nature biotechnology

Autoantigen discovery with a synthetic human peptidome

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

Immune responses targeting self-proteins (autoantigens) can lead to a variety of autoimmune diseases. Identification of these antigens is important for both diagnostic and therapeutic reasons. However, current approaches to characterize autoantigens have, in most cases, met only with limited success. Here we present a synthetic representation of the complete human proteome, the T7 peptidome phage display library (T7-Pep), and demonstrate its application to autoantigen discovery. T7-Pep is composed of >413,000 36-residue, overlapping peptides that cover all open reading frames in the human genome, and can be analyzed using high-throughput DNA sequencing. We developed a phage immunoprecipitation sequencing (PhIP-Seq) methodology to identify known and previously unreported autoantibodies contained in the spinal fluid of three individuals with paraneoplastic neurological syndromes. We also show how T7-Pep can be used more generally to identify peptide-protein interactions, suggesting the broader utility of our approach for proteomic research.

© 2011 Nature America, Inc. All rights reserved.

Vertebrate immune systems have evolved sophisticated genetic mechanisms to generate T-cell receptor and 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[1]. In humans, loss of tolerance to self-antigens results in a number of diseases including type I diabetes, multiple sclerosis, lupus and rheumatoid arthritis. Knowledge of the self-antigens involved in autoimmune processes is important not only for understanding disease etiology but also for developing accurate diagnostic tests. In addition, physicians may someday use antigen-specific therapies to destroy or disable autoreactive immune cells.

Traditional approaches for identifying autoantibody targets have relied largely on the expression of fragmented cDNA libraries as polypeptides fused to the capsid proteins of bacteriophage. Notable technical limitations of this method include the small fraction of clones expressing coding sequences in the correct reading frame (with a lower bound of 6%)[2], and the highly skewed representation of differentially expressed cDNAs. Nevertheless, this approach has led to the discovery of many important autoantigens[3–5]. Strides have been made to improve peptide display systems[6,7], but there remains an unmet need for better display libraries and methods to analyze binding interactions.

We have therefore constructed a synthetic representation of the complete human proteome, which we engineered for display as peptides on the surface of T7 phage. This library (T7-Pep) was extensively characterized and found to be both faithful to its *in silico* design and uniform in its representation. Autoantibodies within a complex mixture can thus be used to enrich for phage-displaying autoantigenic peptides. The population of enriched phage can then be analyzed by high-throughput DNA sequencing. This approach provides several advantages over traditional methods, including comprehensive and unbiased proteomic analysis, and peptide enrichment quantification using a streamlined, multiplexed protocol requiring just one round of enrichment. We have applied phage immunoprecipitation sequencing (PhIP-Seq) to interrogate the autoantibody repertoire in the spinal fluid of patients with neurological autoimmunity and identified both known and novel autoantigens. Furthermore, to demonstrate how this approach can be used more generally to identify peptide-protein interactions, we used it to recover proteins known to interact with replication protein A2 (RPA2).

RESULTS

Construction and characterization of the T7-Pep library

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 ORFs; 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 *Escherichia coli*, the T7 host, with more abundant synonymous codons. We then divided each ORF into overlapping sequences of 108 nucleotides encoding 36-amino acid peptides (**Fig. 1a**). Consecutive peptides overlapped by seven residues, the estimated size of a linear epitope. Finally, the

[1]Harvard-MIT Division of Health Sciences and Technology, Cambridge, Massachusetts, USA. [2]Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [3]Department of Genetics, Harvard University Medical School, and Division of Genetics, Howard Hughes Medical Institute, Brigham and Women's Hospital, Boston, Massachusetts, USA. [4]Department of Mathematics, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [5]Department of Genetics, Harvard University Medical School, Boston, Massachusetts, USA. [6]Division of Neuro-Oncology, Department of Neurosciences, University of California, San Diego, Moores Cancer Center, La Jolla, California, USA. [7]Agilent Technologies, Genomics, Santa Clara, California, USA. [8]Present address: Biogen Idec, Cambridge, Massachusetts, USA. Correspondence should be addressed to N.L.S. (nsolimini@rics.bwh.harvard.edu) or S.J.E. (selledge@genetics.med.harvard.edu).

Received 15 November 2010; accepted 28 March 2011; published online 22 May 2011; doi:10.1038/nbt.1856

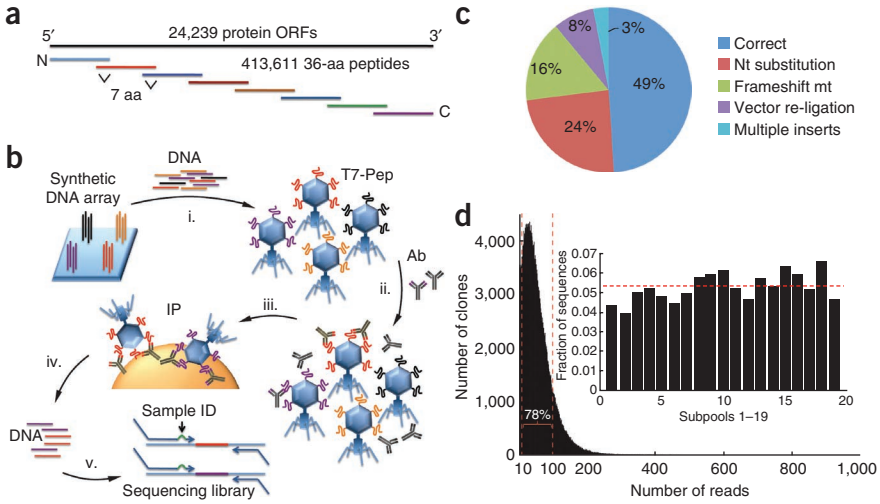


Figure 1 Construction and characterization of T7-Pep and the PhIP-Seq methodology. (**a**) The T7-Pep library is made from 413,611 DNA sequences encoding 36-amino acid peptide that span 24,239 unique ORFs from build 35.1 of the human genome. Each peptide overlaps its neighbors by seven amino acids on each side. (**b**) The DNA sequences from **a** are printed as 140-mer oligos on releasable DNA microarrays. (**i**) After oligo release, the DNA is 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 beads coated with Protein A and G. (**iv**). DNA from the immunoprecipitated phage is recovered. (**v**). Library inserts are PCR-amplified with sequencing adapters. A single nucleotide change (arrow) is introduced for multiplex analysis. (**c**) Results of plaque sequencing of 71 phage from T7-Pep Pool 1 and T7-CPep Pool 1. (**d**) Histogram showing results from Illumina sequencing of the T7-Pep library. Seventy-eight percent 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 the T7-Pep library (inset) compared to expected (horizontal red line). aa, amino acid; nt, nucleotide; mt, mutation.

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 (**Fig. 1b**, **i** and Online 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, for use in future studies.

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 then amplifying and Illumina sequencing the final library at a median depth of 45-fold coverage (**Fig. 1d**), we detected 91.2% of the expected clones. Chao1 analysis[8] was done to estimate the actual library complexity (assuming infinite sampling), which predicted that >91.8% of the library was represented (**Supplementary Fig. 1**). In addition, the T7-Pep library is highly uniform, with 78% of the library members within tenfold 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**).

