Direct Sequencing of Phage Display Products Presents an Unbiased Tool for Analysis of Protein Interactions

A thesis

submitted by

Andrew J. Barry

In partial fulfillment of the requirements

for the degree of

Master of Science

In

Biomedical Engineering

TUFTS UNIVERSITY

November, 2012

Adviser: David Kaplan, PhD

Adviser: Nathaniel Cosper, PhD

Committee member: Qiaobing Xu, PhD

Cross-department committee member: Hyumin Yi, PhD

Abstract

DNA sequencing of phage display products presents a combination of two technologies capable of providing a high throughput means of screening protein-protein and protein-substrate interactions. Phage display presents a system for the controlled interrogation of molecular interactions at the expressed protein level, yet uses the encoding DNA as the analyte. Recent advances in DNA sequencing technologies have rendered this technology a tractable readout for a variety of applications, including phage display. The method described circumvents the need for any infection of phage into bacteria with the intent of removing amplification as a source of bias. In order to understand the utility of this technology for screening and discovery of amino acid motifs involved in extracellular matrix adhesion, highly diverse phage libraries were screened against silk, collagen, and polystyrene substrates. Products were PCR amplified and prepared for Illumina sequencing. Recovered sequences were translated *in silico* to peptide sequences and aligned to known extracellular matrix protein sequences to identify regions of substrate interaction.

Acknowledgements

I would like to thank my advisor, Professor David Kaplan for supporting and encouraging me throughout the project.

I also wish to thank Olena Rabotyagova for guidance on the direct PCR approach, and Carmen Preda and the Kaplan lab for supplying silk protein.

I would also like to acknowledge John Stalker for assistance in sequence data transfer, and James Meldrim for guidance on utilization of bioinformatics tools, and introduction to the AWK programming language.

Most of all I want to thank my wife, Laura Barry, who was Laura McKeon when the project started and two children later, is still supporting my efforts in this and all things.

Finally I would like to thank the Broad Institute Sequencing platform for enabling this work.

Table of Contents

List of Tables

List of Figures:

INTRODUCTION

The ability of new technologies to enhance scientists' understanding of the fundamental aspects of human biology hinges largely upon the development of novel applications for which to leverage these advances. Completion of large scale projects, such as the human genome project (Lander, 2001; Venter, 2001) , in addition to providing a framework for the organization and execution of these efforts, has also provided the necessary attention to the technologies that underpin and enable the acceleration of such advances.

The tools to understand the functional and regulatory components of human biology have been revolutionized through advances in the high throughput readout of information through DNA sequencing technologies. As DNA sequencing technology has advanced extremely rapidly, it has been further leveraged for applications studying RNA expression (Marionni, 2008) and expression-regulating proteins (Bernstein, 2005).

There exists today a fundamental gap in our broad-scale understanding of the physical interactions that exist among proteins (Lyne, 2002). This is largely the result of the lack of tools required to broadly screen these interactions that exist in complex environments such as the extracellular matrix (ECM).

PROTEIN SCREENING TECHNOLOGIES AND CURRENT LIMITATIONS

A system for the screening of protein interactions in complex environments would ideally enable complete control of substrate and target material while requiring little to no prior knowledge about either at the molecular level. Additionally, the ability to screen extremely complex libraries of molecules, and receive information about the specific interaction at the amino acid level would be valuable. Finally, the ability to utilize such a system in both *in vitro* as well as *in vivo*, would allow results to be validated and more efficiently translated to native environments.

Several tools exist that allow interrogation of peptide-protein interactions including immunochromatography and mass spectrometry, dual polarization interferometry, and surface plasmon resonance. These tools are limited in their utility mainly due to the small number of analytes that can be screened. For this purpose, there exists relatively few that can be used for high throughput screening. Common approaches for high throughput screening include yeast two hybrid screens, peptide microarray approaches, tandem affinity purification, and phage display.

Protein microarrays have proven a valuable tool for screening protein interactions (Uetz, 2000) The major limitation to this technology is the physical limit of the number of features that can be spotted on such arrays, as well as the requirement that the specific sequences of the arrayed peptides must be directed and known in advance (Walter, 2000)

In vivo techniques such as yeast-two hybrid screens and tandem affinity purification also suffer from similar pitfalls, both in the requirement of knowledge over the substrate and their limitation in the population and complexity of the screen. (Brukner, 2009)

Phage display has several intrinsic qualities that render it a tractable approach for elucidation of protein interaction at a large scale. The most obvious is the relatively large number of analytes that can be screened as typical phage display libraries can exceed one billion discrete peptides. Opposed to alternative methods, phage display-based screens can be designed with little to no knowledge of the recombinant phage library. In addition, the substrate can be almost anything from purified proteins, cells, organic molecules, or tissues. Phage display has demonstrated success for in vitro and in vivo studies, which has demonstrated utility for the discovery of therapeutic agents (Whitney, 2010).

	Phage Display	Yeast two-hybrid	Peptide Arrays	Tandem Affinity Purification
Screening Population	Combinatorial peptides, cDNA fragments	Combinatorial peptides, cDNA fragments	Peptides, cDNA fragments	Proteins
Targets	Proteins, cell surfaces, tissues, small organic molecules	Protein	Synthetic Peptides	Proteins
Screening format	*In vivo, in vitro*	*In vivo*	*In vitro*	*In vivo*
Screening population diversity (of clones)	Billions	Millions	Tens of thousands	hundreds
Control over substrate?	Y	Y	Y	N
Control over ligand	Y	Y	Y	N
Previous knowledge about substrate required?	N	Y	Y	N
Previous knowledge about Ligand Required	N	N	Y	N
Output readout	DNA sequencing, ELISA	Binary clone analysis	Fluorescence, radiolabeling	SDS PAGE / Mass Spectroscopy

Table 1. Comparison of commonly used techniques for screening protein-protein interactions

PHAGE DISPLAY AND CURRENT LIMITATIONS

Phage display is a technology widely used for a variety of applications focusing on molecular interactions. First developed in 1985 by George Smith, phage display provides an *in vitro* system for expression of peptides as a surface coat protein in filamentous bacteriophage. The specific amino acid sequences are programmed by encoding libraries of DNA sequences. This is accomplished through insertion of foreign DNA fragments into a filamentous phage vector containing filamentous phage gene III to create a fusion protein, which, when the phage virions infect bacterial host cells, displays the phage coat protein on the surface of the infected cell.

The unique ability to create a direct link between genotypic and phenotypic information has rendered phage display a practical tool for a variety of applications. The most common application for phage display technology is the use for in vitro affinity selection for the discovery of receptor ligands. Traditional applications include the epitope mapping of monoclonal antibodies (Fack,1997), elucidation of peptide structure recognized by major histocompatibility (MHC) molecules (Hansen, 2001) and improve understanding of cell surface integrin proteins. (Koivunen, 2003; Deshayes, 2002).

The utility of phage display for these functions is not however limited to *in vitro* applications, as *in vivo* receptor ligand studies have proven successful in identification of tissue specific markers (Arap, 1998), as well as identification of receptor-ligand pairs specific to disease.

Phage display technology has specific application in the field of tissue engineering. Identification of specific protein ligands and binding motifs allows for engineering of cell, tissue, or tissue analog specific adhesion, the directional delivery of growth factors, and the adhesion of cells and tissues to synthetic scaffolds (Sreejalekshmi, 2010). Additionally, motifs demonstrating material specific affinity are used in the production of synthetic peptides for use in therapeutic applications (Nettles, 2010; Nomura, 2011). These peptides are synthesized to contain regions compromising different motifs and allow control over the specificity of therapeutic agents for biological substrates (Tan, 2006).

More recently, phage display technology has been applied for more direct therapeutic applications. Screening studies have proven success in the identification of peptide mimotopes, which have demonstrated utility in development of novel vaccines. Peptides obtained from phage display screening have been used as gene delivery vectors (Barry, 1996), and directed drug delivery agents (Maxwell, 2004). Finally, phage display is a commonly used tool for the directed evolution of enzymes (Fernandez-Gacio, 2003) and other biologically active proteins and antibodies (Rader, 1997) for research purposes. Phage display derived peptides, and peptides in general, are attractive as direct therapeutic agents due to their low production cost, low immunogenicity, high efficacy, and ability to penetrate tissues (Vlieghe, 2010). These applications are particular relevant in the field of tissue engineering, where biomimetic materials have great utility in mediating cell specific interactions (Sreejalekshmi, 2010). or specifically obtained DNA fragments are used as the source genetic material.

The term biopanning is used to describe the process of exposing surface peptides expressed by the phage library to a substrate. The nature of the substrate dictates the application of this technology, for which several exist. Phage display is typically employed to understand protein-protein, protein-peptide, and protein-DNA interactions. The general protocol involves the immobilization of target material to a surface, incubation of the phage display library with the target material, a series of washes to remove unbound phage molecules, and elution of the bound phage molecules for analysis.

Despite the widespread use of phage display technology, there are intrinsic qualities have created serious bottlenecks, limiting the utility of the methodology. The first is the need for extensive amplification of phage particles demonstrating target affinity. The need to infect bacterial cells and grow large quantities of eluted products creates amplification bias, which can lead to a misinterpretation of true specificities (Dias-Neto, 2009). The bias incurred is likely due to selectivity towards surviving phage in intracellular environments suffering from compromised infectivity, and can result in the loss of valuable peptide sequences Larocca, 2004; Molenaar, 2004; Molenaar, 2005). Moreover, factors such as phage growth rates can further skew the representation of derived sequences, resulting in high false positive calls (Nolan, 2007).

Large scale efforts to recognize the full potential of phage display have demonstrated the scalability of the upfront screening process. Groups have undertaken major efforts to automate and scale the biopanning process, while still

relying on individual clone selection for characterization by sequence analysis (Konthur, 2002; Shofield, 2007) This is problematic both in terms of the cost associated, as well as the limited number of clones that can be selected for sequencing (Rahim, 2006).

SEQUENCING OF PHAGE DISPLAY PRODUCTS

Recent advances in DNA sequencing technologies have resulted in 10,000 fold reduction in the cost per basepair of sequence data (Mardis, 2007). These same advances have obviated the need for cloning into bacterial cells as a means of separating molecules discreetly, instead relying on techniques for polyclonal amplification at the molecular level.

Due to the limitation of Sanger sequencing as a process bottleneck for phage display screening, researchers have applied the latest in sequencing technology to this application. Pyrosequencing was applied to phage display as early as 2008 (Rahim, 2008), and interest has continued as these technologies evolve. A comprehensive study was published in 2009 to explore the viability of NGS with phage display, specifically with regard to liminating the plating of recovered phage-infected bacteria as a rate limiting step for final quantification before sequencing. These results concluded that a qPCR-based approach eliminated the need for plating and overnight growth. This publication clearly obviated the need for a bacteria-free system, and implicitly called for an alternative strategy for enrichment that eliminated the need for excessive rounds of amplification through infection into host bacteria. (Dias-Neto, 2009). A more recently published study

used a synthetic oligonucleotide library to demonstrate that the application of this technology is not limited to screening, but that targeted libraries can be used, leveraging recent advances in the production of large pools of directed specific oligonucleotides (Larmin, 2011) The application of vast amounts of sequence data has also been studied as a means to eliminate the need for excessive amounts of amplification, which has additional implications for the necessity to infect into bacterial cells. This 2011 study demonstrated that specificity for targets can be achieved without excessive rounds of amplification, and that through the use of next generation sequencing, we can derive meaningful results due to the depth and analytical filtering of the data. This study, however, used the single round of bacterial infection as the baseline for amplification, and to date there is no published account of a system that does not require any bacterial infection (Hoen, 2011).

The need for amplification is a direct function of the specificity for the interactions that are being studied. For applications such as tissue engineering, where there is a need to elucidate complex protein interactions that take place in intricate networks of extracellular matrix proteins, a broad-scale approach would be advantageous as it would survey of the functional regions of multiple components of the system.

SIGNIFICANCE

In order to understand the feasibility of this approach towards a practical application, experiments were conducted aimed at the discovery of peptide

ligands with tissue engineering utility, and developing a method that is capable of performing high throughput analysis of protein-protein interactions in a complex environment such as the extracellular matrix.

BACKROUND INFORMATION

OVERVIEW OF PHAGE DISPLAY

Since first introduced in 1985 by George Smith, phage display technology has become a general purpose tool for a variety of applications involving the screening of proteins and peptides (Willats, 2002)). Smith was able to illustrate that the fusion protein was capable of being expressed, and was accessible and capable of being enriched for through immunological affinity purification. (Smith, 1985).

The method presents a system for the display of peptides on the surface of a filamentous bacteriophage, which are capable of infecting bacterial cells without causing fatal effects. This is achieved through the fusion of segments of DNA to a gene encoding a pIII or pVIII coat protein, allowing the translated peptide to be expressed as a N- or C- terminal fusion on the surface of the M13 bacteriophage. As a result, a direct link between the genotypic information encoded in the library and the phenotypic result of interactions of individual library components can be derived (Paschke, 2005).

Phage Biology

Filamentous bacteriophages used for phage display applications belong to the genus *Inovirus*, class Ff, and include strains M13, f1, fd, and ft (3). They are known for their ability to infect and reproduce in gram-negative bacteria, and are characterized by a circular, single-stranded DNA genome, 6400 nucleotides in length that is encapsulated in long cylindrical capsid. The origin of the Ff name is derived from the bacterial receptor pilus that is specific for F plasmid containing E.coli. The Ff class includes the f1, fd, and M13 species which contain homologous genome sequences. The phage itself is roughly 6.5 nm in diameter and 930 nm in length. The genome encodes only 10-11 well characterized genes, two of which encode the 50 amino acid coat protein pVIII, as well as the minor coat protein pIII, which are used for the gene fusion to display peptides on the surface of the phage (Silverman, 2001; Marvin, 1998). The major coat protein is present at about 2700 copies per phage, while 5 copies of the minor coat protein are expressed at the proximal end of the phage (Kay, 2005). (Figure 1)

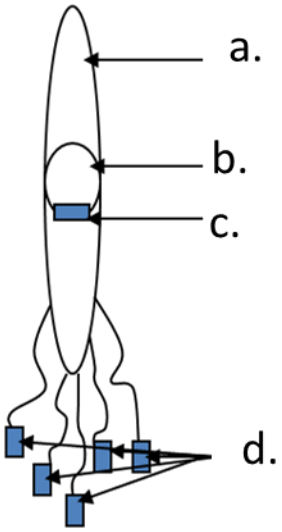

Figure 1. Overview of Phagemid Structure
a. Phagemid coat. b. M13 vector. c. random peptide-encoding 12bp insert.
d. surface expressed PIII minor coat protein

Infection occurs when one of the five copies of the pIII minor coat protein attaches to the F pilus receptor of the bacteria. The infection does render the phage particles intact, allowing replication of the phage through a process where the host secretes phage particles into growth media. This is accomplished first through conversion of the phage genome to a double stranded plasmid *via* the host cell's own replication process. Single stranded copies of this are created through rolling circle amplification, which in turn are used to express the coat proteins for the newly formed phage (Arap, 2005)

Biopanning Process

Selection for peptides demonstrating desired phenotypic effects is typically achieved through a process known as "biopanning," where the phage library in

question is exposed to a target molecule or molecules, and allowed to interact under controlled conditions. Phage library-substrate complexes are typically subjected to varying amounts of washing, where unbound molecules are physically removed, and molecules that remain bound can be selectively eluted from the targeted substrate (Figure 2). Washing is typically performed using a detergent such as Tween 20, which reduces nonspecific hydrophobic interactions between the phage and the target of interest. Typically the concentration of the detergent is increased in order to increase the stringency of the washing. In addition, salt concentrations are important for controlling nonspecific hydrophobic interactions. To control this, and avoid high levels of background binding, a high salt buffer such as TBS (Tris-buffered saline) is used in conjunction with the detergent (Silverman, 2001).

