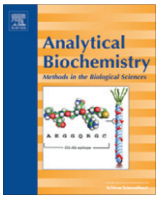
Analytical Biochemistry 421 (2012) 622–631

Contents lists available at SciVerse ScienceDirect

Analytical Biochemistry

journal homepage: www.elsevier.com/locate/yabio



Phage display screening without repetitious selection rounds

Peter A.C. 't Hoen [a,*], Silvana M.G. Jirka [a], Bradley R. ten Broeke [a], Erik A. Schultes [a], Begoña Aguilera [b], Kar Him Pang [b], Hans Heemskerk [a], Annemieke Aartsma-Rus [a], Gertjan J. van Ommen [a], Johan T. den Dunnen [a]

[a] *Center for Human and Clinical Genetics and Leiden Genome Technology Center, Leiden University Medical Center, 2300 RC Leiden, The Netherlands*
[b] *Prosensa Therapeutics, 2333 CH Leiden, The Netherlands*

ARTICLE INFO

Article history:
Received 15 September 2011
Received in revised form 4 November 2011
Accepted 5 November 2011
Available online 13 November 2011

Keywords:
Phage display
Next generation sequencing
Bone
Drug targeting

ABSTRACT

Phage display screenings are frequently employed to identify high-affinity peptides or antibodies. Although successful, phage display is a laborious technology and is notorious for identification of false positive hits. To accelerate and improve the selection process, we have employed Illumina next generation sequencing to deeply characterize the Ph.D.-7 M13 peptide phage display library before and after several rounds of biopanning on KS483 osteoblast cells. Sequencing of the naive library after one round of amplification in bacteria identifies propagation advantage as an important source of false positive hits. Most important, our data show that deep sequencing of the phage pool after a first round of biopanning is already sufficient to identify positive phages. Whereas traditional sequencing of a limited number of clones after one or two rounds of selection is uninformative, the required additional rounds of biopanning are associated with the risk of losing promising clones propagating slower than nonbinding phages. Confocal and live cell imaging confirms that our screen successfully selected a peptide with very high binding and uptake in osteoblasts. We conclude that next generation sequencing can significantly empower phage display screenings by accelerating the finding of specific binders and restraining the number of false positive hits.

Phage display is a powerful technique for the identification of peptides, proteins, or antibodies with affinity for a specific target [1]. Many different phage libraries have been created and used for different applications, ranging from libraries with short random peptide inserts to complementary DNA (cDNA)[1] and classical and

affinity binders.

Traditionally, peptides enriched after several rounds of selection are identified by DNA sequencing of the inserts of a limited number (tens to hundreds) of clones. Depending on the sequence diversity remaining in the library after selection, the analysis of such a limited number of clones does not necessarily result in the discovery of the most promising candidates. Moreover, phage display screenings are notorious for their identification of false positive hits. These emerge for two important reasons: binding to non-target-related materials used during the selection (e.g., plastics, albumin) and propagation advantages [5]. A well-known example in the latter category is the greatly accelerated propagation of phages displaying the HAIYPRH peptide in the Ph.D.-7 library as a consequence of a mutation in the Shine–Dalgarno box of the phage protein gIIp in this clone [6]. This peptide has been identified in at least 13 inde-

identification of

have been con-

structed: PepBank can be used to search for peptides already published in other experiments [7], whereas SAROTUP searches for peptides binding to unintended materials [8].

With the advance of next generation sequencing (NGS), it is now possible to sequence millions of inserts in parallel [9,10]. Thus, NGS permits a more expedient and higher resolution characterization of the library [11–13]. In this article, we use NGS to examine the contents and enrichment process of the Ph.D.-7 library, the most popular commercially available phage library for peptide ligand screening. In this library, random 7-mer peptides are displayed at the tip of the pIII minor coat protein of the M13 phage. We employed NGS after each round of selection to carefully characterize the enrichment process and show that positive hits can already be found after one round of selection. By comparison of the content of different libraries and by sequencing of the naive library, we have found an efficient way to discriminate true binders and false positives such as target-unrelated peptides.

* Corresponding author. Fax: +31 71 5268285.
E-mail address: p.a.c.hoen@lumc.nl (P.A.C. 't Hoen).

[1] *Abbreviations used:* cDNA, complementary DNA; NGS, next generation sequencing; α-MEM, alpha minimum essential medium; EDTA, ethylenediaminetetraacetic acid; PBS, phosphate-buffered saline; BSA, bovine serum albumin; PCR, polymerase chain reaction; TFA, trifluoroacetic acid; TIS, triisopropyl silane.

0003-2697 © 2011 Elsevier Inc. Open access under CC BY-NC-ND license.
doi:10.1016/j.ab.2011.11.005

Materials and methods

Cell culture

A murine mesenchymal progenitor stem cell line (KS483, Percuros, Enschede, The Netherlands) was cultured into osteoblasts in α-MEM (minimum essential medium, 1×) with GlutaMAX supplemented with 10% fetal bovine serum and 1% penicillin/streptomycin (all from Gibco, Breda, The Netherlands) under 5% CO_2 [14–16]. Cells were passaged by 0.05% trypsin–EDTA (ethylenediaminetetraacetic acid) (Gibco) treatment at 3- to 4-day intervals. The cultured cells were grown to subconfluency. For differentiation, KS483 cells were seeded at a density of 15,000 cells/cm^2 in an 8.6-cm Petri dish. Every 3 to 4 days, the medium was changed. From day 4 of culture, full confluence was reached and L(+) ascorbic acid (50 μg/ml, VWR International, The Netherlands) was added to the culture medium. When compact cell nodules appeared (from day 11 of culture onward), β-glycerolphosphate (5 mmol/L, Fluka) was added to the culture medium. At day 18 of differentiation, the control cells were stained with 3% Alizarin red solution (Sigma–Aldrich, St. Louis, MO, USA) to confirm successful differentiation.

Biopanning

A heptapeptide phage display library (Ph.D.-7 Phage Display Peptide Library Kit, New England Biolabs, Beverly, MA, USA) was used for the in vitro biopanning experiments. KS483 cells, at 4 and 18 days of differentiation, were washed gently with phosphate-buffered saline (PBS) and incubated with 5 ml of α-MEM, containing 0.1% (w/v) bovine serum albumin (BSA), for 1 h at 37 °C under 5% CO_2. The cells were gently washed once with 5 ml of PBS before adding the phage library. Phages (2×10^{11}) were added in 3 ml of α-MEM containing 0.1% BSA. Cells were incubated with the phage for 1 h at 37 °C while shaking at 70 rounds per minute. After the incubation, the cells were gently washed six times by incubating with 5 ml of ice-cold α-MEM, containing 0.1% BSA, for 5 min. Subsequently, the cells were incubated for 10 min on ice with 3 ml of 0.1 M HCl (pH 2.2) to elute cell surface-bound phage. This solution was neutralized by the addition of 0.6 ml of 0.5 M Tris. The cells were then lysed for 1 h on ice in 3 ml of 30 mM Tris–HCl and 1 mM EDTA (pH 8.0) to recover the cell-associated phage fraction. Phages from each fraction were titrated and amplified according to the manufacturer's protocol. Each subsequent round of selection employed 2×10^{11} phage derived from the phage library recovered from the previous round.

