**Chapter Title: Next-generation phage display to identify peptide ligands of deubiquitinases**

Anastasios Spiliotopoulos [1,2,4], Sigrun K. Maurer [1], Maria T. Tsoumpeli [2], Juan A. F. Bonfante [2], Jonathan P. Owen [3], Kevin C. Gough (ORCID ID: 0000-0001-5211-2088) [2]* and Ingrid Dreveny (ORCID ID: 0000-0003-2256-642X) [1]*

[1] Biodiscovery Institute, School of Pharmacy, University of Nottingham, Nottingham NG7 2RD, United Kingdom; [2] School of Veterinary Medicine and Science, Sutton Bonington Campus, College Road, Sutton Bonington, Leicestershire, LE12 5RD, UK; [3] ADAS UK, School of Veterinary Medicine and Science, The University of Nottingham, Sutton Bonington, Loughborough, Leicestershire LE12 5RD, UK. [4] Current address: Vertex Pharmaceuticals. * Corresponding authors.

**Abstract**

Phage display (PD) is a powerful method and has been extensively used to generate monoclonal antibodies and identify epitopes, mimotopes, and protein interactions. More recently, the combination of next-generation sequencing (NGS) with PD has revolutionized the capabilities of the method by creating large data sets of sequences from affinity selection-based approaches (biopanning) otherwise challenging to obtain. NGPD can monitor motif enrichment, allow tracking of the selection process over consecutive rounds, and can highlight unspecific binders. To tackle the wealth of data obtained, bioinformatics tools have been developed that allow for identifying specific binding sequences binders that can then be validated. Here, we provide a detailed account of the use of NGPD experiments to identify ubiquitin-specific protease peptide ligands.

**Key Words:** phage display; next -generation sequencing; biopanning; bioinformatical analysis, deubiquitinase, peptides

**Running head:** Deubiquitinase peptide ligand identification

## 1. Introduction

The human genome encodes 56 ubiquitin-specific proteases (USPs), the largest family of deubiquitinating enzymes (DUBs) that are crucial regulators of ubiquitin-mediated proteasomal degradation and virtually all cellular pathways, including processes as diverse as signaling, immune responses to pathogens, DNA damage repair, receptor trafficking and gene regulation (1). Critical to USP function is a network of protein-protein interactions that determines USP specificity, regulation of the catalytic activity, and recruitment to target substrates. Large-scale proteomics and more focused studies have identified numerously confirmed and putative USP binding partners (2). Linear and short-sequence motif recognition is emerging as a theme for transient interactions central to these signaling pathways (3). For example, USP7's non-catalytic regions, including the TRAF-like and UBL (ubiquitin-like) domains, recognize binding motifs in p53, MDM2, viral proteins (4), and the epigenetic regulators UHRF1 (5) and DNMT1 (6). Phage display (PD) in combination with next-generation sequencing (NGS) is a powerful method that has significantly eased the process of identifying ligands to any given target (7) and minimized the need for repetitive selection rounds (8). We adapted an NGPD protocol followed by the evaluation of obtained sequences using Z scores to identify peptide binding motifs for USP11 (9).

Over the last decade, a significant advance has been made in developing peptide phage display libraries due to their versatility for various purposes such as antibody epitope mapping, mimotope discovery, and protein hot spot identification. Peptides have been displayed mainly on the major coat protein (pVIII) or the pIII protein and include standard linear and cyclic libraries of various lengths (*10, 11*) as well as more complex ones such as bicyclic (*12*) and non-canonical amino acid libraries (*13*) for drug discovery.

Identifying linear antibody epitopes (*e.g.*, for Western Blot antibody reagents) is a relatively straightforward approach, including recent examples such as the simultaneous epitope mapping of two hundred COVID antibodies in a single run using a linear commercial library and NGS (*14*). The same approach has also been applied to identify more complex, structural/discontinuous epitopes for

antibodies or protein-protein interactions, with peptide libraries being used for mimotope discovery, ultimately leading to hotspot identification and/or the discovery of binding partners (*9, 15-17*). Recent computational advances will further facilitate this approach (*18*).

Here, we provide a detailed protocol for discovering peptide binders (schematically depicted in **Figure 1**), which will be applicable to USPs and other targets. The chapter covers the production of His-tagged target proteins for biopanning, phage display peptide library construction and propagation for pVIII display, biopanning with His-tagged target protein, and screening phage-peptide binding events by phage ELISA. Moreover, this chapter overviews screening for peptide-binders by next-generation sequencing, associated data analysis methods and initial evaluation using synthetic peptides in ELISA assays.

## 2. Materials

### 2.1. General

All solutions and buffers should be autoclaved or filter-sterilized as appropriate. Reagents should be analytical reagent grade. Suitable equipment to use includes shaking incubator, sonicator, FPLC, electrophoresis equipment, BioRad Gene Pulser Xcell electroporator, thermal cyclers, Qubit 2.0 Fluorometer 2.0 (Invitrogen, UK), NanoDrop spectrophotometer, Ion Torrent sequencing machine or service provider.

### 2.2. Recombinant protein expression and purification

1. Overexpression plasmid of His-tagged target protein (*e.g.*, in a pET vector)
2. Heat-shock competent cells BL21 (DE3) Codon Plus-RP strain (*E. coli* B $F^-$ *ompT hsdS*($r_B^-$ $m_B^-$) *dcm*$^+$ Tet$^r$ *gal endA* Hte [*argU proL* Cam$^r$])
3. Antibiotics as required for the vector and *E. coli* strain (*e.g.*, kanamycin 50 µg/µL and chloramphenicol 35 µg/µL working concentrations)

4. Lysogeny broth (LB) agar (per liter): 20 g agar, 10 g NaCl, 10 g tryptone, 5 g yeast extract
5. LB microbial growth media (per liter): 10 g NaCl, 10 g tryptone, 5 g yeast extract
6. 2xYT microbial growth media (per liter): 5 g NaCl, 16 g tryptone, 10 g yeast extract
7. Isopropyl ß-D-1-thiogalactopyranoside (IPTG)
8. Bacterial Protein Extraction Reagent (BPER, Thermo Scientific)
9. Broad range unstained protein ladder
10. SDS-PAGE gels
11. Reducing SDS-PAGE sample loading dye 3x: 0.135 M Tris-Cl pH 6.8, 30% (v/v) glycerol, 3% (w/v) SDS, 0.03% bromophenol blue, 0.15 M DTT
12. Tris-glycine SDS-PAGE running buffer 1x: 25 mM Tris base, 192 mM glycine, 0.1% SDS, pH 8.3
13. Coomassie staining solution: 0.25% (w/v) Brilliant Blue R-250, 25% (v/v) isopropanol, 25% (v/v) glacial acetic acid
14. Coomassie de-staining solution: 15% (v/v) isopropanol, 15% (v/v) glacial acetic acid
15. Immobilized metal affinity chromatography (IMAC) Buffer A, filtered and degassed: 50 mM Tris-Cl, pH 8, 300 mM NaCl, 20 mM imidazole, 10% glycerol
16. IMAC Buffer B, degassed: 50 mM Tris-Cl, pH 8, 300 mM NaCl, 500 mM imidazole
17. Size exclusion chromatography (SEC) buffer, filtered and degassed: 50 mM Tris-Cl, pH 7.5, 150 mM NaCl, 1% (v/v) glycerol
18. FPLC protein purification system
19. HiTrap chelating HP column 5 mL (Cytiva) charged with nickel ions or similar
20. Superdex S75 or S200 16/60 Prep grade (PG) chromatography column (Cytiva)
21. Centrifugal concentrators at suitable molecular weight cut-off

