nature protocols

PROTOCOL

https://doi.org/10.1038/s41596-018-0025-6

Corrected: Publisher Correction

PhIP-Seq characterization of serum antibodies using oligonucleotide-encoded peptidomes

Divya Mohan[1,4], Daniel L. Wansley[1,4], Brandon M. Sie[1,4], Muhammad S. Noon[1], Alan N. Baer[2], Uri Laserson[3,4]* and H. Benjamin Larman[1,4]*

The binding specificities of an individual's antibody repertoire contain a wealth of biological information. They harbor evidence of environmental exposures, allergies, ongoing or emerging autoimmune disease processes, and responses to immunomodulatory therapies, for example. Highly multiplexed methods to comprehensively interrogate antibody-binding specificities have therefore emerged in recent years as important molecular tools. Here, we provide a detailed protocol for performing 'phage immunoprecipitation sequencing' (PhIP-Seq), which is a powerful method for analyzing antibody-repertoire binding specificities with high throughput and at low cost. The methodology uses oligonucleotide library synthesis (OLS) to encode proteomic-scale peptide libraries for display on bacteriophage. These libraries are then immunoprecipitated, using an individual's antibodies, for subsequent analysis by high-throughput DNA sequencing. We have used PhIP-Seq to identify novel self-antigens associated with autoimmune disease, to characterize the self-reactivity of broadly neutralizing HIV antibodies, and in a large international cross-sectional study of exposure to hundreds of human viruses. Compared with alternative array-based techniques, PhIP-Seq is far more scalable in terms of sample throughput and cost per analysis. Cloning and expression of recombinant proteins are not required (versus protein microarrays), and peptide lengths are limited only by DNA synthesis chemistry (up to 90-aa (amino acid) peptides versus the typical 8- to 12-aa length limit of synthetic peptide arrays). Compared with protein microarrays, however, PhIP-Seq libraries lack discontinuous epitopes and post-translational modifications. To increase the accessibility of PhIP-Seq, we provide detailed instructions for the design of phage-displayed peptidome libraries, their immunoprecipitation using serum antibodies, deep sequencing-based measurement of peptide abundances, and statistical determination of peptide enrichments that reflect antibody-peptide interactions. Once a library has been constructed, PhIP-Seq data can be obtained for analysis within a week.

Introduction

Development of the protocol

PhIP-Seq is a powerful technology platform that overcomes many previous limitations of comprehensive antibody-binding analysis[1–5]. PhIP-Seq combines OLS[6] with high-throughput DNA sequencing analysis of phage-displayed libraries. The synthetic oligonucleotide libraries are designed to encode peptide tiles that together span a library of protein sequences (entire proteomes, for example). The result is a comprehensive and normalized (uniform in abundance) representation of the encoded peptides, which we refer to as the 'peptidome(s)'. Deep DNA sequencing of phage-displayed peptidomes permits the quantification of each peptide's antibody-dependent enrichment, relative to other library peptides, spike-in standards, other antibody-containing samples, and/or negative-control samples lacking antibodies (Fig. 1). Sample multiplexing is achieved by using barcoded PCR primers, which are employed during preparation of the sequencing library. This markedly reduces the per sample DNA-sequencing cost, thereby enabling the analysis of large sample sets. Importantly, the streamlined protocol presented here can be easily performed by hand or automated for high-throughput sample processing using liquid-handling robotics. Compared with protein microarrays[7,8], PhIP-Seq is not restricted to proteins that have been cloned and can be expressed recombinantly. However, phage-displayed peptidomes lack many of the conformational epitopes present on full-length proteins. Compared with peptide microarrays[9], PhIP-Seq features longer, higher-quality peptides. However, due to the cost and effort of constructing new phage libraries,

[1]Division of Immunology, Department of Pathology, Johns Hopkins University, Baltimore, MD, USA. [2]Division of Rheumatology, Department of Medicine, Johns Hopkins University, Baltimore, MD, USA. [3]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [4]These authors contributed equally: Divya Mohan, Daniel L. Wansley, Brandon M. Sie, Uri Laserson, H. Benjamin Larman. *e-mails: uri@lasersonlab.org; hlarman1@jhmi.edu

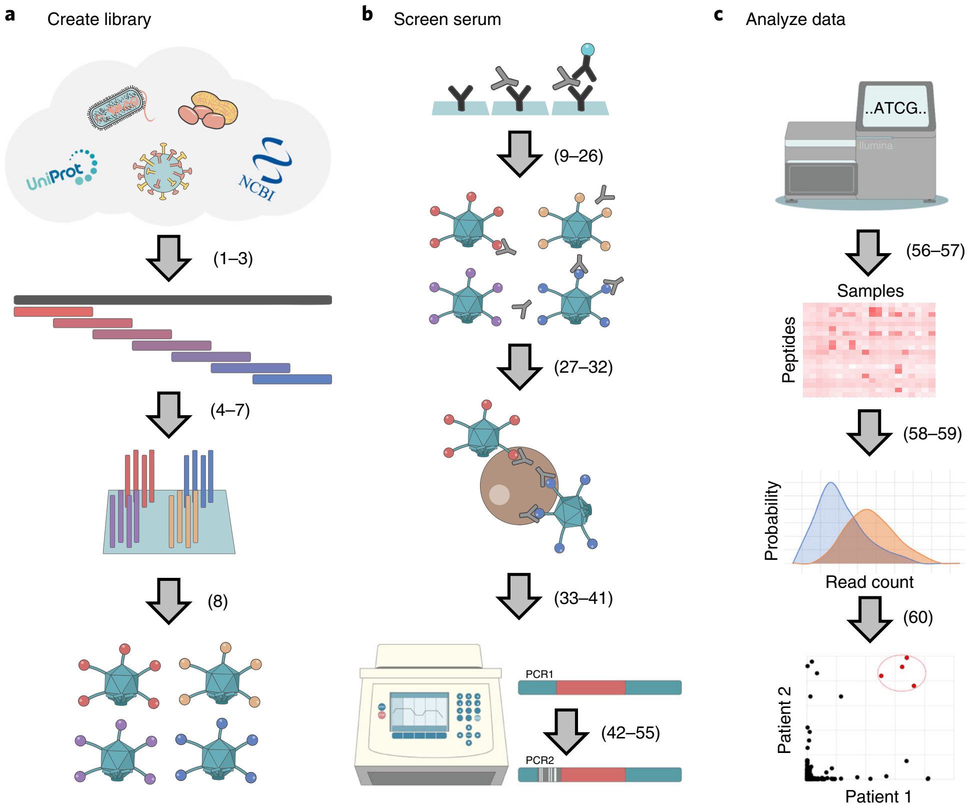

Fig. 1 | Overview of the PhIP-Seq methodology. Procedure step numbers are indicated in parentheses. **a**, A protein database is downloaded or designed. The pepsyn software is used to tile the protein sequences with overlapping peptide sequences. The oligonucleotide library encoding the peptide sequences is synthesized. The oligonucleotide library is PCR-amplified with adapters for cloning into the phage-displayed vector of choice. **b**, ELISA is used to quantify each sample's IgG content for normalizing the amount of antibody input into each phage-binding reaction. Antibodies and their bound phage are captured using protein A/G-coated magnetic beads. The library of peptide-encoding DNA sequences is amplified by PCR directly from the immunoprecipitate. A second round of hemi-nested PCR is used to add sample-specific barcodes and sequencing adapters to the PCR1 product. Barcoded amplicons are pooled for sequencing on an Illumina instrument. **c**, Fastq sequencing files are demultiplexed and aligned to the reference sequences to obtain a count matrix. Statistical analysis of the count matrix is performed to determine peptide enrichments. Project-specific analysis of peptide enrichments (e.g., identification of a common autoantigen) can then be carried out.

programmable peptide arrays may be more appropriate for screening sample-individualized peptide libraries. For projects involving large numbers of samples, the per-sample analysis cost of PhIP-Seq is roughly two orders of magnitude less expensive as compared with microarray-based alternatives.

Overview of the protocol

From start to finish, the key stages involved in a PhIP-Seq project include computational design of the peptidome(s) (Steps 1–6), construction of the phage library (Steps 7 and 8), quantification of the samples' immunoglobulin content (Steps 9–26), phage-antibody complex formation and immunoprecipitation (Steps 27–41), preparation of DNA-sequencing libraries (Steps 42–55), and deep sequencing and data analysis (Steps 56–60). Detailed protocols for the construction and expansion of phage libraries can be found elsewhere (in the Novagen T7Select System Manual (https://www.emdmillipore.com/Web-US-Site/en_CA/-/USD/ShowDocument-File?ProductSKU=EMD_BIO-70550&DocumentId=TB178.pdf&DocumentType=USP&Language=EN&Country=US), for example). Once a phage library is successfully constructed and expanded, it can be used for analysis of large samples sets and re-expanded at almost no cost.

Additional applications of the protocol

Beyond performing PhIP-Seq analysis of serum antibodies, we have also used the protocol to identify the epitopes of monoclonal antibodies[4] and the binding partners of recombinant proteins[1]. In addition to identifying cognate antibody targets, we have used PhIP-Seq to dissect the fine specificity of antibody binding with variant epitope libraries designed to contain informative non-synonymous mutations and/or truncations[5,10]. For antibody-isotype-specific analyses, this protocol can be easily adapted to incorporate streptavidin-coated magnetic beads prepared with biotinylated isotype-specific capture antibodies. These and other related applications may require substantial deviation from the protocol presented here, along with assay-specific optimization.

Limitations

It is important that users of PhIP-Seq understand its limitations. Phage-displayed peptide libraries may lack the conformational structure required to detect important binding specificities because of the limited length of the synthetic oligonucleotide library–encoded peptides (currently up to 90 aa, for example)[1,11]. Disulfide linkages and post-translational modifications will also be absent from typical T7 phage particles, which are produced in the cytoplasm of *Escherichia coli*. As antibodies frequently target conformational and modified epitopes, PhIP-Seq may frequently fail to identify the targets of monoclonal antibodies, compared with those of polyclonal antibodies, which often harbor a subset of specificities that recognize linear epitopes. Depending on the experimental context, PhIP-Seq may therefore perform optimally in combination with alternative methodologies intended to interrogate more 'native' protein antigens, such as protein microarray analysis[12,13], parallel analysis of translated open reading frames (ORFs)[14,15], and/or immunoprecipitation followed by mass spectrometry[16].