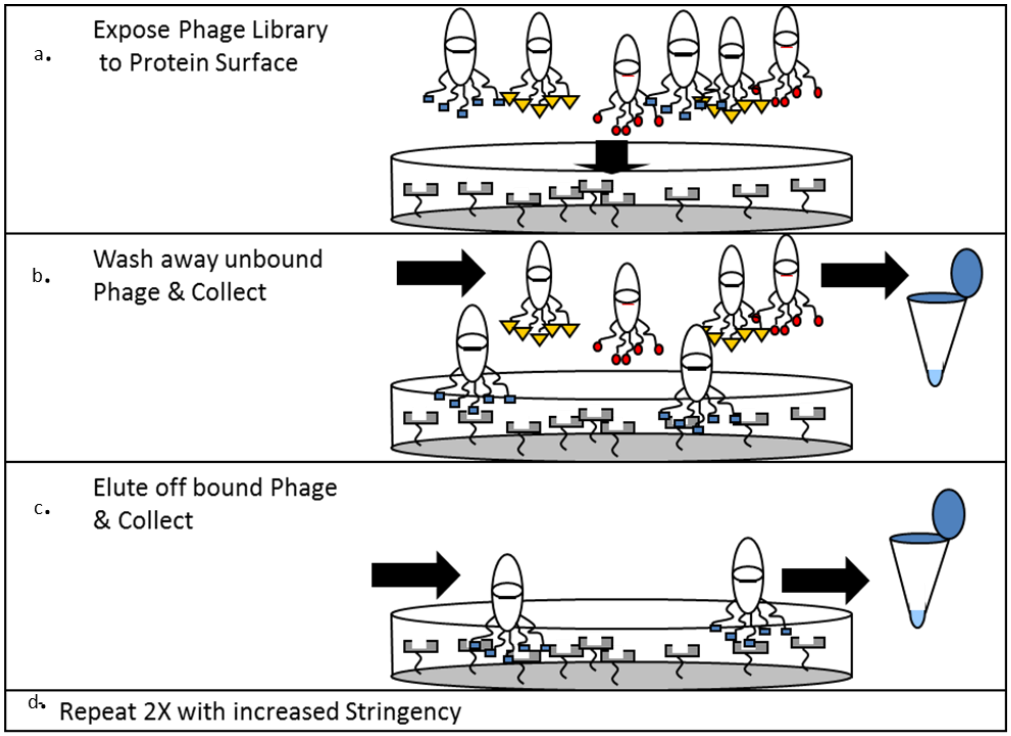

Figure 2. Biopanning Process

a. Exposure of phage library to protein coated surface. b. Washing to remove unbound phagemid. c. Elution of bound phagemid. d. repeat process with increased stringency

Typically, selectivity is further controlled through amplification between rounds of washing, where eluted phage is amplified *via* infection into gram negative bacteria. Amplified phage is then used as the input to the subsequent round of panning, resulting in increased specificity for targets.

Analysis of Recovered Products

Characterization of selected library molecules is achieved by sequencing the bacteriophage vector insert. Using traditional sequencing methods, this is achieved by infecting a bacterial cell, growing colonies of selected molecules, then purifying nucleic acids and applying dye terminator sequencing using capillary electrophoresis.

The output amino acid sequences derived by phage display sequencing are typically used to derive what is known as a consensus sequence. A consensus sequence is defined as the most common amino acid residue at a given position following alignment of multiple sequences. There are several tools available for the position-wise alignment of sequences. The basis for several of these tools is a position weight matrix (PWM) which is defined as a matrix of score values providing a weighted match to any given substring of fixed length (Tatusov, 1994; Ben-Gal, 2005; Beckstette, 2005). There are several tools that have been developed to enable generation of consensus sequences from groups of phage displayed derived peptide sequences (Moreau, 2006). These tools were used in the analysis of recovered peptide sequences.

Alignment of recovered peptides to sequences of proteins known to interact with substrates is a useful tool for identification of substrate binding sites as well as peptide binding motifs (Shlomi, 2007). These motifs are typically 3-4 amino acid residues in length and are difficult to predict using *in silico* models based on 3D composition of complex proteins. Alignment of the peptide sequences recovered through phage display experiments using a highly diverse, randomized library is attractive in that the motif will likely be present in different regions of separate random peptide inserts. This enables analysis of the # of unique peptides for which an alignment is used to add statistical significance to the likelihood that the alignment is representative of regions of reference protein molecules to interact with the substrate.

. Statistical tools have been developed specifically for the purpose of taking into account the multiple factors that can influence the reliability and significance of the alignment and whether what is observed can be associated with biological significance. These factors include the lengths of both the subject and query sequences, and the number of gaps that exist in the alignment. Additional factors specific to the biological context include the charge and polarity of the residues, as well as the hydrophobicity of the segment. (Borodovsky, 2005).

The goal of any statistical test is to compare the observed value to one that is derived by chance, or a random set of occurrences. In the case of sequence alignments, this is measured through the existence of what is known as a high scoring pair (HSP), involving the two sequences being aligned. In order to efficiently measure and communicate the significance, two tools are used, namely the e-value and the Bit Score. These are interrelated measurements that can be thought of as the measurement (e-value) and the unit of measurement (Bit Score). The e-value is derived from the length of the subject sequence (m), the query sequence (n), a measure of natural scale for the search space size (K) and a natural scale for the scoring system (λ). The e-value is defined as the number of high-scoring pairs with a minimum score (S) by the formula:

$$E = Kmn\, e^{-\lambda S}$$

The other required piece of information, the Bit Score, is based on a normalization of the raw e-value through the function:

$$S' = \lambda S - \ln K / \ln 2$$

Application of both the e-value and the Bit-score allow the interpretation of the significance of the HSP based solely on the search space.

Conversion of the e-value to P-value allows alignments to be compared for significance and communicated to broader scientific relevance. The conversion is a simple function defined as:

$$P = 1\text{-}e^{-E}$$

One of the specific challenges that expected in dealing with an amplification-free system is the high degree of background noise that will come as the result of a direct measurement. Similar challenges have been dealt with through analysis of alignments by looking at the fraction of total alignments that are unique relative to the total number of alignments for a given pair. These values can be further augmented to understand the statistical significance through a calculated z-score, which is a measure of the distance of the standard deviation of the unique alignments relative to the mean of a given reference data set (Chlu, 2008).

NEXT GENERATION DNA SEQUENCING TECHNOLOGIES

Overview

Dideoxy-nucleotide dye terminator sequencing technology was invented over 30 years ago in 1977 by Frederick Sanger (Sanger, 1977) the technique involves the creation of many copies of the template, each molecule terminating in a fluorescently tagged dideoxy nucleotide, with one molecule corresponding to each nucleotide position of the template. The automated manifestation of this was through capillary electrophoresis, (Smith, 1986) and it made possible large scale sequencing projects including the sequencing of the human genome (Lander, 2001; Venter, 2001)

Capillary electrophoresis has been largely replaced in the past five years by newer technologies (Chan, 2005). These technologies, including Illumina's Genome Analyzer, Roche's 454 sequencer, and Life Technologies' SOLiD platform, are largely characterized by either sequencing by synthesis (SBS) or sequencing by ligation. The major difference between the two sequence detection schemes lies in method employed to synthesize and detect nucleotides complementary to a single stranded, clonally amplified template sequence. Sequencing by synthesis relies on a DNA polymerase to incorporate individually fluorescently labeled nucleotide residues, while sequencing by ligation relies on DNA ligase to incorporate labeled nucleotides consisting of specific and degenerate nucleotides.

Critical to these achievements is the ability to process multiplexed molecules as a pool throughout the process. In the final steps preceding detection, the pools are divided into single molecules (Mardis, 2007). This concept has been dubbed "massively parallel" and is common to all of the so called "next-generation" sequencing technologies (Schuster, 2008). Starting material are DNA or RNA molecules that are either present at, or sheared to, lengths in the hundreds of basepairs. To the proximal ends of these molecules are ligated universal oligonucleotide duplexes, which allow successive molecular manipulations, including multiplexed PCR amplification of library fragments, and binding of library molecules to a solid substrate (Brenner, 2000)

Polyclonal Amplification

Isolation of individual molecules for sequence analysis is achieved in two distinct ways. The first method, used in 454 and SOLiD platforms, utilizes aqueous-in-oil emulsion PCR amplification of Poisson diluted molecules on the surface of micron-sized beads. Successfully amplified molecules are subsequently enriched by cross-linking these to low density polystyrene beads to which oligonucleotides complementary to the universal oligonucleotide sequences are bound. These complexes are then separated through centrifugation, resulting in the capture of template-bearing beads (Marguiles, 2005) (Figure 3).

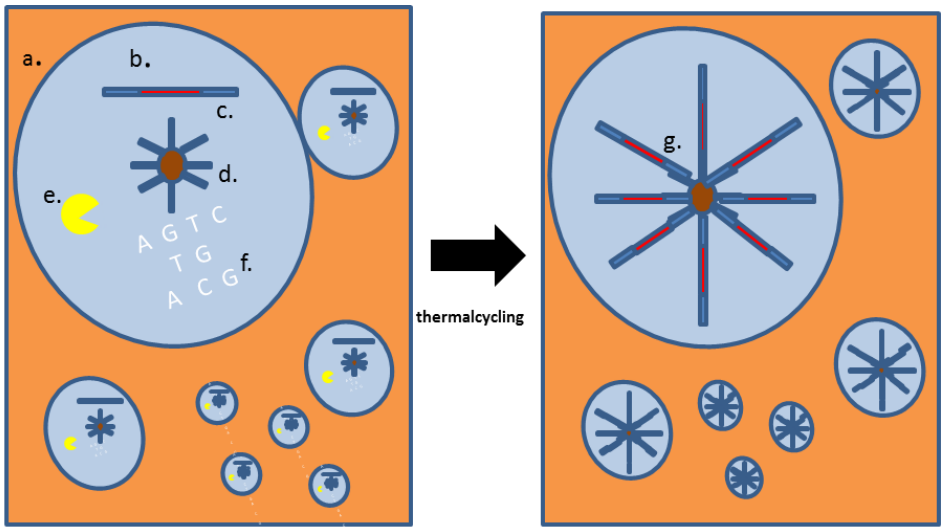

Figure 3. Clonal Amplification using emulsion PCR

a. Oil phase b. Aqueous phase containing PCR buffer c. Template fragment at poisson dilution
d. Oligonucleotide coated straptavidin bead e. Taq DNA Polymerase f. dNTP's
g. Polyclonally amplified templates on bead

The second method, employed by Illumina sequencing, employs a solid phase bridge DNA amplification to achieve molecular segregation. This method consists of surface bound primers to which template DNA flanked by universal sequences is hybridized at low concentrations (Adessi, 2000). In order for a molecule to amplify, it must bend so that the unbound end can find the complementary primer, which is also bound to the solid phase (Mercier, 2003). Steric interactions force molecules to bend outward, forming colony-like

"clusters" of homogenous molecules (Mercier, 2005).

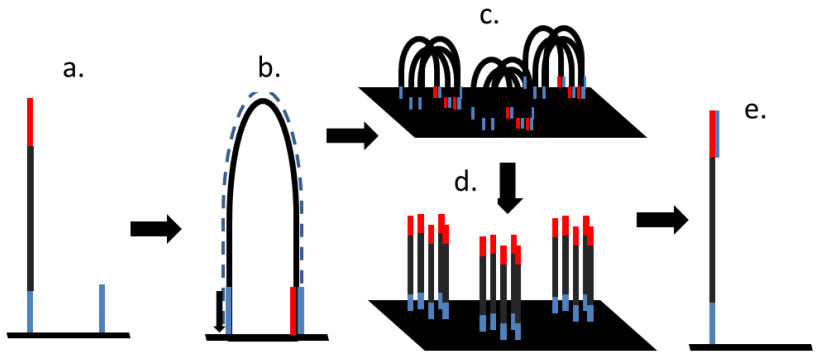

Figure 4. Polyclonal amplification *via* bridge PCR
a.) Hybridization of library fragments to solid substrate pre-seeded with complements to universal adapters. b.) primer annealingand extension of complementary strand binding to second universal pre-seeded oligo- formation of "bridge. C.) Linearization and repeat for several cycles. d.) linearization of amplified fragments e.) hybridization of sequencing primer.

Sequencing By Synthesis

Two of these sequencing methodologies, Illumina (San Diego, CA) and 454 Life Sciences (Branford, CT) employ what is known as "sequencing by synthesis", where nucleic acid bases are labeled, typically with a fluorescent molecule, and interrogated as they are incorporated during complementary strand synthesis. This is achieved by proprietary tagged molecules known as "reversible terminators." These molecules are designed so that chain termination takes place upon their incorporation, allowing analysis of the incorporated molecule, yet they can subsequently be modified, or "unblocked", allowing incorporation of the next nucleotide complementing the template sequence (Bentley, 2008).

Illumina sequencing employs a fluorescent based detection system, allowing all four labeled nucleotides (adenine, guanine, thymine and cytosine) to

be incorporated simultaneously, and differentiation of the fluorophores are determined based on the specific excitation laser and emission filter combinations. 454 sequencing utilizes the luciferase reaction, where incorporation of a nucleotide releases a molecule of ATP that is monitored using firefly luciferase (Ronaghi, 2005). This method requires that the nucleotides be incorporated one at a time, as there is no signal differentiation between the incorporated bases.

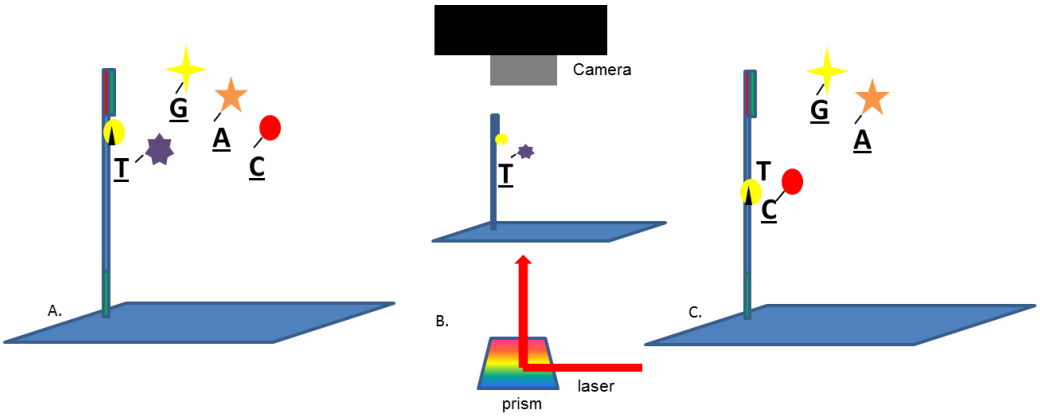

Figure 5. Sequencing by synthesis cycle chemistry
A.)DNA polymerase incorporation of terminating fluorophore labeled nucleotides. B.) Laser excitation and imaging of incorporated nucleotides. C.)Chemical cleavage of fluorophore and reversal of termination to enable subsequent incorporation event

Ligation Based Sequencing

Another widely commercialized method, used by the SoLID technology by Life Technologies (Carlsbad, CA) utilizes a "sequencing by ligation" methodology, where isolated fragments are introduced to fluorescently tagged oligonucleotides containing a mixture of known and degenerate bases, which are ligated to the fragments that have been isolated on beads. Analysis of incorporated oligonucleotides, combined with knowledge of the sequence and

order of attempted incorporation, allow elucidation of the template sequence (Shendure, 2005).

