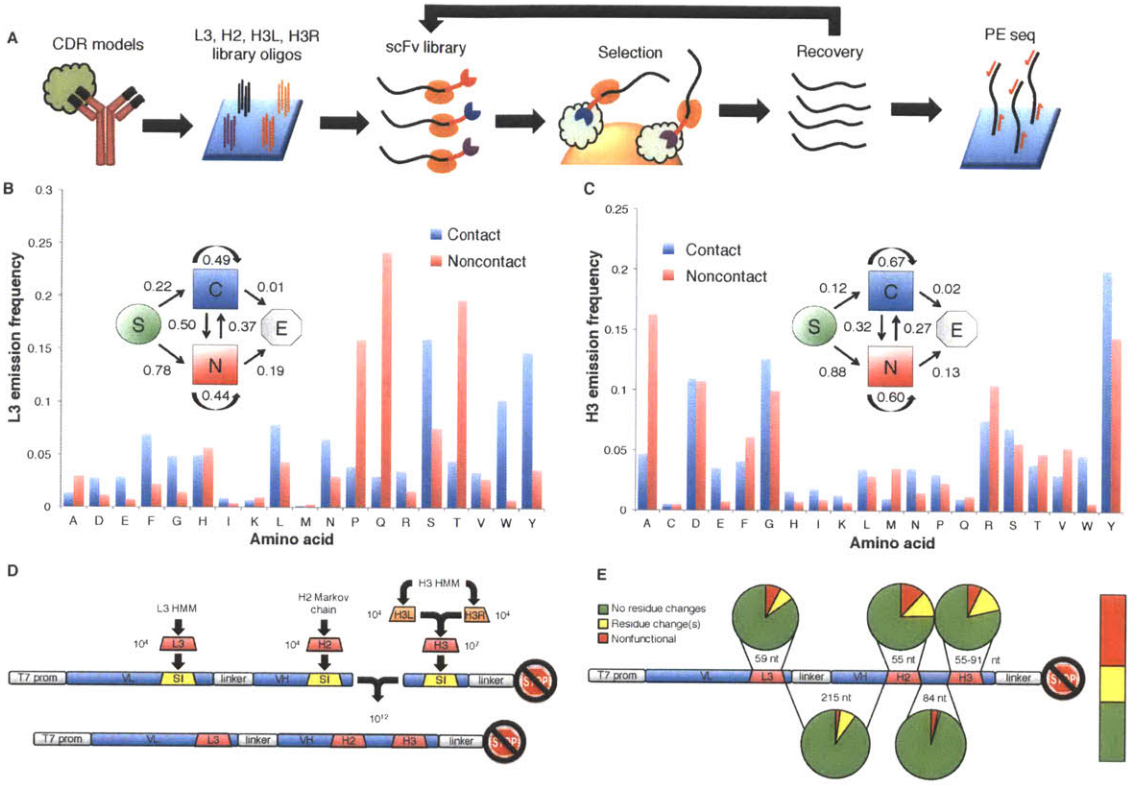


Figure 2.2: HMM antibody library design and synthesis

A. Strategy for design and assembly of the rationally designed scFv library for display on ribosomes. After enrichment for antigen binding clones, library recovery and/or analysis by paired end sequencing can be performed.

B. Model defining parameters for the H3 HMM. Probability of emission for each amino acid corresponding to the two possible states. State transition probabilities are inset: "S" denotes start of a chain, "C" denotes the contact state, "N" denotes the noncontact state, "E" denotes the end of the chain.

C. Model defining parameters for the L3 HMM. Definitions are the same as for **B**.

D. Overview of the scFv display vector and library assembly strategy. "VL" and "VH" are the light and heavy variable domains, respectively. "T7 prom" is the T7 promoter, and the crossed stop sign denotes lack of a stop codon. L3, H2 and H3 are the CDR libraries designed to replace the "SI" suicide inserts. H3L and H3R sublibraries are brought together by shuffling ligation to create H3. Similarly, the L3-H2 fragment is brought together with the H3 fragment in a shuffling ligation.

E. Clonal Sanger sequencing analysis of 93 HMM scFv library members.

We used our HMM to generate >10,000 unique sequences for each of L3 and H3.[48] Whereas the length of L3 sequences was fixed at 13 residues, 1,000 H3 sequences were randomly chosen for each length from 9 to 21 amino acids long. As an analog to VJ recombination, we further expanded the diversity of H3 by separating each sequence into two halves: "H3L" and "H3R", for subsequent combinatorial ligation to form full length H3 sequences (Figure 2.2.D). This was accomplished by placing a type IIS restriction site downstream of H3L and upstream of H3R on their encoding oligos. After PCR, restriction digest and purification, fragments with a 3 nt 5' overhang were brought together for ligation. The reading frame, as determined by H3L, would therefore always remain intact.

The germline-encoded H2 CDR is characterized by structural features not present in L3 or H3 chains, and this is reflected in its heterogeneous contact profile (Figure 2.1B). It has been suggested that H2 contributes to the stability of the variable domain of the heavy chain through interactions among its hydrophobic residues.[49, 50] In order to avoid disrupting stability, we created a first order Markov chain to generate framework-compatible H2 sequences based on the 176 H2-unique IMGT chains. This model was used to generate >10,000 H2 sequences.

Finally, all CDR sequences were passed through a series of three filters in order to maintain their utility. First, all restriction sites to be used during library construction were eliminated by introducing silent codon changes. Second, we sought to minimize the potential immunogenicity of the scFvs by discarding peptides with a high potential for loading onto MHC class II molecules during antigen presentation. We used the ProPred online server to filter our CDR sequences against the four most common HLA-DRB1 alleles (101, 301, 701, and 1501) with a stringency of 45% of the best substrate.[51] This resulted in replacement of about 18% of all H3 sequences by theoretically less immunogenic peptides. The third filter replaced sequences with the potential to interfere with industrial scale production (e.g. methionine oxidation, asparagine deamidation/cyclization), as well as glycosylation motifs.

The final set of 43,803 CDR sequences (L3, H2, H3L, H3R) were flanked by the appropriate restriction sites, as well as sublibrary-specific PCR primer binding sequences, and then synthesized as releasable oligonucleotides on a silicon wafer (Agilent Technologies). The oligo libraries were PCR amplified and cloned into the V_H1-69 and V_L1-44 human heavy and light chain variable fragments and assembled as shown in Figure 2.D (described further in the detailed Methods). In vitro transcription was then performed to create the mRNA template for ribsome display and library selection.

We characterized the HMM scFv library in two ways. First, we cloned a small sample of the library mRNA. This allowed us to perform Sanger sequencing on individual colonies, and thereby estimate the overall fraction of the library expected to contain functional, full length scFvs with no frameshift or nonsense mutations (57% functional, n=93; Figure 2.2E). None of the colonies examined had retained their CDR "suicide insert", and none had multiple CDR insertions. Second, we used our Illumina sequencing data to characterize the length distribution of the H3 loop (Supplementary Figure 2.1). Satisfied that our library was true to its design, we next performed selections against an emerging cancer antigen, PVRL4,[52, 53] and used Illumina sequencing to track the library during selection.

2.3.2 Affinity selections on a protein antigen

We utilized a positive control scFv and bait pair to develop a robust ribosome display selection protocol. These could each be diluted into negative control, nonspecific scFv and nonspecific bait so that enrichment could be monitored while experimental selection parameters were varied. Pluckthun et al. have used ribosome display to affinity mature an scFv (4c11L34Ser, "Pluck-scFv") to high affinity binding (Kd = 40 pM) to a peptide derived from the yeast GCN4 protein (RMKQLEPKVEELLPKN**YHLENEVARLKK**-LVGER, epitope in bold).[54] Our eventual goal was to perform selections on GST-fusion proteins, and so we synthesized DNA to encode the GCN4 peptide, and recombined this into the pDEST15 vector for inducible expression of GST-GCN4 in BL21 E. coli cells. As a negative control scFv, a random clone ("rand-scFv") was picked from a naive human repertoire[55] and expressed in the same ribosome display vector backbone. A negative control peptide, "GST-pep" in the same pDEST15 vector backbone, was used as nonspecific bait. A series of protocol refinement experiments were undertaken to maximize the degree of both enrichment and recovery of the Pluck-scFv that we could attain. For most experiments, Pluck-scFv was diluted 1,000 fold into a background of rand-scFv, while GST-GCN4 was diluted 1,000 fold into a background of GST-pep. Using a Pluck-scFv specific TaqMan probe, enrichments of several hundred fold were routinely obtained. Despite protocol optimization, however, Pluck-scFv recovery efficiency tended to be ~0.2% when GST-GCN4 was present at 100%, most likely reflecting a limitation inherent to ribosome display technology.

In addition to the optimized ribosome display selection protocol, we incorporated a system of quality control measures to ensure that each round of selection was a success. First, we spiked Pluck-scFv into our HMM scFv library and GST-GCN4 into our selection bait, GST-PVRL4, both at a dilution of 1:1,000. In this way, the efficiency of enrichment and recovery for each selection could thus be monitored using TaqMan qPCR probes specific to either Pluck-scFv or to the HMM scFv library. If these measures were below a threshold, then the selection was considered a failure and repeated. Second, degradation of mRNA transcripts is a concern with ribosome diplay, and so we utilized two distinct TaqMan probes targeting either the 3' or 5' ends of the transcript. In the absence of degradation, these two signals arise with equal strength. The 5' signal is differentially diminished by degradation, and so the ratio of the two signals can be used as a proxy for the degree of degradation that occurred during the selection. If the 5':3' signal ratio was below our threshold, the selection was considered a failure and repeated.

2.3.3 Analysis of selected HMM scFv libraries

Four successive rounds of ribosome display selection were performed with the HMM scFv library on GST-PVRL4. By the fourth round of selection on PVRL4, we noticed an increased amount of HMM scFv library recovery, presumably due to the accumulation of binders. To identify non-PVRL4 component binders, the third round selection library was additionally selected on GST only (no PVRL4) so that we could discriminate between scFvs that bind to PVRL4 and those that bind to GST or to some other component of the system.

The minimal region of the HMM scFvs that contains the three diversified CDRs is an appropriate size for analysis by paired end Illumina sequencing, thus these libraries can be conveniently prepared by PCR. A small amount of material from each of the selected libraries, as well as from the starting HMM scFv library were amplified with Illumina sequencing adapters. These adapters include a 7 nt barcode for library identification, thus permitting the multiplex analysis of many different libraries. We analyzed the input and selected libraries after each round of enrichment by pooling these barcode-containing libraries for multiplex sequencing. Because the complexity of the libraries is expected to decrease significantly with each round of selection, we divided the contribution of each library by two for each round of enrichment undergone. For example, if we added 100 ng of input library product to the multiplex pool, then we would add 50 ng of round 1 selected library, 25 ng of round 2 selected library, 12.5 ng of round 3 selected library, and so on.

Our strategy was to perform paired end sequencing in two separate Illumina Hi Seq 2000 flow cell lanes, such that by sequencing L3-H3 pairs in one lane and H2-H3 pairs in the other lane, we could use the extreme diversity of H3 sequences to unambiguously match corresponding L3 and H2 sequences, thereby reconstructing each complete scFv clone (Figure 2.3). We observed a significant degree of PCR chimerism to occur during PCR amplification, which complicated, but in most cases did not prevent the reconstruction of individual scFv clones. Importantly, CDR recombination has been observed to significantly increase scFv affinity during ribosome display selection.[56]

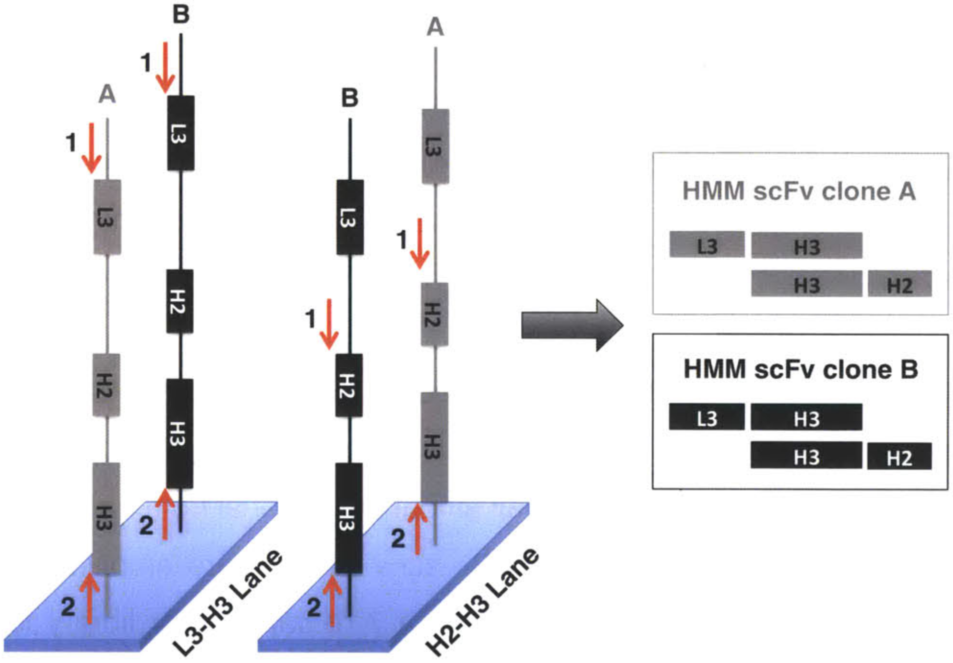


Figure 2.3: Strategy for sequencing reconstruction of HMM scFv clones

100 nucleotide paired end sequencing is performed on the same library in two independent lanes on an Illumina HiSeq 2000. In the "L3-H3" lane, the first sequencing primer lands upstream of L3. In the "H2-H3" lane, the first sequencing primer lands upstream of H2. The H3 sequence is then determined by reading from a common, second primer. L3 and H2 sequences are then paired using their unique H3 identifier to fully define the scFv clone.

We next determined the relative abundance of each clone in the library over the course of four rounds of selection on GST-PVRL4, and compared this to the results of a round 3 PVRL4 selected library that was selected on GST alone (Figure 2.4). The most abundant clones in the library after four rounds of selection all displayed behavior indicative of PVRL4 candidate antibodies, and so a subset of these were rescued from the library for further analysis.

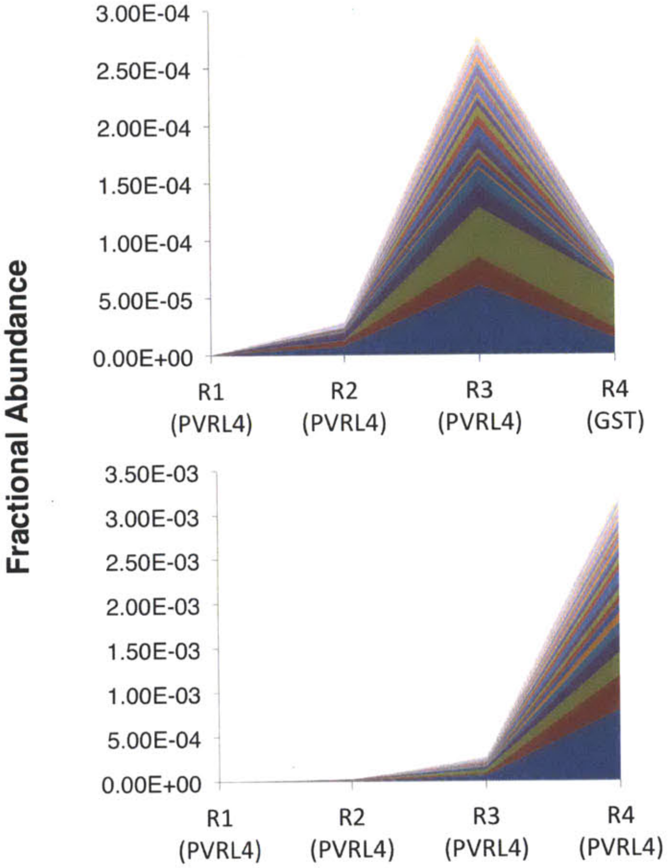


Figure 2.4: Fractional abundance of top 30 HMM scFv clones over 4 rounds of selection

The relative abundances of the 30 most abundant HMM scFv clones after 4 rounds of enrichment on PVRL4 are shown. The data in the first three selections are the same between the two panels.

2.3.4 Binding properties of candidate HMM scFvs

Before characterizing individual scFvs for their ability to bind antigen, they must be cloned. This can be done either by re-synthesizing the CDRs for cloning back into an expression framework, or alternatively by PCR-rescuing the clones using forward and reverse primers specific for L3 and H3, respectively. We chose to recover candidate scFvs by performing PCR with L3/H3-specific primers, which also contained 5' homology arms for subsequent isothermal assembly into an epitope-tag expression vector (Figure 2.5A).

Rescued candidate anti-PVRL4 scFv clones were expressed in vitro as FLAG-tagged proteins. Three of the 25 tested clones were found to have human mammary epithelial cell (HMEC)-expressed PVRL4 binding properties by FACS analysis (Figure 2.5B). The binding affinity of these scFv clones will soon be determined by SPR analysis on a Biacore 3000 instrument.

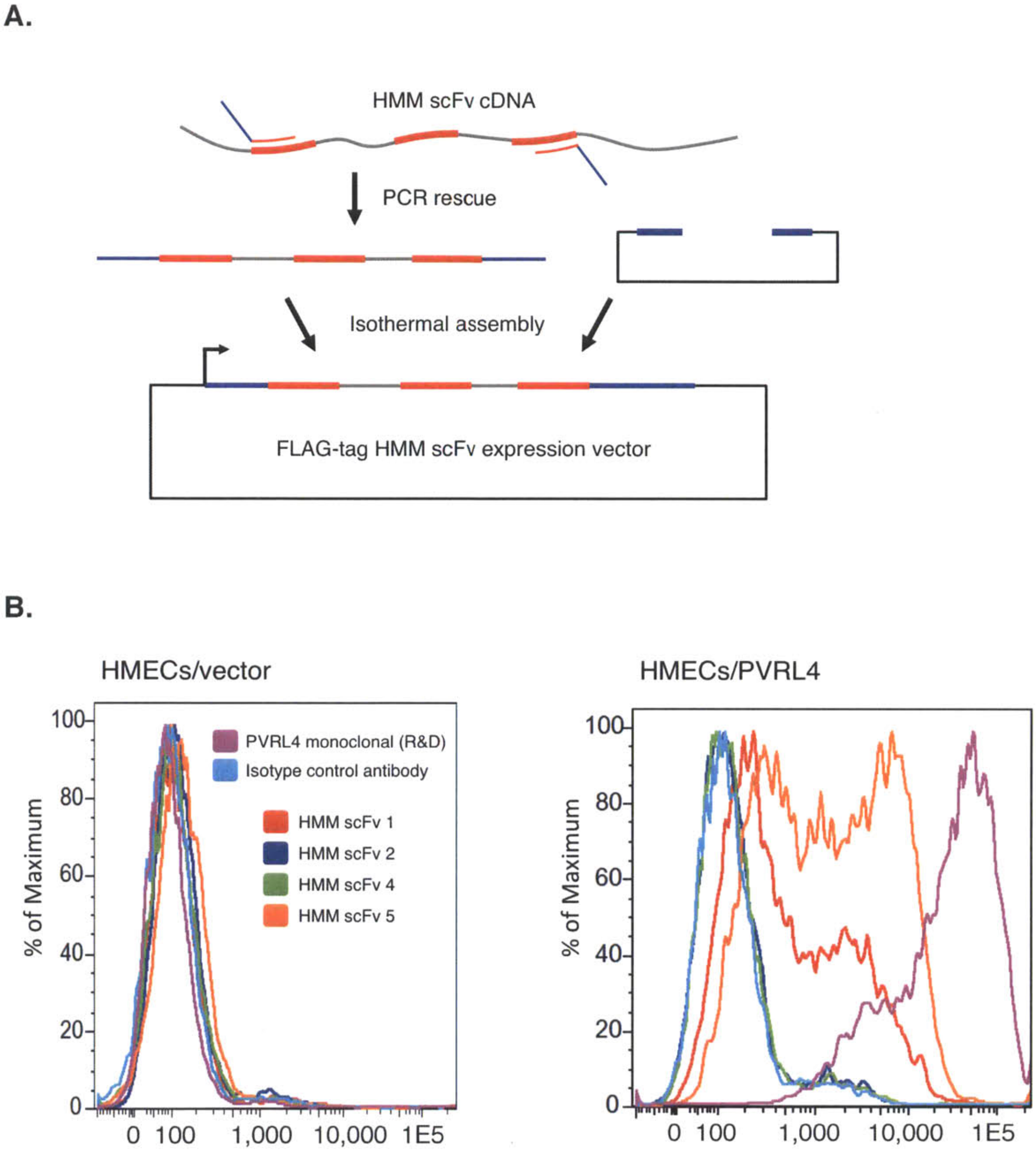


Figure 2.5: HMM scFv rescue strategy and FACS validation

A. Candidate HMM scFv clones are PCR rescued with primers specific for L3 and H3, which also have 5' homology arms for subsequent isothermal assembly into an expression vector with differing codon usage. **B.** Results of FACS experiment to assess binding of candidate scFvs. PVRL4 antigen was overexpressed on HMECs and then stained with control antibody or HMM scFv.

2.4 Discussion

The promise of synthetic biology has yet to deliver antibody production platforms that rival vertebrate immune systems in both product quality and manufacturing convenience.[57] There exist many successful examples of synthetic antibody production pipelines that meet specific industrial needs, most notably in the production of fully human IgG molecules, but these boutique solutions are not widely affordable/accessible.[31] However, we anticipate that along with the maturation of gene synthesis technologies and the affordability of DNA deep sequencing, will also come advances in antibody production pipelines that outperform animal immune systems in all regards. Our work and others provide early evidence that this potential can eventually be brought to fruition.

The development of robust single framework scFv libraries is key to their convenient analysis with massively parallel DNA sequencing, and therefore to the more widespread adoption of synthetic approaches. One immediate benefit of working with single framework libraries is that the entire scFv sequence can be reconstructed from two pairs of short sequencing reads. We found that the hyperdiversity of our H3 CDR library permitted the near unambiguous pairing of L3 and H2 sequences with their shared H3, thus completely defining the repertoire at each round of selection. Another important advantage of single framework libraries is the relative ease with which desirable clones can be rescued. Our method provides a two-step recovery protocol for clone amplification and assembly into a common expression vector. In contrast, combinatorial sets of heavy and light chain frameworks require many more steps for the PCR rescue and reconstruction of library clones.

Massively parallel analysis of evolving antibody repertoires enables several technical innovations not possible with traditional clonal analysis techniques.[58] One powerful application, the discovery of rare binders, was recently demonstrated by two different groups. Ravn et al. used deep sequencing to discover valuable, low abundance, high affinity scFvs, and they demonstrate that these clones are often lost during purifying selections.[59] Zhang et al. took this one step further and performed selections on complex bacterial cell surfaces overproducing a target antigen. They report a method to isolate rare scFvs from the resulting phage populations, which are necessarily dominated by a large number of off-target binders.[60]

An exciting, but as yet unproven application of deep sequencing-facilitated selections is the production of scFv sets targeting multiple antigens in parallel using deconvolution strategies. In one embodiment, an "array" of antigens can be pooled by rows and columns, so that scFvs specific to both a particular row and a particular column can be associated with the single antigen at their intersection. This strategy reduces the number of selections to the square root of the number of antigens. This can be made even more efficient by performing initial rounds of selection on the entire antigen collection in a single pool, prior to selections on the row and column pools. Whereas a collection of 100 antigens would individually require 400 selections (assuming 4 rounds are generally sufficient to identify scFvs by deep sequencing), the same could be accomplished with only 23 selections using pooling strategies (i.e. the first three rounds are performed on a single super pool, and the last on row/column subpools). Multiplex sequencing of the 20 post-selection libraries would permit the rapid identification of antigen-specific scFvs. Future single-pot, massively parallel selections will require the development of robust library-versus-library deconvolution strategies. Preliminary progress has recently been reported.[61]

Another interesting prospect for scFv display technologies in general, which is certainly not possible with traditional affinity reagents, is their usage in a highly multiplexed context. For example, one may wish to study expression levels of a set of interleukin receptors on individual T cells. A defined set of anti-ILR scFv-displaying phage could be used to probe the cell surfaces, and their relative abundance subsequently measured by deep sequencing. This example also highlights the utility of DNA-coupled scFv probes, as they can be PCR amplified from an otherwise undetectable level.

For these reasons we have produced a novel synthetic scFv library within a single framework, which was chosen for its proven performance in producing high affinity antibodies. To compensate for a potential loss of framework-contributed structural diversity, we invested a great deal of effort in the rational design of CDR sequences. Our design is based on a mathematical model that captures subtle amino acid sequence biases that contribute to the formation of good antigen contacts. Many features inherent to this model have been observed by others,[62-65] whereas additional, novel features may provide new implications. In terms of combinatorial complexity, we also developed a method to mimic the junctional diversity of VJ recombination using type IIS restriction cleavage followed by shuffling ligation. Finally, we have introduced a strategy for paired end sequencing-based reconstruction of full length scFv populations. As more sophisticated selection/deconvolution strategies emerge, we anticipate that rapid, low cost production of high quality synthetic scFvs will finally become a reality.

2.5 Methods

Construction of the ribosome display vector

Plückthun and others have optimized vectors capable of reliably accomplishing ribosome display of scFvs.[66, 67] We have adapted components of these and other such vectors to our present purpose. Beginning from the 5' end of the DNA vector, the following parts were assembled as a synthetic gene product (DNA2.0).

1) T7 promoter for in vitro transcription from the DNA library (TAATACGACTCACTATAGGGAGACCACAACGGTTTCCC)
2) 5' mRNA stemloop (5'-GGGAGACCACAACGGTTTCCC-3') to improve transcript stability
3) Ribosome binding site for translation of the library
4) Kozak sequence for potential use in eukaryotic translation systems
5) N-terminal 6xHis tag for detection and potential purification of scFv protein
6) The variable domain of the light chain was encoded N-terminal to the heavy chain so that PCR recovery of the three diversified CDRs (L3, H2, H3) would require the shortest amplicon. (Description of the heavy and light chain sequences are described in the next section.)
7) Between the N-terminal variable light chain (V_L) and C-terminal variable heavy chain (V_H) is a "$(G_4S)_3$" linker with optimized codon usage (5'-ggtggtggtggttctggtggtggtggttctggcggcggcggctccagtggtggtggatcc-3')
8) The C-terminus of V_H is fused to a linker segment derived from the TolA E. coli protein (accession: NP_415267, position 131-214), which provides a spacer between the displayed scFv and the ribosomal tunnel.
9) 3' mRNA stemloop (5'-CCGCACACCTTACTGGTGTGCGG-3') to improve transcript stability

NotI sites flank the 3' and 5' ends of the construct for isolation of the in vitro transcription template. Directional SfiI sites flank the minimal scFv for facile movement of clones into and out of daughter vectors.

HMM scFv library assembly I: framework

We wished to use the J chains most commonly associated with the V_H1-69 and V_L1-44 segments. In a sequenced heavy chain repertoire from an individual, IGHJ4 was the J chain most often recombined with V_H1-69 (Laserson, Church et al. unpublished). We used work by Schofield et al. to determine that in a large pool of selected phage, IGLJ2 was the J chain that most often recombined with V_L1-44.[68] These components were assembled and reverse translated into an E. coli codon preference (Table 2.1).

We introduced silent mutations into the framework regions flanking L3, H2, and H3, for the purpose of cloning in the CDR libraries. We required that at least one of each of these pairs be non-palindromic so as to eliminate the possibility of getting multiple CDR insertions during library cloning. To this end, we introduced a BbsI site 5' and an Acc65I site 3' of L3, a PflMI site 5' and an ApoI site 3' of H2, an AccI site 5' and a BstEII site 3' of H3. These pairs of cloning sites flanked replaceable "suicide inserts," which were designed to contain a stop codon in all reading frames to prevent ribosome display of clones retaining a suicide insert, as well as a XhoI restriction site that could be used to destroy clones with a remaining suicide insert.

Feature	AA Sequence	Nt Sequence	Remark
VL1-44	QSVLTQPPSASGTPGQRVTISCSGSSSNIGSNTVNWYQQLPGTAPKLLI YSNNQRPSGVPDRFSGSKSGTSASLAISGLQSEDEADYYC- L3_suicideInsert	CAATCTGTGCTGACCCAGCCACCGTCGGCCTCGGGTACTCCGGGTCAGCGTGT TACGATCTCCTGCAGCGGTTCTTCCTCTAACATCGGTAGCAACACGGTTAACT GGTATCAACAGCTGCCGGGCACTGCCCCAAAACTGCTGATCTACTCCAACAAC CAGCGTCCAAGCGGCGTTCCGGATCGTTTCAGCGGTAGCAAAAGCGGTACTTC CGCGTCCCTGGCGATCTCTGGCCTGCAGTCCGAAGACGAAGCGGATTATTATT GCtaataactcgagttaataactagttttaataaggtg	Framework in UPPERCASE, suicide insert in lower case
VL1-44_opt	QSVLTQPPSASGTPGQRVTISCSGSSSNIGSNTVNWYQQLPGTAPKLLI YSNNQRPSGVPDRFSGSKSGTSASLAISGLQSEDEADYYC- L3_suicideInsert	CAAAGCGTTCTGACCCAGCCTCCGTCCGCGAGCGGCACCCCGGGTCAGCGTGT TACCATTTCTTGTAGCGGTAGCAGCAGCAACATTGGTAGCAATACCGTCAATT GGTATCAGCAACTGCCGGGCACCGCACCGAAACTGTTGATCTACAGCAACAAC CAGCGCCCGAGCGGCGTCCCAGACCGTTTTTCGGGCAGCAAATCCGGTACGAG CGCCAGCTTGGCGATCAGCGGTCTGCAAAGCGAAGACGAGGCCGATTACTACT GC taataactcgagttaataactagttttaataaggtg	Framework in UPPERCASE, suicide insert in lower case
IGVL-J3	FGGGTKLTVL	TTTGGCGGCGGTACCAAACTGACCGTTCTG	
IGVL-J3_opt	FGGGTKLTVL	TTCGGCGGTGGTACCAAGCTGACGGTGCTG	
VL-VH_Linker	GGGGGSGGGGSGGGGSSGGGS	GGCGGTGGTGGTGGCTCTGGTGGTGGGGGTTCCGGTGGTGGCGGCAGCTCCGG CGGTGGTTCC	
VL-VH_Linker_opt	GGGGGSGGGGSGGGGSSGGGS	GGCGGTGGTGGTGGCTCCGGTGGCGGTGGTTCCGGTGGTGGCGGTTCGAGCGG TGGCGGCAGC	
IGHV1-69	QVQLVQSGAEVKKPGSSVKVSCKASGGTFSSYAISWVRQAPGQGLE- H2_suicideInsert- YAQKFQGRVTITADEATSTAYMELSSLRSEDTAVYYC- H3_suicideInsert	CAGGTGCAGCTGGTGCAGTCCGGTGCGGAAGTTAAGAAACCGGGTTCCTCCGT AAAAGTCTCTTGCAAGGCGAGCGGTGGTACTTTCTCCTCCTACGCGATTTCTT GGGTGCGTCAGGCACCGGGCCAAGGTCTGGAAtgatgactcgagttgatgaga tatcttgatgagTATGCGCAGAAATTTCAGGGCCGCGTAACCATCACTGCCGA TGAGGCGACTTCCACCGCCTACATGGAGCTGTCTAGCCTGCGTTCTGAAGATA CCGCTGTCTACTACTGCtgataattaattaatgactcgagtttgataagg	Framework in UPPERCASE, suicide inserts in lower case
IGHV1-69_opt	QVQLVQSGAEVKKPGSSVKVSCKASGGTFSSYAISWVRQAPGQGLE- H2_suicideInsert- YAQKFQRVTITADEATSTAYMELSSLRSEDTAVYYC- H3_suicideInsert	CAAGTGCAGCTGGTGCAGAGCGGTGCAGAGGTTAAGAAACCGGGCTCTAGCGT AAAGGTGTCTTGTAAGGCCTCCGGTGGTACGTTCAGCAGCTATGCGATTAGCT GGGTTCGCCAAGCACCGGGCCAAGGCCTGGAAtgatgactcgagttgatgaga tatcttgatgagTATGCGCAGAAATTTCAACGTGTCACCATCACCGCTGACGA GGCTACTAGCACGGCGTACATGGAACTGAGCAGCCTGCGTTCTGAGGATACGG CGGTcTACTATTGCtgataattaattaatgactcgagtttgataagg	Framework in UPPERCASE, suicide inserts in lower case
VHJ3	WGQGTMVTVSS	TGGGGCCAGGGCACGATGGTGACCGTGAGCAGC	W is position 131 and varied according to composition
VHJ3_opt	WGQGTMVTVSS	TGGGGTCAGGGTACTATGGTGACCGTCAGCAGC	W is position 131 and varied according to composition
TolA	QKQAEEAAAKAAADAKAKAEADAKAAEEAAKKAAADAKKKAEAEAAKAA AEAQKKAEAAAAALKKKAEAAEAAAAEARKKAATE	CAGAAGCAAGCTGAAGAGGCGGCAGCGAAAGCAGCGGCAGATGCGAAAGCTAA GGCCGAAGCAGATGCTAAAGCTGCGGAAGAAGCAGCGAAAAAGGCGGCTGCAG ATGCAAAGAAGAAGGCAGAAGCAGAAGCCGCCAAAGCCGCAGCCGAAGCGCAG AAAAAAGCCGAGGCAGCCGCCGCGGCACTGAAAAAGAAGGCGGAAGCGGCAGA AGCAGCAGCAGCAGAAGCAAGAAAGAAAGCGGCAACTGAA	
TaqCommF		gcagaagcagaagccg	qPCR primer/probe set for TolA (3') framework
TaqCommProbe		aaaagccgaggcagccgc	
TaqCommR		ttcagttgccgctttctttct	
scFv-LL5'-TaqF		GGGTCAGCGTGTTACGATCT	qPCR primer/probe set for 5' part of HMM framework
RDscFv-LL5'-Taq Probe		CACTGCCCCAAAACTGCTGATCTACTC	
scFv-LL5'-TaqR		GCTTTTGCTACCGCTGAAAC	
PluckTaqManPrimer_F1		gttctggtggtggtggttct	qPCR primer/probe set for Pluck control
PluckTaqManProbe		cggcggctccagtggt	
PluckTaqManPrimer_R1		ggtcctgactcctgaagctg	
preTolA Primer		GGTttcagttgccgctttctttcttg	Ribosome display vector reverse transcription primer
preT7B2 Primer		CAACGGTTTCCCTCTAGAAATAATTTTGTTTAAC	PCR1 Forward primer for remaking the library by PCR. Used with preTolA.
TolA Primer		CCGCACACCAGTAAGGTGTGCGGTttcagttgccgctttctttct	Forward and reverse primers for making IVT template from PCR1
T7B2 Primer		ATACGAAATTAATACGACTCACTATAGGGAGACCACAACGGTTTCCCTCTAGA AATAATTTTG	
LL_RTPCR1_F1		GCCCCAAAACTGCTGATCTA	PCR primers for rescue of the HMM CDRs
LL_RTPCR_R2		gcctcttcagcttgcttctg	
IS7_L3F_PE		ACACTCTTTCCCTACACGACCCTGCAGTCCgaagacga	These primers are for amplification of the CDR fragment for Illumina sequencing. x's are the 7 nts of barcode (set of 96)
IS8_H3R_PE_Multi		GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTcaccatcgtgccctggcc	
IS4_L3F_PE		AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACCCTGCA	
P7 barcoding primers		CAAGCAGAAGACGGCATACGAGATxxxxxxGTGACTGGAGTTCAGACGTGT	

Table 2.1: Vector components and sequences

Sequences of the single framework scFv construct for ribosome display. Shown are original sequences used for the screen, and the codon optimized sequences for protein expression. Real time PCR primer and probe sets are also shown.

HMM scFv library assembly II: combinatorial CDR library cloning

The CDR libraries were released from the microarray as 10 pmol of single-stranded DNA and resuspended in 200 µl water. For each sublibrary (L3, H2, H3L, H3R), 1 µl was then used as input for library-specific PCR using 1 µl Taq polymerase (Takara) according to the manufacturer's instructions (2 µM each primer). The thermal profile was

1. 95 °C 5 m,
2. 94 °C 15 s,
3. 55 °C 30 s,
4. 68 °C 15 s,
5. Go to step 2 24x,

At this point, the reaction was divided in two and primers were replenished.

6. 95 °C 5 m,
7. 94 °C 45 s,
8. 67 °C 7 m,

The H3L library PCR product was first NheI/BssHII subcloned into the pPAO2 vector.[69] About $5x10^6$ transformants were obtained and plasmid DNA collected. In parallel, H3R library PCR product was prepared. From the pPAO2-H3L plasmid pool, ~300 bp of upstream sequence was PCR amplified for subsequent size discrimination of H3L-H3R ligation product. Both pPAO2-H3L and H3R PCR products were digested with SapI for subsequent shuffling ligation of the H3L and H3R libraries by their 5' overhanging codons. High concentration T4 ligation was carried out at 15C overnight, conditions which permit mismatched ligation at a relatively high frequency. Indeed, upon sequencing a large number of H3 clones, we observed many examples of library members with unmatched codons that were ligated together, and importantly without disrupting the reading frame. After H3 ligation, the correct size product was gel purified and PCR amplified. This PCR product and the HMM scFv vector were then digested with AccI and BstEII, so that the final H3 library could replace the vector's H3 suicide insert. If only complementary codons were able to ligate together, the theoretical diversity of the H3 sublibrary would be 1.2x10^{7}. However, we frequently observed non-complementary ligation, thus increasing the expected diversity of H3. During library construction, about 10^{7} H3 clones were obtained.

The L3 sublibrary was cloned into the scFv vector at the BbsI and Acc65I sites. After electrotransformation of DH10B cells, they grew overnight on 15 cm carbenicillin plates. We harvested >10^{7} transformants by scraping, and purified their plasmid DNA. Starting with this HMMscFv-L3 library, the same procedure was then employed to replace the H2 suicide insert with the H2 library PCR product by utilizing the engineered PflMI and ApoI sites. Again, we obtained >10^{7} transformants (HMMscFv-L3-H2 library) and purified the plasmid DNA.

In order to bring together HMMscFv-L3-H2 and HMMscFv-H3 in a final ligation (Figure 2.2D), 60 µg of each of library was first digested with AccI and BbsI and the desired fragment gel purified. In a high concentration T4 ligation at 37 °C, the two fragments were concatemerized. Finally, the product was digested with both NotI (to release the desired in vitro transcription template) and XhoI (to destroy clones retaining a suicide insert) and gel purified. We recovered 2.44 µg of HMMscFv-L3-H2-H3 library DNA at the correct size, which corresponds to 3.07 pmol or 1.85x10^{12}, in theory mostly unique DNA molecules. This material was used as a template for in vitro transcription (RiboMAX Large Scale RNA Production System T7, Promega) to produce mRNA, which was subsequently isolated with TRI reagent (Ambion).

For subsequent rounds of selection, RNA was made from DNA template as above, but purified on a Qiagen RNeasy column according to the manufacturer's instructions.

Ribosome display

Buffers were based on "RD Buffer" (1L: 50 mM Tris Acetate (6.07 g), 150 mM NaCl (8.77 g), pH to 7.5 with acetic acid), which was autoclaved 15 min on liquid cycle and stored at 4 °C.

Before immobilization of antigen-GST fusion protein, MagneGST beads (Promega) were washed 3x in 1x TBST. 5 µl beads were used per IP, and beads were coated with 100 µl of bacterial lysate containing GST fusion protein mixed 1:1 with TBST. 2 µl of 1M dTT were included. Binding to occurred overnight by rotating at 4 °C.