Experimental design

Design and construction of a bacteriophage library

After downloading or constructing the protein sequence database or translated ORF database to be encoded, we use the pepsyn Python package (https://github.com/lasersonlab/pepsyn) to design our oligonucleotide library, including the following processes: (i) splitting the protein sequences into peptide tiles of chosen length and chosen length of overlap, (ii) selection of representative peptide sequences from peptide clusters of similarity greater than a chosen threshold, (iii) reverse translation of the selected peptide tiles with an optimized *E. coli* codon usage algorithm, (iv) addition of forward and reverse PCR-primer binding sequences to the resulting DNA sequences, and (v) removal of restriction cloning sites (aside from those intended) by silent codon substitution (Fig. 2a). The resulting DNA sequences are then provided to a DNA manufacturer for OLS. We have previously purchased libraries from Agilent Technologies and Twist Bioscience.

The optimal lengths and overlaps of the peptide tiles are governed by considerations related in part to the manufacturing of the oligonucleotide library. There are two main tradeoffs in terms of tile length. Longer peptides will contain greater secondary structure, which is an important aspect of many antibody–epitope interactions. However, longer oligonucleotides will contain more mutations per peptide, and thus may reduce the overall quality of the library. A second consideration relates to assessment of polyclonal responses. Observing multiple, non-overlapping enriched peptides from the same protein may provide increased confidence in an antigen-driven response versus a single, potentially cross-reactive antibody specificity. The length of the overlaps determines both the density of the tiles (i.e., how many tiles per length of protein) and the size of the smallest epitopes contained in the library. There is another tradeoff in terms of library size (and thus the cost to construct and sequence it) versus the coverage of antigenic space.

Upon receipt of the synthetic oligonucleotide pool, standard PCR (we prefer the Herculase II DNA polymerase from Agilent) is used for amplification before restriction cloning into a phage vector of choice (we have used a derivative of the T7Select 10-3b, mid-copy system, called T7-FNS2), according to the manufacturer's instructions (Novagen T7Select System Manual). We have preferentially used the T7Select 10-3b mid-copy system for PhIP-Seq, primarily because lytic bacteriophage libraries are expected to exhibit less bias as compared with trans-membrane secretion systems (such as M13, for example). The mid-copy system permits display of up to ~1,000-aa-long peptides at a copy number of 5–15 per particle. The drawback of the T7Select system is that libraries must be packaged by using an expensive extract, which is also much less efficient compared with electroporation into host bacterial cells.

The success of any PhIP-Seq project will depend upon the quality of the starting phage library. Aside from the library design and fidelity of the OLS, the quality of the library is also determined by clonal dropout, skewing, titer, and presence of contaminants. 'Dropout' refers to the loss of peptide library members due to insufficient coverage of the library during construction of the initial,

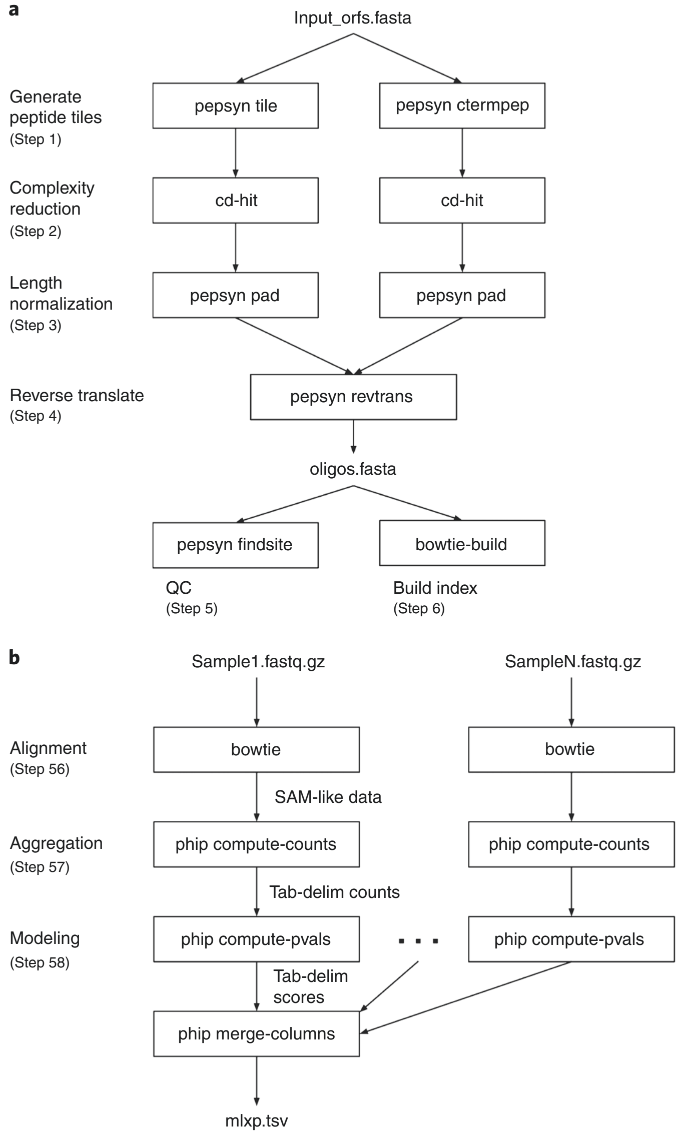

Fig. 2 | Bioinformatics workflows. Procedure step numbers are indicated. **a**, pepsyn workflow. Workflow for designing a peptide library. We provide only an outline of the protocol, as this stage will probably be customized depending on your library/preferences. **b**, phip-stat workflow for PhIP-Seq data analysis.

unexpanded phage library. We recommend scaling the library construction steps such that each library member is always represented on average by at least 100 infectious particles. For example, packaging of the T7 genomic DNA ligation reaction should result in at least 10^7 plaques for a library containing 10^5 unique peptides. This applies to library expansion too. Library skewing refers to changes in relative abundance of individual library members due to differing growth kinetics and stochastic fluctuation. There are a variety of factors that determine how a particular displayed peptide will influence the kinetics of the phage clone's growth. Even small differences in phage-replication efficiency can result in substantial differences in a clone's final representation within the expanded library, especially after serial expansions. Phage clones that express truncated peptides due to nonsense mutations may have a growth advantage over their unmutated counterparts, especially for longer or toxic peptides, thus reducing the representation of the corresponding unmutated library members. Phage library degradation due to skewing is minimized primarily by expanding the library on solid media (rather than in liquid culture) and by avoiding unnecessary serial passage of the library. The expanded screening library should have a plaque-forming unit (PFU) concentration (titer) that provides each library member with a representation of at least 10^5 PFU per ml. This means that for a library of complexity 10^5, an absolute minimal titer of 10^{10} PFU per ml must be achieved in the expanded library. For the T7Select 10-3b mid-copy system, we typically obtain titers ~10^{11} by centrifugally concentrating log-phase host bacterial cells to an optical density (at 600 nm) of ~4 just before performing plate amplification of the library (otherwise according to the manufacturer's instructions). Finally, it is important to remove particulates (including bacterial cells) from the expanded phage library lysate by centrifugation and to prevent further growth of bacteria by the addition of a second antibiotic, to which the host cells are sensitive. We store our final expanded phage peptidome library aliquots indefinitely at −80 °C after the addition of 10% (vol/vol) DMSO.

Library quality control

The quality of each new, or newly expanded, phage library should be assessed in two ways before screening. First, several plaques (we recommend at least 20) should be individually picked, and the inserts should be analyzed by Sanger sequencing to assess the fidelity of the oligonucleotide synthesis (as described in the Novagen T7Select System Manual). Clones expressing mutated or truncated inserts may have a growth advantage over intact inserts, so it is important to pick from a representative range of physical plaque sizes to avoid unintentionally underestimating the quality of the library because of biased plaque selection. Second, the library should be analyzed using deep sequencing[1,5] to assess the baseline distribution of the clonal frequency and the completeness of the library. Ideally, 90% of the library should fall within one log of clonal frequency, and at least 90% of the library should be observed at a sequencing depth of at least ten reads per clone.

Protein A/G–based immunoprecipitation

For convenience and optimal performance, we suggest using protein A/G–coated magnetic beads as the capture matrix for PhIP-Seq experiments. To obtain reproducible inter-sample data, it is important that the amount of IgG antibody input be uniform and below the binding capacity of the capture matrix. Steps 9–26 of this protocol are therefore devoted to the measurement of each sample's IgG concentration, to ensure appropriate IgG input.

Antibody isotype–specific immunoprecipitations

For antibody isotype–specific immunoprecipitations, we recommend pre-coating M-280 streptavidin Dynabeads (Invitrogen, cat. no. 11205D) with an amount of capture antibody that is two to four times the binding capacity of the beads. This will minimize bead aggregation, which can markedly reduce target antibody capture efficiency. For isotype-specific antibody capture experiments, we typically immobilize ~1 μg of capture antibody, for the capture of, at most, 1 μg of target antibody. Aside from these bead-preparation protocol variations and isotype-specific ELISA quantification, the remainder of the Procedure will apply equally well.

Controls

Within-experiment negative controls (and for 96-well plate runs, within plate negative controls) are extremely important for obtaining interpretable results. Such controls depend on the experimental design. For example, profiling antibodies to identify antibody–phenotype associations should include analysis of individuals without the phenotype, but who are otherwise well matched. Technical positive and negative controls are also important to include when possible. Positive controls may include monoclonal antibodies or previously analyzed samples. We strongly suggest including a set of negative controls that lack antibody input ('mock IPs'), to obtain important background binding information for each phage clone. Sequencing of the unenriched input library is required to determine the clonal distribution of the starting library and to use the computational pipeline provided here. To this end, we typically add ~1 × 10^7 PFU of starting library to a PCR1 reaction.