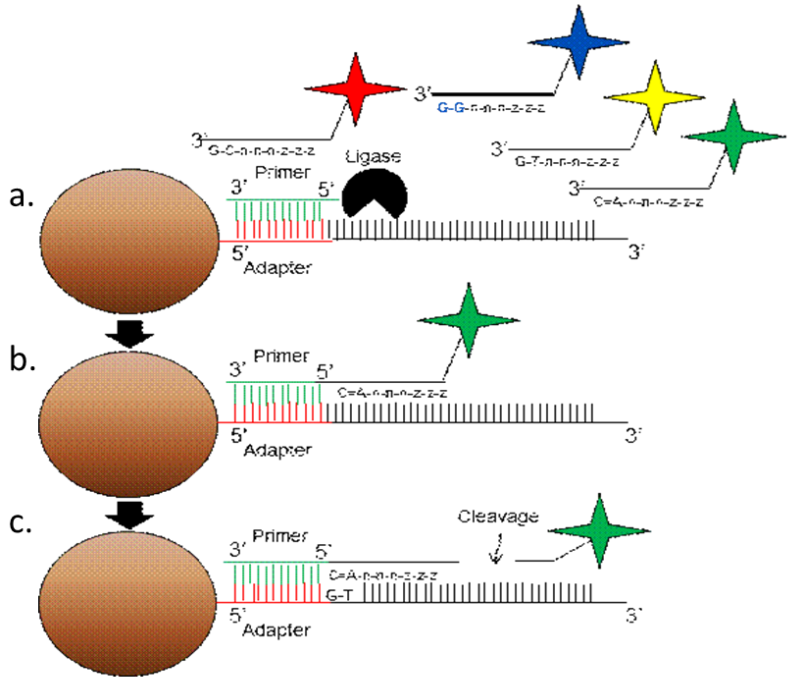

Figure 7. Sequencing by ligation chemistry cycle

a. Exposure of bead bound template to fluorescently tagged probes with dinucleotide combinations followed by degenerate basepairs. b. Hybridization of complementary dinucleotide probe followed by excitation and imaging c. Cleavage of fluorophore followed by successive exposure to probes.

Sample Preparation For Next Generation Sequencing

The process for preparing DNA molecules for solid phase amplification is very similar across sequencing technologies. Molecules are either present at or sheared to a length of somewhere between 150-600bp. The length required is a function of the desired or technological limits of the sequencing run, as well as the application for which the data being generated will be used. Several technologies can be used for shearing of DNA molecules including sonication,

nebulization, and adaptive focused acoustics. These molecules then undergo parallelized processing, where ends are blunted using a combination of T4 DNA polymerase and T4 polynucleotide kinase. Molecules are phosphorylated and a single adenosine residue is added to the terminal ends using exonuclease resistant klenow. The adenylated ends then undergo sticky-end ligation of universal synthetic double stranded fragment of DNA, also referred to as adapters, to the proximal 5' and 3' ends. These adapters have a "Y" shape that includes complementary base pairing at the portion of the molecule interacting with the library fragments, and a non-complementary portion distal to the library fragments. The utility of the "Y" shaped adapter is to enable directional enrichment through PCR, ensuring that amplified fragments are palindromic. This full length sequence is later used for further manipulation of the sample fragments, including PCR amplification, solid phase amplification, and ultimately hybridization of a sequencing primer in the case of sequencing by synthesis, or a complementary sequence to systematically ligate for sequencing by ligation. A general schematic of the sample preparation process is shown in figure 7.

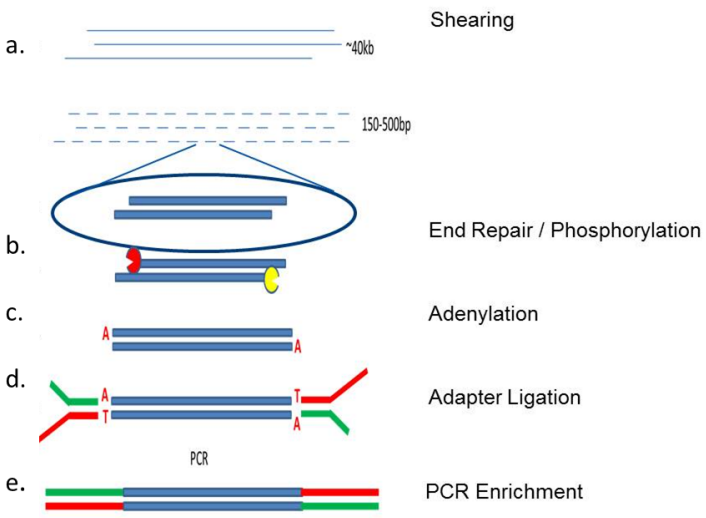

Figure 7. General Construction of Libraries for Massively Parallel Sequencing
a.) Fragmentation of genomic DNA. b.) Blunting of ends through T4 Polynucleotide Kinase and T4 DNA Polymerase followed by phosphorylation. c.) Addition of single dATP residue Klenow exo- enzyme. d.) Ligation of forked universal adapters e.) PCR amplification to create full-length blunted molecules

Multiplexing of Samples for NGS

One of the major advantages and opportunites for the application of next generation sequencing as a readout for phage display screening is the sheer amount of data that is generated. A typical Illumina sequencing run can generate in excess of 100 billion nucleotides of data (Bedard, 2009). In order to recognize the economy of the data generated, strategies have been developed to allow the pooling of samples prior to sequencing. This is particularly useful for the phage display application, as the amount of data produced for a single experiment will require far less than the amount of data generated in a single Illumina experiment. In order to allow pooling of samples or replicates from a given experiment, specific short DNA sequences are incorporated into the universal oligonucleotide portion of the library molecules. These samples can be pooled together prior to

sequencing. Once sequenced, the sequence information can be used to disambiguate the reads associated with a given sample from the pool of sequencing reads (Smith, 2010).

PHAGE DISPLAY SCREENING AND TISSUE ENGINEERING

Collagen I and Tissue Engineering

Collagen I is the most abundant protein present in mammals (Ramshaw, 1996). It is the primary component responsible for structural and morphological integrity of connective tissues. The protein is present in a triple-helical structure, consisting of three individual polypeptide left-handed alpha helices, intertwined into a coiled coil held together by hydrogen bonding. Due to its profusion and varied utility as a connective tissue, Collagen-I has been extensively evaluated for tissue engineering applications (Riesle, 1998) and is used broadly as a biomaterial. Features such as biodegradability, and biocompatibility render it a useful material for scaffolding (Glowacki, 2007).

Silk Fibroin as a Biomaterial

Silkworm (*bombx mori*) silk fibroin protein is of particular interest in the field of tissue engineering for its biocompatibility, tunable rate of degradation, and physical strength. Silk fibroin has been used in a variety of approaches as a scaffold for cell seeding, as it mimics endogenous extracellular matrix

environments by providing physical adhesion structures for cellular recruitment, growth, and proliferation until time points where extracellular matrix environments are autonomously produced (Minoura, 1990). Significant efforts have been dedicated to developing methods for generating scaffolds with optimal morphology, porosity, stability, and degradation profiles (Kim, 2005).

OBJECTIVE

The objective of this project is to test whether next generation sequencing can provide the necessary depth of coverage to enable target specificity in a cell and amplification-free system, and use this as a utility to elucidate protein interactions in the cellular matrix. In order to test this hypothesis, a randomized, commercially available library was panned against three substrates that have utility in tissue engineering applications, namely collagen I, silk fibroin, and tissue culture plastic (polystyrene), using varying washing stringencies. The recovered phage particles were directly lysed and DNA was purified and the peptide-encoding insert was PCR amplified and sequenced using next generation sequencing. The resulting reads were translated from nucleic acid to protein, and aligned to various extracellular matrix (ECM) proteins to look for aligned regions of the recovered peptides that are homologous to ECM proteins.

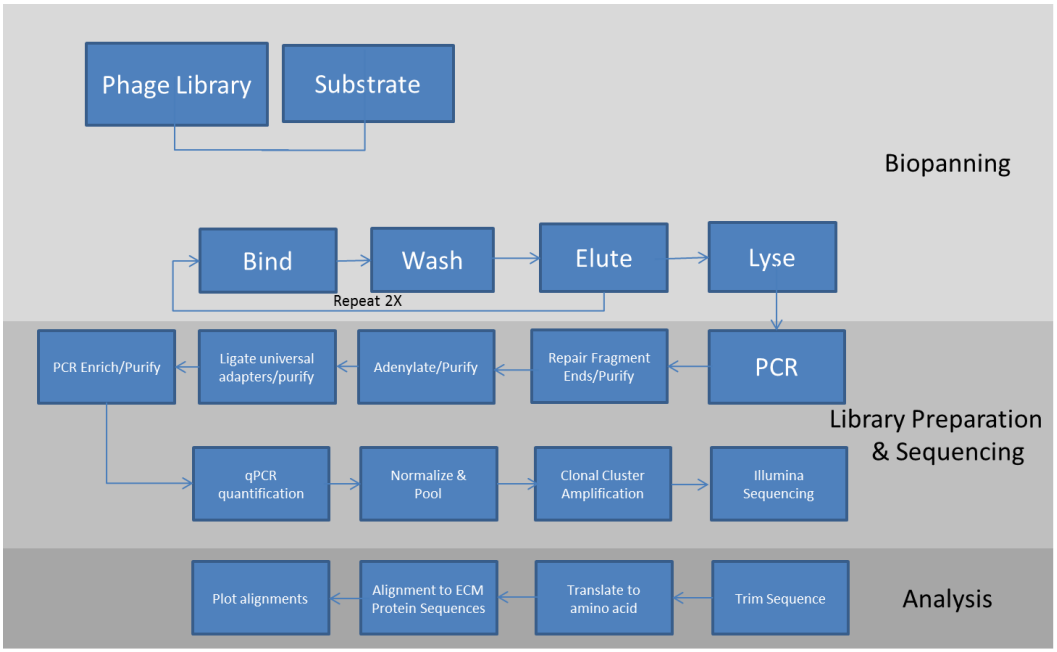

Figure 8. Process workflow for Next Generation Sequencing of Phage Display Products

Methods

OVERVIEW

In order to understand the applicability of next generation sequencing to phage display screening, a randomized phage library (New England Biolabs, Ipswich, MA, USA phD12 part # E8111L) was biopanned against three substrates, rat tail collagen (Roche Diagnostics, Manheim, Germany, part # 11179179001), recombinant silk fibroin protein (Kaplan lab, Medford, MA, USA) and tissue culture plastic. Biopanning was carried out in round bottom tissue culture plastic 96-well microtiter plates (BD Falcon, Franklin Lakes, New Jersey, USA, cat # 353227). Two trials were conducted, the first with a less stringent and the second with a more aggressive wash protocol. All solutions used in the experiment were made fresh and filtered with 0.2 uM filters (VWR, Radnor, PA, USA part # 28143-315)

PREPARATION OF PROTEIN SUBSTRATES

Sources of Protein Substrates

Rat tail collagen was supplied as a lyophilized protein. To reconstitute, 4.95ml of 0.2% acetic acid was added to the container and allowed to incubate overnight at room temperature with no agitation.

Silk fibroin protein solution was derived from silkworm (B. mori) cocoons. Extraction of silk fibroin from cocoons was accomplished by cutting

cocoons into 5g pieces and boiling for 30 minutes in 4.24g Na2CO3 in 2L of Deionized H2O. The supernatant was removed, and the protein precipitate was transferred to a plastic beaker containing 1.5L of deionized H20. The protein was washed 3X total, each time refreshing with fresh diH2O followed by 20 minute incubation at room temperature. Pursuant to the final rinse, the supernatant was removed, and the silk fibroin was allowed to dry in a fume hood for 12 hours. Following the 12 hour drying, a 20% w/v solution was created by adding 1g Silk to 4mL 9.3M LiBr, and allowing the solution to dissolve in an oven at 60ºC for 4 hours. The resulting silk was present as a 7.9 w/v solution (Preda, 2009).

Dilution of Protein Substrates

To dilute the collagen to the working concentration, 25ul of the reconstituted solution was added to 25ml of 0.1M NaHCO3 (pH 8.6) to reach a final concentration of 1 mg/ml. To dilute the silk protein, 126.58ul of supplied protein was added to 9.87 ml of 0.1M NaHCO3. 5ml of this solution was then added to 45ml of 0.1M NaHCO3 to obtain a final concentration of 1 mg/ml.

Substrate Coating of Polystyrene Plates

To coat the target substrate onto solid support, 150ul of each protein was added to wells of separate tissue culture plastic plates, utilizing 1 plate per protein substrate and 1 plate per wash stringency for a total of 6 plates. Solution was swirled to cover the entire well surface, and allowed to incubate for 16 hours at 4 degrees in a plastic box lined with damp paper towels while being agitated on a plate shaker set at 25 rpm. Plates were retrieved from 4 degree cold room and protein coating

solution was removed by slapping the plates face-down on a dry paper towel. Each well was filled completely with blocking buffer (0.1M NaHCO3 pH8.6, 5 mg/ml BSA, 0.02% NaN3) and allowed to incubate at 4 degrees for 1 hour. Blocking buffer was removed by slapping face down on a paper towel. Plate wells were washed 6X by adding 200ul TBST((TBS + 0.1 % v/v Tween 20). Wells were coated completely swirled quickly, then emptied by slapping face down on a dry paper towel.

BIOPANNING PROCESS

Phage Binding

Dilution of randomized phage library was accomplished by diluting pHD-12 phage library to a working concentration of 4e10 phage by adding 300ul of the original library to 2.9 ml TBST. 100ul of diluted phage library was added to each of the plates containing the substrate of interest (plate 1 = Collagen, plate 2 = Silk, plate 3 = tissue culture plastic) and rocked gently on a plate agitator for 60 minutes at room temperature. Nonbinding phage was discarded by pouring off and slapping face down on a clean, dry, paper towel.

Biopanning and Elution of Bound Phage

Plates were washed according to wash stringency by adding the appropriate wash solution per condition (Table 2)

Condition	Substrate	Wash Buffer	# of Wash Cycles
C1	Collagen	TBS + 0.1%[v/v] Tween20	10
C2	Collagen	TBS + 0.5%[v/v] Tween20	20
C3	Collagen	TBS + 0.5%[v/v] Tween20	30
S1	Silk	TBS + 0.1%[v/v] Tween20	10
S2	Silk	TBS + 0.5%[v/v] Tween20	20
S3	Silk	TBS + 0.5%[v/v] Tween20	30
P1	Tissue Culture Plastic	TBS + 0.1%[v/v] Tween20	10
P2	Tissue Culture Plastic	TBS + 0.5%[v/v] Tween20	20
P3	Tissue Culture Plastic	TBS + 0.5%[v/v] Tween20	30

Table 2. Experimental conditions for Trial 1.
Variable wash buffer and number of wash cycles to increase stringency

Condition	Substrate	Wash Buffer	# of Wash Cycles
C1	Collagen	TBS + 0.5%[v/v] Tween20	10
C2	Collagen	TBS + 1.0%[v/v] Tween20	20
C3	Collagen	TBS + 2.0%[v/v] Tween20	30
S1	Silk	TBS + 0.5%[v/v] Tween20	10
S2	Silk	TBS + 1.0%[v/v] Tween20	20
S3	Silk	TBS + 2.0 [v/v] Tween20	30
P1	Tissue Culture Plastic	TBS + 0.5%[v/v] Tween20	10
P2	Tissue Culture Plastic	TBS + 1.0%[v/v] Tween20	20
P3	Tissue Culture Plastic	TBS + 2.0%[v/v] Tween20	30

Table 3. Experimental conditions for Trial 2.
Variable wash buffer and number of wash cycles to increase stringency

Bound phage was eluted from the substrate by adding 100ul 0.2M Glycine HCl pH2.2, and gently rocked on a plate agitator for 60 minutes at room temperature. Eluted phage was transferred into a microcentrifuge tube and neutralized by adding 15ul of 1M Tris pH 9.1.

Phage Lysis and Purification

Phage particles were lysed by adding 150ul of Lysis buffer A (10mM EDTA, 10mM Tric-HCl pH 8.3, 1% Triton X-100) to each microcentrifuge tube containing eluted phage, and allowed to incubate for 10 minutes at 95 C.

To a Microcon YM-100 microcentrifuge filter (Millipore, Billerica, MA, USA Cat # 14422AM) 250ul of lysed phage was added and centrifuged at 14,000 x g for 12 minutes at room temperature. Filters were washed 2X with 500ul Ultrapure DNase and RNase free water (VWR, Radnor, PA cat # BDH4216). Microcentrifuge filters were inverted and purified phage particles were eluted using 30ul Ultrapure DNase and RNase free water.

PCR AMPLIFICATION OF PEPTIDE ENCODING INSERT

Amplification of the peptide insert from the purified phage vector was performed in order to prepare eluted fragments for next generation sequencing library construction. Primers were designed against the single stranded M13KeV vector (reference sequence) using Primer3 (F Primer: 5'-TCCTTTAGTGGTACCTTTCTATTC-3', Reverse Primer: 5'TTTGTCGTCTTTCCAGACGTT 3') to result in a double stranded DNA fragment 144bp long (Figure). This length was optimized for compatibility with protocols used for preparation of samples for illumina sequencing, and to allow for the randomized insert fragment to be sequenced during SBS cycles typically producing high quality bases.

