Application of a synthetic human proteome to autoantigen discovery through PhIP-Seq

Published in final edited form as:
Nat Biotechnol. ; 29(6): 535–541. doi:10.1038/nbt.1856.

Application of a synthetic human proteome to autoantigen discovery through PhIP-Seq

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,*], and **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

[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

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) analysis of 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

[*]Correspondence to: nsolimini@rics.bwh.harvard.edu, selledge@genetics.med.harvard.edu.
[4]Present address: Biogen Idec, Cambridge, Massachusetts, USA

AUTHOR CONTRIBUTIONS

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. The manuscript was prepared by H.B.L. and edited by N.L.S. and S.J.E.

CONFLICT OF INTEREST STATEMENT

The authors declare no competing financial interests.

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.

Keywords

Synthetic biology; proteomics; phage display; humoral autoimmunity; paraneoplastic neurological disorder; protein-protein interactions

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 autoimmunity1. 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 boundary of 6%)2, and the highly skewed representation of differentially expressed cDNAs. Nevertheless, expression cloning has led to the discovery of many important autoantigens3-5. Strides have been made to improve peptide display systems6, 7, 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.

RESULTS

Construction and Characterization of T7-Pep

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 (Fig. 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.

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; Fig. 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 1), clone sequencing (Fig. 1c), and FLAG expression (Supplementary Table 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 1).

After combining 5×10^8 phage from each subpool and amplifying the final library, Illumina sequencing was performed at a median depth of 45-fold coverage (Fig. 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 Fig. 1)8. 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 1).

We next optimized a phage immunoprecipitation protocol for detecting antibody-peptide interactions within complex mixtures (Supplementary Fig. 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 (Fig. 1b).

PhIP-Seq analysis of spinal fluid from a PND patient with NOVA autoantibodies

Cancers often elicit cellular and humoral immune responses against tumor antigens which may limit disease progression9. 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 success3, 10.

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 autoantibodies11. The NOVA autoantigen (neurooncological 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 (Fig. 2a), likely due to weak nonspecific interactions with the beads.

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 data12) and found that this distribution family fits the data well at various input abundances (Fig. 2b). We calculated the GP parameter values for each input abundance level13 and regressed these parameters to form our null model for the calculation of enrichment significance (p-values) of each clone (Fig. 2c and Supplementary Methods). Comparing the two PhIP-Seq replicates revealed that the most significantly enriched clones were the same in both replicas (Fig. 2d), highlighting the assay`s reproducibility. This contrasts dramatically with a comparison of clones enriched by two different patients (Supplementary Fig. 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 (Fig. 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 2).

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 (Fig. 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, 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 software14 to characterize this motif, which represents the likely epitope recognized by Patient A's autoantibodies (Fig. 3b).

PhIP-Seq analysis of spinal fluid from 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 2). Both TGIF2LX and CTAG2 were confirmed by immunoblotting (Fig. 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 (Fig. 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 responses15. Several reports have described both humoral and cellular immune responses targeted against CTAG216, 17. TGIF2LX is also testis restricted18, 19 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 Fig. 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 Fig. 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 patients20, 21 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 (Fig. 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 (Fig. 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 2). Both candidate autoantigens were confirmed by immunoblotting lysates from TRIM9- or TRIM67-overexpressing cells (Fig. 3f). TRIM67 is expressed in some normal tissues (including skin) and is often highly expressed in melanoma19. TRIM9 has recently emerged as a brain-specific E3 ubiquitin ligase and has been implicated in neurodegenerative disease processes22. 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 Fig. 6). TRIM9 and TRIM67 autoreactivity was not detected in the CSF of three patients with non-PND CNS autoimmunity (Supplementary Fig. 4).

In total, 16 of the candidate autoantigens in Table 2 were available to us as full-length Gateway Entry clones from the ORFeome collection23. 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 PhIPSeq dataset (Fig. 3g; Supplementary Fig. 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.

PhIP-Seq can be used to identify protein-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 Fig. 8), PhIP-Seq with GST-RPA2 robustly identified the N-terminal peptide from the known interactor SMARCAL1 ($P < 10^{-14}$, Fig. 4a), among others (Supplementary Table 3). The enriched SMARCAL1 peptide contains a previously identified motif known to bind RPA224, 25. 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 (Fig. 4b). 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.

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 folding26, 27. 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 products28, 29, misfolded proteins or protein aggregates30, 31. 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-restricted19 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 states32. 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.