Sequencing the library

A variety of high-throughput ('next generation') DNA-sequencing platforms now exist for the analysis of DNA libraries. PhIP-Seq can, in principle, be adapted to any such platform, provided that the number of sequencing reads per library member is sufficient for quantification of peptide enrichment (ideally ~10 reads per clone on average). We have primarily used the Illumina HiSeq and NextSeq instruments, as they provide the lowest per read cost. The PCR primer sequences presented here therefore include the Illumina sequencing adapters (suitable for both single and paired-end flow cells). Substitution with different platform-specific adapters should be straightforward. In addition, we present PCR primers that are specific to one of our T7-FNS2 derived libraries (VirScan), but these are easily replaced with alternative, library-specific primers. The sequencing primer provided here is

Table 1 | Primer sequences for PhIP-Seq

Step	Primer name	Sequence
PCR1 (Step 42)	T7-Pep2_PCR1_F	5′-ATAAAGGTGAGGGTAATGTC-3′
PCR1 (Step 42)	T7-Pep2_PCR1_R+ad_min	5′-CTGGAGTTCAGACGTGTGCTCTTCCGATCAGTTACTCGAGCTTATCGTC-3′
PCR2 (Step 45	T7-Pep2_PCR2_F_P5	5′-AATGATACGGCGACCACCGAGATCTACACGGAGCTGTCGTATTCCAGTC-3′
PCR2 (Step 48)	ad_min_**BCX**_P7 (*X* is 1–96)	5′-CAAGCAGAAGACGGCATACGAGAT**GACTGACT**GTGACTGGAGTTCAGACGTGTGCTC-3′
PCR3 (Step 52)	P5	5′-AATGATACGGCGACCACCGA-3′
PCR3 (Step 52)	P7.2	5′-CAAGCAGAAGACGGCATACGA-3′
Sequencing, read 1 (Step 55)	T7-VirScan_SP	5′-GGTGTGATGCTCGGGGATCCAGGAATTCCGCTGCGT-3′
Sequencing, read 2 (Step 55)	Index_SP	5′-GATCGGAAGAGCACACGTCTGAACTCCAGTCAC-3′

Primers for PCRs 1–3 can be used for any PhIP-Seq project that is based on the T7Select 10-3b FNS2 vector[1,5]. Underlined sequences are adapter sequences and thus do not participate in the initial round of PCR. The bold sequence is the barcode ('index') that uniquely defines the sample. We have designed and tested 96 of these sequences, an example ($X = 1$) is shown here. The remaining 95 ad_min_BCX_P7 primer sequences can be found in Supplementary Table 1. The T7-VirScan_SP sequencing primer is specific for analysis of the VirScan library. The Read 2 Illumina Index Read Primer ('Index_SP') is a standard Illumina primer and available for use from most high-throughput DNA-sequencing core facilities free of charge.

also specific to the VirScan library, but any appropriately designed sequencing primer can be used instead. The use of a sequencing primer that results in balanced base incorporation for at least the first five sequencing cycles is necessary for Illumina instruments to resolve clusters. Use of a PhIX spike-in (at >30%) can help with cluster resolution for biased libraries when this is not possible.

A second DNA barcode may be incorporated into the PCR2 forward primer (Table 1) for 'dual indexing' (e.g., AATGATACGGCGACCACCGAGATCTACACXXXXXXXXGGAGCTGTCGTATTCCAGTC, in which the eight Xs represent the eight nucleotides of the i5 index). This allows combinatorial (i5 + i7) barcoding of PCR products, thus increasing the level of potential sample multiplexing[17]. Of note, on certain Illumina instruments (e.g., the NextSeq 500), i5 is sequenced in the reverse direction, such that it requires a custom i5-sequencing primer (AGCATCACACCTGACTGGAATACGACAGCTCC).

Data analysis

Analysis of high-throughput DNA-sequencing data requires an informatics pipeline, which can be implemented on a high-performance computing cluster. The analytical stages include: (i) demultiplexing and alignment to the reference sequence database, (ii) tabulation of aligned sequences, (iii) statistical evaluation of each peptide's enrichment within each sample, and (iv) interpretation of peptide enrichments, the first three steps of which are illustrated in Fig. 2b. For stage iii, we have previously reported the use of a generalized Poisson distribution as a null model[1]. Conceptually, it is important to understand that sequencing-based quantitation of library member abundance is governed by sampling statistics of count data. More abundant clones will be sampled more deeply as compared with less abundant clones, meaning that differences in relative abundance can be measured more accurately for more abundant clones. For example, a tenfold enrichment can be much more reproducibly measured for a clone that is sequenced hundreds of times in the control condition versus a clone that is sequenced only once or twice in the control condition. Our statistical model therefore takes this into account when comparing enrichments among differentially abundant clones.

How can measures of phage clone enrichment be correlated to more familiar concepts such as assay dynamic range, signal-to-noise, and antibody titer? Unfortunately, there is no simple way to convert the statistical assessment of PhIP-Seq enrichments into parameters that are traditionally applied to single-plex assays, which typically produce chemical signals of a continuous (non-discrete) nature. For each individual target peptide, one could envision, for example, constructing a sample dilution standard curve to plot the PhIP-Seq enrichment *P* value against the signal from a corresponding ELISA assay. However, such an exercise may be of little value, as the relative impact of differences in antibody abundance, affinity, or avidity are expected to differ in a non-linear way between these two types of measurements.

After quantifying phage peptide enrichments, project-specific considerations will determine the best approach to the interpretation of their significance. For example, we have used permutation analyses of cross-sectional case–control studies to set false-discovery rate thresholds on lists of candidate disease-associated autoantibodies[2]. Longitudinal studies, on the other hand, may entail intra-patient pairwise sample comparisons. It should also be emphasized that PhIP-Seq is primarily a hypothesis-generating tool, and that absence of peptide enrichment cannot be interpreted as absence of the corresponding antibody specificity. We suggest confirming PhIP-Seq discoveries via at least one or two orthogonal assays. In the case of autoantigen confirmation, we have used mammalian cells that express full-length, epitope-tagged proteins[1,3]. Western blotting for the epitope tag can be used to assess the abundance of the tagged protein in the immunoprecipitate. ELISA assay using commercially available proteins is another possibility. For validation of anti-viral antibodies, clinically validated antibody tests are available for a variety of human pathogens.

Materials

Reagents

▲CRITICAL Prepare all solutions using deionized water. Prepare and store all reagents at room temperature (25 °C), unless otherwise indicated.

- Capture antibody: goat anti-human IgG-UNLB (Southern Biotech, cat. no. 2040-01)
- Detection antibody: goat F(ab′)$_2$ anti-human IgG-HRP (horseradish peroxidase) (Southern Biotech, cat. no. 2042-05)
- Human IgG ELISA standards (Life Technologies, cat. no. 02-7102)
- 1-Step Turbo TMB (3,3',5,5'-tetramethylbenzidine) ELISA Substrate Solution (Thermo Fisher Scientific, cat. no. 34022)
- Stop solution: 2 N or 1 M H_2SO_4 (Sigma Aldrich, cat. no. 258105-100ml)
- Dynabeads Protein A (Invitrogen, cat. no. 10002D)
- Dynabeads Protein G (Invitrogen, cat. no. 10004D)
- Serum samples for study !CAUTION Serum samples must be analyzed in compliance with IRB-approved human subject research guidelines. The example data shown here were obtained from de-identified donors under JHU Human Subject Research exemption IRB00049327. !CAUTION Wear a one-time-use splash shield, a mask, a biohazard suit, and surgical gloves while handling human serum samples. Dispose of protective equipment after a single use. ▲CRITICAL Serum samples should be stored in cryovials at −80 °C until use. The diluted serum samples (1:1 million dilution) can be stored at 4 °C for 2 d.
- PCR primers (Integrated DNA Technologies; see Table 1 and Supplementary Table 1)
- Herculase II polymerase (store at −20 °C; Agilent, cat. no. 600679)
- DNA Clean & Concentrator Kit (store at room temperature; Zymo Research, cat. no. D4005)
- Agarose (type I EEO, store at room temperature; Sigma-Aldrich, cat. no. 9012-36-6)
- KAPA Library Quantification Kit (store at −20 °C; KAPA Biosystems, cat. no. KK4828)
- NucleoSpin Gel and PCR Clean-up (store at room temperature; Macherey–Nagel, cat. no. 740609.50)
- T7Select Packaging Kit (store at −80 °C; EMD Millipore, cat. no. 70014-3), for library construction
- T7Select 10-3b DNA (store at 4 °C; EMD Millipore, cat. no. 70548), for library construction
- $NaHCO_3$ (Sigma-Aldrich, cat. no. S5761)
- Na_2CO_3 (Sigma-Aldrich, cat. no. 223530)
- PBS (Thermo Fisher Scientific, cat. no. 10010023)
- FBS (Corning, cat. no. 35-011-CV)
- 1Kb Plus DNA Ladder (Thermo Fisher Scientific, cat. no. 10787018)
- 150 mM NaCl (Sigma-Aldrich, cat. no. S7653)
- 50 mM Tris-HCl (Sigma-Aldrich, cat. no. RES3098T-B7)
- 0.1% (vol/vol) NP-40 (Sigma-Aldrich, cat. no. 492016)
- Tween 20 (Sigma-Aldrich, cat. no. P1379)
- FBS (Thermo Fisher, cat. no. 16140071)
- dNTPs (deoxyribose nucleoside triphosphates; Applied Byosystems, cat. no. N8080261)

Equipment

- E-max Precision Microplate Reader (SpectraMax iD3, Molecular Devices)
- 96-Well ELISA plates (Thermo Fisher Scientific, cat. no. 3455)
- Thermolyne Labquake Rotator (Barnstead, cat. no. 3.625.485)
- 2.0 ml, polypropylene (PP), pyramid-bottom, non-sterile 96-well plates (Cell Treat, cat. no. CT-229572)