Beads were washed 5x with buffer "RDWB+T" (RD Buffer plus 50 mM Mg Acetate and 0.5% Tween 20) and tubes were changed after every other wash. Beads were blocked in 50 µl "Selection Buffer" (RDWB+T plus 2.5 mg/ml heparin and 1% BSA and 83.3 µg/ml tRNA) plus 1 µl RNasin (Promega) at 4 °C for 2 h.

6.37 µg RNA (1 x 10^{13} RNA molecules) per 14 µl translation reaction were used. Translations were performed using the RTS 100 E. coli Disulfide kit (5 PRIME) according the manufacturer's instructions, except that the feeding solution was not used. Translation was allowed to proceed for 13 min 45 s at 30 °C. Each 14 µl reaction was immediately diluted with 96 µl ice cold Selection Buffer and 3 µl RNasin. Reactions were centrifuged 14K x g for 5 min in 4 °C centrifuge. Supernatant was then moved to a new, cold tube. 50 µl bead solution was added to the ribosome displayed scFv library, and rotated 4 h at 4 °C. Beads were washed 6 times with 500 µl ice-cold RDWB+T. Tubes were changed after every other wash.

Ribosomal complexes were disrupted after the final wash by resuspending beads in 50 µl "EB20" (RD Buffer plus 20 mM EDTA) plus 1 µl RNasin and incubating at 37 °C for 10 min. Released RNA was then cleaned up on Qiagen RNeasy column according to instructions, and eluted into 33 µl nuclease free H2O at max speed 1 min.

Superscript III kit (Invitrogen) was used to reverse translate the selected RNA library from the preTolA primer according to manufacturer's instructions. 1 µl (5 U) of E. coli RNAse H (NEB) was added and incubated at 37 °C for 20 min.

cDNA recovered after selection was first PCR amplified using primers that flank an insert region containing the CDR's (LLF2 and LLR2). PCR amplification was performed with the GC-RICH PCR kit (Roche) using the following the conditions: 1X GC-RICH Buffer, 0.2 mM of dNTP, 0.2 µM LLF2 primer, 0.2 µM of LLR2 primer, 0.5 µM of Resolution Solution, 1uL of enzyme per 50 µL reaction. The thermal profile was:

1. 95°C for 3 min
2. 95°C for 15 sec
3. 55°C for 30 sec
4. 72°C for 1 min
5. Go to step 2 39 times
6. 72°C for 7 min

The resulting PCR product was then double-digested with BbsI and BamHI (NEB), gel extracted, and ligated using T4 Ligase into the pRDscFv2 vector digested with BamHI and BbsI. The ligation product was then PCR amplified using primers specific for the T7 promoter and the TolA linker (T7B2 and TolA). PCR amplification was performed with the GC-RICH PCR kit (Roche) using the following the conditions: 1X GC-RICH Buffer, 0.2 mM of dNTP, 0.2 µM LLF2 primer, 0.2 µM of LLR2 primer, 0.5 µM of Resolution Solution, 1µL of enzyme per 50 µL reaction. The thermal profile was:

1. 95°C for 3 min
2. 95°C for 15 sec
3. 55°C for 30 sec
4. 72°C for 1 min
5. Go to step 2 39 times
6. 72°C for 7 min

The final PCR product was digested with XhoI (NEB) to remove contaminating undigested pRDscFv2 vector, which contains suicide inserts with XhoI sites. The digested product was gel extracted and used as the template for the next round of ribosome display selection. This product can be Illumina sequenced, or it can be used for a subsequent round of selection after in vitro transcription.

Alternatively, cDNA could be amplified in the following way (to avoid subcloning): PCR1 was performed with preTolA and preT7B2 primers, 30 cycles (GC Rich PCR kit, Roche). PCR1 was purified, and used as template for PCR2 with T7B2 and TolA primers (Table 2.1), 10 cycles (GC Rich PCR kit, Roche). 1.2 kb PCR2 product was purified. This product can be Illumina sequenced, or it can be used for a subsequent round of selection after in vitro transcription.

Illumina sequencing

Libraries for Illumina Sequencing were prepared by two rounds of PCR amplification to add on the Illumina adapters and barcode sequences. All libraries except the Round 4 libraries were PCR amplified from the *in vitro* transcription template DNA using the TaKaRa EX HS kit. The conditions for the first round of PCR were: 1X TaKaRa EX HS Buffer, 0.2 mM dNTP, 0.4 µM IS7_L3F_PE primer, 0.4 µM IS8_H3R_PE_Multi primer, 0.5 µL TaKaRa Ex HS enzyme, and 1 µL of template per 50µL reaction. The thermal profile was:

1. 98°C for 10 sec
2. 50°C for 30 sec
3. 72°C for 1 min 30 sec
4. Go to step 1 9 times
5. 72°C for 7 min

The conditions for the second round of PCR were: 1X TaKaRa EX HS Buffer, 0.2 mM dNTP, 0.5 µM of IS4_L3F_PE primer, 0.5 µM of the barcoding primer, 0.5 µL TaKaRa

Ex HS enzyme, and 1 µL of the first round PCR product per 50 µL reaction. The thermal profile was:

1. 98°C for 10 sec
2. 60°C for 30 sec
3. 72°C for 1 min 30 sec
4. Go to step 1 9 times
5. 72°C for 7 min

The Round 4 libraries were PCR amplified from the cDNA using the Phusion HF kit (NEB). The conditions for the first round of PCR were: 1X Phusion High-Fidelity PCR Master Mix with HF Buffer, 0.5 µM of IS7_L3F_PE primer, 0.5 µM of IS8_H3R_PE_Multi primer, and 1 µl of cDNA recovered after library selection per 50 µl reaction. The thermal profile was:

1. 98°C for 30 sec
2. 98°C for 10 sec
3. 55°C for 30 sec
4. 72°C for 30 sec
5. Go to step 2 9 times
6. 72°C for 10 min

The conditions for the second round of PCR were: 1X Phusion High-Fidelity PCR Master Mix with HF Buffer, 0.5 µM of IS4_L3F_PE primer, 0.5 µM of the barcoding primer, and 1 µL of the first round PCR product per 50 µL reaction. The thermal profile was:

1. 98°C for 30 sec
2. 98°C for 10 sec
3. 60°C for 30 sec
4. 72°C for 30 sec
5. Go to step 2 9 times
6. 72°C for 10 min

All of the second round PCR products were gel extracted before sequencing.

Rescue of HMM scFv clones from a selected library

Single HMM scFv clones were rescued from the selected library by PCR with CDR-specific primers followed by assembly into a protein expression vector. Forward primers contained the 5' sequence of the target clone's L3 sequence preceded by a 20 bp adapter sequence for assembly into a protein expression vector. Reverse primers contained the reverse complement of the target clone's H3 sequence preceded by a 20 bp adapter sequence for assembly into a protein expression vector. Longer primers were designed for less abundant clones to increase specificity. PCR amplification was performed with the following conditions: 1X Phusion High-Fidelity PCR Master Mix with HF Buffer, 0.2 μM each of the forward and reverse primers, 1 μl of cDNA recovered after library selection per 50 μl reaction. For the more abundant clones, the thermal profile was:

1. 98°C for 30 sec
2. 98°C for 10 sec
3. 55°C for 30 sec
4. 72°C for 1 min
5. Go to step 2 29 times
6. 72°C for 10 min

For the less abundant clones, the thermal profile was:

1. 98°C for 30 sec
2. 98°C for 10 sec
3. 72°C for 1 min
4. Go to step 2 29 times
5. 72°C for 10 min

PCR products were subsequently gel purified individually and assembled into a protein expression vector using an isothermal assembly method. The protein expression vector contains the RDscFv framework followed by a FLAG tag and two in-frame stop codons instead of the TolA linker. The adapter sequences on the forward and reverse rescue primers are homologous to sequences flanking the L3 and H3 insert regions, respectively, of the protein expression vector. Vector lacking the L3-H2-H3 insert regions was prepared by PCR and gel purification. The isothermal assembly reaction was performed as previously described.[70] Each reaction contained 100 ng of empty vector DNA and 20 ng of the rescue PCR product, and was incubated at 50°C for 1 h. 1 µl of the assembly reaction product was transformed in DH5α cells and colonies were picked for sequence verification. Plasmids from sequence-verified clones were expressed using the RTS 100 Disulfide Kit for coupled *in vitro* transcription and translation (Fisher). 25 µl of lysate and 1 µl of lysate activator were first incubated on a rocker at room temperature. 25 µl of the resulting activated lysate was added to 7 µl of reaction mix, 7 µl of amino acid mix, 1 µl of methionine, and 500 ng of plasmid in 10 µl of distilled water. The reaction was then incubated at 30°C for 3 h and the resulting product was used directly in subsequent experiments.

Live cell FACS analysis

Telomerase-large T-immortalized human mammary epithelial cells (TL-HMECs) were transduced with retroviral constructs expressing human PVRL4 or control (empty vector). For labeling with in vitro-translated scFvs, cells were dissociated from the tissue culture plate with enzyme-free cell dissociation buffer (Invitrogen), resuspended in Stain buffer (BD Biosciences) and filtered through a 35 um nylon mesh cell strainer (BD Biosciences). Cells were incubated with in vitro-translated FLAG-tagged scFvs at a 1:100 dilution or anti-PVRL4 mouse monoclonal antibody (RnD Systems) for 30 min on ice, washed twice with Stain buffer and incubated with M2 anti-FLAG antibody (Sigma) at a 1:100 dilution for 30 min on ice. Labeled cells were washed twice and incubated with Alexa Fluor 488-conjugated goat-anti-mouse secondary antibody (Invitrogen) at 1:500 dilution for 30 min on ice. After a final series of washes, cells were resuspended in Stain buffer. Fluorescent signal was measured on LSR II FACS Analyzer (BD Biosciences) and analyzed with FlowJo software.

3. PhIP-Seq with a Synthetic Human Peptidome

Collaborator affiliations and contributions

H. Benjamin Larman[1,2,3], Zhenming Zhao[3,4], Uri Laserson[1,5,6], Mamie Z. Li[3], Alberto Ciccia[3], M. Angelica Martinez Gakidis[3], George M. Church[6], Santosh Kesari[7], Emily M. LeProust[8], Nicole L. Solimini[3] & Stephen J. Elledge[3]

[1]Harvard-MIT Division of Health Sciences and Technology, Cambridge, MA, USA

[2]Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA

[3]Department of Genetics, Harvard University Medical School, and Division of Genetics, Howard Hughes Medical Institute, Brigham and Women's Hospital, Boston, MA, USA

[4]Present address: Biogen Idec, Cambridge, Massachusetts, USA

[5]Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA, USA

[6]Department of Genetics, Harvard University Medical School, Boston, MA, USA

[7]Division of Neuro-Oncology, Department of Neurosciences, U.C. San Diego, Moores Cancer Center, La Jolla, CA, USA

[8]Agilent Technologies, Genomics, Santa Clara, CA, USA

S.J.E. conceived the project, which was supervised by N.L.S. and S.J.E. Z.Z. designed the DNA sequences for synthesis. Oligo libraries were constructed by E.M.L. Cloning was performed by M.Z.L., M.A.M.G, and N.L.S. The T7-Pep, T7-NPep, and T7-CPep phage libraries were constructed by N.L.S. and characterized by N.L.S. and H.B.L. The PhIP-Seq protocol was developed and implemented by H.B.L. Clinical evaluations and patient sample acquisitions were performed by S.K. Statistical analysis of PhIP-Seq data was conceived by U.L. under supervision of G.M.C. and implemented by H.B.L. PhIP-Seq candidates were confirmed by H.B.L. RPA2 IP was performed by A.C. This chapter was prepared by H.B.L. and edited by N.L.S. and S.J.E.

This work was supported in part by grants from the United States Department of Defense (W81XWH-10-1-0994 and W81XWH-04-1-0197) to S.J.E., and in part by the National Institutes of Health (K08CA124804), The American Recovery and Reinvestment Act (3P30CA023100-25S8), Sontag Foundation Distinguished Scientist Award, and a James S. McDonnell Foundation award to S.K. N.L.S. is a fellow of the Susan G. Komen for the Cure Foundation. S.J.E. is an investigator with the Howard Hughes Medical Institute. We would like to thank S. Gowrisankar, O. Iartchouk, and L. Merrill for assistance with Illumina sequencing, and D. Šćepanović for statistical support.

3.1 Abstract

In this study, we improve on current autoantigen discovery approaches by creating a synthetic representation of the complete human proteome, the T7 “peptidome” phage display library (T7-Pep), and use it to profile the autoantibody repertoires of individual patients. We provide methods for 1) designing and cloning large libraries of DNA microarray-derived oligonucleotides encoding peptides for display on bacteriophage, and 2) analyzing the peptide libraries using high throughput DNA sequencing. We applied phage immunoprecipitation sequencing (PhIP-Seq) to identify both known and novel autoantibodies contained in the spinal fluid of three patients with paraneoplastic neurological syndromes. We also show how our approach can be used more generally to identify peptide-protein interactions and point toward ways in which this technology will be further developed in the future. We envision that PhIP-Seq can become an important new tool in autoantibody analysis, as well as proteomic research in general.

3.2 Introduction

Vertebrate immune systems have evolved sophisticated genetic mechanisms to generate antibody repertoires, which are combinatorial libraries of affinity molecules capable of distinguishing between self and non-self. Recent data highlight the delicate balance in higher mammals between energy utilization, robust immune defense against pathogens, and autoimmunity[71]. In humans, loss of tolerance to self-antigens results in a number of diseases including type I diabetes, multiple sclerosis, and rheumatoid arthritis. Knowledge of the self antigens involved in autoimmune processes is not only important for understanding the disease etiology, but can also be used to develop accurate diagnostic tests. In addition, physicians may someday utilize antigen-specific therapies to target auto-reactive immune cells for destruction or quiescence.

Traditional approaches to identification of autoantibody targets largely rely on expression of fragmented cDNA libraries. Important technical limitations of this method include the small fraction of clones expressing in-frame coding sequences (with a lower bound of 6%)[72], and the highly skewed representation of differentially expressed cDNAs. Nevertheless, expression cloning has led to the discovery of many important autoantigens[73-75]. Strides have been made to improve peptide display systems[76, 77], but there remains an important unmet need for better display libraries and methods to analyze binding interactions.

Here, we have constructed the first synthetic representation of the complete human proteome, which we have engineered for display as peptides on the surface of T7 phage. This T7 "peptidome" library (T7-Pep) was extensively characterized and found to be both faithful to its *in silico* design and uniform in its representation. We combined our T7-Pep library with high-throughput DNA sequencing to identify autoantibody-peptide interactions, a method we call phage immunoprecipitation sequencing (PhIP-Seq). This approach provides several advantages over traditional methods, including comprehensive and unbiased proteome representation, peptide enrichment quantification, and a streamlined, multiplexed protocol requiring just one round of enrichment. We have applied PhIP-Seq to interrogate the autoantibody repertoire in the spinal fluid of patients with neurological autoimmunity and identified both known and novel autoantigens. We further demonstrate how PhIP-Seq can also be used more generally to identify peptide-protein interactions.

3.3 Results

3.3.1 Construction and characterization of the T7-Pep library

We sought to create a synthetic representation of the human proteome. We began by extracting all open reading frame (ORF) sequences available from build 35.1 of the human genome (24,239; 23% of which had "predicted" status). When there were multiple isoforms of the same protein, we randomly selected one representative ORF. We modified the codon usage by eliminating restriction sites used for cloning and by substituting very low abundance codons in *E. coli* with more abundant synonymous codons. We parsed this database into sequences of 108 nucleotides encoding 36 amino acid tiles with an overlap of seven residues between consecutive peptides (Figure 3.1a), the estimated size of a linear epitope. Finally, the stop codon of each ORF was removed so that all peptides could be cloned in-frame with a C-terminal FLAG tag.

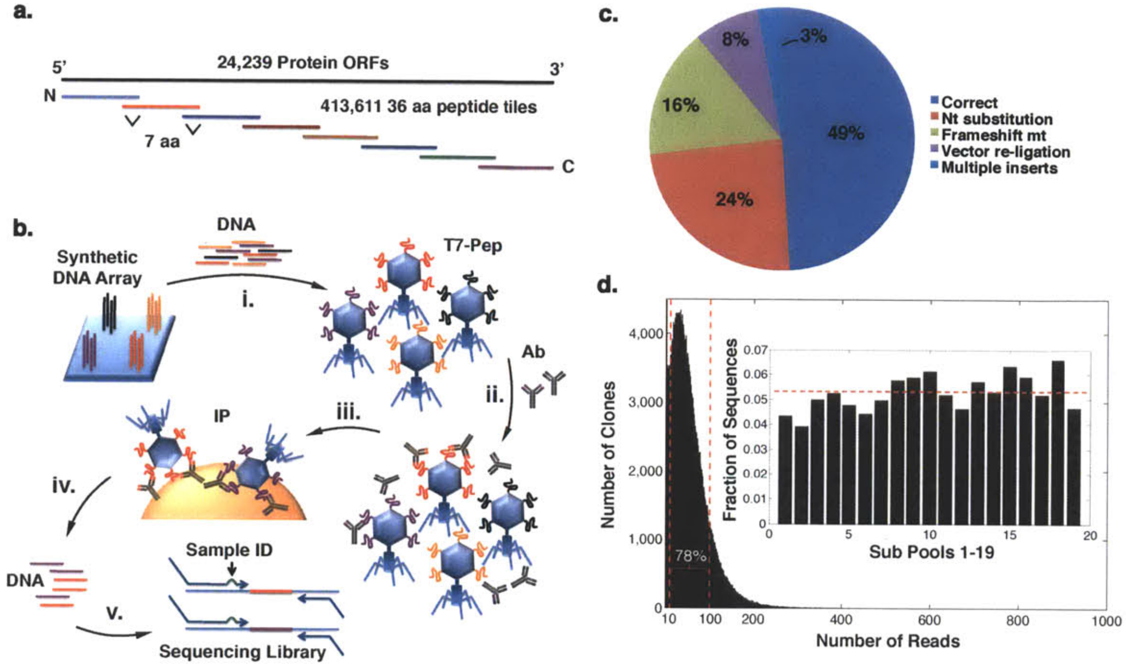


Figure 3.1: Construction and characterization of T7-Pep and the PhIP-Seq methodology

(**a**) The T7-Pep library is made from 413,611 DNA sequences encoding 36 amino acid peptide tiles that span 24,239 unique ORFs from build 35.1 of the human genome. Each tile overlaps its neighbors by seven amino acids on each side.

(**b**) The DNA sequences from (a) were printed as 140-mer oligos on releasable DNA microarrays. (i) After oligo release, the DNA was PCR-amplified and cloned into a FLAG-expressing derivative of the T7Select 10-3b mid copy phage display system. (ii) The T7-Pep library is mixed with patient samples containing autoantibodies. (iii) Antibodies and bound phage are captured on magnetic protein A/G coated beads. (iv) DNA from the immunoprecipitated phage is recovered and (v) library inserts are PCR-amplified with sequencing adapters. A single nucleotide change (arrow) is introduced for multiplex analysis.

(**c**) Pie chart showing results of plaque sequencing of 71 phage from T7-Pep Pool 1 and T7-CPep Pool 1.

(**d**) Histogram plot showing results from Illumina sequencing of T7-Pep. 78% of the total area lies between the vertical red lines at 10 and 100 reads, demonstrating the relative uniformity of the library. Representation of each subpool in T7-Pep (inset) compared to expected (horizontal red line).

The final library design includes 413,611 peptides spanning the entire coding region of the human genome. The peptide-coding sequences were synthesized as 140-mer oligonucleotides with primer sequences on releasable DNA microarrays in 19 pools of 22,000 oligos each, PCR-amplified and cloned into a derivative of the T7Select 10-3b phage display vector (Novagen; Figure 3.1b i and Supplementary Methods). We also generated two additional libraries comprising the N-terminal and C-terminal peptidomes (T7-NPep, T7-CPep), which encode only the first and last 24 codons from each ORF.

The extent of vector re-ligation, multiple insertions, mutations, and correct in-frame phage-displayed peptides was determined by plaque PCR analysis (Supplementary Table 3.1), clone sequencing (Figure 3.1c), and FLAG expression (Supplementary Table 3.2) of randomly sampled phage from all subpools. Sequencing revealed that 83% of the inserts lacked frameshifting mutations. These data indicate that a much greater fraction of in-frame, ORF-derived peptides is expressed by our synthetic libraries compared to those constructed from cDNA (Table 3.1).

After combining 5 x 10^8 phage from each subpool and amplifying the final library, Illumina sequencing was performed at a median depth of 45-fold coverage (Figure 3.1d) and detected 91.2% of the expected clones. Chao1 analysis was performed to estimate the actual library complexity (assuming infinite sampling), which predicted that >91.8% of the library was represented (Supplementary Figure 3.1)[78]. In addition, T7-Pep is highly uniform, with 78% of the library members within 10-fold abundance (having been sequenced between 10 and 100 times). These data suggest that our library encodes a much more complete and uniform representation of the human proteome than can otherwise be achieved with existing technologies (Table 3.1).

We next optimized a phage immunoprecipitation protocol for detecting antibody-peptide interactions within complex mixtures (Supplementary Figure 3.2). By combining this protocol with T7-Pep and deep sequencing DNA analysis, we have developed a new method to quantitatively profile autoantibody repertoires in patients (Figure 3.1b).

Feature	Classic cDNA Phage Display	Protein Array	T7-Pep + PhIP-Seq
Proteome representation	• Incomplete • Highly skewed distribution	• Small fraction • Uniform distribution	• Nearly complete • Uniform distribution
Fraction of clones expressing an ORF peptide in frame	As low as 6%	Up to 100%	~83%
Size of displayed peptides	Up to full-length proteins	Up to full-length proteins	36 amino acid overlapping tiles
Rounds of selection	Requires multiple selection rounds, which favor more abundant and faster growing clones[79]	No selection	Single selection, which eliminates clone growth bias and population bottleneck
Analysis	Individual clone sequencing: • Initial abundance unknown • Requires population bottleneck	Microarray scanning: • Quantitative • Statistical analysis of antibody binding	Deep sequencing of library: • Quantify population before and after a single round of selection • Statistical analysis of enrichments
Determination of antibody polyclonality	Difficult	Not possible	Often straightforward for antigens of known crystal structure
Epitope mapping	Difficult	Not possible	Often straightforward
Effort	Labor intensive	Minimal	Minimal
Sample throughput	Low	Medium	Adaptable to 96 well format
Multiplexing capability	No	No	Yes
Cost	Low	Moderate to high	Moderate

Table 3.1: Comparison between T7-Pep + PhIP-Seq and current proteomic methods for autoantigen discovery

3.3.2 Analysis of a PND patient with NOVA autoantibodies

Cancers often elicit cellular and humoral immune responses against tumor antigens which may limit disease progression[80]. In rare cases, tumor immunity can recognize central nervous system (CNS) antigens, triggering a devastating autoimmune process called paraneoplastic neurological disorder (PND). Clinical presentations of PND are heterogeneous and correlate with the CNS autoantigens involved. PND has served as a model for CNS autoimmunity, and the application of phage display to PND autoantigen discovery has met with much success[73, 81].

To assess the performance of PhIP-Seq for autoantigen discovery, we examined a sample of cerebrospinal fluid (CSF) from a 63-year-old female (Patient A) with non-small cell lung cancer (NSCLC) who presented with a PND syndrome and was found to have anti-NOVA autoantibodies[82]. The NOVA autoantigen (neuro-oncological ventral antigen, or "Ri") is commonly targeted in PND triggered by lung or gynecological cancers, and results in ataxia with or without opsoclonus/myoclonus. A concentration of 2 μg/ml of CSF antibody was spiked with 2 ng/ml of an antibody specific to SAPK4 (positive control) to monitor enrichment of the targeted peptide on protein A/G beads. Despite extensive washing, 298,667 unique clones (83% of the input library) were found in the immunoprecipitate. A significant correlation was observed between the abundance of input clones and immunoprecipitated clones (Figure 3.2a), likely due to weak nonspecific interactions with the beads.

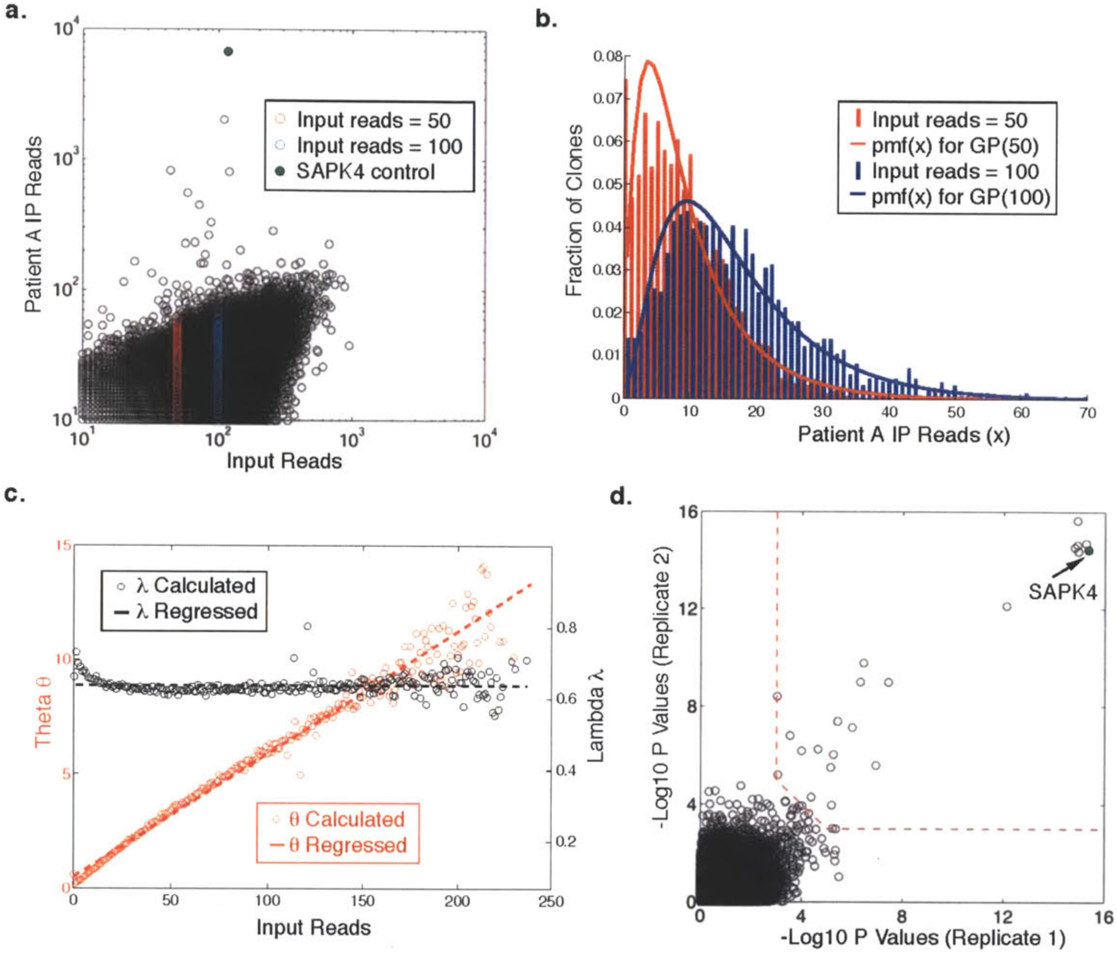


Figure 3.2 Statistical analysis of PhIP-Seq data

(**a**) Scatter plot comparing sequencing reads from T7-Pep input library and from Patient A immunoprecipitated (IP) phage (Pearson coefficient = 0.435; *P* ~= 0). Highlighted are all clones with an input abundance of 50 reads (red), and all clones with an input abundance of 100 reads (blue). The target of the SAPK4 control antibody is highlighted in green.

(**b**) Histogram plot of sequencing reads from the data highlighted in (a) with corresponding colors. The curves are fit with a generalized Poisson (GP) distribution. Pmf is the probability mass function of the corresponding GP distribution and x is the number of IP sequencing reads.

(**c**) Plots of lambda and theta for each input abundance, calculated using the method of Consul et al[83]. Lambda is regressed to its average value (black dashed line) and theta is linearly regressed (red dashed curve).

(**d**) Scatter plot comparing clone enrichment significances (as -Log10 p-value) from two independent PhIP-Seq experiments using CSF from Patient A. Red dashed line shows the cutoff for considering a clone to be significantly enriched, and the SAPK4 control antibody target is highlighted in green.

To approximate the expected distribution of IP'ed clones' abundances, we employed a two-parameter generalized Poisson (GP) model (as recently demonstrated for RNA-seq data[84]) and found that this distribution family fits the data well at various input abundances (Figure 3.2b). We calculated the GP parameter values for each input abundance level[83] and regressed these parameters to form our null model for the calculation of enrichment significance (p-values) of each clone (Figure 3.2c and online methods). Comparing the two PhIP-Seq replicates revealed that the most significantly enriched clones were the same in both replicas (Figure 3.2d), highlighting the assay's reproducibility. This contrasts dramatically with a comparison of clones enriched by two different patients (Supplementary Figure 3.3). Performing PhIP-Seq in the absence of patient antibodies identified phage capable of binding to the protein A/G beads. We thus defined Patient A positive clones as those clones with a reproducible Log10 p-value greater than a cutoff (Figure 3.2d, dashed red line), but not significantly enriched on beads alone ($P < 10^{-3}$). Patient A positives included the expected SAPK4-targeted positive control peptide ($P < 10^{-15}$), the expected NOVA1 autoantigen ($P < 10^{-15}$), and six additional candidate autoantigens (Table 3.2).

Patient Info	-Log10 P value	Protein	Pep-tides	Validation
A: 63 y.o. female with non-small cell lung cancer. Presents with classic cerebellar syndrome. ***CSF positive for anti-NOVA antibodies.***	*15.38*	*NEURO-ONCOLOGICAL VENTRAL ANTIGEN 1 (NOVA1)*	*1*	***WB+***
	14.76	**HYPOTHETICAL PROTEIN LOC26080**	**7**	**DB+**
	14.54	**TGFB-INDUCED FACTOR HOMEOBOX 2-LIKE, X-LINKED (TGIF2LX)**	**1**	**WB+**
	8.00	NEBULIN (NEB)	1	NT
	6.49	**DEBRANCHING ENZYME HOMOLOG 1 (DBR1)**	**1**	**WB-,DB+**
	6.20	**PROTOCADHERIN 1 (PCDH1)**	**1**	**WB-,DB+**
	4.29	INSULIN RECEPTOR (INSR)	1	NT
B: 59 y.o. female with non-small cell lung cancer. Presents with dysarthria, ataxia, head titubation and muscle lock. Paraneoplastic antibody panel is negative.	15.18	SOLUTE CARRIER FAMILY 25 MEMBER 43 (SLC25A43)	1	NT
	13.06	**GLUTAMATE DECARBOXYLASE 2 (GAD65)**	**2**	**RIA+,WB-, IP+**
	12.96	**TESTIS EXPRESSED SEQUENCE 2 (TEX2)**	**1**	**DB+**
	12.11	ATAXIN 7-LIKE 3 ISOFORM B (ATXN7L3)	1	NT
	11.93	ETS-RELATED TRANSCRIPTION FACTOR ELF-1 (ELF1)	1	NT
	11.91	**TGFB-INDUCED FACTOR HOMEOBOX 2-LIKE, X-LINKED (TGIF2LX)**	**1**	**WB+**
	11.34	INSULIN RECEPTOR SUBSTRATE 4 (IRS4)	1	NT
	6.98	HEPATOMA-DERIVED GROWTH FACTOR-RELATED PROTEIN 2 (HDGFRP2)	1	NT
	6.60	TUBULIN, BETA (TUBB)	1	WB-
	6.54	**CANCER/TESTIS ANTIGEN 2 (CTAG2)**	**1**	**WB+**
	6.30	**DENN/MADD DOMAIN CONTAINING 1A (DENDD1A)**	**1**	**WB-,DB+**
	6.09	DOUBLESEX AND MAB-3 RELATED TRANSCRIPTION FACTOR (DMRT2)	1	NT
	5.53	TUDOR AND KH DOMAIN CONTAINING ISOFORM A (TDRKH)	1	NT
C: 59 y.o. female with melanoma. Presents with ataxia, dysarthria, horizontal gaze palsy. Paraneoplastic antibody panel is negative. However, CSF stained brain and cerebellar IHC slides.	**15.72**	**TRIPARTITE MOTIF-CONTAINING 67 (TRIM67)**	**2**	**WB+**
	15.65	**TRIPARTITE MOTIF-CONTAINING 9 (TRIM9)**	**3**	**WB+**
	12.13	**FIBROBLAST GROWTH FACTOR 9 (GLIA-ACTIVATING FACTOR) (FGF9)**	**1**	**WB-,DB+**
	10.18	**DUAL-SPECIFICITY TYROSINE-(Y)-PHOSPHORYLATION REGULATED KINASE 3 (DYRK3)**	**1**	**WB-,DB+**
	6.93	CENTROSOMAL PROTEIN 152KDA (CEP152)	1	NT
	6.57	TITIN (TTN)	1	NT
	6.34	NUCLEOPORIN LIKE 2 (NUPL2)	1	NT
	5.43	**HISTONE DEACETYLASE 1 (HDAC1)**	**1**	**WB-,DB+**
	5.36	**MITOCHONDRIAL RIBOSOMAL PROTEIN L39 (MRPL39)**	**1**	**WB-,DB+**
	5.35	**CHROMOSOME 10 OPEN READING FRAME 82 (C10ORF82)**	**1**	**WB-,DB+**
	5.15	NLR FAMILY, PYRIN DOMAIN CONTAINING 5 (NLRP5)	1	NT
	4.83	TASPASE, THREONINE ASPARTASE, 1 (TASP1)	1	NT
	4.70	KIAA0090	1	NT
	4.55	SERINE (OR CYSTEINE) PROTEINASE INHIBITOR, CLADE A (ALPHA-1 ANTIPROTEINASE, ANTITRYPSIN), MEMBER 9 (SERPINA9)	1	NT
	4.21	**PROTEIN TYROSINE PHOSPHATASE, NON-RECEPTOR TYPE 9 (PTPN9)**	**1**	**WB-,DB+**

Table 3.2: Results of PhIP-Seq for 3 Patients

A previously validated autoantigen is shown in italics. Autoantigens confirmed by any secondary assay are shown in bold. Confirmation of patient antibodies with the full-length protein via western blot is indicated by red type. Average of replicate -Log10 p-values are shown in column 2. If multiple peptides from the same ORF are enriched, the average -Log10 p-value of the most significantly enriched peptide is shown. Secondary validation assay abbreviations: WB = western blot of full-length proteins; IP = immunoprecipitation of full-length proteins followed by western blotting for the fusion tag; RIA = radioimmunoassay; DB = dot blot; NT = not tested. Validation assay is followed by "+" or "-" depending on whether the results were positive or negative.

We tested three of these predictions by expressing full-length TGIF2LX, DBR1 and PCDH1 in 293T cells and immunoblotting with patient CSF. TGIF2LX (TGFB-Induced factor homeobox 2-like, X-linked) was confirmed as a novel autoantigen, as we detected strong immunoreactivity at the expected molecular weight (Figure 3.3a). Full-length DBR1 and PCDH1, while expressed well in 293T cells (not shown), were not detected by CSF antibodies. We observed two bands in the untransfected lysate migrating at approximately 50 and 62 KDa, possibly representing endogenously expressed proteins that correspond either to untested candidates or to false negatives of the PhIP-Seq assay.

Strikingly, the hypothetical protein LOC26080 had seven distinct peptides that were significantly enriched, and they all appeared to share a nine residue repetitive motif. We used MEME software[85] to characterize this motif, which represents the likely epitope recognized by Patient A's autoantibodies (Figure 3.3b).

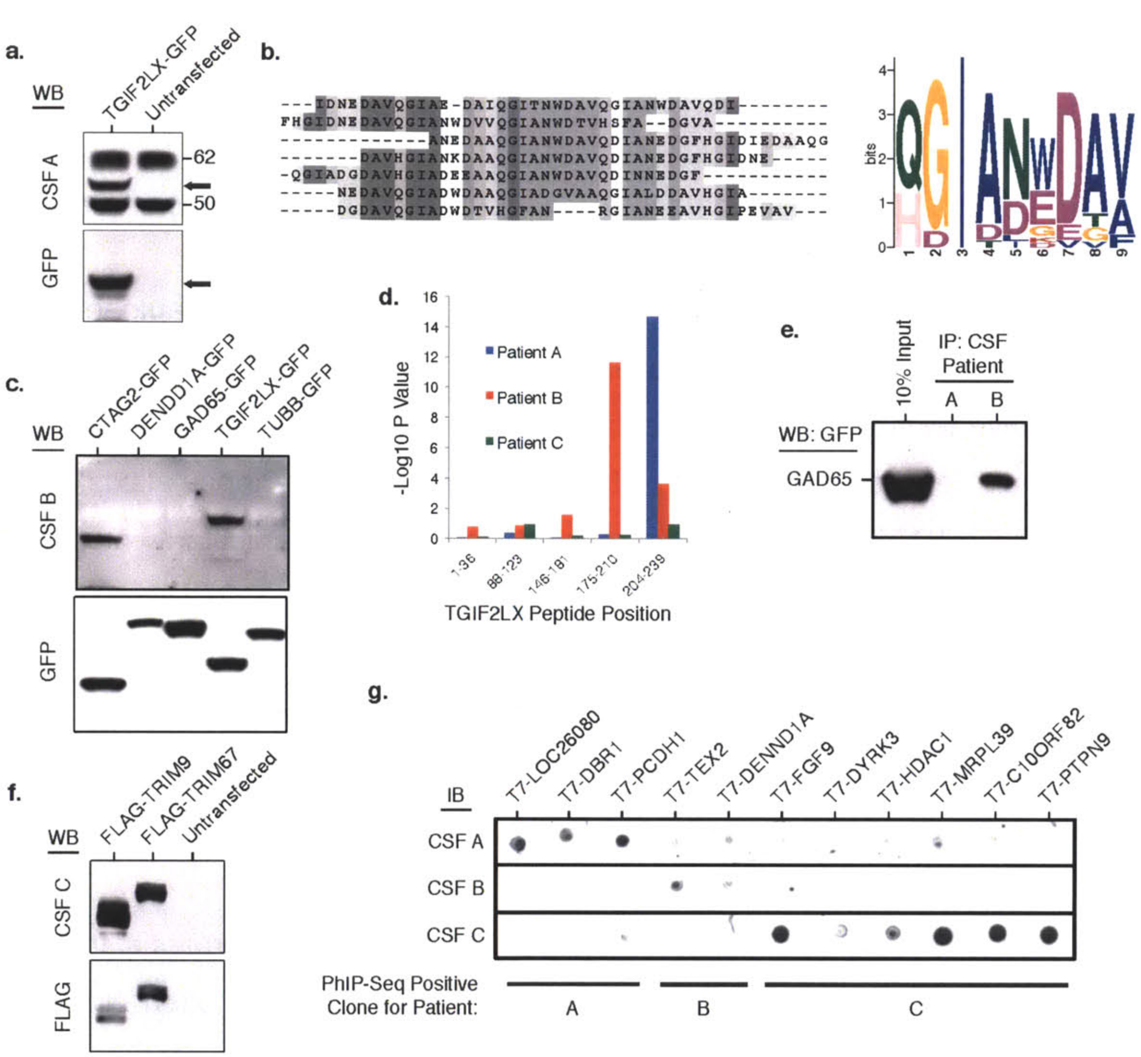


Figure 3.3: Validation of full-length PhIP-Seq candidates

(**a**) Western blot with CSF from Patient A, staining for full-length TGIF2LX-GFP expressed in 293T cells by transient transfection. Bands corresponding to TGIF2LX-GFP are denoted by an arrow.

(**b**) ClustalW alignment of the seven significantly enriched hypothetical protein LOC26080 peptides, and the nine-element MEME-generated recognition motif.

(**c**) Western blot with CSF from Patient B, staining for indicated full-length proteins expressed in 293T cells by transient transfection.

(**d**) Bar graph of -Log10 p-values of enrichment for the indicated TGIF2LX peptides by the three patients.

