December 2015

Optimizing Acyclic Identification of Aptamers

Caitlin M. Miller
Syracuse University

Abstract

The minimal human alpha-thrombin binding aptamer, d(GGTTGGTGTGGTTGG), has previously been successfully identified from a DNA library of randomized hairpin loop, m=15, with the High Throughput Screening of Aptamers (HTSA) technique, later termed Acyclic Identification of Aptamers (AIA). AIA eliminated the need for multiple cycles of *in vitro* evolution typically used for aptamer discovery by employing libraries with an over-representation of all possible sequences and high throughput sequencing. Although the method was successful at identifying the thrombin binding aptamer, improvements to partitioning and sample preparation inconsistencies encountered during replication attempts were necessary in order to maximize its value. Subsequent revisions and variations of the protocol improved and streamlined the work-flow such that the thrombin binding aptamer (TBA) and multiple variants were identified as high affinity sequences using the ligation based m=15 DNA hairpin loop library mentioned above. The revised AIA protocol was also successful for identifying TBA and multiple variants in a nuclease resistant, modified 2'-OMe RNA/DNA chimera library. Sample throughput was increased with the introduction of indexed adapters containing sequence "barcodes" that facilitate multiplexing during high throughput sequencing. To eliminate the constraints of the hairpin loop library structure, a library based on direct amplification, the "adapter" library was designed. The m=15 DNA and m=15 2'-OMe RNA/DNA chimera adapter libraries were unsuccessful at identifying TBA from the library pool, likely due to interference from the flanking adapter regions required for PCR amplification. A secondary PCR product identified during sample work-up was also characterized. This prompted a significant interest in creating the ability to accurately assess whether or not a sample should be sequenced prior to consuming valuable resources. To accomplish this, the capture of a minimally flanked library

(pACAC-m15-CACA) with full length adapters and no requirement for amplification was optimized. The library provides greater flexibility in secondary structure formation and was shown to successfully identify TBA from an over-represented library pool. Amplification-free AIA introduced the unique ability to predict relative maximum sequence frequencies based on the quantity of recovered library, initial degree of over-representation, and anticipated data output. The ability to predict whether or not a sequence could be counted above background was used to assess whether that sample would be sequenced, ultimately saving time and money. To expand the applicability of the tailed libraries and amplification free protocol, a novel partitioning method that eliminated the requirement for protein immobilization was developed. Reversible formaldehyde cross-linking in conjunction with Electrophoretic Mobility Shift Assay was used to successfully identify TBA above background in proof of concept experiments using amplification free sample preparation. The ability to perform AIA partitioning in solution provides greater flexibility in target selection, including mixtures of proteins. The method also effectively reduces the aptamer off-rate to zero by covalently linking high and moderate affinity sequences to the protein target during selection, an advantage over protein immobilization for partitioning where loss of some DNA from a reversibly bound complex is inevitable. The sample preparation techniques that evolved over the course of this work offer superior control and predictability in the outcome of high throughput sequencing data. This was aided by absolute quantification with qPCR CopyCount^{TM} software that effectively improved library quantification, distribution of indices, and cluster quality, which is crucial for maximization of data output on Illumina sequencing platforms. Consistent, high quality data eliminates the potential for costly resequencing. Future experiments will capitalize on the breadth of improvements to the AIA method described in this work.

OPTIMIZING ACYCLIC IDENTIFICATION OF APTAMERS

By

Caitlin M. Miller

B.S. Seton Hall University, 2009

Dissertation

Submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Chemistry

Syracuse University

December 2015

Acknowledgements

I would like to thank my advisor, Professor Philip N. Borer, for his support, guidance and knowledge. Most importantly, I am thankful for the freedom and encouragement he provided, allowing me to become a more independent and thoughtful scientist.

I would like to thank those whose efforts contributed to the success of this work, including Dr. Gillian Kupakuwana and Dr. Lei Chen, for their many years of research on the original high throughput screening projects. I would like to thank Dr. Mark McPike for his collaborative efforts and most helpful knowledge, and James Crill for his insightful advice and friendship. Profound thanks are given to Dr. Damian Allis for assistance in data analysis. His knowledge, effort and patience contributed immensely to the success of this project. Also, members of the Borer group, including Dr. Deborah Kerwood, Dr. Wei Ouyang, Dr. Collin Fischer, Raghuvaran Iyer and Nan Thuzar Myint.

I am also grateful to Dr. Tom Duncan of SUNY Upstate for use of the ForteBio Octet RED96 system and Vicki Lyle of the SUNY Upstate DNA sequencing facility. Also, Dr. Frank Middleton of the SUNY Upstate Molecular Analysis Core Facility for use of the MiSeq and Karen Gentile for running the experiments. Thank you to Dr. Liviu Movileanu for serving as the chair of my dissertation committee and to Dr. Joseph Chaiken, Dr. Robert Doyle, Dr. Bruce Hudson and Dr. James Hougland for serving on my dissertation committee.

I am especially grateful to the Graduate School for the many opportunities afforded to me through the Future Professoriate Program, Women in Science and Engineering Future Professoriate Program, and Teaching Mentor program. I am most grateful for the opportunity to engage my passion for science education while working on this project.

I would like to thank my family and friends for their love and confidence over the years, especially my husband, Sean, for his encouragement, humor, lighthearted sarcasm and patience while I continued my education. Finally, my daughter, Leah, who has inspired me in ways I could never have imagined.

For my daughter, Leah.

Table of Contents

List of Tables and Figures

Chapter 1: Introduction

Aptamers are functional single-stranded oligonucleotides capable of folding into three-dimensional structures that bind their targets with high specificity.[1] These structures provide binding locations specific to molecular targets, including metal ions, chemicals, small molecules, proteins, cell surfaces, and parasites,[1] with applications as biosensors,[1-10] therapeutic agents[1, 11-18], and diagnostic tools.[1, 19-25] Aptamers closely mimic the activity of antibodies and can have similarly high affinities for their targets.[1, 20, 26-28] There are potential applications in all fields currently dominated by antibodies. There are many advantages to utilizing aptamers; unlike antibodies, aptamers are identified in a process that does not involve live tissue, so conditions can be manipulated to suit the target environment.[28] The use of toxic substances or those that elicit a negative immune response is not restricted during the discovery, isolation, and generation of aptamers.[28] Aptamers can be produced at 1-10% of the cost of antibodies using scalable solid-phase DNA or RNA synthesis, a favorable characteristic for commercial applications. Additionally, incorporating modified oligonucleotides, including 2'-fluoro RNA[29], 2'-O-Methyl RNA[30] and L-oligonucleotides[31, 32] limits degradation *in vivo* due to substrate specificity of nucleases. With such a broad range of applications and benefits, the development of a fast, efficient, and widely applicable aptamer discovery protocol is highly desirable. The Acyclic Identification of Aptamers (AIA) method was designed and implemented with those goals in mind and was used successfully to rapidly screen and identify aptamers in a single round of selection.[33] The goal of this work was to improve upon AIA by utilizing varied oligonucleotide library structures and screening techniques, with the intent of streamlining the approach as a universal aptamer discovery method.

The Origin of Aptamers

Many short RNA molecules are known to have functional roles *in vivo*. Viral RNAs were of early interest to those studying functional RNAs, as they bind their targets with high affinity and specificity.[34] In human immunodeficiency virus (HIV), a short mRNA named the trans-activation response element (TAR) binds the viral Tat protein to promote viral replication.[35, 36] Additionally, RNA stem loop 3 (SL3), which is absent in spliced mRNA encoded by the HIV provirus, interacts with the HIV nucleocapsid protein 7 (NcP7) and plays a vital role in viral packaging.[37] The function of an adenovirus mRNA transcript termed virus-associated RNA (VA RNA) that inhibits translation was also extensively studied.[38] The continued study of these naturally occurring functional RNAs led to the development of a method to discover functional nucleic acids artificially. The Systematic Evolution of Ligands by Exponential Enrichment (SELEX) technique, also known as *in vitro* selection, was independently developed by two separate groups led by Larry Gold and Jack Szotstak.[39, 40] The method utilizes oligonucleotide libraries, typically of length 30-80 nucleotides, and capitalizes on the ability of single-stranded oligos to fold into varied secondary structures. These stable conformations include hairpins, internal loops and bulges, multi-branched loops, pseudoknots and G-quadruplexes which create highly specific binding sites for their targets (**Figure 1.1**).[41] These highly specific oligonucleotides are called aptamers, a term credited to Andrew Ellington, who worked alongside Szostak. The term comes from the Latin “aptus” meaning “to fit.”[40]

The SELEX Protocol

Library Design

SELEX has been the standard method for aptamer discovery since its development in 1990. The method begins with a randomized RNA or DNA library of a specific length, m, and the notion that a fraction of these sequences will fold in such a way to yield a binding site specific to the target of interest. The goal is to isolate and identify these sequences through multiple rounds of selection, purification, and amplification. There are four main considerations when selecting a library for SELEX, (1) the chemistry of the library, (2) library length, (3) type of randomization and (4) length and composition of fixed regions. Both DNA and RNA libraries are used in SELEX, each with distinct advantages and disadvantages. DNA libraries are very stable in nuclease-free environments, while RNA libraries are susceptible to hydrolysis due to the 2'-hydroxyl group of the pentose sugar. However, the 2'-hydroxyl group increases the hydrogen bonding potential of RNA, giving it a broader range of secondary and/or tertiary structure formations than a DNA molecule of the same sequence. This may increase the likelihood of discovering an aptamer, as the secondary and/or tertiary structure is critical to creating a binding site for the target. Modified nucleic acids may also be used to provide nuclease resistance, such as 2'-fluoro RNA, 2'-amino RNA, and 2'-O-Methyl RNA. The chemistry and application of these modifications will be discussed in the next section.

The length of the DNA or RNA library used in SELEX is highly varied. If there is no known nucleic acid binding activity for the chosen target, multiple library lengths may need to be explored, as there are benefits to both short and long libraries. For traditional SELEX experiments, library lengths of 30 to 60 nucleotides are most common.

These shorter libraries allow for structural motifs to be easily identified and aligned in sequence data.[42] Libraries as short as 22 and 26 nucleotides have been used to successfully identify an aptamer for isoleucine.[43, 44] Additionally, libraries of 25 nucleotides have been used successfully to identify aptamers for arginine [45] and proteins [46-48]. However, longer libraries may be necessary as they allow for greater structural complexity. In the case of naturally occurring 76-mer tRNA molecules, a library of 80 nucleotides would have a much higher success rate at identifying any tRNA mimics than a shorter library.[42] Longer libraries of 120 [49-53] and 228 [54] nucleotides have been used to successfully identify aptamers. It is possible to identify shorter motifs within long libraries; however, the folding of the motif may be influenced by the context of the remaining library sequence. Truncation of the library to the shorter motif may improve or deplete the function. Additionally, because shorter motifs naturally occur at a higher frequency within the library, selection of these shorter motifs may dominate over a sparsely represented longer binding motif.[55] Conversely, the selection of shorter binding motifs within a long library may be inhibited by the folding of the remaining library sequence.

The diversity of a library used in a SELEX experiment, and consequently the diversity of any shorter n-mer, is limited by both the library length and quantity of library used. DNA libraries are synthesized by solid-phase synthesis, thus the frequency of any unique sequence is ultimately limited by the library length and synthesis scale. **Table 1.1** outlines the combinatorial considerations to be made when selecting a library. The number of possible sequences for a library of length, m, is 4^m. The number of possible sequences increases exponentially as library length increases and the number of unique sequences occurring in a library pool is dependent on

the synthesis scale. For a typical synthesis scale of 1 µmole, approximately $6.02x10^{17}$ library molecules will be synthesized. In order for every possible sequence to be present at least one time and assuming a 100% yield, the library can be no longer than 28 nucleotides. Additionally, only a fraction of the 1 µmole synthesis would be used during a partitioning step. For these reasons, the libraries used in SELEX are always sparsely sampled. In fact, most sequences will not be present at all and any present sequences will occur only once. For example, a 40-mer library contains approximately $1.2x10^{30}$ possible sequences. A 1 µmole scale synthesis is capable of producing only $5.0x10^{-5}$ % of these sequences. In other words, only 1 in every $2.0x10^{6}$ unique sequences is synthesized. A common argument for choosing a longer library in SELEX is the notion that all possible shorter n-mers will exist within the longer library.[42] For example, all possible 30-mers will exist within a 40-mer library. This is true when considering the entire $1.2x10^{30}$ unique sequences in a 40-mer library, however, the entire population cannot be used in a SELEX experiment. If a 100 pmol aliquot of a 40-mer library is used during selection, approximately $6.0x10^{13}$ molecules are used. A 30-mer library contains $1.15x10^{18}$ unique sequences. It is impossible for all of the 30-mers to be present in the 100 pmol aliquot. For much shorter n-mers, the presence of all possible sequences is dependent on the n-mer length and quantity of library. Additionally, the secondary or tertiary structure of the 30-mer will be influenced by the remaining library sequence, which may cause it to behave differently than the same sequence from a 30-mer library. Even if a given 30-mer is present multiple times within the aliquot of 40-mer library, the remaining randomized region will not be identical. This is complicated further by the fact that an identical 1 nmol aliquot of 40-mer library cannot be replicated due to the sparsely represented nature.

The degree and type of randomization in a library can also be varied. In addition to complete randomization, libraries with partial or segmental randomization have been used. Segmental randomization involves complete randomization of a short region or regions within a known aptamer sequence. This method can aid in the discovery of higher affinity binding sequences while illustrating the role of specific nucleotides within the aptamer sequence.[42] Partial randomization, also known as doping, involves the introduction of point mutations into a known sequence. Doping at one or more locations within a selected sequence and reselecting against the target may aid in the identification of critical nucleotides and their role in target binding.[42, 56]

Due to the sparsely represented nature of SELEX libraries, multiple rounds of selection, purification, and amplification are required in order to isolate any high affinity sequences from the starting library pool. To facilitate amplification, SELEX libraries are flanked by fixed, non-complementary primer regions. For RNA libraries, the library is synthesized as DNA and the 5'-primer region is used to facilitate RNA transcription, given it contains a T7 RNA polymerase promoter sequence. Following selection, RT-PCR regenerates and amplifies the DNA library and the cycle can be repeated. The composition of the flanking regions, often called primer or priming regions, should be designed carefully. Ideally, the sequences will be non-complementary and lack any significant secondary structure. The corresponding primers used for amplification should anneal with high efficiency, have suitable melting temperatures, and not form primer dimers.[42] Efficient and specific amplification of the library molecules is critical for regenerating a library used in SELEX.

In Vitro Selection

The SELEX method, as illustrated in **Figure 1.2**, includes multiple cycles of selection, purification, and amplification. If a DNA library is desired, an aliquot of the synthesized library is typically enriched by PCR amplification in order to create multiple copies of each unique sequence in the starting pool. A commonly used method for regeneration of single stranded DNA is the use of a modified primer, such as 5'-biotinylated primer, followed by separation on streptavidin beads.[54, 57-59] The strands may also be separated by electrophoresis.[60] If an RNA library is desired, the DNA library is simply transcribed into RNA; the DNA library may or may not be amplified prior to this step.[40, 61]

During the selection step, the target is usually immobilized in order to separate bound from un-bound library molecules. This is typically achieved using affinity chromatography methods, including nitrocellulose filters, or modified sepharose or magnetic beads.[62, 63] The conditions for the selection step, including target/library concentrations, buffer composition, temperature, and incubation time all affect the selection stringency. The ionic strength of the buffer is an important consideration, as the shielding effect of solvent ions will naturally influence electrostatic interactions between negatively charged DNA and a target with charged regions. Also, solvent cations stabilize oligonucleotide secondary structures.[64, 65] Unbound library molecules are removed by washing with multiple rounds of selection buffer. Bound library molecules are typically collected following elution with a chemical denaturant, such as Guanidine HCl or Urea.[20, 66] Phenol may be used to co-elute protein and library directly from an affinity matrix[67] or to extract the library from complexes co-eluted by another method. Alternatively, competing ligands may be used to elute protein-library complexes[57] and

adjusting ionic strength has been successfully used to elute the library. [46] Ethanol precipitation is most commonly employed to purify and concentrate the eluted or extracted library.

Following selection, the library pool may now contain sequences with affinity for the target, along with some percentage of nonspecific binding sequences. The purified library is amplified in order to create multiple copies of the surviving sequences for use in the next cycle of selection. RNA libraries are subjected to reverse-transcription PCR and subsequent transcription to regenerate the single-stranded RNA library. Single-stranded DNA is regenerated as mentioned previously. As the cycle of selection, purification, and amplification is repeated, typically 5-15 times, the frequency increases for any high affinity sequences originally present in the library pool, or ones that evolved through the amplification process. In order to identify these sequences, SELEX traditionally employs Sanger sequencing of a few dozen to a few hundred cloned sequences. The SELEX method is a labor intensive and time consuming process with minimal data out. Many efforts to improve the method have been published, including the AIA method from which this work was developed.

Improving the SELEX Method

Modified SELEX

The goal of many modified SELEX methods is to simultaneously improve selection efficiency while reducing labor and time demands. One such method, Capillary Electrophoresis SELEX (CE-SELEX), has successfully identified aptamers while reducing the number of selection cycles. The method utilizes the electrophoretic mobility shift of target-library complexes to

separate them from unbound library and target. Carryover of nonspecific library molecules is significantly reduced in the method, allowing the number of necessary selection cycles to be reduced. The CE-SELEX method has successfully identified aptamers in as few as 1 or 2 selection rounds in a matter of days,[68-75] compared to weeks to months for traditional SELEX. One potential drawback of CE-SELEX is the poor separation of small target molecules from target-library complexes, especially when the target is smaller than the individual library members.[76] The method is also limited by the type of target; the target must shift the electrophoretic mobility of the bound oligonucleotide and the application does not work for whole cells.[77] Additional methods aimed at reducing the number of selection rounds include Non-SELEX, M-SELEX, and I-SELEX. Non-SELEX utilizes multiple rounds of capillary electrophoresis but forgoes any amplification steps.[78-80] M-SELEX, or microfluidic SELEX, utilizes chip-based microfluidics and magnetic beads to achieve efficient separation of target bound libraries via magnetic force in combination with high pressure wash steps.[81-84] I-SELEX, or inertial microfluidic SELEX, utilizes curvilinear microfluidic channels to capitalize on the effect of centrifugal acceleration on hydrodynamic forces. Smaller particles travel to the outer wall of the channel, while larger particles travel toward the inner wall. Target bound oligonucleotides are collected while unbound oligonucleotides are diverted to a separate outlet. The method was applied to whole cell SELEX and successfully identified aptamers for surface proteins of malaria parasite-infected red blood cells.[85]

Automating the SELEX process with robotic workstations was also explored. The goal of automated SELEX was to streamline workflow without necessarily reducing the number of selection cycles. The first SELEX workstation included a pipetting robot, thermocycler, and

magnetic bead separator.[86] Several modifications and improvements to the workstation allowed it to perform as many as 12 rounds of selection in only 42 hours.[87] Several aptamers were discovered using automated SELEX,[86, 88-91] and the workstation was even used to perform *in vitro* transcription and translation of the protein target.[92] Semi-automated SELEX utilizing robotic magnetic separation and solid-phase emulsion-PCR has also been successful at identifying aptamers, with 12 selection cycles completed in only 10 days.[93]

An alternative selection method to the traditional affinity chromatography or bead-based methods is the Electrophoretic Mobility Shift Assay (EMSA). Similarly to CE-SELEX, EMSA-SELEX capitalizes on the electrophoretic shift of target bound oligonucleotides. Unbound oligonucleotides travel through the gel faster than target bound oligonucleotides. The method offers the distinct advantage of visualizing the ratio of bound to unbound library, provided autoradiography is utilized when working in minute concentrations. The stringency of selection can then be visualized over a range of selection cycles; the ratio of shifted library increases as the number of selection cycles increases. It can be assumed that high affinity sequences have been selected when the ratio does not increase with additional cycles. The stringency of selection can also be increased simply by decreasing the amount of protein, thus increasing competition for binding sites.[94] The conditions of EMSA-SELEX are limited due to compatibility with gel electrophoresis. Sample volume constraints affect library and target concentration, ultimately limiting the diversity of long SELEX libraries.[77] However, EMSA as a partitioning method was successful at identifying aptamers for Roaz, a rat zinc finger protein, after 9 rounds of selection.[95] In addition to PAGE-EMSA,[95, 96] agarose-EMSA has been used successfully for *in vitro* selection[97].

Capture-SELEX offers yet another unique method of selection. Traditional SELEX is not well suited for the selection of many small organic molecules, as it requires their immobilization to a solid support. Capture-SELEX was designed to immobilize the oligonucleotide library instead of the target during selection. The library was synthesized with a "docking sequence" that is complementary to an immobilized oligonucleotide on a magnetic bead surface. The small molecule target is allowed to bind to the immobilized library in solution; any sequences that show affinity for the target undergo a conformational change for binding, causing them to be released from the beads. The high affinity sequences are then collected from the supernatant. It is possible for some sequences to undergoing a conformational charge that does not include the docking sequence, in which case those sequences are not released from the beads. Aptamers for kanamycin-C were identified after 13 rounds of Capture-SELEX.[98] The method was also used in conjunction with Surface Plasmon Resonance to identify aptamers for tobramycin, an aminoglycoside antibiotic.[99]

Modified Libraries

The composition and functionality of SELEX libraries has been modified is various ways in order to improve selection efficiency, including the introduction of modified nucleotides for enhanced capture, modified nucleic acids offering increased nuclease resistance, or reducing or eliminating fixed flanking regions. Introducing modified nucleotides capable of covalently cross-linking with proteins has been shown to improve selection efficiency. PhotoSELEX, or photochemical SELEX, utilizes photocross-linking of modified oligonucleotides to the aromatic or sulfur-bearing amino acid residues of the protein target. UV-induced covalent bonds between a 5-bromouracil substituted DNA library and recombinant human basic fibroblast growth factor

resulted in the selection of high affinity aptamer sequences.[100] Photocross-linking of a 5'-iodouracil substituted RNA library was used in a dual selection method to identify sequences with high affinity for Rev, a protein from HIV-1.[101] In an effort to improve capture efficiency by reducing the off-rates of aptamer sequences, SomaLogic, Inc. developed SOMAmers (Slow Off-Rate Modified Aptamers). These nucleotides capitalize on the notion that aptamers may exhibit protein-like properties if they contain functional groups similar to amino acid side chains.[102] With only four possible nucleotides, the diversity of a DNA or RNA library depends almost entirely on the length and sequence complexity of the library. Expanding the chemical diversity of the oligonucleotide side chains to include 5'-benzyl, 5'-naphthyl, 5'-tryptamino and 5'-isobutyl groups increases the binding potential of the library.[20, 103]

Introducing nuclease resistance is a popular modification to the SELEX protocol. Retaining high affinity sequences throughout the selection process is critical for efficient aptamer discovery. While DNA libraries are easy to work with *in vitro* and are relatively stable, RNA libraries are much more susceptible to hydrolysis by small amounts of contaminating nucleases. However, both DNA and RNA aptamers are susceptible to hydrolysis *in vivo*, where the presence of nucleases cannot be controlled, thus limiting their applications.

One of the first methods for introducing nuclease resistance was the use of L-nucleotides in place of the naturally occurring D-nucleotides.[31] Libraries synthesized with L-nucleotides, also known as spiegelmers, are not recognized by nucleases due to substrate specificity. In order to select L-aptamers, the D-nucleotide library is partitioned against a mirror image protein containing D-amino acids. Once the D-aptamer sequence is identified, the L-aptamer is

synthesized. The L-aptamers have high affinity for the naturally occurring proteins containing L-amino acids, while remaining nuclease resistant. Several spiegelmers for various small protein targets have been identified [31, 32, 104-106] and applications in therapeutics are promising due to their bio-stability and immunologic passivity.[107, 108] A limitation of the method, of course, is that a mirror image protein cannot be expressed in cells; however the chemical synthesis of many interesting bio-active peptide targets from D-amino acids is simple and economically feasible.

Chemically modified nucleotides such as 2'-fluoro, 2'-amino, and 2'-O-methyl (2'-OMe) also offer nuclease resistance in addition to preventing hydrolysis of RNAs, although incorporation of these modified nucleic acids into the SELEX protocol has its challenges. The structures of these three types of modifications are shown in **Figure 1.3**. Modified nucleotides may be used both during the selection process as modified libraries and post-SELEX as modified aptamers. After identifying an aptamer by SELEX, a nuclease resistant variant containing modified nucleotides can be chemically synthesized. The type and degree of modification are important considerations for *in vivo* applications. 2'-fluoro and 2'-amino modifiers are more costly than 2'-OMe modifiers and pose a greater safety concern. Both 2'-fluoro and 2'-amino nucleotides are not naturally occurring within the human body[30, 41, 109] and they could potentially be incorporated into host DNA with unknown toxicological effects.[110] The introduction of synthetic 2'-OMe nucleotides into biological systems is not a safety concern. The nucleotides exist in the human body as a result of posttranscriptional modification of ribosomal RNA,[111] yet 2'-OMe nucleoside triphosphates are not accepted by human DNA polymerase as substrates.[30, 109, 110] This unique property of 2'-OMe RNA nucleotides eliminates the safety concerns of using

2’-OMe RNA aptamers in therapeutic applications. The effects of 2’F, 2’NH_2, and 2’-OMe nucleotides on aptamer secondary structure must also be considered. Modifications to an aptamer sequence found through SELEX may disrupt the secondary structure, the binding properties and/or inhibitory or catalytic function of the aptamer.[112, 113] In some instances, function is altered to varying degrees depending on which nucleotides are substituted.[113] Partial substitution often allows for retention of binding/function while imparting some degree of nuclease resistance. [113-115]

The issue of altering aptamer structure and/or function by incorporating modified nucleotides can be avoided by utilizing modified libraries during the selection process. In SELEX, the modified library must be generated enzymatically prior to each selection cycle. This poses a challenge because T7 RNA polymerase and DNA polymerase incorporate modified nucleotides with very poor efficiency. A modified library prepared by chemical synthesis can be read efficiently by reverse transcriptase, leaving only the amplification of modified DNA and/or regeneration of modified RNA as the challenge. Efforts to develop mutant T7 RNA polymerases that are capable of incorporating 2’-modified nucleotides have had success with incorporating 2’F and 2’NH_2 nucleotides while retaining the ability to incorporate 2’-deoxy nucleotides.[116, 117] The incorporation of 2’-OMe nucleotides was very inefficient, although enhanced with the introduction of a double mutant T7 polymerase.[117, 118] Burmeister *et al*. were able to carefully optimize transcription conditions utilizing the double mutant T7 RNA polymerase enough to discover aptamers for vascular endothelial growth factor, interleukin 23, and thrombin with high specificity and nuclease resistance.[30, 109] The use of 2’F libraries in SELEX has

seen greater success than 2'NH_2 and 2'-OMe, specifically for libraries comprised of a mixture of 2'-fluoro-pyrimidine and standard purine nucleotides.[119-123]

Altering the structure of the library, specifically the length of the fixed flanking regions, has potential for improving selection efficiency in SELEX. In most applications the two fixed regions required for amplification and regeneration of the library, and cloning for Sanger sequencing are about 20 nucleotides each. This contributes significantly to the overall length of the library. In many instances, the combined fixed regions are longer than the actual library. This brings possibilities for nonspecific binding to the target, inclusion of parts of the fixed region in folding the core binding domains of selected aptamers, and interference between highly abundant fixed sequences with low to moderate affinity and the desirable high affinity sequences that are present in low abundance. Reducing or eliminating the primer regions (complementary to the PCR primers) reduces interactions between the PCR primers and the library region, as well as between the fixed sequences and the target during partitioning.[124-126] Although it has been argued that the fixed regions do not significantly contribute to or reduce binding potential of the library,[127] primer sequences do often contribute to an aptamer's binding sequence.[128] Additionally, genomic SELEX may benefit significantly from minimal primer or primer free libraries, as fixed regions often account for the majority of selected sequences.[129] A 2009 review discusses various methods for reducing the fixed-sequence regions of libraries for aptamer selection, including minimal primer and primer-free SELEX.[125] Minimal primer SELEX utilizes two nucleotide flanking regions on the 5' and 3'-termini of the library, and primer free SELEX utilizes 0 or 2 nucleotides on the 3'-terminus only. The minimal primer method was developed with an m=27 DNA library with fixed primer regions on each side (18

and 19 nucleotides). The 64-mer was PCR amplified to yield the dsDNA product. Restriction enzyme digestion with *Nt.*BstNBI/NotI generated the library with two nucleotide flanking regions on the 5' and 3'-termini. The digested product was analyzed on denaturing PAGE and the library was extracted and purified. Following partitioning against a melanoma cell line, surviving library molecules were recovered and new primer regions ligated. The library was PCR amplified again to yield the dsDNA library fit for a second round of digestion and partitioning.[124] The primer free method works similarly, with varied restriction enzymes to cleave the library at different locations. Eliminating interference from flanking regions with minimal primer and primer free SELEX may improve selection efficiency. Carryover of nonspecific binding sequences would be reduced and any high affinity sequences in the library pool will be constrained minimally.

Aptamer Applications

As a result of the many variations of SELEX, aptamers for over 3000 targets have been discovered. Applications in biosensors, diagnostics and therapeutics have been explored for a number of these aptamers. Most notable is the first aptamer-based drug, Macugen, which was approved by the U.S. Food and Drug Administration in 2004 for the treatment of age-related macular degeneration.[12, 130] The drug was developed from a 27-mer RNA aptamer for anti-vascular endothelial growth factor (VEGF). The aptamer was discovered from a 2'-fluoro-pyrimidine modified RNA library and the final aptamer was substituted with 2'-O-methyl-purines to further improve the aptamers functionality *in vivo*, and was further modified with a terminal polyethylene glycol (PEG).[120] Elevated VEGF contributes to macular degeneration by inducing angiogenesis and increasing vascular permeability and inflammation. The

PEGylated aptamer effectively inhibits VEGF, slowing the progression of the disease.[130] Aptamer therapeutics that have reached Phase III clinical trials include Revolixys (or pegnivacogin), an anticoagulation system utilizing an aptamer for factor IXa[103, 123, 131-133] and Fovista, a drug utilizing an aptamer for anti-platelet-derived growth factor-B for the treatment of age-related macular degeneration.[103, 134] Other aptamer therapeutics include, but are not limited to, an aptamer inhibitor of von Willebrand factor which plays an important role in platelet coagulation,[135-138] and an aptamer inhibitor of nucleolin which was shown to have anti-proliferative effects on advanced carcinomas.[139]

Aptamers have great potential as diagnostic tools, with applications including biosensing, imaging, and disease cell and biomarker discovery (see the beginning of this introduction for references). Aptamers are highly specific to the targets, offering a high level of confidence for their use as diagnostics. One such aptamer for theophylline binds with 10,000 times greater affinity than it does for caffeine, a molecule that differs by only a single methyl group.[140] Aptamers have been shown also to bind with high sensitivity; aptamer-modified quantum dots were shown to detect a mere 7 zeptomoles of C-reactive protein in 150µl of spiked human serum using surface plasmon resonance imaging.[141] An aptamer for tenascin-C, an extracellular matrix protein upregulated by a number of cancers, has been used to successfully image tumor cells utilizing ^{99m}Tc radiolabeled nucleotides.[60, 142] In addition to differentiating between healthy and cancer cells, one study demonstrated that it is possible to differentiate between various cancer stages. 44 potential biomarkers were selected from 813 total proteins measured; from these, a panel of 12 blood proteins were identified and used to distinguish between stages I-III of non-small cell lung cancer.[143] The biomarkers were identified using SOMAscan

technology, a highly selective, multiplexed assay that quantifies proteins by quantifying SOMAmers following selection in solution. In recent publications this technology was capable of measuring 1,129 proteins and has been used to discover biomarkers for many other diseases, including mesothelioma,[144] Alzheimer's[145] and chronic kidney disease.[20] Although aptamers offer great potential as diagnostics, clinical or other market applications are still extremely limited compared to the prevalent use of antibodies. With applications as antitoxins, antivenins, as well as bacterial, viral and cancer treatments, the continued study of aptamer discovery and functionality is highly significant. [146]

While modified bases offer attractive features from a chemical point of view, there are other important issues, especially for aptamer therapeutics. Solid-phase DNA synthesis is such a large business that the monomers are cheap enough that DNA has a substantial cost advantage over all other formulations. This advantage makes it likely that an effective DNA aptamer would be preferred over all of the other possibilities for applications where a great deal of material could be needed, for instance to replace monoclonal antibodies and other "biologicals" in pharmaceutical applications. Monomer precursors for current generation SOMAmers are not even commercially available and will never come close to the economies of scale for simple therapeutics based on DNA or even RNA aptamers. Another issue for aptamer therapeutics is clearance – a biological half-life should be neither too long nor too short. Terminal modification of DNA aptamers along with nanoparticle delivery schemes should give half-lives in the range of hours, which is a typical time-scale for many current drugs. On the other hand, few natural mechanisms exist to degrade spiegelmers and other unusual modified nucleic acids, which could lead to toxicity, off-target effects, and regulatory challenges.

Acyclic Identification of Aptamers

Continued efforts to increase selection efficiency, reduce time and labor demands, and produce stable, functional aptamers will likely spur increased interest in aptamer based technologies from an industry currently dominated by antibodies. One such effort is the Acyclic Identification of Aptamers (AIA), previously named High Throughput Screening of Aptamers (HTSA), originally developed by the Borer lab.[33] The method addresses the bottleneck of multiple cycles of *in vitro evolution*. AIA utilizes libraries with an over-representation of all possible sequences and was used to successfully identify the thrombin binding aptamer (TBA) after a single round of selection and subsequent high throughput sequencing. A schematic representation of the AIA method can be found in **Figure 1.4.** High throughput sequencing (HTS) has been an increasingly popular addition to the SELEX protocol,[83, 99, 147] offering millions of sequence reads in comparison to hundreds from traditional Sanger sequencing. Sanger sequencing requires that the library pool converges to a relatively small number of sequences in order to identify them in the limited sequence space. When HTS is used intermittently throughout the selection process, the breadth of sequence data may allow for fewer cycles because the evolution and frequency of aptamer candidates can be monitored. In combination with HTS, the AIA approach offers a fast, efficient and high throughput method for aptamer discovery. A comparison highlighting the time and labor commitment of AIA versus traditional SELEX can be found in **Figure 1.5.**

Synopsis of Chapters

The work presented in this thesis builds upon the original AIA method by improving sample preparation techniques, introducing a 2'-OMe RNA/DNA chimera library, exploring a variety of library structures that exhibit over-representation of sequences, and implementing a novel

partitioning method with broad applications. In Chapter 2, experiments were aimed at consistently replicating the results of the original, published AIA results, a necessary precursor to optimize and expand the AIA method. Additionally, experiments aimed at observing the effects of target to library ratio on selection efficiency were explored. Inconsistencies in sample preparation and partitioning efficiency were encountered that produced lower than expected frequencies for the canonical thrombin binding aptamer and poor data quality. In chapter 3, a series of troubleshooting experiments and procedural modifications achieved consistency in sample preparation. The improved AIA method successfully identified the canonical thrombin binding aptamer at sufficient frequency above background with improved high throughput sequencing data quality. To illustrate the applicability of 2'-OMe libraries in AIA, a 2'-OMe RNA/DNA chimera library was also used to successfully identify the 2'-OMe RNA analog of the canonical thrombin binding aptamer at high frequency above background. Attempts to adapt the partitioning method to discover novel aptamers for four epigenetic protein targets using hairpin loop libraries of various lengths was not successful. Although the hairpin loop library structure used in these experiments was ideal for selecting the canonical thrombin binding aptamer, it was possible that the hairpin structure was inhibiting aptamer discovery for these protein targets. A requirement for the success of AIA is applicability to a broad range of protein targets and this prompted the introduction of a different library structure that eliminated the hairpin loop conformation. Index adapters were also introduced in this chapter; sequence "barcodes" built into the adapter constructs used in ligating the partitioned library allowed for multiplexing multiple experiments during high throughput sequencing, ultimately increasing sample throughput.

In chapter 4, methods for capture, amplification and sequencing of the "adapter library" were optimized. In addition to eliminating the hairpin loop structure, the adapter library did not require ligation and was captured by PCR amplification. The goal of capturing the library with amplification was to reduce sample loss and shorten the AIA protocol further. However, during methods development, it was found that the variable quantities of library recovered during partitioning complicated the amplification methods. After the initial exponential phase of amplification, a secondary product that appeared larger during gel electrophoresis was formed. This product was identified as heteroduplex DNA and it was found that sequencing heteroduplex DNA resulted in the correct sequencing cassette but with reduced data quality. DNA and 2'-OMe RNA/DNA chimera adapter libraries were screened against thrombin to gauge their functionality in AIA. Neither library was successful at identifying TBA from the library pool. It is possible that the adapter constructs that flanked the variable region were interfering with aptamer selection. To remedy this, methods to capture un-flanked and minimally flanked libraries that would allow complete or nearly complete freedom in secondary structure formation were explored as described in chapter 5. Although the m=15 adapter library failed to identify TBA from the starting pool of library, many of the techniques mastered in this chapter allowed significant progress to be made in chapters 5 and 6 with increased confidence during sample preparation. The experiments also prompted a significant interest in creating the ability to accurately assess whether or not a sample was of sufficient quantity and quality for productive high throughput sequencing prior to consuming valuable resources. This was successfully accomplished in chapter 5.

Chapter 5 outlines experiments aimed at capturing un-flanked libraries with various versions of ligation. Efficient capture of un-flanked libraries was not achieved which prompted the introduction of a minimally flanked library. The library is flanked by two, four-base non-complementary tails that do not constrain the variable region to a hairpin loop, but can be captured by ligation similarly to the hairpin loop library. A benefit to revisiting the ligation-based capture approach is the ability to eliminate the amplification step required with the adapter libraries. Ligating full length adapters to the "tailed" library and proceeding directly to sequencing shortened the AIA protocol and eliminated a potential source of error. Without amplification, sequence data would accurately represent the selected library sequences with zero amplification bias from sample preparation. Amplification free AIA introduced the unique ability to predict relative maximum sequence frequencies based the quantity of recovered library, initial degree of over-representation and anticipated data cluster density. The ability to predict whether or not a sequence could be counted above background was used to assess whether that sample would be sequenced, ultimately saving time and money. Once capture efficiency was optimized and an experiment produced the desired quantity of recovered library, the m=15 DNA tailed library was shown to successfully identify the canonical thrombin binding aptamer above background with frequencies similar to the prior expectations. To expand the applicability of the tailed libraries and amplification free protocol, a novel partitioning method that eliminated the requirement of protein immobilization was developed.

In chapter 6, reversible formaldehyde cross-linking in conjunction with electrophoretic mobility shift assay was developed as a partitioning method. Reversible formaldehyde cross-linking produces covalent links between protein-bound library molecules in solution. In combination

with EMSA under denaturing conditions, this method allows for the separation of high and moderate affinity sequences from low affinity and nonspecific sequences. This method was used to successfully identify the thrombin binding aptamer above background in proof of concept experiments using amplification free sample preparation. The ability to perform AIA partitioning in solution provides greater flexibility in target selection, including mixtures of proteins. These experiments also confirmed the ability to predict maximum sequence frequencies based on the quantity of recovered library and initial level of over-representation. Additionally, the quality of sequencing data was improved with the introduction of qPCR CopyCount™ software. qPCR CopyCount™ eliminates the requirement of a standard curve, which reduces cost and labor demands. It was found that qPCR CopyCount™ provided a more accurate determination of library quantity than the gold-standard KAPA Library Quantification kit; this is crucial for maximization of data output on Illumina sequencing platforms. Consistent, high quality data eliminates the potential for costly resequencing.

The sample preparation techniques that evolved over the course of this work offer superior control and predictability in the outcome of high throughput sequencing data. Chapter 7 outlines future goals for this work including (1) optimization of the ForteBio Octet RED96 interaction analysis system as a partitioning method. (2) Experiments exploring the application of the principles and techniques of AIA to a minimal primer method, that would incorporate moderate cycling and aggressive mutagenesis, look especially promising for the discovery of novel aptamers. (3) Application of the improved AIA methodology to additional protein targets to discover novel aptamers with DNA, 2'-OMe and 2'F libraries. (4) Transition of the GAIIx

sequencing platform to perform post-sequencing protein binding studies as a secondary screening method for aptamer characterization.

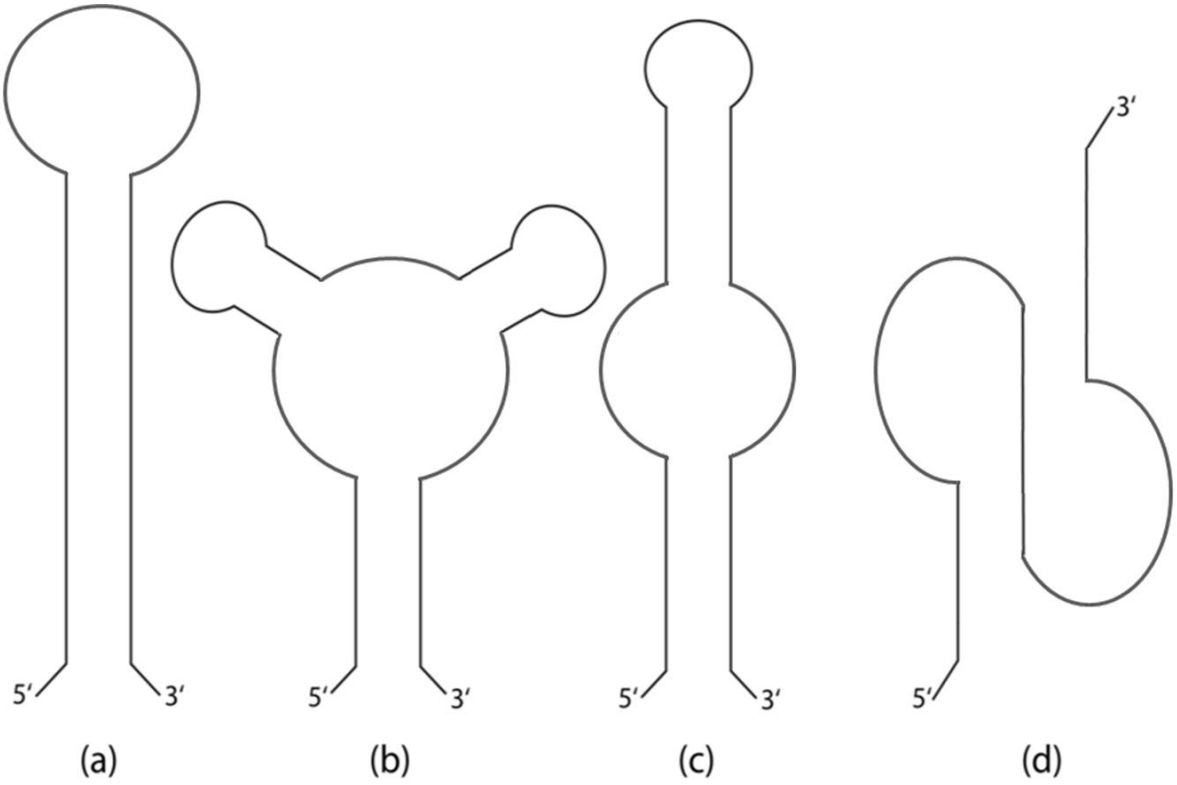


Figure 1.1 Common aptamer conformations

(a) Hairpin loop, (b) multi-branched loop, (c) internal loop/internal bulge, (d) pseudoknot. G-quadruplex structure can be seen in **Figure 2.1.**

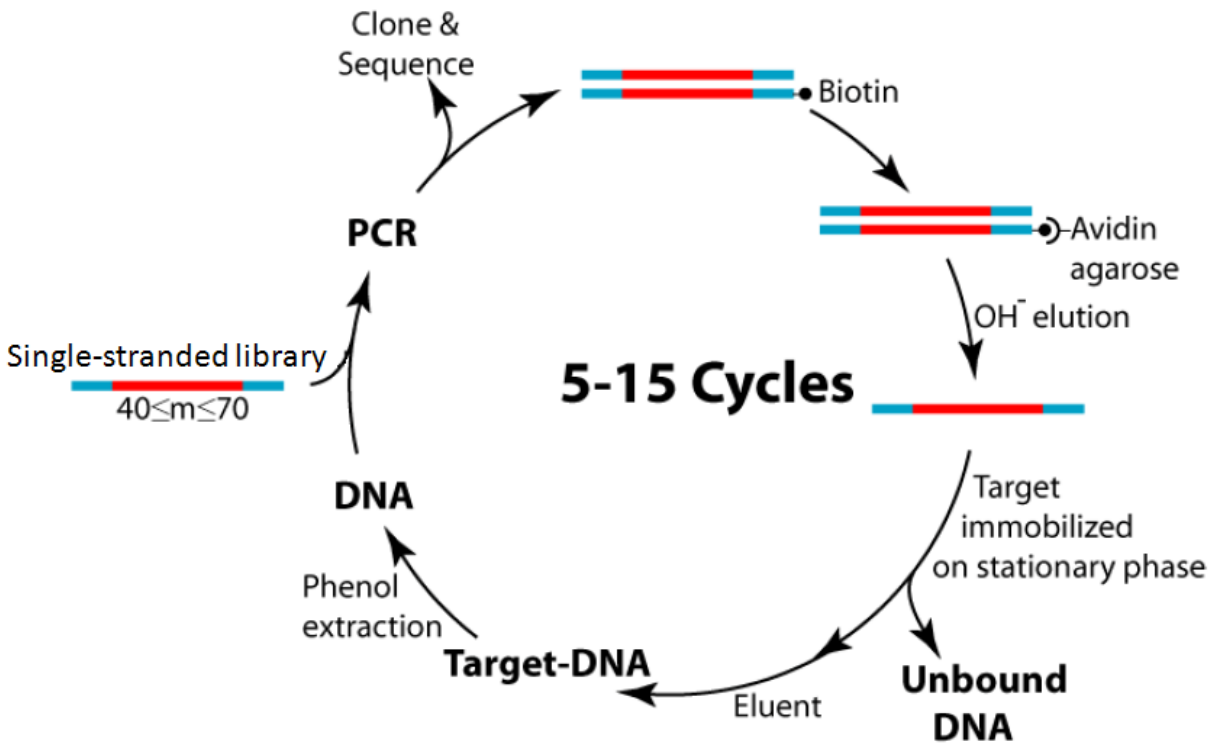


Figure 1.2 SELEX method outline

Schematic representation of the SELEX method with a synthetic DNA library. The outline is similar for RNA libraries, with added T7 Transcription prior to partitioning and Reverse Transcription prior to PCR amplification. (Adapted from the figure by Dr. Philip Borer.)

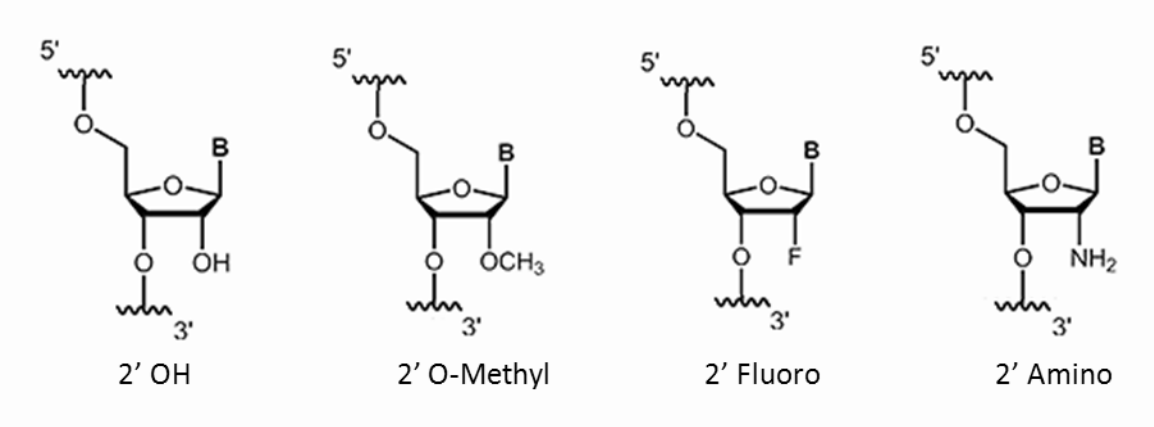


Figure 1.3 2' Modified nucleotides

Chemical structures of abbreviated 2'OH, 2'O-Methyl, 2'-Fluoro and 2'-Amino ribonucleotides where B = A, G, C or U.

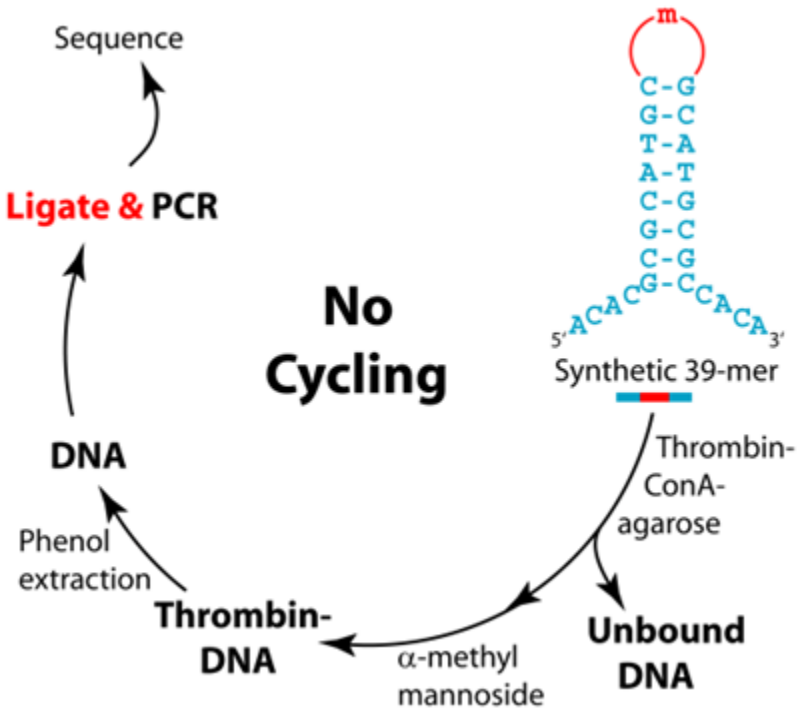


Figure 1.4 AIA method outline

Schematic representation of the AIA method with a synthetic 39-mer hairpin loop DNA library against human α-thrombin immoblized on Concanvilin-A beads. Experimental details are elaborated in Chapter 2. (Figure by Dr. Philip Borer.)

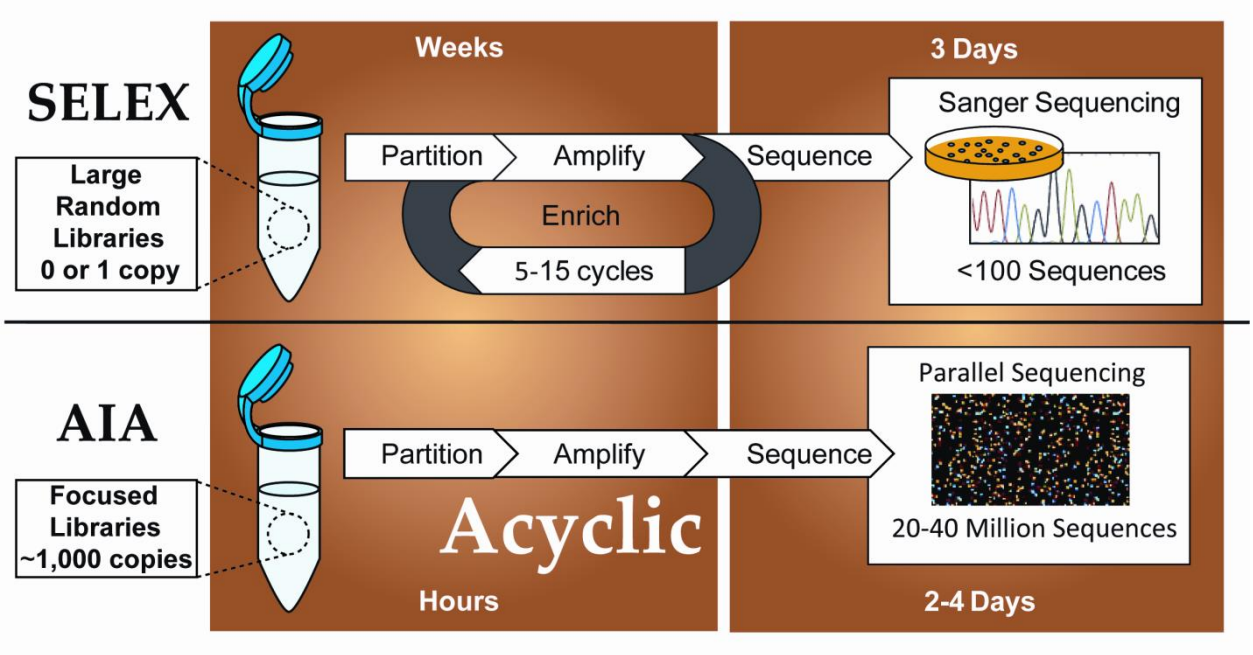


Figure 1.5 SELEX versus AIA

Comparison of SELEX versus AIA workflows. AIA eliminates cycling which drastically reduces time and labor demands. A single experiment is reduced from weeks to days. Sequencing on the Illumina GAIIx requires 2-4 days, while sequencing time for the Illumina MiSeq is significantly reduced to 1-2 days. (Figure by Dr. Mark McPike.)

Table 1.1 Combinatorial considerations for 100 pmol of library

m	P_m	N_p	t_m
1	4	1.50E+13	1.50E+06
3	64	9.40E+11	9.40E+04
5	1,024	5.90E+10	5.90E+03
7	16,384	3.70E+09	3.70E+02
9	262,144	2.30E+08	2.30E+01
11	4.20E+06	1.40E+07	1.40E+00
13	6.70E+07	9.00E+05	9.00E-02
15	1.10E+09	56,066	5.50E-03
16	4.30E+09	14,016	1.40E-03
17	1.70E+10	3,504	3.50E-04
18	6.90E+10	876	8.70E-05
19	2.70E+11	219	2.20E-05
20	1.10E+12	55	5.50E-06
21	4.40E+12	13.7	1.40E-06
22	1.80E+13	3.4	3.30E-07
23	7.00E+13	0.9	8.60E-08
24	2.80E+14	0.2	2.10E-08
25	1.10E+15	0.1	5.50E-09
30	1.20E+18	5.00E-05	5.00E-12
60	1.30E+36	4.60E-23	4.60E-30
120	1.80E+72	3.30E-59	3.30E-66

Combinatorial considerations for 100 pmol of DNA library of variable length, m. Calculations assume precise DNA synthesis with A, T, G, and C occurring at 25% each for each position in the sequence. The number of unique sequences of length, m, is $P_m = 4^m$. The average quantity of each unique sequence is $N_p = 6.0x10^{13}/P_m$, where there is a total of $6.0x10^{13}$ library molecules in the 100 pmol pool. The average number of times a particular sequence will appear if $6.0x10^6$ clusters are sequenced, is t_m, where $t_m=6.0x10^6/P_m$ (See Chapter 2 for details on Illumina clustering).

Chapter 2: Acyclic Identification of Aptamers

Chapter Summary

The HTSA method (High Throughput Screening of Aptamers) described by Dr. Gillian Kupakuwana[148] and later termed AIA (Acyclic Identification of Aptamers)[33] made marked improvements to the traditional aptamer discovery method known as SELEX. AIA eliminated the timely process of cyclic evolution by employing over-represented libraries and deep sequencing. Dr. Kupakuwana successfully identified the minimal human α-thrombin binding aptamer (TBA), originally discovered by Bock *et al* using the SELEX technique[57], with the HTSA/AIA method.[33] "Canonical" TBA is a fifteen residue DNA molecule formed of two stacked G-quartets connected by three T-rich loops, dGGTTGGTGTGGTTGG.[149] A 39-mer DNA library containing a constant stem to the variable m=15 region was designed with the intent of selecting for TBA. The conserved complementary eight base-pair stem and four base non-complementary tails presented the library region in a hairpin loop and allowed for capture of the library via ligation. From a 100 pmol aliquot of this m=15 DNA library containing approximately 56,000 copies of any unique sequence, TBA was successfully selected above background following partitioning against thrombin.[33, 148] This chapter outlines experiments aimed to reproduce Dr. Kupakuwana's data consistently, a necessary precursor to optimize and expand the AIA method. Additionally, experiments to explore the effects of protein to library ratio and library length on identifying high affinity sequences are detailed.