- Full-skirted PCR plate (Bio-Rad, cat. no. HSP9601)
- Silicone 96-well plate sealing gaskets (Thermo Fisher Scientific, cat. no. AB0675)
- MicroAmp optical adhesive film (Thermo Fisher Scientific, cat. no. 4311971)
- 96-Well magnet (Agilent, cat. no. VP 771G-4RM) or magnetic particle concentrator (for 1.5-ml tube; Thermo Fisher Scientific, cat. no. MPC-S)
- Saran wrap (Saran, cat. no. 00140)

Software

- Prism (v6, GraphPad Software: https://www.graphpad.com/scientific-software/prism/)
- Python (pepsyn requires Python 3.6+; phip-stat works with Python 2.7 and Python 3: https://www.python.org/)
- pepsyn for oligo design tools (https://github.com/lasersonlab/pepsyn) ▲CRITICAL The pepsyn package is under active development; check README on the GitHub site for the latest protocols.
- (Optional) cd-hit for clustering oligos to reduce redundancy (http://weizhongli-lab.org/cd-hit/)
- phip-stat for processing of PhIP-Seq raw data (https://github.com/lasersonlab/phip-stat) ▲CRITICAL The phip-stat package is under active development; check README on GitHub for the most up-to-date protocol. Example data are available in the GitHub repository in the examples/directory.
- Bowtie for alignment (http://bowtie-bio.sourceforge.net/index.shtml)
- (Optional) HPC (high performance computing) cluster with batch job scheduler such as LSF (https://www.ibm.com/support/knowledgecenter/en/SSWRJV_10.1.0/lsf_welcome/lsf_kc_ss.html), Grid Engine (https://en.wikipedia.org/wiki/Oracle_Grid_Engine), or SLURM (https://slurm.schedmd.com/)

Reagent setup

ELISA coating buffer

Dissolve 2.93 g of $NaHCO_3$ and 1.5 g of Na_2CO_3 in 900 ml of deionized water, adjust pH to 9.5 (pH indicated is critical), and adjust final volume of the buffer to 1 liter. Store at room temperature for up to 1 month.

ELISA wash buffer

Mix 0.5 ml of Tween 20 in 1 liter of PBS. Store at room temperature for up to 1 month.

ELISA blocking buffer

Mix 5% (vol/vol) FBS in 1× PBS. Make fresh and store at 4 °C for no more than 1 week.

Magnetic bead wash buffer

Supplement 1× PBS with 0.02% (vol/vol) Tween 20. Store at room temperature for up to 1 month.

IP wash buffer

Make a mixture of 150 mM NaCl (8.76 g/l), 50 mM Tris-HCl (7.88 g/liter), and 0.1% (vol/vol) NP-40 (1 ml/liter), and adjust the pH to 7.5 (pH indicated is critical). Store at 4 °C for up to 1 month.

Equipment setup

Input data

This protocol assumes the existence of a text file called `input_orfs.fasta` that contains the full protein library sequences in .fasta format. We recommend that the sequence identifiers for each protein sequence be a simple, unique name, ideally without spaces or other punctuation other than underscores, periods, or dashes. Most tools in the pepsyn package accept '–' as the input and output file, which will read/write .fasta data from `stdin` and `stdout`. This facilitates the modular integration of various tools into processing pipelines using the Unix pipe functionality. The protocol below is just one possible example that illustrates this principle, and the separate processing parts can be easily swapped or varied. The computations are generally fast, allowing rapid iteration on designs. The commands below are executed in a Bash shell.

Software installation and setup

We recommend using the Anaconda/Miniconda Python distribution for all Python work. It is easy to install and comes with a modern package manager (conda) to manage local Python environments. It can also install other non-Python software (such as cd-hit and Bowtie).

Using the following commands, install miniconda into your home directory on your local machine or cluster:

```
curl  https://repo.continuum.io/miniconda/Miniconda3-latest-Linux-
x86_64.sh \
   > miniconda3.sh
bash miniconda3.sh -b -p $HOME/miniconda3
```

(See the ContinuumIO documentation at https://conda.io/docs/installation.html for a Windows-compatible command.) If desired, add the new Python distribution to your PATH to make sure it is set as the default distribution by setting the following command in your `.bash_profile` configuration file (using the correct location of the conda installation):

```
export PATH = "$HOME/miniconda3/bin:$PATH"
```

Using the following commands, install the Python packages required for pepsyn and phip-stat:

```
conda install -y numpy scipy biopython click tqdm
```

Finally, you can use conda to install Bowtie and cd-hit as well. First, add the 'bioconda' channel to your conda installation, using the following commands:

```
conda config --add channels conda-forge
conda config --add channels defaults
conda config --add channels r
conda config --add channels bioconda
```

Then install Bowtie and cd-hit, using the following commands:

```
conda install -y bowtie cd-hit
```

The tools should now be available for use from your `PATH`. Using conda, it is very easy to switch between Python 2 and 3, or different custom environments.

Using phip-stat to process the raw data, many users will be working on an HPC cluster with a job scheduler. Typically on such a cluster, one would submit a job for batch execution:

```
bsub -q expressalloc -W 0:20 my_command
```

This will execute `my_command` somewhere on the cluster, assuming you use the LSF scheduler. ▲**CRITICAL** Consult with your local HPC cluster for guidance on submitting many jobs concurrently. Steps that are relatively computationally intensive and parallelizable are pointed out in the Procedure.

Procedure

Synthetic peptidome library design ● Timing 1 d

1 Generate two sets of peptide sequences: one set that tiles the existence of a text fileacross the whole protein and a second set that consists of C-terminal peptides, using the following commands in the pepsyn software package.

```
TILESIZE=56
OVERLAP=28
cat input_orfs.fasta \
  | pepsyn x2ggsg - - \
  | pepsyn tile -l $TILESIZE -p $OVERLAP - - \
  | pepsyn disambiguateaa - - \
  > orf_tiles.fasta
cat input_orfs.fasta \
  | pepsyn x2ggsg - - \
  | pepsyn ctermpep -l $TILESIZE --add-stop - - \
  | pepsyn disambiguateaa - - \
  >cterm_tiles.fasta
```

Note how the commands are stitched together into a pipeline, each one reading .fasta data and writing .fasta data, allowing for flexible and modular pipelines during the design phase. The first command (`pepsyn x2ggsg`) eliminates stretches of Xs by replacing them with a glycine–serine linker sequence. The next command (either `pepsyn tile` or `pepsyn ctermpep`) chops up each ORF into short tiles of specified length. The `tile` version generates overlapping sequences, whereas `ctermpep` takes only the last amino acids of the sequences (i.e., 'C-terminal peptide'). Finally, `disambiguateaa` removes ambiguous IUPAC amino acid codes (e.g., B for aspartic acid or asparagine) by generating all possible peptides. The peptides are written into `orf_tiles.fasta` and `cterm_tiles.fasta`. Note that we have elected to add amber stop codons to the C-terminal peptides to allow flexibility in whether native stop codons are incorporated into the peptide or not.

▲ **CRITICAL STEP** You can find usage notes by adding `-h` to any command (e.g., `pepsyn -h` or `pepsyn tile -h`). There are numerous additional tools that perform helpful operations in peptide design (e.g., `pepsyn clip` for trimming sequences and `pepsyn builddbg` for building a De Bruijn graph on k-mers).

2 The resulting files may contain peptides that are identical or highly similar to each other. Eliminate some of this redundancy using the cd-hit tool, similar to what is done in the UniProt database, using the following commands:

```
cd-hit -i orf_tiles.fasta -o orf_tiles_clustered.fasta \
  -c 0.95 -G 0 -A 50 -M 0 -T 1 -d 0
cd-hit -i cterm_tiles.fasta -o cterm_tiles_clustered.fasta \
  -c 0.95 -G 0 -aL 1.0 -aS 1.0 -M 0 -T 1 -d 0
```

In this particular case, we are clustering the peptide tiles to 95% (`-c 0.95`) local identity (`-G 0`) while controlling the alignment coverage (`-A 50` requires the alignment to cover at least 50 aa). The C-terminal peptides are aligned more stringently to ensure that the final residues of the ORF are not lost (`-aL 1.0 -aS 1.0` requires 100% of each sequence to be aligned with possible mismatches). Specifying `-M 0` allows unlimited memory, `-T 1` specifies one CPU thread, and `-d 0` ensures that sequence names are not truncated. See cd-hit documentation for more options (http://cd-hit.org). The clustered peptides are written to `orf_tiles_clustered.fasta` and `cterm_tiles_clustered.fasta`.

3 The rest of the peptide processing is the same for the C-terminal and tiled peptides. Use the following commands to first concatenate the files (`cat`). Because the results of the previous step could generate some peptide sequences shorter than 56 aa, also pad the peptides to make them of uniform length (`pad`).

```
cat orf_tiles_clustered.fasta cterm_tiles_clustered.fasta \
  | pepsyn pad -l $TILESIZE --c-term - - \
  > protein_tiles.fasta
```

The final peptide tiles are combined in `protein_tiles.fasta`.

