In vitro compartmentalization and deep sequencing enable discovery of functional cell-binding ligands by phage display

by

Wadim L. Matochko

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Chemistry
University of Alberta

Abstract

Phage display is a technique that accelerated the discovery of peptide and protein-based ligands to numerous targets in academia and industry. Many FDA-approved antibodies and peptides on the market have originated from phage display experiments. However, one of the main drawbacks to this technique is that there are two independent steps in the selection process, an enrichment step for ligands with binding affinity towards a target and an amplification step to amplify the enriched phage clones. The amplification step introduces a bias towards enriching phage clones with a different phenotype based on growth rate. The implication of this bias is that target-binding ligands are enriched from a small subset, with low diversity, of the phage library that contains fast growing phage clones. Furthermore, selection from the fast growing phage population gives rise to a large number of target unrelated ligands. This undesired amplification bias is especially detrimental when selecting for ligands against multi-site binding target, such as cell or tissues. In this thesis, we examine whether the collapse of diversity can be prevented in phage libraries by amplifying these libraries in emulsions. We show that preventing the diversity collapse, we can identify more ligands than the standard selection method and speed up the discovery of ligands to targets with multiple binding sites.

This thesis first describes the development of the emulsion amplification technique. We describe the manufacturing of the microfluidic devices and synthesis of the perfluoro-surfactant needed to maintain stable emulsions throughout the amplification process. To analyze the amplification process, we

develop a method for deep-sequencing of phage display libraries using Illumina and Ion Torrent platforms, as well as MATLAB scripts, to analyze deep-sequencing data. We applied deep sequencing to examine how diversity of peptides in phage display libraries changes as a result of amplification of libraries in bacteria. Using a Ph.D.-12 library as our model library, we observed that amplification enriches ~150 clones, which dominate ~20% of the library. Deep sequencing, for the first time, characterized the collapse of diversity in phage libraries.

We extend the use of next-generation sequencing to characterize the Ph.D.-7 library. Using Illumina and Ion Torrent sequencing and multiple biological replicates of amplification of the Ph.D.-7 library, we identified a focused population of 770 sequences that grow quickly, we term these sequences 'parasites'. In all, 197 sequences from this population have been identified in literature reports that used Ph.D.-7 library. Many of these enriched sequences have confirmed function (*e.g.*, target binding capacity). The bias in the literature, thus, can be viewed as a selection with two different selection pressures: i) target-binding selection, and ii) amplification-induced selection. Enrichment of parasitic sequences could be minimized if amplification bias is removed. Here, we demonstrate that emulsion amplification in libraries of ~10^6 diverse clones prevents the selection of parasitic clones.

We examine if emulsion-amplification can prevent enrichment of parasitic clones in selection against a multi-site target, here we use MDA-MB-231 breast cancer cells. We perform selection using the standard method to amplify phage

libraries in one common bulk solution (bulk amplification). We reproducibly identified peptide ligands for breast cancer cells from a ~0.0001% sub-population of the library, which harbors fast-growing, “parasitic” phage. Replacing bulk with emulsion-amplification dramatically altered the selection landscape and yielded ligands from the regions of the library not accessible to bulk-amplification selection by preventing diversity collapse during amplification. We propose incorporating emulsion-amplification into selection against multi-site targets (cells, antibody mixtures, etc.), can lead to the discovery of ligands missed by conventional selection strategies.

Preface

In Chapter 1 of this thesis, Figures 1.1 was adapted from the following publication: W.L. Matochko, S. Ng, M.R. Jafari, J. Romaniuk, S.K.Y. Tang, and R. Derda, "Uniform amplification of phage display libraries in monodisperse emulsions", *Methods*, 2012, **58**: 18-27. Figures 1.4-1.5 were adapted from the following publication: R. Derda, S.K.Y. Tang, S.C. Li, S. Ng, W. Matochko, and M.R. Jafari, "Diversity of phage-displayed libraries of peptides during panning and amplification", *Molecules*, 2011, **16**, 1776-1803. Figure 1.7 was adapted from the book chapter: W.L. Matochko, R. Derda, "Next generation sequencing of phage displayed peptide libraries". In *Peptide Libraries: Methods in Molecular Biology*; R. Derda, Ed.; Springer: New York, 2015, **1248**, 249-266.

Chapter 2 has been published as W.L. Matochko, S. Ng, M.R. Jafari, J. Romaniuk, S.K.Y. Tang, and R. Derda, "Uniform amplification of phage display libraries in monodisperse emulsions", *Methods*, 2012, **58**: 18-27. I established a methodology for encapsulating and amplifying phage in perfluoro-water emulsions based on the protocol originally published by my advisor, Ratmir Derda, (*Angew. Chem. Intl. Ed.* **2010** 49: 5301-5304). I optimized the emulsion amplification (EmA) protocol by changing the size of the library, the amplification cycle, and recovery of phage. S. Ng, M.R. Jafari, and J. Romaniuk aided in the synthesis and characterization of the perfluoro-surfactant used to stabilize the emulsions. I prepared all figures and partially contributed to the writing of the manuscript.

Chapter 3 has been published as W.L. Matochko, K. Chu, B. Jin, S.W. Lee, G.M. Whitesides, and R. Derda, “Deep sequencing analysis of phage libraries using Illumina platform”, *Methods*, 2012, **58**: 47-55. I established methodology to amplify the variable region of phage display libraries to produce DNA compatible with next generation sequencing (NGS). I first optimized it for the Illumina platform and then Ion Torrent NGS platform. K. Chu developed the first protocol for the preparation of dsDNA for Illumina sequencing. I partially contributed to the preparation of figures and writing of the manuscript. Additionally, I developed the sequencing protocol for the Ion Torrent NGS platform and subsequently re-designed the preparation steps to streamline the PCR amplification and purification protocol. I also developed a new MATLAB script for the processing of data from this NGS. I contributed to all experimental, figure preparation, and writing of this work and is used in subsequent articles and summarized in a book chapter: W.L. Matochko, R. Derda, “Next generation sequencing of phage displayed peptide libraries”. In *Peptide Libraries: Methods in Molecular Biology*; R. Derda, Ed.; Springer: New York, 2015, **1248**, 249-266.

Chapter 4 has been published as W.L. Matochko, S.C. Li, S.K.Y. Tang, and R. Derda, "Prospective identification of parasite sequences in phage display screens" *Nuc. Acid. Res.*, 2014, **42**: 1784-1798. I combined EmA and NGS of phage libraries to illustrate that different phage libraries have populations of fast growing phage clones (“parasites”) which can be traced in phage display literature to-date. I also showed that EmA prevents “parasites” from taking over the library pool. My contribution to the article in *Nucleic Acids Res*earch, includes all experimental work, and processing of the sequencing data. I also partially contributed to the analysis and figure preparation.

Chapter 5 has not been submitted for publication yet. In this chapter, I incorporated EmA into the selection of peptide ligands for receptors on the cell surface of the breast cancer cell line, MDA-MB-231. The manuscript demonstrates that conventional selection identifies ligands from a "parasite" subset of the library. Introducing EmA into the selection procedure allowed the selection of binding ligands from a more diverse set of sequences. My contribution to this work includes all experimental work, processing of the all data, figure and manuscript preparation. I partially contributed to the writing of Matlab processing scripts, with the help of fellow co-authors, Frédérique Deiss and Ratmir Derda.

The Appendix contains two additional published manuscripts. The first manuscript was published as F. Deiss, W.L. Matochko, N. Govindasamy, E.Y. Lin, and R. Derda, "Flow-through synthesis on teflon-patterned paper to produce peptide arrays for cell-based assays". Angew. Chem. Int. Ed., **2014**, 53: 6374-6377. This manuscript describes the patterned deposition of Teflon on paper to create solvent-repelling barriers for parallel organic synthesis and cell-based assays. My contribution includes assisting a post-doctoral fellow, Frédérique Deiss, in designing and optimizing a paper platform in which to perform SPOT synthesis. I developed and performed all cell-adhesion assay experiments and prepared figures that summarized the results. The second manuscript was published as W.L. Matochko and R. Derda, "Error analysis of deep sequencing of phage libraries: peptides censored in sequencing", Comput. Math Methods Med., **2013**: 491612. This manuscript describes error analysis in deep-sequencing of a Ph.D.-7 library sequenced by Illumina. My contribution includes the acquisition of all deep-sequencing data and proof-reading the paper.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

BSA	bovine serum albumin
CH_2Cl_2/DCM	dichloromethane
cfu	colony forming unit
DIAD	diisopropyl azodicarboxylate
DMF	dimethylformamide
DMSO	dimethyl sulfoxide
dsDNA	double stranded deoxyribonucleic acid
E. coli	*Escherichia coli*
ELISA	enzyme linked immunosorbant assay
ESI	electron spray ionization
EtOH	ethanol
FBS	fetal bovine serum
F pilus	fertility pilus
FTIR	Fourier transform infrared spectroscopy
GFP	green fluorescent protein
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
LB	lysogeny broth
LC-MS	liquid chromatography mass spectroscopy
IPTG	isopropyl β-D-1-thiogalactopyranoside
IR	infrared spectroscopy
MALDI	matrix-assisted laser desorption/ionization
MEM	minimum essential media

MeOH	methanol
MWt	molecular weight
NMR	nuclear magnetic resonance
PCR	polymerase chain reaction
PDMS	polydimethylsiloxane
PEG	poly(ethylene glycol)
pfu	plaque forming units
PolyDMAP	poly 4-dimethylaminopyridine
RNA	ribonucleic acid
SDS-page	sodium dodecyl sulfate polyacrylamide gel electrophoresis
ssDNA	single-stranded deoxyribonucleic acid
TBE	tris/borate/EDTA (buffer solution)
TBS	tris-buffered saline (buffer solution)
TFA	trifluoroacetic acid
THF	tetrahydrofuran
X-gal	5-bromo-4-chloro-3-indolyl β-D-galactopyranoside

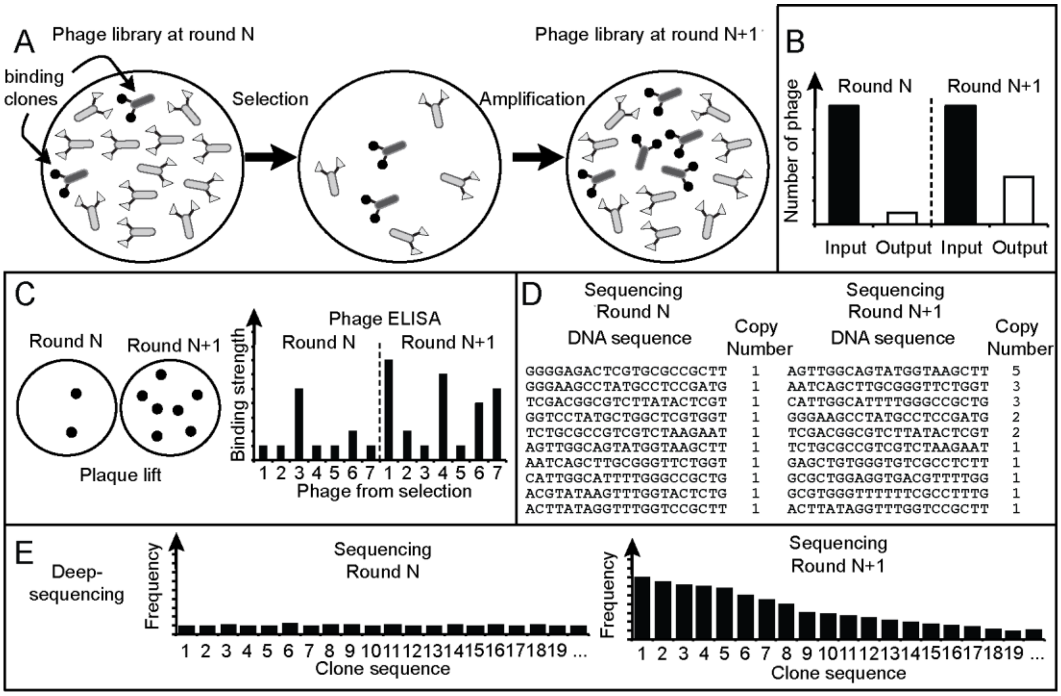
Chapter 1 General Introduction

1.1 Overview and premise

Phage display is a powerful technology for discovering polypeptide-based ligands for targets, such as proteins, cells and tissues, and even inorganic materials[1-6]. Phage display library technology identifies these ligands by testing billions of random polypeptides and narrowing this vast population through selection to a few ligands with useful binding properties. This process, which we refer to as “diversity collapse”, can be productive and yield a collapsed population of enriched useful binders, but in many cases, it can be unproductive and yield a significant fraction of false positives. Unproductive diversity collapse is difficult to identify and characterize. In this introductory chapter, we review the selection process combined with deep-sequencing (a.k.a. deep-panning[7]) and we describe how deep-sequencing can aid in the detection, characterization, and design of strategies to bypass the unproductive diversity collapse.

There are two key steps involved in phage display selection: i) the panning step, during which phage clones are captured on an immobilized target and selects for ligands that have affinity towards a target, and ii) the amplification step, during which phage infect bacteria and replicate inside bacteria, this expands the number of phage clones to be used for further rounds of selection (Figure 1.1A). Successive rounds are performed until selection has enriched binding ligands. There are several methods to determine the success in enriching binding ligands through one or a combination of the following methods. One method is to compare the average binding capacity to the target in a phage population before and after selection. As the target-binding population is enriched, the fraction of library that could be captured by the target increases significantly (Figure 1.1B). A second method is to measure binding capacity of individual clones in the library (Figure 1.1C). This method cannot access all the ligands in the library; instead it extrapolates the property of the library based on the sub-population of the selected library. An alternative method for assessment of diversity collapse is to observe the sequence composition of the library after each round of selection (Figure 1.1D). Typically, 10 to 100 clones are analyzed using Sanger sequencing. Target-binding ligands are identified as those that are found in multiple copies[8]. This method is very simple, when compared to the other analyses and this simplicity made sequencing-based assessment one of the most widely used methods for determining the success of the selection. Sequencing of < 100 clones from a population that contains > 10^5 diverse sequences provides a very shallow representation of the phage population. Conclusions from such analyses can be misleading, as we will show below. Incorporating deep-sequencing allows for analysis of enrichment and identification of consensus motifs with higher precision compared to the shallower Sanger sequencing method[9].

The consequence of performing two independent steps in the selection process introduces two orthogonal biases into the selection process: i) panning selects for phage clones that contain ligands with binding preference toward the target, while ii) amplification selects for phage clones that amplify faster than others, irrespective of the displayed sequence (Figure 1.2)[10]. This growth bias



Sequencing Round N DNA sequence	Copy Number	Sequencing Round N+1 DNA sequence	Copy Number
GGGGAGACTCGTGCGCCGCTT	1	AGTTGGCAGTATGGTAAGCTT	5
GGGAAGCCTATGCCTCCGATG	1	AATCAGCTTGCGGGTTCTGGT	3
TCGACGGCGTCTTATACTCGT	1	CATTGGCATTTTGGGCCGCTG	3
GGTCCTATGCTGGCTCGTGGT	1	GGGAAGCCTATGCCTCCGATG	2
TCTGCGCCGTCGTCTAAGAAT	1	TCGACGGCGTCTTATACTCGT	2
AGTTGGCAGTATGGTAAGCTT	1	TCTGCGCCGTCGTCTAAGAAT	1
AATCAGCTTGCGGGTTCTGGT	1	GAGCTGTGGGTGTCGCCTCTT	1
CATTGGCATTTTGGGCCGCTG	1	GCGCTGGAGGTGACGTTTTGG	1
ACGTATAAGTTTGGTACTCTG	1	GCGTGGGTTTTTTCGCCTTTG	1
ACTTATAGGTTTGGTCCGCTT	1	ACTTATAGGTTTGGTCCGCTT	1

Figure 1.1. (A) Schematic representation for the enrichment of target-binding phage clones in one round of selection. In the initial round of selection (N), a phage library contains target-binding phage clones in a population of non-target binding phage clones. After selection, the target binding phage clones remain part of the library population, while the non-target binding phage clones are largely removed from the library population. Amplification increases the number of all remaining phage clones in the library. In the next round of selection (N+1) the phage library contains an enriched population of target binding phage clones. The rounds are repeated until selection is considered to be successful. The success of selection can be determined through one or a combination of four ways: i) Monitoring the average number of phage before (Input) and after (Output) selection (B); ii) Monitoring binding capacity of pooled phage or individual clones from a sub-population of the phage library either by plaque lift[11] or ELISA[12] (C); iii) Identifying enriched DNA sequences using Sanger sequencing[8]

(D). iv) Identifying phage clones that have increased in frequency (number of reads of each clone / total number of reads) using deep-sequencing[13].

leads to diversity loss, and may present a problem in selection for a single target. Literature reports have characterized specific sequences (*e.g.* GKPMPPM[14-16]) that have high growth rates and were identified from multiple screens to different targets[17,18]. This growth bias would prevent phage library selections from identifying target-binding ligands outside of a small subset of fast growing clones. In Chapter 5, we show that this collapse is most detrimental when the target has multiple binding sites. Examples of such targets are cells or mixtures of proteins (such as antibodies isolated from whole serum).

Our aim is to adapt phage library selection to identify ligands that bind to all receptors on cancer cells. However, in order to identify a large and diverse set of ligands, we first need to understand diversity collapse and overcome the undesired growth bias that takes place during a phage library screening. The research described in this thesis describes improving the amplification process to avoid amplification bias in the selection process. We draw inspiration from the widespread use of emulsion, first used in directed evolution[19] and emulsion PCR[20-22], then expanded to encapsulate bacteria[23,24] and mammalian cells[25-27], in mono-disperse emulsions. We apply this method to the amplification of phage libraries, and describe the following: 1) Application of emulsion amplification to prevent competition between fast and slow growing phage clones. 2) Deep

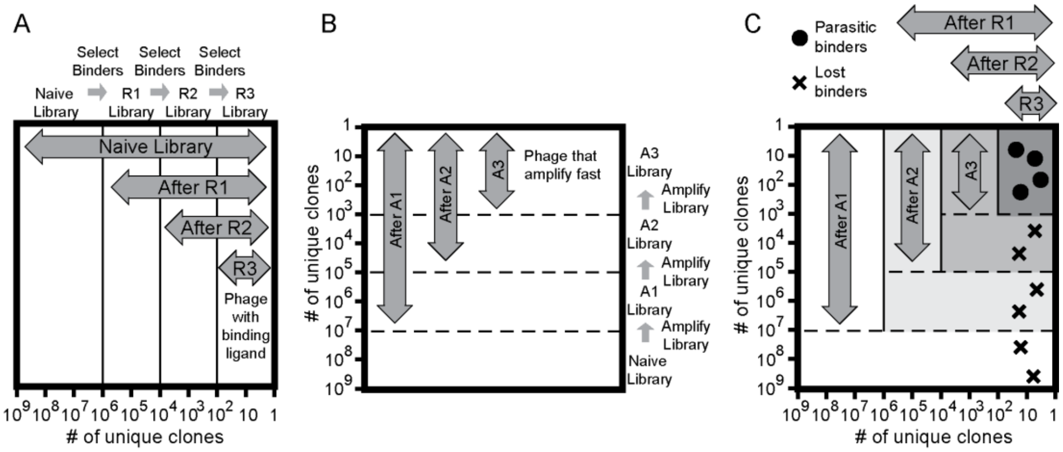


Figure 1.2. (A) Selection from phage display libraries after rounds of binding to the target can be represented as progressive collapse of a naïve library (10^9 diverse sequences) to a smaller number of binding sequences (here, 10^2 sequences). (B) It is known that the naïve library of phage displayed peptides contains sequences that amplify slowly in bacteria and those that amplify faster[17,18]. Repetitive rounds of amplification in bacteria, thus, lead to progressive collapse of diversity from the theoretical 10^9 clones to a smaller number of binding sequences. (C) Collapse due to binding preferences and due to amplification in bacteria is independent of one another. Diversity collapse due to amplification bias occurs in amplified phage libraries regardless of amount of phage (e.i. 10^3-10^9 PFU) or volume (*e.i.* 1 mL – 1 L)[17]. In a selection that involves rounds of binding and re-amplification, the library collapses to a few clones that bind to a target and have high amplification rates, labeled as '•' (referred to as 'parasitic binders'). As a consequence, many binders, labeled as 'x' (referred to as 'lost binders'), cannot be discovered in conventional phage display selection.

sequencing of phage display libraries. 3) Application of deep sequencing to characterize biases in bulk amplification of phage display libraries. 4) Application of emulsion amplification to prevent bias and collapse in library diversity. 5) Incorporation of emulsion amplification into the selection process to identify sequences that could not have been identified using the standard phage library screen. In the introductory chapter, I discuss the NGS technique that has been gaining in popularity for determining the diversity of phage libraries during the course of the selection process. I will present how NGS can detect, characterize, and provide examples of strategies used to bypass unproductive diversity collapse.

1.1.1 Phage display library construction and selection

The development of phage display started from the observation that the minor coat protein (pIII) of the filamentous bacteriophage (M13, fd, f1) allows insertion of foreign polypeptides at its N-terminus as fusion proteins[1]. These fusion proteins are encoded in the phage genome and ‘displayed’ on the phage surface, where they and are accessible to the external environment. Since there is a direct linkage between the displayed protein and its encoding gene, selected binding ligands can be identified through sequencing the phage DNA linked to the ligand. Additionally, clones baring a specific peptide can be amplified by infecting *E. coli* cells[2]. The M13 strain of bacteriophage is the most common strain used in peptide phage display. The M13 genome encodes 11 genes. It comprises of a single stranded circular DNA encased in a protein capsid with an

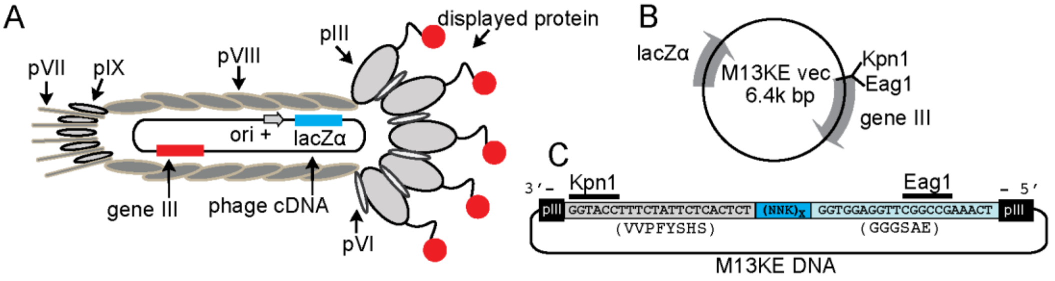


Figure 1.3. (A) Schematic representation of M13 bacteriophage. (B) The cDNA contains two important genes for library construction; the lacZα gene and restriction sites at the N-terminus of gene III, which encodes for protein pIII. (C) The displayed library is inserted after the leader sequence (VVPFYSHS), which transports the pIII protein to the periplasm and is cleaved before secretion from the bacteria.

approximate diameter of 7 nm and length of 900 nm. The N-terminus of the pIII located on the tip of the capsid has been the only cloning site to display peptides in phage display libraries using the commercial M13KE phage vector. While peptide libraries can be produced by restriction cloning or Kunkel mutagenesis techniques[28], the most convenient peptide libraries are those that are commercially available. Examples are the linear and cyclic heptapeptide (Ph.D.-7, Ph.D.-C7C) and linear dodecapeptide (Ph.D.-12) libraries distributed by New England Biolabs Inc. (NEB). These libraries are built on the M13KE phage vector[29], which contains the lacZα reporter and allows production of colored plaques in a lawn of bacteria that contains the lacZΩ fragment of β-galactosidase supported by agar supplemented with the colorimetric probe X-gal. The M13KE vector also contains restriction sites at the N-terminus of the pIII gene that allow for easy insertion of a degenerate DNA encoding a peptide library between the leader sequence (VVPFYSHS) and a linker (GGGSAE). (Figure 1.3) The leader sequence is part of the pIII signal sequence that is cleaved by proteases upon secretion of the pIII coat protein across the inner membrane of the bacterial cell. A library of DNA sequences is incorporated into the M13KE vector, and the resulting library of vectors is transduced into electro-competent cells to yield the original library of phage-displayed peptides. The libraries diversity is determined by the degeneracy of the peptide-encoding DNA sequence and the number of clones produced.

There are multiple strategies for selection, but most of them can be described as binding of the phage library to a target immobilized on some heterogeneous carrier (bead, plate, column, fluidic channel, *etc*.). The first step in selection is the panning process, which involves pre-absorption of the phage library and repeat washing. Pre-absorption of the phage library can be performed in micro-wells containing all of the components of the screen, except the target molecule. This step will remove non-specific binding phage, *e.i.* phage that bind to plastic, streptavidin, or BSA[30]. Repeated washing will remove non-binding phage. Non-binding phage are washed off from this carrier and the binding phage are eluted by applying the conditions that destroy interactions between the phage and the target. The typical panning process decreases the library size from 10^{10} to 10^{4}-10^{6} phage particles (PFU). Prior to a further round of panning, the eluted phages are amplified by using them to infect an excess of *E. coli* cells in one

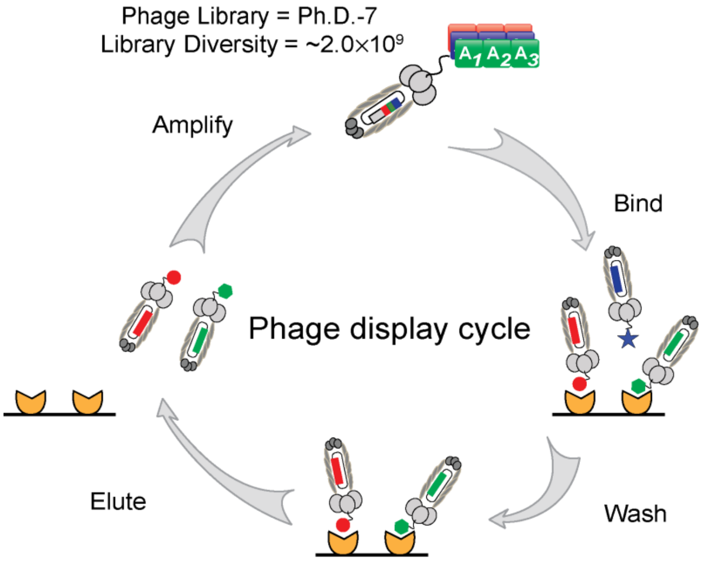


Figure 1.4. Schematic representation of a typical round of phage display selection. Selection involves panning a phage display library, which involves incubation of the library with a target, washing of non-binding phage, and eluting the bound phage. Panning is followed by and amplification of the eluted phage, which can undergo successive rounds of panning and amplification.

common flask (Figure 1.4). In order to enrich the pool for target binding phage to the level detectable by Sanger sequencing, multiple cycles of selection are performed. A survey of published examples of selections with Ph.D.-7, Ph.D.-C7C, and Ph.D.-12 libraries, show that an average of three to four cycles are carried out (Figure 1.5,1.6)[31].

Sanger sequencing was the sole tool for identification of binding sequences in the first two decades after the development of phage display technology. This method is practical for sequencing of 100, and rarely up to 1000 clones. It is most common to sequence less than 100 phage clones. Since Sanger sequencing has

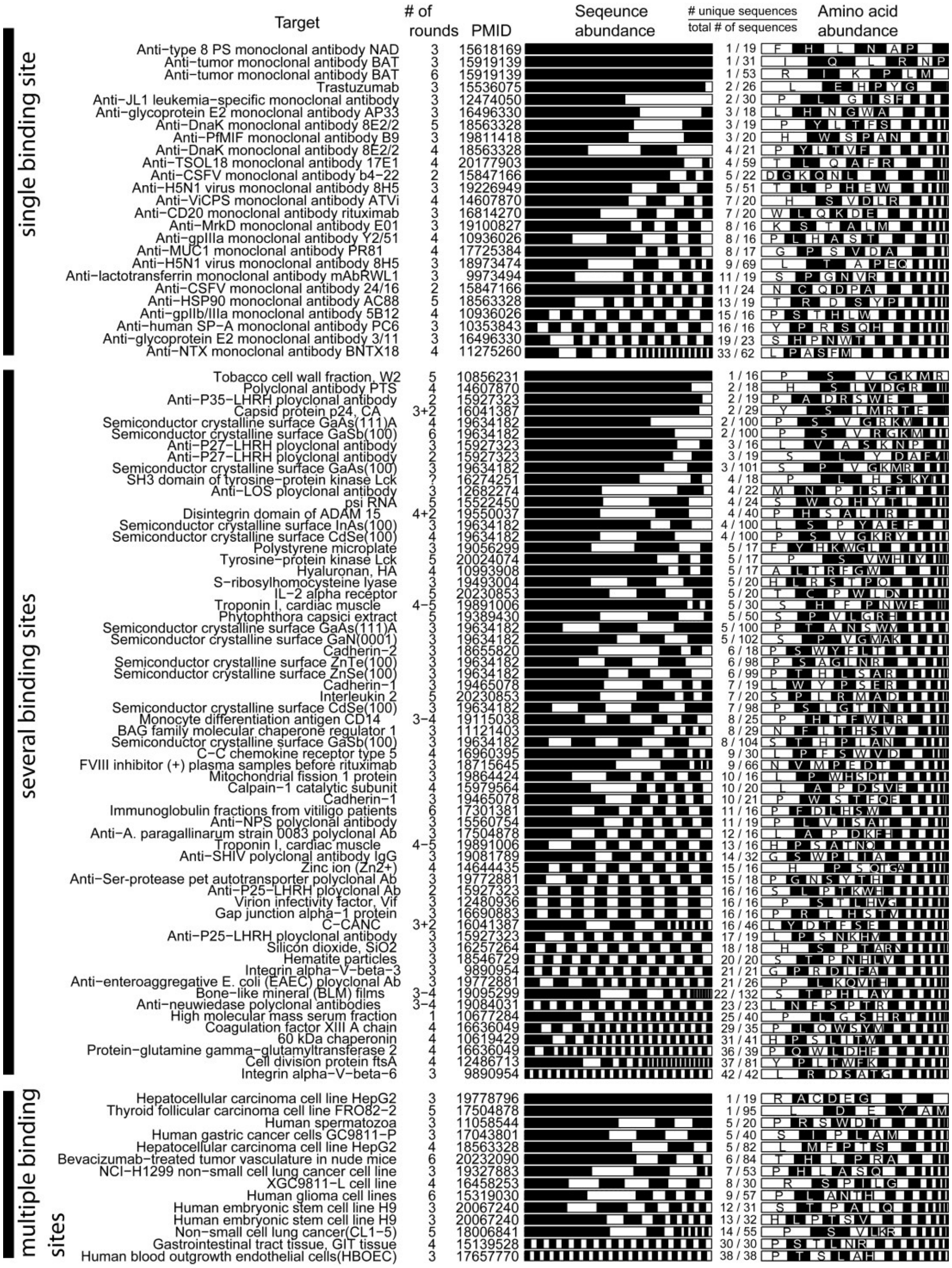


Figure 1.5. Analysis of the diversity of the Ph.D.-12 phage library after screening against various targets from papers that reported >15 DNA sequencing. The data were extracted from a raw MimoBD database using custom Matlab software. PMID is the PubMed ID of each article.

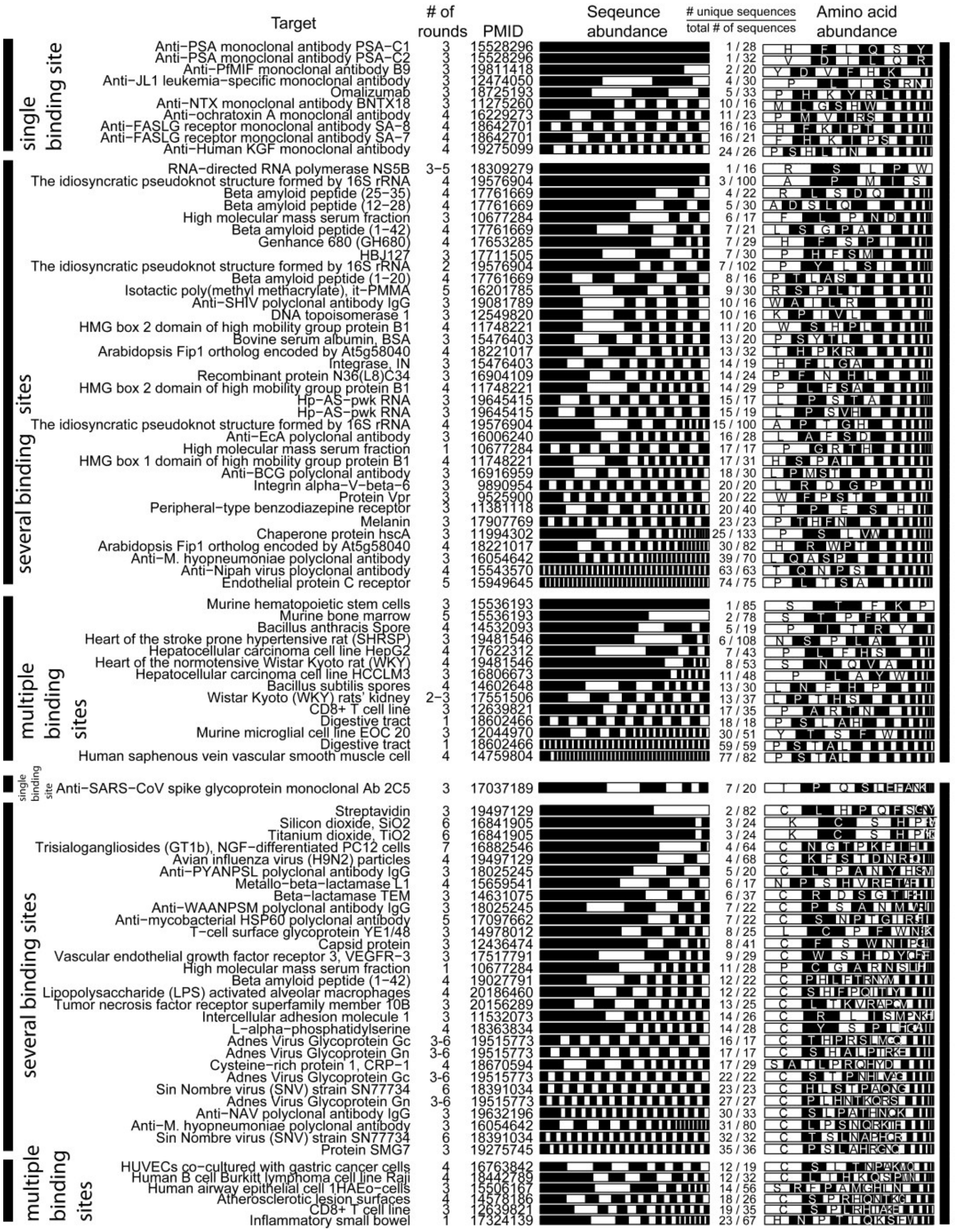


Figure 1.6. Similar analysis as Figure 1.4 for Ph.D.-7 and Ph.D.-C7C libraries.

shallow sequence coverage, the selected set of phage clones do not represent the entire diversity of the selected library.

1.1.2 Limitation in current phage display strategies

Selection from one round to another should narrow the library diversity. Theoretically, this process will enrich the abundance of target-binding phage. However, amplification also narrows the diversity by selecting fast-growing clones (Figure 1.2)[10]. These biases act together to decrease the library diversity to a small sub-population of clones. Increasing evidence suggests that amplification decreases the library diversity and limits the number of binding clones a screen can identify. There have been several reports that have discussed the biological reasons for growth advantage, such as the use of rare codons[32], peptides that favor faster packing or infection[33,34], and mutations in the regulatory regions of phage genes[18,35,36]. Structural properties of peptides displayed on pIII of phage also affect the amplification rate; peptides with β-turn structures amplify faster, whereas those with α-helical structures amplify slower[34,37]. For phage libraries that display peptides on pIII and short (<8-mer) peptides on pVIII these effects on growth rate are small[38-40]. However, phage libraries of peptides displayed on pVIII are more prone to loss of sequence diversity than those displayed on pIII[33,41,42]. This amplification bias is not specific to only phage display libraries. Loss of diversity during amplification also occurs in related phage display techniques such as phagemid-display[43,44] which is used to display natural[45,46] or synthetic antibody[43] fragments and other full length proteins[47]. From a survey of literature reports from 1990-2010 using the phage display peptide libraries from NEB, selections against multi-site targets identified one or a few unique sequences (Figure 1.5,1.6). Therefore, selection against any target type, single or multi-site, always converges to a small number of ligands. If amplification bias did not lead to diversity collapse during selections against multiple targets, there should be more ligands identified[10]. This decrease in diversity due to amplification bias can hinder the identification of useful binding ligands for targets with single binding sites[48,49], or targets with multiple binding sites[12,32-34,41,50].

In general, phage display technology has been successful because the modification of the phage coat proteins has minor effects on the rate of production of phage[1,51,52]. Nevertheless, even small differences in growth rate can have important consequences in the distribution of phage that display different peptides after amplification. This can be simply illustrated by amplifying a mixture of library phage with wild type phage in a common solution. Amplification of these two phages results in a dramatic enrichment of the faster growing wild type phage (Figure 1.7). Ratios change from 1:1 to 100:1 after just 5 hours of amplification.

The small differences in growth rates between clones in phage display libraries are difficult to detect with shallow sequencing methods, such as Sanger. In the past decade, next-generation sequencing (NGS) technologies, such as Illumina and Ion Torrent, were developed and can sequence between 10^6-10^8 DNA sequences per run. Utilizing NGS platforms will make it possible to sequence phage libraries of diversity up to 10^6 and phage libraries from selections, which typically obtains 10^4-10^6 PFU. By enabling large scale sequencing of phage libraries, we can more accurately identify the most enriched

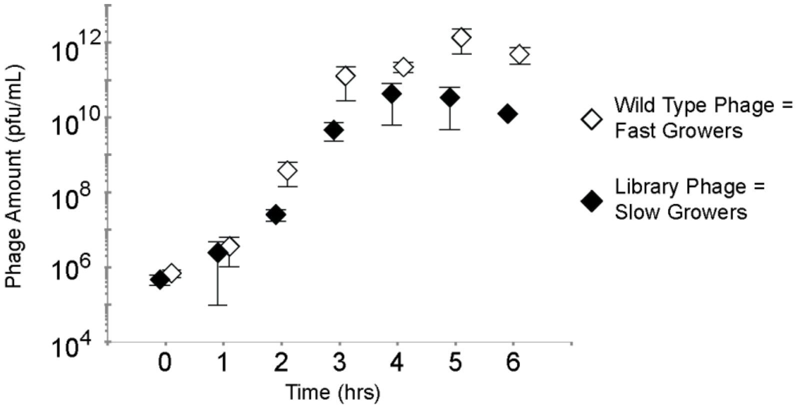


Figure 1.7. Titer for wild type and library phage amplified in the same solution. The input ratio of ~1.2 wild type/library phage was added to the same solution of bacteria and LB media. After amplifying for 6 hours, the ratio was ~65 times higher.

sequences. Additionally, by deep-sequencing the original library or libraries from negative selections, for example, we can eliminate false positives.

1.2 Next-generation sequencing (NGS)

For single targets, analysis of a shallow population of ~100 sequences enriched from selection can be enough to predict one consensus motif[53]. However, in order to accurately and precisely identify multiple binding sequences, or to track enrichment during selection rounds, analysis of a larger population of sequences is needed. NGS technologies, can determine abundance of peptide sequences from selection with higher precision than Sanger sequencing. NGS technologies also allow for tracking of enrichment from an original library to the Nth round of selection. With larger data sets, statistical analyses can be incorporated into selection to identify significantly enriched sequences and multiple consensus motifs[17,54,55].

The early work performed on sequencing phage display libraries was done using the Roche's 454 sequencing platform[56-59]. Since then, Illumina[7,17,60-66], Ion Torrent[56-58,67], and other NGS platforms have been used to characterize phage display libraries[55]. Since each sequencing run has considerable cost, multiplexing allows for pooling of multiple DNA samples from different experiments to be processed and sequenced in the same sequencing run, thus, reducing the sequencing cost per experiment. Multiplexing is achieved using primers with short standardized DNA sequences ("barcodes"), which allow segregating independent samples during the processing of sequencing data. For example, Sidhu and co-workers used barcoded primers for the preparation and sequencing of 22 independent panning experiments in one run[58]. These barcoded primers enable the analysis of peptide sequences present in the library at various steps of selection and enrichment against a specific target to be determined. Our group routinely uses barcoded primers in Ion Torrent[54] and Illumina[17] sequencing.

Pasqualini and co-workers were the first group to incorporate NGS analysis into phage library selection in 2009[67]. They demonstrated the advantages, both in cost and time, of NGS over the standard Sanger sequencing method. In this work, the authors performed *in vivo* panning of a C7C library on an end-of-life patient. Phage clones were recovered from biopsies of various organs such as bone marrow, adipose tissue, muscle and skin, and either sequenced through Sanger or 454 sequencing. Sanger sequencing, identified 3,840 sequences, while 454 identified 319,361 sequences. To sequence 3,840 clones by Sanger sequencing took ~15 days. The authors extrapolated the cost and time required to sequence 100,000 phage clones by Sanger as 9,898 hrs (412 days) and $338,884. Conversely evaluation of 100,000 phage clones by 454 sequencing required only 74.8 hrs and $1,307. Today, a more uniform comparison is used, such as the reagent cost per megabase (Mb) of read length. For the 454 NGS platform, it is $7-12.4, whereas for Sanger it is $1,500. For other NGS platforms, such as Illumina, the reagent cost/Mb has been reduced to $0.04[68]. The low cost and large amounts of data obtained, make it attractive to incorporate NGS into the phage library selection procedure.

Since 2009, the use of NGS platforms in phage library screen has expanded rapidly as more research groups gained access to this technology and analysis software. In the next section, I describe the Illumina and Ion Torrent platforms and their application in examining the diversity of phage libraries. I also present two strategies that are used to incorporate NGS into selection. The first strategy uses NGS platforms to track enrichment over several rounds. The second strategy uses NGS after the first panning event, avoiding the amplification step. This approach minimizes growth bias after panning, but requires careful analysis of the sequence abundance to identify productive enrichment from only one panning event.

1.2.1 NGS platforms used in phage display

1.2.1.1 Illumina sequencing

The Illumina platform is based on an approach called, “sequencing by-synthesis”[69]. In this approach, ssDNA is immobilized on the surface of a glass flow-cell, and is clonally amplified into clusters using bridge amplification. The DNA of each cluster is sequenced using reversible fluorescent dye-terminated nucleotides. All four nucleotides are added simultaneously onto the flow cell during each cycle of the sequencing process. Each nucleotide is labelled with a base-specific fluorescent label at the 3'-OH position, which allows only one nucleotide to be incorporated by DNA polymerase. After the base is incorporated, fluorescence intensity is recorded and 3'-OH blocking group is chemically removed in preparation for the incorporation of the next base[70].

The region of the phage genome that contains the degenerate fragment must be converted to short dsDNA that can hybridize to the surface of the Illumina flow cell. This conversion can be achieved by PCR amplification of the vector using primers that hybridize upstream and downstream from the degenerate region. At each end of the dsDNA, this PCR amplification also introduces Illumina adapters, which contain regions complementary to DNA sequences immobilized on the flow cell (Figure 1.7A-C). Clonal bridge amplification creates the clusters for subsequent sequencing-by-synthesis.

The output from Illumina sequencing is most commonly represented as a plain text file in “FASTQ” format, which contains 10^6-10^9 blocks of four lines:

```
'@SXF2J:00010:00047'
'TCTGGGCAAACGGCCGAACCTCCG'
'+'
'44444-44=-44242442/566,,'
'@SXF2J:00025:00034'
'TTCCTGCTTCACCCGAACCTCCACCAGCAGCCCGACCAAGATGCGTAGAGTGAGAATAGAAA'
'+'
'924899998444464;>A8:;=@@?998>888033341=9;?B;88=;993////,/0,**)'
'@SXF2J:00031:00025'
'TTCGTGATTCCCGAACCTCCACCATGCTGAATATGCATATTATAAGAGTGAGAATAGAAAGG'
......
```

The first line begins with ‘@’ and is followed by information about the sequencing run, which includes the instrument name, flow-cell lane, tile number, X and Y coordinates of clusters. The second line is the interpretation of the sequence from the recorded fluorescence intensities of the flow-cell. The third line contains the ‘+’ character. The forth line encodes the quality scores or Phred scores, denoted as ‘*Q*’, for the sequence in the second line. The Phred score describes the probability (*p*) that the corresponding base is read incorrectly, and is calculated using the formula: $Q = -10 \log_{10}p$. This score ranges from 0-41 in the Illumina platform. In FASTQ format, the score is converted to one of the ASCII symbols minus 33. For example, the symbol ‘=’ has the ASCII code of 61, therefore it has a 61-33 = 28 Phred score, which corresponds to $10^{-28/10} = 0.0016$ (0.16%) chance of being an incorrect read.

This protocol was utilized to identify growth bias in phage libraries in Chapters 3 and in the publication: W.L. Matochko, K. Chu, B. Jin, S.W. Lee, G.M. Whitesides, and R. Derda, “Deep sequencing analysis of phage libraries using Illumina platform”, *Methods*, 2012, **58**: 47-55.

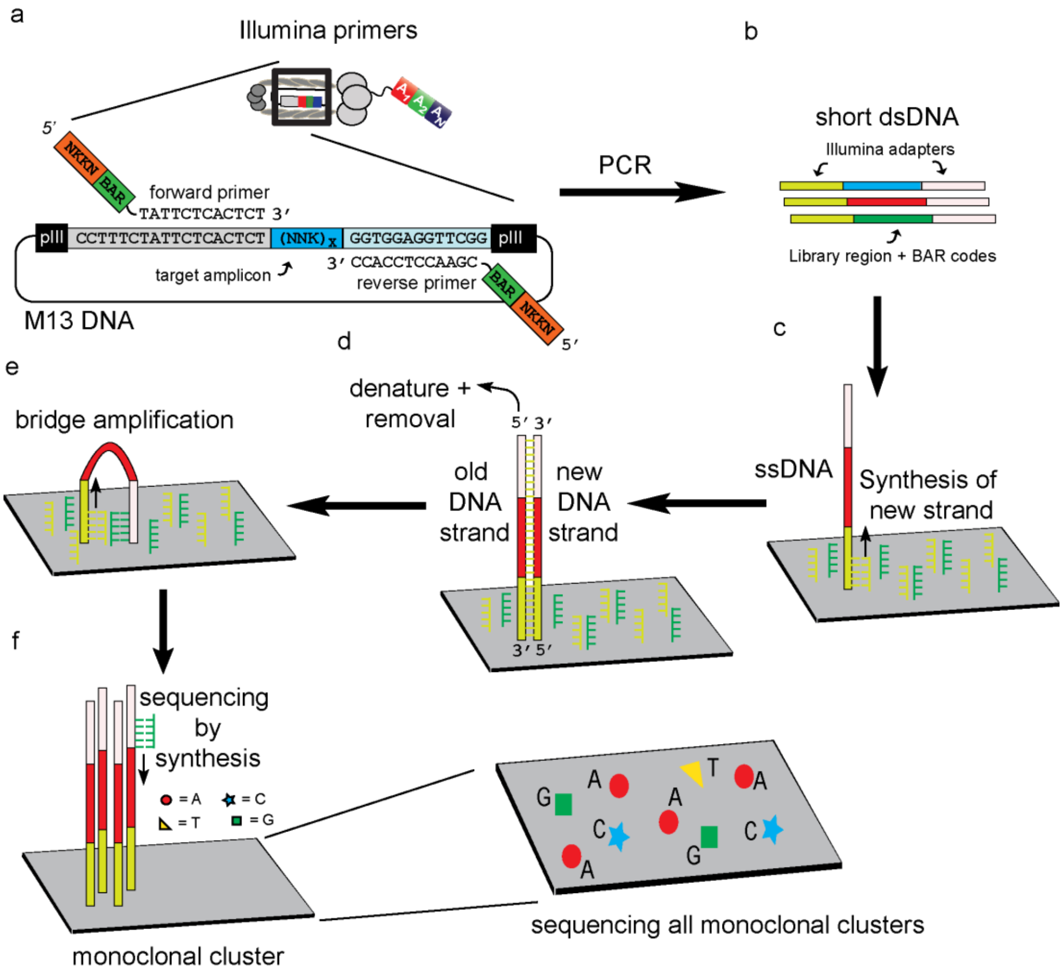


Figure 1.7. Schematic representation of steps involved in Illumina sequencing. (A) The library region is amplified using primers flanking the variable region. (B) Illumina adapters are ligated to the dsDNA fragments. Each adapter contains a complementary region to oligo nucleotides present on the flow cell. (C) Hybridization of the ssDNA to one of the oligos on the flow cell followed by (D) strand extension and removal. (E) The new ssDNA is clonally amplified through bridge amplification and creates a cluster of up to 1,000 identical copies. (F) Each cluster is sequenced using dye-label terminated nucleotides.

1.2.1.2 Ion Torrent sequencing

Ion Torrent uses semiconductor microchips to detect hydrogen ions (H^+) released during the polymerization of DNA[68]. The microchips contain a dense array of more than a million micro-wells. Each micro-well contains an ion sensitive transistor and space for one bead with a clonal population of DNA. The transistors enable real time measurement of the change in pH during polymerization, which is converted into a voltage. In this approach, the four nucleotides are sequentially added to the semiconductor microchips containing the DNA to be sequenced. If the nucleotide is complementary to the DNA strand, hydrogen ions are released; the ion sensor is triggered, which detects an electrical signal proportional to the number of bases incorporated. If a nucleotide is not complementary to the template nucleotide, there is no reaction. In the Ion Torrent technique, neither modified nucleotides nor optical detection are used[68,70].

Unlike the Illumina platform, in which each DNA sequence is amplified *via* bridge amplification on the flow cell, in the Ion Torrent platform, each DNA sequence is amplified through emulsion PCR. Specifically, DNAs are amplified on Ion Sphere Particles (ISPs) that contain specific DNA adapter sequences that capture the target DNA. Emulsification of ISPs and DNA fragments ensures that a single DNA fragment from the library is linked to a single bead in one microdroplet. PCR amplification of DNA library fragments in microdroplets results in a population of beads containing a monoclonal population of single DNA fragment. The templated beads are loaded into proton-sensing wells that are fabricated on a silicon wafer and sequencing is primed from a specific location in the adapter sequence[70] (Figure 1.8). Two strengths of this technology are a rapid sequencing speed and low cost made possible by avoiding the use of modified nucleotides and optical measurements[68].

A common criticism of Ion Torrent technology is the higher insertion and/or deletion (indel) error rate as compared to Illumina technology. Ion Torrent, additionally, contains lower quality reads, ranging from Phred score 10 to 20. The high rate of insertion and deletion (indel) errors contributes to inaccurate nucleotide assignments in Ion Torrent sequencing, most frequently in homopolymeric regions[70,71]. Even correctly called homopolymeric regions are typically assigned lower confidence[72]. Typically, publications use high quality reads with Phred scores ≥ 30[13,73], however, tolerating a Phred score < 18 for three bases per read is not detrimental to the analysis and can provide the most optimal results in identifying binding ligands. In other reports, utilizing lower quality reads, with Phred score 5 and above, did not alter interpretations or results[17,64]. Filters applied inappropriately could remove too many sequences and in this way introduce strong biases[17]. By placing stringent conditions, important sequences could be removed those are needed to identify consensus motifs.

This optimized protocol was utilized to identify growth bias in phage libraries and to characterize the parasite population in Chapters 4, and in Chapter 5, to identify cell binding ligands from selection against breast cancer cells.

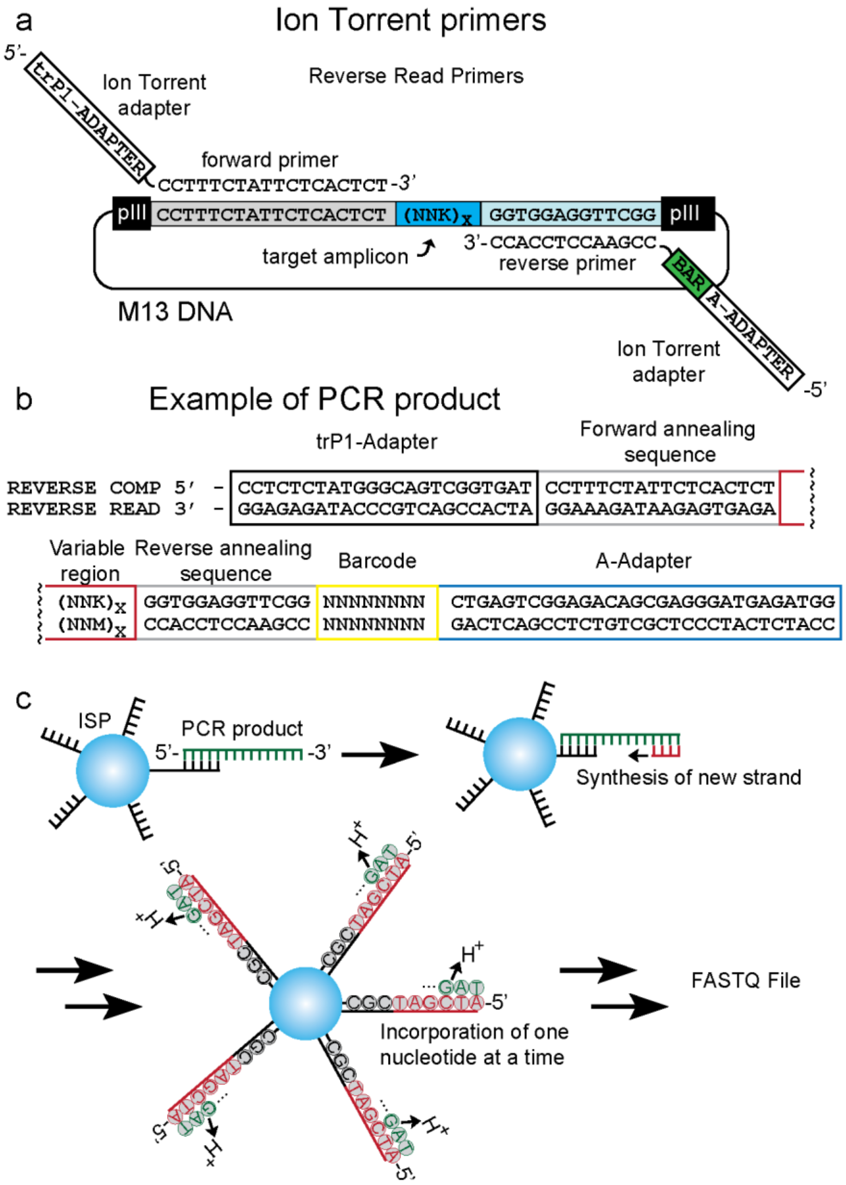


Figure 1.8. Schematic representation of steps involved in Ion Torrent sequencing. (A) Ion Torrent adapters are incorporated directly into the primers, allowing for one PCR and one purification step. (B) Example of a dsDNA fragment that is used directly in emulsion amplification with ISPs. (C) Each ISP is trapped in one well of a semiconductor microchip. One nucleotide is added at a time and the sequence is determined by the release of H^+ ions is the nucleotide was incorporated into the complement strand.

1.2.2 NGS analysis methods for identifying binding ligands

1.2.2.1 Selection with multiple rounds and deep-sequencing

Deep-sequencing the final round of selection can provide information about the true abundance of sequences in the selected library. Sequencing of all rounds of selection, including the original library, can provide additional information to identify potential hits, such as enrichment, structure-activity relationships, and selection of false positives. For example, Birnbaum *et al.* conducted selection over four rounds using a yeast display of a peptide in complex with a major histocompatibility complex molecule (MHC) library targeting 2B4, 226, and 5cc7 T-cell receptors (TCRs). In the first attempt to identify binders, plaque picking was used. However, only three unique clones from 12 were identified and all were related to the wild type (WT) MHC peptide. The Sanger method only identified WT-sequences and prevented identification of alternative, non-homologous sequences[74]. By switching to deep-sequencing and tracking the enrichment over four rounds, Birnbaum *et al.* identified and synthesized a library of 44 peptides to examine their potential to stimulate CD69 upregulation and IL-2 production in T cells. Most of the peptides (36) bound to T-cells and induced CD69 upregulation. For these authors, they were able to use deep-sequencing data to perform structural-activity relationships and identify peptides that bound the MHC molecule and induce CD69 upregulation[74].

Incorporating deep-sequencing into selections, phage libraries can aid in simple analyses, such as the identification of the most abundant sequence(s). Ngubane *et al.* performed selection against *Mycobacterium tuberculosis* using a phage display of a C7C peptide library and employed two sequencing methods[9], Sanger and Illumina, to determine abundant peptide sequences. The peptide library and the enriched libraries after Rounds 3 to 5 of selection were sequenced using Illumina, while ten phage clones were sequenced after the Round 5. Sanger sequencing identified three unique peptide sequences out of the ten selected clones. These three unique sequences, however, were not among the most abundant sequences determined by Illumina. The three unique sequences identified by Sanger included: clone 1 (CHYDGARAC), which represented 0.81% of the library diversity, clone 2 (CDHGYLPSC) represented 0.92%, and clone 3 (CFDTRSLVC) represented 5.05%. The most abundant sequence (CPLHARLPC, which represented 82.49% of library diversity) was identified using Illumina[9]. The most abundant sequence also displayed the highest binding when assayed against *M. tuberculosis*, *M. smegmatis* and *M. bovis*. Utilizing NGS improved the identification of binding sequences that would be missed by clone picking and Sanger sequencing.

With Sanger sequencing, it is difficult to determine what phage clones contain a growth advantage in the original library. By sequencing the original library in addition to phage libraries from selections, phage clones with growth advantages can be determined and discarded as potential hits. t'Hoen *et al.* performed selection of a PhD-7 library against KS483 osteoblast cells, and deep-sequenced the original library and all four rounds of selection. Phage clones with the highest abundance in the original library increased from 0.26% in that library to 10% in Round 4. These clones may contain high propagation potential[66]. For example, HAIYPRH was a peptide in the original library with the highest count (36), by Round 1, it was found 41,257 times (0.26% of total sequences), and 237,535 times after Round 3 (1.8% of all sequences)[66]. Additionally, the authors noticed that the amino acid composition in the original library was different from the theoretical composition based on manufacturer's specification. Amino acid residues such as cysteine, glycine and arginine were underrepresented, whereas proline was overrepresented. Therefore, there was already a bias in the naïve library[66]. Using these observations, the authors removed all selected peptides in Round 4 that contained a copy number of two or more in the original library and that were found in the databases such as PepBank[75] and SAROTUP[76]. For example, the GETRAPL sequence was removed as a potential hit since it was previously published in selections against polystyrene[77]. By cross-referencing selected peptides with those found in databases, false positives or promiscuous binders were removed. The authors synthesized 10 putative hits selected from deep-sequencing results. Of these, four were confirmed to bind to KS483 cells[66].

In the screen against KS483 cells t'Hoen *et al.* also tracked the top sequences after each round of selection[66]. In general, the top 1000 sequences in Round 4 overlap with 60% of the sequences in Round 1. Additionally, out of the top 10 sequences in Round 4, eight make up the top ten sequences in Round 1. The authors used this observation to propose that: 1) further rounds of selection will not lead to the identification of peptides that could not have been found in earlier rounds, and 2) high affinity peptides can already be identified after the first round of selection. The sequence HAIYPRH was identified as the most abundant sequence in all selections, indicating further rounds might be more detrimental to the selection of target binding hits, instead could lead to the identification of false positives[66].

Sidhu and co-workers used phage display with deep sequencing to analyze the co-evolution between a domain and its peptide ligands[58]. The authors produced a library of PDZ domains with diversity introduced at ten positions in the binding site, and the domains were selected for binding against 15 distinct peptide ligands. The binding affinities of the selected domains were compared to 61 previously characterized PDZ domains that were obtained without any selective pressure[78]. After selection of the PDZ library, 162 unique domains were identified, but only 22 were examined for binding to their respective ligand. The authors identified 22 domains that recognize their cognate peptides with higher affinity but lower specificity compared to the 61 unselected domains. Deep-sequencing analysis revealed selected PDZ domains had common features when selected against a common peptide[58].

Affinity selection coupled with deep-panning has been applied to other display techniques, such as T7 display of the WW domain of the human Yes-associated protein 65 (hYAP65). Fowler *et al.* tracked the fate of thousands of variants of the 50 amino acid human WW domain from the original library to the selected library after Round 6 against the peptide sequence GTPPPPYTVG[79]. Using Illumina, the authors observed selection reduced library diversity from ~ 600,000 to ~94,000 variants after Round 6. The authors also observed mutational preferences and evolutionary conservation in amino acid residues of the WW domain by calculating an enrichment ratio for each amino acid at each position in the variable region. The enrichment ratio was calculated as the frequency (the sum in i^{th} position of the variable region bearing the j^{th} amino acid over the sum of all reads) of mutations in the Round 6 library divided by the frequency in the original library. The authors compared the enrichment of each amino acid to a consensus motif. There were distinct regions in the WW domain that were permissive to mutation (had high mutational frequency) and others that were intolerant to mutation (had low mutational frequency). Out of the 25 variable residues, 20 remained the same or similar to the consensus motif. These observations led to an understanding of how mutations can impair protein function in the WW domain[79].

1.2.2.2 Single-round selections

Canonical selection involves multiple rounds of panning and amplification, however, the latter can introduce bias into the selection and result in enrichment of target unrelated binders, or phage clones with high growth potential. By incorporating deep-sequencing after just one round, some of these biases might by minimized or removed from selection.

Heinis and co-workers demonstrated that one round of selection is sufficient to identify ligands to five targets: Sortase A from *S. aureus*, human urokinase-type plasminogen activator, activated human coagulation factor XII, human plasma kallikrein, and streptavidin[55]. The novelty of this approach, compared to t'Hoen's[66] or Sidhu's[58], was the use of multiple replicates. Additionally the authors based their selection on the presence of consensus motifs rather than the abundance of individual sequences. In all selections, at least one or two previously identified target-binding peptide motifs could be found. Namely ‘LPP’ was found for Sortase A, ‘T/S AR’ and ‘K/R F/Y S/T L’ for urokinase-type plasminogen activator[80,81], ‘VxxKCL’ for human coagulation factor XII[82], ‘F/Y/W xxCRV’ for plasma kallikrein and ‘HPQ’ for streptavidin. They also identified a number of different consensus sub-families; 15 for Sortase A, 16 for urokinase-type plasminogen activator, 2 for coagulatin factor XII, 11 for plasma kallikrein and 2 for streptavidin.

Heinis and co-workers proposed that the copy number of the peptides in the selected library is important. If the average copy numbers of peptides are low, the identification of target-binding peptide motifs might be difficult to distinguish from background peptides that were isolated through non-specific interactions. To overcome the problem of selecting a low copy number of peptides, multiple parallel independent selections can be performed. For example by comparing the output of two selections performed in parallel against human coagulation factor XII, three target-binding clusters were identified from background clusters. This approach could identify target-binding ligands from noisy datasets and parasitic sequences.

With a larger number of replicates, more stringent criteria can be used to identify binding ligands. For example, our group performed six independent selections of phage-displayed glycopeptides against one target (Concanavalin A) and three to six replicates of control group selections: i) glycosylated library panned against BSA; ii) unmodified peptide library panned against ConA; iii) methyl-oxime modified library panned against ConA[54]. All outputs from all selections were sequenced using Ion Torrent. The authors then used a volcano plot to identify sequences that were enriched by a factor of five or more, with a significance threshold of $p < 0.05$ between the test selection and the control selections. Applying this analysis to the three control groups yielded three sets of binding sequences. The intersection of all three sets yielded 86 sequences that shared one N-terminal consensus motif WYD. We subsequently tested synthetic peptides predicted by this analysis, and showed that peptides with an N-terminal WYD motif and mannose oxime (Man-WYDRFPPHES) displayed >40-fold enhancement in affinity as compared to the parent mono-mannoside ligand (Man)[54].

It is necessary to consider the copy number and diversity of the library when performing a single round of selection. Ren Sun and co-workers[63] demonstrated that to identify enriched sequences from one round of selection, the copy number of each sequence must be > 1000. The authors reasoned that binding ligands with a Kd ~ 100 nM or better are present at a frequency of ~1 in 10^9 in mRNA libraries of 10^{12}-10^{14} diversity. The authors demonstrated that in order to enrich clones that are present in the selected library in 20-30 copies, enrichment needed to be > 1000 fold above the original library[63]. This copy number for the selected sequences was obtained using Illumina sequencing, which produced ~3.1 x 10^7 reads. Using shallower sequencing methods might result in lower copy numbers for the selected sequences and discrimination between ligands with μM and nM binding constants might not have been possible.

Selection with one round has been used in other display systems, such as SELEX, which typically is conducted with >10 rounds. Morse and co-workers, panned a SELEX library of 30 randomized nucleotides against ZnO nanoparticles for one round. By performing clustering of consensus nucleotide motifs, the top four motifs poly(T), poly(C), poly (D), and poly(A) were selected. The poly(T) and poly(C) motifs were tested for binding to a ZnO surface, and displayed >100 times stronger binding, based on fluorescence intensity, than the starting library. The poly(T) nucleotide sequence facilitated the synthesis of ZnO nanoparticles with a crystalline core at neutral pH[83].

A broader look at all the examples in which deep-sequencing is used in parallel with selection of display libraries can be found in Table 1.1. Selected examples were mentioned in the text above. Those that were not were either similar to the discussed methods (for example Entry #3 is similar to #17), or involved libraries that are not the focus of this thesis (for example Entries #9-15 deal with antibody libraries).

Table 1.1. List of examples in the literature that use deep-sequencing in phage display selections.

Entry #	Target	Library	Rounds of selection	NGS platform	Analysis method	Ref
1	In vivo human selection and biopsies of skin, fat-tissue, bone marrow and skeletal-muscle	CX_7C peptide phage library	1	454	Enrichment	67
2	In vivo human selection and biopsies of skin, fat-tissue, bone marrow and skeletal-muscle	CX_7C peptide phage library	3	454	Enrichment	84
3	PDZ domain	pVIII X_7 phage library	5	454	Consensus motifs	58
4	Concanavalin A	Ph.D.-7 phage library	1	Ion Torrent	Enrichment, Consensus motifs	54
5	KS483 osteoblast cells	Ph.D.-7	4	Illumina	Enrichment	66
6	Mycobacterium tuberculosis	CX_7C peptide phage library	5	Illumina	Enrichment	9
7	mAb GV4H3 and human IgG from HIV-1^+ individuals	pVIII X_7 phage library	3	Illumina	Enrichment	7
8	Sortase A, Urokinase-type plasminogen activator, Coagulation factor XII, Plasma kallikrein and Streptavidin	Bicyclic ACX_mCX_nCG phage library*	1	Illumina	Clustering	55
9	16 targets	Antibody phage library	4	454	-	57
10	Antibody 5E3 and Human interferon gamma (hIFNγ)	Antibody phage library	3	Illumina	-	85
11	ESAT6, Ag85, IgER, ubiquitin	Antibody phage library	2	Ion Torrent	Enrichment	86
12	Antibody 5E3	Antibody phage library	3	Illumina	Enrichment	13
13	Hapten trinitrophenyl	Antibody phage library	3	454	Enrichment	87
14	Human interferon gamma (hIFNγ)	Antibody phage library	3	Illumina	Enrichment	88
15	Retrovirus glycoprotein 41 (gp41) and IL-1 receptor	Antibody phage library	2	454	Enrichment	89
16	Rabbit antibodies immunized	Tpl Phagemid display	3	454	-	90

	with Tpl strain V1					
17	GST peptide and GST-PDZ	PDZ-pVIII phage library	5	Illumina	Consensus motifs	91
18	Thrombin	API - T7 library	5	Ion Torrent	Enrichment	92
19	GTPPPPYTVG peptide	hYAP65 WW - T7 library	6	Illumina	Consensus Motif and Enrichment	79
20	Maltose binding protein and human IgG (Fc)	10Fn3 - mRNA library	1	Illumina	Enrichment	63
21	ZnO	SELEX	1	SOLiD	Consensus motif	83
22	*E. coli* protein Hfq	SELEX	10	454	Consensus motif	93
23	mEphA2 extracellular domain, N202.1A and MDA-MB-231	SELEX	8	454	Enrichment	94

* ‘m’ and ‘n’ range from 3 to 5 amino acids.

1.2.2.3 Analysis methods

Deep-sequencing of phage library screens demand development of analysis tools for processing of the data and identification of target-binding ligands. Many research groups that use deep-sequencing technology develop their own processing software in C++[56], Perl[95], Java[91] and Matlab[17,55] computer languages. Currently there is no universal consensus on criteria for analysis methods and quality control.

Advanced analysis techniques use not only abundance but also information about round-to-round enrichment and emergence of consensus motifs. For example, the software Enrich, used in reports by Fowler and co-workers[96], can analyze enrichment of each amino acid at each position of the variable region from NGS to identify all unique variants of a protein in NGS datasets. Conveniently the software can reduce sequencing errors by using overlapping paired-end reads. Enrich can also use the frequency of each variant before and after selection in protein displayed libraries to calculate an enrichment ratio[96]. Kim *et al.* introduced one of the first programs to analyze large data sets from deep-sequencing of phage selection experiments[97]. The software termed MUltiple Specificity Identifier (MUSI) determines consensus motifs, the data can be visualized as a set of peptide sequence logos. It has already been used in several publications[73,91,97]. Aside from MUSI there are other tools to determine consensus motifs, such as: SLiMSuite[98], MCAFFT[89], and the Multiple EM for Motif Elicitation (MEME) algorithm[7]. These tools are used primarily for the identification of peptide motifs.

Although many groups are developing their own analysis tools, it might be more efficient to improve on original analysis tools to overcome deficiencies and increase the chance of identifying binding sequences. Our group was one of the first to developed MatLab-based software for the analysis of phage-selected peptides sequenced by Illumina[64]. The software was first used in analyzing commercially available phage display libraries from NEB. It provides information about DNA sequences, amino-acid sequences, and abundance. This software contains several procedures, such as sorting of sequences, quality filtering, abundance ranking, translation, and defining library size, start and end positions for expanded use in analyzing different peptide libraries[64]. Using the same software, Scott *et al.* expanded the analysis to determine significantly enriched binding ligands from the original library based on volcano analysis[92]. This analysis method compares the statistical significance (p-value from an ANOVA model) and the magnitude of change (log2 of fold change) between two large datasets composed of replicate data. More recently, Heinis and co-workers expanded on this software to analyze deep sequencing data of phage-selected peptides based on sequence homology[55]. They hypothesized that sequence homology information could provide information about binding interactions and could allow for a identification of sub-groups of consensus sequences. The approach of Heinis et. al. starts from calculating the pair-wise distances among peptides, followed by calculation of a phylogenetic tree, and clustering the peptides in suitable groups.

Currently, analysis programs are not always available from the supporting information of research reports. Providing these programs from a central repository, like TIGR and NCBI, which already provide tools to analyze DNA sequences, could lead to more standardized approaches in analyzing deep-sequencing data.

1.3. Scope of the thesis

Incorporation of deep-sequencing into phage library screens simplified the identification of binding ligands, and in some cases has simplified the selection process itself, by reducing multi-round panning to one round. The methods for deep-sequencing described in Chapters 3 and 4 aided in the rapid preparation and analysis of diversity and frequency of a peptide library before and after selection. These methods can serve as starting points for analysis of sequencing results from other platforms or analysis of other peptide-libraries. While deep-sequencing can detect the undesired collapse of diversity, it cannot be used to fix that collapse. The emulsion amplification method described in Chapter 2 is one of the techniques that we propose to use for mitigating the collapse in diversity. Incorporating emulsion amplification into selection described in Chapter 5 facilitated the selection of a broader class of ligands to a target with multiple binding sites, such as cells, tissues or mixture of antibodies.

Chapter 2 Uniform amplification of phage display libraries in monodisperse emulsions

2.1 Introduction

The amplification process in phage library selection is essential to carry out multiple rounds of selection and enrich for target binding ligands. Additionally, amplification makes it possible to trace and identify peptides even if they are present as a single copy in a mixture of phage-displayed peptides. Ideally, amplification should increase the amount of every clone uniformly. In practice, phage that present different peptide sequences have different growth rates[18]. Each infected bacterial cell produces *ca*. 1000 phage particles[2]. Such rapid, exponential growth makes the amplification process sensitive to minute variations in growth rate. Both experimental results and modeling suggest that phage that amplify in bacteria a mere 10% faster can rapidly outcompete the phage that amplify more slowly[10,99]. Competition among phage clones during amplification leads to undesired collapse of library diversity. This phenomenon was first observed in libraries of peptides displayed on the pVIII coat protein[2,37,38,41]. For pIII-displayed libraries, Rodi and Makowski analyzed the sequences in phage libraries and demonstrated that amplification selects against sequences that interfere with the life cycle of phage[34]. Recently, we analyzed several hundred screens that had used pIII-displayed libraries of peptides[10]. The results suggested that amplification is a major driving force in decreasing the diversity of libraries. We hypothesized that rounds of panning and amplification yield phage that (1) bind the target and (2) amplify faster than other target-binding clones[10].

To eliminate amplification-induced loss of diversity, phage must be separated into different growth chambers. We believe that two conditions are necessary for uniform amplification: (1) growing individual phage separately to a saturating concentration (*i.e.*, when bacteria reach stationary phase and all bacteria are infected); (2) combining equal volumes of phage solution after amplification. Clonal growth could be performed in separate test tubes or multi-well plates. However, this process is practical only for a small number of clones. Some research groups amplify phage clones as separate plaques in agar[46,51,100,101]. This process is laborious for large libraries. Also it is unclear if diversity is preserved in plaque-based amplifications, because phage clones grow in plaques continuously and never reach saturation.

Derda, Tang, and Whitesides demonstrated that requirements (1) and (2) can be satisfied when each phage is encapsulated into a droplet of media suspended in oil (emulsion)[99]. Droplets act as isolated growth chambers for phage and bacteria. After a few hours of growth, destabilization of the emulsion releases the phage clones and yields a library that is uniformly amplified. As expected, the number of phage virions produced per droplet increases with droplet size. It is important, therefore, to amplify phage in emulsions in which droplets have identical size (monodisperse emulsions).

Microfluidic devices can be used to generate micrometer-sized droplets at high speed (100 Hz to 10 kHz) [25,102]. These droplets can serve as compartments for growth of cells and bacteria[23-25]. Droplet size is the function of channel geometry and pressure in the channel; if pressure is constant, the droplets produced in the channel have nearly uniform size (polydispersity of 1.01) [103,104]. Microfluidics is one of the simplest approaches for rapid generation of large number of monodisperse growth compartments (*e.g.* 10^6). Other potentially interesting approaches for large-scale clonal isolation of cells are arrays of micro-wells [105], arrays of micro-pockets [106] or microbeads [107]. Polydisperse emulsions produced by simple mixing are useful for many biochemical applications[19,21,108,109]; however, they are not suitable for phage growth because amplification in polydisperse droplets is not uniform[99].

In the following sections, we describe the steps for generation, handling, and destabilization of monodisperse emulsions. We describe the steps for efficient separation of phage into individual droplets, and identify conditions in which this separation is possible. We also provide a detailed description of perfluorinated surfactants that stabilize or destabilize these emulsions. Unlike previous reports[25,99,110], we describe the synthesis of these surfactants from commercially available materials.