### 2.3. Phagemid library construction, panning, and clone screening

1. Phagemid template (in this protocol, pc89 is used) (*19*) [1]

2. Primers suitable for site-directed mutagenesis (SDM) (forward: 5’-GCGGAGGAGCGCCCAAT-3’ and reverse: 5’-TTCCTCGCTCACTGACTC-3’), phosphorylated at their 5’ ends.
3. Suitable sequencing primer: 5’-CTTTATGCTTCCGGCTCGTATG-3’
4. UltraPure DNase/RNase-free distilled water
5. Q5 High-Fidelity DNA polymerase with 5x concentrated buffer
6. Q5 GC enhancer reagent 5x concentrated
7. Deoxynucleotide (dNTP) solution mix (10 mM)
8. DEPC treated, molecular grade water
9. Agarose
10. Deep well plates
11. Nancy 520 or similar DNA gel stain
12. 6x DNA loading dye
13. TAE buffer (agarose gel running buffer): 2 M Tris–HCl, 1 M acetic acid, 0.05 M EDTA, pH 8 (50x)
14. 1 kb DNA molecular weight marker
15. Gel and PCR Clean-up Kit (Macherey-Nagel™ NucleoSpin™ Gel and PCR)
16. DpnI
17. T4 DNA ligase (4000U)
18. TG1 *Escherichia coli*: F' traD36 proAB lacIqZ ΔM15] supE thi-1 Δ(lac-proAB) Δ(mcrB-hsdSM)5(rK - mK -), (Lucigen)
19. Gene Pulser cuvette (0.1 cm)
20. SOC outgrowth medium: 2% (w/v) tryptone, 0.5% (w/v) yeast extract, 10 mM NaCl, 2.5 mM KCl, 10 mM $MgCl_2$, 10 mM $MgSO_4$, and 20 mM glucose
21. 2YT: 16% (w/v) tryptone, 10% (w/v) yeast extract, 5% (w/v) NaCl
22. 2YT Agar: 15 % (w/v) tryptone, 10% (w/v) yeast extract, 5% (w/v) NaCl, 15 g/L agar
23. Ampicillin: 150 mg/mL

24. DNA isolation kit
25. Phagemid (in this protocol pc89_SapI_free)
26. NNK randomized primers: pc89_forward 5'-GATTGCTCTTCGGATCCCGCAAAAGCGGCCTTTG -3'and pc89_reverse 5'-GGTAGCTCTTCGATC(MNNX16)GAATTCACCCTCAGCAGCGA)-3'
27. Qubit™ dsDNA BR Assay Kit (ThermoFisher Scientific)
28. Bioassay dish (L × W × H: 245 mm × 245 mm × 25 mm)
29. His-tagged target protein and His-tagged non-related control protein
30. Blocking solution: 6 or 3% (w/v) skimmed milk powder in PBS (mPBS)
31. Phage-peptide library
32. Dynabeads™ His-Tag Isolation & Pulldown
33. Magnetic rack (DynaMag™-2 Magnet)
34. Phage elution buffer: 100 mM triethylamine
35. Neutralizing buffer: 1 M Tris, pH 7.4
36. Helper phage: MK13KO7 used at $2x10^{12}$ pfu/mL
37. PEG-8000 20% (w/v) /2.5 M NaCl: 200 g PEG-8000 and 146.1 g NaCl per liter
38. D-(+)-Glucose solution: 25 % (w/v) D-(+)-glucose filter sterilized
39. QIAprep Spin M13 kit
40. Agencourt AMPure XP kit
41. P8-Forward1NGS primer: 5'-**GTAATCCTTGTGGTATCG**GATGCTGTCTTTCGCTGC-3'
42. P8 Reverse1NGS primer: 5'-CTAGAACATTTCACTTACGGTTTTCCCAGTCACG-3'
43. Akey-BC-linker1 (Forward): 5-CCATCTCATCCCTGCGTGTCTCCGACTCAG<u>xxxxxxxxxx</u>**GTAATCCTTGTGGTATCG**-3'. Underlined is the 10 bp barcoded region, with many variants available to use. This region varies by 5 bases from every other primer. Codons in bold are the regions that are identical in primers in both first and second-step PCRs.

44. P1prim-linker2 (Reverse): CCTCTCTATGGGCAGTCGGTGAT**CTAGAACATTTCACTTAC**-3’.

    Codons in bold are the homologous linker regions, identical for first and second-step PCRs.

45. Sterile 96-deep well clear polypropylene plate
46. Nunc MaxiSorp flat-bottom 96-well plates
47. Anti-M13 bacteriophage antibody HRP (Sigma-Aldrich)
48. 3,3’,5,5’-Tetramethylbenzidine (TMB) blotting substrate solution
49. 0.45 µm pore size Millipore filter
50. Minimal media (M9) plates:
    a. Make sterile 5x M9 salts, mix 32 g $Na_2HPO_4.7H_2O$ [OR 21.35 g $Na_2HPO_4.2H_2O$ OR 15 g $Na_2HPO_4$ (anhydrous)], 7.5 g $KH_2PO_4$, 2.5 g $NH_4Cl$, 1.25 g NaCl, and distilled water to a total volume of 500mL.
    b. Prepare sterile 20% (w/v) $MgCl_2$ solution.
    c. Prepare sterile 100 mg/mL Thiamine-hydrochloride solution (filter sterilise).
    d. On the day of use, mix 1.5 g agar powder in 75 mL $H_2O$, autoclave, and allow to cool to 50°C, then add 25 mL 5x M9 salts, 1.6 mL 25% glucose, 100 µL $MgCl_2$, and 50 µL Thiamine Hydrochloride. Pour 20 mL per 9 cm diameter plate, allow to set.

# 3. Methods

## 3.1. Recombinant His-tagged target protein production for panning

A one or two-step protocol comprising of affinity and optional size exclusion chromatography depending on the purity of the sample after the first step can be used. The sample purity requirements (>90%) are not as strict as for many biophysical experiments.

### 3.1.1. His-tagged target protein overexpression

1. Transform 1 µL of the plasmid DNA expression construct (~100 ng/µL) into 50 µL of BL21(DE3) CP competent cells by heat-shock or electroporation, add 500 µL LB broth media and place into shaker incubator at 37°C, 180 rpm for 1 hour.
2. Plate 100 µL of the cell culture onto an LB agar plate with suitable antibiotics and incubate at 37°C for ~16 hours (static).
3. Use a single colony to inoculate 100 mL of LB broth containing chloramphenicol and the construct-specific antibiotic, grow cells at 37°C with 180 rpm shaking overnight.
4. The following day, inoculate 3 x 800 mL (depending on expression levels) 2xYT growth medium containing antibiotics 1:100 with the overnight culture and continue growing at 37°C to an $OD_{600}$ of ~0.6. Make a note of the $OD_{600}$ before induction and take a 1 mL sample (pre-induction sample) in a microcentrifuge tube, centrifuge for 3 min at 5,000 g and freeze small pellet at -20°C.
5. Induce recombinant protein expression by adding IPTG to a final concentration of 0.5 mM. Lower the temperature to 18°C and grow for about 16 hours (*see* **Note 1**).
6. After this time, take a 1 mL post-induction sample in the same way as the pre-induction sample. Harvest the culture in a suitable high-volume centrifuge for 30 min at 4,500 g at 4°C and collect the pellet.
7. Resuspend the cell pellet in 25 mL LB and transfer to a 50 mL centrifugation tube, centrifuge again for 15 min at 4,500 g, discard the supernatant and make a note of the weight of the pellet. The cell pellet can be used immediately or frozen down prior to IMAC purification after checking for successful protein overexpression.

3.1.2. Checking for successful overexpression

1. A simple method to evaluate soluble and insoluble protein overexpression fractions, using B-PER solution (chemical lysis, *see* **Note 2**): Pre- and post-induction 1 mL sample pellets are defrosted and resuspended in B-PER by adjusting their respective $OD_{600}$ using 50 µL B-PER per $OD_{600}$ = 0.6 units. Vortex samples for 1 min before centrifugation for 5 min at 13,000 g.

2. Transfer the supernatant to a fresh microcentrifuge tube (soluble fraction). Resuspend the remaining pellet in the same volume of IMAC buffer A as B-PER used in the previous step (insoluble fraction; this requires rigorous resuspending, do not re-centrifuge).
3. Mix samples to 1x with 3x SDS sample loading buffer. Heat the mixtures on a heat block for 10 min at 97°C and then apply 5 µL sample per lane on SDS-PAGE alongside a protein ladder to establish the correct size of the protein bands.
4. Wash gels in $dH_2O$ and add Coomassie staining solution. Leave for 1 hour on a gel shaker, then wash again with $dH_2O$. Add de-staining solution and leave for 15 min, then repeat. Once the gel is fully de-stained, wash with $dH_2O$.