4 To this point, we have been manipulating amino acid sequences. Now, reverse-translate the peptide sequences into DNA sequences using the `revtrans` command. This command randomly chooses codons according to the *E. coli* frequency table, dropping any codons that are more rare than a given frequency threshold. We exclusively use the amber stop codon. Add `prefix/suffix` sequences that will be used for PCR/cloning. Finally, search for any restriction sites that will be used for cloning within the coding sequence and recode them as necessary. The final oligonucleotides are presented in `oligos.fasta`.

```
PREFIX=AGGAATTCCGCTGCGT
SUFFIX=GCCTGGAGACGCCATC
PREFIXLEN=${#PREFIX}
SUFFIXLEN=${#SUFFIX}
FREQTHRESH=0.01
cat protein_tiles.fasta \
  | pepsyn revtrans --codon-freq-threshold $FREQTHRESH --amber-only
  - - \
  | pepsyn prefix -p $PREFIX - - \
  | pepsyn suffix -s $SUFFIX - - \
  | pepsyn recodesite --site EcoRI -site HindIII --clip-left $PRE-
  FIXLEN \
  --clip-right $SUFFIXLEN --codon-freq-threshold $FREQTHRESH \
  --amber-only - - \
>oligos.fasta
```

? TROUBLESHOOTING

5 Finally, verify that the library is free of any EcoRI or HindIII sites, using the following command:

```
pepsyn findsite --site EcoRI --clip-left 3 oligos.fasta
pepsyn findsite --site HindIII oligos.fasta
```

6 Generate a Bowtie index now, in preparation for aligning of sequencing data later. Generate a reference .fasta file that contains just the DNA tiles without the adaptors, using the following command:

```
  pepsyn clip --left $PREFIXLEN --right $SUFFIXLEN oligos.fasta oligos-
ref.fasta
```

Then create the Bowtie index (or index for whichever aligner you prefer, such as bwa, Bowtie2, or kallisto) called 'mylibrary' as follows:

```
bowtie-build -q oligos-ref.fasta bowtie_index/mylibrary
```

Construction and expansion of the phage screening library ● Timing 3 weeks

7 Send the oligos.fasta file to a DNA synthesis company for manufacture of the oligonucleotide library.
■ PAUSE POINT The oligonucleotide library should be divided into aliquots and stored frozen at −80 °C indefinitely.

8 Perform library PCR using primers with cloning sites and binding sites for the PREFIX/SUFFIX sequences appended to the oligonucleotide library. Follow the procedures for standard cloning of PCR product into bacteriophage for display using published protocols (e.g., that in the Novagen T7Select System Manual)[1,5].
▲ CRITICAL STEP We recommend centrifuging the expanded library for 2 h at 4 °C at 3,000*g* and carefully moving the supernatant to a new container before freezing, in addition to the centrifugation specified in Step 28.
▲ CRITICAL STEP The quality of each new phage library should be assessed by Sanger sequencing of 20 or more randomly selected plaques (to confirm the fidelity of the oligonucleotide synthesis) and by Illumina sequencing to assess the distribution of the clonal frequency (Experimental design).
■ PAUSE POINT The expanded library should be divided into aliquots and stored in 10% (vol/vol) dimethyl sulfoxide (DMSO) at −80 °C indefinitely until used.

Serum IgG quantification by ELISA ● Timing 1 d

9 If working with more than a few samples, randomly assign each sample to a position on a 96-well plate. This will reduce the potential for positional artifacts. Dilute each sample 1:100 in PBS, to a final volume of 200 µl, in a non-tissue-culture-treated round-bottom 96-well plate.
! CAUTION Whenever working with human serum, wear a one-time-use splash shield, a mask, a biohazard suit, and surgical gloves. Dispose of the protective equipment after single use.

10 Dilute the unlabeled IgG capture antibody to a final concentration of 2 µg/ml in ELISA coating buffer and add 50 µl to each well of an enhanced-binding ELISA plate.

11 Wrap the ELISA plate with Saran wrap and incubate on a flat surface at 4 °C overnight.

12 Splash out the capture antibody-coating solution and wash the ELISA plate three times with 150 µl of ELISA wash buffer.

13 Block the ELISA plate by adding 150 µl of ELISA blocking buffer to each well.

14 Wrap the ELISA plate with Saran wrap and incubate on flat surface at 37 °C for at least 1 h.

15 Wash the plate three times with 150 µl of ELISA wash buffer. Just before the addition of samples and standards, blot the plate against a clean paper towel.

16 Dilute the human IgG (hIgG) standard to 100 ng/ml in ELISA blocking buffer and perform six 1:3 serial dilutions in ELISA blocking buffer. At the same time that the samples are added to the ELISA plate (Step 17), add 50 µl of diluted standard per well to the empty wells reserved for use as negative controls. We typically include eight such wells per 96-well plate, one of which serves as a blank and contains blocking buffer only.

17 For total human IgG quantitation, serially dilute sera 1:100 two additional times (for a final dilution of 1:1,000,000) in ELISA blocking buffer. Add 50 µl of the diluted sample to each well of the ELISA plate at the same time as the IgG standards (Step 16).

18 Wrap the ELISA plate with Saran wrap and incubate on flat surface at 37 °C for 1 h.

19 Wash the plate five times with 150 µl of ELISA wash buffer. Just before the addition of detection antibody, blot the plate against a clean paper towel.

20 Dilute the hIgG-HRP detection antibody in the ratio of 1:5,000 in ELISA blocking buffer. Add 50 µl to each well of the ELISA plate.

21 Wrap the ELISA plate with Saran wrap and incubate on flat surface at room temperature for at least 30 min.

22 Wash the plate five times with 150 µl of ELISA wash buffer and blot against a clean paper towel.

23 Add 50 µl of TMB substrate to each well. Incubate from 3 to 30 min at room temperature until the color becomes somewhat dark in the highest IgG standard well, but is not yet colored in the blank standard well.

? TROUBLESHOOTING

24 Add 50 µl of stop solution to each well in the same order as the TMB substrate and at the same speed.

25 Read the optical absorbance for each well at 450 nm, using a plate reader.

26 Prepare an *xy* table in a graphing program and generate a standard curve graph. Interpolate the sample's *x* values by non-linear regression analysis, using a 'one-site binding' model. To do this, the net OD values for all serum samples are obtained by subtracting the OD of a blank well from their original OD values. Then these OD values are analyzed using Prism software with the single site binding saturation regression analysis to estimate the serum IgG concentration (*x* values in nanograms per milliliter) of each sample for the 1:1,000,000 dilution. Multiply the *x* values (nanograms per milliliter) by 10,000 to obtain the sample concentration in the 1:100 dilution plate (created in Step 9). Calculate the volume of the 1:100 sample dilution that will contain 2 µg of IgG.

? TROUBLESHOOTING

■ **PAUSE POINT** After IgG quantification, diluted human serum samples from Step 9 can be stored refrigerated at 4 °C until proceeding to immunoprecipitation, but not for longer than a couple of days.

Peptide–antibody complex formation and immunoprecipitation ● Timing 2 d

▲ **CRITICAL** Steps 27–41 can be performed in single 1.5-ml Eppendorf tubes for smaller numbers of samples or in 96-well plate format for larger numbers of samples. Here, we refer only to the 96-well plate format. If performing screens in 1.5-ml Eppendorf tubes, the magnetic particle concentrator can be used rather than the 96-well plate magnet.

27 Thaw the phage library or libraries from Step 8 and combine them in a large enough vessel for all screens that will be performed in the current run. We recommend peptide–antibody complex formation volumes of 1 ml and an input of 10^5 PFU per individual phage-library member. After addition of the phage library (or libraries), make up the remaining volume by adding PBS. Optionally, add immunoprecipitation (IP) spike-ins (e.g., control antibodies and/or control phage clones) at this time.

28 Centrifuge the phage library mixture for 2 h at 4 °C at 3,000*g* and carefully move the supernatant to a new container, being careful not to disturb any pelleted material (even if there is no visible pellet). Pelleted material may include cell debris or precipitate that may interfere with the assay and so should be discarded.

29 Mix the phage very well by pipetting up and down with a serological pipette and then distribute 1 ml to each well of a 2-ml deep-well plate.

30 Add 2 µg of serum IgG from the 1:100 dilution in PBS (volume calculated in Step 26) to the corresponding wells of the deep-well plate containing the phage mix. We suggest running multiple negative controls without antibody so that antibody-dependent enrichments may be quantified by comparison. Optionally, if screening a large number of samples, this step is best automated to avoid error.

▲ **CRITICAL STEP** It is important that the amount of input antibody be below the binding capacity of the magnetic beads. If in excess, soluble antibody will compete with bound antibody for specific interactions with target phage, reducing enrichment efficiency and thus sensitivity.

31 Rotate the mixtures end-over-end in a cold room at 4 °C for ~18 h. If screening in 96-well plate format, wells must be tightly sealed (e.g., with a 96-well silicone mat gasket seal) to avoid cross-contamination.

▲ **CRITICAL STEP** We have alternatively performed this step at 37 °C for 1 h with roughly similar results.

32 Centrifuge IP mixtures at 1,000*g* for 1 min at room temperature to remove the liquid from the gasket seal.

33 Wash 20 µl of protein A- and 20 µl of protein G-coated magnetic beads per IP three times in bead wash buffer and resuspend in the same volume of 1× PBS.

34 For capture of human IgG, add 20 µl of pre-washed protein A-coated Dynabeads and 20 µl of pre-washed protein G-coated Dynabeads to each tube or well.

35 Again rotate the mixtures end-over-end in a cold room for ~4 h. If screening in 96-well-plate format, wells must be tightly sealed (e.g., with a silicone mat gasket seal) to avoid cross-contamination. If peptide–antibody complex formation was performed at 37 °C, this step can also be performed at 37 °C for 30 min during end-over-end rotation.

36 Centrifuge the mixtures at 900*g* for 2 min at room temperature to remove volume from the gasket seal (and to pellet beads, if necessary, given that the geometry of many magnets would not efficiently pellet beads in deep wells). (Optional) Steps 37–41 can be automated. We have implemented the bead washing steps on the BioMek FX and the Agilent Bravo liquid-handling robots with similar results.

37 Remove and discard the supernatant from the pelleted beads, but leave ~100 µl in each well. Use this volume to resuspend the beads and transfer them to a full-skirted PCR plate.

38 Place the plate on a 96-well magnet and allow the beads to collect. Remove as much supernatant as possible without aspirating the beads.

39 Immediately resuspend the beads in 170 µl of IP wash buffer. If using a liquid-handling robot, bead resuspension is best accomplished by combined pipetting and light vortexing until the bead suspension is uniform. This typically requires ~20 cycles of automated pipetting or about ten cycles of pipetting by hand.

▲ **CRITICAL STEP** Strong vortexing may shear the immunoprecipitated phage particles off the beads.

40 Repeat Steps 38 and 39 once more for a total of two bead washes. Performing additional washes or raising the salt concentration can increase the wash stringency as desired for specific projects.