2.2 Results and discussion

Amplification of phage libraries in water-oil emulsion follows three steps: encapsulation, incubation and destabilization. There are several requirements for successful amplification: (1) Phage must remain in separate droplets with bacteria hosts for the duration of amplification (4-5 hours). It is important to prevent coalescence of droplets during the incubation period. In previous reports, coalescence of droplets was prevented by storing the droplets inside gas permeable tubing[23,24,111]. Here we use a surfactant, developed by Weitz and co-workers [25], which imparts droplets exceptional stability. (2) Each droplet must contain a clonal population of phage. Techniques based on microfluidic sorting allow deposition of one cell in every droplet[23,106]. Such an approach requires complex microfabrication and instrumentation, and is unlikely to be useful for manipulation of the phage, which has dimensions of 0.01 x 1 micron. We use a much simpler approach based on dilution. Specifically, we explore a wide range of regimes, which lead to encapsulation of different numbers of phage per droplet, and investigate the result of amplification in each encapsulation regime. (3) During amplification of phage, the bacterial host requires oxygen and nutrients. While nutrients are supplied from the media in droplets, oxygen must diffuse from the outside. Emulsions, thus, cannot be formed in hydrocarbon oil-water systems, because mineral oil is impermeable to oxygen. Perfluorinated solvents, however, are highly permeable to oxygen. (4) Encapsulation of phage in droplets is necessary only for amplification. Once amplification is complete, viable phage must be recovered from every droplet by destabilizing them. Destabilization of all droplets must be mild and quantitative. We describe several methods for destabilization, among which chemical destabilization is the most robust.

2.2.1 Generation and handling of droplets

The stability of droplets is the most critical aspect of droplet-based amplification. At the time of the publication of this chapter[112], there were several reports of droplet-based culture of bacteria[23,111]; but the droplets described in these reports had limited stability. For example, Ismagilov and co-workers [111], stabilized aqueous droplets by using a triethyleneglycolmonol[1H,1H-perfluorooctyl]ether (RfOEG) surfactant, and used these droplets to culture rare bacteria. Due to the short length of the fluorinated chain in RfOEG, the droplets stabilized by this surfactant were prone to coalesce on contact [113]. To avoid coalescence, the bacteria had to be cultured as a string of droplets inside gas permeable tubing. This approach is cumbersome when a large number of droplets are incubated. For example, in each of our experiments, close to a million 165 μm droplets are generated. Storing these droplets in a tube would require hundreds of meters of tubing (assuming that 2-3 droplets are spaced per mm length of tubing).

Weitz and co-workers tested a variety of perfluorinated compounds and selected several surfactants that produced droplets with exceptional stability [25]. They did not coalesce with one another or with walls of the microfluidic channel. Here we outline additional practical considerations for the generation and handling of these stable emulsions.

Published reports for generating aqueous-perfluoro emulsions contain no discussion about the role of perfluoro-solvent chemical composition in stabilizing emulsions. Perfluorinated solvents drastically differ in their polarity and surface tension. We noticed that the chemical composition of the perfluoro phase used to generate droplets has a dramatic effect on droplet stability. We tested several commercially available perfluorinated solvents with different hydrophobicities and surface tensions (HFE-7500, HFE-7100, PFMD). Stable generation of droplets was achieved only in HFE-7500. In other solvents, droplets could be generated but were prone to coalescence.

During the generation of droplets, we took several precautions to eliminate droplet coalescence. For example, we collected the droplets into a petri dish filled with a perfluorinated solvent "cushion", because droplets coalesce on contact with dry polystyrene. We used FC-40, or HFE-7500 as a "cushion" for storing droplets due to the high boiling points of these solvents. The outlet of the channel generating the droplets was positioned close to the interface of air and fluorous solvent to minimize splashing, contact with walls or droplet breakup [114]. We also minimized exposure to static electricity, by placing a wet filter paper below and above the petri dish.

Overall, droplets produced in HFE-7500 + surfactant and stored atop a perfluorinated "cushion" were amazingly stable. For example they did not coalesce after several weeks of storage at room temperature.

2.2.3 Determining suitable concentrations of phage for uniform amplification

The key for successful uniform amplification of phage is the separation of individual phage into separate compartments. The simplest approach is based on the statistical distribution of phage into droplets during the generation of droplets in a microfluidic channel. When the ratio of phage to droplet is equal to 1, the

Poisson probability for a droplet to contain 0, 1, 2 or 3 phage particles is 37, 37, 18 and 6%, respectively. Thus, a significant fraction of droplets will encase monoclonal populations.

Increasing the phage to droplet ratio is detrimental, because it increases the probability of co-encapsulating of two or more phage clones within the same droplet (Figure 2.1A). Decreasing the ratio increases the fraction of “clonal” droplets, but it also increases the fraction of droplets with no phage (Figure 2.1A). Amplification in low phage/droplet ratio leads to a decreased yield of phage libraries after amplification (Figure 2.1A). There is an optimal window, in which the concentration of the amplified phage is the highest, and the population of droplets containing monoclonal populations is optimal.

To determine whether populations are monoclonal, we amplified a mixture of slow and fast-growing phage. Their rates of growth were characterized in previous publications[10,99]: slow-growing (engineered) phage produce progeny after 90 minutes, while fast-growing (wild type) phage produce progeny after 70 minutes. These subtle differences in growth rate are sufficient to cause a 100-fold difference in phage abundance after amplification. If the mixture of phage is amplified in the same solution, fast-growing wild type phage rapidly outcompete the slow-growing phage displaying peptides (“library phage”). This competition can be easily detected, because library phage contain a LacZ reporter and formed blue plaques in a bacterial lawn on IPTG/Xgal plates. Wild type phage formed white plaques in the same conditions. When a 1:1 mixture of slow and fast phage was amplified in the same solution, the white-blue ratio reached 100:1 or 1000:1

after amplification[99]. Amplification in sub-optimal conditions, when multiple phage are co-encapsulated in one droplet, should lead to increase in competition and increase in white-blue ratio.

Droplets used for amplification had a diameter of 165 μm; 4.25x10^5 drops could be formed from one milliliter of solvent. To generate a 1:1 ratio of phage to droplets, a 4.25x10^5 pfu/mL concentration of phage is required. We used a mixture of wild type and library phage with concentrations above and below this concentration, and determined the ratio of each phage after amplification.

As expected, after amplification of phage encapsulated from solutions of <4.25x10^5 pfu/mL concentration (left of dotted line in Figure 2.1B) the ratio of wild type and library phage was similar to the input ratio. However, a higher proportion of empty droplets resulted in a lower amount of amplified phage.

Encapsulation of greater than one phage per droplet yielded a population with an increased proportion of wild type phage (right of dotted line Figure 2.1B). Figure 2.1B suggests the optimal concentration of library for emulsion amplification. Conveniently, the optimal window, 10^4-10^5 pfu/mL is similar to a typical outcome from phage screens. Outcome with higher concentration should be diluted to an optimal lower concentration. Amplification of phage at much lower concentration does not produce sufficiently high concentration, but this problem could be mitigated by double amplification. For example, emulsion amplification of 4.25x10^2 pfu/mL solution yielded ca. 10^{10} pfu/mL solution (100x lower than standard amplification in shaking culture). Diluting this solution to the

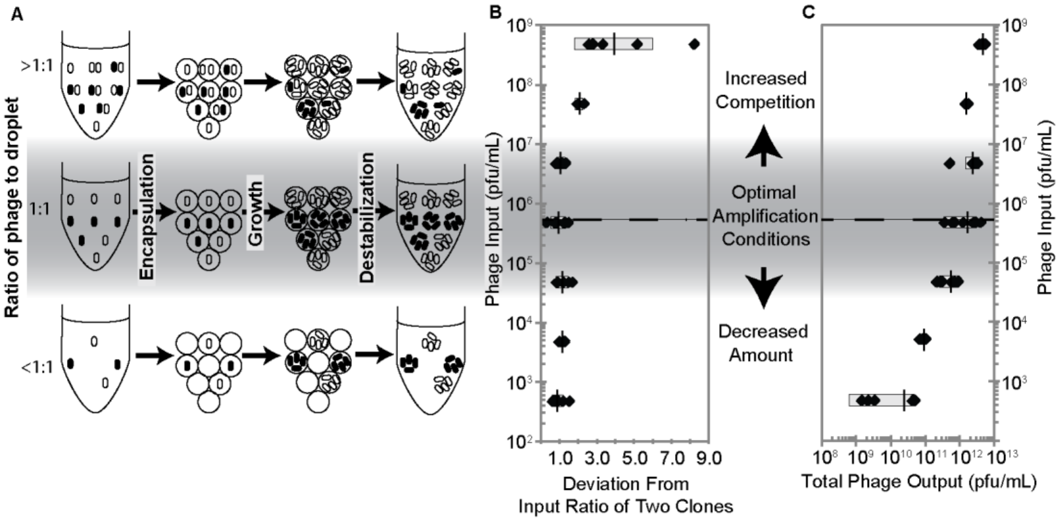


Figure 2.1. In this experiment, we formed an emulsion containing droplets with a diameter of 165 µm; there are 5x10^5 droplets formed from 1 mL of LB. At phage concentration of 5x10^5 pfu/mL, there is 1:1 ratio between phage and droplets (dotted line). (A) Schematic representation of the result of amplification of phage with >, =, or <1 phage per droplet. White particles represent fast-growing, wild-type (WT) phage while the black slow-growing, library (L) phage. (B) At phage concentrations >10^7 pfu/mL, phage-to-droplet ratio exceeds 1 (i.e., the there are > 1 phage per droplet). In this case, we observed competition between phage clones, which was manifested as an increase in WT:L ratio. (C) At phage concentrations <10^5 pfu/mL, phage-to-droplet ratio was significantly lower than 1; majority of the droplets were “empty” and the overall concentration of amplified phage was low. Concentrations of phage 10^5-10^7 pfu/mL represent an optimal window for the amplification in 165 µm droplets. Note: this optimal window will change if the size of droplets is different. (Figure B and C contains all data points with the overlaying gray bar equal to 2 x standard deviation).

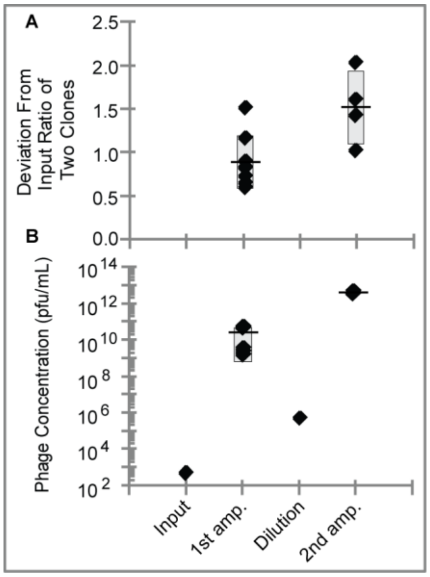


Figure 2.2. (A) Deviation from the initial ratio of wild type to library phage remains relatively constant as (B) a mixture of wild type to library phage are amplified first at a low concentration followed by re-amplification at a higher concentration. (The figure contains all data points with the overlaying gray bar equal to 2 x standard deviation).

optimal 4.25×10^{5} pfu/mL and re-amplifying, yielded desired 10^{12} pfu/mL. Although the phage population was amplified twice, the ratio of slow and fast growing phage in this population did not increase significantly (Figure 2.2).

2.2.3 Survey of destabilization conditions

Destabilization of emulsion is essential for the recovery of amplified libraries. The destabilizing agent must be mild, and should not interfere with the viability of phage or bacteria stored inside the droplets. Previously described reports of destabilization conditions of perfluorinated emulsions rely on proprietary chemical destabiliser (RainDance) of unknown structure. Therefore, we surveyed several destabilization conditions:

Destabilization by washing. This strategy for destabilization emerged from our observation that droplets are not stable in some perfluorinated solvents (e.g. HFE-7100). We hypothesized that if HFE-7500/surfactant + FC-40 is replaced by surfactant-free HFE-7100, stability of droplets will decrease dramatically. To test this approach, we transferred the droplet suspension to a separatory funnel, drained the perfluoro (bottom) layer and washed the droplets 2-3 times with HFE-7100. After several washes, the majority of the droplets coalesced and formed a continuous aqueous layer. This approach was relatively simple, but it rarely yielded complete destabilization. A small volume of residual emulsion was often present at aqueous-perfluorinated interface even after several washes.

Destabilization by electrostatics. Small number of droplets can be easily coalesced using static electricity. Rubbing the dish, in which the droplets are stored, with dry paper led to coalescence of droplets in the dish. This approach, conceptually, is similar to electrofusion used by Weitz and co-workers to coalesce the droplets continuously in microfluidic channels [115,116]. Destabilization by electrostatics was effective only when a small number of droplets were present in a dish (a few thousand droplets, which form less than a monolayer of coverage).

Chemical Destabilization. Review of literature about chemical destabilizers of oil/water emulsions (anti-foaming agents) suggests that destabilizers often have structure similar to that of the surfactant: poly(ethylene glycol) (PEG), and fatty alcohols or acids can be used as anti-foaming agents. In a

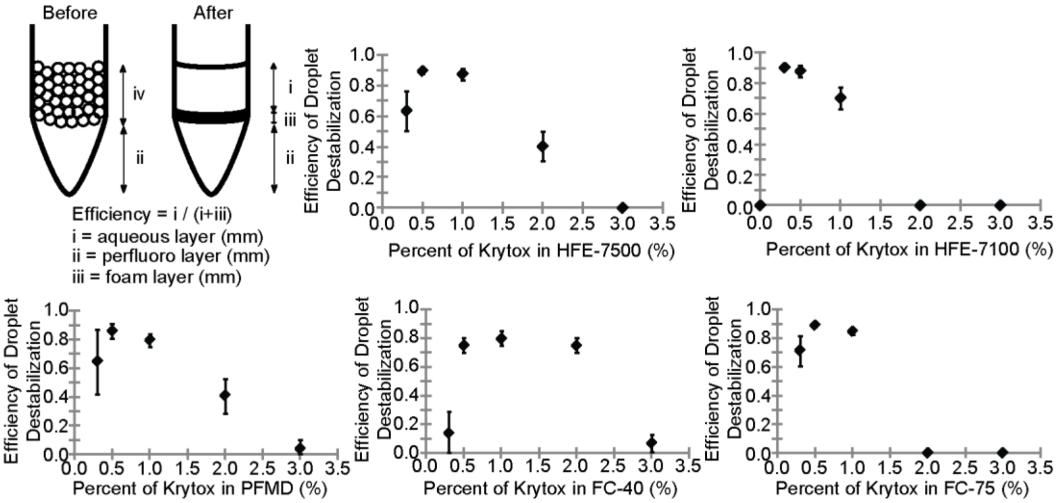


Figure 2.3. Efficiency of droplet destabilization determined by the height of the aqueous layer over the sum of the aqueous and foam layers. Plots of the efficiency of droplet destabilization by Krytox in various perfluorinated solvents include average efficiencies with error bars equal to 2 x standard deviation).

perfluorinated system, the latter chemicals should be replaced by long-chain perfluoroalkyl carboxylic acids or perfluoroalkyl alcohols. We tested PEG as a destabilizer, but it was effective only at very high concentrations (10 wt %). Interestingly, perfluoropolyether carboxylic acid, Krytox, was not effective at high concentration, but it was an excellent destabilizer at a specific, low concentration (Figure 2.3).

To map the best conditions for destabilization, we prepared Krytox at different concentrations in five different perfluorinated solvents: HFE-7500, HFE-7100, PFMD, FC-40, FC-75. To test the destabilization, we mixed a suspension of droplets with an equal volume of Krytox solution, centrifuged the mixture briefly to separate the layers and measured the height of aqueous, perfluorinated and foam layers (Figure 2.3). We defined destabilization efficiency as a ratio of the height of the aqueous layer over a sum of aqueous and foam layers (the foam layer included un-coalesced droplets). The ratio, thus, approached zero when the destabilizing solution was ineffective, and approached 1 when destabilization was complete. The optimal concentration of Krytox for destabilization was 0.5% (w/v). The optimal and most economically viable solvent was HFE-7100.

2.3 Conclusions

Emulsion amplification of phage libraries eliminates the competition between phage clones that have different growth rates. In emulsion amplification, all clones within a library of phage-displayed peptides amplify uniformly. We anticipate that eliminating growth bias will enable identification of rare ligands and improve identification of ligands for multiple targets (i.e., cells, tissues, etc). The elimination of growth-based competition also ensures that selection of binding clones in phage display is driven only by the binding strength of each clone. In such selections, one could potentially predict the affinity from the abundance of the clones and make conclusions based on motifs that are absent from the screen: (i.e., sequences which are not enriched during the selection do not bind to the target). Such structure-activity analysis would be empowered by large-scale sequencing of phage libraries. To accomplish this goal, one could use existing next-generation sequencing tools; specifically, in chapter 3, we describe the use Illumina deep-sequencing to analyze a phage library[64].

By investigating different amplification regimes, we demonstrated that emulsion amplification works best for re-amplification of phage libraries of up to 10^7 clones. It is, thus, ideally suited for amplification of sub-libraries generated during the selection process of panning which typically selects for 10^5 - 10^6 clones. Amplification of phage libraries that contain more than 10^7 clones, such as naïve libraries that contain 10^9 clones, is possible, but it requires improvements in droplet-generation technology. Specifically, re-amplification of a billion-scale library is a challenge because it requires generation of 10^9-10^{10} droplets within 1 hour (prior to generation of the first progeny of phage). Single nozzle microfluidics channel we describe in the paper cannot be used for these purposes, because at a typical droplet generator speed (~1 kHz), less than 4x10^6 droplets are produced in an hour. A droplet generator with multiple droplet-generating nozzles could be used to reduce the production time [117], but over a hundred of parallel nozzles would be required to generate 10^9 droplets. Additionally, the amount of media and surfactants required for this amplification is significant. Specifically, there are 5x10^5 of 165 μm droplets in 1 mL of media. Generation of 10^9 droplets, thus, would require 2 L of growth media, 3-4 L of perfluorinated media with ~50 g of surfactant. Such production is unlikely to be a "daily routine" for an average academic lab. But it might be of interest to biotechnology companies that produce libraries commercially. On the other hand, re-amplification of small number of clones (<10^7) is a routine lab-scale procedure, which can be done with small volumes (1-3 mL) and reasonable times (30-40 min).

Interestingly, the ability to generate $>10^9$ droplets could potentially be used to mitigate the loss of the diversity during the production of phage libraries. The diversity is lost during library production because the rates of protein production, periplasmic export, and secretion from bacteria is different for different sequences[34]. To fix this problem, one could encapsulate bacteria in monodisperse emulsions shortly after electroporation and allow each bacteria to secret phage for a prolonged period of time “in the privacy of its own droplet”.

Microfluidics became a routine technology in many laboratories. Yet, like any technology it has a learning curve. Complexity of droplet microfluidics is similar to that of SDS-page: Albeit it is routine in some labs, it could appear very challenging to other labs. Making a microfluidic channel from a master is similar to pouring a gradient SDS page gel. The electronic infrastructure needed to run SDS page (power source) is similar in its cost and complexity to syringe pump. Our manuscript describes a modular setup for droplet generation. In addition, commercial plug-and-play droplet generation solutions have been appearing on the market. We believe that both modular and integrated solutions are attractive to researchers. For example, some research groups pour gels, and assemble SDS-PAGE setup for each run; other labs use plug-and-play solutions with integrated gel-running modules and pre-cast gels such as eGel. Similarly, both modular droplet generation and integrated droplet generating instruments will be equally important.

2.4 Materials and methods

2.4.1 Synthesis and characterization of surfactant

Clausell-Tormos *et. al* described the use of tri-block co-polymer perfluorocarbon surfactants to encapsulate mammalian cells in microfluidic droplets [25]. Partial synthesis of this surfactant was described by Weitz and co-workers [110]. This synthesis, however, was difficult to replicate because it lacked careful characterization and relied on starting materials that are not available from commercial vendors. We describe the synthesis of a surfactant from commercially available materials as a sequence of three simple steps. Mitsunobu reaction, followed by deprotection converted polyethylene glycol into diamine derivative **2** (Figure 2.4). Reacting excess of diamine **2** with acyl chloride derivative of Krytox **3** produced surfactant **4** in high yields (Figure 2.4). The choice of tetrachlorophthalimide (TCP) protected intermediate was based on our previous experience with TCP groups: we reported that PEG-containing compounds, which are usually viscous liquids, can be easily converted into solids by introducing a TCP group[118]. The solid intermediates can be easily purified by crystallization. Unfortunately, it was difficult to crystalize the TCP intermediate **1** in this synthesis. The diamine derivative **2** was purified by solvent extraction and the surfactant **4** was used as crude to make stable droplets.

Characterization of high-molecular weight perfluorinated tri-block-co-polymer **4** was not trivial, due to its low solubility in organic solvents (for NMR), low ionization capacity (for mass-spectroscopy) and low mobility in conventional thin layer chromatography (TLC) plates/solvents.

Using FluoroFlash® TLC (Fluorous Technologies Inc.) as stationary phase and 5% (v/v) MeOH in HFE-7100 as mobile phase, surfactant **4** and Krytox can be visualized separately to confirm that the crude product does not contain unreacted Krytox. The TLC can be visualized by $KMnO_4$ staining solution. Kyrtox was stained poorly, but gave a distinct color from the background when the TLC was over-heated. (Figure 2.5A)

Using FTIR spectroscopy of neat surfactant **4** or Krytox proved to be the most successful in confirming that -COOH of Krytox (C=O stretch: 1775 cm^{-1}) was quantitatively converted to -CONH-R in compound **4** (C=O stretch: 1701 cm^{-1}) and was used to generate stable droplets. Incomplete reaction of acyl chloride Krytox can be easily distinguished with IR (Figure 2.5B and C). Interestingly, we observed that surfactant synthesized from an incomplete reaction (Fig. 2.5B) cannot form stable droplets because of the leftover Krytox (see Appendix 1.7-1.9 for FTIR spectra).

Using ESI or MALDI mass spectroscopy was unsuccessful with a variety of matrices, including perfluorinated matrix (pentafluorobenzoic acid). MALDI spectra contained only low-molecular weight compounds, which presumably have higher ionization capacity.

H^1-NMR was also successful in hexafluoro-2-isopropanol-d_2. Conversion of diamine derivative **2** to surfactant **4** was accompanied by a downfield shift of a CH_2 triplet from δ3.21 (methylene α to amine) to δ3.65 (methylene α to amide).

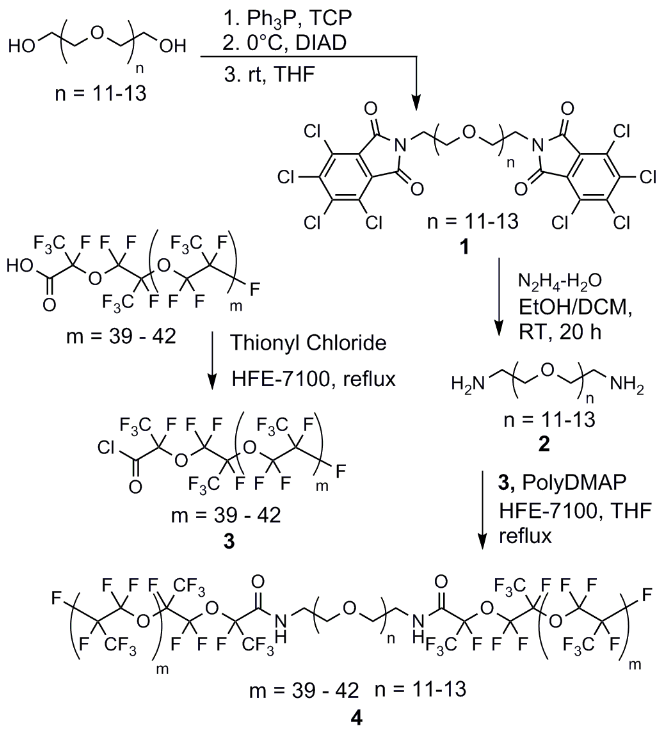


Figure 2.4. Reaction scheme for the preparation of surfactant **4** used in the generation of microfluidic droplets.

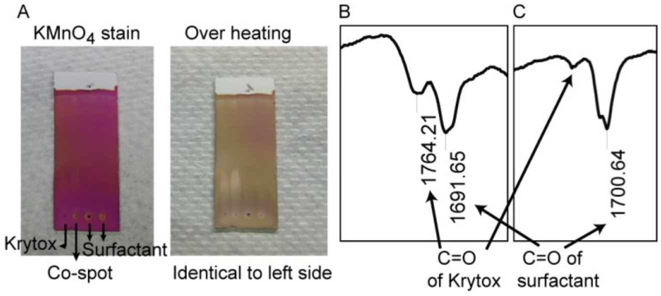
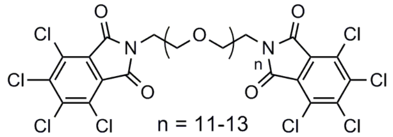


Figure 2.5. Characterization of surfactant **4**. (A) Krytox and surfactant visualized by $KMnO_4$ staining solution on TLC plates. IR spectra represents the C=O region from (B) an incomplete reaction and (C) complete reaction to surfactant **4**.

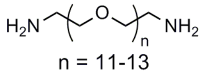
2.4.1.1 Synthesis of bis(tetrachlorophthalimido)-polyethylene glycol (1)



1

Polyethylene glycol (PEG avg MWt 600 g/mol) (3.00 g, 5 mmol) (Sigma 202401) was subjected to azeotropic removal of water with anhydrous THF (3 x 25 mL) (Sigma 196562) prior to use. The resulting compound was dissolved in 25 mL of anhydrous THF under nitrogen, triphenylphosphine (3.95 g, 15 mmol) (Sigma T84409) and tetrachlorophthalimide (4.27 g, 15 mmol) (TCI T0918) were added successively to the reaction mixture. The resulting mixture was cooled to 0 °C and DIAD (2.61 g, 15 mmol) (Sigma 225541) was added drop-wise to the reaction mixture over 10 minutes. The reaction was allowed to warm to room temperature and stirred overnight (15 h). The reaction mixture was then filtered and the filtrate was concentrated. This crude mixture was purified by column chromatography, first using hexane-ethyl acetate (1:1, v/v) to remove the impurities, followed by elution of product with CH_2Cl_2-MeOH (9:1, v/v). The eluent was dried under reduced pressure to yield **1** (3.09 g, 54%) as pale yellow oil. ^{1}H NMR (400 MHz, $CDCl_3$): δ 3.58-3.64 (m, 54H); 3.74 (t, *J* = 5.7 Hz, 4H); 3.90 (t, *J* = 5.7, 4H) (Appendix 1.1). ^{13}C NMR (125 MHz, $CDCl_3$); δ 38.1, 67.5, 70.1, 70.6, 127.7, 129.7, 140.1, 163.5 (Appendix 1.2).

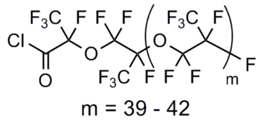
2.4.1.2 Synthesis of diaminopolyethylene glycol (NH_2-PEG-NH_2) (2)



2

To a solution of compound **1** (1.134 g, 1.00 mmol) in a mixture of EtOH (10 mL) and CH_2Cl_2 (3 mL), 64% hydrazine monohydrate was added (0.49 mL, 10 mmol). After stirring at room temperature overnight (20 h), the reaction mixture was filtered through celite (Sigma 22140) and washed with ethanol (30 mL). The filtrate was concentrated and the resulting residue was dissolved in H_2O (20 mL), and washed with diethyl ether (3 × 20 mL). The aqueous layer was collected and concentrated under reduced pressure. Traces of water were removed by azeotropic distillation with toluene using a rotary evaporator. The residue was further dried in *vacuo* to yield **2** (0.570 g, 95%) as pale yellow oil. ^{1}H NMR (500 MHz, D_2O): δ 3.21 (t, *J* = 5.1 Hz, 4H); 3.76 (t, *J* = 5.1 Hz, 4H); 3.71 (m, 48H) (Appendix 1.3). ^{13}C NMR (125 MHz, D_2O); δ 39.2, 66.4, 69.6, 69.7 (Appendix 1.4). HR-ESIMS for $C_{26}H_{57}N_2O_{12}$ Calculated: 589.3906 $[M+H]^+$; Found: 589.3901 (Appendix 1.5).

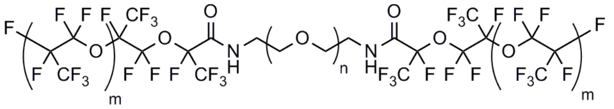
2.4.1.3 Synthesis of (perfluoropolyether)-acid chloride (PFPE-COCl) (3)



3

High viscosity FSH-Krytox (**5**) (MWt according to manufacturer 7000-7500 g/mol, 7250 g/mol was used for calculation of stochiometry) (7.4 g, 1.02 mmol) (Miller-Stephenson Chemical Company 157FSH) was dissolved in HFE-7100 (41 mL) (VWR 98-0211-8941-4). Thionyl chloride (0.37 mL, 5.1 mmol) (Sigma 88952) was added into the solution under nitrogen purge. After the reaction mixture was refluxed for 18 h, the solvent and volatile compounds were removed and the residue was dried in *vacuo*. The resulting pale yellow oil compound **3** was used directly without further purification.

2.4.1.4 Synthesis of bis(perfluoropolyether)-polyethylene glycol (PFPE-PEG-PFPE) (4)



m = 39 - 42 n = 11-13

4

Diaminopolyethylene glycol (**2**) (0.321 g, 0.54 mmol) and PolyDMAP (0.890 g, 5.1 mmol) (Sigma 359882) was dissolved in anhydrous THF (27 mL) under nitrogen protection. Compound **3** (7.41 g, 1.02 mmol) in HFE-7100 (27 mL) was then added into the suspension. The resulting mixture was stirred vigorously under nitrogen and refluxed for 24 h. The reaction mixture was filtered through celite to remove polyDMAP and the volatile solvent was removed in *vacuo* to give a white oil. The surfactant **4** was used without further purification. IR 2874, 1701 cm^{-1}; ^{1}H NMR (400 MHz, $C_3D_2F_6O$): δ 3.65 (m, 4H); 3.73 (m, 5H); 3.78 (m, 62H) (Appendix 1.6). Peaks are referenced to the hexafluoroisopropanol methine deuteron at δ 4.40 (septet).

2.4.2 Fabrication of master

Fabrication of the master was performed using standard soft-lithography as described by Derda and co-workers[99]. Briefly: A photomask was printed by CAD/Art Services Inc. SU8-50 was spin coated to form a 140 µm thin film, soft baked at 95 ºC and allowed to incubate for 24 hr at room temperature. The photomask was used to lithographically transfer the image of the channel onto a silicon wafer coated with SU8-50. The master was hard baked (95 ºC for 30 min), developed in SU8 developer for 15 min, rinsed in isopropanol, dried, and silanized with trichloro(1H,1H,2H,2H-perfluorooctyl)silane. For a visual demonstration of the fabrication of Si/SU8-master, fabrication of the PDMS-channel and droplet generation, we recommend consulting the video-publication of Casadevall and co-workers[119].

2.4.3 Fabrication of channel

We used standard soft-lithography procedure to make PDMS channels, as described in previous publications [120]. The PDMS elastomer base and curing agents were mixed in a 10:1 ratio and degassed for 30 min in a vacuum desiccator. The PDMS was poured in a "petri dish 1" that contained the SU-8 master (to make the top layer of the channel) and a blank "petri dish 2" (to make flat bottom). 10 g of PDMS was able to cover an eight cm petri dish to make ca. 0.5 cm thick layer. After incubation at 60 °C for >1 hour, PDMS of similar sizes were cut out from dishes 1 and 2, and immediately covered with Scotch tape (to protect the PDMS from fingerprints and dust). Holes were punched at the top PDMS layer in locations X, Y, and Z (marked on Figure 2.6B) using a 1.25 mm biopsy puncher. Both PDMS layers were oxidized in the plasma cleaner for 1 min (remove the scotch tape prior to plasma cleaning!), and pressed together immediately after plasma cleaning. Oxidized PDMS sheets form a covalent seal when pressed together. A few minutes after oxidation, however, the surface chemistry of PMDS changes, and it only forms a reversible seal with other PDMS surfaces. A reversible seal is not suitable for microfluidic channels because it fails under pressure. The channel was incubated for at least one hour at 120 °C to strengthen the seal. The channels were then filled with Aquapel for 10 seconds. The Aquapel was thoroughly aspirated using a pipette connected to a vacuum pump. The channels were immediately washed with ethanol by dispensing ethanol at one end, and aspirating at the other. These steps make the channel hydrophobic. Thorough removal of Aquapel is critical because Aquapel solution hydrolyses and forms insoluble precipitates that clog the channel. Channels can be stored at room temperature for up to 3 years and autoclaved prior to use. Channels can also be re-used multiple times. It has to be properly rinsed with EtOH after each use and stored in EtOH to prevent any regrowth of microorganisms.

2.4.4 Amplification of Phage Library in Droplets.

A 3 cm petri dish was placed with 500 µL of FC-40 inside a 14 cm petri dish with a wet paper (Figure 2.6B). The wet paper served a dual purpose: (1) it provided a humid environment for droplets; (2) it minimized the accumulation of static electricity on dry plastic dishes, which can promote coalescence of droplets. One end of a non-sterile PE tube was connected to the outlet of the channel (Figure 2.6B) and the other end was placed at the top surface of the perfluorocarbon layer of FC-40. It is important that the tubing is positioned at the air-liquid interface to ensure that droplets float on top of the perfluoro layer as they exit the tubing. If droplets are pressed against the bottom of the plate or fall into the plate from significant height, they could split into smaller droplets or coalesce into larger droplets. The perfluoro phase was prepared by dissolving 2% (w/v) surfactant (**4**) in HFE-7500. Droplets formed in other perfluorinated solvents were not stable (see Section 2.2.3). A 10 mL syringe was filled with the perfluoro phase, secured onto a syringe pump, and the needle of the syringe was connected to the perfluoro phase (PF) inlet (Figure 2.6B) with PE tubing. The tubing was checked to ensure that no air bubbles were present. Air bubbles can change the fluid pressure as they enter and exit the channel; these variations in pressure can dramatically alter the size of the droplets. The flow of perfluoro liquid was started at the desired flow rate (see below), and the liquid allowed to fill the whole channel. After the channel is filled with perfluoro phase, 2.81 mL of lysogeny broth (LB) medium, 141 μL of log phase bacteria culture (5 x 10^7 CFU/mL) and 10 μL of library phage (*ca.* 10^8 pfu/mL) were mixed in a suitable container (*e.g.* 5 mL Falcon tube). The mixed solution was loaded into a 3 mL syringe, secured to a syringe pump, and the needle of the syringe connected, *via* PE tubing, to the aqueous phase (AQ) inlet (Figure 2.6B). The aforementioned values are representative examples of conditions that produce on average one phage and >10 bacteria per droplet. Other conditions for different droplet sizes and concentrations of bacteria and phage can be easily calculated. It is important to mix phage and bacteria only after securing and connecting a syringe with the perfluoro phase to the channel. Infection of bacteria by phage occurs shortly after mixing, but the first progeny are generated only 1 hour after mixing (at room temperature). It is important that progeny generation occurs only when all phage clones are separated into droplets.

The flow rates of the aqueous and perfluoro phases were maintained at 4 mL/h and 6 mL/h, respectively, using syringe pumps. These flow rates produce droplets of 165 μm in diameter and >1.0x10^6 droplets are generated in 30-40 mins. Other flow rates can be used to vary the size of droplets. See original publication for the theory of droplet generation and list of regimes in which stable generation occurs[103]. Size and monodispersity of the droplets are usually estimated by characterizing the droplets as they emerge from the flow-focusing nozzle. Droplet generation, however, is too fast to be detected with the naked eye or ordinary cameras. We used a Phantom V7.3 ultra-fast camera (Vision Research) to monitor droplet generation at the nozzle. Alternatively, droplets can be observed at the outlet of the channel. At this location, droplets are slow and can be easily characterized without expensive cameras. We recommend discarding the droplets generated in the first 40 seconds of the operation of the microdroplet generator (these droplets have variable sizes because the flow pressure is not stable during the first few seconds; once the flow stabilizes, it takes 10-30 seconds for the polydisperse population to exit the tubing). The remaining droplets were collected into the dish filled with FC-40. Running the aqueous phase at 4 mL/h allows emulsification of 2-3 mL in less than one hour (before the first phage progeny are generated). Once all of the aqueous solution is converted to droplets, the petri dish is closed and placed in a temperature-controlled shaker (60 rpm, 37

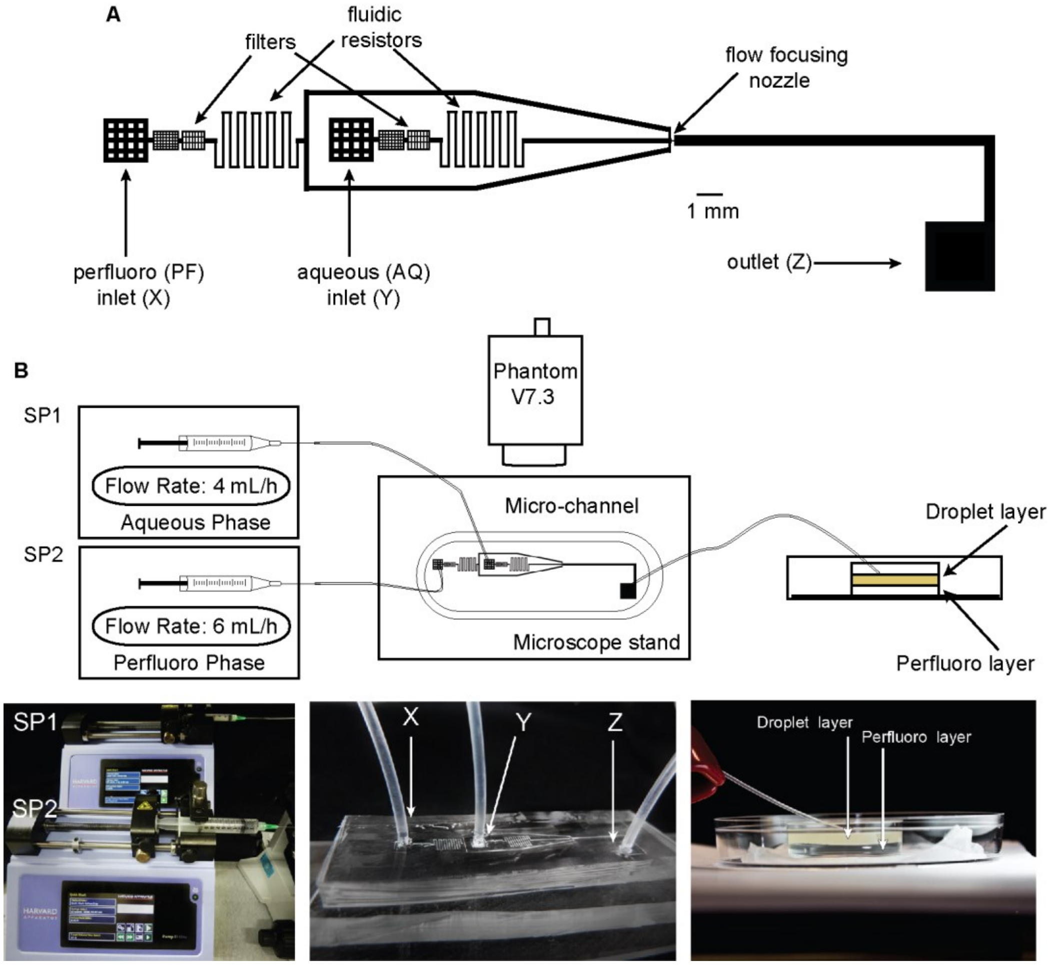


Figure 2.6. Representation of the microfluidic set-up. (A) The design of microfluidic channel. (B) Schematic of microdroplet set-up including; side views of syringe pumps (SP), micro-channel, and outlet from micro-channel positioned in the droplet layer.

°C). The libraries are amlified for a standard time (4-5 hours) and the phage harvested from the droplets as described in “destabilization of droplets”.

2.4.5 Destabilization of droplets and isolation of phage libraries

The mixture of droplets and perfluoro liquid were transferred into a 15 mL conical tube and centrifuged at 1,000 rpm for 2 min to establish discrete layers containing pure perfluoro solvent (bottom phase due to its higher density) and "droplet foam" (top phase due to the lighter density of media). The "droplet foam" was distributed in 500 μL portions into several 1.5 mL microcentrifuge tubes. An equal volume (500 μL) of destabilizing solution (0.5 % (w/v) Krytox in HFE-7100) was added to the droplets. The solution vortexed and centrifuged at 14,000 rpm for 2 min. The top layer contains phage solution, the bottom layer contains the perfluoro phase, and the bacterial pellet is positioned at the interface of the two layers. The top layer was transferred to new 1.5 mL microcentrifuge tubes and centrifuged for 2 min at 14,000 rpm to remove the remaining bacteria. The supernatant was mixed with 40% volume of PEG/NaCl solution (46 g PEG, 44 g NaCl in 300mL of H_2O) to precipitate the phage (2 h or overnight at 4 °C). The phage was pelleted by centrifugation (15 min, 14,000 rpm); the pellet re-suspended in a suitable buffer (e.g. PBS) and used for titering or for panning.

Chapter 3 Deep sequencing analysis of phage libraries using Illumina platform

3.1 Introduction

The confirmation of a successful selection by phage display is the analysis of peptide sequences present in the library at various steps of the screening. Sequences enriched as a result of selection correspond to the specific binders against the target. Conventional Sanger sequencing of clones require isolation of DNA from individual phage clones. It is a labor-intensive process and is rarely used to analyze more than a hundred library clones. The shallow coverage of the library provided by Sanger sequencing does not reflect the true abundance of the clones in the phage library and could identify false binders. Sequencing technologies with throughput higher than 10^4-10^5 could provide more complete coverage of the libraries. Increased throughput could also allow analysis of multiple experiments in a single run. Illumina/Solexa deep-sequencing technology analyzes a library of blunt-ended double-stranded DNA (dsDNA) fragments and generates up to 10^9 base pair (bp) reads in a single run.

In this chapter, we describe a one-step PCR protocol that converts a library of M13KE plasmids isolated from a phage library into a collection of short dsDNA sequences suitable for Illumina sequencing. Using custom MatLab software, we perform large-scale analysis of sequence diversities. Using deep sequencing, we explore the effects of amplification of phage libraries in bacteria on the diversity of peptides in these libraries. In previous publications, the result from sequencing of ~100 phage clones suggested that the amplification process enriches for specific peptide sequences[10,12]. Deep-sequencing, however, can provide observations that could not be interpreted from the sequence of 100 clones [121]. For example, deep sequencing of a library of DNA aptamers demonstrated that repeated amplification does not select for particular sequences. Instead, it enriches DNA sequence motifs that have low stability[93,94]. In this report, we analyzed the peptide diversity of amplified libraries using Illumina based deep-sequencing and observed a collapse of diversity in phage-displayed libraries after a single round of growth in bacteria. The collapse of the 10^6-scale library to a few hundred abundant sequences would not be apparent in small-scale Sanger sequencing[34,121,122]; it also would have been difficult to detect with smaller-throughput 454 Sequencing.

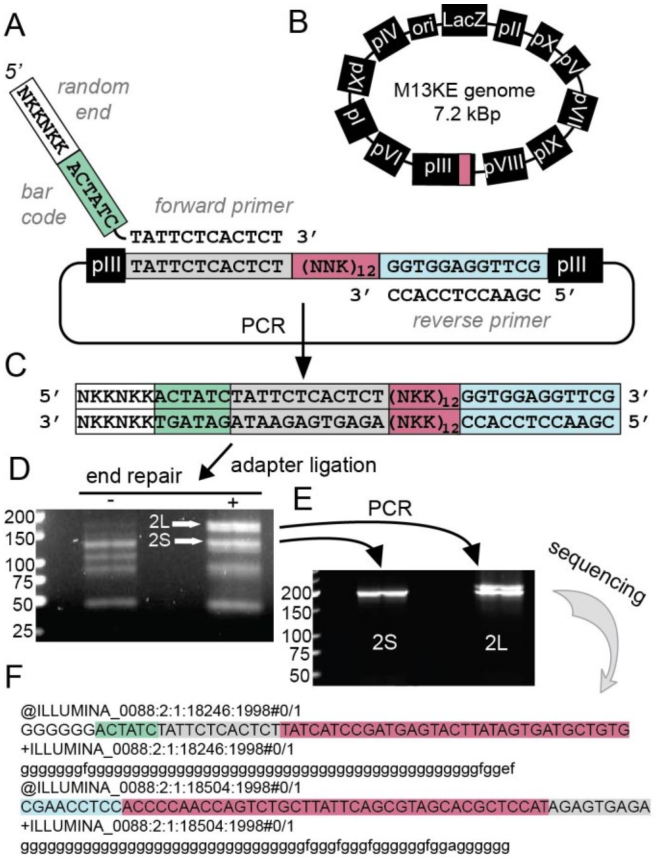
Characterization of sequence diversity is important for phage display technology, which has been used in over 5000 publications and patents in the past 20 years. It has enabled the discovery of ligands for hundreds of targets, yet the literature still contains several poorly-explained observations: (1) identical sequences could emerge from unrelated screens for unrelated target[123,124], and (2) screens that should yield a large number of diverse ligand often yield only one sequence motif (reviewed in[10]). The nearly complete sequence coverage of libraries by deep-sequencing illuminates the origin of these observations. It highlights that the collapse of diversity in amplification is a major limitation of phage-display technology. Deep-sequencing analysis also makes it possible to bypass problems originating from the unwanted collapse of diversity[89]. Large-scale analysis could also help developing methods that preserve diversity of peptide libraries[99,112]. Thus, deep-sequencing can be used to discover ligands that previously have been lost in phage library screens.

3.2 Results and Discussion

3.2.1 Isolation of variable ds DNA fragments from phage libraries.

The majority of the phage display vectors share the same design: they contain a variable sequence flanked by constant regions containing restriction enzyme sequences (used for cloning of the library). We attempted to isolate the library sequences using KpnI and EagI restriction enzymes to isolate variable domains from M13KE vectors [125]. The collection of sticky-end fragments could be filled in to give blunt-ended fragments with identical termini. These fragments, however, could not be reliably sequenced by Illumina because the sequencing algorithm uses differences in terminal nucleotides to distinguish sequence clusters[126]. We attempted to introduce variable termini by ligation of short random nucleotide sequences; this approach, however, gave poor yields and was eventually abandoned. Nevertheless, we expect that excision by restriction nucleases could be useful for other deep sequencing approaches, such as Ion Torrent, which could process fragments with identical termini.

Our successful method for isolation of variable regions used PCR amplification with primers complementary to 12-bp constant regions flanking the variable sequence in the M13KE vector. The forward PCR primer contained a

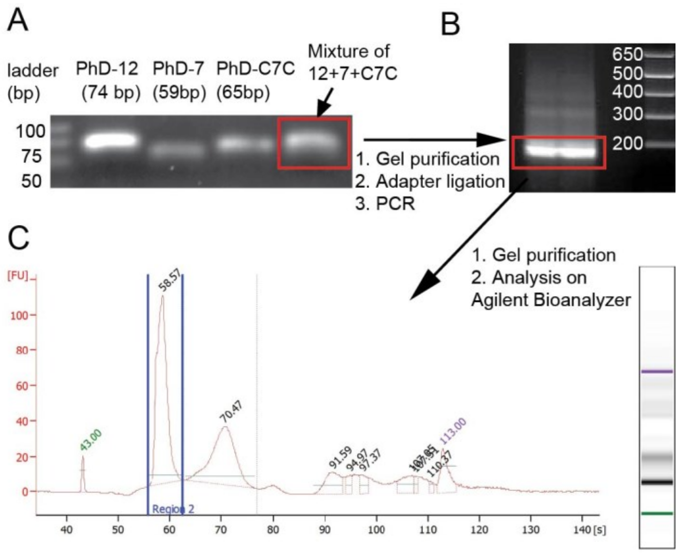


```
@ILLUMINA_0088:2:1:18246:1998#0/1
GGGGGGACTATCTATTCTCACTCTTATCATCCGATGAGTACTTATAGTGATGCTGTG
+ILLUMINA_0088:2:1:18246:1998#0/1
gggggggfggggggggggggggggggggggggggggggggggggggggggfggef
@ILLUMINA_0088:2:1:18504:1998#0/1
CGAACCTCCACCCCAACCAGTCTGCTTATTCAGCGTAGCACGCTCCATAGAGTGAGA
+ILLUMINA_0088:2:1:18504:1998#0/1
ggggggggggggggggggggggggggggggggfgggfgggfggggggfggagggggg
```

Figure 3.1. (A) Alignment of forward and reverse primers to 12-bp sequences flanking the variable region, (NNK)$_{12}$, at the N-terminus of the *pIII* gene in M13KE vector (B). (C) PCR product. The 5' of the forward primer, and one of the 5' of the PCR product contain random sequence NKKNKK, which should facilitate formation of clusters during Illumina sequencing. (D) Ligation of the Illumina single-end primers to fragment (C) with and without end-repair. Ligation after end-repair yields two products—large (**2L**) and small (**2S**)—both have the expected size (~140 bp). (E) PCR amplification of **2L** and **2S** with Illumina primers yields similar products, which yielded similar result after sequencing (see Fig 3). (F) Representative output from the sequencing in FASTQ format depicting forward and reverse sequence. Color-coding of the regions of the sequence is identical to that in panel (B). For details related to sequences, ligation of the adapters and PCR amplification see Appendix 1.10-1.11.

Table 3.1. The optimized conditions for each primer used in the amplification of each variable region of the phage library.

	BAR 1	**BAR 2**	**BAR 3**	**BAR 4**	**BAR 5**
H_2O	24.5 µL				
5x Phusion Buffer	10 µL				
10 mM $MgCl_2$	2.5 µL				
10mMdNTPs	1 µL				
10 µM left-BAR	5 µL				
10 µM right-BAR	5 µL				
M13 phage DNA template (50 ng/µL)	1.5 µL				
Phusion Hot Start DNA polymerase (2U/µL)	0.5 µL				
During amplification.					
Step 1 → 30sec	98°C				
Step 2 → 10sec	98°C				
Step 3 → 20sec	53°C	62°C	62°C	62°C	64°C
Step 4 → 30sec	72°C				
Step 5 → Repeat Steps 2-4, 34x					
Finish → 72°C 5min					
Hold → 4°C					



From [s]	To [s]	Corr. Area	% of Total	Average Size [bp]	Size distr. in CV [%]	Conc. [pg/µl]	Molarity [pmol/l]
55.77	62.50	438.4	46	196	6.8	1,401.89	10,873.9
62.66	76.91	313.5	33	327	10.8	861.46	4,054.0

Figure 3.2. Processing multiple libraries and multiple experiments using barcoded primers. (A) PCR Isolation of variable regions from three different libraries: 12-mer library (Ph.D.-12, New England Biolabs); 7-mer library (Ph.D.-7, New England Biolabs), cyclic 7-mer library (Ph.D.-C7C, New England Biolabs). Each library gives a single band on a gel. After PCR isolation, the libraries can be mixed together and processed as a mixture. (B) 2% agarose gel describing results after gel purification, end repair, adapter ligation and PCR amplification. Strong band at ~200 bp was excised from the gel. The product was analyzed on Agilent Bioanalyzer prior to Illumina sequencing. (C) Agilent trace and "gel view" or the trace (on the right). The table on the bottom describes molecular weight and concentrations for two major peaks.

Table 3.2: Sequences and melting temperatures of each primer. T_{anneal} is the optimal annealing temperature selected for each set of primers.

Primer	Right Primer		Left Primer		T_{anneal} (ºC)
		T_m (ºC)		T_m (ºC)	
BAR1	NKKN GTA CGA ACC TCC ACC	55.9	NKKN GTA TAT TCT CAC TCT	46.1	53.0
BAR2	NKKN GAC CGA ACC TCC ACC	58.7	NKKN GAC TAT TCT CAC TCT	48.7	62.0
BAR3	NKKN TTG CGA ACC TCC ACC	57.2	NKKN TTG TAT TCT CAC TCT	47.0	62.0
BAR4	NKKN TCA CGA ACC TCC ACC	53.4	NKKN TCA TAT TCT CAC TCT	46.5	62.0
BAR5	NKKN CGA CGA ACC TCC ACC	59.0	NKKN CGA TAT TCT CAC TCT	49.2	64.0

unique NKKNKK sequence (where N = any of the four deoxy-nucleotides, and K = G or T nucleotides) at its 5'-position (Figure 3.1A). Each primer, thus, was a mixture of 4x2x2x4x2x2 = 256 different primers. PCR with these primers generates dsDNA with 256 different bunt-end termini; this diversity should be sufficient for the algorithm that finds individual DNA clusters (polonies) during sequencing. We selected the NKKNKK sequence to minimize the possibility for hybridization with $(NNK)_{12}$ motifs in the library. The forward primer also contained a barcode sequence ATCACT. We selected this particular sequence after aligning all 256 (NKKNKK)-(ACTATC)-TATTCTCACTCT sequences to the (+) and (-) strands of the M13KE vector. For all sequences, we observed hybridization of <7 bp, which should not interfere with the PCR conditions optimized for 12 bp-long adapter sequences (Figure 3.1C). We used a similar algorithm to identify other barcode sequences (Table 3.1 and 3.2). The use of multiple barcodes allows the processing of multiple phage libraries in a single run (Figure 3.2).

Successful PCR amplification of variable fragments was confirmed as a single band on 2 % agarose gel. Amplification using primers with shorter variable regions or other barcode sequences yielded similar results (Table 3.1). Due to differences in melting temperatures of the primers, PCR conditions had to be re-optimized for each barcode sequence (Table 3.2). The fragments amplified from libraries of different size, such as 12-mer, 7-mer or 9-mer, gave dsDNA fragments of the expected sizes. For example, the protocol described in Figure 3.1A was validated using three different libraries: (1) Ph.D-12^{TM}, a library of 12-mer peptides encoded by a 36 bp degenerate region; (2) Ph.D-7^{TM}, a library of 7-mer peptides encoded by a 21-bp degenerate region, and (3) Ph.D-C7C^{TM}, a library of 7-mer peptides flanked by Cys residues, and encoded by a 27 bp variable region. We used two primers with a total length of 38 bps and observed PCR products close to the expected (1) 74, (2) 59, and (3) 65 bp (Figure 3.2A).

3.2.2 Preparation of Illumina-compatible dsDNA Fragments

Ligation of DNA adapters that enable Illumina sequencing was performed according to the protocols supplied with Illumina Paired-end Adapter Kit. Successful ligation of Illumina adapter sequences to the blunt-ended PCR product occurred only after end-repair of the product (Figure 3.1D). Ligation yielded two products, referred to as **2L** and **2S**, with length similar to that of the expected product (*e.g.*, 140 bp for the 12-mer peptide library, PhD-12). To enrich the DNA

A **Script:** **rawseq.m** 2h 28 min	**Sample time log for the processing:** Mac Air (laptop), 2.13 GHz Intel Core 2 Duo, 4 GB 1067 MHz DDR3 RAM, Mac Os X 10.6.8 file: s_1_sequence.txt 6.07 GB wrote rawseq0001 Load=46s Convert=0.45s Save=13s wrote rawseq0002 Load=56s Convert=0.71s Save=13s … wrote rawseq0129 Load=55s Convert=0.77s Save=12s
B **Script:** **parseq.m** 2h 43 min	rawseq0001.txt F:17.0s R:31.1s N:4.69s Mut:13.0s Trun:4.55s Rem:0.14s rawseq0002.txt F:17.1s R:31.3s N:4.24s Mut:11.9s Trun:4.33s Rem:0.166s … rawseq0129.txt F:17.8s R:30.5s N:4.48s Mut:13.0s Trun:4.83s Rem:0.191s
C **Script:** quaseq.m 1h 23 min	parseqF0001.txt 55484/85820 seq (64.7%). Rescued 0 seq (0%). Time: 11.8 sec parseqF0002.txt 58061/86134 seq (67.4%). Rescued 0 seq (0%). Time: 12.3 sec … parseqF0129.txt 68524/90520 seq (75.7%). Rescued 0 seq (0%). Time: 14.6 sec parseqR0001.txt 121852/162056 seq (75.2%). Rescued 0 seq (0%). Time: 25 sec parseqR0002.txt 136798/161763 seq (84.6%). Rescued 0 seq (0%). Time: 27 sec … parseqR0129.txt 139021/157190 seq (88.4%). Rescued 0 seq (0%). Time: 25.4 sec
D **Script:** uniseq.m 0h 47 min	quaseqQF0001.txt loaded. Found 22476 nuc in 52794 reads in 1.35 sec quaseqQF0002.txt loaded. Found 37023 nuc in 107998 reads in 1.57 sec quaseqQF0003.txt loaded. Found 48962 nuc in 164152 reads in 1.61 sec quaseqQF0004.txt loaded. Found 59797 nuc in 224116 reads in 1.8 sec … quaseqQF0129.txt loaded. Found 652813 nuc in 7902401 reads in 4.59 sec Wrote 652813 unique seq in 644 sec **Total time: 7h 21 min**

Figure 3.3. Processing of a typical FASTQ file by MatLab scripts. The scripts generate detailed output, which outlines the processing time for each step. Although we optimized the script to minimize the processing time, we believe some steps could be further optimized. (A) The script *rawseq* breaks the 6 GB FASTQ file into 129 files (rawseq0001.txt to rawseq0129.txt). Loading, conversion and saving time for each file is noted. (B) The *parseq* script loads files rawseq000N.txt files and searchers for forward (F), reverse (R) adapters, as well as adapters with unknown nucleotides (N), mutations (Mut), truncations (Trun) and saves unidentified, remaining sequences (Rem). Time (sec) used for each search is noted. (C) The *quaseq* script loads parseqF000N.txt and parseqR000N.txt files and assesses the quality of sequences. The output indicates

the number of high-quality sequences/total sequences. The script could also rescue low-quality reads with errors in N*3rd position, and saves those sequences that could be unambiguously translated. The rescued sequences comprise ~1.5% of total sequences, and are similar to high-quality reads. The rescue option was turned “off” for this particular run. If this option is turned one, the rescued sequences are saved in separate file with tags “EF” (erroneous forward) or “ER” (erroneous reverse). (D) The *uniseq* script loads quaseqQF000N.txt files (high-quality forward reads) and identifies unique nucleotide sequences. The nucleotides are then translated to peptides and all sequences are saved in uniqueQF.txt file. The same steps are repeated for quaseqQR000N.txt, (high-quality reverse reads). For clarity, the output for these files is not shown.

fragments, which successfully ligated to the adapters, we PCR amplified purified **2L** and **2S** fragments using primers that complement the Illumina adapters. Both **2L** and **2S** yielded products of correct size after this amplification (Figure 3.1E), confirming that both the **2L** and **2S** fragments contained correctly ligated adapters. Both products were subjected to Illumina sequencing (single-read, 50 bp reads on HiSeq) yielding similar sequence abundances and diversities (see Figure 3.5B below).

3.2.3 Design of the analysis software.

Sequencing by Illumina generates a ~4-10 Gigabyte text file. It is difficult to handle because, for example, most desktop computers cannot open the file in a standard text editor. Additionally, Illumina is used primarily for genome sequencing, and most available software is written for assembly of genomes. Therefore, we wrote software in MATLAB tailored for the analysis of phage libraries. The basic feature of the software is batch processing. The program first breaks the original 4-5 Gb FASTQ file into text files of ~100 Mb each. The subsequent processing, thus, requires less operational memory. Analysis proceeds in several steps: (i) conversion of one FASTQ file into smaller plain text files, (ii) identification of constant complementary regions and parsing of sequences, (iii) analysis of sequence quality, (iv) analysis of diversity of sequences, (iv) translation of sequences, (vi) log-log plots of sequence abundance. After each step, the program saves intermediate files in plain text (*.txt) format. Any intermediate text files can be opened and inspected in a standard text editor. Software written in MatLab was effective in analyzing a 4-5 Gb FASTQ file in 6-8 hours on an average desktop or laptop computer (Figure 3.3). We anticipate that re-writing the same script in a lower-level language (e.g. C++) could further accelerate the processing.

3.2.4 Overview of the scripts

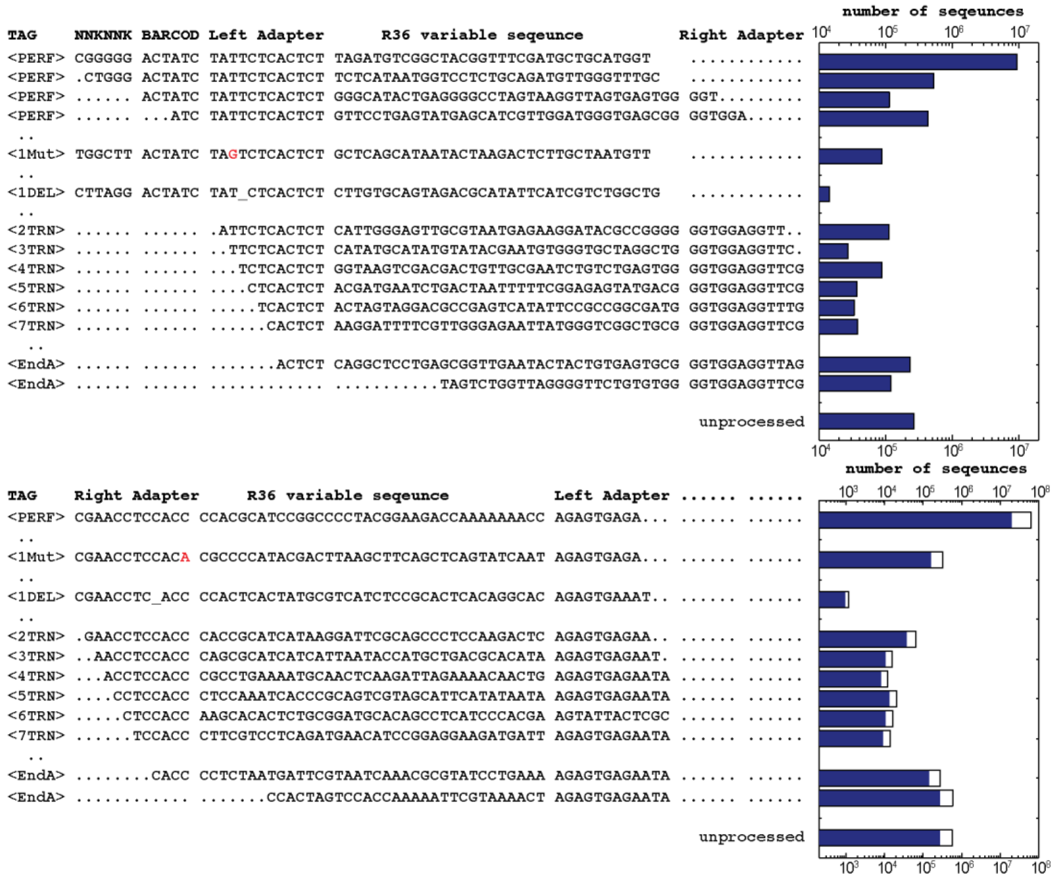
Although the length of the dsDNA construct depicted in Figure 3.1C is 72 bp, single-end sequencing yielded reads of only 57 bp and contained complete sequences for only one constant region: either from the forward or the reverse primer. We designed the algorithm to use one constant adapter region to map the functional portions of the sequence: (1) NKKNKK portion, (2) barcode portion, (3) left adapter, (4) R36, and (5) right adapter (see Figures 3.1A, C, F).

The process starts from *rawseq.m* scripts, which breaks the original FASTQ file into smaller text files, 250000 lines each. The *parseq.m* script then searchs for forward or reverse adapter sequences (highlighted grey or blue in Figure 3.1C). We used a multi-step algorithm to identify the adapters. The majority of the sequences were mapped by perfect alignment to full-length adapter sequence (<PERF> in Figure 3.4). 1% of sequences contained adapters with one mutation (<1MuT> in Figure 3.4; mutation is highlighted in red). A few adapters had one internal deletion (<1Del> in Figure 3.4; deletion is underscored, Figure 3.4). A significant fraction of adapters had terminal truncations (lines tagged as <2TRN> to <7TRN> in Figure 3.4). Truncated reads contained sequences of nucleotides from the i^{th} to the $(56+i)^{th}$ position (i=2-25). Finally, primers with excessive truncations in one complementary region could be identified by alignment with the complementary sequence at the opposite end of the variable region (lines labeled as <EndA> in Figure 3.4). This algorithm mapped the majority of the forward and reverse reads (Figure 3.4A forward and Figure 3.4B for reverse search). Approximately 1.6% of sequences (0.5 million) could not be mapped because they contained a large number of low-quality base calls or reads with multiple mutations or deletions in the adapter regions.

The parsed files were then processed by *quaseq.m* script that assessed the quality of the R36 region containing the $(NNK)_{12}$ sequences. We selected only high-quality output in which all nucleotides had a Phred Quality Score above 5 (this value could be changed in *quaseq.m* script on demand). High-quality sequences were then analyzed by *uniseq.m* script to generate abundances of nucleotides and their corresponding peptide sequences. The results were saved to *uniqueN_QF.txt* and *uniqueN_QR.txt* file (where F and R designate analysis of forward and reverse reads). These files are available as a part of the supporting information.

In summary, from 32 million raw reads, the software identified ~11.1 million forward and 20.2 million reverse reads from which R36 sequences could be extracted. From R36 motifs with NNK structure, the software extracted 8.5 and 17.8 million peptide sequences from forward and reverse reads respectively. In current analysis of 12-mer libraries, the majority of the forward reads were truncated at the 11th amino acid (see *uniqueN_QF.txt*). Reverse reads, however, contained sequences covering the full-length of the 12-mer peptide region (see *uniqueN_QR.txt*). We focused the remaining analysis on the 17.8 million reverse reads.

The script had options to retain or discard the sequences that did not follow the NNK format (*i.e.*, sequences with G or T in the third position of every codon). If non-NNK sequences were retained, the results contained a significant fraction of sequences with TGA stop codons. M13KE vectors with a stop codon in the N-terminal region of the *pIII* gene would lack N-terminal leader sequence and would not produce viable phage.[125] We concluded that TGA codons and other non-NNK codons are sequencing errors.



```
TAG     NNKNNK BARCOD Left Adapter       R36 variable seqeunce          Right Adapter
<PERF> CGGGGG ACTATC TATTCTCACTCT TAGATGTCGGCTACGGTTTCGATGCTGCATGGT    ...........
<PERF> .CTGGG ACTATC TATTCTCACTCT TCTCATAATGGTCCTCTGCAGATGTTGGGTTTGC   ...........
<PERF> ...... ACTATC TATTCTCACTCT GGGCATACTGAGGGGCCTAGTAAGGTTAGTGAGTGG GGT........
<PERF> ...... ...ATC TATTCTCACTCT GTTCCTGAGTATGAGCATCGTTGGATGGGTGAGCGG GGTGGA.....
  ..
<1Mut> TGGCTT ACTATC TAGTCTCACTCT GCTCAGCATAATACTAAGACTCTTGCTAATGTT    ...........
  ..
<1DEL> CTTAGG ACTATC TAT_CTCACTCT CTTGTGCAGTAGACGCATATTCATCGTCTGGCTG   ...........
  ..
<2TRN> ...... ...... .ATTCTCACTCT CATTGGGAGTTGCGTAATGAGAAGGATACGCCGGGG GGTGGAGGTT..
<3TRN> ...... ...... ..TTCTCACTCT CATATGCATATGTATACGAATGTGGGTGCTAGGCTG GGTGGAGGTTC.
<4TRN> ...... ...... ...TCTCACTCT GGTAAGTCGACGACTGTTGCGAATCTGTCTGAGTGG GGTGGAGGTTCG
<5TRN> ...... ...... ....CTCACTCT ACGATGAATCTGACTAATTTTTCGGAGAGTATGACG GGTGGAGGTTCG
<6TRN> ...... ...... .....TCACTCT ACTAGTAGGACGCCGAGTCATATTCCGCCGGCGATG GGTGGAGGTTTG
<7TRN> ...... ...... ......CACTCT AAGGATTTTCGTTGGGAGAATTATGGGTCGGCTGCG GGTGGAGGTTCG
   ..
<EndA> ...... ...... .......ACTCT CAGGCTCCTGAGCGGTTGAATACTACTGTGAGTGCG GGTGGAGGTTAG
<EndA> ...... ...... ............ ...........TAGTCTGGTTAGGGGTTCTGTGTGG GGTGGAGGTTCG

                                                                unprocessed
```

```
TAG     Right Adapter     R36 variable seqeunce            Left Adapter ...... ......
<PERF> CGAACCTCCACC CCACGCATCCGGCCCCTACGGAAGACCAAAAAAACC AGAGTGAGA... ...... ......
  ..
<1Mut> CGAACCTCCACA CGCCCCATACGACTTAAGCTTCAGCTCAGTATCAAT AGAGTGAGA... ...... ......
  ..
<1DEL> CGAACCTC_ACC CCACTCACTATGCGTCATCTCCGCACTCACAGGCAC AGAGTGAAAT.. ...... ......
  ..
<2TRN> .GAACCTCCACC CACCGCATCATAAGGATTCGCAGCCCTCCAAGACTC AGAGTGAGAA.. ...... ......
<3TRN> ..AACCTCCACC CAGCGCATCATCATTAATACCATGCTGACGCACATA AGAGTGAGAAT. ...... ......
<4TRN> ...ACCTCCACC CGCCTGAAAATGCAACTCAAGATTAGAAAACAACTG AGAGTGAGAATA ...... ......
<5TRN> ....CCTCCACC CTCCAAATCACCCGCAGTCGTAGCATTCATATAATA AGAGTGAGAATA ...... ......
<6TRN> .....CTCCACC AAGCACACTCTGCGGATGCACAGCCTCATCCCACGA AGTATTACTCGC ...... ......
<7TRN> ......TCCACC CTTCGTCCTCAGATGAACATCCGGAGGAAGATGATT AGAGTGAGAATA ...... ......
   ..
<EndA> ........CACC CCTCTAATGATTCGTAATCAAACGCGTATCCTGAAA AGAGTGAGAATA ...... ......
<EndA> ............ .......CCACTAGTCCACCAAAAATTCGTAAAACT AGAGTGAGAATA ...... ......

                                                                unprocessed
```

Figure 3.4. Parsing of the full-length reads into mapped regions containing right adapter, left adapter, R36 variable region, NNK and Barcode regions preceding left adapter. Alignment was performed by searching for constant forward (A) or reverse (B) adapters. Tags at the beginning of each line describe the algorithm by which the adapter was identified. <PERF> perfect alignment; <1Mut> one mutation in the adapter; <1Del> one deletion in the adapter; <2TRN> <3TRN>, etc are truncation to 2nd, 3rd, etc nucleotide in the adapter; <EndA> alignment to the adapter at the opposite end of R36 region. The log-scale plots above describe the relative abundance of sequences identified by specific algorithm.

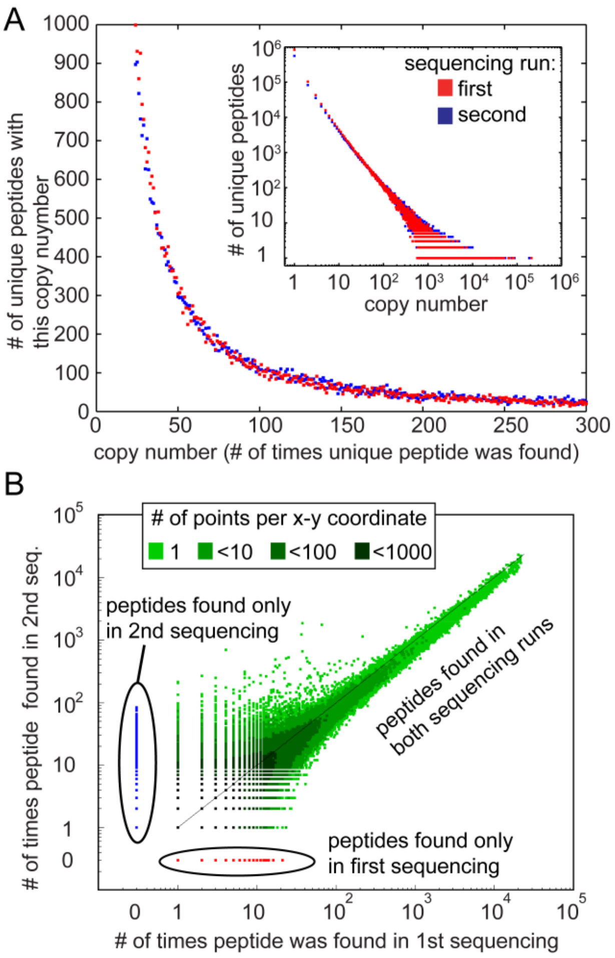


Figure 3.5. (A) Abundance of peptides in the library; each point represents a peptide sequence. Red and blue colors represent two independent sequencing runs where red data correspond to **2S** and blue data correspond to **2L** library (Figure 1D-E) prepared from the same amplified PhD-12 library. The insert describes a log-log plot of the same data. (B) Reproducibility of peptide abundances in two sequencing runs. The abundance of peptides at copy number >100 is highly reproducible between two runs. Peptides found in only run 1 (red dots) or run 2 (blue dots) have low relative abundance. Darker shades of green represent <10, <100 or <1000 data points in the same (x,y) coordinate.

3.2.5 Preliminary analysis of sequence diversity in the library

Complete analysis of sequence diversities obtained using Illumina sequencing is beyond the scope of this manuscript. Here, we present the preliminary analysis of the sequences, and we confirm that the sequencing runs were reproducible. Figure 3.5 describes the distribution of sequence abundances in the library obtained by sequencing of two library preparations (band **2S** and **2L** in Figure 3.1D). The abundance of sequences in the two runs were similar (see Figure 3.5): 150 unique peptides were found in copy number of 10^4 and higher; nearly 10^6 peptide sequences were found in low copy number. The abundances of specific peptide sequences were highly reproducible between the two runs (Figure 3.5B). Peptides that were observed 10^2-10^5 times in sequencing run 1, were observed at similar copy number in the second sequencing run. Deviation from 1:1 correlation was observed at copy number <100. Some peptides, observed at copy numbers of 10-100 in the first run, were present at much lower copy numbers in the second run, or were completely absent from the second sequencing run.

The distribution of sequence abundance was dramatically different from the predicted Poisson distribution with an expected value of 20. It could not be modeled as a Poisson distribution at any expected value. A mere 20 clones constituted 8% of the size of the library and were present at a copy number of >30,000 (Figure 3.6). On the other hand, 500-800 thousand diverse sequences constituted another 8% and were present at copy number of <10.

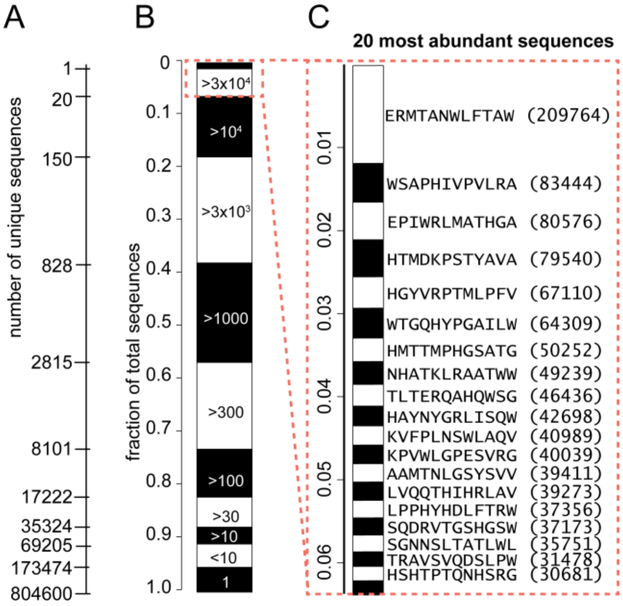


Figure 3.6. Distribution of (A) the number of unique peptide sequences and (B) fraction of total peptide sequences in the library. Black-and-white stacked bar or "zebra-bar" describes the library in this and two subsequent figures (Figure 3.7, 3.8, and 3.9). The height of each segment is proportional to the fraction that each sub-population occupies in the library. For example, ~5% of the library is occupied by 20 sequences, present at abundance of >30,000 copies. 20% of the library is occupied by 150 sequences, present at >10,000 copies, etc. (C) Zoomed-in zebra-bar describes top 20 sequences. The height of each segment is proportional to the fraction of the library occupied by each sequence. For example, top sequence occupies 1.2 % of the library.