3.1.3. IMAC Nickel affinity chromatography purification of His-tagged DUB

1. Resuspend protein pellets in IMAC buffer A (use 10 mL buffer per gram pellet weight) on ice.
2. Sonicate the cell suspension for 8 min at an amplitude of 18 kHz, switching between 10 sec on 5 sec off (*see* **Note 3**) on ice. Transfer the solution to centrifugation tubes and clear the lysate by centrifugation for 1 h at 24,000 g at 4°C.
3. During the centrifugation step, prepare the Hi-Trap Chelating HP column (*see* **Note 4**) with 5 column volumes (CV) $dH_2O$, 1 CV 50 mM $NiCl_2$, 5 CV $dH_2O$, 1 CV IMAC buffer B (500 mM imidazole), 5 CV IMAC buffer A (20 mM imidazole) using a peristaltic P-1 pump.
4. After centrifugation, filter the lysate using a 0.45 µm filter and load the lysate on the pre-equilibrated Hi-Trap column using a sample loop or peristaltic pump. Wash the column with an additional 5 CV of buffer A.
5. Prepare the FPLC system with IMAC buffer A and B using the appropriate buffer lead. Attach the column and wash with an additional 5 CV buffer A to remove further contaminations.
6. Elute the protein using a linear gradient of imidazole over 20 CV from 0-100% IMAC buffer B using a set flow rate of 2 mL/min and collect ~2.5 mL fractions. Add 1 mM DTT (*see* **Note 5**) to peak fractions to reduce cysteines and keep fractions at 4°C.

7. Analyze the peak fractions by SDS-PAGE (include lysate, pellet, and flow-through) and measure protein concentration using a Nano spectrophotometer.
8. Collect samples that contain the target protein and concentrate the fractions to ~2.5 mL (depending on concentration) using a centrifugal concentrator at a suitable MW cut-off. This sample can be used straight away or further purified using SEC if required (*see* **Note 6**).

3.1.4. SEC purification of His6-tagged target protein

1. Equilibrate a Superdex S75, S200 16/60 PG, or alternative column in a suitable MW range with 1.5 CV SEC buffer.
2. Centrifuge the ~2.5 mL protein sample from the IMAC step for 10 min at 13,000 g prior to loading the sample onto the column via a sample loop.
3. Use a constant flow rate suitable for the column and collect fractions.
4. Analyze the peak fractions by SDS-PAGE and determine the concentration.
5. Using a centrifugal protein concentrator, concentrate pure protein fractions to the desired concentration, flash-freeze aliquots in liquid $N_{2,}$ and store them at -80°C until further use.

## 3.2. Inverse PCR removal of a SapI restriction enzyme site within the phagemid vector

The library production method uses the SapI enzyme to cleave inverse PCR products and facilitate high-efficiency ligation. If your phagemid vector carries a SapI site, it can be removed using inverse PCR as follows. The site-directed mutagenesis (SDM) primers, as listed in Materials are designed for blunt-end ligation, and the example used here is for the pc89 phagemid (*19*). See Figure 2 for the site-directed mutagenesis (SDM) strategy.

1. Isolate the pc89 phagemid vector from TG1 *E. coli* using a plasmid DNA isolation kit. Determine DNA concentration using Qubit. The Qubit working solution is prepared by diluting the BR Reagent 1 in 200 in the BR buffer. 10 µl of the standards (0 ng and 100 ng) are added to 190

µl of working solution. 2 µl of the samples are added to 198 µl of working solution. Allow the incubation with the samples and the BR buffer for a minimum of 5 min at RT.

2. Dilute pc89 phagemid template to 10 ng/µl with UltraPure DNase/RNase-Free distilled water.
3. Set up PCR reaction in 50 µl as follows using the SDM primers: 10 µl Q5 buffer (5x), 10 µl GC enhancer, 1 µl dNTPs (10 mM), 1 µl of forward primer (10 µM), 1 µl of reverse primer (10 µM), 0.5 µl of Q5 enzyme, 1 µl of 10 ng/µl template and 25.5 µl of UltraPure water.
4. Carry out PCR amplification for 95°C for 5 min, then 30 cycles of 95°C for 30 sec, 55°C for 50 sec and 72°C for 5 min, followed by a single incubation at 72°C for 5 min.
5. Add 12 µl of 6x loading to the PCR reaction. Resolve the PCR amplicon on a 1% agarose (w/v) gel, containing 6 µL of Nancy 520 alongside a 1 kb DNA molecular weight marker.
6. Excise and purify the band of the correct size (around 3400 bp for pc89) using a gel clean-up kit. Elute in 20 µl prewarmed (37°C) UltraPure DNase/RNase-Free Distilled Water.
7. Determine DNA concentration using Qubit, as described above.
8. Set up a digestion reaction using DpnI restriction enzyme (to remove template DNA). 1 µg of DNA is digested for 2 h at 37°C, followed by 20 min enzyme heat inactivation at 80°C.
9. Purify DNA using a Nucleospin kit and measure the DNA concentration with Qubit.
10. Set up a ligation reaction as follows: 1 µg of DNA ligated with 0.4 µl of T4 DNA ligase, 4 µl of T4 buffer (10x) in 40 µl total reaction (using UltraPure DNase/RNase-Free Distilled Water) for 18h at 4°C, followed by 20 min at 80°C incubation for enzyme heat inactivation.
11. Again, purify DNA using a Nucleospin kit and measure the DNA concentration.
12. Mix 0.5 µg of purified and ligated DNA with 50 µl of TG1 electrocompetent cells, transfer to an electroporator cuvette and pulse using an electroporator at 1.8 kV, and recover for 1 h at 37°C, 220 rpm in 1 mL of SOC outgrowth medium. Include a transformation of the pre-ligated DNA to estimate the presence of the template DNA.
13. Plate 100 µl of the transformation volume onto 2YT agar plates with 150 µg/mL ampicillin and 1% (w/v) glucose and incubate at 37°C, static for ~16 h.

14. Pick single colonies and inoculate 5 mL of 2YT medium with 150 µg/mL ampicillin and 1% (w/v) glucose each, incubate at 37°C, 220 rpm for ~16 h.
15. Isolate DNA according to manufacturer's instructions.
16. Sequence DNA using the listed sequencing primer to ensure accurate site-directed mutagenesis (in this protocol, the product is designated pc89_SapI_free phagemid).

**3.3. Peptide library construction in a phagemid for pVIII display.**

The method is based on library construction within the pc89 phagemid vector that displays the peptide on pVIII. Alternative phagemid vectors are available, and the method can be adapted for pIII display. The example given here is for a 16-mer peptide display library but can be adapted for other sized peptides, linear or constrained.

1. Isolate the pc89_SapI_free phagemid vector from *E. coli* using a plasmid DNA isolation kit. Determine DNA concentration using Qubit.
2. Dilute pc89_SapI_free phagemid template to 10 ng/µl using UltraPure DNase/RNase-Free distilled water.
3. Set up PCR reaction as follows using the NNK randomized primers (the given volumes are for 1x reaction, 40x reactions are carried out): 10 µl Q5 buffer, 10 µl GC enhancer, 1 µl dNTPs (10 mM), 1 µl of forward primer (10 µM), 1 µl of reverse primer (10 µM), 0.5 µl of Q5 enzyme, 1 µl of 10 ng/µl template and 25.5 µl of UltraPure water.
4. Carry out PCR amplification for 95°C for 3 min, then 30 cycles of 95°C for 30 sec, 57.5°C for 30 sec and 72°C for 2 min and 20 sec, followed by a single incubation at 72°C for 5 min.
5. Add 12 µl of 6x loading dye to every PCR reaction. Resolve the PCR amplicons on a 1% agarose (w/v) gel, containing 6 µl of Nancy 520 and TAE buffer to confirm amplification; include 1 kb DNA molecular weight marker.