DNA preparation and sequencing

Sequencing was performed with the Illumina Whole Genome Analyzer WG2 (San Diego, CA, USA). Phage DNA was isolated from the amplified phage stocks and the naive (unselected) library. For this, 10 μl of a 1000 times diluted phage stock was added to 1 ml of a 100 times diluted overnight culture of ER2738 bacteria and grown for 4.5 h at 37 °C while shaking at 200 rpm. Bacteria were centrifuged for 30 s at 15,700g. Then, 500 μl of the top 80% of the supernatant was precipitated with 200 μl of polyethylene glycol (PEG)/NaCl for 10 min at room temperature, and the DNA was further isolated according to the manufacturer's protocol. The final pellet was dissolved in 25 μl of Milli-Q water, and the DNA concentration was determined by Nanodrop before freezing at −20 °C.

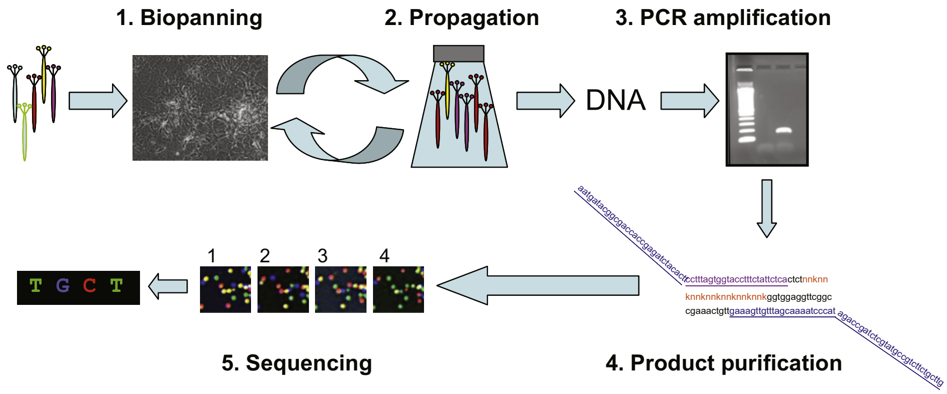

1. Biopanning
2. Propagation
3. PCR amplification
DNA
1
2
3
4
T G C T
aatgatacggcgaccaccgagatctacactcctttagtggtacctttctattctcactctnnknn
knnknnknnknnknnkggtggaggttcggc
cgaaactgttgaaagttgtttagcaaaatcccatagaccgatctcgtatgccgtcttctgcttg
5. Sequencing
4. Product purification

Fig.1. Overview of method. (1) In the biopanning phase, cells are incubated with the phage library. After washing away nonbound phages, binders are isolated and amplified in bacteria. The amplified libraries can be reused in a subsequent round of biopanning. (2) Alternatively, the DNA of the phages can be isolated and (3) amplified with PCR primers complementary to sequences flanking the variable region of the M13 phage DNA. The PCR primers contain tails with the adapter sequences necessary for Illumina sequencing. (4) A 160-bp fragment is purified from gel (lane 1: size marker; lane 2: negative PCR control; lane 3: phage DNA) and (5) sequenced by the Illumina sequencer.

Table 1
Overview of sequencing results phage display experiments.

Selection round	Differentiation	Internalized/ surface	Number of sequences	Number of unique sequences	Most abundant sequence	Number of counts for most abundant
3	4 Days	Surface	11,192,802	274,666	GETRAPL	1,279,913
4	4 Days	Surface	12,858,902	123,972	GETRAPL	3,113,643
3	4 Days	Internalized	12,357,279	358,844	GETRAPL	397,669
3	18 Days	Surface	13,146,497	196,615	GETRAPL	3,599,322
1	18 Days	Internalized	15,595,055	1,913,785	HAIYPRH	41,257
2	18 Days	Internalized	17,092,798	1,318,281	GETRAPL	547,210
3	18 Days	Internalized	13,217,869	350,379	RHEPPLA	543,337
4	18 Days	Internalized	13,880,199	282,266	RHEPPLA	1,413,696
No selection	–	–	6,688,401	3,887,498	HAIYPRH	36

Phage DNA was amplified with the following polymerase chain reaction (PCR) primers:

> Forward: AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT TCC TTT AGT GGT ACC TTT CTA TTC TC∗A
> Reverse: CAA GCA GAA GAC GGC ATA CGA GAT CGG TCT ATG GGA TTT TGC TAA ACA ACT TT∗C,

where an asterisk (∗) indicates phosphorothioate bond. PCR primers used to amplify the phage DNA contained a subsequence that recognized the sequence flanking the 21-nucleotide unknown insert sequence and the adapters necessary for binding to the Illumina flow cell. The final product of the PCR was 160 bp long. The PCR protocol applied was the following: 1 ng of phage DNA was incubated with 2.625 U of high-fidelity Taq polymerase (Roche Diagnostics, The Netherlands), 20 pM of primers in 1 × high-fidelity PCR buffer containing 15 mM $MgCl_2$, and amplified for 20 cycles, each consisting of an incubation for 30 s at 94 °C, 30 s at 60 °C, and 30 s at 72 °C. The PCR was stopped in the exponential phase to mitigate PCR-induced sequence biases. The final PCR product was purified with the QIAquick PCR Purification Kit (Qiagen, Valencia, CA, USA). Concentrations, as well as the correct length, of the PCR product were established with an Agilent 2100 Bioanalyzer and a DNA 1000 assay.

Each PCR product was applied to a single lane of an Illumina flow cell and subjected to solid phase amplification in the cluster station following the manufacturer's specifications. Single end sequencing for 27 to 35 cycles (27 cycles are sufficient, but runs were sometimes extended to 35 cycles due to requirements for samples in other lanes of the same flow cell) was performed with a custom sequencing primer that started exactly at the first position of the unknown insert sequence (ACA CTT CCT TTA GTG GTA CCT TTC TAT TCT CAC TC∗T).

Data analysis

All sequenced lanes were run through the initial Illumina Genome Analyzer Pipeline (Firecrest → Bustard → Gerald) for image analysis, quality control, and base calling. Only sequences with the expected 6-nucleotide sequence after the insert (GGTGGA) were used (~95% of the sequences remaining after this filter step). DNA sequences were translated to amino acid sequences with a custom Perl script using conventional amino acid codon tables. All reported numbers and sequences refer to the translated peptide sequences. An asterisk (∗) indicates the presence of a stop codon, whereas a minus symbol (−) indicates the presence of an unknown nucleotide in the triplet.

For plotting of phage abundance, a square root transformation was applied on the number of counts in the library, a commonly applied data transformation to stabilize the variance in count data [17].

Simulation of random clone picking

We randomly selected 50 amino acid sequences from the round 1, round 2, round 3, and round 4 phage libraries selected for internalization into KS483 cells at day 18 of differentiation. We counted how frequently we selected the most abundant or 10 most abundant peptide sequences, as identified after complete sequencing of the round 4 library. We repeated this 20 times and report the average percentages and standard deviations for the 50 random picks in Fig. 3D (see Results).