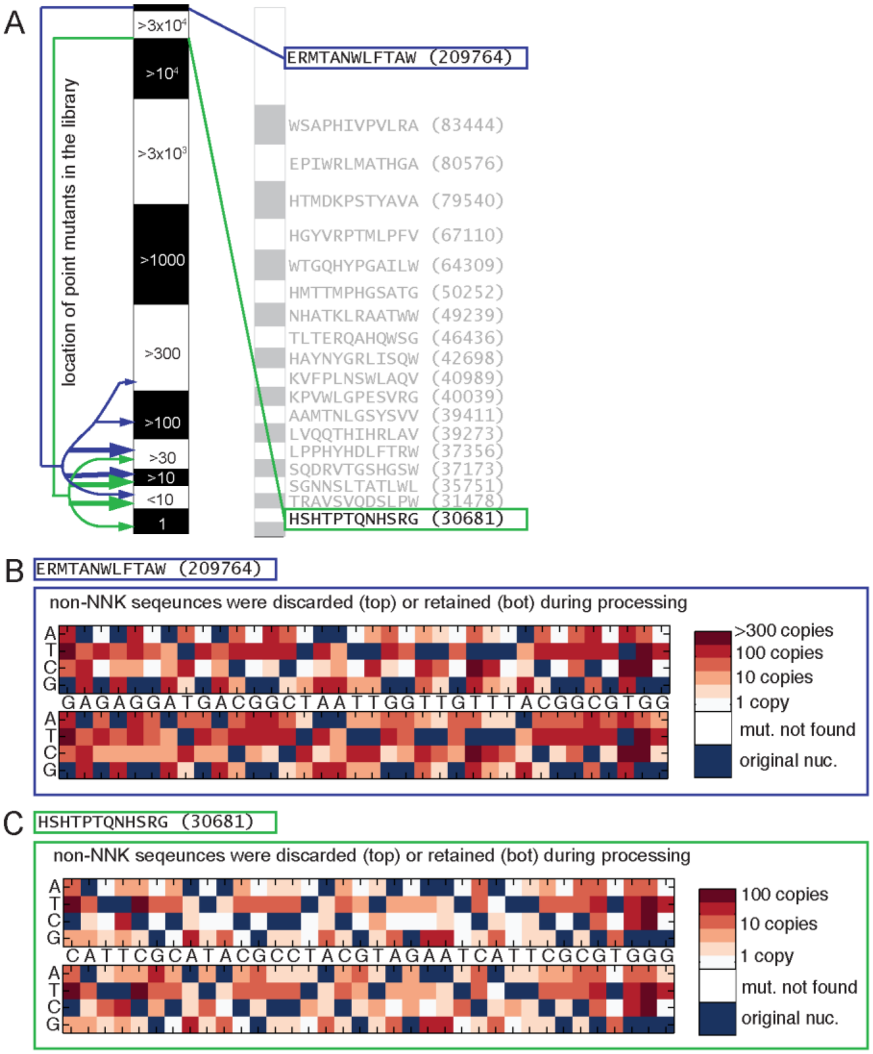


Figure 3.7. Analysis of point mutations in the library. We selected two abundant sequences (the most abundant and 19th most abundant), generated point mutations of these sequences and searched for these point mutations in the library. The approximate locations of these sequences in the library is showed by green and blue arrows. The size of the arrows qualitatively indicate abundances or mutants in each region. (B) and (C) indicate positional abundance for each mutation. For example, the copy number of nucleotide that has G to C substitution in the 1st position is ~30; whereas abundance of G to T mutation in the same position is >300. We find significantly more point mutations than one would expect to have

in a sparse library (total number of clones is 10^6 while potential diversity is 10^{18}). In fact, for the top sequence we find all possible point mutations, including K to M mutations in positions 3, 6, 9 etc. To see these mutations, we re-analyzed the library and included non-NNK sequences in our analysis (see bottom heat plots in B and C). The median abundance of point mutations is ~200 (B) and ~20 in (C), which is ~0.1% of the abundance of their original sequences. We concluded that most point mutations, thus, correspond to sequencing errors. We could thus assume that the region of the library that has abundances of >100 is free of sequencing errors.

The distribution of sequence abundances followed the power-law distribution, producing a linear plot on a log-log scale (Figure 3.5A, insert). We observed a deviation from this distribution for the low copy number peptides. Extrapolation of a log-log plot predicts that the number of single copy-number sequences should be 3-5x10^5. The observed deviation suggested that a significant fraction of the low-copy-number peptides resulted from sequencing errors. Errors are abundant in Illumina sequencing[127], but we anticipate that many of these errors could be easily identified. One possible algorithm could be based on the assumption that the library is sparse. In other words, a library of nucleotides with structure $(NNK)_{12}$ has $(4 \times 4 \times 2)^{12} = 10^{18}$ potential members, and in a pool of 10^6 sequences, the probability of finding a mutant is low.

Despite this prediction, the search for point mutations of most abundant sequences yielded ~100 point-mutants for high-copy-number sequences (Figure

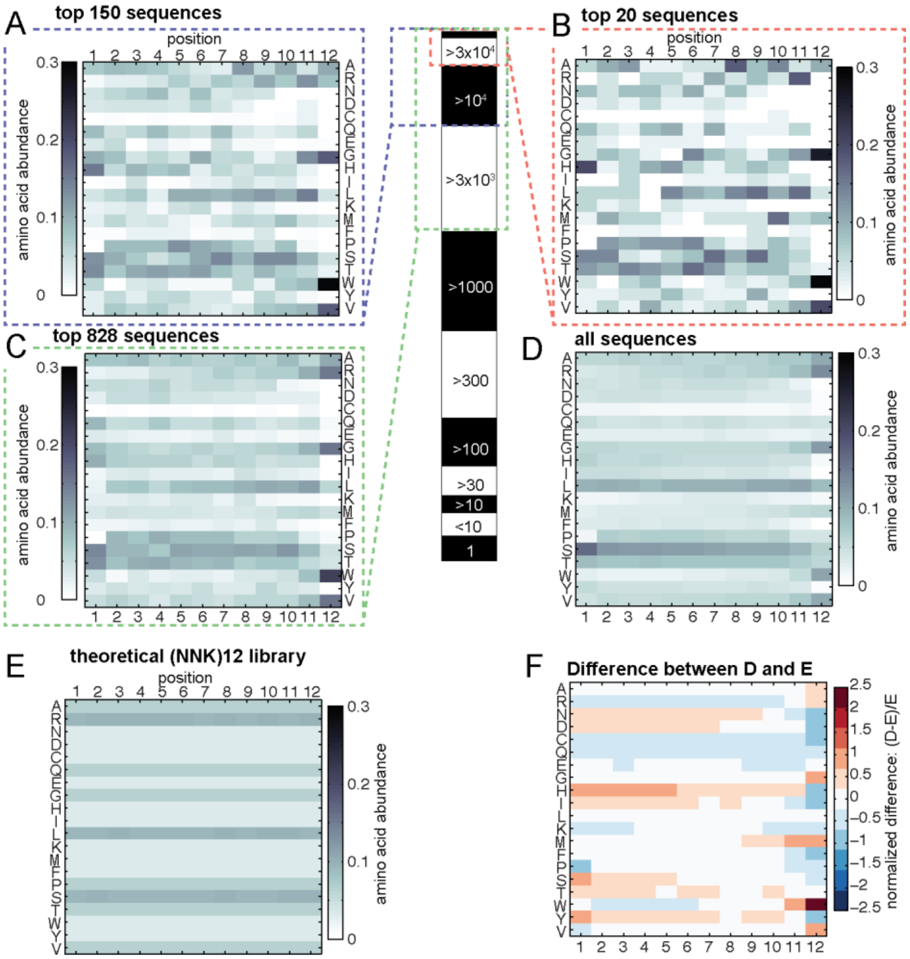


Figure 3.8. (A-D) Positional abundance of amino acids in the top 20 sequences (B) is very different from abundance of amino acids in all peptides in the library (D). Abundance in top 150 sequences (A) and top 20 sequences (B) were similar. On the other hand, the abundance in the top 850 sequences (C) resembled that of the whole library. The sequences present at copy number of >30,000 are different from the rest of the sequences in the library. (D-F) Comparison of the distribution of the amino-acid in the entire library (D) and theoretical distribution of amino-acids in (NNK)12 library (E) reveals differences in positional abundances of

individual amino acids. The plot in (F) describes fold-increase (red) or decrease (blue) in abundance of specific amino acids in specific position.

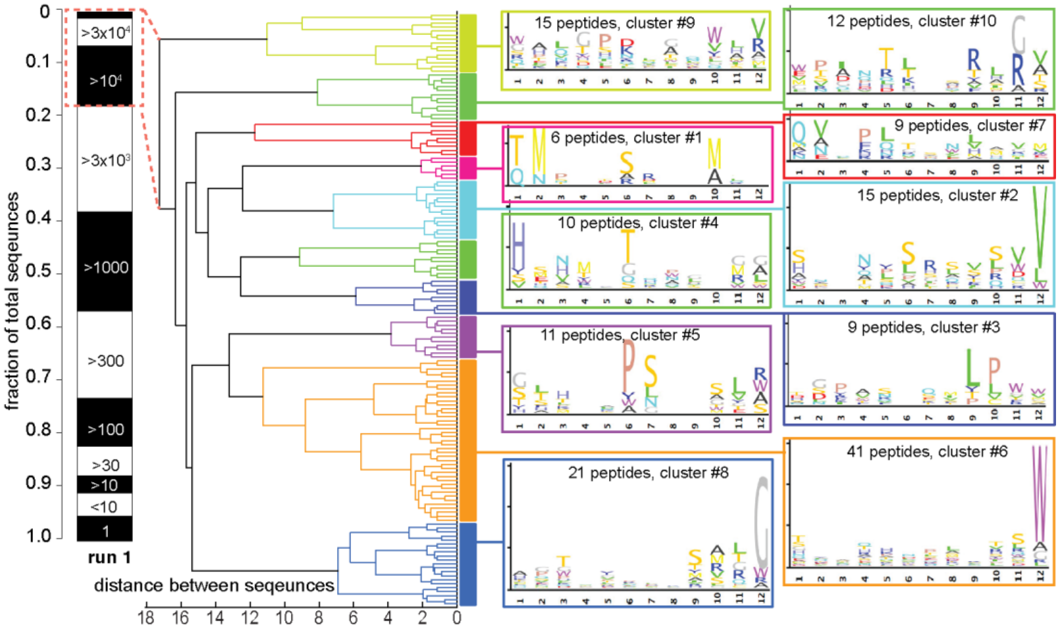


Figure 3.9. Clustering analysis of the top 150 sequences (highlighted as dotted rectangle) based on sequence similarity. We observed 10 distinct clusters, which contained distinct consensus sequences. Calculation of distance and clustering was performed using Euclidian metric in MatLab. Consensus motifs were generated using protein LOGO (pLOGO)[128].

3.7). The majority of these mutated sequences was present at low abundance (Figure 3.7); average abundance was ~1%, which is similar to the frequency of point mutations in adapter sequences (compare <PERF> and <1Mut> in Figure 3.2). This preliminary analysis suggests that sequences with abundance of >100 copies contain no errors. Those with an abundance of <100 and differ at one or two positions could be corrected to the sequence of the more abundant clone. Validation of the error analysis and repair algorithm, however, is beyond the scope of this manuscript.

Positional analysis of amino acid abundances (Figure 3.8) demonstrated that the distribution of amino acids in the top 150 sequences, present at copy number of >10,000, was different from that of the remaining library. The distribution of amino acids in sequences present at copy number <10,000 was similar to those in the overall library. The overall distribution of amino acids in peptides in the library was similar to those observed in earlier reports[34,121,129]. The library had abundant Ser/Thr in all positions. Abundance of Cys was low in all positions. The N-terminus exhibited significant preference for some amino acids, presumably due to proteolytic preference of the signal peptidase, which cleaves between the leader peptide sequence and the displayed peptide[34,122].

Clustering analysis identified ten distinct sequence patterns in the top 150 fastest growing clones. Figure 3.9 describes the clustering tree diagram and protein LOGO[128] display of the conserved sequence within each sub-sequence. Remarkably, a rare amino acid, Trp, appeared as a consensus amino acid in many sub-sequences, and it was present as the N-terminal amino acid in 50 out of 150 peptides. Our simple clustering analysis could be replaced by that of more advanced software packages, such as MUltiple Specificity Identifier (MUSI)[97], which is designed to identify distinct families of consensus sequence motifs within deep sequencing data. Such an analysis could potentially identify conserved peptide motifs emerging as the results of growth-induced selection.

3.3 Conclusion

Illumina sequencing, for the first time revealed strong amplification bias to a small number of sequences. The scale at which this bias is apparent is difficult to attain by other next-generation sequencing techniques. The reason for this bias remains unknown, but we strongly believe that the bias has resulted from growth preferences of individual phage. It is unlikely to be the result of simple bias in PCR preparation; the latter bias is unlikely to give abundances of 10,000-fold. PCR also does not favor specific sequences but rather classes of sequences with a specific melting point and/or specific GC-content[93,94]. The bias we observed is unlikely to be present in the original library, which should contain up to 10^9 clones according to the manufacturer (New England Biolabs). Indeed, sequencing of original (non-amplified) libraries demonstrated that there is little bias to specific sequences in the library[66].

Deep sequencing of phage libraries also leaves a few open questions. One of them is general error analysis of random libraries. A growing body of literature has confirmed that a large number of errors is present in Illumina results [127], but reliable identification of errors in random libraries is not trivial. The other unexplained observation is the dramatic abundance of reverse reads when compared to forward reads (Figure 3.5). The preparation based on dsDNA should give equal numbers of forward and reverse strands; the reason for the observed bias towards reverse strands is unclear. It is unlikely that the reads were lost in the analysis because our analysis maps accounted for mutations and frame shifts of constant primer regions and, thus, could map up to >99% of reads. We hypothesize that hybridization to the Illumina chip and on-chip sequencing might be biased to one read (or one type of DNA sequence). On-chip sequencing is known to discriminate against specific classes of sequences and introduce specific errors (frame shifts, *etc.*) [127]. The analysis of sequence bias in different reads and comprehensive error analysis will be described in Chapter 4. Overall, we foresee that Illumina sequencing and analysis similar to the one outlined in this manuscript will provide many advantages to the analysis of phage-library screens. Furthermore, analysis of the biological origin of sequences emerging from amplified libraries will enable identification of a mechanism that promotes or interferes with selection of useful binding sequences in phage display.

3.4 Materials and methods

3.4.1 Experimental Design and choice of library:

In this report, we sequence a commercially-available library of random 12-mers from New England Biolabs (Ph.D-12). As of 2012, this library has been used in ~800 publications (source of estimate: PLoMics database and MimoDB database[123,124]). According to the manufacturer (NEB), the original library contains up to 10^9 different clones. Since this number is beyond the sequencing capabilities of Illumina, we worked with a 1/1000th portion of the library containing 10^6 different sequences. If the sequencing run produces 20 million sequences, the observed frequency of sequences could be approximated by a Poisson distribution with an expectation value of 20. For the above uniform library of 10^6 clones, the distribution predicts that every sequence will be observed at least five times. Over 99% of the library should be observed within 3 standard deviation of the expectation value (sqrt(20)x3=13). The majority of the clones, thus, should be present at 7 to 33 copies.

To explore the effect of amplification on library diversity, we amplified a pool of 10^6 clones to 10^{12} PFU and isolated ssDNA from the combined pool of phage. Specifically, 10^6 PFU from the original library were mixed with 10^7 CFU of *E. coli* in 1 mL of LB. The mixture was shaken at 200 rpm for 5 h at 37 °C. Amplification yielded ~10^{12} PFU. Approximately 10^6 copies of each clone should be present in this pool. If relative abundances of clones were not changed during amplification, abundances of clones observed after deep sequencing should follow the Poisson distribution described above. In reality, we observed that a distribution of clones was dramatically different from the Poisson distribution, suggesting that growth preference of individual clones led to enrichment of some clones and depletion of others.

3.4.2 Isolation of DNA from phage libraries

DNA was isolated using standard NaI/EtOH precipitation method. The steps below are for 500 μL of solution containing 10^{12}-10^{13} pfu/mL of phage. A phage solution was mixed with PEG/NaCl solution (200 μL) and incubated on ice for two hours. The solution was centrifuged (14,000 rpm, 4 °C, 15 min), the supernatant discarded and the pellet was thoroughly dissolved in NaI solution (63 μL). Ethanol (100%, 156 μL) was added and the solution incubated on ice for two hours to precipitate DNA[125]. The solution was centrifuged (14,000 rpm, 4 °C, 15 min) to yield DNA as white or translucent pellet. The DNA pellet was re-suspended in 70% ethanol (200 µL) to remove residual salt. The EtOH-DNA solution was centrifuged (14,000 rpm, 4 °C, 15 min), the ethanol supernatant discarded and the pellet dried for 15-20 min at room temperature. The DNA sample was further purified using phenol-chloroform extraction. The DNA pellet was dissolved in RNAse free water (400 µL). An equivalent amount of phenol-chloroform (1:1 v/v) was added, shaken thoroughly, and centrifuged (14,000 rpm, r.t., 1 min). The aqueous layer was transferred into a separate 1.5 mL microfuge tube and extracted again with an equivalent volume of phenol-chloroform. Again, the aqueous layer was transferred into a separate 1.5 mL microfuge tube and an equivalent amount of chloroform was added. Finally, the aqueous layer (400 µL) was transferred into another 1.5 mL microfuge tube and sodium acetate solution (3 M, 40 µL), 100% ethanol (800 µL), and glycogen (2 µL) was added. The solution was incubated at -20 °C for two hours to precipitate the DNA. The DNA was centrifuged (14,000 rpm, 4 °C, 15 min) and 70% ethanol (400 µL) was added to remove residual salt. The solution was centrifuged a final time (14,000 rpm, 4 °C, 15 min) and the ethanol supernatant removed. The pellet was air dried and re-suspended in RNAse free water (~20 µL).

3.4.3 Preparation of phage library DNA for Illumina sequencing

The following protocol was our first iteration for preparing phage DNA for Illumina sequencing. This approach was used only in this chapter to identify growth bias in phage libraries.

DNA isolated from the Ph.D.^{TM}-12 Phage Display Peptide Library was subjected to PCR amplification with primers flanking the variable region. A list of optimized reaction conditions for PCR amplification is found in Table 3.1 along with cycling conditions specific for each primer listed in Table 3.2. The PCR product was concentrated using ethanol precipitation. If multiple barcoded primers were used, all PCR products were pooled together. The PCR product was run on a 2% (w/v) agarose gel in TBE buffer. The band corresponding to the expected product was excised (Figure 3.3A), and DNA extracted from the gel using the QIAEX II Gel Extraction Kit. The extracted DNA fragment was purified and concentrated using phenol-chloroform and ethanol precipitation as described above. The resulting dsDNA fragment was blunt-end repaired using Illumina Paired-End DNA Sample Prep Kit protocol, and the repaired fragments were purified using the QIAquick PCR Purification Kit protocol. An 'A' base was added to the 3' end at each fragment using the Klenow fragment (Illumina Kit), and purified using the MinElute PCR Purification Kit protocol. Illumina adapters (Illumina Kit) were liganted to each fragment and purified according to the QIAquick PCR Purification Kit protocol. The samples were loaded and run on a 2% agarose gel in TBE buffer, and bands that correspond to fragments with adapters (Figure 3.1D) were purified using QIAquick Gel Extraction Kit protocol. The fragments with adapters were enriched through PCR amplification using PCR Primer PE 1.0 and 2.0 (Illumina Kit) and purified according to QIAquick PCR Purification Kit protocol. To purify the final product, the samples were loaded and run on a 2% agarose gel in TBE buffer and the corresponding bands (Figure 3.1E) were purified using the QIAquick Gel Extraction Kit protocol.

3.4.4 Sequencing of the library

The concentration of dsDNA with ligated Illumina adapters was estimated using the Qubit Fluorimeter (Invitrogen) or Agilent Bioanalyzer using the manufacturer’s protocol. The sample was diluted to 10 nM and submitted for sequencing to the Harvard FAS sequencing facility (Boston, Massachusetts). Sequencing was performed using the Illumina HiSeq and 50 bp single end reads.

Chapter 4 Prospective identification of parasitic sequences in phage-display screens

4.1 Introduction

In recent years, deep sequencing approaches have been employed to assist the analysis of phage-displayed library selection[85], and in many cases, selections against multi-site targets[66,67,89]. Our group has employed deep sequencing to detect convergence, which occurs in phage library screens without any selection (Figure 1.1B). We amplified 10^6 sequences from a naïve library in bacteria, and observed that amplification alone enriched a few hundred motifs by 10-100 fold and depressed the remaining 10^6 motifs[64]. This experiment, for the first time quantified the collapse of the library during growth in bacteria. As this collapse is observed in the absence of targets, it is independent of (or orthogonal to) the collapse induced by target-binding selection[10]. A typical phage library screen procedure involves multiple rounds of panning and amplification in bacteria is thus driven by two orthogonal selection pressures (Figure 1.1C). There are two fundamental predictions from Figure 1C: (i) selection could identify only a small number of available binding clones (dots in Figure 4.1C); (ii) most of the selections should co-cluster with fast growing clones, which from here on are referred to as "parasitic clones". Figure 1.1C is a theoretical prediction[10], which we confirm in this chapter.

Despite abundant evidence of growth-induced bias, it is often viewed as an experimental inconvenience that could be overcome by improvements in the target-binding procedure (*e.g.* more washing steps). In this paper, we show that growth-induced bias is ubiquitous in phage library screens during the amplification step using the Ph.D.-7, Ph.D.-12, and Ph.D.-C7C libraries as examples. Parasitic or fast growing-clones are abundant in the original libraries. These clones dominate the screens and they cannot be eliminated by any improvements in the target-binding procedure. There are only two strategies to avoid growth bias: (i) avoid amplification; (ii) use amplification that enriches all phage clones uniformly[99,112]. In this report, we confirm that the latter strategy can remove sequence bias and avoid enrichment of parasitic clones.

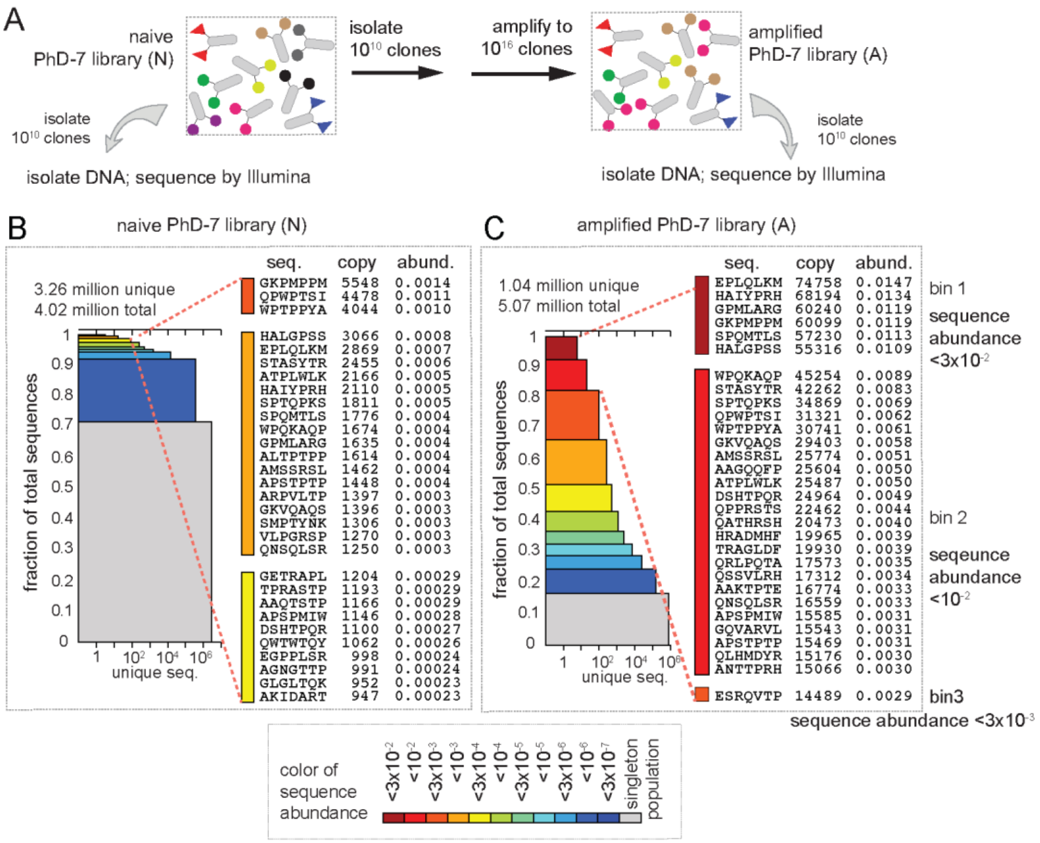
4.2 Results and Discussion

4.2.1 Identification of parasitic sequences using deep-sequencing of naïve and amplified PhD-7 library

Our report focuses on the library of 7-mer peptides (Ph.D.-7^{TM}) because the reported diversity of the library (10^9) approaches the theoretical diversity of peptide X_7-motifs ($1.3x10^9$) and it covers most amino-acid diversity. To assess the diversity of the naïve library, we isolated DNA from 10^{10} PFU from the original Ph.D.-7 library (Figure 4.1A); this number should yield, on average, 10 copies of each available sequence, if the library was uniform. Sequencing of DNA by Illumina yielded $4x10^6$ reads (Figure 4.1B). Although sequence coverage was not complete, it was sufficient for our analysis here. If the original library contains

10^9 sequences in equal abundances, the expected value of abundance of each sequence in a sub-sample of $4\text{x}10^6$ reads is $4\text{x}10^6/10^9$ =0.004 or 0.4% of the total. For this expected value, the Poisson probabilities to find a sequence with copy number 1, 2, 3, or 4 is 0.996, 0.002, $3\text{x}10^{-5}$, $3\text{x}10^{-9}$ respectively. Over 99% of the population, thus, should have a single copy number ("singleton population"). In $4\text{x}10^6$ reads, we expect at most one sequence with copy number strictly above three. In reality, we found that only 72% of the library comprised a singleton population (grey segment, Figure 4.1B), 20% comprised sequences with copy numbers of 2 or 3 (blue segment, Figure 4.1B) and 8% of the library had copy numbers of >3. Some sequences (26) had a copy number higher than 1000 (Figure 4.1B, list of top 30 sequences).

We hypothesized that library members present at higher than theoretical abundance are the rapid-growing clones. Their number, thus, must increase if the library is re-amplified in bacteria. To validate this hypothesis, we amplified 10^9 PFU from the original library in bacteria to yield 10^{15} PFU (expected amplification by a factor of 10^6 for each clone) under amplification Condition 2 (Methods). Isolation of DNA from the amplified population and Illumina sequencing yielded ~$5\text{x}10^6$ reads. We observed that sequences that had high copy number in the original library **N** (e.g., GKPMPPM: copy number 5548, abundance 0.0014, or 0.14%, Figure 4.1B) were further enriched in the amplified library **A** (GKPMPPM: copy number 60099, abundance 0.012, Figure 4.1C). Copy numbers of some sequences in the amplified libraries reached > 50,000; this



seq.	copy	abund.
GKPMPPM	5548	0.0014
QPWPTSI	4478	0.0011
WPTPPYA	4044	0.0010
HALGPSS	3066	0.0008
EPLQLKM	2869	0.0007
STASYTR	2455	0.0006
ATPLWLK	2166	0.0005
HAIYPRH	2110	0.0005
SPTQPKS	1811	0.0005
SPQMTLS	1776	0.0004
WPQKAQP	1674	0.0004
GPMLARG	1635	0.0004
ALTPTPP	1614	0.0004
AMSSRSL	1462	0.0004
APSTPTP	1448	0.0004
ARPVLTP	1397	0.0003
GKVQAQS	1396	0.0003
SMPTYNK	1306	0.0003
VLPGRSP	1270	0.0003
QNSQLSR	1250	0.0003
GETRAPL	1204	0.00029
TPRASTP	1193	0.00029
AAQTSTP	1166	0.00029
APSPMIW	1146	0.00028
DSHTPQR	1100	0.00027
QWTWTQY	1062	0.00026
EGPPLSR	998	0.00024
AGNGTTP	991	0.00024
GLGLTQK	952	0.00023
AKIDART	947	0.00023

seq.	copy	abund.
EPLQLKM	74758	0.0147
HAIYPRH	68194	0.0134
GPMLARG	60240	0.0119
GKPMPPM	60099	0.0119
SPQMTLS	57230	0.0113
HALGPSS	55316	0.0109
WPQKAQP	45254	0.0089
STASYTR	42262	0.0083
SPTQPKS	34869	0.0069
QPWPTSI	31321	0.0062
WPTPPYA	30741	0.0061
GKVQAQS	29403	0.0058
AMSSRSL	25774	0.0051
AAGQQFP	25604	0.0050
ATPLWLK	25487	0.0050
DSHTPQR	24964	0.0049
QPPRSTS	22462	0.0044
QATHRSH	20473	0.0040
HRADMHF	19965	0.0039
TRAGLDF	19930	0.0039
QRLPQTA	17573	0.0035
QSSVLRH	17312	0.0034
AAKTPTE	16774	0.0033
QNSQLSR	16559	0.0033
APSPMIW	15585	0.0031
GQVARVL	15543	0.0031
APSTPTP	15469	0.0031
QLHMDYR	15176	0.0030
ANTTPRH	15066	0.0030
ESRQVTP	14489	0.0029

Figure 4.1. (A) We selected 10^6 unique clones from Ph.D.-7 library, amplified in bacteria, isolated the phage genome, amplified the library portion by PCR, and obtained 4-5 million sequences using Illumina HiSeq. (B-C) To visualize all sequences, we generate a sacked-bar in which each segment contains all sequences with specific copy number (color-coded); the width of each segment is equal to the number of unique sequences per segment. Prior to amplification (B) the majority of the clones in naïve library have low copy number. After amplification (C), ~8% of the library is occupied by 6 sequences (crimson segment), ~20% of the library is occupied by 35 sequences (red + crimson segments), etc.

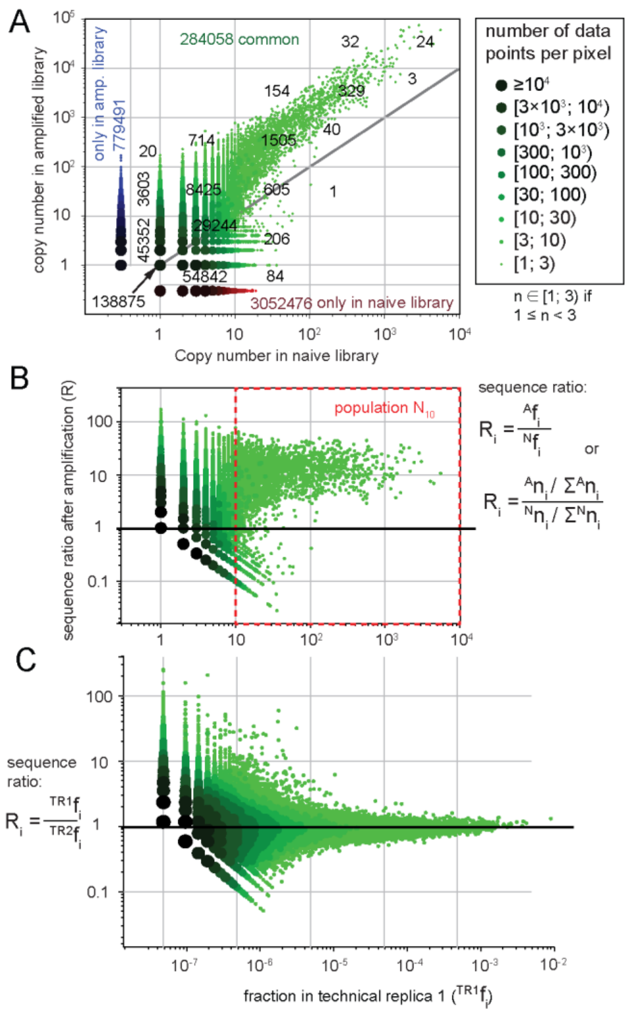


Figure 4.2. (A) Scatter plot describing naïve (**N**) and amplified (**A**) Ph.D.-7 library (condition 2, see Methods section). Each dot is a unique sequence; multiple data at the same (x,y) coordinate are bigger, darker dots (see legend). Numbers represent the number of data points within each cell of the rectangular grid. Green data is observed both in **N** and **A**, while blue and red data is unique to **N** or **A**. (B) Ratio plot compares normalized ratio of each sequence between naïve and amplified library and copy number in naïve library. Copy number of many sequences present in the naïve library at copy number n_{naive} >10 (red box, N_{10}) increased during re-amplification. (C) is the ratio plot similar to (B) comparing

the same phage library sequences by Illumina twice (data from reference[64]). Distribution of the ratios of two technical replicas TR1 and TR2 is symmetric around 1.

number, when normalized to the total number of reads ($5x10^6$) corresponds to 1% of the abundance in the library (sequences in the crimson segment of Figure 4.1C). Comparing **N** and **A** multisets by scatter plot (Figure 4.2A) and ratio plot (Figure 4.2B) traced the fate of all parasitic sequences during amplification. It suggested that most sequences with a copy number >10 in the original libraries have been further enriched during re-amplification (Figure 4.2A-B). Previously, we showed that Illumina sequencing of the same amplified population of phage yielded reproducible copy numbers[64]. Figure 4.2C shows the ratio plot of re-sequencing data and suggests that the increase in copy numbers in amplifications is not the result of sequencing bias. We sought to validate that the observed data are not the result of the biological variability in amplification or technical variability in sample preparation for deep sequencing.

4.2.2 Variability of sequence abundances during phage amplification

Copy numbers in deep sequencing only approximate the true sequence abundance. Variability in copy numbers in re-sequencing of the same DNA samples could be modeled by Poisson distribution[130]; variability in sequencing of closely related biological samples follows a Poisson distribution with Gaussian noise[131]. Variability of the amplification process in phage libraries, however, has

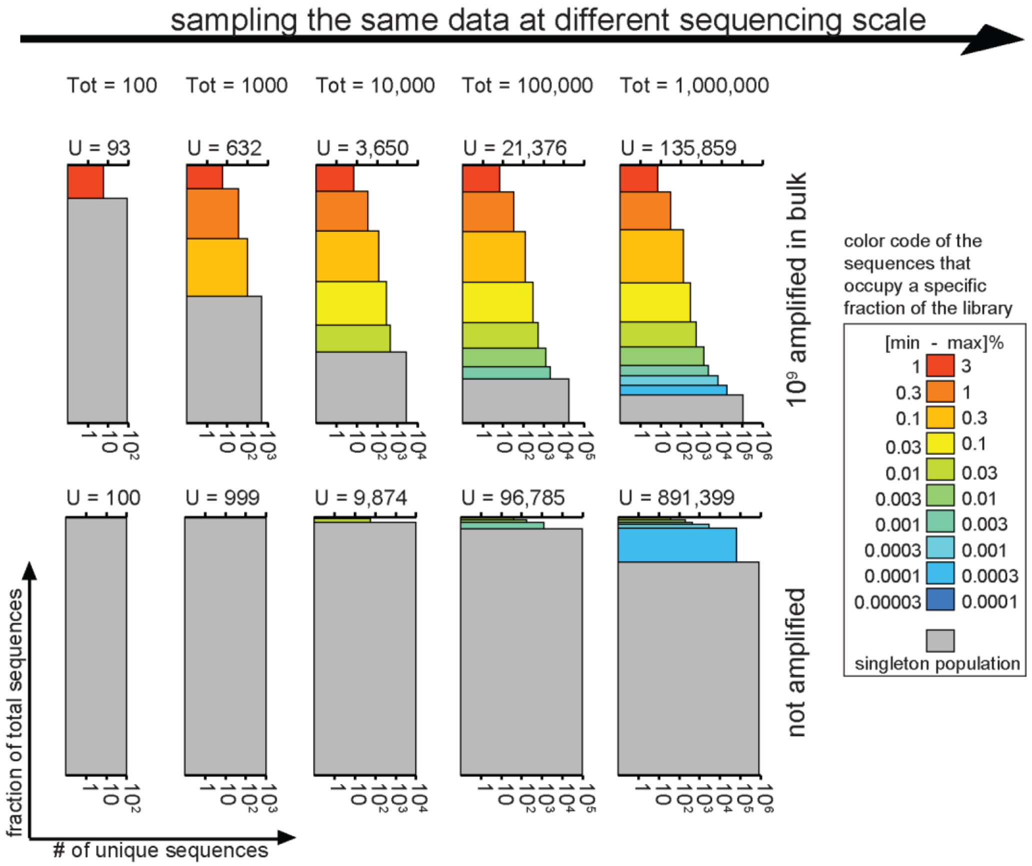


Figure 4.3. This figure displays how the same sequencing data looks at lower sequencing resolution (tot stands for "total number of reads). The figure is generated by random sampling of Illumina data. Original ("not amplified") library is practically "invisible" to sequencing below 10,000 total reads (<100 reads have copy number 2-3, the rest are singleton reads). Amplified library could be reliably analyzed with depth of sequencing of 10,000 to 100,000 reads. In 1000 total reads, one could see a few hundred "parasites" with copy number 2-3 (orange). There are only 10 sequences with copy number above 3-10 (orange-red segments). Observations, thus, do not have high confidence. Sequencing the amplified library at 100-reads scale (typical Sanger sequencing scale) could uncover only a few "parasites" with unreliable confidence (copy number of 2).

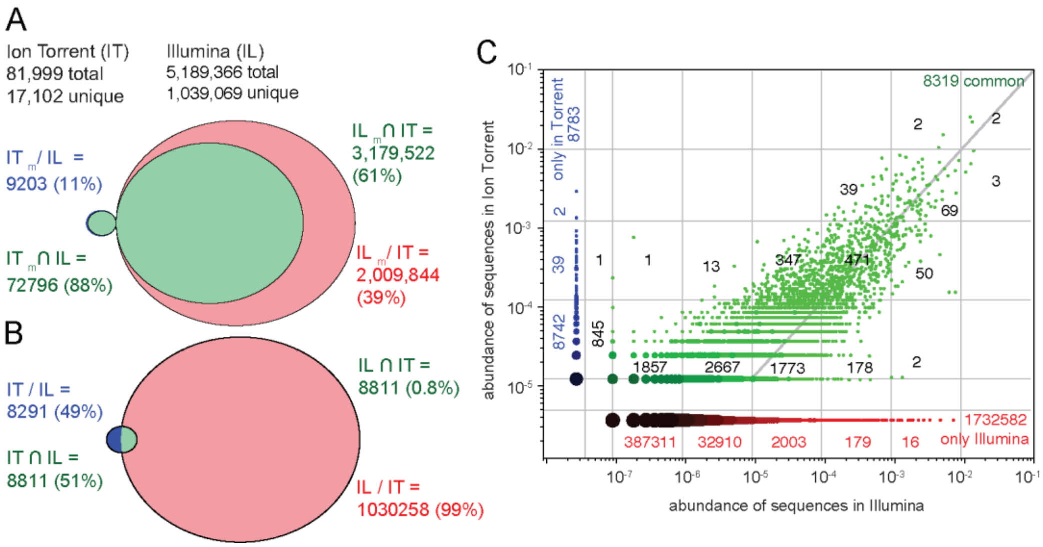


Figure 4.4. (A) Venn diagram describing multiset-specific m-intersect and m-difference between Ph.D.-7 amplified library sequenced by Ion Torrent (**IT**) and Illumina (**IL**). Over 60% of sequences found in the Illumina multiset are also present in the Ion Torrent multiset. The m-intersect for **IL** $_{\mathbf{m}}\cap$ **IT** contains all elements within **IL** that are also found in **IT**. Similarly **IT** $_{\mathbf{m}}\cap$ **IL** contains 88% of sequences. (B) The set **IT** $\cap$ **IL** and **IT** $\cap$ **IL** as it would be constructed without the consideration of frequency of each sequence found in both **IT** and **IL** sets. (A) and (B) are drawn to scale in relation to the size of each set. (C) Scatter plot describing Ph.D.-7 amplified library sequenced by Ion Torrent (**IT**) and Illumina (**IL**). Each dot is a unique sequence; multiple data at the same (x,y) coordinate are bigger, darker dots (see legend). Green data describe m-intersect, while blue and red describe m-difference population (data unique to **IL** or **IT**). Parasitic sequences are found in both sequencing methods. The **IL** and **IT** sets are in good agreement as sequences concentrate around the identity line (1:1 line drawn

through the center of the scatter plot). Numbers represent the number of data points within each cell of the rectangular grid.

never been characterized. To this end, we analyzed multiple biological replicas of phage amplification using the lower cost (and lower throughput) Ion Torrent instrument the sequencing. We estimated how naïve and amplified libraries would look at lower sequencing resolutions (Figure 4.3). The analysis suggested that the high-copy-number sequences in amplified libraries should be readily identified from amplified libraries by Ion Torrent. Indeed, most of the high-copy-number sequences visible in amplified libraries by Illumina (**IL**) were also identified by Ion Torrent (**IT**) sequencing (Figure 4.4). Figure 4.5A describes the sampling process: five biological replicas (BR) originated from five independent samples of the library, 10^8 PFU each. Every population of phage was amplified by a factor of 10^6 in bacteria and sequenced independently. Additionally, we generated five technical replicas (TR) by isolating the DNA from the same amplified library five times and sequencing it separately. We examined how copy number in each replica deviated from average values, and indeed observed higher variance in BR than in TR (Figure 4.5B). We calculated the Pearson's cumulative test statistic from five replicas (Figure 4.5C) and compared it to a chi-square distribution with 4 degrees of freedom[130]. QQ-plot confirmed that the distribution of copy numbers belong to a normal distribution class and the variance of BR and TR is 1.5 and 3-x higher than the variance predicted by Poisson distribution (Figure 4.5C).

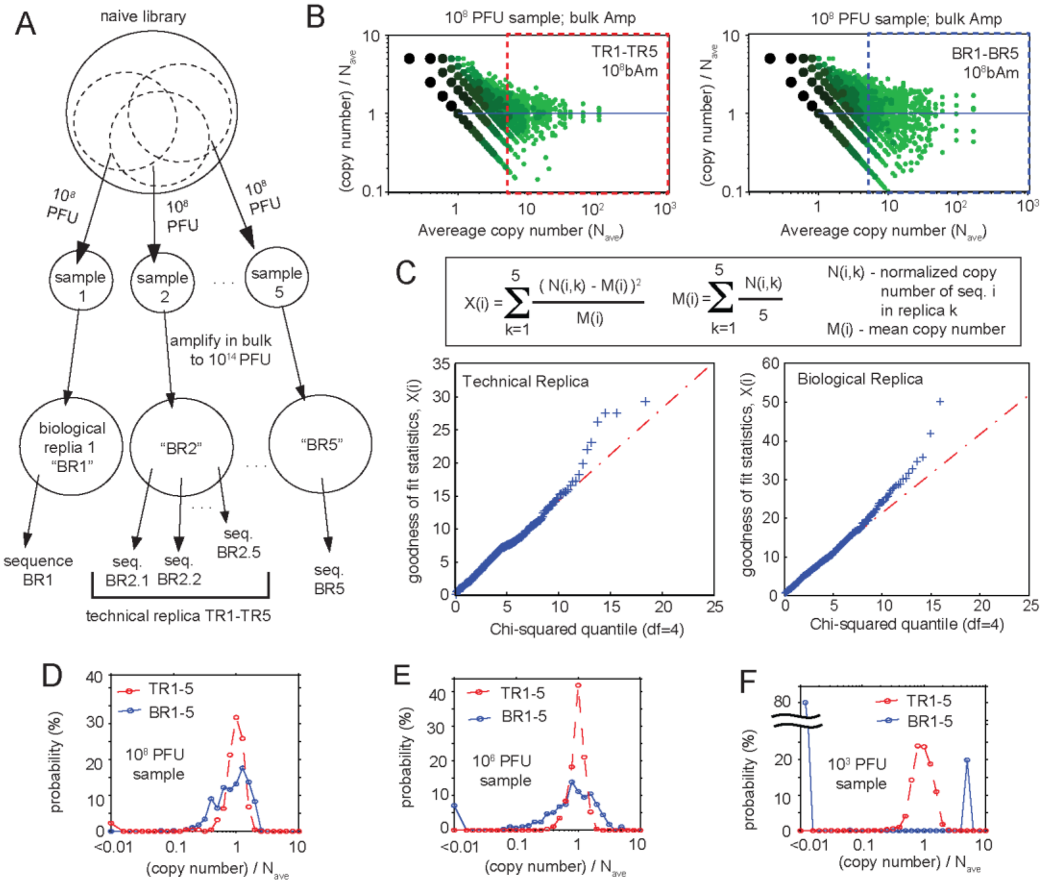


Figure 4.5. (A) Scheme describing generation of biological replicas (BR) and technical replicas (TR). (B) Scatter plot of copy numbers in five replicas normalized by the mean copy number. (C) QQ-plots comparing goodness-of-fit statistics X(i), assuming Poisson distribution [130] and χ^2 distribution with 4 degrees of freedom. The slopes of 1.4 for TR and 2 for BR suggested that both technical and biological replicas are distributed approximately normally but their variance is 1.4-x and 2-x greater than Poisson distribution. A small number of clones deviate from normal distribution in biological replica. Increased variance emanates from the noise during PCR or re-amplification of phage in bacteria. (D-F) Comparison of the distributions of the normalized copy numbers in BR and TR originating from different sample sizes. BR that start from 10^6 PFU (E, blue line)

have higher variance than BR that start from 10^8 PFU (D) while BR that start from 10^3 PFU are not reproducible; all technical replicas are reproducible and have similar variance (red line).

Our technical replicas contained three sources of noise: (i) DNA isolation; (ii) PCR amplification and (iii) sequencing. Deviation from Poisson distribution caused by PCR re-amplification and re-sequencing has been observed previously[131]. The biological replicas contained (iv) variability in phage amplification and (v) variability in the composition of the initial sample. The latter increased as the sample size decreased from 10^8 PFU to 10^6 PFU (Figure 4.5D-E). Decreasing the sample to 10^3 PFU made all five replicas completely irreproducible (no common sequences were observed among five BR, Figure 4.5F). In conclusion, when sample size is sufficiently large (here 10^8 PFU), the biological variance is only two fold higher than the technical variance, and observable copy numbers are reproducible and normally distributed. Low-PFU samples are theoretically attractive because they could be sequenced with high coverage by medium-throughput sequencing; but for a library with 10^9 theoretical clones, biological replicas based on 10^3 PFU yield misleading and irreproducible biological replicas.

4.2.3 Statistically-significant definition of the fast-growing (parasite) sequences

Using multiple biological and technical replicas, we established the limits of the variance in ratios of copy numbers in repeated amplification experiments. If deep-sequencing data were filtered to remove copy numbers <10, the 99th percentile of the distribution of ratios was 2.4-3.0 in technical or biological replicas on Illumina and IonTorrent platforms (Figure 4.6A-B). The deep sequencing data acquired by high-throughput hiSeq Illumina, thus, could be analyzed by these two criteria—n(naïve)>10 and n(amp)/n(naïve)>3—to define a population of parasites significantly enriched during the amplification process (Figure 4.6C). As this definition does not use true biological replicas, only extrapolated variance, we call this population P_{1R} (parasites based on one replica).

In lower-throughput methods, such as Ion Torrent, significance based on cutoff in copy numbers is unreliable because very few reads have n(naïve)>10 (Appendix 1.12). For IonTorrent, the significance of increase could be determined from *k* biological replicas (here *k*=5) generated by sampling and amplifying 10^8 PFU and *m* re-sequencing instances of the naïve library (here *m*=8). For the i^{th} sequence, we calculate the fold-increase as $f_i = \langle n_{ik}(amp)\rangle / \langle n_{im}(naïve)\rangle$ where ⟨..⟩ denotes averaging over replicas, and estimating the statistical significance t_i of this increase using one-sided unequal variance Student's t-test. The resulting f_i-t_i plot ("volcano plot") for ~10^5 sequences appears in Figure 4.6D (each dot is a unique sequence). We identified 996 parasites above the 95% confidence interval and termed this population P_{BR} or "parasites based on biological replicas". While

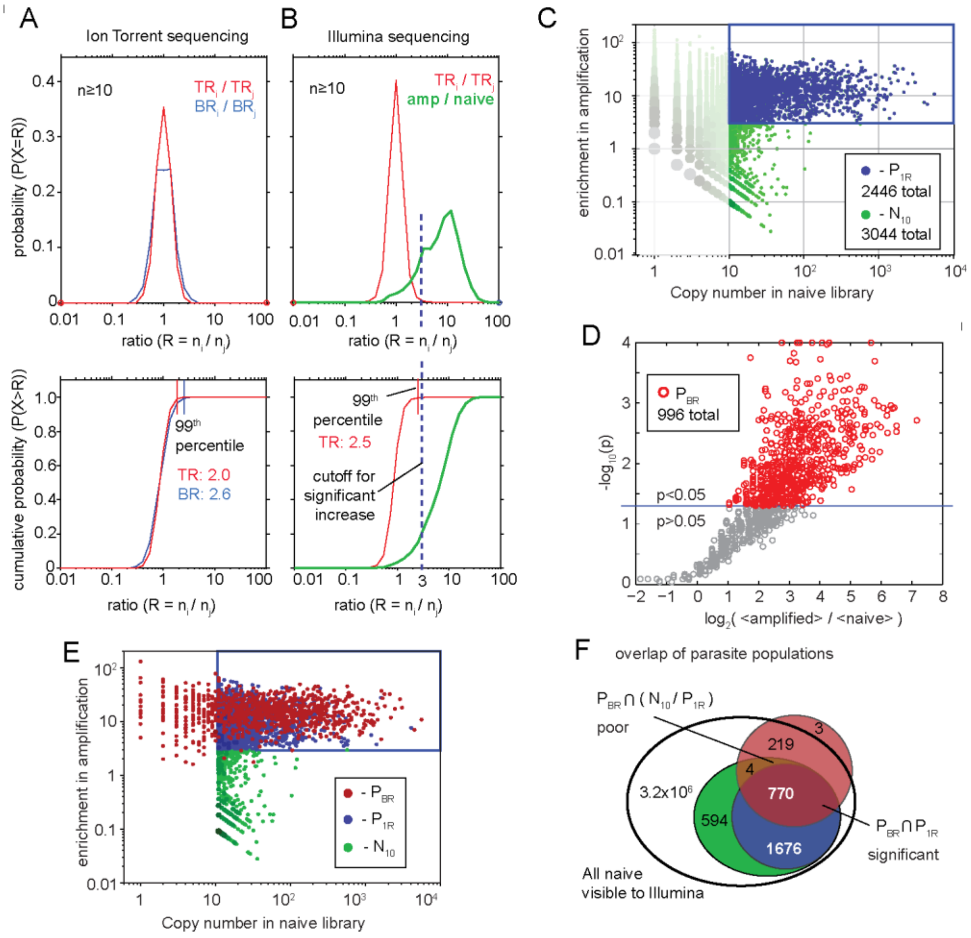


Figure 4.6. (A) Distribution and cumulative distributions of ratios observed between technical replicas (TR) or biological replicas (BR) described in figure 4A-D. Less than 1% of sequences increased by >2.6 fold in BR. (B) Distribution of ratios in technical replicas or amplified and naïve libraries from Figure 2C. Both A and B used reads with copy number >10. (C) The 99th percentile of replica in (A-B) suggested the use of 3-fold increase in n(amp)/n(naïve) ratio to define parasite populations, referred to as P_{1R}. (D) More rigorous definition of parasite population, denoted as P_{BR}, used five biological replicas of the amplified population. Volcano plot highlights 996 sequences that increased significantly ($p<0.05$) in amplification. 99% of sequences increase by >3-fold. (E) Mapping the

P_{BR} population onto a parasite population defined by one replica of Illumina Sequencing (P_{1R}). Some sequences identified in P_{1R} have copy number <10 in naïve library, but all of them increase in amplification (as predicted by Illumina). (C) Venn diagram description of the overlap between naïve, P_{10}, P_{1R} and P_{BR} populations.

P_{BR} originates from a different platform and a different type of statistical analysis, 80% of P_{BR} can be found in the P_{1R} population (Figure 4.6E). The remaining 20% of P_{BR} were found in the population with n(naïve)<10, but the majority of these sequences (~99%) exhibited an increase in copy number by Illumina sequencing (n(amp)/n(naïve) > 3, Figure 4.6E). Identification of a similar parasitic population from two separate sequencing platforms and two types of analysis confirmed that the increase in ratio of copy numbers is neither the result of sequencing artifacts nor of biological noise.

4.2.4 Parasitic sequences in the literature

The hypothesis formulated in Figure 1.1 predicts that selection could identify only a small number of binding clones and that fast-growing sequences should be commonly identified during panning against any target. To test this hypothesis, we used MimoDB to extract sequences found in most peer-reviewed literature reports that used the Ph.D.-7 library (**Lit**) to date[123]. Four observations are important: (i) 382 out of 2000 **Lit** peptides could be identified in the entire Naïve library (Figure 6A). (ii) The "hit rate"—that is, the probability of finding peptides in the naïve library—increased as we focused on sub-populations with higher copy numbers (Figure 4.7B). The “hit rate” changed from 0.01% in the entire **N** to 4.3% in P_{10}, in a sub-population of ~3000 peptide sequences with a copy number n>10. (iii) From 129 literature hits in the P_{10} population, 127 resided in a parasite population P_{1R} identified from one round of Illumina sequencing (hit rate: 5.3%). (iv) Parasites defined by IonTorrent and biological replicas P_{BR} contained 95 results from the literature (hit rate: 9.5%). (iv) From 770 sequences in $P_{1R} \cap P_{BR}$ population, which contained parasites found by both sequencing platforms, 85 were found in the literature (hit rate: 11%).

Statistical significance of the observations above can be validated using bootstrapping simulations and a series of null hypotheses (H_0). To test this observation (i) the null hypothesis was: “the intersection of literature population (**Lit**) and any random library of 3.2 million peptides ($\mathbf{Rnd}^{3200000}$) yields 382 common peptides” (H_0 : ($\mathbf{Lit} \cap \mathbf{Rnd}^{3200000}$)=382). To test it, we generated random uniform libraries of 3.2×10^6 $(NNK)_7$ encoded peptides *in silico* and calculated $\mathbf{Lit} \cap \mathbf{Rnd}^{3200000}$. The average value of intersection between $\mathbf{Lit} \cap \mathbf{Rnd}^{3200000}$ followed Poisson statistics with an expectation value of 15 (Appendix 1.12A). The probability (p) to observe ≥382 common sequences was $p << e^{-382}$. This result suggested that the much larger observed overlap between $\mathbf{Lit} \cap \mathbf{N}$ is not due to chance, but may instead be the result of diversity collapse via similar forces. Testing a general hypothesis for sample size *m* assessed the expected overlap between the literature and any sample $\mathbf{Lit} \cap \mathbf{Rnd}^{m}$ (Figure S4E). For example, $\mathbf{Rnd}^{770}$ had the same size as the “focused parasite population” ($P_{1R} \cap$

P_{BR}, Figure 4.6F). The probability to find a population of 770 random peptides that contained even _one_ literature hit was 0.4% (one in 250 populations contained one literature "hit", the rest contained none). It was highly improbable ($p<<e^{-85}$) to "guess" a population of 770 peptides that contained 85 sequences from the literature. Observations (ii) through (iv) could also be tested as another hypothesis: "parasites are a random subpopulation of naïve library". Specifically, for (ii) H_0 : ($\mathbf{Lit} \cap \mathbf{N}^{3000}$)=382, where $\mathbf{N}^{3000}$ are any 3000 peptides from **N**. We generated $\mathbf{N}^{3000}$ libraries by random sampling of the **N** library and observed that $\mathbf{Lit} \cap \mathbf{N}^{3000}$ followed Poisson distribution with an expectation value of 0.4. The probability to observe overlap of 130 peptides was $p<<e^{-130}$ (Appendix 1.12B). It is therefore essentially impossible to "guess the parasite sequences at random" from a sequenced set.

To provide additional "replicas" for the literature search experiment, we selected 770 peptides from the most stringent parasite population ($P_{TR} \cap P_{BR}$), eliminated 85 peptides found in MimoDB and searched for the remaining 685 peptides on the open Web using Google (see Methods). Interestingly, we found 112 matching peptides in various peer-reviewed and non-reviewed publications (Figure 4.7E). Specifically, 33 originated from PubMed-indexed, peer-reviewed publications, 15 were from published theses and the rest were from patent literature. All publications used the PhD-7 library. References to all publications are available in the supporting information. In summary, nearly 197 peptides could be found in a small 770-peptide population ($P_{TR} \cap P_{BR}$). Using the size of the MimoDB database, we estimated that every tenth peptide in the literature is found in a subset of parasite peptides that constitute <10^{-7} of available diversity. (We believe that there is a correlation between the NEB library lot number and the probability to identify a parasite. Unfortunately, it was impossible to map the lot origin of the libraries used in the literature because very few publications report the lot number).

Some parasitic sequences we identified have been already characterized. Norren and co-workers identified that the HAIPYRH sequence is associated with phage that have mutations in the regulatory regions [35]. This sequence has a copy number of >2000 in the naïve library and >68,000 in the amplified library (Figure 4.1B, C). In addition, it appeared in screens against thirteen unrelated targets[123], and has been confirmed as a weak binder for many targets. Other sequences have similar properties: GETRAPL (#21 in Figure 4.1C) was found in four independent screens; six independent screens identified sequence SILPYPY and eleven screens identified LPLTPLP (found in reference #41)[123,132]. Sequences, such as EPLQLKM (#1 in Figure 4.1C) has been identified in over six screens[133-135], annotated in databases and flagged as "suspicious". Other sequences, such as sequence #8 STASYTR has not been annotated in any databases yet, but it has been found in two published screens[136,137] and our own unpublished results see chapter 5. The parasite population has no common sequence motif. Aside from the small bias to Pro and Ser/Thr amino acids, we could not detect any sequence similarity in "parasites". The sequences did not correlate with motifs that occur due to non-specific binding to polystyrene[30]. The designation "parasite" is different from "non-specific binder". In many publications the binding properties

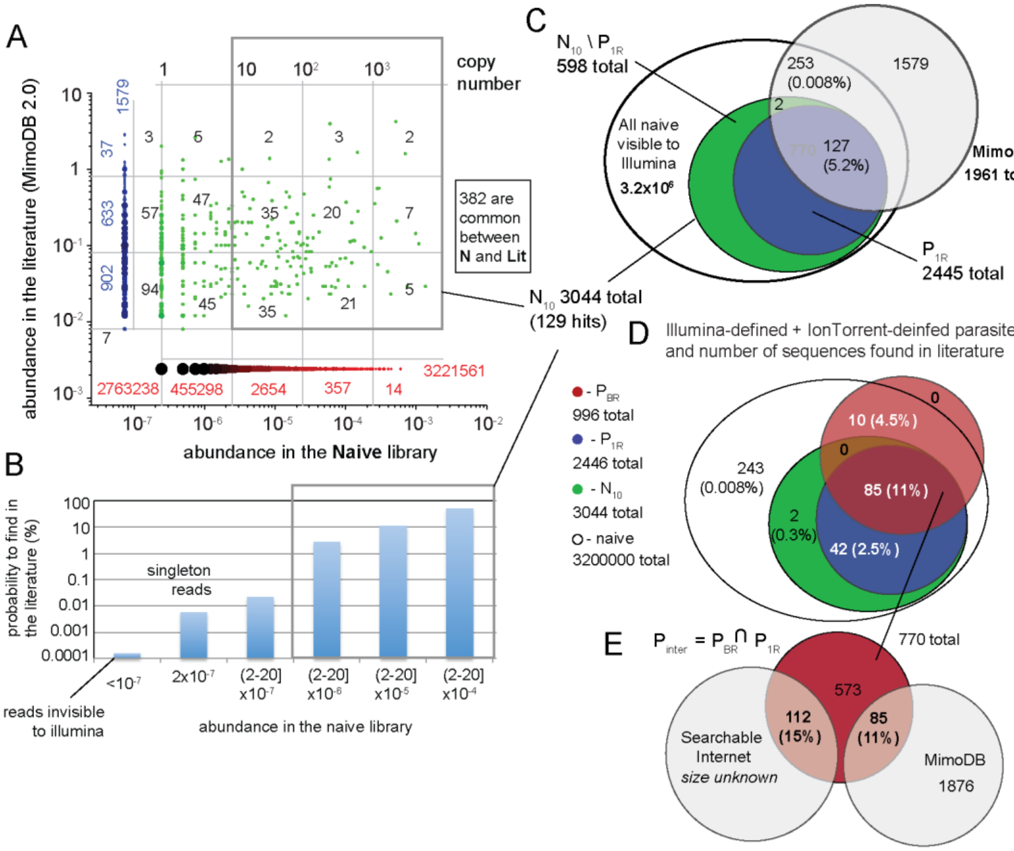


Figure 4.7. (A) Scatter plot comparing the abundance of sequences found in the literature (MimoDB database) and the naïve library sequenced by Illumina. Each dot is a unique sequence; multiple data at the same (x,y) coordinate are bigger, darker dots. Numbers represent the number of data points within each cell of the rectangular grid. Green data describes common sequences, while blue and red describe data unique to the MimoDB database or the naïve library. (B) Abundance of a sequence in the naïve library is correlated with the probability of finding this sequence in the literature. Abundance is reported as range: (2-20] means that abundance is >2 and ≤20. The second bar represents singleton reads, hence, abundance is not reported as range; the first bar represents the reads that were not found in the Illumina run. They are calculated as a difference between

all possible 7-mer peptides and observed peptides (X7 \ Illumina)” (C) Overlap between MimoDB and two putative parasite populations defined by Illumina. P_{1R} population (see Figure 5) has the most significant overlap with literature. The overlap is >1000-fold higher than overlap between MimoDB and 3000 random sequences, see Appendix 1.12). (D) Overlap between MimoDB and parasite populations defined by Illumina (P_{1R} and P_{10}) and IonTorrent (P_{BR} from volcano plot, Figure 5). The $P_{BR} \cap P_{1R}$ population (crimson) has the highest overlap with the literature. (F) From 770 peptides in $P_{BR} \cap P_{1R}$ population, we found 85 in MimoDB; we performed an exhaustive Google search using 685 remaining peptides and found additional 112 peptides in the patent literature, published thesis work and peer-reviewed publications not yet included in MimoDB.

of these sequences have been confirmed to be in the micromolar range. These observations confirm that the parasitic sequences are selected because they have both target binding capacity and high amplification rate (in line with our prediction in Figure 1.1).

4.2.5 Bypassing selection of parasite sequences

Enrichment of parasites occurs due to competition between phage clones during amplification in bacteria (Figure 1.1B). If competition between clones could be avoided, emergence of “parasites” could be suppressed. Previously, in Chapter 2, we described a technology that allows performing uniform amplifications in emulsions. We demonstrated that emulsions can be used to amplify a mixture of fast- and slow-growing phage clones uniformly[99,112]. Here we demonstrate that emulsion amplification can bypass the biased overselection of parasitic sequences from large libraries. We have previously demonstrated that this technique is well-suited for amplification of 10^6 PFU [112]; we also observed that amplification experiments based on samples of 10^6 PFU yields reproducible, albeit noisy biological replicas (Figure 4.5E). We selected 10^6 PFU from the naïve library and amplified them to 10^{12} copies using bulk or emulsion amplification (Figure 4.8A) (for details, see Conditions 1 and 3, in the Methods section). The library after bulk amplification of 10^6 PFU (Figure 4.8B, D) was similar to the library after bulk amplification of the entire 10^9-scale library (Figure 4.1C, 4.2A). It contained the same parasitic sequences and >50% of them had been enriched beyond the variance of biological replicas (>3 fold, Figure 4.8E); small deviations originated from a limited sampling in a 10^6 PFU set. In contrast, the emulsion amplification maintained the abundance of the sequences (Figure 4.8C). The abundance of high-copy-number clones in the phage library amplified in emulsion was suppressed (Figure 4.8D). The abundance of the majority of the parasitic sequences from P_{1R} and P_{TR} populations remained within the variance of the biological replica. Their ratio increased by <3 fold (Figure 4.8F).

We emphasize that the use of emulsion amplification **cannot fix** the skewed diversity already present in the naïve libraries; it **can maintain** this diversity and minimize any further selection of fast growing clones. We have used emulsion amplification in selection to show that such selection allows identification of sequences that cannot be identified by conventional phage display (‘**x**’ in Figure

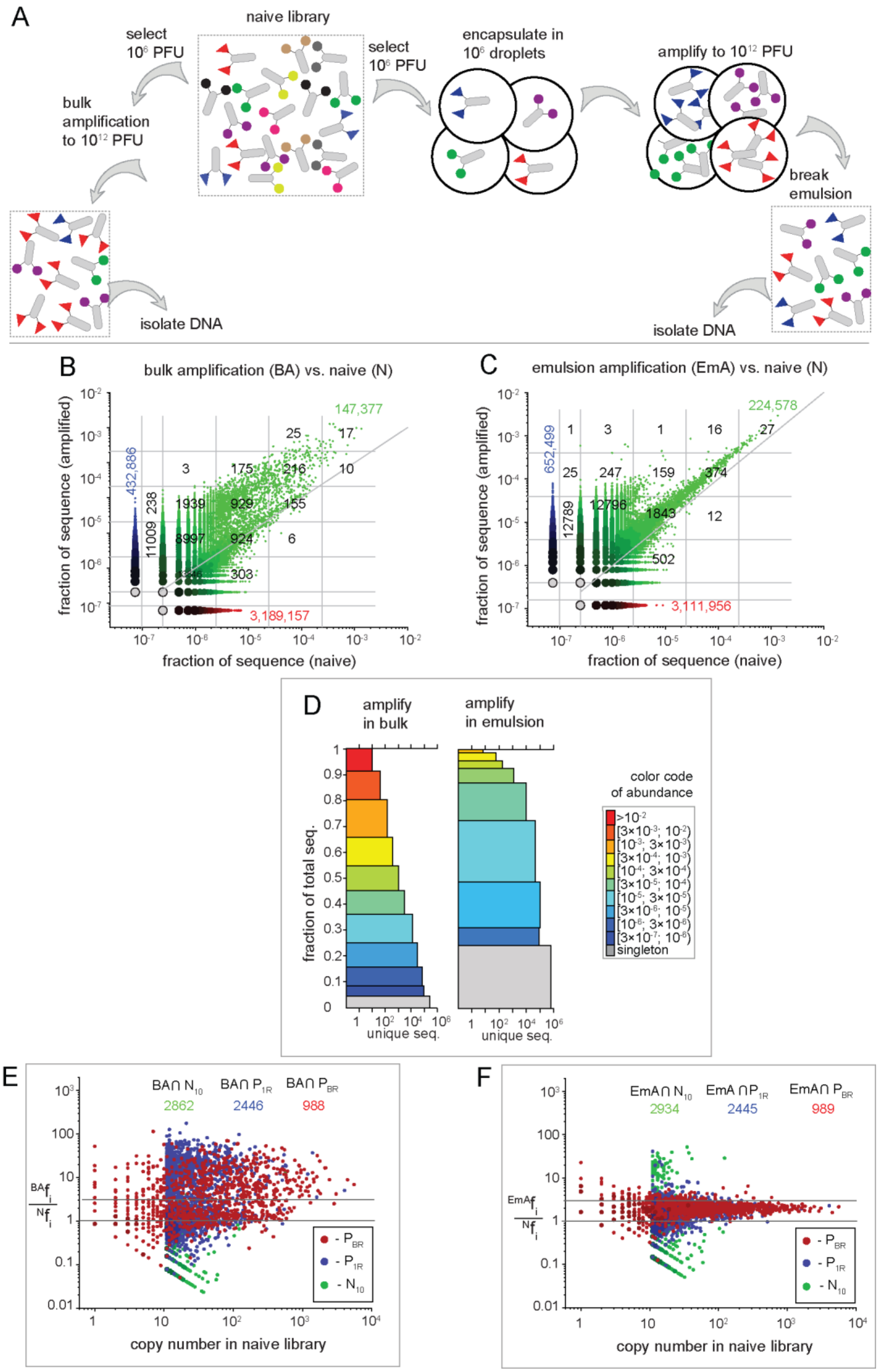

A
naive library
select 10^8 PFU
bulk amplification to 10^{12} PFU
isolate DNA
select 10^6 PFU
encapsulate in 10^6 droplets
amplify to 10^{12} PFU
break emulsion
isolate DNA
B
bulk amplification (BA) vs. naive (N)
fraction of sequence (amplified)
fraction of sequence (naive)
147,377
432,886
3,189,157
25
17
3
175
216
10
238
11009
1939
929
155
8997
924
6
303
C
emulsion amplification (EmA) vs. naive (N)
fraction of sequence (amplified)
fraction of sequence (naive)
224,578
652,499
3,111,956
1
3
1
16
27
25
247
159
374
12789
12796
1843
12
502
D
amplify in bulk
amplify in emulsion
fraction of total seq.
unique seq.
unique seq.
color code of abundance
>10^{-2}
[3×10^{-3}; 10^{-2})
[10^{-3}; 3×10^{-3})
[3×10^{-4}; 10^{-3})
[10^{-4}; 3×10^{-4})
[3×10^{-5}; 10^{-4})
[10^{-5}; 3×10^{-5})
[3×10^{-6}; 10^{-5})
[10^{-6}; 3×10^{-6})
[3×10^{-7}; 10^{-6})
singleton
E
BA∩ N_{10} 2862
BA∩ P_{1R} 2446
BA∩ P_{BR} 988
$\frac{^{BA}f_i}{^{N}f_i}$
copy number in naive library
- P_{BR}
- P_{1R}
- N_{10}
F
EmA∩ N_{10} 2934
EmA ∩P_{1R} 2445
EmA∩ P_{BR} 989
$\frac{^{EmA}f_i}{^{N}f_i}$
copy number in naive library
- P_{BR}
- P_{1R}
- N_{10}

Figure 4.8. (A) Scheme of the amplification of 10^6 PFU taken from Ph.D.-7 naïve library. Amplification was performed either in bulk or emulsion (as described in Conditions 1 and 3 in the Methods section). (B) Bulk amplification or “BA” shows significant enrichment of parasitic sequences when compared to emulsion amplification “EmA” (C). (D) The sequences with high abundance ($f_i > 10^{-4}$, orange-red segments) constitute ~35% of the population after bulk amplification; these highly-abundant sequences are largely constitute <1% of the emulsion-amplified library (E-F). We monitored the fate of parasites (P_{BR} and P_{1R} populations). Both parasite populations are enriched during BA (E). (F) In EmA, the majority of the clones from the parasite populations increased by <3 (within the 99% confidence interval, as defined in Figure 5A).

1.1C). These results, however, extend beyond the scope of this chapter and they will be presented in Chapter 5.

4.2.6 Other libraries

We observed similar results to those described above in other libraries: in Ph.D.-C7C (Figure 4.9A-B) and Ph.D.-12 (Figure 4.10A-B); namely, the diversity in naïve libraries was skewed, and it collapsed upon re-amplification. We used these libraries to demonstrate that emulsion amplification is reproducible. The collapse of diversity in PhD-C7C and PhD-12 libraries was mitigated by emulsion amplification (Figure 4.9C-G and 4.10C-G). We anticipate that the diversity of other phage libraries could be maintained by this method.

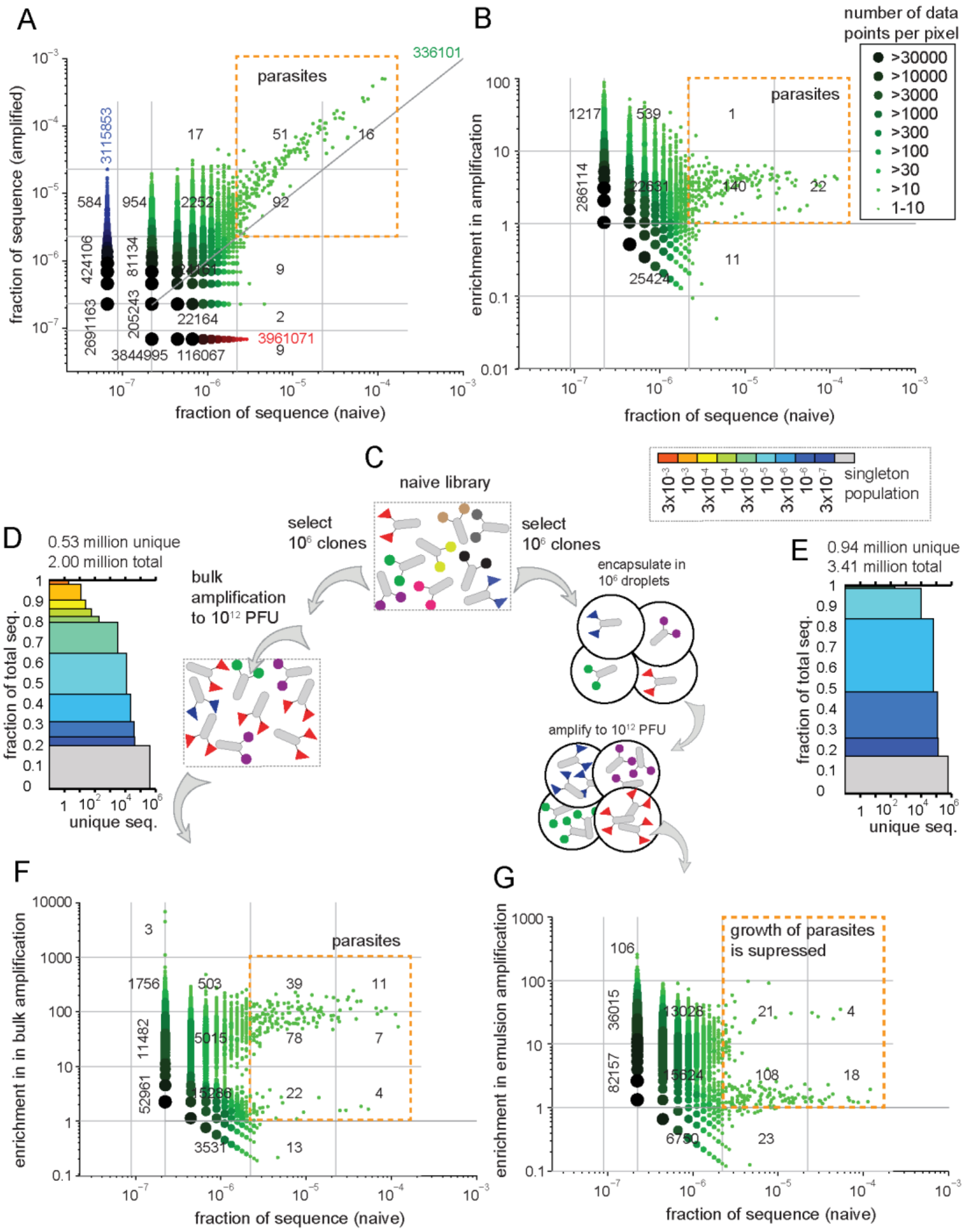


Figure 4.9. (A) Scatter plot describing naïve (**N**) and amplified (**A**) Ph.D.-C7C library. Each dot is a unique sequence; multiple data at the same (x,y) coordinate are bigger, darker dots (see legend). Numbers represent the number of data points within each cell of the rectangular grid. Green data describe m-intersect, while blue and red describe m-difference population (data unique to **N** or **A**). (B) We

calculated sequence enrichment as f_{amp}/f_{naive} and plotted it vs. f_{naive}, where f is a fraction of sequences (copy number normalized by total number of reads). (c) Schematic for the amplification of 10^6 clones taken from Ph.D.-C7C naïve library. Amplification was performed either in bulk or emulsion (as described in Conditions 1 and 3 in the Methods section). (D-F) Amplification in bulk shows significant enrichment of parasitic sequences compared to amplification in emulsion sequences (>100 fold increase respectively from the original fraction in the naïve library). (E-F) Amplification in emulsion yields uniform library without high-copy-number reads. The enrichments of parasitic sequences is suppressed in emulsion amplification.