METHODS

Detailed methods are available in the supplementary online materials at http://www.nature.com/naturebiotechnology.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported in part by grants from the DOD (W81XWH-10-1-0994 and W81XWH-04-1-0197) to S.J.E., and in part by the NIH (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.

REFERENCES

1. Graham AL, et al. Fitness correlates of heritable variation in antibody responsiveness in a wild mammal. Science. 2010; 330:662–665. [PubMed: 21030656]
2. Faix PH, et al. Phage display of cDNA libraries: enrichment of cDNA expression using open reading frame selection. BioTechniques. 2004; 36:1018–1022. 1024, 1026–1019. [PubMed: 15211753]
3. Albert ML, Darnell RB. Paraneoplastic neurological degenerations: keys to tumour immunity. Nat Rev Cancer. 2004; 4:36–44. [PubMed: 14708025]
4. Wang X, et al. Autoantibody signatures in prostate cancer. N Engl J Med. 2005; 353:1224–1235. [PubMed: 16177248]
5. Anderson KS, et al. A Protein Microarray Signature of Autoantibody Biomarkers for the Early Detection of Breast Cancer. J Proteome Res. 2010
6. Zacchi P, Sblattero D, Florian F, Marzari R, Bradbury ARM. Selecting open reading frames from DNA. Genome Research. 2003; 13:980–990. [PubMed: 12727911]

7. Kim Y, et al. Identification of Hnrph3 as an autoantigen for acute anterior uveitis. Clin Immunol. 2011; 138:60–66. [PubMed: 20943442]
8. Hughes JB, Hellmann JJ, Ricketts TH, Bohannan BJ. Counting the uncountable: statistical approaches to estimating microbial diversity. Appl Environ Microbiol. 2001; 67:4399–4406. [PubMed: 11571135]
9. Swann JB, Smyth MJ. Immune surveillance of tumors. J. Clin. Invest. 2007; 117:1137–1146. [PubMed: 17476343]
10. Darnell RB, Posner JB. Paraneoplastic syndromes involving the nervous system. N Engl J Med. 2003; 349:1543–1554. [PubMed: 14561798]
11. Musunuru K, Kesari S. Paraneoplastic opsoclonus-myoclonus ataxia associated with non-small-cell lung carcinoma. J Neurooncol. 2008; 90:213–216. [PubMed: 18618225]
12. Srivastava S, Chen L. A two-parameter generalized Poisson model to improve the analysis of RNA-seq data. Nucleic Acids Research. 2010
13. Consul P, Shoukri M. Maximum likelihood estimation for the generalized poisson distribution. Communications in Statistics - Theory and Methods. 1984; 13:1533–1547.
14. Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol. 1994; 2:28–36. [PubMed: 7584402]
15. Almeida LG, et al. CTdatabase: a knowledge-base of high-throughput and curated data on cancer-testis antigens. Nucleic Acids Research. 2009; 37:D816–D819. [PubMed: 18838390]
16. 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. 2000; 165:7253–7261. [PubMed: 11120859]
17. Chen YT, 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. 1998; 95:6919–6923. [PubMed: 9618514]
18. Blanco-Arias P, Sargent CA, Affara NA. The human-specific Yp11.2/Xq21.3 homology block encodes a potentially functional testis-specific TGIF-like retroposon. Mamm Genome. 2002; 13:463–468. [PubMed: 12226713]
19. Berglund L, et al. A Genecentric Human Protein Atlas for Expression Profiles Based on Antibodies. Molecular & Cellular Proteomics. 2008; 7:2019–2027. [PubMed: 18669619]
20. Li L, Hagopian WA, Brashear HR, Daniels T, Lernmark A. Identification of autoantibody epitopes of glutamic acid decarboxylase in stiff-man syndrome patients. J Immunol. 1994; 152:930–934. [PubMed: 7506741]
21. Schwartz HL, et al. High-resolution autoreactive epitope mapping and structural modeling of the 65 kDa form of human glutamic acid decarboxylase. Journal of Molecular Biology. 1999; 287:983–999. [PubMed: 10222205]
22. 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. 2010; 38:210–218. [PubMed: 20085810]
23. Rual J-F, et al. Towards a proteome-scale map of the human protein-protein interaction network. Nature. 2005; 437:1173–1178. [PubMed: 16189514]
24. Ciccia A, et al. The SIOD disorder protein SMARCAL1 is an RPA-interacting protein involved in replication fork restart. Genes Dev. 2009; 23:2415–2425. [PubMed: 19793862]
25. Mer G, et al. Structural basis for the recognition of DNA repair proteins UNG2, XPA, and RAD52 by replication factor RPA. Cell. 2000; 103:449–456. [PubMed: 11081631]
26. Barlow DJ, Edwards MS, Thornton JM. Continuous and discontinuous protein antigenic determinants. Nature. 1986; 322:747–748. [PubMed: 2427953]
27. Jin L, Fendly BM, Wells JA. High resolution functional analysis of antibody-antigen interactions. J Mol Biol. 1992; 226:851–865. [PubMed: 1380563]
28. Miyazaki K, et al. Analysis of in vivo role of alpha-fodrin autoantigen in primary Sjogren's syndrome. Am J Pathol. 2005; 167:1051–1059. [PubMed: 16192640]
29. 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. 2005; 142:148–154. [PubMed: 16178869]