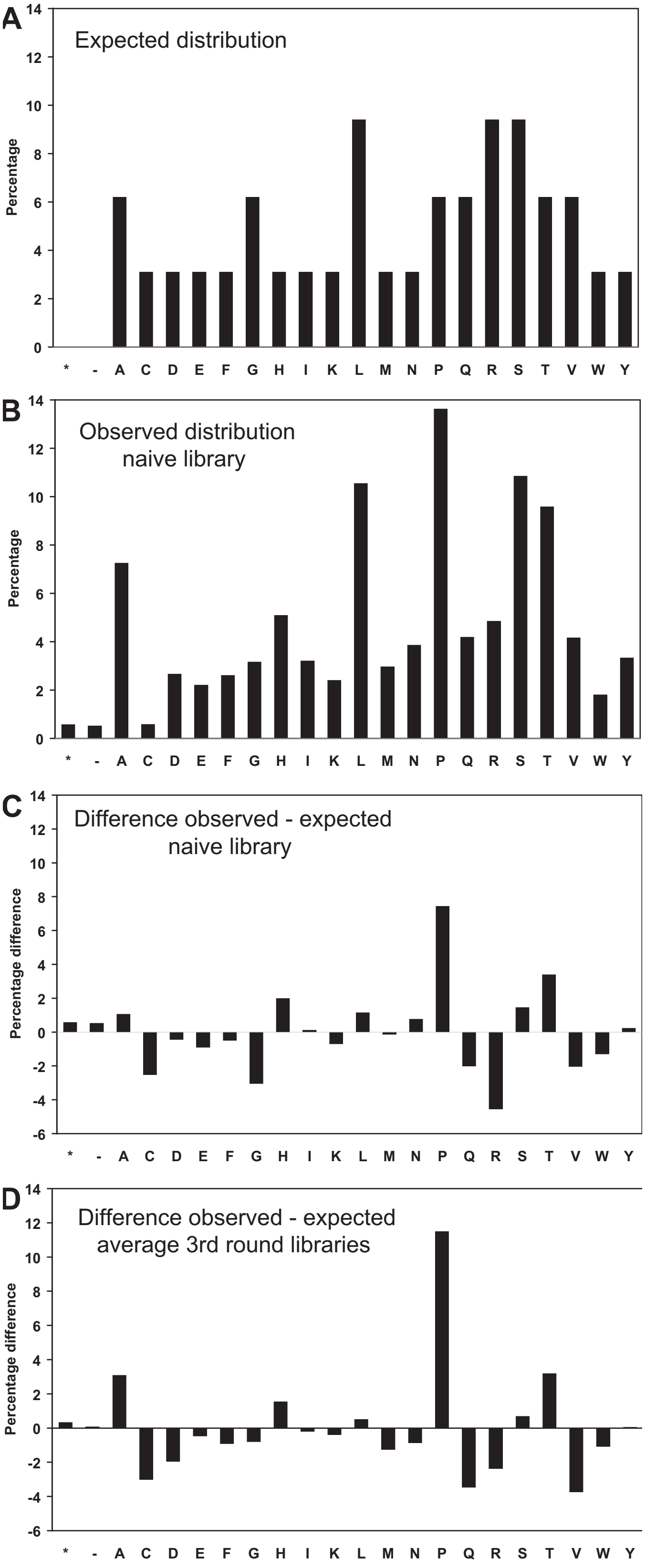


Fig.2. Analysis of amino acid composition of Ph.D.-7 library: Overview of average amino acid composition (%) of all unique peptides sequenced. (A) Theoretical amino acid composition based on the translation of random $(NNK)_7$ inserts. (B) Observed composition in naive library. (C) Difference between the observed composition in the naive library and the theoretical composition. (D) Difference between the average composition in all round 3 libraries and the theoretical composition. Standard amino acid one-letter codes are used; ∗, one of the nucleotides in the triplet is unknown (N); −, stop codon.

Peptide synthesis

The 7-mer peptides were synthesized by standard Fmoc solid phase peptide chemistry on a PS3 or Tribute Peptide synthesizer (Protein Technologies) using HCTU (5 eq) as activating reagent and *N,N′*-diisopropylethylamine (DIPEA, 10 eq) as base. Resin-bound peptides were manually coupled to the fluorescent label by treatment with 5(6)-carboxyfluorescein *N*-hydroxysuccinimide ester and triethylamine in dimethyl formamide (DMF). After cleavage (trifluoroacetic acid [TFA]/triisopropyl silane [TIS]/H_2O 95:2.5:2.5 [v/v] or TFA/thioanisol/TIS/H_2O 90:2.5:2.5:5 [v/v] for peptide sequences containing Met), filtration, precipitation over cold ether, and centrifugation, crude FAM-labeled peptides were obtained. Peptides were purified by reverse phase high-performance liquid chromatography (RP–HPLC) on a Shimadzu Prominence HPLC system (Alltima C_{18} column [5 μm,10 × 250 mm], solvent A [0.1% TFA CH_3CN/H_2O 5:95], and solvent B [0.1% TFA CH_3CN/H_2O 80:20]). Peptides were analyzed by electrospray ionization mass spectrometry (ESI–MS, positive mode) on an Agilent LC ion trap mass spectrophotometer. Fluorescent peptide concentrations were determined by spectrophotometric analysis at 490 nm and pH 7.5.

Fluorescent imaging of KS483 cells

KS483 cells were seeded at a density of 15,000 cells/cm^2 in 6-well plates with glass cover slides (21 × 26 mm, Menzel Glaser, Germany) or in Mattek glass-bottom dishes with a diameter of 14 mm and a glass thickness of 1.5 mm in culture medium described above. Cells were gently washed with PBS before adding 2.25 μM FAM-labeled peptide in the medium without serum for 24 or 48 h. After incubation, cells at 4 days of differentiation were washed three times with PBS, fixed in ice-cold methanol for 5 min, and air-dried for 5 min. Subsequently, the cells where embedded on microscope slides with VECTASHIELD (Vector Laboratories), and the slides where analyzed with a Leica TCS SP5 DMI6000 confocal microscope (HCX PL APO 63×/1.4 oil immersion objective, 8 bit resolution, 512 × 512 pixels, 400 Hz speed, Leica Microsystems). Cells at day 18 of differentiation were gently washed with PBS (three times) and supplied with fresh medium before analysis with live cell imaging using a Leica Af6000LX inverted microscope (HCX PL FLUOTAR 63x/1.25 oil immersion objective, 12 bit resolution, 1392 × 1040 pixels, Leica Microsystems).

Results

In the current study, we employed NGS technology to characterize the phage display screening process during successive rounds of selection. We used the combination of phage display and NGS to select for peptides that are binding to the surface of and/or internalized by KS483 cells in different stages of differentiation. KS483 cells are osteoblastic cells that can be efficiently differentiated into mature osteoblasts and form nodules depositing mineralized calcium material within 18 days [14,18]. The identified peptides may ultimately be used for the targeting of pharmaceutical formulations to bone or for enhancing the intracellular uptake of drugs into osteoblasts.

Sequencing of phage display libraries

The preparation of the phage libraries for sequencing is a simple and short procedure, as depicted in Fig. 1. After DNA isolation from the entire phage pool, the fragment containing the insert was amplified and equipped with the linkers necessary for NGS in a one-step PCR. The unknown inserts of 21 nucleotides (7 amino acids) were sequenced on the Illumina Whole Genome Analyzer. After quality control and translation of the DNA sequences to amino acid sequences, we obtained approximately 13 million peptide sequences for each phage display library (Table 1). This provides an ultra-deep profiling of the content of the phage display libraries.

Characterization of naive library

Before analyzing the phage display libraries selected against biological targets, we screened for potential sequence biases introduced by the propagation of the phages in bacteria by sequencing the naive (unselected) library after one round of bacterial amplification. The library of 2×10^{11} phages theoretically contains all of the 2×10^9 heptapeptide sequences. By sequencing millions of peptides, only a fraction of the entire library was analyzed. However, we confirmed that the peptide diversity in the library was high given that more than 50% of the peptide sequences in the naive library were found only once (Table 1). Nevertheless, some peptides were found at higher frequencies than expected by chance. The peptide HAIYPRH was found most frequently (36 times) and is known for its accelerated propagation in phage display experiments due to a mutation in the Shine–Dalgarno box of the phage protein gIIp [6]. Hence, growth advantages unrelated to the target selection emerge even after a single round of bacterial amplification of the naive phage library. A large number of additional peptides were found more than twice in the naive library, whereas the chance of this event based on a Poisson distribution with the current number of sequenced peptides would be only 5.5×10^{-9}. Presumably, these peptides also have growth advantages. We provide a list of these nearly 700,000 peptides and their frequencies in the naive library in Supplementary Table 1 (see supplementary material). It seems wise not to choose peptides that are found more than twice in the naive library for follow-up studies even when they demonstrate high enrichment after several rounds of selection.