Library Design and Target Selection

The 39-mer DNA library was designed by Dr. Mark McPike and James Crill II (Borer Lab) and was synthesized by Integrated DNA Technologies, Coralville, IA. The m=15 variable region is flanked by a conserved stem and tails: complementary eight base-pair stem and four base non-complementary tails, dACACGCGCATGC-m15-GCATGCGCCACA (**Figure 2.1**). The eight base-pair stem presents the variable region in a hairpin formation, while the non-complementary tails serve as sticky ends to capture the library with adapters for ligation. The hairpin structure and m=15 length were deliberate design choices, as the 15-mer TBA's structural motif is two stacked quartets, which is presented similarly to a hairpin (**Figure 2.1**). During synthesis, hand mixing of the four possible nucleotides ensures that the variable m=15 region contains each base at equal probability, 25%. The incorporation rate of the four bases during DNA synthesis can vary due to efficiency of incorporation, and this varies for all synthesizers. If unaccounted for, this can create libraries with skewed ratios of the four nucleotides. It may also alter the profile of the library pool and may limit the complexity of the library due to sequence specific effects.[150]

The human α-thrombin binding aptamer has been well characterized by both NMR and X-ray crystallography, [65, 149, 151-154] which made human α-thrombin an ideal target for the development of HTSA and continued use in improving AIA protocols. Human α-thrombin is a serine protease that converts fibrinogen into active fibrin and is involved in the coagulation cascade in blood.[155] There is no known physiological binding between thrombin and nucleic acids, however, TBA inhibits the activity of thrombin when bound. With potential for anti-clotting therapy, the structural properties of the thrombin-binding aptamer are important and have been studied extensively. The 15-mer aptamer was used in a Phase I clinical trial by

Archemix Corporation and Nuvelo, Inc. in 2005 as an anti-coagulation agent for coronary artery bypass graft surgery. The pure form of TBA did not enter Phase II due to a high dosage requirement,[156] however, a modified 26-mer aptamer for thrombin was discovered and is currently undergoing a Phase II clinical trial.[157] The nature of thrombin-TBA binding and the correct conformation of TBA have been reconciled by comparing both NMR and X-ray crystallography data (**Figure 2.2**). It has been well established that the thrombin-binding aptamer forms a three dimensional structure of two stacked guanine quartets in solution. The guanine quartets are connected by two T-T loops on one end, and by a T-G-T loop on the opposite end. The orientation of these loops differs between the NMR and X-ray crystal structures of the thrombin-binding aptamer. Although the 5'-syn-anti-3'-orientation of the guanine residues along the edges of the quartets is conserved in both the NMR and X-ray crystal structures, the orientation of the two TT and TGT loops and directionality of the guanine bases differs.[154] The NMR structure determined by Macaya *et al*. has the two TT loops crossing the two narrow grooves and the TGT loop crossing a wide groove.[153] Conversely, the X-ray crystal structure indicates that the two TT loops cross the two wide grooves and the TGT loop crosses the narrow groove. These two arrangements are non-superimposable. In order to reconcile the differences between the NMR and X-ray structures and determine the most likely structural orientation of the thrombin-binding aptamer, Kelly *et al*. re-examined the crystallographic data to determine if it was consistent with the NMR solution structure.[154] They compared eight different directional and structural orientations of the NMR model to the X-ray crystal structure, which was based on the D_4 symmetry of the guanine quartets. Given their criteria including strand directionality, the number of close contacts between the aptamer and thrombin, and steric clashes in the crystalline environment, the optimally oriented structure was found to be consistent with

the directionality of the NMR solution structure. This specific orientation consists of the two TT loops across the narrow grooves and the TGT loop across the wide groove. Thrombin in complex with an aptamer of this orientation results in the two TT loops in close proximity to the fibrinogen recognition site (exosite I) of thrombin, a specific anion binding exosite that is distinct from the catalytic site. Additional crystollagraphic evidence from Russo *et al.* confirms this orientation.[65]

Reconciliation of the correct thrombin-TBA binding orientation sheds new light on the conclusions made from the earliest AIA results. The first six bases of TBA, GGTTGG were 99% conserved for the top 108 TBA sequence variants (54,140 total counts). The middle TGT loop saw moderate variability, while G8, G10, G14 and G15 saw the largest variability. Although affinity decreased with a substitution in the TGT loop, the sequences were still recovered during partitioning. Variation in the two TT loops was minimal, with ≥99.00% conservation, indicating that fewer sequences with substitutions in these loops were recovered during partitioning. This is consistent with the reconciled binding nature of TBA; substitutions within the two TT loops may cause binding affinity to drop so drastically that these sequences would not survive partitioning. The binding affinity of TBA for thrombin was estimated by Dr. Kupakuwana via Surface Plasmon Resonance (SPR). The estimated K_d of 12 nM is similar to previously determined affinities.[33] The human α-thrombin used in the proceeding experiments was purchased from Hematologic Technologies, Essex Junction, VT. The quality and purity of this product was analyzed by Dr. Kupakuwana via MALDI TOF as >90% pure with minimal degradation products.[33]

Replicating Aptamer Selection for Thrombin

The ratio of 60:1, protein: library used by Bock *et al.* [57] was used in the original AIA protocol. A partitioning procedure with agarose-Concanavilin-A beads was used in order to screen the m=15 DNA library against glycosylated human α-thrombin. The DNA library was applied to thrombin immobilized on Con-A beads. After several wash cycles, the protein-DNA complexes were eluted. Phenol extraction and ethanol precipitation resulted in a partitioned DNA library containing sequences with an affinity for human α-thrombin. An outline of the AIA method can be found in **Figure 1.4.**

Partitioning of the DNA library was performed using agarose-Concanavilin-A beads (Glycoprotein Isolation Kit, Pierce Biotechnology) at room temperature. The Con-A beads contained in spin columns were pre-equilibrated in partitioning buffer (20 mM Tris-HCl, 140 mM NaCl, 5 mM KCl, 1 mM $CaCl_2$, 1 mM $MgCl_2$, pH 7.4) by three wash cycles. The buffer was removed by centrifugation at 1,500 rpm for 1 minute. A 100 pmol aliquot of the DNA library in 750 µl partitioning buffer was applied to 500 µl of pre-equilibrated Con-A beads. Following a 20 minute incubation with end-over-end rotation at 25 rpm, the DNA was recovered by centrifugation. Nonspecific binding resulted in the negative selection of the DNA library. The negatively selected DNA library was then applied to 1 ml of pre-equilibrated Con-A beads with 6 nmol of immobilized thrombin. Following a 30 minute incubation with end-over-end rotation at 25 rpm, the unbound DNA was removed by centrifugation. Three wash cycles (5 minutes, rotation at 25 rpm) with 750 µl partitioning buffer ensured thorough removal of unbound DNA. Thrombin-DNA complexes were eluted from the Con-A beads with 750 µl elution buffer (α-

methyl mannoside, Glycoprotein Isolation Kit, Pierce Biotechnology) during a 5 minute incubation with rotation at 25 rpm and collected by centrifugation.

Phenol extraction and ethanol precipitation

Phenol extraction was used to recover the partitioned DNA libraries. An equal volume of Tris-buffered, pH 8.0, 0.1mM EDTA, 50% phenol, 48% chloroform, 2% isoamyl alcohol (Sigma Aldrich) was added to the eluted protein-DNA mixture and vortexed for 30 seconds. Centrifugation at 13,000 rpm for 5 minutes resulted in an aqueous top layer (containing the library) and an organic bottom layer. The aqueous layer contains the anionic DNA while the denatured protein is contained in an intermediate layer. The aqueous layer was removed and the library was subsequently extracted twice with equal volumes of 100% chloroform. The library was then purified by ethanol precipitation. One tenth the volume of 3M sodium acetate and three times the volume of cold ethanol and were added to the library and mixed briefly. Following a 30 minute incubation at -80°C, centrifugation at 13,000 rpm for 30 minutes resulted in a library pellet. After decanting the liquid, the pellet was washed with 1 ml 70% ethanol. Centrifugation at 13,000 rpm for 20 minutes resulted in a library pellet. The pellet was air dried at room temperature in a Speed-Vac and then resuspended in dH_2O.

Modifying the library for sequencing

The procedure for modifying the library for sequencing is outlined in **Figure 2.3**. The DNA library was ligated with sequencing adapters and their complements, compatible with the Illumina SR (Single Read) Format. Adapter 1, Adapter 1 Complement, Adapter 2, and Adapter

2 Complement (See **Appendix 2** for sequences) at 50 µM were added in 10 µl volumes to the 20 µl library. The mixtures were incubated at 90°C for 3 minutes. After cooling to room temperature, 7 µl of 10X ligation buffer (300 mM Tris-HCl, 100 mM $MgCl_2$, 100 mM Dithiothreitol, 10 mM ATP, pH 7.8), (Promega), 2 µl dH_2O and 1 µl T4 DNA ligase (50% glycerol stock, 3 u/µl), (Promega) were added. Incubation at 25°C for 30 minutes completed the ligation. The ligated library was purified with a QIAquick PCR Purification Kit (Qiagen) to reduce the volume from 70 µl to 30 µl, more suitable for gel electrophoresis in a single well. The kit also removes protein; however, this is not a requirement for gel electrophoresis. The product was visualized on a 2% TBE agarose gel alongside a 50 bp DNA ladder. (See **Figure 3.3** in the next chapter for illustration of the product size.) The product is 105 nucleotides long with 66 of those being base-paired with Adapter Complements and runs between the 100 and 150 bp ladder bands. The product band was excised and purified with a MiniElute Gel Extraction Kit (Qiagen), resulting in pure ligated DNA library. PCR amplification with Forward and Reverse Primers (See **Appendix 2** for sequences) extends the 5'-length of the ligated product to allow for annealing to the immobilized complement of the Illumina flow cell, while simultaneously filling in the complement to the variable region. The 10 µl purified adapter ligated library was combined with 1 µl each of the Forward and Reverse Primers at 10 µM (IDT, Coralville, IA), 1 µl of 250 µM dNTPs (Stratagene), 2 µl 10X Pfu Turbo Polymerase buffer (200 mM Tris-HCl, 20 mM $MgSO_4$, 100 mM KCL, 100 mM $(NH_4)_2SO_4$, 1% TritonX-100, 1 mg/ml nuclease free BSA, pH 8.8), (Stratagene), 4 µl dH_2O, and 1 µL Pfu Turbo polymerase (50% glycerol), (Stratagene). The 20 µl reaction mixture was amplified with the following conditions:

Denaturation	94°C	2 min.
PCR Amplification (18 cycles)	94°C	1 min.
	61°C	1 min.
	72°C	1 min.
Final Extension	72°C	10 min.

The final 130 bp product was size checked on a 2% TBE agarose gel alongside a 50 bp DNA ladder (See **Figure 3.4** for illustration of product size).

Sanger sequencing

The partitioned, ligated, and modified DNA libraries were screened using Sanger sequencing prior to sequencing on the Illumina GA to ensure the presence of the correct ligated and modified DNA sequences: the sequencing cassette. PCR of a small aliquot of the sample using a 5'-phosphorylated Forward and Reverse Primers (IDT) adds a 5'-phosphate to the library, allowing ligation into a plasmid. Using a CloneSmart HCKan cloning kit (Lucigen), 100 ng of the 5'-phosphorylated DNA library (determined by quantification with the NanoDrop Spectrophotometer) was ligated into a pSmart-HCKan vector at room temperature for 30 minutes. Ligation was terminated at 70°C for 15 minutes, cooled to room temperature for 15 seconds, then to 4°C for 15 seconds in a thermocycler. To 25 µl Lucigen E. Cloni 10G electrocompetent cells, 2 µl of the ligated pSmart-HCKan vector was added. The mixture was carefully transferred to a pre-chilled electroporation cuvette. Electroporation in an Eppendorf Model 2510 pulser at 1800V and addition of 975 µl CloneSmart Recovery Media (11.8 g Bacto-tryptone, 23.6 g yeast extract, 9.4g anhydrous K_2HPO_4, 2.2g anhydrous KH_2PO_4, 0.4% glycerol in 1L dH_2O) completed the transformation. After incubated for 1 hour at 37°C with shaking at 250 rpm in 5 ml falcon culture tubes, 40 µl was plated on a kanamycin (100 µg/µl) agar plate.

Overnight incubation produced a high colony density, and 8 individual colonies were picked and inoculated in 5 ml Superbroth (32g tryptone, 20g yeast extract, 5g NaCl, 5ml 1.0M NaOH in 1L dH_2O) with kanamycin (50 µg/ml). After overnight incubation at 37°C with shaking at 150 rpm, cells were collected with centrifugation at 1000 rpm for 10 minutes. The plasmids were extracted and purified with a QIAprep Spin Miniprep Kit (Qiagen), and then Sanger sequenced using the CloneSmart SL1 Primer (See **Appendix 2**) at the Core Facility for DNA Sequencing at SUNY Upstate Medical University, Syracuse, New York. Data is viewed using Chromas Lite software (**Figure 2.4**), (Technelysium Ptt, Ltd., South Brisbane, QLD, Australia).

Quantitative PCR

After the correct sequencing cassette was identified via Sanger sequencing, an additional quality control method was implemented in order to maximize data quality. The precise quantity of amplifiable DNA was determined with the KAPA Library Quantification Kit on the BioRad iCycler. The KAPA kit was specifically designed to reproducibly and accurately determine sample concentration for sequencing on Illumina platforms. The KAPA kit is capable of amplifying single or double stranded products containing the Illumina P5 and/or P7 motifs (See **Appendix 2** for sequences). The accurate quantification of a library is crucial for maximization of data output on Illumina sequencing platforms. If the amount of amplifiable library is overestimated, the result is lower than expected cluster density. If the amount of amplifiable library is underestimated, the result is crowded clusters and poor resolution. Any non-library DNA that is not flanked by the adapter motifs is irrelevant as it will not be amplified. Quantification by UV absorption is not a reliable method because it accounts for all DNA species as well as other contaminants that absorb UV similarly, such as residual ethanol. The

KAPA kit requires the use of six standard solutions and the acquisition of a standard curve. The quantification of unknowns is inferred based on this curve.[158]

The concentration of the sample was estimated on the BioTek Synergy 3 Microplate reader and diluted to 20 nM. The sample is diluted further to 1:1000 in library dilution buffer (10 mM Tris-HCl, pH 8.0, 0.05% Tween 20). The dynamic range of the assay extends from 20 pM to 0.0002 pM via the six DNA standards. At 20 pM, the 1:1000 dilution is at the upper limit of the standard curve; however, concentrations were typically overestimated by UV absorption and diluting to the upper limit of the curve ensured that data points would fall within the range of the standard curve. The library and DNA standards were prepared in triplicate as follows: 12 µl KAPA SYBR FAST qPCR Master Mix containing Primer Premix, 4 µl dH_2O and 4 µL diluted library or DNA standard. qPCR cycling was as follows: (1) 5 minutes at 95 °C and (2) 35 cycles of 30 seconds at 95°C and 45 seconds at 60°C. The average concentration (pM) of the triplicate data points is adjusted to compensate for the size of the DNA standards and the 1:1000 dilution:

Avg. Conc. (pM) x [452 bp/library length] x 1000 = Conc. of library stock (pM)

There is no need to account for the 4 µl volume of the library used in the 20 µl reaction because the volume is the same for the library and standards. If one of the triplicate data points is an outlier, it is discarded. If more than one is an outlier, the data cannot be considered reliable and the assay must be repeated. If the quantity of amplifiable DNA was sufficient (5-20 nM) and Sanger sequencing data was agreeable, the sample was considered for sequencing.

Partitioning at Other Protein: Library Ratios

The m=15 DNA library was also partitioned at ratios of 30:1 (3 nmol thrombin: 100 pmol library), 15:1 (3 nmol thrombin: 500 pmol library) and 3:1 (3 nmol thrombin: 1 nmol library), using 500 µl Con-A beads for partitioning and 250 µl beads for negative selection. These experiments were aimed at increasing the selection pressure for high affinity sequences while providing benchmark data for optimizing the minimal ratio that provides usable AIA counts. By decreasing the quantity of thrombin relative to the quantity of library, the competition to occupy binding sites naturally increases. If all of the high affinity binding sequences occupy binding sites on the protein target, any excess protein increases the likelihood of carryover of nonspecific binding sequences. In order to identify high affinity binding sequences, they must be counted above background; a high carryover of nonspecific sequences decreases the ratio of captured high affinity sequences to background, which would be reflected in the data. Dr. Kupakuwana explored this hypothesis with 6:1 (600 pmol thrombin: 100 pmol library) and 1:1 (100 pmol thrombin: 100 pmol library) ratio experiments. Although she observed TBA at 28,389 counts out of a total 5,719,989 counts (frequency of 0.496 %) for the 6:1 experiment, she only observed 642 counts of TBA out of a total 7,788,985 counts (frequency of 0.008%) for the 1:1 experiment. The original 60:1 experiment produced 48,671 counts of TBA out of a total 1,728,220 counts (frequency of 2.3 %).[148] Here, "count" refers to the number of times a specific sequence occurs as a "good read," which is defined in the next section. These results do not reflect the hypothesis that a decreased ratio of protein: library increases selection pressure. In fact, they demonstrate that the selection pressure is decreased. The aforementioned additional experiments were aimed at probing this hypothesis further.

Additionally, m=16 and m=19 DNA libraries were partitioned against thrombin in 60:1 ratio experiments (6 nmol thrombin: 100 pmol library). The number of TBA molecules in each of these pools decreased from ~56,000 for an m=15 library to ~14,000 and ~219, respectively. However, due to their increased length, any 15-mer may occur within these libraries at multiple positions within the variable region. As suggested by Marshall and Ellington,[42] complete sampling of a library of length, *m*, may not be necessary, as the library would be completely represented, if not over-represented for all *n*-mers of length <*m*. Given that the m=16 and m=19 libraries are over-represented at 100 pmol, any 15-mer sequence would also be over-represented. These experiments were aimed at exploring the capabilities of under-represented libraries for identifying the known 15-mer, TBA, in the context of longer variable regions.

High Throughput Sequencing

Sequencing of the experiments in this chapter was performed on our in-house Illumina Genome Analyzer IIx (GAIIx). All Illumina sequencing platforms require immobilization of the library sequences to the surface of a proprietary glass flow cell. The GAIIx utilizes an eight lane microfluidic flow cell which allows the user to input 8 different sample solutions for sequencing. On the flow cell, surface bound templates complementary to the prepared libraries allow for capture of the library sequences as they flow through the microfluidic channel (lane). These immobilized oligonucleotides are referred to as the "primer lawn" (See **Appendix 2** for sequences). Following NaOH denaturation, the library molecules are captured by a complementary sequence on the primer lawn. An extension step generates the full length, double-stranded product. The original library molecule is washed away following a NaOH

denaturation. The complement to the original library molecule is now covalently linked to the flow cell surface. Next, the strand "bridges" to anneal to its complement on the primer lawn and the double stranded product is generated. This bridge amplification of the captured library sequences creates millions of colonies, each containing up to 1,000 copies of the original sequence, known as clusters. The clustering process is performed on the free standing Cluster Station. The clustered flow cell is transferred to the GAIIx where it is sequenced by Sequencing by Synthesis (SBS) technology. The dye-terminating chemistry utilizes four fluorescently-labeled nucleotides. After the initial annealing of the sequencing primer, single fluorescent dNTPs are incorporated during each cycle. High resolution imaging through four filters captures fluorescence emission to determine the specific location of base incorporation that corresponds to individual clusters. The fluorescent dye is then enzymatically cleaved to allow for incorporation of the next base. Prior to sequencing, a diol linker within the P5 flow cell oligo is cleaved with periodate, leaving only the strands connected to the flow cell at their 5'-ends.[159] This allows for the sequencing primer to anneal to the immobilized strand such that SBS occurs from the top, down. This ensures that the fluorescence emission is observed at the same distance from the flow cell for each cycle of base incorporation. The number of base incorporations is user defined to reflect the length of the library or any sequence portion of interest. The experiments in this chapter utilized a 36 base, Single Read (SR) kit. The workflow is outlined in **Figure 2.5**.

Data analysis

Sequence data is reported in FASTQ format, a text file that identifies each sequence and the corresponding quality scores. The FASTQ file identifies all 36 base incorporations, meaning

additional processing is required in order to analyze the internal region. To accomplish this, a Perl script was created by Dr. Huitao Sheng of Dr. Chluhuri's bioinformatics group at Syracuse University (See **Appendix 2**). The Perl script has since been annotated and modified by Dr. Damian Allis as our data analysis needs changed over time, as seen in future chapters. The Perl script identifies and qualifies the constant tail and stem regions independently of the variable region. The quality of the tail and stem regions identifies sequences as "good", "candidate" or "bad" reads. Candidate reads contain ≤2 mismatches or a single gap or insertion within the eight nucleotides from the 5'-end of the variable region, or the five nucleotides from the 3'-end. Of these sequences, those containing a variable region of exactly 15 nucleotides (or any user defined number) are good reads. Sequences that did not meet the criteria of candidate reads are considered bad reads. The stringency of the qualifiers for candidate and good reads, as well as the length of the head and tail under scrutiny are user defined and may be changed to suit a specific experiment. These three files are searchable text files and the full length sequences are reported. After the Perl script has parsed the full length sequences, the occurrence of each "good" read is tallied to generate the "n-mer count" file which reports only the user specified variable region and the frequency of each sequence. The nmer count data can be reformatted into FASTA format using an additional Perl script for use in sequence alignment or drawing phylogenic trees (See **Appendix 2**). ClustalX2 allows for sequence alignment of up to 1000 sequences while TreeView plots the alignments to create phylogenetic trees.

Results and Discussion

Of the four initial AIA experiments (60:1, 30:1, 15:1, and 3:1 ratio) aimed at duplicating and expanding the original AIA results, the 60:1 experiment was disqualified for sequencing due to poor qPCR results. This sample also failed to produce the correct sequencing cassette as determined by Sanger sequencing. The remaining three samples were sequenced on the Illumina GAIIx. The raw data files were parsed based on the last eight bases of the head (dGCGCATGC) and the first five bases of the tail (dGCATG). The top twenty sequences for the published 60:1 AIA experiment are reported in **Table 2.1**, which shows the canonical TBA 15-mer, GGTTGGTGTGGTTGG, at highest abundance. The top twenty sequences from this work are reported in **Tables 2.2-2.4** for comparison. Dr. Kupakuwana identified two major motifs: sequences similar to TBA and a carbohydrate binding sequence. Motif Ia has TBA variants with mainly T↔G substitutions, for which the decreasing counts going down **Table 2.1** correlate with the decreasing DNA-thrombin binding constant as determined by SPR. Motif Ib contains weaker binding TBA variants which adds T→C and/or G→A substitutions. The carbohydrate binding aptamer (Carb1) weakly binds the glycan component of thrombin, which is a post-translational modification,[160] and binds glucose with a K_D of ~1.4 μM, which is found in the Con-A storage buffer. A third sequence of note, a systematic artifact, is termed the "jump" sequence. The jump sequence "skips over" the m=15 variable region and a portion of the tail to include a section of Adapter 2 in the reported sequence. This systematic artifact was hypothesized to originate during PCR amplification of the ligated library. Sequence similarity between Adapter 2 and the tail region allows the Perl script to identify these sequences as "good" reads and is illustrated in **Figure 2.7**. Multiple variations on this "jump" behavior are seen in the candidate and bad reads files as well. This sequence and its variants can be excluded from sequence

alignments and further data analysis by confirming the "jump" behavior by looking at the 15-mer in context within the good reads file.

Table 2.5 compares pertinent statistical data from the three AIA experiments in comparison with the original 60:1 AIA experiment. Acyclic selection does not allow the library pool to converge into a small number of sequences; therefore, selection efficiency is determined by the presence of specific sequences above the background or noise threshold. When considering the success of an AIA experiment, it is critical to look at the frequency of a sequence in addition to the count. The count represents the number of times a specific sequence occurs, however, the total number of sequences per experiment varies. High throughput sequencing provides an enormous sequence space and as such, there can be a large variation in the quantity of data. Thus, the frequency of a sequence is a more accurate determination of selection efficiency. It is also useful to compare sequence frequency with respect to good reads and total reads. The overall percentage of good reads is indicative of sample quality. An AIA experiment can have any combination of high/low selection efficiency and good/poor data quality. An ideal result is a high percentage of good reads which contain high frequency aptamer candidates with a low background. In the remainder of this dissertation, the "frequency" for a sequence or a set of sequences is taken with respect to the total of good reads.

The counts of TBA for the 30:1, 15:1 and 3:1 experiments were 10, 78, and 56, respectively. The frequency of TBA for the three experiments and the original 60:1 experiment with respect to good reads and total reads is reported in **Table 2.5**. As mentioned, the frequency of TBA for the original 60:1 experiment was 2.37 %. The frequencies of TBA for the 30:1, 15:1 and 3:1

experiments are 0.00012 %, 0.00094 % and 0.00067 %, respectively. Although the frequency increases for the 15:1 experiment compared to the 30:1, it does not increase for the 3:1 compared to the 15:1. Overall, the frequencies are very much lower than the original 60:1 experiment. There is also a lack of TBA variants within the three experiments. This can be visualized by comparing phylogenetic tree diagrams showing similarity as the result of sequence alignment (**Figure 2.6**). Unlike the many TBA variants found in the original 60:1 experiment data, only the consensus TBA sequence is found in the three AIA experiments.

It is imperative to compare the frequency rather than counts when the starting quantity of library differs, as with the 30:1, 15:1 and 3:1 ratio experiments (100 pmol, 500 pmol, 1 nmol of library, respectively), because the library is over-represented to different extents. If the starting count per unique sequence is considered solely, the frequency of TBA would be expected to increase 5-fold for the 15:1 experiment and 10-fold for the 3:1 experiment compared to the 30:1 experiment. Although the frequency increases 8-fold for the 15:1 experiment, it only increases ~5-fold for the 3:1 experiment. In combination with a hypothesized increase in selection pressure due to the relative decrease in protein, frequency is expected to diminish further. The new data display an overall low frequency and lack of a clear trend in frequency with respect to the protein: library ratio. More sequencing runs might allow testing the hypothesis that a decreased ratio of protein: library increases selection pressure, but a sequencing kit for each run on the GAIIx cost about $8,000, so we decided not to pursue this further until the overall lower than expected frequency of TBA was addressed. The samples also produced lower than expected frequencies for the carbohydrate binding sequence.

The small number of sequences above background in these experiments suggests an issue with the partitioning method; at a 100 pmol scale with approximately 56,000 starting copies of each sequence, counts for high affinity sequences were expected to be higher. There are at least three possible explanations for the low frequencies: (1) the partitioning was not stringent enough, meaning large quantities of non-specific sequence were carried over. This would prevent any high affinity sequences from appearing significantly above background. (2) It is also possible that too-stringent washing could elute bound TBA and its weaker binding relatives as well as the desired removal of non-specifically bound DNA. Using the value of K_d = 12 nM reported by Kupakuwana, *et al*.,[33] for the TBA-Thrombin complex and assuming a diffusion controlled on-rate, $k_{on} = 1\times10^7$ typical for the interaction of fairly large molecules in aqueous solution, the definition that $K_d = k_{on}/k_{off}$ can be used to predict that $k_{off} = 0.12\ sec^{-1}$, and the lifetime of the complex $\tau = 1/k_{off} = 8$ sec. Lifetimes for more weakly bound variants of TBA can be lower by orders of magnitude. Of course, the on-rate may be slower for a bead-bound complex and the on/off processes may be more complex kinetically than assumed in this simple model. However, it does emphasize that loss of some DNA from a reversibly bound complex is inevitable using immobilization for partitioning. This is part of the rationale for adopting the covalent formaldehyde crosslinking approach for partitioning described in Chapter 6. Although the present work closely emulated the Kupukuwana protocols for washing the bead-bound library, the possibility remains that there were differences. (3) During a negative selection step, it was observed that the beads were absorbing much of the library. If the overall pool of library is decreased significantly, fewer copies of each sequence are present, making it difficult to select high affinity binders above the background. This background binding varied considerably from lot-to-lot of agarose Concanavalin-A beads and was explored in the next chapter.

The jump sequence and/or variants of it are present in all three experiments. The counts are similar to TBA and Carb1 but do not dominate the good reads. Also, a large percentage of the total reads were qualified as good reads, indicating that there is not an issue with the quality of data from the sequencer. For the experiments with m=16 and m=19 libraries, the m=16 experiment was disqualified for sequencing due to poor qPCR results. This sample also failed to produce the correct sequencing cassette as determined by Sanger sequencing. The m=19 sample was sequenced on the Illumina GAIIx. The top twenty sequences are reported in **Appendix 2**. The sequences that occur above background are exclusively jump sequence variants. Any non-jump sequences occur at counts of two or one, rendering that data unusable for determining any high affinity sequences. The poor data quality could be attributed to unsuccessful partitioning and/or flawed sample preparation. It is possible that the m=19 library did not bind to thrombin with any specificity, which would produce very little correctly ligated product for amplification. This could allow jump sequences to dominate during amplification. The library's hairpin loop structure allows the m=15 variable region to form the desired G-quadruplex structure with ease, however, the added bases in the m=16 and m=19 libraries may prevent this. However, inconsistent sample preparation for the m=15 library, 60:1 experiment aimed at duplicating the original 60:1 experiment indicates that the failure of the m=16 and m=19 libraries is also due in part to flawed sample preparation.

Although the three experiments summarized in **Tables 2.2 – 2.5** identified TBA and the Carb1 sequences above background, frequencies are much lower compared to the original 60:1 experiment. The unexpectedly low sequence counts for the variable ratio experiments led to several troubleshooting experiments that are detailed in the next chapter. Additionally, steps to

improve sample preparation aimed at eliminating the jump sequence and reliably producing the correct sequencing cassette were employed.

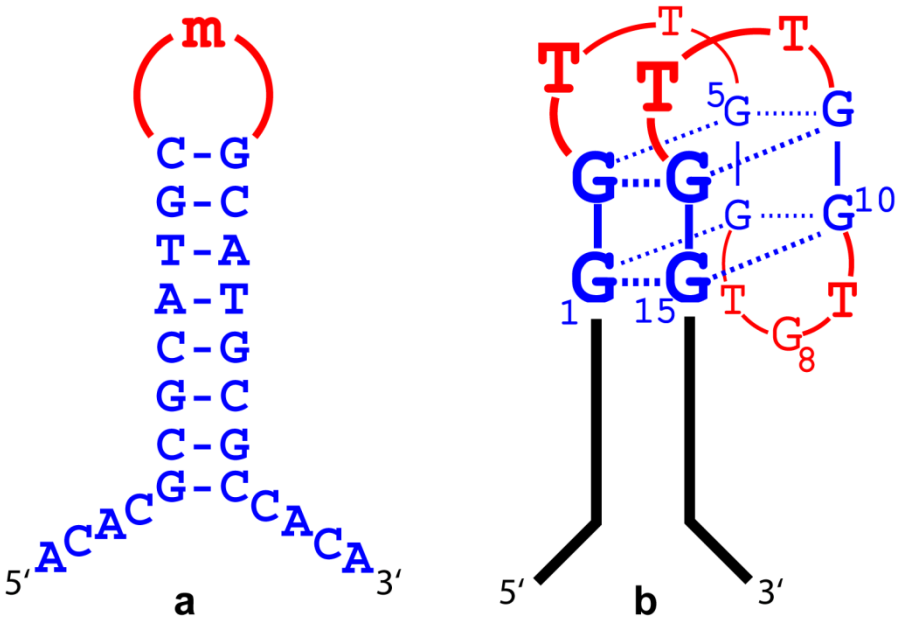


Figure 2.1 Hairpin loop library structure and thrombin binding aptamer

(a) Structure of the m=15 hairpin loop DNA library. Four base non-complementary tails and 8 base complementary stem regions shown in blue. Variable region loop, *m*, shown in red. The constant region at the 5'-end of the variable loop is termed the "head." The constant region at the 3'-end of the variable loop is termed the "tail." **(b)** Structure of the canonical G-quadruplex thrombin binding aptamer in the structural context of the library. Head and tail regions designated by black lines. The stacked eight stacked G's form a quadruplex structure, shown in blue. The two TT and single TGT loops are shown in red. (Figure by Dr. Philip Borer.)

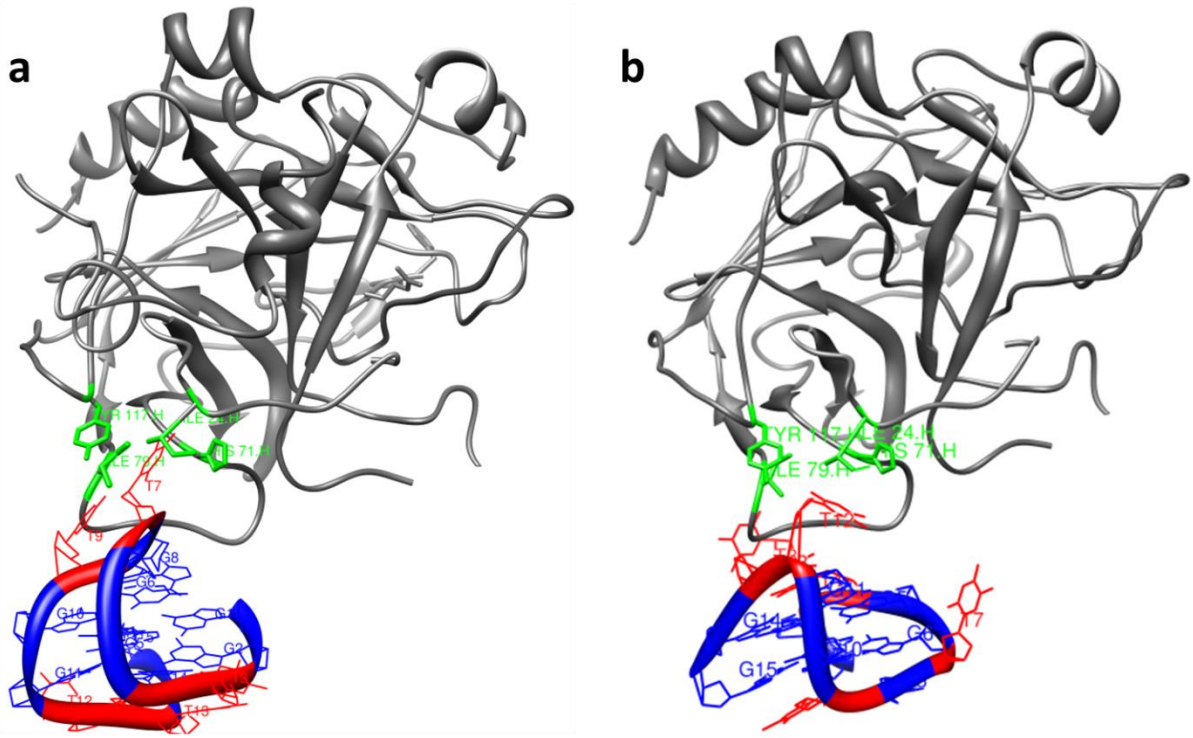


Figure 2.2 Two proposed interactions of the thrombin binding aptamer and thrombin as determined by X-Ray crystallography and NMR

(a) Thrombin binding aptamer (blue/red) in complex with the anionic exosite I of thrombin, the fibrinogen exosite. The T7 of the TGT loop is shown within the hydrophobic pocket formed by Ile24, His71, Ile79 and Tyr117 (green) near the fibrinogen binding site. Hydrogen bonding and ionic interactions of the TGT loop with thrombin are based on the original X-ray model from Padmanabhan, *et al in* 1993.[151] Protein Data Bank ID Number 1HAP. **(b)** Thrombin binding aptamer (blue/red) in complex with the fibrinogen exosite. T12 of the TT loop is shown in the hydrophobic pocket formed by Ile24, His71, Ile79 and Tyr 117 (green), placing the TT loop in close proximity to the fibrinogen binding exosite. This interaction is based on the most optimally oriented NMR model.[65, 152, 154] Protein Data Bank ID Number 1HAO.

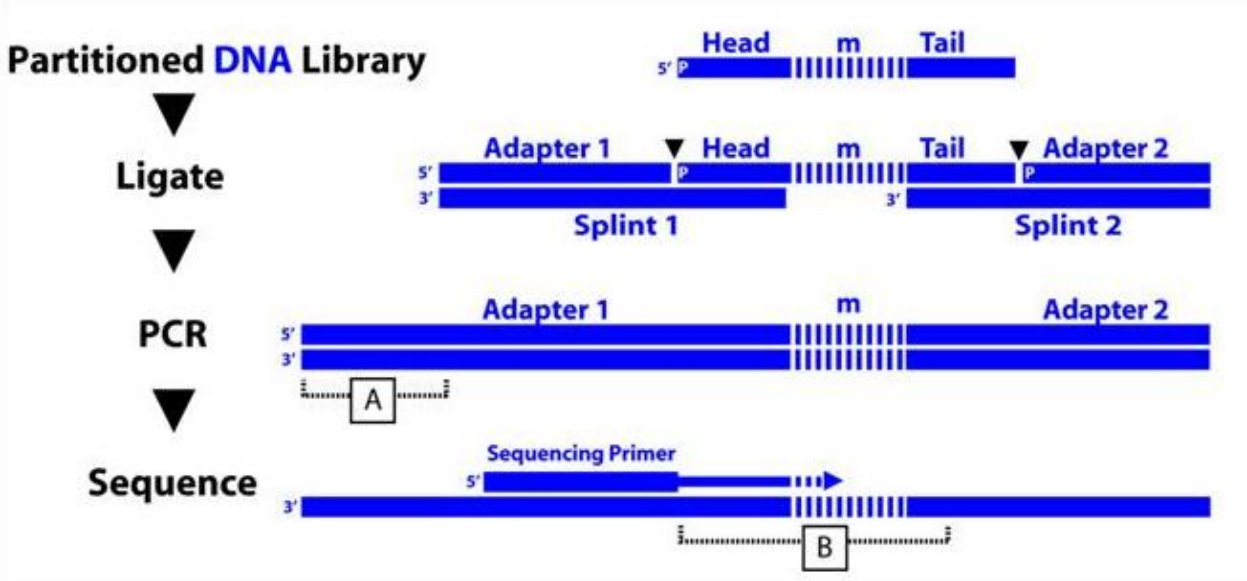


Figure 2.3 Preparing the library for sequencing in AIA

The partitioned DNA library is ligated with Adapter 1 and Adapter 2 following annealing of the head and tail regions with overhangs on Adapter 1 Complement and Adapter 2 Complement (Splint and Splint 2). PCR amplification adds a 5'-extension of 25 bases (A), which contains the P5 sequence necessary for annealing to the Illumina flow cell. PCR amplification simultaneously fills in the gap across from the variable region, *m*. During sequencing on the Illumina platform, the sequencing primer anneals complementary to a portion of Adapter 1 Complement. Sequencing by Synthesis across the bottom strand begins at the 3'-end of the 39-mer library insert and continues for a user defined number of base incorporations (B). The reported sequence is the top strand, which corresponds to the original library sequence. (Figure by Dr. Philip Borer.)

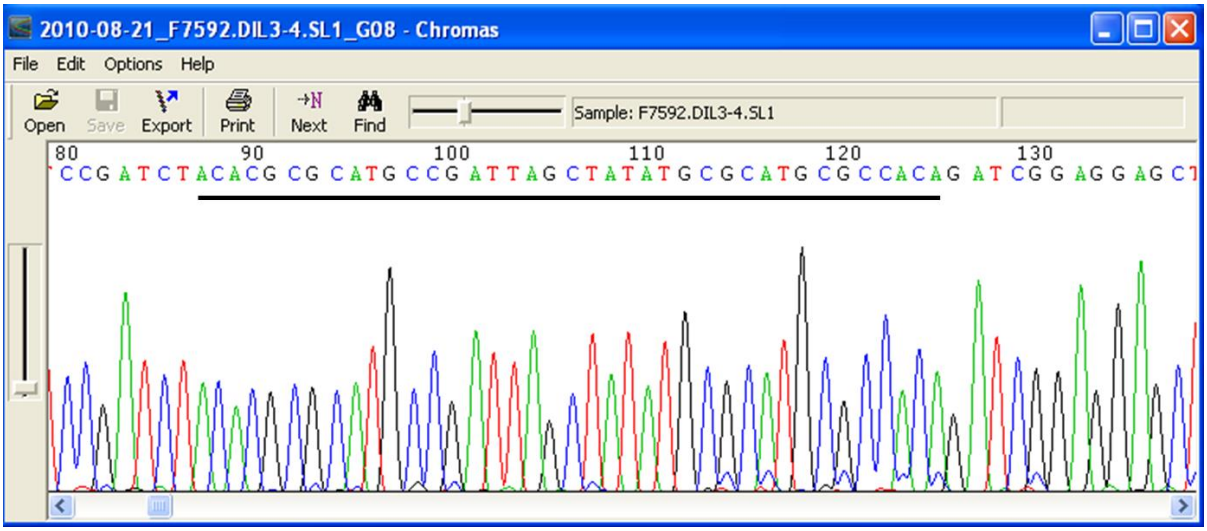


Figure 2.4 Sanger sequencing sample data

Screenshot of successful Sanger sequencing of an m=15 DNA hairpin loop library ligated with Adapter 1 and Adapter 2. The 39mer library sequence is underlined. The correct Adapter 1 and Adapter 2 sequences flank the library (See **Appendix 2** for sequences).

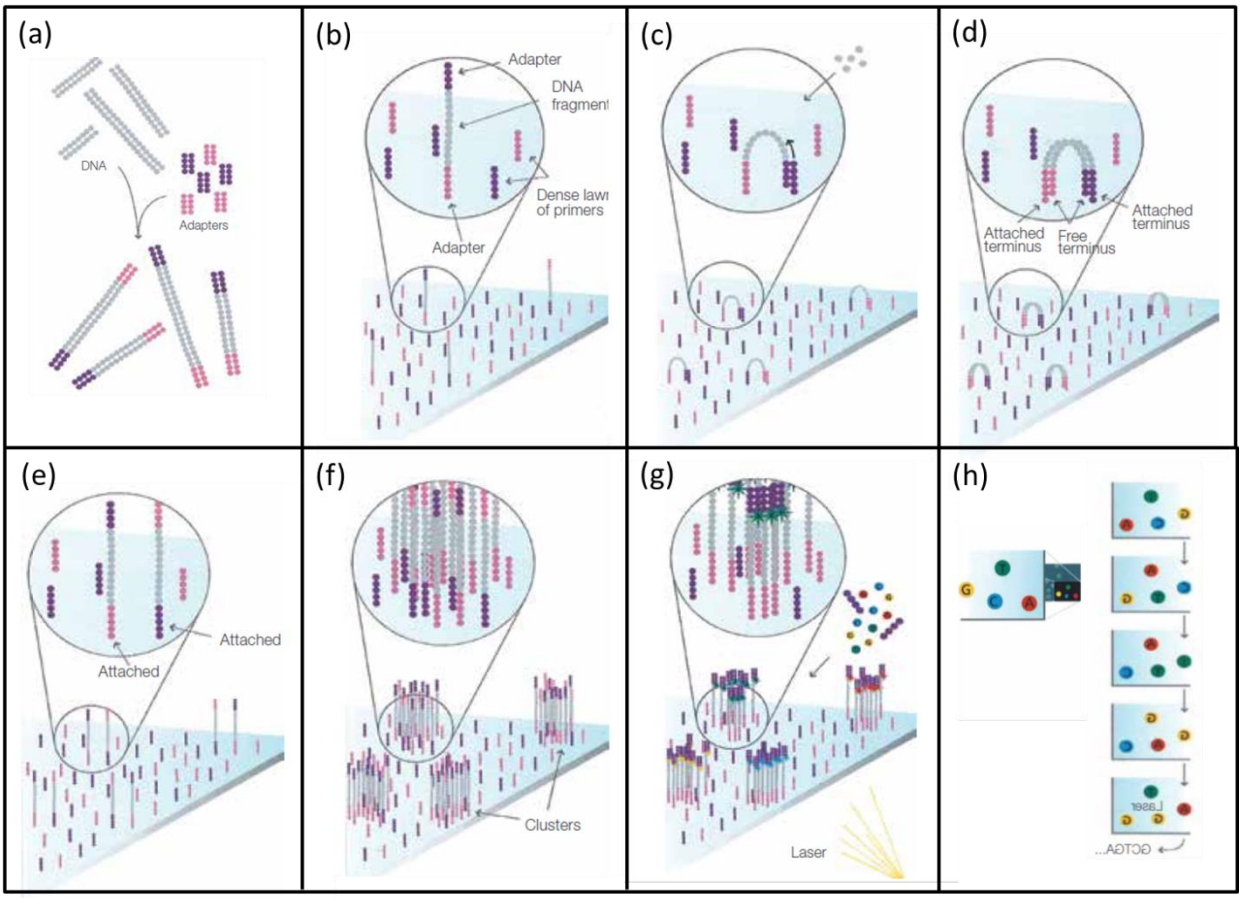


Figure 2.5 Overview of Illumina sequencing

(a) Preparation of library for sequencing (image shows blunt end adapter ligation only). **(b) S**ingle-stranded library fragments anneal randomly to the flow cell surface, complement strand is generated. NaOH denaturation and washing leaves only the single-stranded, newly generated complement strands attached to the flow cell. **(c)** PCR reagents are added to initiate bridge amplification **(d)** One cycle of bridge amplification. **(e)** NaOH denaturation leaves only the single-stranded templates attached to the flow cell. **(f)** Cyclic bridge amplification generates millions of clusters with up to 1,000 copies each. **(g)** Following cleavage of fragments attached at the 3'-end, the sequencing primer is annealed and the first sequencing cycle begins by adding a mixture of the four fluorescently labeled nucleotides. **(h) (***Left)* Bird's eye view of four

clusters. Emission fluorescence identifies the first base incorporation. (*Right)* Repetition of the sequencing cycles determines the sequence of each fragment. Figure adapted from [159].

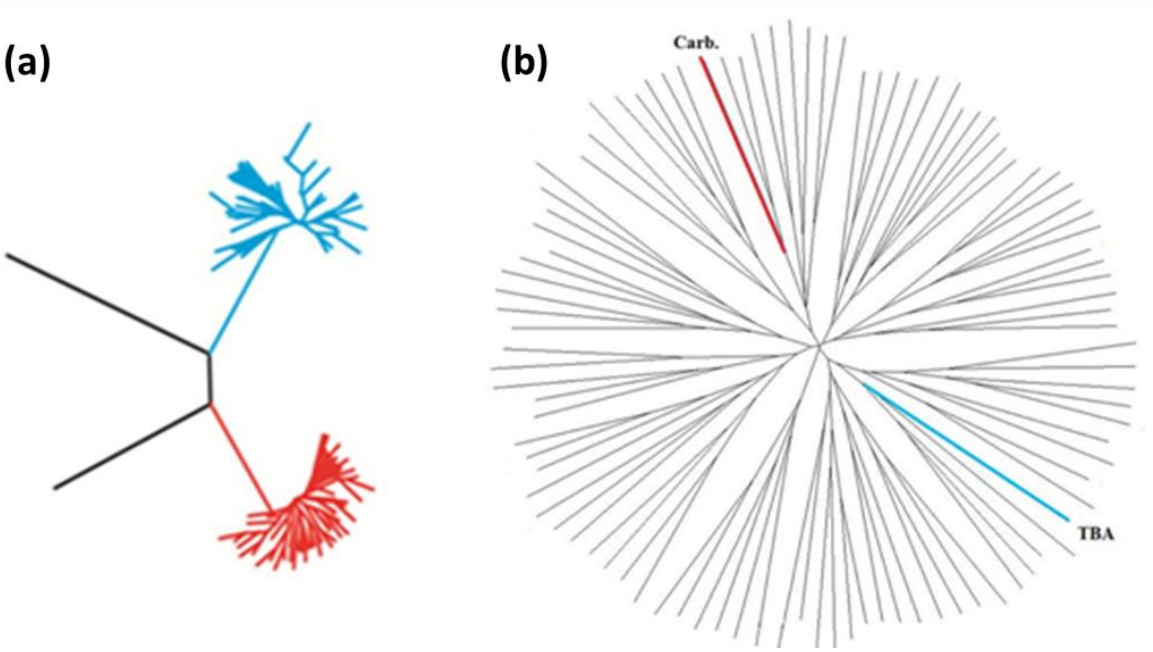


Figure 2.6 Phylogeny trees of original 60:1 and new 30:1 AIA for thrombin

(a) Original 60:1 experiment for thrombin. Phylogenic tree of sequences occurring at least 7 times (top 150 sequences). TBA and variants (motif Ia) in blue. Carbohydrate binding sequence variants in red. Adapted from [148] **(b)** 30:1 experiment for thrombin, phylogenic tree of top 100 sequences. TBA in blue, carbohydrate binding sequence in red. Little correlation is seen between sequences.

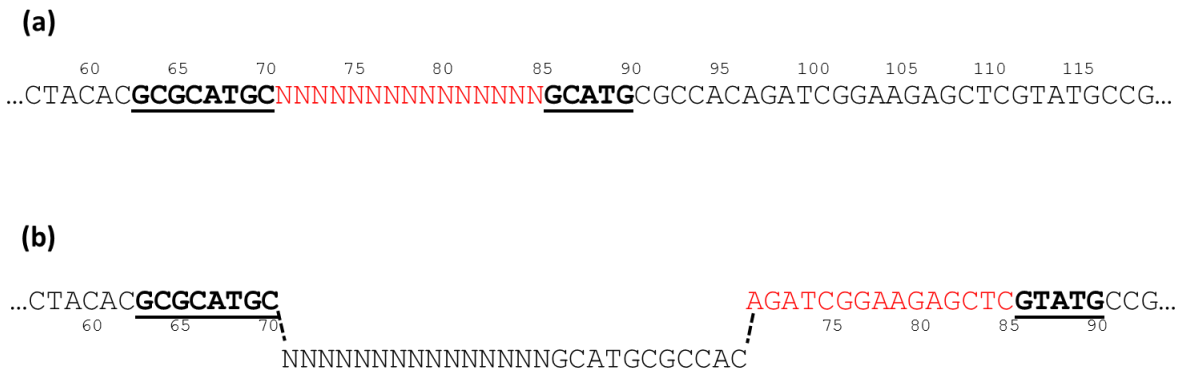


Figure 2.7 The jump sequence

(a) Snapshot of the 39-mer library (bases 59-90) ligated to Adapter 1 and Adapter 2 (shown 5' to 3'). The Perl script recognizes the two underlined sections as the head and tail. The reported 15-mer is identified in red.

(b) Snapshot of the most common jump sequence, a systematic artifact of library preparation. The lowered section, including the entire randomized region and 11 of 12 tail bases, is jumped over. The Perl script recognizes the correct head, but recognizes a portion of Adapter 2 as the tail (underlined). The portion of Adapter 2 that is 15 bases downstream from the jumped section differs from the original tail by a single mismatch, allowing the sequence to be reported as a good read. The reported 15-mer is consequently the last base of the tail and the first 14 bases of Adapter 2. Variation on the exact number of bases "jumped" and insertions/deletions can alter the sequence of the reported 15-mer, however, all jump sequences qualified as good reads are recognized due to sequence similarity between the tail and Adapter 2.

Table 2.1 Top 15 aequences, original 60:1

Rank	Sequence	Count	Identifier
1	GGTTGGTGTGGTTGG	46444	TBA (Thb1)
2	gctatcatcgcaacg	29405	Carb1
3	GGTTGGTGTGGTTTG	2451	Ia
4	gctatcatcgccacg	1040	II
5	AGATCGGAAGAGCTC	710	Jump
6	gctatcatcgcaccg	678	Ia
7	GGTTGGTGTTGTTGG	647	Ia
8	GGTTGGTGTGGTTGT	591	Ia
9	GGTTGGTTTGGTTGG	419	Ia
10	gctctcatcgcaacg	354	II
11	GGCTGGTGTGGTTGG	255	Ib
12	gctatcatcgcaacc	220	II
13	GGTTGGTGTGTTTGG	215	Ia
14	gctatcctcgcaacg	199	II
15	GGTTGGCGTGGTTGG	195	Ib
16	gctatcatcgcaacg	160	II
17	GGTTGGTGTTGTTTG	159	Ia
18	gctatcatccaacg	153	II
19	GGTTGGTGTGGCTGG	125	Ib
20	GGTTTGTGTGGTTGG	124	Ia

Top twenty sequences from the 60:1, thrombin: m=15 DNA partitioning experiment of Dr. Kupakuwana.[148] A total of 1,959,748 good reads out of a total 2,231,235 reads. A total of 1,728,220 unique sequences occur in the good reads file. The canonical TBA sequence (rank = 1) is the parent of motif, Ia, where mainly T↔G substitutions occur, and the lower ranking motif Ib, which has T→C and/or G→A substitutions; these motifs are shown in upper case with the variant base underlined. A novel carbohydrate binding sequence and its relatives were identified as motif II, where the highest ranking sequence is referred to as Carb1 (sequences are shown in lower case with variations from Carb1 underlined). The background threshold was identified as 4 counts. 296 unique sequences occur above background. A PCR artifact is labeled as "Jump;" see text for explanation.

Table 2.2 Top 15 sequences, 30:1

Rank	Sequence	Count	Identifier
1	CGGATGCATTTATTC	195	
2	GCCACAGATCGGAAG	40	Jump*
3	TAGTGCATGCGCCAC	10	
4	GGTTGGTGTGGTTGG	10	TBA
5	TTCCGATCTACACGC	8	
6	AGATCGGAAGAGCTC	5	Jump
7	gctatcatcgcaacg	5	Carb1
8	ATAGCGTTCTATCGA	4	
9	TGACAGGAGAGAAAT	4	
10	TTTATCTTCGGTGCG	4	
11	CATTGTATGAGTTTT	4	
12	AACAGAGGCTATAAA	4	
13	AGTCAGTACTTCGGA	4	
14	ATTCGAATCTGCACA	4	
15	TATTGCTTTTCAGGT	3	
16	TGTTAAGGTTAACCT	3	
17	ATCCGGTAGTCACTC	3	
18	GAAGCTTAGTGAACG	3	
19	ATCGGGCGGTTACGC	3	
20	ATATGAGAGGGCTTA	3	

Top twenty sequences from the 30:1, thrombin: m=15 DNA partitioning. A total of 8,721,035 good reads out of a total 8,984,277 reads. A total of 8,639,971 unique sequences occur in the good reads file. The background threshold was identified as 3 counts. Only 7 unique sequences occur above background.

*Jump sequence variant.

Table 2.3 Top 15 sequences, 15:1

Rank	Sequence	Count	Identifier
1	GGTTGGTGTGGTTGG	78	TBA
2	CGGATGCATTTATTC	49	
3	TAGTGCATGCGCCAC	34	
4	gctatcatcgcaacg	25	Carb1
5	GCCACAGATCGGAAG	24	Jump*
6	AGGTGTGGAGCCATT	15	
7	CAGCTACGGAATGAC	8	
8	AATGTTGGATAATGG	6	
9	CGCGCCAGGTTAAAC	5	
10	CTCGTACTGGTATTC	5	
11	TGCAAGTCAAAGAGG	5	
12	TGAACTGGGGAAGAT	4	
13	GCTTTCGAGTAAACG	4	
14	ACCGTATTGAATTCA	4	
15	AGCAATGTATTCTAG	4	
16	GGCACCGTGGTATAA	4	
17	AGTAAATTCCACTGG	4	
18	TGCATGGCCTGACCT	4	
19	GCTATCGGGTCCTTG	4	
20	CTTAGGTCACGCGCA	4	

Top twenty sequences from the 15:1, thrombin: m=15 DNA partitioning. A total of 8,481,474 good reads out of a total 9,168,389 reads. A total of 8,267,894 unique sequences occur in the good reads file. The background threshold was identified as 4 counts. Only 11 unique sequences occur above background.

*Jump sequence variant.

Table 2.4 Top 15 sequences, 3:1

Rank	Sequence	Count	Identifier
1	CGGATGCATTTATTC	111	
2	GGTTGGTGTGGTTGG	56	TBA
3	GCCACAGATCGGAAG	31	Jump*
4	gctatcatcgcaacg	17	Carb1
5	CAGCTACGGAATGAC	12	
6	AGGTGTGGAGCCATT	10	
7	TTCCGATCTACACGC	9	
8	TAGTGCATGCGCCAC	8	
9	AATGGTCTAACAGAG	5	
10	AGTCAGTACTTCGGA	5	
11	AGATCGGAAGAGCTC	5	
12	TGGGCAGTTTACGGC	4	
13	ATGACTTTGTCAAGC	4	
14	TAGCGATCTATATAG	4	
15	AAGCCAATTATGCCG	4	
16	GGTGTTAATGGAGGA	4	
17	CTCAGGATGTAGAAA	4	
18	TGACAGTTCAAGTCT	4	
19	TTCAGTTTGGCACTT	4	
20	GGCCGGGCTGTCAGG	4	

Top twenty sequences from the 15:1, thrombin: m=15 DNA partitioning. A total of 8,484,318 good reads out of a total 8,785,571 reads. A total of 8,322,934 unique sequences occur in the good reads file. The background threshold was identified as 4 counts. Only 11 unique sequences occur above background.

*Jump sequence variant.

Table 2.5 Statistical sequencing data for original 60:1 and variable ratio AIA experiments

Experiment	Original 60:1 AIA	30:1 AIA	15:1 AIA	3:1 AIA
Quantity of Protein	6 nmol	3 nmol	3 nmol	3 nmol
Quantity of Library	100 pmol	100 pmol	200 pmol	1 nmol
Starting Copies per Sequence	56,000	56,000	112,000	560,000
Total Reads	2,142,146	8,984,277	9,168,389	8,785,571

Total Good Reads	1,959,748	8,721,035	8,481,474	8,484,318
Total Unique Sequences	1,728,220	8,639,971	8,267,894	8,322,934
% Unique Sequences Occurring Once	88.2 %	99.1 %	97.4 %	98.1 %
Background/Noise Threshold	4	3	4	4
Sequences Above Background	296	14	11	11
TBA[a]	1_46,444	4_10	1_78	1_56
Frequency of TBA per Good Reads[b]	2.37 %	0.00012 %	0.00094 %	0.00067 %
Frequency of TBA per Total Reads[c]	2.17 %	0.00011 %	0.00085 %	0.00064 %
TBA Variants Above Background	100+	0	0	0
Carb1 Sequence[a]	1_29,405	7_5	4_25	4_17
Carb1 Variants Above Background	50+	0	0	0

Statistical analysis of sequencing data for the original 60:1 experiment and 30:1, 15:1 and 3:1 AIA experiments, m=15 DNA library partitioned against thrombin. **(a)** For canonical TBA, the rank *(left)* and count *(right)* are separated by an underscore. **(b)** Counts of TBA divided by the total good reads. **(c)** Counts of TBA divided by the total reads.

Chapter 3: AIA Improvements and Application of a 2'-OMe RNA/DNA Chimera Library

Chapter Summary

The first section of this chapter focuses on improvements to the AIA method. Issues related to consistency in sample preparation and partitioning efficiency were encountered while aiming to reproduce the original AIA results. Many experiments failed the qPCR and Sanger sequencing quality control tests and were not sequenced. For the three experiments with m=15 hairpin loop library against thrombin that were sequenced, the target sequence (TBA) was found at lower than expected counts. The frequency of these counts within the entire collected data was also much lower than anticipated. Additionally, variants of TBA were not found within the data. To address these issues, a number of control experiments were sequenced and experimental changes were implemented. The m=19 hairpin loop library against thrombin experiment produced a high percentage of jump sequences, indicating substantially poor data quality. The experimental changes were aimed at improving the consistency of sample work-up, including establishing protocols to ensure quality qPCR data as well as minimizing or eliminating the jump sequence artifacts. For the first control experiment, an unpartitioned aliquot of the m=15 DNA library was sequenced. The purpose was to visualize how ligation and PCR affect sequence frequency, specifically the jump sequence, and to identify possible PCR champions. Second, aliquots of the m=15 DNA library negatively selected against Con-A beads from two manufacturers, Pierce Biotechnology and GE Healthcare Life Sciences, were sequenced in order to visualize the effect of negative selection on sequence frequency. Third, the m=15 DNA library was partitioned against thrombin without the negative selection step using Con-A beads from both Pierce

Biotechnology and GE Healthcare. The purpose was to visualize the effects of negative selection on partitioning efficiency.