© 2011 Nature America, Inc. All rights reserved.

Table 1 Comparison between T7-Pep + PhIP-Seq 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	Mostly full-length proteins	36-amino-acid overlapping peptides
Rounds of selection	Requires multiple selection rounds, which favor more abundant and faster-growing clones[33]	No selection	Single selection, minimizing 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	Possible for antigens of known crystal structure
Epitope mapping	Difficult	Not possible	Resolution of 36 amino acids
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

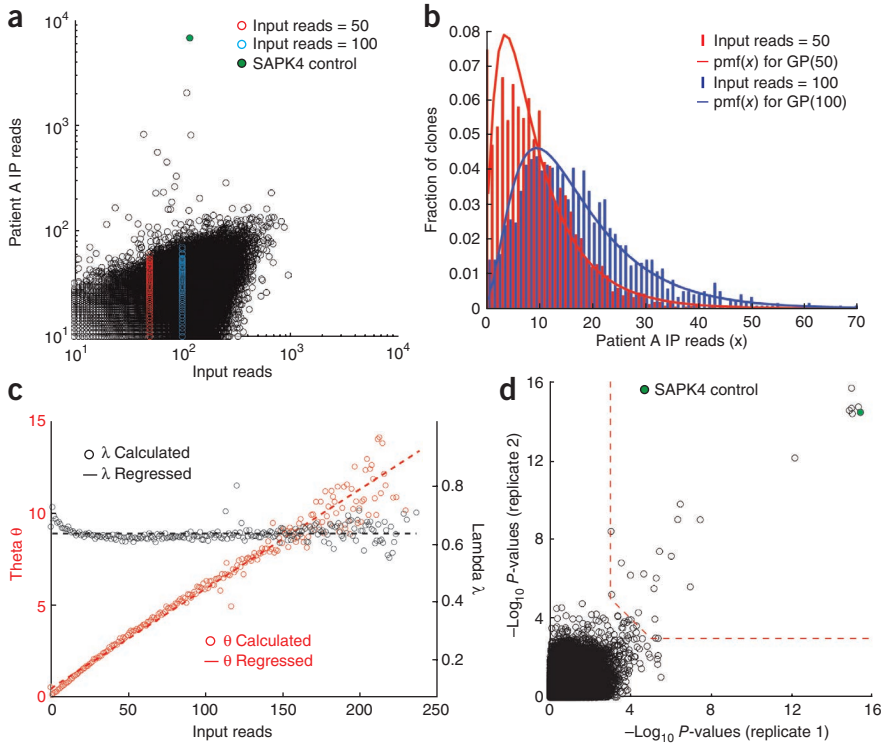


Figure 2 Statistical analysis of PhIP-Seq data. (**a**) Comparison of sequencing reads from T7-Pep input library and from patient A immunoprecipitated (IP) phage (Pearson coefficient = 0.435; $P \cong 0$). Highlighted are all clones with an input abundance of 50 reads (red), and all clones with an input abundance of 100 reads (blue). The target of the SAPK4 control antibody is in green. (**b**) Histogram of sequencing reads from the data highlighted in **a** with corresponding colors. The curves are fit with a generalized Poisson (GP) distribution. Pmf, probability mass function of the corresponding GP distribution. *x*, number of immunoprecipitated clones' sequencing reads. (**c**) Poisson distribution parameters lambda and theta for each input abundance, calculated as previously described[13]. Lambda is regressed to its average value (black dashed line), and theta is linearly regressed (red dashed curve). (**d**) Comparison of clone enrichment significances (as $-\log_{10}$ *P*-value) from two independent PhIP-Seq experiments using cerebrospinal fluid from patient A. Red dashed line shows the cutoff for considering a clone to be significantly enriched, and the SAPK4 control antibody target is in green.

We next optimized a phage immunoprecipitation protocol for detecting antibody-peptide interactions within complex mixtures by systematically varying parameters such as the relative concentrations of antibody and phage, number of washes and type of beads. (**Supplementary Fig. 2**). Our streamlined procedure consists of incubating T7-Pep with unmodified patient cerebrospinal spinal fluid, immunoprecipitating phage-antibody complexes on beads with Protein A and Protein G, washing away unbound phage and then subjecting the enriched phage population to PCR amplification and deep-sequencing DNA analysis (**Fig. 1b**).

Analysis of a PND patient with NOVA autoantibodies

Tumor antigens often elicit cellular and humoral immune responses, which may limit disease progression[9]. In rare cases, such immune responses recognize central nervous system (CNS) antigens, triggering a devastating autoimmune process called paraneoplastic neurological disorder (PND). Clinical symptoms of PND are heterogeneous and often 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[3,10].

To assess the performance of PhIP-Seq for autoantigen discovery, we examined a sample of cerebrospinal fluid from a 63-year-old female (patient A) with non-small cell lung cancer (NSCLC) who had a PND syndrome and was previously found to have NOVA (neuro-oncological ventral antigen) autoantibodies[11]. The NOVA autoantigen is commonly targeted in PND triggered by lung or gynecological cancers, and results in ataxia with or without opsoclonus or myoclonus. A final concentration of 2 μg/ml of patient cerebrospinal fluid antibody was spiked with 2 ng/ml of a commercial antibody specific to SAPK4, which was used as a positive control to monitor enrichment of the SAPK4 peptide library member. 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 owing to weak nonspecific interactions with the beads.

We used a two-parameter, generalized Poisson model to approximate the distribution of the abundance of immunoprecipitated clones, as recently demonstrated for RNA-Seq data[12]. Because the sample preparation and sequencing steps introduce similar biases, this distribution family fits the data quite well (**Fig. 2b**). We calculated the generalized Poisson parameter values for each input abundance level[13] and regressed these parameters to form our null model for the calculation of enrichment significance (*P*-values) of each clone (**Fig. 2c**). Comparing the two PhIP-Seq replicates for patient A revealed that the most significantly enriched clones were the same in both replicas (**Fig. 2d**), highlighting the reproducibility of the assay. This contrasts dramatically with a comparison of two different patients, as the enrichment of peptides is highly patient-specific (**Supplementary Fig. 3**). Performing a negative-control PhIP-Seq experiment identified phage capable of binding to the slurry of Protein A and Protein G magnetic beads in the absence of patient antibodies. We thus defined patient A–positive clones as those clones with a reproducible $-\log_{10}$ *P*-value greater than a cutoff (**Fig. 2d**, dashed red line) but not significantly enriched on beads alone ($P < 10^{-3}$). Patient A–positive clones 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 predicted autoantigens (TGIF2LX, DBR1 and PCDH1) by expressing full-length protein in 293T cells and immunoblotting with patient cerebrospinal fluid. 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**). Although full-length DBR1 and PCDH1 were expressed well in 293T cells (data not shown), they were not detected by cerebrospinal fluid antibodies. We observed two bands in the untransfected lysate migrating at ~50 and 62 kDa, representing endogenously expressed proteins that correspond either to untested candidates or to false negatives of the PhIP-Seq assay.