Amino acid composition of peptides from naive and enriched libraries

The naive library is generated from a degenerate $(NNK)_7$ oligonucleotide library, where K represents an admixture of 50% T and 50% G. The expected amino acid frequencies resulting from this degenerate code is given in Fig. 2A. Already after sequencing 70 random clones, the manufacturer noticed considerable differences between the actual and expected amino acid compositions of the naive library (see New England Biolab's documentation). A much more refined distribution of amino acid frequencies was obtained after sequencing more than 6 million inserts (Fig. 2B and C). The naive library suffers from a considerable depletion of cysteine residues (frequency <1% of all amino acids). Glycine and arginine residues are also underrepresented but are still found at frequencies higher than 2.5%. Proline is the most overrepresented amino acid. After selection, these trends were even stronger (Fig. 2D), suggesting that cysteine residues impede, whereas proline residues enhance, the propagation of phages. Some major positional effects were also observed (see Supplementary Fig. 1 in supplementary material); the most important is the restriction of the overrepresentation of proline residues to positions 2 to 7, whereas the first position is actually depleted of proline residues, as noted previously [19].

Characterization of enrichment process

To identify phages with high binding affinity for undifferentiated and differentiated osteoblast (KS483) cells, we sequenced phage libraries extracted from the surface of the cells. Internalizing phages were identified by the sequencing of libraries extracted

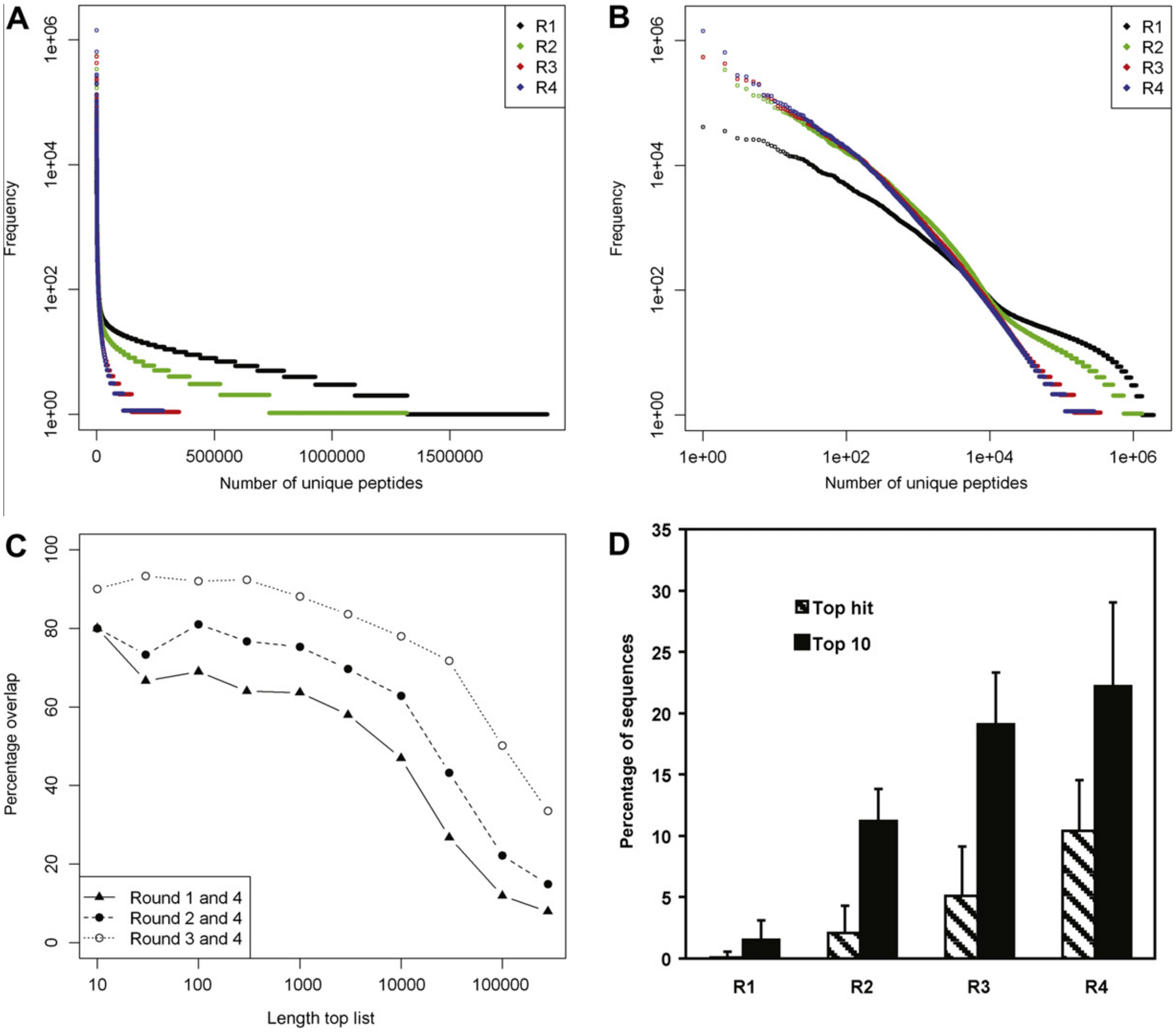


Fig.3. Comparison of round 1 (R1), round 2 (R2), round 3 (R3), and round 4 (R4) libraries. (A, B) Frequency distribution (A: linear scale; B: ^{10}log scale) of all unique peptide sequences (ordered from high to low abundance) after one (black), two (green), three (red), or four (blue) rounds of selection for phages internalizing into differentiated KS483 cells. The horizontal lines at the bottom of the plot represent the peptide sequences that occur once, twice, three times, and so on, with higher frequencies being more difficult to discriminate. The length of such a line represents the number of peptides with a particular frequency. (C) Percentage overlap in the lists of most abundant peptides obtained after different rounds of selection. (D) Simulation of traditional picking of 50 clones after round 1 (R1), round 2 (R2), round 3 (R3), and round 4 (R4). The percentages of sequences belonging to the most abundant phages (hashed bars) and 10 most abundant phages (black bars) identified by next generation sequencing of the round 4 library are shown. Standard deviations refer to the variance observed in 20 independent random picks of 50 clones. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

from cell lysates after removing surface-bound phages. We initially sequenced libraries after three or four rounds of selection (Table 1) because we observed a significant enrichment, based on titration of output/input ratios, after these rounds (data not shown). To further characterize the enrichment process and the potential for identification of interesting phages from earlier rounds of selection, we also analyzed the phage libraries isolated after one, two, three, and four rounds of selection for the internalization of phages on differentiated KS483 cells.

During subsequent rounds of selection, the overall diversity decreased while the frequency of the most enriched peptides steadily increased (Table 1 and Fig. 3A and B). To illustrate this, the number of unique sequences decreased from 1,913,785 in round 1 to 282,266 in round 4, whereas the frequency of the highest abundant peptide increased from 0.26% to 10% and the number of sequences that were found only once decreased from 594,379 to 170,592 (Fig. 3A and B). Thus, libraries converge toward certain peptide sequences in later rounds, consistent with expectations behind (traditional) phage display experiments.