The first procedural modification was aimed at eliminating the jump sequence. It was hypothesized that gel purifying the final product prior to sequencing may eliminate the jump sequence. The m=15 DNA library was partitioned against thrombin immobilized on Con-A beads from both Pierce Biotechnology and GE Healthcare in duplicate; the first sample in each set was prepared following the standard protocol while the second sample was purified on agarose gels following PCR. The second procedural modification included two changes to the adapter formats; (1) a 12 base portion of Adapter 2 was altered to eliminate sequence similarity between the fragment and a portion of Adapter 1, and (2) introduction of a multiplexing Adapter 1. The goal of altering the 12 base portion of Adapter 2 was to eliminate the potential for binding between Adapter 1 and Adapter 2 that could reduce ligation efficiency. The goal of introducing a multiplexing Adapter 1 was to increase the number of samples that can be clustered per lane of the Illumina flow cell while utilizing the Single Read Illumina format. The third set of changes includes improvements to the ligation scheme; (1) a lower concentration of adapters and (2) overnight incubation at 16°C which resulted in a cleaner gel product. Superior gel purification techniques were also introduced, including (1) a positive ligation control to serve as a size marker for lower concentration samples and (2) use of disposable gel excision tips to reduce cross contamination while excising gel bands. Finally, details including the length of library to protein binding time and use of negative selection were explored.

The second section of this chapter focuses on the introduction of a 2'-OMe RNA/DNA chimera library. RNA aptamers are often preferred over DNA aptamers due to their presumably more versatile secondary structures and wide range of possible modifications. However, standard RNA is susceptible to nuclease degradation and hydrolysis. Modified nucleotides such as 2'-OMe RNA offer increased nuclease resistance and their application in the AIA method is straightforward. The acyclic method eliminates the requirement to regenerate 2'-OMe RNA, a major obstacle in using 2'-OMe RNA in SELEX. Additionally, screening directly with a 2'-OMe RNA as opposed to a post-SELEX 2'-OMe modification eliminates the potential for altered aptamer structure or function. The 2'-OMe RNA/DNA chimera library was designed such that the non-complementary tails and constant stem regions consist of DNA nucleotides and the variable region consists of 2'-OMe RNA nucleotides. This format allows for a simple DNA ligation strategy while retaining the ability to screen for 2'-OMe RNA aptamers. The library was screened against human α-thrombin using several of the procedural modifications mentioned in first section of this chapter.

The third section of this chapter briefly discusses AIA for four Epigenetic protein targets: WDR5, RbBP5, Ash2L and DPY-30. WDR5 (tryptophan-aspartate repeat protein-5), RbBP5 (retinoblastoma-binding protein-5), Ash2L (absent-small-homeotic-2-like), and DPY-30 (Dumpy-30) compose the sub-complex, WRAD, that combines with MLL1 (mixed lineage leukemia protein-1) to form a H3K4 (histone H3 lysine 4) methyltransferase core complex in eukaryotes.[161] The epigenetic maintenance of transcriptionally active states of chromatin in eukaryotes is dependent on the conversion from mono- to dimethylation of H3K4, although the mechanism is not well characterized.[162] When MLL1 is in complex with WRAD, a 600-fold

increase in dimethylation was observed compared to MLL1 alone.[163] The interaction of MLL1 with WRAD is critical for many biological processes, including development, hematemesis, postnatal neurogenesis and tissue homeostasis.[162] Improved understanding of the mechanism of the MLL1 core complex would aid in understanding the role of mixed lineage leukemia proteins in human development disorders.[162] Identifying aptamers for the molecular surfaces of the four sub-units may aid in the development of potential therapeutics. A variety of DNA and 2'-OMe RNA/DNA chimera libraries of length m=15, m=17, m=18, m=19, m=20, m=21 and m=22 were screened against the four targets.

Part 1: Control Experiments and Implementing Procedural Modifications

Control Experiments

Although the initial AIA experiments identified TBA and the Carb1 sequences above background, frequencies were much lower compared to the original 60:1 AIA experiment. The unexpectedly low sequence counts for the variable ratio experiments in combination with the qPCR and Sanger sequencing failures for numerous experiments prompted an examination of the AIA protocol. The first troubleshooting experiment was to create a "baseline" to compare subsequent experiments to. A 100 pmol aliquot of the m=15 library was ligated and PCR amplified as described in chapter 2. The sample passed both the Sanger sequencing and qPCR quality control tests and was sequenced on the Illumina GAIIx. The top 20 sequences are listed in **Table 3.1**.

The second series of troubleshooting experiments was aimed at evaluating the performance of Con-A beads as the protein immobilization method of choice. After monitoring a negative selection step with Con-A beads from Pierce Biotechnology (scaled up to quantities of DNA that could be detected by UV absorption on a Nanodrop spectrophotometer, ~2.0 pmol in 2µl), it was found that the beads were binding most of the DNA library (flow-through was not quantifiable). Although the quantities of library used during aptamer screening are too small to be quantified via UV absorption, it was assumed that only a fraction of the original 100 pmol library would be applied to the immobilized thrombin given this information. This would result in a significantly smaller pool of library with lower initial counts per unique sequence. It would be more difficult to isolate high affinity sequences from the background in under-represented libraries, and this result is counterproductive to the goals of AIA. Dr. Kupakuwana had previously experienced similar problems with the Con-A beads, but they were remedied by purchasing new beads. This was not the case for several different lots of Pierce Biotechnology Con-A beads that were tested. Con-A beads from GE Healthcare Life Sciences were tested in a similar manner and absorbed ~10% of 0.005 µmoles of DNA in 500 µl buffer. This is a more favorable quantity of recovered DNA, however, it could not be determined if the ratio applies for 100 pmol.

Aliquots of the m=15 DNA library (100 pmol) negatively selected against Con-A beads from the two manufacturers, Pierce Biotechnology and GE Healthcare Life Sciences, were sequenced in order to visualize the effect of negative selection on sequence frequency. Following the negative selection step, the library was phenol extracted, ethanol precipitated, ligated and PCR amplified as described in chapter 2. The top twenty sequences for the Pierce Biotechnology beads experiment are listed in **Table 3.2**. Data from the experiment with GE Healthcare beads could

not be reliably analyzed due to a sequencing error. For the third set of control experiments, 10 pmol aliquots of the m=15 DNA library were partitioned against thrombin immobilized on Con-A beads from the two manufacturers, without the negative selection step. The sample was prepared by binding, phenol extraction, ethanol precipitation, ligation and PCR amplification as described in chapter 2. Data for these two experiments could not be reliably analyzed due to a sequencing error and is therefore not shown.

Results and discussion

Of the five control experiments that were sequenced, only data from the unpartitioned m=15 DNA library and negatively selected m=15 DNA library against Con-A beads from the two manufacturers was of sufficient quality to be parsed and compiled using the Perl script. The raw data files were parsed based on the last eight bases of the head (dGCGCATGC) and the first five bases of the tail (dGCATG). A comparison of the highest frequency sequences from the three experiments reveals two sequences that recur in all three files, dCATGGCCAGAGTATA and dAACAGGACCCCATTC. These two sequences occur 3-4 times above background in **Table 3.1** and at least 50 times above background in **Table 3.2**. These sequences have no clear significance and could be contaminants in the original library or sequences that are preferentially amplified during PCR. The three experiments were sequenced on the same GAIIx flow cell and could also be contaminants from sample preparation or qPCR. An Aptamatrix, Inc. customer sample was also run on the same flow cell; however, this was eliminated as a source of contamination because these two sequences occur within the correct complete head and tail regions of the library structure. With the exception of these two sequences and two jump sequence variants occurring 13 and 8 times in **Table 3.1**, the unique sequences occurring in the

data are evenly distributed. Considering the large sequence space of 11,953,616 good reads, this indicates that the library is not skewed in any visible manner. Analysis of the data in **Table 3.2** for the negatively selected library against Pierce Biotechnology Con-A beads reveals that negative selection skews the library pool. It was hypothesized that negative selection would eliminate sequences with a high affinity for Concanavalin-A, while the remainder of the library pool would be equally represented. At a length of m=15, the library contains approximately 1.07 billion unique sequences; eliminating a small fraction of these sequences would not be visible within the experimental sequence space of 11.6 million total reads. The data shows high counts for many sequences, with 42 unique sequences occurring above the background threshold of seven counts. This, in combination with the variability of Con-A beads for absorbing the library during negative selection, led to the decision to eliminate the negative selection step from the AIA protocol. While this would be a problem for traditional SELEX because negative selection must be repeated in each round to prevent substrate-binders from dominating the evolving pool, it is a minor problem in a single round of selection in AIA.

Implementing Procedural Modifications

Gel purification of the final product

The first procedural modification was aimed at eliminating the jump sequence in order to maximize data output. For every jump artifact that is sequenced, one library sequence is not, reducing the quantity of usable data. Although jump sequence variants can be identified and ignored during data analysis, eliminating them would improve data quality. Purifying the final PCR product via agarose gel electrophoresis allows for only the correct size product to be

applied to the flow cell. If the jump sequence artifacts are a byproduct of PCR amplification and run at a different length than the correct products on the gel, they would be eliminated in the final purification step. Also, gel purification eliminates excess primers or primer dimers, which can contribute to loss of data. By design, the forward and reverse primers contain the P5 and P7 sequences required for annealing to the flow cell (see end of **Appendix 2** for sequences). Excess primers occupy valuable space on the oligo lawn, yet fail bridge amplification, therefore reducing the number of possible clusters.

The m=15 DNA library was partitioned against thrombin immobilized on Con-A beads from both Pierce Biotechnology and GE Healthcare in duplicate; the first sample in each set was prepared following the protocol as described in chapter 2 while the second sample was agarose gel purified following PCR. A 60:1 ratio of protein: library (6 nmol thrombin: 100 pmol library) was used for these experiments. The experiments were performed in parallel with the control experiments; therefore, the negative selection step was still performed. In both cases, the partitioning using the Pierce Con-A beads failed Sanger sequencing and qPCR analysis, again indicating an issue with the quality and consistency of the Con-A beads from this manufacturer. The non-gel purified and gel purified samples using the GE Con-A beads were sequenced on two different flow cells on the GAIIx. The top twenty sequences and additional run statistics are provided in **Tables 3.3** and **3.4**.

Although the two experiments were performed under identical partitioning conditions, the frequency of TBA per total good reads is 0.18% (no gel purification) versus 0.0012% (gel purification). Also of note is the drastic difference in frequency of the jump sequence between

experiments. In the non-gel purified sample, jump sequence variants dominate the high counts and a relatively small percentage of total reads were qualified as good reads (5.1%). For the gel purified sample, fewer jump sequence variants are found above background and a higher percentage of reads were qualified as good reads (33.6%). This indicates that the gel purified sample produced higher quality data, even though it was a less stringent partitioning. The difference in TBA frequency is most likely due to variations in the performance of Con-A beads, while the quality of the data is likely due to the final gel purification. Although gel purification did not eliminate the jump sequence, its frequency was greatly reduced and the overall data quality was improved. The next goal was to improve the ligation and purification of the ligated sample in order to improve data quality.

Ligation scheme improvements

The second procedural modification included two changes to the adapter formats; (1) a 12 base portion of Adapter 2 was altered to eliminate sequence similarity between the fragment and a portion of Adapter 1, and (2) introduction of a multiplexing Adapter 1. The Adapter 1 and Adapter 2 sequences used in AIA were based on Illumina's SR adapter structure. In the Illumina model for library preparation, Adapter 1 and Adapter 2 contain a 12 nucleotide complementary region that allows the two strands to anneal and form a "forked adapter." This complex is ligated to both ends of the library via a 3'-T overhang and PCR amplification with appropriate primers fills in the "forked" regions. The ligation scheme used in AIA utilizes four Adapter constructs in order to anneal to the head and tail regions of the library, rather than a single base overhang. Adapter 1 anneals to Adapter 1 Complement and Adapter 2 anneals to Adapter 2 Complement. For this reason, eliminating the "forked adapter" is essential for efficient ligation. By altering this

12 base portion of Adapter 2, the complementarity with Adapter 1 is eliminated, ensuring only Adapter1/Adapter1 Complement and Adapter2/Adapter2 Complement complexes are formed. The changes are outlined in **Figure 3.1** and sequences are listed in **Appendix 3**. Another change made to the adapter format was the introduction of an indexed Adapter 1. A 6 base index was added to the 3'-end of Adapter 1 (with the appropriate complement added to the 5'-end of Adapter 1 Complement) in order to multiplex several samples in a single flow cell lane. The goal was to increase the number of samples that can be clustered per lane of the Illumina flow cell while still utilizing the Single Read Illumina format. The data contains the 6 base index followed by the head, library and tail, respectively. A Perl script then parses the data from each lane based on the indices in the raw data (complete Perl script is available from Dr. Damian Allis, Syracuse University, upon request). Each indexed data file is then parsed with the original Perl script to qualify each sequence as a good, candidate or bad read. The location of the indices is outlined in **Figure 3.2** and a complete list of indices can be found in **Table 3.4**.

The third set of procedural modifications includes improvements to the ligation scheme; (1) a lower volume and concentration of adapters and (2) overnight incubation at 16°C. Superior gel purification techniques were also introduced, including (1) a positive control for ligation to serve as a size marker for lower concentration samples and (2) use of disposable gel excision tips to reduce cross contamination while excising gel bands. The large volume and large quantity of adapters used in the ligation step was cause for concern. At 70 µl, the ligation product required concentration using the QIAquick PCR Purification Kit (Qiagen) in order to load the entire sample in a single electrophoresis well. Eliminating this step by reducing the reaction volume eliminated a potential source of sample loss. In combination with the reduced volume, the

concentration of adapters was also reduced. The ligation efficiency and subsequent PCR amplification of the gel purified products for a set of control partitioning analyses was analyzed for three different reaction conditions: (1) 70 µl volume, original 0.5 nmol of each adapter, (2) varied volume, 3:1 ratio of adapter: library (determined via UV absorption) and (3) 20 µl volume, 0.01 nmol of each adapter. Even after purification with the QIAquick PCR Purification Kit, condition (1) produces overblown adapter bands on the gel, making resolution difficult. For condition (2), quantifying the library is not accurate at the low concentrations that result from partitioning and introduces an opportunity for product loss. It was found that condition (3) produced the cleanest resolution of ligation products via gel electrophoresis and in turn, a larger quantity of the correct PCR product. **Figure 3.3** compares the ligation product of conditions (1) and (3) on separate agarose gels. It was found that 0.01 nmol of each adapter per reaction was sufficient for ligation of 1.0 pmol of m=15 DNA library and resulted in a much cleaner gel product than a ligation with 0.5 nmol of each adapter. These conditions are sufficient to ligate the small quantity of library recovered following a partitioning. Additionally, the ligation temperature and time was changed from room temperature for 30 minutes, to overnight at 16°C as recommended by Promega. In order to ensure that the correct ligated product is gel extracted with minimal contamination, the use of a disposable gel excision tip was introduced. Previously, the region of interest was cut from the gel using a razor blade. Some flexibility was allowed in collecting the correct band due to the nature of working with a razor blade, especially if the product band was not visible due to low concentration. The gel excision tips (GeneCatcher) allow for a 4.0 mm by 1.0 mm gel slice to be collected. With a gel slice of this size, the exact location of the product band must be known. A positive control of 1-10 pmol library ligated with adapters was included on each gel to ensure that a ligation product was visible. The location of

sample bands was extrapolated using the location of the control. The gel tips also minimize the risk of cross contamination and the indexed Adapter 1 allowed possible contaminants from the positive control size marker to be differentiated from the samples following sequencing.

Applying Procedural Modifications to Variable Ratio Partitioning

The previous 60:1 protein: library ratio requires a large quantity of protein (6 nmol on a 100 pmol DNA library scale). Reducing the amount of protein needed increases cost efficiency and conserves materials. Gold *et al.* demonstrated a successful aptamer selection procedure using <100 pmol of various protein targets,[20] at least 60 times less than the AIA procedure, giving confidence that the AIA protocol could be successful in the 20-100 pmol protein range. At a 100 pmol scale, the over-represented m=15 DNA library contains approximately 56,000 copies of each possible sequence (**Table 1.1**). In a 1:1, 100 pmol thrombin: 100 pmol library screen, the ratio of protein to each unique sequence is 1:56,000. Compared to a screen with 1:10 protein: library ratio with 1 nmol library (100 pmol thrombin), the ratio of protein to each unique sequence decreases to 1:560,000. The same effect occurs when the protein quantity is decreased and the library quantity remains constant. It is hypothesized that this decreased ratio would produce a lower quantity of recovered library, and that with more library per protein, a larger, scaled percentage of binding sequences should be higher affinity sequences. Additionally, skewing the ratio from excess protein relative to library toward excess library relative to protein introduces inter-aptamer competition for target binding site. With excess library relative to protein there is no effectice inter-aptamer competition. This may influence the distribution and frequency of aptamer hits. This hypothesis was tested in previous AIA experiments as discussed in chapter 2; however, results were inconsistent. The hypothesis was tested again utilizing the

improvements in sample preparation mentioned in this chapter and a smaller quantity of protein to conserve materials.

Partitioning the m=15 DNA library

The m=15 DNA library was partitioned at ratios of 1:1 (100 pmol thrombin: 100 pmol library), 1:10 (100 pmol thrombin: 1 nmol library) and 1:100 (100 pmol thrombin: 10 nmol library) with no negative selection. Partitioning of the DNA library was performed using agarose-Concanavilin-A beads from GE Healthcare Life Sciences at room temperature. The Con-A beads contained in spin columns were pre-equilibrated in partitioning buffer (20 mM Tris-HCl, 140 mM NaCl, 5 mM KCl, 1 mM $CaCl_2$, 1 mM $MgCl_2$, pH 7.4) by three wash cycles. A 100 pmol aliquot of glycosylated human α-thrombin (Haematologic Technologies) in 750 µl partitioning buffer was immobilized on 100 µl of pre-equilibrated Con-A beads during a 30 minute incubation with end-over-end rotation at 25 rpm. The DNA library in 750 µl partitioning buffer was then applied to the immobilized thrombin. Following an overnight incubation at 4°C with end-over-end rotation at 25 rpm, the unbound DNA was removed by centrifugation. Three wash cycles (5 minutes, rotation at 25 rpm) with 750 µl partitioning buffer removed unbound DNA. Thrombin-DNA complexes were eluted from the Con-A beads with 200 µl elution buffer (α-methyl mannoside, Glycoprotein Isolation Kit, Pierce Biotechnology) during a 5 minute incubation with rotation at 25 rpm and collected by centrifugation.

Phenol extraction and ethanol precipitation

Phenol extraction was used to recover the partitioned DNA libraries. An equal volume of Tris-buffered, pH 8.0, 0.1mM EDTA, 50% phenol, 48% chloroform, 2% isoamyl alcohol (Sigma

Aldrich) was added and vortexed for 30 seconds. Centrifugation at 13,000 rpm for 5 minutes resulted in an aqueous top layer (containing the library) and organic bottom layer. The aqueous layer was removed and the library subsequently extracted twice with equal volumes 100% chloroform. The library was then purified by ethanol precipitation. One tenth the volume (20 µl) of 3M sodium acetate and three times the volume (600 µl) of cold ethanol and were added to the library and mixed briefly. Following an overnight incubation at 20°C, centrifugation at 13,000 rpm in an AccuSpin Micro 17R microcentrifuge (Fisher Scientific) for 30 minutes resulted in a library pellet. After decanting the liquid, the pellet was washed with 1 ml 70% ethanol and pelleted by centrifugation at 13,000 rpm for 20 minutes. Increased precipitation time at -20°C rather than -80°C was implemented in order to improve recovery. The pellet was dried in a SpeedVac and resuspended in 10 µl dH_2O.

Modifying the library for sequencing

The DNA library was ligated with the modified sequencing adapters and their complements using the improved ligation conditions and Indexed Adapter 1. Adapter 1, Adapter 1 Complement, Adapter 2, and Adapter 2 Complement (IDT) at 10 µM were added in 1 µl volumes to the library. The index on Adapter 1 is dictated by the number of samples prepared and the number of samples desired per Illumina flow cell lane. The 1:1, 1:10 and 1:100 samples were ligated with Indices 1, 3 and 2, respectively. The mixtures were incubated at 90°C for 3 minutes. After cooling to room temperature, 2 µl of 10X ligation buffer (300mM Tris-HCl (pH 7.8), 100mM MgCl2, 100mM DTT and 10mM ATP; Promega), 3 µl dH_2O and 1 µl T4 DNA ligase (10mM Tris-HCl (pH 7.4), 50mM KCl, 1mM DTT, 0.1mM EDTA and 50% glycerol; Promega) were added. Incubation of the 20 µl reaction overnight at 16°C completed the ligation.

The product was visualized on a 2% TBE agarose gel alongside a 25 bp DNA ladder. The product is 105 nucleotides long with 66 of those being base paired with Adapter Complements and runs between the 100 and 125 bp ladder bands. The product band was excised with a 4.0 mm by 1.0 mm disposable gel excision tip (GeneCatcher) and purified with the MiniElute Gel Extraction Kit (Qiagen), resulting in a pure, ligated DNA library. PCR amplification with Forward and modified Reverse Primers (See **Appendix 3**) extends the 5'-length of the ligated product to allow for annealing to the immobilized complement of the Illumina flow cell, while simultaneously filling in the complement to the variable region. The 10 µl purified adapter ligated library was combined with 1 µl each of the Forward and modified Reverse Primers at 10 µM (IDT, Coralville, IA), 1 µl of 100 µM dNTPs (Agilent), 2 µl 10X Paq5000 Hotstart DNA polymerase buffer (proprietary to Agilent), 4 µl dH_2O, and 1 µL Paq5000 Hotstart DNA polymerase (Agilent). The 20 µl reaction mixture was amplified with the following conditions:

Denaturation	94°C	2 min.
PCR Amplification (18 cycles)	94°C	30 sec.
	61°C	30 sec.
	72°C	30 sec.
Final Extension	72°C	5 min.

The final product was purified on a 2% TBE agarose gel alongside a 25 bp DNA ladder. The ~130 base pair band was excised with a disposable gel excision tip and extracted with the MiniElute Gel Extraction Kit (Qiagen). The purified library was resuspended in 10 µl dH_2O.

Control experiments

During the experiment, a control ligation was used as a size marker in the event that the ligated experimental sample had a concentration too low to be visualized. A 10.0 or 1.0 pmol aliquot of

the m=15 DNA library was ligated under identical conditions. The product was gel purified alongside the experimental samples and PCR amplified. This product was run alongside the amplified experimental samples to serve as a ~130 bp marker and is illustrated in **Figure 3.4.**

Sanger sequencing

The AIA protocol previously utilized Sanger sequencing prior to sequencing on the GAIIx to ensure the presence of the sequencing cassette: the correctly ligated and modified DNA sequences. This step was time consuming and if a sample failed the Sanger sequencing analysis, it always failed qPCR analysis. This step was ultimately removed from the protocol in favor of only using qPCR analysis.

Quantitative PCR and high throughput sequencing

The precise quantity of amplifiable DNA was determined with the KAPA Library Quantification kit on the BioRad iCycler as described in chapter 2. The concentration of the samples was estimated on the BioTek Synergy 3 Microplate reader and diluted to 20 nM. The quantity of amplifiable DNA in the 20 nM dilution for the three variable ratio experiments (1:1, 1:10 and 1:100 thrombin: library) was determined as 0.6 nM, 4.8 nM and 5.6 nM, respectively. The samples were multiplexed alongside samples from AptaMatrix, Inc. and sequenced on the GAIIx. The 1:1 and 1:100 samples were sequenced on a different lane than the 1:10 sample. Sequencing data was parsed with the Perl script; the last 8 bases of the head and first 5 bases of the tail were used to qualify the data.

Results

Sequencing results for the 1:100 DNA library screen were not viable due to an error in multiplexing. It was determined that a sample from AptaMatrix, Inc. was incorrectly indexed, resulting in a library of the same length and index as the 1:100 experiment in that lane. The top 20 counts and other run statistics for the 1:1 and 1:10 thrombin: library experiments are presented in **Tables 3.6** and **3.7**. A comparison of run statistics is presented in **Table 3.9**. TBA was the highest ranked non-jump related sequence for both experiments and several TBA variants exist above background. The background was estimated based on the presence of "signal," or TBA-like sequences, in comparison to unrelated sequences of the same frequency. For the 1:1 experiment, TBA was the highest counted sequence at 977. With a background count of 3, TBA appears at approximately 325 times background. For the 1:10 experiment, TBA was the highest counted sequence at 5,506, ranking above the jump sequence. With a background of 3, TBA appears at approximately 1,800 times background. Dr. Kupakuwana observed TBA at approximately 10,000 times background; however, the new results are still excellent. The frequency of TBA in the good reads for the 1:1 and 1:10 experiments is 0.06% and 0.18%, respectively, while Dr. Kupakuwana observed TBA at 2.37%. This is a dramatic improvement from the very low frequencies observed in chapter 2. Both the 1:1 and 10:1 data sets contain dozens of TBA variants, and the highest counted sequences are dominated by these variants. The data from these two experiments reflects the ratios of protein: library, as the frequency of TBA and TBA variants increased as the number of starting copies of each sequence increased during partitioning. The top 50 sequences of each experiment were aligned with ClustalX2 and the phylogeny output is shown in **Figure 3.5**. A distinct cluster of TBA like sequences is seen in each tree, indicating that partitioning was efficient.

Neither CarbI nor related sequences were found in either experiment summarized in **Figure 3.5**. Although CarbI was found in the **Table 2.2**, it is at a low count and there are no variants. The absence of CarbI may be a result of switching to Con-A beads from GE Healthcare rather than Pierce Biotechnology. The Pierce Biotechnology storage solution contains glucose, which was identified as a target for CarbI with a K_d of ~1.4 μM.[33] The GE storage solution does not contain glucose which leaves only thrombin as the binding target for CarbI. The CarbI sequence does appear above background for the 60:1 partitioning with GE beads in **Table 3.3**. Just as the counts for TBA vary from experiment to experiment, the behavior should be expected for CarbI. It is likely the case that CarbI was recovered in this experiment as a result of weak binding to glycosylation sites on thrombin. The absence of CarbI in the later experiments was not investigated further, as it only has weak affinity for the glycan content of thrombin and does not have any effect on the clotting activity.

The data quality of the two experiments was excellent, with a total of 95.9% and 96.4 % of the total reads qualified as good reads by analyzing the last 8 bases of the head and first 5 bases of the tail. Also promising is the minimal number of jump sequence variants. Only one jump sequence variant is found above background in the 1:1 experiment at 15 counts. No jump sequence variants are found above background in the 1:10 experiment. This indicates that the overall quality of the sample was high. This can be attributed to improved sample preparation and gel purification of the final product. The jump sequence in the new AIA experiments varies from that in the previous experiments (sequence can be found in **Table 2.1**) due to the sequence change in Adapter 2; however, the location of the jump sequence within the ligated library has

not changed and it is still qualified as a good read based on similarity between the tail and Adapter 2.

Overall, the modifications made to the AIA protocol produced sequence data with a low percentage of jump artifacts and a high percentage of good reads. The partitioning successfully identified TBA and several TBA variants as high affinity sequences well above background. Additionally, the frequency of TBA increased as the selection pressure and number of starting copies of TBA was simultaneously increased.

Part II: 2'-OMe/DNA Chimera Library

While DNA libraries and aptamers are easy to work with *in vitro* and are relatively stable, RNA libraries and aptamers are more susceptible to degradation by nucleases and spontaneous hydrolysis. However, both DNA and RNA aptamers are susceptible to hydrolysis *in vivo*, where the presence of nucleases cannot be controlled. Modified nucleotides such as 2'-OMe RNA offer increased nuclease resistance and their application in the AIA method is straightforward. The acyclic method eliminates the requirement to regenerate 2'-OMe RNA, a major obstacle in using 2'-OMe RNA in SELEX. Additionally, screening directly with a 2'-OMe RNA as opposed to a post-SELEX 2'-OMe modification eliminates the potential for altered aptamer structure or function. For these reasons, we chose to investigate the use of 2'-OMe RNA libraries in the AIA method. The 2'-OMe RNA/DNA chimera library was designed such that the non-complementary tails and constant stem regions consist of DNA nucleotides and the variable region consists of 2'-OMe RNA nucleotides. This format allows for a simple DNA ligation strategy while retaining

the ability to screen for 2'-OMe RNA aptamers. The m=15 2'-OMe RNA/DNA chimera library was synthesized in-house by Dr. Deborah Kerwood on the ABI 394 synthesizer.

The library was screened against human α-thrombin using the improved AIA protocol, similarly to the variable ratio experiments described in this chapter. The m=15 2'-OMe RNA/DNA chimera library was partitioned at ratios of 1:10 (100 pmol thrombin: 1 nmol library) and 1:5 (20 pmol thrombin: 100 pmol library). Following ligation, the 2'-OMe RNA/DNA chimera library was RT-PCR amplified using the SuperScript III One-Step RT-PCR System with Platinum Taq DNA polymerase (Invitrogen). The library was amplified using the following conditions: 25 µl 2X Master mix, 10 µl library, 1 µl 10 µM Forward Primer, 1 µl 10 µM Reverse Primer, 2 µl SuperScript III RT/Platinum Taq Mix.

cDNA Synthesis	50 °C	15 min.
Denaturation	94°C	2 min.
PCR Amplification (18 cycles)	94°C	15 sec.
	61°C	30 sec.
	68°C	1 min.
Final Extension	68°C	5 min.

The forward PCR Primer (IDT) contains an extension sequence that produced a 5'-overhang for annealing to the immobilized complement of the Illumina flow cell. The volume of the PCR and RT-PCR products were reduced to 10 µl with the QIAquick PCR Purification Kit (Qiagen) prior to gel purification. A 10 pmol aliquot of the m=15 2'-OMe RNA/DNA chimera library was ligated and RT-PCR amplified as a positive control alongside each partitioning experiment. qPCR analysis determined the concentration of 20 nM dilution as 1.8 nM for the 1:5 experiment and was sequenced on the Illumina GAIIx. The 1:10 experiment did not amplify well and the qPCR data was unreliable, therefore it was not sequenced.

Results and discussion

The top 20 sequences for the 1:5 m=15 2'-OMe RNA/DNA chimera library screen are reported as DNA sequences as listed in **Table 3.8**. TBA is the highest ranked sequence at 58,781 counts and the frequency of TBA as a percentage of good reads is 3.50%. A large number of TBA variants are also present above background. The top 50 sequences were aligned with ClustalX2 and the phylogeny output is shown in **Figure 3.6.** The figure illustrates that only 10 of the top 50 sequences are non-TBA like. The background count was determined as 13, which places TBA approximately 4,500 times above background. As compared to the variable ratio m=15 DNA library partitioning in **Tables 3.6 and 3.7**, significantly more TBA variants are present in the 2'-OMe RNA/DNA experiment and the counts for these variants drops off much slower. TBA variants are found throughout the entire data set, with several occurring at counts of 1. These are not considered above background, as they are dominated by unrelated sequences of the same count. This is the case up to counts of 13. The increase in counts for TBA and TBA variants in this 100 pmol starting pool compared to the 100 pmol starting pool of the 1:1 DNA experiment may be due to the nature of short, single-stranded RNA, which may fold into more complex three dimensional structures than DNA and lead to high affinity aptamers beyond the G-quadruplex structure of TBA. However, the high selectivity for the TBA sequence is consistent with this hairpin molecule folding into the quadruplex in the 2'-OMe-ribose context. Additionally, the pattern of TBA variant sequences underlined in **Table 3.8** (2'-OMe) is very different from that in the DNA shown in **Table 2.1** (DNA). The 2'-OMe variants from canonical TBA are more tolerant of substitution in the central TGT sequence and T,G $\rightarrow$ C,A substitutions. These differences suggest subtle differences in the details of how 2'-OMe-TBA binds to thrombin compared to canonical DNA-TBA. The high frequency of TBA and large number of TBA

variants occurring well above background indicate that the partitioning was highly selective. CarbI and related sequences are not present in the data set, again due perhaps to the absence of glucose in the storage buffer for Con-A beads.

In addition to exceptional selection efficiency, the experiment yielded high quality data. Approximately 69.2% of the total reads were qualified as good reads, and of these only two variants of the jump sequence are present above background, the most abundant at 54 counts. The data quality can be attributed to improved sample preparation and gel purification of the final product. A comparison of the 2'-OMe RNA/DNA chimera library screen, 1:1 and 1:10 screens, as well as the 30:1 screen from chapter 2 and the original 60:1 experiment from Dr. Kupakuwana can be found in **Table 3.9.**

The application of an m=15 2'-OMe RNA/DNA chimera library in the AIA for thrombin protocol yielded exceptional results. Although it is reported as DNA, the data represents 2'-OMe RNA in the m=15 variable region. The 1:5 experiment produced TBA as the most abundant sequence, a large number of TBA variants, and few jump sequences. This experiment demonstrates that a 2'-OMe RNA library is well suited to the AIA method. Additionally, this experiment demonstrated a successful partitioning utilizing only 20 pmol of thrombin, a quantity significantly less than the 6 nmol used in the original AIA method. Additional experiments with the m=15 DNA and m=15 2'-OMe RNA/DNA chimera library under identical conditions would be required in order to directly compare the performance of the two libraries. These experiments, along with additional variable ratio experiments that would be needed to confirm the findings of

the 1:1 and 1:10 experiments, were not performed in favor of exploring the new library structure described in chapter 4.

Part III: AIA for Epigenetic protein targets

The AIA protocol was adapted for four epigenetic protein targets, WDR5, RbBP5, Ash2L and DPY-30, provided by Dr. Michael Cosgrove, SUNY Upstate Medical University, Syracuse, NY. With the exception of the following, the modified AIA protocol described in this chapter was followed. His-tag modified forms of the proteins were immobilized on Dynabeads His-Tag Isolation and Pull down magnetic beads (Invitrogen). Protein variants without his-tags were biotinylated with the ChromaLink Biotin Labeling Kit (SoluLink) and immobilized on NanoLink Streptavidin magnetic beads (SoluLink). The binding capacity of the beads, (60 pmol/µl and 20 pmol/µl, respectively) was used in an effort to saturate the beads and eliminate the negative selection step. The his-tagged proteins were stored in a TCEP buffer (20 mM Tris, pH 7.5, 300 mM NaCl, 1 µM $ZnCl_2$, 1 mM TCEP) and were bound to beads equilibrated in this solution. The immobilized protein was washed 3X with the TCEP buffer, washed a final time with SBIT buffer (40 mM HEPES, pH 7.5, 125 mM NaCl, 5 mM KCl, 1 mM $MgCl_2$, 1 mM$CaCl_2$, 0.05% TWEEN-20)[20] and resuspended in SBIT buffer. The biotinylated proteins were buffer exchanged during the biotinylation procedure and were stored in 1X PBS, which was used in place of the TCEP buffer. Following an overnight partitioning of the library against the immobilized protein with 0.5% BSA, the beads were washed 2X with SBIT buffer containing 0.1 mg/ml salmon sperm DNA (Invitrogen), and an additional 2X with SBIT buffer.[20] His-tagged protein-library complexes were eluted with 2M Guanidine HCl in SBIT buffer for 30 minutes at

room temperature. Biotinylated protein-library complexes were eluted with 8M Urea for 3 minutes at 94°C.

Although a multitude of protein target and library combinations were explored, only a fraction were sequenced. The m=15, m=17, m=19 and m=21 2'-OMe RNA/DNA chimera library against his-tagged RbBP5, his-tagged Ash2L and his-tagged DPY-30 as well as the m=17, m=19 and m=22 DNA libraries against biotinylated RbBP5 were sequenced. The m=15 library was screened with approximately 56,000 starting copies of each sequence. The m=17, m=19 and m=21 and libraries were screened with approximately 10,000 starting copies of each sequence. The m=22 library was screened with approximately 300 starting copies of each sequence. The ratio of target: library for the 2'-OMe RNA/DNA chimera libraries was 1:10 with varied quantities of protein to accommodate the copy numbers above. The ratio of target: library for the DNA libraries was 1:3, 1:50, and 1:80 for the m=17, m=19 and m=22 libraries, respectively. None of the experiments produced sequences above background, indicating that no high affinity sequences were present in the starting pools and/or the partitioning was not stringent enough to isolate them from the background. Additional experiments with varied partitioning conditions may reveal whether or not the stringency could be increased in order to isolate any aptamer candidates. However, these experiments were not performed in favor of exploring the new library structure described in chapter 4.

Figure 3.1 Illumina SR Adapters versus AIA Adapters

(Top) Illumina SR Adapter sequences, 12 base complementary regions highlighted in red. When annealed, the Adapters form a forked structure with a single T overhang for ligation. The Adapter complex ligates to both the 5' and 3'-ends of the adenylated library. **(Middle)** AIA Adapter 1 (Index 1) and Adapter 2 sequences, 12 base complementary regions highlighted in red.

The complementary regions allow the Adapters to form the forked structure, which competes with the desired Adapter1/Adapter 1 Complement and Adapter 2/Adapter 2 Complement structures. **(Bottom)** AIA Adapter 1 (Index 1) and revised Adapter 2 sequences, original complementary regions highlighted in red, revised bases highlighted in blue. In the revised Adapter 2 sequence, 9 of the 12 complementary bases are altered, preventing the forked structure shown. This eliminates competition with the desired Adapter1/Adapter 1 Complement and Adapter 2/Adapter 2 Complement structures required for ligation of the library.

Figure 3.2 Location of AIA Adapter 1 Index

Adapter 1 and Adapter 1 Complement annealed, with 12 base overhang required for ligation to the library. The index required for multiplexing with the SR format is positioned at the 3'-end of Adapter 1. The sequencing primer sits in front of the index. During data processing, the first 6 bases represent the index and are parsed independently of the head, library and tail. The index files are then processed with the original Perl script.

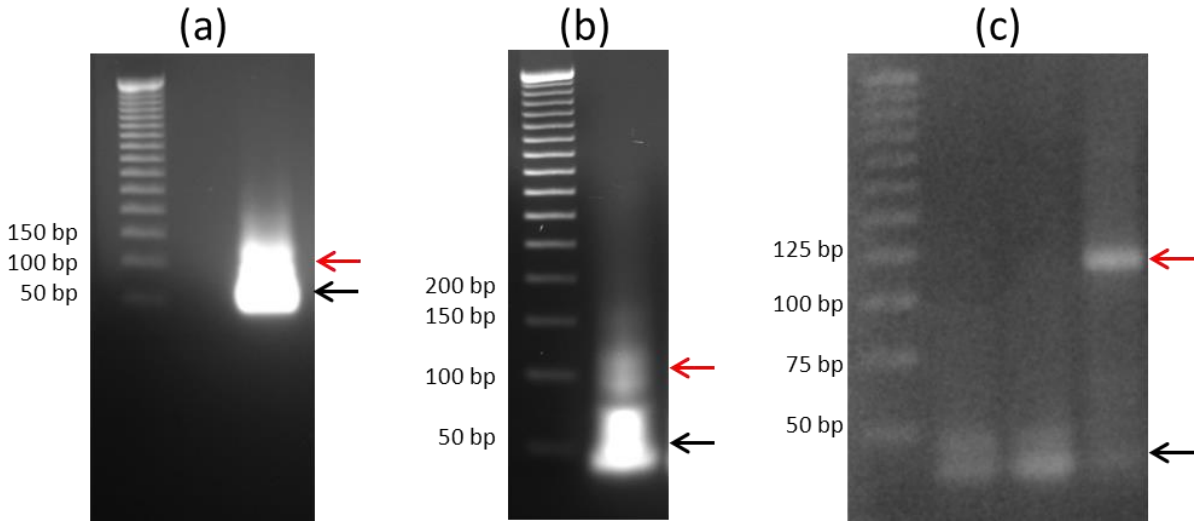


Figure 3.3 Original versus modified ligation conditions

2% Agarose gels of ligated m=15 DNA hairpin loop libraries after partitioning. Black arrows indicate location of excess Adapters. Red arrows indicate ligated product. **(a)** Lane 1: 50 bp DNA ladder. Adapters at original concentration, total quantity is 0.5 nmol of each Adapter. This gel demonstrates the difficulty in purifying the ligated product from excess Adapters if the gel is not run for a sufficient time. The excess Adapter band is overblown and interferes with the ligated product. **(b)** Lane 1: 50 bp DNA ladder. Adapters at original concentration, total quantity is 0.5 nmol of each Adapter. This demonstrates separation of ligated product from excess Adapters. Adapter bands are overblown and ligated product does not run in a clean band. **(c)** Lane 1: 25 bp DNA ladder. Modified ligation conditions. Total quantity of Adapters is 0.01 nmol each, 16°C overnight incubation. Lane 4 is a positive control, 10 pmol of library ligated. Note the conversion of free Adapters to ligated product. Lanes 2 and 3 are partitioned samples, ligated product bands are present but difficult to visualize without adjusting image contrast. Ligated products are clearly separated from excess Adapters. Adapter bands are not overblown and ligated product runs as a concise band.

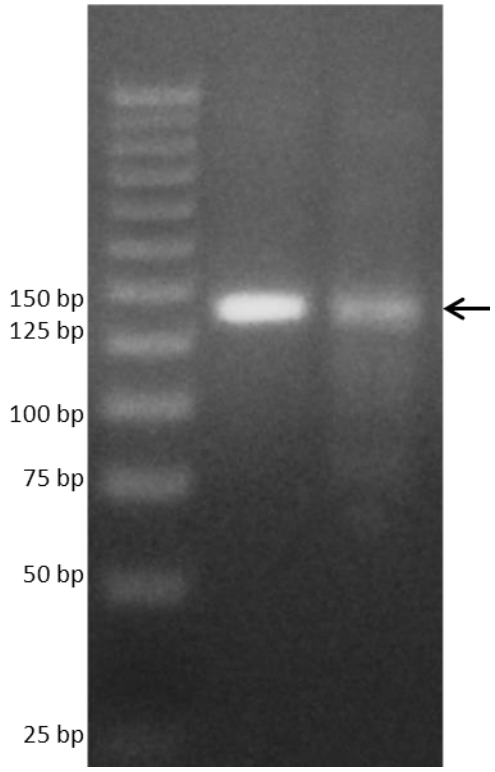


Figure 3.4 PCR product of ligated m=15 DNA hairpin loop library

2% Agarose gel of PCR amplified ligated m=15 DNA hairpin loop library. Lane 1: 25 bp DNA ladder. Lane 2: Positive control, PCR product of gel purified 10 pmol ligated library. Lane 3: PCR product of gel purified ligated partitioning sample. Arrow indicates the location of the correct product bands.

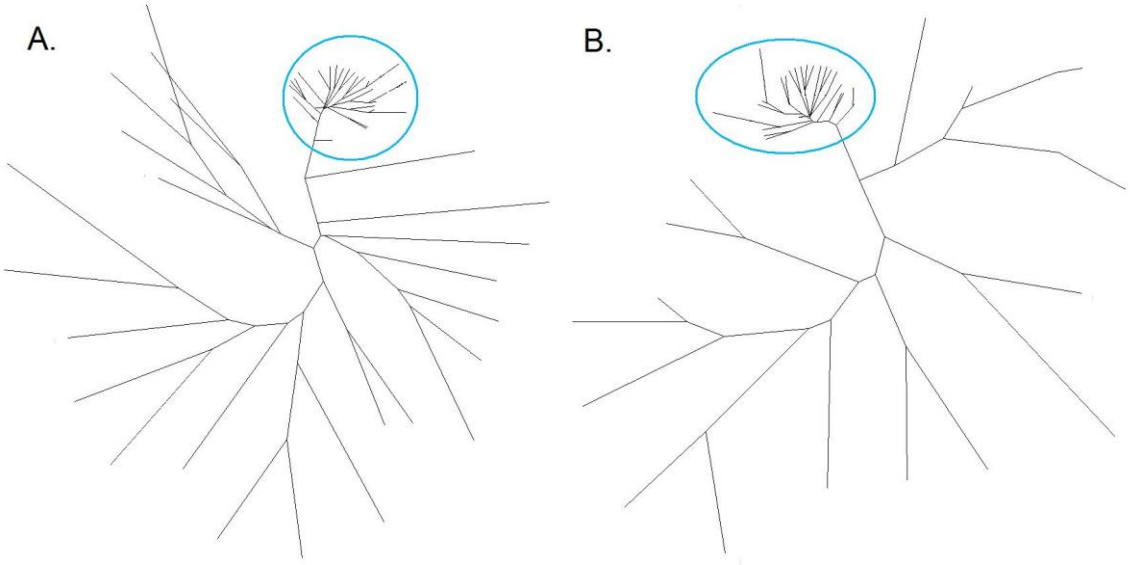


Figure 3.5. Phylogeny trees of the top 50 sequences for the 1:1 and 1:10 ratio experiments

Phylogeny trees created following sequence alignment with ClustalX**. (A)** Top 50 sequences, 1:1 thrombin: m=15 DNA library. TBA-like sequences circled in blue. (**B)** Top 50 sequences, 10:1 m=15 DNA library: thrombin. TBA-like sequences circled in blue.

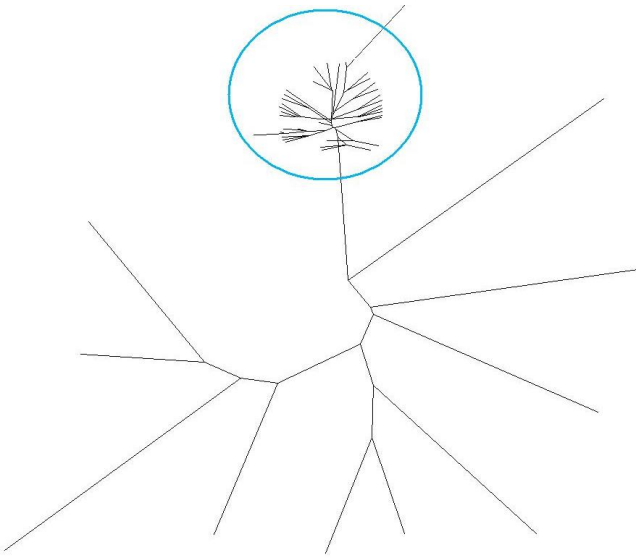


Figure 3.6 Phylogeny tree of the top 50 sequences for the 1:5 2'-OMe RNA/DNA chimera library experiment

Phylogeny tree created following sequence alignment with ClustalX**.** Top 50 sequences, 1:5 thrombin: m=15 DNA library. TBA-like sequences circled in blue. Only 10 of the top 50 sequences are unrelated to TBA.

Table 3.1 Top 20 sequences, m=15 DNA library

Rank	Sequence	Count	Identifier
1	GCCACAGATCGGAAG	13	Jump*
2	CATGGCCAGAGTATA	12	
3	AACAGGACCCCATTC	10	
4	ATGCGCCACAGATCG	8	Jump*
5	ACTGCATGCGCCANN	4	
6	GAGGGTACATCCGGG	4	
7	CATAACGAAAGCACT	4	
8	TGATCTTCGCACAAT	4	
9	ATGTCCTTACACGGC	4	
10	GGGTTGGTTGCACTA	3	
11	AGTACAAGGCACGCA	3	
12	GGTCTATAGCCCCTT	3	
13	ATGGGTCTACTTGTT	3	
14	CATGCTCAGTAAGAA	3	
15	GGTTTGGCTTAGGGT	3	
16	CTAATAGAGAAGTCA	3	
17	GTAGACGGGTCATTG	3	
18	ATCATCGCATAACAT	3	
19	CGGCATATTGTTTCT	3	
20	GACATGGCGCAGTAT	3	

Top twenty sequences from a 100 pmol aliquot of the unpartitioned m=15 DNA library. A total of 11,953,616 good reads out of a total 12,940,575 reads. A total of 11,864,123 unique sequences occur in the good reads file. The background threshold was identified as 3 counts. 9 unique sequences occur above background.

*Jump sequence variant.

Table 3.2 Top 20 sequences, m=15 DNA library negatively selected against Pierce Biotechnology Con-A beads

Rank	Sequence	Count	Identifier
1	AACAGGACCCCATTC	460	
2	ACGATGGGATAGGAA	391	
3	CGGAAGAGCGGTTCA	280	Jump*
4	CATGGCCAGAGTATA	260	
5	CACATAGGATAACCT	186	
6	CGCGCTACGGTGGTG	152	
7	TCACGCTCCTACTCG	107	
8	CCACATCATAACGAT	94	
9	GTTCTAATAAGTCGC	89	
10	CATGCGCCACAGATC	77	Jump*
11	CATGCGCCACCGATC	52	Jump*
12	ATGCGCCACAGATCG	35	Jump*
13	CTAGCAGCTCATCAT	34	
14	ACGTAGGGTCGCCGC	25	
15	TTGCGGACGAGCCTA	24	
16	AAGTAACGCTCAGGC	23	
17	CCGTAGAGCATCCAA	23	
18	GCTATCATCGCAACG	23	
19	AGGCTTGACTCAGTT	21	
20	CTGTCGGTCAGGGAT	16	

Top twenty sequences from a 100 pmol aliquot of the m=15 DNA library negatively selected against Con-A beads from Pierce Biotechnology. A total of 7,281,704 good reads out of a total 11,672,270 reads. A total of 6,216,093 unique sequences occur in the good reads file. The background threshold was identified as 7 counts. 42 unique sequences occur above background.

*Jump sequence variant.

Table 3.3 Top 20 sequences, m=15 DNA library partitioned against thrombin, no gel purification

Rank	Sequence	Count	Identifier
1	AGATCGGAAGAGCTC	119007	Jump*
2	GGTTGGTGTGGTTGG	2054	TBA
3	CGATCGGAAGAGCTC	1075	Jump*
4	AGATCGGANGAGCTC	1032	Jump*
5	AGATCGGAANAGCTC	910	Jump*
6	AGATCGGAAGGGCTC	667	Jump*
7	GTTAGCCATTAGTTT	654	
8	AGATCGGAAGCGCTC	577	Jump*
9	AGATCGGAACAGCTC	386	Jump*
10	AGATCGGAGGAGCTC	380	Jump*
11	AGATCGGAAGAGATC	322	Jump*
12	AGATCGGAAGAGCTA	307	Jump*
13	GAGGGTCGAGGATTA	288	
14	GCTATCATCGCAACG	285	Carb1
15	AAAGAGGCTAGCTTG	281	
16	AGATCGGACGAGCTC	272	Jump*
17	CGGATGCATTTATTC	250	
18	AGATAGGAAGAGCTC	246	Jump*
19	TATCGCTATAGAATG	223	
20	AGATCGGATGAGCTC	222	Jump*

60:1, GE Con-A beads. A total of 1,135,020 good reads out of a total 22,145,341 reads. A total of 597,402 unique sequences occur in the good reads file. The background threshold was identified as 11 counts. 197 sequences occur above background**.** Several of the top hits are variations of the jump sequence.

Table 3.4 Top 20 sequences, m=15 DNA library partitioned against thrombin, gel purification of final product

Rank	Sequence	Count	Identifier
1	AACAGGACCCCATTC	946	
2	CATGGCCAGAGTATA	295	
3	CCACATCATAACGAT	235	
4	TTGCGGACGAGCCTA	214	
5	GTTCTAATAAGTCGC	190	
6	ACGATGGGATAGGAA	146	
7	TCACGCTCCTACTCG	87	
8	CACATAGGATAACCT	49	
9	CCGTAGAGCATCCAA	44	
10	AAGTAACGCTCAGGC	35	
11	GGTTGGTGTGGTTGG	34	TBA
12	AAAGAGGCTAGCTTG	27	
13	CGCGCTACGGTGGTG	24	
14	TGTCTATGGGGACTT	23	
15	CTCGACACAAGTTGA	21	
16	TATCAGATCGGAAGA	16	
17	CGCTCAAGTCAAATC	12	
18	AGGAGTACACGCATG	12	
19	ACCAGGACCCCATTC	11	
20	TAGACAACACGTTAG	10	

60:1, GE Con-A beads, gel purified PCR product. A total of 2,871,584 good reads out of a total 8,536,557 reads. A total of 5,683,961 unique sequences occur in the good reads file. The background threshold was identified as 8 counts. 25 sequences occur above background**.**

Table 3.5 List of Adapter 1 Indices

Index	Sequence
1	ATCGTA
2	GATAGC
3	CGATCT
4	TCGCTA
5	TACGTC
6	AGTAGT
7	GCATGA
8	CTGCTG
9	TCAGTA
10	ATGAGC
11	GACTAT
12	CGTCCA
13	TCAGAG
14	ATGCGA
15	GACTCG
16	CGTATC

List of twelve Adapter 1 Indices used for multiplexing compatible with the Illumina SR flow cell format.

Table 3.6 Top 20 sequences, 1:1 thrombin: m=15 DNA library

Rank	Sequence	Count	Identifier
1	GGTTGGTGTGGTTGG	977	TBA
2	AGTTGGTGTGGTTGG	262	
3	TGTTGGTGTGGTTGG	22	
4	AGACAGAGGTCAGTC	15	Jump*
5	GGTTGGTATGGTTGG	13	
6	GGTTGATGTGGTTGG	11	
7	GGTTGGTGTGATTGG	10	
8	GGTTGGTGAGGTTGG	9	
9	GGTTGGTGTGGTTTG	9	
10	GGTTGGTGGGGTTGG	9	
11	GGTTGGTGCGGTTGG	8	
12	GGTTGGGGTGGTTGG	8	
13	GGTTGGTGTGGCTGG	7	
14	AGTTGATGTGGTTGG	7	
15	GATTGGTGTGGTTGG	6	
16	GGTCGGTGTGGTTGG	6	
17	GGTTGGTGTGGTTGA	6	
18	GGTTAGTGTGGTTGG	6	
19	GGTTGGTGTGGTTGT	6	
20	GGTTGGTGTGGTTAG	6	

1:1 thrombin: m=15 DNA library (100 pmol thrombin: 100 pmol library). A total of 1,520,709 good reads out of a total 1,568,269 reads. A total of 1,503,898 unique sequences occur in the good reads file. The background threshold was identified as 3 counts. 31 sequences occur above background**.** Several of the top hits are variations of TBA, with 26 variants above background.

*Jump sequence variant

Table 3.7 Top 20 sequences, 1:10 thrombin: m=15 DNA library

Rank	Sequence	Count	Identifier
1	GGTTGGTGTGGTTGG	5506	TBA
2	GGTTGGGGTGGTTGG	57	
3	AGTTGGTGTGGTTGG	33	
4	GGTTGGCGTGGTTGG	32	
5	GGTTAGTGTGGTTGG	27	
6	GGTTGGTGGGGTTGG	26	
7	GGTTGGTGCGGTTGG	22	
8	GGTTGGTGTGGTTAG	22	
9	GGTTGGTGAGGTTGG	21	
10	NNNTNCNNNNNGNAN	18	
11	GGTCGGTGTGGTTGG	16	
12	GGTTGGTGTGATTGG	16	
13	GATTGGTGTGGTTGG	14	
14	GGTTTGTGTGGTTGG	13	
15	GGTTGGTGTGGTCGG	13	
16	GGTTGGTATGGTTGG	13	
17	GGTTGGTGTAGTTGG	13	
18	GGTTGGTGTGGTTTG	12	
19	NNNTNCTNNNNGNAA	12	
20	NNNNNCNNNNNGNAN	11	

1:10 thrombin: m=15 DNA library (100 pmol thrombin: 1 nmol library). A total of 3,015,490 good reads out of a total 3,127,147 reads. A total of 2,981,529 unique sequences occur in the good reads file. The background threshold was identified as 3 counts. 104 sequences occur above background**.** Several of the top hits are variations of TBA, with 38 variants above background.

*Jump sequence variant

Table 3.8 Top 20 sequences, 1:5 thrombin: m=15 2'-OMe/DNA chimera library

Rank	Sequence	Count	Identifier
1	GGTTGGTGTGGTTGG	58781	TBA
2	GTTTTGTTTGGTTGG	391	
3	GGTTGGGGTGGTTGG	345	
4	GGTTGGTGGGGTTGG	175	
5	GGTTGGTGCGGTTGG	118	
6	GGTTGGCGTGGTTGG	116	
7	GGTTGGAGTGGTTGG	90	
8	GGTTGGTGTGGCTGG	70	
9	GGTTGGTGTGGTCGG	57	
10	AGACAGAGGTCAGTC	54	Jump*
11	GGTTGGTGAGGTTGG	53	
12	GGATGGTGTGGTTGG	39	
13	GGTTGGTGTGGTTGA	38	
14	GGTTGGGGGGGTTGG	37	
15	GGCTGGTGTGGTTGG	33	
16	GGTTGGTGTGGTTCG	32	
17	GCTTGGTGTGGTTGG	31	
18	GGTTGGTGTGGTTGC	31	
19	GGTCGGTGTGGTTGG	30	
20	GGTTGGTGTGGTTAG	30	

1:5 thrombin: m=15 2'-OMe RNA/DNA chimera library (20 pmol thrombin: 100 pmol library). Data is reported as DNA but represents the 2'-OMe RNA analog. A total of 1,681,274 good reads out of a total 2,431,077 reads. A total of 66,934 unique sequences occur in the file. The background threshold was identified as 13 counts. 160 sequences occur above background**.** Several of the top hits are variations of TBA, with 42 variants above background; underlined bases in variants differ from canonical TBA.

*Jump sequence variant.