© 2011 Nature America, Inc. All rights reserved.

Table 2 PhIP-Seq analysis on spinal fluid from patients with PND and secondary validation

Patient information	$-\text{Log}_{10}$ *P*-value[a]	Protein[b]	No. peptides	Validation
A. Female, 63 years old, with NSCLC as well as classic cerebellar syndrome. Cerebrospinal fluid positive for 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. Female, 59 years old, with NSCLC as well as 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. Female, 59 years old, with ataxia, dysarthria, horizontal gaze palsy. Paraneoplastic antibody panel is negative. However, cerebrospinal fluid stained brain and cerebellar immunohistochemistry 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+**

[a]Average of replicate $-\log_{10}$ *P*-values. If multiple peptides from the same ORF are enriched, the average $-\log_{10}$ *P*-value of the most significantly enriched peptide is shown. [b]Previously validated autoantigens are shown in italics. Autoantigens confirmed by any secondary assay are shown in bold. Confirmation with the full-length protein is indicated by underlining. 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 positive; –, validation assay negative.

© 2011 Nature America, Inc. All rights reserved.

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[14] to characterize this motif, which represents the likely epitope recognized by patient A's autoantibodies (**Fig. 3b**).

Analysis of two PND patients with uncharacterized autoantibodies

Having established that PhIP-Seq could reliably identify known and previously unreported autoantigens, we examined cerebrospinal fluid from two additional patients who had suggestive PND symptoms but tested negative for a panel of commercially available PND autoantigens. Patient B was a 59-year-old female with NSCLC, who had 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**). We were surprised to find that patient B, like patient A, was also strongly autoreactive against TGIF2LX. The enriched peptide was distinct from but its sequence overlapped that of the TGIF2LX peptide enriched in patient A's cerebrospinal fluid (**Fig. 3d**).

CTAG2 is a member of a family of cancer/testis antigen (CTAG) proteins that are normally restricted to germ cells of the developing gonads and/or the adult testis but are frequently expressed in cancers and often elicit antitumor immune responses[15–17]. TGIF2LX, like CTAG2, is normally restricted to testis[18,19] but can also be expressed in cancer cells in either sex, suggesting that it may be a new CTAG family member. As a negative control, we found TGIF2LX reactivity to be absent in the cerebrospinal fluid of three patients with non-PND CNS autoimmunity and oligoclonal immunoglobulin bands (**Supplementary Fig. 4**). Having confirmed TGIF2LX autoreactivity in two individuals with NSCLC, we wondered whether it could be a biomarker for this disease. However, the serum of 15 other 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 neither is likely to explain the neurological syndrome experienced by patient B. GAD65, however, is the rate-limiting enzyme in the synthesis of the inhibitory neurotransmitter GABA and is also a well-characterized autoantigen targeted in the autoimmune disorder stiff person syndrome. Two nonoverlapping GAD65 peptides derived from the domain known to be targeted by pathogenic autoantibodies in individuals with this syndrome[20,21] were enriched in patient B's cerebrospinal fluid. A commercial radioimmunoassay confirmed the presence of high-titer GAD65 autoantibodies (5.12 nmol/L; >250-fold above the reference range). Surprisingly, however, direct immunoblotting with cerebrospinal fluid from patient B

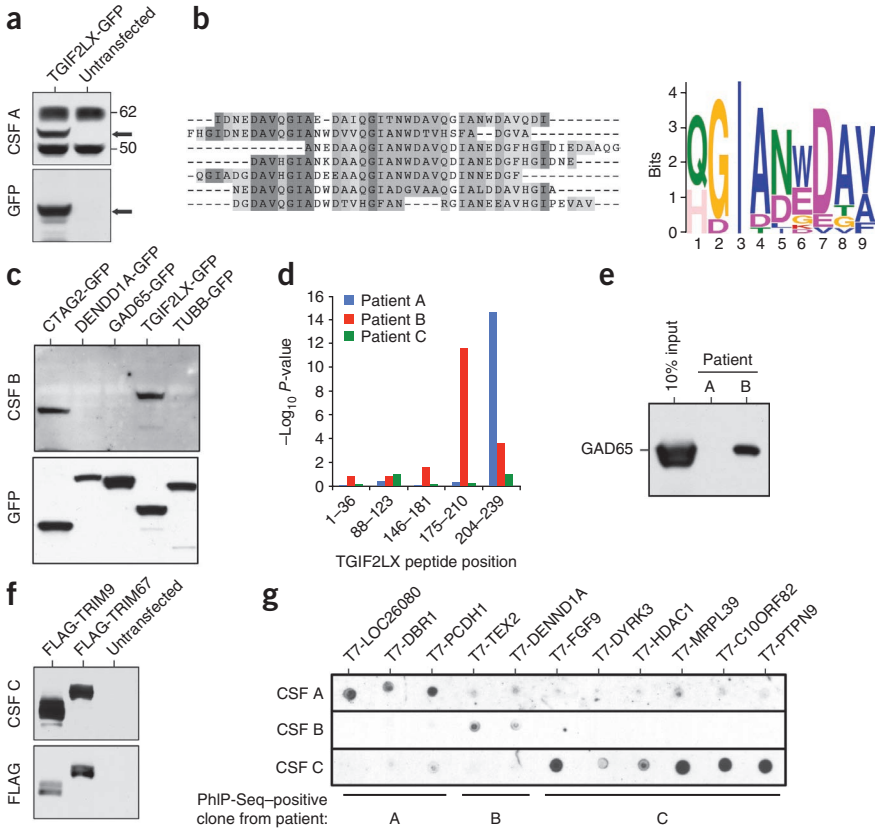


Figure 3 Validation of PhIP-Seq candidates. (**a**) Western blot with cerebrospinal fluid (CSF) from patient A, staining for full-length TGIF2LX-GFP expressed in 293T cells by transient transfection. Bands corresponding to TGIF2LX-GFP are denoted by an arrow. (For full-length blots, see **Supplementary Fig. 9**.) (**b**) ClustalW alignment of the seven significantly enriched hypothetical protein LOC26080 peptides, and the nine-element MEME-generated recognition motif. (**c**) Western blot with cerebrospinal fluid from patient B, staining for indicated full-length proteins expressed in 293T cells by transient transfection. (**d**) Bar graph of $-\log_{10}$ *P*-values of enrichment for the indicated TGIF2LX peptides in the three patients. (**e**) Immunoprecipitation (IP) of the GAD65-GFP from **c** by CSF from patient B (but not patient A). (**f**) Western blot with cerebrospinal fluid from patient C, staining for indicated full-length proteins 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 cerebrospinal fluid.

did not show 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 patient B's cerebrospinal fluid confirmed this hypothesis (**Fig. 3e**).