▲ **CRITICAL STEP** It is important that the DNA polymerase used for PCR1 be insensitive to residual detergent in the wash buffer. If a sensitive polymerase must be used, a final bead wash lacking detergent should be performed.

41 Repeat Step 38 once more so that only the collected beads remain in the wells.

■ **PAUSE POINT** Beads can now be stored frozen (−20 to −80 °C) indefinitely until proceeding to PCR1.

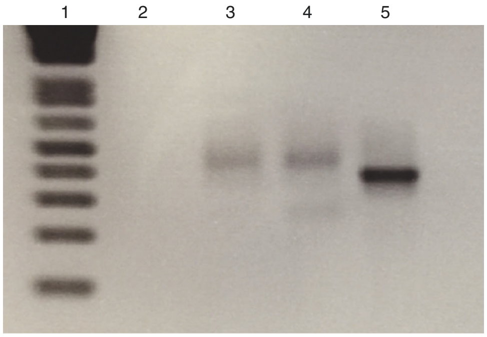

Fig. 3 | Primer-depleted, pooled VirScan PCR2 products run at a higher molecular weight than expected (shown on a 2% (wt/vol) agarose gel in lithium borate). Lane 1: 1Kb Plus DNA Ladder (Invitrogen). Lane 2: empty. Lane 3: primer-depleted VirScan PCR2 product. Lane 4: product of VirScan PCR3 without replenishment of primers. Lane 5: product of VirScan PCR3 with replenishment of primers (P5 and P7.2).

Preparation of peptidome library DNA for sequencing ● Timing 1 d

42 Make enough 1× PCR1 master mix for one 20-µl reaction (19-µl reaction plus ~1 µl of residual bead wash volume) per IP as follows:

Component	Volume (µl)	Final concentration
H_2O	14.5	
5× Herculase buffer	4	1×
dNTPs	0.2	1 mM
T7-Pep2_PCR1_F	0.05	0.25 µM
T7-Pep2_PCR1_R+ad_min	0.05	0.25 µM
Herculase II polymerase	0.2	
Total	19	

▲ **CRITICAL STEP** In addition to the sample and mock IPs, devote at least one PCR1 reaction to sequencing of the input library. To this end, use 1 µl of the input phage library as the template for PCR1. This will generate the set of input library read counts used later in the analysis.

43 If frozen, remove the IP plate or tubes from Step 41 from the freezer and allow the beads to come to room temperature. Resuspend the beads (~1 µl, the residual bead wash volume) in 19 µl of PCR1 master mix from Step 42 and transfer the mixture to a full-skirted PCR plate.

44 Perform thermocycling as follows:

Cycle number	Denature	Anneal	Extend
1	95°C, 2 min		
2–21	95°C, 20 s	58°C, 30 s	72°C, 30 s
22			72°C, 3 min

■ **PAUSE POINT** Either proceed immediately to PCR2 or store PCR1 reactions at −80 °C indefinitely.

45 Make enough PCR2 master mix for one 20-µl reaction (once primers and template have been added in Step 48) per IP, as follows:

Component	Volume (µl)	Final concentration
H_2O	8.55	
5 × Herculase buffer	4	1×
dNTPs	0.2	1 mM
T7-Pep2_PCR2_F_P5	0.05	0.25 µM
Herculase II	0.2	
Total	13	

46 Distribute 13 µl of PCR2 master mix to each well of a full-skirted 96-well plate.

47 If frozen, thaw PCR1 product from Step 44 and keep on ice.

48 To each 13 µl of PCR2 master mix from Step 46, add 2 µl of PCR1 from Step 44 or 47, and 5 µl of the appropriate ad_min_BCX_P7 barcoding reverse primer (from a 1 µM stock concentration). Mix well.

49 Perform thermocycling as follows:

Cycle number	Denature	Anneal	Extend
1	95 °C, 2 min		
2–21	95 °C, 20 s	58 °C, 30 s	72 °C, 30 s
22			72 °C, 3 min

Each DNA library now contains a unique, sample-specifying DNA barcode (or 'index').

■ **PAUSE POINT** Stop and store PCR2 product at −80 °C indefinitely.

50 Pool the barcoded PCR2 products. If PCR2 proceeds to primer depletion (as evidenced by a laddering effect observed on a 2% (wt/vol) agarose gel), then it can safely be assumed that the amount of PCR2 product will be relatively uniform across all samples. In this case, mix the same volume of PCR2 from each sample (e.g., 5 µl). If, however, the amount of PCR2 is expected to be substantially different between samples, one might want to normalize each sample's representation in the final pool, so as to ensure uniform sequencing depth of each library. For accurate quantification of PCR2 products (before or after pooling), we recommend using the KAPA Library Quantification Kit according to the manufacturer's instructions.

? TROUBLESHOOTING

51 Perform PCR2 product cleanup using DNA Clean & Concentrator columns from Zymo Research according to the manufacturer's instructions. Repeat this step to clean up the PCR2 product a second time.

52 If gel purification is desired before deep sequencing, perform a PCR3 (a single PCR cycle with replenished primer) to produce a single clear band on the gel. To do so, prepare enough PCR3 master mix for ten 20-µl reactions as follows:

Component	Volume (µl)	Final concentration
H_2O	12.6	
5 × Herculase buffer	4	1×
dNTPs	0.2	1 mM
P5	0.5	1 µM
P7.2	0.5	1 µM
Herculase II	0.2	
Template (pooled, column-purified PCR2 product from Step 50)	200 ng	
Total	20	

53 Perform thermocycling as follows:

Cycle number	Denature	Anneal	Extend
1	95 °C, 2 min		
2	95 °C, 20 s	58 °C, 30 s	72 °C, 30 s
3			72 °C, 3 min

54 All library DNA should now be non-laddered double-stranded DNA (dsDNA), which will appear as a single sharp band on a 2% (wt/vol) agarose gel (Fig. 3). Extract the PCR3 product from the gel using the NucleoSpin Gel and PCR Clean-up Kit from Macherey-Nagel according to the manufacturer's instructions. If multiple peptidome libraries are being analyzed simultaneously (e.g., the human peptidome and the human virome), it may be desirable to sequence them separately or to differing depths. In this case, gel electrophoresis of PCR3 may separate different-sized libraries. PCR3 products can then be separately isolated and quantified before mixing together in a ratio that will determine their relative sequencing depth. Solid-phase reversible immobilization (SPRI) beads could be used as an alternative to column purification; however, this is not a method we have tested.

? TROUBLESHOOTING

55 Submit the purified PCR2 or PCR3 libraries to a core facility for quantification and deep sequencing. Be sure to provide the custom sequencing primer (e.g., T7-VirScan_SP). For libraries that lack diversity in the first several bases downstream of the sequencing primer (because of adapter sequence, for example), it may be necessary to spike in a base-balanced library, such as the PhiX standard control that is used by Illumina for quality control (at up to a 30% molar ratio). The library can be sequenced using a 50-cycle, single-end protocol. It is essential that the sequencing run include the i7 index read ('Read 2') and that it be of sufficient length (we use 8-nt barcode sequences) to link the sample identity with each peptidome library sequence (Fig. 4).

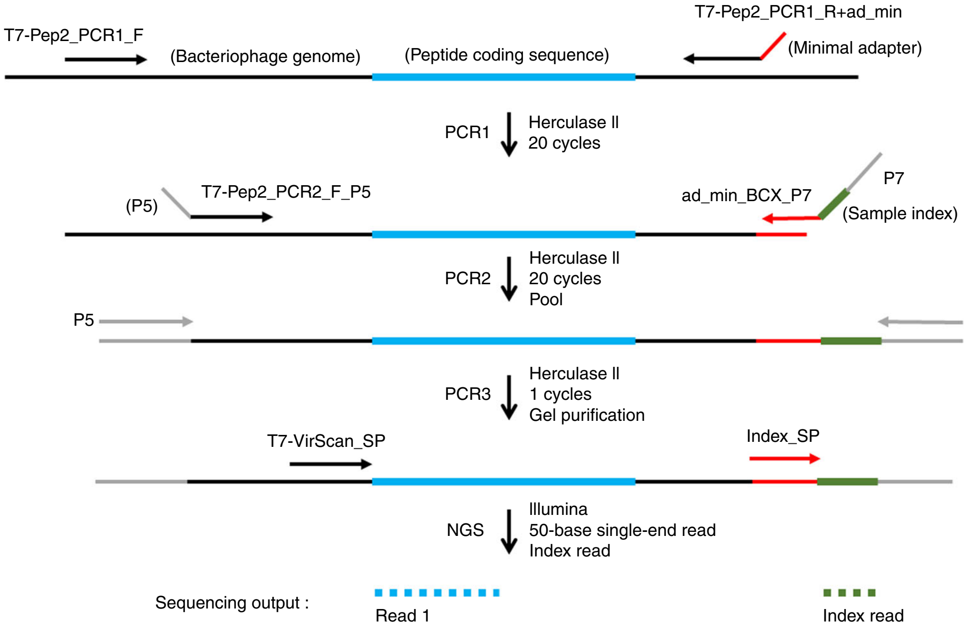

Fig. 4 | Organization of bacteriophage genome, primer binding sites, and PCR products. The peptide coding sequence, originally derived from the oligonucleotide library, is cloned into the T7 genome as a C-terminal fusion with the 10B capsid protein. PCR1 (Steps 42–44): T7-Pep2_PCR1_F is used as the outside PCR1 primer; T7-Pep2_PCR1_R+ad_min is used as the reverse PCR1 primer and to add the minimal adapter required for subsequent addition of the sample barcode during PCR2. PCR2 (Steps 45–49): the product of PCR1 is used as the template for PCR2. T7-Pep2_PCR2_F_P5 is used as the forward PCR2 primer and to add the required Illumina P5 adapter (and, optionally, the i5 dual index (not shown)). The set of primers called ad_min_BCX_P7 (where X represents the sample-specific DNA barcode) are used individually as the reverse PCR2 primers, to add the sample-specific DNA barcode, and to add the required Illumina P7 adapter. After pooling PCR2 products from all the samples, a single round of PCR3 is performed (Steps 52 and 53) using the P5 and P7.2 primers, which ensures the DNA libraries are fully double-stranded. Illumina sequencing (Step 55): T7-VirScan_SP is the read 1 sequencing primer used to obtain the peptide coding sequence. Index_SP is the standard Illumina Multiplex Single End read 2 sequencing primer used to obtain the sample-specific barcode. The sequences generated from the Illumina sequencing run are shown as dashed lines: read 1 obtains the first 50 bases of the peptide coding sequence, and read 2 is the eight-cycle index read.