Table 3.9. Statistical sequence data for improved variable ratio and 2'-OMe RNA AIA experiments

Experiment	Original 60:1 AIA	1:1 AIA	1:10 AIA	1:5 2'-OMe
Quantity of Protein	6 nmol	100 pmol	100 pmol	20 pmol
Quantity of Library	100 pmol	100 pmol	1 nmol	100 pmol
Starting Copies per Sequence	56,000	56,000	560,000	56,000
Total Reads	2,142,146	1,568,269	3,127,147	2,431,077
Total Good Reads	1,959,748	1,520,709	3,015,490	1,681,274
Total Unique Sequences	1,728,220	1,503,898	2,981,529	667,934
% Unique Sequences Occurring Once	88.2 %	99%	99%	39%
Background/Noise Threshold	4	3	3	13
Sequences Above Background	296	31	104	60
TBA[a]	1_46,444	1_977	1_5,506	1_58,781
Frequency of TBA per Good Reads[b]	2.37 %	0.062 %	0.18 %	3.50 %
Frequency of TBA per Total Reads[c]	2.17 %	0.062 %	0.17 %	2.42 %
TBA Variants Above Background	100+	26	38	42
Carb1 Sequence[a]	1_29,405	N/A	N/A	N/A
Carb1 Variants Above Background	50+	N/A	N/A	N/A

Statistical data for the original 60:1 experiment, 1:1 and 1:10 variable ratio experiments and the 1:5 2'-OMe RNA/DNA chimera library experiments. **(a)** The rank *(left)* and count *(right)* are separated by an underscore. **(b)** Counts of TBA divided by the total good reads. **(c)** Counts of TBA divided by the total reads.

Chapter 4: AIA with Adapter Libraries

Chapter Summary

The hairpin loop library design described in chapters 2 and 3 was used to successfully identify the thrombin binding aptamer from pools of m=15 DNA and m=15 2'-OMe RNA/DNA chimera libraries. Consistency in sample preparation and data quality were achieved with a number of procedural modifications. However, the library design was unsuccessful at identifying any aptamer sequences for the Epigenetic targets mentioned in chapter 3 from over-represented libraries of varied lengths. It is possible that no aptamer candidates existed in the pools, however, the limiting nature of the hairpin loop design may have been a factor. The library design explored in this chapter eliminated the constraints of the hairpin loop structure. In place of the eight base complementary stem and 4 base non-complementary tails, the variable regions was flanked by portions of the Illumina TruSeq Adapter 1 and Adapter 2 constructs. This library design is employed similarly to SELEX libraries with the Adapter regions serving as priming regions in PCR. By eliminating the ligation step and proceeding directly to PCR amplification, the AIA method becomes more efficient and even less time consuming. Additionally, retention of sequences selected from libraries is improved. An unknown percentage of selected library sequences is likely lost during ligation due to partial or failed ligations. It may also be possible to "capture" the bound library molecules off of the immobilized protein target by PCR without phenol extraction and ethanol precipitation, further simplifying the AIA protocol. This library based on direct amplification is termed an "adapter library." The work-up and experimental details for both DNA and 2'-OMe RNA/DNA chimeric adapter libraries were determined in simulated partitioning using dilute concentrations of library. Following the determination of

optimal PCR amplification conditions, the libraries were screened against thrombin to gauge their functionality in AIA. Although written in this order, several experiments were completed simultaneously and therefore conclusions were made simultaneously. Some experiments may seem redundant when presented in the order written, however, this manner allows for an organized presentation of methods and results.

Library Design

The sequence arrangement for the adapter library design is illustrated in **Figure 4.1**. The variable region of the adapter libraries ranges from m=15 to m=22 which allows for a sufficient number of copies per sequence for a single round of partitioning. A 2'-OMe RNA/DNA chimera adapter library was also designed, with a 2'-OMe RNA library region and DNA adapters, which is RT-PCR amplified following partitioning similarly to the hairpin loop ligation library. The adapter sequences that flank the library region are based on the Illumina TruSeq adapter format. Illumina's TruSeq technology is in improvement on the Single Read versus Paired End Read versus Multiplex Paired End Read sequencing technologies. The three older technologies employed unique adapter constructs that were not interchangeable. Single Read samples could not be sequenced on Paired End Read flow cells and vice versa. The Single Read format operated as described in chapter 2, with the immobilized oligo sequenced from the top down. The Paired End Read format sequenced the oligo from the top, down, and then from the bottom, up. The multiplexed Paired End format utilized an index within Adapter 2. With TruSeq technology, Single Read, multiplexed Single Read, Paired End Read, or multiplexed Paired End Read formats use the same adapter constructs and flow cell chemistry. This allows greater

flexibility in planning and executing sequencing runs. The full length TruSeq adapter sequences and modified primer sequences for use with the adapter libraries are listed in **Appendix 4**. In the adapter libraries, the last 18 bases of TruSeq Adapter 1 flank the 5'-terminus of the variable region. The first 28 bases of TruSeq Adapter 2 flank the 3'-terminus of the variable region, with bases 2 through 9 modified to include an internal index. Indicated in green in **Figure 4.1** is the eight base internal index that was designed to identify individual libraries. During a multiplexed sequencing run, libraries of different lengths, i.e. m=15 vs. m=21, can easily be separated during data processing in the event of a multiplexing error due to the six base difference. When the difference in length is smaller, i.e. m=15 vs. m=17, separating the data can be more difficult. A sequence of length m=16 could be either an m=15 library with single base insertion or a m=17 library with single base deletion. Without the internal index, it would be impossible to identify the origin of said 16-mer and the sequence would necessarily be discarded. Of greater importance is the need to differentiate DNA and 2'-OMe RNA libraries of the same length, as well as fixed sequences used as controls in a number of experiments. By assigning a different index to each library or fixed sequence, contaminating sequences can be eliminated from the data set.

Workflow Summary

The adapter libraries are modified for sequencing using a two-step amplification scheme. The amplification scheme was designed as a two-step process in order to achieve a cleaner product. The quantity of recovered library is typically too low to quantify via UV absorbance, therefore the required number of amplification cycles is unknown. In SELEX, aliquots of recovered library

are often amplified with varied cycle numbers to determine the optimum number of cycles for preparative PCR without nonspecific products.[164] Although this practice reduces the amount of library in the recovered pool, the cyclic enrichment of SELEX eliminates the concern over losing a fraction of the library. In AIA, it is critical to retain the entire recovered library because high affinity sequences cannot be enriched in subsequent selection cycles. Following the first amplification step, the library can be accurately quantified and the concentration of template can be controlled for the second round of amplification. This would allow for a predetermined number of amplification cycles in order to produce a predictable quantity and quality of product. The products of each amplification step are illustrated in **Figure 4.2** and all primer sequences are listed in **Appendix 4.** The first primers are termed CAP1 and CAP2 for "capture" primers, used to capture the surviving library molecules following partitioning. During amplification, the CAP primers fill in the gap across from the library region while adding minimal length to the product: 15 bases to the 5'-termini and 6 bases to the 3'-termini which provide priming locations of similar melting temperature for the second round of PCR. The final product for an m=15 adapter library is 82 base-pairs in length. Following gel purification of the CAP amplification adapter library, a second round of amplification with Primer 1 (P1) and Primer 2 (P2) adds two important features. Indicated in purple and orange in **Figure 4.2** are regions complementary to oligonucleotides on the Illumina flow cell for bridge amplification. Primer 2 also adds an additional six base index termed the "Illumina index," the underlined portion of Adapter 2 in **Figure 4.2**. Each Primer 2 is termed for the appropriate index (Primer 2.1, Primer 2.2, etc.) and the sequences are taken directly from the Illumina TruSeq Adapter indices. This index allows multiple samples of the same library (with identical internal eight base index) to be sequenced on the same flow cell lane. A complete list of P2 Illumina Indices is found in **Table 4.2.** It was

originally thought that using full length primers to achieve the final 137 base-pair product in one step would be difficult due to the comparably small complementary regions between the primers and adapter library. It was later found that a one-step amplification scheme using the full length primers functions well. However, it was not pursued due to the results mentioned in future sections.

In addition to the implementation of the TruSeq adapter format, a new sequencing system was used for the experiments in this chapter. The Illumina MiSeq Personal Sequencer is a bench top next-generation sequencer with faster sample prep, sequencing, and data analysis times than the Illumina GAIIx. Compared to the eight lane flow cell of the GAIIx, the single lane flow cell of the MiSeq requires fewer samples to financially justify a sequencing run. This is beneficial when developing a new library design because sequence data can be analyzed for a single lane (typically 1-12 samples) before valuable materials are used in filling an entire GAIIx flow cell (typically 7-64 samples). This was a major obstacle during troubleshooting and experimental sample preparation in chapters 2 and 3. It was often the case that data from a single experiment was needed in order to dictate changes made to the subsequent experiments. However, sequencing a single sample on a GAIIx flow cell was not financially responsible. Sequencing fewer samples at a time would allow capitalization on the findings of previous experiments in a more efficient manner. Prior to sequencing on the MiSeq, we attempted to modify the GAIIx to function similarly to a MiSeq. With the assistance of Dr. Borer and Dr. Mark McPike, we attempted to adjust the microfluidics and reagent volumes of the GAIIx to run the eight lanes independently of each other. We installed eight individual valves on both the cluster station and GAIIX that allowed for reagents to be directed to any number and combination of lanes. We

were successful in reprogramming the automated clustering and sequencing programs to deliver appropriate reagent volumes depending on the number of lanes in use. Although we were able to cluster and sequence the lanes independently from one another, data produced from GAIIx runs with fewer than eight lanes was unusable. The pump assemblies and reagent chemistry were designed for use with all eight lanes and performed very poorly when used otherwise, producing very low cluster densities. The project was abandoned in favor of using the Illumina MiSeq at SUNY Upstate's Microarray core facility, run by Frank Middleton, Ph.D. and Karen Gentile. \

Prior to sequencing, the precise quantity of amplifiable DNA was determined with the KAPA Library Quantification kit on a BioRad iCycler. A 20 nM dilution was made from the final gel purified product and used for qPCR quantification. If a sufficient quantity of amplifiable DNA (5-20 nM in 10 µl) was confirmed by qPCR, the samples qualified for sequencing. Depending on the kit version, the Illumina MiSeq requires 10 µl of 2 nM sample DNA or 5 µl of 4nM DNA. Depending on the number of samples per sequencing run and starting concentration, the volume of each sample varies in order to achieve equal representation of multiplexed samples. During sequencing, data is streamed in real-time to Illumina's cloud-based data storage and analysis platform called BaseSpace. The data is parsed by BaseSpace using the Illumina Index added in the second round of PCR. The Perl script was used to parse the data further, both by Internal Index and library length. The adapter library does not contain a "head" region to facilitate parsing of the data, so qualifications were based only on the quality of the internal index found in Adapter 2 and length of the variable region. A 10 base sequence was added to the raw data files to function as the "head" in order for the Perl script to function properly, but had no impact on the determination of good, candidate or bad reads. Multiplexed TruSeq Single Read and

multiplexed TruSeq Paired End Read formats were used in data analysis in this chapter. **Figure 4.3** illustrates the functionality of the Read 1 Primer, Read 2 Primer, and Index Read Primers. The TruSeq Single Read format reads only in the 5' → 3' (termed Read 1) direction while the TruSeq Paired End Read format reads in both the 5' → 3' direction (Read 1) and the 3' → 5' direction (termed Read 2). The data output for Read 1 contains the library, Internal Index, and Adapter 2 of the parent strand, respectively. The Sequencing Primer for Read 1 is complementary to the complement strand. During sequencing, fluorescence data for incorporated bases is captured, thus the data output is of the parent strand. The data output from Read 2 contains the Internal Index, library and Adapter 1 of the complement strand, respectively. When using Illumina Indices, the Index Read data output contains only the Illumina Index, read 5' → 3' of the parent strand. The Read 1 data is associated with the Index Read data via cluster coordinates on the flow cell. Read 2 is associated with Read 1 data in the same manner.

Simulated Partitioning

To determine optimal PCR conditions for a partitioning experiment, simulated conditions using very dilute concentrations of the m=15 DNA adapter library were PCR amplified with CAP1/CAP2 primers. PCR reactions containing between 1 pmol and 1 fmol of library were amplified in a gradient of cycle numbers in order to identify any nonspecific products. Following a partitioning step, the concentration of surviving library molecules is too small to quantify via UV absorbance and varies per experiment. Therefore, is it impossible to know how many PCR cycles are required for sufficient amplification and if/when nonspecific products would amplify. Primer concentration, annealing temperature, and salt concentration were also varied. It was

found that after a certain number of amplification cycles, depending on initial concentration, the product shifts to a larger size. Dilute concentrations of the CAP1/CAP2 PCR products were amplified with P1/P2 under similar conditions and two different size products were also apparent. These bands were characterized by high throughput sequencing on the Illumina MiSeq.

Two-Step amplification and characterization of dual PCR products

To simulate partitioning, a serial dilution series from 1 pmol to 1 fmol of DNA library was amplified using CAP1 and CAP2 primers. DNA libraries were PCR amplified using Paq5000 DNA polymerase Master Mix (Agilent Technologies): 10 µl 2X Master mix, 8 µl library, 1 µl 4 µM CAP1, 1 µl 4 µM CAP2 with the following conditions.

Denature	95 °C	2 min.
Variable # of Cycles	95 °C	30 sec.
	60 °C	30 sec.
	72 °C	30 sec.
Final Extension	72 °C	10 min.

Following the first experiment to evaluate the amplification of the m=15 DNA adapter library, it was found that a second PCR product of greater size appears at increased cycle number, as illustrated in **Figure 4.4**. The appearance is gradual, but over 4-6 cycles the product size completely shifts from a visual 80 base-pairs to 90 base-pairs (expected is 82 base-pairs). These bands have been termed “bottom” and “top”. In order to achieve maximum resolution of the two bands, 4% TAE agarose gels made from 3% NuSieve GTG agarose and 1% standard high melt agarose were implemented. NuSieve GTG Agarose (Lonza) is a high purity and low melting agarose that finely resolves to 10 base-pairs. The product bands were excised from the gel with a

4.0 mm by 1.0 mm disposable gel excision tip (GeneCatcher) to ensure minimal risk of cross-contamination between gel lanes. The product was extracted with the MiniElute Gel Extraction Kit (Qiagen), eluted in 10 µl dH_2O, and quantified via UV absorbance on the Synergy 2 Microplate spectrophotometer (BioTek).

The second round of amplification used Primer 1, indexed Primer 2 and 0.1 pmol of the gel purified CAP PCR product. The 0.1 pmol concentration was determined from a dilution series of starting concentrations that indicated 0.1 pmol as suitable for the range of desired amplification cycles. With template concentrations above 1.0 pmol, amplification was inhibited by the large concentration of template. The product was PCR amplified using Paq5000 DNA polymerase Master Mix (Agilent Technologies): 10 µl 2X Master mix, 8 µl gel purified product, 1 µl 4 µM P1, 1 µl 4 µM P2.X.

Denature	95 °C	2 min.
Varied # of Cycles	95 °C	30 sec.
	65 °C	30 sec.
	72 °C	30 sec.
Final Extension	72 °C	10 min.

Similarly to the CAP PCR product, a second product appears with increased cycle number. The product shifts from a visual 135 base-pairs to approximately 150 base-pairs (137 bp expected) over the course of 4-6 cycles. **Figure 4.5** illustrates the amplification of a gel purified bottom band CAP PCR product with P1/P2 Primers. Additional experiments determined that both the bottom and top bands from the CAP PCR products produce both the bottom and top bands in P1/P2 PCR. These four bands have been termed “bottom-bottom” (for a bottom P1/P2 product produced from a bottom CAP product), “bottom-top,” “top-bottom,” and “top-top,” respectfully.

It was found that DNA and 2'-OMe RNA/DNA chimera libraries of any length variable region produce both bands in CAP and P1/P2 PCR.

From initial experiments to test the PCR amplification of the DNA adapter libraries, it was found that the number of cycles needed to amplify enough product to visualize varies. At 0.1 pmol of starting product, the bottom band is visible at six cycles and the top band is visible at twelve cycles. The shift is gradual, with both bands visible at cycles eight to ten during some 0.1 pmol amplification experiments. Regardless of when the bottom band becomes visible (not until 24 cycles for 1 fmol, etc.), the shift always occurs gradually over a period of four to six cycles. This is true for both the DNA and 2'-OMe RNA/DNA chimera libraries.

It was initially believed that the top band was a nonspecific product and experimental conditions were altered in an attempt to prevent the formation of two different products. The primer concentration for the aforementioned experiments is 0.2 µM. It was found that reducing the primer concentration delays the formation of the top band, but does not eliminate it. Also, altering the Mg^{2+} or KCl concentration in an attempt to influence amplification specificity did not eliminate the top band, but increasing the concentrations caused the top band to be produced faster. The length of the CAP1 and CAP2 primers was also varied in an attempt to produce a single band. The CAP1 primer adds an additional fifteen bases to the 5'-terminus of the library and CAP2 primer adds an additional six bases to the 3'-terminus. Shorter versions of each primer that did not add any additional bases (completely complementary to the library adapters) were tested in all combinations with the longer versions. Sequences can be found in **Appendix 4.** It

was found that no combination of short/long CAP1 and short/long CAP2 produced a single PCR product. In an attempt to determine the source of the second band, the fixed TBA scrambled sequence, dGGTGGTTGTTGTGGT (termed TBAsc), with the adapter constructs was amplified with CAP primers. Over a gradient of cycle numbers, the TBAsc product was a single band of similar size to the bottom band. From this, it was determined that the dual PCR products were a result of the variable library region. Also supportive of this hypothesis is the fact that the size difference between the bottom and top bands increases with an increase in library size; for an m=15 library the difference is approximately ten base-pairs and for an m=40 library, the difference is larger at approximately 25 base-pairs. The annealing temperature was also varied in an attempt to eliminate the second PCR product. A temperature gradient from 50°C to 72°C was used with no change in product formation.

In order to characterize the difference between the bottom and top bands, melt curve analysis was used. A 0.1 fmol starting quantity of m=15 DNA adapter library was amplified using the SYBR Select Master Mix (Life Technologies) and CAP primers with a gradient of cycle numbers. The 0.1 fmol starting quantity was selected from a dilution series used to determine the optimal quantity for visualization with SYBR Select Master Mix. Samples were pulled from the thermocycler at even cycle numbers from 2-30 cycles during the following protocol:

UDG Activation	50°C	2min.
Polymerase, UP Activation	95°C	2 min.
Variable Cycle # of	95°C	15 sec.
	60°C	1 min.

The samples were replaced and melt curve data was collected over a gradient of 95°C to 50°C at a rate of 0.5 °C per 30 seconds. The dissociation curves of the product from cycle numbers 18,

20, 22, and 24 are shown in **Figure 4.6.** The samples were then analyzed on a 4% TAE agarose gel to visualize the association between band size and melting temperature as shown in **Figure 4.7**. Based on the gel analysis, it can be concluded that the product in the bottom band has a melting temperature of 82°C and the product in the top band has a melting temperature of 76°C. The gradual shift in product size seen on the gel correlates with the gradual change in melting temperature seen in **Figure 4.6**. The average melting temperature of the CAP PCR product was calculated as 78°C under the SYBR Select reaction conditions, which does not correlate with either band. The dissociation curves of specific products versus primer dimers are similar to those of the top and bottom bands observed in CAP and P1/P2 PCR. However, it is unlikely that primer-dimers are responsible for the shift in product size. Primer-dimers are typically shorter than the desired product and are amplified in addition to the desired product. If the bottom bands were primer-dimers, they would not disappear with increased cycle number and would not have a higher melting temperature. Although primer-dimers compete for PCR reagents, if the bottom band was the desired product, it would not disappear with the amplification of primer-dimers with increased cycle number. Although primer-dimers are not responsible for the shift in band size, identifying the cause of differing melting temperatures was important to ensure that the correct product was sequenced. If the majority of the quantifiable product was PCR artifacts, a lesser percentage of the desired product would be sequenced, decreasing the sequence space. A requirement for the success of AIA is deep sequencing, so determining the nature of the dual products was critical.

In order to determine any structural differences in the bottom and top bands, an m=15 DNA library was prepared in a simulated partitioning for sequencing. The four band combinations

were sequenced: bottom-bottom, bottom-top, top-bottom, and top-top. A quantity of 0.1 pmol of adapter library was PCR amplified using CAP1 and CAP2 primers following the above protocol. Individual PCR reactions were amplified is a gradient of cycles numbers (10, 12, 14, 16, 18, 20, 22, 24 cycles). The bottom band was gel excised for 10 cycles and the top band excised for 24 cycles. The bands were quantified via UV absorbance and 0.1 pmol was amplified using P1 and indexed P2. Each CAP product was amplified in a gradient of cycle numbers in order to generate both the bottom and top bands (8, 10, 12, 14, 16 cycles). In order to distinguish the four types of bands, two sets of PCR reactions per CAP product band were made, each with a different Illumina Index. The four combinations were indexed as followed:

Bottom-Bottom	Illumina Index 4
Bottom-Top	Illumina Index 3
Top-Bottom	Illumina Index 2
Top-Top	Illumina Index 1

The four products were sequenced using a TruSeq Single Read format and run on an Illumina MiSeq in a 1:1:1:1 ratio as determined by qPCR. A 10 µl volume of 2 nM DNA was prepared. The sequence data was parsed based on the assigned Illumina Indices and run statistics are presented in **Table 4.3.** Good reads were identified by qualifying the length of the variable region and first 5 bases of Adapter 2, which includes a single dA and the first four bases of the internal index. The average percentage of good reads per total reads was approximately 85%, indicating relatively high quality data. There is no significant variation in the percentage of good reads for each band type, with a ratio of 1.054:1.020:1.026:1. This indicates that the sequencing primer binding region, the last 32 bases of Adapter 1 (see **Figure 4.3**), is intact for each of the four band types. From the limitations of a Single Read reagent kit, the samples were sequenced 50 bases from the end of Adapter 1 and include the entire library region (m=15), the internal

index and a portion of Adapter 2 (26 bases). The raw data for each of the four band types contains the correct sequence arrangement, in the expected order. **Figure 4.8** shows a snapshot of the raw data from the bottom-bottom band. Minor sequence errors are visible in the form of deletions that shift the correct sequence. These types of errors are present in all four data files and typically account for the sequences other than "good reads". Although the data quality was similar for all four bands, the quantity of data is dominated by the bottom-bottom sequences (higher total reads). This indicates either inaccurate qPCR calculations and/or a higher percentage of successful bridge amplification for the bottom-bottom sample. If the bottom-top, top-bottom, and top-top samples contain some sequence anomalies that were quantified with qPCR, but failed bridge amplification on the flow cell, it would explain the higher number of total reads for the bottom-bottom samples.

Data from the Single Read run did not provide any explanation for the apparent difference in size of the two bands. The 50 cycle read only provided the sequence arrangement of the library region and a portion of Adapter 2. It was possible that a structural difference contributing to the size and melting temperature difference was present in the Adapter 1 region or farther into Adapter 2. For this reason, a Paired End run with reagent kit capable of sequencing the entire 137 base product was used to sequence the four bands. The sample preparation was repeated due to low sample volume of the first set of experiments. The samples were prepared in a similar manner, with the following associated Illumina Indices:

Bottom-Bottom	Illumina Index 1
Bottom-Top	Illumina Index 2
Top-Bottom	Illumina Index 3
Top-Top	Illumina Index 4

The reversal of associated Illumina Index is of no consequence. The products were sequenced using the Paired End Read format (Illumina MiSeq with a TruSeq kit) in a 1:1:1:1 ratio as determined by qPCR. A 5 µl volume of 4 nM DNA was prepared.

The sequence dataset was parsed based on the assigned Illumina Indices and run statistics are presented in **Table 4.4.** Good reads were identified by qualifying the length of the variable region and first 5 bases of Adapter 2, which includes a single dA and the first four bases of the internal index. For Read 1, there is no significant variation in the percentage of good reads. The Read 1 raw data for all four bands contains the library region (m=15), internal index, entire Adapter 2 (55 bases), and 5 bases into the oligo-T lawn for a total of 84 bases. There is no apparent difference in sequence arrangement for each band, indicating that the quality of reported data is equivalent. For Read 2, there is no significant variation in the percentage of good reads. The Read 2 raw data should contain the internal index, library region (m=15), entire Adapter 1 (58 bases), and 5 bases into the oligo-T lawn for a total of 87 bases. The Read 2 data for all four bands contains the correct sequence arrangement, in the expected order, again indicating that the quality of reported data is equivalent for all bands. **Figure 4.9** shows a sample of the raw data from the Read 1 and Read 2 bottom-bottom band. Minor sequence errors are visible in the form of deletions that shift the correct sequence. These types of errors are present in all four data files and account for the sequences other than "good reads." Similarly to the Single Read run data, the total reads for R1 and R2 are dominated by the bottom-bottom band sequences. According to qPCR quantification, the four samples were sequenced in equal

concentrations. For such a disparity between total read numbers, a 2.56:1.54:1.00:1.38 ratio, a data quality issue that cannot be visualized by high throughput sequencing must be occurring.

Based on the Paired End sequencing data from the four bands, it was determined that the shift in product size was a result of the formation of heteroduplex DNA. For the first few rounds of PCR, the correct size product is amplified. After the first several rounds of PCR, primer concentration begins to deplete. Once the primer concentration has depleted significantly, amplification is saturated and the full length single-stranded fragments are no longer amplified to the double stranded product. Instead, the full length single-stranded fragments simply anneal, denature and reanneal to each other during each amplification cycle. The highly randomized nature of the library region prevents the single-stranded fragments from finding their complements in solution while the highly conserved adapter regions allow the fragments to improperly anneal. After a certain threshold cycle number, the product will be composed solely of this partially double-stranded heteroduplex DNA. The heteroduplex DNA migrates slower than perfectly annealed DNA during gel electrophoresis.[165] This was observed during gel purification of the CAP and P1/P2 PCR products. This explanation was validated by the melt curve analysis; the top band was determined to have a lower melting temperature than the bottom band, which is consistent with the partially single-stranded nature of the heteroduplex DNA of the top band.

The formation of heteroduplexes is common in the amplification of highly complex mixtures of genomic DNA.[165] However, this is not a concern when quantifying libraries via qPCR because the method involves complete denaturation of heteroduplex DNA and amplification of

only perfectly complementary double-stranded products during the log phase of amplification.[158] This also indicates that heteroduplex DNA should not affect sequencing because the product is denatured to single-stranded DNA prior to clustering and still contains the proper library sequence. Additionally, it has been shown that amplifying beyond the log phase and into the plateau phase does not introduce any biases.[166] Although this behavior should not impact the quality of qPCR or sequencing data, the quantity of total reads for the bottom-bottom band was highest for both the Single Read and Paired End Read runs. From the data, it was concluded that the four bands could be sequenced and the results would include the correct sequence arrangement. Any aptamer candidates found from an AIA partitioning would likely be at high enough counts to be identified using any of the four products bands. However, considering the superior performance of the bottom-bottom band, controlling the amplification in order to collect the higher quality bottom band in both CAP and P1/P2 PCR was ideal.

Preliminary experiments monitoring the amplification of a DNA or 2'-OMe RNA/DNA chimera library in real-time indicated that it would be possible to control the formation of the top band. The option of a single step amplification scheme using P1 and P2 was considered for this application in order to further simplify the AIA protocol. The P1/P2 primers were used to amplify 0.1 pmol, 0.1 fmol, and 0.1 amol of m=15 DNA adapter library using Paq5000 DNA polymerase Master Mix (Agilent): 10µl 2X Master Mix, 8µl library, 1µl 4 µM P1, 1µl 4 µM P2.1.

Denature	95 °C	2 min.
18 Cycles of	95 °C	30 sec.
	65 °C	30 sec.
	72 °C	30 sec.
Final Extension	72 °C	10 min.

Once it was determined that the adapter library was successfully amplified with P1/P2 in single step amplification, the SYBR Select Master Mix (Applied Biosystems, Life Technologies) and a SYBR Green One-Step qRT-PCR Kit (SuperScript III Platinum, Invitrogen, Life Technologies) were used to monitor amplification in real time. A benefit of the two step amplification scheme was to control the amount of template in P1/P2 PCR in order to achieve a sufficient quantity of product for qPCR and sequencing without over-amplifying, causing the formation of heteroduplex DNA. Following partitioning, the quantity of surviving library is unknown and too small to quantify by UV absorbance. By monitoring the one-step amplification of the library with P1/P2 in real time, the number of PCR cycles can be controlled on an individual reaction basis. Amplification can be ended early to avoid the formation of heteroduplex DNA or PCR champions, or extended if needed. The procedure was first tested with CAP1 and CAP2 primers for simplicity and to reserve the more costly indexed P2 primers. A dilution series of the m=15 DNA adapter library from 100 fmol to 100 zmol was amplified with the SYBR Select Master Mix using the following conditions: 10 µl 2X Master mix, 8 µl library, 1 µl 4 µM CAP1, 1 µl 4 µM CAP2.

UDG Activation	50 °C	2min.
Polymerase, UP Activation	95°C	2 min.
Variable Cycle # of	95°C	15 sec.
	60°C	1 min.

A dilution series of the adapter library using m=15 chimeric 2'-OMe RNA/DNA covering 100 fmol to 100 zmol was amplified with the SYBR Green One-Step qRT-PCR Kit (SuperScript III Platinum) using the following conditions: 10 µl 2X Master mix, 7.5 µl library, 1 µl 4 µM CAP1, 1 µl 4 µM CAP2, 0.5 µl enzyme.

cDNA Synthesis	50 °C	3 min.
Polymerase, UP Activation	95°C	5 min.
Variable Cycle # of	95°C	15 sec.
	60 °C	1 min.

All samples were run in duplicate. Amplification was ended for the first of each sample when the amplification curve had visually begun to increase and was still exponential. Amplification of the second sample was ended once the curve had plateaued. Both samples were characterized with melt curve analysis from 96°C to 50°C with a gradient of 0.5°C per 30 seconds. Once saturation has been reached (curve plateaus), heteroduplex DNA has formed. By monitoring amplification in real-time, it can be ended at a predetermined intensity, void of heteroduplex DNA. Preliminary experiments using this method indicated that it would be possible to monitor sample amplification on an individual basis to obtain the desired product, in this case, the bottom band only. Samples removed when the amplification curve was still exponential produced the desired product; samples removed once the amplification curve had plateaued produced heteroduplex DNA. This method would be beneficial for AIA partitioning with a single step amplification scheme because the quantity of surviving library from a partitioning step is too low to accurately quantify and differs for each experiment. The ability to monitor amplification and end the protocol prior to heteroduplex product formation would be useful. However, based on results from AIA for thrombin presented in next section, the adapter libraries and the one-step amplification were not pursued further.

AIA for Thrombin

To gauge the functionality of adapter libraries in AIA, they were screened against thrombin using agarose-Concanavalin-A beads. Both m=15 DNA and m=15 2'-OMe RNA/DNA chimera libraries were screened in various ratios. Many previous attempts to PCR amplify the surviving library molecules from bead-based partitioning required trial and error to amplify enough product to visualize. It was found that even experimentally identical partitioning produced highly variable quantities of surviving library. However, when an adequate amount of product could be detected with qPCR, multiple thrombin screens were successfully prepared and sequenced on the Illumina MiSeq.

Experimental methods

Partitioning of the m=15 DNA adapter library was performed using agarose-Concanavalin-A beads (GE Healthcare Life Sciences) at room temperature. The Con-A beads contained in spin columns were pre-equilibrated in partitioning buffer (20 mM Tris-HCl, 140 mM NaCl, 5 mM KCl, 1 mM $CaCl_2$, 1 mM $MgCl_2$, pH 7.4) by three wash cycles. The following represents a 1:1 library: protein screen. A 100 pmol aliquot of glycosylated human α-thrombin (Haematologic Technologies) in 750 µl partitioning buffer was immobilized on 100 µl of pre-equilibrated Con-A beads during a 30 minute incubation with end-over-end rotation at 25 rpm. Unbound thrombin was removed by centrifugation. A 100 pmol aliquot of m=15 DNA adapter library in 750 µl partitioning buffer was applied to the immobilized thrombin. Following one hour incubation with end-over-end rotation at 25 rpm, the unbound DNA was removed by centrifugation. Three wash cycles (5 minutes, rotation at 25 rpm) with 750 µl partitioning buffer ensured thorough removal of unbound DNA. Thrombin-DNA complexes were eluted from the Con-A beads with 200 µl

elution buffer (α-methyl mannoside, Glycoprotein Isolation Kit, Pierce Biotechnology) during a 5 minute incubation, rotation at 25 rpm. The m=15 DNA and 2'-OMe RNA/DNA chimera libraries were screened at ratios of 60:1, 10:1, 1:1 and 1:10 protein: library, with the library quantity maintained at 100 pmol. Based on qPCR data, the following m=15 DNA library: thrombin ratios were sequenced on the Illumina MiSeq: 60:1 (6 nmol thrombin: 100 pmol library) and 1:1 (100 pmol thrombin: 100 pmol library). The following m=15 2'-OMe RNA/DNA chimera library: thrombin ratios were sequenced on the Illumina MiSeq; 60:1 (6 nmol thrombin: 100 pmol library) and 1:1 (100 pmol protein: 100 pmol library).

Phenol extraction was used to recover the partitioned DNA and 2'-OMe RNA/DNA chimera libraries. An equal volume of Tris-buffered, pH 8.0, 0.1mM EDTA, 50% phenol, 48% chloroform, 2% isoamyl alcohol (Sigma Aldrich) was added and vortexed for 30 seconds. After centrifugation at 13,000 rpm for 5 minutes the aqueous layer (top) was removed and extracted twice with equal volumes 100% chloroform. The library was then purified by ethanol precipitation. One tenth the volume (20 µl) of 3M sodium acetate and three times the volume (600 µl) of cold ethanol and were added to the library and mixed briefly. Following a minimum 12 hour incubation at -20 °C, centrifugation at 13,000 rpm for 30 minutes resulted in a library pellet. After decanting the liquid, the pellet was washed with 1 ml 70% ethanol. Centrifugation at 13,000 rpm for 10 minutes resulted in a library pellet. The pellet was air dried at room temperature in a PCR hood, and then resuspended in dH_2O.

The recovered library was amplified using CAP1 and CAP2 primers. DNA libraries were PCR amplified using Paq5000 DNA polymerase Master Mix (Agilent Technologies): 10 µl 2X Master Mix, 8 µl resuspended library, 1 µl 4µM CAP1, 1 µl 4µM CAP2.

Denature	95 °C	2 min.
18 Cycles of	95 °C	30 sec.
	60 °C	30 sec.
	72 °C	30 sec.
Final Extension	72 °C	10 min.

It was determined that 18 cycles was sufficient for the majority of partitioning samples to visualize a product and quantify via UV absorbance. The 2'-OMe RNA/DNA chimera libraries were RT-PCR amplified using the One-Step RT-PCR System with Platinum Taq DNA polymerase (SuperScript, Invitrogen): 10 µl 2X Master Mix, 7.5 µl resuspended library, 1 µl 4 µM CAP2, 0.5 µl RT/Platinum Taq HiFi Mix.

cDNA Synthesis	50 °C	15 min.
Pre-denaturation	94 °C	2 min.

Addition of 1 µl 4 µM CAP1.

Denature	94 °C	30 sec.
18 Cycles of	94 °C	30 sec.
	60 °C	30 sec.
	68 °C	30 sec.
Final Extension	72 °C	30 sec.

The 82 base-pair PCR product (whether present as bottom, top, or both, depending on the experiment) was purified by a 4% TAE agarose gel. Excision of the product band from the gel with a 4.0 mm by 1.0 mm disposable gel excision tip (GeneCatcher) ensured minimal risk of cross-contamination between gel lanes. The product was extracted from the gel with the

MiniElute Gel Extraction Kit (Qiagen), eluted in 10 µl dH_2O, and quantified via UV absorbance on the Synergy 2 Microplate spectrophotometer (BioTek).

The second round of amplification used P1 and indexed P2 and 0.1 pmol of the gel purified CAP PCR product. The product was PCR amplified using Paq5000 DNA polymerase Master Mix (Agilent Technologies): 10 µl 2X Master Mix, 8 µl gel purified product, 1 µl 4µM P1, 1 µl 4µM P2.X.

Denature	95 °C	2 min.
Variable Cycle # of	95 °C	30 sec.
	65 °C	30 sec.
	72 °C	30 sec.
Final Extension	72 °C	10 min.

The 137 base-pair product (whether present as bottom, top, or both, depending on experiment) was gel purified as described above.

During each experiment, a control reaction was used as a size marker in the event that the experimental sample was too dilute to be visualized. A 0.1 pmol aliquot of a size appropriate DNA library, 2'-OMe RNA/DNA chimera library, or fixed sequence was amplified under identical conditions. With 0.1 pmol starting library or CAP PCR product, 18 cycles produced the top band size marker, while controls amplified at six to ten cycles produced the bottom band. These controls help identify which band is present more accurately than comparison with a DNA ladder, as the starting library concentration for CAP PCR is typically too low to be accurately quantified. If a fixed sequence was used (e.g. TBAsc), the bottom band was produced regardless of cycle number.

Results and data analysis

The results from five successfully prepared partitioning experiments were sequenced on the Illumina MiSeq. The experiments qualified as successful based on qPCR results. At least 5 nM (in 2 µl – 10 µl volumes) was sufficient for a MiSeq run. The first experiment was a 60:1 ratio of 6 nmol thrombin: 100 pmol DNA adapter library that was sequenced simultaneously with the Single Read run mentioned in the previous section. The goal was to sequence the bottom and top bands from a partitioning experiment and observe any difference in distribution of aptamer candidates. Following 18 cycles of CAP PCR, the bottom band was gel purified and quantified. Both the bottom and top bands from P1/P2 PCR were selected from five PCR reactions using 0.1 pmol of the CAP product each. The five PCR reactions were performed in duplicate with P2.5 and P2.6 in order to index the bottom and top bands as shown in **Figure 4.10.** A summary of sequence data is presented in **Table 4.5**. For both bands, the highest frequency sequences are variations of TBAsc, dGGTGGTTGTTGTGGT. This fixed sequence was used as a size marker during the CAP PCR step and has a different Internal Index than the m=15 DNA library. The raw data revealed that these sequences contained the Internal Index 1, ACACAGCA, the index associated with TBAsc. The high number of bad reads associated with each sample is attributed to this contamination. The Perl script identifies good read based on the first four bases of the Internal Index, thus qualifying the majority of the contaminating sequences as bad reads. The few TBAsc sequences that qualified as good reads and were counted in the nmer file had sequence mismatches in the Internal Index, qualifying them as good reads. The desired TBA sequence with the correct Internal Index 15, dATGCCTGG, was not found in either sample, indicating that the partitioning of the m=15 DNA library against thrombin was not successful in identifying TBA from the library pool. Due to contamination, the experiment was repeated but

with an unpartitioned library in place of TBAsc as the size marker. The following experiments were prepared without regard to the top and bottom band based results from the previous section.

The next four thrombin partitioning experiments were sequenced on a multiplexed Paired End run, including 60:1 (6 nmol thrombin: 100 pmol library :) and 1:1 (100 pmol thrombin: 100 pmol library) ratios for both the m=15 DNA and m=15 2'-OMe RNA/DNA chimera libraries. Following partitioning, 18 cycles of PCR with CAP primers resulted in the bottom band for all four experiments. Following quantification, 0.1 pmol of each was amplified using a single indexed P1/P2 PCR reaction. The resulting products from 18 cycles were the top band for all four experiments. **Table 4.6** contains a summary of relevant sequencing statistics.

After parsing the data based on library region length and quality of the internal index, it was apparent that an error in the 2'-OMe samples had occurred. The expected Internal Index 8 was not found at any significant counts in the 2'-OMe samples. Instead, the expected Internal Index of the DNA library, Index 15, was found to dominate the data. It is clear from the results that either the m=15 2'-OMe library was synthesized with the incorrect Internal Index or the incorrect library was used during partitioning. The library was not synthesized by IDT, but in-house by Dr. Mark McPike on the ABI 394 synthesizer, making it possible for an error in sequence input. If it were an issue of contamination during PCR and gel purification, it would be expected that the Internal Index 8 would be found at some significant frequency. It would be possible to identify the Internal Index of the library by Sanger sequencing. However, the results from the four experiments do not necessitate any further action with the libraries. In all four experiments, the thrombin binding aptamer was not identified from the library pool. Regardless of whether the 2'-

OMe library contains the incorrect Internal Index or if the DNA library was actually used, the TBA sequence is still expected for a successful partitioning.

Conclusions

Of the five adapter library for thrombin partitioning that were sequenced, none were successful at identifying TBA from the library pool. The partitioning method was successful in identifying TBA using the hairpin ligation libraries, both DNA and 2'-OMe. As demonstrated in chapters 2 and 3, if the TBA quadruplex structure was made available for binding by the construct of the library, it is highly likely that the sequence would be recovered using Concanavalin-A beads as an immobilization method. Therefore, it is likely that the adapter constructs inhibit aptamer discovery for short libraries such as m=15. The basic structure of the adapter library is not unusual for aptamer selection and has significant success in SELEX for both DNA and RNA aptamers using various partitioning methods. Unlike the long libraries typically used in SELEX, the adapter constructs of the AIA adapter libraries used in this chapter constitutes the majority of the library size. It is hypothesized that this interferes with selection of the variable region. Even if the adapter constructs influence aptamer selection in SELEX, multiple rounds of enrichment and selection may allow SELEX to overcome this obstacle whereas the selection is simply not stringent enough for a single round in AIA. It is possible that redesigning the adapter libraries with shorter adapters, variations in adapter sequence, or with no internal indices could improve the performance of the libraries. However, the adapter libraries were not pursued further in AIA in favor of exploring the new library structures described in chapter 5.

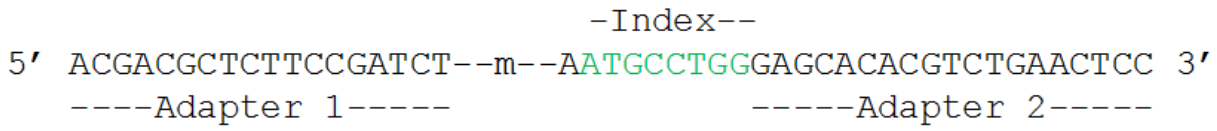


Figure 4.1 Sequence arrangement for an "adapter library."

Library region denoted by "m." Eight base internal index in green. The m=15 DNA library contains Internal Index 15, sequence dATGCCTGG. The m=15 2'-OMe RNA/DNA chimera library contains Internal Index 8, sequence dAGCTAACG. See **Table 4.1** for full list of internal Index sequences.

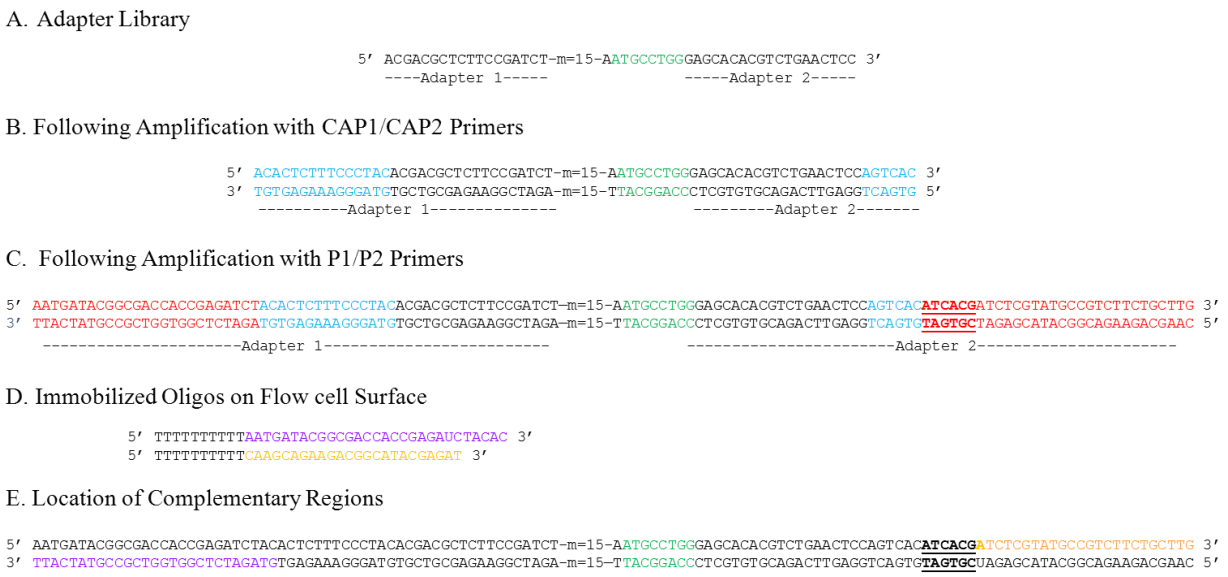


Figure 4.2 Adapter library amplification products

(A) Single-stranded m=15 DNA adapter library. Eight base internal index in green. **(B)** 82 base-pair CAP1/CAP2 PCR product. Regions in blue are added by CAP1 and CAP 2 primers. **(C)** 137 base-pair P1/P2 PCR product. Regions in red are added by P1 and P2 primers. Six base Illumina index is underlined. **(D.)** Immobilized oligos on Illumina flow cell, priming regions in purple and orange. **(E.)** In orange, region of the parent strand necessary for hybridization to the flow cell and complementary to the orange strand in D. In purple, region of the complement strand necessary for hybridization to the flow cell and complementary to the purple stand in D.

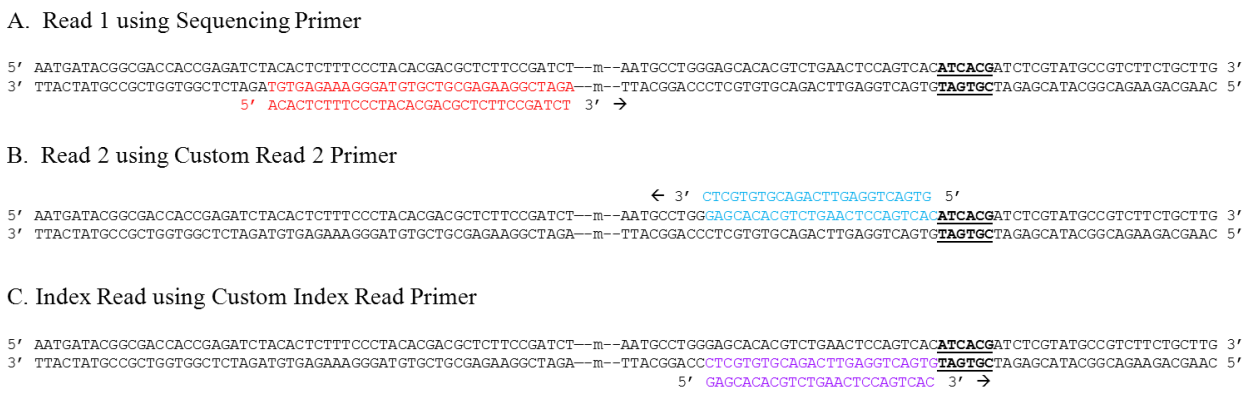


Figure 4.3 Priming locations for Read 1, Read 2 and Index Read

Coordinating colors indicate complementarity. Arrows indicate directionality of Sequencing by Synthesis (SBS). **(A)** Read 1 using Sequencing Primer. The primer sits complementary to Adapter 1 of the complement strand. SBS proceeds across the library, Internal Index and Adapter 2 regions of the complement strand. Data output is the library, Internal index and Adapter 2 region of the parent strand, respectively. **(B)** Read 2 using Custom Read 2 Primer. The primer sits complementary to the Adapter 2 of the parent strand. SBS proceeds across the Internal Index, library and Adapter 1 of the parent strand. Data output is the Internal index, library and Adapter 1 of the complement strand, respectively. Read 2 data represents the reverse complement of the parent strand. **(C)** Index Read using Custom Index Read Primer. The primer sits complementary to Adapter 2 of the complement strand. SBS proceeds across the Illumina Index of the complement strand. Data output is the Illumina Index of the parent strand.

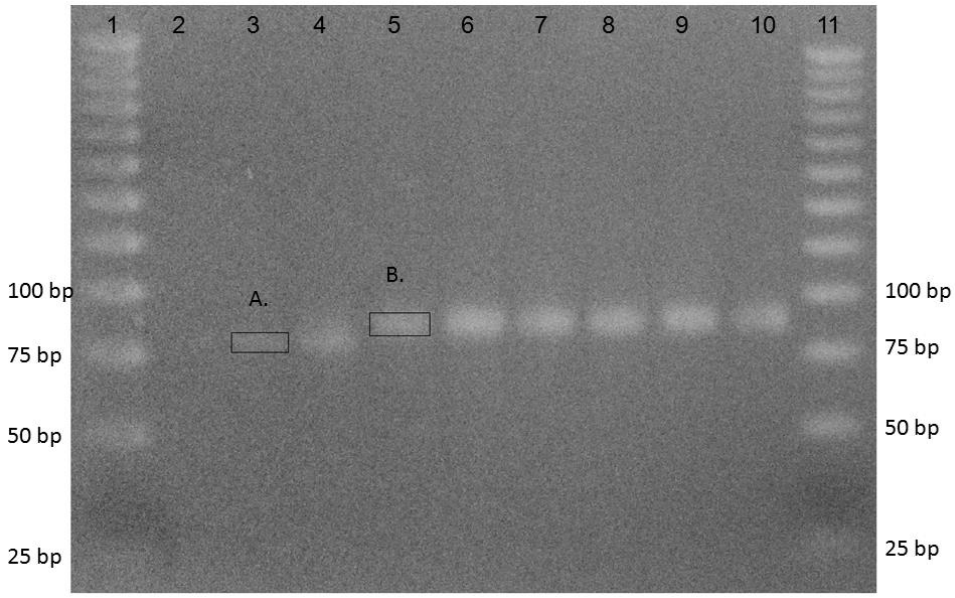


Figure 4.4 CAP1/CAP2, dual PCR products

4% TAE agarose gel. Lanes 1 and 11: 25 bp DNA ladder. Lanes 2-10: 0.1 pmol starting quantity of template, PCR cycles 6, 8, 10, 12, 14, 16, 18, 20, and 22, respectively. **(A)** CAP PCR "Bottom" band for 10 cycles, ~80 base-pairs **(B)** CAP PCR "Top" band for 14 cycles, ~90 base-pairs.

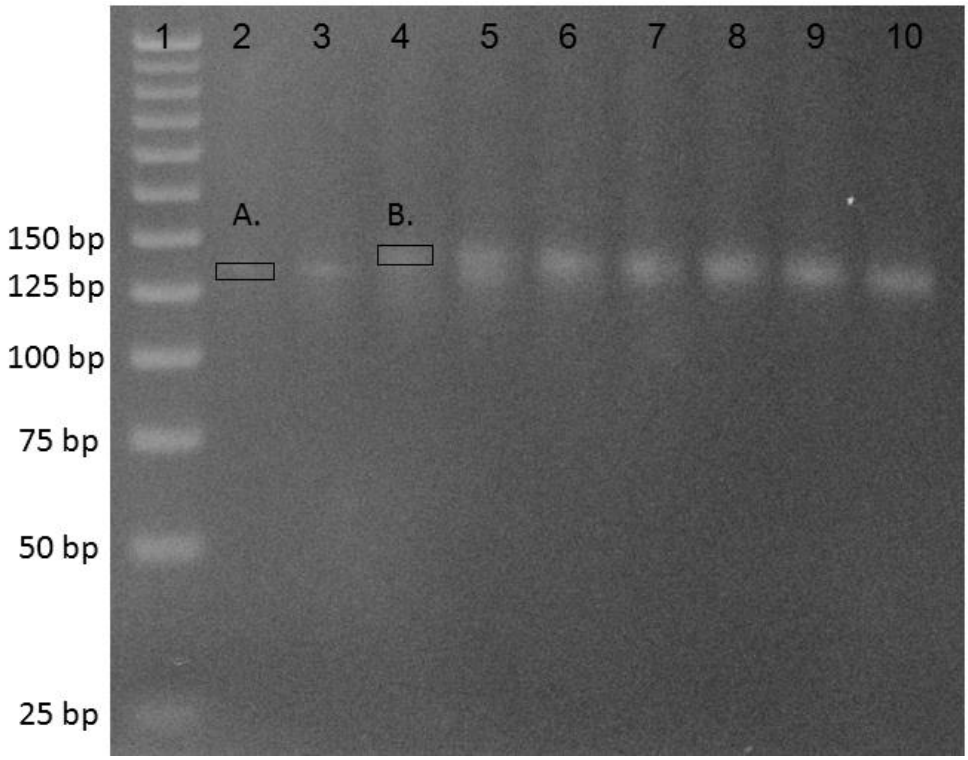


Figure 4.5. P1/P1, dual PCR products

4% TAE agarose gel. Lane 1: 25 bp DNA ladder. Lanes 2-10: PCR cycles 10, 12, 14, 16, 18, 20, 22, 24, 26, respectively, of 0.1 pmol gel purified bottom band, CAP PCR product **(A)** P1/P2 PCR "Bottom-Bottom" band for 10 cycles, ~135 base-pairs **(B)** P1/P2 PCR "Bottom-Top Band" for 14 cycles, ~150 base-pairs.

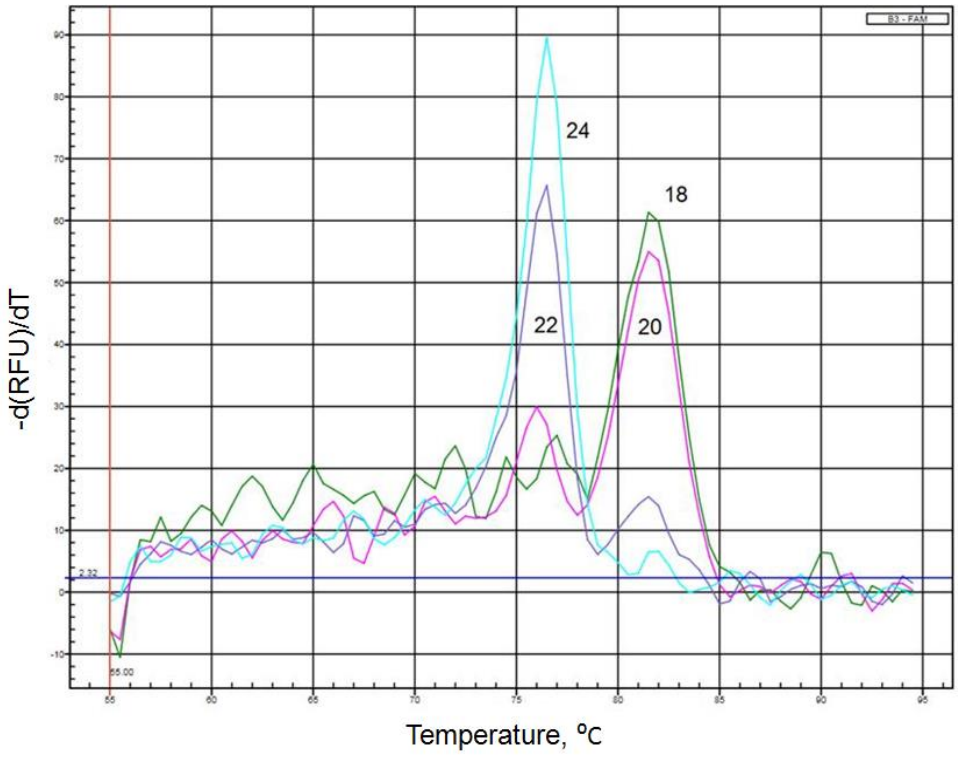


Figure 4.6 Melting temperature analysis of dual PCR products

Dissociation curves for 0.1 fmol of m=15 DNA adapter library at amplification cycles 18, 20, 22, and 24. The melting temperature for 18 cycles is approximately 82°C. The melting temperature for the majority of the 20 cycle product is 82°C, with a smaller portion with melting temperature of 76 °C. The melting temperature for the majority of the 22 cycle product is 76°C, with a smaller portion with melting temperature of 82°C. The melting temperature for 24 cycles is approximately 76°C.

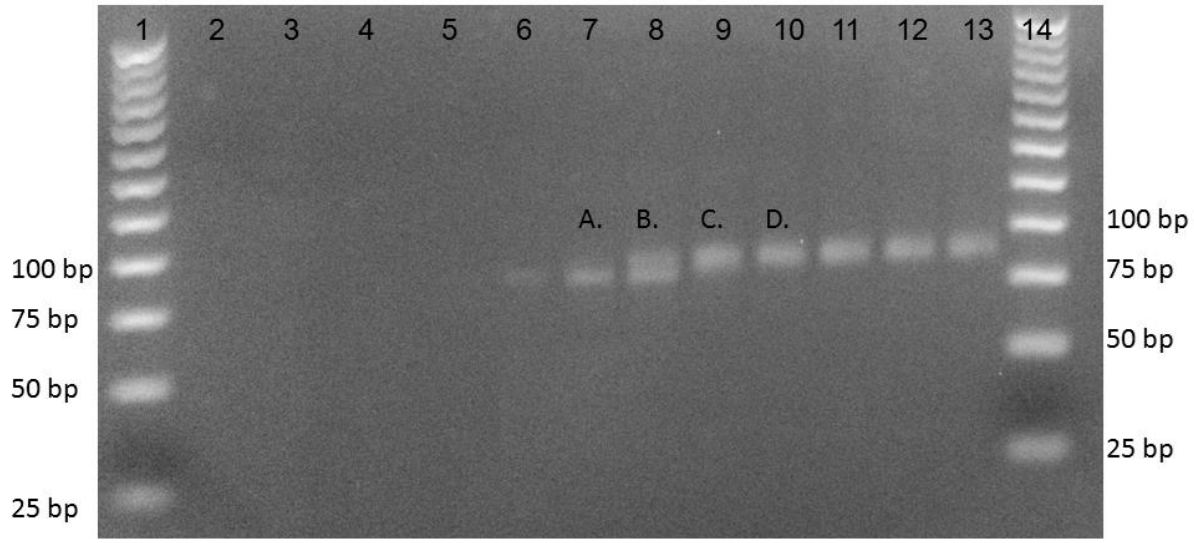


Figure 4.7 Melting temperature and size comparison of dual PCR products

4% TAE agarose gel. Lanes 1 and 14: 25 bp DNA ladder. Lanes 2-13: Even cycle numbers 8-30, respectively. **(A)** 18 cycles, bottom band. **(B)** 20 cycles, bottom and top band. **(C.)** 22 cycles, top band. **(D.)** 24 cycles, top band.