Patient C, a 59-year-old female with PND that arose after she developed melanoma, had unusual symptoms that included horizontal gaze palsy. PhIP-Seq analysis of patient C's cerebrospinal fluid yielded five significantly ($P < 10^{-4}$ to 10^{-15}) 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 melanoma[19]. TRIM9 has recently emerged as a brain-specific E3 ubiquitin ligase and has been implicated in neurodegenerative disease processes[22]. Based on their high degree of homology, our data suggest the possibility that immune responses to the cancer, targeting TRIM67, might have spread to or cross-reacted with TRIM9 in the CNS (**Supplementary Fig. 6**). Neither TRIM9 nor TRIM67 autoreactivity was detected in the cerebrospinal fluid 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 collection[23]. Of these, ten were not confirmed by immunoblotting or immunoprecipitation of the full-length protein. We wondered whether this indicated 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 made nine of these ten candidate T7 clones, plus two additional high-confidence T7 clones for which no full-length ORF was available. 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 data set (**Fig. 3g** and **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 when displayed on the T7 coat.

PhIP-Seq can identify peptide-protein interactions

The utility of the T7-Pep library is not limited to autoantigen identification. To explore a more general application, we have used the library in an *in vitro*, peptide-protein, 'two-hybrid' interaction experiment with GST-RPA2 (replication protein A2) as bait for the T7-Pep library. We again used 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**), GST-RPA2 robustly identified the correct

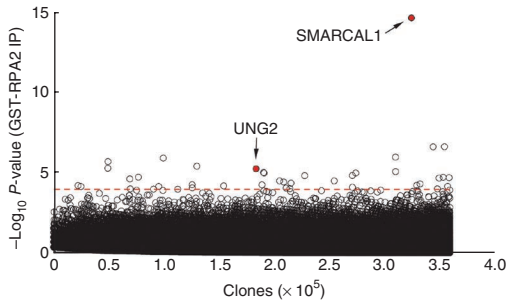


Figure 4 PhIP-Seq can identify peptide-protein interactions. GST-RPA2 was used to precipitate phage from the T7-Pep library on magnetic glutathione beads. $-\text{Log}_{10}$ *P*-values of enrichment were calculated using the generalized Poisson method. Clones are arranged in increasing input abundance from left to right. The experiment identified two of the known RPA2 binding partners SMARCAL1 ($P < 10^{-14}$) and UNG2 ($P < 10^{-5}$), shown in red.

© 2011 Nature America, Inc. All rights reserved.

Table 3 Dependence of peptide–RPA2 interaction on integrity of RPA2 binding motif

Gene	T7-Pep clone	Aligned peptide	$-\text{Log}_{10}$ *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

Aligned phage peptides containing the RPA2-binding motif (underlined) are shown next to their $-\log_{10}$ *P*-value of enrichment. Significantly enriched peptides are shown in bold.

peptide from the known interactor SMARCAL1 ($P < 10^{-14}$; **Fig. 4**), among others (**Supplementary Table 3**). The enriched SMARCAL1 peptide contains a previously identified motif that binds RPA2 (refs. 24,25). Four other proteins are known to contain this motif (UNG2, TIPIN, XPA and RAD52), but most motifs were significantly disrupted by the positions of the breaks between peptides (**Table 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 interaction motifs.

DISCUSSION

We have developed a proteomic technology based on a phage library that uniformly expresses the complete human peptidome. Combining the T7-Pep phage library with high-throughput DNA sequencing enables a variety of proteomic investigations. In addition to applications in autoimmune disease, our approach can be used 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 a 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 binding surfaces on natively folded proteins tend to be dominated by 'discontinuous' epitopes, which are patches of ~4 to 14 amino acid side chains formed by two or more noncontiguous peptides brought into proximity during protein folding[26,27]. If the protein is divided into its constituent peptides, antibody affinity is expected to decrease owing to (i) the loss of contacts contributed by noncontiguous residues, and (ii) the increased entropic costs of binding a free peptide as opposed to the natively constrained peptide. The degree to which individual peptides interact with a given antibody is difficult to predict, and is expected to vary widely. Although our study demonstrates the utility of 36-amino acid peptides, further work will be required to define the precise false-negative discovery rate inherent in the use of the T7-Pep library. Autoantibodies that target normally inaccessible epitopes have also been reported, such as those that recognize proteolytic cleavage products[28,29], misfolded proteins or protein aggregates[30,31]. In circumstances such as these, antigen discovery with full-length, folded proteins may be less sensitive than with shorter peptides.

By performing PhIP-Seq with cerebrospinal fluid from a PND patient (patient A) with a known NOVA autoantibody, we identified this protein as well as a previously unreported, testis-restricted[19] autoantigen (TGIF2LX). As we also found 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 may warrant further investigation as a biomarker for PND. PhIP-Seq analysis of cerebrospinal fluid from two PND patients with uncharacterized antibodies (patients B and C) uncovered likely neuronal targets of their autoimmune syndromes. In patient B, high-titer GAD65 antibodies bound two distinct peptides from the region of the protein associated with stiff person syndrome. Notably, GAD65 targeting in stiff person syndrome occurs more often in people without cancer, raising the possibility that at least part of this neurological syndrome may have been unrelated to the patient's cancer. Because management of stiff person syndrome differs from that of PND, this finding underscores the utility of unbiased antibody profiling to distinguish between deceptively similar disease states[32]. 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, as the autoantibodies might be incidental in nature, arise because of epitope spreading or might simply exhibit noncognate cross-reactivity.

Several notable features of our approach 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 protein, indicating that an important degree of conformational information is retained in the peptide library. Second, for proteins with known crystal structures, using short 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 represents the antigenic epitope.

The T7-Pep library 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.

In conclusion, we have taken a synthetic biological approach to develop a proteomic reagent useful in translational medicine. Combining this reagent with high-throughput DNA sequencing permits unbiased and quantitative analysis of autoantibody repertoires in humans. PhIP-Seq thus complements existing proteomic technologies in the study of autoimmune diseases for which the relevant autoantigens remain unknown.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/nbt/index.html.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

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

© 2011 Nature America, Inc. All rights reserved.

US 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.