(**e**) Immunoprecipitation of the GAD65-GFP from 293T cell transfected lysate by CSF from Patient B (but not Patient A).

(**f**) Western blot with CSF from Patient C, staining for indicated full-length proteinss expressed in 293T cells by transient transfection.

(**g**) Phage lysates from candidate T7 clones were spotted directly onto nitrocellulose membranes, which were subsequently immunoblotted with patient CSF.

3.3.3 Analysis of two PND patients with uncharacterized autoantibodies

Having established that PhIP-Seq could reliably identify known and novel autoantigens, we examined CSF from two additional patients who had suggestive PND presentations but tested negative for a panel of commercially available PND autoantigens. Patient B was a 59-year-old female with NSCLC, who presented with dysarthria, ataxia, head titubation and muscular rigidity. PhIP-Seq analysis yielded three particularly interesting candidate autoantigens: TGIF2LX, CTAG2 (cancer/testis antigen 2), and GAD65 (glutamate decarboxylase 2) (Table 3.2). Both TGIF2LX and CTAG2 were confirmed by immunoblotting (Figure 3.3c). Surprisingly, Patient B, like Patient A, was auto-reactive against TGIF2LX. The enriched peptide was distinct from, but overlapped the peptide enriched by Patient A (Figure 3.3d).

CTAG2 is a member of a family of cancer/testis antigen (CTAG) proteins that are normally germ cell restricted, but frequently expressed in cancers and often elicit anti-tumor immune responses[86]. Several reports have described both humoral and cellular immune responses targeted against CTAG2[87, 88]. TGIF2LX is also testis restricted[89, 90] and may be a new CTAG family member. As a negative control, we found TGIF2LX reactivity to be absent in the CSF of three patients with non-PND CNS autoimmunity and oligoclonal Ig bands (Supplementary Figure 3.4). Having confirmed TGIF2LX autoreactivity in two NSCLC patients, we wondered whether it could be a new biomarker for this disease. However, the serum of 15 additional NSCLC patients without PND did not contain TGIF2LX antibodies detectable by immunoblotting (Supplementary Figure 3.5).

Neither CTAG2 nor TGIF2LX is expressed in the brain, and thus are unlikely to explain the neurological syndrome experienced by Patient B. GAD65, however, is the rate-limiting enzyme in the synthesis of the inhibitory neurotransmitter GABA. GAD65 is also a well-characterized autoantigen targeted in the autoimmune disorder Stiff Person Syndrome (SPS; OMIM ID 184850). Two non-overlapping GAD65 peptides derived from the domain known to be targeted by pathogenic autoantibodies in SPS patients[91, 92] were enriched by Patient B's CSF. A commercial radioimmunoassay (RIA 81596; Mayo Medical Laboratories), confirmed the presence of high titer anti-GAD65 autoantibodies (5.12 nmol/L; >250 fold above the reference range). Surprisingly, direct immunoblotting with Patient B's CSF did not demonstrate reactivity (Figure 3.3c), suggesting that denatured GAD65 epitopes are not recognized by Patient B's antibodies. Successful immunoprecipitation of GAD65 from the same cell lysate with CSF confirmed this hypothesis (Figure 3.3e).

Patient C, a 59-year-old female with PND secondary to melanoma, had an unusual presentation that included horizontal gaze palsy. PhIP-Seq analysis of Patient C's CSF yielded five significantly enriched peptides from two homologous members of the tripartite motif (TRIM) family, TRIM9 and TRIM67 (Table 3.2). Both candidate autoantigens were confirmed by immunoblotting lysates from TRIM9- or TRIM67-overexpressing cells (Figure 3.3f). TRIM67 is expressed in some normal tissues (including skin) and is often highly expressed in melanoma[90]. TRIM9 has recently emerged as a brain-specific E3 ubiquitin ligase and has been implicated in neurodegenerative disease processes[93]. Based on their high degree of homology, our data suggest the possibility that tumor immunity targeting TRIM67 might have spread to, or cross-reacted with TRIM9 in the CNS (Supplementary Figure 3.6). TRIM9 and TRIM67 autoreactivity was not detected in the CSF of three patients with non-PND CNS autoimmunity (Supplementary Figure 3.4).

In total, 16 of the candidate autoantigens in Table 3.2 were available to us as full-length Gateway Entry clones from the ORFeome collection[8]. Of these, 10 were not confirmed by immunoblotting or immunoprecipitation of the full-length protein. We wondered whether this reflected a high rate of false positive discovery inherent to PhIP-Seq, or rather a requirement that the peptides be presented with intact conformation, as was the case for GAD65. We synthesized 9 of these 10 candidate T7 clones, plus 2 additional high confidence T7 clones, for validation in a dot blot assay. Each of these clones exhibited immunoreactivity above background with the appropriate patients' spinal fluid as predicted by the PhIP-Seq dataset (Figure 3.3g; Supplementary Figure 3.7). This finding indicates that PhIP-Seq analysis can have a low rate of false positive discovery, and supports the hypothesis that the 36 amino acid peptides retain a significant amount of secondary structure during display on the T7 coat.

3.3.4 PhIP-Seq can identify peptide-protein interactions

The utility of T7-Pep is not limited to autoantigen identification. To explore more general interactions, we have used the library in an *in vitro* peptide-protein "two-hybrid" interaction experiment with GST-RPA2 (replication protein A2) as bait for T7-Pep. We were again able to utilize the generalized Poisson method for determining the significance of phage clones' enrichment. Whereas GST alone did not significantly enrich any library clones ($P < 10^{-4}$; Supplementary Figure 3.8), PhIP-Seq with GST-RPA2 robustly identified the N-terminal peptide from the known interactor SMARCAL1 ($P < 10^{-14}$, Figure 3.4), among others (Supplementary Table 3.3). The enriched SMARCAL1 peptide contains a previously identified motif known to bind RPA2[94, 95]. Peptides from four proteins known to contain this motif (UNG2, TIPIN, XPA and RAD52) were significantly disrupted by the positions of the breaks between peptides (Table 3.3). One peptide from UNG2 retained most of the motif and that peptide was correspondingly enriched ($P < 10^{-5}$), demonstrating the power of this approach to identify linear interaction motifs.

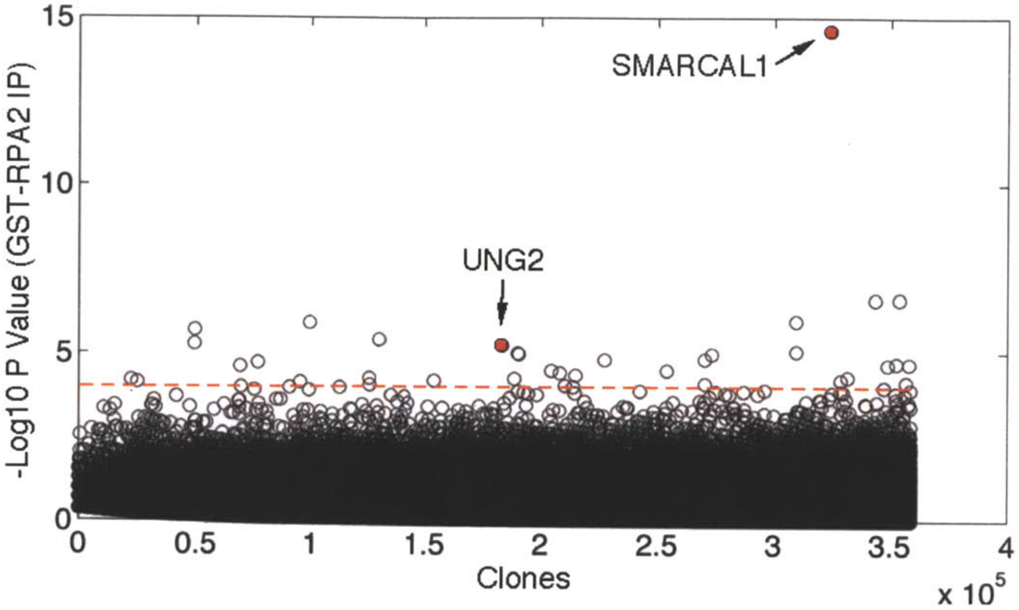


Figure 3.4: PhIP-Seq can identify protein-protein interactions

GST-RPA2 was used to precipitate phage from the T7-Pep library on magnetic glutathione beads. -Log10 p-values of enrichment were calculated using the generalized Poisson method. Clones are arranged in increasing input abundance from left to right. The experiment identified two of the known RPA2 binding partners SMARCAL1 ($P < 10^{-14}$) and UNG2 ($P < 10^{-5}$), highlighted in red.

Gene	T7-Pep Clone	Aligned Peptide	-Log10 P Value
SMARCAL1	**NP_054859.2_1**	**MSLPLTEEQRK-KIEENRQK--ALARRAEKLLAEQHQRT**	**14.6**
UNG2	NP_003353.1_2	...PSSPLSAEQLD-RI--	0.1
	NP_003353.1_3	**AEQLD-RI--QRNKAAAL----LRLAARNVPV...**	**5.2**
TIPIN	NP_060328.1_7	...LSRSLTEEQQR-RIE--RNKQLA	1.1
	NP_060328.1_8	E--RNKQLALERRQAKLLSNSQTL...	0.4
XPA	NP_000371.1_1	...QPAELPASVRA-SIERKRQRAL	0.3
	NP_000371.1_2	RKRQRALML--RQARLAARPYSA...	0.1
RAD52	NP_002870.2_9	...SLSSSAVESEATHQRKLRQKQLQQQF	1
	NP_002870.2_10	KQLQQQFR-ERMEKQQVRV...	0.1

Table 3.3: Dependence of peptide-RPA2 interaction on integrity of RPA2 binding motif

Aligned phage peptides containing the RPA2-binding motif (underlined) are shown next to their -Log10 p-value of enrichment. Significantly enriched peptides are shown in bold.

3.4 Discussion

We have developed a new proteomic technology called Phage Immunoprecipitation Sequencing (PhIP-Seq), which is based on a synthetic phage library (T7-Pep) made to uniformly express the complete human peptidome on the coat of T7 phage particles. Combining T7-Pep with high throughput DNA sequencing enables a variety of innovative proteomic investigations. In addition to applications in autoimmune disease, PhIP-Seq can be utilized to identify peptide-protein interactions and can be a viable alternative to two-hybrid analyses. From a methodological perspective, the robust single-round enrichment signals and the ability to adapt the assay to 96-well format suggests the feasibility of performing automated PhIP-Seq screens on large sets of samples.

Antibodies bind protein antigens by a variety of mechanisms and several studies have uncovered some general themes underlying these interactions. For instance, antibody combining surfaces on natively folded proteins tend to be dominated by "discontinuous" epitopes, which are patches of ~4-14 amino acid side chains formed by two or more noncontiguous peptides brought into proximity during protein folding[96, 97]. If the protein is divided into it's constituent peptides, antibody affinity is expected to decrease due to 1) the loss of contacts contributed by noncontiguous residues, and 2) the increased entropic costs of binding a free peptide as opposed to the natively constrained peptide. The degree to which individual peptides are still able to interact with a given antibody is difficult to predict, and is expected to vary widely. While our study demonstrates the utility of 36 amino acid tiles, further work will be required to define the true false negative discovery rate inherent to the use of T7-Pep. Autoantibodies that target normally inaccessible epitopes have also been reported, such as those that recognize proteolytic cleavage products[98, 99], misfolded proteins or protein aggregates[100, 101]. Antigen discovery with full-length, folded proteins may thus be less sensitive than tiled peptides in some such circumstances.

In our study, performing PhIP-Seq with CSF from a well characterized PND patient (Patient A), identified a known (NOVA1) and a novel, testis-restricted[90] autoantigen (TGIF2LX). Since we also found anti-TGIF2LX antibodies in the spinal fluid of a second PND patient with NSCLC, this protein may represent a new cancer-testis antigen family member, and should be further investigated as a biomarker for PND. PhIP-Seq analysis of CSF from two PND patients with uncharacterized antibodies (Patients B and C) uncovered likely neuronal targets of their autoimmune syndromes. In Patient B, high titer anti-GAD65 antibodies bound two distinct peptides from the region of the protein associated with Stiff Person Syndrome (SPS). Interestingly, GAD65 targeting in SPS occurs more often in patients without cancer, raising the possibility that at least part of this neurological syndrome may have been unrelated to the patient's cancer. This finding highlights the utility of unbiased antibody profiling to distinguish between deceptively similar disease states[102]. In Patient C, we identified TRIM9 as a likely neuronal autoantigen and suggest the possibility of epitope spreading from tumor-derived TRIM67 as a potential mechanism. It should be noted that demonstration of a protein's autoreactivity is not evidence for its role in disease pathogenesis, since the autoantibodies might be incidental in nature, arise due to epitope spreading, or might simply exhibit non-cognate cross-reactivity.

Several interesting features of the T7-Pep + PhIP-Seq platform emerged during this proof-of-concept study. We found that patient antibodies targeting GAD65 robustly recognized two 36 amino acid peptides, but not the corresponding denatured full-length proteins, indicating that an important degree of conformational information is retained in the peptide library. Second, for proteins with known crystal structures, using tiled peptides can facilitate determination of the antibody clonality, as well as the location of the targeted epitope. Finally, the simultaneous quantification of a large number of peptide enrichments permits the discovery of epitope motifs. Autoantibodies from Patient A targeted seven peptides from a repetitive hypothetical protein, and we were thus able to calculate a motif that most likely represented the antigenic epitope, a task less easily performed with alternative technologies.

T7-Pep could be improved in several ways. The generation of longer oligos will decrease the complexity of the library, thereby increasing the sampling depth and making it possible to generate domain libraries that capture more protein-folding units. In addition, PhIP-Seq with libraries of peptides from human pathogens could permit rapid analysis of antibodies to infectious agents, thus aiding vaccine research and the diagnosis of infectious diseases.

We have taken a synthetic biological approach to develop a proteomic resource useful in translational medicine. When combined with high throughput DNA sequencing, our methodology permits unbiased and quantitative analysis of autoantibody repertoires in human patients. PhIP-Seq thus complements existing proteomic technologies in the study of autoimmune processes for which the relevant autoantigens remain unknown.

3.5 Methods

Design of T7-Pep, T7-CPep and T7-NPep ORF sequences. We first downloaded all human protein and cDNA sequences available from the RefSeq database at build 35.1 of the human genome. Accession numbers between a protein and its cDNA were matched, and the paired sequences were used to construct the library. All the ATG start codons in the cDNAs were compared to the corresponding protein sequences until the correct ORF sequence was found. Seventy-two nucleotide (nt) fragments were then separated and overlapped with adjoining sequences by 21 nt (7 amino acids). Each DNA fragment was then scanned for the eight relatively rare codons in *E. coli* (CTA, ATA, CCC, CGA, CGG, AGA, AGG, GGA), and they were replaced by more abundant, synonymous codons (selected randomly if there was more than one replacement available). After that, each DNA fragment was rescanned for the four restriction sites (EcoRI, XhoI, BseRI, MmeI), and they were eliminated by replacement of one codon with a different, abundant, synonymous codon. Sequences were scanned iteratively to ensure the final ORF fragments were free of both rare codons and restriction sites. Finally, common primer sequences were added.

Cloning of T7-Pep. The proteome-wide library (19 pools of 22,000 synthetic oligos per pool) and N/C-terminal libraries (two pools each of 18,000 synthetic oligos per pool) were PCR-amplified as 23 independent pools with common primer sequences using the following conditions: 250 mM dNTPs, 2.5 mM MgCl2, 0.5 μM each primer, 1 μl Taq polymerase and ~350 ng oligo DNA per 50 μl reaction. The thermal profile was

1. 95 °C 30 s,
2. 94 °C 35 s,
3. 50 °C 35 s,
4. 72 °C 30 s,
5. Go to step 2 3x,
6. 72 °C 5 min,
7. 95 °C 30 s,
8. 94 °C 35 s,

9. 70 °C 35 s,
10. 72 °C 30 s,
11. Go to step 8 29x,
12. 72 °C 5 min

The PCR product was then digested and cloned into the EcoRI/SalI sites of the T7FNS2 vector with an average representation of at least 100 copies of each peptide maintained during each cloning step. The T7FNS2 vector is a derivative of the T7Select 10-3b vector (Novagen), which is a lytic, mid-copy phage display system, and displays 5-15 copies as C-terminal fusions with the T7 capsid protein. We modified the T7Select 10-3b vector to generate T7FNS2 by inserting a sequence encoding a FLAG epitope in the NotI and XhoI sites to generate an in-frame FLAG C-terminal fusion with the inserted peptide. Cloning of the synthetic peptide libraries into the T7FNS2 vector results in a C-terminal fusion of the ORF fragments with the T7 10B capsid protein, followed by a C-terminal FLAG epitope tag and stop codon (except for those in T7-CPep, which retain the native stop codons).

Patient samples. Collection and usage of human specimens from consenting patients were approved by the Brigham and Women's Hospital Institutional Review Board (protocol no. 2003-*P*-000655). Cerebrospinal fluid was aliquoted and kept at −80 °C until used, and freeze-thawing was avoided as much as possible after that. Neurological evaluations were performed by a board-certified neurologist. Serum samples from patients with confirmed NSCLC were from Bioserve.

Detailed PhIP-Seq protocol. The following were the multiplex barcode-introducing forward primers. The common P5 sequence for Illumina sequencing is in bold. The underlined segment was where the sequencing primer annealed. The 3-nt barcode is in italics.

HsORF-FL-mmBC1-F

AATGATACGGCGACCACCGAAGGTGTGATGCTCGGGGATCCAGGAATTCC*ACT*GCGC

HsORF-FL-mmBC2-F

AATGATACGGCGACCACCGAAGGTGTGATGCTCGGGGATCCAGGAATTCC*GCC*GCGC

HsORF-FL-mmBC3-F

AATGATACGGCGACCACCGAAGGTGTGATGCTCGGGGATCCAGGAATTCC*CCT*GCGC
HsORF-FL-mmBC4-F
AATGATACGGCGACCACCGAAGGTGTGATGCTCGGGGATCCAGGAATTCC*TCT*GCGC
HsORF-FL-mmBC5-F
AATGATACGGCGACCACCGAAGGTGTGATGCTCGGGGATCCAGGAATTCC*GAT*GCGC
HsORF-FL-mmBC6-F
AATGATACGGCGACCACCGAAGGTGTGATGCTCGGGGATCCAGGAATTCC*GGT*GCGC
HsORF-FL-mmBC7-F
AATGATACGGCGACCACCGAAGGTGTGATGCTCGGGGATCCAGGAATTCC*GTT*GCGC
HsORF-FL-mmBC8-F
AATGATACGGCGACCACCGAAGGTGTGATGCTCGGGGATCCAGGAATTCC*GCG*GCGC
P7-T7Down (this is the common reverse primer):
CAAGCAGAAGACGGCATACGAC ACTG AACCCCTCAAGACCCGTTTA
mmBC-FL_seq_prim (for sequencing the barcode and the library insert at P5 in forward direction):
AGGTGTGATGCTCGGGGATCCAGGAATTCC

Immunoprecipitation wash buffer consisted of 150 mM NaCl, 50 mM Tris-HCl, 0.1% NP-40 (pH 7.5).

Procedure: 1.5 ml tubes were blocked (including under cap) with 3% fraction V bovine serum albumin (BSA) in tris-buffered saline with 0.5% tween-20 (TBST) overnight at 4 °C rotating. Positive control SAPK4 C-19 antibody (Santa Cruz, sc-7585) was added (2 ng/ml final concentration; 1/1,000 of patient antibody) to phage stock (5 x 10^{10} pfu T7-Pep/ml final concentration) and mixed before being added to patient antibody (2 µg/ml final concentration). Each IP reaction was brought to a final volume of 1 ml using M9LB (Novagen).

Note: replicas were independent after this point (that is, there were two IP reactions as above for each sample).

Tubes were rotated at 4 °C for 24 h. 40 µl of 1:1 mix of Protein A and Protein G coated magnetic Dynabeads (Invitrogen, 100.02D and 100.04.D) slurry was added to each tube. Tubes were rotated for 4 more hours at 4 °C. Beads were washed 6 times in 500 µl IP wash buffer by pipetting up and down eight times per wash. Tubes were changed after every second wash. As much wash buffer as possible was removed and beads were resuspended in 30 µl H2O. IP was then heated at 90 °C for 10 min to denature phage and release DNA. 50 µl PCR reactions were prepared with TaKaRa HS Ex polymerase (TAKARA BIO), using the entire 30 µl of IP: 9.5 µl H2O, 5 µl 10x TaKaRa buffer, 4 µl dNTP (2.5 mM each), 0.5 µl P7-BC-T7Down (200 µM), 0.5 µl P5-mmBCn-F (100 µM), 0.5 µl TaKaRa HS Ex enzyme mix, 30 µl phage IP. The thermal profile was

1. 98 °C 10 s,
2. 56 °C 15 s,
3. 72 °C 25 s,
4. Go to step 1 39x,
5. 72 °C 7 min

The number of cycles can optionally be increased to 45.

PCR products were gel purified individually. Concentration was measured and then 500 ng of each barcoded sample was mixed together and Illumina sequencing was then performed on final material, using mmBC-FL_seq_prim as sequencing primer.

The first seven nt calls arose from the DNA barcode, and were used to parse the data by sample. Remaining sequence was aligned against the reference file. The reference sequences were truncated to the length of the reads and alignment was constrained to the appropriate strand.

RPA2-peptide interaction screen

Full-length, sequence-verified RPA2 was recombined from an available entry vector into pDEST-15 for inducible expression in *E. coli* as an N-terminal GST-fusion protein. A pDEST-15 clone expressing GST alone was used as a negative control. Protein expression was induced with 0.1 µM IPTG for 5 hours at 30°C. Protein lysate from 50 ml of bacterial culture was prepared in 1.5 ml of lysis buffer (50 mM tris pH 7.5, 500 mM NaCl, 10% glycerol, 1% triton, 10 mg/ml lysozyme) and sonicated before removing insoluble material by centrifugation. 40 µl of MagneGST Glutathione beads (Promega, V8611) were incubated in 1 ml of undiluted bacterial lysate for 2 hours. Beads were then washed 3 times with PBS. 1 ml of M9LB containing $5\text{x}10^{10}$ pfu of T7-Pep was then used to resuspend the beads (now coated with GST or GST-RPA2). The mix was rotated 24 hours at 4°C. At this point the beads were washed 6 times in 500 μl IP wash buffer, and the remaining protocol for PhIP-Seq given above was followed precisely.

Estimation of general Poisson model parameters and regressions

We assessed several distribution families for their ability to appropriately model the PhIP-Seq enrichment data, and found the two-parameter generalized Poisson distribution to be the best:

$$pmf(x) = \theta(\theta + x\lambda)^{x-1}e^{-\theta - x\lambda} / x!$$

For each value of input read number that had at least 50 corresponding clones, we used the following maximum likelihood estimators to calculate the values of lambda (λ) and theta (θ) for the corresponding distribution of n IP reads (x_i).[103]

$$\sum_{i=1}^{n} \frac{x_i(1-x_i)}{X + (x_i - X)\lambda} - nX = 0 \quad \text{where} \quad X = \sum_{i=1}^{n} \frac{x_i}{n} \quad \text{and} \quad \theta = X(1-\lambda)$$

Upon calculation of λ across all the input read numbers, we found it to be approximately constant. For each experiment, we thus regressed this parameter to be equal to the mean of all calculated λ's (Figure 3.2c). Calculation of θ's for all input values revealed the near linearity of this parameter, and so we linearly regressed this parameter prior to calculating the p-values.

Western blot validation of candidate autoantigens

We utilized the ORFeome collection of full-length proteins, which was generated by PCR and Gateway recombinational cloning,[104] as a source for testing autoantigen candidates by immunoblot. Entry vectors were recombined into the appropriate mammalian expression vector (CMV promoter driving ORF expression with either C-terminal GFP fusion or N-terminal FLAG epitope tag) and miniprepped for transient transfection.

293T cells were plated 24 hours before transfection at a density of 0.8 million cells per well of a 6-well plate and grown in DMEM containing10% FBS. TransIT-293T transfection reagent (Mirus, MIR 2700) was mixed with 2 μg expression plasmid per well, and added to the cells. After 24 hours, cells were harvested in 200 μl standard 1x RIPA-based laemmli/DTT sample buffer with Complete protease inhibitor cocktail (Roche) and sonicated for 30 seconds. Insoluble material was removed by centrifugation. 2-20 μl of lysate was run on 4-20 Bis-Tris polyacrylamide gels and transferred onto nitrocellulose using the iBlot system (Invitrogen). Membranes were blocked 1 hr in 5% milk and then stained with either patient CSF (1:250 to 1:1,000) or the appropriate primary anti-GFP (JL-8 monoclonal antibody; Clontech, 632381) or anti-FLAG (M2 monoclonal antibody; Sigma-Aldrich, F9291) antibody in 2.5% milk, TBST. Human antibody from CSF was detected with 1:3,000 peroxidase-conjugated goat affinity purified anti-Human IGG (whole molecule) secondary antibody (MP Biomedicals, 55252) in 2.5% milk, TBST.

For IP-western blotting, cell lysate was harvested in standard RIPA buffer with Complete protease inhibitor cocktail and sonicated for 30 seconds. Insoluble material was removed by centrifugation. 150 μl of lysate was mixed with 1 μg of patient antibodies and rotated overnight at 4°C. A 40 μl slurry of 1:1 mix of Protein A coated magnetic Dynabeads and Protein G coated magnetic Dynabeads was added to each tube. Tubes were rotated 4 hours at 4°C. Beads were washed 3 times in 500 μl RIPA buffer, and then harvested in 25 μl of laemmli/DTT sample buffer. The IP'ed protein and 10% of the input lysate were subject to SDS-PAGE analysis as above, and protein was detected by staining for the protein tag (e.g. GFP).

Dot blot validation of candidate autoantigens

Individual clones were made by synthesizing the peptide-encoding insert as a single, long DNA oligo (IDT, Ultramer™) that was PCR amplified and then cloned into T7FNS2 in the same way as described for the library. Clones were sequence verified and titered. 2 μl of each clone, after normalizing for titer, was spotted directly onto a nitrocellulose membrane and allowed to dry for 30 minutes. Membranes were blocked with 5% milk, TBST for 1 hour at room temperature, and then stained overnight at 4°C with 1 μg/ml of CSF antibody diluted in a solution containing a 1:1 mix of 5% milk, TBST and T7 10-3b-FLAG phage lysate. Human antibody from CSF was then detected with 1:3,000 peroxidase-conjugated goat affinity purified anti-Human IGG (whole molecule) secondary antibody (MP Biomedicals, 55252) in 2.5% milk, TBST. Quantification was performed by scanning developed films and analyzing the .tiff file with Image J software.

4. Defining autoantibody repertoires in health and disease

Collaborator affiliations and contributions

H. Benjamin Larman[1,2,3], Uri Laserson[1,4,5], George Xu[1,3,6], George M. Church[5], Nicole L. Solimini[3], Paul L. Klarenbeek[6], Robert M. Plenge[6], Peter A. Nigrovic[7], Philip L. De Jager[11], Kevin C. O'Connor[8], David A. Hafler[8], Geert A. Martens[9,10] & Stephen J. Elledge[3]

[1]Harvard-MIT Division of Health Sciences and Technology, Cambridge, MA, USA

[2]Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA

[3]Department of Genetics, Harvard University Medical School, and Division of Genetics, Howard Hughes Medical Institute, Brigham and Women's Hospital, Boston, MA, USA

[4]Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA, USA

[5]Department of Genetics, Harvard University Medical School, Boston, MA, USA

[6]Division of Genetics, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA.

[7]Children's Hospital, Boston, Massachusetts

[8]Department of Neurology, Yale School of Medicine, New Haven, CT

[9]Diabetes Research Center, Vrije Universiteit Brussel (VUB), Brussels, Belgium

[10]Department of Clinical Chemistry and Radioimmunology, Universitair Ziekenhuis Brussel, Brussels, Belgium

[11]Program in Translational NeuroPsychiatric Genomics, Institute for the Neurosciences, Department of Neurology and Psychiatry, Brigham and Women's Hospital, Boston, MA

S.J.E. conceived and supervised the project. N.L.S. constructed the T7-Pep library. H.B.L. performed the screen and together with U.L. analyzed the data. U.L. developed the informatics pipeline that was used to analyze the data. P.A.N. and R.M.P. provided the RA samples. K.C.O. and P.L.D.J. provided MS and control samples. G.A.M. provided T1D samples and matched controls, and assisted with analysis of the T1D dataset. P.I.W.D.B. provided statistical support. This chapter was prepared by H.B.L. and edited by S.J.E.

We would like to thank Stewart Rudnicki at the Institute of Chemistry and Cell Biology (ICCB) at Harvard Medical School for helping to automate PhIP-Seq. Paul I W de Bakker provided valuable statistical discussion. We also thank the Dana-Farber/Harvard Cancer Center (DF/HCC) Specialized Programs of Research Excellence (SPORE) in breast cancer for providing valuable breast cancer patient sera. This work was supported in part by grants from the United States Department of Defense (W81XWH-10-1-0994 and W81XWH-04-1-0197), and a HITI/Helmsley Trust Pilot Grants in Type 1 Diabetes to S.J.E. N.L.S. is a fellow of the Susan G. Komen for the Cure Foundation. S.J.E. is an investigator with the Howard Hughes Medical Institute.

4.1 Abstract

Autoimmune disease results from a loss of tolerance to self antigens. Understanding this process requires knowledge of the molecular targets, and so a number of techniques have been developed to determine immune receptor specificities. We have discussed in Chapter 3 the construction of a synthetic human peptidome displayed on T7 phage and à method to analyze its interactions with antibody repertoires using high throughput DNA sequencing (phage immunoprecipitation sequencing, "PhIP-Seq"). Here we present data from the first large-scale PhIP-Seq screen of 289 independent antibody repertoires. We screened sera from 72 healthy donors, resulting in an extensive set of enriched peptides, the majority of which composed each individual's unique "autoantibodyome", and a small number of which are recurrently enriched in the general population. Sera from 39 type 1 diabetes (T1D) patients were screened, revealing an accelerated polyautoreactivity phenotype compared to their matched controls, together with a set of novel candidate T1D autoantigens. Screening a collection of cerebrospinal fluids and sera from 56 multiple sclerosis patients uncovered novel, as well as previously reported specificities. Finally, a screen of synovial fluids and sera from 60 rhuematoid arthritis patients uncovered recurrent autoantibodies independent from seropositivity status. In sum, this work demonstrates the utility of performing PhIP-Seq screens on large numbers of individuals and is a step toward defining the full complement of autoimmunoreactivities in health and disease.

4.2 Introduction

Predisposing inherited alleles can conspire with stochastic and environmental events during development of the immune system and result in a failure of immune tolerance, often with catastrophic health consequences. Our understanding of autoimmune diseases has been limited by available technologies, which cannot capture the molecular complexity of intact immune systems. To address these limitations, we have recently developed an unbiased proteomic technology, phage immunoprecipitation sequencing (PhIP-Seq), with the capacity to quantitatively measure interactions between an individual's antibody repertoire and each of over 400,000 overlapping 36 mer peptides that together span the open reading frame of the human genome.[105] In this work, we have improved the PhIP-Seq method in two ways. First, all aspects of sample processing were made compatible with a 96-well plate format, and we employed a Biomek FX liquid handling robot to perform the immunoprecipitations. Second, we capitalized on recent improvements in Illumina sequencing, and developed a method to perform 96-plex analysis of individual PhIP-Seq libraries.[106] Over 150 million alignable reads per lane are routinely obtained on the Illumina HiSeq 2000 instrument, and so we were thus able to use just 2-3 lanes of a flow cell to analyze a complete 96-sample PhIP-Seq screen. This level of multiplexing reduces the cost to about $25 per sample, thereby enabling cohort-scale repertoire screening projects, even for smaller budget labs.

There are several autoimmune diseases of relatively high incidence for which the role of adaptive humoral autoimmunity is appreciated but not understood. Of these, we selected type 1 diabetes (T1D), multiple sclerosis (MS) and rheumatoid arthritis (RA) for autoantibody repertoire analysis by high throughput PhIP-Seq screening. Strong genetic linkage to class II HLA alleles in each of these diseases supports the view that there is an important role for antigen presentation and subsequent activation of CD4+ helper T cells with self-specificity.[107] The role of B cells in these diseases is less clear, but several observations indicate that analysis of secreted antibodies may provide insight into disease pathogenesis. For example, beta islet cell destruction in T1D is thought to be largely a consequence of CD8+ T cell activity, yet autoantibodies targeting islet-associated antigens are routinely used for diagnosis and risk stratification. In MS, oligoclonal IgG bands of unknown specificity are frequently found in cerebrospinal fluid (CSF), and secondary lymphoid tissue with germinal center activity often forms in the meninges of patients with advanced disease.[108] Patients with RA are classified as seropositive or seronegative depending on the presence of rheumatoid factor and/or anti-citrullinated protein antibodies. Beneficial clinical response to CD20+ B cell depletion therapy in RA has prompted the adoption of rituximab as a second line therapy for patients with high disease activity and features of a poor prognosis.[109, 110] In the treatment of MS and T1D, several studies have demonstrated efficacy of B cell depletion, but to a lesser extent, and with more elusive optimal dosing regimens.[111, 112]

We have performed high throughput PhIP-Seq screening on 39 sera obtained from newly diagnosed T1D patients, 40 synovial fluid samples and 20 sera from RA patients, 27 CSF samples and 35 sera from MS patients (including 6 sets of matching CSF/serum samples). Additionally, 72 sera from healthy donors, including a set of 41 age/sex-matched controls for the T1D cohort, were screened. To control for differences in fluid composition, we screened synovial fluid samples from 19 individuals with gout or non-rheumatoid osteoarthritis, as well as CSF from 9 patients with non-MS associated meningitis, subacute sclerosing panencephalitis, or paraneoplastic neurological disorder. Finally, we had previously screened a collection of 28 sera from patients with estrogen and progesterone receptor positive breast cancer (BC), and while analysis of the BC dataset is not presented here, it was included in all antigen-disease specificity tests. Table 4.1 provides a summary of these samples. A more detailed description can be found in Supplementary Table 4.1.

Class	Subclass	Fluid	Total
Type 1 Diabetes		serum	39
Multiple Sclerosis	(RRMS/SPMS/PPMS)	serum	35
		CSF	27
Rheumatoid Arthritis	Seropositive	serum	10
		synovial fluid	24
	Seronegative	serum	10
		synovial fluid	16
Healthy Controls		*serum*	*72*
Non MS CSF Controls	*SSPE, PND, Meningitis*	*CSF*	*9*
Non RA synovial fluid controls	*Gout, OA*	*synovial fluid*	*19*
Breast Cancer	*ER+/PR+*	*serum*	*28*
		Total	**289**

Table 4.1: Summary of the samples screened by high throughput PhIP-Seq

Control samples are italicized. Many samples were screened in duplicate, but only unique samples are shown in this summary. RR, relapse remitting MS; SP, secondary progressive MS; PP, primary progressive MS; SSPE, subacute sclerosing panencephalitis; PND, paraneoplastic neurological disorder; OA, osteoarthritis; ER+, estrogen receptor positive; PR+, progesterone receptor positive. Six sets of MS CSF/serum samples are patient matched.

4.3 Results

4.3.1 Polyautoreactivity and assay sensitivity

We used samples screened in duplicate to perform an analysis of peptide enrichment reproducibility (see Methods, Supplementary Figure 4.1), which led us to consider peptides with a -log10 P-value equal to 5 or greater as scoring positively above background. This threshold was used in all analyses unless otherwise stated. To exclude peptides that immunoprecipitated nonspecifically, we ignored those that displayed enrichment with -log10 P-values equal to 3 or greater in two or more out of 8 negative control (no antibody) IPs.

We first turned our attention to the data from the 72 healthy donors. In sum, 11,533 unique peptides were enriched. An overwhelming majority (10,122) of these autoreactivities were "personal" in the sense that they were observed to occur in only one individual (Figure 4.1A). At the other extreme, we observed a small number of peptides that were frequently enriched by healthy individuals. For example, we found that serum from 39% of individuals significantly enriched a single peptide from the activin receptor type IIB (ACVR2B), and serum from 43% of individuals had reactivity against a peptide from melanoma antigen family E, 1 (MAGEE1). The reactivities were not correlated with each other and did not depend on age, suggesting that these antibodies arise independently and at an early time. As it is not immediately clear whether these common autoantibodies were "on-target" or rather simply cross-reactive, we looked for evidence of epitope spreading within the database. Whereas we did find convincing examples of epitope spreading, likely due to CD4+ T cell help (e.g. CENPC1, Figure 4.2C), this was not true for ACVR2B and only weakly suggestive for MAGEE1. We therefore conclude that these recurrent anti-peptide antibodies are most likely cross-reactive and because they occur frequently in the serum of healthy individuals are unlikely to have a pathological consequence.

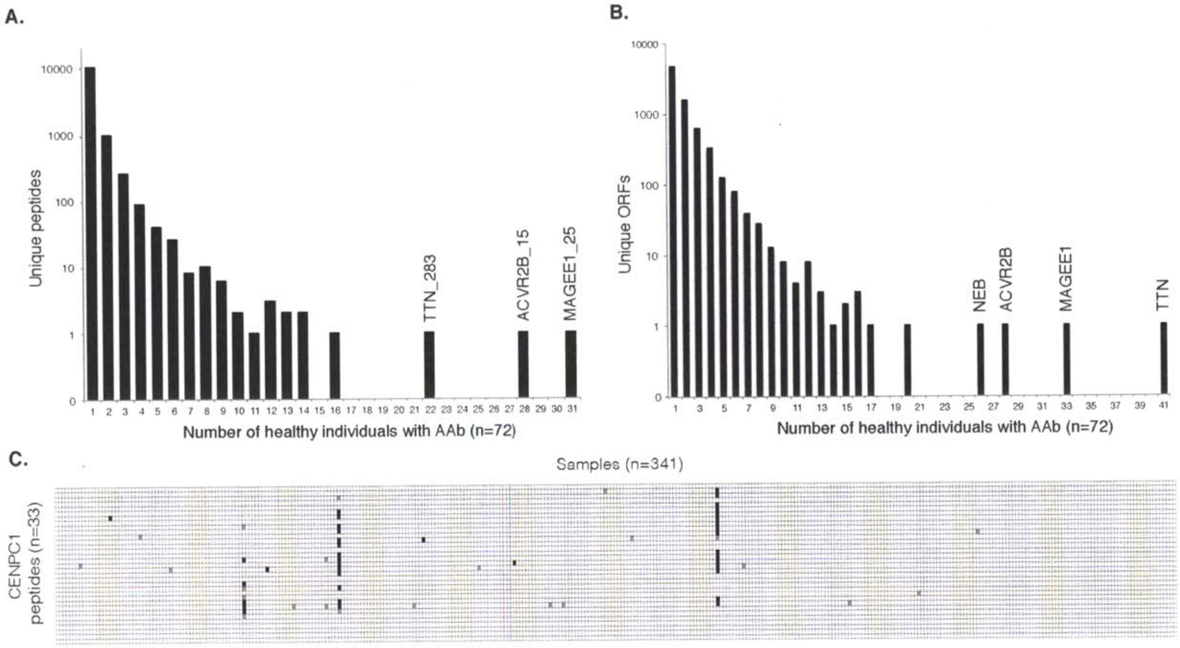


Figure 4.1: Enrichment distribution and evidence of T cell help

A. Frequency distribution of the 11,533 unique peptides enriched by greater than -log10 P value of 5 in healthy individuals. Peptides enriched above a threshold of -log10 P value of 3 or greater in 2 or more negative controls were considered nonspecific and removed from the analysis. TTN, titin; ACVR2B, activin receptor type-2B; MAGEE1, melanoma antigen family E, 1; NEB, nebulin.

B. Frequency distribution of the 7619 unique ORFs enriched by greater than -log10 P value of 5 in healthy individuals.

C. Heat map representation of the CENPC1 peptide enrichment data matrix. Enrichment -log10 P values greater than 5 are colored black, greater than 3 are colored gray, and less than 3 are white. Three individuals exhibit distinctive evidence of multi-epitope immune responses.