There is a high correlation between the counts observed after the different selection rounds (Supplementary Fig. 2). In all comparisons, the correlations are high (Pearson correlation ranging from 0.47 [round 1 vs. round 2] to 0.94 [round 3 vs. round 4]). Moreover, the correspondence in the top 1,000 most abundant peptides between the different rounds is high, ranging from approximately 60% between rounds 1 and 4 to more than 80% between rounds 3 and 4 (Fig. 3C). Of the 10 most abundant peptides from round 4, 8 are also in the top 10 after the first or second selection round (Fig. 3C). This means that, with the current sequencing depth, (i) further rounds of selection will not lead to the identification of peptides that could not have been found in earlier rounds and (ii) high-affinity peptides can already be identified after the first round of selection.

We performed a simulation study to illustrate that this result is in striking contrast to a phage display experiment with traditional clone picking. In Fig. 3D, we show that, after the fourth round, approximately 10% (say, 5 of 50 clones picked) would be derived from the most abundant phage, as identified by sequencing of

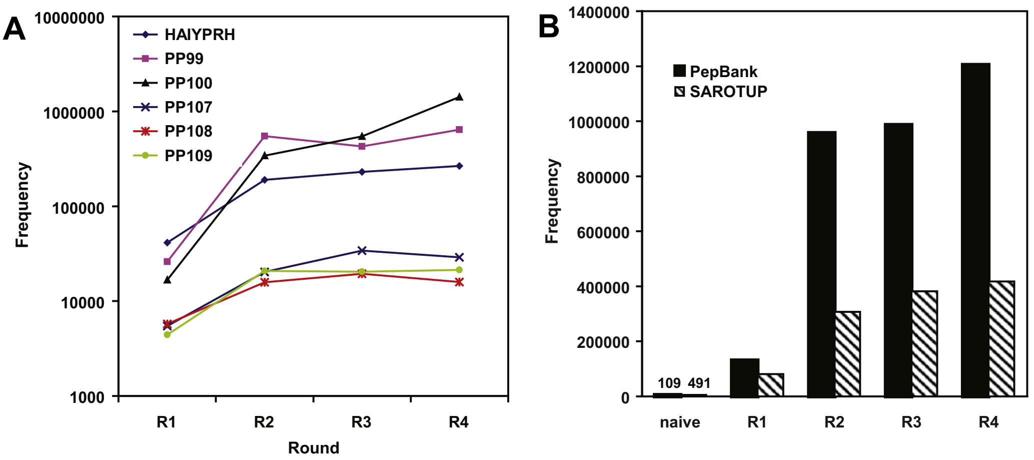


Fig.4. Counts of non-target-specific and target-specific peptide sequences during subsequent rounds of selection. (A) Counts of non-target-specific peptides HAIYPRH, PP99, and PP100 and of target-specific peptides PP107, PP108, and PP109 after four subsequent rounds of selection for phages internalizing into differentiated KS483 cells. (B) Percentages of sequences recorded in PepBank (black bars, database of short peptide sequences previously reported in literature) and SAROTUP (hashed bars, database of peptides that bind unintended targets) identified after sequencing of the naive library and the round 1 (R1), round 2 (R2), round 3 (R3), and round 4 (R4) libraries.

the complete round 4 library. A further 12% of sequenced clones (say, 6 of 50 clones picked) would be derived from other phages from the top 10 of the complete round 4 library. When sequencing only 50 clones, these top 10 peptides would most likely be sequenced only once, and some of them would not be sequenced at all. The remaining of the 50 randomly picked sequences (78%) are from phages that are not found in the top 10 after NGS analysis of the round 4 library. It is clear from Fig. 3D that these results are even worse when randomly picking 50 clones from the round 3 library, where only 5% of the sequences would be derived from the top phage. Sequencing 50 randomly selected phages after rounds 1 and 2 would be completely uninformative.

Identification of false positives

The selection of one sequence or a few sequences from a library of 10^{11} molecules is an inherently noisy process. There are two principal causes for the high abundance of irrelevant peptide sequences in phage display experiments. On the one hand, phages may have lower than average replication time in bacterial hosts, as illustrated by the HAIYPRH example. The impact of these propagation advantages increases with every round of selection and amplification. As an example, HAIYPRH was found 41,257 times after one round of selection and amplification (0.26% of all sequenced peptides) and 237,535 times after three rounds of selection (1.8%) (Fig. 4A). These false positives can be identified by sequencing of the naive library after one round of amplification in bacteria. On the other hand, peptides may bind to nontarget substrates and propagate during subsequent selection rounds without having any affinity to the intended target. These are less easy to identify because their enrichment pattern may be similar to that of target-specific peptides (Fig. 4A). A notorious example is the GETRAPL peptide identified in many independent experiments [7]. Most likely, the peptide binds to plastics in general given that it was found at a frequency of 35% in phage display selection for polystyrene binding peptides and demonstrated considerable affinity for polystyrene [20]. GETRAPL was the most abundant in 6 of 9 of our libraries selected on KS483 cells and among the top-ranked peptides in the other libraries. Libraries selected against in vivo targets did not have an overabundance of GETRAPL (data not shown), consistent with this peptide's presumed affinity for polystyrene plastics not present in in vivo screens. Two databases, PepBank [7] and SAROTUP [8], are very helpful in identifying non-target-specific peptides. Fig. 4B shows that subsequent

Table 2
Selected peptide sequences.

Group	Peptide sequence	Peptide number (PP)	Counts in naive library	Counts in day 4–surface	Counts in day 4–internalized	Counts in day 18–surface	Counts in day 18–internalized	SAROTUP	PepBank
0: Control peptides	LPLTPLP	98	16	13,006	29,175	20,988	20,003	Yes	Yes
	GETRAPL	99	15	1,279,913	397,669	3,599,322	427,153	Yes	Yes
	RHEPPLA	100	0	6	0	0	543,337	No	No[a]
1: Abundant in all selected libraries	AMSSRSL	101	0	15,301	54,629	11,517	134,645	No	No
	YRAPWPP	102	1	34,250	23,765	56,436	55,732	No	No
	ASSSHRS	103	0	73,075	5,033	57,476	24,902	No	No
2: Surface specific	DLKIPLR	104	0	37,802	0	55,342	59	No	No
	IEFSPLM	105	1	4,765	0	43,697	0	No	No
3: Day 4 specific (internalized)	NPWTTRP	106	0	0	50,155	0	0	No	No
4: Day 18 specific (internalized)	ALPQIVR	107	0	0	210	0	34,034	No	No
	DERHQHY	108	0	0	0	0	19,404	No	No
	WQSVPTI	109	0	0	0	0	20,390	No	No

Note. The following selection criteria were applied. Control peptides had more than two counts in the naive library and were included in PepBank or SAROTUP. Peptides in group 1 had more than 1,000 counts in all round 3 libraries. Peptides in group 2 had more than 1,000 counts in all surface libraries but not in internalized libraries. The peptide in group 3 had more than 1,000 counts, but only in the internalizing phages at day 4 of differentiation. Peptides in group 4 had more than 1,000 counts in internalizing phages at day 18 only.

[a] Mentioned as a polystyrene binder in Ref. [20].