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 the supervision of G.M.C. and implemented by H.B.L. PhIP-Seq candidates were confirmed by H.B.L. The RPA2 experiment was performed by A.C. The manuscript was prepared by H.B.L. and edited by N.L.S. and S.J.E.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/nbt/index.html.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Graham, A.L. *et al.* Fitness correlates of heritable variation in antibody responsiveness in a wild mammal. *Science* **330**, 662–665 (2010).
2. Faix, P.H. *et al.* Phage display of cDNA libraries: enrichment of cDNA expression using open reading frame selection. *Biotechniques* **36**, 1018–1022 (2004).
3. Albert, M.L. & Darnell, R.B. Paraneoplastic neurological degenerations: keys to tumour immunity. *Nat. Rev. Cancer* **4**, 36–44 (2004).
4. Wang, X. *et al.* Autoantibody signatures in prostate cancer. *N. Engl. J. Med.* **353**, 1224–1235 (2005).
5. Anderson, K.S. *et al.* A protein microarray signature of autoantibody biomarkers for the early detection of breast cancer. *J. Proteome Res.* **10**, 85–96 (2011).
6. Zacchi, P., Sblattero, D., Florian, F., Marzari, R. & Bradbury, A.R.M. Selecting open reading frames from DNA. *Genome Res.* **13**, 980–990 (2003).
7. Kim, Y. *et al.* Identification of Hnrph3 as an autoantigen for acute anterior uveitis. *Clin. Immunol.* **138**, 60–66 (2011).
8. 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).
9. Swann, J.B. & Smyth, M.J. Immune surveillance of tumors. *J. Clin. Invest.* **117**, 1137–1146 (2007).
10. Darnell, R.B. & Posner, J.B. Paraneoplastic syndromes involving the nervous system. *N. Engl. J. Med.* **349**, 1543–1554 (2003).
11. Musunuru, K. & Kesari, S. Paraneoplastic opsoclonus-myoclonus ataxia associated with non-small-cell lung carcinoma. *J. Neurooncol.* **90**, 213–216 (2008).
12. Srivastava, S. & Chen, L. A two-parameter generalized Poisson model to improve the analysis of RNA-seq data. *Nucleic Acids Res.* **38**, e170 (2010).
13. Consul, P. & Shoukri, M. Maximum likelihood estimation for the generalized poisson distribution. *Comm. Statist. Theory Methods* **13**, 1533–1547 (1984).
14. Bailey, T.L. & Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **2**, 28–36 (1994).
15. Almeida, L.G. *et al.* CTdatabase: a knowledge-base of high-throughput and curated data on cancer-testis antigens. *Nucleic Acids Res.* **37**, D816–D819 (2009).
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.* **165**, 7253–7261 (2000).
17. 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. USA* **95**, 6919–6923 (1998).
18. 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).
19. Berglund, L. *et al.* A genecentric human protein atlas for expression profiles based on antibodies. *Mol. Cell. Proteomics* **7**, 2019–2027 (2008).
20. 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).
21. Schwartz, H.L. *et al.* High-resolution autoreactive epitope mapping and structural modeling of the 65 kDa form of human glutamic acid decarboxylase. *J. Mol. Biol.* **287**, 983–999 (1999).
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. *Neurobiol. Dis.* **38**, 210–218 (2010).
23. Rual, J.-F. *et al.* Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**, 1173–1178 (2005).
24. 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).
25. 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).
26. Barlow, D.J., Edwards, M.S. & Thornton, J.M. Continuous and discontinuous protein antigenic determinants. *Nature* **322**, 747–748 (1986).
27. Jin, L., Fendly, B.M. & Wells, J.A. High resolution functional analysis of antibody-antigen interactions. *J. Mol. Biol.* **226**, 851–865 (1992).
28. 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).
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.* **142**, 148–154 (2005).
30. 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).
31. Papachroni, K.K. *et al.* Autoantibodies to alpha-synuclein in inherited Parkinson's disease. *J. Neurochem.* **101**, 749–756 (2007).
32. 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).
33. Derda, R. *et al.* Diversity of phage-displayed libraries of peptides during panning and amplification. *Molecules* **16**, 1776–1803 (2011).

© 2011 Nature America, Inc. All rights reserved.

ONLINE 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 $MgCl_2$, 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 3×,
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 29×
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 (5×10^{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 H_2O. 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 H_2O, 5 µl 10× 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 39×
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, which is available for download upon request. The reference sequences were truncated to the length of the reads and alignment was constrained to the appropriate strand. Analysis of enrichment significance was performed using MATLAB scripts that can be provided as .m files. A list of clones enriched by beads alone can also be provided upon request.

© 2011 Nature America, Inc. All rights reserved.

doi:10.1038/nbt.1856

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 h 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. MagneGST glutathione beads (Promega; 40 μl) were incubated in 1 ml of undiluted bacterial lysate for 2 h. Beads were then washed three times with phosphate buffered saline (PBS). M9LB (1 ml) containing 5×10^{10} pfu of T7-Pep was then used to resuspend the beads (now coated with GST or GST-RPA2). The mix was rotated 24 h 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 null distribution, 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 immunoprecipitation reads (x_i)[13].

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

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 λs (**Fig. 2c**). Calculation of θs for all input values revealed the near linearity of this parameter, and so we linearly regressed this parameter before calculating the *P*-values.

Validation of full-length candidate autoantigens. We used the ORFeome collection of full-length proteins, which was generated by PCR and Gateway recombinational cloning[34], as a source for testing autoantigen candidates by immunoblot. Entry vectors were recombined into the appropriate mammalian expression vector (cytomegalovirus promoter driving ORF expression with either C-terminal GFP fusion or N-terminal FLAG epitope tag) and mini-prepped for transient transfection.

293T cells were plated 24 h before transfection at a density of 0.8 million cells per well of a 6-well plate and grown in DMEM containing 10% FBS. TransIT-293T transfection reagent (Mirus, MIR 2700) was mixed with 2 μg expression plasmid per well, and added to the cells. After 24 h, cells were harvested in 200 μl standard 1× RIPA-based laemmli/DTT sample buffer with complete protease inhibitor cocktail (Roche) and sonicated for 30 s. Insoluble material was removed by centrifugation. Lysate (2–20 μl) was run out on 4–20% Bis-Tris polyacrylamide gels and transferred onto nitrocellulose using the iBlot system (Invitrogen). Membranes were blocked 1 h in 5% milk and then stained with either patient cerebrospinal fluid (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 cerebrospinal fluid 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 the commercial GAD65 radioimmunoassay (RIA 81596; Mayo Medical Laboratories), spinal fluid was diluted eightfold in PBS and shipped on dry ice to the Mayo Clinic Medical Laboratories.

For immunoprecipitation–western blotting, cell lysate was harvested in standard RIPA buffer with complete protease inhibitor cocktail and sonicated for 30 s. Insoluble material was removed by centrifugation. Lysate (150 μl) 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 h 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 immunoprecipitated protein and 10% of the input lysate were subjected 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. Two microliters of each clone, after normalizing for titer, were spotted directly onto a nitrocellulose membrane and allowed to dry for 30 min. Membranes were blocked with 5% milk, TBST for 1 h at 25 °C, and then stained overnight at 4 °C with 1 μg/ml of cerebrospinal fluid antibody diluted in a solution containing a 1:1 mix of 5% milk, TBST and T7 10-3b-FLAG phage lysate. Human antibody from cerebrospinal fluid was then detected with 1:3,000 peroxidase-conjugated goat affinity-purified secondary anti-Human IGG (whole molecule; MP Biomedicals, 55252) in 2.5% milk, TBST. Quantification was performed by scanning developed films and analyzing the .tiff file with ImageJ software.

34. 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).

© 2011 Nature America, Inc. All rights reserved.