Patterns of disease-associated autoreactivity may only become apparent in the context of full-length proteins, since different individuals may produce antibodies that recognize different epitopes of a shared protein. We therefore collapsed the peptide enrichment matrix onto an ORF enrichment matrix by taking the most significant value from the set of peptides corresponding to each ORF. Again, if this -log10 P-value was greater than 5, the ORF was considered enriched by the individual. Analysis of ORF enrichments by healthy individuals resulted in a distribution similar to the peptide enrichments, with the majority of significantly enriched ORFs (62%) arising in just one person (Figure 4.1B).

This analysis is biased toward larger proteins being commonly enriched, and indeed significant reactivity against at least one peptide from titin (TTN, the largest ORF in our library) was observed in 41 of the 72 healthy individuals (Supplementary Discussion).

We screened a collection of serum samples obtained from 39 newly diagnosed type 1 diabetic (T1D) patients. As controls for comparison, we also screened sera from 41 healthy donors that were carefully matched for age and gender. These samples were all screened in the same automated run, and their positions on the 96-well plate were interspersed and randomized so as to avoid any technical artifact. Titers of clinically utilized autoantibody biomarkers (islet cell cytoplasmic antibody, "ICA"; insulin autoantibody, "IAA"; glutamic acid decarboxylase 2 antibodies, "GADA"; protein tyrosine phosphatase, receptor type, N antibodies, "PTPRNA" or "IA2A"; zinc transporter, member 8 antibodies, "ZnT8A") were measured for each of the T1D patients and controls. In order to determine the false negative discovery rate (sensitivity) inherent to our high throughput PhIP-Seq method, we compared radioimmunoassay (RIA) titers for each individual to the PhIP-Seq -log10 P-values of the corresponding ORFs. No PhIP-Seq enrichment was observed in any of the patients or controls for insulin or ZnT8A, whereas GAD2 and PTPRN enrichment was observed only in a small fraction of the T1D patients who had the highest RIA titers for those antigens (1 GAD2 PhIP-Seq positive out of 32 GAD2 RIA positives and 4 PTPRN PhIP-Seq positives out of 27 PTPRN RIA positives). In addition, one healthy individual was PhIP-Seq positive for PTPRN, despite having a negative titer by RIA (Figure 4.2A and Supplementary Figure 4.2).

We reasoned that if the amount of antibody-self peptide cross-reactivity in any way reflected the complexity of the antibody repertoire, then older individuals should IP more unique peptides compared to their younger counterpart. Comparing ages 12 and under ("young") with those 18 and older ("adult"), we observed a significant difference in the number of enrichments between young and adult healthy controls (Figure 4.2B). However, when we performed the same analysis of the T1D cohort, we found young

T1D patients to be significantly precocious in their development of autoreactive antibodies (P = 0.009; Student's t test, 1 tail). There was no significant difference in the amount of autoreactivity between healthy and T1D adults, or between young and adult T1D patients.

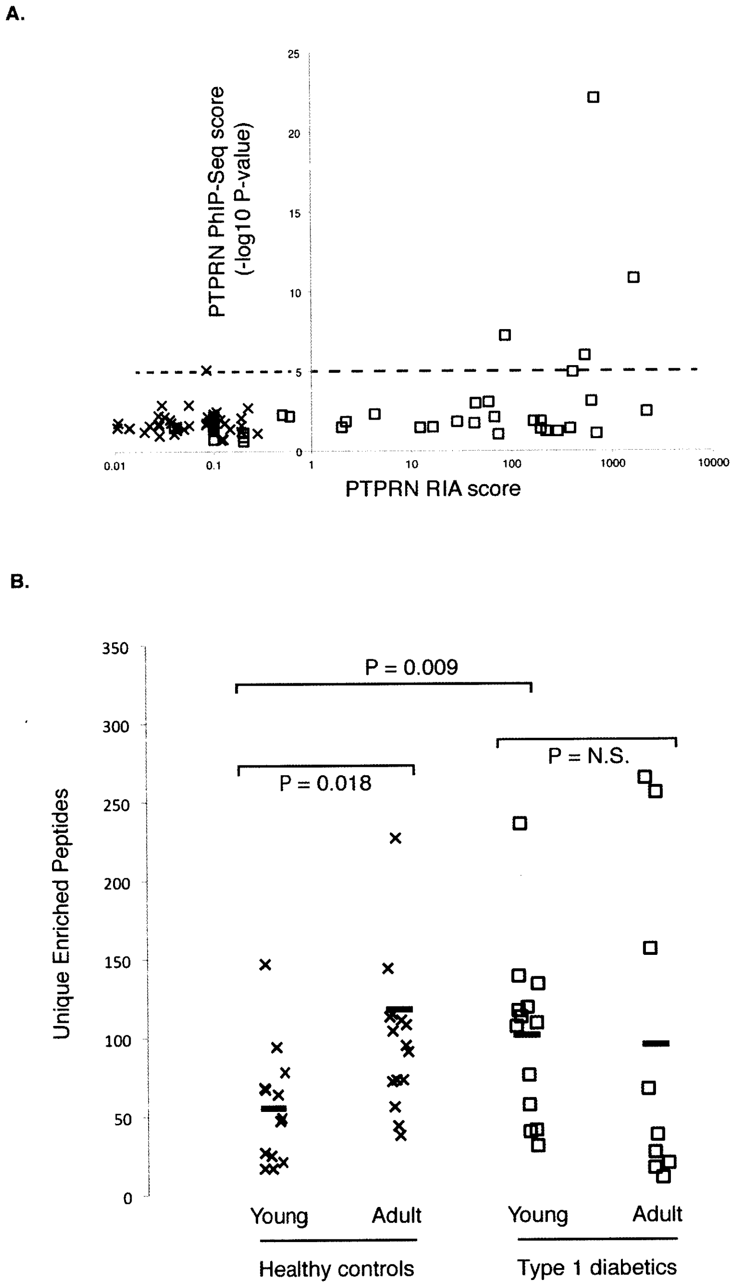


Figure 4.2: Analysis of T1D and healthy control sera

A. Sensitivity of PTPRN (IA2) autoantibody detection by PhIP-Seq compared to RIA in T1D patients (boxes) and healthy controls (x). PhIP-Seq values correspond to the most enriched peptide from the ORF. A value of >0.5 is considered positive for RIA.

B. Comparison of the total number of unique peptides enriched by individuals of different age groups, and T1D disease status. "Young" individuals are 12 years old or younger; "adult" individuals are 18 years old or older. Statistical comparisons of the means were performed using the Student's t test, with one tail.

4.3.2 Disease-specific autoantibodies

We next identified peptide and ORF autoreactivities specifically associated with each of the autoimmune diseases under investigation. For this analysis, each disease group was compared to all other samples, in the form of a Fisher's exact test to determine the significance of association. This analysis was performed for each peptide in the library, and so a distribution of Fisher P values was therefore obtained. To account for multiple hypothesis testing, we created a null distribution of "expected" Fisher P values by randomly permuting the sample labels 10,000 times (details can be found in the Methods section). We compared the distribution of expected significance values to that which was actually observed, and then set a threshold for 10% false positive discovery. All autoreactivities that exhibited disease association with this level of confidence are reported in Table 4.2.

			Summary						
			39	60	56	72	28	19	7
	Gene name associated with peptide or ORF	Cluster	T1D	RA	MS	HC	BC	OA/Gout	CSF Ctrl
T1D	protein tyrosine phosphatase, receptor type, N		3	0	0	0	0	0	0
	chromodomain helicase DNA binding protein 7		3	0	0	0	0	0	0
	arginine decarboxylase		3	0	0	0	0	0	0
	oxoeicosanoid (OXE) receptor 1		3	0	0	0	0	0	0
	ring finger protein 180		3	0	0	0	0	0	0
RA	BCL-6 interacting corepressor isoform 2	RA1	0	8	0	1	0	1	0
	keratin 33B	RA2	0	5	0	0	0	0	0
	septin 8	RA2	0	5	0	0	0	0	0
	ADAM metallopeptidase domain 33	RA2	0	6	0	0	0	0	0
	LPS-responsive vesicle trafficking, beach and anchor containing	RA2	0	5	0	0	0	1	0
	cAMP responsive element binding protein 3-like 1	RA2	0	5	0	0	0	0	0
	ring finger protein, LIM domain interacting; similar to ring finger protein (C3H2C3 type) 6	RA1	0	5	1	0	0	0	0
	ataxin 2	RA2	0	4	0	0	0	0	0
	PTK2 protein tyrosine kinase 2	RA2	0	4	0	0	0	0	0
	ATPase family, AAA domain containing 5	RA1	0	4	0	0	0	0	0
	KIAA0565		0	4	0	0	0	0	0
	hornerin		0	6	0	0	1	0	0
	NACC family member 2, BEN and BTB (POZ) domain containing		0	7	2	1	0	0	0
MS	keratin 75	MS1	0	0	8	0	0	0	0
	triple functional domain (PTPRF interacting)	MS1	0	0	8	0	0	0	0
	FLJ42289	MS1	0	0	7	0	0	0	0
	methyltransferase like 23	MS1	0	0	7	0	0	0	0
	DENN/MADD domain containing 4C	MS1	0	1	8	0	0	0	0
	SRY (sex determining region Y)-box 17		0	1	11	4	1	0	0
	KIAA1045	MS1	0	0	6	0	0	0	0
	integrator complex subunit 1	MS1	0	0	6	0	0	0	0
	regulating synaptic membrane exocytosis 2	MS1	0	1	7	0	0	0	0
	splicing factor, arginine/serine-rich 16	MS1	0	0	5	0	0	0	0
	bromodomain adjacent to zinc finger domain, 2A	MS1	0	0	5	0	0	0	0
	FERM domain containing 4B	MS1	0	0	5	0	0	0	0
	protein phosphatase 1, regulatory (inhibitor) subunit 10		1	3	15	9	4	1	0
	ubiquitin specific peptidase 11		0	0	6	0	2	0	0

Table 4.2: Peptide/ORF enrichments associated with disease

All disease-associated autoantigens with a false positive discovery rate of 10% are listed. ORF-only association is shown in italics. If the peptide is among a nonrandomly assorted cluster, the name of that cluster is provided in the third column. The summary of enrichments provides the total number of individuals from each group that displayed immunoreactivity against the peptide/ORF. Shown at top are the total number of individuals from the group.

While there were only 5 T1D associated peptides at an FDR of <10%, we were encouraged to find the validated T1D autoantigen PTPRN (IA2) in this list. These immunoreactivities assorted randomly among the patients, and so are unlikely to depend upon a common epitope. We used additional "circumstantial" data (e.g. pancreatic-specific gene expression and protein abundance, evidence of epitope spreading, etc.) to generate a set of candidates for follow up validation from a less stringent list (Supplementary Tables 4.2 and 4.3). Based on this analysis, and clone availability in our sequence-verified ORFeome collection,[113] we selected GNAS, RNF180, ZNF345, REG1B and BANF2 for follow up RIA studies using the full-length proteins. These experiments are ongoing.

We next examined peptide/ORF immunoreactivities specifically associated with RA (Figure 4.3). Of the 12 peptides found to have an FDR of <10%, 10 of them assorted nonrandomly as two clusters of 3 and 7 peptides. 16 out of the 60 RA patients exhibited immunoreactivity against at least one peptide from one of these clusters, compared to 4 out of 221 non RA individuals. Whereas the strength of RA2 enrichments were equal between the synovial fluid and the serum samples, RA1 enrichments were significantly stronger in the synovial fluid ($P = 1.1x10^{-5}$; Student's t test, 2 tails). Only one of the 19 synovial fluid control samples had immunopositivity for a peptide from RA1 (compared to 9 out of 40 RA synovial fluid samples), suggesting that RA1 reactivity is unlikely to be an artifact of the fluid composition. Interestingly, none of the RA-associated enrichments appeared to correlate with seropositivity (i.e. reactivity against rheumatoid factor and/or citrullinated peptide; Figure 4.3.B). Despite attempts to uncover a shared motif among RA1- and RA2-clustered peptides using blastp and MEME algorithms, none could be identified.

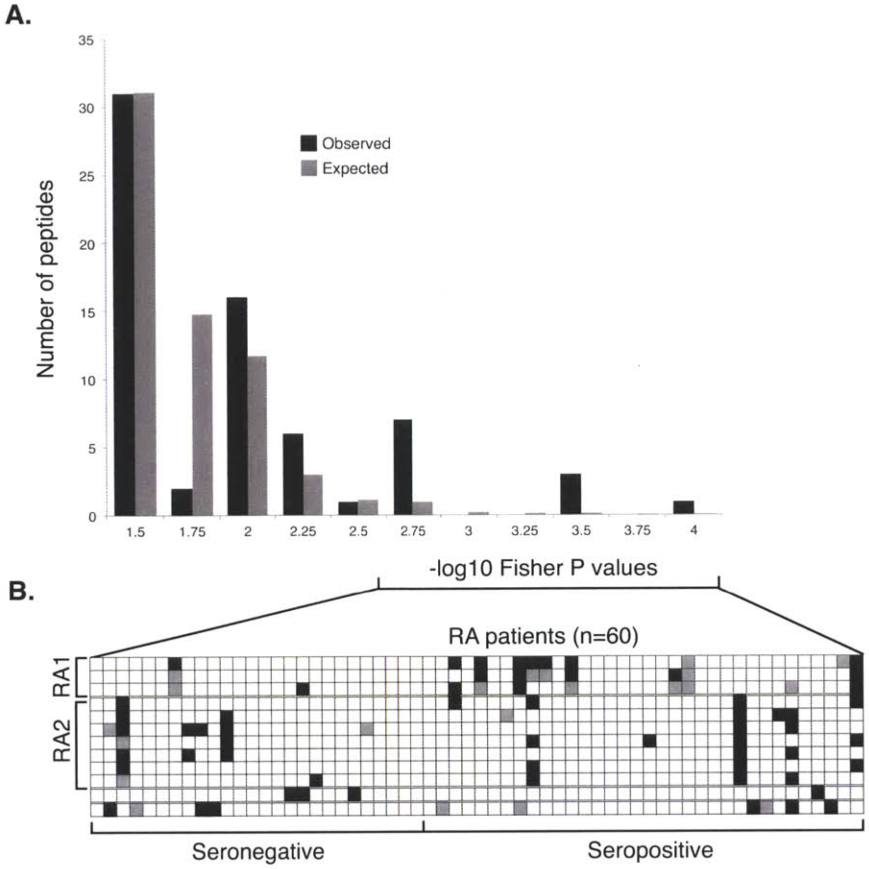


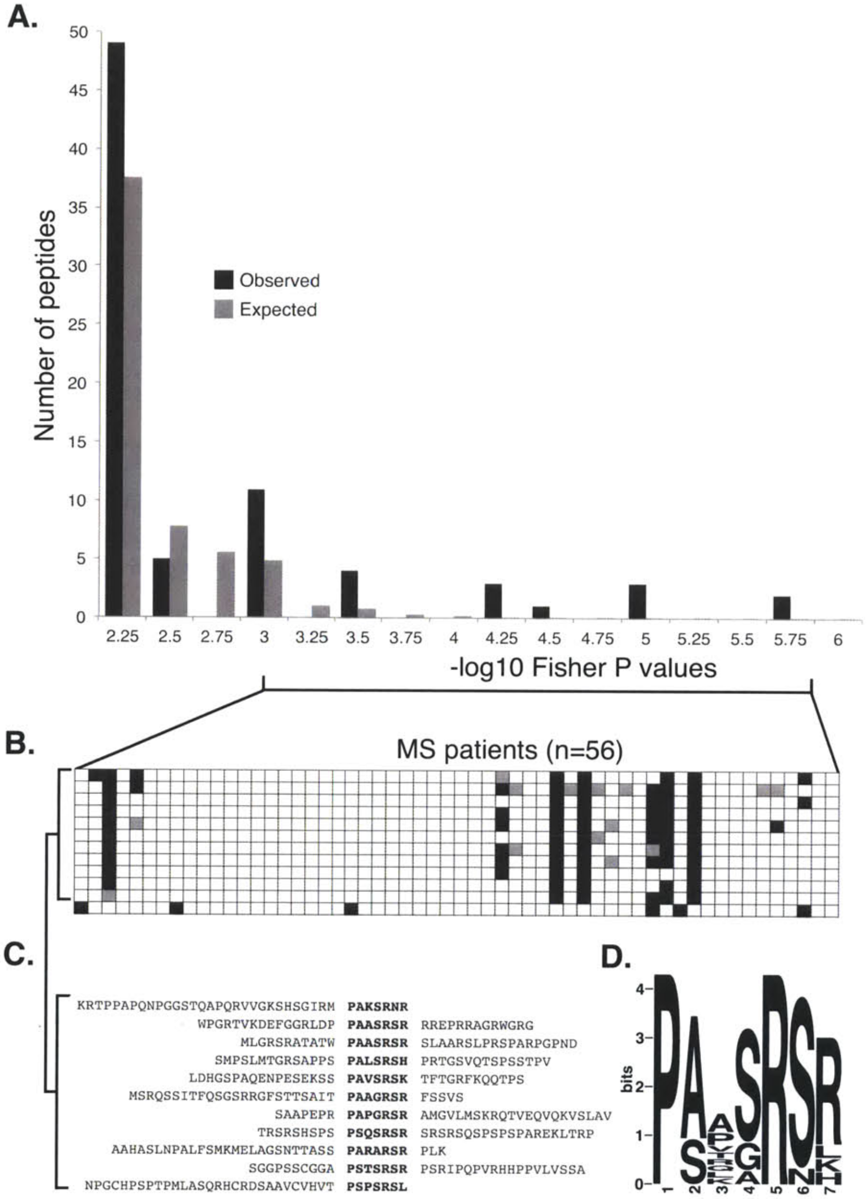
Figure 4.3: RA associated peptides and their clusters

A. Permutation analysis of peptide enrichments associated with RA. "Observed" bars indicate the number of peptides associated with RA at a given P-value by Fisher's exact test. "Expected" bars show the number of peptides expected to have a -log10 Fisher P-value at least that extreme due to chance alone.

B. Peptide enrichment heat map (as in Figure 4.1) illustrating nonrandom segregation of peptide enrichments (rows) and RA patients (columns). Peptides are organized by the RA1 and RA2 clusters. Patients are organized by their seropositivity. -Log10 P-values less than 3 are white, between 3 and 5 are gray, and greater than 5 are black.

MS patients are frequently found to have oligoclonal immunoglobulin in their CSF, which is resolvable by isoelectric focusing. As the presence of these oligoclonal IgGs is the most consistent laboratory abnormality in MS (detectable in about 95% of patients compared with 10%-15% of controls), it has long been assumed that the specificities of these intrathecally-produced antibodies harbor clues to the pathogenesis of the disease.

We therefore screened 27 CSF samples and 35 serum samples from patients with clinically definite multiple sclerosis. As additional negative controls, we screened 10 CSF samples from individuals with subacute sclerosing panencephalitis (SSPE), paraneoplastic neurological disorder (PND) and meningitis. At a cutoff of -log10 Fisher P-value of 3, the false discovery rate is ~10%. We thus examined the set of 14 peptides with this degree of disease association. After removing two peptides that were enriched in more than two non-MS samples, we were left with 12 peptides, the enrichments of which are displayed for each of the 56 MS patients in Figure 4.3B. Eleven of these peptides assorted non-randomly among a subset of MS patients, and motif discovery revealed a 7 amino acid sequence contained in all of them.[85] By dot blot with purified biotinylated peptide, we confirmed the presence of these antibodies in an MS patient that scored highly in the screen (Supplementary Figure 4.3). Notably, a motif nearly identical to MS1 was discovered by Cepok et al. in a similar screen of MS CSF samples,[114] and they reported an alignment with the BRRF2 protein of the Epstein-Barr virus, a pathogen repeatedly implicated in MS pathogenesis. We therefore performed an alignment of the MS1 motif against the UniProt database of all proteins from viruses with human tropism, collapsed onto 90% identity clusters (7,546 UniRef sequences; 656 unique taxa). This search confirmed the best alignment to be with the EBV BRRF2 protein (E value = 1.2; sequence: PAASRSK).



```
KRTPPAPQNPGGSTQAPQRVVGKSHSGIRM PAKSRNR
             WPGRTVKDEFGGRLDP PAASRSR RREPRRAGRWGRG
                  MLGRSRATATW PAASRSR SLAARSLPRSPARPGPND
                SMPSLMTGRSAPPS PALSRSH PRTGSVQTSPSSTPV
              LDHGSPAQENPESEKSS PAVSRSK TFTGRFKQQTPS
        MSRQSSITFQSGSRRGFSTTSAIT PAAGRSR FSSVS
                      SAAPEPR PAPGRSR AMGVLMSKRQTVEQVQKVSLAV
                     TRSRSHSPS PSQSRSR SRSRSQSPSPSPAREKLTRP
      AAHASLNPALFSMKMELAGSNTTASS PARARSR PLK
                   SGGPSSCGGA PSTSRSR PSRIPQPVRHHPPVLVSSA
 NPGCHPSPTPMLASQRHCRDSAAVCVHVT PSPSRSL
```

Figure 4.3: MS associated peptides share a common motif

A. Permutation analysis of peptide enrichments associated with MS. "Observed" bars indicate the number of peptides associated with MS at a given P-value by Fisher's exact test. "Expected" bars show the number of peptides expected to have a -log10 Fisher P-value at least that extreme due to chance alone.

B. Peptide enrichment heat map (as in Figure 4.1) illustrating nonrandom segregation of peptide enrichments (rows) and MS patients (columns). Peptides had a -log10 Fisher P-value >3 and were enriched by less than 3 non-MS patients. -Log10 P-values of enrichment less than 3 are white, between 3 and 5 are gray, and greater than 5 are black.

C. Alignment of the co-segregated peptides reveals a shared epitope.

D. MS associated epitope (MS1) motif logo, calculated from the peptides in **C** (MEME software).

4.3.3 Analysis of matched MS samples

As part of our collection, we obtained six sets of MS CSF/serum pairs. Each of these samples was screened in duplicate, and we considered only those peptides that were reproducibly enriched with a -log10 P-value greater than 3 in both replicates. For each of these MS patient pairs, we plotted the average -log10 P-value for each peptide's serum enrichment against the average CSF enrichment (Figure 4.4). In all cases we observed a strong correlation in the enrichment profiles between these two compartments. An overwhelming majority of the enrichments were found in both compartments, with a strong trend toward higher significance in the serum, suggesting that most of the autoreactivity we detected in the CSF could be attributed to antibody leakage from the serum compartment. In several cases, however, we did find peptides that were specifically enriched in the CSF compartment. For example, CSF from patient 9292 enriched two highly similar peptides from interferon alpha 5 and 14 much more significantly than serum from the same patient. Since many MS patients are administered therapeutic interferon beta, we wondered if this enrichment might reflect cross-reactivity of inhibitor antibodies. This is unlikely to be the case, however, as the homologous peptide from interferon beta was not enriched in either compartment.

We then examined all the CSF-specifically enriched peptides (enriched by both CSF replicas with -log10 P-value > 3, and neither serum replica with -log10 P-value > 3) that were identified in three of the six patients (Table 4.3). Motif discovery was performed for each set of CSF-specific peptides, and one motif was identified for patient 10894 (Figure 4.4B and Table 4.3). This motif was searched into the database of human viruses, and a significant alignment was found with the major capsid protein VP1 of the

JC polyomavirus (JCV; E value = 0.03; sequence: RRVKNP). Similar to EBV, JCV infection is highly prevalent, infecting 70 to 90 percent of humans. Also of note, JCV can cross the blood-brain barrier into the central nervous system, where it infects oligodendrocytes and astrocytes, possibly through the 5-HT2A serotonin receptor.[115]

Patient 8911 had serum samples drawn on two separate occasions within one year, which allowed us to examine the correlation of PhIP-Seq autoreactivities over time. The scatterplot (Figure 4.4D) of these autoreactivities reveals near identity.

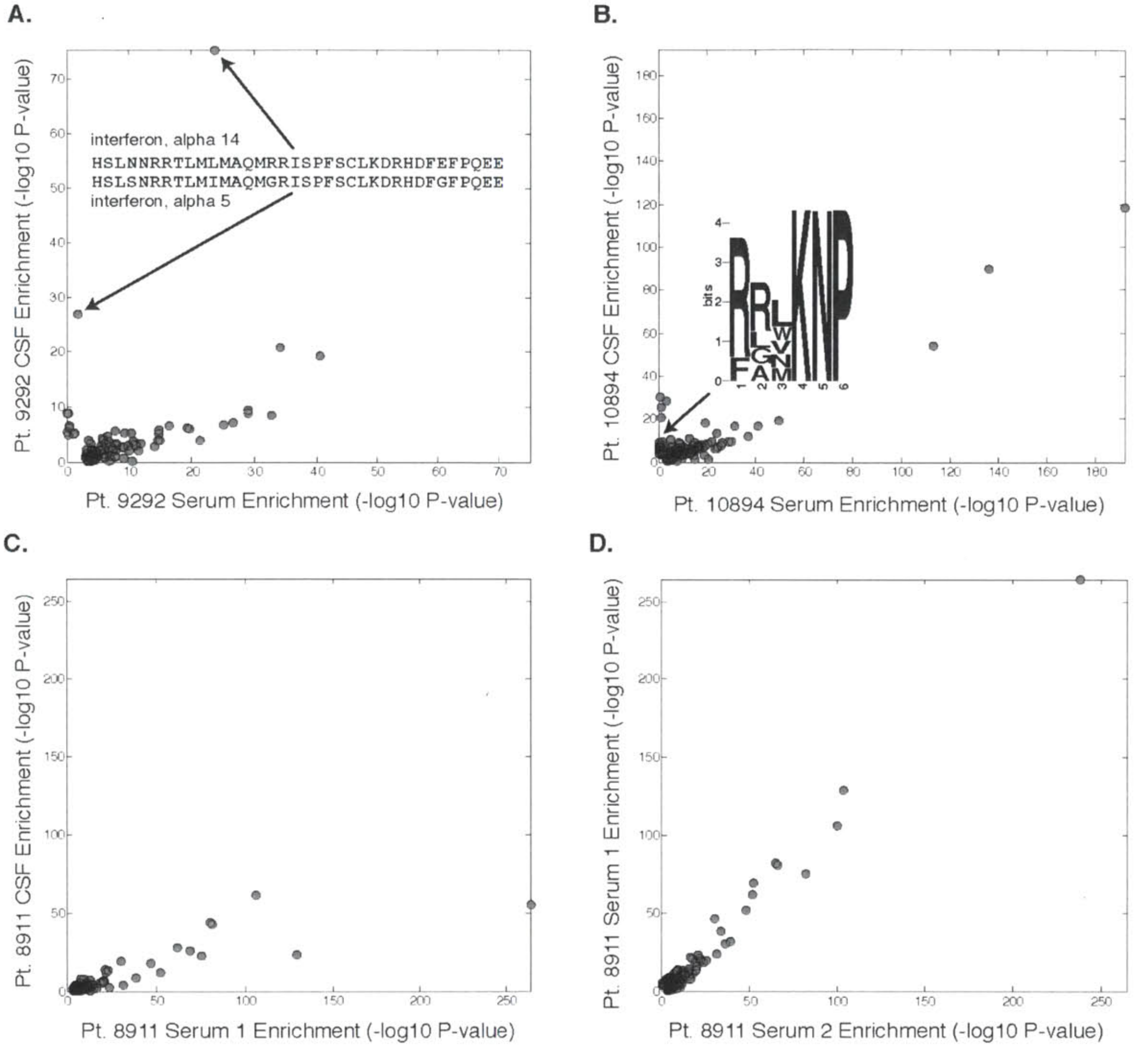


Figure 4.4: Analysis of MS patient CSF/serum pairs.

Scatter plots of matched samples from the same individuals. Each sample was analyzed in duplicate and the average -log10 P-value is plotted for peptides enriched by more than -log10 P value of 3 in both duplicates.

A. Patient 9292 enrichments in CSF versus serum, and enrichment of nearly identical peptides from IFN-α5/14 specifically in the CSF.

B. Patient 10894 peptide enrichments in CSF versus serum, and CSF specific enrichment of the motif shown.

C. Scatter plot of patient 8911 peptide enrichments in CSF versus serum is characteristic of most pairs - no CSF specific peptide enrichments, and a general depletion of autoreactivity compared to serum.

D. The serum PhIP-Seq profile in patient 8911 is unchanged over time.

We were intrigued by the observation that the ACVR2B_15 peptide, which was very frequently enriched in the sera of healthy individuals (Figure 4.1A), was enriched specifically in the CSF compartment of MS patient 9358 (-log10 P-value of 17.8 in CSF compared to 0.8 in serum; Table 4.3). In hopes of identifying an epitope within the ACVR2B_15 peptide, we searched for peptide enrichments that were highly correlated with ACVR2B_15. MEME analysis revealed a motif shared by 3 peptides that were only enriched when ACVR2B was also enriched (12/12, 11/11 and 10/10 enrichments). Interestingly, the most significant viral peptide alignment was again found within the proteome of EBV, but this time with the latent membrane protein-1 (LMP-1; E value = 0.03; sequence: LTEEVANK).

	-log10 P-value				
Patient	CSF	Serum	Peptide sequence	Symbol	Gene name
9292	27.0	1.7	HSLSNRRTLMIMAQMGRISPFSCLKDRHDFGFPQEE	IFNA5	interferon, alpha 5
	8.9	0.1	IIANALSSEPACLAEIEEDKARRILELSGSSSEDSE	PRKDC	similar to protein kinase, DNA-activated, catalytic polypeptide
	8.8	0.2	RPLTTQKLILRVESLLEVRPGNTRQKKQEDHSSGSL	LOC283682	hypothetical protein LOC283682
	6.5	0.4	AFDVQASPNEGFVNQNITIFYRDRLGLYPRFDSAGR	HYAL2	hyaluronoglucosaminidase 2
	5.4	0.0	QFQLLEQEITKPVENDISKWKPSQSLPTTNSGVSAQ	NEDD9	neural precursor cell expressed, developmentally down-regulated 9
	4.9	0.2	AESLLEAGDMLQFHDVRDAAAEFLEKNLFPSNCLGM	KLHL25	kelch-like 25 (Drosophila)
	5.2	1.2	MVLGKVKSLTISFDCLNDSNVPVYSSGDTVSGRVNL	ARRDC3	arrestin domain containing 3
	5.0	1.1	NKVLIAQKLHECARCGKNFSWHSDLILHEQIHSGEK	ZNF311	zinc finger protein 311
9358	17.0	0.8	VDEYMLPFEEEIGQHPSLEELQEVVVHKKMRPTIKD	ACVR2B	activin A receptor, type IIB
	10.1	0.6	TDTSLTMDIYFDENMKPLEHLNHDSVWNFHVWNDCW	TGM1	transglutaminase 1 (K polypeptide epidermal type I, protein-glutamine-gamma-glutamyltransferase)
	8.9	2.0	QIQVTHGKVDVGKKAEAVATVVAAVDQARVREPREP	TTN	titin
	6.9	0.6	KSQLQKVSGVFSSFMTPEKRMVRRIAELSRDKCTYF	RIN2	Ras and Rab interactor 2
	6.2	0.1	ALGEFVLVEKDVKISKKGKIYNLNEGNAKYFDRAVT	N/A	N/A
	7.2	1.7	PFPSSPPFPSSPPFPSSPPFPSSPPFPSSPPFPSSP	N/A	N/A
	6.3	0.9	MAELQQLQEFEIPTGREALRGNHSALLRVADYCEDN	ABI3	ABI family, member 3
	6.5	1.4	LVNSLKVWGKKRDRKSAIQDIRISPDNRFLAVGSSE	EML6	echinoderm microtubule associated protein like 6
	5.8	1.1	IRMPPLRNVGAGGVSGAIRTPRPMGQEASVTTGLGR	ELFN1	extracellular leucine-rich repeat and fibronectin type III domain containing 1
	6.0	2.1	IRLPSLYHVLGPTAADAGPESEKGDEEVCEPAVSPP	POPDC2	popeye domain containing 2
	5.0	1.4	KKRSLWDTIKKKKISASTSHNRRVSNIQNVNKTFSV	ASPM	asp (abnormal spindle) homolog, microcephaly associated (Drosophila)
	5.2	1.7	GKDRVVSLSEKNFKQVLKKYDLLCLYYHEPVSSDKV	CASQ2	calsequestrin 2 (cardiac muscle)
	5.3	1.9	VEKDENYDPKTEDGQASQSRYSKRRIWRSVKLKDYK	ZBTB24	zinc finger and BTB domain containing 24
10894	30.4	0.9	RAACYFTMGLYEKALEDSEKALGLDSESIRALFRKA	ZC3H7B	zinc finger CCCH-type containing 7B
	25.6	1.0	AEDLEDVRAEGTEDVGTEGTEDVGAEDSEDIRAESS	N/A	N/A
	20.6	1.2	LRLEAPSPKAIVTRTALRNLSMQKGFNDKFCYGDIT	PDZD8	PDZ domain containing 8
	8.9	0.2	EDGGSEITNYIVDKRETSRPNWAQVSATVPITSCSV	TTN	titin
	9.5	1.6	SLLPEGEDTFLSESDSEEERSSSKRRGRGSQKDTRA	GTF3C1	general transcription factor IIIC, polypeptide 1, alpha 220kDa
	7.8	0.6	KVDEYTDTDLYTGEFLSFADDLLSGLGTSCVAAGRS	ASTN2	astrotactin 2
	7.0	0.1	VSDVSRDSVNLTWTEPASDGGSKITNYIVEKCATTA	TTN	titin
	6.9	0.2	RPVPGCVNTTEMDIRKC**RRLKNP**QKVKKSVYGVTEE	RGS6	regulator of G-protein signaling 6
	6.4	0.0	LLDTQRDGLQNYEALLGLTNLSGRSDKLRQKIFKER	UNC45B	unc-45 homolog B (C. elegans)
	7.6	1.9	RSRSKDEYEKSRSRSRSRSPKENGKGDIKSKSRSRS	SFRS6	splicing factor, arginine/serine-rich 6; similar to arginine/serine-rich splicing factor 6
	6.5	0.7	GEDGSRRFGYCRRLLPGGKGKRLPEVYCIVSRLGCF	DENND2A	DENN/MADD domain containing 2A
	6.7	1.2	EQKLKLE**RLMKNP**DKAVPIPEKMSEWAPRPPPEFVR	PRKRIP1	PRKR interacting protein 1 (IL11 inducible)
	5.7	0.2	LLPRTKGFTTAVKCLRGTVAAVYDVTLNF**RGNKNP**S	AGPAT3	1-acylglycerol-3-phosphate O-acyltransferase 3
	5.8	0.5	PEFEDSEEVRRIWNRAIPLWELPDQEEVQLADTMFG	C10ORF2	chromosome 10 open reading frame 2
	5.6	0.2	SPGEWQQASAGPLHLSVPEPG**RAWKNP**ERGSKSRWS	BCYRN1	brain cytoplasmic RNA 1 (non-protein coding)
	4.3	0.0	QKTCQEQELLKQEDISMTNLGSMACPIMEPLHLENT	CCDC168	coiled-coil domain containing 168
	4.7	0.7	SVGKQDKSGLLMKLQNLCTRLDQDESFSQRLPLNIE	APAF1	apoptotic peptidase activating factor 1
	4.2	0.4	DAVYLDSEEERQEYVLTQQGFIYQGSAKFIKNIPWN	TGM2	transglutaminase 2 (C polypeptide, protein-glutamine-gamma-glutamyltransferase)
	3.8	0.2	CADMYLENPKEYLTLVQGEENFSEVY**GFRLKNP**YQC	ADAMTS20	ADAM metallopeptidase with thrombospondin type 1 motif, 20
	3.6	0.1	AEDPNLNQPVWMKPCRINSSYF**RRVKNP**NNLDEIKS	CEP350	centrosomal protein 350kDa

Table 4.3: Peptides specifically enriched in CSF compared to serum

Average -log10 P-value peptide enrichments are reported if they were greater than 3 in both duplicates of CSF sample and less than 3 in both duplicates of serum sample. The CSF specific motif in patient 10894 is shown in bold.

4.4 Discussion

We report the first large scale PhIP-Seq screen of a population of individuals with different autoimmune diseases for direct comparison to healthy controls and to each other. These data provide an unbiased, proteomic-scale assessment of precise autoreactivities found within 289 independent antibody repertoires. The vast majority of autoreactivities were individually unique, lending support to the notion that each person possesses a unique "autoantibodyome", of which the impact on phenotype remains to be explored. It is interesting to note that as our database of enriched peptides grows, so will the number of peptides recurrently enriched by a small fraction of the population - a situation analogous to the ongoing identification of progressively less common alleles in sequenced genomes. Screening large numbers of genotyped individuals will additionally reveal correlations between recurrent autoreactivities and HLA haplotypes, antibody variable domain alleles, and other immunogenetic modifiers. An interesting therapeutic possibility for the highly recurrent anti-peptide antibodies would be to "repurpose" them by fusing the antigenic peptide (e.g. ACVR2B_15) to a therapeutic biologic or to a targeting molecule. Such a strategy could extend serum half-life of a biologic,[116] decrease a molecule's immunogenicity,[24] or redirect antibody dependent cellular cytotoxicity toward a malignant target.[117]

Our unbiased method revealed a large number of novel peptide autoreactivities, but when compared to RIA-determined titers of known autoantibodies, appears to suffer from a relatively high rate of false negative discovery. We detected no anti-insulin antibodies in the T1D patients, with the important caveat that we did not acid-extract insulin from the serum prior to performing PhIP-Seq, which is standard practice for the RIA assay. It is therefore possible that the anti-insulin antibodies were occupied by endogenous or injected insulin and therefore not accessible for peptide binding. Additionally, ZnT8 RIA titers were obtained using a fusion protein consisting of two allelic variants of the immunodominant epitope, and so the single consensus sequence in T7-Pep may have contributed to the low sensitivity. The most important source of our overall high false negative discovery rate, however, is most likely the limited amount of conformational structure inherent to 36 amino acid peptide tiles. We know from previous work that T7-Pep displayed peptides contain significantly more secondary structure than what is retained after reducing and denaturing polyacrylamide gel electrophoresis of the same full-length protein.[105] However, the more comprehensive findings presented here highlight the need for improved display libraries with even more structural integrity.

While we did not observe a large number of recurrent T1D-associated enrichments beyond that expected by chance, we did observe a significantly increased degree of polyreactivity inherent to the younger T1D patients, as compared to their age matched controls. To our knowledge, this finding has not been explicitly reported in the literature. Several possible factors may contribute to this finding. Perhaps most obvious is the role that HLA haplotype could play, since T1D genetic risk is tightly linked to MHC class II alleles. It would therefore be interesting to explore the effect of risk and protection-conferring alleles on PhIP-Seq polyreactivity in a sufficiently powered study. Given that T1D patients are precocious in their acquisition of polyreactivity, it is somewhat surprising that adult T1D patients were essentially equivalent to their matched counterparts. It is therefore interesting to consider the possible existence of a "risk window", during which increased polyreactivity provides more opportunities to acquire pathogenic autoreactivity.