We propose that it should be possible to map parasitic sequences in other libraries using two simple steps. If diversity of the library is 10^k for some k>1: (i) isolate the DNA from ~10^k clones in the naïve library and sequence them to obtain several replicas of the naïve library (N). (ii) Amplify separate samples of at least 10^{k-1} clones from the naïve library by factor of 10^6 and sequence them to get amplified libraries (A). Then, compare multisets A and N using statistical analysis (*e.g.* similar to volcano plot in Figure 4.6) to identify parasitic populations. We strongly believe that performing prospective identification of parasitic populations will be critical for selecting functional sequences from these libraries. This identification should become a standard protocol/practice for the researchers using these libraries, as well as commercial providers of these libraries. Both high-throughput methods like Illumina HiSeq and lower-throughput technique

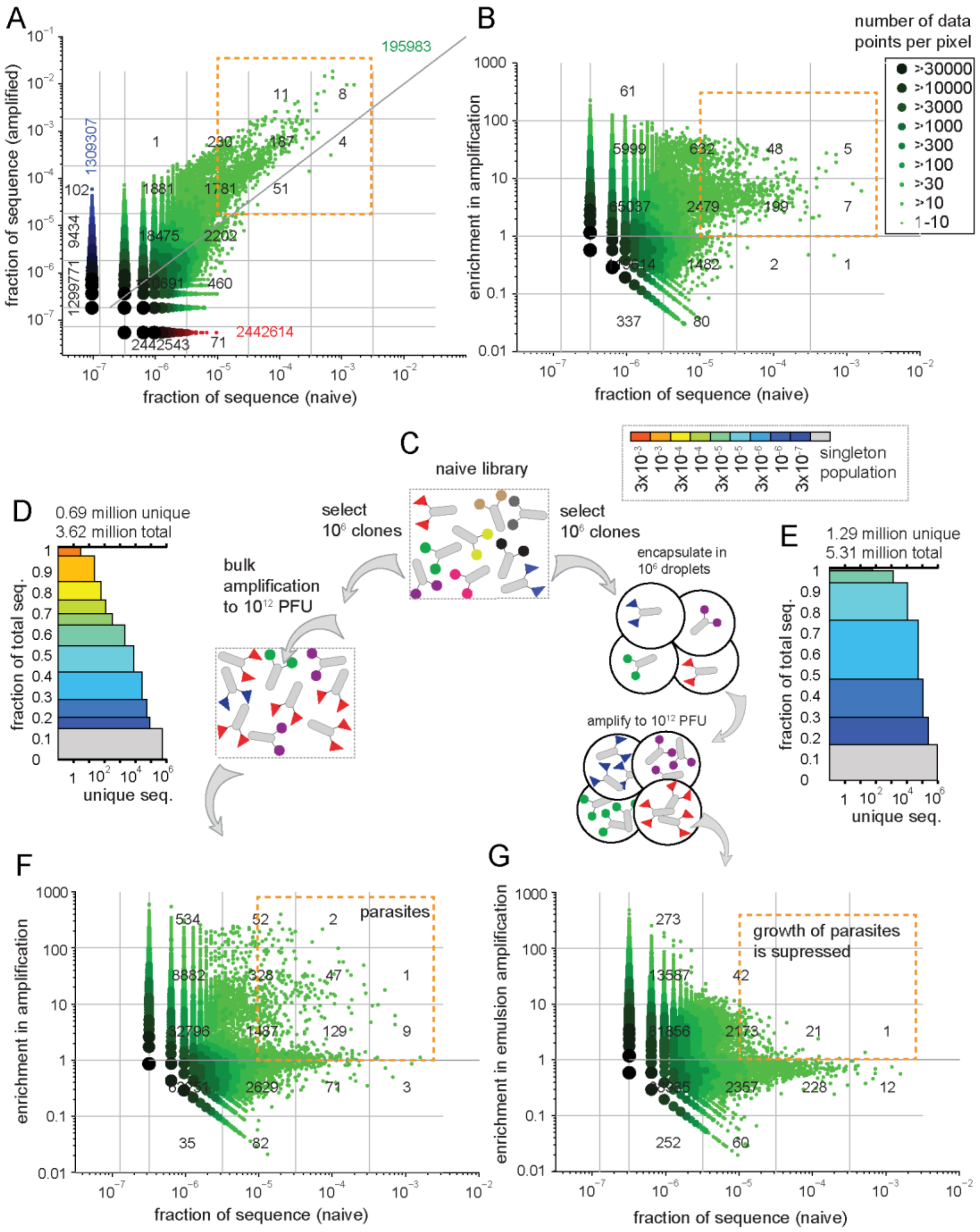


Figure 4.10. (A) Scatter plot describing naïve (**N**) and amplified (**A**) Ph.D.-12 library. Each dot is a unique sequence; multiple data at the same (x,y) coordinate are bigger, darker dots (see legend). Numbers represent the number of data points within each cell of the rectangular grid. Green data describe m-intersect, while blue and red describe m-difference population (data unique to **N** or **A**). (B) We

calculated sequence enrichment as f_{amp}/f_{naive} and plotted it vs. f_{naive}, where f is a fraction of sequences (copy number normalized by total number of reads). (c) Schematic for the amplification of 10^6 clones taken from Ph.D.-12 naïve library. Amplification was performed either in bulk or emulsion (as described in Conditions 1 and 3 in the Methods section). (D-F) Amplification in bulk shows significant enrichment of parasitic sequences compared to amplification in emulsion sequences (>100 fold increase respectively from the original fraction in the naïve library). (E-F) Amplification in emulsion yields uniform library without high-copy-number reads. The enrichments of parasitic sequences is suppressed in emulsion amplification.

like Ion Torrent could provide statistically significant results with high predictive power.

4.3 Conclusion

For libraries made from 10^9 transformants of randomized DNA vectors, the expected abundance of each sequence is 0.0000001%[125]. However, our data indicates that as the DNA is translated and the naïve library is produced in bacteria, the abundance of parasitic sequences rose from 0.0000001% to >0.01% (over five orders of magnitude). Additional amplification of this library in bacteria increases the abundance of parasites to 1%. To our knowledge, this is the first time naïve libraries have been characterized at this level. The analysis of diversity as a result of amplification provides an explanation to several problems commonly observed in the phage library literature: (1) the majority of published screens could identify only a small number of binding clones; (2) binding ability of phage rarely correlates with its abundance in the screen; (3) screens against targets with multiple binding sites (cells and tissues) identify only a few hits. These observations were summarized in several recent reviews [30,50]. To explain these observations, we proposed a two-dimensional selection model (Figure 1.1)[50], which describes how phage display selection and amplification drive collapse of diversity and lead to identification of only a subset of binding sequences (Figure 1). The deep sequencing data presented in this report strengthens this model.

Loss of useful binding clones cannot be mitigated by improved selection procedures: if multiple binders have an equal selection pressure in binding (equal K_d) [138-140] and have unequal selection pressures in amplification (different phage propagation rates), the "slow growing" binder always disappears from the selection and the "parasite" is always selected. Such loss presents no problem if the screen aims to identify only one lead. Loss of binders, however, precludes simultaneous identification of ligands for multi-site targets, such as mixtures of antibodies, and surfaces of cells and tissues. To select diverse sequences for these targets, one must re-engineer amplification (e.g. use emulsion amplification [141]) or avoid amplification entirely and use deep sequencing to run selections without amplification[66]. We note that for some targets, the properties of the sequence that generate stronger binding could be identical to those that enhance amplification. Such a possibility has been proposed for peptide libraries [27].

4.3.1. Parasites and censored clones

Makowski and co-workers, among others, introduced the term “censorship” to describe that some sequences are improbable to find in the library [27]. They linked censorship to a specific pattern of amino acids at specific positions and they hypothesized that censored sequences displayed on phage inhibit infection and production of phage. Makowski also attempted to predict fast growing sequences using the same positional abundance algorithm [27]. Our report uncovers “parasites”, which do not have a specific amino-acid sequence. Their high abundance cannot be predicted from positional abundance of amino acids. For example, if positional abundance was important, most of the point mutants of the parasites should have high copy numbers as well (this hypothesis could be easily rejected by searching for any mutants of sequences in Figure 3C-D, see Figure S8). The biological mechanism that makes some sequences “parasitic” is already known: they emerge due to mutation in the regulatory region of the phage genome[35]. This mechanism was first verified for the parasitic clone HAIYPRH, but has since been characterized in 27 additional parasites that carry 14 unique mutations[18,142]. Since the displayed sequence is not related to mutation in the regulatory region, it might not be possible to predict parasitic sequences. Although Jian Huang did publish a paper in which he speculated that some fast growing sequences can be identified based on their displayed sequence[67]. While reports by Marilena Hall and Jian Huang are contradictory, it is possible that parasites might be sequence dependent and sequence independent. Regardless of their origin, parasites can be reliably mapped prospectively for each batch of the produced library by sequencing a portion of the naïve and amplified library.

Smith and co-workers predicted the existence of "parasites" but they hypothesized that the incidence of mutations that yield parasitic clones are rare and such mutations occur only after serial amplification[36]. Our large-scale sequencing results suggest the opposite: parasitic clones exist in the library immediately after generation; however, they become visible to small-scale sequencing only upon serial re-amplification of the library. Deep-sequencing and appropriate statistical analysis could identify these parasites directly in naïve libraries using only one round of amplification.

4.3.2 Prospective mapping of parasitic clones in all libraries

Our analysis of parasitic clones in this report is based on one lot of the phage library. New England Biolabs (NEB) produced and sold over 10 independent lots of their phage libraries (NEB; personal communication). As these lots could contain different sequences, our analysis does not contain all possible parasitic clones. This fact could explain the incomplete overlap of "parasitic clones" with literature clones in Figure 4.7. Sequencing of all lots of all libraries produced to date could provide a powerful bioinformatics resource for analysis of past and future phage screens. Importantly, this sequencing could be completed using only 1-2 deep-sequencing runs of pooled libraries tagged by barcoded primers [62,143].

The examples presented here were related to peptide libraries identified via phage display. Identical steps can be used to analyze polypeptide libraries from other screens (e.g. RNA-, DNA-, ribosome-, bacteria- or yeast-display) and RNA/DNA aptamers. The molecular mechanisms that generate “parasitic” sequences in RNA or DNA libraries [144,145] are different from the mechanism that leads to emergence of parasitic phage; the phenotypic outcome—enrichment in amplification—can be readily detected by deep sequencing. The online version of our visualization software (chem-derda-web.chem.ualberta.ca) can be expanded to allow for linking to existing databases that contain peptide or nucleotide sequences. We anticipate that the analysis techniques described in this report will improve analysis of selection and amplification from all genetically-encoded libraries.

4.3.3 Emulsion amplification and generation of parasite-free libraries

We believe that it should be possible to use emulsion amplification to repair the collapse of diversity that occurs during the generation of libraries in bacteria. The transformation of bacteria in emulsions has been reported[101]. Large-scale emulsion-generation techniques to produce 10^8-10^9 droplets are also known[146]. This large-scale transformation-in-emulsion could be used to generate naïve libraries with uniform sequence diversity. Due to rapid development of techniques for generation of monodisperse emulsions and their popularization in biotechnology [147], we anticipate that such capabilities could be achieved in a few years.

4.4 Materials and methods

4.4.1 Phage libraries and their amplification

All libraries used in this report were purchased from New England Biolabs. Lot numbers were Ph.D.-7 (# 0061101), Ph.D.-12 (# 0101002), Ph.D.-C7C (# 3). Reported diversity for each library was 10^9 sequences. Each library was amplified under 3 different conditions:

Condition 1, bulk amplification of a 10^6-subset of the library: 10^6 PFU from the original library were mixed with 10^7 CFU of *E. coli* in 1 mL of LB. The mixture was shaken at 200 rpm for 5 h at 37 °C. Amplification yielded ~10^{12} PFU (each initial PFU should have been amplified by a factor of 10^6).

Condition 2, bulk amplification of the entire original library: 10^9 PFU from the original library were mixed with 10^{10} CFU of *E. coli* in 1 L of LB. The mixture was shaken at 200 rpm for 5 h at 37 °C. Amplification yielded ~10^{15} PFU (each initial PFU should have been amplified by a factor of 10^6).

Condition 3, emulsion amplification of a 10^6-subset of the library: 10^6 PFUs from the original library were mixed with 10^7 CFU of *E. coli* in 3 mL of LB and emulsified using microdroplet generator as described previously [112]. The microdroplet generator produces ~4x10^5 droplets/mL, 3 mL of LB was used to ensure each clone was encapsulated into individual compartments and to avoid growth bias between clones. The emulsion was shaken at 40 rpm for 5 h at 37 °C and then destabilized to combine all amplified phage. Amplification yielded ~10^{12} PFU (each initial PFU should have been amplified by a factor of 10^6).

The phage population from each condition was processed for deep sequencing as described below.

4.4.2 Illumina sequencing

The steps for deep sequencing of phage libraries and analysis of the results were similar to those described in our previous report [143]. In short, we isolated ssDNA from M13 phage using NaI/EtOH precipitation and purified it using phenol-chloroform extraction. The variable regions were isolated from the library and amplified by PCR using barcoded primers (see Appendix 1.13, 1.14). The DNA was pooled together and processed for Illumina sequencing using the manufacturer's protocols for end-repair, adenylation, adapter ligation, and PCR amplification of the product. The samples were sequenced on HiSeq Illumina instrument using a 50 bp single-end run. FASTQ files were analyzed using custom MATLAB scripts (Appendix 1.14). The software generated plain text-based lists of sequences and their abundances (Appendix 1.14). These text files were used by MATLAB scripts to generate Figures 4.2, 4.5-4.8 (see "Data Visualization" section below). Raw FASTQ (>10 Gb of data) and Matlab files are not included in this thesis and are available upon request.

4.4.3 Illumina analysis

Sequences emanating from each amplification condition were identified using their respective barcodes (Appendix 1.14). Abundances of the sequences and their quantities are described in Appendix 1.15. In short, ~98% of the sequences could be mapped to a specific barcode. In the mapped sequences, 60% of the sequences contained all nucleotide locations with Phred Score>30. From these sequences, 80% contained nucleotides with $(NNK)_n$ structure. We selected only sequences that had NNK-structure and a Phred>30 for each position in the sequence. We note that Illumina sequencing yielded both forward (**F**) and reverse (**R**) sequences originating from the (+) and (-) strand of the vector. The ratio of sequence abundances in **F** and **R** multisets varied from 40 to 60% (Appendix 1.15). Forward and reverse sequences represent two independent sampling of the same DNA population and the abundances in **F** and **R** should be within the (*ave*±*ste*) range, where *ave* is average expected value and *ste* is standard error of the true sequence abundance. Specifically, highly abundant sequences should be identified in both **F** and **R** pools at similar abundances, whereas sequences present in the **F** pool with copy number of 1-5, could be absent from the **R** pool (and vice versa). In our processing, after removing non-NNK sequences and Phred<30 sequences, we observed significant overlap in sequence identity in **F** and **R** populations and similar sequence abundances in these populations.

4.4.4 Mathematical representation of sequence uniqueness and their abundance

A given list of sequences $[s_1, s_2 \ldots s_n]$ can be conveniently represented as mathematical multisets, (S, m), where S is the set of all unique sequences, and m is the count of each sequence element. We combine the multiset of all forward sequences **F** with the multiset of all reverse sequences **R** by union for analysis.

The union of all forward and reverse sequences is the list of all unique sequence from either F or R, where the count of each sequence is equal to the maximum number of its appearances in either the forward or reverse population max(f,r). For multisets **F**=(F,f) and **R**=(R,r), $\mathbf{C}_{\mathbf{F \cup R}}$=(F ∪ R, max(*f*,*r*)) is the union of **F** and **R** (SI Scheme 2).

4.4.5 Ion Torrent sequencing

We isolated ssDNA from M13 phage libraries using QIAprep Spin M13 kit (#27704). Isolated phage ssDNA was subjected to PCR amplification with primers flanking the variable region. To avoid a second round of PCR amplification, the primers contained Ion Torrent adapters at the 5' ends. The concentration of PCR fragments that resulted from amplification of phage libraries was determined by analytical gel (2% (w/v) agarose gel in TBE buffer using a low molecular weight DNA ladder as a standard (NEB, #N3233S). Multiple PCR-amplified phage libraries amplified with different barcoded primers were pooled together before running E-gel. The band corresponding to the expected dsDNA product was purified on an E-gel SizeSelect 2% gel (Invitrogen). The dsDNA fragments were extracted with RNAse-free water and the concentration determined by Qubit Fluorimeter (Invitrogen) using manufacturer's protocol. The dsDNA fragments were ligated onto Ion Sphere Particles (ISPs) and amplified by emulsion PCR according to Ion Torrent protocol. The concentration of ISPs with ligated dsDNA fragments after emulsion PCR was determined using Qubit Fluorimeter (Invitrogen) according to manufacturer's protocol. The ISPs with ligated dsDNA fragments were enriched for and loaded on an Ion 316 chip. The sequencing was performed using an Ion Torrent system (Life Technologies) with an Ion OneTouch 200 Template Kit.

4.4.6 Data visualization

There are no standard tools for the effective analysis of 10^3-10^4 peptide sequences from a pool of 10^9 sequences. Tabular presentation of sequences is ineffective because it hinders direct comparison of important information and patterns emerging in large data sets. Sequence logo [128,148] or combination of logos [58,97] works only if results converge on a defined sequence motif(s). It does not work when common motifs are frame-shifted[58], or when common motifs do not exist [143]. The main challenge in data visualization is the simultaneous analysis of sequence and its copy number.

A result from the selection can be represented mathematically as a *multiset*, a set in which members can appear more than once [143]. There are few standard visualization techniques for multisets. 1D-stacked bars describe both the number of unique sequences and their copy number in the multiset (Appendix 1.16B) [50,143]. To describe large multisets, the set elements could be grouped by their copy number and represented in 2D: the width of each segment illustrates the number of the unique sequences in this segment; its height represents the fraction of these sequences in the library (Appendix 1.16D). Multisets can be compared using Venn diagrams. Appendix 1.16E describes examples of two multisets, **X** and **Y,** and the results of intersection (**X**∩**Y**) and difference (**X** \ **Y**) operators (Appendix

1.16F). We define the multiset operations m-intersection $\mathbf{X}_m \cap \mathbf{Y}$, and m-difference $\mathbf{Y}_m \backslash \mathbf{X}$, (Appendix 1.16G) in order to describe weighted contribution of common sequences. M-intersect is defined as $\mathrm{X_m} \cap \mathrm{Y} = \{(\mathrm{s}, \mathrm{n}) |\ (\mathrm{s}, \mathrm{m}) \in \mathrm{X} \bigcap \mathrm{Y} \wedge (\mathrm{s}, \mathrm{n}) \in \mathrm{X}\}$ (i.e., **X** m-intersect **Y** is equal to the multiset of elements (s,n) such that that s exists in **X** intersect **Y** and n is the count of element s in **X**). That is, m-intersect contains every unique element in the intersection of **X** and **Y** at multiplicities of the element's original count in **X**. We define the m-difference as the remainder: $\mathbf{X}_m \backslash \mathbf{Y} = \mathbf{X} - \mathbf{X}_m \cap \mathbf{Y}$. An intuitive tool for multiset comparison is a scatter plot, which describes pairwise differences in abundances of the individual elements (Appendix 1.16H). These plots could be equipped with color gradients and quantification grids (akin to those used in flow-cytometry software).

We believe that both research data and visualization techniques used to represent the data will be of benefit to the reader. With increasing use of deep-sequencing techniques for the analysis of *in vitro* selection procedures, the visualization techniques described here will be especially useful for analysis and comparison of deep-sequencing results.

4.4.7 Generation of stacked bars and scatter plots

All images described in Figures 4.2, 4.4-4.10 were generated from plain text files describing identity and abundance of peptide/nucleotide sequences. Specifically, stacked bars, Venn diagrams, and scatter plots in Figures 4.2, 4.4-4.8 were generated by one MatLab script '*command_center.m*', which contains user-friendly graphic user interface (scripts can be downloaded from the derda group website www.chem.ualberta.ca/~derda/parasitepaper or the journal supporting information nar.oxfordjournals.org/content/suppl/2013/10/29/gkt1104.DC1).

To calculate the dimensions of the 2D stacked bars segments, we used an algorithm similar to that described in Appendix 1.16. We started from a library that contained S^{all} total sequences and U^{all} unique sequences. We binned the library by copy number (N): sequence belongs to bin [N_1 N_2] if the copy number of sequence is $>N_1$ and $\leq N_2$. We used bins [0 1], [1-3], [3-10], [10-30], [30-100], [100-300] etc, because this binning was uniform on log-scale. In some cases, we converted copy number range to sequence abundance as N_i / N_{reads}; where N_i is sequence copy number and N_{reads} is total number of reads. In that case, binning was performed as [0.03 0.1] [0.1 0.3] [0.3 1], etc. Each bin was represented by segment of specific color. We calculated the total number of sequences and the total number of unique sequences in each bin (S^{bin} and U^{bin}). The height *h* and width *w* of the segment representing each bin was calculated as $h^{bin} = S^{bin} / S^{all}$, and $w^{bin} = \log_{10}(U^{bin})$.

Specifically, in Figure 4.1C as an example, the top crimson segment contains six unique peptides ($U^{crimson} = 6$). Each sequence has abundance ≤0.03 and >0.01. Due to their large abundance in the library, these six peptides constitute 8% of the library ($S^{crimson} / S^{all} = 0.08$). The peptides in the bottom blue segment also constitute 8% of the library. This segment, however, contains 100,000 unique peptides ($U^{blue} = 100,000$). Each peptide has an abundance ≤0.0000003 and >0.0000001. Bottom grey segment represents singleton populations (sequences were observed only once). The number of sequences and their identities in the singleton population should be interpreted with caution because this segment contains the highest number of sequencing errors. However, singleton populations cannot be discarded because in some cases they constitute over 70% of the library (Figure 4.1C).

4.4.8 Analysis Ph.D.-7, Ph.D.-12, and Ph.D.-C7C library screens

The literature data of phage display screens that used Ph.D.-7, Ph.D.-12 and Ph.D.-C7C libraries were extracted from raw MimoDB 2.0 database. The MimoDB is a database of all peptides identified by phage display screens [149]. We used database provided by Jian Huang, from which we extracted hits for each library. The files were used by *command_center* script to generate Figure 4.9 and 4.10 (scripts can be downloaded from the Derda group website www.chem.ualberta.ca/~derda/parasitepaper or the journal supporting information nar.oxfordjournals.org/content/suppl/2013/10/29/gkt1104.DC1).

Chapter 5 Emulsion amplification levels the selection landscape in phage display panning and uncovers 'lost' binders

5.1 Introduction

Genetically-encoded libraries displayed on phage[51], yeast[150], or RNA[151] are powerful technologies for the discovery of ligands for virtually any molecular target, including many therapeutically-relevant targets[152]. They also permit unbiased selection of ligands that bind multi-target entities such as cells and organs (reviewed in ref. 5)[153], and the human antibody repertoire (see ref. 6 and references within)[154]. Functional ligands emanating from the multi-target screens can give rise to therapeutic candidates[152,155] or targeting probes[153,156] and instructive materials that control stem cell differentiation[157] and self-renewal[12]. All *in vitro* selection strategies start from a diverse library of ligands (10^9 or higher) and increase the abundance of target-binding 'hits' in the sequence pool via rounds of panning—retention of binding and removal of non-binding ligands—and re-amplification of recovered ligands. These steps exert two orthogonal selection pressures: (i) panning selects for ligands that bind to the target; (ii) re-amplification selects the library clones that exhibit higher rates of amplification than the population average. The latter bias has been characterized in oligonucleotide libraries[93,158] and in phage-displayed libraries of peptides[17,34]. In phage, growth enhancement arises from mutations in the Shine-Dalgarno sequence of *pII* protein[18] or in the (+)-origin[36]. The phage clones with high rates of amplification can be identified in libraries by deep sequencing and can be traced in >100 of published screens[17]. We hypothesize that the amplification-induced bias is a major detriment in selection of ligands for multi-target baits, such as cells or mixtures of antibodies. In such a screen, many target-binding ligands exhibit equal selection pressure due to binding, and a few clones with a higher amplification rate can dominate the screen and suppress identification of other target-binding ligands.

5.2 Results and discussion

To minimize the amplification bias in a selection for a multi-target bait, we employed *in vitro* compartmentalization to replace the standard 'bulk-amplification' method (BA) with 'emulsion-amplification' (EmA)[99], while maintaining the selection method and target constant (Figure 5.1A). EmA can prevent growth-induced bias[99] and avoid the collapse of diversity in a phage library[17]. The advantages of compartmentalization are well established in PCR, and emulsion-PCR is a standard method for bias-free amplification of nucleic acids[20,21]. In this report, we show that integration of EmA into the phage display selection process dramatically altered the selection landscape and led to the discovery of new classes of binding ligands that cannot be identified in traditional phage display selection using the same library. While this report focusses on phage-display libraries of peptides in which we previously characterized

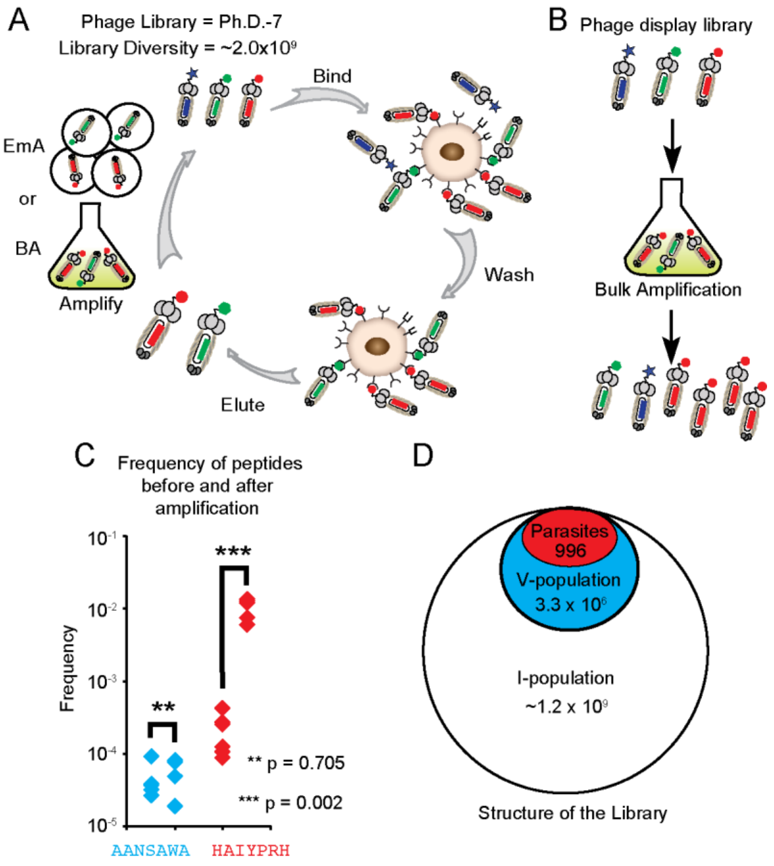


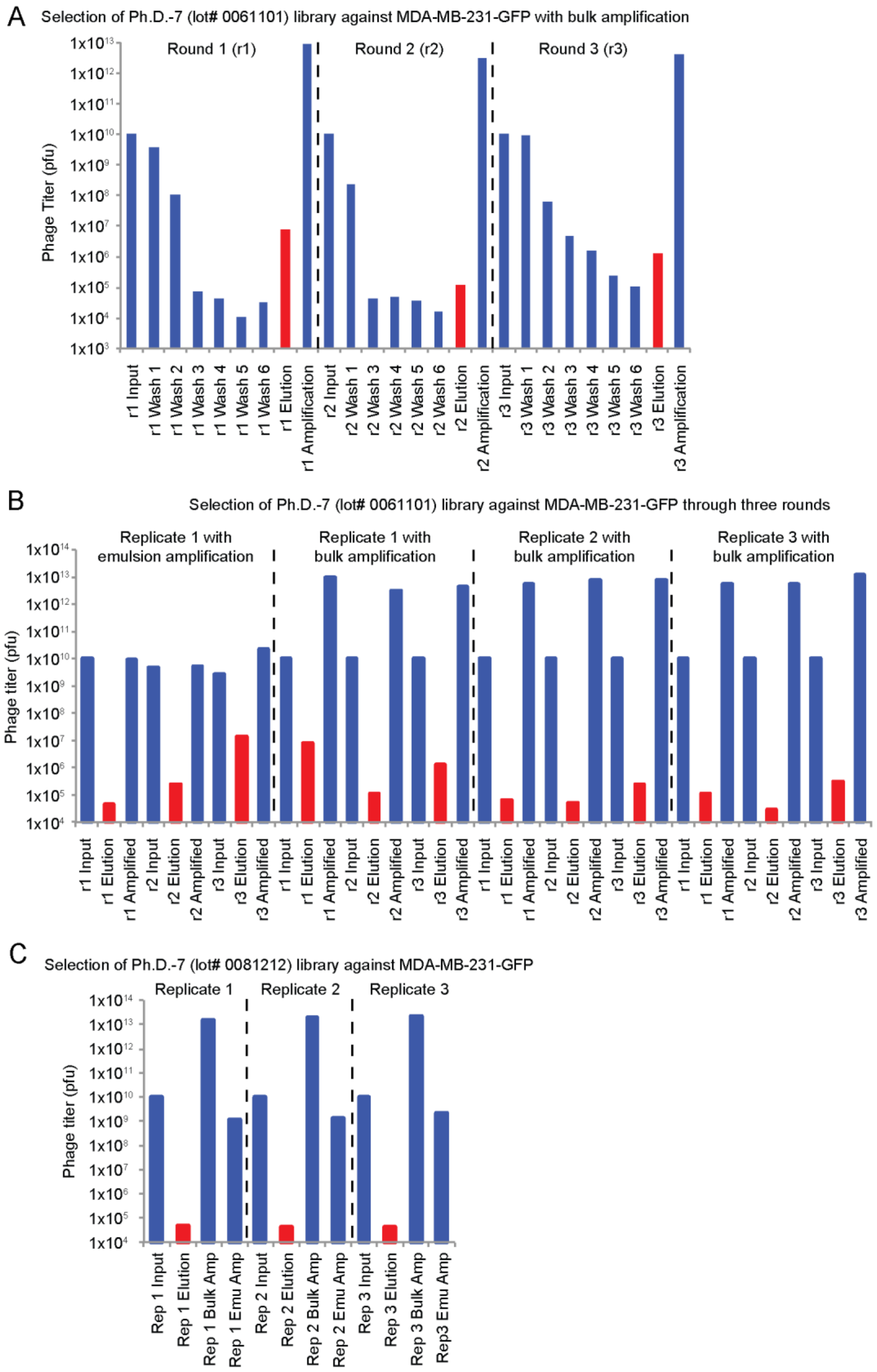
Figure 5.1. (A) Schematic representation of two independent selections performed with a Ph.D.-7 library against breast cancer cell line, MDA-MB-231. In one selection, the eluted library is amplified in a bulk solution (BA), in the other; the eluted library is amplified in emulsion (EmA). Phage display libraries contain clones with high growth rates (parasites). To identify parasites, (B) a phage library was amplified in a common solution (BA). The amplified and naïve phage libraries were sequenced using Illumina. We identified the parasite population as sequences that significantly ($p < 0.05$) amplified more from the naïve population. We used volcano analysis to analyze all members of the naïve and amplified libraries. (C) The abundance of the representative non-parasite sequence AANSAWA and the parasite sequence HAIYPRH before and after

amplification. The latter significantly increases as a result of amplification. (D) To determine the identity of sequences obtained from selections, the phage library was divided into three groups of sequences; the invisible (I)-population, visible (V)-population, and parasites (P-population).

amplification-induced bias[17], we hypothesize that results can be universally applicable to other *in vitro* selection strategies.

5.2.1 EmA expands the diversity landscape

To explore the changes in the selection landscape as a function of amplification bias, we performed selection of peptides that bind to the breast cancer cell line MDA-MB-231. We used either EmA or BA after each round of panning and proceeded for a total of three rounds (Figure 5.1A and Figure 5.2). To facilitate tracing of ligands during the selection, we analyzed the library by deep sequencing. Working with a complete library of 7-mer peptides—10^9 theoretical and experimental diversity as reported by the manufacturer—allowed us to classify the peptide sequences of the phage library into three groups (Figure 5.1D): (i) the 'visible population' (V) defined as 10^6 sequences identified by Illumina sequencing, (ii) the 'invisible population' (I) corresponding to the set of all possible 7-mer peptides excluding the visible population; (iii) within the V-population, we mapped the 'parasite population' (P) as ~10^3 sequences that increased significantly during re-amplification in the absence of selection[17] (Figure 5.1B-D and Figure 5.3). Three rounds of phage display selection with



A Selection of Ph.D.-7 (lot# 0061101) library against MDA-MB-231-GFP with bulk amplification

B Selection of Ph.D.-7 (lot# 0061101) library against MDA-MB-231-GFP through three rounds

C Selection of Ph.D.-7 (lot# 0081212) library against MDA-MB-231-GFP

Figure 5.2. Monitoring phage titers during selection. The phage titer was monitored for input, washes 1 through 6, eluted and amplified (in BA) phage over three rounds of selection against MDA-MB-231-GFP cells (A). We observe that four washes in enough to remove non-specific cell binding phage. (B) We monitor the input, output and amplified phage titers for the remaining replicates of selection using BA and EmA. Replicate 1 with bulk amplification is the same as panel A. We observe the output phage titer increases during each round of selection with EmA. Output phage titer for each replicate of selection with BA is less consistent. Selection in A and B was performed with each replicate separate from one another, and half of the amplified library was used for deep-sequencing. C Selection was performed with a second lot of the Ph.D.-7 library. In this method, each replicate was panned against MDA-MB-231-GFP cells, with the eluted phage mixture separated into two samples and amplified either by BA or EmA.

BA identifies mainly parasite sequences, as seen in tracing of the top 20 (Figure 5.4A) or top 50 sequences (Figure 5.4B) from the third round of panning back to the naïve library. We note that the top 20 and top 50 sequences comprise 84 and 95% of the sequence population after the third round of selection. Between 75-95% of the sequences enriched after round three originated from the P-population of the library (Figure 5.4A and Figure 5.5). To explore how the composition of library changes over rounds of panning, we compared the identity of the top 50 enriched sequences from the BA-screen over three rounds of panning (Figure

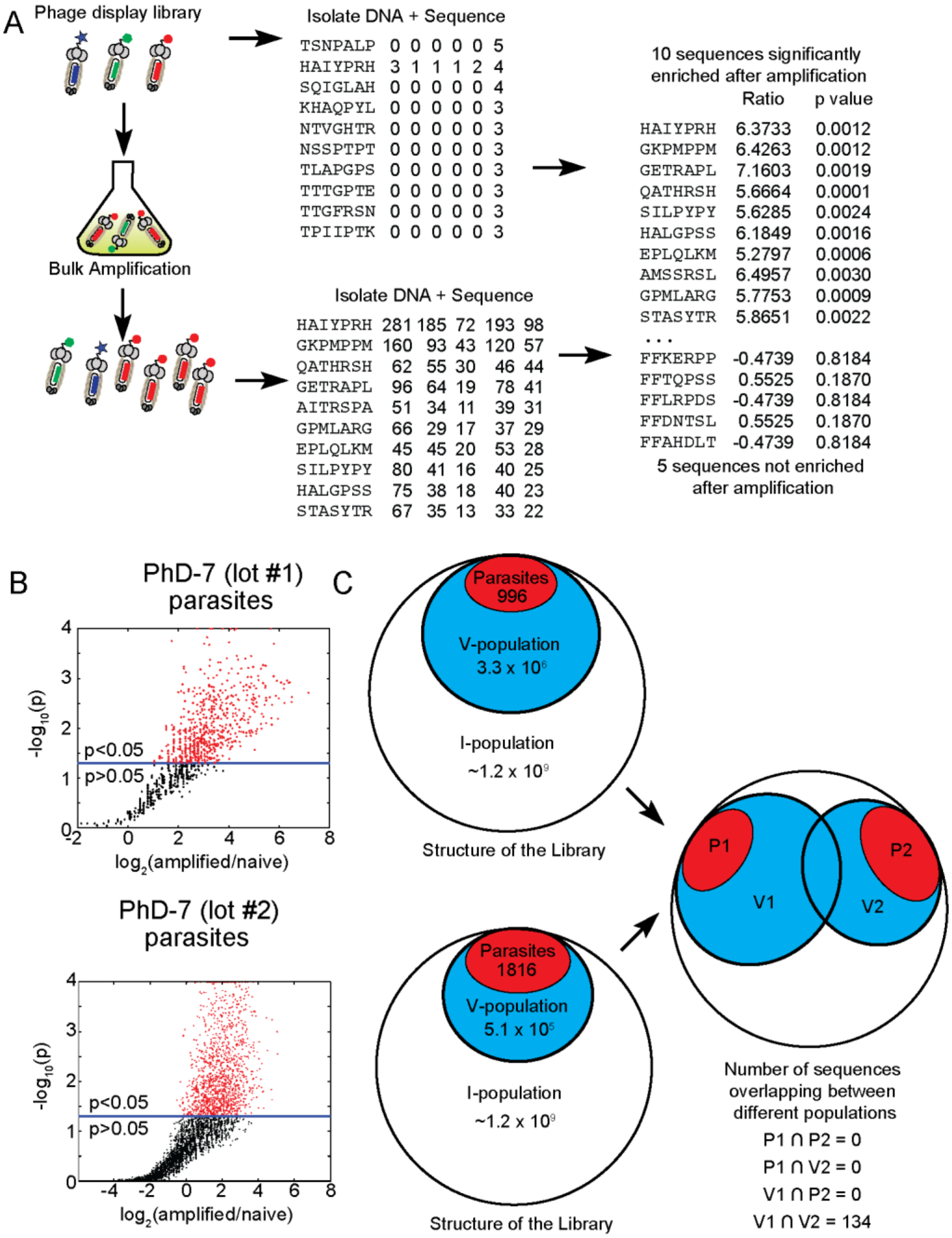


Figure 5.3. (A) Schematic representation for the determination of parasite sequences in a phage display library. Briefly, the phage display library was added to a flask of lysogeny broth media and bacteria. The library was amplified and the phage isolated. The cDNA of the naïve and amplified libraries were extracted and the variable library region was amplified by PCR and sequenced by Illumina. After sequencing, the data was processed and organized. We provide a

representative snapshot illustrating the top 10 sequences organized by the descending copy number of the last replicate (replicate 6 for the naïve library and replicate 5 for the amplified library). For each sequence, we determine the frequency found in each replicate and determine the average ratio before and after amplification and the p-value through a t-test. The P-population is identified as sequences that amplify significantly ($p < 0.05$) more from the naïve population using volcano analysis. (B) A volcano analysis is used to test all sequences. (C) Representative structure of the two Ph.D.-7 libraries used. The "invisible population" (I-population) is composed of all possible sequences of a 7-mer peptide library. The theoretical diversity is ~1.2 billion sequences. Within the I-population is the "visible population" (V-population), which contains all sequences that was identified by Illumina sequencing[17]. The subset of V-population that significantly amplify more than the rest of the population are made up of parasites sequences. Out of all sequences identified through Illumina for both Ph.D.-7 libraries (V-populations), there were only 112 sequences identified as overlapping. There was no overlap between the two P-populations nor between P- and V-populations.

5.4B). In all three rounds of the BA-screen, >78% of the top 50 sequences originated from the P-population. Additionally, to investigate whether the BA-screen can enrich populations of sequences beyond the P-population, we

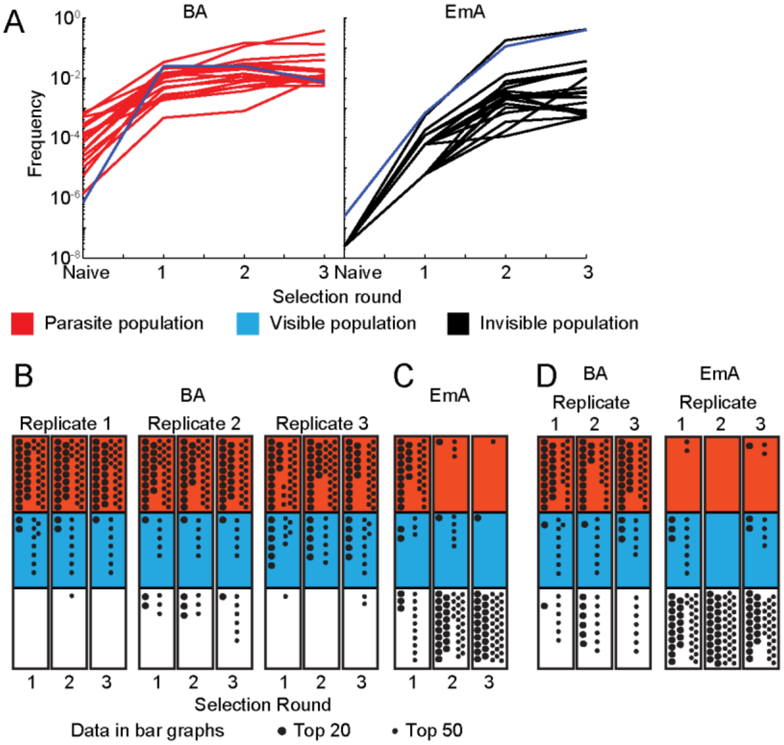
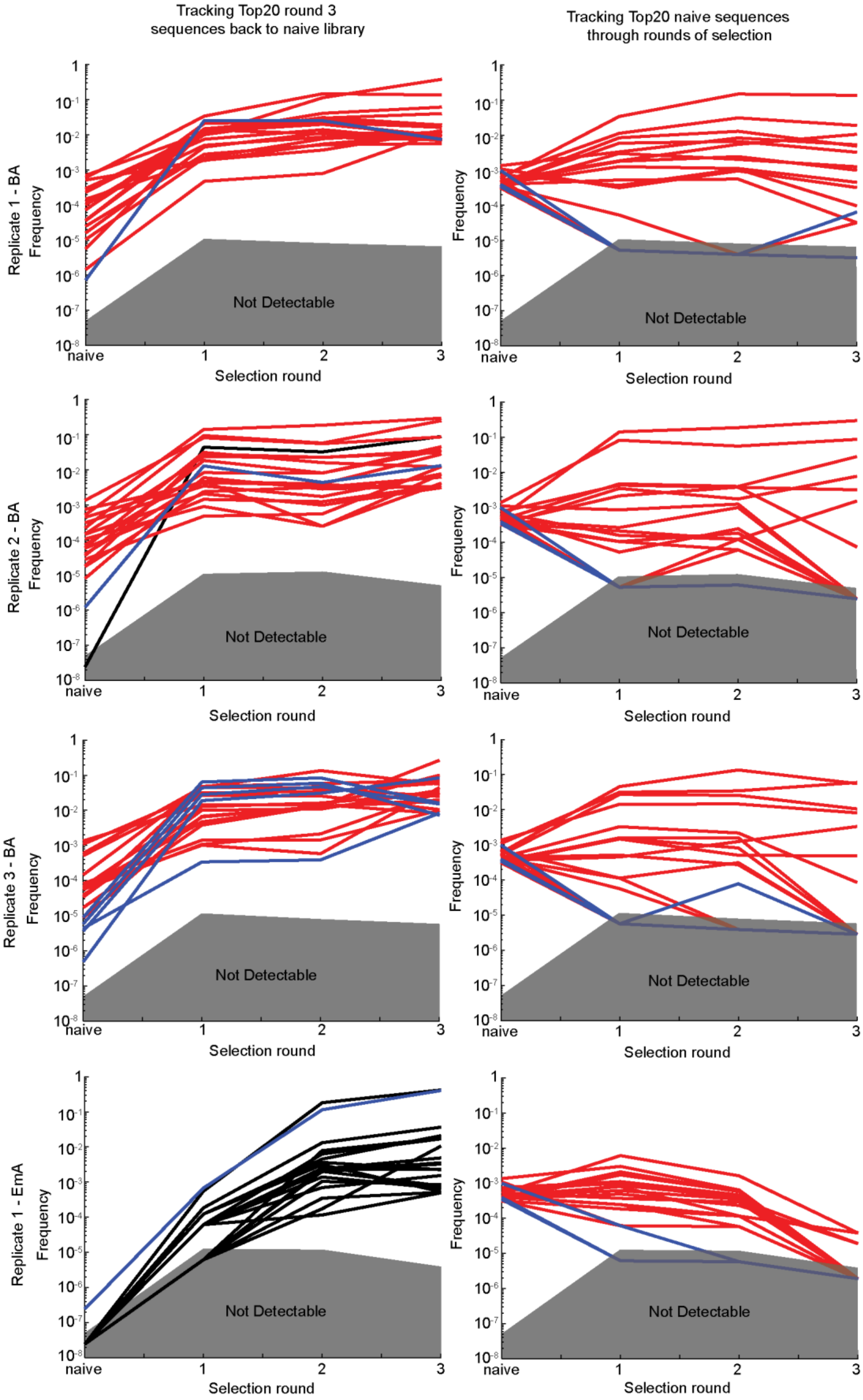


Figure 5.4. (A) Tracing the origin of the top 20 hits after 3 rounds of BA and EmA selections. Each trace describes a unique peptide sequence. The color of the sequences describes its origin (red – 'parasite'; blue – 'visible'; black – 'invisible'). Each sequence in the third round of selection is placed into one of three segments of a stacked bar representing the classification of the library. The darker and lighter spots denote the top 20 and 50 identified sequences respectively. (B) The origin of the top 50 sequences enriched in three separate replicates of BA and one replicate of EmA selection. Each 'dot' represents an individual sequence, the 'big dots' correspond to top 20 sequences, and the 'small dots' correspond to the remaining top 50 sequences. (C) Selection was repeated with a different PhD-7 library (lot #0081212) for one round and the classification of the top 50 sequences was determined. The EmA-screen enriches the majority

of sequences from the I-population, whereas the BA-screen enriches sequences mainly from the P-population.

compared three independent selections with three rounds of panning and amplification. The identity of the enriched sequences changed from screen to screen (see Appendix 1.17 and 1.18 for sequences of peptides and copy number in each round); however, we consistently observed that 68-84% of sequences originated from the P-population (Figure 5.4B).

Replacing BA with EmA in phage display selection resulted in enrichment of a different class of sequences; we uncovered only 1 parasite in the top 50 sequences (Figure 5.4D). More than 95% of peptides originated from the invisible population of the phage library (Figure 5.4A and Figure 5.5). By extending the analysis to earlier rounds of selection with EmA, we observed that 32 out of the top 50 sequences in Round 1 were parasites, but this number decreased to 4/50 in Round 2 and 1/50 after Round 3 (Figure 5.4C). Depletion of parasites in EmA-selection can also be visualized by tracing the fate of the top 20 sequences in the naïve library through three rounds of selection. In the EmA-screen, the top 20 sequences are suppressed to levels not detectable by deep-sequencing (Figure 5.5). In contrast, in BA-screens, the top sequences persist through multiple rounds of selection. As there was little variation in the fractions of parasites over three individual replicates of three rounds of the BA-screens (Figure 5.4B), we hypothesize that selection driven by BA cannot identify a broader class of binders


Tracking Top20 round 3 sequences back to naive library

Tracking Top20 naive sequences through rounds of selection

Figure 5.5. Left panels identify the origin, in the naïve library, of the top 20 sequences from the third round of selection. The colors for each line indicate the identity of each sequence (red – 'parasite'; blue – 'visible'; black – 'invisible'). Parasites end up being enriched when using the BA method, while sequences from the I-population are enriched when using the EmA method. Right panels illustrate tracking the top 20 sequences in the naïve library through the selection rounds. When BA-screen is used, sequences in the naïve library follow any one of three paths from the naïve library to the third round of selection; become enriched, remain at equilibrium, or become depleted. When EmA-screen is used, all top 20 sequences in the naïve library are depleted by the third round of selection.

cannot identify a broader class of binders even if repeated multiple times. We hypothesize that conventional (BA) selection, thus, operates in a limited population of parasites that constitute ~0.0001% of the entire Ph.D.-7 library (Figure 5.3B-C). In contrast, selection with EmA identifies ligands from a larger population of the library.

To confirm that the EmA-screen reproducibly selects hits from a greater diversity space, we repeated BA and EmA selections with a different lot of a Ph.D.-7 library (lot #0081212, or for simplicity 'lot #2') (Figure 5.2C). Growth enhancement in parasites is not related to displayed sequences; rather it originates from a random mutation in a regulatory region of the phage genome[18]. As a result, the identity of the sequences displayed on parasite phages changes when the library is re-expressed *de novo*. As a result, lot #2 of the library contained a parasite population that was completely different from the P-population in lot #1 used above (Figure 5.1D, P2: 1816 parasites are in Figure 5.3B-C). We performed BA and EmA selection in triplicate using one round of panning. As in previous experiments, the top 50 enriched sequences from lot #2 and enriched via a BA-screen were predominately parasites (34/50, 29/50, 35/50, Figure 5.4D). The EmA-screen yielded 2/50, 0/50, and 3/50 sequences from the P-population in three independent screens. Moreover, 76% or more sequences originated from the I-population not accessible to BA-screen (Figure 5.4D). To demonstrate that parasite bias is universal for any multi-receptor target, we screened HEK293 cells, commonly used for protein expression, against lot #2 of the Ph.D.-7 library. The selection was repeated three times using one round of BA-selection, and revealed the same bias toward parasites. From the top 50 sequences identified in the screen, 31/50, 36/50, and 33/50 sequences originated from the P2-population (Appendix 1.29). These observations suggest the generality of the phenomena: all M13 libraries contain a unique small P-population and in all multi-receptor screens the origin of the most abundant sequences is strongly biased towards a parasite population. These observations are in line with our previously reported analysis of published reports that use M13-displayed peptide libraries of the same origin. From 1961 peptides identified from Ph.D.-7 libraries, 95 belong to the P-population[17].

5.2.2 Validation of cell binding ligands

To validate the cell-binding hits from the deep sequencing analyses, we synthesized the peptide sequences on Teflon-patterned paper arrays and performed a cell-binding assay as described previously[159]. Specifically, we studied short-term adhesion of breast cancer cells MDA-MB-231-GFP to the top 20 peptides identified from every round and every replicate of the BA- and EmA-screens (Figure 5.6A and Appendix 1.17). For each peptide, we validated the presence of cells on the peptide-modified paper by confocal fluorescent microscopy (Figure 5.6H and Appendix 1.26 and 1.27), and used a fluorescent gel scanner for high-throughput imaging of cells on each zone of the peptide array. We converted this intensity to the number of cells (Figure 5.6B and Appendixes 1.19, 1.21-1.25) using a calibration curve (Appendix 1.20). Binding of cells to peptides on peptide arrays was reproducible over multiple replicates (10) for each batch (Appendix 1.25). We noticed, however, that different synthetic batches of peptide-arrays can exhibit varying levels of cell adhesion. This variability was correlated with the surface density of the peptides (Appendix 1.28). In subsequent sections, we tested cell-binding ability of all peptides using 4-8 replicates and two different batches, unless stated otherwise. We classified cell-binding peptides as confirmed hits when the peptides supported adhesion of significantly ($p < 0.05$) more cells than the negative control (GGRDS peptide). Both BA and EmA-screens yielded cell-binding peptides from every round and replicate; from the top 20 sequences in 18 screens and replicates (360 peptides total, 209 unique), we identified 56 unique cell-binding hits (Figure 5.6 C-D). The fraction of validated

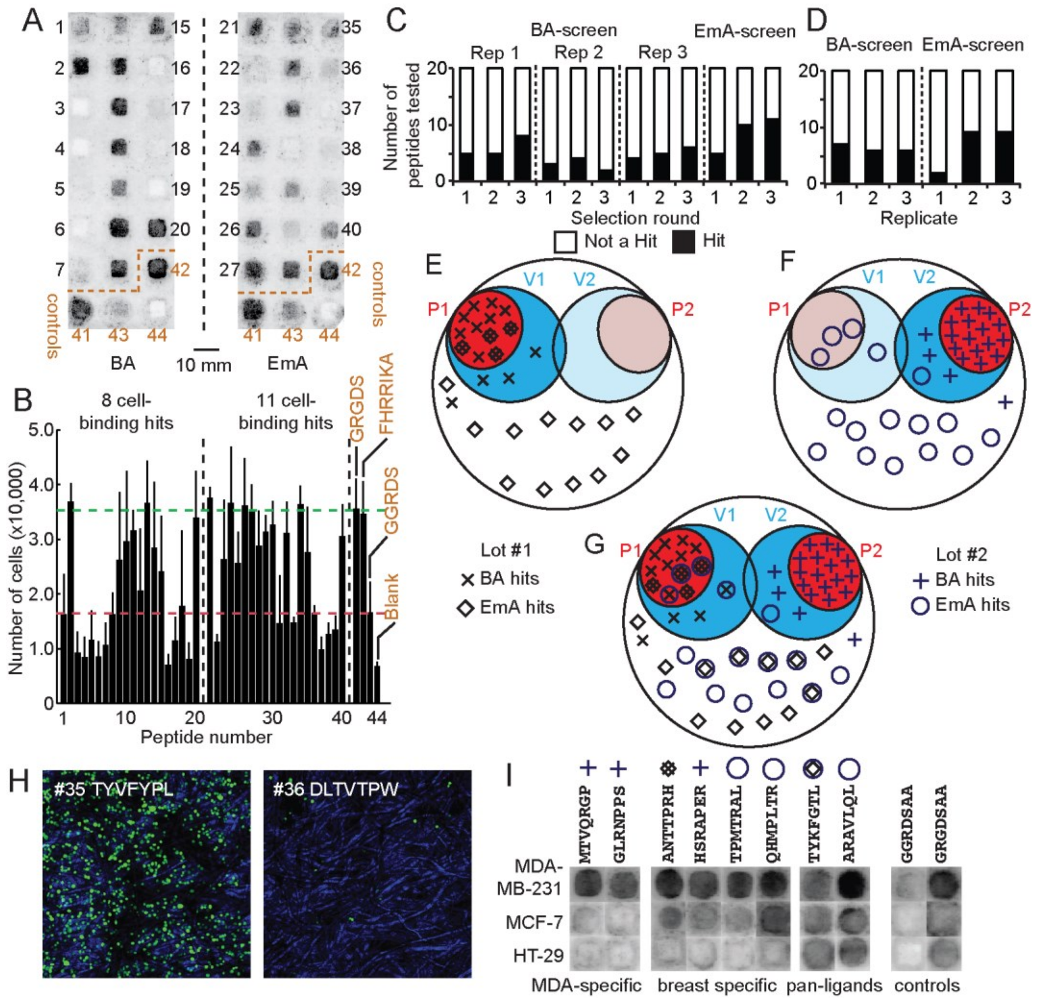
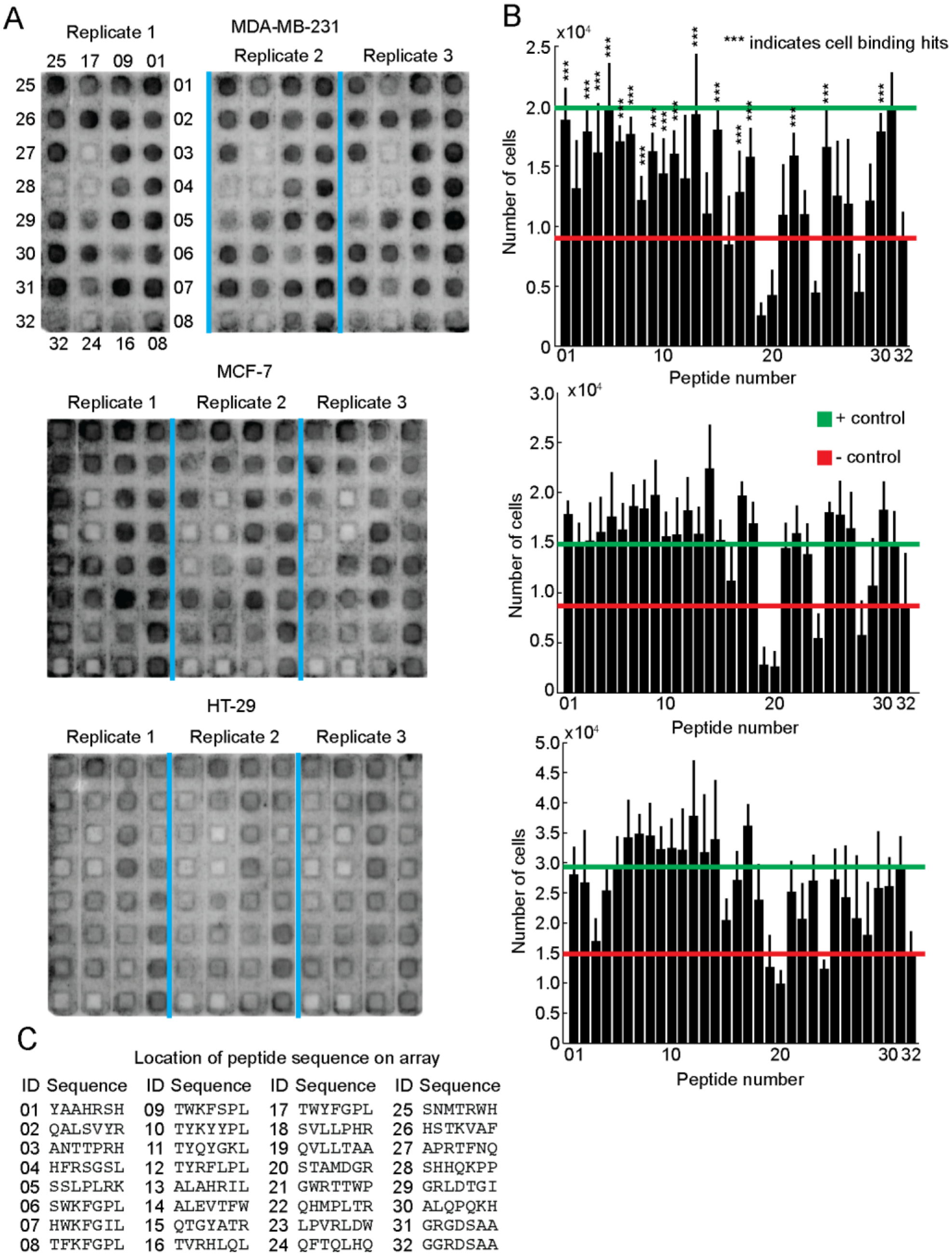


Figure 5.6. (A) Schematic representation of the workflow for validating the top 20 sequences from each round and replicate of selection for cell adhesion by synthesis on Teflon patterned peptide arrays. MDA-MB-231-GFP cells are seeded on paper peptide arrays and visualized by a fluorescent gel scanner. The peptide zones with cells will appear dark; peptide zones without cells will appear light. (B) Representative fluorescent gel scanner image of the top 20 peptide sequences after round 3 of BA and EmA selection, synthesized on paper and tested for short term adhesion with MDA-MB-231-GFP cells (see Appendix 1.20) for list of sequences). The darker the peptide zone, the more cells there are. Integrin- and

Heparin-binding peptides are the positive controls (GRGDS and FHRRIKA, respectively), and paper bearing no peptide (blank) and scrambled GRGDS (GGRDS) are used as negative controls. The scale bar represents 10 mm. (C) The greyscale intensity for each peptide is correlated to a standard curve of known amounts of cells and plotted. Cell-binding hit peptides are determined as binding significantly more ($p < 0.05$) cells than the negative control (GGRDS, red line). (D-E) Plot of the number of hits and non-hits for each selection relicate and round from the BA- and EmA-screens. The number of hits is identified for selection performed with two PhD-7 libraries, lot #0061101 (D) and #0081212 (E). (F) Placement of all hits identified from BA and EmA selection methods into the three population groups of the PhD-7 library determined as originating from the P-population (red), V-population (blue), or I-population (white). Hits from the BA-screen originate primarily from the P-population, whereas hits from the EmA-screen are selected throughout the diversity space, primarily from the I-population. (G) The cell binding hits are re-tested for cell adhesion to two different cancer cell lines, MCF-7 and HT-29. Representative images of selected peptides binding to MDA-MB-231, MCF-7, and HT-29. The peptides include: two controls (GRGDSAA, +ve ctl, and GGRDSAA, -ve ctl), two bind to all cell lines (ARAVLQL and TYKFGTL), four bind to MDA-MB-231 and MCF-7, but not HT-29 (ANTTPRH, QHMPLTR, TPMTRAL, and HSRAPER), and two bind MDA-MB-231, but no MCF-7 and HT-29 (GLRNPPS and MTVQRGP).



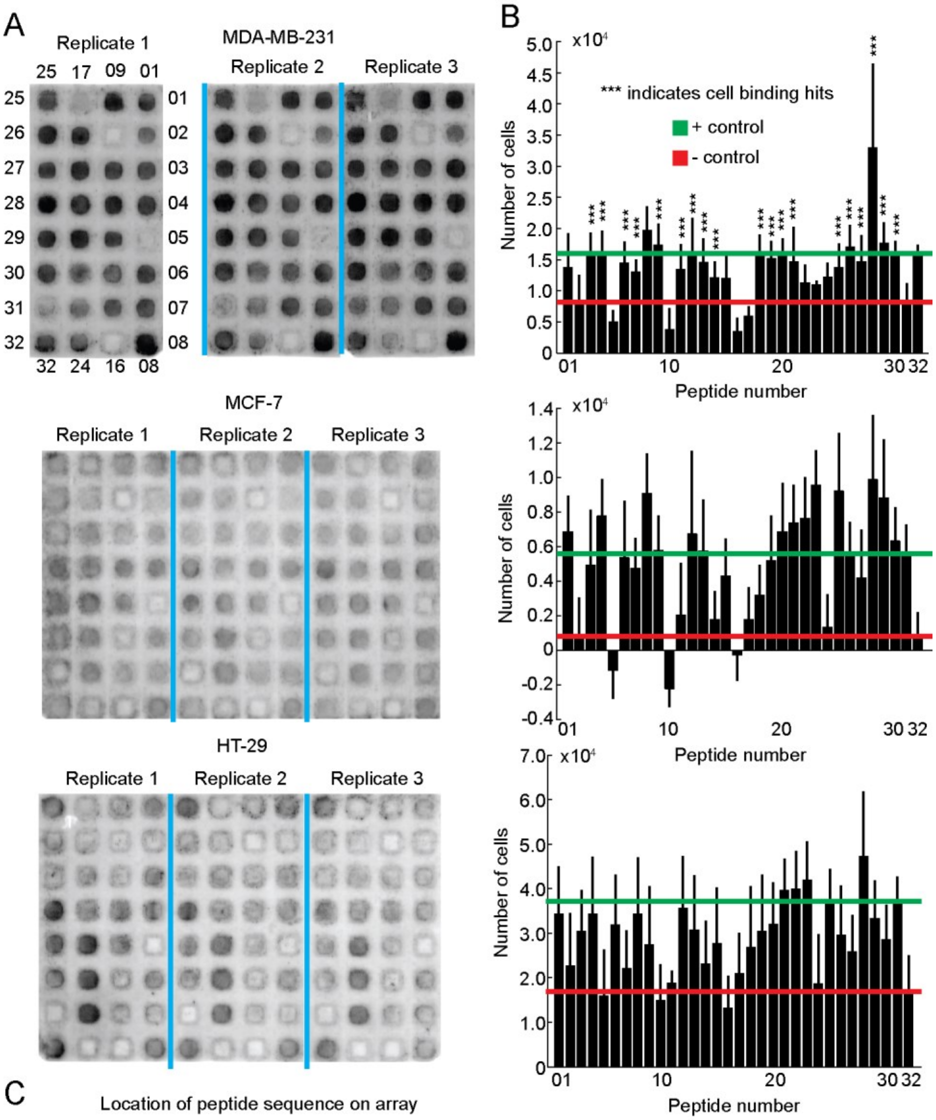
ID	Sequence	ID	Sequence	ID	Sequence	ID	Sequence
01	YAAHRSH	09	TWKFSPL	17	TWYFGPL	25	SNMTRWH
02	QALSVYR	10	TYKYYPL	18	SVLLPHR	26	HSTKVAF
03	ANTTPRH	11	TYQYGKL	19	QVLLTAA	27	APRTFNQ
04	HFRSGSL	12	TYRFLPL	20	STAMDGR	28	SHHQKPP
05	SSLPLRK	13	ALAHRIL	21	GWRTTWP	29	GRLDTGI
06	SWKFGPL	14	ALEVTFW	22	QHMPLTR	30	ALQPQKH
07	HWKFGIL	15	QTGYATR	23	LPVRLDW	31	GRGDSAA
08	TFKFGPL	16	TVRHLQL	24	QFTQLHQ	32	GGRDSAA

Figure 5.7. (A) Representative images of the **first** set of cell-binding hits tested for cell-adhesion with MDA-MB-231, MCF-7, and HT-29 cells stained with Cell Tracker Green. Cell lines were stained for one hour with 4 μM Cell Tracker green in MEM media prior to use in the cell adhesion assay. Labeled peptides (***) indicate hits binding to cells significantly more than the negative control

(GGRDSAA). (B) Plots of the average number of cells adhering to different peptides on the peptide array. The number of cells was extrapolated from a standard curve (Appendix 1.21) for each cell line. Data represent an average from 6 experiments; error bar is 1 standard deviation. Cell-binding hit peptides are determined as binding significantly more ($p < 0.05$) cells than the negative control (GGRDS, red line). (C) Location of each peptide sequence on the peptide arrays.

cell-binding peptides varied from 20 to 40% in BA-screens and up to 60% in EmA-screens (Figure 5.6 C-D).

Figure 5.6 E-G maps the origin of the confirmed cell-binding ligands selected from BA and EmA-screens in two independent lots of the Ph.D.-7 library. 75% of cell-binding hits from the BA-screen reside in the P-population and only 6% in the I-population. In contrast, 75% of hits from the EmA-screen were in the I-population (Figure 5.6 E-F). Comparison of hits that originated from lot #2 or lot #1 (Figure 5.6 G) demonstrated that EmA-screen starting from two lots of library yielded common hits (see overlapping symbols in Figure 5.6 G). In contrast, BA-screens that used lot #1 and lot #2 were completely non-overlapping; the hits from these screens were localized to their respective P-populations (P1 and P2) (Figure 5.6 E-F). Traditional BA-screens, thus, are very ineffective in searching for hits that reside outside of the P- and V-populations, which encompass abundant and fast amplifying library members (Figure 5.6 G). Simple



C

Location of peptide sequence on array

ID	Sequence	ID	Sequence	ID	Sequence	ID	Sequence
01	NAYGGRI	09	ESRVMSR	17	AFTTSYM	25	HWKYWPL
02	TKTVLER	10	AFPPVTA	18	STASYTR	26	AMSSRSL
03	WNQRATG	11	MTVQRGP	19	HAIYPRH	27	TPMTRAL
04	LPGNRLL	12	QLPFTIK	20	GVKALST	28	ARAVLQL
05	QLYREFN	13	FSLQTTR	21	SWQYGKL	29	AWSAVMR
06	YRNHVTY	14	GLRNPPS	22	TYKFGTL	30	IPAPLRS
07	HSRAPER	15	TEKFRVT	23	TYRFGPL	31	GGRDSAA
08	FHRRIKA	16	MQLPLAT	24	VRPMPLQ	32	GRGDSAA

Figure 5.8. (A) Representative images of the **second** set of cell-binding hits tested for cell-adhesion with MDA-MB-231, MCF-7, and HT-29 cells stained with Cell Tracker Green. Cell lines were stained for one hour with 4 μM Cell Tracker green in MEM media prior to use in the cell adhesion assay. Labeled peptides (***) indicate hits binding to cells significantly more than the negative control

(GGRDSAA). (B) Plots of the average number of cells adhering to different peptides on the peptide array. The number of cells was extrapolated from a standard curve (Appendix 1.21) for each cell line. Data represent an average from 6 experiments; error bar is 1 standard deviation. Cell-binding hit peptides are determined as binding significantly more ($p < 0.05$) cells than the negative control (GGRDS, red line). (C) Location of each peptide sequence on the peptide arrays.

incorporation of EmA into selection expanded the diversity space accessible to a panning procedure within a single library to a broader set of cell-binding sequences.

5.2.3 Selection against different receptors

As a final step, we sought to demonstrate that the peptide hits bind to different receptors on MDA-MB-231 cells. The most accurate method for receptor identification is pull-down and proteomic analysis. This method, when applied to over 50 distinct peptides can be resource demanding. We selected another approach and tested whether peptides, support binding to a related epithelial breast cancer line (MCF-7) and a distally related epithelial colon cancer cell line (HT-29). We defined, MCF-7 and HT-29 binding peptides as those that bound significantly more ($p \leq 0.05$) cells than the negative control (GGRDS) (Figure 5.7-5.9). Out of 36 tested peptides, 30 supported adhesion of all three cell lines, four peptides supported adhesion of MDA-MB-231 and MCF-7 cells, but

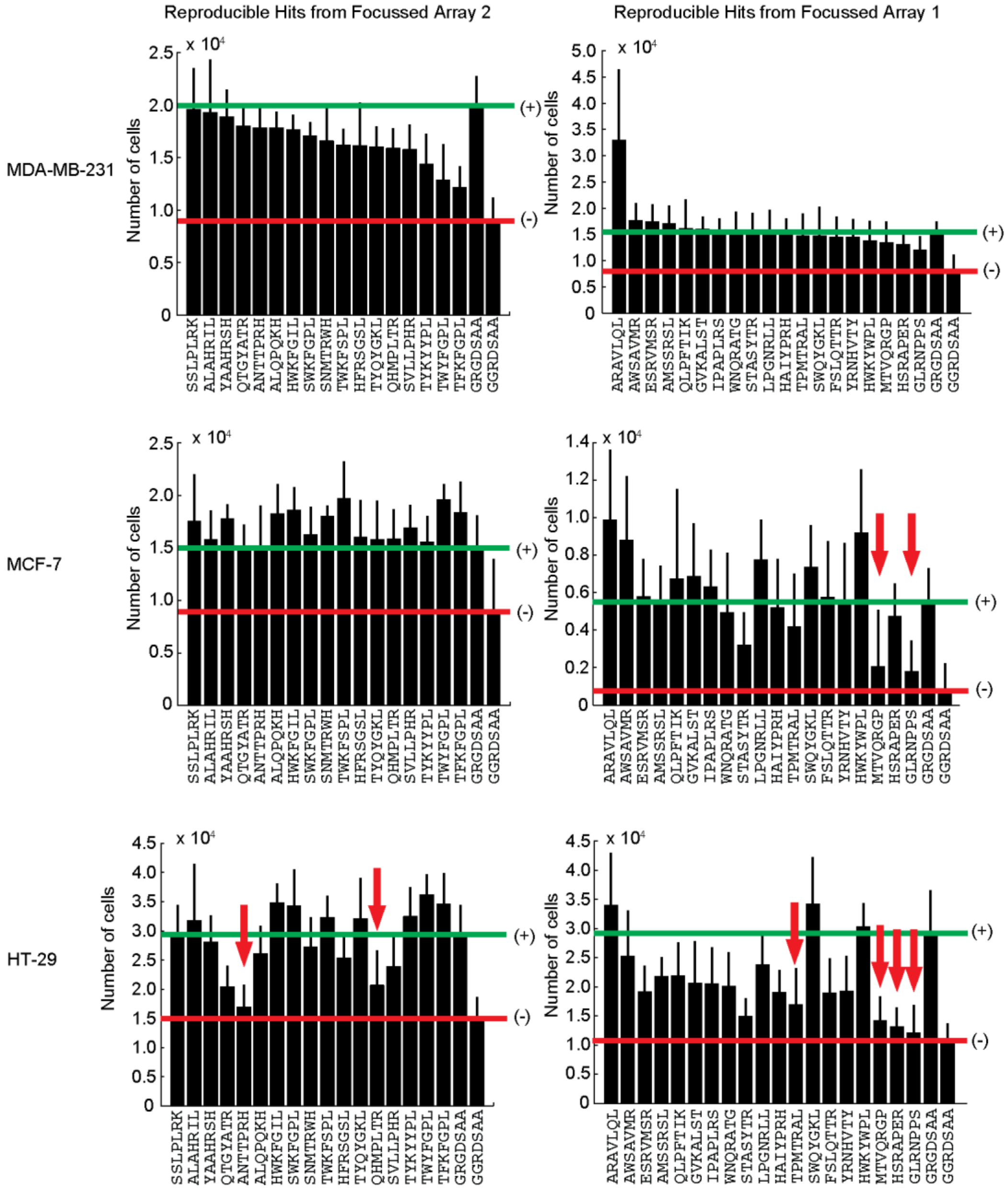


Figure 5.9. Comparison between MDA-MB-231, MCF-7, and HT-29 cell binding to reproducible hits. The reproducible cell-binding hits against MDA-MB-231 are ordered for binding the most to least number of cells. The positive (G*RGD*SAA, highlighted by green lines) and negative (G*GRD*SAA, highlighted by red lines) controls are placed at the end of each focused array. Binding of MCF-7 and HT-29 to the same order of peptides. Red arrows indicate peptides that bind

significantly less cells than the negative control. MCF-7 binds to 34, and HT-29 binds to 29 out of 36 reproducible peptide hits.

not HT-29, and two peptides were specific for MDA-MB-231, but not to MCF-7 or HT-29 cells (Figure 5.6I and Figure 5.7-5.9). Importantly, some of these MDA-MB-231 specific peptides (QHMPLTR and TPMTRAL) originated from EmA-screen and they would not be discovered by the classical BA approach. We envision that the expanded populations of cell-binding clones will be instrumental in addressing different types of receptors on the surface of a specific cell type (here, MDA-MB-231) and fuel subsequent phenotypic screens such as screens for peptide sequences that control differentiation in stem cells[12,157,160].

5.3 Conclusion

An average cell contains several thousand molecularly distinct receptors. Deep-sequencing identified ~2,000 potential cell-binding ligands from a single panning experiment (209 peptides were tested to yield 56 cell-binding hits). As these ligands target different receptors, it is theoretically possible to discover ligands for a large fraction of the receptors in a single screen. Our report, however, uncovers fundamental limitations of phage display selection using conventional bulk amplification (BA). Due to growth bias, the cell binding hits identified by a conventional panning originate predominantly from a small sub-population of the library that comprises <0.0001% of the available diversity. Introducing one subtle change in the screen–replacing bulk with emulsion amplification–enabled the discovery of an expanded set of binding ligands that were lost in the conventional screen. We anticipate that EmA can serve as a general technique for expanding the accessible diversity in screens against multi-site targets (cells, organs, mixtures of antibodies isolated from serum).