Supporting Online Material for:

Autoantigen discovery with a synthetic human proteome

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

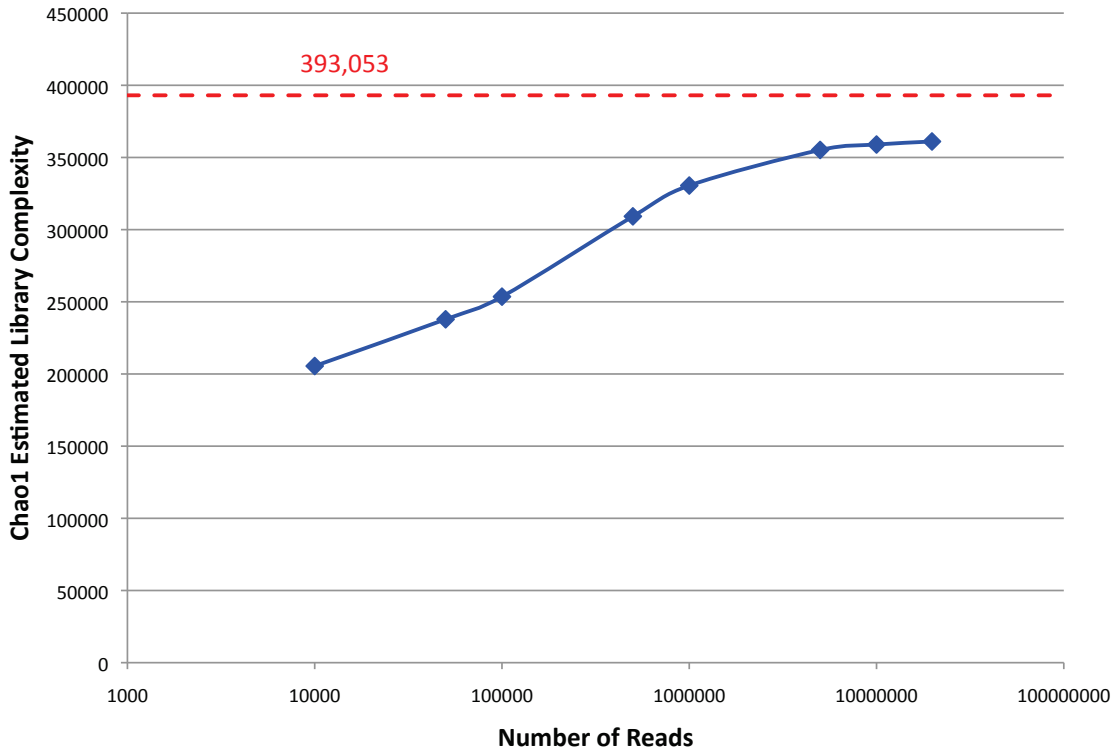
*Correspondence to:
nsolimini@rics.bwh.harvard.edu, selledge@genetics.med.harvard.edu

This file contains:

Supplementary Figures 1-9
Supplementary Tables 1-3

Nature Biotechnology: doi:10.1038/nbt.1856

Supplementary Figure 1.



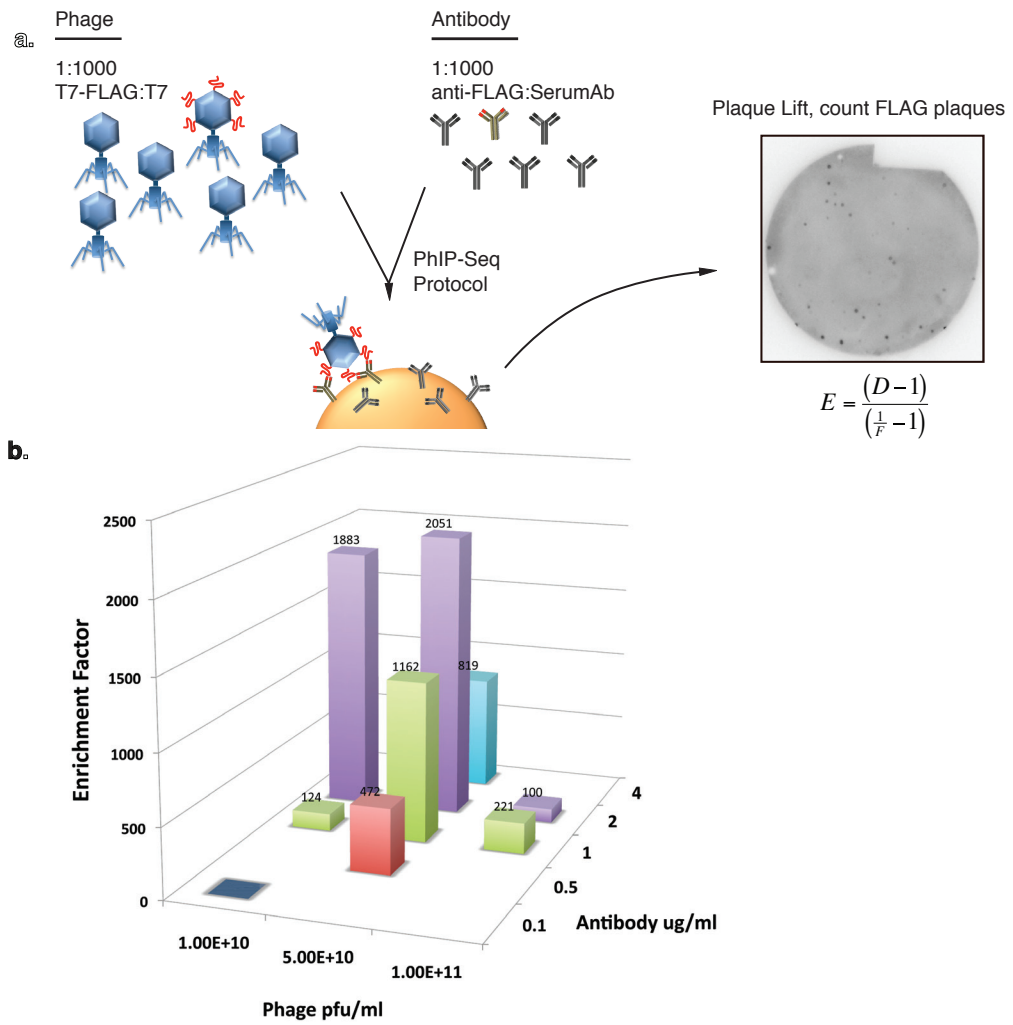
Supplementary Figure 1. The effect of sequencing depth on estimated library complexity.
Chao1 estimates of library complexity given by

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

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).