In the late 1990's, several investigators attempted to identify the specificities of oligoclonal bands in the CSF of MS patients. Dybwad et al. screened a single oligoclonal band with three different libraries and reported the discovery of two motifs that aligned with collagens, a neurofilament protein, versican, and several viral proteins.[118] These motifs were contained within several peptides in our library, but none of them were enriched in our screens. In another study, Cortese et al. used a library of constrained nonamers to find mimitopes for CSF antibodies in 2 MS patients. One of the sequences (KPPNP) is contained within several of our library peptides. Of them, one peptide from XP_499190.1 (SQQWRENPRTQNQSAVER**KPPNP**EPVSSGEKTPEPR), was enriched by 6 MS patients and 9 non-MS patients, and so was weakly associated with MS (Fisher's P value = 0.05). Perhaps most notable of these studies, Rand et al. used a small collection of CSF samples from MS patients to screen a phage library of random hexamers.[119] They uncovered a recurrently enriched sequence (RRPFF) in several individuals, and reported alignment with the heat shock protein αB crystallin and the Epstein-Barr virus nuclear antigen (EBNA-1). The T7-Pep library contains 4 instances of the RRPFF sequence, and two of these occur in peptides recurrently enriched by many individuals in our screen, regardless of disease status. The most robustly and frequently enriched is MAGEE1_25 (RAFAEGWQALPHF-**RRPFF**EEAAAEVPSPDSEVSSYS), which is the most commonly enriched peptide by our healthy controls (31/72; Figure 4.1A). We detected MAGEE1_25 immunoreactivity in 10 of the 27 MS CSF samples and in 1/7 non-MS CSF controls. MAGEE1_25 was enriched with equal frequency in the serum of MS patients compared to healthy controls (17/29 MS and 17/29 HC). The other peptide containing the RRPFF sequence is derived from ZNF335, and was enriched by the same MAGEE1_25-enriching subset of MS and healthy donors, but to a lesser degree. Of the MS patients for which we had matching CSF and serum samples, two had MAGEE1_25 antibodies. Both of them exhibited stronger enrichment in their serum than in their CSF. Taken together, we believe our PhIP-Seq data are consistent with a scenario in which RRPFF antibodies occur with equal frequency in the serum of MS and healthy individuals, and suggest that they are unlikely to be produced specifically within the CNS. Finally, we searched for this sequence in the database of viral proteins. As reported by Rand et al. it occurs in the EBNA-1 protein of EBV, but is additionally present in the 65 kDa early nonstructural protein of human cytomegalovirus.

Similar to MAGEE1_25, the vast majority of autoreactivities observed in MS patient CSF were also observed in the serum of the same individuals, though usually to a lesser extent. This result is somewhat surprising, given that the total IgG concentration in CSF tends to be dominated by intrathecal secretion. The simplest explanation is that the majority of these intrathecal antibodies do not bind epitopes contained within T7-Pep.

We did find several non-recurrent CSF-specific autoreactivities (Table 4.3), including against interferon alpha 5/14. Naturally occurring IgG anti-IFNα autoantibodies have been reported to occur with high prevalence in patients with acute viral hepatitis (~50% in both HAV and HBV),[120] and in pharmaceutically prepared human IgG.[121] It is possible that the serum counterparts of these autoantibodies were saturated with endogenous interferon in the serum, thereby appearing (falsely) enriched in the CSF. The significance of the anti-interferon alpha, and the other seemingly CSF-specific, non-recurrent autoantibodies reported here is unknown.

In addition to ACVR2B_15, a recurrently enriched peptide from the protein ZC3H7B was enriched by the CSF, but not by the serum of patient 10894 (Table 4.3). ZC3H7B, also known as RoXaN (rotavirus X protein associated with NSP3), is involved in translation regulation and interacts directly with the rotavirus nonstructural protein NSP3.[122] Interestingly, immunohistochemical staining of this protein reveals strong cytoplasmic positivity in neuropil of the CNS, while remaining tissues stained weakly or not at all.[90]

The findings presented here point to the accumulating value of large-scale, low cost PhIP-Seq screening. As our database grows, so will our ability to detect rare, yet significantly disease-associated autoantibodies. Quantitative elucidation of disease associated polyautoreactivities will be essential to a basic understanding of complex, heterogeneous autoimmune disease pathogenesis. In a background of numerous "personal" autoreactivities, low frequency signals will only begin to emerge with data from large numbers of individual patients. An important strength of PhIP-Seq data is its universality: diverse datasets can be immediately compared, thus dramatically increasing the power of any single screen. This feature of genome-wide SNP association data has been instrumental to its success, and we imagine the same will hold true for PhIP-Seq screening. This study should be considered a prelude of what is to come, since we have uncovered only a very small fraction of all autoreactivities associated with health and disease. As DNA sequencing costs continue to decline, and as the length of synthetic DNA library oligos continues to grow, high throughput PhIP-

Seq screening will become an increasingly important approach to unraveling the immense complexity of the immune system.

4.5 Methods

4.5.1 Patient samples and autoantibody titers

T1D patient samples and matched controls

Type 1 diabetic patients (n=39, <40 years at diagnosis, male/female ratio = 1.18, average age 18±2 years, range 3-37 years) were consecutively recruited by a Belgian network of endocrinologists between May 2004 and January 2006. Blood was sampled within 7 days from clinical onset/diagnosis by the Belgian Diabetes Registry. Only diabetic patients with three or more samples during yearly follow-up by the Registry were included in this study. Age/sex-matched healthy control samples (n=41, male/female ratio = 1.18, average age 18±2 years, range 4-37 years) were obtained from patients undergoing elective minor surgery. Controls were verified to be negative for all known type 1 diabetic autoantibodies. Samples were used with respect for patient anonymity after approval by the BDR Steering group, and procedures were approved by the Biomedical Ethical Committee at VUB/University Hospital Brussels.

Insulin, GAD65, PTPRN and ZnT8 Autoantibody Radioimmunoassay

After acid charcoal extraction of the endogenous and/or injected insulin, serum was incubated with radioactive labeled human recombinant insulin (mono-125I-tyrosin-A14-insulin) in the presence and absence of an excess of unlabeled insulin. Immune complexes were precipitated using polyethylene glycol (PEG). After washing (to remove the unbound 125I-insulin), radioactivity of the PEG precipitate was measured. The IAA concentration is expressed as specific 125I-insulin binding capacity of the serum (% tracer bound of the total amount of tracer added). Sera with insulin binding ≥ 0.6% were considered IAA positive.

GAD65, PTPRN (aminoacids 603-980), and ZnT8 (gene SEC30A8 is a chimeric construct of two peptides, amino acids 268-369), were produced in-house using in vitro transcription/translation of pEX9 (cDNA) using the Promega L4600 TnT-Kit. For ZnT8, the CR variant carries 325Arg while the CW variant carries 325Trp. The chimeric CW-CR construct contains both CR and CW.[123] Precipitations were performed as above.

Islet Cell IgG Cytoplasmic Autoantibodies

Indirect immunofluorescence was performed on non-fixed cryosections of human O+ donor pancreas, calibrated to a Juvenile Diabetes Foundation (JDF)-standard (assigned arbitrarily an ICA titer of 200 JDF-units). Pancreas sections were incubated with a serial dilution of the unknown serum, washed with phosphate buffer, and attached anti-islet IgG visualized by FITC-labeled rabbit anti-human IgG gamma chain antibody. When islet immunoreactivity was detected, the exact ICA titer was determined by further serial dilution (2-fold step), and samples with titers ≥ 12 JDF-units are considered ICA+.

MS and encephalitis patient samples

A detailed clinical intake form was collected from outside investigators, summarizing the patient's neurological history, relapse features, neurological examination, MRI and CSF findings. For samples collected at the Brigham and Women's Hospital, the same information was obtained from the MS Center's clinical database. Patients were diagnosed with relapsing-remitting MS according to the McDonald criteria.

Viral encephalitis serum samples were provided by the New York State Department of Health. Sera from patients infected with West Nile virus or St. Louis Encephalitis virus were reactive in ELISA tests and were confirmed by cross species plaque reduction neutralization tests with paired acute and convalescent sera. Sera from patients with enteroviral infection were collected on the same day as spinal fluids for which PCR tests for enteroviruses were positive. Healthy control samples were collected at Brigham and

Women's Hospital from subjects self-reported to be free of MS or other autoimmune disease. All serum and CSF samples were stored in aliquots at -80°C.[124]

Human synovial fluids

Human knee synovial fluids were obtained as discarded material from patients with various arthritides undergoing diagnostic or therapeutic arthrocentesis. Arthritis diagnosis was ascertained by an American Board of Internal Medicine certified Rheumatologist and/or by review of laboratory, radiologic and clinic notes and by applying ACR classification criteria. All studies received Institutional Review Board approval.[125]

Breast cancer patient sera

Breast cancer patient serum samples were obtained from the Dana-Farber/ Harvard Cancer Center (DF/HCC) Breast SPORE Blood Bank. These samples were originally collected under Protocol #93-085 at the DF/HCC.

4.5.2 Phage Immunoprecipitation

The T7-Pep library was prepared as described previously[105] and stored at -80 °C until used. For all samples, the final amount of Ig added to each 1 ml IP mix was approximately 2 μg. Serum/plasma samples were assumed to have 10 μg/ul of Ig, and so were diluted 10x in PBS before addition of 2 μl to the IP mix. If patient samples were derived from a different fluid compartment, their protein content was measured by Bradford assay and converted to an Ig concentration in the following way. For CSF the Ig fraction was assumed to be 29% of the total protein concentration. For synovial fluid, we used the following conversion: [Ig conc] = 0.154 x [total protein conc]+0.098. Sample dilutions were performed in a 96 well polystyrene PCR plate that had been blocked overnight with 1% fraction V or agarose purified BSA (Invitrogen) in PBS to minimize the amount of Ig lost to nonspecific binding of the polystyrene plate.

Each 1 ml IP mix contained 5x10^{10} T7-Pep phage particles and 2 ng of Positive control SAPK4 C-19 antibody (Santa Cruz, sc-7585) diluted in M9LB (for 1L: 46.7 ml 20X M9 salts, 18.7 ml 20% glucose (filtered), 0.93 ml 1 M $MgSO_4$, 934 ml LB) with 100 µg/ml ampicillin. 1 ml IP mixes were placed in each well of a 96 deep well plate (Cole-Parmer, EW-07904-04). At this point, each patient sample or control was randomly assigned to a position on the IP plate and the appropriate volume for 2 µg of Ig was added to each IP. The plate was then carefully sealed with adhesive optical tape (Applied Biosystems) and placed on a rotator for 20 hours, mixing at 4 °C.

The plate was briefly centrifuged to collect volume. 40 µl of 1:1 Protein A / Protein G slurry (Invitrogen, 100-02D, 100-04D) was added to each well. The re-sealed plate was then placed on rotator for 4 hours at 4 °C.

The plate was briefly centrifuged. At this point the beads were subjected to an automated IP protocol, which was carried out on a BioMek FX liquid handling robot. Briefly, IPs were washed in 440 µl IP Wash Buffer (150 mM NaCl, 50 mM Tris-HCL, 0.1% NP-40, pH 7.5) by pipetting up and down 30 times, for a total of 3 washes. Wash buffer was removed after magnetic separation on a 96 well magnet. Beads were moved to a new, clean plate after the second wash. After the final wash, IPs were resuspended in 40 µl of pure water and transferred to a new polystyrene PCR plate. This plate was heated to 95 °C for 10 minutes and then frozen at -80 °C until next step.

4.5.3 Preparation of immunoprecipitated T7-Pep sequencing libraries

Primers used (underlined sequences anneal with initial template, x's are the index barcode):

PCR1 forward: "IS7_HsORF5_2"

ACACTCTTTCCCTACACGACTCCAGTCAGGTGTGATGCTC

PCR1 reverse: “IS8_HsORF3_2”

GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCCGAGCTTATCGTCGTCATCC

PCR2 forward: “IS4_HsORF5_2”

AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACTCCAGT

PCR2 reverse: "index N" (set of 96)

CAAGCAGAAGACGGCATACGAGAT**xxxxxx**GTGACTGGAGTTCAGACGTGT

P5_Primer:

AATGATACGGCGACCACCGA

P7_Primer_2:

CAAGCAGAAGACGGCATACGA

Internal HsORF3' “Taqman” FAM Probe:

GCCGCAAGCTTGTCGAGCGATG (modified with 5' 6-FAM-ZEN-3' Iowa Black FQ)

T7-Pep Library Sequencing Primer “T7-Pep_96_SP”:

GCTCGGGGATCCAGGAATTCCGCTGCGC

Standard Illumina Multiplex Index Sequencing Primer “Index SP”:

GATCGGAAGAGCACACGTCTGAACTCCAGTCAC

We tested the sensitivity of several DNA polymerases to residual NP-40 detergent from the wash buffer. Some of these enzymes performed poorly in the presence of this contaminant. We found the Herculase II Fusion DNA Polymerase (Agilent) to perform the most efficiently under all conditions, and so developed the following PCR protocol to recover IP'ed T7-Pep libraries. For each 50 μl PCR1 reaction, the following components were mixed with 30 μl from each IP: 8.75 μl water, 10 μl 5x Herculase Buffer, 0.5 μl of 100 mM dNTP, 0.125 μl of 100 uM IS7_HsORF5_2 forward primer, 0.125 μl of 100 uM IS8_HsORF3_2 reverse primer, and 0.5 μl of Herculase II enzyme. The reaction was then brought to 95 °C for 2 min, and cycled 30 times with the following thermal profile.

1. 95 °C, 20s
2. 58 °C, 30s
3. 72 °C, 30s

and then subjected to a final extension for 3 min at 72 °C.

A set of 96, 7 nucleotide barcode-containing primers for PCR2 were designed using the method of Meyer et al.[106] to 1) be compatible with standard Illumina multiplex sequencing, 2) be base-balanced to maximize Illumina cluster definition, and 3) have no fewer than 3 nucleotide differences between them to minimize misalignment.[106] This set of oligos was purchased from Invitrogen in 10 µl 25 uM aliquots and then diluted to a final concentration of 2.5 uM by adding 90 µl of water.

For each 50 µl PCR2 reaction, the following components were mixed with 5 µl of the appropriate index primer and 1.5 µl of unpurified PCR1 product: 27.9 µl water, 10 µl 5x Herculase Buffer, 0.5 µl of 100 mM dNTP, 0.125 µl of 100 uM IS4_HsORF5_2 forward primer, and 0.5 µl of Herculase II enzyme. The reaction was then brought to 95 °C for 2 min, and then cycled 10 times with the following thermal profile.

1. 95 °C, 20s
2. 58 °C, 30s
3. 72 °C, 30s

and then subjected to a final extension for 3 min at 72 °C.

Unpurified PCR2 product was next quantified using real time quantitative PCR on a 7500 Fast PCR-System (Applied Biosystems). Each PCR2 product was serially diluted 100 fold to a final 10,000x dilution in water. 4 µl of this dilution was added to 16 µl of master mix composed of: 4 µl water, 10 µl Universal TaqMan 2X PCR Master Mix (Applied Biosystems, PO4475), and 2 µl of a P5/FAM Probe/P7_2 mix (5 uM P5, 5 uM P7_2, and 2.5 uM FAM Probe). The thermal profile was:

1. 50 °C, 2m
2. 95 °C, 10m
3. 95 °C, 15s
4. 60 °C, 2m

and steps 3 and 4 were repeated 35 times. We estimated the DNA concentration (in ng/ul) by [Conc] = 5000*10^((Ct-3.0964)/-4.5781). 300 ng of each PCR2 product were then combined in a single tube, mixed, and run on a 2% agarose gel. The dominant band at 316 bp was cut out and column purified twice (QIAGEN).

This 96-plex pooled library was sequenced on 2 or 3 lanes of an Illumina HiSeq 2000 using 93+7 single end cycles (93 cycles from the "T7-Pep_96_SP" primer, and 7 cycles from the "Index SP" primer) to obtain between 300 and 450 million reads.

4.5.4 High throughput PhIP-Seq informatics pipeline

We developed an informatics pipeline for processing the single end, 100 nucleotide sequencing data generated from high throughput PhIP-Seq experiments. Unless otherwise noted, scripts were written in python, and are available online for download from: https://github.com/laserson/phip-stat

This pipeline was implemented on Harvard Medical School's Orchestra Shared Research Cluster. The pipeline assumes that the initial data set is a single .fastq file (not "de-multiplexed") and that the barcode is in the header of each read. If reads have been de-multiplexed one can skip fastq2parts.py and proceed to bowtie_parts_with_LSF.py. Note that these commands are for dispatch to the LSF job scheduler.

The count data for each IP was then analyzed one sample at a time by comparison to the counts obtained by sequencing the un-enriched T7-Pep library. We used our generalized Poisson significance assignment algorithm[105] to compute -log10 P-values for each peptide/sample pair. Briefly, the IP count distribution for each input count was fitted to a generalized Poisson (GP) distribution. The two GP parameters, λ and θ were then regressed to form a joint distribution between the IP counts and the GP parameters such that each IP count could be evaluated for its likelihood of enrichment.

4.5.5 Analysis of high throughput PhIP-Seq screen data

All computational analysis was performed in MATLAB software (MathWorks). Reproducibility between each replica pair was assessed as follows. Scatter plots of the log10 of the -log10 P-values were generated, and a sliding window of width 0.05 was moved in steps of 0.05 from -2 to 3 across the x-axis. The mean and standard deviation of the values within this window were calculated at each step and plotted as a function of -log10 P-values (see Supplementary Figure 4.1.A for example). For all such plots, at low -log10 P-values the standard deviation is larger than the mean. At high -log10 P-values, however, the reverse is true. For each pair, we determined the -log10 P-value at which the mean was equal to the standard deviation (analogous to the "signal" being equal to the "noise"). A histogram plot of these values are given as Supplementary Figure 4.1.B. Based on this data, we chose a -log10 P-value of 5 to be our cutoff for considering a peptide to be significantly enriched. Within each 96-well plate screened, several samples were run in duplicate so that the reproducibility of each run's automated IPs could be assessed. We found that occasionally, sequences from random clones were amplified dramatically only in one of the replicas. The cause of these potential false positives is under investigation, but they seemed to follow no particular pattern so did not contribute to disease association of enriched clones. They are unlikely to be due purely to spurious PCR amplification, as the same clones were amplified from the same wells with two independent PCR reactions using two different enzymes.

For analyses of peptide/ORF-disease association, we set all -log10 P-values less than 5 equal to 0, and -log10 P-values greater than 5 equal to 1. This allowed us to sum the "hits" for each peptide/ORF in each disease category and then to compute the P value for association using Fisher's exact test. To correct for multiple hypothesis testing, we performed a permutation analysis by randomly permuting the sample names and then calculating the "null" Fisher P-values for each peptide/ORF. This was repeated 10,000 times and a final histogram of null Fisher P-values was constructed. Finally, an "expected" Fisher P-value distribution could be calculated for each P-value by summing the null distribution from that P-value to infinity. This expected distribution indicates how many peptide/ORF associations with a P-value at least as extreme, one would expect to observe by chance alone, given the same dataset with randomly permuted sample names. We corrected for bias due to differences in the total number of hits between samples by requiring that the difference in total number of hits after permutation is less than 1% compared to before permutation.

5. Conclusions and Future Directions

This chapter provides the outlines of future and ongoing studies involving the HMM scFv and T7-Pep libraries. In section 5.1 considerations related to optimal display platforms for the HMM scFv library are discussed, and presented in the context of a recently developed system for phage assisted continuous evolution (PACE). With improved methods for performing scFv selection, may come the ability to parallelize selections on multiple antigens simultaneously. In section 5.2, benefits of this advance are considered, as well as the library-versus-library technologies that may be utilized for antibody-antigen deconvolution. Section 5.3 details a new approach to the analysis of library evolution data. We have posed the problem in terms of Bayesian inference of library members' fitness. Computationally solving this problem for a complex library is a difficult algorithmic challenge, and we are in the process of "crowdsourcing" a solution in collaboration with the Harvard Catalyst. In addition to improving our statistical methods, we are in the process of creating a new "T7-Pep2 " peptidome library, which is based on Agilent's latest 300 mer technology (section 5.4). T7-Pep2 will display 90 amino acid peptide tiles which overlap adjacent tiles by 45 amino acids. The larger folding domains are certain to improve our detection sensitivity. To complement PhIP-Seq, we have developed a method to display and then deep sequence full-length ORFs in ribosome display format. Section 5.5 describes this technology, and provides some preliminary, proof-of-principle data. Section 5.6 presents ideas about how to combine massively parallel DNA synthesis and sequencing for the discovery of T cell epitopes. Such technologies might someday synergize with B cell epitope discovery methods like PhIP-Seq. Finally, section 5.7 includes a description of the most recent high throughput PhIP-Seq screen, as well as some considerations for the design of these studies.

5.1 An alternative display platform for the HMM scFv library

When we set out to construct and display a rationally designed scFv library that could conveniently integrate with deep sequencing, we examined available molecular display platforms for their appropriateness. Expecting that our library might suffer from the lack of framework diversity, we prioritized platforms with the greatest attainable diversity. The complexity of *in vivo* systems are limited by their transformation efficiency, making it a formidable challenge to create libraries with diversities $>10^9$, and so we chose to utilize a purely *in vitro* technology. Of these, ribosome display, mRNA display,[126] *in vitro* compartmentalization,[127] STABLE,[128] CIS,[129] and CAD[130] DNA display methods have all demonstrated feasibility for selecting scFv binders from extremely complex libraries. Each technology has its own associated strengths and weaknesses, mostly related to convenience and genotype-phenotype coupling efficiency. We adopted ribosome display for its proven track record in selecting high affinity scFvs, and its relative simplicity of implementation.

Most reports of successful ribosome display involve improving upon affinities of either immune libraries or previously selected binders. Indeed this method resulted in the affinity maturation of a previously selected anti-bovine insulin scFv to 82 pM.[46] Certainly affinity maturation applications leverage the strengths of ribosome display: 1) high complexity scFvs libraries can be rapidly constructed or moved from any display platform into ribosome display format, and 2) mutagenic library amplification and/or recombination[56] is straightforward and so repeated rounds of selection can be performed in rapid succession. The drawbacks of ribosome display, however, are also numerous. Since the mRNA must remain intact throughout the entire selection cycle (normally hours), sensitivity to nuclease contamination is an important concern. For target antigens produced in bacterial cells (such as the GST-fusion proteins in Chapter 2), this is especially germane. Second, antigenic targets containing DNA/RNA-binding domains are problematic, as the background binding of the library mRNA may prohibit successful scFv enrichment. The third, and perhaps most important caveat, however, is the efficiency of binder recovery. Indeed we and others can achieve respectable enrichments with each round of selection (several hundred fold), but the fractional recovery of positive controls is never more than ~0.2% in our hands. This fact is not often discussed in the literature, but is probably the main reason why reports of successful ribosome display selections starting with nonimmune or unenriched synthetic libraries are relatively uncommon.

Well aware of the first two limitations of ribosome display, we constructed our HMM scFv library in this format, and produced DNA template with combinatorial CDR diversity of ~10^{12}. In our standard selection protocol, we begin with an input library of ~10^{13} molecules of mRNA (~6 μg). Given the low efficiency noted above (~10^{-3}), we are in actuality only sampling ~10^{10} molecules roughly once on average. To ensure minimal sampling of each clone, we should like to have ~10 copies of each,[131] thereby bringing the effectively sampled library down to ~10^{9}. Therefore, our HMM scFv library of 10^{12} complexity, in reality behaves more like a minimally sampled library of ~10^{9}. Because complexity was our primary motivation to adopt a cell free display system, we might have in turn chosen a suboptimal platform for our purposes.

For ease of subcloning, the synthetic ribosome display vector into which we assembled the HMM scFv library contains directional, flanking SfiI restriction sites. These sites can be used to easily move the library into an alternative display format. A particularly exciting recent development in M13 phage display technology argues that perhaps the library might best be served in this format. Esvelt et al. have reported the development of a system that enables the continuous directed evolution of M13 phage-borne molecules. During phage-assisted continuous evolution ("PACE"), evolving genes are transferred between host cells during a modified phage life cycle that takes place inside a chamber subjected to continuous flow. Their approach has an attractive feature that various rates of mutagenesis can be injected into the system, simulating the hypermutation phase of antibody affinity maturation. The authors use PACE to evolve the activity of T7 RNA polymerase in several interesting ways. It is not immediately clear how their system might be used to evolve high affinity scFvs, but there is no reason to believe that this goal should be out of reach. Fusing antigen bait to the F' pilus or engineering a variant of selectively infectious phage ("SIP") technology are two possible strategies. One can easily envision how deep sequencing of the HMM scFv library might combine synergistically with PACE. Sampling the system at defined time points with multiplex Illumina sequencing would immediately reveal the trajectories of evolving library subpopulations. It is also interesting to speculate on the possible affinities that could be achieved with PACE, given the possibility to continuously mutate and select clones.

5.2 Library-versus-library scFv selections and deconvolution

As noted in Chapter 2, the marriage of DNA deep sequencing with antibody selection strategies is certain to become an important source of innovation. Perhaps most exciting is the prospect of performing massively parallel antigen-specific selections in a single reaction vessel. From an industrial manufacturing perspective, such an approach might dramatically reduce the costs of affinity reagent production. One can imagine performing large, parallel selections and then rescuing or re-synthesizing monoclonal antibodies for individual customers. However, rather than making single monoclonal antibodies, one might consider instead using collections of monoclonal scFvs for multiplex assays. Applications range from parallel analysis of immobilized complex protein mixtures, to single cell analyses of surface protein expression. The latter example is of particular relevance, since the PCR amplification of bound scFv-expressing phage allows analysis of protein expression at levels not normally accessible to affinity reagents. Multiplex protein analysis with collections of scFvs will most efficiently be performed by deep sequencing the H3 CDR, which will require *a priori* knowledge of the antigenic targets that correspond to each H3 sequence. This means that at some point during scFv collection development, deconvolution of antibody-antigen pairs is required. Strategies for "library-versus-library" interaction analysis have been developed in a number of different systems, and will play an essential role in the deconvolution phase of highly multiplex selections, if not in the selections themselves.

Perhaps the most widely utilized platform for the discovery of protein-protein interactions is the yeast two hybrid (Y2H) assay. It was originally developed by splitting the GAL4 transcription factor into two domains - one for binding to DNA, and the other for transcriptional activation. These protein fragments are fused to bait and prey, rendering host cells white unless GAL4 is reconstituted by bait-prey binding to drive the transcription of β-galactosidase, resulting in blue colonies on X-gal.[8, 132] A number of variations on this system have been reported, including bacterial two hybrid,[133] yeast cytoplasmic[134] or extracellular interaction systems.[135, 136] A particularly interesting innovation utilized a Cre recombinase to physically link the interacting pairs of ORF DNA, thus paving the way for true library-versus-library Y2H screens that can be analyzed by high throughput sequencing.[137]

Further developments in protein-protein interaction screens utilize enzymatic complementation strategies that fuse bait and prey to two separately inactive halves of a split enzyme.[138, 139] This type of protein fragment complementation assay (PCA) was first demonstrated with a split β-lactamase construct that could be used in a colorimetric assay (either *in vitro* or *in vivo*).[140] An improved *in vivo* PCA based on murine dihydrofolate reductase (mDHFR) has proven particularly successful in a number of studies.[141-143] In this strategy, reconstitution of essential dihydrofolate reductase activity in the presence of the E. coli DHFR inhibitor, trimethoprim, confers survival to an E. coli host cell. Importantly, selection of scFvs, as well as cognate antigen-scFv pair matching has successfully been demonstrated in this system.[144, 145] A more flexible, life-death selection system based on complementation of the yeast cytosine deaminase (yCD) has also been reported.[146] Several additional, alternative PCA strategies have been developed and deserve mention. Split luciferases (both firefly and renilla) can be used *in vitro* and *in vivo* to recover interacting proteins.[147, 148] As an alternative, reconstitution of protein splicing activity of DnaE intein enzyme produces a functional transcription factor which then drives luciferase expression.[149] Fluorescence or FRET complementation can also be used effectively in conjunction with cell sorting[150-153].

Several variants of phage display have potential for library-versus-library applications. Selectively infectious phage (SIP) utilizes dependence of M13 infectivity on the reconstitution of a split pIII protein, part of which is expressed by the bacterial host cell.[154, 155] Some novel variation on SIP might be well suited for use with the type of continuous evolution system described by Esvelt et al.[156] Others have reported mating libraries of scFv-expressing M13 phage with yeast or T7-expressing antigen libraries.[157, 158]

Finally, fully cell free systems have been described, which may be suitable for high throughput library-versus-library selections. McGregor et al. have developed a platform they termed interaction-dependent PCR (IDPCR), which could in theory be adapted to the analysis of interacting protein pairs displayed in one of the DNA display formats (see 5.1).[159] Other cell free systems rely on emulsion methods to link the genes of interacting proteins for subsequent analysis after breaking the emulsions.[160]

For multiple reasons, the library-versus-library platforms described above are better suited to the deconvolution of existing antibody-antigen pairs, than they are to the production of high affinity antibodies. One can thus envision a two-step production process to create complex sets of scFvs. In the first step, selections are performed on a mixture of antigens, and since each target antigen is relatively dilute compared to components they might share (e.g. GST, linkers, common domains, etc), great care must be taken to avoid the preferential expansion of scFvs that bind these "off target" components. Negative selections and decoy strategies will thus be critical to the success of parallel selections. In a second phase of production, the selected scFvs and antigens would be introduced into an appropriate library-versus-library platform for the deconvolution of binding partners. A particularly promising platform is the mDHFR system described above, as it has been optimized by Mossner et al, to identify cognate scFv-antigen pairs at an extremely high level of signal to noise (~10^7).[145] One caveat of the *in vivo* systems is the transformation barrier to library complexity >10^9. Looking forward to truly proteomic scale scFv selections, one can imagine ~10^4 antigens being used to select a set of at least 10^5 scFvs (at a minimum to ensure maximal coverage). The set of pairwise combinations (10^9) will therefore be largely undersampled with a transformation-limited library of ~10^9. One possibility would be to prepare a library of E. coli expressing the antigen-mDHFR(N) on one plasmid, and then using a library of lambda phage to bring in the scFv-mDHFR(C). By expressing Cre recombinase in the bacteria, engineered plasmids could be linked together[137] and if successful complementation occurs, host cells will survive treatment with trimethoprim. The linked

DNA can then be rescued and undergo paired end sequencing for deconvolution of the antibody-antigen pairs.

5.3 Algorithmic developments and the TopCoder competition

As demonstrated in this thesis, large libraries of DNA molecules can now be routinely analyzed using next generation DNA sequencing. If the total number of sequencing reads is sufficiently large, one can estimate each member's relative abundance within the library population. Specific sequences become preferentially enriched or depleted in the presence of a selective pressure, depending on their "fitness" relative to the rest of the population. In addition, sequence abundance can change due to stochastic fluctuation, and uncertainty is compounded by imprecision in the measurements. Current approaches to estimating fitness rely on modeling the underlying process, and then fitting the data to a distribution. The weakness of this approach is that the true underlying distribution is only very rarely known, and so the goodness of fit becomes a subjective, "good enough" assessment. In this case of PhIP-Seq data (Chapters 3 and 4), count data was fit with a generalized Poisson distribution, and required a computationally intensive algorithm.

We have initiated a project with the "crowdsourcing" company TopCoder, Inc. (Glastonbury, CT; www.topcoder.com) to run a public algorithmic development contest. This approach leverages algorithm and software developers from around the world to compete to solve problems for prize money. The Harvard Catalyst has chosen to sponsor our project as a proof-of-principle that such crowdsourced algorithms can bring value to the biomedical community.

We are hopeful that the outcome of this contest will be improved statistical methods for analyzing count data that describes an evolving population. The competition requires representative and appropriate test data, and so we have developed a general Bayesian model to describe changes in clonal abundance due to an underlying fitness distribution. The population is then sampled to some depth, resulting in a distribution of counts that reflects relative clonal abundance. Competitors will be given input counts (Z) and output counts (X) that were generated using the Bayesian model below, and asked to infer the fitness values (w).

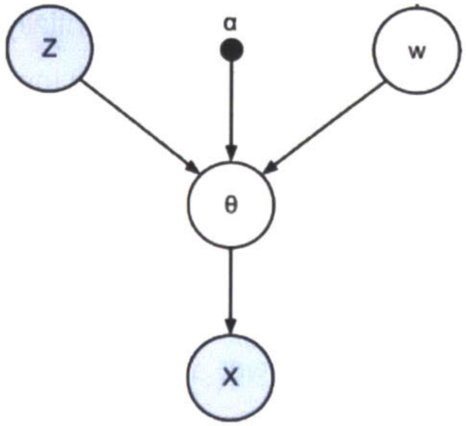


Figure 5.3.1: Bayesian inference model of clonal fitness

Model parameters are defined as follows:

Zi are the observed pre-enrichment input counts

Xi are the observed post-enrichment output counts

X ~ Multinomial(theta)

Theta ~ Dirichlet(Zi,wi)

Theta integrates the pre-enrichment clone i abundance Zi and "fitness" wi to produce post-enrichment abundance Xi.

Alpha can be used to inject noise into the post-enrichment counts.

The distribution of wi reflects the composition of the population, as well as the type and strength of selective pressure which has been applied to it. Approximating wi is the goal of the exercise, as this determines which clones to prioritize for follow up analysis. Simulation data can easily be generated from the model, and will serve as test data for the competition. Some subset of the overall test data will be provided to the competitors to help them understand it, while the rest of the test data will be held back and used for final evaluation purposes.

We used a Markov chain Monte Carlo Gibbs sampling method to simulate a competitor's potential solution. Figure 5.3.2 is the result of several simulations at each value of total input reads (n_i) and total output reads (n_o). The simulated challenge consisted of 1,000 rows of data, and reads per row ranged from 1 to 10,000. The score of the solution is calculated as the mean of the Spearman and Pearson correlation coefficient between the top and bottom 2% of the inferred wi's compared to the true wi's. The Markov chain was iterated 5,000 times.

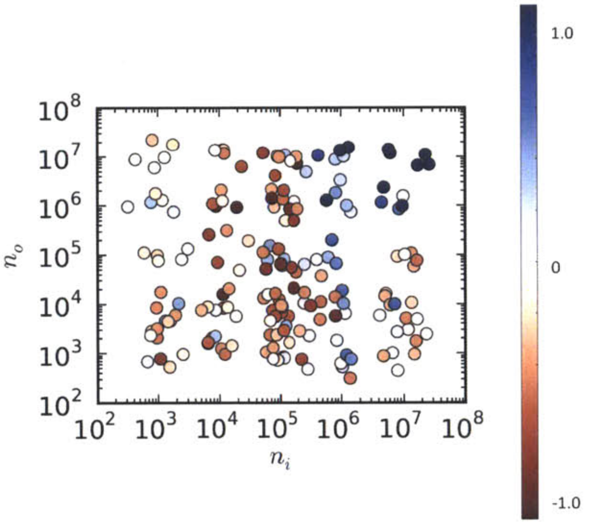


Figure 5.3.2: Simulated TopCoder challenge solution and scoring

Mean of Spearman and Pearson correlation for solutions after Gibbs sampling 1,000 rows of model data with 5,000 iterations

To our knowledge, the type of problem described here has not been adequately addressed in the scientific literature. The recent availability of both massively parallel DNA sequencing and high throughput DNA synthesis ensures the proliferation of the types of library-based selection screens described in this thesis. Our lab in particular has an urgent need for robust analysis of these screening datasets, which range from autoantigen discovery screens to screens for drug targets in breast and prostate cancer.

5.4 T7-Pep2 and beyond

Our T7-Pep library is the first synthetic representation of a normalized, complete human peptidome. It is composed of 36 amino acid peptide tiles, each of which overlaps its neighbors by 7 amino acids. Many (but not most) autoantibodies recognize short, linear motifs, and so the library peptides are sufficient for their detection even at peptide junctions (Chapter 3). We also discovered that the 36 amino acid peptides retain a significant degree of conformational information, since we could robustly detect GAD65 autoantibodies that did not recognize the fully reduced and denatured protein by PAGE and western blot analysis. However, when we examined the overall sensitivity of T7-Pep to detect a set of previously measured autoantibodies in a collection of T1D patients, the false negative discovery rate was found to be quite high (Chapter 4). This latter finding should not be a surprise, though, given that antibody combining surfaces on natively folded proteins tend to be dominated by "discontinuous" epitopes, which are patches of ~4-14 amino acid side chains formed by two or more noncontiguous peptides brought into proximity during protein folding[96, 97]. When the protein is artificially split up into its constituent 36 amino acid peptides, relative antibody affinity is expected to decrease due to 1) the loss of contacts contributed by noncontiguous residues, and 2) the increased entropic costs of binding a free peptide as opposed to the natively constrained surface.

In the six years that have elapsed since the production of the first T7-Pep oligo libraries, inkjet printing of DNA has significantly advanced in terms of coupling chemistry and spot density, thereby allowing the production of even longer, higher quality oligo libraries of increased complexity.[161] In a recent collaboration, we have designed and obtained a 300 mer peptidome library (Figure 5.4.1). This library, "T7-Pep2 ", provides a number of display advantages, and significantly reduces the sequencing depth required for analysis of the library.

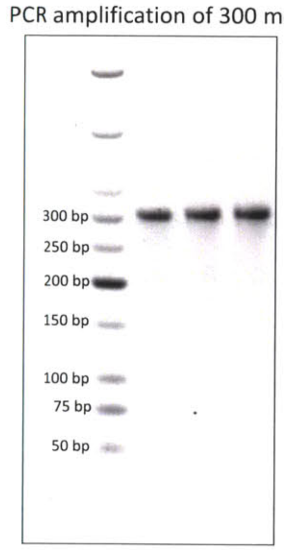


Figure 5.4.1: Results from PCR amplification of a 300 mer oligo library
(image courtesy of supplier)

T7-Pep2 oligos encode 90 amino acid peptide tiles that overlap each neighbor by 45 amino acids. The increase in protein-protein interaction detection sensitivity should be substantial. If one considers our attempt to identify interactors of the DNA damage protein RPA2, where we identified only 2 of 5 known binding partners due to disruption of the interaction motif in the 3 that were missed, the benefits can be immediately appreciated. In this example, because the T7-Pep2 overlaps (45 aa) are larger than the RPA interaction motif, in principle none of the binding partners would be missed. In addition, the increased degree of structural information inherent to 90 amino acid peptides suggests that discontinuous epitopes will often be present in the library.

In addition to the conformational benefits expected in T7-Pep2 , we have incorporated sequence design improvements. At the protein level, the new library is based on the latest build of the consensus RefSeq database (as of June, 2011), and so is composed of peptides far more likely to be expressed. We have also included all splice junctions and exon products not initially included in the library's selected full length isoforms. The full complexity of the T7-Pep2 library is ~250,000. At the nucleotide level, in addition to removing rare codons and restriction sites for library cloning, we have randomized the codon usage in T7-Pep2 for two reasons. First, homologous proteins often contain stretches of highly similar sequences, and so short read sequencing may not completely resolve two homologous peptides. In the case that the codons are randomized, however, sequencing just a short distance into the library insert will resolve even proteins sharing identical stretches of peptide sequence. Whereas we sequenced T7-Pep with 100 nt reads, we expect to be able to sequence T7-Pep2 with a maximum of 50 nt reads or less. The second reason we've utilized randomized codons is because it is expected to improve our success rate during an oligo assembly process that links together adjacent 300 mer oligos into a library of 570 mers. This strategy is described below.

The aim of our assembly strategy is to perform massively parallel splinted ligation of adjacent 300 mer tiles. We have therefore encoded the peptidome on 2 sublibraries - an "A" sublibrary and a "B" sublibrary - such that the B sublibrary can serve as the "guide" strand for splinted ligation of adjacent A tiles. A and B sublibraries have different primer sequences so that they can be separately amplified. Preparation of the DNA libraries for assembly requires the mixture of essentially three engineered sublibraries (Figure 5.4.2):

The A sublibrary is prepared as two different subpools, "A1" and "A2" (these begin as simply two aliquots from the PCR amplified A library). A1 is PCR amplified with a reverse primer that includes a SapI restriction (type IIS) site. Digestion with SapI removes the reverse primer sequence, and leaves behind a 5' phosphate on the "bottom" strand of the A1 sublibrary. Treatment with ExoI exonuclease digests away the reverse strand, leaving behind the A1 forward strand minus the reverse (3') primer sequence. A1 is now ready for assembly. The A2 sublibrary is amplified with a forward primer that contains a U base (instead of a T) at the terminal 3' position, and a reverse primer that is 5' phospohorylated. Once again digesting with ExoI removes the bottom strand, and treatment with uracil DNA glycosylase releases the forward primer sequence containing the U. The product of these reactions is the A2 sublibrary, which is again the top strand of sublibrary A, but this time minus the forward (5') primer sequence.