6. Excise the correctly sized band (around 3400 bp in this example) and purify using the NucleoSpin™ Gel and PCR clean-up kit. Elute in 20 µl prewarmed at 37°C UltraPure™ DNase/RNase-Free distilled water.
7. Pool elution fractions and determine DNA concentration using Qubit.
8. Set up a digestion reaction using SapI and DpnI restriction enzymes. 1 ug of DNA per reaction was digested for 2 h at 37°C, followed by 20 min enzyme heat inactivation at 80°C.
9. Purify DNA using a Nucleospin kit and measure the concentration using Qubit.
10. Set up ligation reaction as follows: 1-2 ug of DNA ligated with 2 µl of T4 DNA ligase, 20 ul of T4 buffer (10x) in 200 µl total reaction volume incubated for 6 h at 4°C, 18 h at 16°C and 6 h at RT, and then followed by 20 min at 80°C incubation for enzyme heat inactivation. Keep aside 1 µl of DNA after DpnI/SapI digestion and before the ligation step to carry out a separate ligation to estimate the presence of the remaining DNA template.
11. Purify DNA using a Nucleospin kit and measure DNA concentration using Qubit.
12. Mix 1 µg of purified and ligated DNA with 50 µl of TG1 electrocompetent cells, transfer to an electroporator cuvette, and pulse using the electroporator at 1.8 kV and then recover for 1 h at 37°C, 220 rpm in 1 mL of SOC Outgrowth Medium. Ten to 20 transformations should be carried out in total. Include transformation of the pre-ligated DNA to estimate the presence of the template DNA.
13. Prepare 10-fold serial dilutions from 10 µl of transformation reaction, each at a final volume of 100 µl in 2YT with 150 µg/mL ampicillin (w/v). Spread each dilution (containing 10 µl of transformation reaction and a further eight further 10-fold serial dilutions) onto 2YT agar plates with 150 µg/mL ampicillin and 1% (w/v) glucose.
14. Spread the remainder of the transformations onto square bioassay plates containing 2YT agar with 150 µg/mL ampicillin and 1% (w/v) glucose. Incubate all Petri and bioassay dishes at 37°C, static for 16 h.

15. Count the colonies on the serial dilution plates to estimate the library size. Library size should be consistent across at least two consecutive serial dilutions.
16. Add 1-3 mL of 2YT medium with 150 µg/mL ampicillin and 1% (w/v) glucose onto each bioassay plate, and carefully scrape all colonies using a cell scraper (*see* **Note 7**).
17. Collect all the medium-containing bacterial cells, mix them from all bioassay plates by rotation for 5 min, add 20% (w/v) glycerol, prepare 500 µl aliquots and store at -80°C.

**3.4. Phage-peptide library propagation**

The method described uses a 16mer peptide library within the pc89_SapI_free phagemid but is suitable for use with other phage-display peptide libraries using a phagemid system. See Figure 3 for the cloning strategy.

1. Streak minimal media plates with TG1 from glycerol stocks. Incubate at 37°C overnight. As a control, streak the TG1 onto a 2YT agar plate containing ampicillin and glucose and grow overnight at 37°C. Any growth observed indicates contamination of the TG1 stock.

2. Grow the TG1-16mer phagemid library until mid-log phase to a 2 L flask, add 500 µL of glycerol library stock to 500 mL of 2YT containing 150 µg/mL ampicillin and 1% (w/v) glucose and grow to $OD_{600}$ = 0.4 to 0.6 in a shaking incubator at 180 rpm.
3. Add M13KO7 helper phage to a multiplicity of infection (MOI) of 10 and incubate for 45 min at 37 °C static.
4. Centrifuge culture at 3,000 g for 10 min and discard the supernatant. Remove all traces of glucose-containing supernatant with a second quick spin (30 sec) to collect supernatant adhering to the walls of the centrifuge vessel. Discard any supernatant.
5. Re-suspend cell pellet in double the initial volume of 2YT containing 150 µg/mL ampicillin and 50 µg/mL kanamycin (no glucose), shaking at 200 rpm overnight at 30°C.

6. Pick a single TG1 colony from a minimal agar plate and inoculate 10 mL of 2YT medium. Incubate at 37°C with shaking for approximately four hours to reach $OD_{600}$ = 0.5-1.0. These cells are for determining phage titer and will also be used to confirm that no contamination is present within this culture.
7. Plate 100 µL of the TG1 culture on a 2YT agar plate containing 150 µg/mL ampicillin and 1% (w/v) glucose to check for contamination.
8. Centrifuge the phage library culture for 20 min at 8,000 g. Collect the supernatant containing phage. The supernatant can be used directly in phage display experiments or can be further purified/concentrated with PEG. PEG precipitated phage is usually used in the first round of panning, and phage present in culture supernatant used for subsequent rounds.
9. Precipitate the phage by adding 250 mL 20% (w/v) PEG-8000/2.5 M NaCl for every liter of phage supernatant.
10. Place on ice for 30-60 minutes, then centrifuge at 8,000 g for 30 min at 4°C. Discard the supernatant and remove residual PEG/NaCl by centrifuging two more times at 8,000 g for 30 sec, making sure to remove any liquid at the bottom of the centrifuge vessel after each spin.
11. Re-suspend the phage pellet by adding PBS (typically 10 mL total for each 1 L of initial supernatant) and allow the PBS resuspension to soak 5-10 min on ice.
12. Pipette gently to re-suspend the purified phage pellet. Then mix by rotation for 30 min at RT, spin at 13,000 g for 10 min and collect the supernatant.
13. Titre phage immediately. Dilute 10 µL of phage to 1 mL in PBA, then carry out four further 100-fold serial dilutions into PBS. Add 20 µL of the dilutions to 180 µL of the TG1 culture. Also, include diluent PBS buffer only control. Leave at 37°C static for 30 min for the phage to infect bacteria. Spread 100 µL of the infected TG1 onto 2YT agar plates containing 150 µg/mL Ampicillin and 1% (w/v) glucose. Incubate plates at 37°C overnight. Count colonies for each dilution and calculate the original number of phages in the stock, taking into account

the dilution of the phage used. Phage present in culture supernatant should be at ~$1x10^{12}$ per mL, and PEG precipitated phage at least $1x10^{13}$ per mL.

14. Phage should be stored short-term 7-10 days at 4°C or long-term at -80°C as a glycerol stock.

**3.5. Biopanning with phage display targeting his-tagged target protein**

Biopanning is performed in solution against a poly-His tagged protein. Magnetic beads are used to pull the antibody-phage complexes from the solution (*see* **Note 8**).

1. Block the PEG- precipitated phage library in a 1.5 mL microcentrifuge tube with a 1:1 dilution in 6% (w/v) mPBS and incubate for 1 hour at RT on a roller mixer.
2. Target protein (100 µg) is bound to Dynabeads His-Tag Isolation & Pulldown in 1.5 mL microcentrifuge tubes, following the manufacturer's instructions. Block target-coated beads with 3% (w/v) mPBS for 1 hour at RT using a roller mixer. Following blocking, discard the supernatant using a magnetic rack and wash beads four times with 1 x binding/wash buffer (50 mM sodium phosphate, 300 mM NaCl, 0.01% (v/v) Tween 20, pH 8).
3. Add 1 mL of the blocked library [containing around $10^{12}$ phage] to the beads coated with the target and incubate for 1 hour at RT on a roller mixer.
4. Beads are then washed ten times with PBS containing 0.1% (v/v) Tween and ten times with PBS. Bound phage is eluted from the beads using 1 mL of a 100 mM triethylamine solution (incubated for 10 min at RT), and the eluted phage is neutralized by adding 500 µL of 1 M Tris-Cl pH 7.4.
5. Mid-log phase TG1 cells (10 mL) are infected by adding half of the eluted phage library (store the other half at 4°C) directly to the culture in suspension and incubating statically for 30 min at 37°C.

6. Pellet bacteria by spinning at 3,000 g for 10 min and then resuspend pellet in 1 mL 2YT. Remove 10 µL of cells and plate the remaining cells onto 2YT-agar bioassay dishes with 150 µg/mL ampicillin and 1% (w/v) glucose to select for infected cells: Incubate overnight at 37°C.
7. The 10 µL of cells are used to titrate the eluted phage. Carry out a 10-fold serial dilution of the bacteria in 1% (v/v) glucose in 2YT medium. Plate out onto 9 cm 2YT-agar plates with 150 µg/mL ampicillin and 1% (w/v) glucose. Incubate overnight at 37°C. Between $10^4$ and $10^7$ phage eluted from the panning rounds is expected.
8. Make glycerol stocks of each selection by scraping the lawn of bacteria from the bioassay dishes using a cell spreader and 4.5 mL of 2YT media with 150 µg/mL ampicillin and 1% (w/v) glucose. Add glycerol to 20% (v/v), mix on a rotating mixer for 15 min and then divide into 1 mL aliquots and store at -80°C.
9. For further panning rounds, phage particles can be prepared from glycerol stocks as described above for the original library. Panning experiments usually consist of between 3 and 5 rounds of panning (*see* **Notes 8 and 9**).

### 3.4 Screening phage peptide-binding events

Next generation sequencing (NGS) to detect binding events can reveal additional binders compared to the traditional random picking of clones (*7, 8*). The following describes both random picking and then NGS analysis methods.