Nature Biotechnology: doi:10.1038/nbt.1856

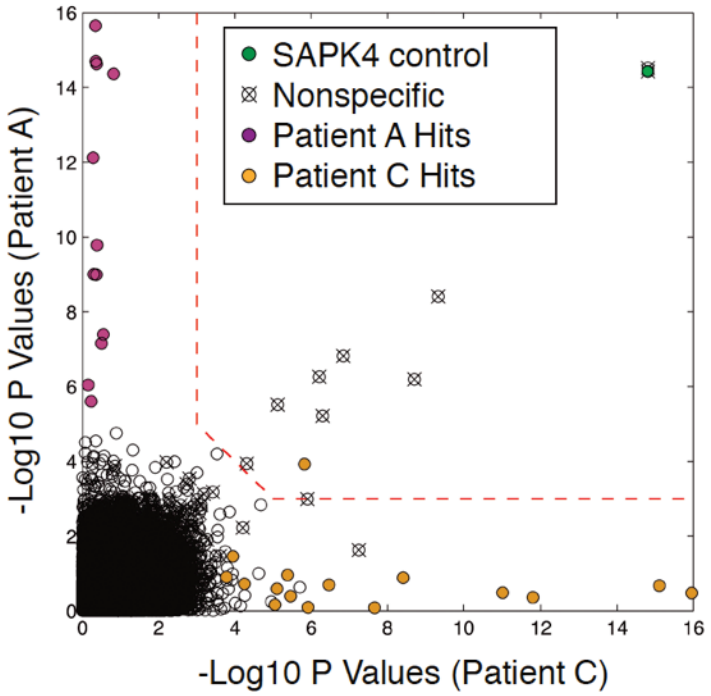
Supplementary Figure 2.



Supplementary Figure 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 5×10^{10} pfu/ml phage and 2 μg/ml total antibody.

Nature Biotechnology: doi:10.1038/nbt.1856

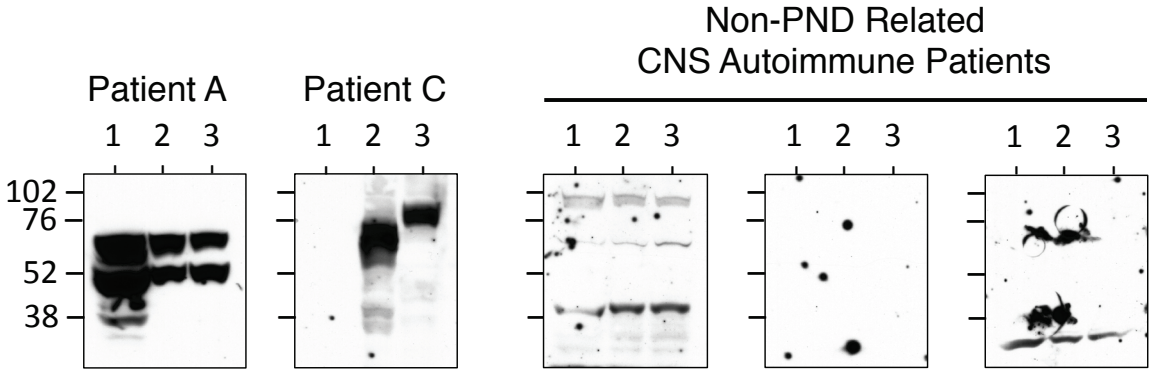
Supplementary Figure 3.



Supplementary Figure 3. Comparison of PhIP-Seq experiments with different patients.
Scatter plot as in Fig. 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.

Nature Biotechnology: doi:10.1038/nbt.1856

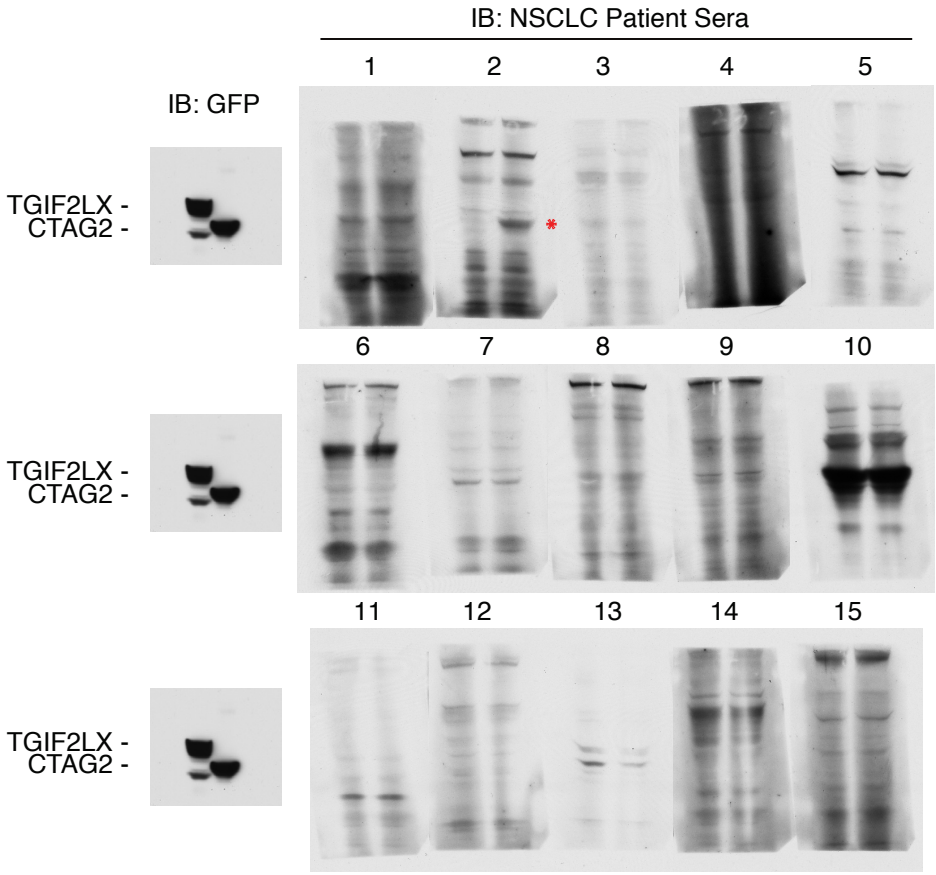
Supplementary Figure 4.



Supplementary Figure 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.

Nature Biotechnology: doi:10.1038/nbt.1856

Supplementary Figure 5.



Supplementary Figure 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.

Nature Biotechnology: doi:10.1038/nbt.1856

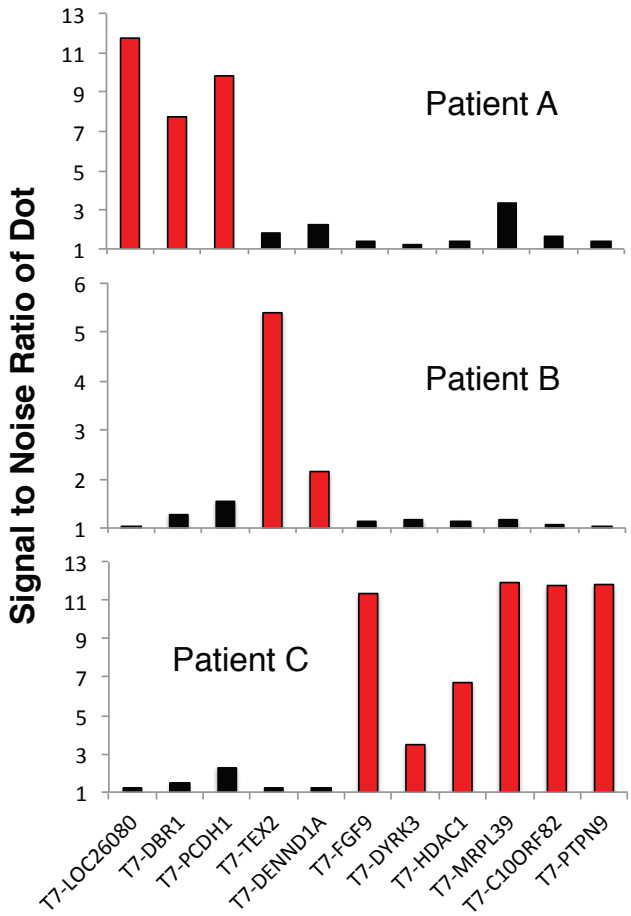
Supplementary Figure 6.