Library		Index	Adapter 2
NGAGCACTGTACAGG	A	ATGCCTGG	GAGCACACGTCTGAACTCCAGTCACT
NAGGGCGTGTTTGCC	A	ATGCCTGG	GAGCACACGTCTGAACTCCAGTCACT
NTGTAGCTTTGCCGT	A	AT CTGG	GAGCACACGTCTGAACTCCAGTCACTG
ATGGCCGCTTGGCGC	A	ATGCCTGG	GAGCACACGTCTGAACTCCAGTCACT
ATGGTTTAACCGTAC	A	ATGCCTGG	GAGCACACGTCTGAACTCCAGTCACT
CGGTGTATGCTAGAT	A	ATGCCTGG	GAGCACACGTCTGAACTCCAGTCACT
ATCGGGAGCTAATAC	A	ATGCCTGG	GAGCACACGTCTGAACTCCAGTCACT
TAGGGTGCTGAAGTT	A	ATGCCTGG	GAGCACACGTCTGAACTCCAGTCACT
CTTGCACGTTGAACC	A	ATGCCTGG	GAGCACACGTCTGAACTCCAGTCACT
GAGGGACACCTGCGA	A	ATGCCTGG	GAGCACACGTCTGAACTCCAGTCACT
AGTTGCCTCTTACTT	A	ATGCCTGG	GAGCACACGTCTGAACTCCAGTCACT
GTTGGGTCGATAGTC	A	ATGCCTGG	GAGCACACGTCTGAACTCCAGTCACT
CAAGGCATTATGCTT	A	ATGCC G	GAGCACACGTCTGAACTCCAGTCACTG
CGGCCGAGGCTTGTC	A	ATGCCTGG	GAGCACACGTCTGAACTCCAGTCACT
GGCGTGCGTCTACTA	A	ATGCCTGG	AGCACACGTCTGAACTCCAGTCACTG
CCCCGTCTCAGCTGA	A	ATGCCTGG	GAGCACACGTCTGAACTCCAGTCACT
TAATAAGATTATGTC	A	ATGCCTGG	GAGCACACGTCTGAACTCCAGTCACT
TAGGATCTACGCATT	A	ATGCCTGG	GAGCACACGTCTGAACTCCAGTCACT
CGGTGCTTGTTAACG	A	ATGCCTGG	GAGCACACGTCTGAACTCCAGTCACT
GTGGCTCTACAGTCC	A	ATGCCTGG	GAGCACACGTCTGAACTCCAGTCACT
GCACGTCACACTATT	A	ATGTCTGG	GAGCACACGTCTGAACTCCAGTCACT
GATCTGCTGCTCTAG	A	ATGCCTGG	GAGCACACGTCTGAACTCCAGTCACT
CTCCTGGCCTCAGGT	A	ATGCCTGG	GAGCACACGTCTGAACTCCAGTCACT
TCTTCCGTCGGGATC	A	ATGCCTGG	GAGCACACGTCTGAACTCCAGTCACT
ACACTAACCCCG		TTGCCTGG	GAGCACACGTCTGAACTCCAGTCACTGACC
CTATTGTGTTCTTGG	A	ATGCCTAG	GAGCACACGTCTGAACTCCAGTCACT
TGTCGGTCGCGACAC	A	ATGCCTGG	GAGCACACGTCTGAACTCCAGTCACT
ACAGTTACTGACGCA	A	ATGCATGG	GAGCACACGTCTGAACTCCAGTCACT
GACCGCCCCCGCTCT	A	ATGCCTGG	GAGCACACGTCTGAACTCCAGTCACT

Figure 4.8 Read 1 raw data: Single Read band characterization

Snapshot of Read 1 raw data from the bottom-bottom band. Read 5' → 3' of the parent strand: m=15 library, eight base Internal Index, 26 bases of Adapter 2.

A. Read 1 Data, 5' → 3'

Library		Index	Adapter 2
GTCTCCAGTGTCAGA	A	ATGCCTGG	GAGCACACGTCTGAACTCCAGTCACATCACGATCTCGTATGCCGTCTTCTGCTTGAAAAA
GCCGCACAATCCATC	A	ATGCCTGG	GAGCACACGTCTGAACTCCAGTCACATCACGATCTCGTATGCCGTCTTCTGCTTGAAAAA
ATTGAGCCGGCCAAT	A	ATGCCTGG	GAGCACACGTCTGAACTCCAGTCACATCACGATCTCGTATGCCGTCTTCTGCTTGAAAAA
GGAACCCAAGTTGGT	A	ATGCCTGG	GAGCACACGTCTGAACTCCAGTCACATCACGATCTCGTATGCCGTCTTCTGCTTGAAAAA
TAACGTCAAACTACT	A	ATGCCTGG	GAGCACACGTCTGAACTCCAGTCACATCACGATCTCGTATGCCGTCTTCTGCTTGAAAAA
CATTCTGGCATGACT	A	ATGCCTGG	GAGCACACGTCTGAACTCCAGTCACATCACGATCTCGTATGCCGTCTTCTGCTTGAAAAA
ATCAATGGCGAAGTC	A	TGCCTGGG	AGCACACGTCTGAACTCCAGTCACATCACGATCTCGTATGCCGTCTTCTGCTTGAAAAAA
GATGTGTCGAATCCT	A	ATGCCTGG	GAGCACACGTCTGAACTCCAGTCACATCACGATCTCGTATGCCGTCTTCTGCTTGAAAAA
TCACTGTTTATACTG	A	ATGCCTGG	GAGCACACGTCTGAACTCCAGTCACATCACGATCTCGTATGCCGTCTTCTGCTTGAAAAA
GGCCGGCGGGGTACC	A	ATGCCTGG	GAGCACACGTCTGAACTCCAGTCACATCACGATCTCGTATGCCGTCTTCTGCTTGAAAAA
TTGTGCGGGACAATT	A	ATGCCTGG	GAGCACACGTCTGAACTCCAGTCACATCACGATCTCGTATGCCGTCTTCTGCTTGAAAAA
AGGGCCTTTTGATGC	A	ATGCCTGG	GAGCACACGTCTGAACTCCAGTCACATCACGATCTCGTATGCCGTCTTCTGCTTGAAAAA
GGGGCGGTCCCGACC	A	ATGCCTGG	GAGCACACGTCTGAACTCCAGTCACATCACGATCTCGTATGCCGTCTTCTGCTTGAAAAA
TTTGTTCGCCAT	A	ATTCCTGG	GAGCACACGTCTGAACTCCAGTCACATCACGATCTCGTATGCCGTCTTCTGCTTGAAAAAAAA
CATGACAGGCTCTAA	A	ATGCCTGG	GAGCACACGTCTGAACTCCAGTCACATCACGATCTCGTATGCCGTCTTCTGCTTGAAAAA
GTTCATCTCATCGAT	A	ATGCCTGG	GAGCACACGTCTGAACTCCAGTCACATCACGATCTCGTATGCCGTCTTCTGCTTGAAAAA
TGGGTCTCTCGGTGC	A	ATGCCTGG	GAGCACACGTCTGAACTCCAGTCACATCACGATCTCGTATGCCGTCTTCTGCTTGAAAAA
AGACGCCTGCGGCT	A	ATGCCTGG	GAGCACACGTCTGAACTCCAGTCACATCACGATCTCGTATGCCGTCTTCTGCTTGAAAAAA
TATCAATAAGTGGAG	A	ATGCCTGG	GAGCACACGTCTGAACTCCAGTCACATCACGATCTCGTATGCCGTCTTCTGCTTGAAAAA
TCGCCTCCCGTGTTG	A	ATGCCTGG	GAGCACACGTCTGAACTCCAGTCACATCACGATCTCGTATGCCGTCTTCTGCTTGAAAAA

B. Read 2 Data, Reverse Complement of Parent Strand, read 3' → 5'

Index		Library	Adapter 1
CCAGGCAT	T	GAGGAGGTATTCAGG	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATTAAAAA
CCAGGCAT	T	CATCACGTTGTCTGT	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATTAAAAA
CCAGGCAT	T	CAGGCGACGTCCACC	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATTAAAAA
CCAGGCAT	T	ATAATGCCGGTCCAA	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATTAAAAA
CCAGGCAT	T	CGCAATATGGGACG	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATTAAAAAA
CCAGGCAT	T	GGACCGACTGATACT	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATTAAAAA
CCAGGCAT	T	CCCTGGATAGTCACA	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATTAAAAA
CCAGGCAT	T	GGTTACGGGCGGCGT	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATTAAAAA
CCAGGGAT	T	ACACCGTCTTGATTC	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATTAAAAA
CCAGGCAT	T	TTCGGATAAACTGGG	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATTAAAAA
NCAGGCAT	T	CCGCGCCCACATTCT	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATTAAAAA
CCAGGCAT	T	GATGCTTTGGCAGAT	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATTAAAAA
CCAGGCAT	T	TATACAGACTATCTC	AGATCGGAAGAGCGTCGTGTAGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATTAAAAAA
NCAGGCAT	T	ATATTGCGTCGAGGT	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATTAAAAA
CCAGGCAT	T	GGCCTAGACCGCACG	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATTAAAAA
CCAGGCAT	T	CGCTGTTCTAACCAG	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATTAAAAA
CCAGGCAT	T	GCATTCCCCTTCCAT	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATTAAAAA
NCAGGCAT	T	AAGATGGGGTGAGGC	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATTAAAAA
CCAGGCAT	T	TCCACCCTGCTACCT	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATTAAAAA
CCAGGCAT	T	TGTCGCGAGACACCA	AGATTGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATTAAAAA

Figure 4.9 Read 1 and Read 2 raw data: Paired End band characterization

Snapshot of Read 1 and Read 2 sequencing data **(A)** Read 1 raw data from the bottom-bottom band. Read 5' → 3' of the parent strand: m=15 library, eight base Internal Index, 55 bases of Adapter 2, and 5 bases into the oligo-T lawn. **(B)** Read 2 raw data from the bottom-bottom band. Read 3'→ 5' of the complement strand: Internal Index, m=15 library, 58 bases of Adapter 1, and 5 bases into the oligo-T lawn.

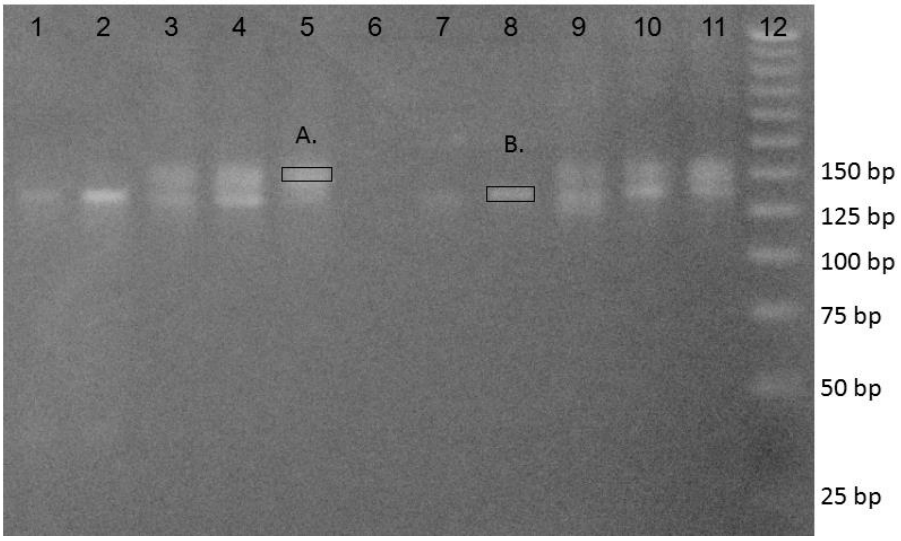


Figure 4.10 P1/P2, dual PCR products: AIA for thrombin, 60:1 ratio

4% TAE agarose gel. Lane 12: 25 bp DNA ladder. Lanes 1-5: Cycles 8 – 16, respectively, P1/P2.5. Lane 6: blank. Lanes 7-11: Cycles 8-16, respectively, P1/P2.6. **(A)** Top band gel extracted for qPCR. Contains Illumina Index 5. **(B.)** Bottom band gel extracted for qPCR. Contains Illumina Index 6.

Table 4.1. List of adapter library eight base internal indices

Internal Library Index	Sequence
1	ACACAGCA
2	ACAGTGCT
3	ACCGTAGC
4	ACGTCAAC
5	ACGTTGCC
6	ACTTAGCG
7	AGCAACAT
8	AGCTAACG
9	AGGCATCT
10	AGTAGCAT
11	AGTCAGTC
12	ATAGCGAC
13	ATAGTTGC
14	ATGCACGT
15	ATGCCTGG
16	ATGTCACT

Eight base internal indices are numbered 1 through 16. Sequences were designed by Dr. Damian Allis with minimum overlap for ease of de-multiplexing.

Table 4.2 List of Illumina Indices and associated P2 Primers.

Illumina Index	Sequence	P2 Primer
1	ATCACG	P2.1
2	CGATGT	P2.2
3	TTAGGC	P2.3
4	TGACCA	P2.4
5	ACAGTG	P2.5
6	GCCAAT	P2.6
7	CAGATC	P2.7
8	ACTTGA	P2.8
9	GATCAG	P2.9
10	TAGCTT	P2.10
11	GGCTAC	P2.11
12	CTTGTA	P2.12

Six base Illumina indices are numbered 1 through 12. Sequences are identical to those from the Illumina TruSeq Adapter 2 Indices. The associated P2 primers for use with the adapter library are named in the far right column.

Table 4.3 Statistical sequencing data: Single Read band characterization

Sample Name	Illumina Index	Total Reads	Good Reads	% Good Reads
Bottom-Bottom	4	670,553	586,471	87.46%
Bottom-Top	3	456,005	386,061	84.66%
Top-Bottom	2	474,495	404,044	85.15%
Top-Top	1	525,605	436,118	82.97%

Read 1 results for a simulated partitioning of m=15 DNA adapter library (Illumina MiSeq with TruSeq kit, Single Read). Total reads, good reads, and good reads as a percentage of the total reads.

Table 4.4 Statistical sequencing data: Paired End band characterization

Sample Name	Illumina Index	R1 Total Reads	R1 Good Reads	R1 %Good Reads	R2 Total Reads	R2 Good Reads	R2 %Good Reads
Bottom-Bottom	1	1,388,196	1,332,461	95.99	1,388,522	1,300,941	93.69
Bottom-Top	2	832,491	796,885	95.72	832,625	778,411	93.49
Top-Bottom	3	541,303	521,063	96.26	541,412	508,058	93.83
Top-Top	4	749,228	714,576	95.37	749,378	698057	93.15

Read 1 and Read 2 results for a simulated partitioning of m=15 DNA adapter library (Illumina MiSeq with TruSeq kit, Paired End). Total reads, good reads, and good reads as a percentage of total reads for Read 1 and Read 2.

Table 4.5. Summary of statistical sequencing data: AIA for thrombin, 60:1 ratio

Sample Name	Illumina Index	R1 Total Reads	R1 Good Reads	R1 Bad Reads
Bottom Band	6	555,626	255,388	265,504
Top Band	5	482,065	300,865	130,777

MiSeq results for 60:1, thrombin: m=15 DNA adapter library (TruSeq Single End Read). The high percentage of bad reads is due to contamination from the TBAsc sequence. TBA was not found in either data set. The top twenty sequences for each sample can be found in **Appendix 4.**

Table 4.6. Summary of statistical sequencing data: AIA for thrombin, DNA and 2'-OMe

Sample Name	Illumina Index	Internal Index	R1 Total Reads	R1 Correct Internal Index	R2 Total Reads	R2 Correct Internal Index
DNA 60:1	5	ATGCCTGG	429,493	412,197	429,617	401,838
DNA 1:1	7	ATGCCTGG	441,140	415,268	441,223	401,424
2'-OMe 60:1	6	AGCTAACG	468,343	55	468,448	189
2'-OMe 1:1	8	AGCTAACG	139,269	21	139,289	30

MiSeq sequencing results for 60:1 and 1:1 thrombin: library experiments with m=15 DNA and m=15 2'-OMe RNA/DNA chimera adapter libraries (TruSeq Paired End Read). The incorrect internal index was identified for both 2'-OMe RNA/DNA chimera library samples. TBA was not found in either the correctly indexed DNA library samples or the incorrectly indexed 2'-OMe RNA samples.

Chapter 5: Alternative Ligation Approaches for Amplification Free-AIA

Chapter Summary

The failure of the adapter library design in combination with the success of the hairpin loop libraries in AIA for thrombin prompted the return to a ligation based approach for library design and capture. It was concluded that the structure of the adapter library was interfering with the binding of TBA to thrombin during partitioning. As demonstrated in chapters 2 and 3, if the TBA quadruplex structure was made available for binding by the construct of the library, it is highly likely that the sequence would be selected above background using Concanavalin-A beads as an immobilization method. Therefore, it is likely that the adapter constructs inhibited aptamer discovery for short libraries such as m=15. In order to investigate this hypothesis, methods to capture un-flanked or minimally flanked libraries were explored. Minimally flanked libraries allow substantial flexibility in secondary structure formation while un-flanked libraries allow complete freedom in secondary structure formation. Failure of the variable length hairpin loop DNA and 2'-OMe RNA/DNA chimera libraries in the selection against multiple Epigenetic targets indicated that a greater freedom in secondary structure formation may be necessary if AIA is to be successful as a universal aptamer discovery method. Compared to the length of the short variable regions of AIA libraries, the adapter regions contributed to the bulk of the molecule (45 of the 61 bases in an m=15 library are adapter constructs). Consider the size comparisons illustrated in **Figure 5.1**.

A second benefit to reconsidering a ligation based capture approach is the ability to eliminate the amplification step required with the adapter libraries. Ligating full length adapters to un-flanked or minimally flanked libraries and proceeding directly to sequencing shortens the AIA protocol and eliminates a potential source of error. Without amplification, sequence data would accurately represent the selected library sequences with zero amplification bias during sample preparation. Primer-free or minimal-primer protocols have been used in SELEX.[124, 126] However, multiple rounds of selection require a lengthy procedure of primer ligation, PCR amplification, and enzymatic digestion to regenerate the primer-free or minimal-primer library. This method is suitable for long libraries where only one or a few copies exist for each sequence and enrichment is required in order to identify aptamer candidates. However, the short libraries used in AIA, ranging in length of m=15 to m=22, allow for thousands of copies of each possible sequence which eliminates the cycling requirement. Using short un-flanked or minimally flanked libraries allows for one cycle of selection and adapter ligation prior to sequencing.

This chapter outlines preliminary methods development for an un-flanked m=15 library with various capture techniques, including randomized adapter splints and T4 DNA ligase capture, hairpin adapters and T4 DNA ligase capture, single-stranded adapters and T4 RNA ligase/5'APP DNA/RNA ligase capture, and chemical ligation. This chapter also outlines methods development for minimally flanked libraries with non-complementary four base tails and their application in AIA for thrombin.

Ligation of Un-flanked Libraries

Utilizing randomized splints and T4 DNA Ligase capture

In this variation on ligation-based AIA, un-flanked libraries were captured with full length adapters ligated at the 5' and 3'-termini with randomized splints to facilitate annealing. The experimental design is outlined in **Figure 5.2A**. Specific sequences have been omitted for simplicity and can be found in **Appendix 5**. The Illumina TruSeq adapter sequences were used in designing Adapter 1 (58 bases), Adapter 1 Complement (59-62 bases), Adapter 2 (64 bases), and Adapter 2 Complement (65-68 bases) which were ordered from IDT. The six base Illumina Index is found in Adapter 2 and Adapter 2 Complement similarly to the P1/P2 primers used for amplification of the adapter libraries. During protocol development, Illumina Index 1 was used. In order to ligate an un-flanked library, the "sticky ends" used to capture the library were randomized as well. The length of the randomized sticky ends on the complement strands that were tested had n=1, 2, 3 or 4. For fixed sequences, a 4 base splint would have higher ligation efficiency due to more stable annealing of the two strands. However, combinatorial considerations must be made for libraries. When n=4, there are 4^8 (65,536) possible sequence combinations. As "n" is decreased, the number of possible sequences decreases. When n=3 there are 4^6 (4,096), when n=2 there are 4^4 (256) and when n=1 there are 4^2 (16) possible sequence combinations. If a library pool becomes significantly skewed towards a specific family of sequences following a partitioning (i.e. G/T rich sequences for thrombin aptamers), the ratio of available complementary adapters changes. For a skewed library, the percentage of complementary adapter sequences decreases when the length of the complementary region increases. Whether or not a balance between hybridization efficiency and the percentage of complementary adapters can be found for this experiment is unknown, as the current partitioning

methods skew the library in an unpredictable fashion. It does not appear that the skewness of an AIA partitioned library is substantial enough to be a limiting factor in library capture based on TBA and TBA variant frequencies presented in chapters 2 and 3. Although the quantity of library recovered following a partitioning step varies, the adapters are used in enough excess to likely overcome any combinatorial limitations for n=1, 2, 3 or 4.

After a successfully ligated library is agarose gel purified, the sample would be quantified with qPCR. Although the ligated product is not completely double-stranded, qPCR and high throughput sequencing are still possible. Both the parent and complement strands are not needed at the start of either protocol. During the first cycle of qPCR, the complement strand is generated and data output is adjusted to compensate for the single-stranded template. During cluster amplification on the Illumina flow cell, clusters would be generated from annealing of the parent strand alone. The complement strand required for Read 1 and the Index Read would be generated during bridge amplification. The Adapter 1 Complement containing the P5 region is capable of hybridizing to the Illumina flow cell, however, the entire sequence length would not be generated and bridge amplification would fail. Nevertheless, it may interfere with cluster generation of the desired product by occupying sites on the oligo lawn. A second design scheme using a shorter Adapter 1 Complement (shortened from 62 to 37 bases when n=4) lacking the region complementary to the flow cell oligo lawn is shown in **Figure 5.2.B**.

Preliminary experiments to ligate an m=15 un-flanked library were successful to some measure. The experimental design included a conventional one step ligation approach using the four full

length adapters (complement length varied: n=1, 2, 3, 4) and m=15 insert. A 10 pmol quantity of m=15 library, TBA (dGGTTGGTGTGGTTGG) and TBAsc (dGGTGGTTGTTGTGGT), (IDT) as well as a no insert control were ligated to adapters in a 1:1 ratio. The oligonucleotide mixture was denatured at 95°C for 3 minutes. After cooling to room temperature, 7 µl of 10X ligation buffer (300 mM Tris-HCl, 100 mM $MgCl_2$, 100 mM Dithiothreitol, 10 mM ATP, pH 7.8), (Promega), 1 µl T4 DNA ligase (50% glycerol stock, 3 u/µl), (Promega) and dH_2O to 20 µl were added. Following an overnight incubation at 16°C, the products were visualized on a 4% TAE agarose gel. The ligation products of the m=15 library, TBA and TBAsc with full length adapters with n=1, n=2, n=3 and n=3 are presented in **Figure 5.3**. The desired product length is 137 base-pairs. The band at approximately 125 base-pairs is found for all three inserts and the no insert control, indicating that it is not the desired ligation product. Excess adapters are present between 50 and 75 base-pairs.

To determine if self-ligation of the adapters was causing the 125 base product, the m=15 library was ligated to Adapter 1/Adapter 1 Complement and Adapter 2/Adapter Complement independently of each other. Ligations conditions, including concentrations, were maintained from **Figure 5.3** for **Figure 5.4.A**. When ligated independently from each other, the source of the 125 base-pair product is identified as the self-ligation of the double stranded Adapter 2/Adapter Complement. **Figure 5.4.A** shows that the m=15 insert was successfully ligated to the Adapter1/Adapter 1 Complement complex when n=2. In an attempt to improve ligation efficiency, the Adapter 1 ligation was repeated with samples cooled in a thermocycler at a rate of 0.5°C per second rather than simply cooling on the bench top at room temperature following heat

denaturation. **Figure 5.4.B** shows that when n=1, Adapter 1 and the insert did not ligate efficiently. For n≥2, a ligated product is visible, indicating that controlled cooling aids in annealing of the splints when n≥2. To resolve the problem of self-ligation of Adapter 2/Adapter 2 Complement, the use of a 3'-blocking group was implemented. An amino modifier replaced the free hydroxyl group at the 3'-end of Adapter 2 Complement (Ordered from IDT). **Figure 5.5.A** illustrates how the self-ligated product forms, while Figure 25.B illustrates the location of the 3'-amine.

Ligation of the m=15 library, TBA, and TBA in the hairpin loop format to Adapter 1/Adapter 1 Complement and Adapter 2/Adapter 2 Complement (3'-amine) independently was performed under identical conditions to **Figure 5.4.B** with controlled cooling following denaturation. TBA in the hairpin loop format was chosen as a control against the un-flanked molecules. Results from the ligation with Adapter 2 and Adapter 2 Complement (3'-amine) indicate that the 3'-terminal amine successfully prevents self-ligation of the Adapter 2/Adapter 2 Complement complex, as shown in **Figure 5.6.** By eliminating the self-ligation product, a greater quantity of the Adapter 2/Adapter 2 Complement complex is available to ligate the library or control sequence. Although the self-ligation product was eliminated, ligation results were still poor for both Adapter 1 and Adapter 2. The desired product is slightly visible in lanes 3 and 4, while un-ligated insert is clearly visible in lanes 5 and 9 at approximately 25 bp. To determine whether or not consistency could be achieved, the temperature and reaction time were varied from the previous 16°C for 18 hours/overnight. As temperature decreases, molecular movement slows and the time of association increases. The melting temperature of the 1-4 base complementary region of the adapter and library varies based on length and sequence, but it is always small (<12°C). As

indicated by the results in **Figure 5.7**, as the ligation temperature is decreased and reaction time is increased, the quantity of correctly ligated product increases. However, additional experiments at 4°C with complementary regions of varied length (n=1, 2, 3, or 4) for m=15, TBA and TBAsc illustrated a lack of consistency. For this reason, additional methods of capturing the un-flanked library were explored.

Hairpin adapters and T4 DNA Ligase capture

A hairpin adapter structure was explored in an attempt to stabilize the Adapter/Adapter Complement complex. Although the full length adapters are highly specific, it was hypothesized that the hairpin format may improve accuracy of the Adapter/Adapter Complement hybridization, which would ensure the correct overhang (n=1, etc.). The corresponding Adapter/ Adapter Complement pairs were designed as single-stranded molecules, with the two fragments joined by a 5 base loop (dT-dT-rU-dT-dT). The loop allows the Adapter/Adapter Complements to anneal in a hairpin to form the correct adapter constructs for the 5' and 3'-ligations. Adapter 2 contains a 3'-terminal amine to prevent self-ligation as discussed in the previous section. The ligation scheme is illustrated in **Figure 5.8.A.** Following ligation, the "bottom" complement strands of the hairpins would need to be removed to eliminate large interference with hybridization to the Illumina flow cell. Treatment with 1M NaOH successfully cleaves the complement strands by hydrolyzing the single RNA base as confirmed with IE-HPLC. The remaining bases of the hairpin would not likely interfere will hybridization to the Illumina flow cell. The cleaved Adapter 1 fragments could be separated from the ligated product via denaturing PAGE purification to prevent competition on the flow cell oligo lawn. To test the applicability of

the hairpin adapters, shortened adapters with n=4 splints were synthesized on the ABI 394 synthesizer (Adapter 1 = 75 bases and Adapter 2 = 59 bases). The full length adapters would range in size from 132 – 145 bases and the synthesis would be difficult in-house. If the ligation scheme was successful with the shortened adapters, the more costly full length adapters would have be ordered from IDT. Several ligation experiments with varied conditions were unsuccessful. A control experiment using the previous hairpin loop library and hairpin adapters with corresponding fixed complementary splints was also unsuccessful. This is likely due to insufficient purity of the adapters synthesized on the ABI 394. The method was not investigated further following these initial experiments.

Single-stranded adapters and T4 RNA ligase I /5'APP DNA/RNA ligase capture

This capture strategy employed ligation of a single stranded Adapter 1 and Adapter 2 to the un-flanked m=15 DNA library with T4 RNA ligase I and Thermostable 5'APP DNA/RNA ligase (New England BioLabs) as illustrated in **Figure 5.8.B.** RNA ligase is capable of ligating ss-RNA and/or ss-DNA by blunt end ligation. Eliminating the adapter complements eliminates any combinatorial or hybridization limitations introduced by the 1 to 4 base complementary splint. Although an expensive and tedious procedure, if the ligation with RNA ligase were efficient, it would be an ideal method for capturing the partitioned un-flanked library. The two step ligation first requires the generation of AppDNA Adapter 2 from pDNA Adapter 2 with the 5'DNA Adenylation kit (New England BioLabs). In the first ligation step, Thermostable 5'APP DNA/RNA ligase (New England BioLabs) joins the 5'App of Adapter 2 to the free 3'-OH of the library. The enzyme is heat denatured to terminate the ligation. Gel purification of the single

stranded product would be difficult and is not required due to the use of two different enzymes. In the second ligation step, T4 RNA Ligase I joins the 5'-Phosphate of the library to the free 3'-OH of Adapter 1. Excess Adapter 2 will not ligate to free Adapter 1 due to the 5'App. Experimental details are omitted for simplicity, as the protocol did not yield the correct ligated product. The method was not pursued further do to the costly nature of the reagents and initial failures.

Chemical ligation

The final capture strategy employed the double-stranded adapter splints in combination with chemical ligation, "Click" DNA ligation, at the 5'-and/or 3'-ends of the insert. Copper(II) catalyzed azide-alkyne cycloaddition (CuAAC) is a form of click chemistry used to ligate DNA.[167, 168] The trizole linkage that forms between labeled oligonucleotides can be read by DNA polymerases,[167, 169] making it a viable option for applications with downstream qPCR and high throughput sequencing. An important feature of click ligation is that the 5'-terminal azide and 3'-terminal alkyne are not found in natural systems, eliminating issues of non-specific or self-ligation of adapters. The click ligation could be used in combination with an enzymatic ligation to eliminate self-ligation of the Adapter 2/Adapter 2 Complement complex and perhaps improve ligation efficiency. For example, ligation with DNA ligase of Adapter 1 to 5'-phosphorylated library and click chemistry for ligation of 3'-terminal alkyne library and 5'-terminal azide Adapter 2. **Figure 5.8.C** shows a schematic representation of how click chemistry was applied to the ligation of un-flanked libraries in preliminary experiments. The adapters and libraries were synthesized using terminally labeled amidites from Glen Research. Additionally,

the fixed m=15 TBA and TBAsc sequences were synthesized with click modifications along with corresponding adapter complements with fixed four base complementary overhangs. Specific synthesis information has been omitted for simplicity because the protocol was not continued after preliminary experiments. Sequences can be found in **Appendix 5**.

The experimental protocol followed the method by Lumiprobe, Corp. utilizing a Cu-TBTA complex as the reaction catalyst.[170] Several problems were encountered during preliminary experiments. In order for the azide-modified oligos to remain soluble, DMSO is required in the reaction buffer. While loading the ligated samples onto an agarose gel, the high concentration of DMSO caused the solution to precipitate into the 1X TAE running buffer. Desalting columns pre-equilibrated with 50% DMSO and ethanol precipitation were used independently to exchange the reaction buffer to successfully prevent precipitation of the sample during loading. The click ligation was unsuccessful for all attempts with un-flanked libraries and fixed sequences. Several measures were taken to establish why the click ligation was not successful, including IE-HPLC purification of the oligos (not shown) and ESI/LC Mass Spectrometry Analysis (Novatia, LLC, Newton, PA) of the TBA with 5'-terminal azide and 3'-terminate alkyne (**Figure 5.9**). It was critical to determine if the chemistry involved in synthesizing the modified oligos was successful. From the mass spectroscopy data, it was concluded that the chemistry was successful and that the TBA oligo contained both the 5'-terminal azide and the 3'-terminal alkyne. The click ligation protocol was not investigated any further due to simultaneous improvements mentioned next, however, a variety of protocols utilizing different reagents are available and may be investigated in the future.

Ligation of Minimally Flanked Libraries

Due to inconsistencies encountered with ligation of un-flanked libraries, a library flanked with a short fixed regions on both the 5' and 3'-termini was introduced. The library contains the same "ACAC/CACA" tails of the hairpin loop library but without the eight base complementary stem. The adapter design with specific sequences modeled after the Illumina TruSeq Adapters is shown in **Figure 5.10**. The adapter complements capture the library via the ACAC/CACA tails with a four base fixed sequence. Adapter 2 Complement does not require the 3'-amino modifier to prevent self-ligation because the fixed four base overhangs are not complementary.

This "tailed" library format was first tested in a two-step approach: ligation of the m=15 insert to Adapter 1/Adapter 1 Complement followed by agarose gel purification and subsequent ligation of Adapter 2/Adapter 2 Complement and vice versa. The ligation was successful with either adapter used in the first step; however, a significant amount of product was lost during the gel purification. A two-step approach without the intermediate gel purification step of the first product was successful, indicating that it is not necessary. This suggested that a one-step protocol may also work efficiently. A second version of the adapter complements was also tested in an attempt to improve the ligation efficiency and consistency. The fixed four base overhangs were extended to a total length of eight bases with the addition of four randomized bases. The sequences are provided in **Figure 5.10.** The additional n=4 overhangs were hypothesized to improve ligation efficiency by providing additional structural integrity to the annealed products without lengthening the fixed tails of the library. The Adapter 2 Complement contains a 3'-amino modifier to prevent self-ligation because the 5'-NNNNGTGT-3' of Adapter 1 Complement and the 5'-TGTGNNNN-3' of Adapter 2 Complement are capable of hybridizing.

Sequence differences between the “Fixed” and “Fixed/n=4” adapter complements are highlighted in red in **Figure 5.10.**

A comparison of the efficiency the “Fixed” and “Fixed/n=4” adapter complements in a one-step ligation is shown in **Figure 5.11**. A 10 pmole quantity of the m=15 tailed library as well as two fixed control sequences in the same minimally flanked format, TBA and TBAsc, were ligated to the four corresponding adapters at a 1µM final concentration (1:1 molar ratio). The oligonucleotide mixtures were denatured at 95°C for 3 minutes and cooled in a thermocycler at a rate of 0.5°C per second to 4°C prior to the addition of 7 µl of 10X ligation buffer (300 mM Tris-HCl, 100 mM $MgCl_2$, 100 mM Dithiothreitol, 10 mM ATP, pH 7.8), (Promega), 1 µl T4 DNA ligase (50% glycerol stock, 3 u/µl), (Promega) and dH_2O to 20 µl. Following a 24 hour incubation at 4°C, the products were analyzed on a 4% TAE agarose gel as shown in **Figure 5.11**. The correct 145 bp product is seen for m=15, TBA and TBAsc with the Fixed/n=4 complement and for m15 and TBA with the Fixed complement. The intensity of the ligated m=15 library is significantly greater for the Fixed/n=4 complement structure. This indicates that the Fixed/n=4 complement format is superior and the structural integrity introduced by the additional four base randomized overhang increases ligation efficiency.

The high efficiency of the Fixed/n=4 complement format is particularly important during AIA because the quantity of recovered library is much less than the 10 pmoles used in the control experiments. To demonstrate whether or not this ligation scheme can successfully capture much smaller quantities of library, a dilution series of the m=15 library was ligated with the Fixed/n=4

complement format in a one-step protocol. The oligonucleotide mixtures were denatured at 95°C for 3 minutes and cooled in a thermocycler at a rate of 0.5°C per second to 4°C prior to the addition 7 µl of 10X ligation buffer (300 mM Tris-HCl, 100 mM $MgCl_2$, 100 mM Dithiothreitol, 10 mM ATP, pH 7.8), (Promega), 1 µl T4 DNA ligase (50% glycerol stock, 3 u/µl), (Promega) and dH_2O to 20 µl. Following a 24 hour incubation at 4°C, the products were visualized on a 4% TAE agarose gel as shown in **Figure 5.12.** The correct 145 bp product is clearly visible for 10 pmoles and 1 pmole of library. A faint band is present for 0.1 pmoles and no product is seen for 0.01 pmoles. As the concentration of insert decreases, the presence of single-ligation products diminishes prior to the correct double-ligation product. This indicates that a partitioned library with concentration too low to be visualized would be ligated correctly. The visible and non-visible products were excised from the gel and purified with the MiniElute Gel Extraction Kit (Qiagen). The purified products were quantified in triplicate by qPCR with the KAPA Library Quantification kit. The purpose of this was twofold: to determine the average recovery and to also confirm that the ligated product is in fact amplifiable with the Illumina platform. The average recovery values for the four dilutions (10, 1.0, 0.1 and 0.01 pmoles) were calculated as 8.47nM, 0.37nM, 0.36nM and 0.11nM, respectively. This correlates to recovery values of 0.08 pmoles, 0.0037 pmoles, 0.0036 pmoles and 0.001 pmoles. The lower than expected calculated recovery is due in part to incomplete excision of the product from the gel. A 4.0 mm by 1.0 mm disposable gel excision tip (GeneCatcher) is used to excise the band and only excises the core of the band. Also, the Qiagen MiniElute Gel Extraction Kit has a relatively short shelf life and product may have been lost due to this. It is impossible to calculate a percent recovery without also knowing the quantity of library that was successfully ligated, however, the quantity of product recovered follows the decreasing trend of the dilution series. Additional ligated samples

used in the optimization of the ligation scheme (including those from **Figure 5.12**) were also quantified by qPCR and confirm that the gel purified product can be successfully quantified with the Illumina platform via the KAPA Library Quantification kit.

In order to apply the new library design to AIA, multiple indexed adapters were needed. By multiplexing samples on a single Illumina flow cell lane (for the MiSeq, a single run), a higher rate of sample/data output can be achieved. For the initial studies, four Illumina Indices (2, 4, 6 and 12) were ordered from IDT, in addition to Illumina Index 1 used in the initial design experiments. The selection of the four indices was based on guidelines from Illumina for multiplexing small numbers of samples, taking into considering the level of sequence similarity. **Table 5.1** illustrates the best options for multiplexing with Illumina Indices. An additional change was made to the Adapter 2 and Adapter 2 Complements: the Adapter design in **Figure 5.10** includes a single dA base at the 5'-terminus immediately prior to the start of the Illumina TruSeq Universal Adapter sequence. This dA was carried over during adapter design from the adapter library from chapter 3 and is unnecessary for this format. The single base was removed from the sequence of the four new adapter sets. This change produced an unexpected benefit in the analysis of sequence data in experiments mentioned in the next section. The original format with the extra dA, Index 1, was used as a size marker for gel purification of small quantities of ligated, partitioned library that were typically too dilute to visualize. During data processing, sequences containing Illumina Index 1 can be counted to reveal the level of contamination (if any) from the size marker. By looking at the raw data, the additional dA helps to confirm that the Index 1 sequence was in fact contamination from the size marker.

Although the sequence differences were not suspected to affect the outcome of ligation, the four indices were tested for consistency by ligating 10 pmoles of library under identical conditions. There was no discernable difference in the ligation efficiencies of the four new adapter sets and the original adapter set used in the experiment design. Based on the consistency at which the new library format is ligated and the success in multiple qPCR tests, the m=15 DNA tailed library was tested in AIA for thrombin.

Application of the Tailed m=15 DNA Library to AIA

Workflow summary

Following confirmation that the Fixed/n=4 adapter splints ligate the tailed m=15 DNA library successfully, the library was tested in AIA for thrombin. The goal was to determine if the library structure allowed for TBA to be selected from the over-represented library pool using Concanavalin-A as an immobilization method. If successful, additional partitioning methods would be explored. The m=15 DNA library was partitioned against thrombin immobilized on agarose-Concanavilin-A beads as described in chapter 4. Phenol extraction and ethanol precipitation resulted in a partitioned DNA library containing sequences with an affinity for human α-thrombin. Multiplexed adapters were ligated to the partitioned library and the product was agarose gel purified. The product was quantified via qPCR analysis in order to determine the precise quantity required for clustering on the Illumina MiSeq Desktop Sequencer (SUNY Microarray Core Facility, Upstate Medical University). The adapters are the full length required for sequencing on the Illumina platform and thus require no PCR amplification. Elimination of the amplification step prompted a more in depth analysis of the quantification data relative to the

starting quantity of library. Sample preparation with the hairpin loop and adapter libraries employed qPCR following amplification, so the quantity of recovered and ligated library was unknown. In the absence of amplification, it was found that a superior level of control and predictability in the outcome of high throughput sequencing data could be achieved. Calculations based on the quantity of ligated library in combination with the starting quantity of library pool and the average cluster output are capable of determining if it is possible for any sequence to appear above background. If the quantity of recovered library is too large and/or the number of expected clusters is too small, it may be statistically impossible for any sequence to appear above background. As our knowledge of the Illumina sequencing platforms improved, we have been able to predict whether or not a sample is worth sequencing based on the qPCR data and expected cluster output. We can predict a relative maximum frequency of aptamer candidates using these calculations. This capability is in line with the goals of AIA, which include quickly and efficiently identifying aptamers while conserving resources.

Partitioning the DNA Library

Partitioning of the DNA library was performed using agarose-Concanavalin-A beads (GE Healthcare Life Sciences) at room temperature The Con-A beads contained in spin columns (ThermoScientific) were pre-equilibrated in partitioning buffer (20 mM Tris-HCl, 140 mM NaCl, 5 mM KCl, 1 mM $CaCl_2$, 1 mM $MgCl_2$, pH 7.4) by three wash cycles. The following represents a 5:1 DNA library: protein screen. A 20 pmol aliquot of glycosylated human α-thrombin (Haematologic Technologies) in 200 µl partitioning buffer was immobilized on 10 µl of pre-equilibrated Con-A beads during a 30 minute incubation with end-over-end rotation at 25

rpm at room temperature. Unbound protein was removed by centrifugation at 1,500 rpm for 1 minute. A 100 pmol aliquot of the DNA library in 200 µl partitioning buffer was then applied to the immobilized thrombin. Following a 1 hour incubation at room temperature with end-over-end rotation at 25 rpm, the unbound DNA was removed through centrifugation. Three wash cycles (5 minutes, rotation at 25 rpm) with 200 µl partitioning buffer ensured thorough removal of unbound DNA. Thrombin-DNA complexes were eluted from the Con-A beads with 200 µl 1M α-methyl mannoside (Sigma Aldrich) during a 5 minute incubation, rotation at 25 rpm.

Phenol Extraction and Ethanol Precipitation

Phenol extraction was used to recover the partitioned DNA libraries. Equal volume Tris-buffered, pH 8.0, 0.1mM EDTA, 50% phenol, 48% chloroform, 2% isoamyl alcohol (Sigma Aldrich) was added and vortexed for 30 seconds. After centrifugation at 13,000 rpm for 5 minutes the aqueous layer was removed and the library subsequently extracted twice with equal volumes of 100% chloroform. The library was then purified by ethanol precipitation. One tenth the volume (20 µl) of 3M sodium acetate and three times the volume (600 µl) of cold ethanol were added to the library and mixed briefly. Following an overnight incubation at 20 °C, centrifugation at 13,000 rpm for 30 minutes resulted in a library pellet. After decanting the liquid, the pellet was washed with 1 ml 70% ethanol, and the pellet was dried in a SpeedVac and resuspended in 10 µl dH_2O.

Modifying the Library for Sequencing

The libraries were ligated with full length adapters and their complements. Adapter 1, Adapter 1 Complement, Adapter 2 (indexed), and Adapter 2 Complement (indexed) at 10 µM were added in 1 µl volumes to the recovered library. The indexed Adapter 2 was dictated by the number of samples prepared and the number of samples desired per Illumina flow cell lane. The mixtures were incubated at 90°C for 3 minutes and cooled in a thermocycler at a rate of 0.5°C per minute to 4°C. At 4°C, 2 µl of 10X ligation buffer (300mM Tris-HCl (pH 7.8), 100mM MgCl2, 100mM DTT and 10mM ATP), (Promega), 3 µl dH_2O and 1 µl T4 DNA ligase (10mM Tris-HCl (pH 7.4), 50mM KCl, 1mM DTT, 0.1mM EDTA and 50% glycerol), (Promega) were added. Incubation of the 20 µl reaction for 24 hours at 4°C completed the ligation. The ligated library was purified by gel electrophoresis. Excision of the ~144 base-pair band from 4% TAE agarose gel (3% NuSieve GTG Agarose, 1% high melt agarose) with a 4.0 mm by 1.0 mm disposable gel excision tip (GeneCatcher) ensured minimal risk of cross-contamination between gel lanes. The library was extracted from the gel with the MiniElute Gel Extraction Kit (Qiagen) and resuspended in 10 µl of dH_2O.

Control Experiments

During each experiment, a control ligation was used as a size marker in the event that the ligated experimental sample was too low of a concentration to be visualized. A 10.0 pmol aliquot of the m=15 DNA is ligated under identical conditions with Adapter 2/Adapter 2 Complement (Index 1). The product was gel purified alongside the experimental samples as a ~144 bp size marker and quantified via qPCR as the positive control.

Quantitative PCR and *High Throughput Sequencing*

The precise quantity of amplifiable DNA was determined with the KAPA Library Quantification kit on the BioRad iCycler. Unlike previous methods, the concentration of the sample was not estimated prior to qPCR analysis to reduce loss. A 1 µl quantity of each sample was required for qPCR and was diluted to 1000 µl for analysis. All samples were quantified in triplicate to ensure accuracy. The data is reported as pM concentration and any outliers were eliminated prior to averaging the data points. The quantity of recovered DNA was calculated by the following method:

Average concentration x 2 = Concentration of single-stranded product (to account for single-stranded template)

Concentration of single-stranded product x (452/144) = Concentration of 1000 µl dilution (Size adjustment for the KAPA control)

Concentration of 1000 µl dilution in pM x (1000/1000) = Concentration of Stock solution in nM

Concentration of Stock solution in nM / 100 = Total pmoles recovered

Total pmoles recovered x 0.9 = pmoles available for use in sequencing (to account for 1 µl used in qPCR)

The quantity of sample still available for sequencing is important for determining whether or not a sample should be sequenced. The MiSeq allows for a range of concentrations of oligonucleotide to be applied to the flow cell. For this application, the maximum recommended concentration was used due to the short length of the sequences (144 base-pairs) and resulting small clusters as compared to commonly used genomic DNA. A 5 µl volume of 4 nM sample is required and from this, a 20 pM denatured solution was made and loaded onto the flow cell. The

4 nM concentration assumes that the modified libraries are double-stranded. To account for the single-stranded nature of the modified library, an 8 nM concentration was supplied. This indicates that 0.04 pmoles of single-stranded library is required. When multiplexing the four indexed adapters currently in use, a maximum recovery of 0.011 pmoles of each sample is desired. **Table 5.2** demonstrates why it is desired to recover 0.011 pmoles or less of each sample per four-plex sequencing run using a hypothetical partitioning experiment using 100 pmol m=15 DNA library with varied recovery.

Table 5.2 demonstrates that for a sample with a total recovery of 0.02 pmoles and a perfect recovery of the 56,000 TBA molecules in the starting pool, only 28,000 of these will be loaded on the flow cell. If 5,000,000 clusters are formed from this sample, only 23 of the original 56,000 copies of TBA will be clustered. Based on prior experiments, this is enough to select an aptamer from the background, however, the number of expected counts for any unique sequences becomes much lower as the quantity of library recovered increases. This indicates that the partitioning method must be stringent in eliminating non-specific sequences to reduce background. The ideal case for a four-plex sequencing run is to have 0.01 pmoles or less of each sample available, with lower quantities producing the most favorable result. This means that ~0.011 pmoles or less of ligated library is the ideal for combined use in pPCR and clustering.

Data Analysis

The goal of the first set of AIA experiments was to demonstrate that TBA could be identified from the tailed library structure. A 100 pmol aliquot of the m=15 DNA tailed library was

selected against 20 pmol of thrombin immobilized on Con-A (1:5, thrombin: library ratio) in four replicates to demonstrate consistency in sample preparation. Following indexed ligation, the samples were purified on a 4% TAE agarose gel, shown in **Figure 5.13**. The samples were quantified via qPCR and the quantity of library recovered and the expected number of counts of TBA (or any unique sequence) were calculated. Calculations are provided in **Table 5.3**. The results indicate that too large a quantity of library was recovered and that the frequency of any unique sequence would not likely be found above background. The estimated number of clusters per sample for a four-plex sequencing run (5,000,000) is optimistic. Data from previous MiSeq runs indicates that 1,000,000 – 2,000,000 may be more appropriate, which would reduce the counts even further. With 100% recovery of the 56,000 starting copies of TBA and 5,000,000 clusters, TBA is expected to be sequenced 11, 3, 31 or 7 times for samples 1-4 , respectively. If only 1,000,000 clusters are sequenced, this number drops to 5, 0, 6 and 1, respectively. With such low counts expected for the samples, they were not sequenced and modifications were made to the partitioning method in an attempt to improve stringency.

Improving Partitioning Stringency

If it is assumed that all possible aptamer candidates exist in the pool of partitioned library, the only manner of increasing the expected counts is to reduce the background – the number sequences randomly carried forward with the specifically bound molecules in partitioning. Several changes were made to the partitioning method in order to accomplish this. The first included re-evaluating the quantity of Con-A beads. The quantity of beads required for a 1:5 ratio screen utilizing 20 pmoles of thrombin was determined by evaluating the binding capacity

of the Con-A beads. Based on data from previous chapters demonstrating variability in the behavior of beads, this determination would be required for each new lot of Con-A beads. A series of volumes of the slurried beads sold by the commercial supplier (100 µl, 50 µl, 20 µl, 10 µl, 5 µl and 1µl) were pre-equilibrated in partitioning buffer by three wash cycles. A 100 pmol aliquot of thrombin was applied to the beads in 200 µl of partitioning buffer and allowed to bind for 30 minutes with end-over-end rotation at 25 rpm at room temperature. The beads were then centrifuged for 1 minute at 1,500 rpm. The quantity of protein in the flow-through was determined by absorption at 280 nm with the NanoDrop UV/Vis spectrophotometer. It was found that no significant difference in binding capacity was seen for 20 µl or less of Con-A beads, with an average of 80% of the 100 pmoles of thrombin binding for these volumes. According to the manufacturer, the Con-A beads have a binding capacity of 30-67 nmoles/ml.[171] Using this stated value, to bind 20 pmoles of thrombin at the beads' minimum binding capacity, 0.66 µl of Con-A beads are required. Based on the ratio of protein: beads used in the original AIA protocol, 3.3 µl of Con-A beads are required for 20 pmoles of thrombin and were used for the next two sets of experiments (reduced from 10 µl). By decreasing the quantity of beads void of immobilized thrombin, the probability of retaining sequences with affinity for Con-A and non-specific binding sequences was reduced. The second modification was a wash step after the protein has been immobilized to the Con-A beads. It is possible that unbound thrombin is still present after the initial centrifugation. Although this step may not necessarily reduce the quantity of library recovered, it did ensure that any thrombin present in the partitioning step is bound to Con-A. The third modification was increasing the binding time of the library from 1 hour to overnight (~18 hours). The fourth modification was increasing the number of wash steps from three to six. This step was hypothesized to make the most difference in reducing the background.

A fifth modification used in two of the four samples was the reintroduction of the negative selection step. This step was eliminated in previous methods because the Con-A beads were absorbing high percentages of the library prior to application to the protein. The lot of Con-A beads used for this set of experiments was found to absorb approximately 20% of a 100 pmol aliquot of tailed m=15 DNA library. By negatively selecting the library against these Con-A beads, it may be possible to reduce the number of non-specific sequences recovered without significantly reducing the number of potential aptamer candidates. Four samples were prepared with the above modifications as described in **Table 5.4**. The first two samples compared the effect of negative selection on selection stringency while the third sample compared the effect of target: library ratio. The fourth sample (no thrombin) served as a negative control. Following ligation, the samples were gel purified as shown in **Figure 5.14**. As compared to **Figure 5.13**, the location of the 144 bp product band for these samples is not as intense, indicating a smaller quantity of recovered library, which was initially promising. Following qPCR of the gel purified samples, the quantity of library recovered and the expected number of counts of TBA (or any unique sequence) were calculated**.** The specific calculations are provided in **Table 5.5**. The quantity of library recovered for this sets of samples is much closer to the desired 0.01 pmoles of each than in the first set of experiments. To compensate for the fact that sample 1 produced less than 0.01 pmoles of library available for sequencing, the quantity of the other 3 samples in each set was increased to reach a total of 0.04 pmoles. This has been accounted for in the data table. The four samples were sequenced on a multiplexed TruSeq Single Read run on the MiSeq.

Data Analysis

Following sequencing on the Illumina MiSeq, the data was parsed based on the Illumina Indices. These data files were then parsed using a modified Perl Script that identifies the four base ACAC head and four base CACA tail flanking the m=15 library region. The Perl script identified sequences with perfect matches for both the head and tail as "candidate" reads and those with an m=15 variable region were qualified as "good" reads. The Perl script was originally designed to function optimally with the longer head and tail of the hairpin loop library. For this reason, the Perl script must parse the data with strict qualification in order to function correctly. The counts per unique sequences were tallied and this data is presented in **Table 5.6**. A brief search of the "candidate" and "bad" reads does not indicate that the counts for the top sequences would be altered drastically if a revision to the Perl script were made to allow for mismatches or insertions/deletions in the head and tail. The total number of clusters, the number of clusters that pass the instrument's filter (PF) and the total number of reads identified in the data files are listed in **Table 5.7**. The total number of reads per index and the quantity of good reads are also summarized in **Table 5.7**. The counts of TBA expected (or any unique aptamer sequence) were recalculated based on these numbers and are compared to the counts of TBA observed.

As indicated in **Table 5.7**, approximately 16 million original clusters were formed and approximately 50% of the clusters did not pass filter. Of the clusters that passed filter, the data was parsed by Illumina's BaseSpace software into individual data files based on the Index. During this process the software eliminates any sequence data that cannot be read properly (does not contain the correct Index information), reducing the percentage of quality clusters further. A total of approximately 41% of the original clusters were converted into readable data. Using this

information and the specific number of reads identified per index, the number of expected TBA counts (or of any unique aptamer sequence) was recalculated.

The observed counts of TBA for samples 2 and 3 are similar to what is expected. TBA was counted 4 times in sample 2 and 24 times in sample 3, with expected counts of ~3.8 and ~2.6, respectively. Although TBA was not identified in sample 1, it cannot be concluded whether or not this experiment failed due to the low number of expected counts of TBA. The expected counts of TBA assumes a 100% recovery during partitioning, ligation and gel purification. The effect of reintroducing the negative selection step in samples 2 and 3 cannot be accurately determined without duplicate experiments. Additional experiments are required to determine whether or not the negative selection step is beneficial. However, this method of AIA utilizing the tailed library and ligation of full length adapters was successful at identifying TBA from a starting pool. The two negative controls (no thrombin) did not offer any conclusive data. The sample was only expected to produce ~1.5 copies of any unique sequence. Any aptamer candidates would likely not be observed for this sample. The identification of TBA above background in combination with a decrease in the quantity of recovered library indicates that the procedural modifications improved selection stringency.

Consistent with the hairpin loop library and adapter library data, the jump sequence is present in the data set. In the context of the tailed library, the sequence is dCACAGATCGGAAGAG, which is the result of skipping over the first 19 nucleotides of the library molecule (dACACNNNNNNNNNNNNNNN). The CACA tail and the first 11 nucleotides of Adapter 2

compose the jump sequence. This sequence is captured by the Perl script because the next four nucleotides of Adapter 2 are CACA, qualifying the jump sequence as a good read. It was originally hypothesized that the jump sequence was a systematic artifact resulting from PCR amplification. However, the jump sequence was found regardless of whether the various AIA libraries were prepared with or without ligation or PCR amplification. This suggests that the jump sequence is a product of bridge amplification during clustering. Although the data quality for this sample set was moderately good, it provided proof of concept for the tailed libraries in AIA, illustrating that the thrombin binding aptamer could be successfully identified from an over-represented pool of the m=15 DNA tailed library.

Although bead-based partitioning was successfully used in this chapter, Con-A beads have been historically unpredictable and often require trial and error to produce the desired result. The frequency of TBA with respect to good reads was 0.00036% and 0.0077% for samples 2 and 3. This is much lower than the frequencies observed in successfully AIA partitioning with hairpin loop libraries described in chapters 2 and 3. The original 60:1 AIA experiment identified TBA at a frequency of 2.3% and the improved AIA method identified TBA at frequencies of 0.062% and 0.18%. For targets with unknown aptamer sequences, a more reliable and reproducible partitioning method is highly desirable. With the ability to predict counts for high affinity sequences as described in this chapter, determining whether an AIA experiment failed due to the partitioning method or because there were no aptamer candidates in the pool is critical. It is easy to determine if a thrombin selection failed due to the partitioning method (assuming the library structure has been successful previously) because the thrombin binding aptamer sequence is

known. The next chapter discusses reversible formaldehyde crosslinking in combination with EMSA with the intent of developing a consistent and more universal partitioning method.

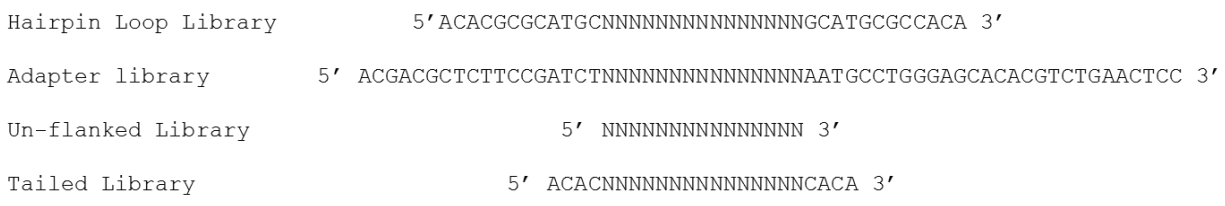


Figure 5.1 Size comparison of library designs

Visual size comparison of m=15 hairpin loop library, m=15 adapter library, m=15 un-flanked library and m=15 tailed library.