Polymerase chain reactions were carried out by adding 1ul of each 200uM primer and 10ul of each purified biopanning product to 25ul 2X Phusion High Fidelity Mastermix (New England Biolabs, Ipswich, MA, USA, cat# F-531B) and 12.5ul Ultrapure Water to wells of a 96 well reaction plate (VWR, Radnor, PA, USA, Eppendorf Twintec Part# 47744-116) . Reactions were incubated at 95°C for 2 minutes, underwent 35 cycles of 95°C for 30 s, 65°C for 30s, 72°C for 60s, and incubated at 72°C for 10 minutes before ramping to 4°C.

AGAROSE GEL ELECTROPHORESIS OF PCR PRODUCTS

Amplified products were visualized by 2.2% agarose gel electrophoresis using Lonza Flashgel system (Lonza Inc., Basel, Switzerland, Cat# 57031) by adding 1ul of sample to 4ul loading dye (Lonza Inc., Basel, Switzerland, Cat# 50462) and 3ul 50bp ladder (Lonza Inc., Basel, Switzerland, Cat# 50475) Gels were allowed to run at 85V for 10 minutes (Figure 9)

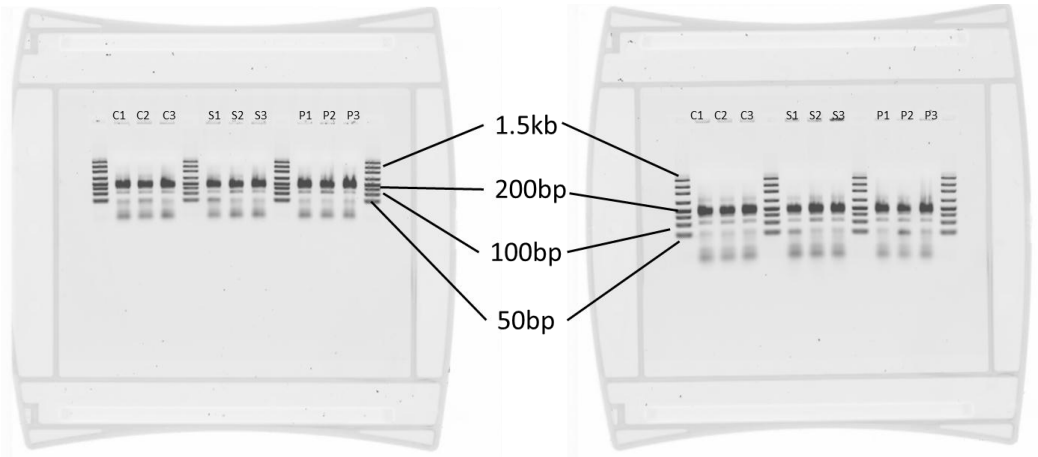

Figure 9. Agarose Gel Electrophoresis for Visualization of PCR Amplified Phage Insert

NEXT GENERATION SEQUENCING

Preparation of Libraries for Next Generation Sequencing

Illumina library construction was performed in order to prepare amplified products for solid phase PCR (cluster amplification) prior to sequencing. These protocols are characterized by four discrete enzymatic reactions in which fragment ends are repaired and phosphorylated, a single adenine residue is added to the 3' end of the fragments, a universal DNA duplex is ligated to the proximal ends of the fragments, and the amplification of library products by PCR. Between each of these steps, reaction products are purified. In order to maximize the efficiency of the sequencing, a molecular indexing strategy is applied where different samples are tagged with a DNA barcode that allows samples to be pooled prior to sequencing.

Separate protocols were used for the construction of the libraries for each of the trials with the protocol with the discrepancies surrounding the enzymatic cleanup methodology and strategy for affixing DNA index barcodes to the library products.

In Trial 1, the enzymatic cleanups were performed using microcentrifuge columns (MinElute PCR Purification Kit, Part# 28006, Qiagen, Hilden, Germany). These were replaced in Trial 2 by a magnetic bead based purification that had been optimized for DNA yield in the time between trials (Fisher, 2011). This protocol does not require samples to be transferred between reaction vessels and significantly increases the yield of entire process. Binding of reaction products to

magnetic beads is mediated by pH and is molecular weight cutoffs are controlled by adding varying volumes of binding buffer to reaction products. Individual reaction cleanups are optimized for the specific molecular weight cutoff requirements of each reaction.

Trial 1 employed a 6bp DNA index barcode, which was added as a tailed primer during the final PCR amplification, and required a separate PCR primer for each sample. In Trial 2, an 8-base barcode was added to the fragments as part of the DNA duplex that is ligated to the amplicons. This required separate DNA duplexes for each sample, but could utilize a universal primer for PCR amplification.

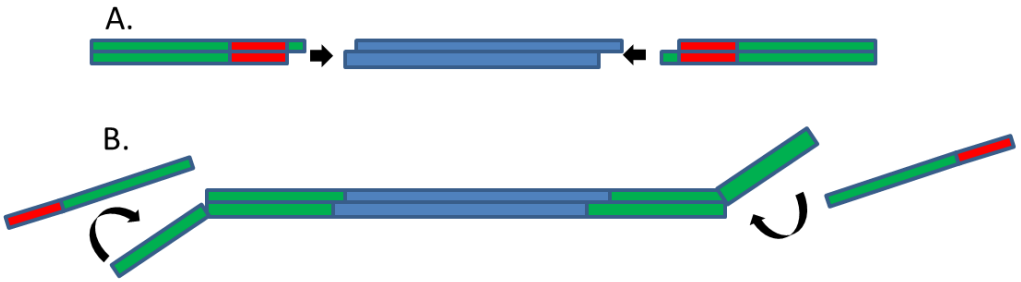

Figure 10. Molecular Indexing Strategies
Addition of molecular indices for pooling of libraries is achieved by two ways.
A. Index is contained within universal adapter and directly ligated to the library molecules
B. Index is added in PCR enrichment and is contained within PCR primer. This approach requires a discrete sequencing read.

Condition	Top adapter strand (5'-3')	Bottom Strand (5'-3')	Unique 8 base index
C1	/5Phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACGCACAC GAATCTCGTATGCCGTCTTCTGCTTG	ACACTCTTTCCCTACACGAC GCTCTTCCGATC*T	GCACACGA
C2	/5Phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACCAGGAG CCATCTCGTATGCCGTCTTCTGCTTG	ACACTCTTTCCCTACACGAC GCTCTTCCGATC*T	CAGGAGCC
C3	/5Phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACGCACATC TATCTCGTATGCCGTCTTCTGCTTG	ACACTCTTTCCCTACACGAC GCTCTTCCGATC*T	GCACATCT
S1	/5Phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACAGTTGCT TATCTCGTATGCCGTCTTCTGCTTG	ACACTCTTTCCCTACACGAC GCTCTTCCGATC*T	AGTTGCTT
S2	/5Phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACTACTTAG CATCTCGTATGCCGTCTTCTGCTTG	ACACTCTTTCCCTACACGAC GCTCTTCCGATC*T	TACTTAGC
S3	/5Phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACAACTTGA CATCTCGTATGCCGTCTTCTGCTTG	ACACTCTTTCCCTACACGAC GCTCTTCCGATC*T	AACTTGAC
P1	/5Phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACGGTCCA GAATCTCGTATGCCGTCTTCTGCTTG	ACACTCTTTCCCTACACGAC GCTCTTCCGATC*T	GGTCCAGA
P2	/5Phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACAGCAATT CATCTCGTATGCCGTCTTCTGCTTG	ACACTCTTTCCCTACACGAC GCTCTTCCGATC*T	AGCAATTC
P3	/5Phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACTTCGCTG AATCTCGTATGCCGTCTTCTGCTTG	ACACTCTTTCCCTACACGAC GCTCTTCCGATC*T	TTCGCTGA

(/5Phos/= 5' phosphorylated *=Phosphorothioate linkage)

Table 4. Trial 1 Library construction adapter ligation based indexing strategy

Condition	Bottom, index containing , "Tag" primer (5'-3')	6 base index
C1	CAAGCAGAAGACGGCATACGAGATCGTGATGTGACTGGAGTTC	ATCACG
C2	CAAGCAGAAGACGGCATACGAGATACATCGGTGACTGGAGTTC	CGATGT
C3	CAAGCAGAAGACGGCATACGAGATGCCTAAGTGACTGGAGTTC	TTAGGC
S1	CAAGCAGAAGACGGCATACGAGATTGGTCAGTGACTGGAGTTC	TGACCA
S2	CAAGCAGAAGACGGCATACGAGATCACTGTGTGACTGGAGTTC	ACAGTG
S3	CAAGCAGAAGACGGCATACGAGATATTGGCGTGACTGGAGTTC	GCCAAT
P1	CAAGCAGAAGACGGCATACGAGATGATCTGGTGACTGGAGTTC	CAGATC
P2	CAAGCAGAAGACGGCATACGAGATTCAAGTGTGACTGGAGTTC	ACTTGA
P3	CAAGCAGAAGACGGCATACGAGATCTGATCGTGACTGGAGTTC	GATCAG

Table 5. Trial 2 Library construction oligonucleotide PCR based indexing strategy

Reaction products were purified by adding 50ul of AMPureXP (Cat # A63880 Beckman Coulter Genomics, Danvers, MA, USA) SPRI reagent to PCR products, mixing 20X, placing reaction plate on a magnetic particle collector (Cat #, Dynal MPC-96S Life Technologies, Carlsbad, CA, USA), allowing plate to incubate for 5 minutes at RT while on the magnet before removing supernatant. Magnetic

beads were washed by adding 60ul of 70% ethanol, and eluted from the beads by adding 40ul 0.1mM Tris-HCl pH8.0 and mixing 20X.

In Trial 1, the plate was placed back on the magnetic particle collector, and the PCR product containing eluate was transferred to individual 1.5ml tubes (Cat# C03217-1, Bioexpress, Kaysville, Utah, USA).

In Trial 2, the beads were allowed to remain in the reaction plate for the remainder of the library construction protocol.

Purified reaction products underwent end-repair in which T4 DNA polymerase is allowed to remove any 3' overhanging bases and fill in any 5' underhanging nucleotides, and T4 polynucleotide kinase is allowed to add a phosphate to any 5' ends and remove any 3' phosphate groups. This was accomplished by adding 5ul of 10X T4 DNA Ligase Buffer, 5ul 1mg/ml BSA, 5ul 10mM ATP, 2ul 10mM each dNTP, 5ul T4 PNK and 5ul T4 Polymerase (New England Biolobs Cat# E6050S/L) to 40ul purified reaction products, capping with an optical stripcap (Life Technologies cat.#4323032) and incubated at 12°C for 15 minutes followed by incubation at 25°C for 15 minutes.

Reaction products were cleaned up in order to remove any residual nucleotides and enzymes, as well as to concentrate reaction products using either Qiagen minelute purification (Trial 1) or SPRI purification (Trial 2) Qiagen minelute was carried out by adding 335ul of supplied PB buffer (Qiagen Cat # 19066) and 67ul of reaction product to a 1.5ml tube. This was added to the Minelute column and

centrifuged at 12,200 rpm for 1 minute. Waste was discarded and the waste tube was returned to the column. Column was washed by adding 750ul of supplied buffer PE (Qiagen Cat # 19065) and centrifuged at 13,200 rpm for 1 minute. Flow through was again discarded and waste tube was dried and returned to the column. Columns were spun for an additional 1 minute at 13,200 rpm to remove any residual ethanol from the buffer PE. Waste tube was discarded and replaced by a 1.5ml collection tube. 40ul of supplied buffer EB (Qiagen Cat# 19086)was then added to the center of the column, which was allowed to stand at RT for 2 minutes. Column was then centrifuged at 13,200 rpm for 1 minute. Columns were discarded and the purified product containing flow-through remained in the collection tube.

SPRI purification (Trial 2) was carried out by adding 147ul of 20% polyethylene glycol 8000 2.5M NaCl pH5.5 to the magnetic bead containing reaction products. This solution was mixed 20X, placed on a magnetic particle collector (Dynal MPC-96S Life Technologies, Carlsbad, CA, USA), allowed to incubate for 5 minutes at room temperature while on the magnet before removing supernatant. Magnetic beads were washed by adding 60ul of 70% ethanol, and eluted from the beads by adding 40ul 0.1mM Tris-HCl pH8.0 and mixing 20X.

Phosphorylated fragments undergo the addition of a single dATP to the 5' end in order to facilitate downstream adapter ligation. This is achieved using Klenow exo- and dATP (New England Biolabs Cat # E60503S/L) . To 40ul of purified, 5'phosphorylated fragments, 5ul of 10X Klenow Buffer 10ul 1mM dATP, 3ul Klenow exo- (NEB part #) and 2ul nuclease free water was added and pipette

mixed 20X. Reaction products were incubated at 37°C for 30 minutes and cleaned up using Qiagen Minelute protocol or 132ul of SPRI binding buffer in Trials 1 and 2, respectively. Purified products were eluted in 40ul Tris-HCL pH8.0.

Adenlylated fragments underwent ligation of synthetic dsDNA adapters based on their trial. In Trial 1, a universal adapter was ligated to all samples, where in Trial 2, adapters for each condition utilized an adapter that contained a unique 8bp region that allowed samples to be pooled following library construction. In both cases, ligation is facilitated by a 3' thymine overhang on the synthetic adapter, which allows for sticky end ligation with library fragments *via* DNA ligase. Double stranded adapter fragments are designed, to have the outside ends, or the region that is not meant to ligate to fragments to have non-complementary ends. This is designed to prevent adapters from annealing in the incorrect orientation. During PCR amplification, the molecules are amplified in such a way that they become double stranded throughout the length of the library fragment.

For Trial 1, 12.5ul 2X DNA QuickLigase Buffer and 3ul each of adapters from Illumina Multiplexing Sample Preparation Oligonucleotide Kit (Illumina part # PE-400-1001) was added to 40ul adenylated fragments. To this was added 2.5ul 1U/ul DNA Ligase (New England Biolabs Cat# E6056S/L). Reactions were allowed to incubate at 25°C for 15 minutes. Reaction products were cleaned up using Qiagen Minelute purification products and eluted in 40ul Tris-HCl pH8.0.

For Trial2 specific adapters were synthesized by Integrated DNA Technologies (Coralville, Iowa, USA). Tables 5 lists ssociations of specific sample conditions with custom adapters.

Oligonucleotides were purified using standard desalting and annealed by heating to 65° for 15 minutes. Ligation of duplexed oligos to adenylated fragments was accomplished by adding 12.5ul 2X DNA QuickLigase Buffer and 3ul custom adapters was added to 40ul adenylated fragments. To this was added 2.5ul 1U/ul DNA Ligase (New England Biolabs Cat # E6065S/L). Reactions were allowed to incubate at 25°C for 15 minutes. Following ligation, reaction products were purified by the addition of 41ul of SPRI binding buffer (20% PEG-8000, 2.5M NaCL pH5.5) to reactions which already contained magnetic beads from prior steps. Purification was carried out as previously specified and desired products were eluted in 40ul and transferred to a new PCR plate.

PCR enrichment of libraries is required to make full length double stranded molecules to be used as input for solid-phase PCR or "cluster amplification." For Trial 1, it is required to add the 6 base index. This is accomplished by using a 3 primer PCR, two of which are universal, and one of which contains the unique 6-base region. During the first round of amplification, the first universal primer is annealed and molecules are extended. Double stranded products of the first cycle are denatured and the second universal primer is annealed. This primer is tailed with a non-complementary region, which results in a 5' overhang. The 3' end of this extension creates the annealing site for the third, index containing primer. The extension products of this extend through the 5' overhang, which creates, for

the first time, the full length molecule needed for Illumina cluster amplification and sequencing.