5.4 Materials and methods

5.4.1 Isolation and preparation of DNA from phage libraries for Next-Generation Sequencing.

Phage cDNA was isolated from phage libraries using QIAprep Spin M13 kit. Isolated phage cDNA was subjected to PCR amplification with primers flanking the variable region. To avoid a second round of PCR amplification, the primers contained Ion Torrent or Illumina adapters at the 5’ ends. Ion Torrent (Life Technologies) was used to sequence phage libraries obtained from selections. The cDNA was amplified using the following primers:

Forward primer: 5’-CCTCTCTATGGGCAGTCGGTGATCCTTTCTATTCTCACTCT

Reverse primer: 5’-CCATCTCATCCCTCGCTGTCTCCGACTCAG$\mathbf{N}_8$CCGAACCTCCACC

Illumina was used to sequence naïve phage libraries (lot #1 and #2). The ssDNA was amplified using the following primers:

Forward primer: 5’-

CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT$\mathbf{N}_4$CCTTTCTATTCTCACTCT, and Reverse primer: 5'-

AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCG

ATCT**N_4**ACAGTTTCGGCCGA,

where N_X denotes the barcode sequence. The temperature cycling protocol was as follows: 95 °C for 30 s, followed by 25 cycles of 95 °C for 10 s, 60.5 °C for 15 s and 72 °C for 30 s, and then a final extension at 72 °C for 5 min before holding at 4°C. The concentration of the PCR fragments that resulted from amplification of phage libraries was determined by analytical gel (2 % w/v agarose gel in TBE buffer) using a low molecular weight DNA ladder as a standard (NEB, Cat# N3233S). Multiple PCR products amplified with different barcoded primers were pooled and purified using E-gel with a SizeSelect 2 % gel (Invitrogen). The fragments were extracted with RNAse free water. The concentration of purified product was determined using a Qubit Fluorimeter (Invitrogen). PCR amplification of the phage library region and data processing is further described in our previous publication[17].

5.4.2 Next generation sequencing of the library.

Prior to Ion Torrent sequencing, the library was clonally amplified on Ion Sphere Particles (ISPs). The dsDNA fragments (3 pmol) were ligated onto ISPs and amplified by emulsion PCR according to Ion Torrent protocol. The concentration of ISPs with ligated dsDNA fragments after emulsion PCR was determined using Qubit Fluorimeter according to manufacturer's protocol. The ISPs with ligated dsDNA fragments were enriched for and loaded on an Ion 316 chip. The sequencing was performed using an Ion Torrent system (Life Technologies) with an Ion OneTouch 200 Template Kit. Ion Torrent sequencing was performed at the Molecular Biology Service Unit at the University of Alberta. The Donnelly Sequencing Center at the University of Toronto performed Illumina sequencing. Processing of deep-sequencing data and statistical analysis used for prospective identification of the parasite population was performed as described in our previous report[17].

5.4.3 Cell culture

Cells were cultured in Minimum Essential Media (MEM) (HyClone) supplemented with 10 % fetal bovine serum (FBS) (HyClone), 1% Non-essential amino acids (HyClone), and 1 % GlutaMAX (Gibco) at 37 °C in a 5 % CO_2 incubator. Cells were passaged every 2-3 days using trypsin (HyClone). All culture reagents were acquired from Thermo Scientific.

5.4.4 Cell panning

A commercially available library of M13KE phage displaying a random 7-mer peptide on the pIII protein was used in all panning experiments (Ph.D.^{TM}-7 kit, New England Biolabs, complexity of 1.1x10^9 individual clones, Lot #0061101and Lot #0081212). An aliquot of the Ph.D.-7^{TM} library (1 μL from 10^{13} pfu/mL stock) was combined with 100 μL of CM-HBS-3% BSA (1 mM $CaCl_2$, 1 mM $MgCl_2$, 150 mM NaCl, 50 mM HEPES, 3% BSA, pH=7.0) and incubated for 30 min at 4 °C. A suspension of ~10^7 live MDA-MB-231-GFP cells in 100 μL of

CM-HBS-3% BSA was combined with the phage solution and the mixture was rocked on a rotisserie for 1 h at 4 °C. The cell-phage suspension was centrifuged for 5 min at 2,000 rpm, and the supernatant removed. To remove unbound phage, 5-7 rounds of washing were performed. Each round of washing involved the following steps: 1) re-suspending the cell pellet in 50 mL of CM-HBS-1% BSA, 2) incubating for 5 min on ice, 3) centrifugation for 5 min at 2,000 rpm, and 4) discarding the supernatant. To elute the bound phage, the cell pellet was re-suspended in 200 μL of "Elution Buffer" (0.2 M Glycine-HCl, 0.1% BSA, pH=2.2) for no more than 10 min. 30 μL of "Neutralization Buffer" (1 M Tris-HCl, pH=9.1) was added to neutralize the solution and prevent loss of phage viability. 3-5 μL from each wash and elution solutions were used to determine the phage titer (Figure 5.2). The eluted phage was amplified either in emulsion or in bulk solutions. The input and output titers were monitored in all selection procedures and they are summarized in Figure 5.2.

5.4.5 Amplification of phage library in common solution

The eluted phage solution was combined with 25 mL of Lysogeny broth (LB) and 250 μL of *Escherichia coli* K12 ER2738 (Ph.D.-7 kit, New England Biolabs) in log phase growth. The phage were amplified for 4.5 h at 37 °C at ~225 rpm. The culture broth was centrifuged at 5,000 rpm for 15 min at 4 °C and the supernatant was combined with 12 mL of PEG/NaCl solution (15% (w/v) $PEG_{8,000}$, 1 M NaCl) and incubated for 2 h on ice to precipitate the phage. The phage was pelleted by centrifugation at 14,000 rpm for 15 min at 4 °C, and the pellet was re-suspended in 1 mL of CM-HBS-3% BSA buffer for further rounds of panning, Ion Torrent sequencing, and titering.

5.4.6 Amplification of phage library in emulsions

The phage solution and 115 μL of *Escherichia coli* K12 ER2738 in log phase growth were combined with LB to a final volume of 3 mL and the mixture was emulsified using a microfluidics flow-focusing device as previously described[17,99]. The emulsion was incubated for 4.5 h at 37 °C at ~40 rpm. The emulsion was destabilized using 0.5% Krytox in HFE-7100 to combine all amplified phage clones into the aqueous layer. The aqueous-perfluoro solution was centrifuged at 14,000 rpm for 2 min. The aqueous layer was removed and combined with PEG/NaCl in a 1:1 ratio. The solution was incubated for 2 h on ice to precipitate the phage. The phage solution was pelleted at 14,000 rpm for 15 min at 4 °C, and the pellet was re-suspended in 100 μL of CM-HBS-3% BSA buffer for further rounds of panning, Ion Torrent sequencing, and titering.

5.4.7 Adhesion studies on peptide arrays on cellulose support (paper)

Arrays were synthesized as described in our previous publication[159]. The peptide functionalized on paper was soaked in MilliQ H_2O for 30 min in a Nunc Omni-Tray. The paper was then washed twice with 13 mL of MEM media, followed by two washes with 13 mL of MEM media (2 x 5 min at 45 rpm). A custom made insert (for design of insert, see Appendix 1.30) was added to hold the paper submerged and the paper and insert were washed twice with 13 mL of binding-media (0.5 % BSA-MEM media). A suspension of live MDA-MB-231-GFP cells (0.3 × 10^5 cells/mL) in 25 mL of binding media was added to the array and incubated for 3 h at 37 °C in a CO_2 incubator. The array was subsequently washed with MEM (3 x 13 mL) and imaged using a fluorescent gel scanner (GE Healthcare, Typhoon FLA9500) and a confocal fluorescent microscope (Zeiss LSM 700).

5.4.8 Quantification of peptides on modified cellulose support

Three replicates of a peptide array were treated with 50% TFA:DCM for 5 min to remove cells after cell-adhesion assay. The arrays were washed with DCM (3 x 10 mL), MeOH (3 x 10 mL) and air-dried. The middle of each peptide zone was punched out and treated with NH_3 gas overnight. The peptides were dissolved in 50 μL of H_2O and subjected to LC-MS. The amount of each peptide was determined by comparing the peak areas of each peptide to the peak area of an internal standard peptide.

Chapter 6 Conclusion

6.1 Summary of the thesis

Selection using phage display peptide libraries carries a growth bias that is independent to any binding preferences of polypeptides displayed on phage clones. This growth bias limits the identification of target-binding ligands to a ~0.0001% sub-population of the library diversity. The focus of this thesis was to prevent growth bias and expand selection of ligands from a larger diversity space of a phage display library.

To overcome amplification bias we use emulsion amplification to uniformly amplify phage-displayed libraries of peptides. In Chapter 2, we described the synthesis of a perfluoro-surfactant used to maintain stability of emulsions during amplification, and protocols to emulsify phage libraries and recover them after amplification. In Chapter 3, we developed software and protocols to analyze deep-sequencing data. In Chapter 4 we used deep-sequencing to demonstrate that emulsion amplification can prevent diversity collapse in phage libraries. Lastly, in Chapter 5, we incorporated emulsion amplification into phage display library selection against breast cancer cells and identified new binding ligands from a greater diversity space. These ligands would not have been able to be discovered using the standard selection method.

Ligands that are lost due to conventional selection, using bulk amplification, may be more important than those that are selected. In selections for therapeutic candidates, phage display library selections are most frequently screened against a single antigen in order to obtain a neutralizing ligand. However, the ligand with the highest affinity may be irrelevant because it binds to a region of the antigen that is not a target for neutralization. Additionally, any therapeutic advantage may be lost in selections against a single target. The physiological environment, which more accurately represents the disease state in which the target is located in, is lost and may be more relevant for selection. Therefore, it will be essential to perform selections against multi-site targets, such as cells, in order to identify multiple and diverse ligands that can give rise to therapeutic candidates or targeting probes.

Ligands that are discovered as diagnostic and therapeutic agents can also be used as ligands that can induce cellular effects. Given that these ligands bind to receptors, they have the potential to induce a phenotypic response. However, there is no guarantee the discovered ligands will trigger the cellular response of interest. Phenotypic screens are performed to identify usable ligands that initiate phenotypes in cancer cells such as, arrest of migration, proliferation, angiogenesis, and viability. Current and future work in our group is focussed on phenotypic screens to arrest division of cancer cells to cancer stem cells. We believe that emulsion amplification can improve *in vitro* selection and enable the discovery of functional ligands for use as materials that can control stem cell differentiation and self-renewal in the drug discovery field.

6.2 Future directions

Discovery of many binding ligands can aid us in identifying novel ligands that allow for further investigation into asymmetric division in cancer cells. Asymmetric division is a process in which a cell divides asymmetrically into two different daughter cells. Normally, asymmetric division allows stem cells to differentiate and maintain tissue homeostasis, however, defects in asymmetric division give rise to cancer and creates a small population of cancer cells that can form tumors[161,162]. Most of the proteins regulating this division were identified by genetic knock-out approaches that lead to defects in asymmetric division[163,164]. We intend to use an opposite approach, in which asymmetric distribution of the protein is induced. We propose controlling the location of the components of the cell membrane by growing cancer cells on self-assembled monolayers (SAMs) of peptides identified from these screens. We will validate increases or decreases in cancer stem cell populations via phenotypic screens. We will use breast cancer cell lines MDA-MB-231 and MCF-7 as models systems. We are developing three phenotypic assays: 1) We will monitor the population of CD24-/CD44+ cancer cells using flow cytometry. An increase in this population has been linked to the cancer stem cell phenotype[165]. 2) We will monitor mammosphere forming capacity. 3) Finally, we will use RT-PCR to monitor gene expression of E-cadherin, smooth muscle actin (SMA), Vimentin, and SMAD proteins[166]. This discovery-based approach might identify new molecules and materials that regulate differentiation of cells and provide new targets for development of therapeutics for cancer and regenerative medicine.

References

(1) Smith, G. P. *Science* **1985**, *228*, 1315.

(2) Smith, G. P.; Petrenko, V. A. *Chem. Rev.* **1997**, *97*, 391.

(3) Kehoe, J. W.; Kay, B. K. *Chem. Rev.* **2005**, *105*, 4056.

(4) Funke, S. A.; Willbold, D. *Mol. Biosyst.* **2009**, *5*, 783.

(5) Deutscher, S. L. *Chem. Rev.* **2010**, *110*, 3196.

(6) Baneyx, F.; Schwartz, D. T. *Curr. Opin. Biotechnol.* **2007**, *18*, 312.

(7) Ryvkin, A.; Ashkenazy, H.; Smelyanski, L.; Kaplan, G.; Penn, O.; Weiss-Ottolenghi, Y.; Privman, E.; Ngam, P. B.; Woodward, J. E.; May, G. D.; Bell, C.; Pupko, T.; Gershoni, J. M. *Plos One* **2012**, *7*.

(8) Li, J. W.; Feng, L.; Jiang, X. G. *Amino Acids* **2015**, *47*, 401.

(9) Ngubane, N. A. C.; Gresh, L.; Ioerger, T. R.; Sacchettini, J. C.; Zhang, Y. J. J.; Rubin, E. J.; Pym, A.; Khati, M. *Plos One* **2013**, *8*.

(10) Derda, R.; Tang, S. K. Y.; Li, S. C.; Ng, S.; Matochko, W.; Jafari, M. R. *Molecules* **2011**, *16*, 1776.

(11) Hu, W. G.; Jager, S.; Chau, D.; Mah, D.; Nagata, L. P. *Appl. Biochem. Biotechnol.* **2010**, *160*, 1206.

(12) Derda, R.; Musah, S.; Orner, B. P.; Klim, J. R.; Li, L.; Kiessling, L. L. *J. Am. Chem. Soc.* **2010**, *132*, 1289.

(13) Ravn, U.; Didelot, G.; Venet, S.; Ng, K. T.; Gueneau, F.; Rousseau, F.; Calloud, S.; Kosco-Vilbois, M.; Fischer, N. *Methods* **2013**, *60*, 99.

(14) Serizawa, T.; Sawada, T.; Kitayama, T. *Angew. Chem. Int. Ed.* **2007**, *46*, 723.

(15) Gearhart, D. A.; Toole, P. F.; Beach, J. W. *Neurosci. Res.* **2002**, *44*, 255.

(16) Caprini, A.; Silva, D.; Zanoni, I.; Cunha, C.; Volonte, C.; Vescovi, A.; Gelain, F. *New Biotechnol* **2013**, *30*, 552.

(17) Matochko, W. L.; Li, S. C.; Tang, S. K. Y.; Derda, R. *Nucleic Acids Res.* **2014**, *42*, 1784.

(18) Nguyen, K. T. H.; Adamkiewicz, M. A.; Hebert, L. E.; Zygiel, E. M.; Boyle, H. R.; Martone, C. M.; Melendez-Rios, C. B.; Noren, K. A.; Noren, C. J.; Hall, M. F. *Anal. Biochem.* **2014**, *462*, 35.

(19) Tawfik, D. S.; Griffiths, A. D. *Nat. Biotechnol.* **1998**, *16*, 652.

(20) Dressman, D.; Yan, H.; Traverso, G.; Kinzler, K. W.; Vogelstein, B. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 8817.

(21) Williams, R.; Peisajovich, S. G.; Miller, O. J.; Magdassi, S.; Tawfik, D. S.; Griffiths, A. D. *Nat. Methods* **2006**, *3*, 545.

(22) Vogelstein, B.; Kinzler, K. W. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 9236.

(23) Boedicker, J. Q.; Li, L.; Kline, T. R.; Ismagilov, R. F. *Lab Chip* **2008**, *8*, 1265.

(24) Marcoux, P. R.; Dupoy, M.; Mathey, R.; Novelli-Rousseau, A.; Heran, V.; Morales, S.; Rivera, F.; Joly, P. L.; Moy, J. P.; Mallard, F. *Colloids Surf., A* **2011**, *377*, 54.

(25) Clausell-Tormos, J.; Lieber, D.; Baret, J. C.; El-Harrak, A.; Miller, O. J.; Frenz, L.; Blouwolff, J.; Humphry, K. J.; Koster, S.; Duan, H.; Holtze, C.; Weitz, D. A.; Griffiths, A. D.; Merten, C. A. *Chem. Biol.* **2008**, *15*, 427.

(26) Du, G. S.; Pan, J. Z.; Zhao, S. P.; Zhu, Y.; den Toonder, J. M. J.; Fang, Q. *Anal. Chem.* **2013**, *85*, 6740.

(27) Lee, Y. Y.; Narayanan, K.; Gao, S. J.; Ying, J. Y. *Nano Today* **2012**, *7*, 29.

(28) Huang, R. H.; Fang, P. T.; Kay, B. K. *Methods* **2012**, *58*, 10.

(29) Devlin, J. J.; Panganiban, L. C.; Devlin, P. E. *Science* **1990**, *249*, 404.

(30) Menendez, A.; Scott, J. K. *Anal. Biochem.* **2005**, *336*, 145.

(31) Bonnycastle, L. L. C.; Mehroke, J. S.; Rashed, M.; Gong, X.; Scott, J. K. *J. Mol. Biol.* **1996**, *258*, 747.

(32) Umlauf, B. J.; Mercedes, J. S.; Chung, C.-Y.; Brown, K. C. *Bioconj. Chem.* **2014**, *25*, 1829.

(33) Kuzmicheva, G. A.; Jayanna, P. K.; Sorokulova, I. B.; Petrenko, V. A. *Protein Eng. Des. Sel.* **2009**, *22*, 9.

(34) Rodi, D. J.; Soares, A. S.; Makowski, L. *J. Mol. Biol.* **2002**, *322*, 1039.

(35) Brammer, L. A.; Bolduc, B.; Kass, J. L.; Felice, K. M.; Noren, C. J.; Hall, M. F. *Anal. Biochem.* **2008**, *373*, 88.

(36) Thomas, W. D.; Golomb, M.; Smith, G. P. *Anal. Biochem.* **2010**, *407*, 237.

(37) Malik, P.; Tarry, T. D.; Gowda, L. R.; Langara, A.; Petukhov, S. A.; Symmons, M. F.; Welsh, L. C.; Marvin, D. A.; Perham, R. N. *J. Mol. Biol.* **1996**, *260*, 9.

(38) Iannolo, G.; Minenkova, O.; Petruzzelli, R.; Cesareni, G. *J. Mol. Biol.* **1995**, *248*, 835.

(39) Legendre, D.; Fastrez, J. *Gene* **2002**, *290*, 203.

(40) Petrenko, V. A.; Smith, G. P.; Gong, X.; Quinn, T. *Protein Eng.* **1996**, *9*, 797.

(41) Iannolo, G.; Minenkova, O.; Gonfloni, S.; Castagnoli, L.; Cesareni, G. *Biol. Chem.* **1997**, *378*, 517.

(42) Li, Z. P.; Koch, H.; Dubel, S. *J. Mol. Microbiol. Biotechnol.* **2003**, *6*, 57.

(43) Barbas, C. F.; Bain, J. D.; Hoekstra, D. M.; Lerner, R. A. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 4457.

(44) Breitling, F.; Dubel, S.; Seehaus, T.; Klewinghaus, I.; Little, M. *Gene* **1991**, *104*, 147.

(45) Bass, S.; Greene, R.; Wells, J. A. *Proteins: Struct., Funct., Genet.* **1990**, *8*, 309.

(46) Marks, J. D.; Hoogenboom, H. R.; Bonnert, T. P.; McCafferty, J.; Griffiths, A. D.; Winter, G. *J. Mol. Biol.* **1991**, *222*, 581.

(47) Lofblom, J.; Feldwisch, J.; Tolmachev, V.; Carlsson, J.; Stahl, S.; Frejd, F. Y. *FEBS Lett.* **2010**, *584*, 2670.

(48) Kelly, K. A.; Carson, J.; McCarthy, J. R.; Weissleder, R. *Plos One* **2007**, *2*.

(49) Li, M.; Duc, A.-C. E.; Klosi, E.; Pattabiraman, S.; Spaller, M. R.; Chow, C. S. *Biochemistry* **2009**, *48*, 8299.

(50) Ma, C.; Yin, G.; Yan, D.; He, X.; Zhang, L.; Wei, Y.; Huang, Z. *J. Pept. Sci.* **2013**, *19*, 730.

(51) Scott, J. K.; Smith, G. P. *Science* **1990**, *249*, 386.

(52) Roth, T. A.; Weiss, G. A.; Eigenbrot, C.; Sidhu, S. S. *J. Mol. Biol.* **2002**, *322*, 357.

(53) Karasseva, N. G.; Glinsky, V. V.; Chen, N. X.; Komatireddy, R.; Quinn, T. P. *J. Protein Chem.* **2002**, *21*, 287.

(54) Ng, S.; Lin, E.; Kitov, P. I.; Tjhung, K. F.; Gerlits, O. O.; Deng, L.; Kasper, B.; Sood, A.; Paschal, B. M.; Zhang, P.; Ling, C.-C.; Klassen, J. S.; Noren, C. J.; Mahal, L. K.; Woods, R. J.; Coates, L.; Derda, R. *J. Am. Chem. Soc.* **2015**.

(55) Rebollo, I. R.; Sabisz, M.; Baeriswyl, V.; Heinis, C. *Nucleic Acids Res.* **2014**, *42*.

(56) Mannocci, L.; Zhang, Y. X.; Scheuermann, J.; Leimbacher, M.; De Bellis, G.; Rizzi, E.; Dumelin, C.; Melkko, S.; Neri, D. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 17670.

(57) Glanville, J.; Zhai, W. W.; Berka, J.; Telman, D.; Huerta, G.; Mehta, G. R.; Ni, I.; Mei, L.; Sundar, P. D.; Day, G. M. R.; Cox, D.; Rajpal, A.; Pons, J. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 20216.

(58) Ernst, A.; Gfeller, D.; Kan, Z. Y.; Seshagiri, S.; Kim, P. M.; Bader, G. D.; Sidhu, S. S. *Mol. Biosyst.* **2010**, *6*, 1782.

(59) Lee, S. M.; Lee, E. J.; Hong, H. Y.; Kwon, M. K.; Kwon, T. H.; Choi, J. Y.; Park, R. W.; Kwon, T. G.; Yoo, E. S.; Yoon, G. S.; Kim, I. S.; Ruoslahti, E.; Lee, B. H. *Mol. Cancer Res.* **2007**, *5*, 11.

(60) Buller, F.; Steiner, M.; Scheuermann, J.; Mannocci, L.; Nissen, I.; Kohler, M.; Beisel, C.; Neri, D. *Bioorg. Med. Chem. Lett.* **2010**, *20*, 4188.

(61) Livnah, O.; Stura, E. A.; Johnson, D. L.; Middleton, S. A.; Mulcahy, L. S.; Wrighton, N. C.; Dower, W. J.; Jolliffe, L. K.; Wilson, I. A. *Science* **1996**, *273*, 464.

(62) McGregor, D. P. *Curr. Opin. Pharmacol.* **2008**, *8*, 616.

(63) Olson, C. A.; Nie, J.; Diep, J.; Al-Shyoukh, I.; Takahashi, T. T.; Al-Mawsawi, L. Q.; Bolin, J. M.; Elwell, A. L.; Swanson, S.; Stewart, R.; Thomson, J. A.; Soh, H. T.; Roberts, R. W.; Sun, R. *Angew. Chem. Int. Ed.* **2012**, *51*, 12449.

(64) Matochko, W. L.; Chu, K. K.; Jin, B. J.; Lee, S. W.; Whitesides, G. M.; Derda, R. *Methods* **2012**, *58*, 47.

(65) Rajotte, D.; Arap, W.; Hagedorn, M.; Koivunen, E.; Pasqualini, R.; Ruoslahti, E. *J. Clin. Invest.* **1998**, *102*, 430.

(66) t Hoen, P. A. C.; Jirka, S. M. G.; ten Broeke, B. R.; Schultes, E. A.; Aguilera, B.; Pang, K. H.; Heemskerk, H.; Aartsma-Rus, A.; van Ommen, G. J.; den Dunnen, J. T. *Anal. Biochem.* **2012**, *421*, 622.

(67) Dias-Neto, E.; Nunes, D. N.; Giordano, R. J.; Sun, J.; Botz, G. H.; Yang, K.; Setubal, J. C.; Pasqualini, R.; Arap, W. *Plos One* **2009**, *4*.

(68) Glenn, T. C. *Mol. Ecol. Resour.* **2011**, *11*, 759.

(69) Wang, F. Y.; Zhang, T. Y.; Luo, J. X.; He, G. A.; Gu, Q. L.; Xiao, F. *Biosci. Biotechnol., Biochem.* **2006**, *70*, 2035.

(70) Quail, M. A.; Smith, M.; Coupland, P.; Otto, T. D.; Harris, S. R.; Connor, T. R.; Bertoni, A.; Swerdlow, H. P.; Gu, Y. *Bmc Genomics* **2012**, *13*.

(71) Loman, N. J.; Misra, R. V.; Dallman, T. J.; Constantinidou, C.; Gharbia, S. E.; Wain, J.; Pallen, M. J. *Nat. Biotechnol.* **2012**, *30*, 434.

(72) Bragg, L. M.; Stone, G.; Butler, M. K.; Hugenholtz, P.; Tyson, G. W. *PLoS Comput. Biol.* **2013**, *9*.

(73) Ivarsson, Y.; Arnold, R.; McLaughlin, M.; Nim, S.; Joshi, R.; Ray, D.; Liu, B.; Teyra, J.; Pawson, T.; Moffat, J.; Li, S. S. C.; Sidhu, S. S.; Kim, P. M. *Proc. Natl. Acad. Sci. U.S.A.* **2014**, *111*, 2542.

(74) Birnbaum, M. E.; Mendoza, J. L.; Sethi, D. K.; Dong, S.; Glanville, J.; Dobbins, J.; Ozkan, E.; Davis, M. M.; Wucherpfennig, K. W.; Garcia, K. C. *Cell* **2014**, *157*, 1073.

(75) Duchrow, T.; Shtatland, T.; Guettler, D.; Pivovarov, M.; Kramer, S.; Weissleder, R. *BMC Bioinf.* **2009**, *10*.

(76) Huang, J.; Ru, B. B.; Li, S. Y.; Lin, H.; Guo, F. B. *J. Biomed. Biotechnol.* **2010**.

(77) Serizawa, T.; Techawanitchai, P.; Matsuno, H. *Chembiochem* **2007**, *8*, 989.

(78) Ernst, A.; Sazinsky, S. L.; Hui, S.; Currell, B.; Dharsee, M.; Seshagiri, S.; Bader, G. D.; Sidhu, S. S. *Sci. Signal.* **2009**, *2*.

(79) Fowler, D. M.; Araya, C. L.; Fleishman, S. J.; Kellogg, E. H.; Stephany, J. J.; Baker, D.; Fields, S. *Nat. Methods* **2010**, *7*, 741.

(80) Rebollo, I. R.; Angelini, A.; Heinis, C. *Medchemcomm* **2013**, *4*, 145.

(81) Chen, S. Y.; Rebollo, I. R.; Buth, S. A.; Morales-Sanfrutos, J.; Touati, J.; Leiman, P. G.; Heinis, C. *J. Am. Chem. Soc.* **2013**, *135*, 6562.

(82) Baeriswyl, V.; Calzavarini, S.; Gerschheimer, C.; Diderich, P.; Angelillo-Scherrer, A.; Heinis, C. *J. Med. Chem.* **2013**, *56*, 3742.

(83) Bawazer, L. A.; Newman, A. M.; Gu, Q.; Ibish, A.; Arcila, M.; Cooper, J. B.; Meldrum, F. C.; Morse, D. E. *Acs Nano* **2014**, *8*, 387.

(84) Staquicini, F. I.; Cardo-Vila, M.; Kolonin, M. G.; Trepel, M.; Edwards, J. K.; Nunes, D. N.; Sergeeva, A.; Efstathiou, E.; Sun, J.; Almeida, N. F.; Tu, S. M.; Botz, G. H.; Wallace, M. J.; O'Connell, D. J.; Krajewski, S.; Gershenwald, J. E.; Molldrem, J. J.; Flamm, A. L.; Koivunen, E.; Pentz, R. D.; Dias-Neto, E.; Setubal, J. C.; Cahill, D. J.; Troncoso, P.; Do, K. A.; Logothetis, C. J.; Sidman, R. L.; Pasqualini, R.; Arap, W. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 18637.

(85) Ravn, U.; Gueneau, F.; Baerlocher, L.; Osteras, M.; Desmurs, M.; Malinge, P.; Magistrelli, G.; Farinelli, L.; Kosco-Vilbois, M. H.; Fischer, N. *Nucleic Acids Res.* **2010**, *38*.

(86) Ferrara, F.; Naranjo, L. A.; D'Angelo, S.; Kiss, C.; Bradbury, A. R. M. *J. Immunol. Methods* **2013**, *395*, 83.

(87) Saggy, I.; Wine, Y.; Shefet-Carasso, L.; Nahary, L.; Georgiou, G.; Benhar, I. *Protein Eng. Des. Sel.* **2012**, *25*, 539.

(88) Venet, S.; Kosco-Vilbois, M.; Fischer, N. *Mabs* **2013**, *5*, 690.

(89) Zhang, H. K.; Torkamani, A.; Jones, T. M.; Ruiz, D. I.; Pons, J.; Lerner, R. A. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 13456.

(90) Rosander, A.; Guss, B.; Frykberg, L.; Bjorkman, C.; Naslund, K.; Pringle, M. *Vet. Med.* **2011**, *153*, 315.

(91) McLaughlin, M. E.; Sidhu, S. S. In *Methods in Protein Design*; Keating, A. E., Ed. 2013; Vol. 523, p 327.

(92) Scott, B. M.; Matochko, W. L.; Gierczak, R. F.; Bhakta, V.; Derda, R.; Sheffield, W. P. *Plos One* **2014**, *9*.

(93) Zimmermann, B.; Gesell, T.; Chen, D.; Lorenz, C.; Schroeder, R. *Plos One* **2010**, *5*.

(94) Thiel, W. H.; Bair, T.; Thiel, K. W.; Dassie, J. P.; Rockey, W. M.; Howell, C. A.; Liu, X. Y. Y.; Dupuy, A. J.; Huang, L. Y.; Owczarzy, R.; Behlke, M. A.; McNamara, J. O.; Giangrande, P. H. *Nucleic Acid Ther.* **2011**, *21*, 253.

(95) Ge, X.; Mazor, Y.; Hunicke-Smith, S. P.; Ellington, A. D.; Georgiou, G. *Biotechnol. Bioeng.* **2010**, *106*, 347.

(96) Fowler, D. M.; Araya, C. L.; Gerard, W.; Fields, S. *Bioinformatics* **2011**, *27*, 3430.

(97) Kim, T.; Tyndel, M. S.; Huang, H. M.; Sidhu, S. S.; Bader, G. D.; Gfeller, D.; Kim, P. M. *Nucleic Acids Res.* **2012**, *40*.

(98) Davey, N. E.; Edwards, R. J.; Shields, D. C. *BMC Bioinf.* **2010**, *11*.

(99) Derda, R.; Tang, S. K. Y.; Whitesides, G. M. *Angew. Chem. Int. Ed.* **2010**, *49*, 5301.

(100) Rodi, D. J.; Janes, R. W.; Sanganee, H. J.; Holton, R. A.; Wallace, B. A.; Makowski, L. *J. Mol. Biol.* **1999**, *285*, 197.

(101) McConnell, S. J.; Uveges, A. J.; Spinella, D. G. *Biotechniques* **1995**, *18*, 803.

(102) Garstecki, P.; Fuerstman, M. J.; Whitesides, G. M. *Phys. Rev. Lett.* **2005**, *94*.

(103) Garstecki, P.; Gitlin, I.; DiLuzio, W.; Whitesides, G. M.; Kumacheva, E.; Stone, H. A. *Appl. Phys. Lett.* **2004**, *85*, 2649.

(104) Garstecki, P.; Stone, H. A.; Whitesides, G. M. *Phys. Rev. Lett.* **2005**, *94*.

(105) Binkert, A.; Studer, P.; Voros, J. *Small* **2009**, *5*, 1070.

(106) Kobel, S.; Valero, A.; Latt, J.; Renaud, P.; Lutolf, M. *Lab Chip* **2010**, *10*, 857.

(107) Tumarkin, E.; Tzadu, L.; Csaszar, E.; Seo, M.; Zhang, H.; Lee, A.; Peerani, R.; Purpura, K.; Zandstra, P. W.; Kumacheva, E. *Integr. Biol.* **2011**, *3*, 653.

(108) Shao, K. K.; Ding, W. F.; Wang, F.; Li, H. Q.; Ma, D.; Wang, H. M. *Plos One* **2011**, *6*.

(109) Griffiths, A. D.; Tawfik, D. S. *Trends Biotechnol.* **2006**, *24*, 395.

(110) Holtze, C.; Rowat, A. C.; Agresti, J. J.; Hutchison, J. B.; Angile, F. E.; Schmitz, C. H. J.; Koster, S.; Duan, H.; Humphry, K. J.; Scanga, R. A.; Johnson, J. S.; Pisignano, D.; Weitz, D. A. *Lab Chip* **2008**, *8*, 1632.

(111) Liu, W. S.; Kim, H. J.; Lucchetta, E. M.; Du, W. B.; Ismagilov, R. F. *Lab Chip* **2009**, *9*, 2153.

(112) Matochko, W. L.; Ng, S.; Jafari, M. R.; Romaniuk, J.; Tang, S. K. Y.; Derda, R. *Methods* **2012**, *58*, 18.

(113) Marcoux, P. R.; Dupoy, M.; Mathey, R.; Novelli-Rousseau, A.; Heran, V.; Morales, S.; Rivera, F.; Joly, P. L.; Moy, J. P.; Mallard, F. *Colloids Surf. Physicochem. Eng. Aspects* **2011**, *377*, 54.

(114) Link, D. R.; Anna, S. L.; Weitz, D. A.; Stone, H. A. *Phys. Rev. Lett.* **2004**, *92*.

(115) Ahn, K.; Agresti, J.; Chong, H.; Marquez, M.; Weitz, D. A. *Appl. Phys. Lett.* **2006**, *88*.

(116) Tan, W. H.; Takeuchi, S. *Lab Chip* **2006**, *6*, 757.

(117) Li, W.; Greener, J.; Voicu, D.; Kumacheva, E. *Lab Chip* **2009**, *9*, 2715.

(118) Derda, R.; Wherritt, D. J.; Kiessling, L. L. *Langmuir* **2007**, *23*, 11164.

(119) Solvas, X. C. I.; Niu, X. Z.; Leeper, K.; Cho, S.; Chang, S. I.; Edel, J. B.; deMello, A. J. *J. Visualized Exp.* **2011**.

(120) McDonald, J. C.; Whitesides, G. M. *Acc. Chem. Res.* **2002**, *35*, 491.

(121) Makowski, L.; Soares, A. *Bioinformatics* **2003**, *19*, 483.

(122) Makowski, L. *Phage Nanobiotechnology* **2011**, 33.

(123) Huang, J.; Ru, B. B.; Zhu, P.; Nie, F. L.; Yang, J.; Wang, X. Y.; Dai, P.; Lin, H.; Guo, F. B.; Rao, N. N. *Nucleic Acids Res.* **2012**, *40*, D271.

(124) Ru, B. B.; Huang, J. A.; Dai, P.; Li, S. Y.; Xia, Z. K.; Ding, H.; Lin, H.; Guo, F. B.; Wang, X. L. *Molecules* **2010**, *15*, 8279.

(125) Noren, K. A.; Noren, C. J. *Methods* **2001**, *23*, 169.

(126) Bentley, D. R.; Balasubramanian, S.; Swerdlow, H. P.; Smith, G. P.; Milton, J.; Brown, C. G.; Hall, K. P.; Evers, D. J.; Barnes, C. L.; Bignell, H. R.; Boutell, J. M.; Bryant, J.; Carter, R. J.; Cheetham, R. K.; Cox, A. J.; Ellis, D. J.; Flatbush, M. R.; Gormley, N. A.; Humphray, S. J.; Irving, L. J.; Karbelashvili, M. S.; Kirk, S. M.; Li, H.; Liu, X. H.; Maisinger, K. S.; Murray, L. J.; Obradovic, B.; Ost, T.; Parkinson, M. L.; Pratt, M. R.; Rasolonjatovo, I. M. J.; Reed, M. T.; Rigatti, R.; Rodighiero, C.; Ross, M. T.; Sabot, A.; Sankar, S. V.; Scally, A.; Schroth, G. P.; Smith, M. E.; Smith, V. P.; Spiridou, A.; Torrance, P. E.; Tzonev, S. S.; Vermaas, E. H.; Walter, K.; Wu, X. L.; Zhang, L.; Alam, M. D.; Anastasi, C.; Aniebo, I. C.; Bailey, D. M. D.; Bancarz, I. R.; Banerjee, S.; Barbour, S. G.; Baybayan, P. A.; Benoit, V. A.; Benson, K. F.; Bevis, C.; Black, P. J.; Boodhun, A.; Brennan, J. S.; Bridgham, J. A.; Brown, R. C.; Brown, A. A.; Buermann, D. H.; Bundu, A. A.; Burrows, J. C.; Carter, N. P.; Castillo, N.; Catenazzi, M. C. E.; Chang, S.; Cooley, R. N.; Crake, N. R.; Dada, O. O.; Diakoumakos, K. D.; Dominguez-Fernandez, B.; Earnshaw, D. J.; Egbujor, U. C.; Elmore, D. W.; Etchin, S. S.; Ewan, M. R.; Fedurco, M.; Fraser, L. J.; Fajardo, K. V. F.; Furey,

W. S.; George, D.; Gietzen, K. J.; Goddard, C. P.; Golda, G. S.; Granieri, P. A.; Green, D. E.; Gustafson, D. L.; Hansen, N. F.; Harnish, K.; Haudenschild, C. D.; Heyer, N. I.; Hims, M. M.; Ho, J. T.; Horgan, A. M. *Nature* **2008**, *456*, 53.

(127) Kircher, M.; Heyn, P.; Kelso, J. *Bmc Genomics* **2011**, *12*.

(128) Crooks, G. E.; Hon, G.; Chandonia, J. M.; Brenner, S. E. *Genome Res.* **2004**, *14*, 1188.

(129) Lee, S. J.; Lee, J. H.; Kay, B. K.; Dreyfuss, G.; Park, Y. K.; Kim, J. K. *J. Microbiol.* **1997**, *35*, 347.

(130) Marioni, J. C.; Mason, C. E.; Mane, S. M.; Stephens, M.; Gilad, Y. *Genome Res.* **2008**, *18*, 1509.

(131) Balwierz, P. J.; Carninci, P.; Daub, C. O.; Kawai, J.; Hayashizaki, Y.; Van Belle, W.; Beisel, C.; van Nimwegen, E. *Genome Biol.* **2009**, *10*.

(132) Shtatland, T.; Guettler, D.; Kossodo, M.; Pivovarov, M.; Weissleder, R. *BMC Bioinf.* **2007**, *8*.

(133) Heemskerk, J. A.; Van Deutekom, J. C. T.; Van Kuik-Romeijn, P.; Platenburg, G. J.; Prosensa Technologies B V: 2012; Vol. US20120322724 A1.

(134) Kim, S. N.; Kuang, Z. F.; Slocik, J. M.; Jones, S. E.; Cui, Y.; Farmer, B. L.; McAlpine, M. C.; Naik, R. R. *J. Am. Chem. Soc.* **2011**, *133*, 14480.

(135) Llano-Sotelo, B.; Klepacki, D.; Mankin, A. S. *J. Mol. Biol.* **2009**, *391*, 813.

(136) Leclerc, D. 2010; Vol. WO2008058369 A1.

(137) Sawada, T.; Mihara, H. *Mol. Biosyst.* **2012**, *8*, 1264.

(138) DeLano, W. L.; Ultsch, M. H.; de Vos, A. M.; Wells, J. A. *Science* **2000**, *287*, 1279.

(139) Rodi, D. J.; Makowski, L.; Kay, B. K. *Curr. Opin. Chem. Biol.* **2002**, *6*, 92.

(140) Lancet, D.; Sadovsky, E.; Seidemann, E. *Proc. Natl. Acad. Sci. U. S. A.* **1993**, *90*, 3715.

(141) Derda, R.; Tang, S. K. Y.; Whitesides, G. M. *Angew. Chem. Int. Ed.* **2010**, *49*, 5301.

(142) Olson, C. A.; Adams, J. D.; Takahashi, T. T.; Qi, H. F.; Howell, S. M.; Wu, T. T.; Roberts, R. W.; Sun, R.; Soh, H. T. *Angew. Chem. Int. Ed.* **2011**, *50*, 8295.

(143) Matochko, W. L.; Chu, K.; Jin, B.; Lee, S. W.; Whitesides, G. M.; Derda, R. *Methods* **2012**, *58*, 47.

(144) Breaker, R. R.; Joyce, G. F. *Proc. Natl. Acad. Sci. U. S. A.* **1994**, *91*, 6093.

(145) Zimmermann, B.; Gesell, T.; Chen, D.; Lorenz, C.; Schroeder, R. *Plos One* **2010**, *5*, e9169.

(146) Li, W.; Greener, J.; Voicu, D.; Kumacheva, E. *Lab Chip* **2009**, *9*, 2715.

(147) Theberge, A. B.; Courtois, F.; Schaerli, Y.; Fischlechner, M.; Abell, C.; Hollfelder, F.; Huck, W. T. S. *Angew. Chem. Int. Ed.* **2010**, *49*, 5846.

(148) Schneider, T. D.; Stephens, R. M. *Nucleic Acids Res.* **1990**, *18*, 6097.

(149) Ru, B.; Huang, J.; Dai, P.; Li, S.; Xia, Z.; Ding, H.; Lin, H.; Guo, F.-B.; Wang, X. *Molecules* **2010**, *15*, 8279.

(150) Boder, E. T.; Wittrup, K. D. *Nat. Biotechnol.* **1997**, *15*, 553.

(151) Roberts, R. W.; Szostak, J. W. *Proc. Natl. Acad. Sci. USA* **1997**, *94*, 12297.

(152) Nelson, A. L.; Dhimolea, E.; Reichert, J. M. *Nat. Rev. Drug Discov.* **2010**, *9*, 767.

(153) Gray, B. P.; Brown, K. C. *Chem. Rev.* **2014**, *114*, 1020.

(154) Ballew, J. T.; Murray, J. A.; Collin, P.; Maki, M.; Kagnoff, M. F.; Kaukinen, K.; Daugherty, P. S. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 19330.

(155) Keefe, A. D.; Pai, S.; Ellington, A. *Nat. Rev. Drug Discov.* **2010**, *9*, 537.

(156) Barnhart, K. F.; Christianson, D. R.; Hanley, P. W.; Driessen, W. H. P.; Bernacky, B. J.; Baze, W. B.; Wen, S. J.; Tian, M.; Ma, J. F.; Kolonin, M. G.; Saha, P. K.; Do, K. A.; Hulvat, J. F.; Gelovani, J. G.; Chan, L.; Arap, W.; Pasqualini, R. *Sci. Transl. Med.* **2011**, *3*.

(157) Xie, J.; Zhang, H. K.; Yea, K.; Lerner, R. A. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 8099.

(158) Breaker, R. R.; Joyce, G. F. *Proc. Natl. Acad. Sci. USA* **1994**, *91*, 6093.

(159) Deiss, F.; Matochko, W. L.; Govindasamy, N.; Lin, E. Y.; Derda, R. *Angew. Chem. Int. Ed.* **2014**, *53*, 6374.

(160) Klim, J. R.; Li, L. Y.; Wrighton, P. J.; Piekarczyk, M. S.; Kiessling, L. L. *Nat. Methods* **2010**, *7*, 989.

(161) Morrison, S. J.; Spradling, A. C. *Cell* **2008**, *132*, 598.

(162) Morrison, S. J.; Kimble, J. *Nature* **2006**, *441*, 1068.

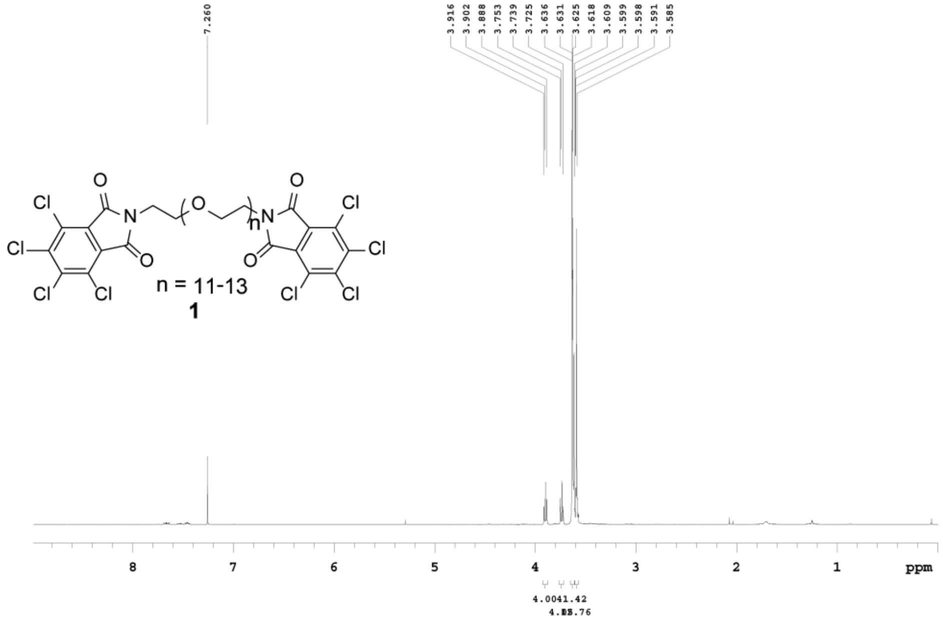
(163) Knoblich, J. A. *Cell* **2008**, *132*, 583.

(164) Siller, K. H.; Doe, C. Q. *Nat. Cell Biol.* **2009**, *11*, 365.

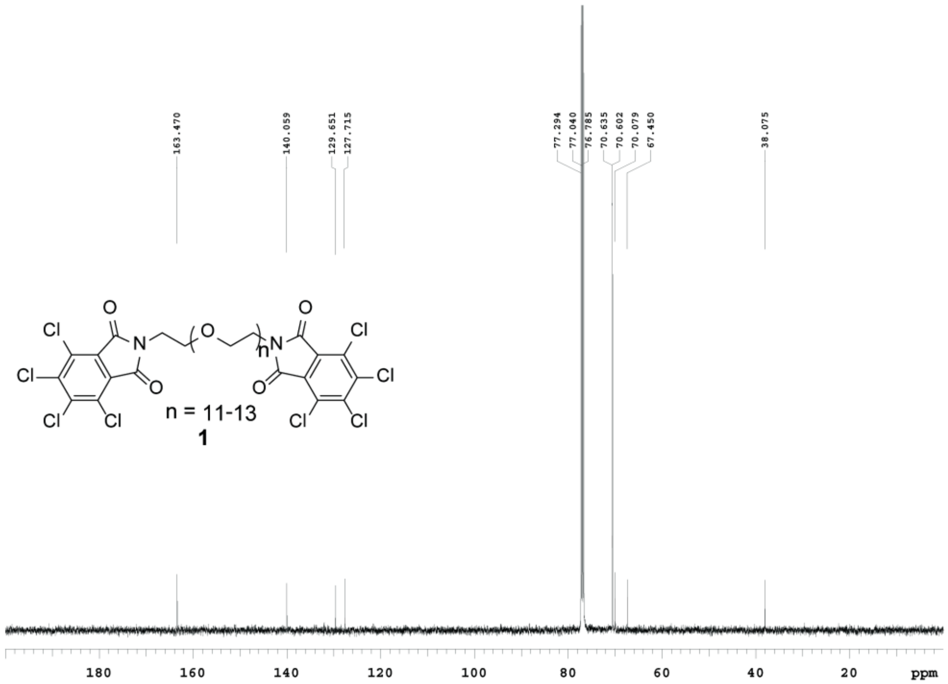
(165) Mani, S. A.; Guo, W.; Liao, M. J.; Eaton, E. N.; Ayyanan, A.; Zhou, A. Y.; Brooks, M.; Reinhard, F.; Zhang, C. C.; Shipitsin, M.; Campbell, L. L.; Polyak, K.; Brisken, C.; Yang, J.; Weinberg, R. A. *Cell* **2008**, *133*, 704.

(166) Quail, D. F.; Taylor, M. J.; Walsh, L. A.; Dieters-Castator, D.; Das, P.; Jewer, M.; Zhang, G. H.; Postovit, L. M. *Molecular Biology of the Cell* **2011**, *22*, 4809.

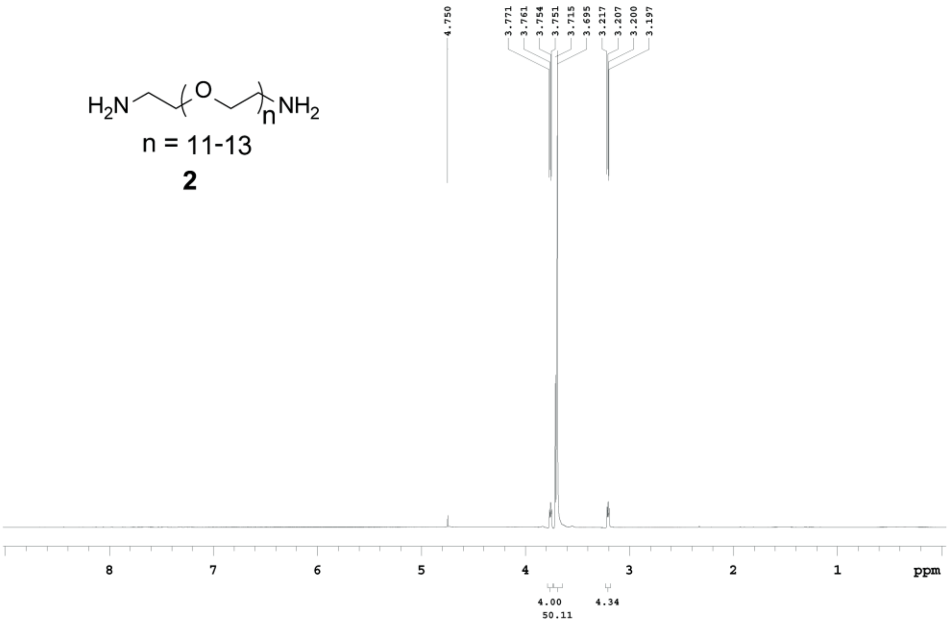
Appendix



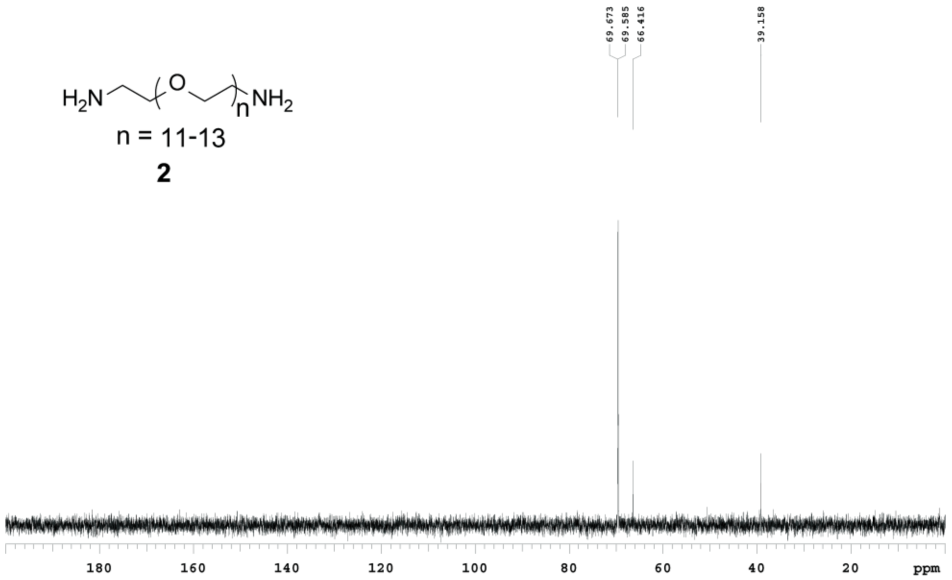
Appendix 1.1. ^{1}H-NMR (400 MHz, $CDCl_3$) of bis(tetrachlorophthalimido)-polyethylene glycol (**1**).



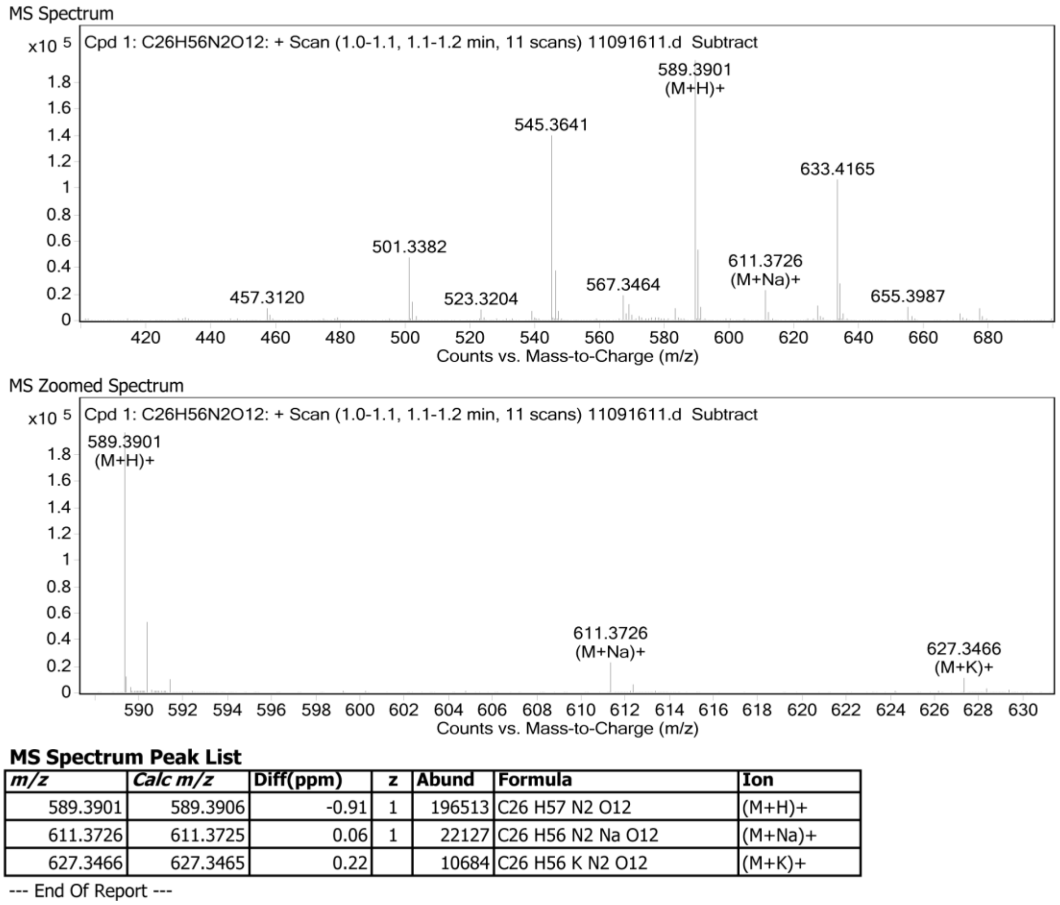
Appendix 1.2. ^{13}C-NMR (125 MHz, $CDCl_3$) of bis(tetrachlorophthalimido)-polyethylene glycol (**1**).



Appendix 1.3. ^{1}H-NMR (500 MHz, D_2O) of diaminopolyethylene glycol (NH_2-PEG-NH_2) (**2**).



Appendix 1.4. ^{13}C-NMR (125 MHz, D_2O) of diaminopolyethylene glycol (NH_2-PEG-NH_2) (**2**).



MS Spectrum

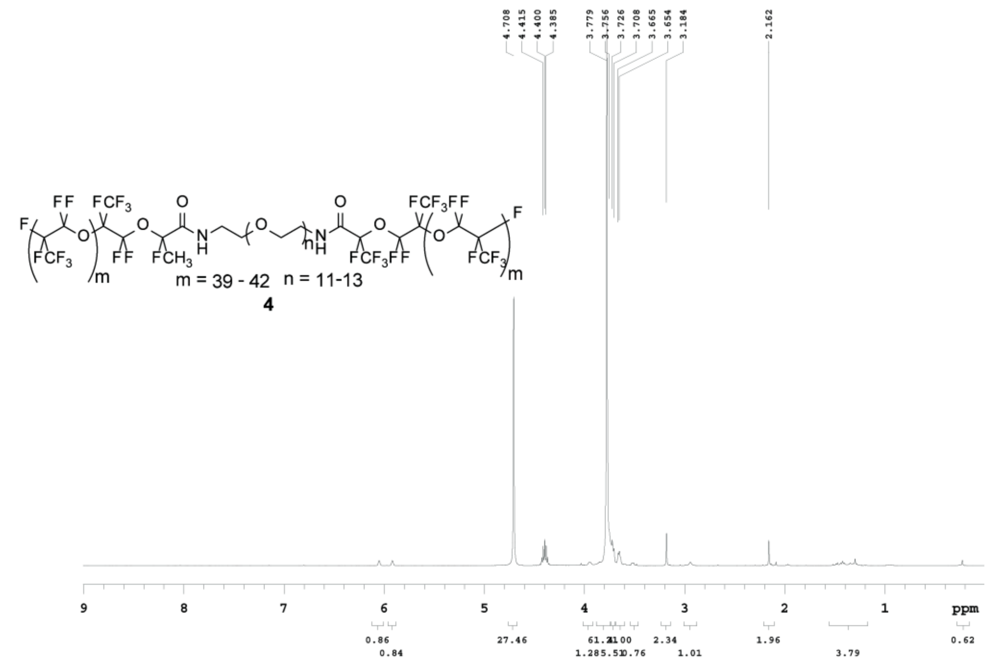
MS Zoomed Spectrum

MS Spectrum Peak List

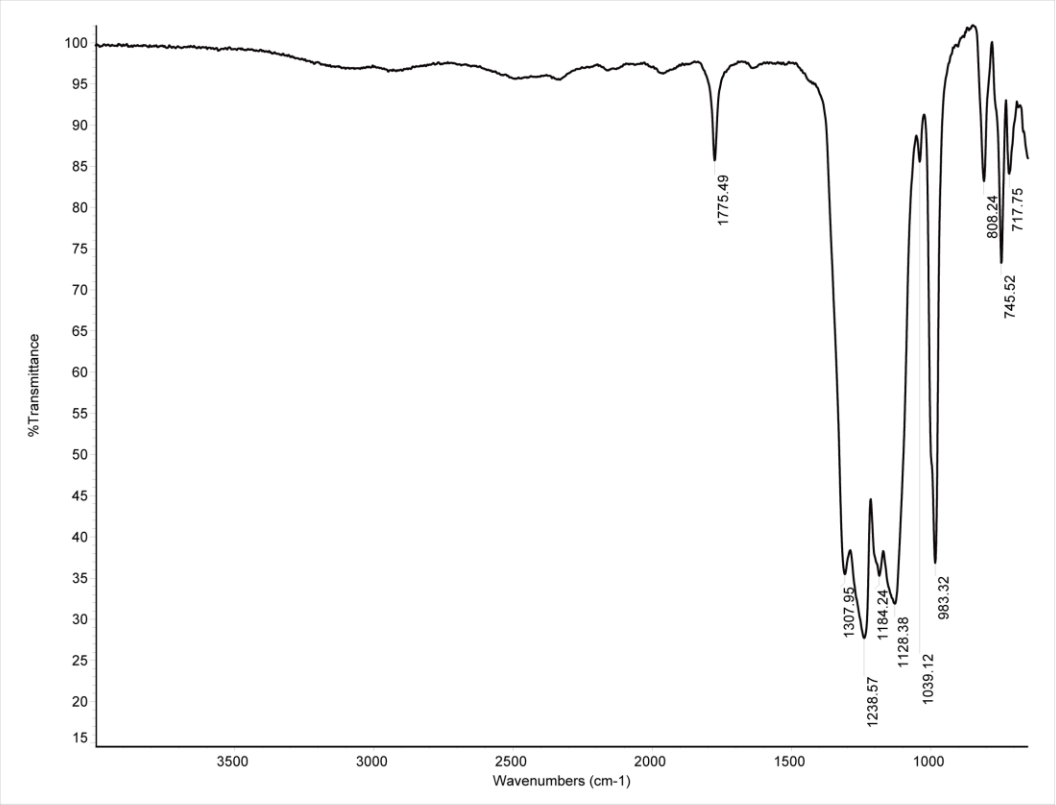
m/z	*Calc m/z*	Diff(ppm)	z	Abund	Formula	Ion
589.3901	589.3906	-0.91	1	196513	C26 H57 N2 O12	(M+H)+
611.3726	611.3725	0.06	1	22127	C26 H56 N2 Na O12	(M+Na)+
627.3466	627.3465	0.22		10684	C26 H56 K N2 O12	(M+K)+

--- End Of Report ---

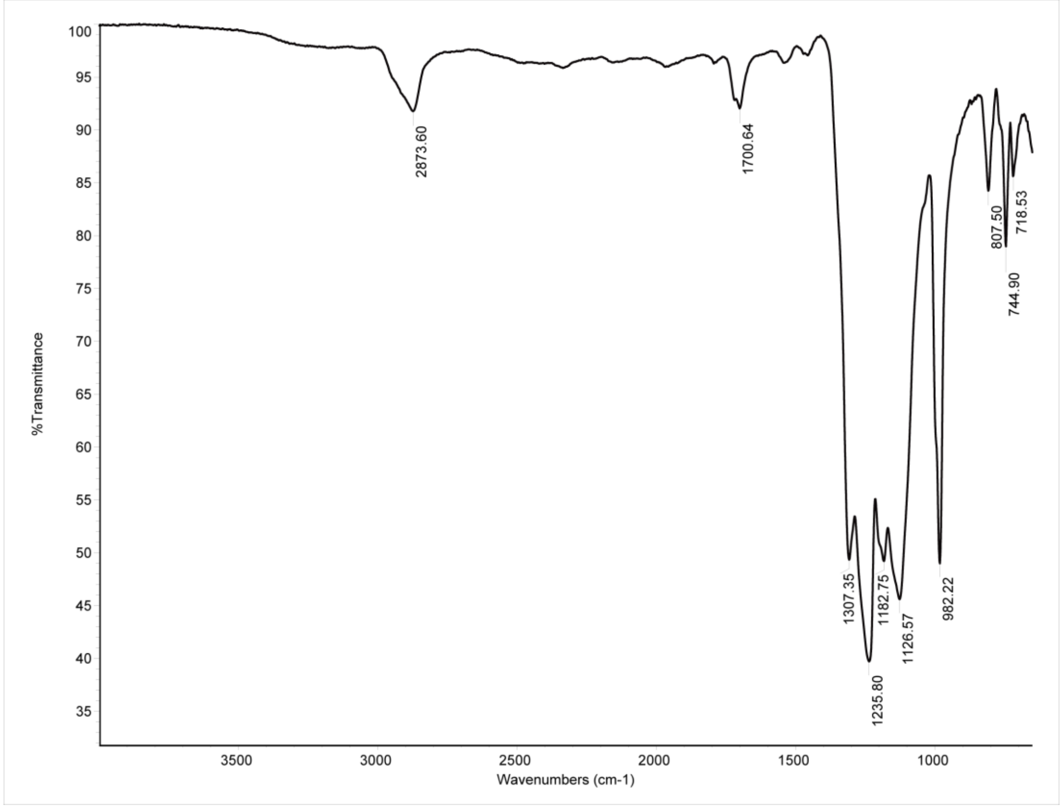
Appendix 1.5. ESI-MS of diaminopolyethylene glycol (NH_2-PEG-NH_2) (**2**).



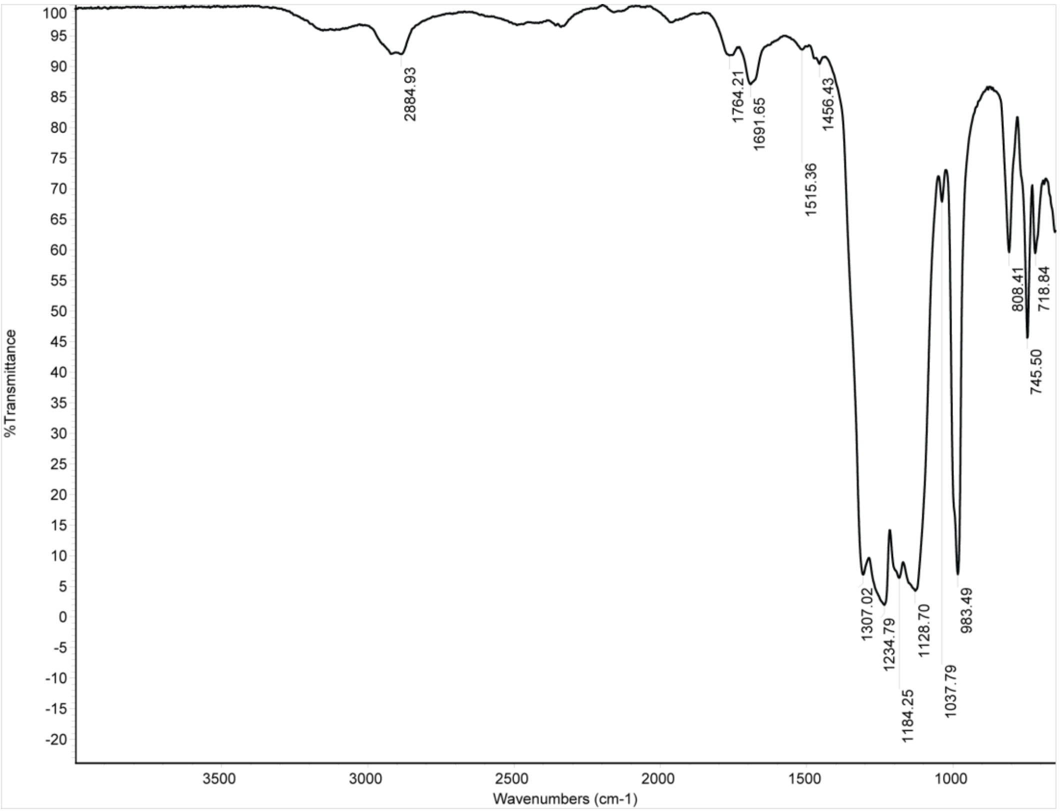
Appendix 1.6. ^{1}H-NMR (400 MHz, $C_3D_2F_6O$) of bis(perfluoropolyether)-polyethylene glycol (PFPE-PEG-PFPE) (**4**).



Appendix 1.7. FTIR of FSH-Krytox (**5**).



Appendix 1.8. FTIR of bis(perfluoropolyether)-polyethylene glycol (PFPE-PEG-PFPE) (**4**).



Appendix 1.9. FTIR of incomplete reaction between (perfluoropolyether)-acid chloride (PFPE-COCl) (**3**) (acid chloride of FSH-Krytox) and bis(perfluoropolyether)-polyethylene glycol (PFPE-PEG-PFPE) (**4**).

Region from M13KE vector with variable insert

```
1   GTG GTA CCT TTC TAT TCT CAC TCT                         24
25  NNK NNK NNK NNK NNK NNK NNK NNK NNK NNK NNK NNK          60
61  GGT GGA GGT TCG GCC GAA                                         78

Abbreviated as— TAT TCT CAC TCTR36GGT GGA GGT TCG
```

Primers

```
L1:   5'- NKK NKK ACT ATC TAT TCT CAC TCT -3'

R1:   5'- CGA ACC TCC ACC -3'

R2:   5'- TTC GGC CGA ACC TCC ACC -3'
(longer complimentary region)

(ACTATC is the 6-mer barcode)
(NKKNKK is a random hexamer that will facilitate cluster formation)
```

*********** **primer alignment for L1+R1 pair** ******************

```
5' NKK NKK ACT ATC TAT TCT CAC TCT                  3'   (L1)

5’                 TAT TCT CAC TCTR36GGT GGA GGT TCG 3’
3’                 ATA AGA GTG AGAR36CCA CCT CCA AGC 5’

3'                                   CCA CCT CCA AGC 5'  (R1)
```

Result after PCR (72 bp fragment):

```
5’ NKK NKK ACT ATC TAT TCT CAC TCTR36GGT GGA GGT TCG 3’
3’ NKK NKK TGA TAG ATA AGA GTG AGAR36CCA CCT CCA AGC 5’
```

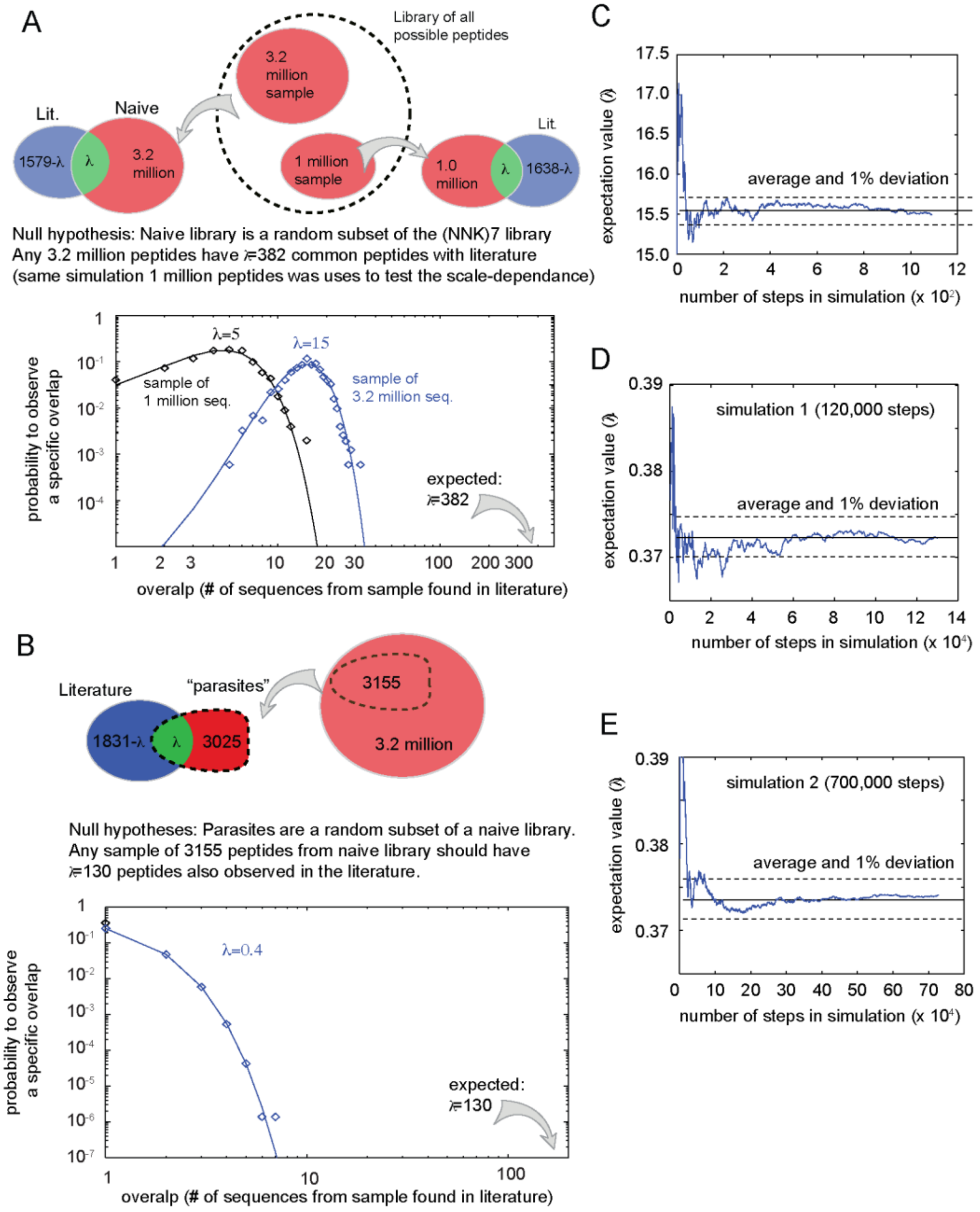
After ligation of the adapters and PCR (190 bp fragment):

```
5’ ILL-LEFT-NKK NKK ACT ATC TAT TCT CAC TCTR36GGT GGA GGT TCG-ILL-RIGHT 3’
3’ ILL-LEFT-NKK NKK TGA TAG ATA AGA GTG AGAR36CCA CCT CCA AGC-ILL-RIGHT 5’

ILL-LEFT- = 5’-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCxT-
ILL-LEFT- = 3’-TTACTATGCCGCTGGTGGCTCTAGATGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGxA-

-ILL-RIGHT = -AxGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT-3’
-ILL-RIGHT = -TxCTAGCCTTCTCGCAGCACATCCCTTTCTCACATCTAGAGCCACCAGCGGCATAGTAA-5’
```

Appendix 1.10. First generation design of primers. Alignment of primers to the (+) and (-) strands of the M13KE vector and expected products after PCR. We designed and tested two right primers: shorter R1 and longer R2. Both primers yielded expected product after amplification (see Supporting Figure 1). We selected the R1, because we were aiming to find the primer of the shortest possible length. Complete sequence of M13KE is available from New England Biolabs.

```
5'- NKKNBARTAT TCT CAC TCT -3'   (left-BAR1-NKKN)

5′- NKKNBARCGA ACC TCC ACC -3′   (right-BAR1-NKKN)

*********** primer alignment ******************

5'- NKKNBARTAT TCT CAC TCT -3'   (left)

5′         TAT TCT CAC TCT(NNK)12GGT GGA GGT TCG -3′

3′         ATA AGA GTG AGA(NNK)12CCA CCT CCA AGC -5′

3'                               CCA CCT CCA AGCRABNKKN -5'(right)

**************** fragment *********************

5′- NKKNBARTAT TCT CAC TCT(NNK)12GGT GGA GGT TCGRABNKKN -3′

3′- NKKNBARATA AGA GTG AGA(NNK)12CCA CCT CCA AGCRABNKKN -5′
```

Appendix 1.11. Second generation design of primers for the amplification of phage libraries. Each primer contains the NKKN region, barcode region (BAR), randomized nucleotide sequence ((NNK)$_{12}$) which corresponds to random amino acid sequence for Ph.D-12^{TM} phage, compliment sequence at the 5' end of the variable region (TAT TCT CAC TCT), and the reverse compliment at the 5' end (CCA CCT CCA AGC). Different primers used were only different at the BAR code region only.

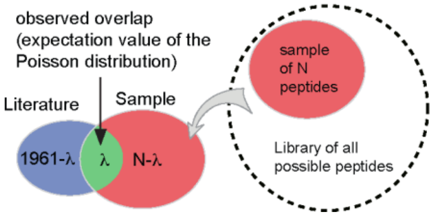


Null hypothesis: Naive library is a random subset of the (NNK)7 library
Any 3.2 million peptides have λ=382 common peptides with literature
(same simulation 1 million peptides was uses to test the scale-dependance)

Null hypotheses: Parasites are a random subset of a naive library.
Any sample of 3155 peptides from naive library should have
λ=130 peptides also observed in the literature.

F General sampling problem (see illustration on the right): Given a random sample of N peptides from a random peptide library encoded by (NNK)7 codons, how many peptides are expected to be common between this N-sample and a literature?