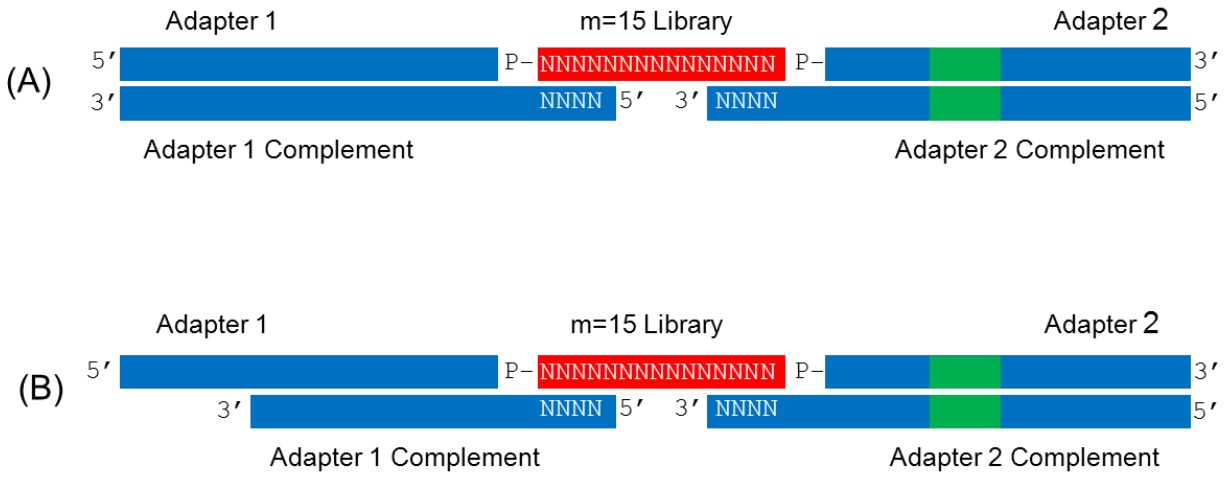


Figure 5.2 Randomized splints and T4 DNA Ligase capture

Sequences listed in **Appendix 5**. Figure not to scale. **(A)** Ligation scheme for un-flanked libraries using double-stranded Adapter/Adapter Complements with randomized splints. The randomized splint sequences, N_n, tested for annealing varies from n=1 to 4. Illumina Index region shown in green. **(B)** Ligation scheme for un-flanked libraries with shortened Adapter 1 Complement to prevent annealing to the Illumina flow cell oligo lawn.

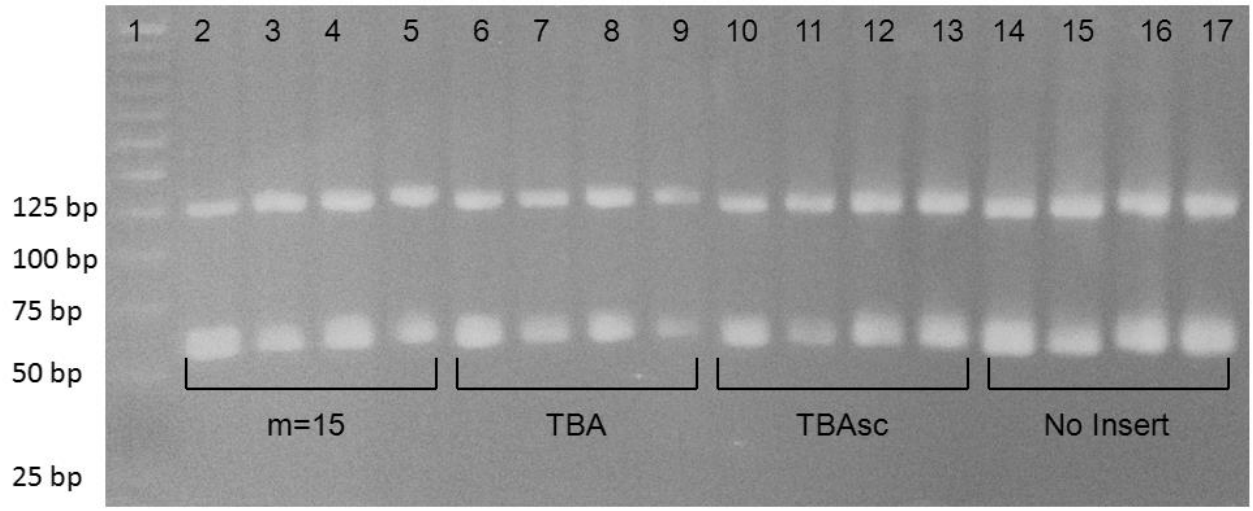


Figure 5.3 Randomized splints and T4 DNA Ligase capture of un-flanked m=15, TBA and TBAsc

4% TAE agarose gel. Lane 1: 25 bp DNA ladder. Ligation of full length adapters to un-flanked m=15 library, TBA, TBAsc, and no insert control. Each set of ligation reactions in brackets includes adapters with n=1, 2, 3, or 4, respectively. Excess adapters run between 50 and 75 bp, ligated product at approximately 125 bp is present in all cases including the no insert control and represents the self-ligation of Adapter 2/Adapter 2 Complement.

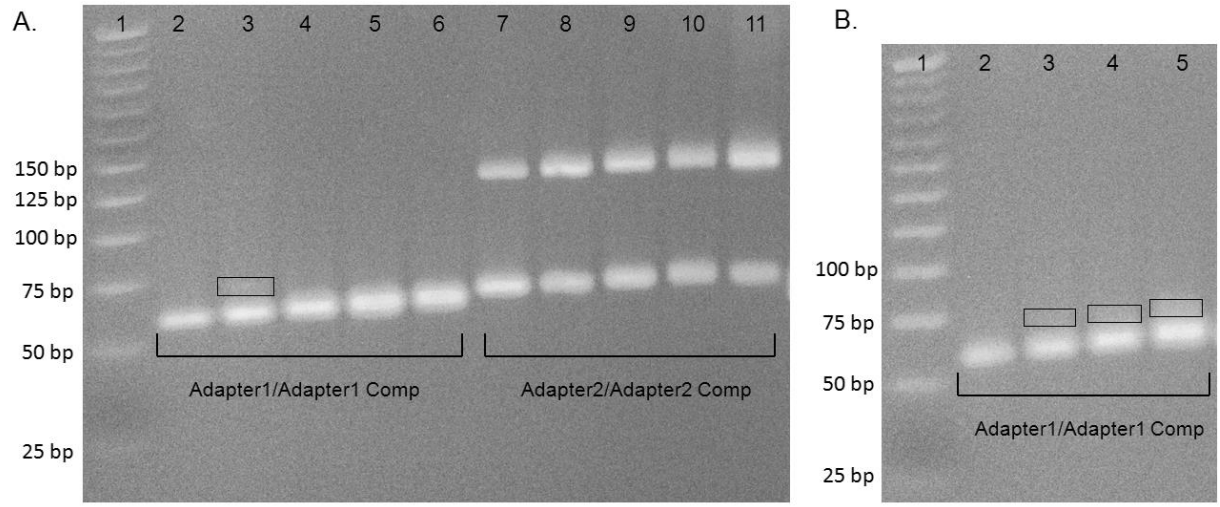


Figure 5.4 Independent ligation of Adapter 1 and Adapter 2

4% TAE agarose. **(A)** Lane 1: 25 bp DNA ladder. Ligation of full length Adapter1/Adapter1 Complement and Adapter2/Adapter2 Complement to un-flanked m=15 library independently of each other. Each set of ligation reactions in brackets includes adapters with n=1, 2, 3, or 4 and n=4 with no insert, respectively. Correctly ligated product outlined in lane 3. Excess Adapters between 50 and 75 bp. Product at approximately 125 bp is self-ligation of Adapter 2/ Adapter 2 Complement. **(B)** Lane 1: 25 bp DNA ladder. Repeat of Adapter1/Adapter1 Complement ligation with n=1, 2, 3 or 4 respectively. Following heat denaturation at 95°C for 3 minutes, samples were cooled in a thermocycler at a rate of 0.5°C per second rather than simply cooling on the bench top at room temperature. Correctly ligated product outlined in lanes 3, 4, and 5 for n=2, 3 and 4, respectively.

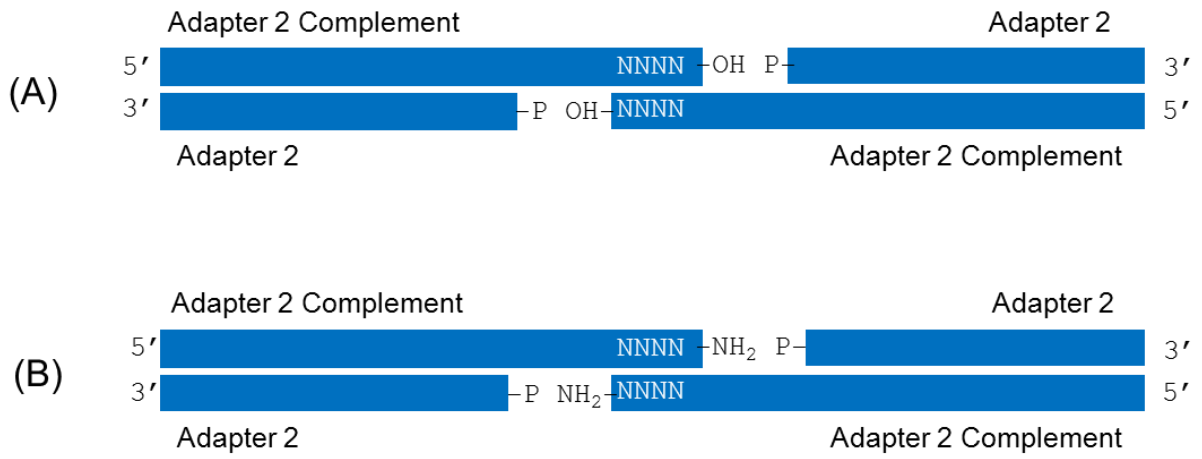


Figure 5.5 Self-ligation of Adapter 2/ Adapter 2 Complement

Sequences listed in **Appendix 5**. Figure not to scale. **(A)** Illustration of how the Adapter2/Adapter 2 Complement complex is capable of self-ligation. Randomized splints anneal, placing the 5'-phosphate of Adapter 2 adjacent to the 3'-hydroxyl of Adapter 2 Complement, facilitating self-ligation of the complex. **(B)** 3'-terminal amine used to prevent self-ligation.

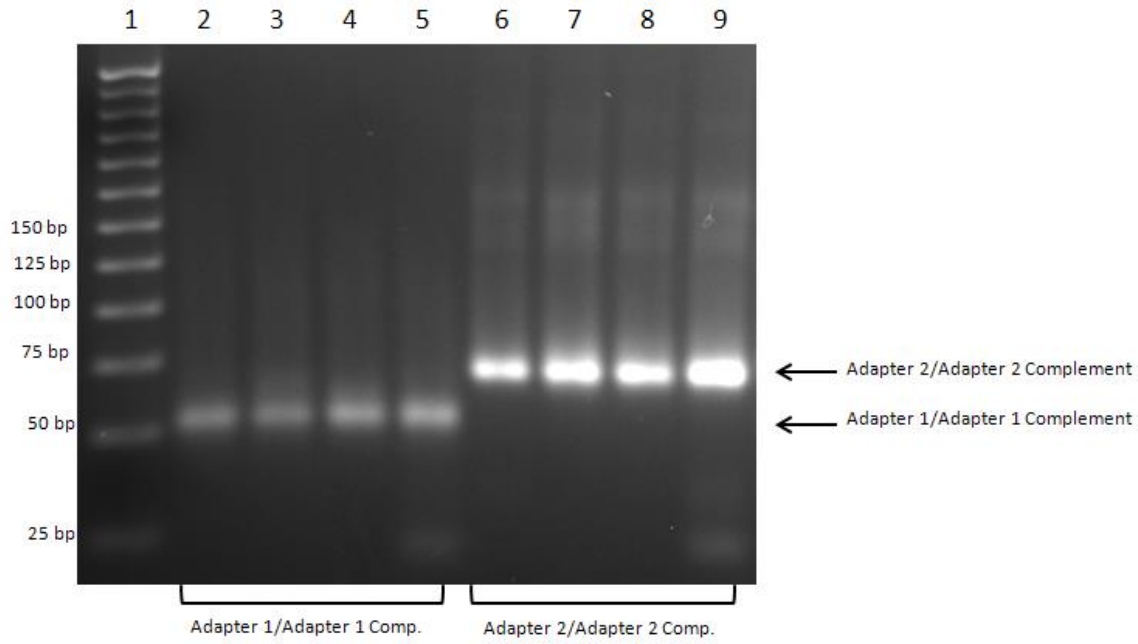


Figure 5.6 3'-Amino modifier prevents self- ligation of Adapter 2/ Adapter 2 Complement complex

4% TAE agarose**.** Lane 1: 25 bp DNA ladder. Adapter 1 Complement and Adapter 2 Complement (3'amine) have an n=4 splint. Lanes 2-5 show ligation results of Adapter 1/Adapter 1 Complement in a no insert control, to m=15 un-flanked library, to the 15-mer TBA and to TBA in the 39 base hairpin loop stem format, respectively. Lanes 6-9 show ligation results of Adapter 2/Adapter 2 Complement (3'-amine) in a no insert control, to m=15 un-flanked library, to the 15-mer TBA and to TBA in the 39 base hairpin loop format, respectively. The desired product is slightly visible in lanes 3 and 4, while un-ligated insert (TBA in hairpin loop) is clearly visible in lanes 5 and 9 at approximately 25 bp.

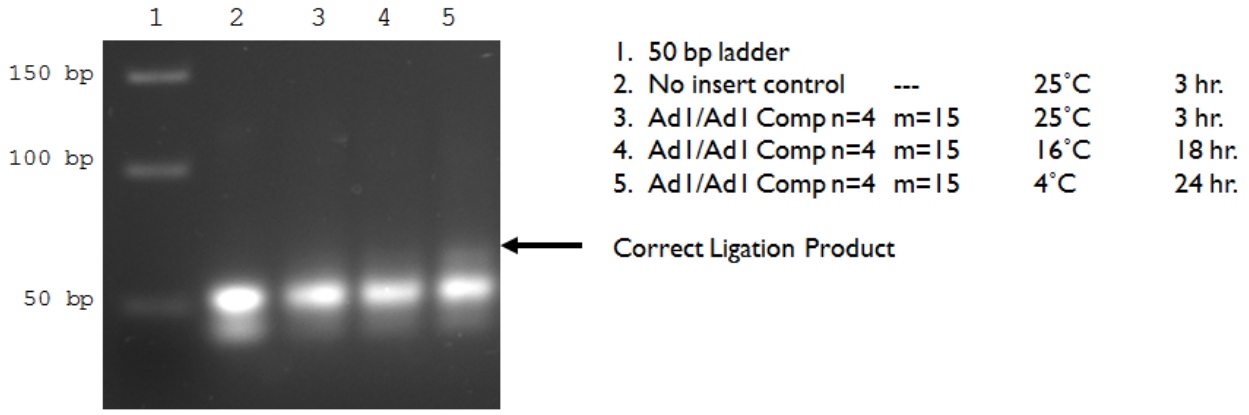


Figure 5.7 Variable ligation conditions

4% TAE agarose. Lane 1: 50 bp DNA ladder. Ligation of Adapter 1/Adapter 1 Complement n=4 to 10 pmoles un-flanked m=15 DNA library under various conditions. Ratio of adapters to library is 1:1. Arrow indicates correct product. No product is visible in the no template control. The ligated product is visible in lanes 3, 4 and 5. Intensity of the ligated product increases with decreased temperature and increased time.

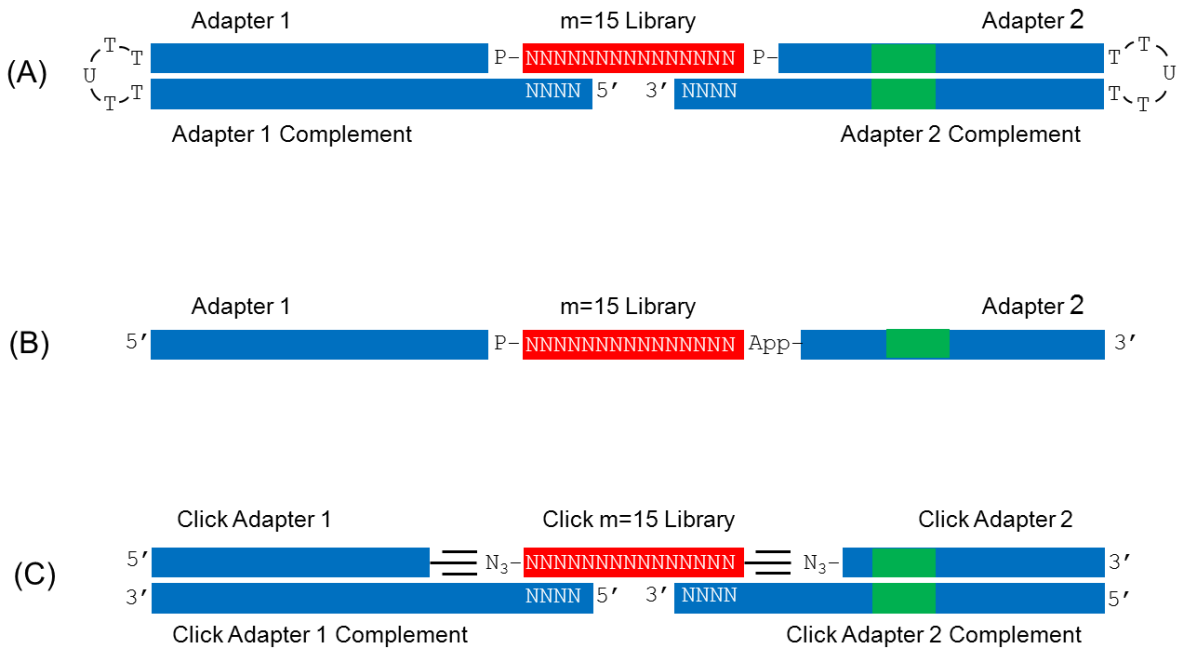


Figure 5.8 Various ligation methods for un-flanked libraries

Sequences listed in Ap**pendix 5**. Figure not to scale**. (A)** Schematic representation of the hairpin adapter format. Each adapter is a single strand of DNA, with complementary regions joined by a 5-base loop to impose the desired hairpin structure necessary to ligate the m=15 library. The sequence of the 5-base loop is dT-dT-rU-dT-dT. The single RNA base is cleaved via NaOH hydrolysis to release the adapter complement strands prior to sequencing. **(B)** Ligation Scheme of Adapter 1 and 5'App Adapter 2 to the un-flanked m=15 DNA library with two forms of RNA ligase. **(C)** Click ligation scheme. Click Adapter 1 with 3'-terminal alkyne. Click m=15 library with 5'-terminal azide and 3'-terminal alkyne. Click Adapter 2 with 5'-terminal azide.

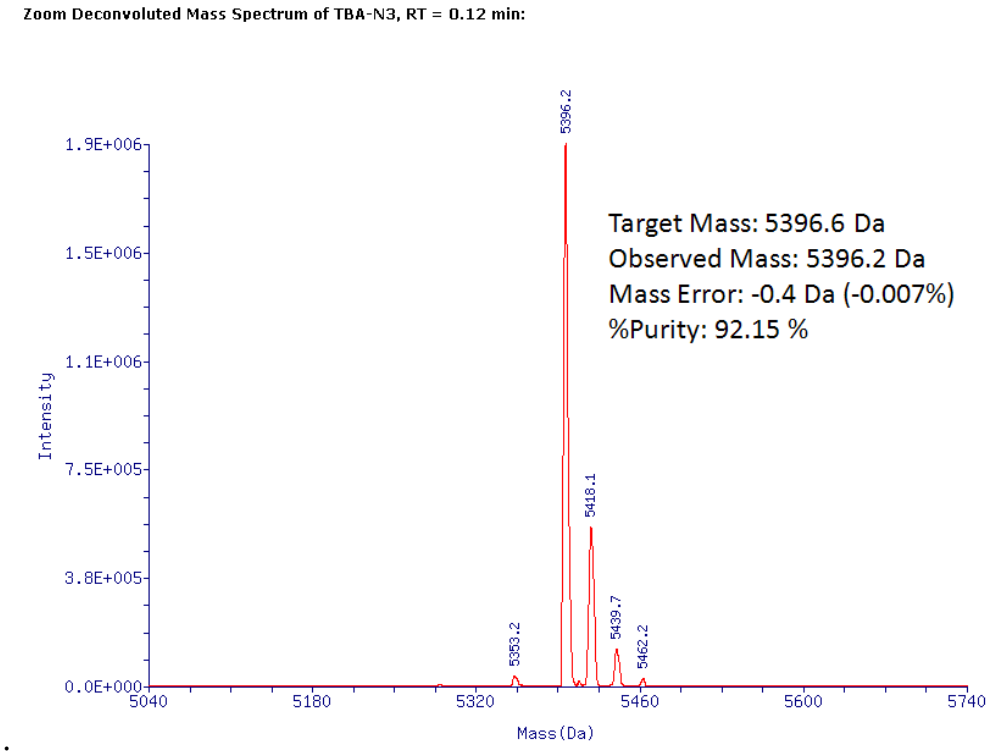


Figure 5.9 ESI/LC Mass spectrum of Click TBA

ESI/LC Mass spectrum of TBA with 5'-terminal azide and 3'-terminal alkyne. The observed mass is 5396.2 Da, a difference of only 0.4 Da form the target mass, 5396.9 Da indicating that the Click TBA was synthesized with the correct 5' and 3'-modifiers.

Figure 5.10 Ligation Scheme for tailed m=15 libraries with Fixed versus Fixed/n=4 adapter splints

(Top) Sequences of m=15 DNA tailed library. **(Middle)** Sequences of Adapter 1, Adapter 1 Complement Fixed, Adapter 2 (Index 1) and Adapter 2 Complement Fixed (Index 1), synthesized by IDT. Abbreviated structure of the ligated product shows the library in green. Complementary overhangs of the adapters in bold**. (Bottom)** Sequences of Adapter 1, Adapter 1 Complement Fixed/n=4, Adapter 2 (Index 1) and Adapter 2 Complement Fixed/n=4, (Index 1), synthesized by IDT. Abbreviated structure of the ligated product shows the library in green. Complementary overhangs of the adapters in bold.

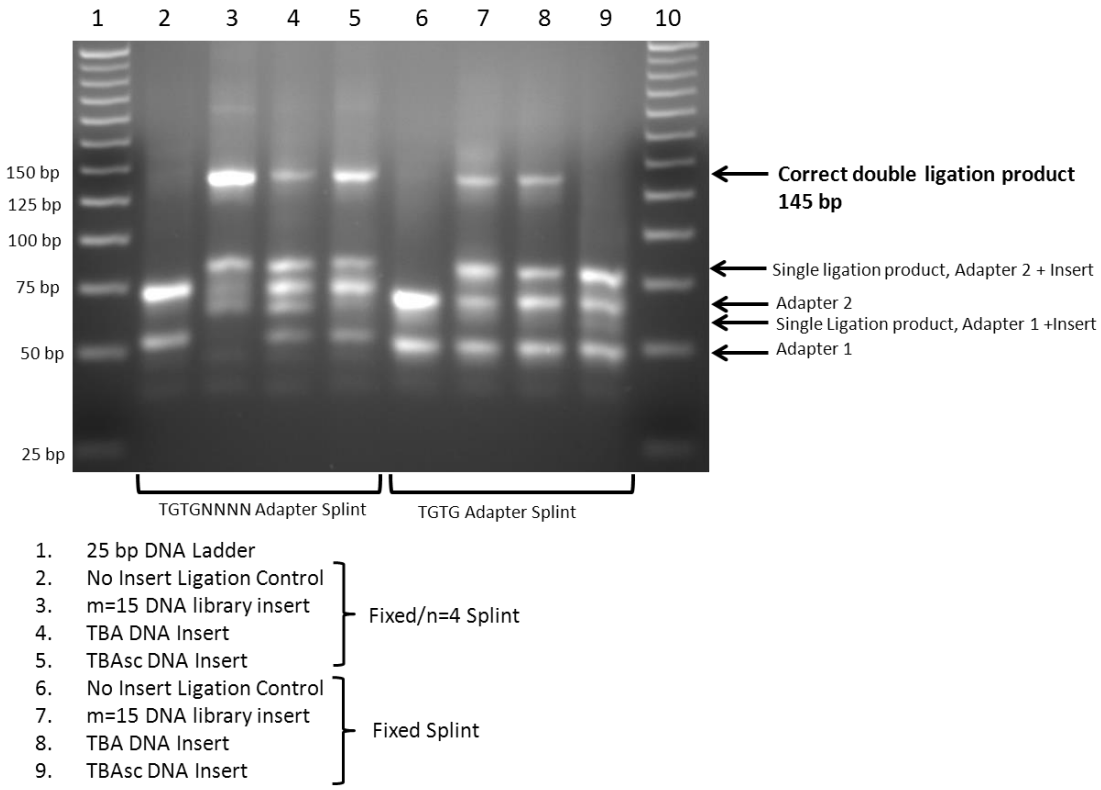
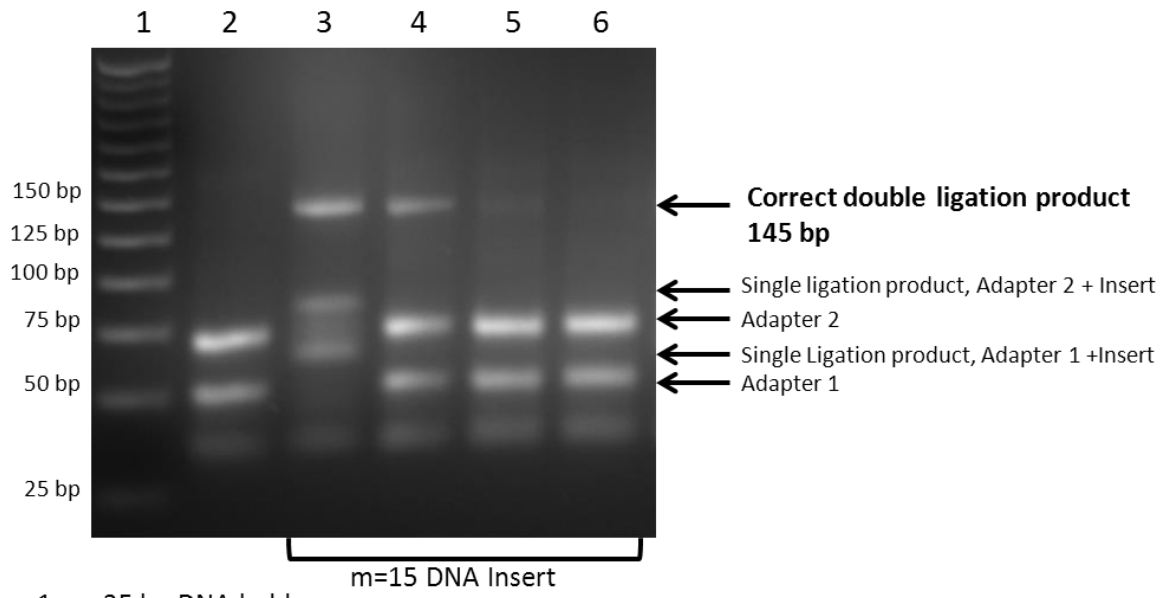


Figure 5.11 Ligation comparison of Fixed versus Fixed/n=4 adapter splints

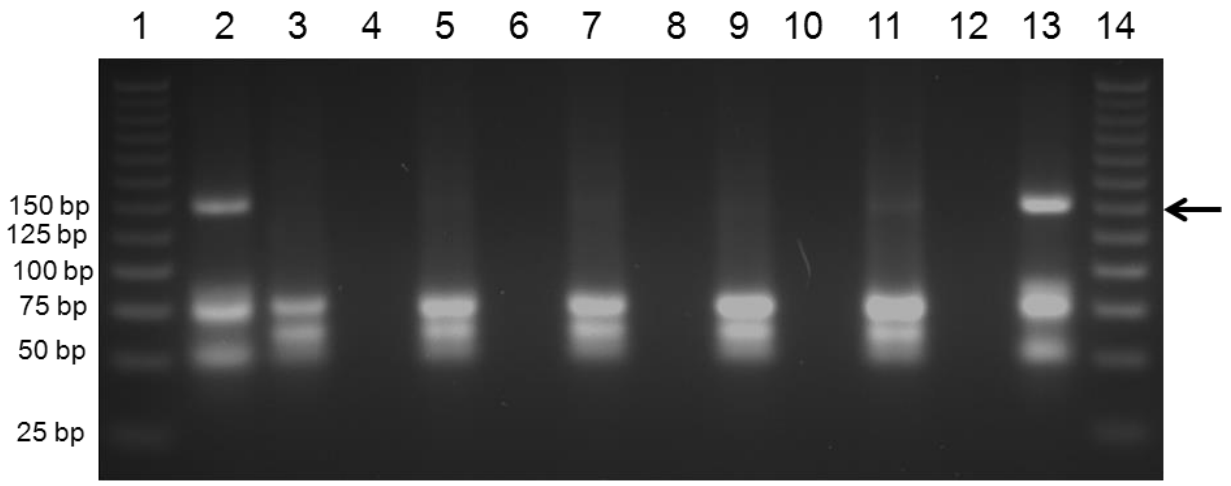
4% TAE agarose gel. Comparison of the "Fixed" and "Fixed/n=4" adapter complements. Ligation of the no stem flanked m=15 DNA library as well as two fixed control sequences: TBA (dACACGGTTGGTGTGGTTGGCACA) and TBAsc (dACACGGTGGTTGTTGTGGTCACA.) The no insert control demonstrates a lack of non-specific products; Adapter 1 appears less intense than Adapter 2 due to the partially single stranded fragment. All adapters ligate the insert independently, resulting in the single-ligation products identified by appropriate arrows. The correct, double-ligation product is of greater intensity for all inserts with the Fixed/n=4 splint.



1. 25 bp DNA ladder
2. No Insert Ligation Control
3. 10 pmol m=15 DNA library
4. 1.0 pmol m15 DNA library
5. 0.1 pmol m=15 DNA library
6. 0.01 pmol m=15 DNA library

Figure 5.12 Dilution series ligation with Fixed/n=4 adapter splints

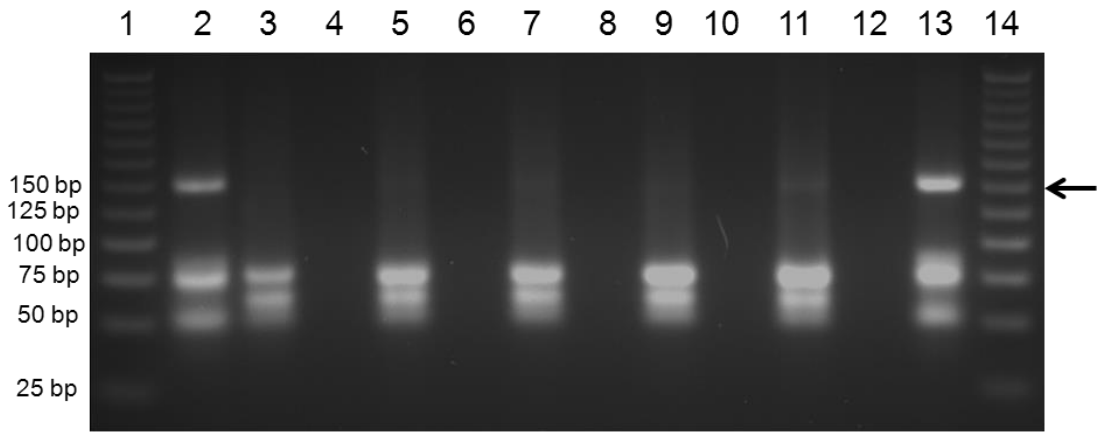
4% TAE agarose gel. Comparison of visual intensity of ligated products of a dilution series of m=15 DNA library. As the concentration of insert decreases, the presence of single-ligation products diminishes prior to the correct double-ligation product. This indicates that a partitioned library with concentration too low to be visualized would be ligated correctly.



1. 25 bp ladder
2. Size Marker, Positive Control
3. No Insert, Negative Control
5. Sample 1
7. Sample 2
9. Sample 3
11. Sample 4
13. Size Marker, Positive Control
14. 25 bp ladder

Figure 5.13. Ligation of 1:5, thrombin: m=15 DNA tailed library

4% TAE agarose gel purification of ligated samples after partitioning. Lanes 1 and 8: 25 bp DNA ladder. Experimental samples were loaded in alternating wells to reduce cross contamination. The correct 144 base-pair products are indicated by the arrows. The positive controls show the correct product as well as single ligation products. The negative controls show no ligation products. Sample replicates 1, 2, 3 and 4 show the correctly ligated product.



1. 25 bp ladder
2. Size Marker, Positive Control
3. No Insert, Negative Control
5. Sample 1
7. Sample 2
9. Sample 3
11. Sample 4
13. Size Marker, Positive Control
14. 25 bp ladder

Figure 5.14. Ligation of partitioning results aimed at improving stringency

4% TAE agarose gel purification of ligated partitioned samples. Lanes 1 and 14: 25 bp DNA ladder. Samples are loaded in alternating wells to reduce cross contamination. The correct 144 base-pair product is indicated by the arrow. The positive controls show the correct product as well as single ligation products. The negative control shows no ligation products. The ligated product is slightly visible for sample 4 only.

Table 5.1 Recommended multiplexing combinations

TruSeq Index	Index Sequence	Pool 2 samples	Pool 3 samples	Pool 6 samples	Pool 12 samples
1	ATCATG				
2	CGATGT				
3	TTAGGC				
4	TGACCA				
5	ACAGTG				
6	GCCAAT				
7	CAGATC				
8	ACTTGA				
9	GATCAG				
10	TAGCTT				
11	GGCTAC				
12	CTTGTA				

The first 12 TruSeq Indices. A total of 27 indices exist for the Illumina TruSeq Sample prep kits. Shaded boxes indicate which indices to select when pooling the indicated number of samples. Custom AIA adapters with indices 2, 4, 6 and 12 were purchased from IDT with the intent of multiplexing four or fewer samples per lane.

Table 5.2 Sample calculations for anticipated sequence frequency

Sample	pmoles of Library Recovered	pmoles of Library Available	pmoles of Library Used	Percentage of Total	Fraction of Original 56,000 TBA Copies Used
1	0.05	0.045	0.01	20%	11,200
2	0.04	0.036	0.01	25%	14,000
3	0.03	0.027	0.01	30%	16,800
4	0.02	0.018	0.01	50%	28,000

Sample	Number of Molecules loaded onto flow cell	Expected Number of Clusters	Clusters as a Percentage of Molecules	Number of Surviving TBA molecules
1	6,020,000,000	5,000,000	0.083 %	9.30
2	6,020,000,000	5,000,000	0.083 %	11.62
3	6,020,000,000	5,000,000	0.083 %	13.95
4	6,020,000,000	5,000,000	0.083 %	23.25

Sample calculations for anticipated sequence frequency assuming 100% recovery of the starting counts and 5,000,000 clusters per sample. “Pmoles of library recovered” is determined by qPCR. “pmoles of library available” represents the remaining quantity of sample following qPCR analysis. “pmoles of library used” totals 0.04 pmoles (requirement of clustering on the MiSeq). “Percentage of total is pmoles of library used/pmoles of library recovered*100. The “fraction of original 56,000 TBA copies used” represents the number of copies in 0.01 pmoles as a fraction of the total amount recovered. “Expected number of clusters” is based on an ideal 20 million clusters divided by four samples. The “clusters as a percentage of molecules” is the number of clusters/ the number of molecules loaded on the flow cell. The “number of surviving TBA molecules” is calculated as 0.083% of the number of TBA copies used.

Table 5.3 Calculations for anticipated sequence frequency, 1:5 replicate partitioning experiments

Sample	pmoles of Library Recovered	pmoles of Library Available	pmoles of Library Used	Percentage of Total	Fraction of Original 56,000 TBA Copies Used
1	0.040	0.036	0.01	25%	14,000
2	0.150	0.135	0.01	7%	3,733
3	0.015	0.0135	0.01	67%	37,333
4	0.065	0.0585	0.01	15%	8,615

Sample	Number of Molecules loaded onto flow cell	Expected Number of Clusters	Clusters as a Percentage of Molecules	Number of Surviving TBA molecules
1	6,020,000,000	5,000,000	0.083 %	11.62
2	6,020,000,000	5,000,000	0.083 %	3.10
3	6,020,000,000	5,000,000	0.083 %	31.00
4	6,020,000,000	5,000,000	0.083 %	7.15

Calculations for anticipated sequence frequency assuming 100% recovery of the starting counts and 5,000,000 clusters per sample. “Pmoles of library recovered” is determined by qPCR. “pmoles of library available” represents the remaining quantity of sample following qPCR analysis. “pmoles of library used” totals 0.04 pmoles (requirement of clustering on the MiSeq). “Percentage of total is pmoles of library used/pmoles of library recovered*100. The “fraction of original 56,000 TBA copies used” represents the number of copies in 0.01 pmoles as a fraction of the total amount recovered. “Expected number of clusters” is based on an ideal 20 million clusters divided by four samples. The “clusters as a percentage of molecules” is the number of clusters/ the number of molecules loaded on the flow cell. The “number of surviving TBA molecules” is calculated as 0.083% of the number of TBA copies used.

Table 5.4 Sample information for partitioning experiments aimed at improving stringency

Sample	Ratio	Protein	Library	Index	Negative Selection
1	1:5	20 pmol	100 pmol	2	No
2	1:5	20 pmol	100 pmol	4	Yes
3	60:1	6 nmol	100 pmol	6	Yes
4	N/A	20 pmol	100 pmol	12	No

Quantity of m=15 DNA library and protein used in partitioning as well as corresponding Illumina Indices. Samples 1 and 2 differ by the reintroduction of the negative selection step. Samples 2 and 3 differ by the ratio of target: library. Sample 4 serves as a negative control with the absence of protein.

Table 5.5 Calculations for anticipated sequence frequency, partitioning experiments aimed at improving stringency

Sample	pmoles of Library Recovered	pmoles of Library Available	pmoles of Library Used	Percentage of Total	Fraction of Original 56,000 TBA Copies Used
1	0.0234	0.02106	0.0103	44%	24,650
2	0.0279	0.02511	0.0103	37%	20,674
3	0.0101	0.00909	0.0091	90%	50,455
4	0.1122	0.10098	0.0103	9%	5,141

Sample	Number of Molecules loaded onto flow cell	Expected Number of Clusters	Clusters as a Percentage of Molecules	Number of Surviving TBA molecules
1	6,202,000,000	5,000,000	0.083 %	20.47
2	6,202,000,000	5,000,000	0.083 %	17.17
3	5,472,000,000	5,000,000	0.0913 %	51.13
4	6,202,000,000	5,000,000	0.083 %	4.26

Calculations for anticipated sequence frequency assuming 100% recovery of the starting counts and 5,000,000 clusters per sample. “Pmoles of library recovered” is determined by qPCR. “pmoles of library available” represents the remaining quantity of sample following qPCR analysis. “pmoles of library used” totals 0.04 pmoles (requirement of clustering on the MiSeq). “Percentage of total is pmoles of library used/pmoles of library recovered*100. The “fraction of original 56,000 TBA copies used” represents the number of copies in 0.01 pmoles as a fraction of the total amount recovered. “Expected number of clusters” is based on an ideal 20 million clusters divided by four samples. The “clusters as a percentage of molecules” is the number of clusters/ the number of molecules loaded on the flow cell. The “number of surviving TBA molecules” is calculated as 0.083% of the number of TBA copies used.

Table 5.6. Statistical sequencing data for partitioning experiments aimed at improving stringency

Run	Total Number Clusters	Number of Clusters PF	Total Reads in Data Files
Samples 1-4	16,317,725	8,381,792	6,777,855

Experiment	1:5	1:5, Neg. Sel.	60:1, Neg. Sel.	Neg. control
Quantity of Protein	20 pmol	20 pmol	20 pmol	N/A
Quantity of Library	100 pmol	100 pmol	100 pmol	100 pmol
Starting Copies per Sequence	56,000	56,000	56,000	56,000
Total Reads	1,611,487	1,516,747	1,090,909	2,265,735
Total Good Reads	1,143,847	1,116,372	313,767	1,776,285
Background/Noise Threshold	3	3	3	3
Sequences Above Background	1	2	2	2
Expected Counts of TBA	4.7	3.8	2.6	0 (1.5)[d]
Observed Counts of TBA[a]	N/A	2_4	2_24	N/A
Frequency of TBA per Good Reads[b]	N/A	0.00036 %	0.0077 %	N/A
Frequency of TBA per Total Reads[c]	N/A	0.00026 %	0.0022 %	N/A

(Top) Total number of clusters produced per sequencing run. The number of clusters that passed filter (PF) is approximately 51%. The total reads are approximately 43% of the original clusters. **(Bottom)** Total number of reads for each sample generated by Illumina's BaseSpace. A total of 292,720 reads had undetermined Illumina Indices and 254 reads contained Index 1 for the positive control size marker. The number of good reads is generated by the Perl Script. The expected number of TBA counts was calculated based on the quantity of library recovered, the quantity of library applied to the flow cell and the number of good reads assuming a 100% recovery of the original 56,000 copies of TBA. **(a)** The rank *(left)* and count *(right)* are separated by an underscore. **(b)** Counts of TBA divided by the total good reads. **(c)** Counts of TBA

divided by the total reads. **(d)** TBA is not expected for the no thrombin control, however, any unique sequence was expected at a maximum of 1.5 counts.

Table 5.7 Top 10 Sequences for partitioning experiments aimed at improving stringency

Sample 1: 1:5		
Sequence	**Count**	**Identifier**
CACAGATCGGAAGAG	51	Jump*
CAGAGACAGGTATGT	3	
TACGTATGCGACCGT	3	
CACATCTGCGGCACG	3	
CGGCTATGGCGGCGC	3	
GAGAAGAGAGTATTC	3	
CCGATTCCGAGCAGA	3	
ATGGACAACAGTCAA	3	
GAGAGATACGACATC	3	
CGTCAGACCTCGGCC	3	

Sample 2: 1:5, Negative Selection		
Sequence	**Count**	**Identifier**
CACAGATCGGAAGAG	47	Jump*
GGTTGGTGTGGTTGG	4	TBA
CAGAAAGAAGGTCGA	3	
TAGAATTGGCATCGT	3	
CTTACTGTTTGAAAT	3	
GAGATACCCAGCAGG	3	
TAGTAACCAAACAAT	3	
CGCTTGCGGGCAAAG	3	
CCAGAGCTAAACAAT	3	
CATTAGCTTATCCGG	3	

Sample 3: 60:1, Negative Selection		
Sequence	**Count**	**Identifier**
CACAGATCGGAAGAG	29	Jump*
GGTTGGTGTGGTTGG	24	TBA
TTATAGGCATTATAT	3	
CTAGAATGACAGTCT	3	
AACGATGGTATGCTT	3	
TGCCTGAGTTAGTCC	3	
GTCGTATAGCCGCAG	3	
ATGGTCGTAAGGTGT	3	
GAATTCCGACCGGCC	3	
GAATCGCGACCTTCG	3	

Sample 4: No thrombin Control		
Sequence	**Count**	**Identifier**
CACAGATCGGAAGAG	43	Jump*
AAAGACCGGGTGGAA	4	
CCAAGACGACATTGT	3	
CCCCATGGATGTCAT	3	
GCGTCGCGTCTGCAC	3	
GAGGATACACTCGTG	3	
ACTGGGTGAGCGCAC	3	
GCAGGGGTTAGAATG	3	
AGGTGCCACTCCAGT	3	
AATCCCCGTCAGCAA	3	

Top 10 sequences and corresponding counts for Samples 1-4. The thrombin binding aptamer (TBA) was identified at counts of 4 and 24 for samples 2 and 4. The jump sequence is present in all four sample as the highest frequency sequence. The background for all four experiments is 3. Excluding the jump sequence, the only sequence above background for samples 1-3 is TBA. One other sequence occurs above background in the no thrombin control.

*Jump sequence variants.

Chapter 6: Reversible Formaldehyde Cross-linking and EMSA

Chapter Summary

To expand the applicability of AIA using the tailed libraries, a partitioning method utilizing reversible formaldehyde cross-linking in conjunction with Electrophoretic Mobility Shift Assays (EMSA) was investigated. Formaldehyde (HCHO) produces DNA-protein and protein-protein cross-links when the molecules come in close contact with each other. DNA-protein cross-links (DPCs) can be formed when the amino groups of cytosine, guanine or adenine react with HCHO to form a Schiff base, which reacts with the amino groups of specific amino acid side chains (e.g. Lys, Arg). [172-175] DNA-protein cross-links can also be formed when HCHO reacts with the amino acid side chains initially as well. Imino groups, including that of thymine, are also capable of forming cross-links.[174] **Figure 6.1** illustrates the mechanism of cross-linking between two amino groups. These covalent cross-links are stable at room temperature but are reversible at increased temperature. At 50°C, slow degradation of the cross-links occurs and at 95°C, the cross-links have been shown to completely degrade within 2 minutes.[173] Formaldehyde cross-linking has been used extensively to explore the association of chromosomal DNA or RNA with nucleosomes, transcription factors and protein with Chromatin immunoprecipitation (ChIP) assays.[172, 175, 176] Randomly fragmented DNA or RNA is equilibrated with protein in solution and protein bound oligos are cross-linked with formaldehyde. The protein is immunoprecipitated with antibodies, cross-links are reversed by heating, and the selected DNA or RNA molecules are sequenced. This method was adapted to AIA with the incorporation of EMSA following cross-linking.

EMSA is a standard method for characterizing the affinity of nucleic acids to a protein target and has been used as a partitioning method in SELEX.[94] For a library of DNA, high affinity sequences remain bound to the protein while unbound sequences travel freely through the gel and moderately bound sequences may present as a smear. The use of SDS-PAGE or a denaturing agent in the sample buffer destroys any non-covalent interactions between the protein and library. EMSA with denaturing conditions would allow HCHO cross-linked complexes to be isolated from the remaining pool of free library. Additionally, HCHO cross-linking would facilitate capture of sequences with moderate affinity that would otherwise dissociate during EMSA alone. Using HCHO cross-linking and EMSA (or gel filtration chromatography, GFC) may eliminate the need for protein immobilization for a wide variety of targets and could greatly simplify aptamer selection. Immobilization can be problematic, especially for post-translationally modified proteins and peptides isolated from serum or tissues. Some other proteins may take considerable effort to isolate with terminal His-tags or other affinity tags, and immobilization schemes that involve amino groups on proteins may cause conformational changes. Other important applications that could benefit from aptamer selection via HCHO cross-linking without immobilization could involve complex mixtures of proteins (e.g. cellular lysates) and small molecule or other non-protein targets with amino labels.

This chapter details the development of a reversible HCHO cross-linking and EMSA protocol that offers a fast, efficient and high throughput method for aptamer discovery. Additionally, the library capture and sequencing methods described in chapter 5 were expanded further with the introduction of qPCR CopyCount^{TM} software as a means of quantifying samples prior to sequencing. The accurate quantification of a library is crucial for maximization of data output on

Illumina sequencing platforms. qPCR CopyCount™ ultimately reduces costs while improving data output compared to the gold-standard KAPA Quantification kit.

Tracking cross-linked DNA with fluorescent libraries

A variety of conditions and experimental designs were explored during the development of the HCHO cross-linking protocol. The first concern was facilitating visualization of the cross-linked library during EMSA. As illustrated in chapter 5, a requirement of amplification-free AIA is the recovery of low quantities of library that have been partitioned with high stringency. To this end, the desired quantity of cross-linked library is small, on the order of 0.01 pmoles or less (dependent on the level of over-representation of the library). In order to visualize the shifting of such small quantities of library, a fluorescent label was introduced into the library structure. Pyrrolo-dC is a fluorescent nucleotide capable of base pairing with dG similarly to dC and is incorporated by DNA and RNA polymerases. The small structure of Pyrrolo-C does not interfere with the DNA helix formation, making it suitable for use in AIA.[177] Excitation of Pyrrolo-C occurs at 347 nm with emission at 470 nm. The fluorescence is visible under UV light, making it an ideal marker for EMSA. Tailed libraries and fixed sequences (m=15, TBA and TBAsc) were synthesized on the ABI 394 with Pyrrolo-dC CE phosphoramidite (Glen Research) substituted for a single dC (see **Appendix 6)**. The sequences were purified via DMT purification using Glen-Pak DNA Purification Cartridges according to the manufacturer's instructions and purified by IE-HPLC. The TBA-PyC and TBAsc-PyC molecules were correctly identified by ESI/LC Mass Spectrometry. Based on the imaging results of preliminary HCHO cross-linking experiments and a dilution series of the oligos, it was determined that the fluorescence intensity of TBA-PyC and TBAsc-PyC was too low to be visualized for partitioning experiments. In the

absence of any staining methods, the oligonucleotides were not visible at quantities below 10 pmol. A library with a total of four Pyrrolo-C substitutions was also synthesized; however, the increase in fluorescence intensity was minimal. This disqualified the PyC oligos as a method to track cross-linked versus free library in EMSA. In the absence of a fluorescent oligonucleotide, locating the cross-linked DNA-protein complexes too dilute to be visualized was performed with the use of a positive control size marker in adjacent lanes.

Preparative gel-shift conditions

Reversible HCHO cross-linking and preparative gel-shift conditions were optimized using traditional staining methods, including ethidium bromide and SYBR gold. Although EMSA is typically performed with native polyacrylamide gels, several combinations of polyacrylamide and agarose gels in both native and denaturing conditions were tested. Agarose gels offer many advantages, including ease of preparation, shorter running times, and simpler extraction methods. However, polyacrylamide gels can provide greater resolution for shorter sequences and may have advantages for some systems. Preparative gel shift conditions were optimized by cross-linking a TBA molecule flanked by ACAC/CACA tails (ordered from IDT) to thrombin. In order to isolate the covalently cross-linked DNA-protein complexes from non-covalently linked complexes, a denaturing agent is required. To illustrate the shifting differences between the two types of complexes, native PAGE and native agarose were employed. The denaturing agent (SDS) was added to the sample loading buffer if isolating the cross-linked complexes was desired. Native tris-glycine PAGE with SDS in the sample buffer was not a suitable option; free DNA traveled with the salt front and the denatured protein smeared. Native TAE agarose with SDS in the sample buffer was also not an ideal option. Due to the larger pore size of agarose

gels, both the full length protein and any degraded protein fragments traveled through the gel at the same rate, resulting in a single band. Similarly, any protein cross-linked with DNA traveled in that same band. Both horizontal and vertical agarose gels were tested. Vertical agarose gels are thin (1.5 mm) and stain much quicker than horizontal slab agarose gels (~0.7 mm-10.0 mm). The reduced thickness in combination with limited diffusion during staining resulted in a sharper image for vertical agarose gels. However, due to the nature of the well shape, smearing of free DNA occurred in vertical gels, which may lead to contamination while extracting cross-linked products. Increased glycerol concentration and loading the sample while applying current did not resolve the issue. SDS-agarose gels (SDS in gel buffer and/or running buffer) in the horizontal and vertical orientations were also tested. Both SDS-agarose and SDS-PAGE provided distinct separation of free DNA from cross-linked DNA. While both require the use of dialysis membranes to extract the desired product, SDS-PAGE provides separation of full length protein from protein fragments, ensuring that only the desired cross-linked products are collected. Ultimately, SDS-PAGE was chosen as the preferred preparative method for separating free DNA and cross-linked DNA. SDS-PAGE is most commonly used for protein analysis; therefore nucleic acid staining procedures are not widely used. The preferred method for staining nucleic acid in SDS-PAGE is washing the gel in dH_2O for 10 minutes, rinsing once with dH_2O, then staining in ethidium bromide at 1 µg/ml in dH_2O for 20 minutes.[178] This staining method is superior to staining with ethidium bromide in running buffer and staining with SYBR gold, both of which resulted in a cloudy image.

Reversible formaldehyde cross-linking conditions

While determining the best preparative gel-shift option, the reaction conditions for HCHO cross-linking were also optimized. The detailed procedure was modeled after the technique by Brodolin.[172]

(1) Combine DNA and thrombin with 2 µl 5X CLB buffer (0.25 M HEPES-NaOH, pH 8.0, 0.5 M NaCl, 25 mM $MgCl_2$, 25% glycerol) to a total volume of 10 µl.

(2) Gently mix, then heat to 37°C for 30 minutes.

(3) Add 1 µl of 0.2 M formaldehyde, freshly prepared from 37% formaldehyde (Sigma-Aldrich).

(4) Vortex briefly and incubate for 30 minutes at 37°C.

(5) Add 11 µl of 2X SB buffer (125 mM Tris-HCl pH 6.8, 2% SDS, 10 mM DTT, and 10% glycerol) to quench the reaction.

(6) The 22 µl final volume is used immediately for gel electrophoresis or frozen at -80°C.

(7) If desired, heat sample at 95°C for 5 minutes or 65°C for a minimum of 4 hours to reverse cross-links.

(8) Pre-run 12% SDS-PAGE with 0.1% SDS Tri-Glycine running buffer for 1 hour at 80V at 4°C. The gel apparatus was placed in an ice bath in a 4°C chromatography refrigerator to prevent heating which may reverse the DNA-protein cross-links.

(9) Run the gel for 1-1.5 hours under the same conditions.

(10) Rinse gel in 200 µl dH_2O for 10 minutes with agitation. Rinse briefly with dH_2O. Stain gel with 1 µg/ml ethidium bromide in dH_2O for 20 minutes with agitation. Visualize under UV light.

(11) If desired, stain protein with Coomassie Blue or GelCode Blue Safe Protein Stain.

Many of the parameters outlined above, including reaction temperature, cross-linking time, and HCHO concentration were varied to determine the effects on cross-linking efficiency. The quantity of cross-linked products is greater at 37°C than at 25°C and at 200 mM compared to lower concentrations. It was also determined that at 37°C, the quantity of cross-linked product for 100 pmol of thrombin with 100 pmol of TBA is saturated after approximately 30 minutes of reaction time. **Figure 6.2** illustrates the result of cross-linking thrombin in the presence or absence of TBA at 37°C and the effect of heating at 95°C for 5 minutes. Following staining with ethidium bromide in **Figure 6.2.A**, it was apparent that both the nucleic acid and protein could be visualized. Vincent and Scherrer[178] noted that proteins can also be detected by ethidium bromide and visualization of proteins can be intensified by reducing the fluorescence of the gel itself by washing in water prior to staining. Although protein staining was used for some preliminary experiments to confirm the location of thrombin (see **Figure 6.2.B**), this effect essentially eliminates the need for protein staining. Lanes 1 through 4 show the effect of cross-linking time on thrombin in the absence of TBA. A very slight increase in size is noticed as reaction time is increased from 30 seconds in lane 1 to 20 minutes in lane 4. In lane 5, the sample has been heated to 95°C for 5 minutes following 20 minutes of cross-linking time. After reversal of the cross-links, a slight shift back to the smaller product is noticed. This shift in size is likely the result of protein-protein cross-links. Lanes 6 through 10 show the effect of cross-linking time on thrombin in the presence of TBA. An obvious shift in size occurs with increasing intensity as time is increased from 30 seconds in lane 6 to 20 minutes in lane 9. After reversal of the cross-links, the larger product disappears in lane 10. **Figure 6.2.B** shows the same gel with coomassie blue protein staining. The visible shift to a larger band correlates with the shifting observed in **Figure 6.2.A** and only occurs in the presence of TBA. The results from this experiment indicate

that the band that appears following the cross-linking of thrombin and TBA is a TBA-thrombin complex and that the cross-links are visibly reversed after 5 minutes at 95°C.

Application of HCHO cross-linking to AIA

Based on sequencing data presented in chapter 5, in order for HCHO cross-linking to be an effective method of partitioning with a 100 pmol aliquot of a tailed library and no amplification, the quantity of recovered library must be approximately 0.011 pmoles or less. Based on calculations from **Table 5.2**, if 0.01 pmoles is recovered with a 100% retention rate of a unique sequence from a staring pool of 100 pmoles (~56,000 copies of each sequence) and optimal cluster efficiency of a 4-plex sequencing run is achieved, that unique sequence should appear approximately 46 times in 5,000,000 clusters. This is sufficient to distinguish aptamer hits against a low background. However, an even smaller recovery with lower background is favorable to accommodate less than optimal cluster density.

To test reversible HCHO cross-linking as a partitioning method, the m=15 tailed library (pACAC-m15-CACA) was screened against thrombin in ratios of 1:1, 1:5, 1:10 and 1:50 (thrombin: library) as indicated in **Table 6.1**. The quantity of thrombin was held constant at 100 pmol while the library quantity was varied at 100 pmol, 500 pmol, 1 nmol and 5 nmol. The samples were cross-linked at 37°C for 30 minutes following the aforementioned protocol and analyzed via SDS-PAGE (**Figure 6.3**). Lane 1 shows the location and intensity of 100 pmol of free m=15 library (cross-linked). Lane 2 serves as a negative control (absence of DNA) to show the location and intensity of 100 pmol of cross-linked thrombin. Lanes 3-6 are the cross-linking reactions in order of increasing library concentration for 100 pmol of thrombin. Due to the large

quantity of free DNA and because thrombin is also stained by ethidium bromide, it is difficult to estimate the quantity of DNA that is cross-linked. The gel image illustrates that as the concentration of library increases, a greater quantity of DNA-protein cross-links are formed. Also evident, is the formation of a second, larger cross-linked product that appears with increasing concentration of library. This band appears clearly in lanes 4 and 5 and is likely protein cross-linked with >1 library molecules. The bands corresponding to DNA-protein cross-links and what is assumed to be protein cross-linked with >1 library molecules were excised for lanes 3-6 with gel excision tips (GeneCatcher). The DNA-thrombin complexes were eluted from the gel slices using D-Tube Dialyzers (Millipore, MW cut-off 3.5 kDa). The gel slices were placed in pre-hydrated dialyzers with 800 µl of running buffer. The dialyzers were submerged in running buffer in a horizontal electrophoresis unit and current was applied perpendicular to the membrane windows at 110 V for 2 hours. The current was reversed for 1 minute to release the protein and DNA from the membrane. The current was reapplied in the initial direction for an additional 30 minutes followed by a final 2 minutes with reversed current. It is possible that the HCHO cross-links were reversed during electro-elution, as warming of the buffer was not prevented. The molecular weight cut-off of the membrane is 3.5 kDa, which is lower than the average molecular weight of the library at ~7.0 kDa, so any free DNA would not be lost. After pipetting up and down 5 times, the buffer was removed from the dialysis tube. An aliquot of 200 µl running buffer was used to rinse the inside of the dialysis tube and was collected in the same manner. To ensure complete reversal of the cross-links, the combined 1,000 µl sample was heated at 65°C overnight. Phenol extraction was used to recover the partitioned DNA libraries. Equal volume Tris-buffered, pH 8.0, 0.1mM EDTA, 50% phenol, 48% chloroform, 2% isoamyl alcohol (Sigma Aldrich) was added and vortexed for 30 seconds. Centrifugation at 13,000 rpm

for 5 minutes resulted in an aqueous top layer (containing the library) and organic bottom layer. The aqueous layer was removed and the library was subsequently extracted twice with equal volumes 100% chloroform. The library was then purified by ethanol precipitation. One tenth the volume of 3M sodium acetate and three times the volume of cold ethanol and were added to the library and mixed briefly. Following an overnight incubation at -20°C, centrifugation at 13,000 rpm for 30 minutes resulted in a library pellet. After decanting the liquid, the pellet was washed with 1 ml 70% ethanol. Centrifugation at 13,000 rpm for 20 minutes resulted in a library pellet. The pellet was air dried in a SpeedVac, and then resuspended in 10 µl dH_2O. The recovered DNA was not quantified to reduce loss.