PCR enrichment was carried out by adding 1ul each of PCR primer InPE 1.0 , PCR primer InPE 2.0 and Tag Primer and 6ul 10X PFU Buffer (Agilent Technologies-Stratagene products, cat.# 200532), 1ul 10mM dNTPs (400ul @ 100mM; Agilent Technologies, cat. # 200415), 2ul PFU enzyme (Agilent Technologies, cat.# 929674) to 40ul of adapter ligated fragments. Reactions underwent incubation at 95°C for 2 minutes, followed by 14 cycles of 95°C for 30 seconds, 65°C for 30 seconds, and 72°C for 1 minute, followed by a 10 minute incubation at 72°C for 10 minutes prior to a ramp to 4°C.

Indexed adapter ligated products from Trial 2 were amplified by adding 1ul each of Illumina PCR Primer PE 1.0 and Illumina PCR PE 2.0, 1ul 10mM dNTPs (400ul @ 100mM; Agilent Technologies, cat. # 200415), 2ul PFU enzyme (Agilent Technologies, cat.# 929674) to 40ul of adapter ligated fragments. Reactions underwent incubation at 95°C for 2 minutes, followed by 6 cycles of 95°C for 30 seconds, 65°C for 30 seconds, and 72°C for 1 minute, and concluded with a 10 minute incubation at 72°C for 10 minutes prior to a ramp to 4°C.

PCR amplified molecules from both trials were purified by adding 90ul of SPRI AMPure reagent (Cat # A63880 Beckman Coulter Genomics, Danvers, MA, USA) to reactions, mixing 20X, placing reaction vessels on a magnetic particle concentrator (Life Technologies part# MPC96S), and allowing to incubate on the magnet for 2 minutes. Reaction supernatants were discarded and beads were

washed with 70% ethanol. Plates were removed from the magnet, and 40ul of 1mM Tris-HCL pH 8.0 was added to each well containing sample and mixed. Plate was placed back on the magnet, and eluates were transferred to fresh tubes.

Quantitation of Libraries

Accurate quantitation of the number of molecules present within the sample that are capable of binding to the Illumina flowcell surface is critical to predicting the density of clusters following cluster amplification. A quantitative PCR assay was performed on all libraries to accomplish this. A kit used from KAPA Biosystems (KAPA Part#:KK8200) that combines priming oligonucleotides specific for portions of the universal adapters ligated to library fragments with SYBR Green, a fluorescent dye that intercalates specifically with DNA molecules, and PCR reagents. Reactions were placed on the Applied Biosystems 9700 qPCR instrument and subjected to qPCR cycling. Reactions were analyzed using the system software.

	Enriched Lib	Index BC	Index Name	frag size	date		Concentration(nM)
C1	Solexa-17868	91976405	629	220	3.16.10 Set 2	3.16.10	73.254537
C2	Solexa-17869	91976741	357	220	3.16.10 Set 2	3.16.10	55.6152212
C3	Solexa-17870	91976740	630	220	3.16.10 Set 2	3.16.10	62.8535133
S1	Solexa-17876	91976719	236	220	3.16.10 Set 2	3.16.10	61.3475938
S2	Solexa-17877	91976720	800	220	3.16.10 Set 2	3.16.10	58.6133384
S3	Solexa-17878	91976721	34	220	3.16.10 Set 2	3.16.10	60.8056423
P1	Solexa-17879	91976718	726	220	3.16.10 Set 2	3.16.10	77.4297457
P2	Solexa-17880	91976717	190	220	3.16.10 Set 2	3.16.10	0.19909766
P3	Solexa-17881	91976716	954	220	3.16.10 Set 2	3.16.10	73.892315

Table 6. Library Quantification using SYBR qPCR assay

CLUSTER AMPLIFICATION AND SEQUENCING

Prior to clonal amplification, libraries were quantified using a qPCR protocol with specific probes for the ends of the adapters. The qPCR assay measures the quantity of fragments properly adapter ligated that are appropriate for sequencing. Based on the qPCR quantification, libraries were normalized to 2nM and then denatured using 0.1 N NaOH. Cluster amplification of denatured templates occurred according to manufacturer's protocol (Illumina) using V2 Chemistry and V2 Flowcells (1.4mm channel width). Sybr Green dye was added to all flowcell lanes to provide a quality control checkpoint after cluster amplification to ensure

optimal cluster densities on the flowcells. Flowcells were sequenced on the Genome Analyzer IIx , using v3 Sequencing-by-Synthesis kits.

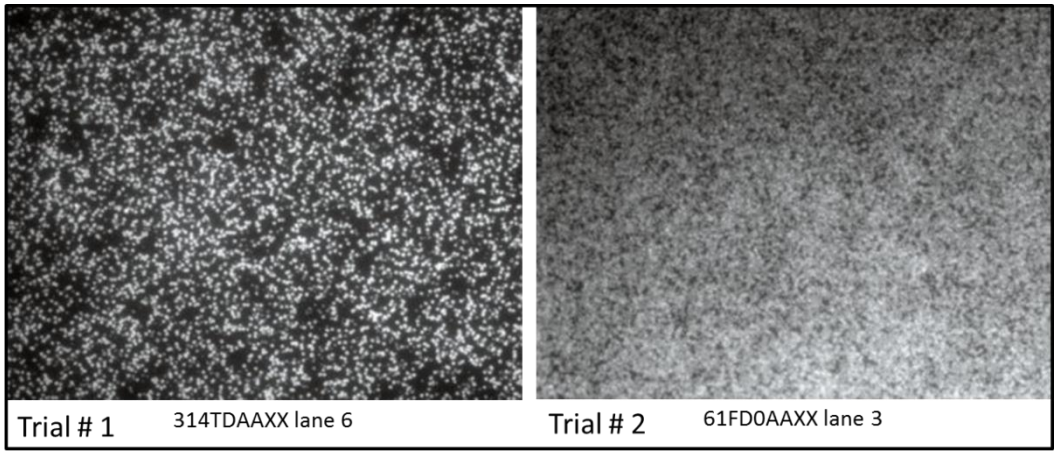

Figure 11. SYBR Green Fluorescence of Cluster Amplified Fragments

RESULTS

PROCESSING AND ANALYSIS OF SEQUENCING DATA

Preparation of Sequencing Reads for Translation

Analysis of generated data was performed through manipulation of .fastq files generated as outputs of the Bustard basecaller (Kircher, 2009) in order to convert within the Illumina Analysis pipeline the Illumina v1.3.4 pipeline, and for Trial 2 using the Illumina 1.6.0a16 pipeline. The output of each sequencing run resulted in large numbers of passing filter reads, with a total of 516,704 Reads in Trial 1, and 26,631,090 Reads in Trial 2. The significant increase in reads generated between the two trials is the result of improved efficiencies of image processing software, allowing for increased ability to discern between polyclonal clusters, which in turn allows denser packing of molecules on the surface of the flowcell.

	Trial 1	Trial 2
C1	90303	2451080
C2	24579	3549480
C3	69021	3191972
S1	28083	3231661
S2	52450	3494358
S3	48772	3477076
P1	85478	3329503
P2	32433	1752
P3	85585	3904208

Table 7. Comparison of Passing Filter Sequence reads between trials per condition

These files contain the raw sequence reads, as well as quality data and cluster identity information. Files were first parsed to create separate files containing sequences only using the AWK programming language.

In order to identify reads derived from the phage vector, sequences were aligned to the M13keV vector sequence using the eland aligner (Illumina).

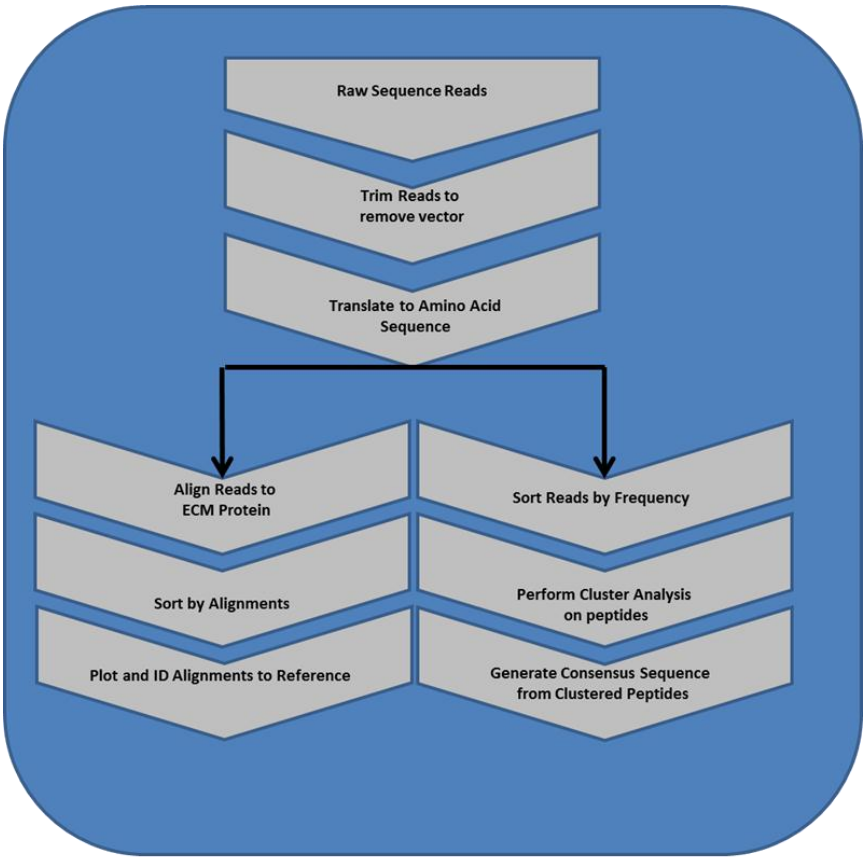

Figure 12. Overview of Analysis Workflow

Sorting of Reads by Molecular Index

Resulting reads were sorted according to the index containing portion of the read using UNIX commands and separate files were created and associated with the experimental condition.

Each of these files was trimmed, removing the leading and trailing portions of the read outside of the random 36bp insert, again using a script written the AWK language.

These files were then converted to FASTA format in preparation of translation into amino acid sequences.

Translation of Reads to Amino Acid Sequences

Once in FASTA format, the sequences were translated to amino acid residues using the BioPerl version 1.6.1 bp_translate_seq.pl program (Stajich, 2002), which resulted in FASTA formatted files containing the translated, 12 amino acid residue, peptide sequences.

The resulting sequences were sorted by number of observances, and a tally of the number of observances for each sequence was added. Further analysis shows that while the vast number of peptide sequences were obtained were singletons, as an expected result of the removal of mid-biopanning round amplification, there are several sequences that occur up to several thousand times within a condition (Figure 13). The top 100 sequences in terms of frequency of occurrences are

listed along with peptide sequences shared across wash stringencies are listed in the appendix. Additionally, the peptides representing the 100 most representative recovered were analyzed for their amino acid content (Figure 14)

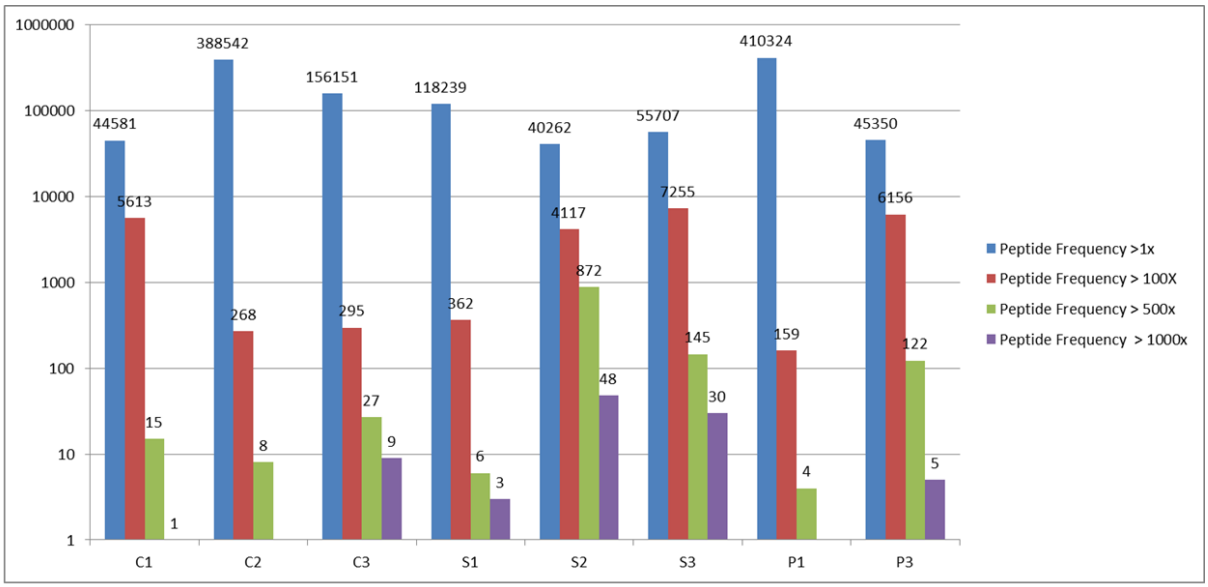

Figure 13. Peptide Frequency Plot

Frequency of individual peptides were calculated and binned for visualization of distributions.

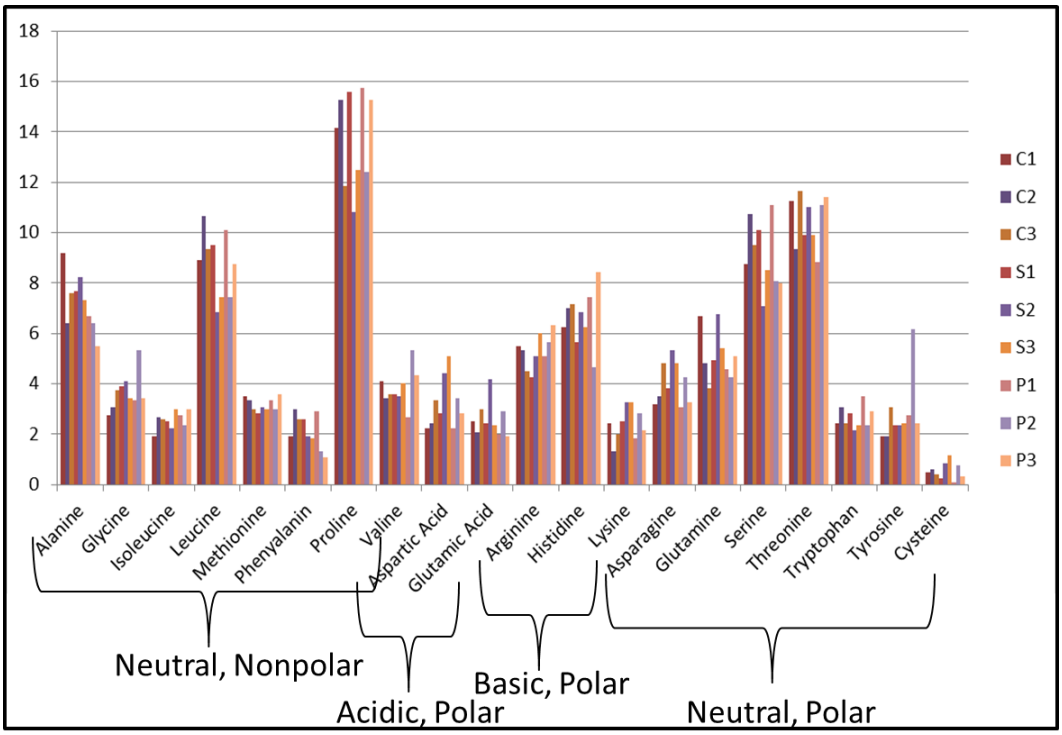

Figure 14. Amino Acid Frequencies of top 100 Peptides by Condition

Amino acid representation within 100 most common peptides associated with each experimental condition grouped by pH and chemical polarity

Cluster Analysis and Derivation of Consensus Sequences

The sequences in Table 8 , which should represent the strongest binding amino acid motifs to various substrates of interest, were filtered first through the Immune Epitope Database (IEDB) cluster analysis tool which clusters a group of sequences that has a sequence similarity greater than the minimum sequence identity threshold specified (Peters, 2005). This clustering step is based on the assumption that these sequences could demonstrate homology to a number of natural protein binders. The identity threshold of 30% was chosen to demonstrate the greatest number of clusters in order to separate these by the

various components capable of binding to the substrate of interest. An example of the result of this clustering is found in Figure 15. Consensus sequence logos were derived from cluster populations using the software tool Weblogo (Schneider, 1999; Crooks, 2004). This generates the consensus sequence logo that provides a visual representation of the consensus sequence where the letter codes are scaled by their frequency in the list of input sequences. An example logo using the largest cluster generated in the IEDB tool is also shown in Figure 15.