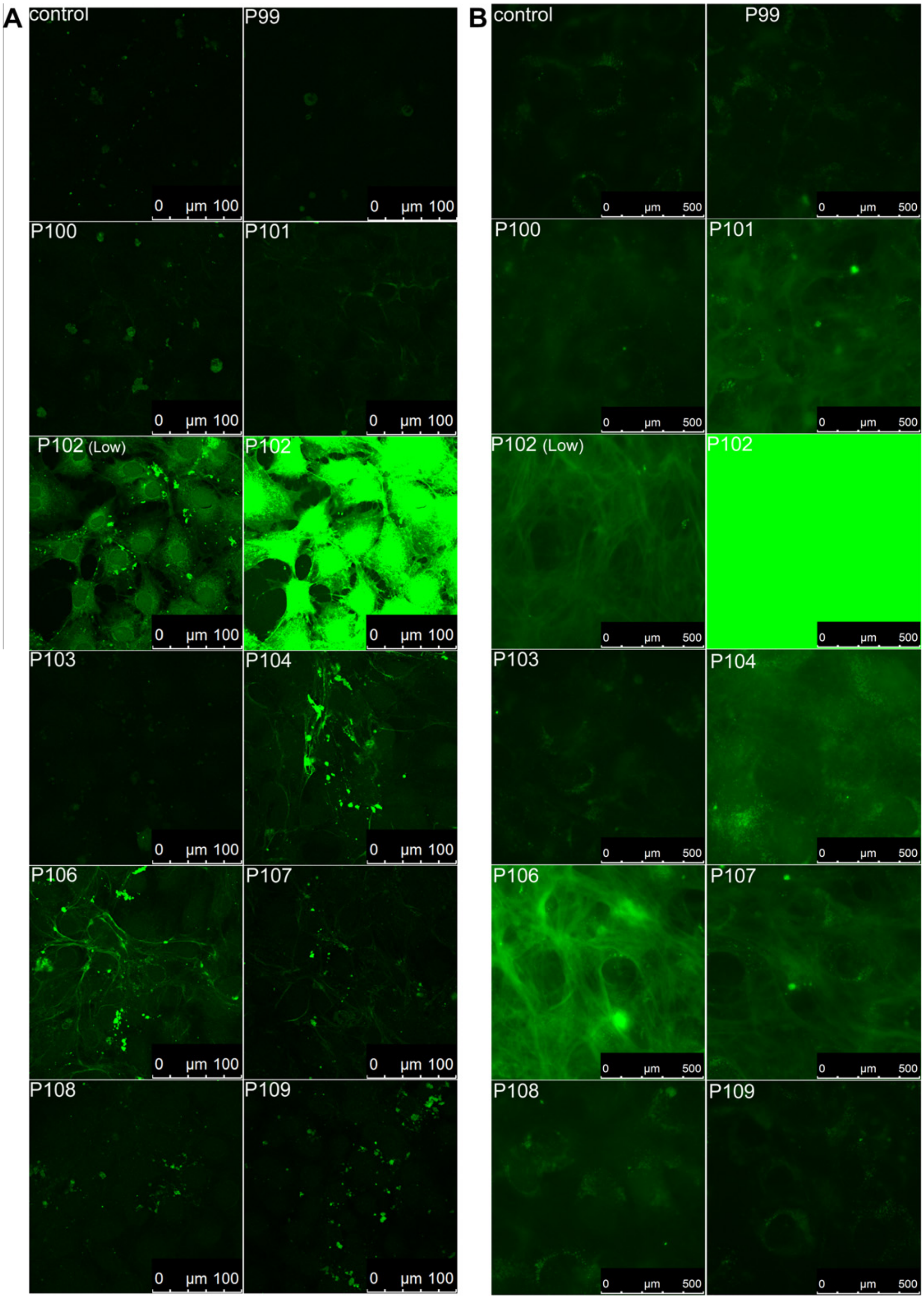


Fig.5. Uptake of fluorescently labeled peptides by KS483 cells. KS483 cells at day 4 (A) and day 18 (B) of differentiation were incubated with 2.25 µM FAM-labeled peptide in serum-free medium and analyzed with a confocal microscope (A) or by live cell imaging with an inverted microscope (B). Because the incubation with PP102 resulted in high uptake and very bright fluorescence, images obtained with two different gains (panel A: 528 [also used for the other peptides] and 424 [labeled with "Low"]) and exposure times (panel B: 220 ms [also used for the other peptides] and 25 ms ([labeled with "Low"]) are shown.

rounds of biopanning resulted in gradual enrichment for known target-unrelated peptides, amounting to approximately 10% of all sequences found after round 4. Through comparisons of many deep-sequenced libraries, it is possible to filter for commonly found parasitic peptides.

Selection of candidate peptide sequences

We were interested in identifying peptides that specifically bind to and/or are internalized by KS483 cells. We applied the filter criteria described below to identify candidate peptide sequences from the different libraries, and we searched for peptides with different apparent specificities. To exclude any putative false positives, we discarded all peptides with a frequency of 2 or higher in the naive library and hits in the PepBank and SAROTUP databases. From the remaining list of peptides, we selected 9 putative target-related peptides (PP101–PP109) with apparent differences in specificities (Table 2). PP101 to PP103 were most highly enriched in all libraries, whereas PP104 to PP109 showed at least 100-fold differences in abundance between libraries and were selected for their apparent specificities. PP104 and PP105 were selected as potential binders to the surface of cells at day 4 and day 18 of differentiated cells. We included peptides PP107 to PP109 with apparent specificity toward fully differentiated (day 18) cells. These peptides were gradually enriched during subsequent rounds of selection (Fig. 4A) and were consistently ranked in the top 120. No other phages in the top 120 displayed similar degrees of specificity for internalization in cells at day 18 of differentiation. As negative controls, we selected three non-target-related peptides (PP98–PP100). LPLTPLP (PP98) and GETRAPL (PP99) demonstrated a frequency of 2 or higher in the naive library and had a hit in PepBank. RHEPPLA (PP100) had been found to bind to polystyrene in a previous study [20] but was unexpectedly found only in the libraries selected for internalization at day 18, where it was the most highly enriched peptide (Table 1), and not in any of the other libraries. The presumed non-target-related peptides demonstrated similar enrichment profiles as PP107 to PP109 (Fig. 4A). Unfortunately, solubility of PP98 and PP105 was too limited to allow further analysis.

Fluorescent imaging of peptide binding and uptake by KS483 cells

Selected peptides, equipped with a fluorescent FAM label, were tested for binding to and uptake by KS483 cells at day 4 and day 18 of differentiation. The experiments with cells at day 4 of differentiation were performed three times with duplicated wells in each experiment, using cells with high and low passage numbers and incubation times of 24 and 48 h, all generating similar results. Day 4 cells were analyzed with confocal microscopy to be able to discriminate between surface binding and internalizing peptides. Confocal microscopy was not possible for day 18 cells because the cells grow on top of each other and deposit thick matrix structures. Therefore, day 18 cells were analyzed by live cell imaging. Representative fluorescent images obtained with the two different technologies are shown in Fig. 5 and Supplementary Fig. 3. As expected, cells incubated with the PP99 control peptide, which has been found in many other phage display selection experiments according to SAROTUP and PepBank, do not show any fluorescent signal. Similarly, PP100, which was found to bind to polystyrene previously [20], demonstrated only very weak cellular staining. PP102, which was highly abundant in all selected libraries, indeed displayed strong binding and uptake in day 4 and day 18 cells. Spectrophotometric analysis of the cell lysates confirmed that the majority of the presented peptides were internalized (data not shown). The intracellular distribution of PP102 differed between day 4 and day 18 cells, with more prominent staining of the nucleus on day 4 and more prominent staining of the extracellular matrix on day 18. Surprisingly, PP101 and PP103, also abundant in all selected libraries, displayed very weak or no signal. In line with the selection of PP104 as a surface binding peptide, PP104 demonstrated strong binding to the extracellular matrix. A similar staining pattern was observed for PP106, which was unanticipated given that this peptide sequence was discovered in the pool of internalized phages. Peptides PP107 and PP108 were identified in day 18 cells only and, indeed, appeared to give stronger staining in day 18 cells than in day 4 cells. Data for PP109 were inconclusive because the peptide was highly aggregation prone.