The libraries were modified with sequencing adapters and their complements. Adapter 1, Adapter 1 Complement, Adapter 2 (indexed), and Adapter 2 Complement (indexed) at 10 µM were added in 1 µl volumes to the recovered library. The mixtures were incubated at 90°C for 3 minutes and cooled in a thermocycler at a rate of 0.5°C per minute to 4°C. At 4°C, 2 µl of 10X ligation buffer (300mM Tris-HCl (pH 7.8), 100mM MgCl2, 100mM DTT and 10mM ATP), (Promega), 3 µl dH_2O and 1 µl T4 DNA ligase (10mM Tris-HCl (pH 7.4), 50mM KCl, 1mM DTT, 0.1mM EDTA and 50% glycerol), (Promega) were added. Following a 24 hour incubation at 4°C, the ligated products were visualized on a 4% TAE agarose gel alongside negative and positive controls. The gel image is shown in **Figure 6.4.** The ~144 bp product is identified in the positive control lanes 3 and 8. The ligated product is visible for all four experimental samples with increasing intensity as the starting concentration of library increases from 100 pmol to 5 nmol. This indicates that a fraction of the starting library was successfully cross-linked to thrombin, eluted from the gel slice, and purified by phenol extraction and ethanol precipitation

prior to ligation. From previous experiments, it is known that a visible ligation product contains too large a quantity of DNA to be suitable for sequencing. The quantity of nonspecific sequences in the four samples would be too high to distinguish any aptamer candidates. For this reason, quantitative PCR was not performed. However, the results from this experiment provide confidence that the reversible formaldehyde cross-linking procedure and the purification methods described above will result in the correct product. This confidence is critical because the quantity of DNA that will produce favorable sequencing data is not visible during SDS-PAGE and agarose gel electrophoresis.

Increasing selection stringency

To determine the concentration of thrombin in a 30 minute cross-linking that would produce fewer than 0.01 pmoles of recovered library, a dilution series of thrombin was cross-linked to the m=15 tailed library. A 100 pmol aliquot of the library was combined with 100, 20, 10, 5, 1, 0.1, or 0.01 pmoles of thrombin (see **Table 6.2, top)** and HCHO cross-linked for 30 minutes at 37°C. The cross-linked library was separated from free library by SDS-PAGE. **Figure 6.5** shows the cross-linked products separated by SDS-PAGE. Lane 1 serves as a negative control (absence of DNA) to indicate the location of thrombin. Lanes 2-8 are the cross-linking reactions in order of decreasing thrombin concentration. Lane 9 is a duplicate of lane 2, containing 100 pmoles of m=15 DNA cross-linked to 100 pmoles of thrombin, and serves as a size marker for excising the non-visible bands across the gel. The cross-linked products from lanes 3-8 (20 – 0.01 pmol thrombin) were excised using 4.0 mm by 1.0 mm disposable gel excision tips (GeneCatcher) to minimize cross contamination between samples. The DNA-protein cross-links were eluted from the gel slices using D-Tube Dialyzers, cross-links were reversed at 65°C overnight, and the

library was phenol/chloroform extracted and ethanol precipitated as described above. The samples were ligated with indexed adapters similarly to the previous samples and visualized on a 4% TAE agarose. Although no bands were visible, gel slices were excised at the corresponding location as identified by a positive control size marker. The ligated library was extracted from the gel with the MiniElute Gel Extraction Kit (Qiagen) and resuspended in 10 µl dH_2O. The precise quantity of amplifiable DNA was determined with the KAPA Library Quantification kit on the BioRad iCycler. **Table 6.2** outlines sample information and statistical data. The number of pmoles of library recovered was calculated by qPCR and was used to determine the potential number of TBA counts in a 4-plex sequencing run with optimal cluster efficiency. Although all six samples were quantifiable by qPCR, the expected trend in the quantity of recovered DNA is not seen. Because the concentration of library was too low to be visualized during ligation for several of the samples, the ligated products were characterized by melt curve analysis to determine whether the quantified DNA was the full length product or primer-dimers. The samples were amplified with the SYBR Select Kit (Life Technologies). Melting curve data was collected over a gradient of 95°C to 50°C at a rate of 0.5°C per 30 seconds. All six ligated samples were characterized with the correct ~82.5°C melting temperature. The primer-dimer that occurs in the negative control was characterized by a melting temperature of ~80.5°C. qPCR data suggests that ≤20 pmoles of thrombin cross-linked for 30 minutes at 37°C will produce a low enough quantity of library to allow high affinity binding sequences to be counted above background.

To demonstrate consistency in sample preparation prior to a commitment to sequencing, a large pool of samples with variations of protein and library concentration and cross-linking times were

analyzed by EMSA on SDS-PAGE. A combination of four samples thought to provide the greatest breadth of understanding on the dynamics of the cross-linking reaction was chosen for sequencing. The goal was to recover a small quantity of library that contains primarily high and moderate affinity binding sequences and low background. The m=15 tailed library was cross-linked to a dilution series of thrombin consisting of 100, 10, 1, 0.1, and 0.01 pmoles. The m=15 tailed library was used in 1 nmol quantities to ensure high copy number of approximately 560,000 in the starting pool. If 0.01 pmoles of library were recovered from a 1 nmol pool with 100% recovery of the 560,000 counts of TBA and optimal cluster efficiency of a 4-plex sequencing run was achieved, TBA should appear approximately 460 times in 5,000,000 clusters. Each ratio was cross-linked for times of 30 seconds, 3 minutes and 30 minutes at 37°C. The samples were analyzed on SDS-PAGE as illustrated in **Figure 6.6** (only 100, 10 and 1 pmol are shown). The thrombin-library complex is visible for 100 pmoles of thrombin at 3 minutes and 30 minutes and for 10 pmoles of thrombin at 30 minutes. The four samples chosen for sequencing were 100 pmoles of thrombin at 30 seconds and 1 pmole of thrombin at 30 seconds, 3 minutes and 30 minutes. This combination of samples provided a comparison of protein concentration and cross-linking time on selection efficiency and stringency. The cross-linked products from lanes 2, 8, 9 and 10 were excised using 4.0 mm by 1.0 mm disposable gel excision tips (GeneCatcher) and eluted using D-Tube Dialyzers, cross-links were reversed at 65°C overnight and the library was phenol/chloroform extracted and ethanol precipitated as described above. The samples were ligated with indexed adapters similarly to the previous samples and visualized on a 4% TAE agarose. Although no bands were visible, gel slices were excised at the corresponding location as identified by a positive control size marker. The ligated library was extracted from the gel with the MiniElute Gel Extraction Kit (Qiagen) and resuspended in 10 µl

dH_2O. The precise quantity of amplifiable DNA was determined with the KAPA Library Quantification kit on the BioRad iCycler. Additionally, the raw amplification data was used to calculate the quantity of amplifiable DNA using qPCR CopyCount^{TM} software as a means of verifying the performance of the KAPA kit to ensure maximum cluster quality.

Absolute quantification using qPCR CopyCount^{TM}

The accurate quantification of a library is crucial for maximization of data output on Illumina sequencing platforms. If the amount of amplifiable library is overestimated, the result is lower than expected cluster density. If the amount of amplifiable library is underestimated, the result is crowded clusters and poor resolution. The KAPA Library Quantification kit has been used extensively for quantification of sequencing samples in AIA. The KAPA kit requires the use of six DNA standards and the acquisition of a standard curve. The quantification of unknowns is inferred based on this curve. In contrast, qPCR CopyCount^{TM} software is capable of analyzing the shape of any qPCR curve in the absence of standards to determine the absolute number of amplifiable copies of DNA at cycle zero. qPCR CopyCount^{TM} is built upon the Mass Action Kinetic model with 2 parameters (MAK2) that account for the concentration at cycle zero and changes in amplification efficiency by cycle to determine the relative concentration of a target in solution[179]. pPCR CopyCount^{TM} is three times more accurate than MAK2 and is automated for all qPCR instrumentation. These developments led to the principle of cPCR (counting PCR), where the fluorescence of a single copy of DNA is determined from the amplification curve to determine the number of copies of DNA at cycle zero. The software boasts quantification with 20% absolute accuracy and 1-5% relative accuracy without calibration. The implementation of a calibration plate (cPCR for 96-384 well plate with ~1.5 copies per well) for a given assay

improves the absolute accuracy to less than 5% error.[180] As mentioned, the HCHO cross-linking experiments intended for sequencing were quantified using the qPCR CopyCount™ software in addition to the KAPA Library Quantification kit. The results from both methods of quantifications are detailed in **Table 6.3.** The quantity of recovered library was calculated with the KAPA data as described in chapter 5. The raw amplification data from the KAPA data (including amplification data for the standards) was analyzed with qPCR CopyCount™. The average copy number from triplicate wells was used to calculate the concentration of the reaction in pM, which was used to calculate the pmoles of library in the stock solution.

Absolute quantification with the qPCR CopyCount™ software gave higher concentrations for all samples. The data from both methods is in approximate agreement for samples 3 and 4. Sample 2 was estimated to be ~1.38 times higher by qPCR CopyCount™. Sample 1 was estimated to be ~6.12 times higher by qPCR CopyCount™. In order to provide insight into the large variation in sample 1, the raw fluorescence data of the six KAPA standards was analyzed with qPCR CopyCount™. The qPCR CopyCount™ determination for the six standards was also higher than the assumed concentrations, ranging from ~1.3 – 3.2 times higher. The discrepancies observed in the absolute quantification of the standards explain the varied results for the unknowns. **Figure 6.7** illustrates the amplification variability of the six KAPA standards as determined by qPCR CopyCount™ for six experiments. The log of the starting concentration (pM) is plotted versus the standard number, creating an illustration of the linear standard curve. For all six experiments, the amplification does not fit the linear standard curve. It is overestimated by varying degrees in all instances, with individual data points varying from approximately 1.06 – 6.6 times higher. This variability between individual data points within a standard curve and the variability

between standard curves of separate experiments illustrates the potential inaccuracies introduced by the standard curve. The qPCR CopyCount™ software has been validated on over 100,000 samples and the website details digital droplet qPCR data compared to qPCR CopyCount™ data with an R^2 value of 0.99991.[180] A calibration plate for the KAPA kit assay was not successful (unreliable due to poor Chi-squared distribution), indicating that the qPCR CopyCount™ data should have a maximum of 20% absolute error. When compared to the KAPA determinations of the experimental samples and standards, the qPCR CopyCount™ data varied well outside this margin. If the qPCR CopyCount™ data is accurate within 20%, this suggests that experimental concentrations determined from the overestimated standard curves would be underestimated. Underestimating the quantity of amplifiable DNA results in over-loading the flow cell which can cause over-clustering and poor resolution.

There are multiple metrics that can be used to help identify over-clustering, including the percentage of clusters that pass filter, Q30 scores, and signal intensity per tile. If over-clustered, clusters begin to overlap which leads to poor template generation. Especially for low diversity libraries, distinguishing between overlapped clusters is difficult and results in poor signal purity. This reduces the overall percentage of clusters that pass filter, ultimately reducing data output. The same effect is observed during demultiplexing, where signal purity of low diversity indices is compromised by over-clustering. Additionally, if a flow cell is over-clustered, the overall fluorescence intensity of the flow cell increases, making it more difficult to analyze signal intensity. This can lead to low Q30 scores; the Q score identifies the probability of incorrect base calling and a score of >Q30 indicates 99.9% accuracy in base calling. The MiSeq flow cell contains 28 tiles with both top and bottom surfaces. The average signal intensity per tile allows

the user to identify regions of the flow cell that are over-clustered. If a tile is over-clustered, image extraction may fail which results in an intensity of zero for that tile, rendering intensity data for that tile unusable. The most recent sequencing run on the Illumina MiSeq was loaded with the suggested maximum concentrations of DNA (0.04 pmoles of single-stranded library as determined by the KAPA kit) and the number of clusters totaled ~16 million. A lower than expected percentage of clusters passed filter (~51%), indicating poor cluster quality. Furthermore, qualified data was reduced to 41% during demultiplexing. The percentage of clusters with >Q30 scores was 85.5% and multiple tiles were identified with image extraction failure. Analysis of the raw qPCR amplification data for the samples sequenced on this run with qPCR CopyCount^{TM} produced higher concentrations for all samples, confirming that they may have been underestimated by the KAPA kit. For the sequencing run of the HCHO cross-linking samples, the qPCR CopyCount^{TM} determinations were used to prepare the sample for loading onto the flow cell. No changes to the sequence structure were made, so if cluster quality improved when utilizing the qPCR CopyCount^{TM} determination for the HCHO cross-linked samples, it would indicate a more accurate determination. Additionally, a PhiX control (generated from the PhiX174 bacteriophage genome) was spiked at 5% during clustering. The PhiX control introduces high diversity clusters to aid in creating cluster diversity among the low diversity tailed libraries.

Results and Discussion

Statistical analysis of sequencing data for the four HCHO cross-linking experiments with varied thrombin concentration and reaction times is outlined in **Table 6.4.** The total number of clusters

is ~17 million with approximately 79% passed filter. Following demultiplexing, this was reduced to ~61%. The percentage of clusters with >Q30 scores was 90.9% and there are no indications of over-clustering based on this data. The four samples were loaded in a 1:1:1:1 ratio during clustering according to the qPCR CopyCount™ determination. The resulting ratio of PF clusters for samples 1-4 was 1.35:1.06:1.00:1.03. Samples 3-4 are in approximate agreement while sample 1 was clustered at a higher frequency. This suggests that the concentration of sample was slightly underestimated in comparison to samples 3-4. Approximately 29% of sample 1 was loaded onto the flow cell, however, 90% would have been required using the KAPA kit determination which would have drastically skewed the ratio of indices to sample 1. Samples 3-4 would have been loaded with an approximate 10% increase, although the ratio remains relatively the same. If the KAPA kit determinations had been used, the distribution of indices would have been significantly more skewed and the flow cell may have been over-clustered. Clustering with the qPCR CopyCount™ data produced a favorable distribution of indices and prevented over-clustering.

The data was parsed with the Perl script to identify sequences with perfect matches for both the head and tail as "candidate" reads and those with an m=15 variable region as "good" reads. The counts per unique sequences were tallied and the top ten sequences are presented in **Table 6.5**. The percentage of good reads for all four samples was lower than expected. A brief search of the "candidate" and "bad" reads does not indicate that the counts for the top sequences would be altered drastically if a revision to the Perl script was made to allow for mismatches or insertions/deletions in the head and tail. The majority of the bad reads were jump sequence variants that would not contribute to the nmer count data. Considering the improved cluster

quality for this sequencing run, it was unclear why the percentage of good reads was low. The expected counts of TBA (or any unique aptamer sequence) for each sample were recalculated based on the total number of reads identified after demultiplexing and an assumed 100% recovery of the ~560,000 starting counts.

The presence of TBA above background indicates that reversible HCHO-crosslinking in conjunction with EMSA was a successful partitioning method. The expected counts were determined as 85, 184, 103 and 96, respectively, while the observed counts of TBA were 123, 15, 0 and 342, respectively. As indicated in **Table 6.5,** the only non-TBA molecules above background are jump sequence variants. This indicates a high level of stringency during partitioning. The results also confirm the applicability of the tailed libraries in amplification-free AIA and the ability to predict relative sequence frequency. The similarity of observed counts/frequency of TBA to the expected counts/frequency confirms the selection efficiency. Cross-linking 1 nmol of m=15 tailed library against 100 pmol of thrombin for 30 seconds produced 123 counts of TBA, an increase of 145% over the expectation. Cross-linking for the same duration with 1 pmol of thrombin produced 15 counts, only 8% of the expected. Increasing the cross-linking time to 3 minutes for 1 pmol of thrombin produced 0 counts of TBA. If the cross-linking was comparably efficient to the 30 second cross-linking time that recovered 8% of the expected, ~8.24 counts of TBA would be expected for sample 3. Although the samples were prepared carefully, slight variations in sample preparation are always possible, including pipetting techniques. For example, a 10% variation in library volume would reduce the starting counts of expected counts of TBA from 103 to 92. Variations in HCHO or protein volume are also possibilities. Increasing the cross-linking time to 30 minutes produced 342 counts of TBA,

an increase of 356% over the expectation. The calculations for the expected counts are based on the assumption that the 1 nmol library pool contained ~560,000 copies of TBA. It is possible that the concentration of the stock solution was underestimated, resulting in a larger than expected quantity of starting pool. This would explain the higher than anticipated counts of TBA observed for samples 1 and 4. Regardless, this data suggests that for 1 pmol of thrombin and ~1 nmol of library, 30 seconds and 3 minutes are not sufficient to reliably select for TBA above background. A 30 second cross-linking time is sufficient for 100 pmol of thrombin; however, cross-linking for 30 minutes with 1 pmol of thrombin was superior. In both experiments, a single TBA variant is also identified above background.

To confirm the results of cross-linking for various times, samples 2-4 were repeated and sequenced on the Illumina Seq. **Table 6.6** details the sample information and anticipated sequence frequency according to qPCR data from both the KAPA kit and qPCR CopyCount^{TM}. The calculated recovery for each sample was below the maximum 0.01 pmoles for a 4-plex sequencing run, therefore, the samples were loading in their entirety (90% of the recovered library after qPCR analysis) in a 1.6:1.1:1 ratio. The overall reduced quantity of recovered library could be attributed to variations in sample preparation, inducing efficiency of electro-elution, phenol extraction, ethanol precipitation or ligation. Statistical analysis of sequencing data for the three HCHO cross-linking experiments with varied reaction times is outlined in **Figure 6.7.** The total number of clusters is ~13.5 million with approximately 79% passed filter. Following demultiplexing, this was reduced to ~57%. A fourth, unrelated sample was also sequenced and 88% of the readable data corresponds to the three HCHO cross-linking samples. The percentage of clusters with >Q30 scores was 90.2% and there are no indications of over-

clustering based on this data. The resulting ratio of PF clusters for samples 2.1-4.1 was 1.45:1.06:1, which is in approximate agreement with the ratio of recovered library as determined by qPCR CopyCount™. The ratio as determined by the KAPA kit is 0.42:1.25:1, indicating that the qPCR CopyCount™ determination was more accurate.

The data was parsed with the Perl script as described above and the counts per unique sequences were tallied and the top ten sequences are presented in **Table 6.8**. The percentage of good reads for all three samples was also lower than expected and the majority of the bad reads were jump sequence variants that would not contribute to the nmer count data. The expected counts of TBA (or any unique aptamer sequence) for each sample were recalculated based on the total number of reads identified after demultiplexing and an assumed 100% recovery of the ~560,000 starting counts.

The expected counts of TBA for 30 seconds, 3 minutes, and 20 minutes of cross-linking were determined as 400, 448 and 451, respectively. The observed counts of TBA were 28, 4, and 272, respectively. The results correlate with the findings in the first set of samples: a 30 minute cross-linking time with 1 pmol of thrombin and 1 nmol of library is capable of selecting TBA well above background, while 30 second and 3 minute cross-linking times are not as efficient. The 30 minute cross-linking time also selected four TBA variants above background, while one was counted in the prior run reported in **Tables 6.4** and **6.5**. Although the K_d's of these sequences are not known, it can be assumed that they are lower affinity binders. This supports the hypothesis that both high and moderate affinity aptamers can be selected with reversible HCHO cross-linking in conjunction with EMSA.

A comparison of the frequency of TBA observed in these seven experiments is not straightforward due to the low percentage of good reads and varied quantity of recovered library. A comparison of frequency within a sequencing run may offer a different conclusion than a comparison of frequency across sequencing runs. If a comparison is made across sequencing runs, it could be concluded that the conditions for sample 1 (30 seconds, 100 pmol thrombin) and sample 2.1 (30 seconds, 1 pmol thrombin) offer similar partitioning efficiencies because they have the same frequency of 0.015%. However, a comparison of the frequencies for sample 1 and sample 2 from the same sequencing run (0.015% and 0.039%, respectively) indicates that sample 1 offers more efficient partitioning conditions. In previous chapters, the frequency of TBA was compared across sequencing runs as a metric for determining partitioning efficiency. However, amplification free AIA introduces a higher level of understanding of the relationship between the quantities of recovered library and sequence frequency. For example, the frequency of TBA for samples 2 and 2.1 (30 seconds, 1 pmol thrombin) are 0.0039% and 0.015%, which correlates with the quantity of recovered library of 0.0127 pmoles and 0.0065 pmoles, respectively. A larger quantity of recovered library contains a larger percentage of nonspecific sequences which reduces the expected frequency of any high affinity sequence.

Conclusions

The results from two MiSeq sequencing runs for reversible formaldehyde cross-linking in conjunction with EMSA illustrate that the method was successful as a partitioning method for the selection of the thrombin binding aptamer against thrombin in solution. The sequencing data also

confirms the applicability of the tailed libraries in amplification-free AIA and the ability to predict relative sequence frequency. Additionally, the sequencing data suggests that quantification with qPCR CopyCount^{TM} software is more accurate and produces a more favorable distribution of indices compared to the KAPA Library Quantification kit. The accurate quantification of libraries is crucial for maximization of data output on Illumina sequencing platforms. Maximizing high quality data output while decreasing costs and labor demands are consistent with the goals of AIA. Implementation of qPCR CopyCount^{TM} eliminates the need for standard curves, which saves both time and money. Additionally, producing higher quality sequencing data with the desired distribution of indices eliminates costly resequencing.

For the discovery of novel aptamers to protein targets, the cross-linking conditions would need to be determined empirically from multiple sequencing experiments. The cross-linking efficiency for individual protein targets will differ based on the amino acid sequence and availability of amino/imino groups as well as the relative affinity of any aptamer candidates. The cross-linking conditions used for the m=15 DNA tailed library and thrombin may serve as a starting point for these experiments. **Figure 6.8** shows the EMSA of a preliminary reversible HCHO cross-linking experiment of SL3 RNA with NCp7. NCp7 was also shown to cross-link with an m=40 minimal primer library (described in the next section) and the m=15 DNA tailed library. Sequencing the results of a reversible HCHO cross-linking partitioning of a DNA or 2'-OMe RNA/DNA chimera library against NCp7 would illustrate the applicability of the method for other protein targets.

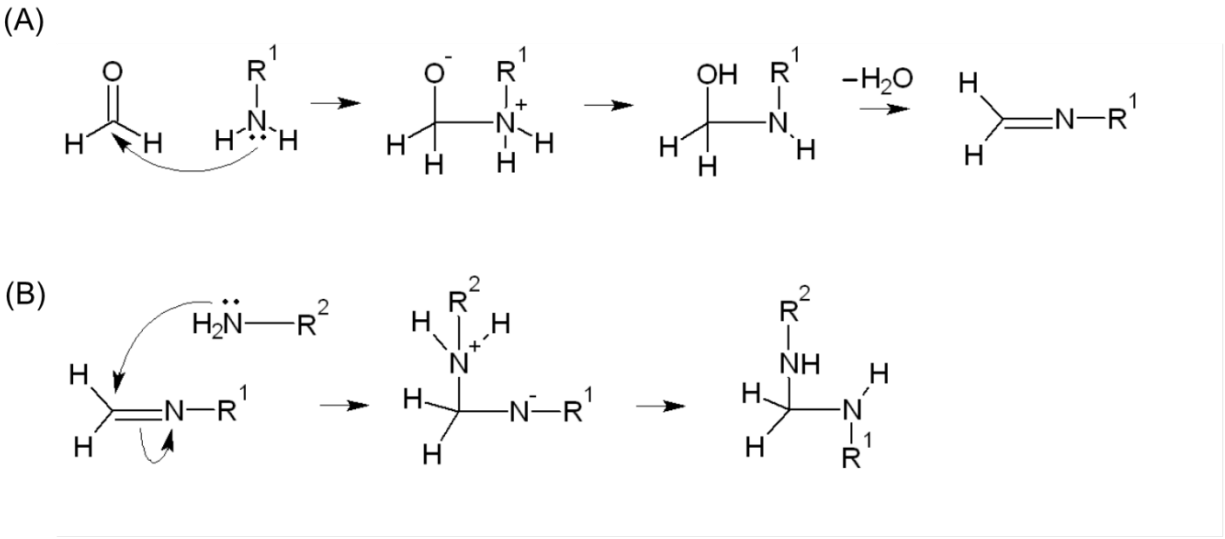


Figure 6.1 Formaldehyde cross-link formation

Formaldehyde crosslinking mechanism. **(A)** Reaction of formaldehyde with an amino group to form a Schiff bae. **(B)** The Schiff base can then react with another amino group to form the cross-link.

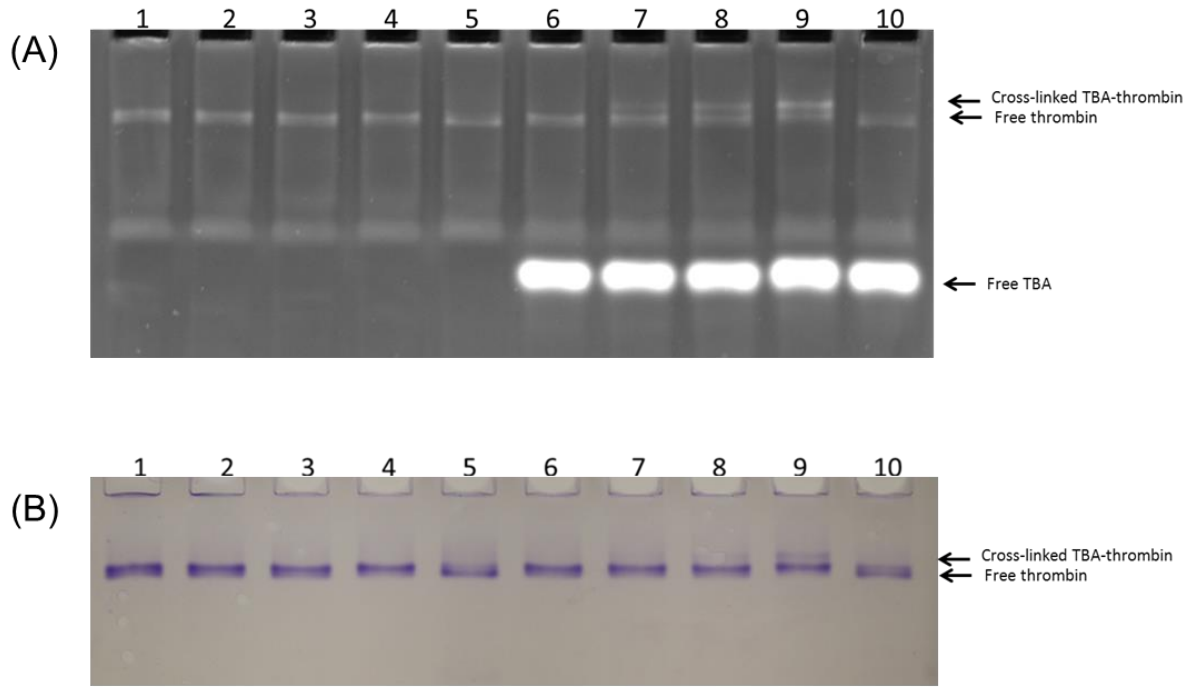


Figure 6.2 EMSA of reversible HCHO cross-linking of thrombin and TBA-thrombin

(A) 12% SDS-PAGE stained with ethidium bromide. Lanes 1-5: 10 pmol of thrombin. Lanes 1-4: HCHO cross-linked for 30 seconds, 5 minutes, 10 minutes and 20 minutes respectively. Lane 5: HCHO cross-linked for 20 minutes and reversed with heat. A slight shift in band size is observed with increased cross-linking time. Lanes 6-10: 10 pmol of thrombin and 100 pmol of TBA. Lanes 6-9: HCHO cross-linked for 30 seconds, 5 minutes, 10 minutes and 20 minutes, respectively. Lane 10: HCHO cross-linked for 20 minutes and reversed with heat. The thrombin-TBA complex is visible after 5 minutes of cross-linking time. Thrombin-TBA cross-links are visibly reversed after heating. **(B)** Same gel stained with Coomasie Blue protein stain. The visible shift to a larger band correlates with the shifting observed in (A) and only occurs in the presence of TBA.

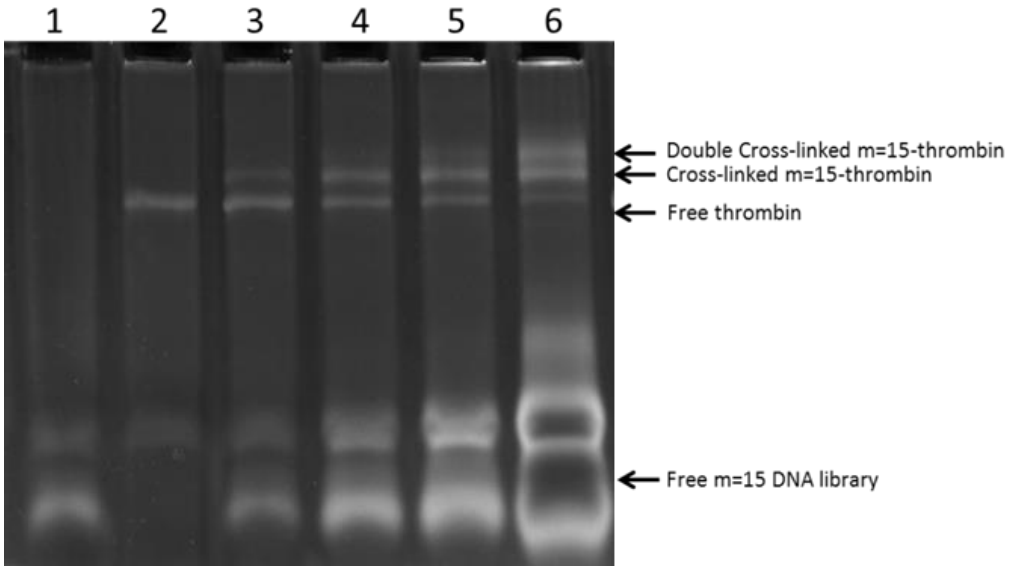


Figure 6.3 EMSA of reversible HCHO cross-linking of m=15 DNA tailed library to thrombin

12% SDS-PAGE stained with ethidium bromide. Lane 1: 100 pmol of cross-linked m=15 DNA library. Lane 2: 100 pmol of cross-linked thrombin. Lanes 3-6: 100 pmol thrombin cross-linked with m=15 DNA library in 100 pmol, 500 pmol, 1 nmol and 5 nmol quantities, respectively. Thrombin-library complexes are visible for all samples in lanes 3-6, with an increase in intensity as the concentration of library is increased. A second, larger band appears with increased library concentration as the result of crosslinking >1 library molecule.

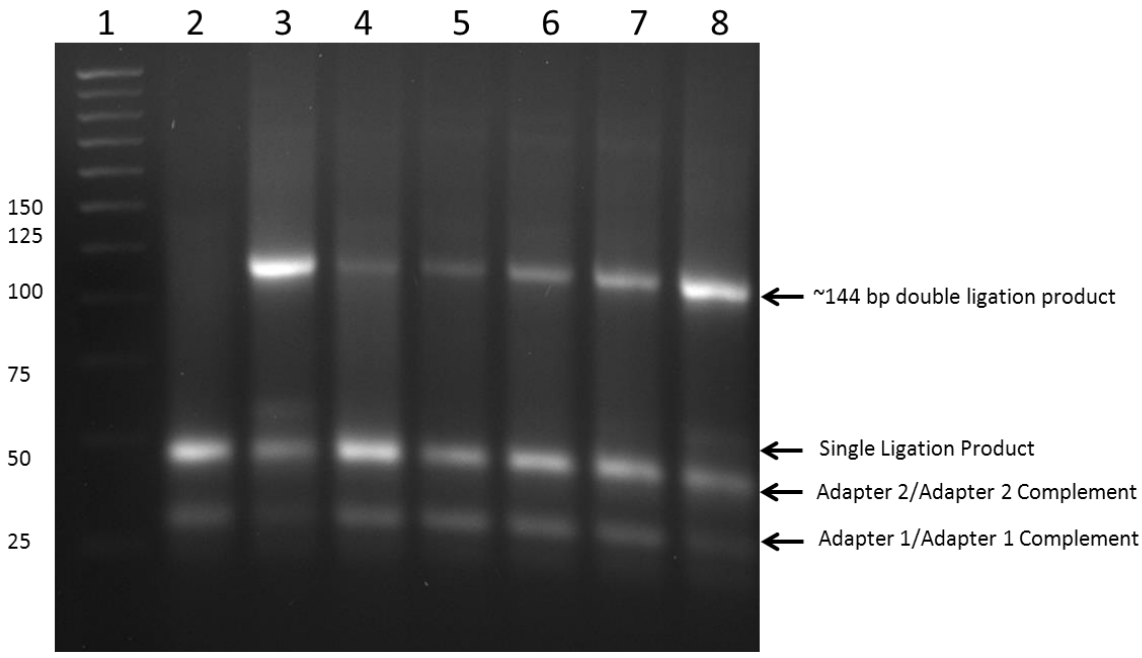


Figure 6.4. Ligation of library recovered by HCHO cross-linking and EMSA

Lane 1: 25 bp ladder. Lane 2: No template control. Lanes 3 and 8: positive controls, 10 pmol of m=15 DNA tailed library ligated to adapters. Lanes 4-7: ligated samples #1-4 (as named in **Table 6.1**). Correct double ligation product is seen for all four samples with increasing intensity as the concentration of library used in cross-linking increases.

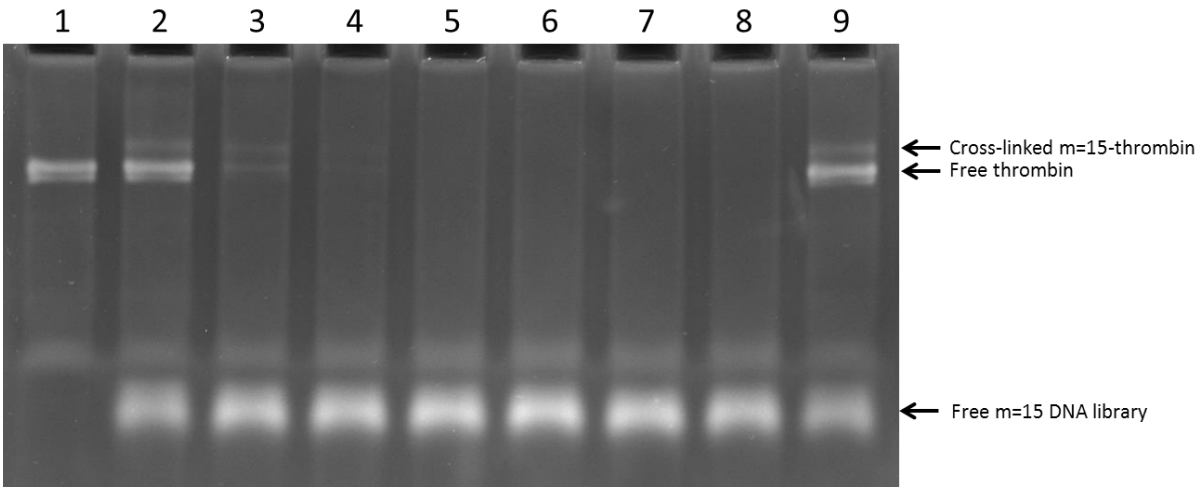


Figure 6.5 EMSA of reversible HCHO cross-linking of the m=15 DNA tailed library to a dilution series of thrombin

12% SDS-PAGE stained with ethidium bromide. 100 pmol of m=15 DNA tailed library HCHO cross-linked to a dilution series of thrombin for 30 minutes at 37°C. Lane 1 is a no DNA control. Lanes 2-8: 100, 20, 10, 5, 1, 0.1 and 0.01 pmoles thrombin, respectively. Lane 9 is a duplicate of Lane 2 to serve as a size marker. The thrombin-m=15 library complex is visualized for 100, 20 and 10 pmoles of thrombin.

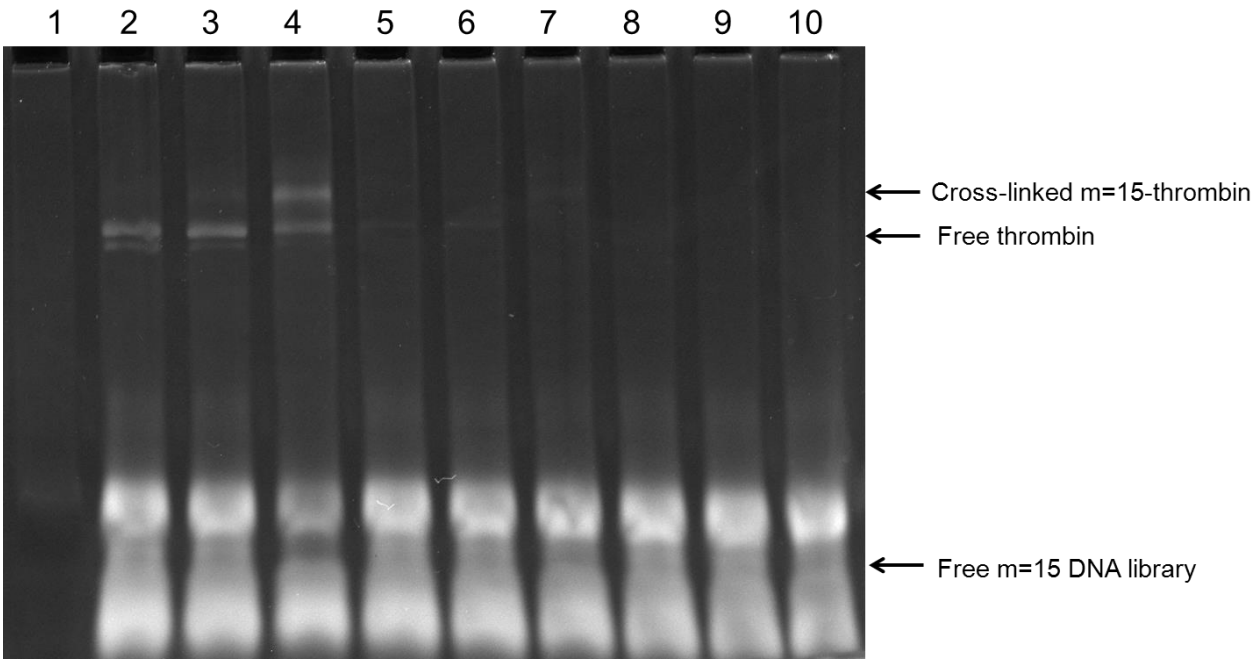


Figure 6.6 EMSA of reversible HCHO cross-linking the m=15 DNA tailed library to a dilution series of thrombin for 30 seconds, 3 minutes and 30 minutes

12% SDS-PAGE stained with ethidium bromide. 1 nmol of m=15 DNA tailed library HCHO cross-linked to a dilution series of thrombin for 37°C. Lane 1 is contains sample buffer only. Lanes 2-4: 100 pmoles of thrombin cross-linked at 30 seconds, 3 minutes and 30 minutes, respectively. The m=55-thrombin complex is visualized for 3 minutes and 30 minutes. Lanes 5-7: 10 pmoles of thrombin cross-linked at 30 seconds, 3 minutes and 30 minutes, respectively. The m=15-thrombin complex is visualized for 30 minutes. Lanes 8-10: 1 pmole of thrombin cross-linked at 30 seconds, 3 minutes and 30 minutes, respectively. The m=15- thrombin complex is not visible.

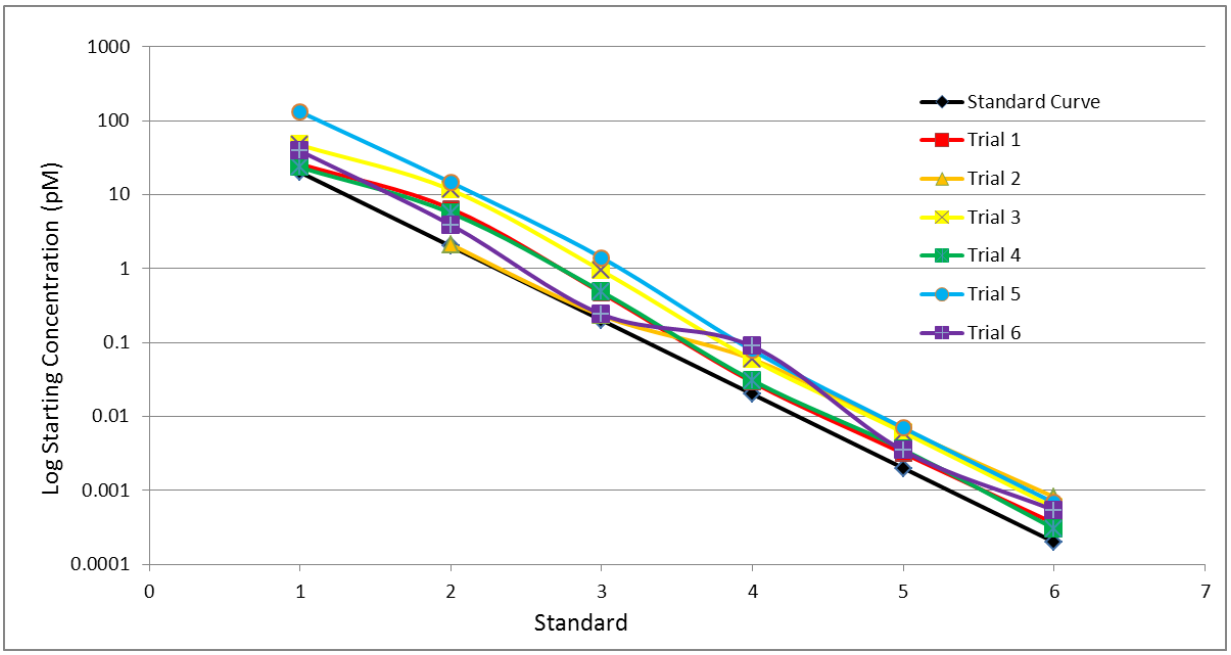


Figure 6.7 Comparison of KAPA Standard Curve versus qPCR CopyCount™ determination

The log of the starting concentration (pM) is plotted versus the standard number, creating an illustration of the linear standard curve of the KAPA Library Quantification kit. "Standard Curve" represents the six DNA standards with assumed concentrations of 20, 2, 0.2, 0.02, 0.002 and 0.0002 pM. The qPCR CopyCount™ determination for the six standards was determined for six experiments and is plotted versus standard number. The amplification data does not fit the linear standard curve and is overestimated by varying degrees in all instances, with individual data points varying from approximately 1.06 – 6.6 times higher.

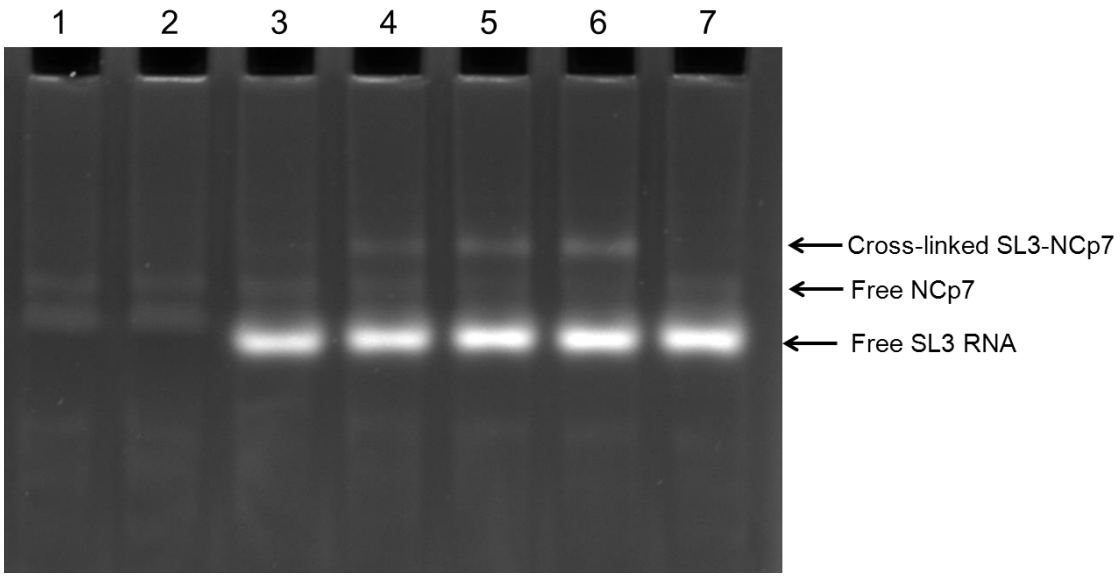


Figure 6.8 EMSA of reversible HCHO cross-linking NCp7 and SL3-NCP7

12% SDS-PAGE stained with ethidium bromide. Lanes 1-2: 10 pmol of NCp7 cross-linked for 20 minutes. Lane 2: reversed with heat. Lanes 3-7: 10 pmol of NCp7 and 100 pmol of SL3 RNA. Lanes 3-6: cross-linked for 30 seconds, 5 minutes, 10 minutes and 20 minutes, respectively. Lane 7: HCHO cross-linked for 20 minutes and reversed with heat. The SL3-NCp7 complex is visible after 30 seconds and cross-links are visibly reversed after heating.

Table 6.1 Reversible HCHO cross-linking sample information

Sample	Ratio (protein: library)	Protein	Library	Index
1	1:1	100 pmol	100 pmol	2
2	1:5	100 pmol	500 pmol	4
3	1:10	100 pmol	1 nmol	6
4	1:50	100 pmol	5 nmol	12

Sample information for four HCHO cross-linking experiments used to confirm the preparative methods.

Table 6.2 Sample information and anticipated sequence frequency for HCHO cross-linking the m=15 DNA tailed library to a dilution series of thrombin

Sample	Ratio (protein: library)	Protein	Library	Index
3	1:5	20 pmol	100 pmol	6
4	1:10	10 pmol	100 pmol	12
5	1:20	5 pmol	100 pmol	2
6	1:100	1 pmol	100 pmol	4
7	1:1,000	0.1 pmol	100 pmol	6
8	1:10,000	0.01 pmol	100 pmol	12

Sample	pmoles of Library Recovered	pmoles of Library Available	Number of Molecules	Expected Number of Clusters	Clusters as a Percentage of Molecules	Number of Surviving TBA molecules
3	0.00118	0.00106	6,380,000,000	5,000,000	0.78 %	436
4	0.00026	0.00024	1,440,000,000	5,000,000	3.47 %	1943
5	0.00040	0.00036	2,170,000,000	5,000,000	2.30 %	1288
6	0.00073	0.00066	3,970,000,000	5,000,000	1.25 %	700
7	0.00103	0.00093	5,600,000,000	5,000,000	0.89 %	498
8	0.00070	0.00063	3,790,000,000	5,000,000	1.32 %	739

(Top) Sample information for six HCHO cross-linking experiments of m=15 DNA tailed library cross-linked to thrombin. Sample name corresponds to the lanes in **Figure 6.5. (Bottom)** "pmoles of Library Recovered" was determined by qPCR analysis. "pmoles of Library Available" is the quantity of remaining sample after preparing the qPCR samples. "Number of molecules" is the number of molecules loaded onto the flow cell for a four-plex MiSeq run. "Expected number of clusters" is based on 20 million clusters for a four-plex MiSeq run. "Clusters as a percentage of molecules" is 5,000,000 as a percentage of the total molecules per sample. "Number of surviving TBA molecules" is the maximum number of TBA counts expected to be seen at 5,000,000 clusters with 100% recovery of the initial 56,000 copies of TBA in the 100 pmol pool of library.

Table 6.3 Sample information and anticipated sequence frequency for HCHO cross-linking the m=15 DNA tailed library with varied thrombin concentration and reaction times

Sample	Ratio (protein: library)	Protein	Library	Index	Cross-linking time
1	1:10	100 pmol	1 nmol	2	30 sec.
2	1:1000	1 pmol	1 nmol	4	30 sec.
3	1:1000	1 pmol	1 nmol	6	3 min.
4	1:1000	1 pmol	1 nmol	12	30 min.

	Sample	pmoles of Library Recovered	pmoles of Library Used	Percentage of Total	Number of Molecules Used	Expected Number of Clusters	Clusters as a % of Molecules	Number of Surviving TBA molecules
KAPA	1	0.0057	0.0051	90%	3,074,775,200	2,000,000	0.07%	328
	2	0.0092	0.0082	90%	4,965,898,000	2,000,000	0.04%	203
	3	0.0220	0.0125	57%	7,525,000,000	2,000,000	0.03%	84
	4	0.0213	0.0125	59%	7,525,000,000	2,000,000	0.03%	87
qPCR CopyCount™	1	0.0349	0.0100	29%	6,020,000,000	2,000,000	0.03%	53
	2	0.0127	0.0100	79%	6,020,000,000	2,000,000	0.03%	147
	3	0.0212	0.0100	47%	6,020,000,000	2,000,000	0.03%	88
	4	0.0235	0.0100	43%	6,020,000,000	2,000,000	0.03%	79

(Top) Sample information for four HCHO cross-linking experiments of m=15 DNA tailed library cross-linked to thrombin at varied protein concentration and cross-linking times. **(Bottom)** "pmoles of Library Recovered" was determined by qPCR analysis with the KAPA kit and QPCR CopyCount^{TM} software. "pmoles of Library used" is the quantity of sample required to equal a total of 0.04 pmoles for clustering. "Percentage of total" is pmoles of library used as a fraction of pmoles of library recovered. "Number of molecules" is the number of molecules loaded onto the flow cell for a four-plex MiSeq run. "Expected number of clusters" is based on cluster quality from previous data for a four-plex MiSeq run. "Clusters as a percentage of molecules" is 2,000,000 as a percentage of the total molecules per sample. "Number of surviving

TBA molecules" is the maximum number of TBA counts expected to be seen at 2,000,000 clusters with 100% recovery of the initial 560,000 copies of TBA in the 1 nmol pool of library.

Table 6.4. Statistical analysis of sequencing data for HCHO cross-linking of the m=15 DNA tailed library with varied thrombin concentration and reaction times

Run	Total Number Clusters	Number of Clusters PF	Total Reads in Data Files
Samples 1-4	17,076,298	13,446,773	10,462,729

Experiment	100 pmol thb 30 sec.	1 pmol thb 30 sec.	1 pmol thb 3 min.	1 pmol thb 30 min.
Quantity of Protein	100 pmol	1 pmol	1 pmol	1 pmol
Quantity of Library	1 nmol	1 nmol	1 nmol	1 nmol
Starting Copies per Sequence	560,000	560,000	560,000	560,000
Total Reads	3,182,769	2,504,860	2,354,637	2,420,463
Total Good Reads	813,673	387,470	188,277	480,985
Background/Noise Threshold	3	3	3	3
Sequences Above Background	4	3	2	4
Expected Counts of TBA	85	184	103	96
Observed Counts of TBA[a]	1_123	2_15	0	1_342
Frequency of TBA per Good Reads[b]	0.015%	0.0039%	0	0.07%
Frequency of TBA per Total Reads[c]	0.0039%	0.00060 %	0	0.0040%

(Top) Total number of clusters produced per sequencing run. The number of clusters that passed filter (PF) is approximately 79%. After demultiplexing, the total reads are approximately 61% of the original clusters. **(Bottom)** Total number of reads for each sample generated by Illumina's BaseSpace. A total of 7,827 reads contained Index 1 of the positive control size marker. The number of good reads was generated by the Perl Script. The expected number of TBA counts was calculated from the quantity of library recovered, the quantity of library applied to the flow cell and the number of good reads assuming a 100% recovery of the original 56,000 copies of TBA. **(a)** The rank *(left)* and count *(right)* are separated by an underscore. **(b)** Counts of TBA divided by the total good reads. **(c)** Counts of TBA divided by the total reads.

Table 6.5. Top 10 Sequences for HCHO cross-linking the m=15 DNA tailed library with varied thrombin concentration and reaction times

Sample 1: 100 pmol thb, 30 sec.		
Sequence	**Count**	**Identifier**
GGTTGGTGTGGTTGG	123	TBA
CACAGATCGGAAGAG	71	Jump*
GTCTGATCGGAAGAG	8	Jump*
GGTTGGTGTTGTTGG	5	Ia
TGCGTGGGCCGGCGC	3	
GTGTTATCGAATCCT	3	
TAGATAGCCAACTTA	3	
GTAAACCCAACTGCG	3	
TATATGAACCCGGAA	3	
CTGGGAAGGCGCAGC	3	

Sample 2: 1 pmol thb, 30 sec.		
Sequence	**Count**	**Identifier**
CACAGATCGGAAGAG	28	Jump*
GGTTGGTGTGGTTGG	15	TBA
GTCTGATCGGAAGAG	14	Jump*
TGAGGCGAGTGAACG	3	
ATCATGAACGTCAGT	3	
AATGTTGTGCAGAAC	3	
GATCTGATGTAACCC	3	
CTCAACCTGTTTGCG	3	
CCCGCTTCTAAATTG	3	
GTAACCCGGGGCCGA	3	

Sample 3: 1 pmol thb, 3 min.		
Sequence	**Count**	**Identifier**
CACAGATCGGAAGAG	11	Jump*
GTCTGATCGGAAGAG	7	Jump*
CCCACACGGGTATTT	3	
CAGCCAGAAACTACC	3	
CAGGAATACGTATCC	3	
ATAAGTCCTGATACC	3	
TTAAAGATATGCTGA	3	
GATATATCGGCTTAT	3	
ATCACACCCCTACCC	3	
CAGGCTTACTTCGCG	3	

Sample 4: 1 pmol thb, 30 min		
Sequence	**Count**	**Identifier**
GGTTGGTGTGGTTGG	342	TBA
CACAGATCGGAAGAG	30	Jump*
GTCTGATCGGAAGAG	11	Jump*
GGTTGTTGTGGTTGG	5	Ia
CGATGTACGCCATGA	3	
TACCGCGCTGACTCC	3	
CCAAACAAGATAAGG	3	
ATACCACGTCGGGCG	3	
CCCCTTGACCCGAGT	3	
ACCAGCTTAAACTTA	3	

Top 10 sequences and corresponding counts for Samples 1-4. The thrombin binding aptamer (TBA) was identified at counts of 123, 15, 0, and 342, respectively. Jump sequence variants are present in all four samples and account for the only non-TBA variants above background. The background for all four experiments is 3. Variants of TBA occur in samples 1 and 4 as the 4th ranked sequences.

*Jump sequence variants.

Table 6.6 Sample information and anticipated sequence frequency for repeat of HCHO cross-linking the m=15 DNA tailed library to thrombin with varied reaction times

Sample	Ratio (protein: library)	Protein	Library	Index	Cross-linking time
2.1	1:1000	1 pmol	1 nmol	2	30 sec.
3.1	1:1000	1 pmol	1 nmol	6	3 min.
4.1	1:1000	1 pmol	1 nmol	12	30 min.

	Sample	pmoles of Library Recovered	pmoles of Library Used	Percentage of Total	Number of Molecules Used	Expected Number of Clusters	Clusters as a % of Molecules	Number of Surviving TBA molecules
KAPA	2.1	0.0030	0.0026	90%	1,619,982,000	2,000,000	0.12%	622
	3.1	0.0089	0.0080	90%	4,827,438,000	2,000,000	0.04%	209
	4.1	0.0072	0.0064	90%	3,906,378,000	2,000,000	0.05%	258
qPCR CopyCount™	2.1	0.0065	0.0059	90%	3,543,372,000	2,000,000	0.06%	284
	3.1	0.0043	0.0038	90%	2,308,068,000	2,000,000	0.09%	437
	4.1	0.0040	0.0036	90%	2,172,618,000	2,000,000	0.09%	464

(Top) Sample information for three HCHO cross-linking experiments of m=15 DNA tailed library cross-linked to thrombin with varied cross-linking times. **(Bottom)** "pmoles of Library Recovered" was determined by qPCR analysis with the KAPA kit and QPCR CopyCount™ software. "pmoles of Library used" is the quantity of library required to equal a total 0.04 pmoles for clustering. "Percentage of total" is pmoles of library used as a fraction of pmoles of library recovered. "Number of molecules" is the number of molecules loaded onto the flow cell for a four-plex MiSeq run. "Expected number of clusters" is based on cluster quality for previous data from a four-plex MiSeq run. "Clusters as a percentage of molecules" is 2,000,000 as a percentage of the total molecules per sample. "Number of surviving TBA molecules" is the maximum number of TBA counts expected to be seen at 2,000,000 clusters with 100% recovery of the initial 560,000 copies of TBA in the 1 nmol pool of library.

Table 6.7. Statistical analysis of sequencing data for repeat of HCHO cross-linking the m=15 DNA tailed library to thrombin with varied reaction times

Run	Total Number Clusters	Number of Clusters PF	Total Reads in Data Files
Samples 2.1-4.1	13,525,058	10,738,484	7,712,596

Experiment	1 pmol thb 30 sec.	1 pmol thb 3 min.	1 pmol thb 30 min.
Quantity of Protein	1 pmol	1 pmol	1 pmol
Quantity of Library	1 nmol	1 nmol	1 nmol
Starting Copies per Sequence	560,000	560,000	560,000
Total Reads	2,815,007	2,053,062	1,944,873
Total Good Reads	186,221	391,049	382,900
Background/Noise Threshold	3	3	3
Sequences Above Background	4	4	8
Expected Counts of TBA	400	448	451
Observed Counts of TBA[a]	1_28	3_4	1_272
Frequency of TBA per Good Reads[b]	0.015%	0.0010%	0.071%
Frequency of TBA per Total Reads[c]	0.00099%	0.00019%	0.014%

(Top) Total number of clusters produced per sequencing run. The number of clusters that passed filter (PF) is approximately 79%. After demultiplexing, the total reads are approximately 57% of the original clusters. **(Bottom)** Total number of reads for each sample generated by Illumina's BaseSpace. A total of 6,382 reads contained Index 1 of the positive control size marker. The number of good reads was generated by the Perl Script. The expected number of TBA counts was calculated based on the quantity of library recovered, the quantity of library applied to the flow cell and the number of good reads assuming a 100% recovery of the original 560,000 copies of TBA. **(a)** The rank *(left)* and count *(right)* are separated by an underscore. **(b)** Counts of TBA divided by the total good reads. **(c)** Counts of TBA divided by the total reads.