#### 3.4.1 Random clone picking to screen for peptide-binders

After 3-5 rounds of panning, random colonies from the titration plates can be picked to assess peptide binding/enrichment against targets. This step is achieved by screening through a phage ELISA.

a) Colony picking

   1. To a sterile 96 deep well plate, add 100 µL of 2YT media containing 150 µg/mL ampicillin and 1% (w/v) glucose per well.

2. Pick colonies from the titration plate and inoculate the deep well plate. One well should be left empty as a blank control. Usually, one to four plates are screened per panning.
3. Incubate the culture plate with shaking (300 rpm) at 37˚C for around 2-4 hours, or until all inoculated wells have reached an estimated $OD_{600}$ of 1.

   Once this stage is reached, the experiment can proceed, or the plates can be stored at 4˚C overnight; alternatively, 50 µL of 2YT media with 50% glycerol can be added to each well and flash frozen with dry ice for long term storage at -80˚C.
4. Take 5 µL of culture at $OD_{600}$ of ~1 and inoculate a new culture in a 96 deep-well plate containing 200 µL per well of 2YT media with 150 µg/mL ampicillin and 1% (v/v) glucose using a multichannel pipette.
5. Incubate the deep-well plate with shaking (300 rpm) at 37˚C for around 2.5 hours, or until the cultures have reached the mid-log phase ($OD_{600}$ = 0.5).
6. Next, mix 4 mL of 2YT media with 350 µL of helper phage stock (at around $10^{12}$ cfu/mL). Add 10 µL of the diluted helper phage solution to each well of the 96 deep-well plate. Incubate statically for 30-60 min at 37˚C.
7. Spin the plates at 3,000 g for 10 min and remove the supernatant (around 200 µL).
8. Resuspend infected bacteria pellets with 400 µL of 2YT media with 150 µg/mL ampicillin and 50 µg/mL kanamycin (no glucose). Incubate overnight with shaking (~300 rpm) at 30˚C.
9. The next day, harvest the phage-containing supernatant (around 400 µL) through 10 min centrifugation at 3,000 g.
10. Block the phage solutions in mPBS, by making a 1:1 dilution of the supernatant with 6% (w/v) mPBS and incubate for at least 1 hour at RT.

    The blocked phage solution will be used as primary antibodies for ELISA screening.

b) Monoclonal ELISA

1. His-tagged target proteins and controls (100 µL) are coated on MaxiSorp plates at a 10-50 µg/mL concentration, through an overnight incubation at 4˚C.

   MaxiSorp well coating should be performed the same day as overnight helper phage incubation (point a8) to have blocked antibody-phage and coated plates on the same day. As a control, use a protein of approximately the same molecular weight (MW) and affinity tag as the target protein.
2. The next day, remove excess solution from the wells and wash the plate twice in 300 µL of PBS.
3. Block the coated and washed MaxiSorp wells for 1 hour at RT, with a 3% (w/v) solution of mPBS, using 350 µL per well.
4. Remove mPBS from the wells and wash the plate 3x in 300 µL of PBS containing 0.1% (v/v) Tween 20 and then 3x in 300 µL of PBS.
5. The blocked phage solution is then added to the MaxiSorp wells (100 µL) and incubated for 1 hour at RT. It is recommended to test each phage clone within the same plate (i.e., by coating wells with the target protein and control adjacent to each other) in order to minimize effects arising from plate-to-plate variability.
6. Blocked phage is then removed, and the plate washed four times in 300 µL of PBS containing 0.1% (v/v) Tween 20 and then 3x in 300 µL of PBS.
7. Next, 100 µL of anti-M13 bacteriophage antibody-HRP (1/2000) in 3% (w/v) mPBS is added to each well and incubated for 1 hour at RT.
8. Unbound anti-phage antibody is removed, and the plate washed 4x in 300 µL of PBS containing 0.1% (v/v) Tween 20, then 3x in 300 µL of PBS and once with water.
9. Finally, add 100 µL of TMB substrate per well and incubate for at least 5 min until a color change is noted.
10. Absorbance is measured at 650 nm using a plate reader.

#### 3.4.2 Screen for peptide-binders by next-generation sequencing

This methodology requires additional steps within the panning process to maximize true peptide binders' selection to the target. At the final selection round, the panning is carried out against the target protein and, in parallel, against a control protein (similar molecular weight with the same affinity tag). In addition, panning against the target and control proteins are carried out in multiple replicates (for example, five each). Each panning step produces a phage eluate and a concomitant TG1 glycerol stock of the eluted phagemid pool (as described for round 1 panning above)

a) Preparation of samples for NGS

1. For each phage sample for the final round of panning (selected against both target and control proteins), single-stranded DNA (ssDNA) is extracted and purified using the QIAprep Spin M13 kit (*see* **Note 10**). The ssDNA is ethanol precipitated and stored as a dry pellet to be used as a PCR template.
2. Carry out PCR amplification in a two-step process, with primers designed to give a final product of around 335 bp and containing the required tags for the NGS step (here described for the Ion Torrent platform; Figure 4).
3. For the first step, a primer set specific for phagemid vector is used (here, the pc89 vector). The primers should be designed to amplify the 16-mer peptide sequence flanked by two small linker regions to a size of 275 bp (P8-ForwardNGS and P8-ReverseNGS primers). Primer regions homologous to the primers used in the second step PCR are also included. PCR reactions are 1 µl of the template (10 ng) mixed with 25.5 µl DEPC water, 10 µl Q5 buffer, 10 µl GC enhancer, 1 µl dNTPs (10 mM), 1 µl of forward primer (10 µM), 1 µl of reverse primer (10 µM) and 0.5 µl of Q5 enzyme. The PCR conditions are 95°C for 3 min, [95°C for 30s, 60°C for 30s, 68°C for 30s] x30 and last 72°C for 5 min.

4. Purify amplicons using a PCR Clean-up Kit after resolution on a 3% (w/v) agarose gel.
5. For the second step PCR, purified amplicon (1 µl) from the first step PCR is used as a template with a new primer set. One of these primers contains a distinct barcode (Akey-BC-linker1), the second contains a universal P1 primer compatible with Ion Torrent sequencing (P1prim-linker2). Differently barcoded primers are used to amplify each different sample. 1 µl of amplicon from the step 1 PCR is amplified as described before, with the PCR conditions 95°C for 3 min, [95°C for 30 sec, 63°C for 30 sec, 68°C for 30 sec] x12 and 72°C for 5 min.
6. Analyze amplicons on a 3% (w/v) agarose gel to confirm amplification of the correct size and estimate amplicon concentration using Qubit. Pool all samples in equal amounts.
7. Run the pooled PCR amplicon on a 3% (w/v) agarose gel and purify the amplicon using a gel extraction kit. The expected size of the amplicon is 335 bp.
8. Further purify the amplicon with a PCR clean up kit and run on a 3% (w/v) agarose gel to confirm the presence of a single band.
9. Finally, the concentration of the pooled DNA is measured with Qubit and diluted as required by the commercial Ion Torrent service provider.

b) Next Generation Phage Display data analysis.

1. Deep sequencing of all samples can be performed using an Ion Torrent commercial service on machines ranging from a PGM to an S5 and using chips ranging from 318 to 540 chips. Which combination is used depends on the depth of sequence required and how many samples are multiplexed. Manufacturer estimates of the read numbers

expected from these chips ranged from 4 - 5.5 million reads for a 318 chip, up to 60 – 80 million reads for a 540 chip.