Discussion

In the studies described in the current article, we have gained significant insight into the phage display selection process. This was achieved by sequencing millions of independent phages after different selection rounds using the latest NGS technology. We showed that the enrichment factor of positive clones gradually increases during subsequent rounds of selection. We did not observe differences in the enrichment kinetics between target-specific and target-unrelated peptides. This is concordant with the observed absence of a correlation between peptide affinity and abundance in other reported phage display experiments [21].

Multiple rounds of selection will be helpful in reducing the background of nonbinders and, therefore, is essential when sequencing a limited number of clones (cf. our simulation study in Fig. 3D). In contrast, significant enrichment can already be observed after the first round of selection when sequencing millions of clones. The high correlation between abundances after the first round and subsequent rounds suggests that further rounds of enrichment might not have additional value. The large effects of small differences in propagation rates on the final abundances, nicely illustrated in Ref. [21], may even result in exclusion or downweighting of interesting phages from the pool. Instead of performing multiple rounds of selection, it may be preferred to perform the first round of selection in duplicate or triplicate to account for variability in the stringency of binding conditions. Thus, our results question the validity of the traditional phage display screening approach; the high number of selection rounds required to arrive at peptides observed multiple times in tens to a few hundred of sequenced clones will come with the risk of an overreduction in the library complexity and the loss of potentially interesting phages. Moreover, NGS-based characterization of phage display libraries is more cost-effective than traditional clone picking because of the lower number of selection rounds required. It is also less laborious than methods previously developed to mitigate amplification biases employing phage amplification in isolated compartments such as monodisperse droplets [22].

Although full characterization of the round 1 library would require sequencing more than 10 million reads, identification of the most abundant phages requires substantially lower sequencing depth. The required sequencing depth depends on the complexity of the library and the enrichment factor. Generally speaking, sequencing of 1 million reads should be sufficient with relatively low enrichment factors obtained after one round of biopanning given that they would be represented more than 50 times according to simulated random draws of 1 million sequences from our round 1 library. This would allow a high level of multiplexing of different phage display libraries in a single lane of the Illumina sequencer and, thus, further reduce sequencing costs.

Previous characterization of phage display libraries were done with Roche 454 sequencing. Dias-Neto and coworkers sequenced approximately 70,000 phages isolated from human tissues after phage infusion [11]. In their experiment, the number of unique phages recovered was much lower than in our experiment—probably due to the distribution of the phages over different organs and the small biopsies taken—and sequencing of approximately 40,000 phages was sufficient to fully characterize the library. In an independent study, a cDNA phage display library selected for binding to transglutaminase was characterized by Roche 454 sequencing [23]. After sequencing of 120,000 phages, no decrease in diversity compared with the naive library was observed. This may have been caused by a lack of high-affinity binders in the phage pool and/or an insufficient sequencing depth. Illumina sequencing technology, like SOLiD and Helicos technologies, provides millions of reads.

This may be an important reason why we observed a clear decrease in library complexity already after the first round of sequencing.

It is not yet feasible to sequence the complete naive library given that it is claimed to contain 2 billion different phages. We sequenced a fraction of approximately 0.3%. Still, the sequencing of the naive library has proven to be extremely useful for two reasons. First, it revealed nonrandom occurrences of the different amino acids at the different positions. We confirmed the findings by Krumpe and coworkers obtained by random sequencing of 100 to 400 phages from similarly constructed CX7C and X12 M13 phage display libraries [19], and we demonstrated that proline residues are overrepresented at all positions except the first position and that cysteine residues are significantly underrepresented. Based on the fact that this was already apparent in the naive library, which went through one round of amplification in bacteria, we tend to explain this by differences in availability of the amino acids, nonrandom incorporation probabilities during translation, and suboptimal codon use in bacteria. The nonrandom representation of amino acids reduces the chance to identify the highest affinity binders; therefore, some optimization of the identified peptide sequences may still be useful. Second, sequences that were already found at higher frequencies than expected by chance are likely to have selective and/or growth advantages. Indeed, the GETRAPL peptide (PP99), which was already found 15 times in the naive library, was not binding to KS483 cells or effective as a viral targeting peptide [24]. The lack of signal in our plastic dishes also suggests that GETRAPL is not a polystyrene binding peptide as claimed previously [20]. Its identification in a screening for plastic binding phages may have been caused by the same selection advantage. Moreover, the PCR amplification step applied during sample preparation may introduce biases that would be revealed by sequencing of the naive library. Therefore, we strongly recommend sequencing naive phage display libraries after one round of amplification in bacteria.

Despite our elimination of possible false positives caused by amplification or PCR biases or nonspecific binding, other factors appear to contribute to the high false positive rate in peptide phage display screenings and were also observed in our experiment. A likely explanation for this is that the binding of synthetic 7-mer peptides may be largely different from the binding of a phage displaying a 7-mer peptide sequence as part of the phage pIII protein, giving rise to substantial differences in secondary and tertiary structures.

We identified one peptide (PP102) with exceptionally strong binding and uptake by KS483 cells and two peptides with strong binding to the extracellular matrix. Further research should demonstrate the potential of these peptides to target bioactive agents or drugs to osteoblasts and/or to shuttle these across the plasma membrane. PP102 was ranked 14 after the first round of selection and remained in the top 20 during subsequent rounds of selection. Interestingly, there were two independent phage clones coding for the PP102 amino acid sequence; TATAGGGCTCCTTGGCCGCCT was the most prominent one, followed by TATCGGGCTCCTTGGCCGCCT, which was observed at frequencies of approximately 10% of the former. The other selected peptides, which displayed less strong binding to KS483 cells, were not represented by multiple independent phages as these DNA variants were present at frequencies less than 1%. This suggests that the identification of independent phages displaying the same peptide has positive predictive value for the affinity of the peptide. Clearly, more research evaluating larger sets of peptides needs to be done to confirm this.

The 1% threshold is our current estimate of the total sequencing error rate for a 21-mer sequence. This estimate is based on the steep increase of the number of DNA variants with a one mismatch difference at frequencies less than or equal to 1% of the more abundant sequence. Another positive predictive feature would be the identification of related peptides with slightly different amino acid compositions. Single amino acid substitutions at frequencies lower than 1% can be explained by sequencing errors and should not be considered. However, for PP102 we found a second peptide, SRAPWPP, differing by just one amino acid, at DNA sequence-based frequencies ranging between 3% and 20% of the frequency of YRAPWPP (PP102). It proved to be difficult to perform a systematic search for clusters of related peptide sequences because existing peptide sequence clustering software is not designed for the clustering of very large sets of very short peptides.

In our experiment, reads of only 21 nucleotides were required to identify unknown 7-mer peptide sequences. Other types of phage display libraries, such as cDNA and antibody phage display libraries, are also amenable to NGS-based characterization, aided by the much longer read length currently obtained with next generation sequencers. Likewise, NGS was recently applied to characterize the antibody variable regions of bone marrow plasma cells, and this proved to be a very efficient method for monoclonal antibody selection [13]. Taken together, NGS can significantly empower phage display screenings and accelerate the finding of specific binders while restraining the number of false positive hits.

Acknowledgments

This research was supported by a grant from the Dutch Ministry of Economic Affairs (IOP–Genomics grant IGE7001), the Centre for Medical Systems Biology within the framework of the Netherlands Genomics Initiative (NGI)/Netherlands Organisation for Scientific Research (NWO), and the Center for Biomedical Genetics. We thank Annelies Boonzaier-van der Laan (Department of Molecular and Cellular Biology, Leiden University Medical Center) for microscopy assistance; Ineke Oudshoorn (Department of Endocrinology, Leiden University Medical Center), Karien de Rooij (Department of Endocrinology, Leiden University Medical Center, and Percuros), and David. de Gorter (Department of Molecular and Cellular Biology, Leiden University Medical Center) for assistance with cell culturing; and Rinse Klooster (Department of Molecular and Cellular Biology, Leiden University Medical Center) for comments on the manuscript.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ab.2011.11.005.