30. Robbins DC, Cooper SM, Fineberg SE, Mead PM. Antibodies to covalent aggregates of insulin in blood of insulin-using diabetic patients. Diabetes. 1987; 36:838–841. [PubMed: 2438179]

31. Papachroni KK, et al. Autoantibodies to alpha-synuclein in inherited Parkinson's disease. J Neurochem. 2007; 101:749–756. [PubMed: 17448146]

32. Dalakas MC, Fujii M, Li M, McElroy B. The clinical spectrum of anti-GAD antibody-positive patients with stiff-person syndrome. Neurology. 2000; 55:1531–1535. [PubMed: 11094109]

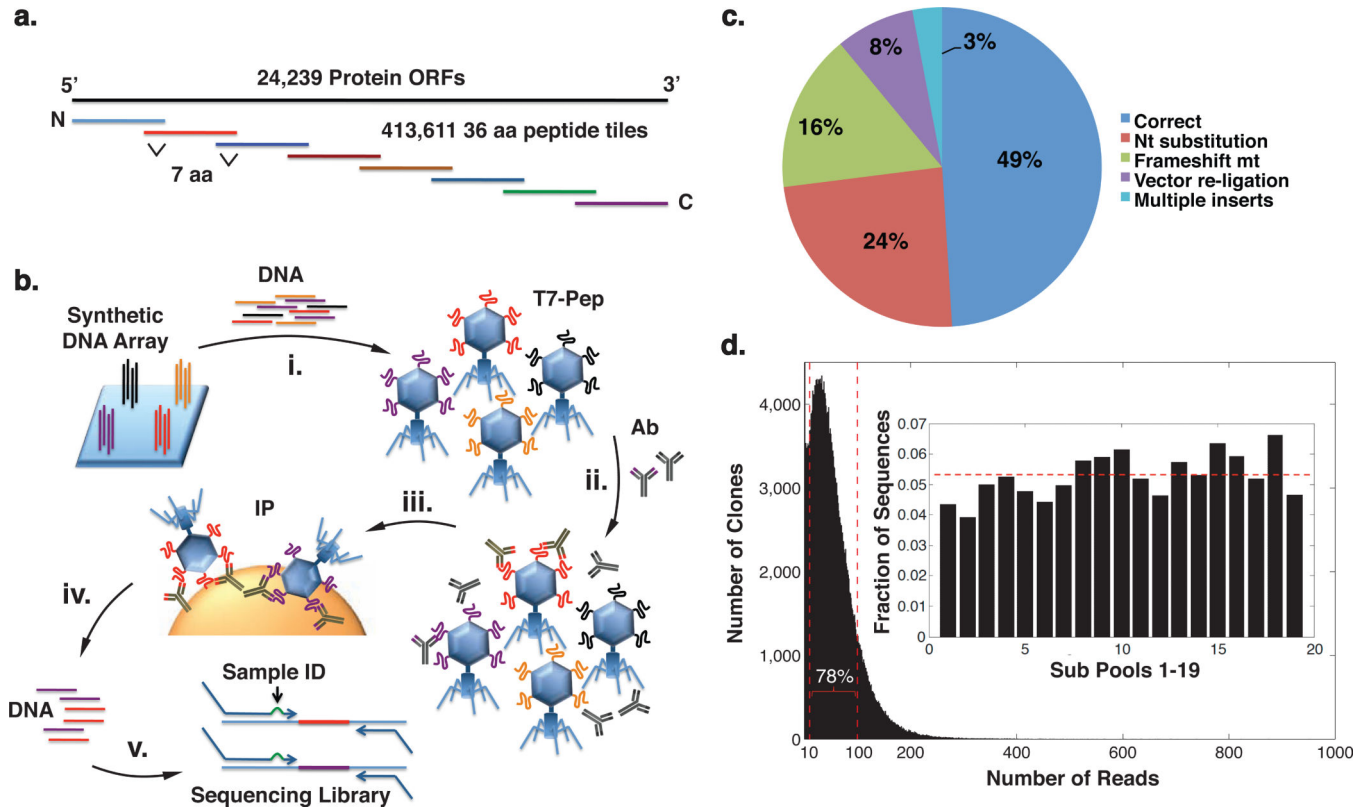