2. 200 base read lengths are used for each chip. Sample requirements for the commercial service used were for amplicon DNA to be supplied in a 30-50µL volume containing a total of 300-400 ng DNA if quantified using a Bioanalyzer or Qubit fluorometry.
3. Data are downloaded as ".gz" suffixed files, compressed using the standard GNU zip (gzip) compression algorithm. The sequencing data are in FASTQ format and are processed on a PC running Ubuntu Linux. Decompression of the compressed FASTQ file is performed at the terminal using gunzip:

```
gunzip -dkf INPUT.fastq.gz
```

4. Conversion of the resultant FASTQ format sequence data file into FASTA format can be performed using several methods with the caveat that sequences must be maintained in single-line format, or if the data is supplied in multi-line format, is converted into single-line FASTQ format at this stage. The following "sed" or "awk" commands typed at the terminal in the directory containing the input file can be used to convert single-line FASTQ format sequencing data files into single-line FASTA format files. Both programs come preinstalled in a Linux environment without any further installation requirements. Alternatively, a suite of bioinformatics tools, such as those found in SeqKit (*20*), can be installed to provide this and further functionality for processing FASTQ and FASTA files.

```
sed -n '1~4s/^@/>/p;2~4p' INFILE.fastq > OUTFILE.fasta
```

```
awk 'NR % 4 == 1 || NR % 4 == 2' INFILE.fastq | sed -e 's/^@/>/' > OUTFILE.fasta
```

```
seqkit fq2fa -w 0 -o OUTFILE.fasta INFILE.fastq
```

5. Demultiplexing is performed on the resultant FASTA file by searching the 5’ region for exact matches to the Ion Torrent barcode sequences used in the second round PCR amplification of the panning-derived DNA samples “binning” FASTA sequences into separate files dependent on which barcode has been identified. This demultiplexing step can be accomplished using freely available software such as FastX-Toolkit; in this case, three files are needed, each containing Ion Torrent barcode sequences of identical length. Given that the standard set of Ion Torrent barcodes are either 10, 11, or 12 nucleotides in length, the first file will contain those of 10 nucleotides in length in tab-separated format with the first column containing the barcode identifier, which will be used to label the appropriate output FASTA file and the second column containing the barcode sequence. The second and third files will contain those barcode sequences, which are 11 and 12 nucleotides long, respectively, along with their corresponding barcode identifiers. Lines starting with ‘#’ are ignored. For example:

```
# ion barcodes length 10nt
BC01	CTAAGGTAAC
BC02	TAAGGAGAAC
BC03	AAGAGGATTC
BC04	TACCAAGATC
BC05	CAGAAGGAAC
```

6. Invocation is carried out at the terminal from the folder in which the data is located, our preference being to set the number of allowed mismatches within the DNA sequence compared to the barcode sequence to zero to ensure only exact matches to the barcode sequence are considered a match:

```
cat INPUT.fasta | fastx_barcode_splitter.pl –mismatches 0 --bc BARCODES10.txt --bol --prefix "fastx_" --suffix ".fasta" >barcodes10.log
```

7. Using this method, an input file of “INPUT.fasta” generates output files of “fastx_BC01.fasta”, “fastx_BC02.fasta”, and so on are generated, generating one demultiplexed FASTA file for each barcode identified. Barcode DNA sequences were removed from FASTA sequences at this stage by deleting any nucleotides up to and including the barcode DNA sequence. Demultiplexing software can be obtained online, such as SeqKit (*20*), FastX-Toolkit (*21*), or flexbar (*22*). Care should be taken to ensure that output FASTA files from whichever software is used are in single line format; i.e., line wrapping is avoided. If sequencing data is supplied in line-wrapped format or subsequently demultiplexed into such a format, conversion into single line format must be performed.

8. FastX_barcode_splitter.pl only identifies barcodes at the exact start (or end) of a sequence (*see* **Note 11**) and does not remove the barcode from the sequence. An alternate method of identifying an exact match to a barcode sequence located within five nucleotides of the 5’ end of a DNA sequence and subsequently removing everything up to and including the barcode sequence CTAAGGTAAC when found can be performed using the following command at the terminal:

```
cat INPUT.fasta | grep --no-group-separator -E -B 1 '^.{0,5}CTAAGGTAAC'
| sed 's/^.\{0,5\}CTAAGGTAAC//' > BC01sequence.fasta
```

9. A convenient method to process all Ion Torrent barcode sequences sequentially is to create a text file containing one such line similar to the above for each barcode to be processed, changing the barcode sequence at both positions in the line and redirecting the output to an appropriately labeled file. Saving the text file with the appendix ".sh", for example, as demultiplex.sh and running by typing the following at the terminal:

```
bash demultiplex.sh
```

Executing the above command will process all lines within the script, attempting to run each new line after the preceding line has finished processing (*see* **Note 12**).

10. For certain library constructs where sequencing was performed in the reverse direction, FASTA sequences were reverse complemented at this stage, and a new file generated from each demultiplexed FASTA file generated in step 8 above. Reverse complementing sequences can be performed using SeqKit specifying the -r and -p options to reverse and complement, respectively, turning off linewidth control and specifying the sequence type, or alternatively using a perl script to perform the process. For example:

```
seqkit seq -v -t DNA -w 0 -r -p -o OUTPUT.fasta INPUT.fasta
```

or

```
perl reverse_complement.pl --infile INPUT.fasta --outfile OUTPUT.fasta
```

11. Translation of DNA sequences into amino acid sequences in each of the three forward frames was performed using a custom Perl script in which amber stop codons were translated to ‘q’ instead of the standard translation ‘*’ symbol, to allow for identification of peptides where translational readthrough could occur in amber-suppressor strain (SupE) TG1 and hence produce viable phage displaying full-length peptide with glutamine (Q) encoded at the stop codon position. Similarly, translation of ochre and opal stop codons were custom coded to characters ‘o’ and ‘p’ respectively (with the -opq option at the command line), although significant translational readthrough would not be expected at these stop codons.

    For each DNA fasta file, the script is run 3 times, specifying the required frame e.g. -f 1 or -f 2 or -f 3 as an argument on each occasion. Output is redirected to a file using the standard Linux redirector ‘>’.

    ```
    perl translate.pl -i input.FASTA -s opq -f 1 >frame1.fasta
    ```

    At this stage, each of the three translated frame files was concatenated together so that subsequent identification of P8 amino acid sequences N-terminally and C-terminally flanking the variable peptide region could be performed by reading from a single protein FASTA file irrespective of which frame the desired sequences were located. The “cat” command is used to join the three translated frame files together:

```
cat frame1.fasta frame2.fasta frame3.fasta >OUTPUT.fasta
```

12. Peptide sequences located between flanking N-terminal and C-terminal vector amino acid sequences (AEGEF and DPAKA respectively for the pc89 pVIII derived libraries) were identified by a further Perl script which must be supplied with the FASTA file generated in the previous step as its input file, along with two files containing respectively a list of the N- and C-terminal conserved motifs to be identified, with the additional option to discard peptide sequences that contain matching amino acid flanking regions if the insert length is shorter than a specified cut off (the default being to retain all sequences with at least 1 amino acid between the motifs).

    Output files are suffixed with ".LR.fa" to indicate the sequences within the file were identified as having matched both the N-terminal and C-terminal motif sequences, with those which matched only the N-terminal (output file suffixed .L.fa) or C-terminal (suffixed .R.fa) sequences.

    The script is executed as follows, and in this example specifies a minimum insert size of 5 amino acids:

```
perl iterate_motifs.pl --infile INPUT.fasta --outfile OUTPUT --peptide 5 --left AEGEF.txt
--right DPAKA.txt
```

An example input sequence and output sequence resulting from this processing step would be:

<u>Input</u>

>G2HKA:08736:10824

SLWYRMLSFAAEGEFRGSLSQPSPLMSRSNLDPAKAAFDSLQATEY……

<u>Output</u>

>G2HKA:08736:10824

RGSLSQPSPLMSRSNL

13. It should be noted that the files describing flanking sequences or motifs may contain multiple target sequences in each file, each of which will be paired with all target sequences in the opposing file when searching the protein FASTA file. As such, if 5 motifs are specified in the N-terminal file and 4 in the C-terminal file, all 20 combinations will be used as search patterns against the translated sequences. Particular amino acids can be excluded using the [^X] pattern, for example AE[^G]EF will identify all motif sequences which do not contain a glycine in position 3 (*see* **Note 13**).

14. A subsequent Perl script counted and ranked identical peptide sequences in two FASTA files to tabulate frequencies on a per peptide basis, normalized frequencies relative to the total number of peptide sequences which passed the length and framework motif filtering in each file, and relative enrichment of any particular peptide from a panning against a target compared to its relative enrichment against a control panning (or library reads where no control panning was used). This was determined using a two proportion Z test to compare the sequence populations between two barcodes using the equation:

$$\left( z = \frac{\mathrm{p1} - \mathrm{p2}}{\sqrt{\frac{p1(1-p1)}{n1} + \frac{p2(1-p2)}{n2}}} \right)$$

This reflects the ratio between the control and target proteins as well as the absolute frequencies and ranks them by relative statistical importance (n1= sample 1 size, n2= sample 2 size, p1= proportion 1, p2= proportion 2).