Processing the raw PhIP-Seq data ● Timing several hours

56 Separately align each sample's. fastq file using the Bowtie short read aligner, with the following command:

```
mkdir -p workdir/alns
bowtie -n 3 -l 100 --best --nomaqround --norc -k 1 -p 4 --quiet \
bowtie_index/mylibrary workdir/reads/sample1.fastq.gz \
workdir/alns/sample1.aln
```

! CAUTION Take care to correctly specify the path to the Bowtie index. If the Bowtie index is called `index/mylibrary.1.ebwt` (along with the additional files), then you should specify `index/mylibrary`. Note that the backslashes above mean line continuation.

▲ CRITICAL STEP See Bowtie's documentation (http://www.bowtie-bio.sourceforge.net) for additional alignment options. With many samples, the commands can be submitted to a batch job scheduler such as LSF, Grid Engine, or SLURM, which are commonly available in scientific computing environments.

▲ CRITICAL STEP This step is computationally expensive, and we recommend submitting a job for each file to a batch system on a cluster.

? TROUBLESHOOTING

57 Aggregate each alignment file into a sample count vector using the following command:

```
phip compute-counts -i workdir/alns -o workdir/counts \
-r path/to/input/counts.tsv
```

The command-line flags are: `-i`, input directory; `-o`, output directory; and -r, reference file containing the input counts for the library. Each sample count file generated by the command above will contain one column for the input counts (specified with `-r`) and another column for the sample counts (specified with `-i`). Therefore, this step requires the input counts generated in Step 42. Alternatively (and if input counts are not available), aggregated counts from negative-control samples can also be used with the `-r` flag. The input counts are necessary for the statistical model used to compute the enrichment scores. As this current step is computationally inexpensive, it is performed locally.

? TROUBLESHOOTING

58 Generate ($-\log_{10}$) P values from the counts by fitting a generalized Poisson model and computing a significance score for each pair of count values. Specifically, we model the count value Y_i for peptide i as

$$Y_i \sim \mathrm{GeneralizedPoisson}(\lambda(X_i), \theta(X_i)), \quad (1)$$

in which the functions $\lambda(x) = a\,x + b$ and $\theta(x) = c$ are fit empirically to the observed data. For each possible input value x, we compute the maximum likelihood estimates for λ and θ using the counts of all peptides with x reads, and regress the λ and θ values against the input counts to get estimated λ and θ as functions of x. The scores can be generated by running the following command:

```
phip compute-pvals -i workdir/counts/sample1.tsv \
-o workdir/mlxp/sample1.mlxp.tsv
```

Here, `-i` is a file containing sample counts and `-o` is the destination file containing the MLXP values (Note: 'mlxp' is short for 'minus log10 p-val').

▲CRITICAL STEP This step is computationally expensive, and we recommend submitting a job for each file to a batch system on a cluster.

? TROUBLESHOOTING

59 Alternatively, merge the count values into a single tab-delimited file to make it easier to analyze as a single matrix with the following command:

```
phip merge-columns -i workdir/mlxp -o mlxp.tsv -p 1
```

Here, `-i` is a directory containing MLXP files and `-o` points to the merged MLXP file containing the full matrix.

This will merge the second column (zero-indexed) of each file together; it assumes that the first column is the join key. This step can also be parallelized on a batch scheduler as is described in the alignment step.

60 Load the resulting tab-delimited file into Python or R as a dataframe for further analysis (e.g., the Python pandas library (https://pandas.pydata.org/) or the R tidyverse (https://www.tidyverse.org/). In Python, the command would be:

```
import pandas as pd
 df = pd.read_csv('mlxp.tsv', sep='\t', header=0)
```

Troubleshooting

Troubleshooting advice can be found in Table 2.

Table 2 | Troubleshooting table

Step	Problem	Possible Reason	Solution
4	Oligo library sequences not generated	Illegal characters in header or protein sequence	Remove illegal characters
23	Low ELISA signal	Poor binding of capture antibody; TMB expired	Make sure proper ELISA plates are being used and that the TMB is not expired

Table continued

Table 2 (continued)

Step	Problem	Possible Reason	Solution
26	ELISA data not within the dynamic range	TMB was developed either for too long or not long enough	Increase or reduce the TMB development time
50, 54	No PCR product	Incorrect primers were used; incorrect thermocycling conditions were used; dNTPs were expired	Carefully repeat with fresh reagents
56	Crash or program freeze	This step essentially rewrites the entire dataset and thus may use more disk space than is available	Ensure your file system has enough disk space
57	Too much of library is missing	Library was bottlenecked during construction or became too skewed during expansion	Reconstruct the library
58	Extreme, unreproducible enrichments	Contamination by host bacterial cells	Use more stringent centrifugation to remove the cells; add antibiotic to prevent growth
	Unreproducible enrichments	Sample cross-contamination. This can usually be determined by examining the overlapping hits between samples	Avoid antibody cross-contamination, PCR1 cross-contamination, and PCR2 barcode cross-contamination

Timing

Steps 1–6, peptide library design: 1 d
Steps 7 and 8, construction and expansion of the phage screening library: 3 weeks
Steps 9–26, IgG quantification by ELISA: 1 d
Steps 27–41, antibody binding and immunoprecipitation: 2 d
Steps 42–55, DNA sequencing library preparation: 1 d
Steps 56–60, PhIP-Seq data processing: several hours

Anticipated results

The results of any PhIP-Seq experiment completely depend on the samples and libraries used for the analysis. We have observed that the number of both autoantibody and viral antibody specificities increases with the age of the donor. Nearly all human serum samples we have analyzed contain antibodies to rhinovirus A peptides, and most adults harbor antibodies that bind several Epstein–Barr peptides. Known autoantibodies can be detected with variable success. For example, TRIM21 ('Ro52') antibodies can be detected by PhIP-Seq in ~90% of Sjögren's syndrome patients who are seropositive by the clinical ELISA assay, whereas we tend not to detect clinically confirmed anti-insulin antibodies present in type 1 diabetics. PhIP-Seq is in many cases less sensitive than optimized single-plex assays but has the advantage of being much more comprehensive in assessing antibody-binding specificities.

The sections below provide additional details about key results generated in the course of a PhIP-Seq project.

Output of the peptide library-design software

The ultimate output of the library design phase is a .fasta file containing oligo sequences (Step 4) that will be sent out for oligonucleotide synthesis. The design phase can use multiple pepsyn tools in a pipeline, and we recommend inspecting the results of intermediate steps to ensure that they correspond to the expected results. For example, are all of the expected ORF sequences represented? Are the peptides/oligos of the expected length? Are the numbers of oligos allowable, given your budget? It is critical to check the final library design, as synthesis is costly and will waste time if it has to be repeated. It is especially important to ensure that the length of the resulting oligos is as expected and to test whether the designed oligos contain restriction sites that may pose a problem during cloning (Step 5).

Quality control of the input phage library

After deep sequencing of the input library and read count tabulation (Step 57), ideally, 90% of the library should fall within one log of clonal frequency. If a sequencing depth of ≥10 reads per library member is achieved, >90% of the library should be sequenced at least once.