Figure 1. Construction and characterization of T7-Pep and the PhIP-Seq methodology

(**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).

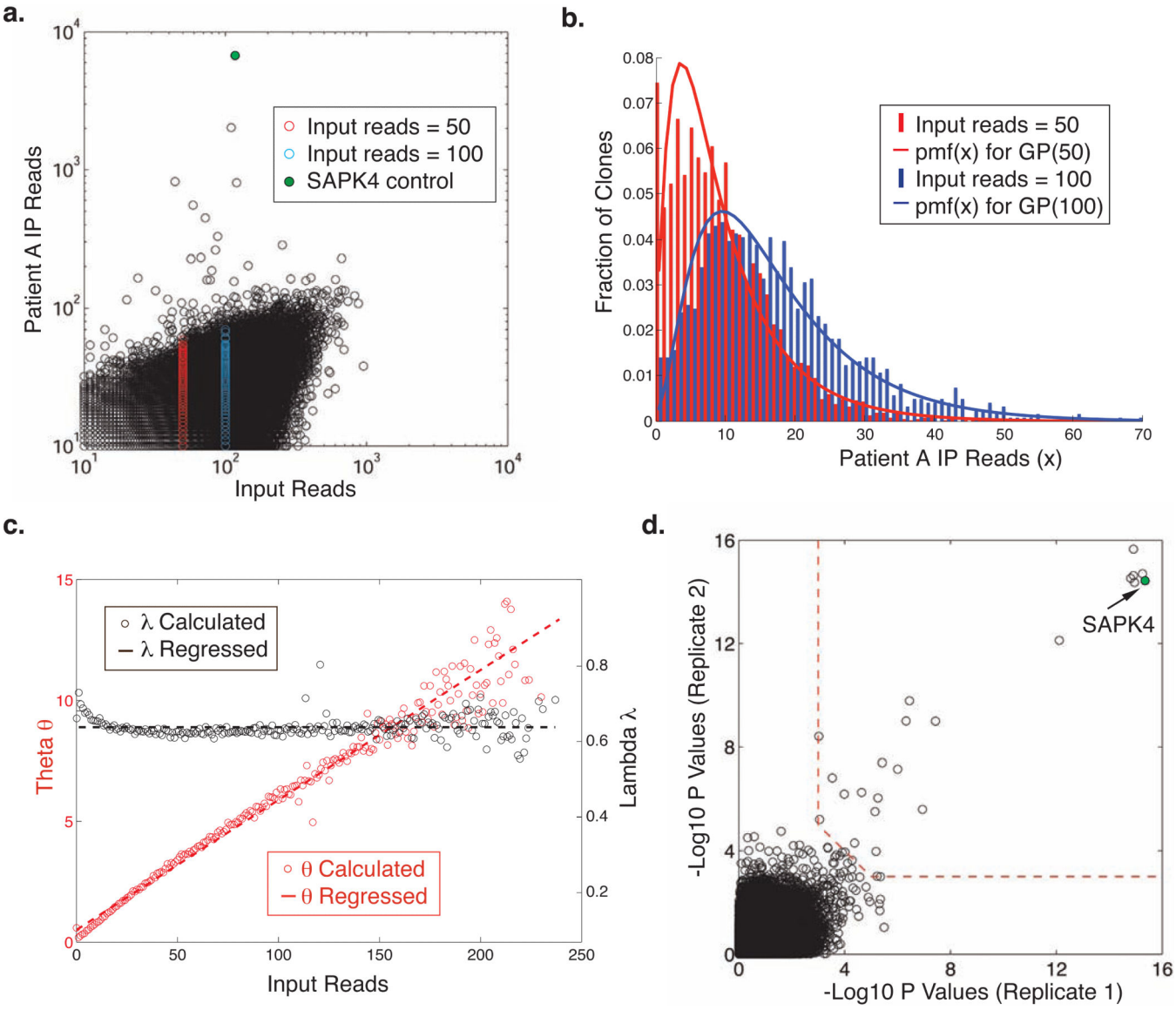


Figure 2. Statistical analysis of PhIP-Seq data

(**a**) Scatter plot comparing sequencing reads from T7-Pep input library and from Patient A immunoprecipitated (IP) phage (Pearson coefficient = 0.435; $P \sim= 0$). 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 al13. 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.