N	3,200,000	1,000,000	500,000	100,000	3,000	770
exp. value (λ)	15	5	2.5	0.5	0.015	0.004
maximum in in X trials	35 X = 1000	18 X = 5000	10 X = 5000	5 X = 5000	1 X = 10000	1 X = 20000

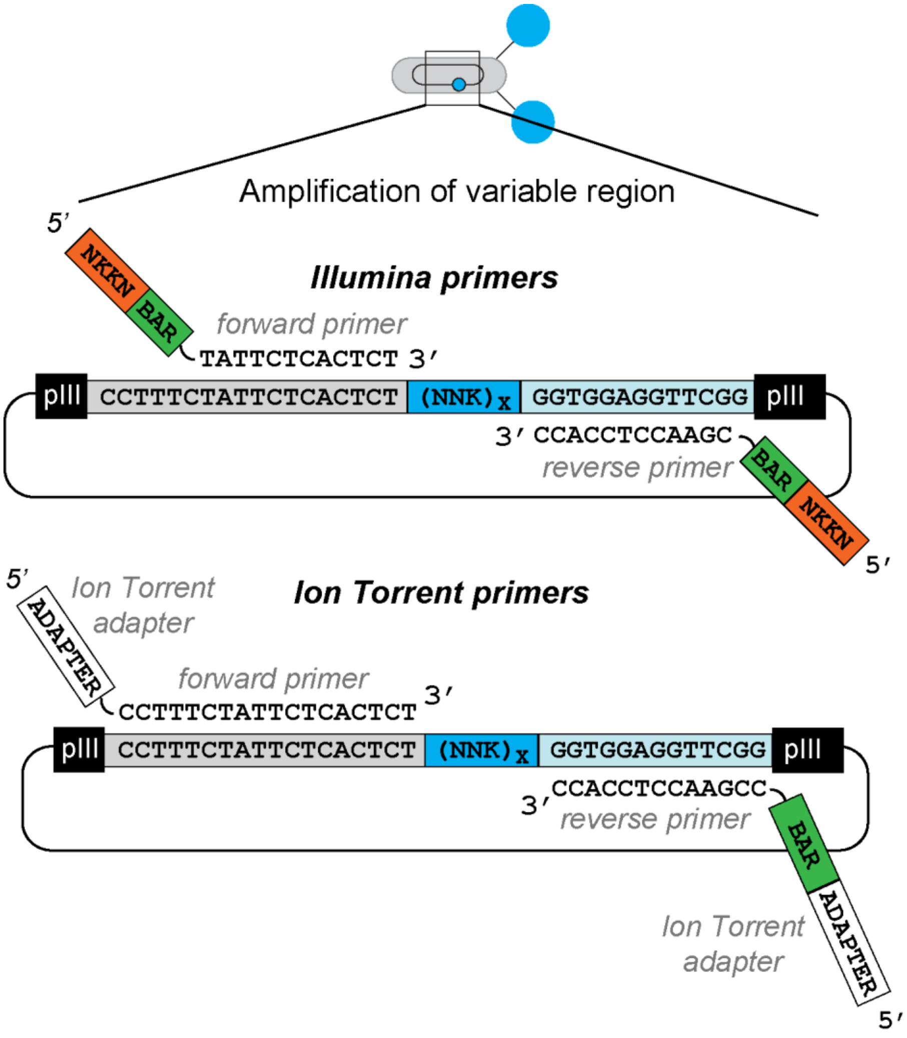


Appendix 1.12. (A) Overlap between literature and naïve library (382 common sequences) cannot occur at random. To prove it, we formulated null hypothesis and sampled 3.2 million peptides from a random peptide library based on NNK codons. In 5000 trials, the average number of common sequences was 15. The extrapolated probability to observe 382 common sequences was $p<<e^{-382}$. The same hypothesis for non-singleton population of the naïve library (ca. 10^6 seq.; ~220 literature hits): a random sample has 5 hits. The probability to observe 220 hits was $p << e^{-200}$ (B) Testing the significance of the overlap between parasite population P_{10} and literature (130 hits). In 10 million random trials, we selected 3155 random sub-sets from naïve library. Average overlap was 0.4 sequences. The probability to observe 130 common sequences was $p<<10^{-130}$. (C) Convergence of the bootstrapping simulation used to generate blue curve in (A). (D-E) convergence of simulations used to generate blue curve in (B). Running two separate simulations with 120,000 or 700,000 steps yields similar results. (E) The null hypothesis tested in (A) could be formulated more broadly for sequencing results that contain <1 million reads.

PhD7: ..TAT TCT CAC TCT (NNK)7 GGT GGA GGT TCG GCC..
Tyr Ser His Ser Rnd Gly Gly Gly Ser Ala

PhD12: ..TAT TCT CAC TCT (NNK)12 GGT GGA GGT TCG GCC..
Tyr Ser His Ser Rnd Gly Gly Gly Ser Ala

PhDC7C: ..TAT TCT CAC TCT GCT TGT (NNK)7 TGC GGT GGA GGT TCG GCC..
Tyr Ser His Ser Ala Cys Rnd Cys Gly Gly Gly Ser Ala



Appendix 1.13. Design of Illumina and Ion Torrent primers. Primer sequences are universal for all libraries made by New England Biolabs (Ph.D.-7^{TM}, Ph.D.-12^{TM}, and Ph.D.-C7C^{TM}) because all three libraries contain the same flanking regions. Note that Tyr-Ser-His-Ser is part of the pIII leader sequence, which is removed during the periplasmic export of the phage.

1. Read FASTAQ file. Parse the reads and create a tagged file containing reads in which the adapters can be identified. Discard non-tagged reads (~2%)
Example of the tagged read:

```
Library
 |  Adapter
 |   |   Adapter
 |   |   quality
 |   |    |       barcode
 |   |    |    NKKN  |  adapter seq.    variable sequences                      quality string parsed in the
same way as sequence string
 |   |    |      |   |     |                          |
Ph7 <F> <PERF> ATGC TTG TATTCTCACTCT ACTACGGTTACGGGGGAGTCTGGTGGAGGTTCG.......  CCCF FFF EHHHHHJJJJJJ
JJJJJJJIIJJJHIJJGJFHHJJHHHHHFFFFD.......

Ph7 <F> <PERF> CTTC TTG TATTCTCACTCT GTTGATAGGCTTCTGTTGATGGGTGGAGGTTCG.......  CCCF FFF FHHHHHJJJJJJ
JJJ+AFHIJJJJJJJIIIFHIJJDGIGIJIIJJ.......

P12 <F> <PERF> GTGC TCA TATTCTCACTCT GCGGAGCCGACGGATTGGCTTACTGTTCGTGAT.......  CBBF FFF FHHHHHJJJJJJ
JJJ)BFHJJJJJIIJJJJJJJIJJJJJJJHHHF.......

P12 <F> <PERF> GGGG TTG TATTCTCACTCT TGGTCTCCGGGTCGTGGTCCGATTGAGTCGCAG.......  BCCF DDF DHHHHHJJJJJJ
JJJHIIJJJJJGHIHIJHIJJJJIIGHHHIIHH.......

Ph7 <F> <PERF> CTTG CGA TATTCTCACTCT CGGACTGCGCCTACTGTGATGGGTGGAGGTTCG.......  B@@F FFD FHHHHHJJJJJJ
JJJJJIIIJIJJHIIJGGDIGIJFHIIJJJJJG.......

C7C <F> <PERF> TGGC TTG TATTCTCACTCT GCTTGTCATTGGTCTCCGGCTTCTCCGTGCGGT.......  ???; 4B1 AD?DDDEEEEEE
C@<E9<::9E9<2AEECEADDA)?9?D/??AD#.......

P12 <F> <PERF> GTGG GAC TATTCTCACTCT ACGCCGAATGCTGTTTTGCCGACTGGGCGTACG.......  @@@D FBD FFHHHHHIEIIII
IGHIIIIIIHIIDGHFHGIIIIAHGCGFHIHHF.......

Ph7 <F> <PERF> GGGC TTG TATTCTCACTCT AGGCCTGCGTCTCTGCCTCCGGGTGGAGGTTCG.......  CBCD FFF FGHHHHJJJJJJ
IJJJIIEGIHGGIIFIJJJJIIJFHJ;FH@EHI.......

Ph7 <F> <PERF> CGGG TTG TATTCTCACTCT CAGCGGCATCATGCGGATGTTGGTGGAGGTTCG.......  B@BF DDF FHHHHHJJJJJJ
JGJJJJJJIJJJFHIJHJDGHJJEGIIJJJJHH.......

P12 <F> <PERF> AGGG TCA TATTCTCACTCT GTTCCTGCTTGGGCGGTGGAGGTTCGTACGCCT.......  BCCF DDF FHHHHHJJJJJJ
JJJJJJJJJJJHIJJGHGHIJFHIJHIIHHHF.......

Ph7 <F> <PERF> AGGG TCA TATTCTCACTCT ATTGAGTCGACGCGGCATCTGGGTGGAGGTTCG.......  BCCF DDF FHHHHHJJJJJJ
JJJ+AFFHIJJJJJJJJJJJJJJ@HIHHHFFFF.......

C7C <F> <PERF> CGGT CGA TATTCTCACTCT GCTTGTCATCAGGCTGGTGCTCCGCATTGCGGT.......  @@CD FFF FHHHHHJJJJJJ
JGHGHIIJJGGHHIIIF?DFGGIIJ0DFHHIHF........

C7C <R> <PERF> CGTC GTA CGAACCTCCACC GCAAAAACTCGACGCCTGATTCTTACAAGCAGA.......  CCCF FFF FHHHHHJJJJJJ
JJJJJ00:?BG)?FHIJJJJJJJJJJHHHHHFF.......

Ph7 <R> <PERF> CTTG TTG CGAACCTCCACC ATGAAGACGCATACTAAACGCAGAGTGAGAATA.......  CCCF FFF FHGHHHJJJJJJ
JIJIJCIJJJJJJJJJJJGJIJJJJJJJJJJJG.......

C7C <R> <PERF> AGGT GTA CGAACCTCCACC GCACAGATTCCGCTGCAACGGACTACAAGCAGA.......  BCCD DEF FHHHHHJJJJJJ
JJJJJ<GHJHJJJJJJJJIJJJJJJJIJJJHHH.......

P12 <R> <PERF> AGGC GTA CGAACCTCCACC AGAAGTCTCCCTCCTATGCAGCTGCGAATCATG.......  @CCF FDF FFFHHFIJJJJJ
GGGHIGIIJIJJGIJJIEIIJJJJJGIJJIJII.......

C7C <R> <PERF> GTGT TTG CGAACCTCCACC GCACTTCTGCACATTATGCCCCCCACAAGCAGA.......  CBBF FFF FHHHHHJJJJJJ
JJJJJJJJJJJ9DHIJJJJJJJJJJHHHHFFFF.......

P12 <R> <PERF> TGGG GAC CGAACCTCCACC AACCTTAGAACTAGGCATAGTAAGCGAAATATC.......  CCCF FFF FHHHHHJJJIJJ
JJJJJCHG4BHHFIJJJJGIGHIJJJJJJJJJH.......

P12 <R> <PERF> TGGG GAC CGAACCTCCACC CCTAGGCACAACCTCATACGCCTGCCGATACGG.......  CCCF FFF FHHHHHJJJJJJ
JJJJJJJJJJJJJJJJJIGI8FHJJJJHHFFDD.......

Ph7 <R> <PERF> CTGG CGA CGAACCTCCACC AGGAAACGAATGCCCAAGCGCAGAGTGAGAATA.......  CCCF FFF FHHHHHJJJJJJ
JJJJJ000)0?FGIJJJJFIGIHHHHFFFFFFC.......

C7C <R> <PERF> CGGG GTA CGAACCTCCACC GCACAGCTCAATAGACTTACTCAGACAAGCAGA.......  CCCF FFF FHHHHHJJJJJJ
JJJJJIJJJJJJJJJJJJJJJJJJJJJJJJJJJ.......
                                                   Etc
```

>50,000,000 lines

2. Process tagged file, filter the Phred<30 reads, and reads that contain NNM codons, sort the reads by libraries and experiments (barcodes).

3. Process F and R reads separately. Translate. Save into separate files for inspection.
Example of one of the processed files (F read on the left, R read on the right)

GAGCCGCTGCAGCTGAAGATG	EPLQLKM	74758	GAGCCGCTGCAGCTGAAGATG	EPLQLKM	67210
CATGCTATTTATCCGCGTCAT	HAIYPRH	68194	CATGCTATTTATCCGCGTCAT	HAIYPRH	64004
GGGAAGCCTATGCCTCCGATG	GKPMPPM	60099	GGTCCTATGCTGGCTCGTGGT	GPMLARG	60240
TCGCCGCAGATGACTCTTTCG	SPQMTLS	52692	TCGCCGCAGATGACTCTTTCG	SPQMTLS	57230
CATGCTCTGGGTCCGTCTTCG	HALGPSS	51036	CATGCTCTGGGTCCGTCTTCG	HALGPSS	55316
TGGCCGCAGAAGGCTCAGCCT	WPQKAQP	45254	GGGAAGCCTATGCCTCCGATG	GKPMPPM	54771
TCGACGGCGTCTTATACTCGT	STASYTR	42262	GCGACGACTGTTCCAGCTTCG	ATTVPAS	44445
CAGCCTTGGCCGACGAGTATT	QPWPTSI	31321	TGGCCGCAGAAGGCTCAGCCT	WPQKAQP	37943
TGGCCTACGCCGCCTTATGCG	WPTPPYA	30741	AGTCCGACGCAGCCTAAGTCG	SPTQPKS	34869

AGTCCGACGCAGCCTAAGTCG	SPTQPKS	27246	GGTAAGGTGCAGGCGCAGTCG	GKVQAQS	29403
GCTATGTCGTCTCGTTCGCTT	AMSSRSL	25774	GCGGCTGGTCAGCAGTTTCCT	AAGQQFP	25604
GCGACTCCGCTTTGGCTTAAG	ATPLWLK	25487	CAGCCTTGGCCGACGAGTATT	QPWPTSI	25588
CAGCCTCCTCGTTCGACGTCG	QPPRSTS	22462	GATTCGCATACTCCGCAGAGG	DSHTPQR	24964
CAGGCTACGCATCGTTCGCAT	QATHRSH	20473	CATAGGGCGGATATGCATTTT	HRADMHF	19965
GATTCGCATACTCCGCAGAGG	DSHTPQR	20319	ACGCGGGCTGGTCTGGATTTT	TRAGLDF	19930
GCGACGACTGTTCCAGCTTCG	ATTVPAS	19646	TGGCCTACGCCGCCTTATGCG	WPTPPYA	19884
		Etc >1,000,000 lines			Etc >1,000,000 lines

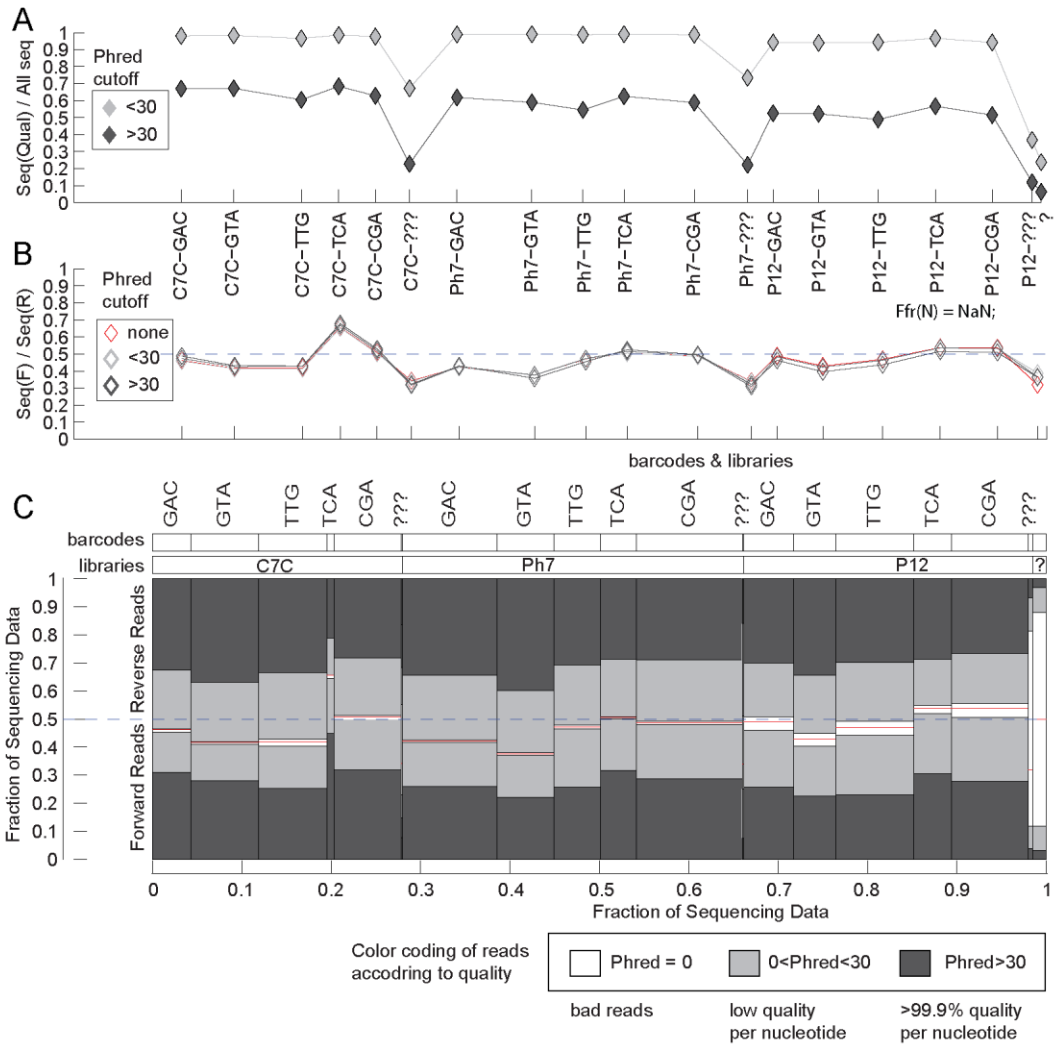
Combine the F and R reads using multiset union definition.

GAGCCGCTGCAGCTGAAGATG	EPLQLKM	74758
CATGCTATTTATCCGCGTCAT	HAIYPRH	68194
GGTCCTATGCTGGCTCGTGGT	GPMLARG	60240
GGGAAGCCTATGCCTCCGATG	GKPMPPM	60099
TCGCCGCAGATGACTCTTTCG	SPQMTLS	57230
CATGCTCTGGGTCCGTCTTCG	HALGPSS	55316
TGGCCGCAGAAGGCTCAGCCT	WPQKAQP	45254
GCGACGACTGTTCCAGCTTCG	ATTVPAS	44445
TCGACGGCGTCTTATACTCGT	STASYTR	42262
AGTCCGACGCAGCCTAAGTCG	SPTQPKS	34869
CAGCCTTGGCCGACGAGTATT	QPWPTSI	31321
TGGCCTACGCCGCCTTATGCG	WPTPPYA	30741
GGTAAGGTGCAGGCGCAGTCG	GKVQAQS	29403
GCTATGTCGTCTCGTTCGCTT	AMSSRSL	25774
GCGGCTGGTCAGCAGTTTCCT	AAGQQFP	25604
GCGACTCCGCTTTGGCTTAAG	ATPLWLK	25487
GATTCGCATACTCCGCAGAGG	DSHTPQR	24964
CAGCCTCCTCGTTCGACGTCG	QPPRSTS	22462
CAGGCTACGCATCGTTCGCAT	QATHRSH	20473
CATAGGGCGGATATGCATTTT	HRADMHF	19965
ACGCGGGCTGGTCTGGATTTT	TRAGLDF	19930
CAGCGGCTGCCTCAGACGGCG	QRLPQTA	17573
CAGTCTAGTGTTCTGCGGCAT	QSSVLRH	17312
GCTGCTAAGACGCCTACGGAG	AAKTPTE	16774

Etc >1,000,000 lines

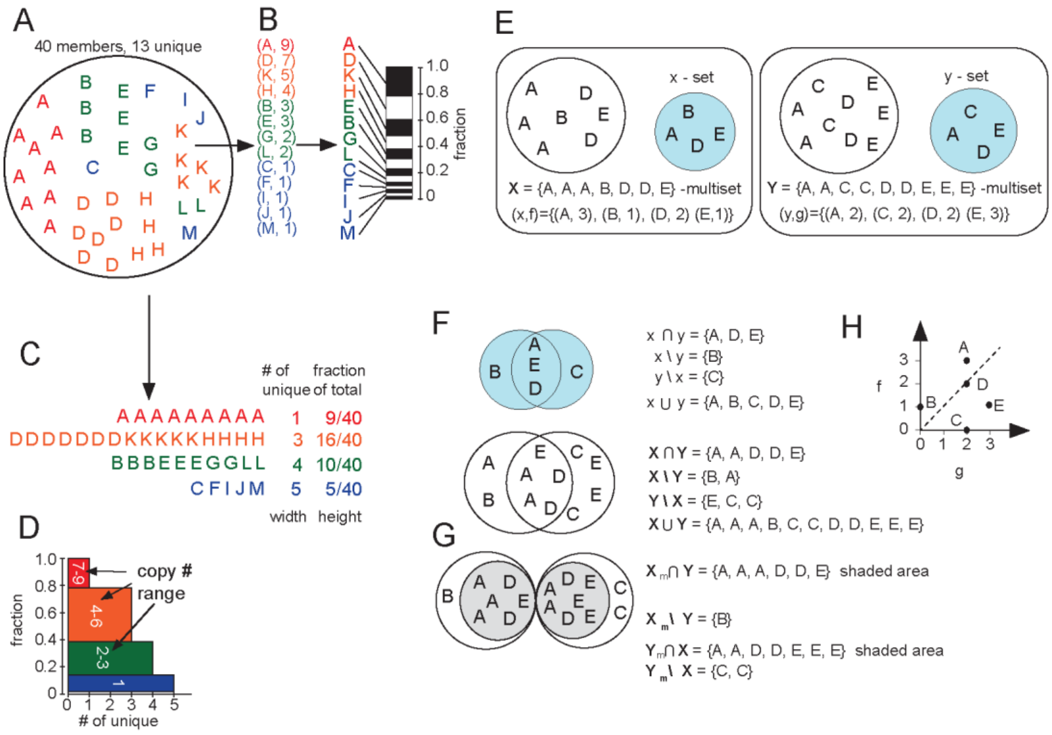
The files were saved as *.txt and used for all graphical processing by the command_center.m script.

Appendix 1.14. Illumina Analysis workflow.

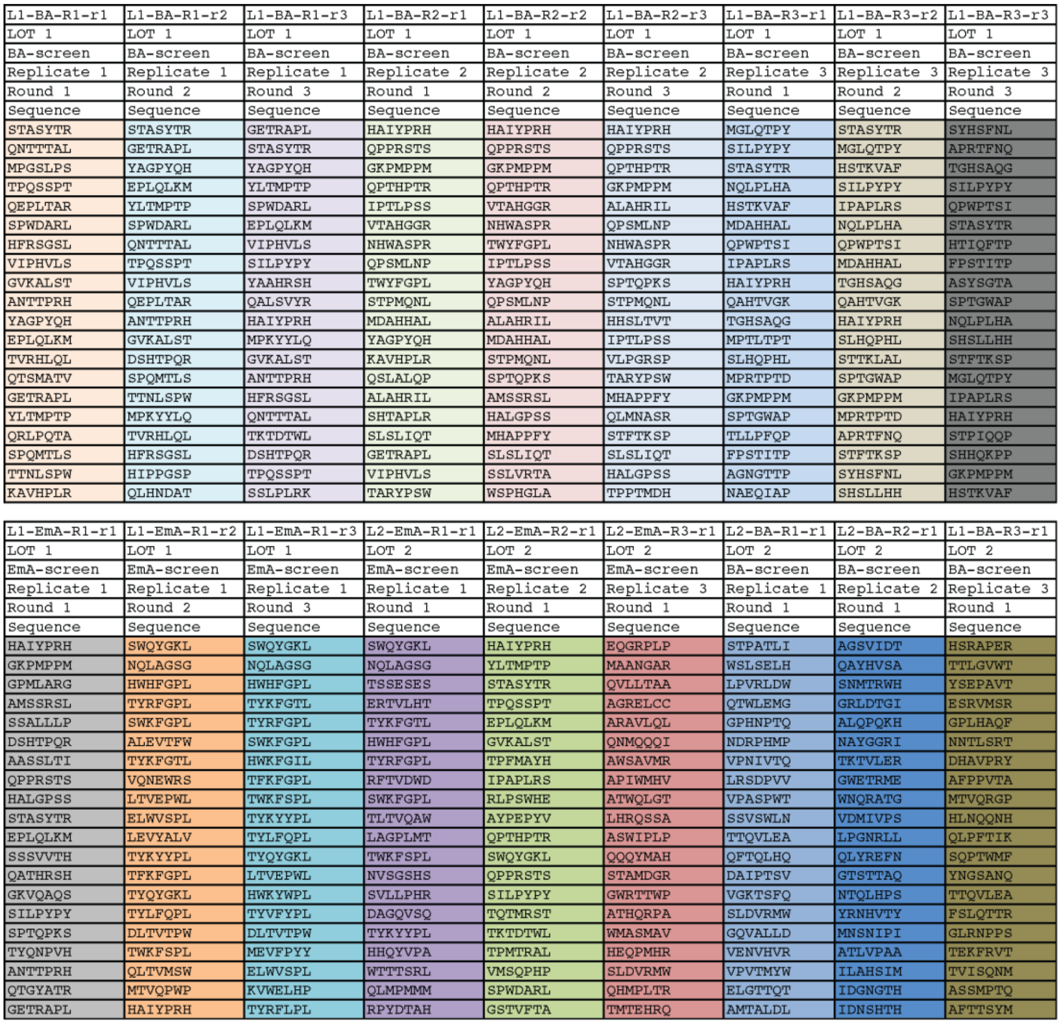


Appendix 1.15. One Illumina sequencing run was used to analyze 3 different libraries and 5 experiments within each library. Experiments were identified using barcodes; libraries were identified by sequence structure. (A) We processed the data using three different cutoffs. Only Phred>30 sequences were used in this paper. The fraction of Phred>0, >13 and >30 sequences in each experiment was consistent. Sequences in which barcodes were damaged (labeled by "???") had significantly lower quality. Damaged barcodes are sequences that do not correspond to any sequences in the original list of barcode sequences. (B) Each experiment contained forward and reverse reads. Their ratio was skewed in reads with damaged barcodes (i.e. low quality). (C) Overall view of the library. Each rectangle represents an experiment. The area of each rectangle is proportional to the fraction of this experiment in the overall pool of sequences. Color represents sequences of certain quality. For example, the first vertical stacked bar represents C7C library tagged by GAC-barcode. Sequences constitute ~4% of the overall sequence space. Ratio of forward sequences is ~45%. Reads with Phred>13 and Phred>30 cutoffs constitute ~80% and ~60% respectively. This graph shows that

~3% of the sequences could not be mapped to any library or any barcode. Majority of those sequences are bad reads (i.e. they contain at least one unknown nucleotide with Phred=0).



Appendix 1.16. (A) Example of multiset or set with multiple elements. (B) A multiset could be represented as a stacked bar. (C) Isolation of sub-sets with different copy numbers could be used to represent the multiset as 2D stacked-bar (D). (E) Comparison of two multisets X and Y, which consist of sets *x* and *y* and multiplicity functions *f* and *g*. (F) Venn diagram describing intersection, difference and union in sets and multisets. (G) Multiset-specific m-intersect and m-difference operators. Note that $X_m \cap Y$ and $Y_m \cap X$ are non-commutative and they are different from intersect $X \cap Y$. (H) Scatter plot of f vs. g multiplicity functions describes the abundances of elements in X in Y.



L1-BA-R1-r1	L1-BA-R1-r2	L1-BA-R1-r3	L1-BA-R2-r1	L1-BA-R2-r2	L1-BA-R2-r3	L1-BA-R3-r1	L1-BA-R3-r2	L1-BA-R3-r3
LOT 1	LOT 1	LOT 1	LOT 1	LOT 1	LOT 1	LOT 1	LOT 1	LOT 1
BA-screen	BA-screen	BA-screen	BA-screen	BA-screen	BA-screen	BA-screen	BA-screen	BA-screen
Replicate 1	Replicate 1	Replicate 1	Replicate 2	Replicate 2	Replicate 2	Replicate 3	Replicate 3	Replicate 3
Round 1	Round 2	Round 3	Round 1	Round 2	Round 3	Round 1	Round 2	Round 3
Sequence	Sequence	Sequence	Sequence	Sequence	Sequence	Sequence	Sequence	Sequence
STASYTR	STASYTR	GETRAPL	HAIYPRH	HAIYPRH	HAIYPRH	MGLQTPY	STASYTR	SYHSFNL
QNTTTAL	GETRAPL	STASYTR	QPPRSTS	QPPRSTS	QPPRSTS	SILPYPY	MGLQTPY	APRTFNQ
MPGSLPS	YAGPYQH	YAGPYQH	GKPMPPM	GKPMPPM	QPTHPTR	STASYTR	HSTKVAF	TGHSAQG
TPQSSPT	EPLQLKM	YLTMPTP	QPTHPTR	QPTHPTR	GKPMPPM	NQLPLHA	SILPYPY	SILPYPY
QEPLTAR	YLTMPTP	SPWDARL	IPTLPSS	VTAHGGR	ALAHRIL	HSTKVAF	IPAPLRS	QPWPTSI
SPWDARL	SPWDARL	EPLQLKM	VTAHGGR	NHWASPR	QPSMLNP	MDAHHAL	NQLPLHA	STASYTR
HFRSGSL	QNTTTAL	VIPHVLS	NHWASPR	TWYFGPL	NHWASPR	QPWPTSI	QPWPTSI	HTIQFTP
VIPHVLS	TPQSSPT	SILPYPY	QPSMLNP	IPTLPSS	VTAHGGR	IPAPLRS	MDAHHAL	FPSTITP
GVKALST	VIPHVLS	YAAHRSH	TWYFGPL	YAGPYQH	SPTQPKS	HAIYPRH	TGHSAQG	ASYSGTA
ANTTPRH	QEPLTAR	QALSVYR	STPMQNL	QPSMLNP	STPMQNL	QAHTVGK	QAHTVGK	SPTGWAP
YAGPYQH	ANTTPRH	HAIYPRH	MDAHHAL	ALAHRIL	HHSLTVT	TGHSAQG	HAIYPRH	NQLPLHA
EPLQLKM	GVKALST	MPKYYLQ	YAGPYQH	MDAHHAL	IPTLPSS	MPTLTPT	SLHQPHL	SHSLLHH
TVRHLQL	DSHTPQR	GVKALST	KAVHPLR	STPMQNL	VLPGRSP	SLHQPHL	STTKLAL	STFTKSP
QTSMATV	SPQMTLS	ANTTPRH	QSLALQP	SPTQPKS	TARYPSW	MPRTPTD	SPTGWAP	MGLQTPY
GETRAPL	TTNLSPW	HFRSGSL	ALAHRIL	AMSSRSL	MHAPPFY	GKPMPPM	GKPMPPM	IPAPLRS
YLTMPTP	MPKYYLQ	QNTTTAL	SHTAPLR	HALGPSS	QLMNASR	SPTGWAP	MPRTPTD	HAIYPRH
QRLPQTA	TVRHLQL	TKTDTWL	SLSLIQT	MHAPPFY	STFTKSP	TLLPFQP	APRTFNQ	STPIQQP
SPQMTLS	HFRSGSL	DSHTPQR	GETRAPL	SLSLIQT	SLSLIQT	FPSTITP	STFTKSP	SHHQKPP
TTNLSPW	HIPPGSP	TPQSSPT	VIPHVLS	SSLVRTA	HALGPSS	AGNGTTP	SYHSFNL	GKPMPPM
KAVHPLR	QLHNDAT	SSLPLRK	TARYPSW	WSPHGLA	TPPTMDH	NAEQIAP	SHSLLHH	HSTKVAF

L1-EmA-R1-r1	L1-EmA-R1-r2	L1-EmA-R1-r3	L2-EmA-R1-r1	L2-EmA-R2-r1	L2-EmA-R3-r1	L2-BA-R1-r1	L2-BA-R2-r1	L1-BA-R3-r1
LOT 1	LOT 1	LOT 1	LOT 2	LOT 2	LOT 2	LOT 2	LOT 2	LOT 2
EmA-screen	EmA-screen	EmA-screen	EmA-screen	EmA-screen	EmA-screen	BA-screen	BA-screen	BA-screen
Replicate 1	Replicate 1	Replicate 1	Replicate 1	Replicate 2	Replicate 3	Replicate 1	Replicate 2	Replicate 3
Round 1	Round 2	Round 3	Round 1	Round 1	Round 1	Round 1	Round 1	Round 1
Sequence	Sequence	Sequence	Sequence	Sequence	Sequence	Sequence	Sequence	Sequence
HAIYPRH	SWQYGKL	SWQYGKL	SWQYGKL	HAIYPRH	EQGRPLP	STPATLI	AGSVIDT	HSRAPER
GKPMPPM	NQLAGSG	NQLAGSG	NQLAGSG	YLTMPTP	MAANGAR	WSLSELH	QAYHVSA	TTLGVWT
GPMLARG	HWHFGPL	HWHFGPL	TSSESES	STASYTR	QVLLTAA	LPVRLDW	SNMTRWH	YSEPAVT
AMSSRSL	TYRFGPL	TYKFGTL	ERTVLHT	TPQSSPT	AGRELCC	QTWLEMG	GRLDTGI	ESRVMSR
SSALLLP	SWKFGPL	TYRFGPL	TYKFGTL	EPLQLKM	ARAVLQL	GPHNPTQ	ALQPQKH	GPLHAQF
DSHTPQR	ALEVTFW	SWKFGPL	HWHFGPL	GVKALST	QNMQQQI	NDRPHMP	NAYGGRI	NNTLSRT
AASSLTI	TYKFGTL	HWKFGIL	TYRFGPL	TPFMAYH	AWSAVMR	VPNIVTQ	TKTVLER	DHAVPRY
QPPRSTS	VQNEWRS	TFKFGPL	RFTVDWD	IPAPLRS	APIWMHV	LRSDPVV	GWETRME	AFPPVTA
HALGPSS	LTVEPWL	TWKFSPL	SWKFGPL	RLPSWHE	ATWQLGT	VPASPWT	WNQRATG	MTVQRGP
STASYTR	ELWVSPL	TYKYYPL	TLTVQAW	AYPEPYV	LHRQSSA	SSVSWLN	VDMIVPS	HLNQQNH
EPLQLKM	LEVYALV	TYLFQPL	LAGPLMT	QPTHPTR	ASWIPLP	TTQVLEA	LPGNRLL	QLPFTIK
SSSVVTH	TYKYYPL	TYQYGKL	TWKFSPL	SWQYGKL	QQQYMAH	QFTQLHQ	QLYREFN	SQPTWMF
QATHRSH	TFKFGPL	LTVEPWL	NVSGSHS	QPPRSTS	STAMDGR	DAIPTSV	GTSTTAQ	YNGSANQ
GKVQAQS	TYQYGKL	HWKYWPL	SVLLPHR	SILPYPY	GWRTTWP	VGKTSFQ	NTQLHPS	TTQVLEA
SILPYPY	TYLFQPL	TYVFYPL	DAGQVSQ	TQTMRST	ATHQRPA	SLDVRMW	YRNHVTY	FSLQTTR
SPTQPKS	DLTVTPW	DLTVTPW	TYKYYPL	TKTDTWL	WMASMAV	GQVALLD	MNSNIPI	GLRNPPS
TYQNPVH	TWKFSPL	MEVFPYY	HHQYVPA	TPMTRAL	HEQPMHR	VENVHVR	ATLVPAA	TEKFRVT
ANTTPRH	QLTVMSW	ELWVSPL	WTTTSRL	VMSQPHP	SLDVRMW	VPVTMYW	ILAHSIM	TVISQNM
QTGYATR	MTVQPWP	KVWELHP	QLMPMMM	SPWDARL	QHMPLTR	ELGTTQT	IDGNGTH	ASSMPTQ
GETRAPL	HAIYPRH	TYRFLPL	RPYDTAH	GSTVFTA	TMTEHRQ	AMTALDL	IDNSHTH	AFTTSYM

Appendix 1.17. List of the top 20 sequences from each round and replicate of selection using both BA and EmA methods. The abbreviation and color for each selection is used to distinguish the origin of hits in Appendices 1.19,1.21-1.25.

	Lot1-BA-Rep1-Round1	Lot1-BA-Rep1-Round2	Lot1-BA-Rep1-Round3	Lot1-BA-Rep2-Round1	Lot1-BA-Rep2-Round2	Lot1-BA-Rep2-Round3	Lot1-BA-Rep3-Round1	Lot1-BA-Rep3-Round2	Lot1-BA-Rep3-Round3	Lot1-EmA-Rep1-Round1	Lot1-EmA-Rep1-Round2	Lot1-EmA-Rep1-Round3	Lot2-EmA-Rep1-Round1	Lot2-EmA-Rep2-Round1	Lot2-EmA-Rep3-Round1	Lot2-BA-Rep1-Round1	Lot2-BA-Rep2-Round1	Lot2-BA-Rep3-Round1	Lot1-Parasites	Lot2-Parasites
STASYTR	645	3722	4203	1	4	0	807	3490	1957	17	6	1	38	1695	53	0	0	0	1	0
QNTTTAL	482	629	229	0	0	0	0	0	0	3	0	0	7	575	32	0	0	0	0	0
LPGSLPS	416	143	23	4	0	0	0	0	0	0	0	0	0	151	38	0	0	0	1	0
TPQSSPT	400	612	207	0	0	0	0	0	0	3	8	1	22	1408	50	0	0	0	1	0
QEPLTAR	317	549	97	0	1	0	0	0	0	0	0	0	10	413	5	0	0	0	0	0
SPWDARL	281	705	593	3	6	0	0	1	0	1	0	0	41	697	39	0	0	0	1	0
HFRSGSL	263	266	264	0	0	0	0	0	0	4	0	1	0	563	12	0	0	0	1	0
VIPHVLS	262	569	503	101	29	7	1	0	0	4	0	0	27	207	44	0	0	0	1	0
GVKALST	236	497	295	42	7	22	0	0	0	6	1	0	5	1139	10	0	0	0	1	0
ANTTPRH	236	523	289	0	0	0	2	8	1	12	0	3	0	475	32	0	0	0	1	0
YAGPYQH	215	1028	1907	190	141	63	1	2	0	4	2	0	4	437	16	0	0	0	1	0
EPLQLKM	214	777	586	3	2	0	2	0	0	16	6	0	7	1159	6	1	0	0	1	0
TVRHLQL	213	273	124	9	3	0	0	0	0	2	0	0	5	266	8	0	0	0	1	0
QTSMATV	200	203	47	35	5	0	0	0	0	2	0	0	0	0	22	0	0	0	1	0
GETRAPL	191	2868	11859	111	36	1	5	0	0	11	6	14	29	478	40	0	0	0	1	0
YLTMPTP	191	749	1254	4	4	0	71	51	0	7	4	1	18	2296	15	0	0	0	1	0
QRLPQTA	166	178	21	1	1	0	31	27	0	2	2	0	0	73	0	0	0	0	1	0
SPQMTLS	159	322	151	5	16	0	58	56	3	10	6	0	0	7	4	0	0	0	1	0
TTNLSPW	157	309	147	0	0	0	0	0	0	1	0	0	0	4	0	0	0	0	1	0
KAVHPLR	152	220	118	173	32	39	0	0	0	2	0	0	6	133	20	0	0	0	1	0
STASYTR	645	3722	4203	1	4	0	807	3490	1957	17	6	1	38	1695	53	0	0	0	1	0
GETRAPL	191	2868	11859	111	36	1	5	0	0	11	6	14	29	478	40	0	0	0	1	0
YAGPYQH	215	1028	1907	190	141	63	1	2	0	4	2	0	4	437	16	0	0	0	1	0
EPLQLKM	214	777	586	3	2	0	2	0	0	16	6	0	7	1159	6	1	0	0	1	0
YLTMPTP	191	749	1254	4	4	0	71	51	0	7	4	1	18	2296	15	0	0	0	1	0
SPWDARL	281	705	593	3	6	0	0	1	0	1	0	0	41	697	39	0	0	0	1	0
QNTTTAL	482	629	229	0	0	0	0	0	0	3	0	0	7	575	32	0	0	0	0	0
TPQSSPT	400	612	207	0	0	0	0	0	0	3	8	1	22	1408	50	0	0	0	1	0
VIPHVLS	262	569	503	101	29	7	1	0	0	4	0	0	27	207	44	0	0	0	1	0
QEPLTAR	317	549	97	0	1	0	0	0	0	0	0	0	10	413	5	0	0	0	0	0
ANTTPRH	236	523	289	0	0	0	2	8	1	12	0	3	0	475	32	0	0	0	1	0
GVKALST	236	497	295	42	7	22	0	0	0	6	1	0	5	1139	10	0	0	0	1	0
DSHTPQR	141	349	208	12	9	0	11	24	3	21	8	0	0	4	0	0	0	0	1	0
SPQMTLS	159	322	151	5	16	0	58	56	3	10	6	0	0	7	4	0	0	0	1	0
TTNLSPW	157	309	147	0	0	0	0	0	0	1	0	0	0	4	0	0	0	0	1	0
MPKYYLQ	85	284	327	16	8	0	0	0	0	0	0	0	0	560	0	0	0	0	1	0
TVRHLQL	213	273	124	9	3	0	0	0	0	2	0	0	5	266	8	0	0	0	1	0
HFRSGSL	263	266	264	0	0	0	0	0	0	4	0	1	0	563	12	0	0	0	1	0
HIPPGSP	134	264	123	12	2	0	15	10	0	3	1	1	16	38	0	0	0	0	1	0
QLHNDAT	133	239	89	0	0	1	0	0	0	2	0	0	9	23	0	0	0	0	1	0
GETRAPL	191	2868	11859	111	36	1	5	0	0	11	6	14	29	478	40	0	0	0	1	0
STASYTR	645	3722	4203	1	4	0	807	3490	1957	17	6	1	38	1695	53	0	0	0	1	0
YAGPYQH	215	1028	1907	190	141	63	1	2	0	4	2	0	4	437	16	0	0	0	1	0
YLTMPTP	191	749	1254	4	4	0	71	51	0	7	4	1	18	2296	15	0	0	0	1	0
SPWDARL	281	705	593	3	6	0	0	1	0	1	0	0	41	697	39	0	0	0	1	0
EPLQLKM	214	777	586	3	2	0	2	0	0	16	6	0	7	1159	6	1	0	0	1	0
VIPHVLS	262	569	503	101	29	7	1	0	0	4	0	0	27	207	44	0	0	0	1	0
SILPYPY	36	92	409	16	10	12	857	1488	2536	13	8	1	10	847	22	0	0	0	1	0
YAAHRSH	9	20	390	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0
QALSVYR	42	192	380	0	1	1	0	0	0	2	4	0	20	17	7	0	0	0	1	0
HAIYPRH	35	137	328	2659	3032	11887	481	661	370	99	28	2	148	8381	19	0	0	10	1	0
MPKYYLQ	85	284	327	16	8	0	0	0	0	0	0	0	0	560	0	0	0	0	1	0
GVKALST	236	497	295	42	7	22	0	0	0	6	1	0	5	1139	10	0	0	0	1	0
ANTTPRH	236	523	289	0	0	0	2	8	1	12	0	3	0	475	32	0	0	0	1	0
HFRSGSL	263	266	264	0	0	0	0	0	0	4	0	1	0	563	12	0	0	0	1	0
QNTTTAL	482	629	229	0	0	0	0	0	0	3	0	0	7	575	32	0	0	0	0	0
TKTDTWL	95	216	219	2	1	0	3	2	0	11	5	1	21	725	34	0	0	0	1	0
DSHTPQR	141	349	208	12	9	0	11	24	3	21	8	0	0	4	0	0	0	0	1	0
TPQSSPT	400	612	207	0	0	0	0	0	0	3	8	1	22	1408	50	0	0	0	1	0
SSLPLRK	53	130	175	5	2	0	0	0	0	2	1	0	0	474	15	0	0	0	1	0
HAIYPRH	35	137	328	2659	3032	11887	481	661	370	99	28	2	148	8381	19	0	0	10	1	0
QPPRSTS	23	33	11	1775	938	10032	8	4	0	19	5	0	0	847	13	0	0	0	1	0
GKPMPPM	63	208	165	1540	899	3479	254	375	287	49	11	0	3	24	0	0	0	0	1	0
QPTHPTR	0	0	0	827	524	3585	3	0	0	0	0	0	0	889	0	0	0	0	0	0
IPTLPSS	1	3	8	589	259	485	86	190	221	4	2	0	7	239	0	0	0	4	1	0
VTAHGGR	0	0	3	524	367	1240	2	1	5	1	0	0	0	0	0	0	0	0	1	0
NHWASPR	0	0	0	443	365	1431	72	189	17	9	2	0	0	233	0	0	0	0	1	0
QPSMLNP	2	0	0	348	132	1457	74	87	121	0	0	0	0	0	0	0	0	0	1	0
TWYFGPL	0	0	0	297	364	0	76	30	0	0	1	0	6	0	16	0	0	0	0	0
STPMQNL	1	0	0	242	71	528	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MDAHHAL	2	0	0	235	97	8	702	870	4	10	3	0	0	0	0	0	0	0	1	0
YAGPYQH	215	1028	1907	190	141	63	1	2	0	4	2	0	4	437	16	0	0	0	1	0
KAVHPLR	152	220	118	173	32	39	0	0	0	2	0	0	6	133	20	0	0	0	1	0
QSLALQP	9	9	1	163	20	1	2	0	0	0	0	0	0	0	0	0	0	0	1	0
ALAHRIL	0	0	0	156	130	1817	0	1	0	4	1	0	0	0	0	0	0	0	1	0
SHTAPLR	92	144	97	112	19	5	1	0	0	3	2	0	27	0	0	0	0	0	1	0
SLSLIQT	2	0	0	112	46	136	0	2	0	1	1	0	0	0	0	0	0	0	1	0
GETRAPL	191	2868	11859	111	36	1	5	0	0	11	6	14	29	478	40	0	0	0	1	0
VIPHVLS	262	569	503	101	29	7	1	0	0	4	0	0	27	207	44	0	0	0	1	0
TARYPSW	78	147	33	89	4	301	19	5	0	2	0	0	0	0	0	0	0	0	1	0

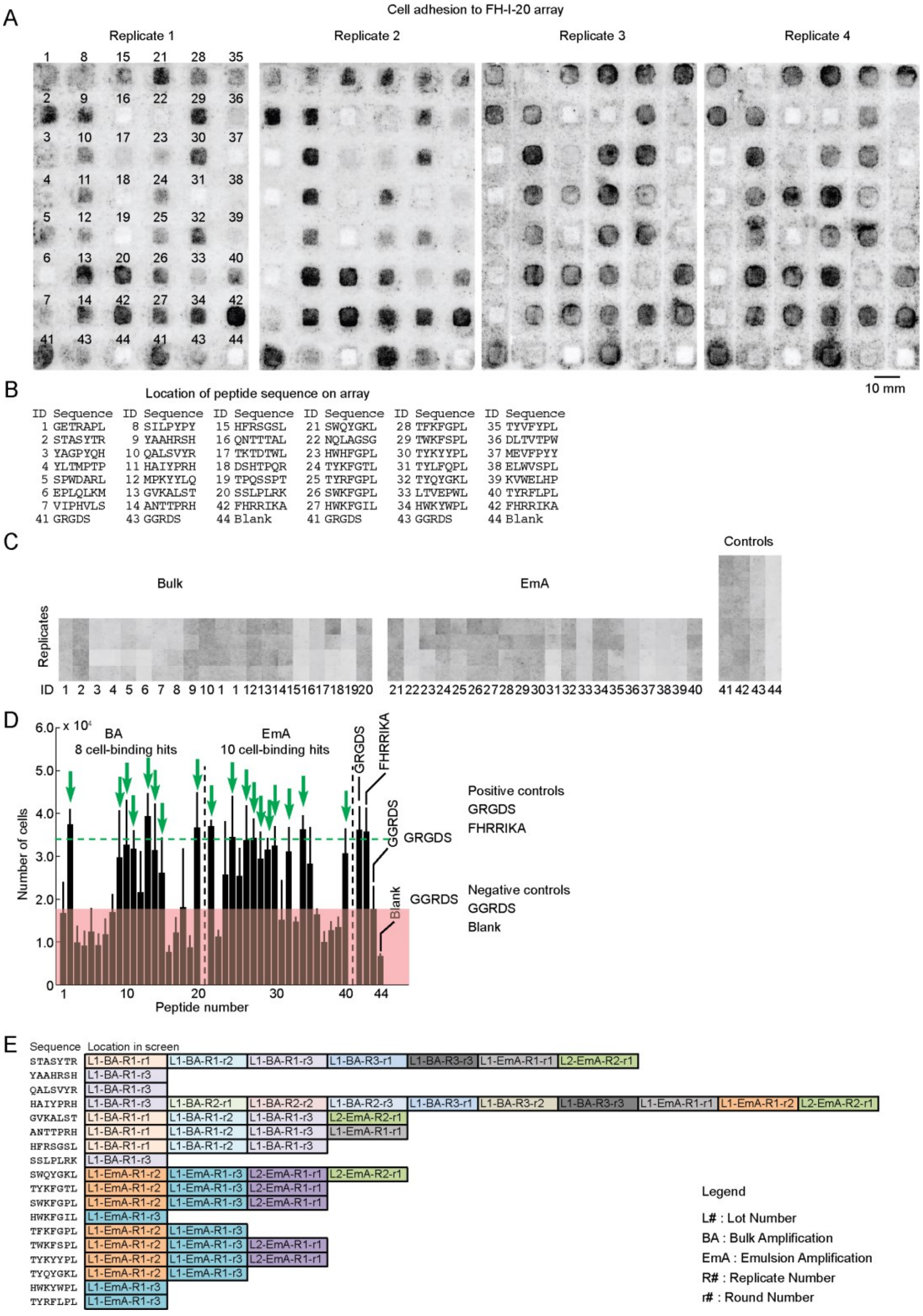
	Lot1-BA-Rep1-Round1	Lot1-BA-Rep1-Round2	Lot1-BA-Rep1-Round3	Lot1-BA-Rep2-Round1	Lot1-BA-Rep2-Round2	Lot1-BA-Rep2-Round3	Lot1-BA-Rep3-Round1	Lot1-BA-Rep3-Round2	Lot1-BA-Rep3-Round3	Lot1-EmA-Rep1-Round1	Lot1-EmA-Rep1-Round2	Lot1-EmA-Rep1-Round3	Lot2-EmA-Rep1-Round1	Lot2-EmA-Rep2-Round1	Lot2-EmA-Rep3-Round1	Lot2-BA-Rep1-Round1	Lot2-BA-Rep2-Round1	Lot2-BA-Rep3-Round1	Lot1-Parasites	Lot2-Parasites
HAIYPRH	35	137	328	2659	3032	11887	481	661	370	99	28	2	148	8381	19	0	0	10	1	0
QPPRSTS	23	33	11	1775	938	10032	8	4	0	19	5	0	0	847	13	0	0	0	1	0
GKPMPPM	63	208	165	1540	899	3479	254	375	287	49	11	0	3	24	0	0	0	0	1	0
QPTHPTR	0	0	0	827	524	3585	3	0	0	0	0	0	0	889	0	0	0	0	0	0
VTAHGGR	0	0	3	524	367	1240	2	1	5	1	0	0	0	0	0	0	0	0	1	0
NHWASPR	0	0	0	443	365	1431	72	189	17	9	2	0	0	233	0	0	0	0	1	0
TWYFGPL	0	0	0	297	364	0	76	30	0	0	1	0	6	0	16	0	0	0	0	0
IPTLPSS	1	3	8	589	259	485	86	190	221	4	2	0	7	239	0	0	0	4	1	0
YAGPYQH	215	1028	1907	190	141	63	1	2	0	4	2	0	4	437	16	0	0	0	1	0
QPSMLNP	2	0	0	348	132	1457	74	87	121	0	0	0	0	0	0	0	0	0	1	0
ALAHRIL	0	0	0	156	130	1817	0	1	0	4	1	0	0	0	0	0	0	0	1	0
MDAHHAL	2	0	0	235	97	8	702	870	4	10	3	0	0	0	0	0	0	0	1	0
STPMQNL	1	0	0	242	71	528	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SPTQPKS	34	61	31	88	64	1120	9	7	0	13	2	0	0	0	0	0	0	0	1	0
AMSSRSL	0	0	0	40	61	3	8	32	117	28	10	0	14	245	0	0	0	0	1	0
HALGPSS	7	27	3	81	60	127	0	2	0	18	8	0	0	0	10	0	0	0	1	0
MHAPPFY	0	0	1	45	52	249	2	1	0	1	0	0	0	0	0	0	0	0	1	0
SLSLIQT	2	0	0	112	46	136	0	2	0	1	1	0	0	0	0	0	0	0	1	0
SSLVRTA	13	38	35	40	44	32	26	33	0	1	1	0	0	209	22	0	0	0	1	0
WSPHGLA	0	0	0	36	43	0	10	4	0	0	0	0	0	0	0	0	0	0	0	0
HAIYPRH	35	137	328	2659	3032	11887	481	661	370	99	28	2	148	8381	19	0	0	10	1	0
QPPRSTS	23	33	11	1775	938	10032	8	4	0	19	5	0	0	847	13	0	0	0	1	0
QPTHPTR	0	0	0	827	524	3585	3	0	0	0	0	0	0	889	0	0	0	0	0	0
GKPMPPM	63	208	165	1540	899	3479	254	375	287	49	11	0	3	24	0	0	0	0	1	0
ALAHRIL	0	0	0	156	130	1817	0	1	0	4	1	0	0	0	0	0	0	0	1	0
QPSMLNP	2	0	0	348	132	1457	74	87	121	0	0	0	0	0	0	0	0	0	1	0
NHWASPR	0	0	0	443	365	1431	72	189	17	9	2	0	0	233	0	0	0	0	1	0
VTAHGGR	0	0	3	524	367	1240	2	1	5	1	0	0	0	0	0	0	0	0	1	0
SPTQPKS	34	61	31	88	64	1120	9	7	0	13	2	0	0	0	0	0	0	0	1	0
STPMQNL	1	0	0	242	71	528	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HHSLTVT	0	0	2	9	9	492	0	0	0	2	0	0	0	0	0	0	0	0	1	0
IPTLPSS	1	3	8	589	259	485	86	190	221	4	2	0	7	239	0	0	0	4	1	0
VLPGRSP	0	0	0	65	28	309	0	0	0	1	0	0	0	0	0	0	0	0	1	0
TARYPSW	78	147	33	89	4	301	19	5	0	2	0	0	0	0	0	0	0	0	1	0
MHAPPFY	0	0	1	45	52	249	2	1	0	1	0	0	0	0	0	0	0	0	1	0
QLMNASR	0	0	0	17	4	167	2	0	0	2	2	0	0	0	0	0	0	0	1	0
STFTKSP	1	0	4	27	20	141	83	335	626	7	3	0	0	0	0	0	0	0	1	0
SLSLIQT	2	0	0	112	46	136	0	2	0	1	1	0	0	0	0	0	0	0	1	0
HALGPSS	7	27	3	81	60	127	0	2	0	18	8	0	0	0	10	0	0	0	1	0
TPPTMDH	28	46	40	39	16	125	1	1	0	4	3	0	0	67	0	0	0	0	1	0
MGLQTPY	0	0	0	0	0	0	1133	2162	564	0	1	0	0	0	0	0	0	0	0	0
SILPYPY	36	92	409	16	10	12	857	1488	2536	13	8	1	10	847	22	0	0	0	1	0
STASYTR	645	3722	4203	1	4	0	807	3490	1957	17	6	1	38	1695	53	0	0	0	1	0
NQLPLHA	24	54	34	2	3	0	806	1085	879	10	4	1	0	312	12	0	0	0	1	0
HSTKVAF	0	0	0	0	0	0	772	1534	256	0	0	0	0	0	0	0	0	0	0	0
MDAHHAL	2	0	0	235	97	8	702	870	4	10	3	0	0	0	0	0	0	0	1	0
QPWPTSI	6	24	10	16	20	0	556	888	2104	1	1	0	0	0	19	0	0	1	1	0
IPAPLRS	0	0	0	0	0	0	496	1219	535	0	0	0	0	983	0	0	0	0	0	0
HAIYPRH	35	137	328	2659	3032	11887	481	661	370	99	28	2	148	8381	19	0	0	10	1	0
QAHTVGK	0	0	0	0	0	0	452	783	35	0	0	0	0	84	0	0	0	0	0	0
TGHSAQG	0	0	0	0	0	0	336	790	2946	1	0	0	25	78	0	0	0	0	0	0
MPTLTPT	0	0	0	0	0	0	330	147	251	2	0	0	0	0	0	0	0	0	1	0
SLHQPHL	0	0	0	0	0	0	316	651	9	6	0	0	0	0	0	0	0	0	1	0
LPRTPTD	0	0	0	0	0	0	260	355	44	2	1	0	3	0	0	0	0	0	1	0
GKPMPPM	63	208	165	1540	899	3479	254	375	287	49	11	0	3	24	0	0	0	0	1	0
SPTGWAP	0	0	0	1	0	0	226	417	1088	1	0	1	2	29	7	0	0	0	1	0
TLLPFQP	0	0	0	0	0	0	216	124	55	0	0	0	8	195	0	0	0	0	0	0
FPSTITP	17	8	21	2	5	105	179	280	1309	2	1	0	0	265	12	0	0	0	1	0
AGNGTTP	0	0	0	2	2	0	153	186	133	0	1	0	0	0	0	0	0	0	1	0
QLHMDYR	6	26	29	0	2	0	145	78	1	3	7	0	0	31	0	0	0	0	1	0
STASYTR	645	3722	4203	1	4	0	807	3490	1957	17	6	1	38	1695	53	0	0	0	1	0
MGLQTPY	0	0	0	0	0	0	1133	2162	564	0	1	0	0	0	0	0	0	0	0	0
HSTKVAF	0	0	0	0	0	0	772	1534	256	0	0	0	0	0	0	0	0	0	0	0
SILPYPY	36	92	409	16	10	12	857	1488	2536	13	8	1	10	847	22	0	0	0	1	0
IPAPLRS	0	0	0	0	0	0	496	1219	535	0	0	0	0	983	0	0	0	0	0	0
NQLPLHA	24	54	34	2	3	0	806	1085	879	10	4	1	0	312	12	0	0	0	1	0
QPWPTSI	6	24	10	16	20	0	556	888	2104	1	1	0	0	0	19	0	0	1	1	0
MDAHHAL	2	0	0	235	97	8	702	870	4	10	3	0	0	0	0	0	0	0	1	0
TGHSAQG	0	0	0	0	0	0	336	790	2946	1	0	0	25	78	0	0	0	0	0	0
QAHTVGK	0	0	0	0	0	0	452	783	35	0	0	0	0	84	0	0	0	0	0	0
HAIYPRH	35	137	328	2659	3032	11887	481	661	370	99	28	2	148	8381	19	0	0	10	1	0
SLHQPHL	0	0	0	0	0	0	316	651	9	6	0	0	0	0	0	0	0	0	1	0
STTKLAL	0	0	0	0	1	0	128	456	22	1	0	0	0	0	0	0	0	0	1	0
SPTGWAP	0	0	0	1	0	0	226	417	1088	1	0	1	2	29	7	0	0	0	1	0
GKPMPPM	63	208	165	1540	899	3479	254	375	287	49	11	0	3	24	0	0	0	0	1	0
LPRTPTD	0	0	0	0	0	0	260	355	44	2	1	0	3	0	0	0	0	0	1	0
APRTFNQ	0	0	0	0	0	0	67	343	3528	0	0	0	6	98	0	0	0	0	1	0
STFTKSP	1	0	4	27	20	141	83	335	626	7	3	0	0	0	0	0	0	0	1	0
SYHSFNL	0	0	0	0	0	0	120	334	9361	0	3	0	0	339	0	0	0	0	1	0
SHSLLHH	0	0	0	0	1	0	114	333	691	7	0	0	0	196	4	0	0	0	1	0

	Lot1-BA-Rep1-Round1	Lot1-BA-Rep1-Round2	Lot1-BA-Rep1-Round3	Lot1-BA-Rep2-Round1	Lot1-BA-Rep2-Round2	Lot1-BA-Rep2-Round3	Lot1-BA-Rep3-Round1	Lot1-BA-Rep3-Round2	Lot1-BA-Rep3-Round3	Lot1-EmA-Rep1-Round1	Lot1-EmA-Rep1-Round2	Lot1-EmA-Rep1-Round3	Lot2-EmA-Rep1-Round1	Lot2-EmA-Rep2-Round1	Lot2-EmA-Rep3-Round1	Lot2-BA-Rep1-Round1	Lot2-BA-Rep2-Round1	Lot2-BA-Rep3-Round1	Lot1-Parasites	Lot2-Parasites
SYHSFNL	0	0	0	0	0	0	120	334	9361	0	3	0	0	339	0	0	0	0	1	0
APRTFNQ	0	0	0	0	0	0	67	343	3528	0	0	0	6	98	0	0	0	0	1	0
TGHSAQG	0	0	0	0	0	0	336	790	2946	1	0	0	25	78	0	0	0	0	0	0
SILPYPY	36	92	409	16	10	12	857	1488	2536	13	8	1	10	847	22	0	0	0	1	0
QPWPTSI	6	24	10	16	20	0	556	888	2104	1	1	0	0	0	19	0	0	1	1	0
STASYTR	645	3722	4203	1	4	0	807	3490	1957	17	6	1	38	1695	53	0	0	0	1	0
HTIQFTP	0	0	0	10	14	17	18	15	1564	4	0	0	5	18	0	0	0	0	1	0
FPSTITP	17	8	21	2	5	105	179	280	1309	2	1	0	0	265	12	0	0	0	1	0
ASYSGTA	15	18	23	33	26	9	19	55	1121	8	3	1	3	227	0	0	0	0	1	0
SPTGWAP	0	0	0	1	0	0	226	417	1088	1	0	1	2	29	7	0	0	0	1	0
NQLPLHA	24	54	34	2	3	0	806	1085	879	10	4	1	0	312	12	0	0	0	1	0
SHSLLHH	0	0	0	0	1	0	114	333	691	7	0	0	0	196	4	0	0	0	1	0
STFTKSP	1	0	4	27	20	141	83	335	626	7	3	0	0	0	0	0	0	0	1	0
MGLQTPY	0	0	0	0	0	0	1133	2162	564	0	1	0	0	0	0	0	0	0	0	0
IPAPLRS	0	0	0	0	0	0	496	1219	535	0	0	0	0	983	0	0	0	0	0	0
HAIYPRH	35	137	328	2659	3032	11887	481	661	370	99	28	2	148	8381	19	0	0	10	1	0
STPIQQP	5	6	2	0	0	0	25	37	354	8	4	0	0	246	0	0	0	0	1	0
SHHQKPP	0	0	0	0	0	0	6	10	294	1	0	0	0	0	0	0	0	0	0	0
GKPMPPM	63	208	165	1540	899	3479	254	375	287	49	11	0	3	24	0	0	0	0	1	0
HSTKVAF	0	0	0	0	0	0	772	1534	256	0	0	0	0	0	0	0	0	0	0	0
HAIYPRH	35	137	328	2659	3032	11887	481	661	370	99	28	2	148	8381	19	0	0	10	1	0
GKPMPPM	63	208	165	1540	899	3479	254	375	287	49	11	0	3	24	0	0	0	0	1	0
GPMLARG	112	166	101	0	0	0	0	0	0	34	10	0	0	43	0	0	0	0	1	0
AMSSRSL	0	0	0	40	61	3	8	32	117	28	10	0	14	245	0	0	0	0	1	0
SSALLLP	0	0	0	0	0	0	0	0	0	24	10	2	0	0	0	0	0	0	0	0
DSHTPQR	141	349	208	12	9	0	11	24	3	21	8	0	0	4	0	0	0	0	1	0
QPPRSTS	23	33	11	1775	938	10032	8	4	0	19	5	0	0	847	13	0	0	0	1	0
AASSLTI	0	0	0	0	0	0	0	0	0	19	8	0	0	0	0	0	0	0	0	0
HALGPSS	7	27	3	81	60	127	0	2	0	18	8	0	0	0	10	0	0	0	1	0
STASYTR	645	3722	4203	1	4	0	807	3490	1957	17	6	1	38	1695	53	0	0	0	1	0
EPLQLKM	214	777	586	3	2	0	2	0	0	16	6	0	7	1159	6	1	0	0	1	0
QATHRSH	128	229	147	23	8	1	6	3	0	14	8	1	7	556	0	0	0	0	1	0
GKVQAQS	57	51	39	0	1	0	28	41	0	14	7	0	0	439	0	0	0	0	1	0
SSSVVTH	111	92	0	0	0	0	0	0	0	14	0	0	0	0	0	0	0	0	0	0
SILPYPY	36	92	409	16	10	12	857	1488	2536	13	8	1	10	847	22	0	0	0	1	0
SPTQPKS	34	61	31	88	64	1120	9	7	0	13	2	0	0	0	0	0	0	0	1	0
ANTTPRH	236	523	289	0	0	0	2	8	1	12	0	3	0	475	32	0	0	0	1	0
QTGYATR	0	0	0	0	0	0	0	0	0	12	10	3	0	0	0	0	0	0	0	0
TYQNPVH	0	0	0	0	0	0	0	0	0	12	6	2	0	0	0	0	0	0	0	0
NQLAGSG	0	0	0	1	1	0	1	1	1	11	1984	21310	4055	83	19	0	0	0	0	0
SWQYGKL	0	0	0	1	0	0	0	2	0	9	3143	22060	5814	848	42	0	1	0	0	0
NQLAGSG	0	0	0	1	1	0	1	1	1	11	1984	21310	4055	83	19	0	0	0	0	0
HWHFGPL	0	0	0	0	0	0	1	1	0	3	228	1929	510	0	26	0	0	0	0	0
TYRFGPL	0	0	0	0	0	0	0	0	0	0	136	923	347	12	0	0	0	0	0	0
SWKFGPL	0	0	0	0	0	0	0	2	0	1	109	899	296	15	0	0	0	0	0	0
ALEVTFW	0	0	0	0	0	0	0	0	0	3	80	25	9	0	0	0	0	0	0	0
TYKFGTL	0	0	0	0	0	0	0	0	0	2	78	1107	545	0	7	0	0	0	0	0
VQNEWRS	0	0	0	0	0	0	0	0	0	4	70	21	13	0	0	0	0	0	0	0
LTVEPWL	0	0	0	0	0	0	0	0	0	2	63	118	108	0	0	1	0	0	0	0
ELWVSPL	0	0	0	0	0	0	0	0	0	2	54	28	21	0	0	0	0	0	0	0
LEVYALV	0	0	0	0	0	0	0	0	0	2	45	16	24	0	0	0	0	0	0	0
TYKYYPL	0	0	0	0	0	0	0	0	0	1	43	177	169	1	0	0	0	0	0	0
TFKFGPL	0	0	0	0	0	0	0	0	0	0	41	257	123	0	0	0	0	0	0	0
TYQYGKL	0	0	0	0	0	0	0	0	0	1	40	122	107	0	12	0	0	0	0	0
TYLFQPL	0	0	0	0	0	0	0	0	0	1	35	172	146	12	0	0	0	0	0	0
DLTVTPW	0	0	0	0	0	0	0	0	0	0	35	41	36	0	0	0	0	0	0	0
TWKFSPL	0	0	0	0	0	0	0	0	0	1	34	187	183	0	0	0	0	0	0	0
QLTVMSW	0	0	0	0	0	0	0	0	0	0	34	21	13	0	0	0	0	0	0	0
LTVQPWP	0	0	0	0	0	0	0	1	0	2	32	17	17	0	0	0	0	0	0	0
HAIYPRH	35	137	328	2659	3032	11887	481	661	370	99	28	2	148	8381	19	0	0	10	1	0
SWQYGKL	0	0	0	1	0	0	0	2	0	9	3143	22060	5814	848	42	0	1	0	0	0
NQLAGSG	0	0	0	1	1	0	1	1	1	11	1984	21310	4055	83	19	0	0	0	0	0
HWHFGPL	0	0	0	0	0	0	1	1	0	3	228	1929	510	0	26	0	0	0	0	0
TYKFGTL	0	0	0	0	0	0	0	0	0	2	78	1107	545	0	7	0	0	0	0	0
TYRFGPL	0	0	0	0	0	0	0	0	0	0	136	923	347	12	0	0	0	0	0	0
SWKFGPL	0	0	0	0	0	0	0	2	0	1	109	899	296	15	0	0	0	0	0	0
HWKFGIL	0	0	0	0	0	0	0	0	0	0	3	556	51	0	3	0	0	0	0	0
TFKFGPL	0	0	0	0	0	0	0	0	0	0	41	257	123	0	0	0	0	0	0	0
TWKFSPL	0	0	0	0	0	0	0	0	0	1	34	187	183	0	0	0	0	0	0	0
TYKYYPL	0	0	0	0	0	0	0	0	0	1	43	177	169	1	0	0	0	0	0	0
TYLFQPL	0	0	0	0	0	0	0	0	0	1	35	172	146	12	0	0	0	0	0	0
TYQYGKL	0	0	0	0	0	0	0	0	0	1	40	122	107	0	12	0	0	0	0	0
LTVEPWL	0	0	0	0	0	0	0	0	0	2	63	118	108	0	0	1	0	0	0	0
HWKYWPL	0	0	0	0	0	0	0	0	0	1	12	81	35	0	0	0	0	0	0	0
TYVFYPL	0	0	0	0	0	0	0	0	0	0	18	44	40	0	0	0	0	0	0	0
DLTVTPW	0	0	0	0	0	0	0	0	0	0	35	41	36	0	0	0	0	0	0	0
LEVFPYY	0	0	0	0	0	0	0	0	0	1	24	35	15	0	0	0	0	0	0	0
ELWVSPL	0	0	0	0	0	0	0	0	0	2	54	28	21	0	0	0	0	0	0	0
KVWELHP	0	0	0	0	0	0	0	0	0	0	6	27	17	0	0	0	0	0	0	0
TYRFLPL	0	0	0	0	0	0	0	0	0	1	2	26	11	0	0	0	0	0	0	0

	Lot1-BA-Rep1-Round1	Lot1-BA-Rep1-Round2	Lot1-BA-Rep1-Round3	Lot1-BA-Rep2-Round1	Lot1-BA-Rep2-Round2	Lot1-BA-Rep2-Round3	Lot1-BA-Rep3-Round1	Lot1-BA-Rep3-Round2	Lot1-BA-Rep3-Round3	Lot1-EmA-Rep1-Round1	Lot1-EmA-Rep1-Round2	Lot1-EmA-Rep1-Round3	Lot2-EmA-Rep1-Round1	Lot2-EmA-Rep2-Round1	Lot2-EmA-Rep3-Round1	Lot2-BA-Rep1-Round1	Lot2-BA-Rep2-Round1	Lot2-BA-Rep3-Round1	Lot1-Parasites	Lot2-Parasites
SWQYGKL	0	0	0	1	0	0	0	2	0	9	3143	22060	5814	848	42	0	1	0	0	0
NQLAGSG	0	0	0	1	1	0	1	1	1	11	1984	21310	4055	83	19	0	0	0	0	0
TSSESES	0	0	0	0	0	0	0	0	0	0	0	0	665	0	0	0	0	0	0	0
ERTVLHT	0	0	0	0	0	0	0	0	0	0	0	0	554	0	0	0	0	0	0	0
TYKFGTL	0	0	0	0	0	0	0	0	0	2	78	1107	545	0	7	0	0	0	0	0
HWHFGPL	0	0	0	0	0	0	1	1	0	3	228	1929	510	0	26	0	0	0	0	0
TYRFGPL	0	0	0	0	0	0	0	0	0	0	136	923	347	12	0	0	0	0	0	0
RFTVDWD	0	0	0	0	0	0	0	0	0	0	0	0	309	0	0	0	0	0	0	0
SWKFGPL	0	0	0	0	0	0	0	2	0	1	109	899	296	15	0	0	0	0	0	0
LAGPLMT	0	0	0	0	0	0	0	0	0	0	0	0	192	0	0	0	0	0	0	0
TWKFSPL	0	0	0	0	0	0	0	0	0	1	34	187	183	0	0	0	0	0	0	0
NVSGSHS	0	0	0	0	0	0	0	0	0	0	0	0	180	0	0	0	0	0	0	0
SVLLPHR	0	0	0	0	0	0	0	0	0	0	0	0	177	0	0	0	0	0	0	0
TYKYYPL	0	0	0	0	0	0	0	0	0	1	43	177	169	1	0	0	0	0	0	0
HHQYVPA	0	0	0	0	0	0	0	0	0	0	0	0	161	0	0	0	0	0	0	0
WTTTSRL	0	0	0	0	0	0	0	0	0	0	0	0	161	0	0	0	0	0	0	0
QLMPMMM	0	0	0	0	0	0	0	0	0	0	0	0	159	0	0	0	0	0	0	0
RPYDTAH	0	0	0	0	0	0	0	0	0	0	0	0	157	0	0	0	0	0	0	0
SAGSMHL	0	0	0	0	0	0	0	0	0	0	0	0	154	0	0	0	0	0	0	0
GLYHSAT	0	0	0	0	0	0	0	0	0	0	0	0	153	0	0	0	0	0	0	0
HAIYPRH	35	137	328	2659	3032	11887	481	661	370	99	28	2	148	8381	19	0	0	10	1	0
YLTMPTP	191	749	1254	4	4	0	71	51	0	7	4	1	18	2296	15	0	0	0	1	0
STASYTR	645	3722	4203	1	4	0	807	3490	1957	17	6	1	38	1695	53	0	0	0	1	0
TPQSSPT	400	612	207	0	0	0	0	0	0	3	8	1	22	1408	50	0	0	0	1	0
EPLQLKM	214	777	586	3	2	0	2	0	0	16	6	0	7	1159	6	1	0	0	1	0
GVKALST	236	497	295	42	7	22	0	0	0	6	1	0	5	1139	10	0	0	0	1	0
TPFMAYH	10	15	13	0	0	0	0	0	0	3	0	0	0	1083	0	0	0	0	1	0
IPAPLRS	0	0	0	0	0	0	496	1219	535	0	0	0	0	983	0	0	0	0	0	0
RLPSWHE	34	25	6	0	0	0	6	1	8	4	0	0	0	912	0	0	0	0	1	0
AYPEPYV	0	0	0	0	0	0	0	0	0	1	0	0	0	910	0	0	0	0	0	0
QPTHPTR	0	0	0	827	524	3585	3	0	0	0	0	0	0	889	0	0	0	0	0	0
SWQYGKL	0	0	0	1	0	0	0	2	0	9	3143	22060	5814	848	42	0	1	0	0	0
QPPRSTS	23	33	11	1775	938	10032	8	4	0	19	5	0	0	847	13	0	0	0	1	0
SILPYPY	36	92	409	16	10	12	857	1488	2536	13	8	1	10	847	22	0	0	0	1	0
TQTMRST	92	126	48	0	0	0	0	0	0	0	0	0	11	749	0	0	0	0	0	0
TKTDTWL	95	216	219	2	1	0	3	2	0	11	5	1	21	725	34	0	0	0	1	0
TPMTRAL	4	0	0	0	0	0	0	0	0	0	0	0	0	711	0	0	0	0	0	0
VMSQPHP	0	0	0	0	0	0	0	0	0	0	0	0	0	708	0	0	0	0	0	0
SPWDARL	281	705	593	3	6	0	0	1	0	1	0	0	41	697	39	0	0	0	1	0
GSTVFTA	0	0	0	0	0	0	0	1	0	0	0	0	0	675	0	0	0	0	0	0
EQGRPLP	0	0	0	0	0	0	0	0	0	0	0	0	0	0	349	0	0	0	0	0
MAANGAR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	262	0	0	0	0	0
QVLLTAA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	243	0	0	0	0	0
AGRELCC	0	0	0	0	0	0	0	0	0	0	0	0	0	0	220	0	0	0	0	0
ARAVLQL	0	0	0	0	0	0	0	0	0	0	0	0	0	0	191	0	0	0	0	0
QNMQQQI	0	0	0	0	0	0	0	0	0	0	0	0	0	0	190	0	0	0	0	0
AWSAVMR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	190	0	0	0	0	0
APIWMHV	0	0	0	0	0	0	0	0	0	0	0	0	0	0	186	0	0	0	0	0
ATWQLGT	0	0	0	0	0	0	0	0	0	0	0	0	0	0	181	0	0	0	0	0
ASWIPLP	0	0	0	0	0	0	0	0	0	0	0	0	0	0	172	0	0	0	0	0
QQQYMAH	0	0	0	0	0	0	0	0	0	0	0	0	0	0	164	0	0	0	0	0
STAMDGR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	163	0	0	0	0	0
GWRTTWP	0	0	0	0	0	0	0	0	0	0	0	0	0	0	161	0	0	0	0	0
WMASMAV	0	0	0	0	0	0	0	0	0	0	0	0	0	0	141	0	0	0	0	0
HEQPMHR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	132	0	0	0	0	0
SLDVRMW	0	0	0	0	0	0	0	0	0	0	0	0	4	19	126	931	25	0	0	1
QHMPLTR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	124	0	0	0	0	0
GEELPTL	0	0	0	0	0	0	0	0	0	0	0	0	0	0	118	0	0	0	0	0
HAMWFSV	0	0	0	0	0	0	0	0	0	0	0	0	0	0	117	0	0	0	0	0
APSGLSR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	114	0	0	0	0	0
STPATLI	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3298	10	0	0	1
WSLSELH	0	0	0	0	0	0	0	0	0	0	0	0	3	8	21	3240	151	0	0	1
LPVRLDW	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3042	0	0	0	1
QTWLEMG	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2864	0	4	0	1
GPHNPTQ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2807	0	0	0	1
NDRPHMP	0	0	0	0	0	0	0	0	0	0	0	0	32	0	1	2565	0	0	0	1
VPNIVTQ	0	0	0	0	0	0	0	0	0	0	0	0	48	0	5	1972	38	0	0	1
LRSDPVV	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1921	0	0	0	1
VPASPWT	0	0	0	0	0	0	0	0	0	0	0	0	0	0	31	1753	21	0	0	1
SSVSWLN	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1586	0	0	0	1
TTQVLEA	0	0	0	0	0	0	0	0	0	0	0	0	10	0	103	1413	6	1693	0	1
DAIPTSV	0	0	0	0	0	0	0	0	0	0	0	0	10	40	0	1197	39	38	0	1
VGKTSFQ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1067	1	0	0	1
SLDVRMW	0	0	0	0	0	0	0	0	0	0	0	0	4	19	126	931	25	0	0	1
GQVALLD	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	850	0	0	0	0
VENVHVR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	848	0	0	0	1
VPVTMYW	0	0	0	0	0	0	0	0	0	0	0	0	6	0	0	835	0	10	0	1
ELGTTQT	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	823	0	0	0	0
AMTALDL	0	0	0	0	0	0	0	0	0	0	0	0	0	7	0	745	0	5	0	1
MAVTQKY	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	707	0	0	0	1

	Lot1-BA-Rep1-Round1	Lot1-BA-Rep1-Round2	Lot1-BA-Rep1-Round3	Lot1-BA-Rep2-Round1	Lot1-BA-Rep2-Round2	Lot1-BA-Rep2-Round3	Lot1-BA-Rep3-Round1	Lot1-BA-Rep3-Round2	Lot1-BA-Rep3-Round3	Lot1-EmA-Rep1-Round1	Lot1-EmA-Rep1-Round2	Lot1-EmA-Rep1-Round3	Lot2-EmA-Rep1-Round1	Lot2-EmA-Rep2-Round1	Lot2-EmA-Rep3-Round1	Lot2-BA-Rep1-Round1	Lot2-BA-Rep2-Round1	Lot2-BA-Rep3-Round1	Lot1-Parasites	Lot2-Parasites
AGSVIDT	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4722	396	0	1
QAYHVSA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4028	0	0	0
SNMTRWH	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3593	0	0	1
GRLDTGI	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	3058	0	0	1
NAYGGRI	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2829	0	0	1
TKTVLER	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	98	2578	93	0	1
GWETRME	0	0	0	0	0	0	0	0	0	0	0	0	39	0	0	702	2251	43	0	1
WNQRATG	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1783	0	0	1
VDMIVPS	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	1613	0	0	1
LPGNRLL	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1510	0	0	0
QLYREFN	0	0	0	0	0	0	0	0	0	0	0	0	110	0	2	328	1481	582	0	1
GTSTTAQ	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1137	0	0	1
NTQLHPS	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1124	0	0	1
YRNHVTY	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1100	0	0	0
MNSNIPI	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	865	0	0	1
ATLVPAA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	842	0	0	1
ILAHSIM	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	801	0	0	0
IDGNGTH	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	721	0	0	0
IDNSHTH	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	490	714	0	0	1
WGRISHV	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0	699	0	0	1
HSRAPER	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	13100	0	1
TTLGVWT	0	0	0	0	0	0	0	0	0	0	0	0	0	0	19	0	0	8649	0	1
YSEPAVT	0	0	0	0	0	0	0	0	0	0	0	0	56	0	1	0	0	6577	0	1
ESRVMSR	0	0	0	0	0	0	0	0	0	0	0	0	0	8	10	360	0	4618	0	1
GPLHAQF	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	3220	0	1
NNTLSRT	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2417	0	1
DHAVPRY	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	40	0	2343	0	1
AFPPVTA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2038	0	0
MTVQRGP	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2029	0	0
HLNQQNH	0	0	0	0	0	0	0	0	0	0	0	0	1	0	22	44	0	1929	0	1
QLPFTIK	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1840	0	1
SQPTWMF	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1710	0	1
YNGSANQ	0	0	0	0	0	0	0	0	0	0	0	0	57	0	1	19	0	1694	0	1
TTQVLEA	0	0	0	0	0	0	0	0	0	0	0	0	10	0	103	1413	6	1693	0	1
FSLQTTR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1415	0	1
GLRNPPS	0	0	0	0	0	0	0	0	0	0	0	0	18	0	0	2	0	1114	0	1
TEKFRVT	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	882	0	1
TVISQNM	0	0	0	0	0	0	0	0	0	0	0	0	8	0	0	15	0	854	0	1
AFTTSYM	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	727	0	1
MSVITKP	0	0	0	0	0	0	0	0	0	0	0	0	20	26	6	253	19	688	0	0

Appendix 1.18. List of top 20 sequences from each round and replicate of selection using both BA and EmA methods. The color for each selection is used to distinguish between different screens. The numbers indicate the copy number of each sequence of the top 20. The sequences are ordered by decreasing copy number in the corresponding screen. If sequences are found in other screens, the corresponding copy number is indicated. The last two columns identify if the sequence is a parasite ('1') or not ('0') for the two studied lots.