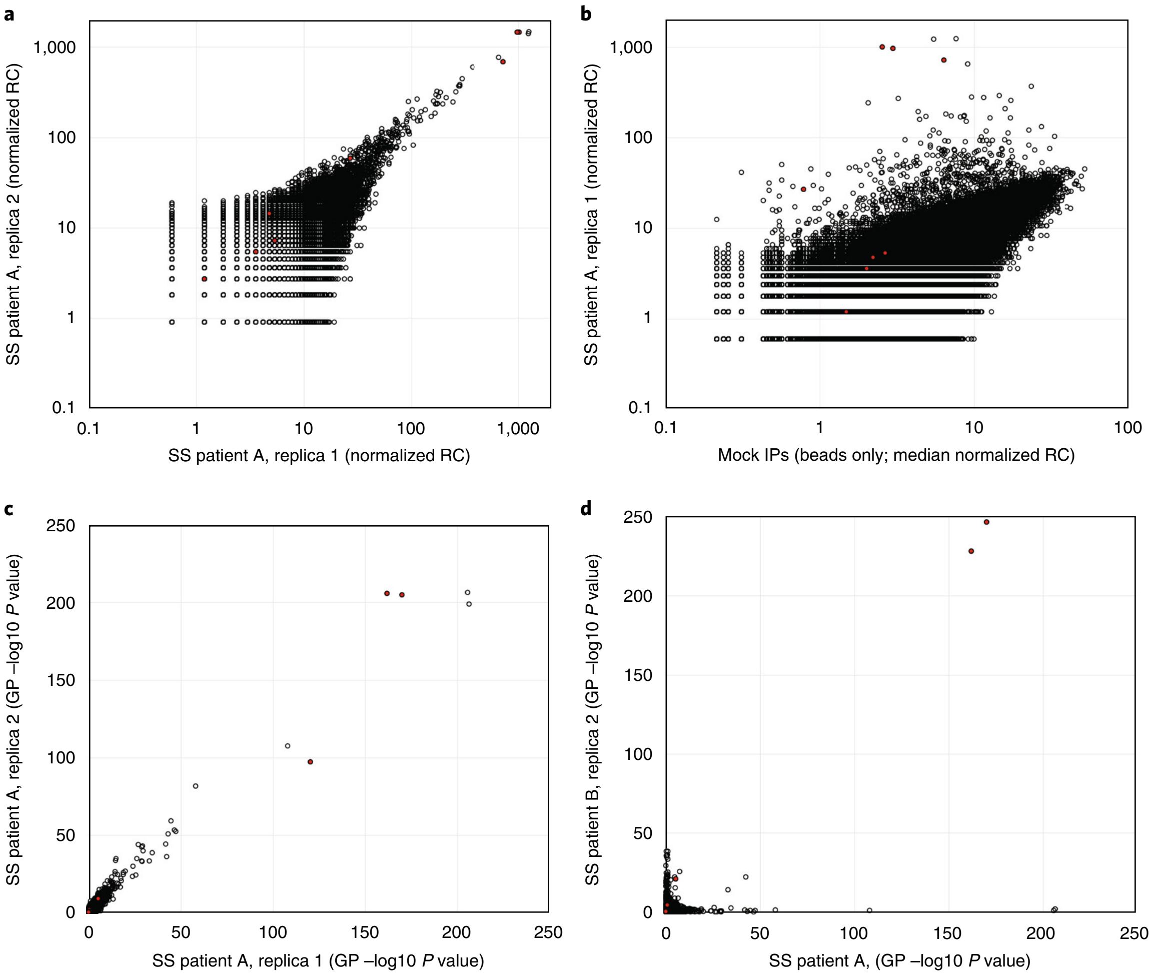

Fig. 5 | Output from the sequencing data analysis pipeline. a, Sjögren's syndrome (SS) patient A's serum sample was screened against the human peptidome library and analyzed in duplicate. Read counts were divided by the total reads, multiplied by 1 × 10^6, and then plotted. The scatter plot illustrates the reproducibility of the post-immunoprecipitation clonal distributions between the two replicas. Red-filled circles highlight peptides from the Ro52 (TRIM21) protein, to which this patient was known to have autoantibodies. **b**, Comparison of patient A's immunoprecipitated clonal distribution to that of a set of mock IPs (no sample input), which illustrates (i) the bias in the starting library and (ii) antibody-dependent enrichment of specific phage clones (including strong enrichment of three Ro52 peptides). **c**, Generalized Poisson (GP)-based *P* values calculated using the data in **a** as input. Background bias has been removed from this distribution, which illustrates reproducible antibody-dependent enrichments. **d**, Comparison of the enrichment scores ($-\log_{10}$ *P* values) of two different individuals illustrates their largely non-overlapping enrichment profiles. However, three peptides from Ro52 are among the shared enrichments. De-identified serum samples were analyzed in accordance with JHU Human Subject Research exemption IRB00049327. RC, read counts.

IgG ELISA data

The standard curve data (Step 26) should be visually inspected along with the data from the samples to ensure that sample measurements are within the dynamic range of the assay (e.g., substantially above background and below saturation of the standard curve). For typical serum samples, IgG concentration should be ~10 µg/µl. However, samples stored frozen for a long time will tend to concentrate because of sublimation, as compared with fresh serum. The volume of the 1:100 diluted serum samples required for 2 µg should therefore be ~20 µl but may be substantially more or less. If the calculation is substantially different than expected, however, there may be a problem with the ELISA, the standards or the calculations.

Amplification of the sequencing libraries

The 20 cycles recommended for PCR1 (Step 44) do not produce high-concentration amplicon. Nevertheless, a weak PCR1 band (507 nt for VirScan) can usually be visualized on an agarose gel after extended exposure. Fewer than the recommended 20 cycles of PCR2 are sufficient to produce enough library for sequencing. However, we typically perform 20 cycles of PCR2 (Step 45-49) to ensure complete primer depletion, and thus equimolar amplification yield, among all samples. There is thus no need to separately quantify the PCR2 amplicons from each reaction before pooling. Such 'over-amplified' libraries, however, run as non-fully double-stranded DNA (dsDNA) structures (including pseudoconcatemers) on agarose gels. Gel extraction thus requires a single round of primer-replenished PCR3 (Step 53), which produces the fully dsDNA product at the expected molecular weight (Fig. 3).

Sequencing data processing

You should expect to successfully align at least 70% of your raw reads (Step 56). A lower alignment rate can indicate the use of the wrong reference file, poor sequencing quality, or a high rate of synthesis error in your oligonucleotide library.

Enrichment analysis

It is important to run several types of control samples, especially during the initial establishment of the PhIP-Seq platform. Negative controls (no antibody input, 'mock IPs') should be run alongside samples in every experiment. We typically reserve four to eight wells on a 96-well plate for such controls. These data should reveal relatively few enriched peptides; reproducible enrichments may reflect peptide-dependent 'background' binding to the beads. Extreme, unreproducible enrichments in these negative controls may indicate contamination of the phage library with host bacterial cells. Replicate IPs should be quantitatively and visually compared for high concordance.

Figure 5 illustrates the analysis of sample data obtained by screening the 90-mer human peptidome library against two Sjögren's syndrome patients (in duplicate). Data normalization and background bias removal using the generalized Poisson model provide antibody-dependent *P* values of enrichment for each peptide (Step 58). These enrichments are reproducible and largely patient-specific. However, peptides from the Ro52 antigen are strongly enriched by both patients.

References

1. Larman, H. B. et al. Autoantigen discovery with a synthetic human peptidome. *Nat. Biotechnol.* **29**, 535–541 (2011).
2. Larman, H. B. et al. PhIP-Seq characterization of autoantibodies from patients with multiple sclerosis, type 1 diabetes and rheumatoid arthritis. *J. Autoimmun.* **43**, 1–9 (2013).
3. Larman, H. B. et al. Cytosolic 5′-nucleotidase 1A autoimmunity in sporadic inclusion body myositis. *Ann. Neurol.* **73**, 408–418 (2013).
4. Finton, K. A. et al. Ontogeny of recognition specificity and functionality for the broadly neutralizing anti-HIV antibody 4E10. *PLoS Pathog.* **10**, e1004403 (2014).
5. Xu, G. J. et al. Viral immunology. Comprehensive serological profiling of human populations using a synthetic human virome. *Science* **348**, aaa0698 (2015).
6. Kosuri, S. & Church, G. M. Large-scale de novo DNA synthesis: technologies and applications. *Nat. Methods* **11**, 499–507 (2014).
7. Atak, A. et al. Protein microarray applications: autoantibody detection and posttranslational modification. *Proteomics* **16**, 2557–2569 (2016).
8. Yu, X. et al. Multiplexed nucleic acid programmable protein arrays. *Theranostics* **7**, 4057–4070 (2017).
9. Henkel, S., Wellhausen, R., Woitalla, D., Marcus, K. & May, C. Epitope mapping using peptide microarray in autoantibody profiling. *Methods Mol. Biol.* **1368**, 209–224 (2016).
10. Finton, K. A. et al. Autoreactivity and exceptional CDR plasticity (but not unusual polyspecificity) hinder elicitation of the anti-HIV antibody 4E10. *PLoS Pathog.* **9**, e1003639 (2013).
11. Xu, G.J. et al. Systematic autoantigen analysis identifies a distinct subtype of scleroderma with coincident cancer. *Proc. Natl. Acad. Sci. USA* **113**, E7526-E7534 (2016).
12. Zhu, H., Luo, H., Yan, M., Zuo, X. & Li, Q. Z. Autoantigen microarray for high-throughput autoantibody profiling in systemic lupus erythematosus. *Genomics Proteomics Bioinformatics* **13**, 210–218 (2015).
13. Miersch, S. & LaBaer, J. Nucleic acid programmable protein arrays: versatile tools for array-based functional protein studies. *Curr. Protoc. Protein Sci.* Chapter 27, Unit 27.2 (2011).
14. Zhu, J. et al. Protein interaction discovery using parallel analysis of translated ORFs (PLATO). *Nat. Biotechnol.* **31**, 331–334 (2013).
15. Larman, H. B., Liang, A. C., Elledge, S. J. & Zhu, J. Discovery of protein interactions using parallel analysis of translated ORFs (PLATO). *Nat. Protoc.* **9**, 90–103 (2014).
16. Jhaveri, D. T. et al. Using quantitative seroproteomics to identify antibody biomarkers in pancreatic cancer. *Cancer Immunol. Res.* **4**, 225–233 (2016).
17. MacConaill, L. E. et al. Unique, dual-indexed sequencing adapters with UMIs effectively eliminate index cross-talk and significantly improve sensitivity of massively parallel sequencing. *BMC Genomics* **19**, 30 (2018).

Acknowledgements

The development of the PhIP-Seq technology platform has been an ongoing collaboration with S.J. Elledge of the Harvard Medical School Genetics Department and the Howard Hughes Medical Institute. Special thanks to T.M. Shi (Department of Art as Applied to Medicine, Johns Hopkins School of Medicine) for creating the artwork used in Fig. 1. Recent improvements in the PhIP-Seq methodology were supported under a U24 Resource-Related Research Projects Cooperative Agreement awarded by the NIH (U24AI118633 to H.B.L. and S. J. Elledge), a grant from the Jerome L. Greene Foundation (to H.B.L. and A.N.B.), and a grant from the Sjögren's Syndrome Foundation (to H.B.L. and A.N.B.). A.N.B. received support from NIH grant R01DE012354 (Rosen, PI).

Author contributions

D.M., D.L.W., and B.M.S. performed experiments related to assay development and optimization. D.M. performed PhIP-Seq screening analysis of the serum samples used in this study. M.S.N. created a draft version of the peptidome design software. A.N.B. provided the Sjogren's syndrome serum samples and disease-specific expertise. U.L. developed the pepsyn and phip-stat software packages. U.L. and H. B.L. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41596-018-0025-6.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to U.L.@.HB.L.1@.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Published online: 6 September 2018

Related links

Key references using this protocol

1. Larman, H. B. et al. *Nat. Biotech.* **29**, 535–541 (2011) https://doi.org/10.1038/nbt.1856
2. Larman, H. B. et al. *Ann. Neurol.* **73**, 408–418 (2013) https://doi.org/10.1002/ana.23840
3. Xu, G. J. et al. *Science* **384**, aaa0698 (2015) https://doi.org/10.1126/science.aaa0698