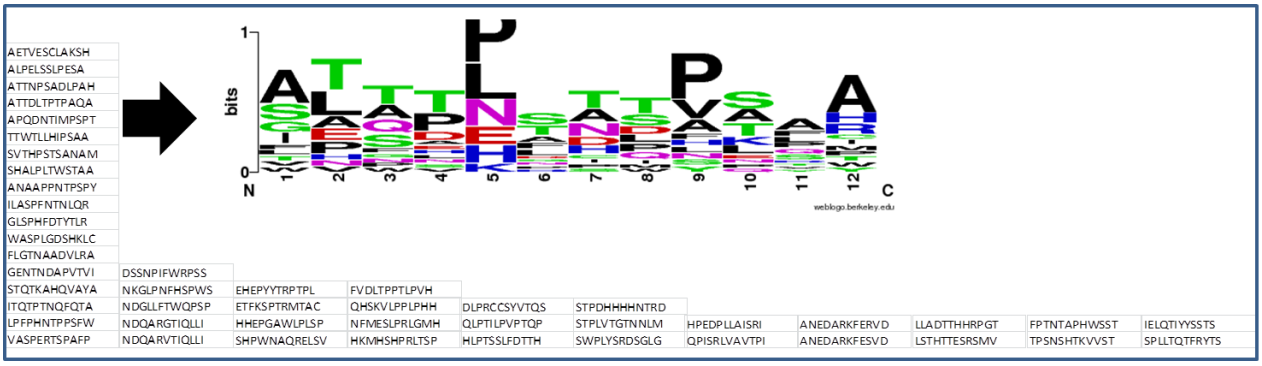

Figure 15. IEDB Cluster Analysis and Consensus Logo

Example of Immune Epitope Database cluster analysis output, and generation of consensus logo using UC Berkeley weblogo tool to identify Motif groupings and consensus sequences as candidates for peptides to be used for further protein binding studies.

Alignment of Peptide Sequences to ECM Proteins

Sequences were aligned to extracellular Matrix proteins using the BLAST software (Basic Local Alignment Search Tool) (Altschul, 1990) using an E-value threshold of 0.1. The E-value provides a direct measure of the probability of the alignment being due to random occurrences, and therefore the lower the E-value, the more confident we are in the alignment. This threshold was set intentionally low due to the short length of the query sequences and the high degree of background expected without multiple rounds of amplification in the biopanning

process. Accession ID's for the ECM protein sequences used in the alignments are listed in Table 9.

ECM Protein References	
Protein	Accession ID
Biglycan	>sp|P47853|PGS1_RAT
Chondroitin Sulfate	>sp|Q9ERQ6|CSPG5_RAT
Collagen I	>UPI000155113E
Collagen II	>sp|P05539|CO2A1_RAT
Collagen IV	>UPI0001B79E8F
Elastin	>tr|D3ZA03|D3ZA03_RAT
Fibronectin	>sp|P04937|FINC_RAT
Heparin	>UPI0000DA215D
HLA	>sp|O35776|HAS2_RAT
Integrin	>sp|P49134|ITB1_RAT
Keratin	>sp|Q8CCX5|KT222_MOUSE
Laminin	>tr|D3ZQN7|D3ZQN7_RAT
Syndecan 1	>sp|P26260|SDC1_RAT

Table 8. ECM Protein Reference Sequences

Accession Id's for the ECP protein sequences used as reference sequence for recovered peptide alignment

Analysis of Sequence Representation Statistics

Analysis of ECM Protein Alignments

The alignments themselves can be graphically displayed (Figure 16) plotting the reference (ECM protein) start and alignment length in gantt charts, which shows the region along the reference protein sequence to which the peptides are aligned. These graphical displays identify regions in our collected peptide fragments that are homologous to regions of the ECM protein we are screening for affinity to our substrates of interest Thus, by looking at loci along

the ECM protein amino acid sequence that show high levels of recovered peptides binding to the substrate, it can be concluded that these regions identify motifs of interest that may be involved in ECM binding to our substrates of interest, and candidates for validation using alternative means such as qPCR analysis, SPR, or other alternatives for low-throughput interrogation of epitope binding (Wegner, 2002).

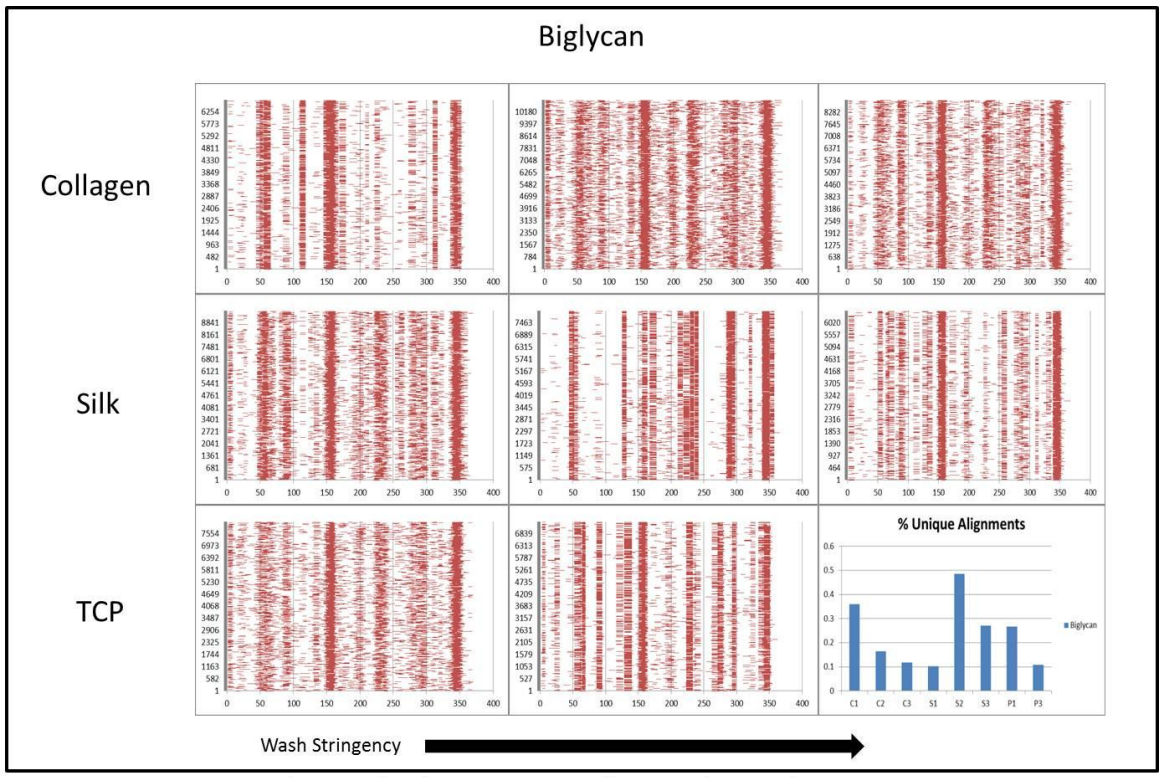

Figure 16. Recovered peptide alignment to substrate by wash stringency.

Unique Alignment Analysis

In order to test ECM specificity for substrates, 100% alignments were compared by wash stringency for numbers of unique alignments. Peptide alignments were filtered for unique reference start and end loci. The frequencies of alignments

were compared for percentage of alignments deriving from unique peptides.

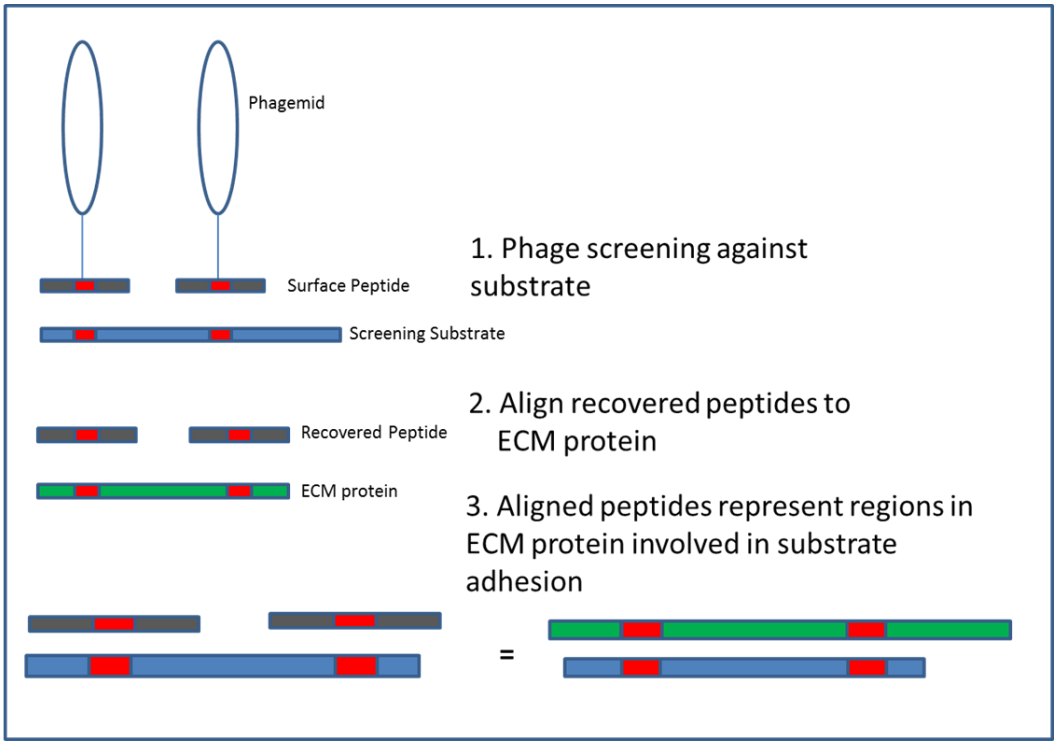

Figure 17. Overview of Sequence Alignment Strategy

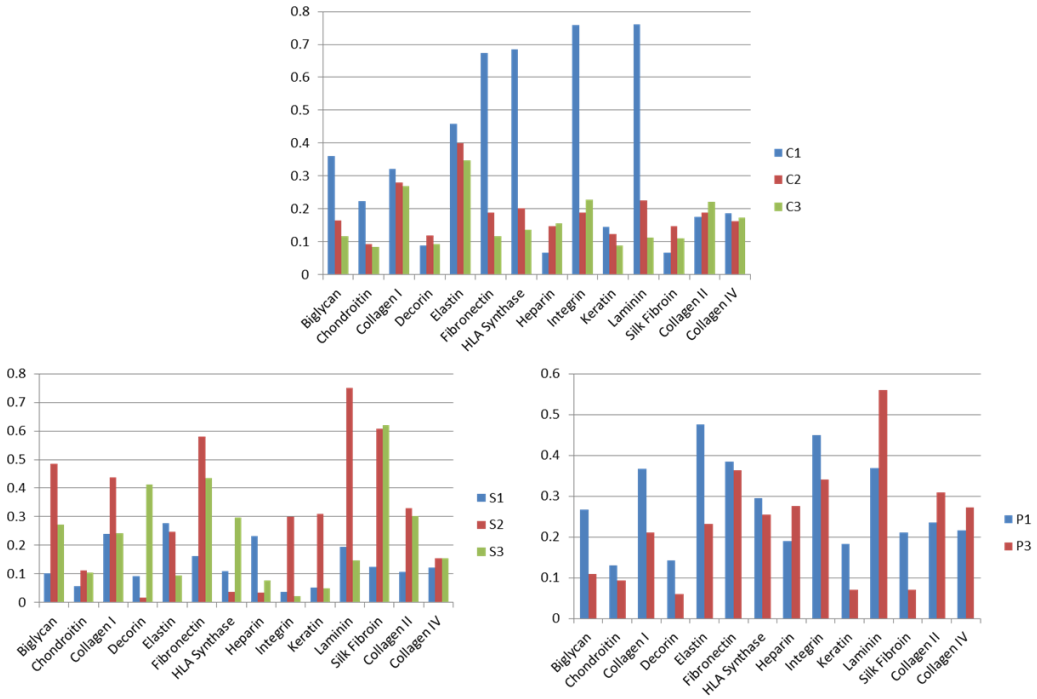

Figure 18. Percent Unique Alignments of 100% aligning peptides
It is hypothesized that percent uniqueness can be an indicator of ECM protein specificity for substrate.

Z-Score Calculation

A further measure of the validity of this approach is the calculation of a z-score. Z-scores are defined as the number of standard deviations from the mean of a given reference set. Ideally, this reference set would be deep sequencing of the native phage library. In the absence of these data, z-scores were calculated using the tissue culture plastic (polystyrene) as the reference relative to the silk and protein alignments. Z-scores were calculated using the below formula and shown in Figure 19.

$$\text{z-score} = \frac{\%\ \text{unique}_{(\text{silk/collagen})} - \text{mean}\ \%\ \text{unique}_{(\text{silk/collagen})}}{}$$

Standard deviation % unique (polystyrene)

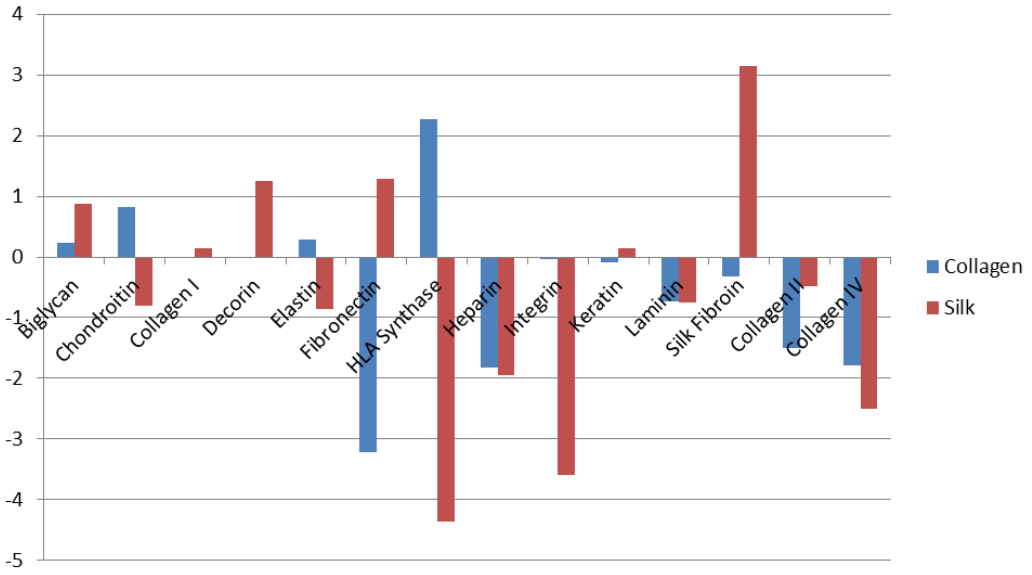

Figure 19. Z-score Calculations Relative to Polystyrene

E-value filtering

An additional level of control for which peptide alignments can be gauged for specificity to substrate is through the filtering of alignments by e-value. As previously described, the e-value is a direct measure of the statistical significance of the alignment, taking into account alignment length as well as the number of gaps present within the alignment. E-values and P-values are directly related, as e-values can be described as the product of P-values and the number of tests performed. The alignment e-value cutoff used for all of the alignments was 0.9, which corresponds to a P-value of 0.094 using the below function.

$$P = 1\text{-}e^{-E}$$

This corresponds to a statistical significance level of 0.006, which indicates high probability that the aligment would not be a function of random interaction. An example of an alignment filtered with e-value of 0.1 versus 0.006 is shown in figure 18.