Finally, the B sublibrary is prepared by simply PCR amplification with a 5' phosphorylated forward primer and then digestion with ExoI. The result is a "B3" sublibrary composed of the bottom strand, and with forward and reverse primer sequences intact on both the 5' and 3' ends.

These three prepared sublibraries (A1, A2, and B3) can then be mixed together in equimolar amounts. They should be heated to 95 C for several minutes and then allowed to anneal at an extremely slow rate (perhaps 1 C per hour) down to about 60 C, and then held for some time. At this point, E. coli DNA ligase can be introduced to the system so that the correctly assembled tri-oligo complexes will enable ligation between adjacent A1 and A2 sublibrary members. It is crucial that the ligation takes place under conditions which foster the maximal amount of specificity. After the ligation is complete, the product may be separated by agarose gel electrophoresis, and the correct size band (570 nt single stranded DNA) isolated. PCR amplification with the same A library PCR primers can then be used to prepare dsDNA for subsequent cloning. The final library should be analyzed by paired end sequencing in order to assess both the specificity of pairing and the fraction of the library which was successfully assembled. The properly assembled A library ("AA") will be composed of 540 nt of coding sequence for 180 amino acid peptide tiles. These will be offset from each other by 90 amino acids (since each A library member is prepared for both 3' and 5' ligation as A1 and A2, respectively), and therefore overlap each neighbor by 90 amino acids. The assembled "BB" library can be prepared in essentially the exact same way as the AA library, by forming tri-oligo complexes between B1, B2 and A3, prior to ligation. A notable feature of our novel assembly strategy is that it is recursive, meaning that adjacent 570 mer tiles can be further brought together to form a 1,110 nt long oligo encoding a 360 amino acid polypeptide. Given the fact that the average full length human protein is about this size, this possibility is extremely exciting.

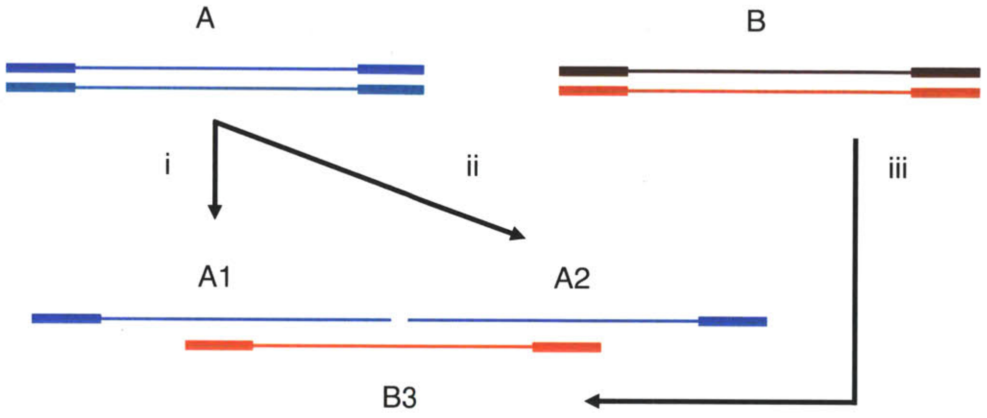


Figure 5.4.2: Assembly strategy for T7-Pep2

A and B sublibraries are shown after independent amplification. The top strands are colored darker and the bottom strands are colored lighter. Process i involves SapI and ExoI treatment. Process ii involves ExoI and glycosylase treatment. Process iii involves ExoI treatment. A1, A2 and B3 are shown after annealing and before ligation.

After amplifying and cloning T7-Pep2 , we sequenced 36 clones, and found about 2/3 of the sequences to be perfectly true to their design (suggesting a chemical coupling efficiency of ~99.9%). Several sequences were difficult to map, as they had large deletions or were likely the result of chimeric crossover PCR. Several sequences had frameshifting indels, but certainly the majority of sequenced clones were functional. This relatively low error rate in the synthesis indeed bodes well for the eventual quality of T7-Pep2 , as well as for the possibility of massively parallel assembly.

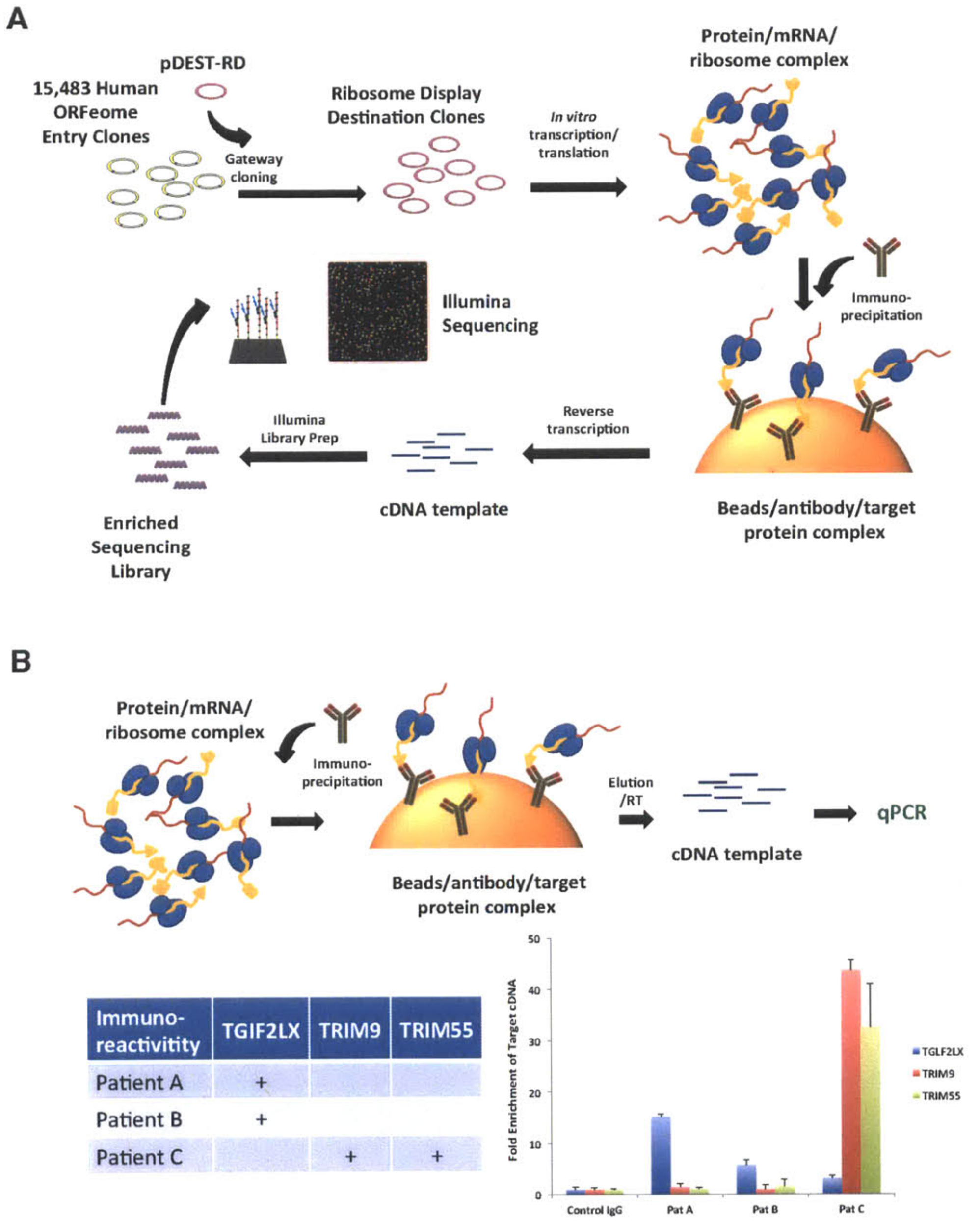
In addition to the human peptidome, we plan to encode a novel library of peptides that tile the proteomes of all viral species with human tropism. Whereas a large number of human cDNA phage display libraries have been prepared in the past, analogous collections of viral libraries have been lacking. With the advent of the complex, high quality, synthetic oligonucleotide libraries now available, it is finally possible to address this need. Applications range from vaccine development to the unbiased diagnosis of infectious disease and even to the study of host-viral protein interaction networks. Our database of viruses with human tropism was taken from UniProt, and after collapse onto 90% identity clusters, contains 7,546 UniRef protein sequences from 656 unique taxa.

This project is the result of a collaboration with Qikai Xu, Ph.D. and Mamie Li, Ph.D. Qikai wrote the scripts to generate oligo sequences from given protein sequences. Mamie cloned the T7-Pep2 library. The assembly strategy was inspired by discussions with Sriram Kosuri, Ph.D.

5.5 Full Length Immunoprecipitation Sequencing (FLIP-Seq)

The T7-Pep synthetic peptidome offers a complete representation all human proteins, but it is certainly not without limitation. As mentioned above, it is well known that most autoantibodies bind to discontinuous epitopes which cannot be represented by the 36 amino acid peptides in T7-Pep. To address this issue, we initiated a project to take advantage of an established collection of full-length cloned genes (the "ORFeome"). In an effort to make high quality clones for the complete set of human open reading frames (ORFs), Vidal and others have used high-throughput PCR and subcloning technologies to generate a collection of 15,483 full length protein-coding expression vectors.[8] Our lab has a copy of this collection in-house, as well as extensive experience working with the ORFeome both as individual clones, and in a pooled format.[162]

Ribosome display has been used extensively to successfully screen antibody libraries against protein antigens to make high affinity binders (Chapter 2). To complement the PhIP-Seq technology, we have assembled a ribosome display Gateway destination vector ("pDEST-RD"), and recombined the pooled ORFeome into this vector. Prior to *in vitro* translation, the library is PCR amplified to produce linear DNA template. The ribosome display protocol of Chapter 2 is then implemented, using patient autoantibodies as bait. ORF enrichment is subsequently amenable to high-throughput analysis by performing RT-PCR on the selected mRNA, followed by deep sequencing of the 3' ORF ends. This region of the transcripts remain bound to the ribosomes even after the considerable amount of degradation that might occur during the enrichment. Pilot experiments with the PND patients' CSF (Chapter 3) have allowed us to track the enrichments of expected targets. Indeed by qPCR we observe robust enrichment of the target cDNAs (Figure 5.5.1), and this enrichment persists in the Illumina sequencing libraries.


A
pDEST-RD
15,483 Human ORFeome Entry Clones
Gateway cloning
Ribosome Display Destination Clones
In vitro transcription/ translation
Protein/mRNA/ ribosome complex
Immuno-precipitation
Illumina Sequencing
Beads/antibody/target protein complex
Reverse transcription
cDNA template
Illumina Library Prep
Enriched Sequencing Library
B
Protein/mRNA/ ribosome complex
Immuno-precipitation
Beads/antibody/target protein complex
Elution /RT
cDNA template
qPCR
Fold Enrichment of Target cDNA
0
10
20
30
40
50
Control IgG
Pat A
Pat B
Pat C
TGLF2LX
TRIM9
TRIM55

Immuno-reactivitity	TGIF2LX	TRIM9	TRIM55
Patient A	+		
Patient B	+		
Patient C		+	+

Figure 5.5.1: Full-length immunoprecipitation sequencing (FLIP-Seq).

A. The human ORFeome collection is pooled and recombined into the ribosome display vector. After *in vitro* expression, patient antibodies are used to immunoprecipitate targeted full-length proteins. Library rescue is then performed followed by massively-parallel DNA sequencing.

B. SYBR green qPCR analysis of cDNA from ribosome displayed ORFeome immunoprecipitated on PND patients' antibodies. Three known target genes for these patients are shown.

In 2011, we received a $160K Helmsley Pilot Grant through the Harvard Catalyst to develop this technology in the context of T1D. Many of the limitations associated with the PhIP-Seq approach (small epitopes, library complexity, etc) are theoretically overcome with FLIP-Seq. We have recently completed the protocol development phase and performed the first panel of FLIP-Seq experiments. These are composed mostly of control samples, and if successful will allow us to proceed to the screening of patient autoantibodies. With an appropriate model of the expected background, we will be able to calculate the P-value for autoantibody-associated enrichment for each ORF in the library.

This project is the result of a collaboration with Jian Zhu Ph.D., who has performed most of the experimental work.

5.6 Ongoing and future PhIP-Seq screens

In the time that has elapsed since construction and analysis of the large PhIP-Seq screening database described in Chapter 4, we have formed several more collaborations and screened additional patient samples from different autoimmune diseases.

In collaboration with Professor Angela Christiano Ph.D. of Columbia University, New York, we have screened a collection of plasma samples taken from patients with alopecia areata ("AA").[163] This autoimmune destruction of the hair follicle results in localized balding on the scalp. In 1–2% of cases, the condition can spread to the entire scalp (alopecia totalis, "AT") or to the entire epidermis (alopecia universalis, "AU"). We have screened 10 plasma samples: 4 are from patients with transient AA, 2 are from patients with patchy, persistent AA, 2 are from patients with alopecia totalis, and 2 are from patients with alopecia universalis. These samples were not run in duplicate, as we will be looking for shared antigen involvement among the patients. It will be interesting to compare candidate autoantigen associations with gene expression patterns within the follicle.

In collaboration with Kai Wucherpfennig M.D., Stephen Hodi Ph.D., and Glen Dranoff M.D., we have screened a collection of serum samples from melanoma patients both before and after successful immunotherapy with the anti-CTLA4 antibody Ipilimumab (trade name Yervoy) in combination with the VEGF antibody Bevacizumab (trade name Avastin).[164] In Chapter 4, we compared MS patient sera drawn at two different time points (Figure 4.4D) and found essentially no time-dependent difference in the autoreactivity profile. This finding suggests that the autoreactive repertoire is relatively stable and perturbations will be readily detectable. Five patients' serum samples were obtained before and after cancer immunotherapy treatment (separated by 14 weeks), and each of these was screened in duplicate (for a total of 20 IPs). Analysis of this dataset will be aimed at detection of newly arising antibodies in the patients' post treatment repertoire, with particular attention paid to evidence of epitope spreading on these candidate autoantigens. In the case that we observe such evidence of T cell help, we may decide to sequence the corresponding tumor cDNAs to identify missense mutations and thus potential T cell neoepitopes.

Vasculitis refers to a heterogeneous group of disorders characterized by inflammatory destruction of blood vessels. They are classified according to the size and location of the vessel they afflict. For example, Takayasu's arteritis, polyarteritis nodosa and giant cell arteritis mainly involve arteries, and can be differentiated by their anatomical predilection. Churg-Strauss syndrome involves destruction of the small and medium vessels of the lung, and is frequently accompanied by anti-neutrophil cytoplasmic antibodies (ANCA) of the p-ANCA (perinuclear staining) variety. Wegener's granulomatosis results in destruction of the blood vessels of the nose, lungs and kidneys, and is characterized by the presence of c-ANCA (cytoplasmic staining) antibodies, frequently targeting proteinase 3 (PRTN3). In collaboration with Paul Monach, M.D., Ph.D.,[165] we have screened 5 sera from patients with Churg-Strauss syndrome, 8 sera from patients with giant cell arteritis, 9 sera from patients with Wegener's granulomatosis, 7 sera from patients with polyarteritis nodosa, and 6 sera from patients with Takayasu's arteritis. Our analysis strategy will seek to uncover common autoantigens among patients in these subcategories of autoimmune vasculitis.

Vaccines to elicit anti-HIV immunity have proven to be a challenge for reasons that are not entirely clear. A number of broadly protective natural antibodies have been reported that protect against infection,[35] but these are not produced commonly in the infected population. A number of studies have suggested that self-reactivity of at least some of these idiotypes prevent them from arising. In collaboration with Roland Strong Ph.D., we have screened scFv versions of two well known anti-HIV monoclonals, B12 and 4E10, for their ability to bind peptides present in T7-Pep for comparison with a negative control scFv, IC6. The lineage of the 4E10 antibody, which is a highly matured (mutated) clone, has been traced back to one of two germline "ancestral" sequences. Both of these scFvs have been also been screened for autoreactivity that might explain their absence from the repertoire of most infected individuals. The scFvs were coupled to epoxy coated magnetic Dynabeads (Invitrogen) with ammonium sulfate to enhance coupling efficiency. The beads were added directly to the T7-Pep library for overnight complex formation and then washed in parallel with the rest of the screening plate. These samples were run in duplicate, and so after identification of reproducibly enriched peptides, we will compare the candidates to each other and to the IC6 control. Understanding the spectrum of autoreactivity inherent to these antibodies may provide insight into improved vaccination strategies.

Inclusion body myositis (IBM) is an inflammatory degenerative disease of the musculature, most significantly afflicting the arms and legs. The etiology of sporadic IBM is currently unknown, but the possibilities include (1) a primary T-cell mediated autoimmune response causing muscle damage, (2) a primary degenerative process involving abnormal protein processing leading to a secondary inflammatory response, and (3) separate and independent immune and degenerative processes caused by an external trigger.[166, 167] We have collaborated with Stephen Greenberg M.D., who has recently identified a 43 kDa protein, which is recognized by autoantibodies of ~50% of IBM patients in a western blot analysis of pooled healthy muscle cell lysates.[168] That these antibodies recognize a fully denatured antigen is especially promising from the perspective of detection by PhIP-Seq. We have therefore screened a set of six IBM patients' sera (not in duplicate) and will search the candidate list for shared antigens that might have protein isoforms expected to migrate at around 43 kDa by PAGE analysis.

In addition to these disease cohort studies, we extended our previous studies of T1D, MS and PND. Within our collection of sera from T1D patients (Chapter 4), were three patients with antibodies that stained islet cells (ICA positivity) but who were classified as antibody negative for all other known T1D autoantibodies. These sera were therefore screened in duplicate, so as to identify with high confidence, all T7-Pep reactivities in these three unusual T1D cases. Philip L. De Jager M.D., Ph.D. has kindly provided us with serum from an MS patient who had an unusual presentation that was accompanied by choreoathetosis. Autoantibodies associated with post-streptococcal neuropsychiatric disease have reported,[169, 170] but the identity of their cognate autoantigens have yet to be identified. This patient's serum was analyzed in duplicate by PhIP-Seq and the results will be compared to those from the collection of 29 previously screened MS sera. Finally, CSF from a patient with PND and a clinical presentation similar to that of Patient C in Chapter 3 has been obtained in a collaboration with Henrikas Vaitkevicius, M.D.

The studies described in this section have all been performed in a single 96 well plate, automated PhIP-Seq run, for analysis in just 2 lanes of an Illumina HiSeq 2000, at a cost of ~$25 per sample. Between receipt of samples and submission of the pooled sequencing libraries, about 4 days of work were required. Indeed no other antibody profiling technology can offer the throughput, the cost, or the comprehensiveness of PhIP-Seq technology. It should also be noted that the studies described in this section are enabled by our "universal" reference database (established in Chapter 4), to which new experiments can immediately and directly be compared.

Looking to the future, several lines of investigation are deemed important to pursue. There are so many mysterious autoimmune diseases for which the target antigen is unknown. For example, we have begun a collaboration with Daniel Brown, M.D., Ph.D. (from the lab of Arlene Sharp M.D., Ph.D.) to screen a collection of serum samples from children with juvenile idiopathic arthritis (JIA), a devastating joint disease suspected to be triggered by viral infection. The list of important autoimmune diseases that stand to benefit from cohort PhIP-Seq screening is very long, and argues in favor of setting up some centralized core facility. It is a technology that, given its price point, should be accessible to all researchers who can benefit from it. A centralized PhIP-Seq database and informatics resource center (analogous to the dbGaP database of human genetic variation) will also be essential to the realization of the technology's full potential. As such resources become available, continued screening of healthy control donors will expand our understanding of the "natural" autoantibody repertoire. It will be interesting to compare these natural repertoires with those that spontaneously develop in the context of a lowered threshold for autoantibody production (such as systemic lupus erythematosus, transplant rejection, and graft-versus-host disease). Combining these studies with expanded analyses of complex, heterogeneous diseases like T1D, RA, and MS will further our understanding of the processes that lead up to loss of tolerance.

References

1. Church, G.M. The personal genome project. *Molecular systems biology* **1**, 2005 0030 (2005).
2. MacArthur, D.G. et al. A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823-828 (2012).
3. Pagani, I. et al. The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res* **40**, D571-579 (2012).
4. Proctor, L.M. The Human Microbiome Project in 2011 and beyond. *Cell host & microbe* **10**, 287-291 (2011).
5. Mitra, R.D. & Church, G.M. In situ localized amplification and contact replication of many individual DNA molecules. *Nucleic Acids Res* **27**, e34 (1999).
6. Harris, T.D. et al. Single-molecule DNA sequencing of a viral genome. *Science* **320**, 106-109 (2008).
7. Schlabach, M.R. et al. Cancer proliferation gene discovery through functional genomics. *Science* **319**, 620-624 (2008).
8. Rual, J.-F. et al. Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**, 1173-1178 (2005).
9. Kosuri, S. et al. Scalable gene synthesis by selective amplification of DNA pools from high-fidelity microchips. *Nat Biotechnol* **28**, 1295-1299 (2010).
10. Carr, P.A. & Church, G.M. Genome engineering. *Nat Biotechnol* **27**, 1151-1162 (2009).
11. Reddy, S.T. et al. Monoclonal antibodies isolated without screening by analyzing the variable-gene repertoire of plasma cells. *Nat Biotechnol* **28**, 965-969 (2010).
12. Weinstein, J.A., Jiang, N., White, R.A., 3rd, Fisher, D.S. & Quake, S.R. High-throughput sequencing of the zebrafish antibody repertoire. *Science* **324**, 807-810 (2009).
13. Klarenbeek, P.L. et al. Inflamed target tissue provides a specific niche for highly expanded T-cell clones in early human autoimmune disease. *Annals of the rheumatic diseases* (2012).
14. McCafferty, J., Griffiths, A.D., Winter, G. & Chiswell, D.J. Phage antibodies: filamentous phage displaying antibody variable domains. *Nature* **348**, 552-554 (1990).
15. Lonberg, N. & Huszar, D. Human antibodies from transgenic mice. *International reviews of immunology* **13**, 65-93 (1995).
16. Knappik, A. et al. Fully synthetic human combinatorial antibody libraries (HuCAL) based on modular consensus frameworks and CDRs randomized with trinucleotides. *J Mol Biol* **296**, 57-86 (2000).
17. Yang, H.Y., Kang, K.J., Chung, J.E. & Shim, H. Construction of a large synthetic human scFv library with six diversified CDRs and high functional diversity. *Mol Cells* **27**, 225-235 (2009).

18. Sanders, J. et al. Crystal structure of the TSH receptor in complex with a thyroid-stimulating autoantibody. *Thyroid : official journal of the American Thyroid Association* **17**, 395-410 (2007).

19. van der Geld, Y.M., Limburg, P.C. & Kallenberg, C.G. Proteinase 3, Wegener's autoantigen: from gene to antigen. *Journal of leukocyte biology* **69**, 177-190 (2001).

20. Lagoumintzis, G. et al. Recent approaches to the development of antigen-specific immunotherapies for myasthenia gravis. *Autoimmunity* **43**, 436-445 (2010).

21. Ziegler, A.-G., Boerschmann, H. & Walter, M. Antigen-based therapy for treating childhood type 1 diabetes. *Curr Diab Rep* **9**, 98-99 (2009).

22. Ludvigsson, J. et al. GAD treatment and insulin secretion in recent-onset type 1 diabetes. *N Engl J Med* **359**, 1909-1920 (2008).

23. Smarr, C.B., Hsu, C.-L., Byrne, A.J., Miller, S.D. & Bryce, P.J. Antigen-Fixed Leukocytes Tolerize Th2 Responses in Mouse Models of Allergy. *The Journal of Immunology*, 1-10 (2011).

24. De Groot, A.S. et al. Activation of natural regulatory T cells by IgG Fc-derived peptide "Tregitopes". *Blood* **112**, 3303-3311 (2008).

25. Brusko, T. & Bluestone, J. Clinical application of regulatory T cells for treatment of type 1 diabetes and transplantation. *European journal of immunology* **38**, 931-934 (2008).

26. Ludemann, J., Utecht, B. & Gross, W.L. Anti-neutrophil cytoplasm antibodies in Wegener's granulomatosis recognize an elastinolytic enzyme. *J Exp Med* **171**, 357-362 (1990).

27. Szabo, A. et al. HuD, a paraneoplastic encephalomyelitis antigen, contains RNA-binding domains and is homologous to Elav and Sex-lethal. *Cell* **67**, 325-333 (1991).

28. Buckanovich, R.J., Posner, J.B. & Darnell, R.B. Nova, the paraneoplastic Ri antigen, is homologous to an RNA-binding protein and is specifically expressed in the developing motor system. *Neuron* **11**, 657-672 (1993).

29. Wenzlau, J.M. et al. The cation efflux transporter ZnT8 (Slc30A8) is a major autoantigen in human type 1 diabetes. *Proc Natl Acad Sci U S A* **104**, 17040-17045 (2007).

30. Lennon, V.A., Kryzer, T.J., Pittock, S.J., Verkman, A.S. & Hinson, S.R. IgG marker of optic-spinal multiple sclerosis binds to the aquaporin-4 water channel. *J Exp Med* **202**, 473-477 (2005).

31. Beck, A., Wurch, T., Bailly, C. & Corvaia, N. Strategies and challenges for the next generation of therapeutic antibodies. *Nat Rev Immunol* **10**, 345-352 (2010).

32. Lowy, I. et al. Treatment with monoclonal antibodies against Clostridium difficile toxins. *N Engl J Med* **362**, 197-205 (2010).

33. Sui, J. et al. Structural and functional bases for broad-spectrum neutralization of avian and human influenza A viruses. *Nat Struct Mol Biol* **16**, 265-273 (2009).

34. Wu, X. et al. Focused evolution of HIV-1 neutralizing antibodies revealed by structures and deep sequencing. *Science* **333**, 1593-1602 (2011).

35. Walker, L.M. et al. Broad neutralization coverage of HIV by multiple highly potent antibodies. *Nature* **477**, 466-470 (2011).

36. Hoogenboom, H.R. Selecting and screening recombinant antibody libraries. *Nat Biotechnol* **23**, 1105-1116 (2005).

37. Holliger, P. & Hudson, P.J. Engineered antibody fragments and the rise of single domains. *Nat Biotechnol* **23**, 1126-1136 (2005).
38. Carmen, S. & Jermutus, L. Concepts in antibody phage display. *Briefings in functional genomics & proteomics* **1**, 189-203 (2002).
39. Vargas-Madrazo, E., Lara-Ochoa, F. & Almagro, J.C. Canonical structure repertoire of the antigen-binding site of immunoglobulins suggests strong geometrical restrictions associated to the mechanism of immune recognition. *J Mol Biol* **254**, 497-504 (1995).
40. Lee, C.V. et al. High-affinity human antibodies from phage-displayed synthetic Fab libraries with a single framework scaffold. *J Mol Biol* **340**, 1073-1093 (2004).
41. Lloyd, C. et al. Modelling the human immune response: performance of a 1011 human antibody repertoire against a broad panel of therapeutically relevant antigens. *Protein Eng Des Sel* **22**, 159-168 (2009).
42. Zhai, W. et al. Synthetic antibodies designed on natural sequence landscapes. *J Mol Biol* **412**, 55-71 (2011).
43. Kaas, Q., Ruiz, M. & Lefranc, M.P. IMGT/3Dstructure-DB and IMGT/StructuralQuery, a database and a tool for immunoglobulin, T cell receptor and MHC structural data. *Nucleic Acids Res* **32**, D208-210 (2004).
44. Ehrenmann, F., Kaas, Q. & Lefranc, M.P. IMGT/3Dstructure-DB and IMGT/DomainGapAlign: a database and a tool for immunoglobulins or antibodies, T cell receptors, MHC, IgSF and MhcSF. *Nucleic Acids Res* **38**, D301-307 (2010).
45. Schlessinger, A., Ofran, Y., Yachdav, G. & Rost, B. Epitome: database of structure-inferred antigenic epitopes. *Nucleic Acids Res* **34**, D777-780 (2006).
46. Hanes, J., Schaffitzel, C., Knappik, A. & Pluckthun, A. Picomolar affinity antibodies from a fully synthetic naive library selected and evolved by ribosome display. *Nat Biotechnol* **18**, 1287-1292 (2000).
47. Ofran, Y., Schlessinger, A. & Rost, B. Automated identification of complementarity determining regions (CDRs) reveals peculiar characteristics of CDRs and B cell epitopes. *J Immunol* **181**, 6230-6235 (2008).
48. Schutz, F. & Delorenzi, M. MAMOT: hidden Markov modeling tool. *Bioinformatics* **24**, 1399-1400 (2008).
49. Bond, C.J., Wiesmann, C., Marsters, J.C., Jr. & Sidhu, S.S. A structure-based database of antibody variable domain diversity. *J Mol Biol* **348**, 699-709 (2005).
50. Lara-Ochoa, F., Vargas-Madrazo, E., Jimenez-Montano, M.A. & Almagro, J.C. Patterns in the complementary determining regions of immunoglobulins (CDRs). *Bio Systems* **32**, 1-9 (1994).
51. Singh, H. & Raghava, G.P. ProPred: prediction of HLA-DR binding sites. *Bioinformatics* **17**, 1236-1237 (2001).
52. Fabre-Lafay, S. et al. Nectin-4 is a new histological and serological tumor associated marker for breast cancer. *BMC cancer* **7**, 73 (2007).
53. Athanassiadou, A.M., Patsouris, E., Tsipis, A., Gonidi, M. & Athanassiadou, P. The significance of Survivin and Nectin-4 expression in the prognosis of breast carcinoma. *Folia histochemica et cytobiologica / Polish Academy of Sciences, Polish Histochemical and Cytochemical Society* **49**, 26-33 (2011).

54. Hanes, J., Jermutus, L., Weber-Bornhauser, S., Bosshard, H.R. & Pluckthun, A. Ribosome display efficiently selects and evolves high-affinity antibodies in vitro from immune libraries. *Proc Natl Acad Sci U S A* **95**, 14130-14135 (1998).
55. Feldhaus, M.J. et al. Flow-cytometric isolation of human antibodies from a nonimmune Saccharomyces cerevisiae surface display library. *Nat Biotechnol* **21**, 163-170 (2003).
56. Chodorge, M., Fourage, L., Ravot, G., Jermutus, L. & Minter, R. In vitro DNA recombination by L-Shuffling during ribosome display affinity maturation of an anti-Fas antibody increases the population of improved variants. *Protein Engineering Design and Selection* **21**, 343-351 (2008).
57. Hoogenboom, H.R. Selecting and screening recombinant antibody libraries. *Nat Biotechnol* **23**, 1105-1116 (2005).
58. Fischer, N. Sequencing antibody repertoires: the next generation. *mabs* **3**, 17-20 (2011).
59. Ravn, U. et al. By-passing in vitro screening--next generation sequencing technologies applied to antibody display and in silico candidate selection. *Nucleic Acids Res* **38**, e193 (2010).
60. Zhang, H. et al. Phenotype-information-phenotype cycle for deconvolution of combinatorial antibody libraries selected against complex systems. *Proc Natl Acad Sci U S A* **108**, 13456-13461 (2011).
61. Bowley, D.R., Jones, T.M., Burton, D.R. & Lerner, R.A. Libraries against libraries for combinatorial selection of replicating antigen-antibody pairs. *Proc Natl Acad Sci U S A* **106**, 1380-1385 (2009).
62. MacCallum, R.M., Martin, A.C. & Thornton, J.M. Antibody-antigen interactions: contact analysis and binding site topography. *Journal of Molecular Biology* **262**, 732-745 (1996).
63. Fellouse, F.A., Wiesmann, C. & Sidhu, S.S. Synthetic antibodies from a four-amino-acid code: a dominant role for tyrosine in antigen recognition. *Proc Natl Acad Sci USA* **101**, 12467-12472 (2004).
64. Fellouse, F., Barthelemy, P., Kelley, R. & Sidhu, S. Tyrosine Plays a Dominant Functional Role in the Paratope of a Synthetic Antibody Derived from a Four Amino Acid Code. *Journal of Molecular Biology* **357**, 100-114 (2006).
65. Fellouse, F.A., Barthelemy, P.A., Kelley, R.F. & Sidhu, S.S. Tyrosine plays a dominant functional role in the paratope of a synthetic antibody derived from a four amino acid code. *J Mol Biol* **357**, 100-114 (2006).
66. Hanes, J. & Pluckthun, A. In vitro selection and evolution of functional proteins by using ribosome display. *Proc Natl Acad Sci U S A* **94**, 4937-4942 (1997).
67. Zahnd, C., Amstutz, P. & Pluckthun, A. Ribosome display: selecting and evolving proteins in vitro that specifically bind to a target. *Nat Methods* **4**, 269-279 (2007).
68. Schofield, D.J. et al. Application of phage display to high throughput antibody generation and characterization. *Genome Biol* **8**, R254 (2007).
69. Zacchi, P., Sblattero, D., Florian, F., Marzari, R. & Bradbury, A.R. Selecting open reading frames from DNA. *Genome Res* **13**, 980-990 (2003).
70. Gibson, D.G. et al. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat Methods* **6**, 343-345 (2009).

71. Graham, A.L. et al. Fitness correlates of heritable variation in antibody responsiveness in a wild mammal. *Science* **330**, 662-665 (2010).

72. Faix, P.H. et al. Phage display of cDNA libraries: enrichment of cDNA expression using open reading frame selection. *BioTechniques* **36**, 1018-1022, 1024, 1026-1019 (2004).

73. Albert, M.L. & Darnell, R.B. Paraneoplastic neurological degenerations: keys to tumour immunity. *Nat Rev Cancer* **4**, 36-44 (2004).

74. Wang, X. et al. Autoantibody signatures in prostate cancer. *N Engl J Med* **353**, 1224-1235 (2005).

75. Anderson, K.S. et al. A Protein Microarray Signature of Autoantibody Biomarkers for the Early Detection of Breast Cancer. *J Proteome Res* (2010).

76. Zacchi, P., Sblattero, D., Florian, F., Marzari, R. & Bradbury, A.R.M. Selecting open reading frames from DNA. *Genome Research* **13**, 980-990 (2003).

77. Kim, Y. et al. Identification of Hnrph3 as an autoantigen for acute anterior uveitis. *Clin Immunol* **138**, 60-66 (2011).

78. Hughes, J.B., Hellmann, J.J., Ricketts, T.H. & Bohannan, B.J. Counting the uncountable: statistical approaches to estimating microbial diversity. *Appl Environ Microbiol* **67**, 4399-4406 (2001).

79. Derda, R. et al. Diversity of phage-displayed libraries of peptides during panning and amplification. *Molecules* **16**, 1776-1803 (2011).

80. Swann, J.B. & Smyth, M.J. Immune surveillance of tumors. *J. Clin. Invest.* **117**, 1137-1146 (2007).

81. Darnell, R.B. & Posner, J.B. Paraneoplastic syndromes involving the nervous system. *N Engl J Med* **349**, 1543-1554 (2003).

82. Musunuru, K. & Kesari, S. Paraneoplastic opsoclonus-myoclonus ataxia associated with non-small-cell lung carcinoma. *J Neurooncol* **90**, 213-216 (2008).

83. Consul, P. & Shoukri, M. Maximum likelihood estimation for the generalized poisson distribution. *Communications in Statistics - Theory and Methods* **13**, 1533-1547 (1984).

84. Srivastava, S. & Chen, L. A two-parameter generalized Poisson model to improve the analysis of RNA-seq data. *Nucleic Acids Research* (2010).

85. Bailey, T.L. & Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology* **2**, 28-36 (1994).

86. Almeida, L.G. et al. CTdatabase: a knowledge-base of high-throughput and curated data on cancer-testis antigens. *Nucleic Acids Research* **37**, D816-D819 (2009).

87. Rimoldi, D. et al. Efficient simultaneous presentation of NY-ESO-1/LAGE-1 primary and nonprimary open reading frame-derived CTL epitopes in melanoma. *J Immunol* **165**, 7253-7261 (2000).

88. Chen, Y.T. et al. Identification of multiple cancer/testis antigens by allogeneic antibody screening of a melanoma cell line library. *Proc Natl Acad Sci U S A* **95**, 6919-6923 (1998).

89. Blanco-Arias, P., Sargent, C.A. & Affara, N.A. The human-specific Yp11.2/Xq21.3 homology block encodes a potentially functional testis-specific TGIF-like retroposon. *Mamm Genome* **13**, 463-468 (2002).

90. Berglund, L. et al. A Genecentric Human Protein Atlas for Expression Profiles Based on Antibodies. *Molecular & Cellular Proteomics* **7**, 2019-2027 (2008).

91. Li, L., Hagopian, W.A., Brashear, H.R., Daniels, T. & Lernmark, A. Identification of autoantibody epitopes of glutamic acid decarboxylase in stiff-man syndrome patients. *J Immunol* **152**, 930-934 (1994).

92. Schwartz, H.L. et al. High-resolution autoreactive epitope mapping and structural modeling of the 65 kDa form of human glutamic acid decarboxylase. *Journal of Molecular Biology* **287**, 983-999 (1999).

93. Tanji, K. et al. TRIM9, a novel brain-specific E3 ubiquitin ligase, is repressed in the brain of Parkinson's disease and dementia with Lewy bodies. *Neurobiology of Disease* **38**, 210-218 (2010).

94. Ciccia, A. et al. The SIOD disorder protein SMARCAL1 is an RPA-interacting protein involved in replication fork restart. *Genes Dev* **23**, 2415-2425 (2009).

95. Mer, G. et al. Structural basis for the recognition of DNA repair proteins UNG2, XPA, and RAD52 by replication factor RPA. *Cell* **103**, 449-456 (2000).

96. Barlow, D.J., Edwards, M.S. & Thornton, J.M. Continuous and discontinuous protein antigenic determinants. *Nature* **322**, 747-748 (1986).

97. Jin, L., Fendly, B.M. & Wells, J.A. High resolution functional analysis of antibody-antigen interactions. *J Mol Biol* **226**, 851-865 (1992).

98. Miyazaki, K. et al. Analysis of in vivo role of alpha-fodrin autoantigen in primary Sjogren's syndrome. *Am J Pathol* **167**, 1051-1059 (2005).

99. Huang, M. et al. Detection of apoptosis-specific autoantibodies directed against granzyme B-induced cleavage fragments of the SS-B (La) autoantigen in sera from patients with primary Sjogren's syndrome. *Clin Exp Immunol* **142**, 148-154 (2005).

100. Robbins, D.C., Cooper, S.M., Fineberg, S.E. & Mead, P.M. Antibodies to covalent aggregates of insulin in blood of insulin-using diabetic patients. *Diabetes* **36**, 838-841 (1987).

101. Papachroni, K.K. et al. Autoantibodies to alpha-synuclein in inherited Parkinson's disease. *J Neurochem* **101**, 749-756 (2007).

102. Dalakas, M.C., Fujii, M., Li, M. & McElroy, B. The clinical spectrum of anti-GAD antibody-positive patients with stiff-person syndrome. *Neurology* **55**, 1531-1535 (2000).

103. P. C, C. & M. M, S. Maximum likelihood estimation for the generalized poisson distribution. *Comm. in Stats. - Theory & Methods* **13**, 1533-1547 (1984).

104. Lamesch, P. et al. hORFeome v3.1: a resource of human open reading frames representing over 10,000 human genes. *Genomics* **89**, 307-315 (2007).

105. Larman, H.B. et al. Autoantigen discovery with a synthetic human peptidome. *Nat Biotechnol* **29**, 535-541 (2011).

106. Meyer, M. & Kircher, M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor protocols* **2010**, pdb prot5448 (2010).

107. Fernando, M.M. et al. Defining the role of the MHC in autoimmunity: a review and pooled analysis. *PLoS Genet* **4**, e1000024 (2008).

108. Serafini, B., Rosicarelli, B., Magliozzi, R., Stigliano, E. & Aloisi, F. Detection of ectopic B-cell follicles with germinal centers in the meninges of patients with secondary progressive multiple sclerosis. *Brain Pathol* **14**, 164-174 (2004).