The script is executed by specifying two .LR.fa input files such as those generated in step 12, i.e. FASTA files where the flanking sequences surrounding the variable peptide sequence have been identified and removed. An optional "-p" switch will generate two further columns in the tabulated output consisting of the percentage abundance of each sequence within the two FASTA files:

```
perl compareZ_2.1.pl INPUT1.fasta INPUT2.fasta >output_table.txt
```

Examples of the first few lines of a typical output are shown in Table 1 (also, *see* **Note 14**).

**3.5 Initial evaluation of candidate peptide binders**

1. Initial evaluation of peptide binders can be performed by coating ELISA MaxiSorp plates with synthetic peptides and controls (100 µL of 50 ug/mL) overnight at a concentration of 10-50 µg/mL at 4 ˚C. Alternatively, biotinylated peptides can be used either by immobilizing them on coated streptavidin on MaxiSorp plates or by coating plates with the protein and using the biotinylated peptides to probe with a streptavidin-poly-HRP antibody.
2. After washing and blocking as described above for the phage ELISA, incubate wells with the His-tagged target protein in blocking buffer.

3. After wash steps, immobilized peptides can be probed with the corresponding His-tagged target protein at 25 µg/mL in PBS. Detect bound protein using 100 µL of mouse anti-tetra His antibody (Qiagen; 1:1000) followed by anti-mouse horseradish peroxidase (HRP) conjugate (1:2000) and TMB.

## 4. Notes

1. Cultures can be scaled up as required. Protein expression conditions can be optimized with regards to parameters such as the use of the medium, $OD_{600}$ at induction, IPTG concentration, expression temperature, and duration.
2. As an alternative to B-PER, micro-sonication (mechanical lysis) can be used to separate the soluble and insoluble fractions.
3. The optimal sonication conditions need to be optimized for the available sonicator. Do not use chemical lysis such as B-PER for the preparative protein purification, as detergents and other reagents contained in such solutions can influence downstream applications.
4. Alternatively, other nickel or cobalt purification columns can be used, or the protein purified in batch.
5. Some target proteins have exposed surface cysteines that require a reducing agent such as DTT or TCEP. Adding these agents to fractions after IMAC works well in our hands and avoids potential issues with a metal reduction during IMAC.
6. If protein is of sufficient purity after the IMAC step perform buffer exchange into PBS or SEC buffer before further use to remove unwanted imidazole, which can lead to protein aggregation.
7. If the cells are too high a density to be scraped into 1 mL of media, add 5-10 mL media to the first plate and use this media consecutively to resuspend cells from each remaining plate.
8. Panning in the solution can be performed with a wide array of variations. If the target protein is not fused to an affinity tag, it can be chemically coupled to one. The most common tag is

biotin using a commercial kit such as the Lightning-Link® Biotin Conjugation Kit (Abcam). Beads such as Dynabeads™ M-280 Streptavidin (ThermoFisher) and SpeedBeads™ Magnetic Neutravidin Coated particles (Merck) can be used to capture the tagged proteins. The biotinylated target can be added directly to the blocked phage library and incubated for 1 hour at room temperature in a rolling machine to interact with peptide. Blocked streptavidin beads can then be used to pull out the target-bound phage through the use of a magnetic rack. Only 5 minutes are needed for the streptavidin-coated beads to bind the target. Subtraction steps with the control biotinylated protein can be carried out in the same stages as described above by direct addition to the blocked phage. Between rounds in the panning experiment, it is also possible to switch the type of magnetic bead (compared to the previous round) used to capture the biotinylated target from streptavidin-coated Dynabeads™ to SpeedBeads™ Magnetic Neutravidin Coated beads. This can help remove any non-specific peptides that bind the beads used to capture the target. However, with NGS, filtering out the background should be possible even if these measures are not taken.

9. A complementary strategy for obtaining more specific peptides uses the subtraction of non-specific binders. After the first round of biopanning, a subtraction step is performed. The subtraction step is achieved by adding a similar-sized protein possessing the same affinity tag as the target (e.g. poly His-tag) coupled to the beads and then added to the phage library. Non-specific binders are then removed from the library through the use of magnetic beads, and the "subtracted" library is then used for further rounds of panning against the target protein. In summary: Dynabeads™ His-Tag Isolation & Pulldown are coated in the control protein and blocked exactly as described for the target protein. The blocked beads are then introduced to the blocked library, and the antibody phage is allowed to bind the unrelated protein for 1 hour. Beads are then discarded using a magnetic rack, and the subtracted supernatant used for a normal round of panning against the target protein.

10. Alternatively, the glycerol stocks of polyclonal phage produced after panning can be used as a source of dsDNA and DAN extracted using the QIAprep Spin Miniprep Kit and following the manufacturer's instructions.
11. Note that FastX_barcode_splitter.pl, which is part of the FASTX-toolkit, does not remove the barcode sequence following the identification, and this must be performed as a separate step. Given that the default Ion Torrent barcodes can be 10, 11 or 12 nucleotides in length, the result is that if the barcode sequence is not removed, the frame in which the correct translation is found will vary depending on the length of the barcode sequence used in the PCR primer for that sample. Similarly, reverse complementing data before translation will result in the frame in which the correct in-frame translation appears being a function of the read length of that particular sequence.
12. Considerable improvements in run times can be obtained by running commands in parallel where suitable, using for example "GNU Parallel" (*23*). Parallel processing can be applied to many processing steps; for example, demultiplexing a FASTA file containing barcoded sequences would rely on a single input file source, with multiple output files each of which is written independently and not reliant on the output from any other process to complete. In the given example of running "demultiplex.sh", where each of the 96 possible Ion Torrent barcodes are searched for and removed during demultiplexing of a DNA FASTA file using the "grep" and "sed" commands, the command to run the script can be adjusted to run on eight processors simultaneously by using the following:

    ```
    parallel -j 8 < demultiplex.sh
    ```

13. While this method of processing Ion Torrent deep sequencing data initially proved very successful, as new Ion Torrent chips were developed, the size of the corresponding data files increased rapidly.

As such, pipeline scripts are written in a modular fashion such as this which rely on their input on text files stored on disk generated during a previous step of the process, soon led to the limiting factor being the rate at which data could be written to and read from storage drives and the corresponding size of files needing to be processed in such a manner at each step.

Running scripts in parallel where possible - for example, demultiplexing barcode sequences from a FASTA file using one processor per barcode - can result in a notable speed improvement only until the limiting factor becomes the rate at which data can be read from hard disk. As an alternative, the entire pipeline can be rewritten so that each sequence in the compressed FASTQ file is read just once from disk into memory and all processing steps performed sequentially on a per-sequence basis. All the processing steps such as barcode identification, translation, and insert identification are performed on that sequence before the resulting data is written out to disk, proved to considerably reduce run times as well as providing a simpler user experience for inexperienced users of the analysis pipeline.

This enhanced pipeline combines all previous scripts for the data processing pipeline into a single program, written in Perl, which carries out all steps from reading in either a compressed or uncompressed FASTQ file, demultiplexing, reverse complementing, translating, identifying N- and C-terminal flanking motifs, and optionally ranking sequences by frequency.

Z test comparisons are still performed separately due to the range of potential target versus control comparisons, which may be considered on a per-experiment basis, and the possibility that normalized frequencies may be the only ranking score an end-user may require for their analysis.

Intermediate files, such as the demultiplexed FASTA files and translations of each of the three frames, are not written to disk unless that option is selected, saving on storage requirements. The user must specify a FASTQ input file and the output file to be created, a

file of barcodes in a tab-separated format consisting of a barcode ID and barcode sequence (one barcode per line), and two files containing the N-terminal and C-terminal amino acid sequences flanking the randomized region of interest. For example:

perl pipeline.pl --infile INFILE.fastq.gz --outfile EXAMPLE --left N_list.txt --right C_list.txt --barcodes BCLIST.txt

The option to reverse complement each DNA sequence prior to translation is given by the "--revcomp" option, the minimum number of amino acids located between the N-terminal and C-terminal motifs to be considered a hit is specified by the "--minimum <integer>" option, with the default being a requirement for at least 1 amino acid to be present between the N- and C-terminal motifs, and generating a text file for each barcode containing ranked frequencies of all identified inserts is done using the "--rank" option.
Full output of all intermediate files, specified using the "--fulloutput" option, is the equivalent to specifying "--fasta", "--translate" and "--rank", and will generate demultiplexed DNA FASTA files, translations of each sequence in all 3 forward frames, and a text file with numerically ranked peptide sequences for each FASTA file.