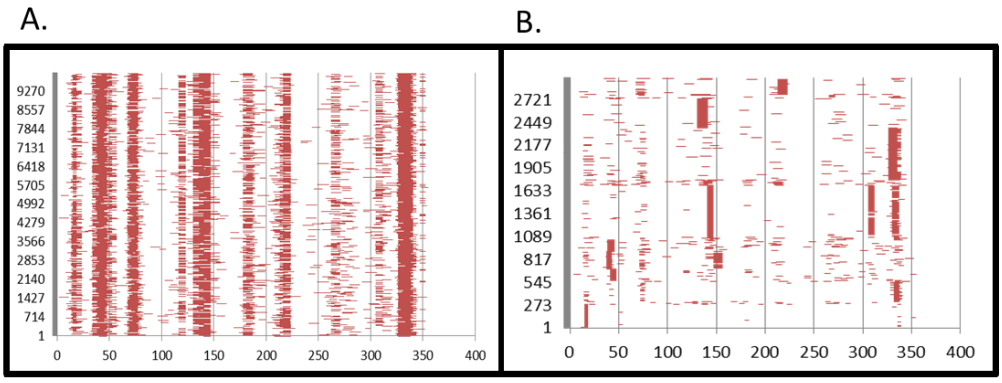

Figure 20. Filtering alignments by e-value

Demonstration of the ability to find higher affinity sequences through bit-score filtration of High Scoring Pairs. A. All sequence alignments obtained through alignments ≤0.9. B. Further filtration of alignments using cutoff of ≤0.099 demonstrates peptide regions with predicted increased affinity.

DISCUSSION

DNA sequencing of the amino acid encoding inserts recovered through phage display screening presents a powerful tool for the broad scale interrogation of the mechanisms of protein-protein interactions. The phage display process is one that can be automated and is flexible to be scaled and applied to a variety of screening approaches. All of the variables used to manipulate the screen are able to be manipulated, including input phage display library composition, reference substrate, and the manner in which the two are permitted to interact. Advances in DNA sequencing and necessary computational tools have removed the previous roadblocks to recognizing the true potential of phage display screening. While this technology shows great promise, there is still much work to be done to optimize for greater control and utility.

The comparison of individual amino acid frequencies of the library between substrates and wash stringencies show similar distributions (Figure 19). This, coupled with the vast numbers of unique peptides obtained through the sequencing indicates that the data contains a high level of background noise present.

SEQUENCE DEPTH

In order to understand the benefit of this approach relative to phage display with mid-round amplification and standard sequencing, results were compared with regard to the number and frequency of recovered peptides. Overall, with a single

lane of Illumina sequencing, 26 million peptides were recovered, with an average of >2.2 million peptides per experimental condition. Of this number, an average of 606,000 are unique peptides, meaning that ~1.4 million are observed at a frequency of >1X. Relative to the total number of possible clones in our PhD-12 library at 1B, and the high frequency of peptides occurring >1X, 100X, and 1,000X (Figure 13), it can be concluded that we are a.) Enriching for peptides likely interact with the substrate, and b.) Generating more sequence data than is needed to observe "high frequency peptides." If a "high frequency peptide" is defined as a peptide that is occurring at a frequency >100X per condition within the data generated in Trial 2, an average of 3028 peptides per condition satisfied this requirement. These peptides represent a small percentage of the overall data at an average of 0.13% of the recovered peptides within a condition. In order to maximize the efficiencies of economy for Illumina sequencing, a ten-fold reduction in the overall amount of sequence data per condition would allow a ten-fold increase in the number of conditions per lane. This is desired over a decrease in the amount of total sequence data, as while lower throughput sequencing is available, it is not the most cost effective option. Applying the percentage of observed high frequency peptides relative to the total amount of peptides recovered to a 10-fold reduction in peptides would produce a total of 90 conditions with a diversity of >220,000 peptides, with 60,600 unique and average of 302 peptides showing up >100X. This would allow us to continue to separate these from background singletons, and produce more manageable list of candidate peptides for downstream validation.

% UNIQUE ALIGNMENT AND PCR DUPLICATION

It is hypothesized that this will provide an indication of substrate protein specificity for ECM proteins. In situations where the ECM protein alignments demonstrate a percentage of unique alignments that is low, we have few different phage peptides aligning to the ECM region. In contrast, where the percentage of unique alignments is high there are several different peptides that contain the aligned motif sequence. This can be confounded by the presence of PCR duplicates, which likely exist within the data. Typically, these are screened and filtered by looking at unique start and end sites, but since the peptide encoding insert required amplification, the start and end sites correspond to the PCR priming sites. An alternative approach would be to randomly shear the phage vector, and ligate NGS library adapters followed by PCR amplification, as this would create fragments with more random start and end sites.

It is hypothesized that where the percentages of unique alignments decrease with increases in wash stringency, ECM protein are more likely to be directly involved in weak binding with the substrate of interest, as this would correspond to peptides being removed as stringencies increase. In contrast, instances where the percentage of alignments that are unique increases with wash stringency would indicate strong adhesion between the ECM protein and substrate of interest. Percentages of unique alignments by condition are shown in Figure 17.

KNOWN BINDING MOTIFS

Leucine Rich Repeats

Interesting observations of both peptides demonstrating the highest frequency of occurrence, as well as presence across substrate types, is the peptide with the sequence ALPELSSLPESA The peptide ranks in the top 2 for three conditions tested (C1, C2, P1) and is present in the top 100 for five of the nine conditions tested. It is hypothesized that this prevalence could be a function of the presence of interspersed leucine residues, as the peptide would resemble the well-characterized leucine-rich repeat (LRR) that has been studied for its role as a protein recognition motif (Kobe, 2001).

Laminin Adhesion to Collagen I

As an example of the utility of this approach for discovery of binding motif's the example of laminin adhesion was tested. Laminin contains a well characterized adhesion site with the amino acid sequence CDPGYIGSR. (Mayer, 1993) This peptide has been studied extensively for its ability to inhibit angiogenesis and solid tumor growth (Iwamoto, 1996).

Alignment metrics for peptides recovered from C1 condition to laminin reveals these alignments having the highest frequency of alignments with an e-value of <0.1 results from the 12 aa peptide spanning residues 993 to 1004. The second most frequent alignment in terms of frequency with an e-value of 0.007 results

from a peptide spanning residues 961 to 970, the region in laminin giving rise to the YIGSR peptide.

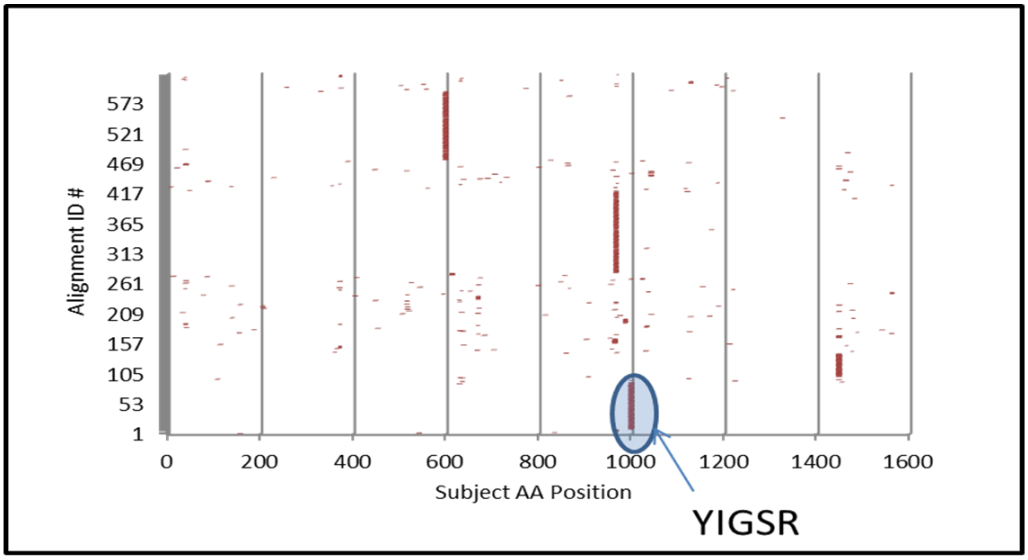

Figure 21. Sequence Alignments of Collagen Screened Peptides to Laminin
Alignment of peptide sequences screened against collagen reveals characterized YIGSR laminin adhesion motif.

Fibronectin Adhesion to Collagen I

Another example demonstrates this utility in the well characterized binding of fibronectin to collagen I (Figure 21). Analysis of aligned peptides reveal two distinct areas if interest located 5 residues long, and centered around the 10th and 30th residue of the fibronectin reference sequence. The amino acid sequence contained in the region of the first area contains the sequence RGNY, a well characterized fibronectin adhesion domain to collagen I (Preciado-Patt, 1994)

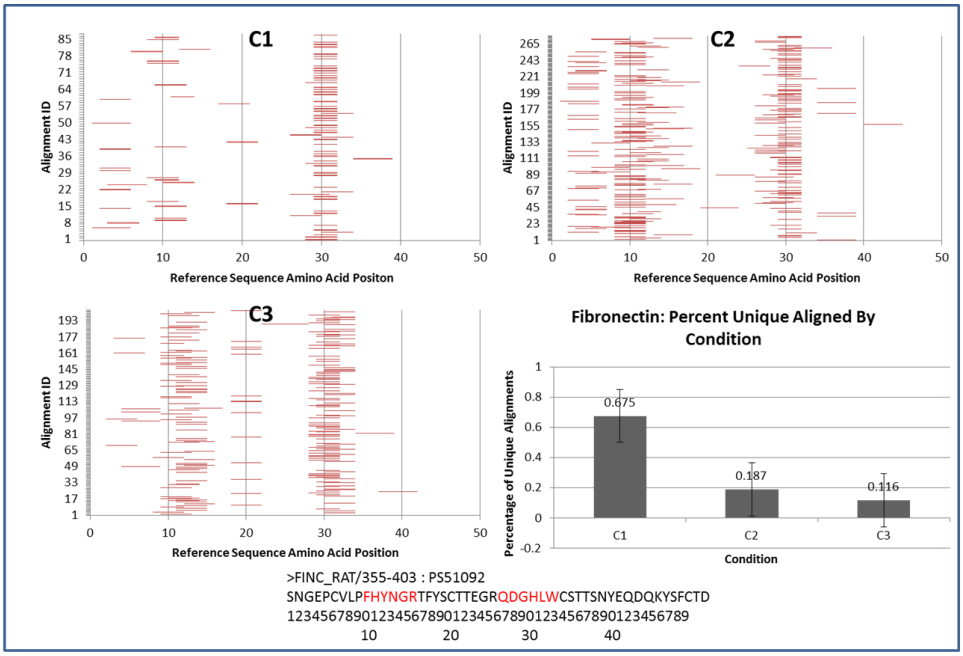

Figure 22. Alignment of Recovered Phage Peptides from Collagen Substrate to Fibronectin

Alignments reveal RGNY sequence as a high scoring pair. This peptide motif has been previously described to play a role in fibronectin adhesion to collagen I.

Future Work:

OPTIMIZATION OF WASH STRINGENCY

Experimental work is needed to optimize the manner in which the library-substrate interaction is controlled. One of the greatest limitations to a cell-free process is the sheer amount of background peptide sequences observed as a result of the omission of amplification followed by additional rounds of panning. While sequencing costs continue to diminish, the amount of sequencing coverage needed to differentiate meaningful contributors to the interactions being studied increases with the amount of background sequences present. This, in turn affects computational facility and power.

The factors controlling stringency of the interaction being interrogated include wash buffer composition, including detergents and surfactants, binding and elution time and temperature, substrate concentration, and the input library concentration based on complexity. These factors need to be fully understood and further optimized in order to reduce the amount of nonspecific binding given the lack of mid-round amplification and subsequent rounds of panning.

SEQUENCING OF NATIVE PHAGE LIBRARY

One of the lacking features of this study was a well-defined reference data set for which to compare alignments. While the use of the tissue culture plastic conditions were used for general assessments, this is not an ideal data set as there

are interactions that are specific to this substrate, but also, because these interactions are likely to occur alongside the other substrate interactions, as these substrates were first coated onto tissue culture plastic. A full sequence analysis of the pHD-12 library would likely have provided a better reference data set upon which to compare unique alignments and calculate more meaningful Z-scores for the data produced. It is noted that as this is an independent study, a future comparison of the data produced in this study with newly produced data would be valid.

COMPARISON TO MID-ROUND AMPLIFICATION

Due to the requirement for this process that no bacterial cells be used, significant challenges are posed in order to measure the efficiency of this approach relative to the standard approach, which uses infection into bacterial for mid-round amplification, and decreases the diversity of the library entering the second round of amplification. A direct comparison between the data produced with this approach relative to approach described in this thesis would help to better understand the benefit of deep sequencing relative to traditional approaches to phage display screening.

DIRECTED LIBRARY DESIGN AND SCALEUP

The application of the advances in DNA synthesis technologies, mainly developed for use in DNA microarrays, will present opportunities for high throughput design of custom peptide libraries. These can be easily cloned and transfected into host phage. The automation of the process for the rapid design, synthesis, and production of these libraries presents a means for the high throughput screening of a customized peptide library against virtually any substrate molecule. Follow-up validation studies will determine the efficacy of derived consensus sequences. Synthesis of consensus sequences and screening of libraries using techniques including ELISA or surface plasmon resonance (SPR) or other methods that while lower throughput, can provide an orthogonal validation approach (Wegner, 2002)

ANTIBODY SCREENING FOR EPIGENETIC RESEARCH

One current area of scientific research standing to benefit from phage display derived antibodies is the growing field of epigenomics, where protein-nucleic acid interactions are studied to discover, understand, and control transcription factors. A highly popular, efficient, and effective means to study this is next generation sequencing of immunoprecipitated chromatin fragments. This technique is based on crosslinking proteins to DNA and using antibodies specific to the protein of interest to pull down the fragments of DNA that are bound to the protein. Recovered DNA fragments are sequenced, aligned to the reference genome, and

counted to identify regions strongly bound to the protein, and likely involved in regulation of gene expression. The technique best used to measure the utility of an antibody for use in ChIPseq is by using the antibodies for immunoprecipitation. A highly valuable system for the screening of antibodies would be phage display screening against transcription factors as described in this text, followed by ChIPseq for validation of the antibody and qualification for use in experimentation.

SOFTWARE FOR PEPTIDE ANALYSIS

User interfaces for high throughput data manipulation and result computation need to be deployed, though the framework for this is contained within this document. In order to facilitate the sharing of information derived from these studies, public databases to store experimental designs and results must be developed (Huang, 2011). Further validation strategies for the results including experimental and *in silico* tools need to be developed in order to confirm the observances elucidated through this approach. This is true both for derived consensus sequences, as well as to further understand the binding motifs uncovered through the alignment of peptides to known protein sequences.

CONCLUSION

In conclusion, the direct sequencing of phage display products produces an overwhelming amount of data. More work is clearly needed to be performed both in terms of the optimization of conditions providing minimal background noise in terms of nonspecific binding products entering the downstream sequencing, as well as the development of user friendly tools in which to visualize and understand the output data

Works Cited

Willats, William G.T. 2002. **Phage display: practicalities and prospects**. Plant Molecular Biology, 50: 837-54.

Paschke, Matthias. 2006. **Phage display systems and their applications**. Applied Microbial Biotechnology, 70: 2-11.

Arap, Marco Antonio. 2005. **Phage display technology – Applications and innovations**. Genetics and Molecular Biology, 28(1):

Burton, D.R., Scott, J.K., Silverman, G.J. 2001. Phage Display, A Laboratory Manual. Laboratory Press, Cold Spring Harbor, NY.

Marvin, D.A., **Filamentous phage structure, infection and assembly**. 2006. Current Opinion in Structural Biology, 8(2): 150-158.

Kehoe, J.W., Kay, B.K**., Filamentous Phage Display in the New Millenium**. 2005. Chemistry Review, 105: 5056-4072.

Smith, G. P., **Filamentous Fusion Phage: Novel Expression Vectors that Display Cloned Antigens on the Virion Surface**. 1085. Science, 228: 1315-1317.