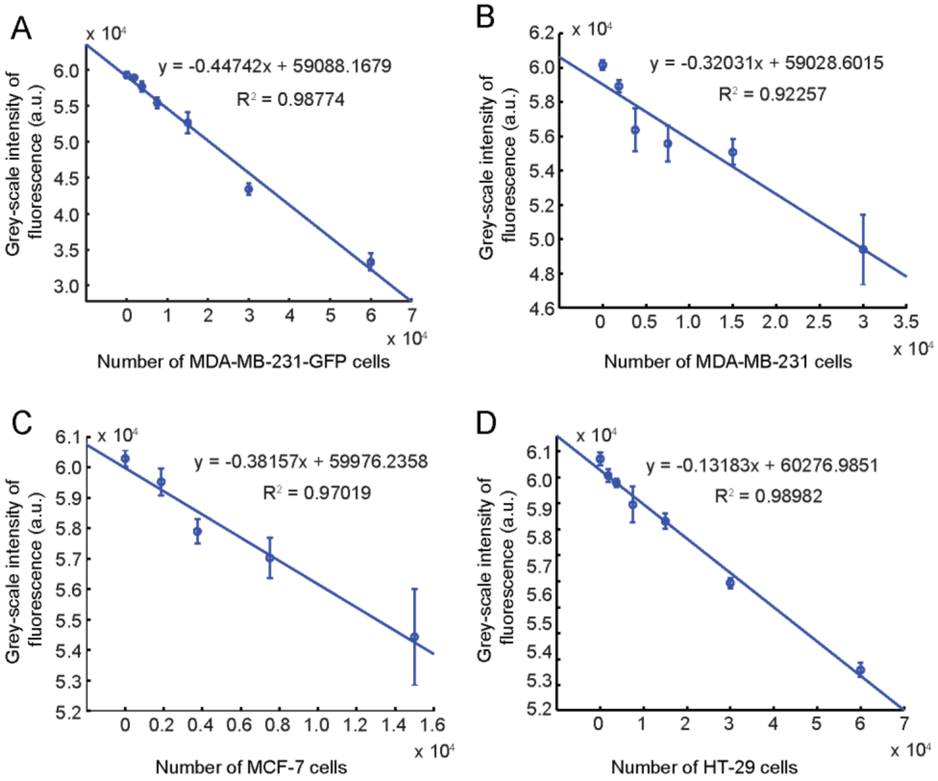
Location of peptide sequence on array

ID	Sequence	ID	Sequence	ID	Sequence	ID	Sequence	ID	Sequence	ID	Sequence
1	GETRAPL	8	SILPYPY	15	HFRSGSL	21	SWQYGKL	28	TFKFGPL	35	TYVFYPL
2	STASYTR	9	YAAHRSH	16	QNTTTAL	22	NQLAGSG	29	TWKFSPL	36	DLTVTPW
3	YAGPYQH	10	QALSVYR	17	TKTDTWL	23	HWHFGPL	30	TYKYYPL	37	MEVFPYY
4	YLTMPTP	11	HAIYPRH	18	DSHTPQR	24	TYKFGTL	31	TYLFQPL	38	ELWVSPL
5	SPWDARL	12	MPKYYLQ	19	TPQSSPT	25	TYRFGPL	32	TYQYGKL	39	KVWELHP
6	EPLQLKM	13	GVKALST	20	SSLPLRK	26	SWKFGPL	33	LTVEPWL	40	TYRFLPL
7	VIPHVLS	14	ANTTPRH	42	FHRRIKA	27	HWKFGIL	34	HWKYWPL	42	FHRRIKA
41	GRGDS	43	GGRDS	44	Blank	41	GRGDS	43	GGRDS	44	Blank

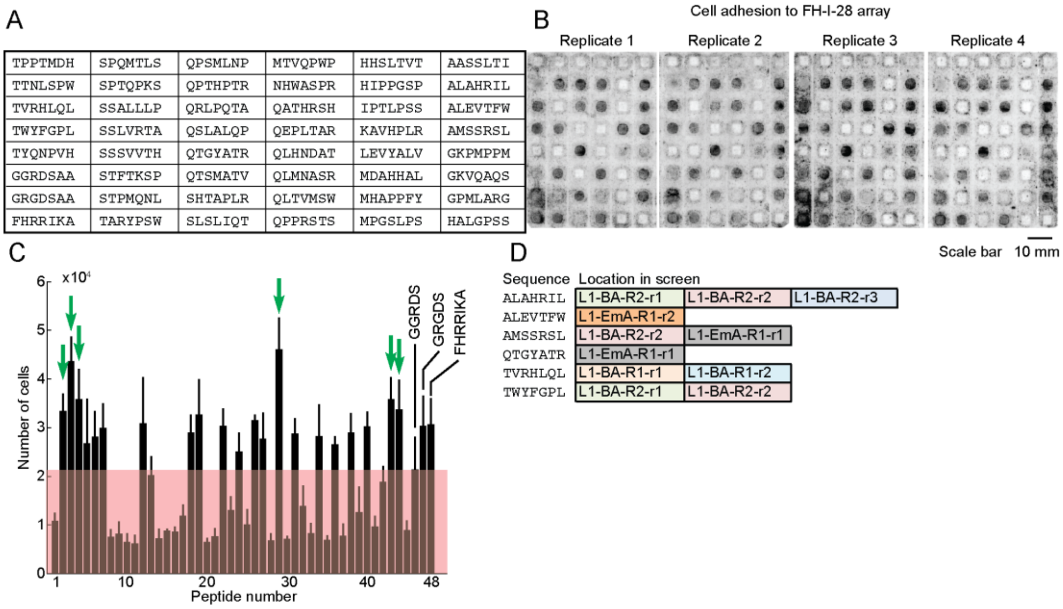
Sequence	Location in screen									
STASYTR	L1-BA-R1-r1	L1-BA-R1-r2	L1-BA-R1-r3	L1-BA-R3-r1	L1-BA-R3-r3	L1-EmA-R1-r1	L2-EmA-R2-r1			
YAAHRSH	L1-BA-R1-r3									
QALSVYR	L1-BA-R1-r3									
HAIYPRH	L1-BA-R1-r3	L1-BA-R2-r1	L1-BA-R2-r2	L1-BA-R2-r3	L1-BA-R3-r1	L1-BA-R3-r2	L1-BA-R3-r3	L1-EmA-R1-r1	L1-EmA-R1-r2	L2-EmA-R2-r1
GVKALST	L1-BA-R1-r1	L1-BA-R1-r2	L1-BA-R1-r3	L2-EmA-R2-r1						
ANTTPRH	L1-BA-R1-r1	L1-BA-R1-r2	L1-BA-R1-r3	L1-EmA-R1-r1						
HFRSGSL	L1-BA-R1-r1	L1-BA-R1-r2	L1-BA-R1-r3							
SSLPLRK	L1-BA-R1-r3									
SWQYGKL	L1-EmA-R1-r2	L1-EmA-R1-r3	L2-EmA-R1-r1	L2-EmA-R2-r1						
TYKFGTL	L1-EmA-R1-r2	L1-EmA-R1-r3	L2-EmA-R1-r1							
SWKFGPL	L1-EmA-R1-r2	L1-EmA-R1-r3	L2-EmA-R1-r1							
HWKFGIL	L1-EmA-R1-r3									
TFKFGPL	L1-EmA-R1-r2	L1-EmA-R1-r3								
TWKFSPL	L1-EmA-R1-r2	L1-EmA-R1-r3	L2-EmA-R1-r1							
TYKYYPL	L1-EmA-R1-r2	L1-EmA-R1-r3	L2-EmA-R1-r1							
TYQYGKL	L1-EmA-R1-r2	L1-EmA-R1-r3								
HWKYWPL	L1-EmA-R1-r3									
TYRFLPL	L1-EmA-R1-r3									

Appendix 1.19. Work flow of cell adhesion analysis. (A) Top 20 sequences identified from selection with BA and EmA are synthesized on paper. MDA-MB-231-GFP cells are seeded onto the array and incubated at 37°C in a CO_2 incubator. After 3 hours we imaged the arrays with a fluorescent gel scanner. Grey-scale intensities were adjusted in this figure to simplify visualization (same

level for all images). Original .gel files were used in processing without any adjustments to grey-scale intensities. (B) The location of each peptide sequence. (C) Matlab script identified the middle of each peptide zone, extracted the grey-scale intensity and organized the replicates. (D) The average greyscale intensity from each replicate is calculated and correlated to the number of cells from a standard curve (Appendix 1.20). Data represent an average from 4-8 experiments; error bar is one standard deviation. Cell-binding hit peptides are determined as binding significantly more ($p < 0.05$) cells than the negative control (GGRDS, red box). Green arrays indicate eighteen cell-binding hits. (E) A list of each cell-binding peptide hit and the corresponding screen(s) in which the peptide was found.



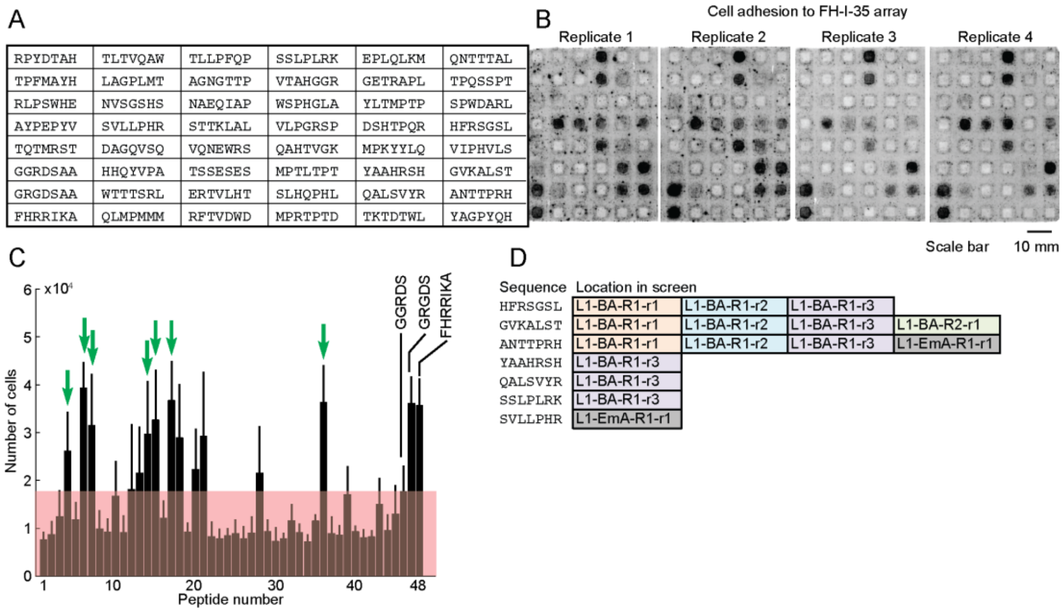
Appendix 1.20. (A) Standard curve for the number of MDA-MB-231-GFP cells and the corresponding grey-scale intensity per peptide zone (A=0.16 cm^2) used in each array. (B-D) Standard curves for the number of MDA-MB-231, MCF-7, and HT-29 cell lines and the corresponding grey-scale intensity per peptide zone (A=0.16 cm^2) used in each array. Cells were stained with 4 µM Cell Tracker Green, 1-hour prior to use in the cell adhesion assay. Cells were scanned using a fluorescent gel scanner; the settings were: LPB filter, 50 µm resolution, 400 V PMT.



TPPTMDH	SPQMTLS	QPSMLNP	MTVQPWP	HHSLTVT	AASSLTI
TTNLSPW	SPTQPKS	QPTHPTR	NHWASPR	HIPPGSP	ALAHRIL
TVRHLQL	SSALLLP	QRLPQTA	QATHRSH	IPTLPSS	ALEVTFW
TWYFGPL	SSLVRTA	QSLALQP	QEPLTAR	KAVHPLR	AMSSRSL
TYQNPVH	SSSVVTH	QTGYATR	QLHNDAT	LEVYALV	GKPMPPM
GGRDSAA	STFTKSP	QTSMATV	QLMNASR	MDAHHAL	GKVQAQS
GRGDSAA	STPMQNL	SHTAPLR	QLTVMSW	MHAPPFY	GPMLARG
FHRRIKA	TARYPSW	SLSLIQT	QPPRSTS	MPGSLPS	HALGPSS

Sequence	Location in screen		
ALAHRIL	L1-BA-R2-r1	L1-BA-R2-r2	L1-BA-R2-r3
ALEVTFW	L1-EmA-R1-r2		
AMSSRSL	L1-BA-R2-r2	L1-EmA-R1-r1	
QTGYATR	L1-EmA-R1-r1		
TVRHLQL	L1-BA-R1-r1	L1-BA-R1-r2	
TWYFGPL	L1-BA-R2-r1	L1-BA-R2-r2	

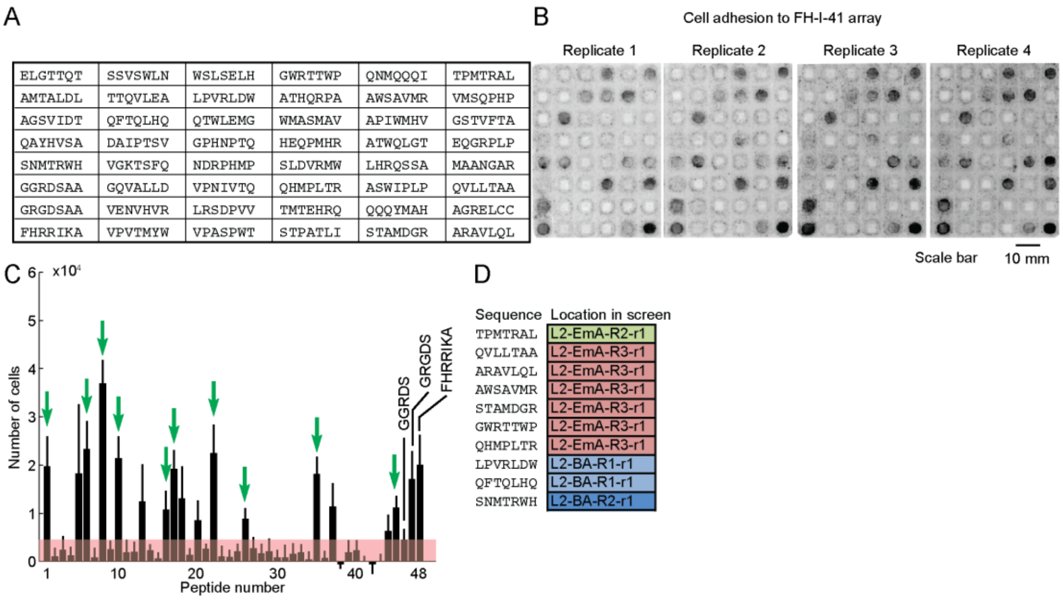
Appendix 1.21. Peptides synthesized on array FH-I-28 (A) and tested for short-term cell adhesion (B). Cell-binding peptide hits are considered as binding significantly more ($p < 0.05$) cells than the negative control GGRDS (red box) (C). Green arrays indicate six cell-binding hits. List of each cell-binding peptide hit and the corresponding screen(s) the peptide was found in (D). The abbreviation and color for each screen is the same as in Appendix 1.17 and Appendix 1.19.



RPYDTAH	TLTVQAW	TLLPFQP	SSLPLRK	EPLQLKM	QNTTTAL
TPFMAYH	LAGPLMT	AGNGTTP	VTAHGGR	GETRAPL	TPQSSPT
RLPSWHE	NVSGSHS	NAEQIAP	WSPHGLA	YLTMPTP	SPWDARL
AYPEPYV	SVLLPHR	STTKLAL	VLPGRSP	DSHTPQR	HFRSGSL
TQTMRST	DAGQVSQ	VQNEWRS	QAHTVGK	MPKYYLQ	VIPHVLS
GGRDSAA	HHQYVPA	TSSESES	MPTLTPT	YAAHRSH	GVKALST
GRGDSAA	WTTTSRL	ERTVLHT	SLHQPHL	QALSVYR	ANTTPRH
FHRRIKA	QLMPMMM	RFTVDWD	MPRTPTD	TKTDTWL	YAGPYQH

Sequence	Location in screen			
HFRSGSL	L1-BA-R1-r1	L1-BA-R1-r2	L1-BA-R1-r3	
GVKALST	L1-BA-R1-r1	L1-BA-R1-r2	L1-BA-R1-r3	L1-BA-R2-r1
ANTTPRH	L1-BA-R1-r1	L1-BA-R1-r2	L1-BA-R1-r3	L1-EmA-R1-r1
YAAHRSH	L1-BA-R1-r3			
QALSVYR	L1-BA-R1-r3			
SSLPLRK	L1-BA-R1-r3			
SVLLPHR	L1-EmA-R1-r1			

Appendix 1.22. Peptides synthesized on array FH-I-35 (A) and tested for short-term cell adhesion (B). Cell-binding peptide hits are considered as binding significantly more ($p < 0.05$) cells than the negative control GGRDS (red box) (C). Green arrays indicate six cell-binding hits. List of each cell-binding peptide hit and the corresponding screen(s) the peptide was found in (D). The abbreviation and color for each screen is the same as in Appendix 1.17 and Appendix 1.19.

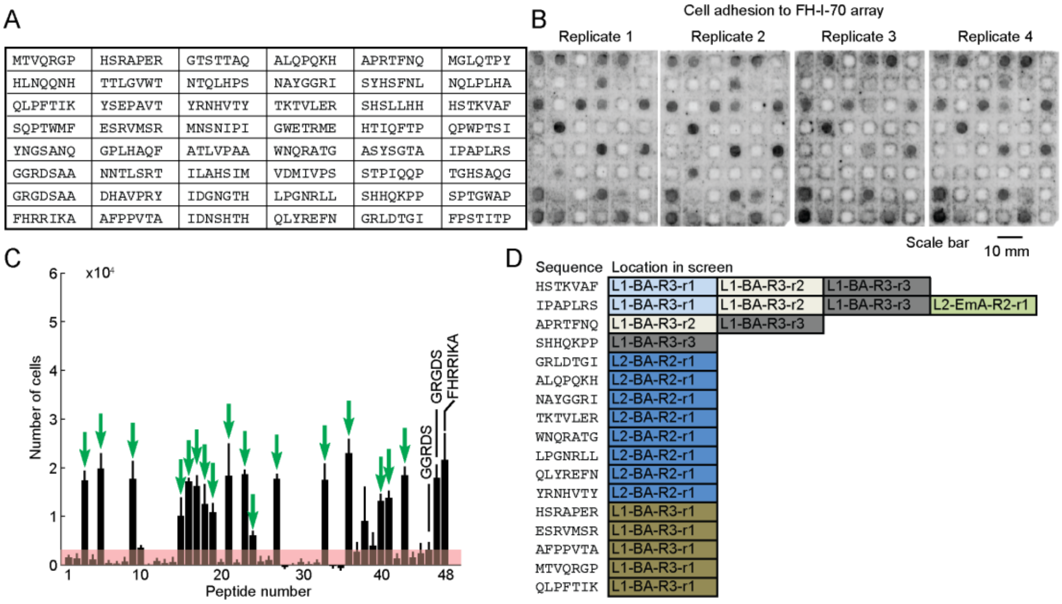


A

ELGTTQT	SSVSWLN	WSLSELH	GWRTTWP	QNMQQQI	TPMTRAL
AMTALDL	TTQVLEA	LPVRLDW	ATHQRPA	AWSAVMR	VMSQPHP
AGSVIDT	QFTQLHQ	QTWLEMG	WMASMAV	APIWMHV	GSTVFTA
QAYHVSA	DAIPTSV	GPHNPTQ	HEQPMHR	ATWQLGT	EQGRPLP
SNMTRWH	VGKTSFQ	NDRPHMP	SLDVRMW	LHRQSSA	MAANGAR
GGRDSAA	GQVALLD	VPNIVTQ	QHMPLTR	ASWIPLP	QVLLTAA
GRGDSAA	VENVHVR	LRSDPVV	TMTEHRQ	QQQYMAH	AGRELCC
FHRRIKA	VPVTMYW	VPASPWT	STPATLI	STAMDGR	ARAVLQL

Sequence	Location in screen
TPMTRAL	L2-EmA-R2-r1
QVLLTAA	L2-EmA-R3-r1
ARAVLQL	L2-EmA-R3-r1
AWSAVMR	L2-EmA-R3-r1
STAMDGR	L2-EmA-R3-r1
GWRTTWP	L2-EmA-R3-r1
QHMPLTR	L2-EmA-R3-r1
LPVRLDW	L2-BA-R1-r1
QFTQLHQ	L2-BA-R1-r1
SNMTRWH	L2-BA-R2-r1

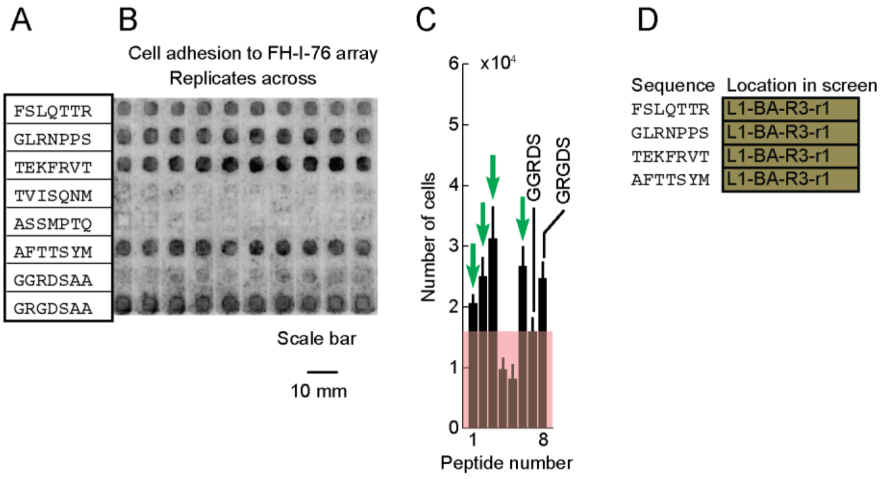
Appendix 1.23. Peptides synthesized on array FH-I-41 (A) and tested for short-term cell adhesion (B). Cell-binding peptide hits are considered as binding significantly more ($p < 0.05$) cells than the negative control GGRDS (red box) (C). Green arrays indicate six cell-binding hits. List of each cell-binding peptide hit and the corresponding screen(s) the peptide was found in (D). The abbreviation and color for each screen is the same as in Appendix 1.17 and Appendix 1.19.



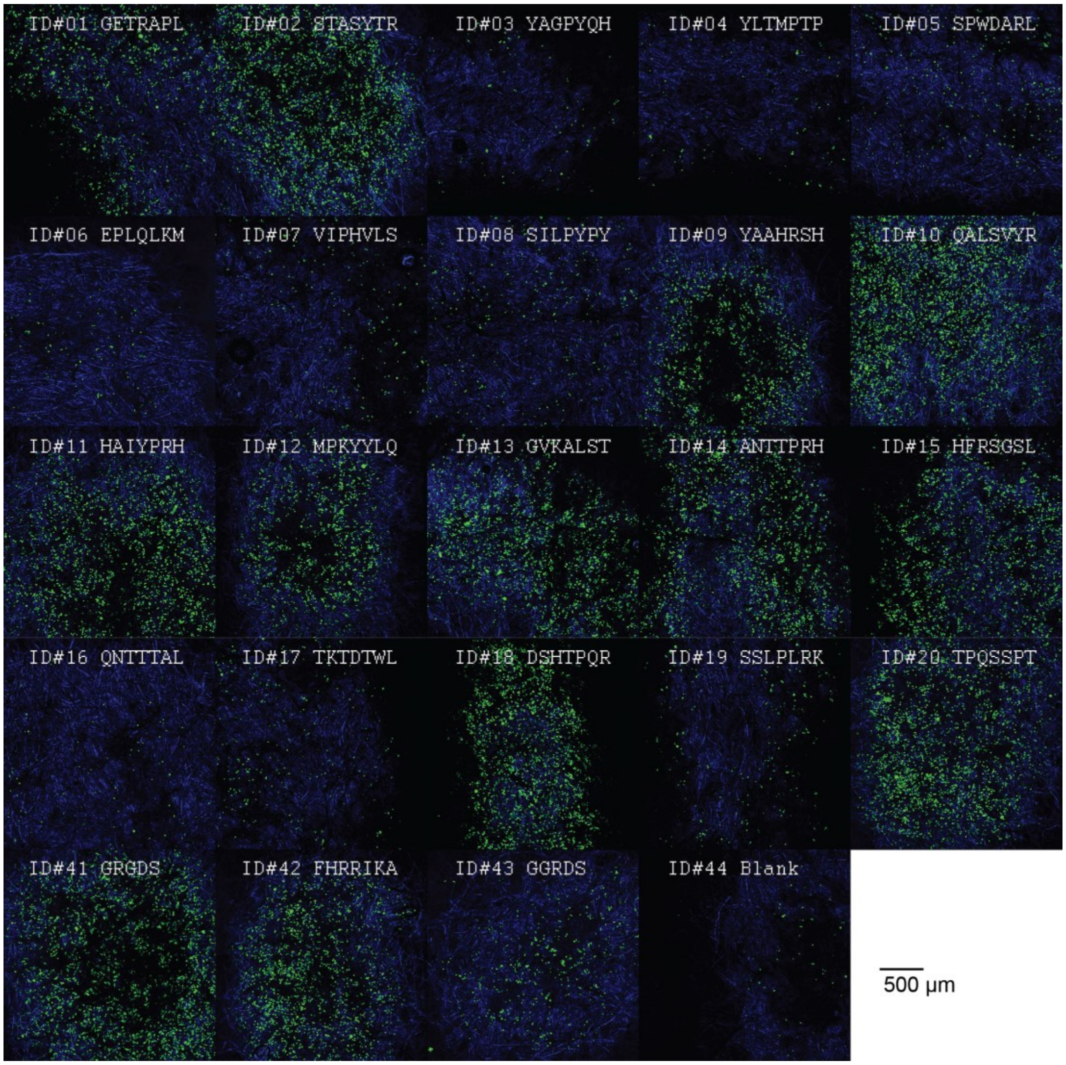
MTVQRGP	HSRAPER	GTSTTAQ	ALQPQKH	APRTFNQ	MGLQTPY
HLNQQNH	TTLGVWT	NTQLHPS	NAYGGRI	SYHSFNL	NQLPLHA
QLPFTIK	YSEPAVT	YRNHVTY	TKTVLER	SHSLLHH	HSTKVAF
SQPTWMF	ESRVMSR	MNSNIPI	GWETRME	HTIQFTP	QPWPTSI
YNGSANQ	GPLHAQF	ATLVPAA	WNQRATG	ASYSGTA	IPAPLRS
GGRDSAA	NNTLSRT	ILAHSIM	VDMIVPS	STPIQQP	TGHSAQG
GRGDSAA	DHAVPRY	IDGNGTH	LPGNRLL	SHHQKPP	SPTGWAP
FHRRIKA	AFPPVTA	IDNSHTH	QLYREFN	GRLDTGI	FPSTITP

Sequence	Location in screen			
HSTKVAF	L1-BA-R3-r1	L1-BA-R3-r2	L1-BA-R3-r3	
IPAPLRS	L1-BA-R3-r1	L1-BA-R3-r2	L1-BA-R3-r3	L2-EmA-R2-r1
APRTFNQ	L1-BA-R3-r2	L1-BA-R3-r3		
SHHQKPP	L1-BA-R3-r3			
GRLDTGI	L2-BA-R2-r1			
ALQPQKH	L2-BA-R2-r1			
NAYGGRI	L2-BA-R2-r1			
TKTVLER	L2-BA-R2-r1			
WNQRATG	L2-BA-R2-r1			
LPGNRLL	L2-BA-R2-r1			
QLYREFN	L2-BA-R2-r1			
YRNHVTY	L2-BA-R2-r1			
HSRAPER	L1-BA-R3-r1			
ESRVMSR	L1-BA-R3-r1			
AFPPVTA	L1-BA-R3-r1			
MTVQRGP	L1-BA-R3-r1			
QLPFTIK	L1-BA-R3-r1			

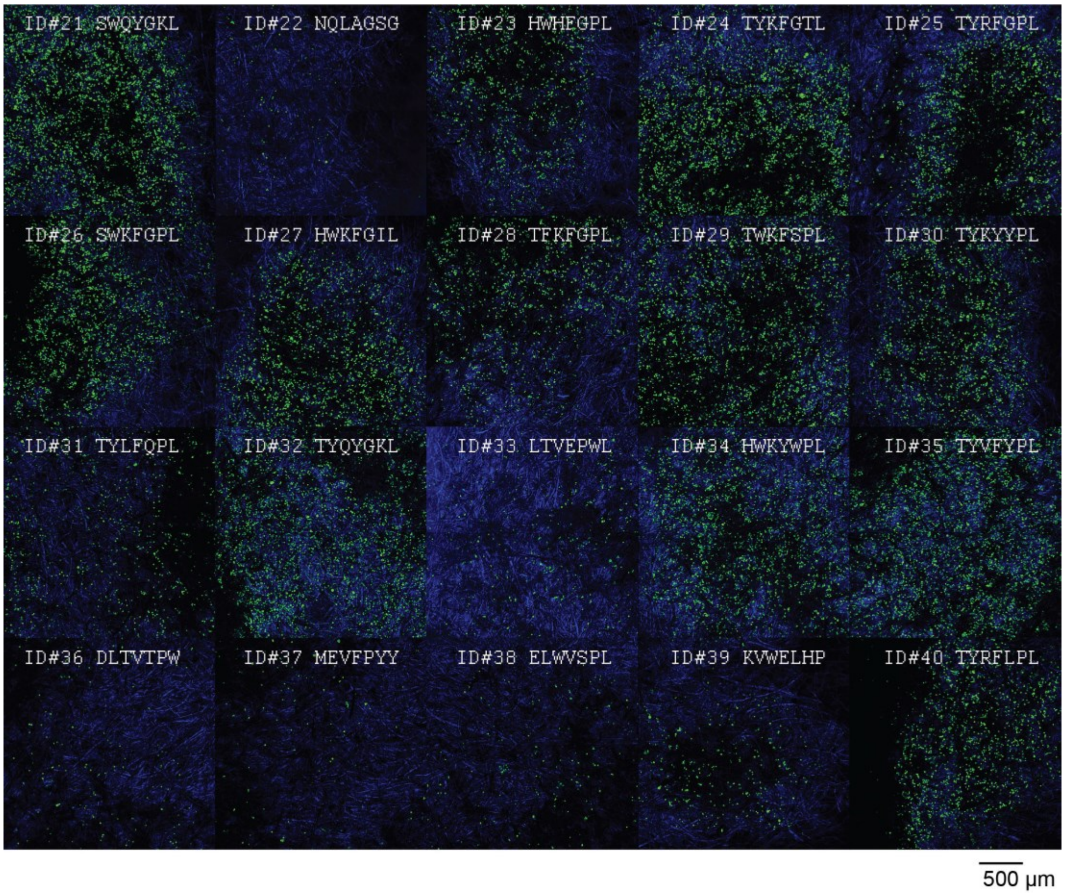
Appendix 1.24. Peptides synthesized on array FH-I-70 (A) and tested for short-term cell adhesion (B). Cell-binding peptide hits are considered as binding significantly more ($p < 0.05$) cells than the negative control GGRDS (red box) (C). Green arrays indicate six cell-binding hits. List of each cell-binding peptide hit and the corresponding screen(s) the peptide was found in (D). The abbreviation and color for each screen is the same as in Appendix 1.17 and Appendix 1.19.



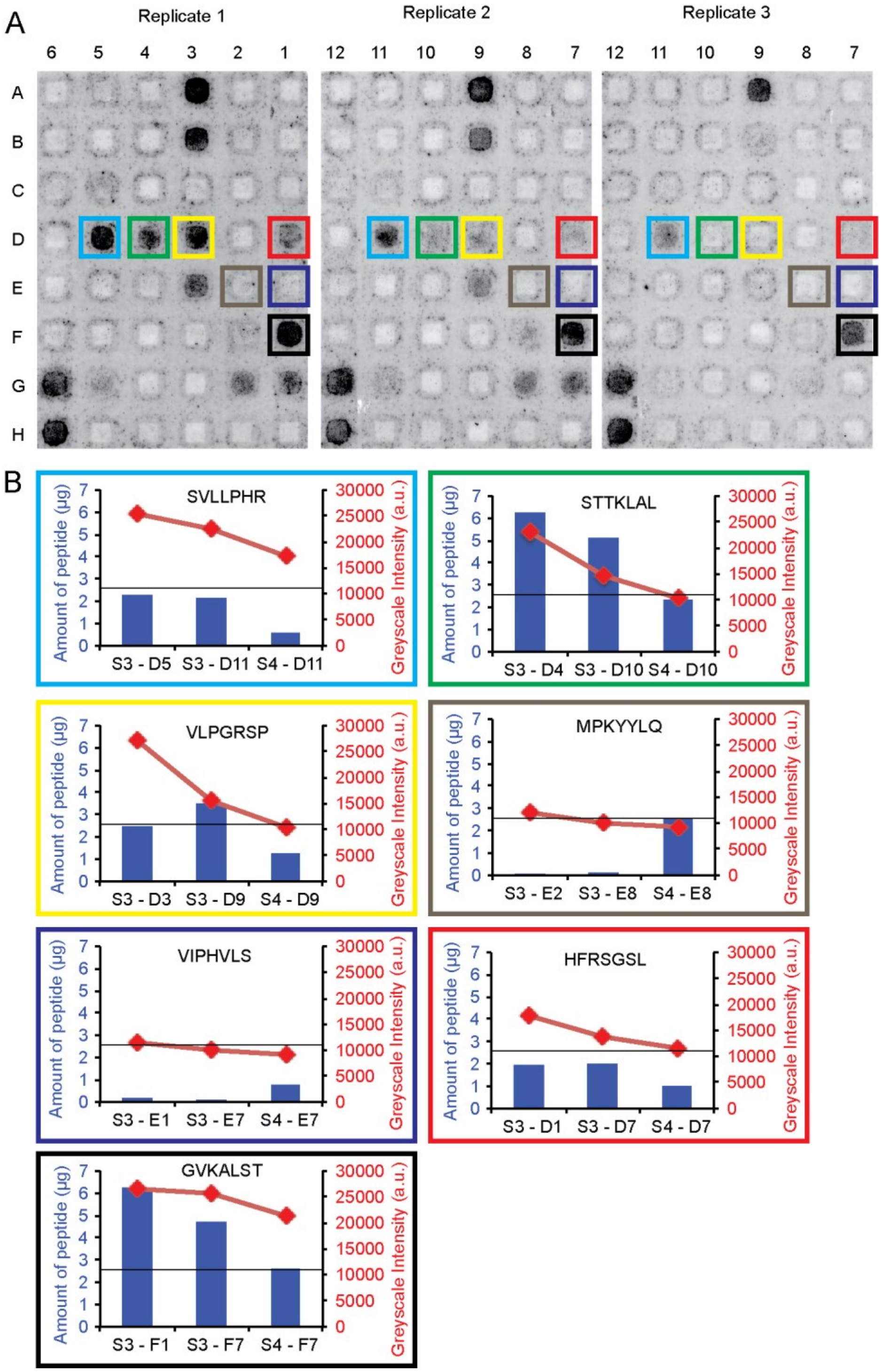
Appendix 1.25. Peptides synthesized on array FH-I-76 (A) and tested for short-term cell adhesion (B). Cell-binding peptide hits are considered as binding significantly more ($p < 0.05$) cells than the negative control GGRDS (red box) (C). Green arrays indicate six cell-binding hits. List of each cell-binding peptide hit and the corresponding screen(s) the peptide was found in (D). The abbreviation and color for each screen is the same as in Appendix 1.17 and Appendix 1.19.



Appendix 1.26. Representative confocal images of MDA-MB-231-GFP cells on array FH-I-20 with peptides selected from the BA screen. Confocal images correspond to the fluorescent gel image represented in Figure 5.1C and Figure 5.6A. The images represent short-term adhesion after 3 hours of incubation on arrays. The scale bar represents 500 μm. The images were acquired using identical microscopy settings (laser intensity, PMT gain). Colors: blue – paper fibers (imaged by reflectance); green – GFP.



Appendix 1.27. Representative confocal images of MDA-MB-231-GFP cells on array FH-I-20 with peptides selected from the BA screen. Confocal images correspond to the fluorescent gel image represented in Figure 5.1C and Figure 5.6A. The images represent short-term adhesion after 3 hours of incubation on arrays. The scale bar represents 500 µm. The images were acquired using identical microscopy settings (laser intensity, PMT gain). Colors: blue – paper fibers (imaged by reflectance); green – GFP.

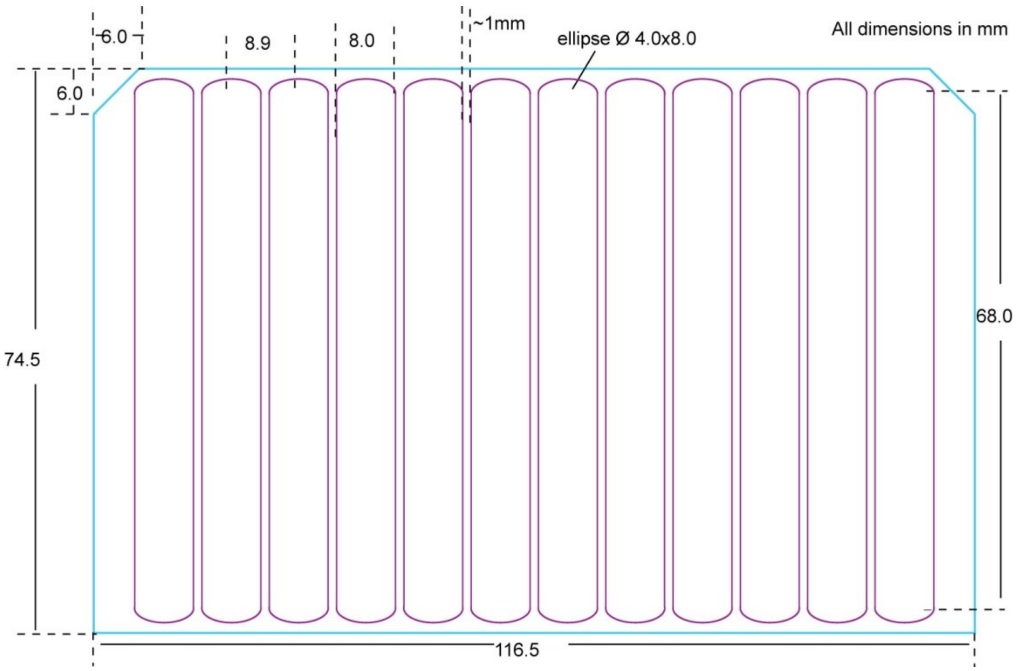


Appendix 1.28. (A) Representative replicates from array FH-I-35 used to quantify the amount of peptide at each peptide zone to the corresponding grey-scale. For illustrative purposes, the grey-scale intensities were adjusted in panel **a** to simplify visualization (same level for all images). (B) The grey-scale of each peptide zone was determined using ImageJ. Original .gel files were used without any adjustments to grey-scale intensities. The amount of peptide and

corresponding grey-scale for each replicate was plotted for each peptide. The bar plot (blue) represents the amount of peptide; the line plot (red) indicates the grey-scale.

	Lot2-BA-Rep1-Round1	Lot2-BA-Rep2-Round1	Lot2-BA-Rep3-Round1	Lot1-Parasites	Lot2-Parasites		Lot2-BA-Rep1-Round1	Lot2-BA-Rep2-Round1	Lot2-BA-Rep3-Round1	Lot1-Parasites	Lot2-Parasites		Lot2-BA-Rep1-Round1	Lot2-BA-Rep2-Round1	Lot2-BA-Rep3-Round1	Lot1-Parasites	Lot2-Parasites
SLLGQTP	2917	1	26	0	1	LSHTLTW	39	7429	5	0	1	SAAWNKS	0	0	14055	0	1
WSSHKNV	1649	0	0	0	0	HSHTLTW	20	4589	3	0	1	WSLSELH	47	168	8439	0	1
AYVARQN	1445	584	0	0	1	TTANVRI	1	2461	0	0	1	SLDVRMW	1	49	5970	0	1
RPPLINT	712	0	0	0	1	GTSIYLH	0	2036	2	0	1	VPWSVTR	0	0	2644	0	0
SYTDLLR	654	0	0	0	0	LRSDPVV	0	1795	670	0	1	ALPAKRE	0	0	2548	0	0
LATLTAC	586	0	0	0	1	AHINVPS	0	1776	11	0	1	TATLPRM	0	0	1920	0	1
IWKQALI	579	0	0	0	0	AGMWNAT	0	1266	0	0	1	LQLAVDT	0	0	1882	0	1
VVYVSFF	480	0	0	0	1	LSHSLTW	8	904	0	0	1	WGRISHV	26	262	1436	0	1
INTLSRT	457	0	0	0	1	VENVHVR	0	705	0	0	1	WDPRVNV	0	0	1311	0	1
TVISQNM	455	0	0	0	1	HSHTLTC	7	650	1	0	1	FQLPWAG	0	0	1260	0	0
AHKHDPT	451	0	0	0	1	FNGSANQ	15	594	2	0	1	GMHETHV	0	0	1146	0	1
ILDKLTH	435	0	0	0	0	AYVARQN	1445	584	0	0	1	LSPSLRN	0	0	1104	0	1
VDIGDSR	376	0	0	0	1	WEIGSGP	0	580	0	0	1	LWTNNWL	0	0	1046	0	0
FQWELYS	359	0	0	0	1	GQGQTIP	0	540	0	0	1	TKPNLYN	54	0	969	0	1
STIANTK	357	0	0	0	0	LLYREFN	33	518	808	0	1	QQLAVDT	0	0	838	0	1
SGAVWPF	326	0	131	0	1	AEIPHRG	0	515	0	0	0	LLYREFN	33	518	808	0	1
LTSVAGA	303	0	0	0	0	YNGSANQ	5	508	4	0	1	MFDMVKL	26	0	792	0	1
NWKQALI	297	0	0	0	0	FMTAPGF	0	371	0	0	0	SAAWNKS	0	0	747	0	1
DVYVSFF	287	0	0	0	0	STMKTGC	0	354	0	0	0	LAAWHFI	0	0	741	0	1
QATLTAC	284	0	0	0	1	FARANAA	0	352	0	0	1	LTWLEMG	72	0	715	0	1
FQVALHS	264	0	0	0	0	HSHTLTR	4	343	0	0	1	AGHNLVP	0	0	709	0	0
LGGVRLY	261	14	49	0	1	GPLHAQF	20	326	0	0	1	LRSDPVV	0	1795	670	0	1
LVVHRTS	257	0	0	0	1	VTAVKTG	0	305	0	0	1	LRAQVTP	0	0	583	0	1
MDGAAGF	234	0	69	0	1	VPQILPY	0	303	21	0	1	WSLSELQ	4	16	578	0	1
NNTLSRT	232	0	0	0	1	AHIYVPS	0	300	2	0	0	SLLSLNF	0	0	546	0	0
GMNTTWT	228	0	79	0	1	LTMPFSG	0	296	23	0	1	HWTNNWL	0	0	489	0	0
NLDKLTH	227	0	0	0	0	AHTSTWP	0	286	0	0	1	KFDMVKL	21	0	473	0	1
IKSAWFL	220	218	0	0	0	QLYREFN	26	283	380	0	1	LHRPANC	0	0	461	0	0
SLLGQTP	214	0	1	0	1	SPTMVNA	0	278	0	0	0	TVNFKLY	2	0	419	0	1
GQHVWWI	213	0	0	0	0	HSHTLTF	1	266	0	0	1	QLYREFN	26	283	380	0	1
VPGLGLN	208	0	0	0	1	GTSFYLH	0	265	0	0	1	SLDVRMC	0	1	357	0	0
VFKINSK	205	0	0	0	0	WGRISHV	26	262	1436	0	1	QTWLEMG	38	0	355	0	1
WSSLKNV	200	0	0	0	0	LLYFAAP	0	257	1	0	1	SAAWNKT	0	0	355	0	0
SFNLPNT	188	0	0	0	1	LANATSL	2	256	0	0	1	TVGNSVG	0	0	352	0	0
AYVARQK	181	61	0	0	1	TTAYVRI	0	244	0	0	0	HAAWHFI	0	0	341	0	1
LHHKVTI	181	0	0	0	0	IAYGGRI	0	243	0	0	1	AKSLMFQ	0	2	317	0	1
YQWELYS	180	0	0	0	1	VQYKPMK	0	241	129	0	1	LGEWIKY	0	0	295	0	1
SSSSHVM	176	0	0	0	1	WQAHGIS	0	241	0	0	0	FASAARV	0	0	293	0	1
VHAGLQV	175	16	0	0	1	ASDGKVA	0	224	0	0	0	SFSRAES	9	16	289	0	1
EPGLGLN	169	0	0	0	1	IKSAWFL	220	218	0	0	0	GKVAKQE	0	0	259	0	1
AIDFARN	168	9	0	0	1	WPNKFMY	0	216	0	0	0	QQLAVDT	0	0	234	0	1
LLADMHA	168	0	0	0	1	TTANVRI	1	212	0	0	1	LPVRLDW	0	0	230	0	1
MTGSVQM	161	0	0	0	0	SYASYYL	0	209	0	0	0	LTPPNDF	0	0	218	0	0
ILLPFAS	161	0	0	0	0	THEDVNL	0	205	0	0	0	VHYVKGP	0	0	198	0	0
IDRTQFM	156	105	6	0	1	MRTSMPH	0	191	0	0	0	IDNSHTH	0	1	189	0	1
LMHREPA	156	0	0	0	1	SHVNVPS	1	187	0	0	1	LKHSHHY	0	0	182	0	0
GLRNPPS	145	27	0	0	1	VKNVPSQ	0	187	0	0	0	TLWGEHP	0	0	181	0	0
IMENSYV	144	0	0	0	0	VSRVMSR	0	185	0	0	1	YASAARV	0	0	181	0	1
VALLSTT	139	0	0	0	0	AYDKFTP	0	184	0	0	1	VPWSVTR	0	0	179	0	0
DHAGLQV	135	9	0	0	1	TTQVLEA	115	183	155	0	1	WSHTPPR	0	0	176	0	0

Appendix 1.29. List of top 50 sequences from three replicates after one round of selection against HEK cell line using the BA method. The color for each selection is used to distinguish between different screens. The numbers indicate the copy number of each of the top 50 sequences. The sequences are ordered by decreasing copy number in the corresponding screen. If sequences are found in other screens, the corresponding copy number is indicated. The color of the font and last two columns identify the sequence as a parasite (red and '1') or not (black and '0'). Asterisks '**' indicate sequences that we identified in the top 50 against HEK cells and the top 20 against the MDA-MB-231 cell line.



Appendix 1.30. Scheme for an aluminium grid insert used to hold peptide paper arrays submerged in a Nunc Omni-Tray for cell adhesion assays.

Flow-Through Synthesis on Teflon-Patterned Paper To Produce Peptide Arrays for Cell-Based Assays**

Frederique Deiss, Wadim L. Matochko, Natasha Govindasamy, Edith Y. Lin, and Ratmir Derda*

Abstract: A simple method is described for the patterned deposition of Teflon on paper to create an integrated platform for parallel organic synthesis and cell-based assays. Solvent-repelling barriers made of Teflon-impregnated paper confine organic solvents to specific zones of the patterned array and allow for 96 parallel flow-through syntheses on paper. The confinement and flow-through mixing significantly improves the peptide yield and simplifies the automation of this synthesis. The synthesis of 100 peptides ranging from 7 to 14 amino acids in length gave over 60%purity for the majority of the peptides (>95% yield per coupling/deprotection cycle). The resulting peptide arrays were used in cell-based screening to identify 14 potent bioactive peptides that support the adhesion or proliferation of breast cancer cells in a 3D environment. In the future, this technology could be used for the screening of more complex phenotypic responses, such as cell migration or differentiation.

Solid-phase synthesis (SPS) is a central technique for producing libraries of lead organic compounds for the pharmaceutical and biotechnology industry. SPOT-synthesis was developed in the 1990s as a method for parallel SPS on a planar support to yield high-density arrays of peptides[1] and other organic molecules.[2] SPOT-synthesis has been adapted in academic and industrial research for the production of functional ligands, epitope mapping, cell-based screens,[3, 4] and the identification of functional materials.[5] To date, the environment of chemical reactions in SPOT synthesis is suboptimal when compared to SPS: the optimal yield and reactivity in SPS is achieved in an actively-mixed solution or in a flow-through reactor in which the solid support is exposed to a continuous flow of reagent. By contrast, in classical SPOT synthesis, a limited amount of reagent is spotted onto paper, thus forming a static spot of liquid with a defined size. Within this spot, flow-through conditions are not possible, thereby limiting mass transfer to diffusion. Furthermore, the fixed relationship between the size of the spot and the volume of the solution limits the volume that can be deposited onto the support.[4] We solved these problems by introducing solvent repelling Teflon barriers into the paper. The patterns confine liquids and thus allow the deposition of an excess volume of reagents, enable parallel flow-through synthesis, and significantly improve the yields of the chemical reactions. While we focus on paper-based supports and peptide synthesis, we believe that an analogous approach could be applied to any planar porous support and other types of reactions.

Herein, we demonstrate that Teflon-patterned paper satisfies all of the criteria for multistep organic synthesis and downstream analysis: 1) Teflon-impregnated paper is stable to prolonged exposure to organic solvents, organic bases (e.g., piperidine), and strong acids (TFA). 2) most of the required solvents and reagents exhibit a high contact angle on paper permeated by Teflon. 3) Teflon-modified paper is suitable for downstream biochemical and cell-based assays since it is neither fluorescent nor toxic. None of the preexisting methods for the patterning of paper[6, 7] satisfy all of the above criteria. We introduced Teflon into the paper according to the protection–deprotection strategy illustrated in Figure 1; a process referred to as "sweet patterning".[8] The outlines of the patterns were first defined by using a solid ink printer[7] and the zones that should not be impregnated by Teflon were "protected" with a solution of sucrose. The entire array was then exposed to a solution of Teflon to impregnate the unprotected zones, with the Teflon forming solvent repelling barriers upon drying. Finally, the sucrose was washed away with water. Given that it requires only two steps of liquid deposition, the process can be readily scaled: we automated it by using a liquid handling robot Precision XS, BioTek (Figure S3, Script S1, and Movie S1 in the Supporting Information). We identified that a pattern with a width of at least 1 mm of Teflon is required to ensure long-term stability of the pattern to organic solvents. (Figure S4 in the Supporting Information).

Teflon-patterned paper confines most solvents that are immiscible with Teflon solution (Figure S5 in the Supporting Information), including weakly polar (dichloromethane, dioxane), polar aprotic (DMF, NMP, DMSO), and protic (alcohols) solvents, as well as aqueous solutions of proteins and surfactants (Figure 2b, c). Examples of liquids that cannot be contained are nonpolar hydrocarbons (hexane) and perfluorinated solvents (HFE-7100). We characterized the pattern by using a "confinement factor" CF=V1V_12, where V1 is the volume contained per unit area of patterned paper and V2 the volume of solvent retained by the same area of nonpatterned paper. For example, when spotting 25 mL of DMF, we obtained CF=12 (Figure 2a). Hydrophobic ink provided no confinement because the ink is both wetted and dissolved by DMF.

The confinement of an excess volume of reagents on each spot induced gravity-driven flow through the paper (Figure 3a, b). Flow rates were reproducible for specific types of solvent and paper (Figure 3 c and Figure S6). The flow rate is nonlinear but it can be characterized by using the time required for half of the volume of the droplet to flow through the paper (t1/2 ; Figure 3d). For filter paper Grade 50, the t1/2 values for DMF and NMP solutions were approximately 5 min. This flow rate allowed approximately 90% of the solvent to flow through the paper during the typical reaction steps of peptide synthesis, such as amino acid coupling (20– 40 min) or deprotection steps (10–15 min). The rate could be accelerated by using paper with a higher porosity (Figure S6), or by placing an aspirating nozzle below the paper (Figure S7). In the following examples, we used 96-zone arrays with a footprint identical to a conventional 96-well plate to perform 96 parallel flow-through syntheses. Flow-through also allows the replication of reactions in a stack of multiple patterned papers (Figure 3e and movie S7). Confinement and dynamic flow-through were advantageous for reactions on paper. The yields of the reactions in flow-through peptide synthesis conditions were higher than those from conventional peptide synthesis on paper (Figure 4). We used Fmoc quantification to demonstrate that the synthesis of the di-β-alanine linker (βAβA) and the subsequent coupling of the first amino acid (to give βAβAA) were significantly more efficient with the Teflon-patterned paper than with nonpatterned paper in multiple independent trials (Figure 4 b). We used the same quantification to show that the conversion at each coupling step in the synthesis of the $(Ala)_{10}$ peptide was higher when using the Teflon-patterned paper, with an estimated 50% conversion of a linker to the final peptide for flow-through synthesis compared to <20% for conventional SPOT (Figure 4c). The identity and purity (correlates to conversion) of more than 100 peptides, ranging in length from 7 to 14 amino acids, that were synthesizedand characterized by one technician are shown in Table S1 in the Supporting Information. Confinement of the solvent and flow-through also simplified the automation of the synthesis and improved the washing steps (Figure S8); it reduced the volume of the washing solutions from 15–30 mL, as required in conventional SPOT synthesis,[4] to 1.5 mL in flow-through washing. We used plate-to-plate transfers with the Precision XS workstation to sequentially deposit solutions of activated amino acids (Figure S9), deprotection and capping agents, and washing solvent.

We previously demonstrated that paper can be used to generate “foldable” 3D tumor models to study 3D cultures of cells in vivo and in vitro,[9] for example, to investigate migration[10] or drug resistance.[11] The use of Teflon-patterned arrays allowed us to characterize surface-immobilized peptides that can support cell adhesion, growth, or differentiation. We synthesized eight reported bioactive peptides that are known to support the self-renewal of stem cells[12, 13] and induce epithelial–mesenchymal transition.[14] When immobilized on paper, five of the eight peptides supported adhesion of MDA-MB-231 breast carcinoma cells at levels similar to or higher than the positive control peptide GRGDS (Figure 5). The other three peptides supported greater adhesion than the scrambled peptide GGRDS. Unmodified cellulose and paper decorated with GGRDS supported minimal binding. Confocal micrographs of the cells in each substrate are shown in Figure S10. None of the components of the Teflon-patterned peptide array exhibited any toxicity toward the cells (Figure S11). Over the long term, the cells on some peptide modified surfaces spread along the fibers and resumed cell division (Figure S12). We then used Teflon-patterned arrays to validate the biological properties of 30 peptides identified de novo by phage-display panning on MDA-MB-231 cells (Table S1). Imaging with a fluorescence gel scanner (Figure 6a and Figure S13) and confocal microscopy (Figure 6b and Figures S14, S15) demonstrated that 14 of the 30 peptides supported cell adhesion at levels higher than the integrin binding peptide GRGDS (Figure 6c); four of the peptides supported adhesion at similar levels, and 10 did not support adhesion. These experiments confirmed that Teflon-patterned paper is an effective platform for the synthesis and cell-based screening of a large number of peptides. Paper is a versatile support for applications such as analytical devices or low-cost diagnostics.[15] We believe that parallel-synthesis capability and the generation of patterns resistant to organic solvents and surfactants will also be beneficial in these areas. The patterning of low-cost paper makes this technology conveniently available; however, we anticipate that future advances in materials production (lithography, 3D-printing, weaving, etc)[16] could yield similar low-cost, self-supported, patterned porous sheets suitable for organic synthesis and bioassays.

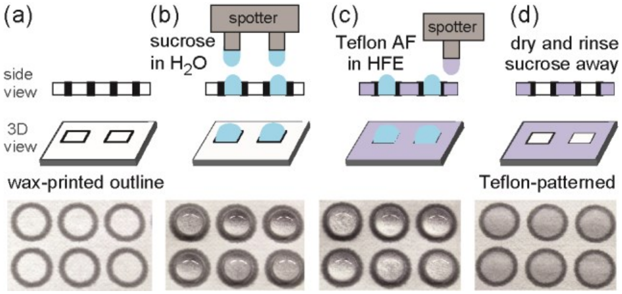


Figure 1. The Teflon-patterning process. a) The pattern is defined by wax-printing. b) Sucrose is spotted into the zones to be kept solvophilic. c) A Teflon solution is deposited onto the remaining regions to form solvophobic barriers. d) After evaporation of the hydrofluoroether (HFE) solvent, the sucrose is washed off.

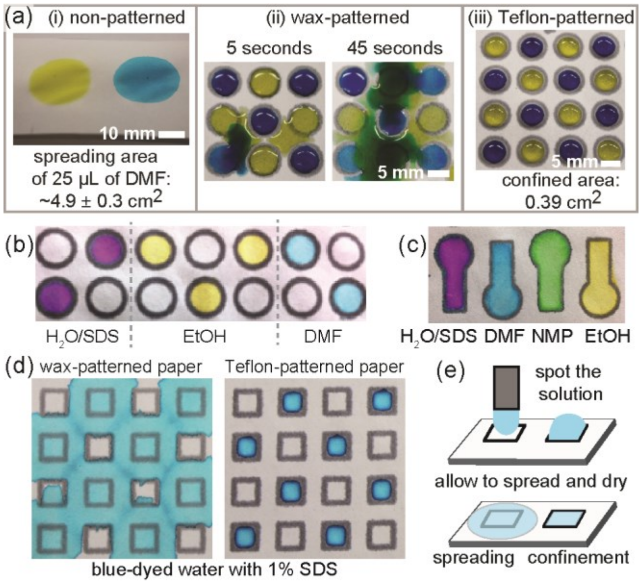
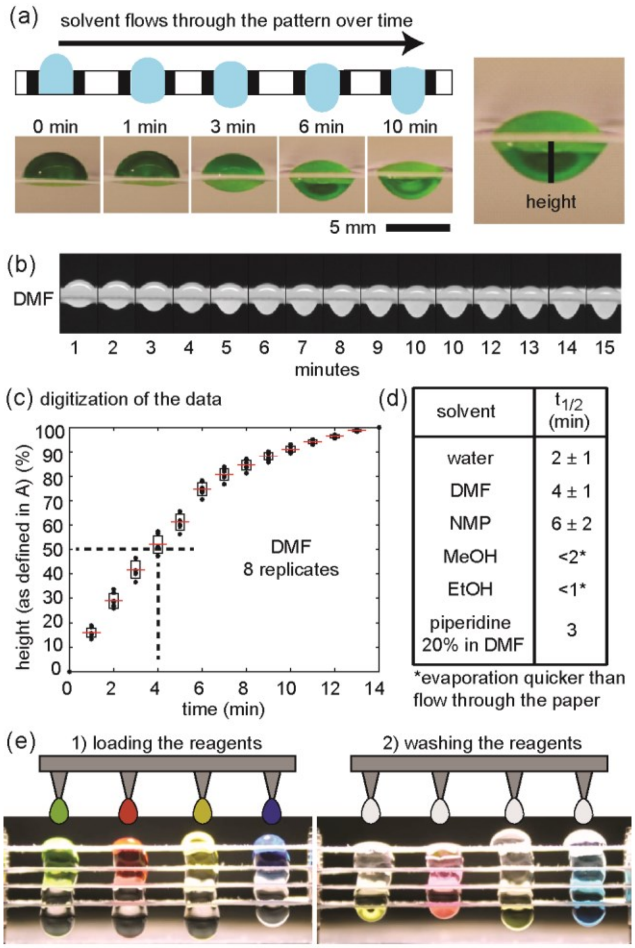


Figure 2. a) Spreading of DMF on non-patterned and wax-patterned paper and confinement on Teflon-patterned paper. b, c) Examples of Teflon-patterned arrays that confined ethanol, DMF, and NMP. d) SDS in water destroys wax-patterned paper; the same solution is confined on Teflon-patterned paper. e) A scheme showing the process of evaluating spreading and confinement. DMF=N,N-dimethylformamide, NMP=N-methyl-2-pyrrolidinone, SDS=sodium dodecylsulfate, DMSO=dimethyl sulfoxide.



solvent	$t_{1/2}$ (min)
water	2 ± 1
DMF	4 ± 1
NMP	6 ± 2
MeOH	<2*
EtOH	<1*
piperidine 20% in DMF	3

*evaporation quicker than flow through the paper

Figure 3. a) The flow of solvent through the paper. The height of hanging droplet was used for the digitization of flow-through data. b) Time-lapse images of DMF flowing through the patterned paper. c) Digitization of eight independent time-lapse experiments. d) A table summarizing the time needed for a liquid to reach half of the maximum height (see the Supporting Information). e) The flow of reagents (1) or washing solutions (2) through a stack of four arrays.

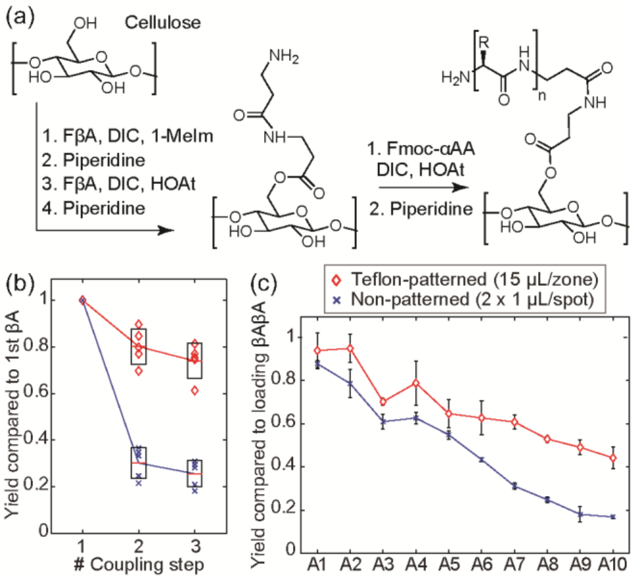


Figure 4. a) Peptide synthesis on paper modified with a βAβA linker. Yields for the synthesis of the bAbAA construct (b) and the βAβA(A)10 peptide (c) on Teflon-patterned paper (◊) and nonpatterned paper (x) are shown. In (b), the five data points represent synthesis on five separately prepared arrays. All yields were estimated as an average of Fmoc loading measured in four independent areas. DIC=diisopropylcarbodiimide, 1-Melm=1-methylimidazole, HOAt=1-hydroxy-7-azabenzotriazole, Fmoc=9-fluorenylmethyloxycarbonyl, FbA=Fmoc-b-alanine, Fmoc–aAA=Fmoc-protected a-amino acid.

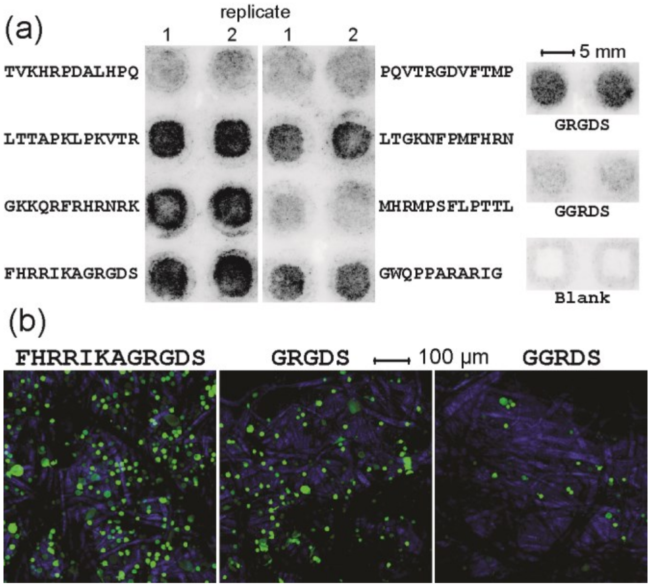


Figure 5. The adhesion of MDA-MB-231-GFP cells to known bioactive peptides synthesized on paper. a) Imaging with a fluorescence gel scanner locates GFP fluorescence (dark areas). b) We confirmed the results by confocal microscopy to validate binding to GRGDS (positive control) and no binding to GGRDS (negative control).

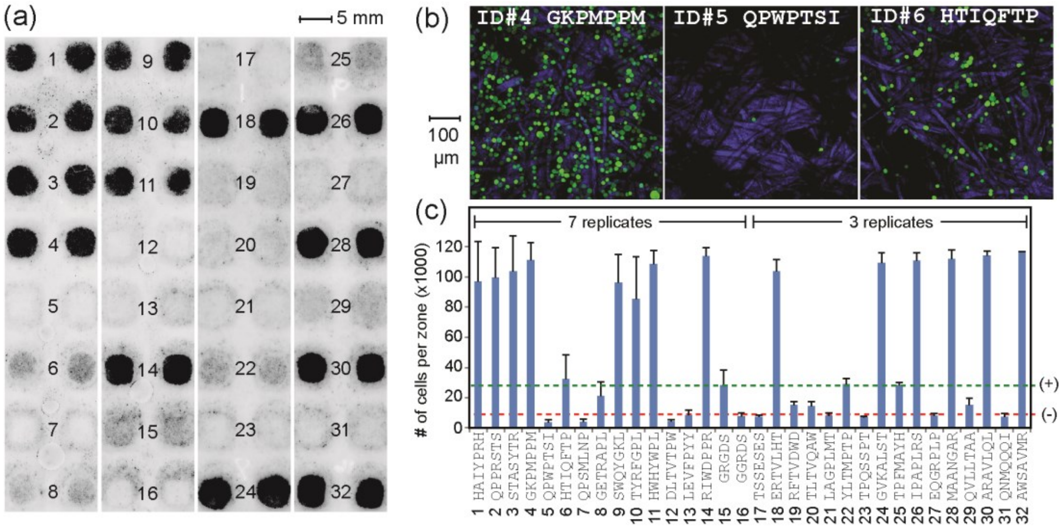


Figure 6. a) We used peptide-arrays to validate the adhesion of MDA-MB-231-GFP cells to 30 peptides identified by phage-display panning. b) Representative confocal images of peptide-modified paper supporting high cell adhesion, no cell adhesion, and moderate cell adhesion. c) The digitization of the number of cells per zone from fluorescence gel scanner images of the cell binding experiments. The (+) and (_) lines represent the intensity levels of the positive (GRGDS) and negative (GGRDS) controls, respectively.

[1] R. Frank, Tetrahedron 1992, 48, 9217 – 9232; R. Frank, J. Immunol. Methods 2002, 267, 13 – 26.

[2] L. Jobron, G. Hummel, Angew. Chem. 2000, 112, 1704 – 1707; Angew. Chem. Int. Ed. 2000, 39, 1621 – 1624; M. D. Bowman, R. C. Jeske, H. E. Blackwell, Org. Lett. 2004, 6, 2019 – 2022; Q. Lin, J. C. O_Neil, H. E. Blackwell, Org. Lett. 2005, 7, 4455 – 4458.

[3] U. Reineke, R. Sabat, R. Misselwitz, H. Welfle, H.-D. Volk, J. Schneider-Mergener, Nat. Biotechnol. 1999, 17, 271 – 275; U. Reineke, C. Ivascu, M. Schlief, C. Landgraf, S. Gericke, G. Zahn, H. Herzel, R. Volkmer-Engert, J. Schneider-Mergener, J. Immunol. Methods 2002, 267, 37 – 51; S. Ahmed, A. S. Mathews, N. Byeon, A. Lavasanifar, K. Kaur, Anal. Chem. 2010, 82, 7533 – 7541; M. D. Bowman, J. C. O_Neill, J. R. Stringer, H. E. Blackwell, Chem. Biol. 2007, 14, 351 – 357; K. Hilpert, R. Volkmer- Engert, T. Walter, R. E.W. Hancock, Nat. Biotechnol. 2005, 23, 1008 – 1012.

[4] K. Hilpert, D. F. H. Winkler, R. E.W. Hancock, Nat. Protoc. 2007, 2, 1333 – 1349.

[5] H. E. Blackwell, Curr. Opin. Chem. Biol. 2006, 10, 203 – 212; M. D. Bowman, M. M. Jacobson, H. E. Blackwell, Org. Lett. 2006, 8, 1645 – 1648.

[6] D. A. Bruzewicz, M. Reches, G. M. Whitesides, Anal. Chem. 2008, 80, 3387 – 3392; A.W. Martinez, S. T. Phillips, M. J. Butte, G. M. Whitesides, Angew. Chem. 2007, 119, 1340 – 1342; Angew. Chem. Int. Ed. 2007, 46, 1318 – 1320; J.