14. Tab-separated tables generated in this manner are amenable to subsequent analysis, for example, filtering the Z test ranked peptide sequences based on which sequences are above an arbitrary cut-off value across replicate panning barcoded samples. These analyses were typically performed using inbuilt Linux tools such as awk, grep, sort, uniq, wc, and nl with additional Perl scripts written on an ad hoc basis for more complicated filtering and collating operations. For example, given a FASTA file containing many peptide sequences such as that generated by the output shown in step 12, a ranked listing of the peptides in descending

numerical order can be produced using the Linux pipeline to pass data from one program to the next using the commands grep, sort and uniq:

```
grep -v '>' INPUT.fasta | sort | uniq -c | sort -nr > OUTPUT.txt
```

In this string of commands the program "grep" is used to identify non-matches (using the -v switch) to lines containing the character '>' which forms the start of a FASTA sequence identifier. This generates a list consisting solely of the corresponding peptides sequences, which are then piped to the "sort" command to be alphabetically sorted. The data is now in the required input format for "uniq" to count adjacent repeated sequence lines (using the -c switch). The resulting data are processed again by sort, this time sorting in reverse numerical order (specified by combining the -r and -n switches), before being redirected to a text file. Chaining simple in-built Linux commands together in this manner can provide highly effective and rapid methods for processing and previewing relevant information from large datasets.

**Acknowledgements**

This work was supported by the University of Nottingham and the Biotechnology and Biological Sciences Research Council (BBSRC) doctoral training studentships [grant number BB/M008770/1].

**References**

1. T. E. T. Mevissen, D. Komander, Mechanisms of Deubiquitinase Specificity and Regulation. *Annu Rev Biochem* **86**, 159-192 (2017).
2. M. E. Sowa, E. J. Bennett, S. P. Gygi, J. W. Harper, Defining the human deubiquitinating enzyme interaction landscape. *Cell* **138**, 389-403 (2009).

3. A. Remenyi, M. C. Good, W. A. Lim, Docking interactions in protein kinase and phosphatase networks. *Current opinion in structural biology* **16**, 676-685 (2006).

4. A. K. Pozhidaeva *et al.*, Structural Characterization of Interaction between Human Ubiquitin-specific Protease 7 and Immediate-Early Protein ICP0 of Herpes Simplex Virus-1. *J Biol Chem* **290**, 22907-22918 (2015).

5. Z. M. Zhang *et al.*, An Allosteric Interaction Links USP7 to Deubiquitination and Chromatin Targeting of UHRF1. *Cell reports* **12**, 1400-1406 (2015).

6. J. Cheng *et al.*, Molecular mechanism for the substrate recognition of USP7. *Protein & cell* **6**, 849-852 (2015).

7. N. A. Ngubane *et al.*, High-throughput sequencing enhanced phage display identifies peptides that bind mycobacteria. *PloS one* **8**, e77844 (2013).

8. P. A. t Hoen *et al.*, Phage display screening without repetitious selection rounds. *Anal Biochem* **421**, 622-631 (2012).

9. A. Spiliotopoulos *et al.*, Discovery of peptide ligands targeting a specific ubiquitin-like domain-binding site in the deubiquitinase USP11. *J Biol Chem* **294**, 424-436 (2019).

10. A. Fagerlund, A. H. Myrset, M. A. Kulseth, Construction and characterization of a 9-mer phage display pVIII-library with regulated peptide density. *Appl Microbiol Biotechnol* **80**, 925-936 (2008).

11. K. A. Noren, C. J. Noren, Construction of high-complexity combinatorial phage display peptide libraries. *Methods* **23**, 169-178 (2001).

12. C. Heinis, T. Rutherford, S. Freund, G. Winter, Phage-encoded combinatorial chemical libraries based on bicyclic peptides. *Nat Chem Biol* **5**, 502-507 (2009).

13. J. M. Tharp *et al.*, An amber obligate active site-directed ligand evolution technique for phage display. *Nature communications* **11**, 1392 (2020).

14. H. Qi *et al.*, Antibody Binding Epitope Mapping (AbMap) of Hundred Antibodies in a Single Run. *Mol Cell Proteomics* **20**, 100059 (2021).

15. J. P. Ting *et al.*, Utilization of peptide phage display to investigate hotspots on IL-17A and what it means for drug discovery. *PloS one* **13**, e0190850 (2018).
16. J. Cao *et al.*, Phage-Display Based Discovery and Characterization of Peptide Ligands against WDR5. *Molecules* **26**,  (2021).
17. R. J. Gubeli *et al.*, In Vitro-Evolved Peptides Bind Monomeric Actin and Mimic Actin-Binding Protein Thymosin-beta4. *ACS Chem Biol* **16**, 820-828 (2021).
18. H. Ashkenazy *et al.*, Motifier: An IgOme Profiler Based on Peptide Motifs Using Machine Learning. *J Mol Biol* **433**, 167071 (2021).
19. F. Felici, L. Castagnoli, A. Musacchio, R. Jappelli, G. Cesareni, Selection of antibody ligands from a large library of oligopeptides expressed on a multivalent exposition vector. *J Mol Biol* **222**, 301-310 (1991).
20. W. Shen, S. Le, Y. Li, F. Hu, SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PloS one* **11**, e0163962 (2016).
21. FastX-Toolkit: http://hannonlab.cshl.edu/fastx_toolkit/.
22. M. Dodt, J. T. Roehr, R. Ahmed, C. Dieterich, FLEXBAR-Flexible Barcode and Adapter Processing for Next-Generation Sequencing Platforms. *Biology (Basel)* **1**, 895-905 (2012).
23. O. Tange, GNU Parallel - The Command-Line Power Tool. *The USENIX Magazine* **Februar 2011**, 32-47 (2011).

## Figure legends

**Fig. 1. Overview of the NGPD method.** A schematic overview of the technique detailing the different steps involved in discovering peptide binders to a given target by combining phage display biopanning with next-generation sequencing.

**Fig. 2. SDM strategy to remove SapI site from the phagemid vector pc89.** The forward primer introduces a site-directed change at a single nucleotide that removes the SapI recognition sequence. After inverse PCR the vector is blunt ligated and transformed. Underlined bases are complementary between the vector and primers. The shaded bases are the SapI recognition site. The SapI cleavage site is also shown.

**Fig. 3. Library cloning strategy for p8 display in phagemid vector pc89.** The P8 forward primer introduces a SapI restriction site that, once cleaved, produces a GAT 5' overhang. The p8 reverse primer introduces 16 NNK codons and a SapI cleavage site that produces a CTA 5' overhang once digested. Upon ligation, the vector is reformed, and the 16 NNK codons replace the original TAG stop codon (boxed). Underlined bases are complementary between the vector and primers. The shaded bases are the SapI recognition site. The SapI cleavage site is also shown. Dotted lines indicate the SapI cleavage of the amplified vector.

**Fig. 4. Overview of the amplification of the 16mer peptide gene region for NGS analysis.** The region AEGEF-16merpeptide-DPAKA is amplified in a primary PCR that introduces linker regions upstream and downstream (A). In the secondary PCR, a primer homologous to the upstream linker 1 introduces a distinct barcode and the 'A' adapter for Ion Torrent sequencing (Akey-BC-linker1), the second primer, homologous to linker 2, introduces a 'P1' adapter for Ion Torrent sequencing (P1prim-linker2).

**Table 1. Example output format from the Perl script used to analyse the NGS data**

| Sequence | INPUT1.fasta | INPUT2.fasta | Z-score | INPUT1.fasta_% | INPUT2.fasta_% |
|---|---|---|---|---|---|
| LTYNARKADPRLSAML | 226640 | 177254 | 23.64 | 48.7150 | 46.1425 |
| HSNLLRCLSSESKRSN | 45315 | 32535 | 20.33 | 9.7402 | 8.4695 |
| MKPAEPMKRSELTVTL | 353 | 9 | 17.88 | 0.0759 | 0.0023 |
| IKPSEPNKVTPTQDSM | 294 | 0 | 17.15 | 0.0632 | 0.0000 |
| TKTELASLSCARAGEW | 127 | 0 | 11.27 | 0.0273 | 0.0000 |