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 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.

Nature Biotechnology: doi:10.1038/nbt.1856

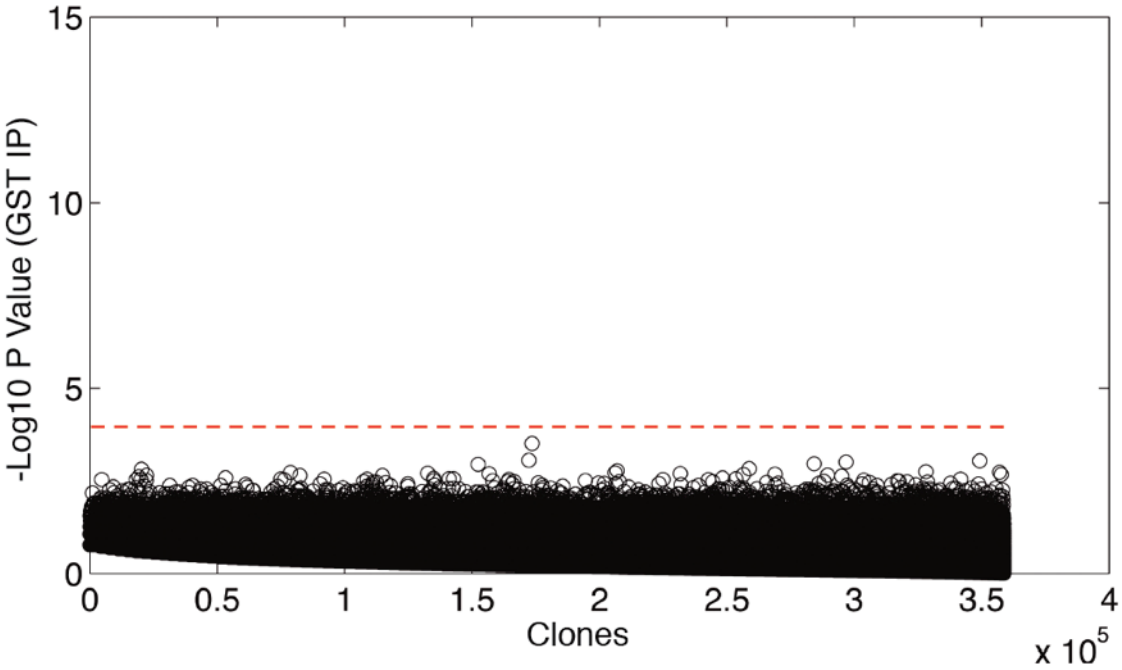
Supplementary Figure 7.



Supplementary Figure 7. Quantification of T7 Candidate Dot Blots.
The dot blots in Fig. 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.

Nature Biotechnology: doi:10.1038/nbt.1856

Supplementary Figure 8

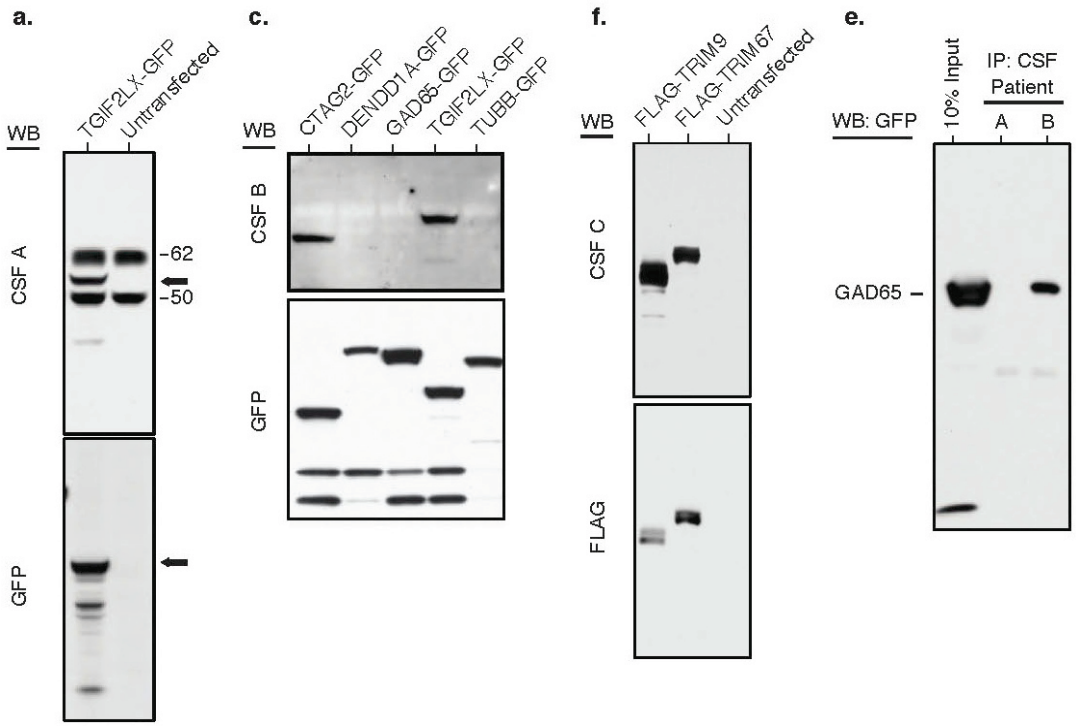


Supplementary Figure 8. PhIP-Seq –Log10 p-values for T7-Pep enrichment by GST alone.

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 ($P < 10^{-4}$).

Nature Biotechnology: doi:10.1038/nbt.1856

Supplementary Figure 9



Supplementary Figure 9. Full blots from Figure 3.
Shown here are the blots corresponding to Figure 3a, c, f and e, which have been expanded to include all detectable bands.

Nature Biotechnology: doi:10.1038/nbt.1856

Supplementary Table 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 1. Subpool analysis of multiple insertions and vector re-ligation 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.

Nature Biotechnology: doi:10.1038/nbt.1856

Supplementary Table 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 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. Great strides have been made to improve the fidelity of this technology[34], resulting in a per-cycle yield of 99.5%. However, since oligonucleotide synthesis for our libraries required as many as 140 cycles, even very small per-cycle losses can result in a significant loss of fidelity.

Nature Biotechnology: doi:10.1038/nbt.1856

Supplementary Table 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. 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.

SUPPLEMENTAL REFERENCES

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

Nature Biotechnology: doi:10.1038/nbt.1856