Wang, M. R. N. Monton, X. Zhang, C. D. M. Filipe, R. Pelton, J. D. Brennan, Lab Chip 2014, 14, 691 – 695.
[7] E. Carrilho, A.W. Martinez, G. M. Whitesides, Anal. Chem. 2009, 81, 7091 – 7095.
[8] R. Derda, A. Laromaine Sagu_, G. M. Whitesides (President and Fellows of Harvard College), EP2265959 A4, 2009.
[9] R. Derda, A. Laromaine, A. Mammoto, S. K. Y. Tang, T. Mammoto, D. E. Ingber, G. M. Whitesides, Proc. Natl. Acad. Sci. USA 2009, 106, 18457 – 18462.
[10] R. Derda, S. K. Y. Tang, A. Laromaine, B. Mosadegh, E. Hong, M. Mwangi, A. Mammoto,D. E. Ingber,G. M.Whitesides, PLoS One 2011, 6, e18940.
[11] F. Deiss, A. Mazzeo, E. Hong, D. E. Ingber, R. Derda, G. M. Whitesides, Anal. Chem. 2013, 85, 8085 – 8094.
[12] Z. Melkoumian, J. L. Weber, D. M. Weber, A. G. Fadeev, Y. Zhou, P. Dolley-Sonneville, J. Yang, Q. Qiu, C. A. Priest, C. Shogbon, A.W. Martin, J. Nelson, P.West, J. P. Beltzer, S. Pal, R. Brandenberger, Nat. Biotechnol. 2010, 28, 606 – 610; J. R. Klim, L. Y. Li, P. J. Wrighton, M. S. Piekarczyk, L. L. Kiessling, Nat. Methods 2010, 7, 989 – 994.
[13] R. Derda, S. Musah, B. P. Orner, J. R. Klim, L. Y. Li, L. L. Kiessling, J. Am. Chem. Soc. 2010, 132, 1289 – 1295.
[14] L. Y. Li, J. R. Klim, R. Derda, A. H. Courtney, L. L. Kiessling, Proc. Natl. Acad. Sci. USA 2011, 108, 11745 – 11750.
[15] A. K. Yetisen, M. S. Akram, C. R. Lowe, Lab Chip 2013, 13, 2210 – 2251; S. M. Z. Hossain, R. E. Luckham, M. J. McFadden, J. D. Brennan, Anal. Chem. 2009, 81, 9055 – 9064.
[16] D. Tian, Y. Song, L. Jiang, Chem. Soc. Rev. 2013, 42, 5184 – 5209.

Error analysis of deep sequencing of phage libraries I: Peptides censored in sequencing

Wadim L. Matochko and Ratmir Derda*

ABSTRACT

Next-generation sequencing techniques empower selection of ligands from phage display libraries because they could detect low-abundant clones and quantify changes in the copy numbers of clones without excessive selection rounds. Identification of errors in deep sequencing data is the most critical step in this process because these techniques have error rates >1%. Mechanisms that yield errors in Illumina and other techniques have been proposed, but no reports to date described error analysis in phage libraries. Our report focuses on error analysis of 7-mer peptide libraries sequenced by Illumina method. Low theoretical complexity of this phage library, as compared to complexity of long genetic reads and genomes, allowed us to describe this library using convenient linear vector and operator framework. We describe a phage library as Nx1 frequency vector $n = || n_i ||$, where n_i is the copy number of the i^{th} sequence, N is the theoretical diversity, i.e., the total number of all possible sequences. Any manipulation to the library is an operator acting on *n*. Selection, amplification or sequencing could be described as a product of a NxN matrix and a stochastic sampling operator (**Sa**). The latter is a random diagonal matrix that describes sampling of a library. In this report we focus on the properties of **Sa** and use them to define the sequencing operator (**SEQ**). Sequencing without any bias and errors is **SEQ = Sa** I_N , where I_N is a NxN unity matrix. Any bias in sequencing changes I_N to a non-unity matrix. We identified a diagonal censorship matrix (**CEN**), which describes elimination, or statistically significant down-sampling, of specific reads during the sequencing process.

INTRODUCTION

In vitro selection experiments—such as phage display [2,51], RNA display, SELEX and DNA aptamer selection [167,168]—employ large libraries, from which 10^2-10^6 active sequences are identified through iterative rounds of selection and amplification. With the recent emergence of deep sequencing, it became possible to extract a large amount of information from the libraries before and after selection [66,67,143,169-171]. Deep examination of the library is a promising technique for direct evaluation of binding capacities of all binding sequences from one panning experiment. Deep sequencing also allows the characterization of unwanted phenomena in selection, such as amplification bias [10,143].

Analysis of 10^6 reads by deep sequencing gave rise to a large number of errors that were not present in the analysis based on the small number of sequences obtained using the Sanger method. Analysis of errors in information-rich datasets is a problem with over 50 years of history; correction of digital data made of bits or words is a topic of intense research in communication theory [172]. As phage display operates with limited digital sets, data analysis techniques from the communication theory could be applied to phage display. For example, Makowski and co-workers used a positional frequency matrix to calculate the informational content or Shannon entropy of each sequence [34]. This approach could be used to distinguish potential fast growing sequences from potential hits [173]. With introduction of deep sequencing, the problem of error analysis in phage display becomes identical to a classical information theory problem: "reproducing at one point, either exactly or approximately, a message selected at another point." (Shannon, "A Mathematical Theory of Communication" 1948) [174]. The "message" is the sequence information stored in the library. Sequencing process transmits this information and makes either stochastic or predictable errors. Understanding the sources of errors during sequencing could provide mechanisms for bypassing them, for correcting the errors, and for maximizing the amount of useful information received from sequencing.

There are over 10,000 published literature reports that contain the terms "deep sequencing" or "next generation sequencing" or any of the trademark names such as "Illumina" (reference: ISI database). Among these reports, less than 10 published reports describe sequencing of phage-displayed libraries [66,67,85,143,169,171,175-177]. Deep-sequencing efforts in the literature are largely focused on genome assembly and metagenomic analyses. The error analysis techniques tailored for genome assembly cannot be used directly for analysis of phage libraries because the data output from phage library sequencing is very different from the genome assembly. In genome assembly, genomic DNA is shredded into random fragments, and sequenced. The genome is then assembled from these fragments *in silico*. Although multiple fragments cover each area of the genome, the probability to observe two identically shredded fragments is very small. Two exact sequences, thus, could be considered amplification artifacts and removed by error analysis software. On the contrary, in phage display sequencing, the reads are exactly the same length. Duplication of the same read is important for validation of the accuracy of this read. Some researchers focus exclusively on reads that have been observed multiple times and discard singleton reads as erroneous [67]. Within each library, the copy numbers of sequences range continuously by six or more orders of magnitude [66,67,143]. Some phage clones are observed in the entire library only a few times; other clones could be present at copy number of 100,000 per sequencing run [66,67,143]. Unlike multiple cells with identical genomes, each screen is unique: Identical set of sequences with identical copy numbers cannot be obtained even if the screen is repeated due to stochastic number of the screen that contains low copy number of binding clones [178].

Metagenomic analyses of microorganisms recovered from environmental samples [179,180], also known as "microbiome" [181] and "viriome" analyses [182], encountered similar problems to those observed in phage library analysis: Concentration of species observed in a particular sample is unequal [183]. Abundance of species might range by a few orders of magnitude [184]. It is possible that error analysis tools developed in the above areas could find use in phage display sequencing. For example, there are multiple published algorithms for removal of errors from low copy number reads to ascertain that low copy number sequences are new species and not sequencing errors (for example, see [185-187] and references within). Metagenomic analysis is usually more complex than analysis of phage display libraries: First, in metagenomics, the bacterial or viral genes must be assembled from short reads *de novo*. Second, there is no simple relationship between phylogenetic classification of "species" and the observed DNA sequence. Third, the exact number of species in the environment is unknown. On the other hand, sequencing of phage-displayed peptide libraries has none of these problems: (i) it requires no assembly steps because each sequence is covered by one read. (ii) A unique DNA sequence defines a unique "species"; (iii) the theoretical complexity in synthetic libraries is known exactly. For small libraries, such as the library of 7-mer peptides, the complexity, $(20)^7$, is within the reach of next-generation sequencing. We see phage displayed peptide libraries as an ideal model playground for development of optimal error analysis and error correction protocols. It is possible that error analysis developed from phage libraries analysis could then be used in other areas such as genomic and metagenomic analyses.

The errors in sequencing could be divided into "annotated" and "invisible". The "annotated" errors that originate from mis-incorporation of nucleotides are annotated using Phred quality score [126]. These annotated errors are removed during the processing (see below). Examples of "invisible" errors are sequence-specific frame shifts that lead to emergence of truncated reads during the Illumina sequencing [188]. Invisible errors could also originate during the preparation of the libraries for sequencing. Examples are removal of AT-rich fragments during purification of dsDNA [189] and erroneous incorporation of nucleotides during PCR [190,191]. Mutations have the most significant impact on the observed diversity of the library. There are 63 ways to misspell a 21-mer-nucleotide sequence with a one-letter-error (point mutation). The large dynamic range in concentrations of clones in the phage library exacerbates the problem. Clones that are present in high abundance—10^5 copies per read—are more prone to yield errors [143]. For example, we observed that random point mutations convert several short sequence with a copy number of 10^5 to a library of sequences with copy numbers ranging from 1-10^2 [10]. In attempt to unify error analysis into one convenient theoretical framework, we generalized all errors as the following: all errors either lead to disappearance of particular sequence or its conversion to another sequence of similar length. Errors, thus, operate within a finite sequence space, and it should be possible to use elementary linear algebra to generalize most processes that lead to errors.

Theoretical Description

Symbols	Meaning
A, a, f, m, n, k	Unless specified otherwise, normal font designates scalars.
A, a, N, P, ^{1}n, ^{13}n	*Italic* font designates vectors. Different vectors can be distinguished by the left-superscript notation
A, a, Abc, Pan,	**Bold** font designates operators or matrices (here all

Sa	operators are matrices)
$^{1}\mathbf{A}$, $^{f}\mathbf{Sa}$, $^{0.9}\mathbf{Sa}$, $^{0.5}\mathbf{Sa}$	Operators can be distinguished by the left-superscript notation. For sampling operator **Sa**, this notation specifies the sampling fraction of the **Sa** operator.
A_1, a_2, A_i, a_j	Normal font with right subscript designate scalar values of the vector
A_{11}, A_{21}, A_{ij}, A_{ii}	Normal font with two right subscripts designate scalar values of the 2D matrix
$\| A_1 \ldots A_5 \|$	Description of the scalar elements in the vector
$\| A_{ij} \ldots A_{ii} \|$	Description of the scalar elements in the matrix
$x \in [A\ B]$	Scalar x belongs to the inclusive scalar interval [A B] i.e., $A \leq x \leq B$
$x \in [A\ B]$	Vector *x* belongs to the "vector interval" [*A B*] i.e., for every element $A_i \leq x_i \leq B_i$
{ A B C … X }	Set where A, B, C, …, X are the unique elements of the set
{ A(a) B(b) … X(x) }	Multiset (2-tuple) where A, B, … X are the unique elements and a, b, x are the scalars describing the copy numbers of the A, B, X elements
$\mathbf{I_N}$	Unity matrix of the Nth order, i.e., NxN matrix $\| A_{ij} \|$, A_{ij}=□$_{ij}$ (Kroneker delta)

Table 1. Symbols and definitions used in the theoretical description section

Operator description of the phage display library and selection process.
In our previous reports, we described the phage library as a *multiset*, or a set in which members can appear more than once [192]. This description also simplifies the analysis of the errors in these libraries. Multiset description represents a library with N theoretical members as an ordered set of N sequences and Nx1 copy number vector (*n*) with positive integer copy numbers (Figure 1A). Any manipulation of a phage library—such as erroneous reading or selection—changes the numbers within the copy number vector. All manipulations to the multiset, thus, could be described by operators (**Op**) that convert vector n_1 to another vector n_2 as: $n_2 = \mathbf{Op}\ n_1$ (Figure 2C). For Nx1 vector, the operator is NxN matrix. If elements are selected or eliminated independently of one another, the NxN matrix is diagonal (Figure 2D). This approach is uniquely convenient for libraries of short reads. For example, a library of 7-mers contains exactly $20^7 = 10^9$ peptides and is described completely using a 10^9-element vector. This size is accessible to the computational capacity of most desktop computers. Extending this approach to libraries made of longer reads, such as antibodies, is possible in theory. In practice, however, other methods might be more effective.

In operator notation, phage display can be described as:

$$Sel = \mathbf{Pan}\ Naive \tag{1}$$

where *Naive* is the copy number vector for naïve library, *Sel* is copy number vector after panning and **Pan** is a panning operator. In standard phage display, the

Pan operator is a complex product of all manipulation steps (binding, amplification, dilutions, etc). If a screen uses no amplification and uses deep-sequencing [66,85], or large-scale Sanger sequencing [193,194] to analyze the enrichment, it might be possible to define the panning process as a simple product of two operators:

$$\mathbf{Pan} = {}^{f}\mathbf{Sa}\ \mathbf{K_a} \qquad (2)$$
$$Sel = {}^{f}\mathbf{Sa}\ \mathbf{K_a}\ Naive \qquad (3)$$

Where $\mathbf{K_a}$ is a deterministic "association" operator, which contains association constants for every phage clone present in the library. Description of such operator is beyond the scope of this report and we recommend consulting other reports that attempted to generalize the selection procedure [178]. Another operator in equation (3) is a sampling operator (${}^{f}\mathbf{Sa}$), which describes stochastic sampling of the library with m sequences to yield a sub-library with f*m-sequences, where f $\in$ [0 1] is a sampling fraction. ${}^{f}\mathbf{Sa}$ operator has the following properties, which emanate from physical properties of the sampling procedure:

I. ${}^{f}\mathbf{Sa_i}\ 0 = 0$ (sampling does not create new sequences from non-existing sequences) (4)

II. ${}^{f}\mathbf{Sa}$ is a diagonal operator with diagonal scalar functions $\|\ Sa_{11}\ Sa_{22} \ldots Sa_{NN}\ \|$, $Sa_i(0)=0$.

III. In $B = {}^{f}\mathbf{Sa}\ A$, B is a vector of positive integers; $B_i \geq 0$ and sum(B)=f*sum(A). Integer values ensure that the observable values of the operator have physical meaning. The clone could be observed once (1), multiple times (2, 3, etc) or not observed at all (0).

IV. **Sa** is non-deterministic operator. When applied to the same vector, operator does not yield the same result, but one of possible vectors that satisfy rules (I-III). In other words, ${}^{f}\mathbf{Sa}\ A \neq {}^{f}\mathbf{Sa}\ A$. Majority of the solutions of the operator, however, reside within a deterministic confidence interval ${}^{f}\mathbf{Sa}\ A \in [\,{}^{lo}C\ {}^{hi}C]$

V. As a consequence from (IV), operator **Sa** is non-linear, non-commutative and non-distributive.

VI. Large sum of sampling operators with same f should "average out" to yield $\mathbf{I_N}$ unity matrix.

$$({}^{f}\mathbf{Sa}_1 + {}^{f}\mathbf{Sa}_2 + {}^{f}\mathbf{Sa}_3 + \ldots\ {}^{f}\mathbf{Sa}_k)/k \rightarrow f * \mathbf{I_N}, \quad \text{as } k \rightarrow \infty \qquad (5)$$

The **Sa** operator is simple to implement as a random array indexing function in any programming language (for example, see Supporting Scheme S1, S2). It might be possible to express ${}^{f}\mathbf{Sa}$ analytically for any f as a diagonal matrix (Figure 1D). In this report, we use numerical treatment by an array sampling function because it is more convenient for multisets of general structure. We tested the random indexing implementation to show that the sampling algorithm yields a normal distribution for a large number of samples (Supporting Figure S1). Despite the simplicity of ${}^{f}\mathbf{Sa}$ implementation—entire code is <30 lines in MatLab—the script allows rapid calculation of the results of ${}^{f}\mathbf{Sa}$ for a multiset of reasonable size (several million sequences, Figure 4-5) on a desktop computer.

We evaluated the behaviors of ${}^{0.5}\mathbf{Sa}$ for several multisets. Even in small multisets, such as { A(1) B(2) C(3) D(4) } made of four unique and 10 total elements, ${}^{0.5}\mathbf{Sa}$ **{** A(1) B(2) C(3) D(4) **}** operation yields large number of solutions that have equal probability, termed "redundant solutions" (e.g. solutions that have equal probability in Figure 3B). Redundancy depends on the structure of the multiset (Figure 2S). This redundancy makes calculations of all probable solutions of **Sa** impractical. For sets even with 5-6 unique elements, identification of all vectors B, which satisfy equation $B = {}^{f}\mathbf{Sa}\ A$ and reside within a 95% interval, requires hundreds of thousands of iterations (Figure S2, S3). On the other hand, calculation of the confidence interval of each element B_i of the vector B converges rapidly. A multiset $\{A_{1000}\}$={ $A_1(1)$ $A_2(2)$ … $A_{1000}(1000)$ } with 1000 unique elements and 1+2+3+..+1000=500,500 total elements is similar to an average deep sequencing data set (Figure 4). Calculation of all probable solutions of ${}^{0.5}\mathbf{Sa}\{A_{1000}\}$ is beyond the capabilities of most computers. On the other hand, the 99.9% confidence interval of all elements of vector $B = {}^{0.5}\mathbf{Sa}\{A_{1000}\}$ can be calculated in ~2 minutes on an average desktop computer. The red dots in Figure 4 are ${}^{lo}C_i$ and ${}^{hi}C_i$ or the 99.9% high and low confidence interval of all elements B_i (Figure 4).

The sampling operator is the most important in phage display because sampling of libraries occurs in every step of the selection and preparation of libraries for sequencing. The stochastic nature of sampling operators makes two identical screens "similar within a confidence interval". Solving equation (1) exactly is not possible, but it should be possible to estimate the solution within a confidence interval.

$$Sel \in [\ {}^{lo}\mathbf{K_a}\ Naive;\ \ {}^{hi}\mathbf{K_a}\ Naive\ ] \qquad (6)$$

Where ${}^{lo}\mathbf{K_a}$ and ${}^{hi}\mathbf{K_a}$ are diagonal matrices of the upper and lower confidence intervals for the association constants. Simulation of the behavior of the **Sa** operator (Figure 3, S3) suggests that the relative sizes of the confidence intervals might be impractically large when the copy numbers of sequences are <10.

Multiple sampling events of the **Sa** operator yield a normal distribution for each element of the vector (Figure 3). Fitting this normal distribution could yield a "true" value of the process. This process is identical to extrapolation of the average from the normal distribution of noisy data. Multiple algorithms for such extrapolation exist for one and multi-dimensional stochastic processes [195,196]. We believe that **Sa** behaves as one-dimensional stochastic process and it might be possible to extrapolate the true value of the sampling from 7-10 repeated instances of **Sa** (i.e., the number of data sufficient to fit a 1D normal distribution). The necessary practical steps towards solving the equation (3) or (10) are the following: (i) Eliminate or account for any bias not related to binding (e.g. growth bias). (ii) Repeat the screen several times. (iii) **Measure** all copy numbers of all sequences, including zero values, with high confidence. Requirement (i) has been an ongoing effort in our group [10,112] and other groups [33-35,197]; for review see [10,198].

Deep sequencing makes it simple to satisfy the requirement (ii) and obtain multiple instances of the same experiment. For example, we described the Illumina sequencing method that allows using barcoded primers to sequence 18 unrelated experiments in one deep sequencing experiment [141]. We recently scaled this effort to 50 primer sets and evaluated the performance replicas of simple selection procedures (in preparation).

Measurement of the copy numbers of sequences is a separate problem, which can be described using the same sampling operators and bias operators that describe how library is skewed by each preparation step. For example, isolation of DNA by gel purification disfavors AT-rich sequences, whereas PCR favors sequence with within specific GC-content range [189]. The *real* sequence abundance in the any phage library (${}^{\mathrm{real}}n$), hence, has to be derived from the *observed* sequence abundance (${}^{\mathrm{obs}}n$) by solving this equation:

$$ {}^{\mathrm{obs}}n = ({}^{f4}\mathbf{Sa\ An}) ({}^{f4}\mathbf{Sa\ Seq}) ({}^{f3}\mathbf{Sa}\ \boldsymbol{PCR}) ({}^{f2}\mathbf{Sa\ Is}) ({}^{f1}\mathbf{Sa}\ \boldsymbol{Gr})\ {}^{real}n \tag{7} $$

In this equation, each operator in brackets describes a bias at a particular step and ${}^{f}\mathbf{Sa}$ describes sampling at that step, and f1-f5, describe the sampling fractions. The bias in growth (**Gr**), isolation (**Is**), PCR amplification (**PCR**), and sequencing (**Seq**) could be related to the nucleotide sequences. The **An** analysis operator is a matrix that describes retaining, discarding or correcting the sequence (Figure 2B). An ideal **An** operator could compensate for the biases introduced by another operator (Figure 2C). To define such operator, the equation (7) could be potentially solved using repeated sequencing of a well-defined model library. In the next, applied section, we examine the real deep sequencing data and identify conditions under which these operators could be, at least partially defined.

Analysis or the error cutoff in deep sequencing reads.

All next generation sequencing techniques provide quality score (Phred Score) for every sequenced nucleotide. In Illumina sequencing, this score is related to probability of the nucleotide to be correct [199]. In low-throughput Sanger sequencing, Phred score monotonously decreases with read length and the mechanisms that yield errors in capillary electrophoresis are well understood. Common practice in Sanger sequencing is to discard all reads after the first nucleotide with Phred score of 0. In next-generation sequencing the filtering the reads is usually more stringent:

A. Discard reads that have at least one read that has score lower than "cutoff".

B. Discard reads that had cumulative Phred score lower than cutoff.

C. Use a combination of A and B (accept reads with minimal cutoff and minimal cumulative score)

Many of the error analyses in the area of deep sequencing are designed for genetic reads, which have variable length and unknown sequence throughout the whole read. Analysis of the reads in a phage display library is a simpler problem because phage-derived constant adapter regions flank the variable reads. Identification of the adapter region is a necessary first step in the analysis. Reads, in which the adapters cannot be mapped, cannot be used and we designed algorithms to recover reads even if adapters were hampered by truncation, deletion or mutation [143]. We observed that the reads flanked by the erroneous adapters had significantly higher error rate than reads flanked by "perfect" adapters (unpublished). Example of mapping of flawed reads in Illumina sequencing is provided in ERROR_TAG_data0001.txt (see Methods section). In the remaining sections, we analyze the population of the sequences preceded by a "perfect" adapter to identify possible sequence-specific biases.

We analyzed a typical library sequenced by Illumina using various cutoffs (Figure 4). We analyzed a 33-bp segment of the library that contained variable seven amino acids and a constant region and GGGS terminus. A simple cutoff that discards reads with Phred<1 nucleotides yields library termed ${}^{1}n$, which had an average 95% accuracy of the 33-nucleotide read. Reads that do not contain Phred=0 nucleotide, rarely contain multiple low-quality reads. The ${}^{1}n$ library was bi-modal: 80% of the reads had overall accuracy of 99%, very few reads with accuracy 5-90% and significant number of reads with accuracy of 1% (Figure 4D-E). These observations suggest that reads can be divided into (i) reads free of errors (ii) reads with multiple errors.

An example of a more stringent cutoff is elimination of reads with Phred<13 nucleotides; this process yielded a library ${}^{13}n$ in which every nucleotide had >95% confidence. The number of total reads in ${}^{13}n$ was 10% less than number of reads in ${}^{1}n$, i.e., sum(${}^{13}n$) = 0.9sum(${}^{1}n$). The observed average read accuracy of the read in the ${}^{13}n$ library was 99.2%. Theoretically, the 0.95 confidence cutoff in a 33-mer nucleotide could yield reads with accuracy as low as $(0.95)^{33}$=18%. In practice, the probability to find reads with multiple nucleotides of 95% accuracy was vanishingly small. Specifically, among 500,000 reads, the lowest observed cumulative accuracy was 77%. Such result, for example could be obtained in a sequence that has 27 "perfect" nucleotides and 5 nucleotides with Phred=13 score: $(1)^{27}(0.95)^{5} = 0.77$. Applying the most stringent cutoff to eliminate all reads with Phred<30, yielded a library ${}^{30}n$ in which every nucleotide had 99.9% confidence. The average confidence of the reads improved subtly from 99.2% to 99.6%. The number of total reads in ${}^{30}n$ was 30% less than number of reads in ${}^{13}n$, i.e., sum(${}^{30}n$) = 0.7sum(${}^{13}n$). It was not clear whether such cutoff is an improvement or a detriment for analysis. In the next section, we examined how frequency of the members of the library changed upon application of each error cutoff.

Example of error analysis: sequence specific censorship during Phred quality cutoff.

If errors occur by random chance, they should be uniformly distributed in all sequences. Removal of erroneous read, in that case, should be identical to sampling of the library by ${}^{f}\mathbf{Sa}$ operator where f is the sampling fraction. For example, consider removal of Phred<13 nucleotides from an unfiltered library (process denoted as ${}^{1}n \rightarrow {}^{13}n$). From experiments, we know that $\Sigma({}^{13}n) = 0.9\Sigma$

(^{1}n); if errors were distributed in sequences at random, the ^{1}n and ^{13}n vectors should be related as:

$$^{13}n = {}^{0.9}\mathbf{Sa}(^{1}n) \qquad (8)$$

The solutions should reside within a confidence interval.

$$^{13}n \in [^{lo}C\ ^{\mathrm{hi}}C] \qquad (9)$$

If errors occur preferentially in specific reads, the frequency of these reads should occur beyond the confidence interval of the $^{0.9}\mathbf{Sa}$. This process could be described by a diagonal matrix **Bias**

$$^{13}n = {}^{0.9}\mathbf{Sa}(\boldsymbol{Bias}\,(^{1}n)\,) \qquad (10)$$

The elements of the diagonal matrix ***Bias*** = $\|\ \mathrm{B_{ii}}\ \|$ could be estimated as following.

$$^{13}n_i \in [^{\mathrm{lo}}C_{\mathrm{i}}\ ^{\mathrm{hi}}\mathrm{C_i}],\ \mathrm{B_{ii}} = 1 \qquad (11)$$

$$^{13}n_i < {}^{\mathrm{lo}}C_{\mathrm{i}},\ \mathrm{B_{ii}} = {}^{13}n_{\mathrm{i}} / (0.9\ ^{1}n_{\mathrm{i}}) \qquad (12)$$

Figure 5C describes the representative solution of the $^{0.9}\mathbf{Sa}(^{1}n)$ (green dots) and confidence interval (blue lines). Supporting Scheme S3 describes the script that calculated this interval from multiset ^{1}n, described as a plain text file PhD7-Amp-0F.txt, using 10,000 iterative calculations of $^{0.9}\mathbf{Sa}(^{1}n)$. This calculation required ~2 hours on a desktop computer. Confidence interval was estimated as the minimum and maximum copy number found after 10,000 iterations. In this approximation of the confidence interval, for sequences with the copy number <10 before sampling, it was impossible to determine whether the sequence disappeared due to random sampling or due to bias. The values of **Bias** operator cannot be defined for these sequences and it could be assumed to be 1 (see equation 11). For copy number >10, however, sequence-specific bias can be readily detected. We observed that removal of Phred<13 reads yielded a multiset in which a large number of sequences deviated beyond the confidence interval (Figure 2D). Their sequences could be readily extracted by comparing the vector ^{13}n with the vector of the lower confidence intervals ^{lo}C (see equation 12). The solution of the **Bias** can be illustrated graphically (Figure 2E). Top 30 censored sequences are listed in Table 1; the other sequences can be found in the supporting information (file PhD7-Amp-0F-13F-CEN.txt) .

We performed similar calculations for $^{1}n \rightarrow {}^{30}n$ and $^{13}n \rightarrow {}^{30}n$ processes. The latter process is the most interesting because ^{13}n library has all nucleotides within acceptable confidence range (>95%) and the distribution of cumulative quality suggested that errors, on average, do not cluster in one read (Figure 5). The $^{13}n \rightarrow {}^{30}n$ conversion eliminated 30% of the reads and copy numbers of many sequences deviated significantly from the random sampling: these sequences are represented by green dots outside the blue confidence interval in Figure 5H. Top 30 sequences are listed in Table 2. The censorship is not only sequence specific but also position specific. In sequences that had been censored during the ${}^{13}n \rightarrow {}^{30}n$ process, lower quality reads clustered around 3-4 specific nucleotides (supporting information Figure S5).

The mechanism that leads to disappearance of censored sequences is not currently clear. We attempted to identify common motifs in censored sequences using clustering and principal component analyses based on Jukes-Cantor distance between sequences or identification of motifs using multiple unique sequence identifier software (MUSI) [175]. These approaches could not detect any property common to censored reads, which would make them significantly different from the other, non-censored reads. Still, we hypothesize that the observed censorship represents sequence specific errors, which occur in every time such sequence passes though the Illumina analyzer. For example, the sequences listed in Table 1, 2 and supporting files were censored in five independent experiments, which were pooled and processed simultaneously in one Illumina run. Analysis of other instances of Illumina sequencing performer by other groups could help prove (or disprove) that censorship is indeed sequence-specific and experiment-independent. Sequence-specific censorship during Illumina analysis has been described in other publications [199]. The observations presented above suggest that reading of some sequences in phage libraries does not yield an accurate copy number. Even if these sequences were enriched due to binding, their apparent copy number in sequencing would be decreased due to sequencing bias. If the magnitude of bias is known, however, such error could be corrected. We anticipate that other biases could be calculated for these and other libraries in similar fashion. Their calculation extend beyond the scope of this manuscript and it will be performed in out next publication.

DISCUSSION

Significance and Transformative Potential of library-wide error correction.

In the Medicinal Chemistry field, structure-activity relationships (S.A.R.) and pharmacophores are built using both positive and negative observations. It is the negative results that bear most significance in these studies because they allow mapping the range of the conditions under which particular structure no longer works. For example, S.A.R. of an R group of a ligand might be built on the following observations: Ligand binds to the target when R group in the specific position is methyl or ethyl; changing R to *iso*-propyl and *tert*-butyl ablates the binding. Conclusion is: the R group must be a small alkyl group. Analogous situation is found in SAR of peptide ligands: the most important information from alanine scan mutagenesis is loss of function, because it helps identifying the important residues. Interestingly, loss-of-binding conclusions are never applied to phage display. Phage display field is driven by positive results. Most publications report and follow up only on sequences enriched in the screen and consider only large copy numbers interesting. All papers focus on sequences that were found. Very few papers in phage display ask why other sequences were **not** found.

One of the reasons why phage display is not used for SAR-type analysis is because negative observations in phage library cannot be determined with high confidence. From practical point of view, measuring zero with high confidence requires the largest number of observation (the highest depth of sequencing). The payoff, however, is immense: one screen with "confident zeros" could potentially yield SAR for every possible substitution of every possible amino acid. We refer to this (theoretical) possibility, as "Instant SAR" and its condensed theoretical form is described in equation (3) or equations (9) and (10). This reports demonstrates that the depth of sequencing is not the only problem towards this goal. Accurate estimate of negative results requires complete characterization of the origins of errors in sequencing which yield false negative values by censoring certain sequencing. Other types of censorship, such as growth bias, should be characterized an eliminated as well. As the phage display field is currently focused on positive results, the need for optimal error corrections and recovery of erroneous reads is low. With the rise of SAR-type applications in phage display, error correction will be recognized as the most significant barrier because it could leads to improper assignment of low frequencies and negative results. Improved error correction strategies could assign a lower confidence to the sequence instead of eliminating the errors and labeling them as confident zero. Proper mathematical framework, possibly similar to the one used in this manuscript, could be then used to carry all confidence intervals through calculations to yield reliable SAR-type data.

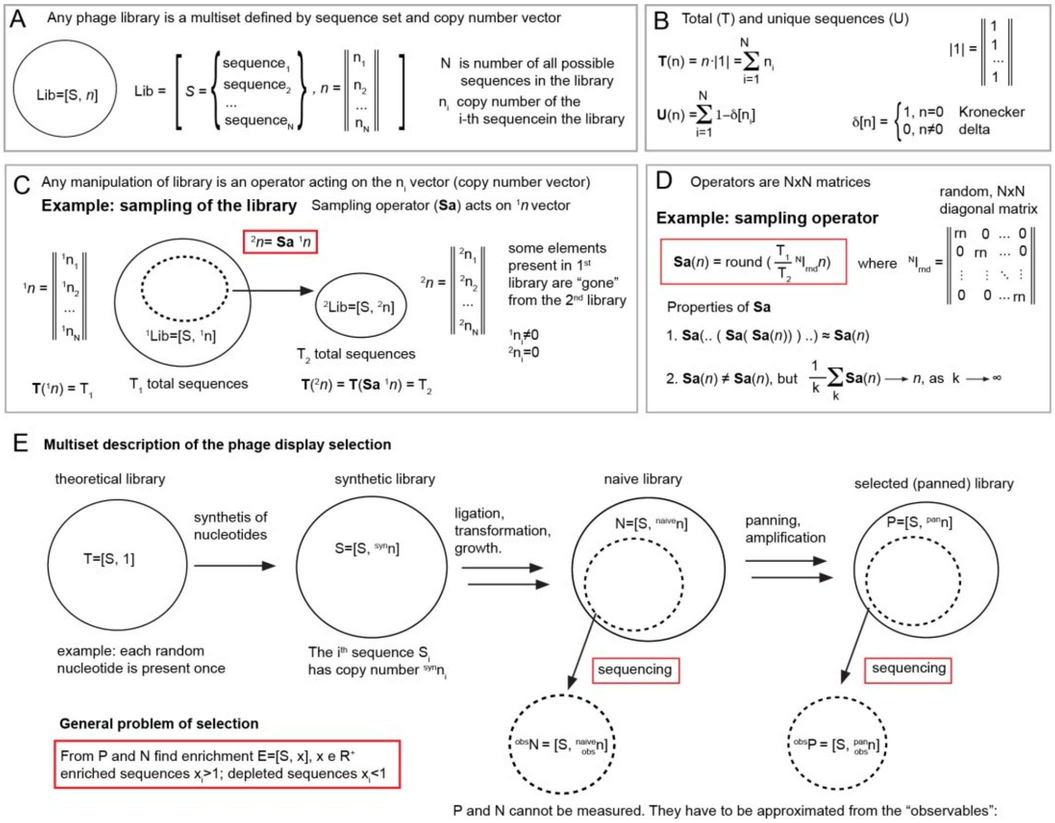

A Any phage library is a multiset defined by sequence set and copy number vector
Lib=[S, n]
Lib = [S = {sequence1, sequence2, ..., sequenceN}, n = ||n1, n2, ..., nN||]
N is number of all possible sequences in the library
ni copy number of the i-th sequencein the library
B Total (T) and unique sequences (U)
T(n) = n·|1| = Σ ni
U(n) = Σ 1−δ[ni]
|1| = ||1 1 ... 1||
δ[n] = {1, n=0; 0, n≠0 Kronecker delta
C Any manipulation of library is an operator acting on the ni vector (copy number vector)
Example: sampling of the library
Sampling operator (Sa) acts on 1n vector
2n= Sa 1n
1n = ||1n1, 1n2, ..., 1nN||
1Lib=[S, 1n]
2Lib=[S, 2n]
2n = ||2n1, 2n2, ..., 2nN||
some elements present in 1st library are "gone" from the 2nd library
1ni≠0
2ni=0
T2 total sequences
T(1n) = T1
T1 total sequences
T(2n) = T(Sa 1n) = T2
D Operators are NxN matrices
Example: sampling operator
random, NxN diagonal matrix
Sa(n) = round (T1/T2 NIrnd n)
where NIrnd = ||rn 0 ... 0; 0 rn ... 0; ⋮ ⋮ ⋱ ⋮; 0 0 ... rn||
Properties of Sa
1. Sa(.. (Sa(Sa(n))) ..) ≈ Sa(n)
2. Sa(n) ≠ Sa(n), but 1/k Σk Sa(n) → n, as k → ∞
E Multiset description of the phage display selection
theoretical library
T=[S, 1]
synthetis of nucleotides
example: each random nucleotide is present once
synthetic library
S=[S, synn]
The ith sequence Si has copy number synni
ligation, transformation, growth.
naive library
N=[S, naiven]
panning, amplification
selected (panned) library
P=[S, pann]
sequencing
sequencing
obsN = [S, naiveobsn]
obsP = [S, panobsn]
General problem of selection
From P and N find enrichment E=[S, x], x e R+
enriched sequences xi>1; depleted sequences xi<1
P and N cannot be measured. They have to be approximated from the "observables":

Figure 1. (A) Phage library can be described by multisets made of S={sequence set} and *n*=||vector of copy numbers||. Any change to library can be described as function/operator acting on the *n*. (B) Relevant functions are calculations of total sequences and unique sequences. (C) Any transformation of library to another library is an operator acting on *n*. Sampling of libraries to yield a sub-library is the most important operator. (D) It can be described as NxN matrix. Specifically, **Sa** is a diagonal matrix of values derived from random distribution. Rounding function is necessary to ensure the physical meaning of the sampling results. **Sa** acting on the same vector yields one of many vectors that have the same number of total elements. As consequence, **Sa** is non-linear, non-distributive and non-commutative operator. Average of many **Sa** operators is a scalar (dilution factor). (E) Any screen of any library can be described as operators acting on the copy number vectors of the naïve (or theoretical) library. Copy number vectors cannot be observed directly. They have to be measured through sequencing. As sequencing contains sampling process (**Sa** operator), result of sequencing is non-deterministic. Sequencing yields one of many possible ***observed*** copy number vectors, none of which are equal to the **real** copy number vector.

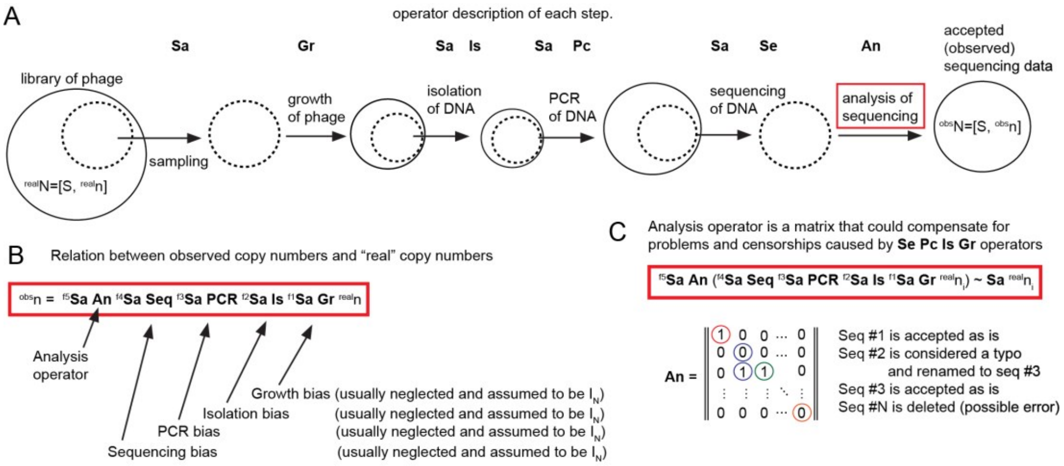

A
operator description of each step.
Sa
Gr
Sa Is
Sa Pc
Sa Se
An
accepted (observed) sequencing data
library of phage
$^{real}N=[S, {}^{real}n]$
sampling
growth of phage
isolation of DNA
PCR of DNA
sequencing of DNA
analysis of sequencing
$^{obs}N=[S, {}^{obs}n]$
B
Relation between observed copy numbers and "real" copy numbers
$^{obs}n = {}^{f5}\mathbf{Sa\ An}\ {}^{f4}\mathbf{Sa\ Seq}\ {}^{f3}\mathbf{Sa\ PCR}\ {}^{f2}\mathbf{Sa\ Is}\ {}^{f1}\mathbf{Sa\ Gr}\ {}^{real}n$
Analysis operator
Growth bias (usually neglected and assumed to be I_N)
Isolation bias (usually neglected and assumed to be I_N)
PCR bias (usually neglected and assumed to be I_N)
Sequencing bias (usually neglected and assumed to be I_N)
C
Analysis operator is a matrix that could compensate for problems and censorships caused by **Se Pc Is Gr** operators
$^{f5}\mathbf{Sa\ An}\ ({}^{f4}\mathbf{Sa\ Seq}\ {}^{f3}\mathbf{Sa\ PCR}\ {}^{f2}\mathbf{Sa\ Is}\ {}^{f1}\mathbf{Sa\ Gr}\ {}^{real}n_i) \sim \mathbf{Sa}\ {}^{real}n_i$
$\mathbf{An} = \begin{Vmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 1 & & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{Vmatrix}$
Seq #1 is accepted as is
Seq #2 is considered a typo and renamed to seq #3
Seq #3 is accepted as is
Seq #N is deleted (possible error)

Figure 2. Operator description of the deep-sequencing process. (A) A library of phage must be processed before deep sequencing. Each step involves sampling, which is either a deliberate partitioning of the sample or random loss of the sample. Each sample preparation state could (and does) introduce bias in sequence abundance. Each step, thus, is an operator chat changes the *n* vector. (B) If we ignore bias during preparation, operators could be approximated as unity vectors, and sequencing could be represented as a product of sampling and analysis operators. (C) Analysis operator (**An**) is a binary decision matrix, which describes what sequences are and are not considered as errors. Decisions, such as removal of sequences or correction of sequences, are the most important because they decide which "*observed*" sequences are considered "*real*". To make analysis of the selection process meaningful, the same **An** operator should be used in all analyses.

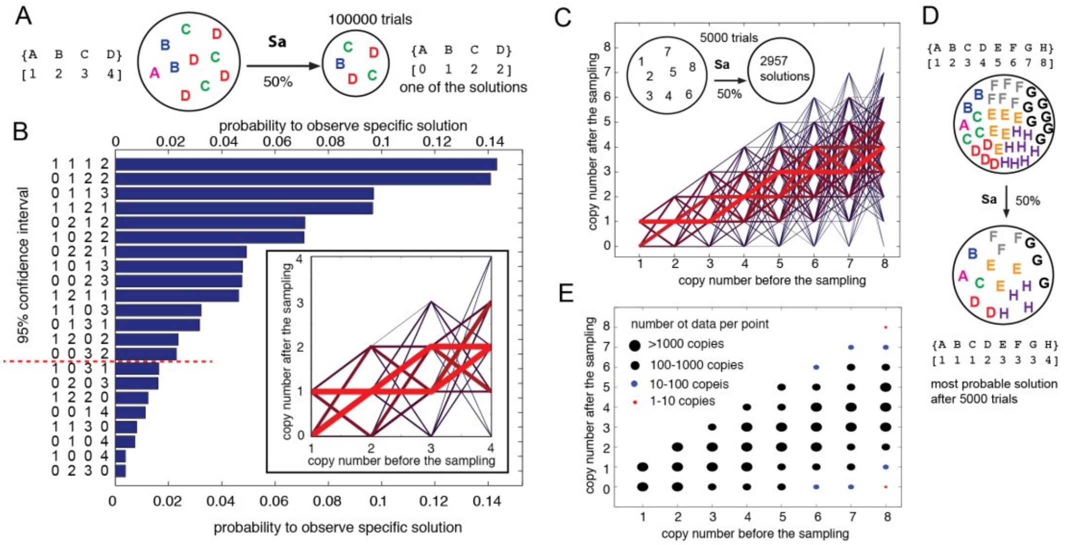


Figure 3. (A) Testing the sampling operator implemented as random indexing function using a model multiset. (B) In 100,000 trials, we observed 22 unique solutions from which 14 resided in a 95% confidence interval. Solutions with 0 and 1 copies of element A were found at equal abundances ("redundant solutions"). Inset describes all solutions as lines. Red thick lines describe the most probable solutions; thinner lines with more blue shade describe less probable solutions. (C) Sampling of larger multisets yields more possible solutions (here, 2957 in 5000 trials). (D) Multiset in graphical form. Panel (C) describes probability to observe a particular solution; panel (E) describes probability to observe a particular copy number after sampling. While B and C are the most accurate representations of the confidence intervals—the thinnest blue lines describe solutions outside the confidence interval—this representation is impractical due to large number of redundant solutions in larger multisets. Confidence interval could be extrapolated from distributions of individual copy numbers (E): red dots are on or outside the confidence interval.

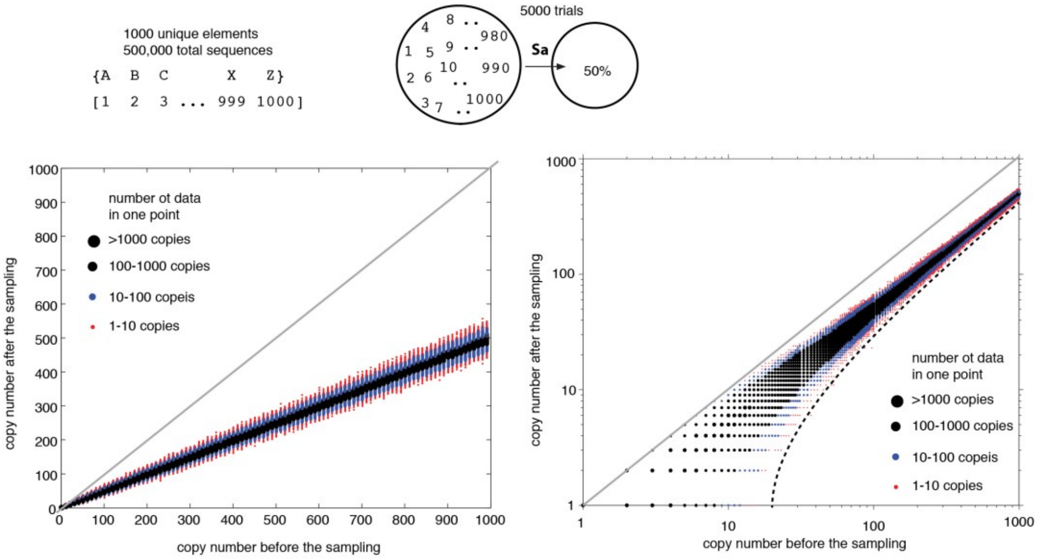
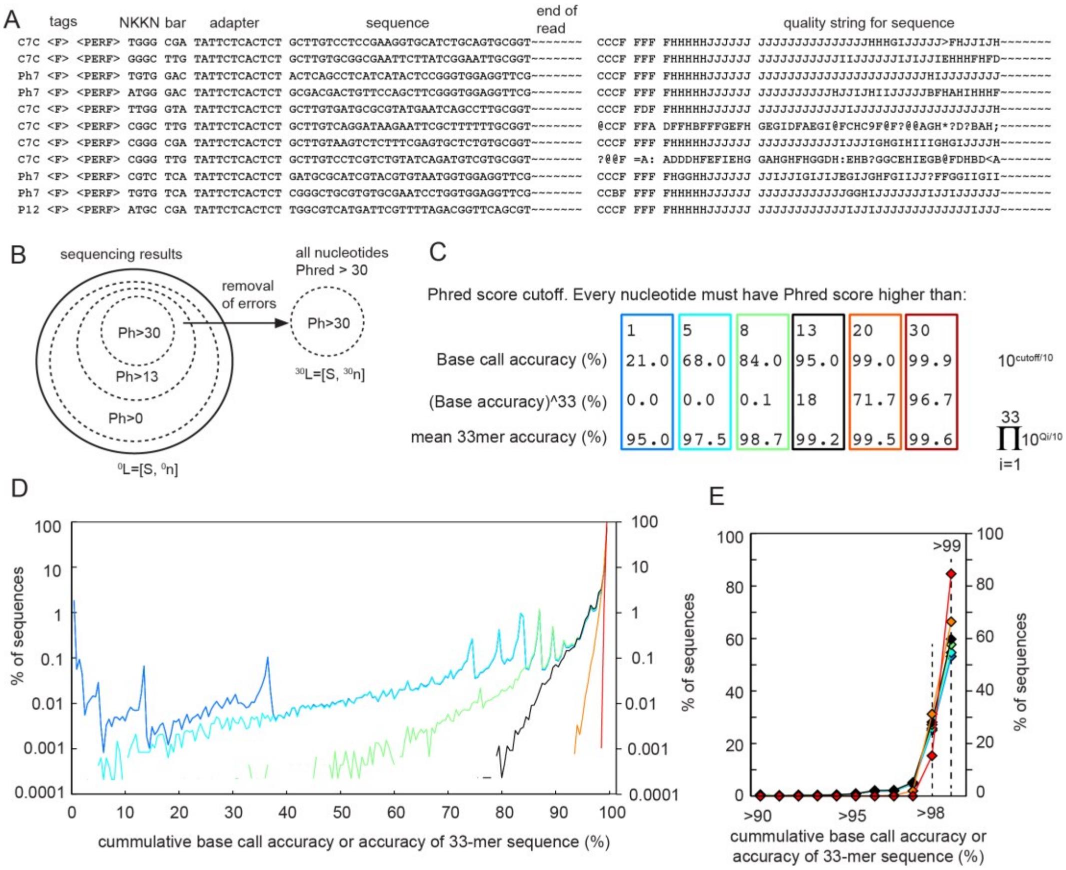


Figure 4. (A) Testing the sampling operator using a large multiset made of 1000 unique elements with 1000 different copy numbers. Images describe linear and log-scale representation of the confidence interval of the sampling operator. Solutions beyond this interval were not observed in 5000 trials. Dotted line represents an overestimate of the 99.9% confidence interval (for details, see Figure S4). Most probable outcomes of the **Sa** operator have either zero or one unique sequence beyond this interval. This line is used in subsequent sections (figure 5, 6). We note that distributions of the copy numbers have well-defined shape; according to central limit theorem, it is a normal distribution. With enough replicas, it should be possible to extrapolate the center of this distribution, define the solutions explicitly and bypass the stochastic nature of the **Sa** operator.



```
C7C <F> <PERF> TGGG CGA TATTCTCACTCT GCTTGTCCTCCGAAGGTGCATCTGCAGTGCGGT~~~~~~~  CCCF FFF FHHHHHJJJJJ JJJJJJJJJJJJJJHHHGIJJJJJ>FHJJIJH~~~~~~~
C7C <F> <PERF> GGGC TTG TATTCTCACTCT GCTTGTGCGGCGAATTCTTATCGGAATTGCGGT~~~~~~~  CCCF FFF FHHHHHJJJJJ JJJJJJJJJJJIIJJJJJJIJIJJIEHHHFHFD~~~~~~~
Ph7 <F> <PERF> TGTG GAC TATTCTCACTCT ACTCAGCCTCATCATACTCCGGGTGGAGGTTCG~~~~~~~  CCCF FFF FHHHHHJJJJJ JJJJJJJJJJJJJJJJJJJJJJHIJJJJJJJJ~~~~~~~
Ph7 <F> <PERF> ATGG GAC TATTCTCACTCT GCGACGACTGTTCCAGCTTCGGGTGGAGGTTCG~~~~~~~  CCCF FFF FHHHHHJJJJJ JJJJJJJJJHJJIJHIIJJJJJBFHAHIHHHF~~~~~~~
C7C <F> <PERF> TTGG GTA TATTCTCACTCT GCTTGTGATGCGCGTATGAATCAGCCTTGCGGT~~~~~~~  CCCF FDF FHHHHHJJJJJ JJJJJJJJJJJIJJJJJJJJJJJJJJJJJJJH~~~~~~~
C7C <F> <PERF> CGGC TTG TATTCTCACTCT GCTTGTCAGGATAAGAATTCGCTTTTTTGCGGT~~~~~~~  @CCF FFA DFFHBFFFGEFH GEGIDFAEGI@FCHC9F@F?@@AGH*?D?BAH;~~~~~~~
C7C <F> <PERF> CGGG CGA TATTCTCACTCT GCTTGTAAGTCTCTTTCGAGTGCTCTGTGCGGT~~~~~~~  CCCF FFF FHHHHHJJJJJ JJJJJJJJJJIJJJIGHGIHIIIGHGIJJJJH~~~~~~~
C7C <F> <PERF> CGGG TTG TATTCTCACTCT GCTTGTCCTCGTCTGTATCAGATGTCGTGCGGT~~~~~~~  ?@@F =A: ADDDHFEFIEHG GAHGHFHGGDH:EHB?GGCEHIEGB@FDHBD<A~~~~~~~
Ph7 <F> <PERF> CGTC TCA TATTCTCACTCT GATGCGCATCGTACGTGTAATGGTGGAGGTTCG~~~~~~~  CCCF FFF FHGGHHJJJJJ JJIJJIGIJIJEGIJGHFGIIJJ?FFGGIIGII~~~~~~~
Ph7 <F> <PERF> TGTG TCA TATTCTCACTCT CGGGCTGCGTGTGCGAATCCTGGTGGAGGTTCG~~~~~~~  CCBF FFF FHHHHHJJJJJ JJJJJJJJJJJGGHIJJJJJJIJJIJJJJJJ~~~~~~~
P12 <F> <PERF> ATGC CGA TATTCTCACTCT TGGCGTCATGATTCGTTTTAGACGGTTCAGCGT~~~~~~~  CCCF FFF FHHHHHJJJJJ JJJJJJJJJJJIJJIJJJJJJJJJJJJJIJJJ~~~~~~~
```

Phred score cutoff. Every nucleotide must have Phred score higher than:

	1	5	8	13	20	30	
Base call accuracy (%)	21.0	68.0	84.0	95.0	99.0	99.9	$10^{cutoff/10}$
(Base accuracy)^33 (%)	0.0	0.0	0.1	18	71.7	96.7	
mean 33mer accuracy (%)	95.0	97.5	98.7	99.2	99.5	99.6	$\prod_{i=1}^{33} 10^{Qi/10}$

Figure 5. (A) Representative lines from the intermediate file from Illumina deep sequencing analysis (for more information see ERROR_TAG_data0001.txt in Methods section and our previous publication [143]). The reads have been parsed to identify adapters and barcodes. Each read has been tagged according to the library type, direction of the read and quality of the adapter regions. We use this intermediate library to identify reads that harbor erroneous nucleotides. (B) Multiset view of the intermediate library. The library contains subsets that have low, medium and high quality reads. Error filtering of this intermediate library to eliminates any read with Phred score below 30 yields a high quality library of reads ^{30}L. (C) Mean accuracy of the reads in the library after error filtering ranges from 95% to 99.6%. Even for very low quality cutoff, Phred>1, the average read quality is 95%. (D) Distribution of cumulative read accuracy in libraries processed using different cutoffs. (E) Linear plot of the data presented in (D) with zoom in on the region with <90% cumulative accuracy.

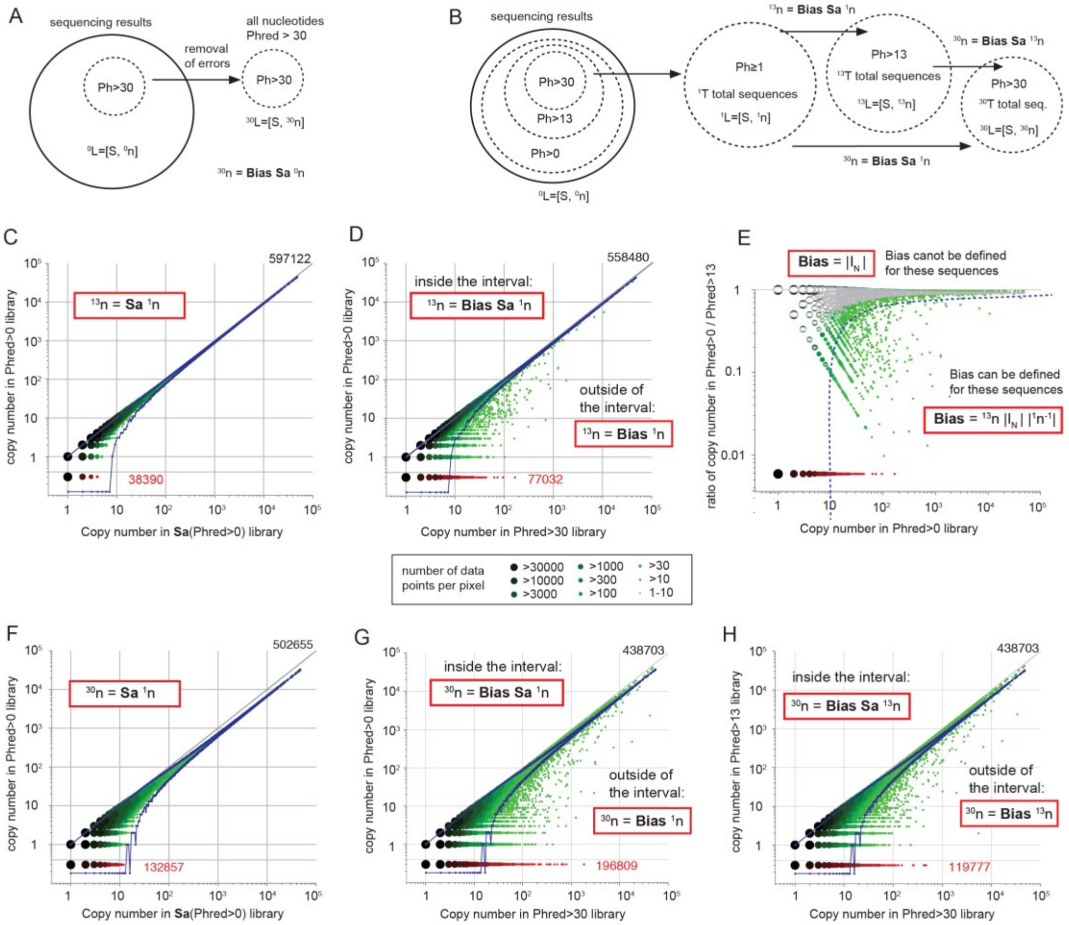


Figure 6. (A) Operator and multiset description of the error filtering procedure. Applying a Phred>30 cutoff to library filtered by Phred>1 cutoff (^{1}n) yields a subpopulation of the library (^{30}n). If errors are sequence-independent, the ^{1}n → ^{30}n process should be identical to random sampling (^{30}n=**Sa** ^{1}n). Any sequence-specific bias (**Bias**) should be detected as deviation from **Sa** ^{1}n. (B) Progressive sampling with more stringent cutoff. (C) Theoretical **Sa** ^{1}n and theoretical 99.9% confidence interval (blue). (D) Observation of statistically-significant deviation from **Sa** operator: dots beyond the blue line represent sequences prone to bias. Red dots represent sequences that disappeared after in ^{1}n → ^{30}n process or during **Sa** ^{1}n sampling. (E) Magnitude of the bias range from 5-100 fold. (F) Bias in sampling of Phred>30 data from Phred>1 data (F is theory, G is observed). (H) Bias upon sampling of Phred>30 data from Phred>13. Many sequences were lost in this sampling and this loss was statistically significant beyond the 99.9% interval. This result shows that some sequences have propensity to harbor low and medium quality reads. Distribution of the errors is sequence specific.

Nucleotide	peptide	^{13}n	^{30}n	^{1o}C	CEN
GGTCCTATGCTGGCTCGTGGT	**GPMLARG**	33518	2340	24732	0.070
GGTAAGGTGCAGGCGCAGTCG	GKVQAQS	24566	13168	18101	0.536
CAGCTGATGAATGCTTCGCGG	QLMNASR	21821	10957	16070	0.502
ATGCTGCCGTCTGTGCTTGAT	MLPSVLD	17619	12446	12955	0.706
GGTACGTGGCTTTCTCGGGGG	GTWLSRG	16583	479	12154	0.029
CAGAGTCCTGATGAGGTTTGG	QSPDEVW	14482	8780	10615	0.606
GCGACGCCGTCGTGGTGGGCT	ATPSWWA	13658	8031	9997	0.588
ACGACGCGTCTTCCGGTTATT	TTRLPVI	11538	8063	8446	0.699
GCGCGTCCGCCTCTGTTTGGT	ARPPLFG	11436	7745	8371	0.677
TGGCCTACGCTGCAGTGGGCG	WPTLQWA	11097	5017	8112	0.452
AGTCAGACGAAGGTGCCGTTG	SQTKVPL	10129	6138	7380	0.606
ACGCTGTTGCAGGCGGCTAGG	**TLLQAAR**	9819	2841	7144	0.289
AATCAGCAGCCGGCTCCTCGG	NQQPAPR	7634	5187	5566	0.679
CGGCTTCCGTCTTGGCATGAG	**RLPSWHE**	7587	3867	5533	0.510
GCTGCTAAGACGCCTACGGAG	**AAKTPTE**	7468	2987	5442	0.400
CTACCTTCATATCATGTGCCT	**LPSYHVP**	7410	4557	5394	0.615
GATGCGGGGTATGTGACTTTG	**DAGYVTL**	7410	4103	5394	0.554
GCGACGACTGTTCCAGCTTCG	ATTVPAS	7287	4514	5290	0.619
AAGCTTCCTGGGTGGTCGGGG	KLPGWSG	6832	340	4970	0.050
GCGTCTACGTTGAAGTGGGCG	ASTLKWA	6776	2279	4928	0.336
AAGCCGGTTCAGCTGGATCAT	KPVQLDH	6744	4687	4906	0.695
GGGGAGACTCGTGCGCCGCTT	**GETRAPL**	6680	4803	4850	0.719
AATCCGATGCAGTCTCGTCCG	NPMQSRP	5928	4135	4297	0.698
TCGTATGCGTCGGAGAAGCGT	SYASEKR	5804	3838	4221	0.661
ACGCCGCAGTGGGCTGGTCAG	**TPQWAGQ**	5602	3638	4063	0.649
ACGCGGGCTGGTCTGGATTTT	TRAGLDF	5538	3275	4007	0.591
CAGCGGCTGCCTCAGACGGCG	**QRLPQTA**	5483	2	3973	0.000
TGGACTGGTTCGTATAGGTGG	**WTGSYRW**	5174	2239	3738	0.433
CATCATGCGCTGCGTTTGGAG	HHALRLE	4993	3196	3610	0.640

Table 1 Top 30 sequences censored during the ^{13}n → ^{30}n process. Bolded sequences could also be found in censorship during the ^{1}n → ^{13}n process (partially described in Table 2). Normal-font sequences are uniquely censored in ^{13}n → ^{30}n process. While typical censorship is a factor of two or three, the highlighted reads are censored by a factor of 10 or more.

Nucleotide	peptide	^{13}n	^{30}n	^{1o}C	CEN
GGTCCTATGCTGGCTCGTGGT	**GPMLARG**	41971	33518	38273	0.799
CATGTGCTTCGTTTTGATACG	HVLRFDT	30513	27073	27804	0.887
CATGTGAAGCCTCTGGTGACG	HVKPLVT	18102	16266	16451	0.899
ACGCTGTTGCAGGCGGCTAGG	**TLLQAAR**	11108	9819	10095	0.884
CAGCGGCTGCCTCAGACGGCG	**QRLPQTA**	10687	5483	9667	0.513
CGGCTTCCGTCTTGGCATGAG	**RLPSWHE**	8794	7587	7966	0.863
GCTGCTAAGACGCCTACGGAG	**AAKTPTE**	8445	7468	7628	0.884
CTACCTTCATATCATGTGCCT	**LPSYHVP**	8442	7410	7640	0.878
GATGCGGGGTATGTGACTTTG	**DAGYVTL**	8241	7410	7452	0.899
GGGGAGACTCGTGCGCCGCTT	**GETRAPL**	7546	6680	6834	0.885
CATGGGCTGTCTCATCGGCTT	**HGLSHRL**	6793	4034	6136	0.594
ACGAGTCCTCGGATTGCGCCT	TSPRIAP	6370	5721	5761	0.898
ACGCCGCAGTGGGCTGGTCAG	**TPQWAGQ**	6244	5602	5641	0.897
TGGACTGGTTCGTATAGGTGG	**WTGSYRW**	5839	5174	5267	0.886
AGTCTGAGGCATGGGTCGTAT	**SLRHGSY**	5401	4425	4882	0.819
TCGGTGGAGTCGGCGTGGAGG	**SVESAWR**	5104	4408	4604	0.864
TCGCCTCATTTGCATGGGGCT	SPHLHGA	4674	4170	4219	0.892
CTGGCGCGTGAGCCTACGTCG	**LAREPTS**	4215	3747	3800	0.889
CATACGGTTCGGACTGGTGAG	**HTVRTGE**	4154	3617	3738	0.871
TCGCGGACTTTGATTGCGCCG	**SRTLIAP**	3620	3236	3258	0.894
GCGGCTGGTCAGCAGTTTCCT	**AAGQQFP**	3510	2790	3151	0.795
GCGACGGGTTGGTCTGCGTTG	**ATGWSAL**	3477	3087	3131	0.888
TCGGAGGCTGAGGCGACGTAT	**SEAEATY**	3389	3023	3039	0.892
CATGTGTATGAGTTTGGGCCG	**HVYEFGP**	3311	2877	2977	0.869
CTTGTGACGACGTGGCCGGCT	**LVTTWPA**	3116	2721	2787	0.873
ACGGGTGTGACGCTTACGGTG	**TGVTLTV**	3111	2437	2791	0.783
GAGTATCGGCTGCTTTATTCG	**EYRLLYS**	2968	1955	2666	0.659
GCGGCGTGGCAGCTTCATAGT	**AAWQLHS**	2801	2491	2515	0.889
TCGGCTACTCAGGCTTCTGTG	**SATQASV**	2791	2356	2501	0.844
CAGGAGCCGCTTCCTGCTTTG	QEPLPAL	2492	2166	2237	0.869
ACGGCGCGGTATCCGTCGTGG	**TARYPSW**	2199	1959	1962	0.891
AATACTGATGTTGCTGGTGGT	**NTDVAGG**	2180	1919	1944	0.880
CAGGCGGGGCTTCTGCGTCAT	**QAGLLRH**	2149	1876	1922	0.873
CGGGCTGATATGTCGACTGTG	**RADMSTV**	2098	1858	1878	0.886
TGGGGGGGGCTGCCTGAGCCT	**WGGLPEP**	2047	1591	1817	0.777
GGTCCTATGCTGGCTCGTGGG	**GPMLARG**	1847	94	1646	0.051

Table 2. Top 30 sequences censored during the ^{1}n → ^{13}n process. Bolded sequences can also be found in censorship during the ^{13}n → ^{30}n process (partially described in Table 1). Red sequences are uniquely censored in ^{1}n → ^{13}n process.

REFERENCES:

1. Scott, J.K. and Smith, G.P. (1990) Searching for Peptide Ligands with an Epitope Library. *Science*, **249**, 386-390.

2. Smith, G.P. and Petrenko, V.A. (1997) Phage display. *Chem. Rev.*, **97**, 391-410.

3. Ellington, A.D. and Szostak, J.W. (1990) Invitro Selection of Rna Molecules That Bind Specific Ligands. *Nature*, **346**, 818-822.

4. Tuerk, C. and Gold, L. (1990) Systematic Evolution of Ligands by Exponential Enrichment - Rna Ligands to Bacteriophage-T4 DNA-Polymerase. *Science*, **249**, 505-510.

5. Dias-Neto, E., Nunes, D.N., Giordano, R.J., Sun, J., Botz, G.H., Yang, K., Setubal, J.C., Pasqualini, R. and Arap, W. (2009) Next-Generation Phage Display: Integrating and Comparing Available Molecular Tools to Enable Cost-Effective High-Throughput Analysis. *Plos One*, **4**.

6. Matochko, W.L., Chu, K., Jin, B., Lee, S.W., Whitesides, G.M. and Derda, R. (2012) Deep sequencing analysis of phage libraries using Illumina platform. *Methods*, **58**, 47-55.

7. Ernst, A., Gfeller, D., Kan, Z., Seshagiri, S., Kim, P.M., Bader, G.D. and Sidhu, S.S. (2010) Coevolution of PDZ domain-ligand interactions analyzed by high-throughput phage display and deep sequencing. *Molecular Biosystems*, **6**, 1782-1790.

8. Kupakuwana, G.V., Crill, J.E., McPike, M.P. and Borer, P.N. (2011) Acyclic Identification of Aptamers for Human alpha-Thrombin Using Over-Represented Libraries and Deep Sequencing. *Plos One*, **6**.

9. t Hoen, P.A.C., Jirka, S.M.G., ten Broeke, B.R., Schultes, E.A., Aguilera, B., Pang, K.H., Heemskerk, H., Aartsma-Rus, A., van Ommen, G.J. and den Dunnen, J.T. (2012) Phage display screening without repetitious selection rounds. *Anal. Biochem.*, **421**, 622-631.

10. Zhang, H., Torkamani, A., Jones, T.M., Ruiz, D.I., Pons, J. and Lerner, R.A. (2011) Phenotype-information-phenotype cycle for deconvolution of combinatorial antibody libraries selected against complex systems. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, 13456-13461.

11. Derda, R., Tang, S.K.Y., Li, S.C., Ng, S., Matochko, W. and Jafari, M.R. (2011) Diversity of Phage-Displayed Libraries of Peptides during Panning and Amplification. *Molecules*, **16**, 1776-1803.

12. Hamming, R.W. (1950) Error Detecting and Error Correcting Codes. *Bell System Technical Journal*, **29**, 147-160.

13. Rodi, D.J., Soares, A.S. and Makowski, L. (2002) Quantitative assessment of peptide sequence diversity in M13 combinatorial peptide phage display libraries. *J. Mol. Biol.*, **322**, 1039-1052.

14. Makowski, L. (2011) Quantitative Analysis of Peptide Libraries. *Phage Nanobiotechnology*.

15. Shannon, C.E. (1948) A Mathematical Theory of Communication. *Bell System Technical Journal*, **27**, 379-423.

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

17. Kim, T., Tyndel, M.S., Huang, H., Sidhu, S.S., Bader, G.D., Gfeller, D. and Kim, P.M. (2012) MUSI: an integrated system for identifying multiple specificity from very large peptide or nucleic acid data sets. *Nucleic Acids Res.*, **40**, e47.

18. Weinstein, J.A., Jiang, N., White, R.A., Fisher, D.S. and Quake, S.R. (2009) High-Throughput Sequencing of the Zebrafish Antibody Repertoire. *Science*, **324**, 807-810.

19. DeKosky, B.J., Ippolito, G.C., Deschner, R.P., Lavinder, J.J., Wine, Y., Rawlings, B.M., Varadarajan, N., Giesecke, C., Doerner, T., Andrews, S.F. *et al.* (2013) High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. *Nat. Biotechnol.*, **31**, 166-169.

20. Levitan, B. (1998) Stochastic modeling and optimization of phage display. *J. Mol. Biol.*, **277**, 893-916.

21. Riesenfeld, C.S., Schloss, P.D. and Handelsman, J. (2004) Metagenomics: Genomic analysis of microbial communities. *Annu. Rev. Genet.*, **38**, 525-552.

22. Sogin, M.L., Morrison, H.G., Huber, J.A., Mark Welch, D., Huse, S.M., Neal, P.R., Arrieta, J.M. and Herndl, G.J. (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc. Natl. Acad. Sci. U.S.A.*, **103**, 12115-12120.

23. Backhed, F., Ley, R.E., Sonnenburg, J.L., Peterson, D.A. and Gordon, J.I. (2005) Host-bacterial mutualism in the human intestine. *Science*, **307**, 1915-1920.

24. Beerenwinkel, N. and Zagordi, O. (2011) Ultra-deep sequencing for the analysis of viral populations. *Current Opinion in Virology*, **1**, 413-418.

25. Huber, J.A., Mark Welch, D., Morrison, H.G., Huse, S.M., Neal, P.R., Butterfield, D.A. and Sogin, M.L. (2007) Microbial population structures in the deep marine biosphere. *Science*, **318**, 97-100.

26. Wilm, A., Aw, P.P.K., Bertrand, D., Yeo, G.H.T., Ong, S.H., Wong, C.H., Khor, C.C., Petric, R., Hibberd, M.L. and Nagarajan, N. (2012) LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Res.*, **40**, 11189-11201.

27. Turnbaugh, P.J., Quince, C., Faith, J.J., McHardy, A.C., Yatsunenko, T., Niazi, F., Affourtit, J., Egholm, M., Henrissat, B., Knight, R. *et al.* (2010) Organismal, genetic, and transcriptional variation in the deeply sequenced gut microbiomes of identical twins. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 7503-7508.

28. Quince, C., Lanzen, A., Curtis, T.P., Davenport, R.J., Hall, N., Head, I.M., Read, L.F. and Sloan, W.T. (2009) Accurate determination of microbial diversity from 454 pyrosequencing data. *Nat. Methods*, **6**, 639-U627.

29. Watson, S.J., Welkers, M.R.A., Depledge, D.P., Coulter, E., Breuer, J.M., de Jong, M.D. and Kellam, P. (2013) Viral population analysis and minority-variant detection using short read next-generation sequencing. *Philosophical Transactions of the Royal Society B-Biological Sciences*, **368**.

30. Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R. *et al.* (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, **456**, 53-59.

31. Nakamura, K., Oshima, T., Morimoto, T., Ikeda, S., Yoshikawa, H., Shiwa, Y., Ishikawa, S., Linak, M.C., Hirai, A., Takahashi, H. *et al.* (2011) Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res.*, **39**.

32. Quail, M.A., Kozarewa, I., Smith, F., Scally, A., Stephens, P.J., Durbin, R., Swerdlow, H. and Turner, D.J. (2008) A large genome center's improvements to the Illumina sequencing system. *Nat. Methods*, **5**, 1005-1010.

33. Zagordi, O., Klein, R., Daumer, M. and Beerenwinkel, N. (2010) Error correction of next-generation sequencing data and reliable estimation of HIV quasispecies. *Nucleic Acids Res.*, **38**, 7400-7409.

34. Schmitt, M.W., Kennedy, S.R., Salk, J.J., Fox, E.J., Hiatt, J.B. and Loeb, L.A. (2012) Detection of ultra-rare mutations by next-generation sequencing. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 14508-14513.

35. Syropoulos, A. (2001), *Proceedings of the Workshop on Multiset Processing: Multiset Processing, Mathematical, Computer Science, and Molecular Computing Points of View*. Springer-Verlag, pp. 347-358.

36. Arap, W., Kolonin, M.G., Trepel, M., Lahdenranta, J., Cardo-Vila, M., Giordano, R.J., Mintz, P.J., Ardelt, P.U., Yao, V.J., Vidal, C.I. *et al.* (2002) Steps toward mapping the human vasculature by phage display. *Nat. Med.*, **8**, 121-127.

37. Blondelguindi, S., Cwirla, S.E., Dower, W.J., Lipshutz, R.J., Sprang, S.R., Sambrook, J.F. and Gething, M.J.H. (1993) Affinity Panning of a Library of Peptides Displayed on Bacteriophages Reveals the Binding-Specificity of Bip. *Cell*, **75**, 717-728.

38. Cox, S., Rosten, E., Monypenny, J., Jovanovic-Talisman, T., Burnette, D.T., Lippincott-Schwartz, J., Jones, G.E. and Heintzmann, R. (2012) Bayesian localization microscopy reveals nanoscale podosome dynamics. *Nat. Methods*, **9**, 195-200.

39. Rosten, E., Jones, G.E. and Cox, S. (2013) ImageJ plug-in for Bayesian analysis of blinking and bleaching. *Nat. Methods*, **10**, 97-98.

40. Matochko, W.L., Ng, S., Jafari, M.R., Romaniuk, J., Tang, S.K.Y. and Derda, R. (2012) Uniform amplification of phage display libraries in monodisperse emulsions. *Methods*, **58**, 18-27.

41. Peters, E.A., Schatz, P.J., Johnson, S.S. and Dower, W.J. (1994) Membrane Insertion Defects Caused by Positive Charges in the Early Mature Region of Protein-Piii of Filamentous Phage-Fd Can Be Corrected Prla Suppressors. *J. Bacteriol.*, **176**, 4296-4305.

42. Brammer, L.A., Bolduc, B., Kass, J.L., Felice, K.M., Noren, C.J. and Hall, M.F. (2008) 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**, 88-98.

43. Kuzmicheva, G.A., Jayanna, P.K., Sorokulova, I.B. and Petrenko, V.A. (2009) Diversity and censoring of landscape phage libraries. *Protein Engineering Design & Selection*, **22**, 9-18.

44. Wilson, D.R. and Finlay, B.B. (1998) Phage display: applications, innovations, and issues in phage and host biology. *Can. J. Microbiol.*, **44**, 313-329.

45. Derda, R., Tang, S.K.Y. and Whitesides, G.M. (2010) Uniform Amplification of Phage with Different Growth Characteristics in Individual Compartments Consisting of Monodisperse Droplets. *Angew. Chem. Int. Ed.*, **49**, 5301-5304.

46. Dohm, J.C., Lottaz, C., Borodina, T. and Himmelbauer, H. (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.*, **36**.