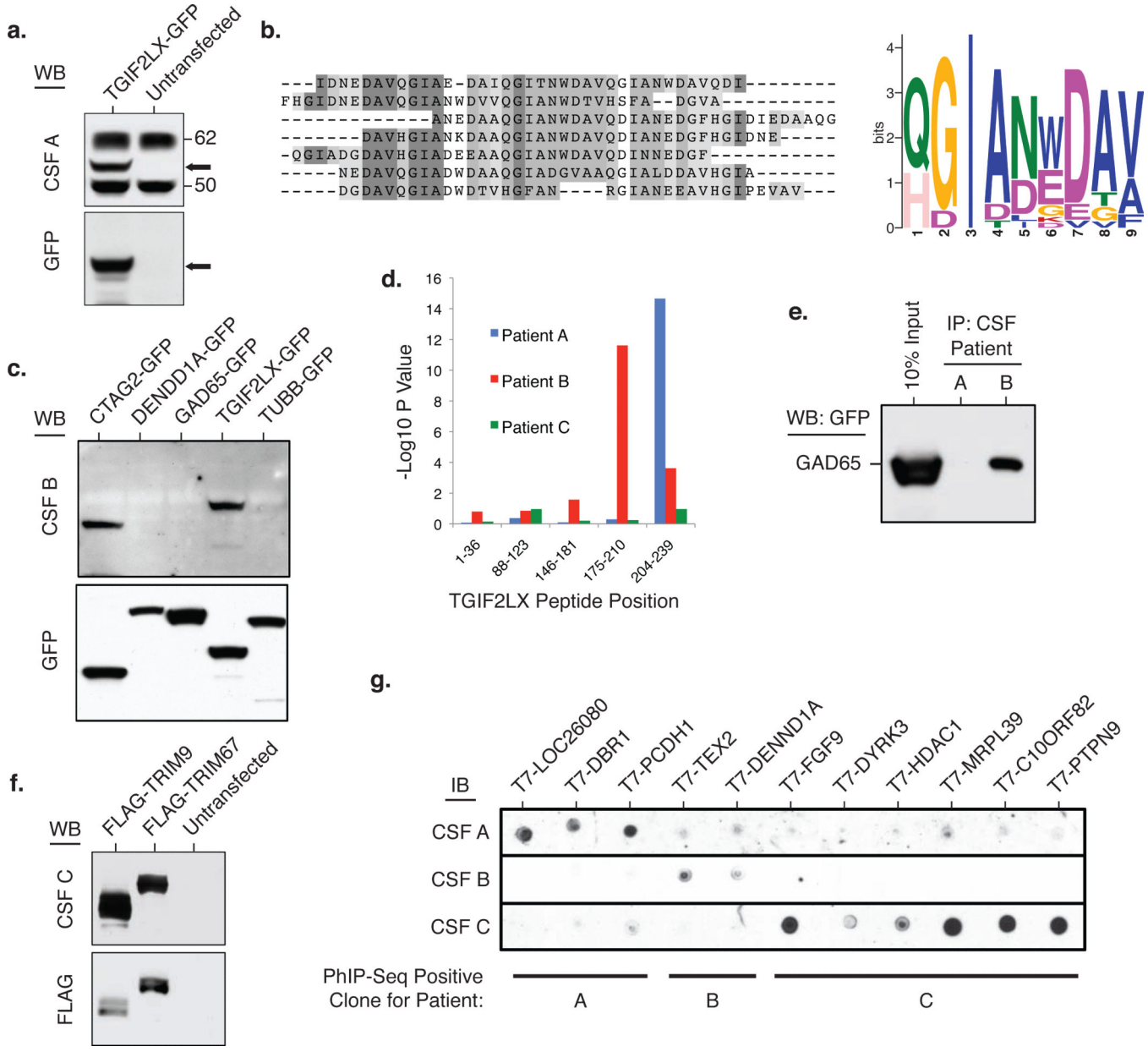
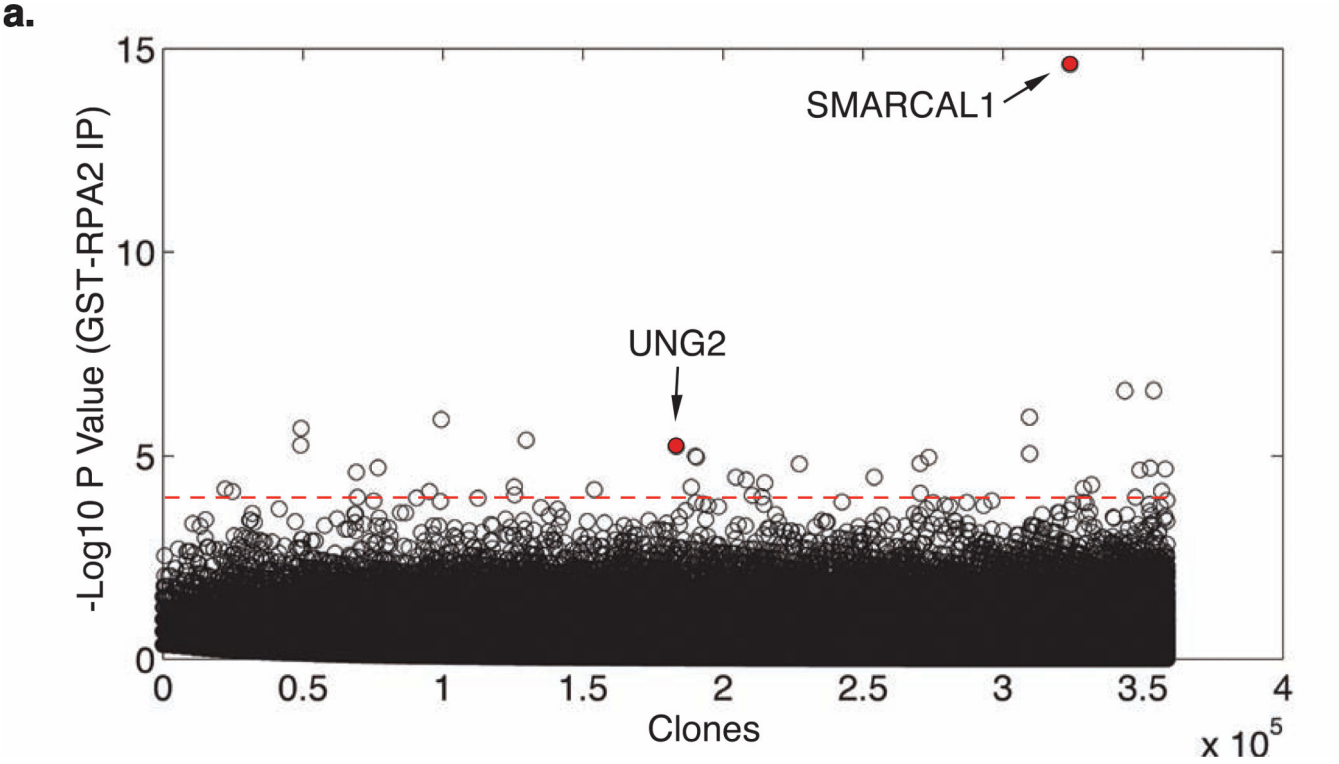


Figure 3. Validation of full-length PhIP-Seq candidates

(**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 (c) 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.

a.



b.

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

Figure 4. PhIP-Seq can identify protein-protein interactions

GST-RPA2 was used to precipitate phage from the T7-Pep library on magnetic glutathione beads. **(a)** –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. **(b)** Dependence of peptide enrichment on integrity

of RPA2 binding motif. Phage peptides containing the RPA2-binding motif (underlined) are shown next to their –Log10 p-value of enrichment.

Table 1

Comparison between the T7-Pep + PhIP-Seq approach and current proteomic methods for autoantigen discovery.

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	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 2
PhIP-Seq analysis on CSF from patients with PND and secondary validation

A previously known autoantigen is shown in green. Autoantigens confirmed by any secondary assay are shown in bold. Confirmation with the full-length protein 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: NT = not tested; WB = western blot of full-length proteins; IP = immunoprecipitation of full-length proteins followed by western blotting for the GFP tag. RIA = radioimmunoassay; DB = dot blot. Validation assay is followed by "+" or "–" depending on whether the results were positive or negative.

Patient Info	-Log10 P value	Protein	# Peptides	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+**