109. Edwards, J.C. et al. Efficacy of B-cell-targeted therapy with rituximab in patients with rheumatoid arthritis. *N Engl J Med* **350**, 2572-2581 (2004).
110. Saag, K.G. et al. American College of Rheumatology 2008 recommendations for the use of nonbiologic and biologic disease-modifying antirheumatic drugs in rheumatoid arthritis. *Arthritis Rheum* **59**, 762-784 (2008).
111. Pescovitz, M.D. et al. Rituximab, B-lymphocyte depletion, and preservation of beta-cell function. *N Engl J Med* **361**, 2143-2152 (2009).
112. Hauser, S.L. et al. B-cell depletion with rituximab in relapsing-remitting multiple sclerosis. *N Engl J Med* **358**, 676-688 (2008).
113. Yang, X. et al. A public genome-scale lentiviral expression library of human ORFs. *Nat Methods* **8**, 659-661 (2011).
114. Cepok, S. et al. Identification of Epstein-Barr virus proteins as putative targets of the immune response in multiple sclerosis. *J Clin Invest* **115**, 1352-1360 (2005).
115. Elphick, G.F. et al. The human polyomavirus, JCV, uses serotonin receptors to infect cells. *Science* **306**, 1380-1383 (2004).
116. Schellenberger, V. et al. A recombinant polypeptide extends the in vivo half-life of peptides and proteins in a tunable manner. *Nat Biotechnol* **27**, 1186-1190 (2009).
117. Dubrovska, A. et al. A chemically induced vaccine strategy for prostate cancer. *ACS Chem Biol* **6**, 1223-1231 (2011).
118. Dybwad, A., Flrre, O. & Sioud, M. Probing for cerebrospinal fluid antibody specificities by a panel of random peptide libraries. *Autoimmunity* **25**, 85-89 (1997).
119. Rand, K.H., Houck, H., Denslow, N.D. & Heilman, K.M. Molecular approach to find target(s) for oligoclonal bands in multiple sclerosis. *Journal of neurology, neurosurgery, and psychiatry* **65**, 48-55 (1998).
120. Ikeda, Y. et al. Naturally occurring anti-interferon-alpha 2a antibodies in patients with acute viral hepatitis. *Clinical and experimental immunology* **85**, 80-84 (1991).
121. Ross, C., Svenson, M., Hansen, M.B., Vejlsgaard, G.L. & Bendtzen, K. High avidity IFN-neutralizing antibodies in pharmaceutically prepared human IgG. *J Clin Invest* **95**, 1974-1978 (1995).
122. Vitour, D., Lindenbaum, P., Vende, P., Becker, M.M. & Poncet, D. RoXaN, a novel cellular protein containing TPR, LD, and zinc finger motifs, forms a ternary complex with eukaryotic initiation factor 4G and rotavirus NSP3. *J Virol* **78**, 3851-3862 (2004).
123. Wenzlau, J.M. et al. A common nonsynonymous single nucleotide polymorphism in the SLC30A8 gene determines ZnT8 autoantibody specificity in type 1 diabetes. *Diabetes* **57**, 2693-2697 (2008).
124. O'connor, K.C. et al. Self-antigen tetramers discriminate between myelin autoantibodies to native or denatured protein. *Nat Med* **13**, 211-217 (2007).
125. Boilard, E. et al. Platelets amplify inflammation in arthritis via collagen-dependent microparticle production. *Science* **327**, 580-583 (2010).
126. Nemoto, N., Miyamoto-Sato, E., Husimi, Y. & Yanagawa, H. In vitro virus: bonding of mRNA bearing puromycin at the 3'-terminal end to the C-terminal end of its encoded protein on the ribosome in vitro. *FEBS Lett* **414**, 405-408 (1997).
127. Tawfik, D.S. & Griffiths, A.D. Man-made cell-like compartments for molecular evolution. *Nat Biotechnol* **16**, 652-656 (1998).

128. Doi, N. & Yanagawa, H. STABLE: protein-DNA fusion system for screening of combinatorial protein libraries in vitro. *FEBS Lett* **457**, 227-230 (1999).

129. Odegrip, R. et al. CIS display: In vitro selection of peptides from libraries of protein-DNA complexes. *Proc Natl Acad Sci U S A* **101**, 2806-2810 (2004).

130. Reiersen, H. et al. Covalent antibody display--an in vitro antibody-DNA library selection system. *Nucleic Acids Res* **33**, e10 (2005).

131. Boder, E.T. & Wittrup, K.D. Optimal screening of surface-displayed polypeptide libraries. *Biotechnol Prog* **14**, 55-62 (1998).

132. Fields, S. & Song, O. A novel genetic system to detect protein-protein interactions. *Nature* **340**, 245-246 (1989).

133. Karna, S.L.R. et al. A bacterial two-hybrid system that utilizes Gateway cloning for rapid screening of protein-protein interactions. *BioTechniques* **49**, 831-833 (2010).

134. Aronheim, A., Zandi, E., Hennemann, H., Elledge, S.J. & Karin, M. Isolation of an AP-1 repressor by a novel method for detecting protein-protein interactions. *Molecular and cellular biology* **17**, 3094-3102 (1997).

135. Hu, X., Kang, S., Chen, X., Shoemaker, C.B. & Jin, M.M. Yeast surface two-hybrid for quantitative in vivo detection of protein-protein interactions via the secretory pathway. *J Biol Chem* **284**, 16369-16376 (2009).

136. Hu, X., Kang, S., Lefort, C., Kim, M. & Jin, M.M. Combinatorial libraries against libraries for selecting neoepitope activation-specific antibodies. *Proceedings of the National Academy of Sciences* **107**, 6252-6257 (2010).

137. Hastie, A.R. & Pruitt, S.C. Yeast two-hybrid interaction partner screening through in vivo Cre-mediated Binary Interaction Tag generation. *Nucleic Acids Research* **35**, e141-e141 (2007).

138. Michnick, S.W., Ear, P.H., Manderson, E.N., Remy, I. & Stefan, E. Universal strategies in research and drug discovery based on protein-fragment complementation assays. *Nat Rev Drug Discov* **6**, 569-582 (2007).

139. Michnick, S.W., Remy, I., Campbell-Valois, F.X., Vallée-Bélisle, A. & Pelletier, J.N. Detection of protein-protein interactions by protein fragment complementation strategies. *Meth Enzymol* **328**, 208-230 (2000).

140. Galarneau, A., Primeau, M., Trudeau, L.-E. & Michnick, S.W. Beta-lactamase protein fragment complementation assays as in vivo and in vitro sensors of protein protein interactions. *Nat Biotechnol* **20**, 619-622 (2002).

141. Tarassov, K. et al. An in Vivo Map of the Yeast Protein Interactome. *Science* **320**, 1465-1470 (2008).

142. 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**, 683-690 (1999).

143. Remy, I., Campbell-Valois, F.X. & Michnick, S.W. Detection of protein–protein interactions using a simple survival protein-fragment complementation assay based on the enzyme dihydrofolate reductase. *Nat Protocols* **2**, 2120-2125 (2007).

144. Koch, H., Grafe, N., Schiess, R. & Pluckthun, A. Direct Selection of Antibodies from Complex Libraries with the Protein Fragment Complementation Assay. *Journal of Molecular Biology* **357**, 427-441 (2006).

145. Mossner, E., Koch, H. & Pluckthun, A. Fast selection of antibodies without antigen purification: adaptation of the protein fragment complementation assay to select antigen-antibody pairs. *Journal of Molecular Biology* **308**, 115-122 (2001).

146. Ear, P.H. & Michnick, S.W. A general life-death selection strategy for dissecting protein functions. *Nat Meth* **6**, 813-816 (2009).

147. Paulmurugan, R. & Gambhir, S.S. Combinatorial library screening for developing an improved split-firefly luciferase fragment-assisted complementation system for studying protein-protein interactions. *Anal Chem* **79**, 2346-2353 (2007).

148. Stefan, E. et al. Quantification of dynamic protein complexes using Renilla luciferase fragment complementation applied to protein kinase A activities in vivo. *Proc Natl Acad Sci USA* **104**, 16916-16921 (2007).

149. Kanno, A., Ozawa, T. & Umezawa, Y. Intein-mediated reporter gene assay for detecting protein-protein interactions in living mammalian cells. *Anal Chem* **78**, 556-560 (2006).

150. Banning, C. et al. A flow cytometry-based FRET assay to identify and analyse protein-protein interactions in living cells. *PLoS ONE* **5**, e9344 (2010).

151. Kodama, Y. & Hu, C.-D. An improved bimolecular fluorescence complementation assay with a high signal-to-noise ratio. *Biotech.* **49**, 793-805 (2010).

152. Wilson, C.G.M., Magliery, T.J. & Regan, L. Detecting protein-protein interactions with GFP-fragment reassembly. *Nat Methods* **1**, 255-262 (2004).

153. You, X. et al. Intracellular protein interaction mapping with FRET hybrids. *Proc Natl Acad Sci USA* **103**, 18458-18463 (2006).

154. Arndt, K.M., Jung, S., Krebber, C. & Plückthun, A. Selectively infective phage technology. *Meth Enzymol* **328**, 364-388 (2000).

155. Nilsson, N., Karlsson, F., Rakonjac, J. & Borrebaeck, C.A. Selective infection of E. coli as a function of a specific molecular interaction. *Journal of molecular recognition : JMR* **15**, 27-32 (2002).

156. Esvelt, K.M., Carlson, J.C. & Liu, D.R. A system for the continuous directed evolution of biomolecules. *Nature*, 1-7 (2011).

157. Bowley, D.R., Jones, T.M., Burton, D.R. & Lerner, R.A. Libraries against libraries for combinatorial selection of replicating antigen-antibody pairs. *Proc Natl Acad Sci USA* **106**, 1380-1385 (2009).

158. Castillo, J., Goodson, B. & Winter, J. T7 displayed peptides as targets for selecting peptide specific scFvs from M13 scFv display libraries. *Journal of Immunological Methods* **257**, 117-122 (2001).

159. McGregor, L.M., Gorin, D.J., Dumelin, C.E. & Liu, D.R. Interaction-dependent PCR: identification of ligand-target pairs from libraries of ligands and libraries of targets in a single solution-phase experiment. *J Am Chem Soc* **132**, 15522-15524 (2010).

160. Nirantar, S.R. & Ghadessy, F.J. Compartmentalized linkage of genes encoding interacting protein pairs. *Proteomics* **11**, 1335-1339 (2011).

161. Leproust, E.M. et al. Synthesis of high-quality libraries of long (150mer) oligonucleotides by a novel depurination controlled process. *Nucleic Acids Research* **38**, 2522-2540 (2010).

162. Yen, H.C., Xu, Q., Chou, D.M., Zhao, Z. & Elledge, S.J. Global protein stability profiling in mammalian cells. *Science* **322**, 918-923 (2008).

163. Petukhova, L. et al. Genome-wide association study in alopecia areata implicates both innate and adaptive immunity. *Nature* **466**, 113-117 (2010).

164. Hodi, F.S. & Dranoff, G. Combinatorial cancer immunotherapy. *Advances in immunology* **90**, 341-368 (2006).

165. Monach, P.A. & Merkel, P.A. Genetics of vasculitis. *Current opinion in rheumatology* **22**, 157-163 (2010).

166. Karpati, G. & O'Ferrall, E.K. Sporadic inclusion body myositis: pathogenic considerations. *Annals of neurology* **65**, 7-11 (2009).

167. Greenberg, S.A. Inclusion body myositis: review of recent literature. *Current neurology and neuroscience reports* **9**, 83-89 (2009).

168. Salajegheh, M., Lam, T. & Greenberg, S.A. Autoantibodies against a 43 KDa muscle protein in inclusion body myositis. *PLoS ONE* **6**, e20266 (2011).

169. Dale, R.C. et al. Post-streptococcal autoimmune neuropsychiatric disease presenting as paroxysmal dystonic choreoathetosis. *Movement disorders : official journal of the Movement Disorder Society* **17**, 817-820 (2002).

170. Hahn, R.G., Knox, L.M. & Forman, T.A. Evaluation of poststreptococcal illness. *American family physician* **71**, 1949-1954 (2005).

171. Su, A.I. et al. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A* **101**, 6062-6067 (2004).

172. Strauss, A.J., van der Geld, H.W., Kemp, P.G., Jr., Exum, E.D. & Goodman, H.C. Immunological concomitants of myasthenia gravis. *Ann N Y Acad Sci* **124**, 744-766 (1965).

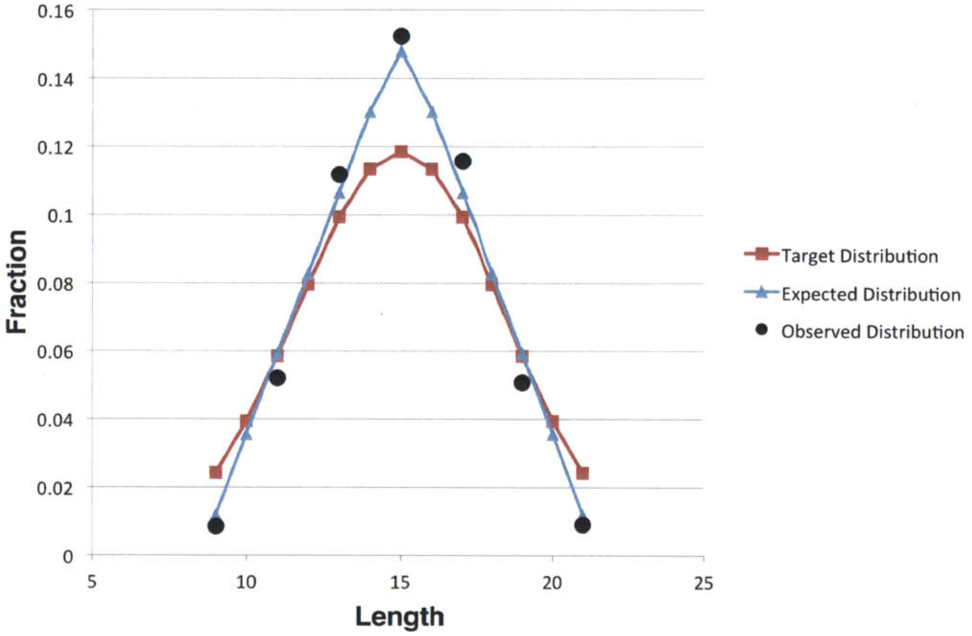
173. Aarli, J.A., Stefansson, K., Marton, L.S. & Wollmann, R.L. Patients with myasthenia gravis and thymoma have in their sera IgG autoantibodies against titin. *Clinical and experimental immunology* **82**, 284-288 (1990).

174. Gautel, M. et al. Titin antibodies in myasthenia gravis: identification of a major immunogenic region of titin. *Neurology* **43**, 1581-1585 (1993).

175. Zelinka, L. et al. Characterization of the in vitro expressed autoimmune rippling muscle disease immunogenic domain of human titin encoded by TTN exons 248-249. *Biochem Biophys Res Commun* **411**, 501-505 (2011).

Appendix 1: Supplementary Materials for Chapter 2

Supplementary Figure 2.1: Length distribution of the H3 CDR library

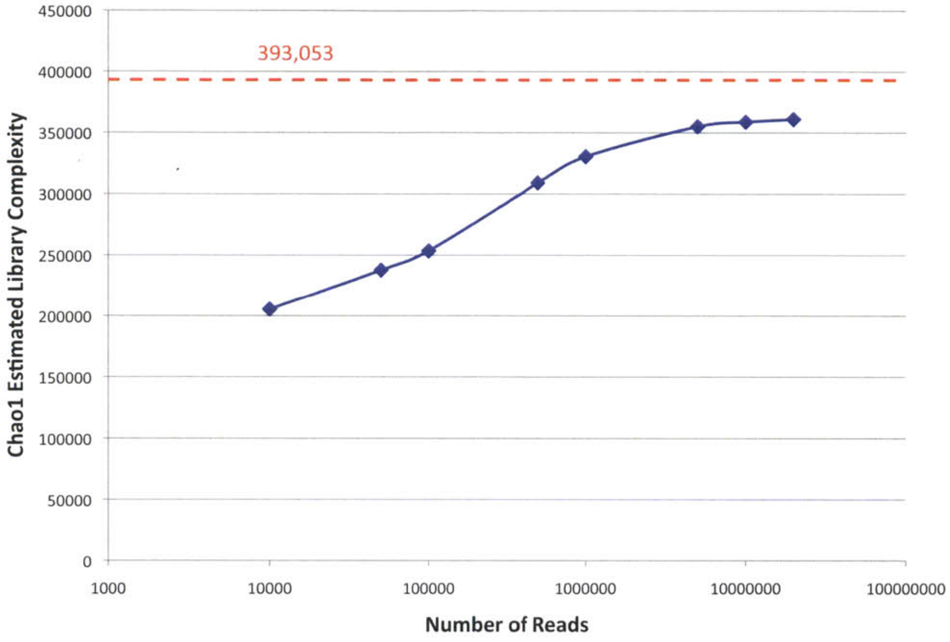


Supplementary Figure 2.1: Length distribution of the H3 CDR library

Target H3 length distribution is based on the high throughput sequencing of an individual's heavy chain repertoire. Expected distribution is the calculated fraction of each length based on random ligation of all H3L sequences with all H3R sequences. The observed distribution is based on the analysis of the Illumina sequencing data from the unselected HMM scFv library.

Appendix 2: Supplementary Materials for Chapter 3

Supplementary Figure 3.1



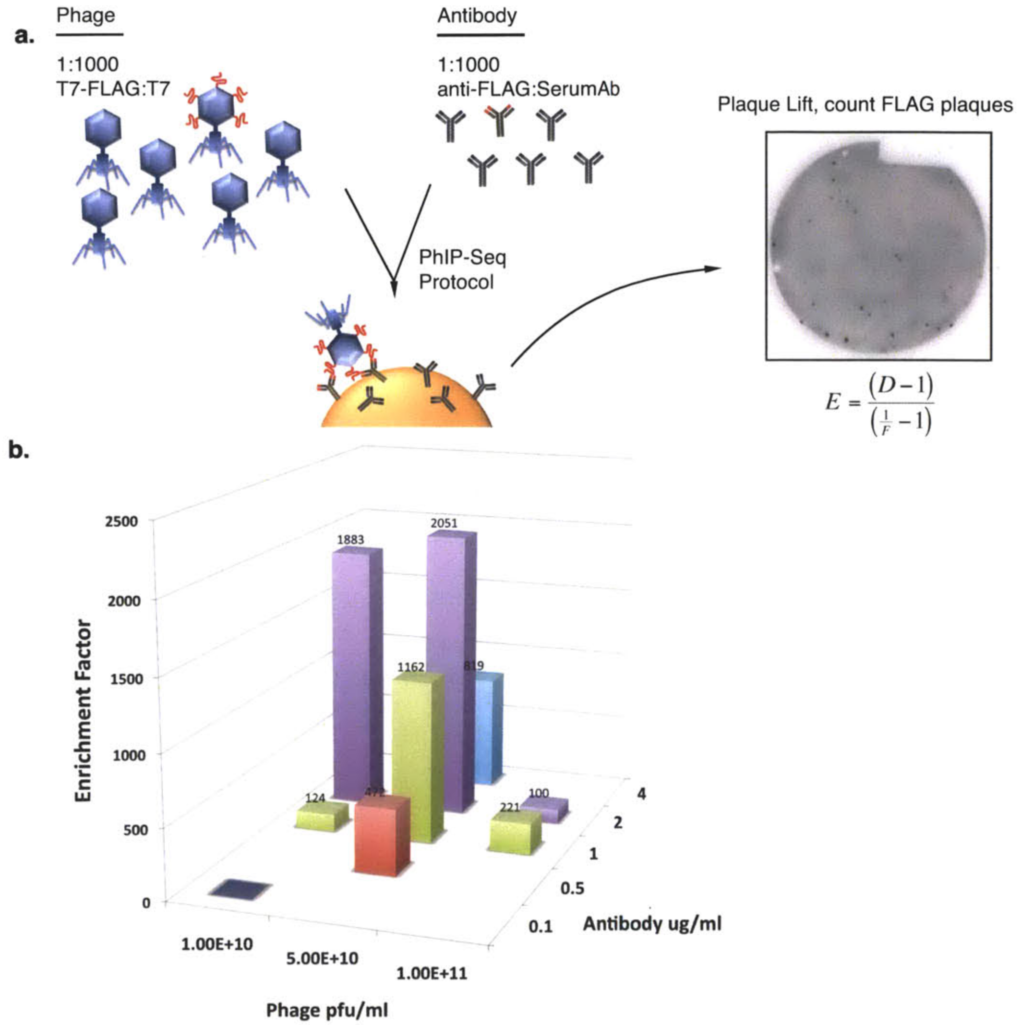
Supplementary Figure 3.1: The effect of sequencing depth on estimated library complexity

Chao1 estimates of library complexity given by

$$S_{Chao1} = S_{obs} + \frac{n_1^2}{2n_2}$$

are shown as a function of simulated T7-Pep library sampling. S_{Chao1} is the estimated complexity, where S_{obs} is the observed library complexity, n_1 is the number of library members observed once, and n_2 is the number of library members observed twice. For the data points shown, S_{obs}, n_1, and n_2 were simulated by randomly sampling the actual sequencing data “Number of Reads” times without replacement. S_{Chao1} was then calculated as above. The sequencing depth actually achieved, ~20 million reads, appears to be near saturating with respect to Chao1 estimate of the library complexity, at 361,070 library members (or ~91.8% of the 393,053 resolvable clones).

Supplementary Figure 3.2

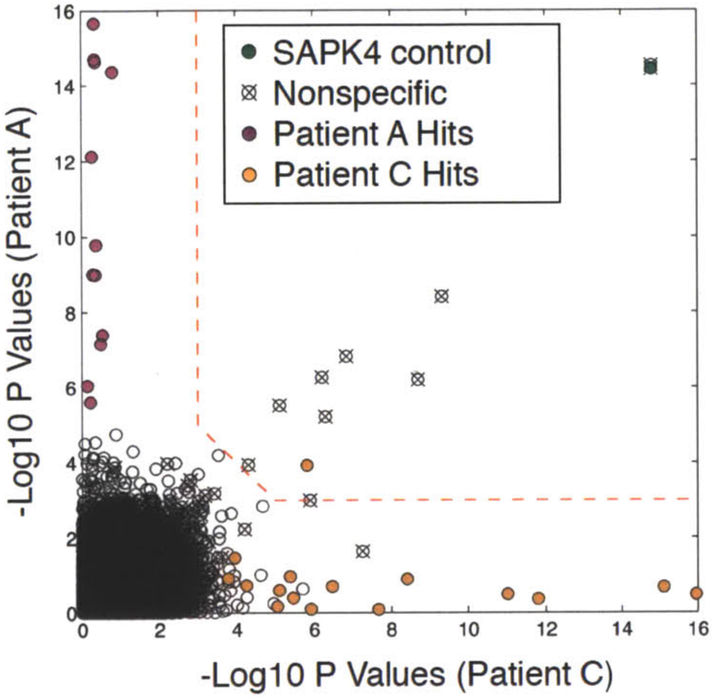


Supplementary Figure 3.2: Optimization of PhIP-Seq target enrichment

A. A FLAG-expressing T7 phage (depicted with red peptide) was diluted at 1:1,000 into native, non-FLAG-expressing T7 phage to mimic a target peptide within the T7-Pep library. An anti-FLAG monoclonal antibody (M2, Sigma-Aldrich; shown with red variable region) was diluted 1:1,000 into human serum antibodies (shown with black variable region) to mimic a rare autoantibody within a patient's antibody repertoire. After performing the IP, plaque lift analysis for FLAG expression was performed to determine enrichment using the equation shown (*E* = enrichment; *D* = dilution factor = 1,000; *F* = fraction of FLAG expressing clones on plaque lift). Enrichment was optimized with respect to type of beads, number of washes, order of antibody-phage/antibody-bead complex formation, and relative concentrations of phage and antibody.

B. Enrichment factor was found to depend on the relative concentrations of phage and antibody during complex formation. We thus varied these parameters independently and found an optimum at about 5x10^{10} pfu/ml phage and 2 mg/ml total antibody.

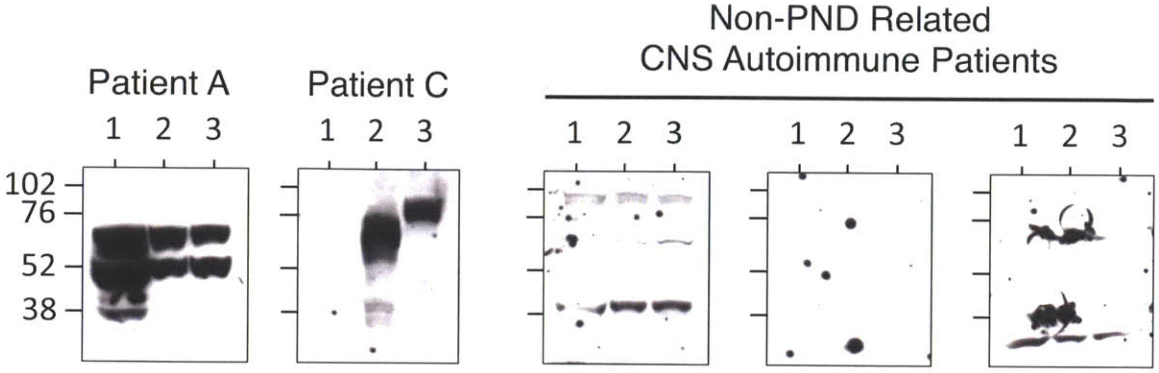
Supplementary Figure 3.3



Supplementary Figure 3.3: Comparison of PhIP-Seq experiments on different patients

Scatter plot as in Figure 2d from text, but comparing clone enrichment p-values from two different patients: Patient A (y-axis) versus Patient C (x-axis). Both experiments included the SAPK4 spike-in antibody. X'ed circles were enriched by beads and SAPK4 antibody alone (no patient antibody in IP). Filled purple and orange circles are the Patient A- and Patient C-specific positives given in Table 2 from the text.

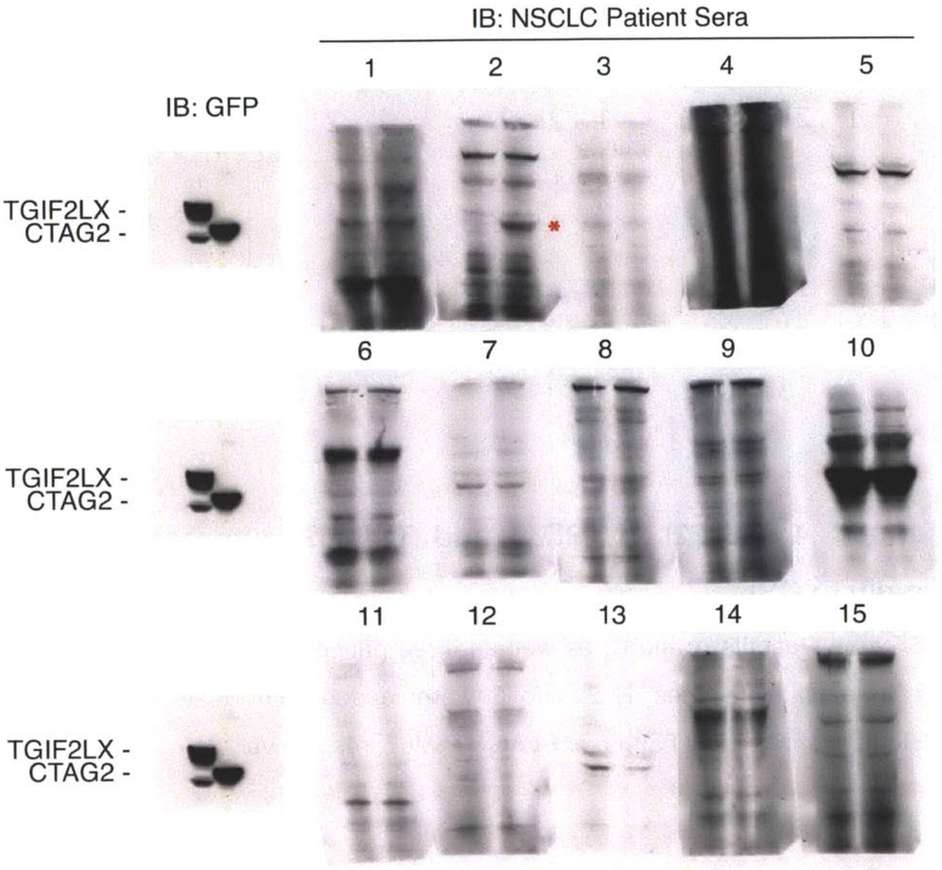
Supplementary Figure 3.4



Supplementary Figure 3.4: TGIF2LX, TRIM9 and TRIM67 autoreactivity is not present nonspecifically in CSF

Western blotting with CSF from Patients A and C, as well as three patients with non-PND related CNS autoimmune syndromes. In each blot, lanes 1, 2, and 3 were loaded with lysate from 293T cells overexpressing either TGIF2LX-GFP, FLAG-TRIM9, or FLAG-TRIM67, respectively.

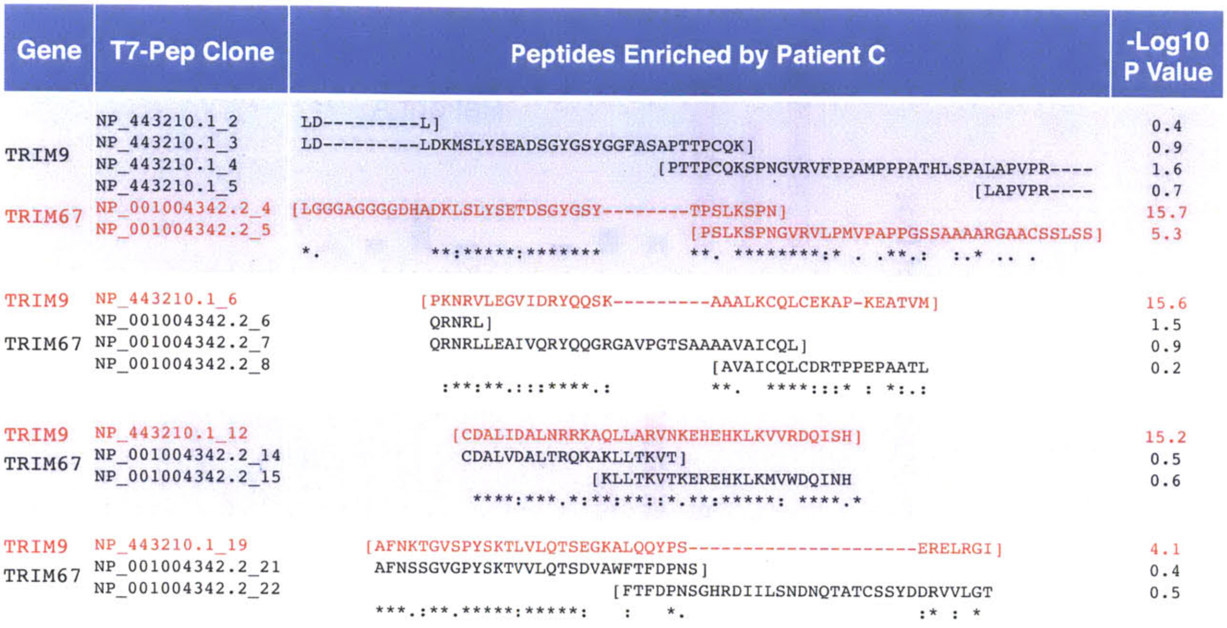
Supplementary Figure 3.5



Supplementary Figure 3.5: Immunoblots for TGIF2LX and CTAG2 reactivity in the serum of NSCLC patients without PND

Sera from fifteen non-small cell lung cancer (NSCLC) patients were used to blot SDS-PAGE separated 293T cell lysate overexpressing either TGIF2LX (left lane) or CTAG2 (right lane), fused with C-terminal GFP. Staining for GFP (left blot) demonstrates overexpression of TGIF2LX and CTAG2 at the expected weights. Only patient 2 was found to have anti-CTAG2 serum antibodies (marked by *). No patients were found to have anti-TGIF2LX serum antibodies.

Supplementary Figure 3.6

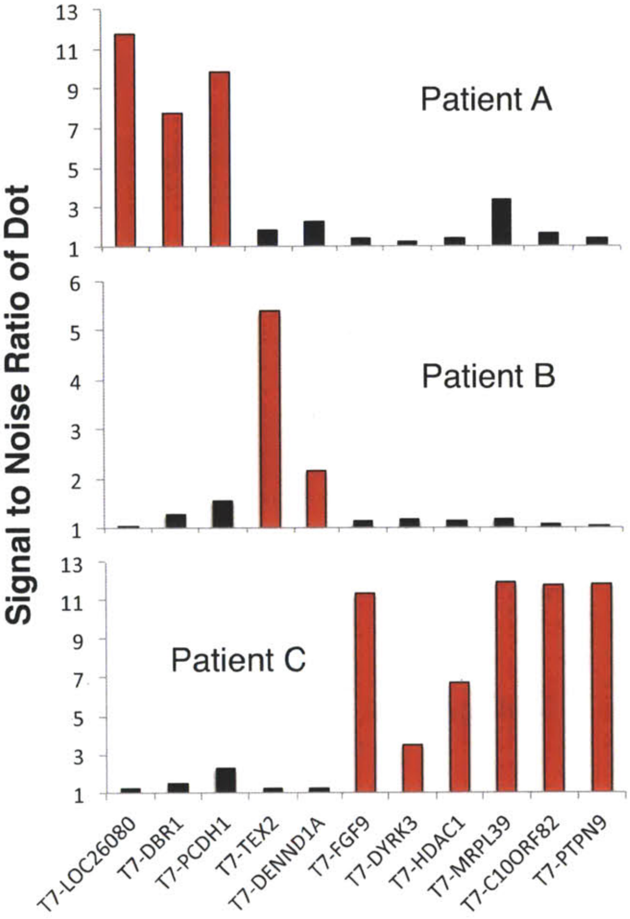


Gene	T7-Pep Clone	Peptides Enriched by Patient C	-Log10 P Value
TRIM9	NP_443210.1_2	LD---------L]	0.4
	NP_443210.1_3	LD---------LDKMSLYSEADSGYGSYGGFASAPTTPCQK]	0.9
	NP_443210.1_4	[PTTPCQKSPNGVRVFPPAMPPPATHLSPALAPVPR----	1.6
	NP_443210.1_5	[LAPVPR----	0.7
TRIM67	NP_001004342.2_4	[LGGGAGGGGDHADKLSLYSETDSGYGSY--------TPSLKSPN]	15.7
	NP_001004342.2_5	[PSLKSPNGVRVLPMVPAPPGSSAAAARGAACSSLSS]	5.3
		*. **:*****:******* **. ********:* . .**.: :.* .. .	
TRIM9	NP_443210.1_6	[PKNRVLEGVIDRYQQSK---------AAALKCQLCEKAP-KEATVM]	15.6
TRIM67	NP_001004342.2_6	QRNRL]	1.5
	NP_001004342.2_7	QRNRLLEAIVQRYQQGRGAVPGTSAAAAVAICQL]	0.9
	NP_001004342.2_8	[AVAICQLCDRTPPEPAATL	0.2
		:**:**.:::****.: **. ****:::* : *:.:	
TRIM9	NP_443210.1_12	[CDALIDALNRRKAQLLARVNKEHEHKLKVVRDQISH]	15.2
TRIM67	NP_001004342.2_14	CDALVDALTRQKAKLLTKVT]	0.5
	NP_001004342.2_15	[KLLTKVTKEREHKLKMVWDQINH	0.6
		****:***.*:**:**::*.**:*****: ****.*	
TRIM9	NP_443210.1_19	[AFNKTGVSPYSKTLVLQTSEGKALQQYPS---------------------ERELRGI]	4.1
TRIM67	NP_001004342.2_21	AFNSSGVGPYSKTVVLQTSDVAWFTFDPNS]	0.4
	NP_001004342.2_22	[FTFDPNSGHRDIILSNDNQTATCSSYDDRVVLGT	0.5
		.:**.**:*****: : *. :* : *	

Supplementary Figure 3.6: Alignment among enriched peptides from TRIM9 and TRIM67

Significantly enriched peptides (in red) from TRIM9 and TRIM67 shown with corresponding ClustalW-aligned peptides from the homologous protein (in black). Boundaries of phage-displayed peptides are denoted with brackets. Peptides are shown next to their –Log10 p-value of enrichment.

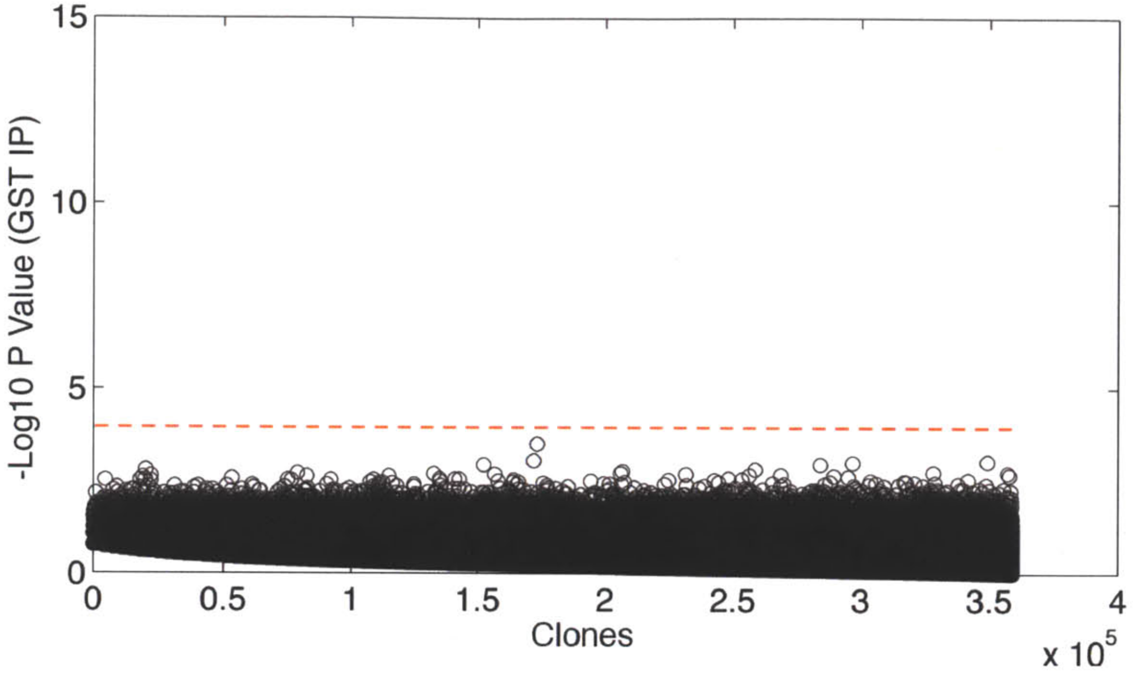
Supplementary Figure 3.7



Supplementary Figure 3.7: Quantification of T7 Candidate Dot Blots

The dot blots in Figure 3.3g were analyzed to determine the signal-to-noise ratio arising from each T7 candidate clone immunoblotted with each of the patients' spinal fluid. The data from the candidates expected to react with a given patient's antibodies are shown in red, whereas that data from the candidates that are expected not to react with a given patient's antibodies are shown in black.