References

[1] G.P. Smith, Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface, Science 228 (1985) 1315–1317.
[2] T. Bratkovic, Progress in phage display: evolution of the technique and its application, Cell Mol. Life Sci. 67 (2010) 749–767.
[3] Y. Georgieva, Z. Konthur, Design and screening of m13 phage display cDNA libraries, Molecules 16 (2011) 1667–1681.
[4] C. Hamers-Casterman, T. Atarhouch, S. Muyldermans, G. Robinson, C. Hamers, E.B. Songa, N. Bendahman, R. Hamers, Naturally occurring antibodies devoid of light chains, Nature 363 (1993) 446–448.
[5] M. Vodnik, U. Zager, B. Strukelj, M. Lunder, Phage display: selecting straws instead of a needle from a haystack, Molecules 16 (2011) 790–817.
[6] L.A. Brammer, B. Bolduc, J.L. Kass, K.M. Felice, C.J. Noren, M.F. Hall, A target-unrelated peptide in an M13 phage display library traced to an advantageous mutation in the gene II ribosome-binding site, Anal. Biochem. 373 (2008) 88–98.
[7] T. Shtatland, D. Guettler, M. Kossodo, M. Pivovarov, R. Weissleder, PepBank—a database of peptides based on sequence text mining and public peptide data sources, BMC Bioinformatics 8 (2007) 280.
[8] J. Huang, B. Ru, S. Li, H. Lin, F.B. Guo, SAROTUP: scanner and reporter of target-unrelated peptides, J. Biomed. Biotechnol. (2010), doi:10.1155/2010/101932.
[9] D.R. Bentley, Whole-genome re-sequencing, Curr. Opin. Genet. Dev. 16 (2006) 545–552.
[10] M. Margulies, M. Egholm, W.E. Altman, S. Attiya, J.S. Bader, L.A. Bemben, J. Berka, M.S. Braverman, Y.J. Chen, Z. Chen, S.B. Dewell, L. Du, J.M. Fierro, X.V.

Gomes, B.C. Godwin, W. He, S. Helgesen, C.H. Ho, G.P. Irzyk, S.C. Jando, M.L. Alenquer, T.P. Jarvie, K.B. Jirage, J.B. Kim, J.R. Knight, J.R. Lanza, J.H. Leamon, S.M. Lefkowitz, M. Lei, J. Li, K.L. Lohman, H. Lu, V.B. Makhijani, K.E. McDade, M.P. McKenna, E.W. Myers, E. Nickerson, J.R. Nobile, R. Plant, B.P. Puc, M.T. Ronan, G.T. Roth, G.J. Sarkis, J.F. Simons, J.W. Simpson, M. Srinivasan, K.R. Tartaro, A. Tomasz, K.A. Vogt, G.A. Volkmer, S.H. Wang, Y. Wang, M.P. Weiner, P. Yu, R.F. Begley, J.M. Rothberg, Genome sequencing in microfabricated high-density picolitre reactors, Nature 437 (2005) 376–380.

[11] E. Dias-Neto, D.N. Nunes, R.J. Giordano, J. Sun, G.H. Botz, K. Yang, J.C. Setubal, R. Pasqualini, W. Arap, Next-generation phage display: integrating and comparing available molecular tools to enable cost-effective high-throughput analysis, PLoS One 4 (2009) e8338.

[12] U. Ravn, F. Gueneau, L. Baerlocher, M. Osteras, M. Desmurs, P. Malinge, G. Magistrelli, L. Farinelli, M.H. Kosco-Vilbois, N. Fischer, By-passing in vitro screening—Next generation sequencing technologies applied to antibody display and in silico candidate selection, Nucleic Acids Res. 38 (2010) e193.

[13] S.T. Reddy, X. Ge, A.E. Miklos, R.A. Hughes, S.H. Kang, K.H. Hoi, C. Chrysostomou, S.P. Hunicke-Smith, B.L. Iverson, P.W. Tucker, A.D. Ellington, G. Georgiou, Monoclonal antibodies isolated without screening by analyzing the variable-gene repertoire of plasma cells, Nat. Biotechnol. 28 (2010) 965–969.

[14] M.M. Deckers, M. Karperien, C. van der Bent, T. Yamashita, S.E. Papapoulos, C.W. Lowik, Expression of vascular endothelial growth factors and their receptors during osteoblast differentiation, Endocrinology 141 (2000) 1667–1674.

[15] Z.C. Dang, R.L. van Bezooijen, M. Karperien, S.E. Papapoulos, C.W. Lowik, Exposure of KS483 cells to estrogen enhances osteogenesis and inhibits adipogenesis, J. Bone Miner. Res. 17 (2002) 394–405.

[16] D.J. de Gorter, M. van Dinther, O. Korchynskyi, P. ten Dijke, Biphasic effects of transforming growth factor beta on bone morphogenetic protein-induced osteoblast differentiation, J. Bone Miner. Res. 26 (2011) 1178–1187.

[17] P.A. 't Hoen, Y. Ariyurek, H.H. Thygesen, E. Vreugdenhil, R.H. Vossen, R.X. de Menezes, J.M. Boer, G.J. van Ommen, J.T. den Dunnen, Deep sequencing-based expression analysis shows major advances in robustness, resolution, and inter-lab portability over five microarray platforms, Nucleic Acids Res. 36 (2008) e141.

[18] T. Yamashita, H. Ishii, K. Shimoda, T.K. Sampath, T. Katagiri, M. Wada, T. Osawa, T. Suda, Subcloning of three osteoblastic cell lines with distinct differentiation phenotypes from the mouse osteoblastic cell line KS-4, Bone 19 (1996) 429–436.

[19] L.R. Krumpe, A.J. Atkinson, G.W. Smythers, A. Kandel, K.M. Schumacher, J.B. McMahon, L. Makowski, T. Mori, T7 lytic phage-displayed peptide libraries exhibit less sequence bias than M13 filamentous phage-displayed peptide libraries, Proteomics 6 (2006) 4210–4222.

[20] T. Serizawa, P. Techawanitchai, H. Matsuno, Isolation of peptides that can recognize syndiotactic polystyrene, ChemBioChem 8 (2007) 989–993.

[21] R. Derda, S.K. Tang, S.C. Li, S. Ng, W. Matochko, M.R. Jafari, Diversity of phage-displayed libraries of peptides during panning and amplification, Molecules 16 (2011) 1776–1803.

[22] R. Derda, S.K. Tang, G.M. Whitesides, Uniform amplification of phage with different growth characteristics in individual compartments consisting of monodisperse droplets, Angew. Chem. Int. Ed. Engl. 49 (2010) 5301–5304.

[23] R. Di Niro, A.M. Sulic, F. Mignone, S. D'Angelo, R. Bordoni, M. Iacono, R. Marzari, T. Gaiotto, M. Lavric, A.R. Bradbury, L. Biancone, D. Zevin-Sonkin, G. De Bellis, C. Santoro, D. Sblattero, Rapid interactome profiling by massive sequencing, Nucleic Acids Res. 38 (2010) e110.

[24] L.M. Work, S.A. Nicklin, N.J. Brain, K.L. Dishart, D.J. Von Seggern, M. Hallek, H. Buning, A.H. Baker, Development of efficient viral vectors selective for vascular smooth muscle cells, Mol. Ther. 9 (2004) 198–208.