Tatusov, R.L., Altschul, S.F., Koonin, E.V. **Detection of conserved segments in proteins: Iterative scanning of sequence databases with alignment blocks**. Proceedings of the National Acadamy of Sciences, USA. 91: 12091-12095.

Ben-Gal, I., Shani, A., Gohr, A., Grau, J., Arviv, S., Shmilovici, A., Posch, S., Grosse, I., **Identification of transcription factor binding sites with variable-order Bayesian networks**. 2005. Bioinformatics, 21(11): 2657-2666.

Beckstette, M., Homann, R., Giegerich, R., Kurtz, S., **Fast index based algorithms and software for matching position specific scoring matrices**. 2006. BMC Bioinformatics, 7: 389.

Moreau, V., Granier, C., Villard, S., Laune, D., Molina, F. 2005. Discontinuous epitope prediction based on mimotope analysis. Structural Bioinformatics, 22(9): 1088-1095.

Mayrose, I., Shlomi, T., Rubinstein, N.D., Gershoni, J.M., Ruppin, E., Sharan, R., Pupko, T. 2006. **Epitope mapping using combinatorial phage-display libraries: a graph-based algorithm.** Nucleic Acids Research, 35(1): 69-78.

Burg, M., Raveym E.P., Gonzales, M., Amburn, E., Ho Faix, P., Baird, A., Larocca, D. 2004. **Selection of Internalizing Ligand-Display Phage Using Rolling Circle Amplification for Phage Recovery**. DNA and Cell Biology, 23(7): 457-462.

Molennar, T.J.M., Michon, I., de Haas, S.A.M., van Berkel, T.JC., Biessen, E.A.L., 2002. **Uptake and Processing of Modified Bacteriophage M13 in Mice: Implications for Phage Display**. Virology, 293(1): 182-191.

Yang, L., Nolan, J.P., **High-Throughput Screening and Characterization of Clones Selected from Phage Display Libraries**. 2007. Cytometry Part A, Journal of the International Society of Analytical Cytology, 71A: 625-631.

Rahim, A.A., **Pyrosequencing of Phage Display Libraries for the Identification of Cell-Specific Ligands**. 2006. Methods in Molecular Biology, 373: 135-146.

Fack, F., Hugle-Dorr, B., Song, D., Queitsch, I., Petersen, G., Bautz, E.K.F. 1997. **Epitope Mapping by phage display: random versus gene-fragment libraries**. 206: 43-52.

Hansen, N.J.V., Pedersen, L.O., Stryhn, A., Buus, S. **Phage display of peptides major histocompatibility class I complexes**. 2001. European Journal of Immunology, 31: 32-38.

Koivunen, E., Restel, B.H., Rajotte, D., Lahdenranta, J., Hagedorn, M., Arap, W., Pasqualini, R. 2003. Integrin-Binding Peptides Derived from Phage Display Libraries. Methods in Molecular Biology, 129: 3-17.

Deshayes, K., Schaffer, M.L. Skelton, N.J. Nakamura, G.R., Kadkhodayan, S., Sidhu, S.S., **Rapid Identification of Small Binding Motifs with High-Throughput Phage Display: Discovery of Peptidic Antagonists of IGF-1 Function**. Chemistry& Biology, 9: 495-505.

Arap, W., Pasqualini, R., Ruoslahti, E. 1998. **Cancer Treatment by Targeted Drug Delivery to Tumor Vasculature in a Mouse Model**. Science, 279: 377-380.

Barry, M.A., Dower, W.J., Johnston, S.A. 1996. **Toward cell-targeting gene therapy vectors: Selection of cell-binding peptides from random peptide-presenting phage libraries**. Nature Medicine, 2(5): 299-305.

Maxwell, D.J., Hicks, B.C., Parsons, S., Sakiyama, S.E. 2004. **Development of rationally designed affinity-based drug systems**. Acta Biomateriala, 1: 101-113.

Fernandez-Gacio, A., Uguen, M., Fastraz, J. 2003. **Phage display as a tool for the directed evolution of enzymes**. Trends in Biotechnology, 21(9): 408-414.

Rader, C., Barbas, C.F. 1997. **Phage display of combinatorial antibody libraries**. Current Opinion in Biotechnology, 8: 503-508.

Uetz, P.Glot, L., Cagney, G., Mansfield, TA., Judson, RS., Knight, JR., Lockson, D., Narayan, V., Srinivasan, M., Pochart, P., Quereshl-Emill, A., Li, Y., Foodwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar., G., Yang, M., Johnston, M., Field, S., Rothberg, J. 2000**. A Comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae.*** Nature, 403: 623-27.

Berstein, BE., Kamal, M., Lindblad-Toh, K., Beklranov, S., Balley, D., Huebert, D., McMahon, S., Karlsson, E., Kulbokas, E., Gingeras, T., Schreiber, s., Lander, E. 2005. **Genomic Maps and Comparticie Analysis of Histone Modifications in Human and Mouse.** Cell, 120: 169-181

Walter, G., Bussow, K., Cahill, D., Lueking, A., Lehrach, H., 2000. **Protein arrays for gene expression and molecular interaction screening.** Current Opinion in Microbiology, 3: 298-302.

Marioni, JC., Mason, CE., Mane, SM, Stephens, M, Gilad, Y. 2008. **RNA-seq an assessment of technical reproducibility and comparison with gene expression arrays.** Genome Research, 18(9): 1509-17.

Lyne, PD. 2002**. Structure-based virtual screening: An overview**. Drug Discovery Today, 7(20): 1047-55.

Vlieghe, P., Lisowski, V., Martinez, J., Khrestchatisky, M. 2010. **Synthetic therapeutic peptides: science and market**. Drug Discovery Today, 15(1/2): 40-56.

Sreejalekshmi, K.G., Nair, P.D., 2010**. Biomimeticity in tissue engineering scaffolds through synthetic peptide modifications-Altering chemistry for enhanced biological response**. Journal of Biomedical Materials Research A, 96(2): 477-491.

Konthur, Z., Walter, G., 2002. **Automation of phage display for high-throughput antibody development**. Targets, 1(1):30-36.

Schofield, D.J., Pope, A.R., Clemente, V., Buckell, J., Chapple, S.D.J., Clarke, K.F., Conquer, J.S., Crofts, A.M., Crowther, S.R.E., Dyson, M.R., Flack, G., Griffin, G.J., Hooks, Y., Howat, W.J., Kolb-Kokocinski, A., Kunze, S., Martin, C.D., Maslen, G.L., Mitchell, J.N., O'Sullivan, M., Perera, R.L., Roske, W., Shadbolt, S.P., Vincent, K.J., Warford, A., Wilson, W.E., Xle, J., Young, J.L., McCafferty, J. 2007**. Application of phage display to high throughput antibody generation and characterization**. Genome Biology, 8(11): 254;1-11.

Sanger, F., Nicklen, S., Coulson, A.R. 1977. **DNA sequencing with chain terminating inhibitors**. Proceedings of the National Academy of Science, USA, 74(12): 5463-5467.

Smith, L.M., Sanders, J.Z., Kaiser, R.J., Highes, P., Dudd, C., Connell, C.R., Heiner, C., Kent, S.B.H., Hood, L. 1986. **Fluorescence detection in automated DNA sequence analysis.** Nature, 321(12): 674-679.

Lander, E.S., et al., 2001. Initial Sequencing and analysis of the human genome. Nature, 409: 860-921.

Venter, J.C., et al. 2001. **The sequence of the human genome.** Science, 291(5507): 1304-1351.

Chan, E.Y. 2004. **Advances in sequencing technology.** Mutation Research, 573: 13-40.

Mardis, E.R. 2007. **The impact of next-generation sequencing technology on genetics.** Trends in Genetics, 24(3): 133-141.

Schuster, S.C., 2008. **Next Generation Sequencing transforms today's biology**. Nature Methods, 5(1): 16-18.

Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D.H., Johnson, D., Luo, S., McCurdy, S., Foy, M., Ewan, M., Roth, R., George, D., Eletr, S., Albrecht, G., Vermaas, E., Williams, S.R., Moon, K., Burcham, T., Pallas, M., DuBridge, R.B., Kirchner, J., Fearon, K., Mao, J., Corcoran, K., 2000. **Gene expression analysis by massively parallel signature sequencing on microbead arrays**. Nature Biotechnology, 18: 630-634.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., et al., 2005. **Genome sequencing in microfabricated high-density picolitre reactors**. Nature, 437(15): 376-380.

Adessi, C., Matton, G., Ayala, G., Turcatti, G., Mermod, J., Mayer, P., Kawashima, E. 2000. **Solid Phase DNA amplification: characterization of primer attachment and amplification mechanisms**. Nucleic Acids Research, 28(20): 1-8.

Mercier, J., Slater, G.W., 2005. **Solid Phase DNA Amplification: A Brownian dynamics study of crowding effects**. Biophysical Journal, 89: 32-42.

Mercier, J., Slater, G.W., Mayer, P., 2003**. Solid Phase DNA Amplification: A Simple Monte Carlo Lattice Model**. Biophysical Journal, 86: 2075-2086.

Bentley, D.R. *et al.* 2008. **Accurate whole human genome sequencing using reversible terminator chemistry**. Nature, 456: 53-59.

Ronaghi, M., Uhlen, M., Nyren, P**. A Sequencing Method Based on Real-Time Pyrophosphate**. Science, 281: 363-365.

Shendure, J., Porreca, G.J., Reppasi, N., Lin, X., McCutcheon, J.P., Rosenbaum, A.M., Wang, M.D., Zhang, K., Mitra, R.D., Church, G.M. 2005. **Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome**. Science, 309: 1728-1732.

Smith, A.W., Heisler, L.E., St.Onge, R.P., Farias-Hesson, E., Wallace, I.M., Bodeau, J., Harris, A.N., Perry, K.M., Giaever, G., Pourmand, N., Nislow, C., 2010. **Highly-multiplexed barcode sequencing: an efficient method for parallel analysis of pooled samples**. Nucleic Acids Research, 38(13): 2-7.

Sreejalekshmi, K.G., Nair, P. 2010. **Biomimeticity in tissue engineering scaffolds through synthetic peptide modifications-Altering chemistry for enhanced biological response**. Journal of Biomedical Materials Research, 96(2):477-491.

Nettles, D.L., Chihkoti, A., Setton, L.A., 2010., **Applications of elastin-like polypeptides in tissue engineering**. Advanced Drug Delivery Reviews, 62(16): 1479-1486.

Nomura, Y., Sharma, V., Yamamura, A., Yokohayashi, Y. 2011. **Selection of silk-binding peptides by phage display**. Biotechnology Letters.

Murphy, A.R., Kaplan, D.L., 2009. **Biomedical applications of chemically-modified silk fibroin**. Journal of Materials Chemistry, 19: 6443-6450.

Tan, S.H., Hugo, W., Sung, W.K., Ng, S.K., 2006. **A correlated motif approach for finding short linear motifs for protein interaction networks**. BMC Bioinformatics, 7:502-518.

Ramshauer, J.A., Werkmeister, J.A., Glattauer, V., 1996. **Collagen-Based Biomaterials**. Biotechnology Genetic Engineering Reviews, 13: 335-382.

Riesle, J., Hollander, A.P., Langer, R., Freed, L.E., Vunjak-Novakovic, G. 1998. **Collagen in Tissue-Engineered Cartilage: Types, Structure, and Crosslinks**. Journal of Cellular Biochemistry, 71: 313-327.

Glowacki, J., Mizuno, S., 2007. Collagen Scaffolds for Tissue Engineering. Biopolymers, 89(5): 338-344,

Minoura, N., Tsukada, M., Nagura, M. 1990. **Physico-chemical properties of silk fibroin membrane as a biomaterial**. Biomaterials, 11(6): 430-434.

Kim, U., Park, J., Kim, H.J., Wadac, M., Kaplan, D., 2005. **Three-dimensional aqueous-derived biomaterial scaffolds from silk fibroin**. Biomaterials, 26(5): 2775-2785.

Instruction Manual, **ph.D. Phage Display Libraries**, version 1.0.

Manual**, Preparing Samples for Paired-End Sequencing**, Illumina, inc., 2008.

Preda, 2009 Protocol for Preparing Silk from Silkworm (bombyx Mori) cocoons.

Fisher, S., Barry, A., Abreu, J., Minie, B., Nolan, J., Delorey, T.M., Young, G., Fennell, T.J., Allen, A., Ambrogio, L., Berin, A.M., Blumensteil, B., Cibulskis, K., Friedrich, D., Johnson, R., Juhn, F., Reilly, B., Shammas, R., Stalker, J., Sykes, S.M., Thompson, J., Zimmer, A., Zwirko, Z., Gabriel, S., Nicol, R., Nusbaum, C., 2011. **A Scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries.** Genome Biology, 12:R1.

Kircher, M., Stenzel, U., Kelso, J., 2009. **Improved base calling for the Illumina Genome Analyzer using machine learning strategies**. Genome Biology, 10: R83.

Stajich, J.E., Block, D., Boulez, K., Brenner, S.E., Chervitz, S.A., Dagdigian, C., Fuellen, G., Gilbert, J.G.R., Korf, I., Lapp, H., Lehvaslaiho, H., Matsalla, C., Mungall, C.J., Osborne, B.I., Pocock, M.R., Schattner, P., Senger, M., Stein, L.D., Stupka, E., Wilkinson, M.D., Birney, E. 2002**. The Bioperl Toolkit: Perl Modules for the Life Sciences**. Genome Research, 12: 1611-1618.

Peters, B., Sidney, J., Bourne, P.J., Bui, H.H., Buus, S., Doh, G., Fleri, W., Kronenberg, M., Kubo R., Lund, O., Nemazee, D., Ponomarenko, J.V., Sathiamurthy, M., Schoenberger, S., Stewart, S., Surko, P., Way, S., Sette,

A. 2005**. The immune epitope datavase and analysis resource: from vision to blueprint**. PLoS Biology, 3(3): 91.

Bui, HH., Sidney, J., Li, W., Fusseder, N., Sette, A.2007**. Development of an epitope conservancy analysis tool to facilitate the design of epitope-based diagnostics and vaccines**. BMC Bioinformatics 8: 361.

Schneider, T.D., Stephens, R.M. 1990. Sequence Logos: **A New Way to Display Consensus Sequences**. Nucleic Acids Research, 18: 6097-6100.

Crooks, G.E., Hon, G., Chandonia, J.M., Brenner, S.E. 2004. **WebLogo: A Sequence Logo Generator. Genome Research**, 14:1188-1190.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J. 1990. **Basic local alignment search tool. Journal of Molecular Biology**, 215(3): 403-410.

Wegner, G.J., Lee, H.J., Corn, R.M. 2002. **Characterization and Optimization of Peptide Arrays for the Study of Epitope-Antibody Interactions Using Surface Plasmon Resonance Imaging**. Analytical Chemistry, 74: 5161-5168.

Kobe, B., Kajava, A.V., 2001. The leucine-rich repeat as a protein recognition motif. Current Opinion in Structural Biology, 11:725-732.

Mayer, U., Mischt, R., Poschl, E., Mann, K., Fukuda, K., Geri, M., Yameda, Y., Timpl, R., 1993. **A single EGF-like motif of laminin is responsible for high affinity nidogen binding**. The EMBO Journal, 12(5): 1879-1993.

Iwamoto, Y., Nomizu, M., Yamada, Y., Ito, Y., Tanaka, K., Sugioka, Y. 1996. **Inhibition of angiogenesis, tunour growth and experimental metastasis of human fibrosarcoma cells HT1080 by a multimeric form of the laminin sequence Tyr-Ile-Gly-Ser-Arg (YIGSR)**. British Journal of Cancer, 73: 589-595.

Preciado-Patt, L., Levartowsky, D., Prass, M., Hershkoviz, R., Lider, O., Fridkin, M. 1994. **Inhibition of cell adhesion to glycoproteins of the extracellular matrix by peptides corresponding to serum amyloid A.** European Journal of Biochemistry, 223: 35-42.

Huang, J., Ru, B., Dai, P. 2011**. Bioinformatics Resources and Tools for Phage Display.** Molecules, 16: 694-709.