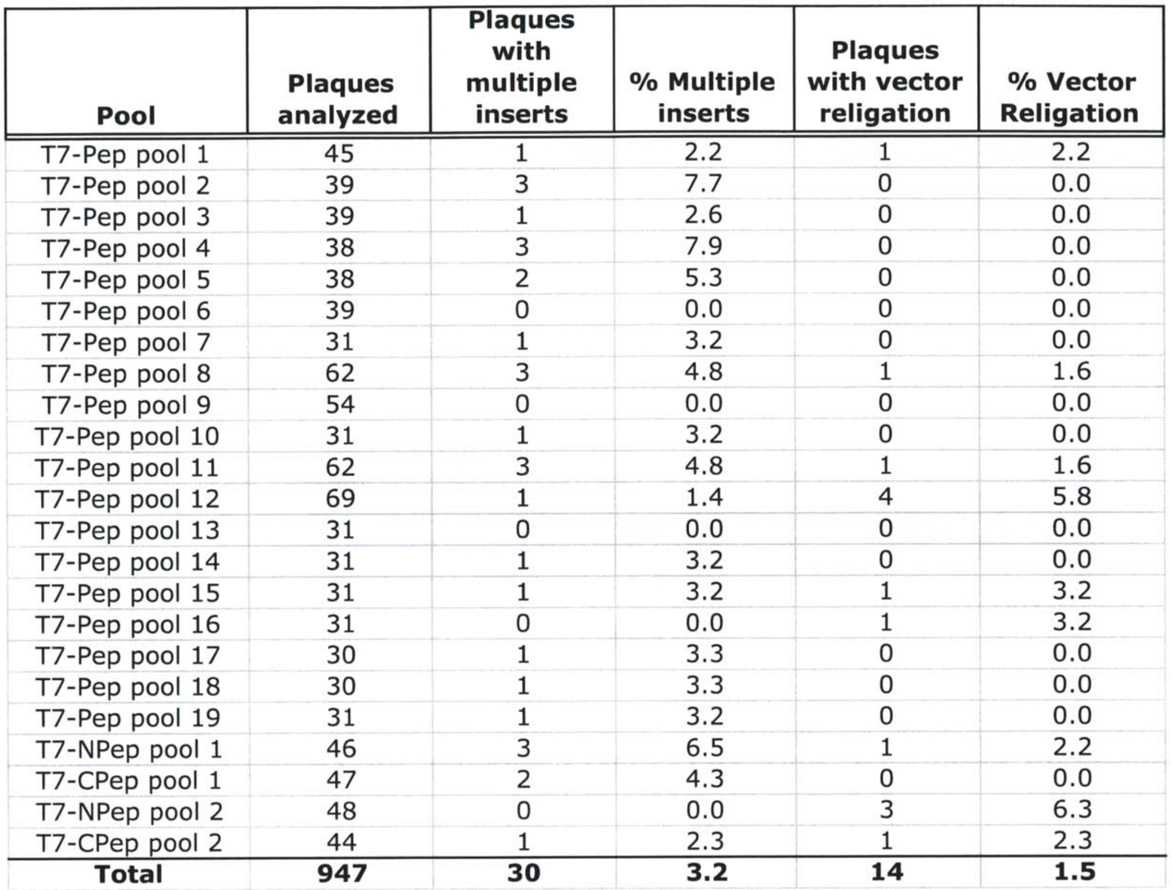
Supplementary Figure 3.8



Supplementary Figure 3.8: PhIP-Seq –Log10 p-values for T7-Pep enrichment by GST alone

GST coated glutathione magnetic beads were used to precipitate phage from the T7-Pep library. Illumina sequencing data was analyzed using the generalized Poisson method. No library members were significantly enriched by GST alone ($P < 10^{-4}$).

Supplementary Table 3.1

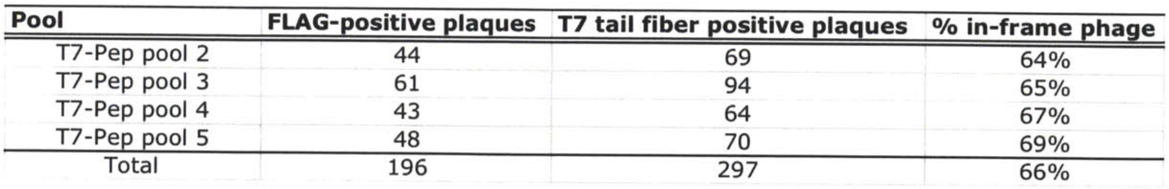


Pool	Plaques analyzed	Plaques with multiple inserts	% Multiple inserts	Plaques with vector religation	% Vector Religation
T7-Pep pool 1	45	1	2.2	1	2.2
T7-Pep pool 2	39	3	7.7	0	0.0
T7-Pep pool 3	39	1	2.6	0	0.0
T7-Pep pool 4	38	3	7.9	0	0.0
T7-Pep pool 5	38	2	5.3	0	0.0
T7-Pep pool 6	39	0	0.0	0	0.0
T7-Pep pool 7	31	1	3.2	0	0.0
T7-Pep pool 8	62	3	4.8	1	1.6
T7-Pep pool 9	54	0	0.0	0	0.0
T7-Pep pool 10	31	1	3.2	0	0.0
T7-Pep pool 11	62	3	4.8	1	1.6
T7-Pep pool 12	69	1	1.4	4	5.8
T7-Pep pool 13	31	0	0.0	0	0.0
T7-Pep pool 14	31	1	3.2	0	0.0
T7-Pep pool 15	31	1	3.2	1	3.2
T7-Pep pool 16	31	0	0.0	1	3.2
T7-Pep pool 17	30	1	3.3	0	0.0
T7-Pep pool 18	30	1	3.3	0	0.0
T7-Pep pool 19	31	1	3.2	0	0.0
T7-NPep pool 1	46	3	6.5	1	2.2
T7-CPep pool 1	47	2	4.3	0	0.0
T7-NPep pool 2	48	0	0.0	3	6.3
T7-CPep pool 2	44	1	2.3	1	2.3
Total	**947**	**30**	**3.2**	**14**	**1.5**

Supplementary Table 3.1: Subpool analysis of multiple insertions and vector religation after cloning of the T7-Pep, T7-NPep, and T7-CPep libraries

Phage plaques from each subpool were randomly selected and PCR analyzed to examine the frequency of multiple insertions and vector religations present within each pool.

Supplementary Table 3.2



Pool	FLAG-positive plaques	T7 tail fiber positive plaques	% in-frame phage
T7-Pep pool 2	44	69	64%
T7-Pep pool 3	61	94	65%
T7-Pep pool 4	43	64	67%
T7-Pep pool 5	48	70	69%
Total	196	297	66%

Supplementary Table 3.2: Subpool analysis of FLAG expression after cloning of T7-Pep

Plaque lifts from four subpools were analyzed by immunoblotting using FLAG and T7 tail fiber antibodies to measure in-frame and total plaques, respectively. Plaques staining positive were counted and a percentage of in-frame, FLAG-expressing phage was determined. The vast majority of frameshifting mutations present in the phage inserts is due to errors in DNA chemical synthesis on the releasable DNA microarrays. In parallel oligonucleotide synthesis, sequence integrity can be compromised by depurination side reactions, inefficient nucleoside coupling, and reversible 5'-hydroxyl deprotection reactions, leading to mutations of the desired oligonucleotide.

Supplementary Table 3.3

Rank	T7-Pep Clone	Peptide	Log P GST	Log P GST-RPA2	Gene Symbol
1	**NP_054859.2_1**	**MSLPLTEEQRKKIEENRQKALARRAEKLLAEQHQRT**	**0.29**	**14.61**	**SMARCAL1**
2	NP_055877.3_31	TPPSMSAALPFPAGGLGMPPSLPPPPLQPPSLPLSM	1.09	6.60	PPRC1
3	NP_006360.3_18	TLSYNGLGSNIFRLLDSLRALSGQAGCRLRALHLSD	2.23	6.59	LRRC41
4	NP_060903.2_28	AVLQQNPSVLEPAAVGGEAASKPAGSMKPACPASTS	0.07	5.95	KDM3A
5	XP_372311.2_13	LTLYDGPNVSSPSYGPYCRGDTSIAPFVASSNQVFI	1.65	5.90	LOC389958
6	NP_060876.2_13	LTPVTTSTVLSSPSGFNPSGTVSQETFPSGETTISS	1.87	5.68	MUC4
7	XP_372592.2_4	AALIHVPPLSRGLPASLLGRALRVIIQEMLEEVGKP	0.28	5.39	PGPEP1L
8	NP_057131.1_2	ITAEEMYDIFGKYGPIRQIRVGNTPETRGTAYVVYE	0.47	5.26	SF3B14
9	**NP_003353.1_3**	**AEQLDRIQRNKAAALLRLAARNVPVGFGESWKKHLS**	**0.20**	**5.23**	**UNG2**
10	NP_443728.2_10	IRPMDDDLLKLLLPLMLQYSDEFVQSAYLSRRLAYF	1.32	5.06	MED12L
11	NP_004981.2_2	ISTVGPEDCVVPFLTRPKVPVLQLDSGNYLFSTSAI	1.60	5.00	MARS
12	NP_078997.2_120	TTSTSQSAASSNNTYPHLSCFSMKSWPNILFQASAR	0.22	4.96	ZFHX4
13	NP_004697.2_27	ITETAGSLKVPAPASRPKPRPSPSSTREPLLSSSEN	2.14	4.96	ARHGEF1
14	NP_003425.2_23	SHLSRHRKTTSVHHRLPVQPDPEPCAGQPSDSLYSL	1.28	4.81	ZNF133
15	NP_996882.1_34	LDRFKNRLKDYPQYCQHLASISHFMQFPHHLQEYIE	0.54	4.80	CNOT1
16	NP_783324.1_3	PHPSALSSVPIQANALDVSELPTQPVYSSPRRLNCA	2.11	4.71	RAB3IP
17	NP_000341.1_5	PESQHLGRIWTELHILSQFMDTLRTHPERIAGRGIR	0.02	4.70	ABCA4
18	NP_689896.1_1	MNRKWEAKLKQIEERASHYERKPLSSVYRPRLSKPE	0.78	4.68	CCDC111
19	XP_095991.7_15	IKTRDICNQLQQPGFPVTVTVESPSSSEVEEVDDSS	1.04	4.65	CEP78
20	NP_741996.1_43	ITNGLAMKNNEISVIQNGGIPQLPVSLGGSALPPLG	1.60	4.60	SALL3
21	NP_997191.1_49	SVYGWATLVSERSKNGMQRILIPFIPAFYINQSELV	0.86	4.48	NUP210L
22	NP_775902.2_9	IHSGERPYECSECGKLFMWSSTLITHQRVHTGKRPY	1.31	4.41	ZNF547
23	XP_496363.1_6	PVRRGYWGNKIGKPHTVPCKVTGRCGSALVHLIPVP	1.66	4.34	RNF11
24	NP_001026.1_92	LSRKLFWGIFDALSQKKYEQELFKLALPCLSAVAGA	0.05	4.29	RYR2
25	NP_006359.3_11	THQWLDGSDCVLQAPGNTSCLLHYMPQAPSAEPPLE	0.78	4.24	CREB3
26	NP_002146.2_6	ITVPAYFNDSQRQATKDAGAIAGLNVLRIINEPTAA	1.37	4.23	HSPA6
27	NP_065987.1_4	ITPTRELAIQIDEVLSHFTKHFPEFSQILWIGGRNP	2.50	4.18	DDX55
28	NP_079279.2_13	SHHDTAVLITRYDICSSKEKCNMLGLSYLGTICDPL	0.24	4.17	ADAMTS20
29	NP_055987.1_47	PKGEPTRRGRGGTFRRGGRDPGGRPSRPSTLRRPAY	1.62	4.13	BAT2L2
30	NP_005712.1_2	IIPSCIAIKESAKVGDQAQRRVMKGVDDLDFFIGDE	1.82	4.13	ACTR3
31	NP_079165.3_2	SPPSQLFSSVTSWKKRFFILSKAGEKSFSLSYYKDH	1.75	4.13	C10ORF81
32	NP_006609.2_26	PPYKYKLRYRYTLDDLYPMMNALKLRAESYNEWALN	0.63	4.08	LOC100133760
33	NP_149163.2_41	INLTIRGHEVVGIVGRTGSGKSSLGMALFRLVEPMA	1.01	4.04	ABCC11
34	NP_001004750.1_11	IHFLFPPFMNPFIYSIKTKQIQSGILRLFSLPHSRA	0.79	4.03	OR51B6

Supplementary Table 3.3: Candidate RPA2 interacting proteins

PhIP-Seq was performed using GST-RPA2 as bait, and enrichment scores (–Log10 p-values estimated by the generalized Poisson method) were compared to enrichment on GST alone.

Appendix 3: Supplementary materials for Chapter 4

Supplementary Table 4.1

Experiment 1						
Class	Subclass	Male	Female	Age	Fluid	Total
Breast Cancer	ER+/PR+	0	28	52.3 (7.0)	serum	28
Multiple Sclerosis	19 RR, 5 SP, 2 PP, 4 ?	0	29	52.8 (6.7)	serum	29
Healthy Controls		0	29	48.1 (9.5)	serum	29

Experiment 2						
Class	Subclass	Male	Female	Age	Fluid	Total
Multiple Sclerosis						
	RR	0	6	39.9 (7.6)	serum	6
	RR	5	11	45.2 (8.9)	CSF	16
	SPMS	10	1	44.4 (7.6)	CSF	11
Controls						
	Meningitis	4	0	50 (6)	CSF	3
	PND	0	2	61 (2)	CSF	2
	SSPE	3	1	17.5 (2.6)	CSF	4

Experiment 3						
Class	Subclass	Male	Female	Age	Fluid	Total
Type 1 Diabetes		21	18	17.4 (9.4)	serum	39
Healthy Controls		21	20	20.1 (10.4)	serum	41

Experiment 4						
Class	Subclass	Male	Female	Age	Fluid	Total
Rheumatoid Arthritis						
	Seropositive				serum	10
	Seronegative				serum	10
	Seropositive	4	20	60.7 (18.3)	synovial	24
	Seronegative	8	8	54.3 (16.6)	synovial	16
Controls						
	Gout	8	2	55.2 (14.3)	synovial	10
	Osteoarthritis	2	7	65.4 (11.1)	synovial	9

Supplementary Table 4.1: Detailed composition of patient cohorts

Each experiment represents a different 96 well plate of samples, whose positions were randomized across the plate. ER+, estrogen receptor positive; PR+, progesterone receptor positive; RR, relapse remitting MS; SP, secondary progressive MS; PP, primary progressive MS; ?, unknown status; PND, paraneoplastic neurological disorder; SSPE, subacute sclerosing panencephalitis.

Supplementary Table 4.2

	T1D	NonT1D		Max P value		Epitope			PROTEIN ATLAS				GNF Expression	
Peptide	(n=39)	(n=248)	Fisher	T1D	NonT1D	Spreading	Symbol	Gene Name	EXOCRINE	ISLET	Specificity	Confidence	Pancreas	Islet
NP_002837.1_22	*3*	*0*	*0.002*	*22.09*	*3.25*	*NO*	*PTPRN*	*protein tyrosine phosphatase, receptor type, N*	*0*	*3*	*10/66*	*Supportive*	*0.87*	*10.25*
NP_683765.1_15	***3***	***0***	***0.002***	***19.77***	***3.34***	***NO***	***OXER1***	***oxoeicosanoid (OXE) receptor 1***					***NE***	***NE***
NP_443724.1_1	*3*	*0*	*0.002*	*18.55*	*4.56*	*NO*	*ADC*	*arginine decarboxylase*	*2*	*1*	*46/66*	*Low*	*NE*	*NE*
NP_848627.1_13	***3***	***0***	***0.002***	***12.7***	***4.31***	***YES***	***RNF180***	***ring finger protein 180***	***1***	***1***	***50/65***	***Uncertain***	***NE***	***NE***
NP_060250.2_52	3	0	0.002	8.19	3.2	NO	CHD7	chromodomain helicase DNA binding protein 7					NE	NE
NP_002560.1_4	3	1	0.008	21.7	39.48	NO	FURIN	furin (paired basic amino acid cleaving enzyme)	3	1	41/66	Uncertain	NE	NE
NP_000029.2_36	3	1	0.008	18.73	22.36	NO	APC	adenomatous polyposis coli	2	1	65/65	Medium	NE	NE
NP_003627.1_11	2	0	0.018	87.4	1.51	NO	KCNAB2	potassium voltage-gated channel, shaker-related subfamily, I	0	0	23/66	Medium	NE	NE
NP_005451.1_23	2	0	0.018	43.86	4.2	NO	SNCAIP	synuclein, alpha interacting protein	1	1	60/64	Uncertain	NE	NE
NP_001930.1_22	2	0	0.018	29.24	2.35	NO	DRP2	dystrophin related protein 2	2	0	10/63	Uncertain	NE	NE
NP_056139.1_3	2	0	0.018	24.13	4.43	NO	RRP8	ribosomal RNA processing 8, methyltransferase, homolog (y	2	2	57/64	Uncertain	NE	NE
NP_597681.1_278	2	0	0.018	22.06	4.35	YES	TTN	titin	0	0	34/66	Medium	NE	NE
NP_848572.1_1	**2**	**0**	**0.018**	**18.55**	**3.54**	**NO**	**BANF2**	**barrier to autointegration factor 2**	**3**		**51/63**	**Uncertain**	**NE**	**NE**
NP_065871.2_34	2	0	0.018	18.07	3.15	YES	PREX1	phosphatidylinositol-3,4,5-trisphosphate-dependent Rac excl	2	1	47/64	Supportive	NE	NE
NP_000029.2_67	2	0	0.018	15.93	4.53	NO	APC	adenomatous polyposis coli	2	1	65/65	Medium	NE	NE
NP_115821.1_29	2	0	0.018	13.99	1.36	NO	MEGF11	multiple EGF-like-domains 11					NE	NE
XP_379774.1_53	2	0	0.018	12.01	2.41	NO	TNRC18	trinucleotide repeat containing 18					NE	NE
NP_061831.1_3	2	0	0.018	11.34	2.59	NO	C15ORF2	chromosome 15 open reading frame 2					NE	NE
NP_689824.1_15	2	0	0.018	10.87	2.72	NO	LRRN4	leucine rich repeat neuronal 4	3	1	45/65	Uncertain	NE	NE
NP_000048.1_49	2	0	0.018	10.69	3.05	NO	BLM	Bloom syndrome, RecQ helicase-like	3	2	61/62	Supportive	NE	NE
NP_001012410.1_18	2	0	0.018	9.38	3.81	NO	SGOL1	shugoshin-like 1 (S	3	1	64/64	Supportive	NE	NE
XP_376567.1_53	2	0	0.018	9.35	3.8	NO	TNRC18	trinucleotide repeat containing 18					NE	NE
NP_055605.2_2	2	0	0.018	9.2	4.15	NO	JAKMIP2	Janus kinase and microtubule interacting protein 2					NE	NE
NP_733750.2_6	2	0	0.018	9.08	1.31	NO	INADL	InaD-like					NE	NE
NP_004386.2_8	2	0	0.018	8.72	2.23	NO	DBN1	drebrin 1	1	1	24/66	Uncertain	NE	NE
NP_036546.2_13	2	0	0.018	8.72	3.6	NO	RAB3GAP2	RAB3 GTPase activating protein subunit 2 (non-catalytic)	2	2	66/66	Medium	NE	NE
NP_694573.1_6	2	0	0.018	8.68	3.19	NO	ZNF75A	zinc finger protein 75a	1	1	63/66	Uncertain	NE	NE
XP_496548.1_24	2	0	0.018	8.65	3.83	NO	CAPN14	calpain 14					NE	NE
NP_054762.2_3	2	0	0.018	8.51	3.31	NO	CHMP2B	chromatin modifying protein 2B	1	0	21/66	Uncertain	1.08	4.18
NP_001442.1_11	2	0	0.018	7.96	3.64	NO	FOXF1	forkhead box F1	0	0	27/65	Uncertain	NE	NE
XP_293354.6_13	2	0	0.018	7.45	3.08	NO	DCAF8L2	DDB1 and CUL4 associated factor 8-like 2					NE	NE
NP_068765.2_34	2	0	0.018	7.21	4.91	NO	BCORL1	BCL6 co-repressor-like 1	2	2	51/65	Low	NE	NE
NP_542166.1_16	2	0	0.018	7.03	3.69	NO	UPF2	UPF2 regulator of nonsense transcripts homolog (yeast)					NE	NE
NP_775106.2_1	2	0	0.018	6.81	3.08	NO	LIN9	lin-9 homolog (C	1	0	42/64	Uncertain	NE	NE
NP_954629.1_5	2	0	0.018	6.71	4.85	NO	LHX6	LIM homeobox 6					NE	NE
NP_078849.1_14	2	0	0.018	6.41	3.33	NO	C6ORF211	chromosome 6 open reading frame 211	2	2	57/66	Medium	NE	NE
NP_003165.1_34	2	0	0.018	6.03	3.08	NO	SVIL	supervillin	2	1	66/66	Low	NE	NE
NP_061027.1_33	2	0	0.018	5.55	3.91	NO	LRP1B	low density lipoprotein-related protein 1B (deleted in tumors)					NE	NE
NP_060509.2_8	2	0	0.018	5.5	3.97	NO	KDM4D	lysine (K)-specific demethylase 4D					NE	NE
NP_849154.1_5	3	2	0.019	35.85	12.76	NO	MORN4	MORN repeat containing 4					NE	NE
NP_002841.2_2	3	3	0.035	11.31	39.61	NO	PTPRS	protein tyrosine phosphatase, receptor type, S					NE	NE
NP_055595.2_39	2	1	0.049	263.48	6.14	NO	CUL7	cullin 7	2	1	66/66	Medium	NE	NE
NP_002408.2_19	2	1	0.049	46.99	10.33	NO	MKI67	antigen identified by monoclonal antibody Ki-67	2	0	42/66	High	NE	NE
NP_005406.3_12	2	1	0.049	43.34	88.72	NO	SLC20A1	solute carrier family 20 (phosphate transporter), member 1	3	2	65/66	Supportive	NE	NE
XP_496720.1_3	2	1	0.049	26.21	6.75	NO	#N/A	#N/A					NE	NE
NP_660330.1_5	2	1	0.049	21.21	11.05	NO	ZNF519	zinc finger protein 519	3	0	20/66	Uncertain	NE	NE
NP_997336.1_7	2	1	0.049	17.46	19.12	NO	#N/A	#N/A					NE	NE
XP_496855.1_1	2	1	0.049	16.93	6.33	NO	#N/A	#N/A					NE	NE
XP_499266.1_1	2	1	0.049	16.9	5.77	NO	#N/A	#N/A					NE	NE
NP_000820.1_12	2	1	0.049	12.3	6	NO	GRIA4	glutamate receptor, ionotrophic, AMPA 4					NE	NE
NP_005954.2_14	2	1	0.049	9.45	41.05	NO	MYH1	myosin, heavy chain 1, skeletal muscle, adult	0	0	2/66	High	NE	NE
NP_597681.1_641	2	1	0.049	8.41	12.03	YES	TTN	titin	0	0	34/66	Medium	NE	NE
NP_056386.1_39	2	1	0.049	7.81	9.19	NO	SENP6	SUMO1/sentrin specific peptidase 6	1	0	47/66	Supportive	NE	NE
NP_056988.2_16	2	1	0.049	7.11	10.04	NO	EIF5B	eukaryotic translation initiation factor 5B	3	2	65/66	Uncertain	2.53	3.56
NP_006531.1_27	2	2	0.090	17.6	37.88	NO	NCOA2	nuclear receptor coactivator 2					NE	NE
NP_004534.1_69	2	2	0.090	17.22	24.35	YES	NEB	nebulin	0	0	5/65	Supportive	NE	NE
NP_940922.1_5	2	2	0.090	8.12	52.83	NO	C12ORF63	chromosome 12 open reading frame 63	2	2	64/64	Medium	NE	NE
NP_005482.1_16	2	2	0.090	7.37	18.35	NO	MAMLD1	mastermind-like domain containing 1	0	0	27/65	Uncertain	NE	NE
XP_497470.1_44	2	2	0.090	7.09	5.97	NO	#N/A	#N/A					NE	NE
XP_294311.1_3	2	2	0.090	6.08	8.55	NO	#N/A	#N/A					NE	NE
XP_380057.1_3	2	2	0.090	5.63	8	NO	#N/A	#N/A					NE	NE
NP_000255.1_47	2	3	0.138	35.59	118.26	NO	PTCH1	patched homolog 1 (Drosophila)	2	2	63/64	Supportive	NE	NE
XP_372194.2_15	2	3	0.138	18.02	79.15	YES	C1ORF170	#N/A					NE	NE
XP_499164.1_20	2	3	0.138	9.04	49.86	NO	#N/A	#N/A					NE	NE
NP_002430.1_24	2	3	0.138	7.77	40.61	NO	MSH3	mutS homolog 3	0	0	5/66	Medium	NE	NE

Supplementary Table 4.2: T1D candidate peptides

Candidates were enriched by at least 2 T1D patients (n=39), and by not more than 3 non-T1D samples (n=248). Peptides reaching global dataset significance by permutation analysis are italicized (FDR = 10%). Epitope spreading is defined as having at least 2 peptides enriched from the same ORF in at least one T1D individual. Protein expression data is from the Protein Atlas database.[171] 0, not expressed; 1, low expression; 2, medium expression; 3, high expression. Confidence refers to confidence in the quality of the antibody used. Specificity refers to the number of tissues with positive expression out of the number of tissues examined. GNF transcript expression[90] ratio is given as whole pancreas value divided by the median of non pancreas tissues ("Pancreas"), and as pancreatic islet value divided by the median of non pancreas tissues ("Islet"). NE, not enriched. Bold candidates were selected for follow up study as full length proteins in a RIA assay.

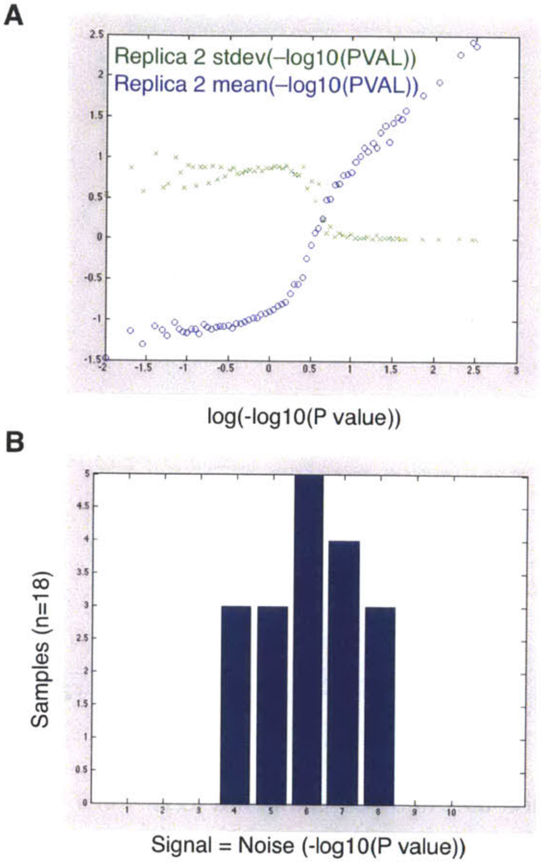
Supplementary Table 4.3

	T1D	NonT1D		Max P value		Epitope			PROTEIN ATLAS				GNF Expression	
ORF	(n=39)	(n=248)	Fisher	T1D	NonT1D	Spreading	Symbol	Gene Name	EXOCRINE	ISLET	Specificity	Confidence	Pancreas	Islet
NP_443724	3	0	0.002	18.55	4.56	NO	ADC	arginine decarboxylase	2	1	46/66	Low	NE	NE
NP_001442	3	0	0.002	7.96	4.04	NO	FOXF1	forkhead box F1	0	0	27/65	Uncertain	NE	NE
NP_683765	3	1	0.008	19.77	5.29	NO	OXER1	oxoeicosanoid (OXE) receptor 1					NE	NE
NP_000618	3	1	0.008	11.44	132.24	NO	LTBP1	latent transforming growth factor beta binding protein	0	0	25/66	Low	NE	NE
NP_689760	2	0	0.018	30.6	3.43	NO	BTNL9	butyrophilin-like 9	1		52/64	Uncertain	NE	NE
NP_060656	2	0	0.018	30.24	4.79	NO	C1ORF112	chromosome 1 open reading frame 112	1	2	44/66	Low	NE	NE
NP_002477	2	0	0.018	28.98	4.76	NO	NCBP1	nuclear cap binding protein subunit 1, 80kDa					NE	NE
NP_061982	2	0	0.018	20.72	3.96	NO	ALG1	asparagine-linked glycosylation 1, beta-1,4-mannosyltr					NE	NE
NP_003410	**2**	**0**	**0.018**	**18.75**	**4.86**	**YES**	**ZNF345**	**zinc finger protein 345**					**NE**	**NE**
NP_848572	**2**	**0**	**0.018**	**18.55**	**3.54**	**NO**	**BANF2**	**barrier to autointegration factor 2**	**3**		51/63	**Uncertain**	**NE**	**NE**
NP_006498	**2**	**0**	**0.018**	**11.86**	**3.17**	**NO**	**REG1B**	**regenerating islet-derived 1 beta**					**1439.63**	**4000.71**
NP_079020	2	0	0.018	11.22	4.43	NO	ALS2CR8	amyotrophic lateral sclerosis 2 (juvenile) chromosome r	3	2	63/65	Uncertain	NE	NE
NP_054762	2	0	0.018	8.51	4.12	NO	CHMP2B	chromatin modifying protein 2B	1	0	21/66	Uncertain	1.08	4.18
NP_078849	2	0	0.018	6.41	4.29	NO	C6ORF211	chromosome 6 open reading frame 211	2	2	57/66	Medium	NE	NE
NP_849154	3	2	0.019	35.85	12.76	NO	MORN4	MORN repeat containing 4					NE	NE
XP_496548	3	2	0.019	21.16	7.69	NO	#N/A	#N/A					NE	NE
XP_496855	3	2	0.019	16.93	8.51	NO	#N/A	#N/A					NE	NE
XP_499266	3	2	0.019	16.9	7.93	NO	#N/A	#N/A					NE	NE
NP_006605	3	2	0.019	6.86	123.79	NO	CHL1	cell adhesion molecule with homology to L1CAM (close	1	1	47/66	Medium	NE	NE
NP_536350	**3**	**3**	**0.035**	**35.7**	**19.08**	**NO**	**GNAS**	**GNAS complex locus**	**0**	**3**	6/66	**Low**	**1.07**	**175.51**
XP_499224	3	3	0.035	12.6	706.24	NO	#N/A	#N/A					NE	NE
XP_496720	2	1	0.049	26.21	6.75	NO	#N/A	#N/A					NE	NE
NP_997336	2	1	0.049	17.46	19.12	NO	#N/A	#N/A					NE	NE
NP_653313	2	1	0.049	15.16	7.12	NO	SLC23A3	solute carrier family 23 (nucleobase transporters), mem	1	1	43/65	Uncertain	NE	NE
NP_115821	2	1	0.049	13.99	5.22	NO	MEGF11	multiple EGF-like-domains 11					NE	NE
XP_496255	2	1	0.049	12.85	67.33	NO	#N/A	#N/A					NE	NE
NP_078791	2	1	0.049	12.31	5.12	NO	WDR25	WD repeat domain 25	2	3	62/64	Uncertain	NE	NE
NP_061831	2	1	0.049	11.34	34.43	NO	C15ORF2	chromosome 15 open reading frame 2					NE	NE
NP_001012410	2	1	0.049	9.38	5.12	NO	SGOL1	shugoshin-like 1 (S	3	1	64/64	Supportive	NE	NE
NP_004938	2	1	0.049	7.02	7.33	NO	DOCK3	dedicator of cytokinesis 3	2	3	57/66	Very low	NE	NE
NP_937797	2	1	0.049	6.93	5.07	NO	TMEM95	transmembrane protein 95					NE	NE
NP_057422	2	1	0.049	6.52	135.85	NO	IPO11	importin 11					NE	NE
XP_373076	2	1	0.049	5.91	107.01	NO	LOC391747	similar to hCG1807616; similar to TBP-associated factor					NE	NE
XP_496294	2	1	0.049	5.17	15.78	NO	#N/A	#N/A					NE	NE
NP_004013	2	2	0.090	78.53	24.75	NO	DMD	dystrophin	0	0	2/66	High	NE	NE
NP_000811	2	2	0.090	50.14	9.82	NO	GAS6	similar to growth arrest-specific 6; growth arrest-specif	0		32/63	Uncertain	NE	NE
NP_060082	2	2	0.090	19.19	13.94	NO	ZCCHC8	zinc finger, CCHC domain containing 8	3	3	65/65	High	NE	NE
NP_005453	2	2	0.090	13.41	14.01	NO	MAGEC1	melanoma antigen family C, 1	0	0	1/65	High	NE	NE
NP_004386	2	2	0.090	8.72	5.6	NO	DBN1	drebrin 1	1	1	24/66	Uncertain	NE	NE
XP_290527	2	2	0.090	8.62	18.67	NO	#N/A	#N/A					NE	NE
NP_005263	2	2	0.090	8.05	10.68	NO	GNAT2	guanine nucleotide binding protein (G protein), alpha tr	2	2	51/65	Uncertain	NE	NE
NP_003893	2	2	0.090	7.49	14.42	NO	FUBP1	far upstream element (FUSE) binding protein 1	3	3	63/63	Supportive	NE	NE
XP_294311	2	2	0.090	6.08	8.55	NO	#N/A	#N/A					NE	NE
XP_380057	2	2	0.090	5.63	8	NO	#N/A	#N/A					NE	NE
XP_380018	2	3	0.138	23.72	59.89	NO	#N/A	#N/A					NE	NE
NP_640339	2	3	0.138	11.25	8.9	NO	TSTD2	chromosome 9 open reading frame 97	3	2	52/66	Low	NE	NE
NP_112220	2	3	0.138	9.95	28.04	NO	SLCO5A1	solute carrier organic anion transporter family, membe	0	1	39/64	Uncertain	NE	NE
NP_116274	2	3	0.138	6.42	11.84	NO	ATG4D	ATG4 autophagy related 4 homolog D (S					NE	NE
NP_057407	2	3	0.138	5.95	345.24	NO	HERC5	hect domain and RLD 5	0	0	12/64	Supportive	NE	NE

Supplementary Table 4.3: T1D candidate ORFs

Candidates were enriched by at least 2 T1D patients (n=39), and by not more than 3 non-T1D samples (n=248). Nomenclature is the same as for Supplementary Table 4.2.

Supplementary Figure 4.1



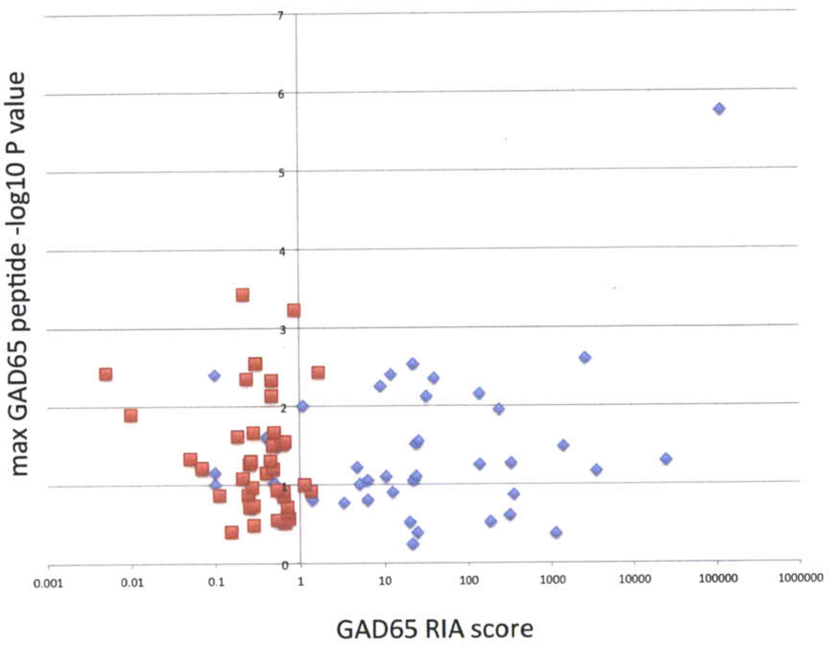
Supplementary Figure 4.1: Dataset reproducibility threshold

P value threshold for reproducibility was established using the data from each sample duplicate pair. Scatter plots of duplicate 1 versus duplicate 2 were used to generate signal to noise analyses.

A. Typical behavior of a duplicate scatterplot. As -log10 P values increase, the average mean (signal) increases while the standard deviation (noise) decreases. The point at which they cross is considered the reproducibility threshold.

B. Histogram plot showing where signal to noise thresholds were achieved for all duplicates in the screen. Based on this analysis, we chose -log10 P value = 5 as the cutoff for reproducibility.

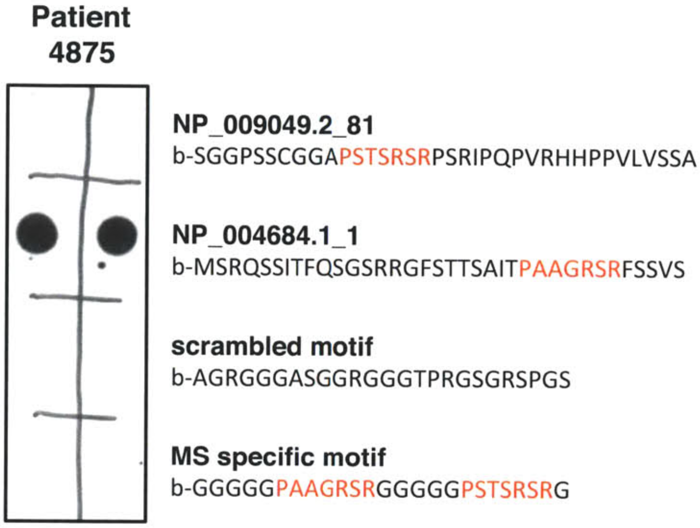
Supplementary Figure 4.2



Supplementary Figure 4.2: False negative PhIP-Seq detection rate of GAD65 autoantibodies

Sensitivity of GAD65 autoantibody detection by PhIP-Seq compared to RIA in T1D patients and healthy controls. PhIP-Seq values correspond to the most enriched peptide from the ORF.

Supplementary Figure 4.3


Patient 4875
NP_009049.2_81
b-SGGPSSCGGAPSTSRSRPSRIPQPVRHHPPVLVSSA
NP_004684.1_1
b-MSRQSSITFQSGSRRGFSTTSAITPAAGRSRFSSVS
scrambled motif
b-AGRGGGASGGRGGGTPRGSGRSPGS
MS specific motif
b-GGGGGPAAGRSRGGGGGPSTSRSRG

Supplementary Figure 4.3: Dot blot confirmation of MS-specific peptide motif

Synthetic, biotinylated peptide (NeoBioSci) dotted onto a streptavidin-soaked nitrocellulose membrane and then probed with CSF from MS patient 4875. In the PhIP-Seq assay, this patient enriched both peptides NP_009049.2_81 and NP_004684.1_1. We interpret the results to mean that the 7 amino acid motif is not sufficient for antibody binding. In addition, steric hindrance likely prevented antibody binding to the NP_009049.2_81 motif.

Supplementary Discussion

Titin autoantibodies have long been investigated in association with myasthenia gravis (MG), since they occur in 20-30% of patients with this autoimmune disease.[172, 173] The main immunogenic region of titin was mapped to a 30 KDa fragment spanning amino acids 7025-7311 of the novex-2 isoform (NP_597681.3).[174] A second, recently discovered, MG-associated immunodominant region was mapped to amino acids 10,319-10,532.[175] Within our dataset, three individuals (one healthy, one MS patient, and one BC patient), demonstrated reactivity against a single peptide from the first region (7,193-7,228), and one BC patient from our study had antibodies targeting a peptide within the second region (10,441-10,476). It would be interesting to determine whether these peptides are in fact the minimal epitopes of the MG-associated titin antibodies. Strikingly, one titin peptide (8,179-8,214), which is not derived from either MG-associated region, was enriched by 85 individuals, making it the third most commonly enriched peptide by healthy controls (Figure 4.1A). To our knowledge, autoreactivity toward this peptide has not been previously described, but due to its prevalence in healthy controls, is unlikely to have pathological consequences.