Table 6.8 Top 10 sequences for repeat of HCHO cross- linking the m=15 DNA tailed library to thrombin with varied reaction times

Sample 2.1: 1 pmol thb, 30 sec.		
Sequence	**Count**	**Identifier**
GGTTGGTGTGGTTGG	28	TBA
CACAGATCGGAAGAG	17	Jump*
GGTGGTTGTTGTGGT	7	TBAsc
GTCTGATCGGAAGAG	7	Jump*
GTGTGCAGTCGAGTT	3	
GTAGCATCTGGTCGA	3	
TTCTATTGGTCATAT	3	
TTAAGAGTCACGCTC	3	
ATTATAATTAGGTTC	3	
GAGCTAATTAGATAA	3	

Sample 3.1: 1 pmol thb, 3 min.		
Sequence	**Count**	**Identifier**
CACAGATCGGAAGAG	36	Jump*
GTCTGATCGGAAGAG	24	Jump*
GGTTGGTGTGGTTGG	4	TBA
TTTGTGCGTACTGTA	4	
TCTATGCACAAATCT	3	
GTATCGGAGCTCTAG	3	
ATCCCGGGGAGTCTG	3	
ATAACGCTGATTAGC	3	
ACCGCGCTTGGAAGA	3	
TCAGCAAATCGGCGA	3	

Sample 4.1: 1 pmol thb, 30 min.		
Sequence	**Count**	**Identifier**
GGTTGGTGTGGTTGG	272	TBA
CACAGATCGGAAGAG	24	Jump*
GTCTGATCGGAAGAG	11	Jump*
GGTTGGTGTTGTTGG	6	Ia
GGTTGGTGTGGTTTG	6	Ia
GGTTGGTTTGGTTGG	6	Ia
GGTTGTTGTGGTTGG	5	Ia
ATGAGGATTAGCGGT	4	
ACATAAGTAGGGGAC	3	
TGGGGTGCGATACAC	3	

Top 10 sequences and corresponding counts for Samples 2.1-4.1. The thrombin binding aptamer (TBA) was identified at counts of 28, 4 and 272, respectively. Jump sequence variants are present in all three samples and account for the majority of non-TBA variants above background. The background for all four experiments is 3. Variants of TBA occur in sample 4.1 as the 4^{th}-7^{th} ranked sequences. The TBAsc sequence is found in sample 1 as the 4^{rd} ranked sequence. It is possible that this is the result on contamination during sample preparation.

*Jump sequence variants.

Chapter 7: Discussion and Conclusions

The purpose of the work described in this thesis was to expand upon the capabilities of AIA to develop a fast, consistent, high throughput and universal aptamer discovery approach. In its infancy, AIA offered a simple and fast method for aptamer discovery that circumvented the cycling bottleneck of SELEX by employing over-represented libraries and high throughput sequencing. The hairpin loop library structure allowed for the isolation of the canonical thrombin binding aptamer as well as several lower affinity binding sequences in one round of selection. An AIA experiment could be completed in 3-6 days, including high throughput sequencing and data analysis, a marked improvement from the weeks to months required for a single SELEX experiment. During preliminary attempts to replicate the results of the original, published AIA results and explore the effects of target to library ratio, inconsistencies in partitioning efficiency, sample preparation and sequencing data quality were encountered. This work successfully expanded the applicability of AIA by (1) remedying sample preparation and sequencing data quality inconsistencies through a series of troubleshooting experiments and procedural modifications. TBA and dozens of variants were successfully identified above background for two variable ratio experiments that demonstrated a correlation between observed frequency of TBA/TBA variants and the degree of over-representation of the starting pool of library. (2) A 2'-OMe RNA/DNA chimera hairpin loop library was used to successfully identify TBA and multiple variants above background. This demonstrated a superior method for incorporation of 2'-OMe RNA in aptamer selection that does not require regeneration of 2'-OMe RNA during selection or risk altering the aptamer's functionality with post-selection modifications. (3) Implementing indexed adapters with sequence "barcodes" that allow

multiplexing of multiple ligated samples in a single flow cell lane which increased sample throughput and lowered costs. (4) Development of an efficient method to capture the minimally flanked, tailed library utilizing adapter splints with partially randomized splints and a 3'-amino modifier to prevent self-ligation. (5) Demonstrated the ability to predict the maximum frequency for a sequence in amplification-free AIA based on the quantity of recovered library as determined by qPCR, the degree of over-representation of the starting pool of library, and the number of clusters generated during high throughput sequencing. (6) Development of a novel partitioning method utilizing reversible formaldehyde cross-linking in conjunction with EMSA. The method was used to successfully identify TBA above background at varying frequencies for multiple reaction conditions. This method allows for partitioning in solution without the requirement of protein immobilization, circumventing the inconsistencies encountered with bead based partitioning. (7) Effectively improved high throughput sequencing data and distribution of indices during multiplexing with absolute quantification with qPCR CopyCount^{TM} software. The software exhibited a higher level of accuracy than the KAPA Library Quantification kit while eliminating the requirement of a standard curve. Ultimately, this directly reduces labor demands and cost while minimizing the potential need for costly resequencing.

Recommendations for Future Work

The sample preparation techniques that evolved over the course of this work offer superior control and predictability in the outcome of high throughput sequencing data. This affords the researcher great flexibility in experimentation with partitioning techniques and modified

libraries. The following are suggestions for future work that would expand the applicability of AIA further.

ForteBio Octet RED96 as a partitioning method

The ForteBio Octet RED96 interaction analysis system is typically used to quantify proteins and small molecules and analyze the kinetics of molecular interactions. The Octet platform boasts advantages such as label-free detection, real-time interaction data, accurate and reproducible results, a non-microfluidic design and simple user interface. The Octet RED96 specifically offers increased sensitivity, enabling the detection of low molecular weight molecules,[181] ideal for use with the tailed libraries of approximately 7 kDa. Two types of biosensors were used during the preliminary experiments considering the Octet RED96 as a partitioning method: Streptavidin (SA) and Amine Reactive Second Generation (AR2G). The streptavidin (SA) biosensors are coated at the tip with streptavidin and are designed for immobilization of biotin labeled proteins for studying protein: protein interactions.[182] The SA biosensors were used as the solid surface for immobilization of biotinylated thrombin for partitioning against multiple library formats. Also, biotinylated Con-A was immobilized on the SA biosensors and used as the solid support for partitioning experiments against native thrombin. The AR2G biosensors are coated at the tip with a carboxy-terminated ligand. An EDC-catalyzed amide bond formation between the carboxylic acids on the biosensor and reactive amines of the protein surface creates a covalent linkage.[183] The protein immobilization is irreversible, allowing for freedom in screening conditions. Preliminary experiments found that the AR2G biosensors may offer a more reliable method of protein immobilization for partitioning over the SA biosensors due to changes

in binding affinity of thrombin following biotinylation. The binding capacity of the AR2G biosensors is on the order of fmoles, which is low for traditional SELEX and for some past AIA experiments (typically tens or hundreds of pmoles of protein). However, aptamers have been successfully identified using fmole quantities of protein in SELEX.[84] If the AR2G biosensors are successful at identifying TBA from a library pool partitioned against immobilized thrombin, the AIA method could see great improvements from their implementation.

In preliminary experiments, the AR2G biosensors were adapted for AIA by immobilizing native thrombin and screening against the m=15 DNA tailed library. The AR2G biosensors were hydrated in dH_2O for 15 minutes prior to the experiment. The instrument utilizes 96-well plates with 200 µl reagent volumes; the biosensors are "dipped" into successive wells across the plate. **Figure 7.1** illustrates the experimental plate set-up and assay steps list. Chemistry in columns 1-5 occurs at a plate rotation of 1,000 rpm. Chemistry in columns 6 and 7 occurs at a plate rotation of 60 rpm. Baseline is achieved in column 1, dH_2O, for 120 seconds. The sensors were activated in a solution of 20 nM EDC (1-Ethyl-3-[3-dimethylsminopropyl] carbodiimide hydrochloride) and 10 nM NHS (N-hydroxysulfosuccinimide) for 300 seconds in column 2. Thrombin in 10 mM acetate buffer, pH 6.0, was loaded onto the activated sensors at 0.5 µM for sensors B-H for 900 seconds in column 3, corresponding to the period marked **A** in **Figure 7.2**. The 0.5 µM concentration of thrombin was determined from preliminary experiments with the AR2G biosensors to be a concentration for sufficient binding. The reactive sites on the sensors were quenched in 1M ethanolamine, pH 8.5, for 300 seconds in column 4, corresponding to the period immediately following the large increase in mass in period **A**. The next two periods between the dashed lines in **Figure 7.2** correspond to equilibration in the AIA partitioning buffer (20 mM

Tris-HCl, 140 mM NaCl, 5 mM KCl, 1 mM $CaCl_2$, 1 mM $MgCl_2$, pH 7.4) for 120 seconds, followed by another 120 seconds in column 6 also containing the partitioning buffer. Period **B** in the figure corresponds to the association of the analyte (m=15 DNA tailed library) in partitioning buffer for 600 seconds in column 7. The library was associated at 5.0 µM, 2.5 µM, and 0.5 µM in duplicate. For the 200 µl used in each well, this corresponds to 1 nmol, 500 pmol and 100 pmol of library, respectfully. This translates to approximately 560,000 copies, 280,000 copies, and 56,000 copies of any unique sequence, respectively. Dissociation in partitioning buffer, column 8, for 600 seconds washed off any unbound sequences in period **C**. The association of thrombin in period **A** produces a large increase in signal due to the large molar mass of thrombin (36.7 kDa). Little to no association is visualized in period **B**; this is thought to be a result of the small size of the library in comparison to thrombin. Typical experiments performed on this instrument employ analytes much larger in size than the immobilized target, where an association curve would be apparent. The sensors were retrieved from the instrument and resuspended in 10 µl dH_2O. There was no discernible difference in the quantity of recovered library between phenol/chloroform extracted samples and those ligated directly off of the biosensors. The libraries were ligated with sequencing adapters directly off of the biosensors, agarose gel purified and quantified with the KAPA Library Quantification kit and qPCR CopyCount™ as described in chapter 6. A single sample from this experiment, sensor B, was sequenced on the Illumina MiSeq and loaded according to the qPCR CopyCount™ determination. Based on the recovered quantity of library, 100% recovery of the ~560,000 copies of TBA in the starting pool and the number of clusters generated, the relative maximum frequency of TBA was expected to be 1,329. TBA was not found in the data set and the only sequences to occur above background were jump sequence variants. The Octet RED96 was also

used to partition the m=15 DNA and m=15 2'-OMe RNA/DNA chimera adapter libraries previous to this experiment and TBA was not isolated in these experiments.

Only a small number of experiments were performed with the Octet RED96 as a partitioning method and there are a number of conditions that would require optimization before it could be disqualified as a partitioning method. Most importantly is optimal loading of the protein target to avoid over-crowding the biosensor while maintaining sufficient target to capture aptamers. One benefit of the Octet RED96 is the consistency in the quantity of recovered library. With a binding capacity on the order of fmoles, the quantity of recovered library ranged from 0.0002 – 0.002 pmoles for numerous experiments. This is well below the desired 0.01 pmoles of a 4-plex sequencing run. If an aptamer sequence was efficiently selected using this method, it would appear well above background. Additionally, this would allow for multiplexed sequencing runs with >4 indices, increasing sample throughput.

The Minimal Primer (MP) method

The tailed libraries described in chapters 5 and 6 are flanked by four fixed bases on each side, reducing the interaction between the randomized library region and the fixed regions necessary for primer annealing. It may be advantageous to apply the concept of minimal fixed regions of over-represented libraries used in AIA to the longer, under-represented libraries traditionally used in SELEX. For under-represented libraries with copy numbers <1,000, it is necessary to perform multiple rounds of partitioning. To accomplish this while retaining minimal fixed regions, restriction enzyme digestion can be used to regenerate the library. The Minimal Primer

(MP) method, as discussed in chapter 1, reduces the length of the fixed regions to 2 nucleotides on each side of the library.[124] The proposed MP protocol was adapted with the AIA ligation scheme and is outlined in **Figure 7.3**. The protocol utilizes advancements in the ligation scheme including a four base randomized overhang that improves capture and a 3'-NH_2 on the Adapter 2 complement to prevent self-ligation of the Adapter 2/Adapter 2 Complement complex.

A preliminary experiment illustrates the ligation efficiency of the MP library design using a 40-mer "library" containing the 15-mer TBA sequence within the randomized region, flanked by two nucleotide non-complementary tails. A 50 pmol aliquot of the 44-mer was ligated to Adapters MP1a, 1b, 2a, and 2b at 20% excess concentration with conditions as described in chapter 6. Results following incubation of the 20 µl reaction for 24 and 48 hours at 4°C are shown in **Figure 7.4** After 24 hours, both a single (~50 bp) and double (~80 bp) ligated product are visible, with an estimated 70-80% capture of the double ligated product. After 48 hours, the intensity of the double ligated product increases to an estimated >90% capture, while the intensity of the single ligated product decreases. The overall intensity of free adapters decreases after 48 hours compared to 24 hours. It was found that a 24 hour ligation yields a visibly equivalent percentage of ligated products when additional adapters, buffer and enzyme are added after the initial 12 hour incubation. This would reduce the time commitment of a 48 hour incubation.

After ligation, PCR primers are used to amplify the ligated library. Restriction enzyme digestion would regenerate the next generation MP library. This method would allow for the cycling

required by under-represented libraries while maintaining the minimally flanked structure of the tailed libraries used in AIA. The library could be prepared for sequencing in two manners: ligation of the full length Illumina adapters in place of MP adapters (this would require no additional PCR after the final partitioning) or ligation with the MP adapters and addition of the critical Illumina adapter sequences by PCR. In keeping with the goals of AIA, simulating the evolution that occurs in multiple rounds of SELEX as a result of errors in PCR amplification would shorten the aptamer discovery process, saving time and money. It is often the case that identified aptamer sequences for under-represented libraries are not present in the initial pool, but evolved due to amplification errors in PCR.[147] This could be accomplished with aggressive chemical mutagenesis, which has been used at low rates in prior studies.[184, 185] Aggressively mutating 1-50% of the library molecules would introduce sequences that were not present in the starting pool, but are relatives of sequences selected in previous rounds. Reversible formaldehyde cross-linking and EMSA partitioning may be ideal for the MP method because it may be possible to observe the convergence of the library pool to high affinity binding sequences via EMSA.

Additional suggestions

Additional experiments to be considered include the introduction of new protein targets to the reversible HCHO cross-linking and EMSA partitioning method. Specifically, partitioning of an m=40 (and possibly, m=60) DNA library in the MP format against *E. coli* cellular lysates via HCHO cross-linking. Additionally, partitioning tailed DNA and 2'-OMe or 2'-F RNA/DNA

chimera libraries against the four Epigenetic protein targets described in chapter 3 with HCHO cross-linking is suggested.

A final recommendation incorporates the high throughput sequencing utilized by AIA with post-sequencing, protein-binding studies. The GAIIx used for sequencing in chapter 2 and 3 would be transitioned for protein-imaging in order to rapidly identify and eliminate aptamer leads that cross-react with off-target proteins. This is specifically useful for libraries selected against mixtures of proteins, including cellular lysates. After clustering and sequencing, restriction enzyme digestion and NaOH denaturation would generate ssDNA features on the flow cell. The flow cell would be washed with binding buffer and subsequently screened with rhodamine 6G labelled thrombin in proof of concept experiments. Software developed at MIT for a similar application[186] would be used to capture images of protein bound DNA and identify the sequences with high affinity for the protein. This method may circumvent secondary screening techniques used during the assessment of aptamer candidates by using the sequenced flow cell as a microarray.

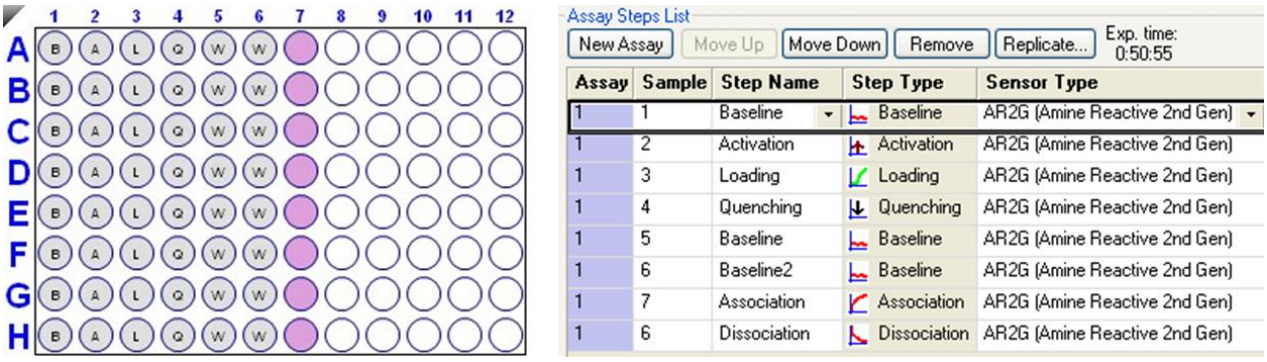


Figure 7.1. Octet RED96 experimental set-up

(left) 96-well experimental plate. Rows A-H correspond to sensors while columns 1-12 correspond to assay steps. Column 1: Baseline in dH_2O. Column 2: Activation in EDC and NHS. Column 3: Loading of thrombin. Column 4: Quenching with ethanolamine. Column 5: Wash in AIA partitioning buffer. Column 6: Baseline in partitioning Buffer and dissociation. Column 7: Association of analyte. **(right)** Assay steps list.

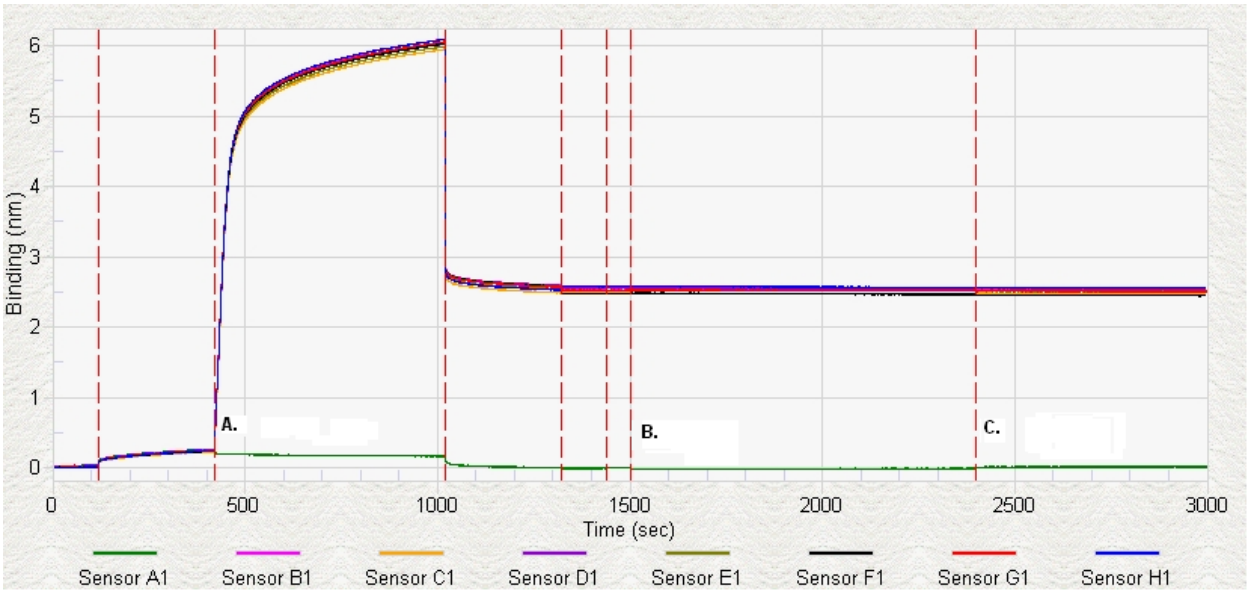


Figure 7.2 Binding curve graph of m=15 tailed library against immobilized thrombin

A. Period for loading of thrombin on sensors B-H: 0.5 µM, sensor A: 0.0 µM. **B.** Period for association of m=15 DNA tailed library: Sensors B-C: 5.0 µM library, sensors D-E: 2.5 µM, sensors F-G: 0.5 µM library, sensor H: 0.0 µM library. **C.** Period for dissociation of complex in partitioning buffer. Baseline drift is observed in green, sensor A, 0.0 µM thrombin and 0.0 µM m=15 DNA tailed library.

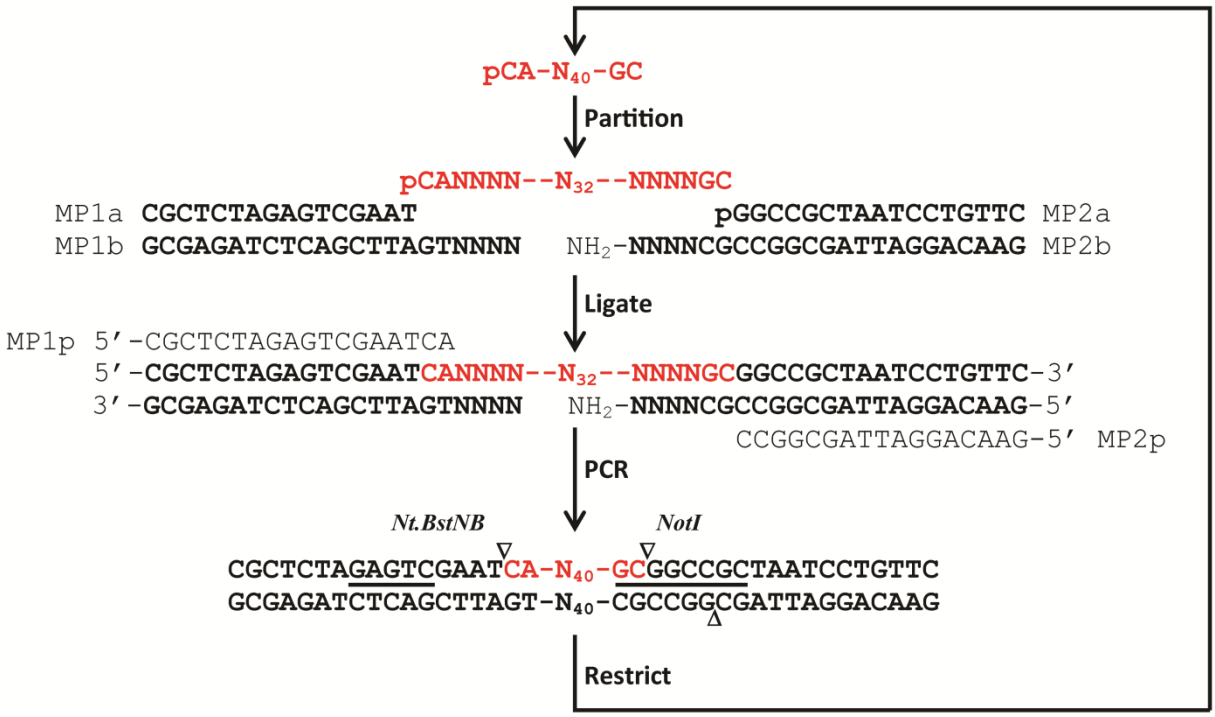


Figure 7.3 Minimal Primer (MP) cycling

The m=40 library (red), flanked by two-nucleotide fixed sequences, is partitioned against a protein target. Surviving sequences are captured by the adapters MP1a, 1b, 2a, and 2b. The 6 nucleotide overhangs contain a 4 nucleotide randomized region to improve annealing efficiency. The NH_2- at the 3′-terminus of MP2b prevents self-ligation of adapters. After ligation, PCR primers MP1p and MP2p are used to amplify the ligated library. Restriction enzyme NotI makes two staggered cuts (triangles) and Nt.BstNB cuts once; enzyme recognition sequences are underlined. The regenerated library can then be used for the next round of partitioning.

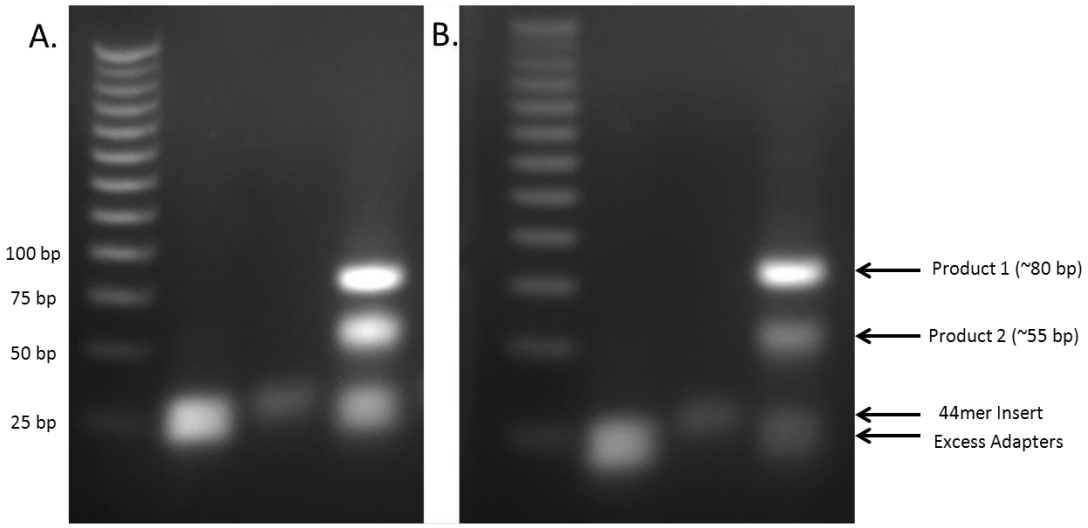


Figure 7.4 Ligation of MP 44-mer

4% Agarose gel stained with Ethidium Bromide. Lane 1: 25 bp DNA ladder. Lane 2: 60 pmol each of Adapters MP1a, 1b, 2a, 2b. Lane 3: 50 pmol of 44-mer library/TBA insert. Lane 4: 50 pmol of 44-mer library/TBA ligated with 60 pmol of Adapters. **(A.)** 24 hour ligation **(B.)** 48 hour ligation.

Appendix 2

DNA Sequences (All sequences written 5' to 3')

m=15 DNA library (39 bases)
Phos/ACACGCGCATGC-m15-GCATGCGCCACA

Adapter 1 (33 bases)
ACACTCTTTCCCTACACGACGCTCTTCCGATCT

Adapter 1 Complement (45 bases)
GCATGCGCGTGTAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT

Adapter 2 (33 bases)
Phos/GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG

Adapter 2 Complement (45 bases)
CAAGCAGAAGACGGCATACGAGCTCTTCCGATCTGTGGCGCATGC

PCR Forward Primer (58 bases)
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT

PCR Reverse Primer (34 bases)
CAAGCAGAAGACGGCATACGAGCTCTTCCGATCT

SL1 Primer (21 bases)
CAGTCCAGTTACGCTGGAGTC

KAPA Primer 1 (20 bases)
AATGATACGGCGACCACCGA

KAPA Primer 2 (21 bases)
CAAGCAGAAGACGCCATACGA

Sequencing Primer (33 bases)
ACACTCTTTCCCTACACGACGCTCTTCCGATCT

Illumina P5 Region, SR Flow Cell (20 bases)
AATGATACGGCGACCACCGA

Illumina P7 Region, SR Flow Cell (21 bases)
CAAGCAGAAGACGGCATACGA

Table 2.6 60:1, m=19 DNA for thrombin

Rank	Sequence	Count
1	CCACAGATCGGAAGAGCTC	283
2	GCCACAGATCGGAAGAGCT	129
3	GCACAGATCGGAAGAGCTC	68
4	GCCACAGATCGGAGAGCTC	38
5	GCCACAGATCGAAGAGCTC	33
6	GCCAAGATCGGAAGAGCTC	27
7	ACACAGATCGGAAGAGCTC	24
8	GCCACAGATCGGAAGGCTC	19
9	GCCCAGATCGGAAGAGCTC	17
10	GCCACAGATCGGAAGACTC	15
11	GCCACGATCGGAAGAGCTC	14
12	GCCACAGATCGGAAAGCTC	12
13	GCCACAGACGGAAGAGCTC	12
14	GCCACAGATCGGAAGAGTC	11
15	TCACAGATCGGAAGAGCTC	9
16	CCACAGATCGGCAGAGCTC	7
17	GCCACAGATCGGCAGAGCT	7
18	GCCACAGATCGGAAGAGCC	7
19	GCCACAGATCGGAAGAGAT	6
20	CACCAGATCGGAAGAGCTC	6

Top twenty sequences from the 60:1, thrombin: m=19 DNA partitioning. A total of 4,644,765 good reads out of a total 5,932,215 reads. A total of 4,640,900 unique sequences occur in the file. A total of 38 of the top 50 sequences are variations on the jump sequence, including all of the top 20. The background threshold was identified as 2 counts. 40 unique sequences occur above background.

Annotated Perl Script

Authored by Dr. Huitao Sheng, annotated by Dr. Damian Allis. The Perl Script is included for readers who may wish to use it. User defined requirements in **BOLD**. All text has been left justified.

```
#!/usr/bin/perl -w
# ---- USER DEFINED ----
#my $fullhead='ACACGCGCATGC';
#my $fulltail='GCATGCGCCACA';
# NN for multiplexing
my $fullhead='ACACGCGCATGC';
my $fulltail='GCATGCGCCACA';
my $nmerlen=15; # length of nmers that we are looking for
# the location of the two multiplexing bases in head sequence
# index starting from 0
my $mp1 = 0;
my $mp2 = 0;
# query length
my $longQueryLength = 8;
my $shortQueryLength = 4;
# penalty and award
my $gapScore = -2;
my $mismatchScore = -1;
my $matchScore = 1;
# threshold: no more than 1/4 mismatches or gaps
my $minLongScore = $longQueryLength / 2;
my $minShortScore = $shortQueryLength / 2;
my $seqId = "S1_L001_R1_001";
my $readsfile = $seqId . ".fastq"; # input fasta file of reads
my $taskId = $seqId . "_"
. $longQueryLength . "_" . $nmerlen. "_" . $shortQueryLength;
my $goodReads = $taskId . "_goodReads.txt";
my $badReads = $taskId . "_badReads.txt";
my $candidateReads = $taskId . "_candidateReads.txt";
my $nmerCount = $taskId . "_nmerCount.txt";
my $locSta = $taskId . "_locationStatistics.txt";
# ----------------------
# relative location
$mp1 -= length($fullhead) - $longQueryLength;
$mp2 -= length($fullhead) - $longQueryLength;
open(GOOD, ">$goodReads");
open(BAD, ">$badReads");
open(CAND, ">$candidateReads");
open(CUNT, ">$nmerCount");
open(LST, ">$locSta");
```

```
open(FILE,$readsfile) or die("Cannot open!");
# looking for {junk}{long head}{N15}{short tail}
my $longHead = substr($fullhead,
length($fullhead) - $longQueryLength, $longQueryLength);
my @longh = split(//, $longHead);
my $shortTail = substr($fulltail, 0, $shortQueryLength);
my @shortt = split(//, $shortTail);
# looking for {short head}{N15}{long tail}{junk}
my $shortHead = substr($fullhead,
length($fullhead) - $shortQueryLength, $shortQueryLength);
my @shorth = split(//, $shortHead);
my $longTail = substr($fulltail, 0, $longQueryLength);
my @longt = split(//, $longTail);
my %mers=();
my $BASE_A = 0;
my $BASE_C = 1;
my $BASE_T = 2;
my $BASE_G = 3;
my $BASE_N = 4;
my $DEL = 5;
my $INS = 6;
my @headBaseCnt = ();
for(my $bIdx = 0; $bIdx < length($longHead); $bIdx++) {
for(my $cIdx = 0; $cIdx < 7; $cIdx++) {
$headBaseCnt[$cIdx][$bIdx] = 0;
}
}
my @tailBaseCnt = ();
for(my $bIdx = 0; $bIdx < length($shortTail); $bIdx++) {
for(my $cIdx = 0; $cIdx < 7; $cIdx++) {
$tailBaseCnt[$cIdx][$bIdx] = 0;
}
}
while (<FILE>) {
chomp;next if(/[^ACTGNactgn]/);
next if(/AAAA/);
my $read = $_;
my $matchLongHead = substr($read,
0, (length($read) - $shortQueryLength - $nmerlen));
my @mlh = split(//, $matchLongHead);
my @lhIdx = ();
my @mlhIdx = ();
my $mlhAlg = "";
my $longHeadScore = &DP_ALIGN(\@longh, \@mlh,
\@lhIdx, \@mlhIdx, \$mlhAlg);
my $lhst = "";
```

```
my $lhstScore = $gapScore * length($read);
if($longHeadScore > $minLongScore) {
my $headMatchIdx = $#lhIdx;
my $readHeadIdx = $lhIdx[$headMatchIdx];
while(($readHeadIdx == -1) &&($headMatchIdx >= 0)) {
$headMatchIdx--;
$readHeadIdx = $lhIdx[$headMatchIdx];
}
if($headMatchIdx >= 0) {
my $matchShortTail = substr($read,
($readHeadIdx + $nmerlen - 1));
my @mst = split(//, $matchShortTail);
my @stIdx = ();
my @mstIdx = ();
my $mstAlg = "";
my $shortTailScore = &DP_ALIGN(\@shortt, \@mst,
\@stIdx, \@mstIdx, \$mstAlg);
if($shortTailScore > $minShortScore) {
my $tailMatchIdx = 0;
my $readTailIdx = $stIdx[$tailMatchIdx];
while(($readTailIdx == -1)
&& ($tailMatchIdx <= $#stIdx)) {
$tailMatchIdx++;
$readTailIdx = $stIdx[$tailMatchIdx];
}
if($tailMatchIdx <= $#stIdx) {
$lhstScore =
$longHeadScore + $shortTailScore;
$lhst = substr($read,
($readHeadIdx + 1),
($readTailIdx +$nmerlen - 2));
if(length($lhst) == $nmerlen) {
my $headMatchNum = 0;
my $headMisNum = 0;
my $headDelNum = 0;
my $headInsNum = 0;
for(my $bIdx = 0; $bIdx < length($longHead); $bIdx++) {
if($lhIdx[$bIdx] == -1) {
$headBaseCnt[$DEL][$bIdx]++;
}
elsif($mlh[$lhIdx[$bIdx]] eq "A") {
$headBaseCnt[$BASE_A][$bIdx]++;}
elsif($mlh[$lhIdx[$bIdx]] eq "C") {
$headBaseCnt[$BASE_C][$bIdx]++;
}
elsif($mlh[$lhIdx[$bIdx]] eq "T") {
```

```
$headBaseCnt[$BASE_T][$bIdx]++;
}
elsif($mlh[$lhIdx[$bIdx]] eq "G") {
$headBaseCnt[$BASE_G][$bIdx]++;
}
elsif($mlh[$lhIdx[$bIdx]] eq "N") {
$headBaseCnt[$BASE_N][$bIdx]++;
}
# an insertion happened before this location
if(($bIdx > 0) && (($lhIdx[$bIdx] - $lhIdx[$bIdx - 1]) > 1) && ($lhIdx[$bIdx - 1] != -1)) {
$headBaseCnt[$INS][$bIdx]++;
$headInsNum++;
}
if($lhIdx[$bIdx] == -1) {
$headDelNum++;
}
elsif(($mlh[$lhIdx[$bIdx]] eq $longh[$bIdx])
|| ($mlh[$lhIdx[$bIdx]] eq "N")
|| ($longh[$bIdx] eq "N")) {
$headMatchNum++;
}
else {
$headMisNum++;
}
}
my $tailMatchNum = 0;
my $tailMisNum = 0;
my $tailDelNum = 0;
my $tailInsNum = 0;
for(my $bIdx = 0; $bIdx < length($shortTail); $bIdx++) {
if($stIdx[$bIdx] == -1) {
$tailBaseCnt[$DEL][$bIdx]++;
}
elsif($mst[$stIdx[$bIdx]] eq "A") {
$tailBaseCnt[$BASE_A][$bIdx]++;
}
elsif($mst[$stIdx[$bIdx]] eq "C") {
$tailBaseCnt[$BASE_C][$bIdx]++;
}
elsif($mst[$stIdx[$bIdx]] eq "T") {
$tailBaseCnt[$BASE_T][$bIdx]++;
}
elsif($mst[$stIdx[$bIdx]] eq "G") {
$tailBaseCnt[$BASE_G][$bIdx]++;
}
elsif($mst[$stIdx[$bIdx]] eq "N") {
```

```
$tailBaseCnt[$BASE_N][$bIdx]++;
}
# an insertion happened before this location
if(($bIdx > 0) && (($stIdx[$bIdx] - $stIdx[$bIdx - 1]) > 1) && ($stIdx[$bIdx - 1] != -1)) {
$tailBaseCnt[$INS][$bIdx]++;
$tailInsNum++;
}
if($stIdx[$bIdx] == -1) {
$tailDelNum++;
}
elsif(($mst[$stIdx[$bIdx]] eq $shortt[$bIdx])
|| ($mst[$stIdx[$bIdx]] eq "N")
|| ($shortt[$bIdx] eq "N")) {
$tailMatchNum++;
}
else {
$tailMisNum++;
}
}
print GOOD $read;
print GOOD "\t";
&PRNMATCH(GOOD, \@lhIdx, \@mlh);
print GOOD "\t";
print GOOD $headMatchNum;
print GOOD "\t";
print GOOD $headMisNum;
print GOOD "\t";
print GOOD $headDelNum;
print GOOD "\t";
print GOOD $headInsNum;
print GOOD "\t";
print GOOD $lhst;
print GOOD "\t";
print GOOD length($lhst);
print GOOD "\t";
&PRNMATCH(GOOD, \@stIdx, \@mst);
print GOOD "\t";

print GOOD $tailMatchNum;
print GOOD "\t";
print GOOD $tailMisNum;
print GOOD "\t";
print GOOD $tailDelNum;
print GOOD "\t";
print GOOD $tailInsNum;
print GOOD "\t";
```

```
print GOOD "\t"
. "$mlh[$lhIdx[$mp1]]"
. "$mlh[$lhIdx[$mp2]]"
. "\n";
if(defined($mers{$lhst})) {
$mers{$lhst}++;
}
else {
$mers{$lhst} = 1;
}
}
else {
print CAND $read;
print CAND "\t";
&PRNMATCH(CAND, \@lhIdx, \@mlh);
print CAND "\t";
print CAND $lhst;
print CAND "\t";
print CAND length($lhst);
print CAND "\t";
&PRNMATCH(CAND, \@stIdx, \@mst);
print CAND "\n";
}
next;
}
}
}
}
my $matchLongTail = substr($read, ($shortQueryLength + $nmerlen));
my @mlt = split(//, $matchLongTail);
my @ltIdx = ();
my @mltIdx = ();
my $mltAlg = "";
my $longTailScore = &DP_ALIGN(\@longt, \@mlt,
\@ltIdx, \@mltIdx, \$mltAlg);
my $shlt = "";
my $shltScore = $gapScore * length($read);
if($longTailScore > $minLongScore) {
my $tailMatchIdx = 0;
my $readTailIdx = $ltIdx[$tailMatchIdx];
while(($readTailIdx == -1) && ($tailMatchIdx <= $#ltIdx)) {
$tailMatchIdx++;
$readTailIdx = $ltIdx[$tailMatchIdx];
}
if($tailMatchIdx <= $#ltIdx) {
my $matchShortHead = substr($read,
```

```
0, ($shortQueryLength + $readTailIdx + 2));
my @msh = split(//, $matchShortHead);
my @shIdx = ();
my @mshIdx = ();
my $mshAlg = "";
my $shortHeadScore = &DP_ALIGN(\@shorth, \@msh,
\@shIdx, \@mshIdx, \$mshAlg);
if($shortHeadScore > $minShortScore) {
my $headMatchIdx = $#shIdx;
my $readHeadIdx = $shIdx[$headMatchIdx];
while(($readHeadIdx == -1)
&& ($headMatchIdx >= 0)) {
$headMatchIdx--;
$readHeadIdx = $shIdx[$headMatchIdx];
}
if($headMatchIdx >= 0) {
$shltScore =
$shortHeadScore + $longTailScore;
$shlt = substr($read,
($readHeadIdx + 1),
($shortQueryLength + $nmerlen + $readTailIdx - $readHeadIdx - 1));
print CAND $read;
print CAND "\t";
&PRNMATCH(CAND, \@shIdx, \@msh);
print CAND "\t";
print CAND $shlt;
print CAND "\t";
print CAND length($shlt);
print CAND "\t";
&PRNMATCH(CAND, \@ltIdx, \@mlt);
print CAND "\n";
next;
}
}
}
}
print BAD "$read\n";
}
foreach $key (sort {$mers{$b} <=> $mers{$a}} keys %mers) {
print CUNT "$key\t$mers{$key}\n";
}
for(my $bIdx = 0; $bIdx <= $#longh; $bIdx++) {
print LST "\t$longh[$bIdx]";
}
print LST "\n";
for(my $cIdx = 0; $cIdx < 7; $cIdx++) {
```

```
if($cIdx == $BASE_A) {
print LST "A";
}
elsif($cIdx == $BASE_C) {
print LST "C";
}
elsif($cIdx == $BASE_T) {
print LST "T";
}
elsif($cIdx == $BASE_G) {
print LST "G";
}
elsif($cIdx == $BASE_N) {
print LST "N";
}
elsif($cIdx == $DEL) {
print LST "DEL";
}
elsif($cIdx == $INS) {
print LST "INS(b4)";
}
for(my $bIdx = 0; $bIdx <= $#longh; $bIdx++) {
print LST "\t$headBaseCnt[$cIdx][$bIdx]"
}
print LST "\n";
}
print LST "\n";
for(my $bIdx = 0; $bIdx <= $#shortt; $bIdx++) {
print LST "\t$shortt[$bIdx]";
}
print LST "\n";
for(my $cIdx = 0; $cIdx < 7; $cIdx++) {
if($cIdx == $BASE_A) {
print LST "A";
}
elsif($cIdx == $BASE_C) {
print LST "C";
}
elsif($cIdx == $BASE_T) {
print LST "T";
}
elsif($cIdx == $BASE_G) {
print LST "G";
}
elsif($cIdx == $BASE_N) {
print LST "N";
```

```
}
elsif($cIdx == $DEL) {
print LST "DEL";
}
elsif($cIdx == $INS) {
print LST "INS(b4)";
}
for(my $bIdx = 0; $bIdx <= $#shortt; $bIdx++) {
print LST "\t$tailBaseCnt[$cIdx][$bIdx]"
}
print LST "\n";
}
close(GOOD);
close(BAD);
close(CAND);
close(FILE);
close(CUNT);
close(LST);
# dynamic programming sequence alignment
sub DP_ALIGN {
# input: the two sequences to be aligned
# array references
my $s1Ref = shift;
my $s2Ref = shift;
my $len1 = $#$s1Ref + 1;
my $len2 = $#$s2Ref + 1;
# output: the index of aligned bases for each input sequence
# array references
my $a1Ref = shift;
my $a2Ref = shift;
# -1 in index means a gap
my $gapIdx = -1;
for(my $sIdx = 0; $sIdx < $len1; $sIdx++) {
$$a1Ref[$sIdx] = $gapIdx;
}
for(my $sIdx = 0; $sIdx < $len2; $sIdx++) {
$$a2Ref[$sIdx] = $gapIdx;
}
my $alignRef = shift;
#       # penalty and award
#       my $gapScore = -2;
#       my $mismatchScore = -1;
#       my $matchScore = 1;
# initialize score talbe
my @scoreTable = ();
for(my $row = 0; $row <= $len1; $row++) {
```

```
for(my $col = 0; $col <= $len2; $col++) {
$scoreTable[$row][$col]= 0; # ignore leading gaps
}
}
# trace direction
my $fromNowhere = 0;
my $fromLeft = 1;
my $fromAbove = 2;
my $fromAboveLeft = 3;
# initialize trace table
my @traceTable = ();
$traceTable[0][0] = $fromNowhere;
for(my $col = 1; $col <= $len2; $col++) {
$traceTable[0][$col] = $fromLeft;
}
for(my $row = 1; $row <= $len1; $row++) {
$traceTable[$row][0] = $fromAbove;
for(my $col = 1; $col <= $len2; $col++) {
$traceTable[$row][$col] = $fromNowhere;
}
}
#fill out score table and traceback table
for(my $row = 1; $row <= $len1; $row++) {
for(my $col = 1; $col <= $len2; $col++) {
my $aboveScore =
$scoreTable[$row - 1][$col] + $gapScore;
my $leftScore = $scoreTable[$row][$col - 1] + $gapScore;
my $aboveLeftScore = $scoreTable[$row - 1][$col - 1];
if(&SAME_BASE($$s1Ref[$row - 1],
$$s2Ref[$col - 1]) == 1) {
$aboveLeftScore += $matchScore;
}
else {
$aboveLeftScore += $mismatchScore;
}
$scoreTable[$row][$col] = $aboveLeftScore;
$traceTable[$row][$col] = $fromAboveLeft;
if($aboveScore > $scoreTable[$row][$col]) {
$scoreTable[$row][$col] = $aboveScore;
$traceTable[$row][$col] = $fromAbove;
}
if($leftScore > $scoreTable[$row][$col]) {
$scoreTable[$row][$col] = $leftScore;
$traceTable[$row][$col] = $fromLeft;
}
}
```

```
}
# ignore tailing gaps
my $maxCol = $scoreTable[0][$len2];
my $maxRowIdx = 0;
for(my $row = 1; $row <= $len1; $row++) {
if($scoreTable[$row][$len2] > $maxCol) {
$maxCol = $scoreTable[$row][$len2];
$maxRowIdx =$row;
}
}
my $maxRow = $scoreTable[$len1][0];
my $maxColIdx = 0;
for(my $col = 1; $col <= $len2; $col++) {
if($scoreTable[$len1][$col] > $maxRow) {
$maxRow = $scoreTable[$len1][$col];
$maxColIdx = $col;
}
}
if($maxRow > $maxCol) {
for(my $col = $len2; $col > $maxColIdx; $col--) {
$traceTable[$len1][$col] = $fromLeft;
}
$scoreTable[$len1][$len2] = $maxRow;
}
else {
for(my $row = $len1; $row > $maxRowIdx; $row--) {
$traceTable[$row][$len2] = $fromAbove;
}
$scoreTable[$len1][$len2] = $maxCol;
}
# get output
my $row = $len1;
my $col = $len2;
my $as1 = "";
my $as2 = "";
while($traceTable[$row][$col] != $fromNowhere) {
if($traceTable[$row][$col] == $fromAboveLeft) {
$$a1Ref[$row - 1] = ($col - 1);
$$a2Ref[$col - 1] = ($row - 1);
$as1 = $$s1Ref[$row - 1] . $as1;
$as2 = $$s2Ref[$col - 1] . $as2;
$row--;
$col--;
}
elsif($traceTable[$row][$col] == $fromAbove) {
$$a1Ref[$row - 1] = -1;
```

```
$as1 = $$s1Ref[$row - 1] . $as1;
$as2 = '-' . $as2;
$row--;
}
elsif($traceTable[$row][$col] == $fromLeft) {
$$a2Ref[$col - 1] = -1;
$as1 = '-' . $as1;
$as2 = $$s2Ref[$col - 1] . $as2;
$col--;
}
}
# test output
#	print "Score table:\n";
#	for(my $row = 0; $row <= $len1; $row++) {
#		for(my $col = 0; $col <= $len2; $col++) {
#			print "\t$scoreTable[$row][$col]";
#		}
#		print "\n";
#	}
#
#	print "Trace table:\n";
#	for(my $row = 0; $row <= $len1; $row++) {
#		for(my $col = 0; $col <= $len2; $col++) {
#			print "\t$traceTable[$row][$col]";
#		}
#		print "\n";
#	}
$$alignRef = $as1 . "\n" . $as2 . "\n";
return $scoreTable[$len1][$len2];
}
# dynamic programming sequence alignment
sub DP_ALIGN_4H {
# input: the two sequences to be aligned
# array references
my $s1Ref = shift;
my $s2Ref = shift;

my $len1 = $#$s1Ref + 1;
my $len2 = $#$s2Ref + 1;
# output: the index of aligned bases for each input sequence
# array references
my $a1Ref = shift;
my $a2Ref = shift;
# -1 in index means a gap
my $gapIdx = -1;
for(my $sIdx = 0; $sIdx < $len1; $sIdx++) {
```

```
$$a1Ref[$sIdx] = $gapIdx;
}
for(my $sIdx = 0; $sIdx < $len2; $sIdx++) {
$$a2Ref[$sIdx] = $gapIdx;
}
my $alignRef = shift;
#       # penalty and award
#       my $gapScore = -2;
#       my $mismatchScore = -1;
#       my $matchScore = 1;
# initialize score talbe
my @scoreTable = ();
for(my $row = 0; $row <= $len1; $row++) {
for(my $col = 0; $col <= $len2; $col++) {
$scoreTable[$row][$col]= 0; # ignore leading gaps
}
}
# trace direction
my $fromNowhere = 0;
my $fromLeft = 1;
my $fromAbove = 2;
my $fromAboveLeft = 3;
# initialize trace table
my @traceTable = ();
$traceTable[0][0] = $fromNowhere;
for(my $col = 1; $col <= $len2; $col++) {
$traceTable[0][$col] = $fromLeft;
}
for(my $row = 1; $row <= $len1; $row++) {
$traceTable[$row][0] = $fromAbove;
for(my $col = 1; $col <= $len2; $col++) {
$traceTable[$row][$col] = $fromNowhere;
}
}
#fill out score table and traceback table
for(my $row = 1; $row <= $len1; $row++) {
for(my $col = 1; $col <= $len2; $col++) {
my $aboveScore =
$scoreTable[$row - 1][$col] + $gapScore;
my $leftScore = $scoreTable[$row][$col - 1] + $gapScore;
my $aboveLeftScore = $scoreTable[$row - 1][$col - 1];
if(&SAME_BASE($$s1Ref[$row - 1],
$$s2Ref[$col - 1]) == 1) {
$aboveLeftScore += $matchScore;
}
else {
```

```
$aboveLeftScore += $mismatchScore;
}
$scoreTable[$row][$col] = $aboveLeftScore;
$traceTable[$row][$col] = $fromAboveLeft;
if($aboveScore > $scoreTable[$row][$col]) {
$scoreTable[$row][$col] = $aboveScore;
$traceTable[$row][$col] = $fromAbove;
}
if($leftScore > $scoreTable[$row][$col]) {
$scoreTable[$row][$col] = $leftScore;
$traceTable[$row][$col] = $fromLeft;
}
}
}
# ignore tailing gaps
#my $maxCol = $scoreTable[0][$len2];
#my $maxRowIdx = 0;
#for(my $row = 1; $row <= $len1; $row++) {
#	if($scoreTable[$row][$len2] > $maxCol) {
#		$maxCol = $scoreTable[$row][$len2];
#		$maxRowIdx =$row;
#	}
#}
my $maxRow = $scoreTable[$len1][0];
my $maxColIdx = 0;
for(my $col = 1; $col <= $len2; $col++) {
if($scoreTable[$len1][$col] > $maxRow) {
$maxRow = $scoreTable[$len1][$col];
$maxColIdx = $col;
}
}
#if($maxRow > $maxCol) {
for(my $col = $len2; $col > $maxColIdx; $col--) {
$traceTable[$len1][$col] = $fromLeft;
}
$scoreTable[$len1][$len2] = $maxRow;
#}
#else {
#	for(my $row = $len1; $row > $maxRowIdx; $row--) {
#		$traceTable[$row][$len2] = $fromAbove;
#	}
#	$scoreTable[$len1][$len2] = $maxCol;
#}
# get output
my $row = $len1;
my $col = $len2;
```

```
my $as1 = "";
my $as2 = "";
while($traceTable[$row][$col] != $fromNowhere) {
if($traceTable[$row][$col] == $fromAboveLeft) {
$$a1Ref[$row - 1] = ($col - 1);
$$a2Ref[$col - 1] = ($row - 1);
$as1 = $$s1Ref[$row - 1] . $as1;
$as2 = $$s2Ref[$col - 1] . $as2;
$row--;
$col--;
}
elsif($traceTable[$row][$col] == $fromAbove) {
$$a1Ref[$row - 1] = -1;
$as1 = $$s1Ref[$row - 1] . $as1;
as2 = '-' . $as2;
$row--;
}
elsif($traceTable[$row][$col] == $fromLeft) {
$$a2Ref[$col - 1] = -1;
$as1 = '-' . $as1;
$as2 = $$s2Ref[$col - 1] . $as2;
$col--;
}
}
# test output
#	print "Score table:\n";
#	for(my $row = 0; $row <= $len1; $row++) {
#		for(my $col = 0; $col <= $len2; $col++) {
#			print "\t$scoreTable[$row][$col]";
#		}
#		print "\n";
#	}
#
#	print "Trace table:\n";
#	for(my $row = 0; $row <= $len1; $row++) {
#		for(my $col = 0; $col <= $len2; $col++) {
#			print "\t$traceTable[$row][$col]";
#		}
#		print "\n";
#	}
$$alignRef = $as1 . "\n" . $as2 . "\n";
return $scoreTable[$len1][$len2];
}
# dynamic programming sequence alignment
sub DP_ALIGN_4T {
# input: the two sequences to be aligned
```

```
# array references
my $s1Ref = shift;
my $s2Ref = shift;
my $len1 = $#$s1Ref + 1;
my $len2 = $#$s2Ref + 1;
# output: the index of aligned bases for each input sequence
# array references
my $a1Ref = shift;
my $a2Ref = shift;
# -1 in index means a gap
my $gapIdx = -1;
for(my $sIdx = 0; $sIdx < $len1; $sIdx++) {
$$a1Ref[$sIdx] = $gapIdx;
}
for(my $sIdx = 0; $sIdx < $len2; $sIdx++) {
$$a2Ref[$sIdx] = $gapIdx;
}
my $alignRef = shift;
#       # penalty and award
#       my $gapScore = -2;
#       my $mismatchScore = -1;
#       my $matchScore = 1;
# initialize score talbe
my @scoreTable = ();
for(my $row = 0; $row <= $len1; $row++) {
for(my $col = 0; $col <= $len2; $col++) {
#$scoreTable[$row][$col]= 0; # ignore leading gaps
$scoreTable[$row][$col]= $gapScore * $row; # ignore leading gaps
}
}
# trace direction
my $fromNowhere = 0;
my $fromLeft = 1;
my $fromAbove = 2;
my $fromAboveLeft = 3;
# initialize trace table
my @traceTable = ();
$traceTable[0][0] = $fromNowhere;
for(my $col = 1; $col <= $len2; $col++) {
$traceTable[0][$col] = $fromLeft;
}
for(my $row = 1; $row <= $len1; $row++) {
$traceTable[$row][0] = $fromAbove;
for(my $col = 1; $col <= $len2; $col++) {
$traceTable[$row][$col] = $fromNowhere;
}
```

```
}
#fill out score table and traceback table
for(my $row = 1; $row <= $len1; $row++) {
for(my $col = 1; $col <= $len2; $col++) {
my $aboveScore =
$scoreTable[$row - 1][$col] + $gapScore;
my $leftScore = $scoreTable[$row][$col - 1] + $gapScore;
my $aboveLeftScore = $scoreTable[$row - 1][$col - 1];
if(&SAME_BASE($$s1Ref[$row - 1],
$$s2Ref[$col - 1]) == 1) {
$aboveLeftScore += $matchScore;
}
else {
$aboveLeftScore += $mismatchScore;
}
$scoreTable[$row][$col] = $aboveLeftScore;
$traceTable[$row][$col] = $fromAboveLeft;
if($aboveScore > $scoreTable[$row][$col]) {
$scoreTable[$row][$col] = $aboveScore;
$traceTable[$row][$col] = $fromAbove;
}
if($leftScore > $scoreTable[$row][$col]) {
$scoreTable[$row][$col] = $leftScore;
$traceTable[$row][$col] = $fromLeft;
}
}
}
# ignore tailing gaps
my $maxCol = $scoreTable[0][$len2];
my $maxRowIdx = 0;
for(my $row = 1; $row <= $len1; $row++) {
if($scoreTable[$row][$len2] > $maxCol) {
$maxCol = $scoreTable[$row][$len2];
$maxRowIdx =$row;
}
}
my $maxRow = $scoreTable[$len1][0];
my $maxColIdx = 0;
for(my $col = 1; $col <= $len2; $col++) {
if($scoreTable[$len1][$col] > $maxRow) {
$maxRow = $scoreTable[$len1][$col];
$maxColIdx = $col;
}
}
if($maxRow > $maxCol) {
for(my $col = $len2; $col > $maxColIdx; $col--) {
```

```
$traceTable[$len1][$col] = $fromLeft;
}
$scoreTable[$len1][$len2] = $maxRow;
}
else {
for(my $row = $len1; $row > $maxRowIdx; $row--) {
$traceTable[$row][$len2] = $fromAbove;
}
$scoreTable[$len1][$len2] = $maxCol;
}
# get output
my $row = $len1;
my $col = $len2;
my $as1 = "";
my $as2 = "";
while($traceTable[$row][$col] != $fromNowhere) {
if($traceTable[$row][$col] == $fromAboveLeft) {
$$a1Ref[$row - 1] = ($col - 1);
$$a2Ref[$col - 1] = ($row - 1);
$as1 = $$s1Ref[$row - 1] . $as1;
$as2 = $$s2Ref[$col - 1] . $as2;
$row--;
$col--;
}
elsif($traceTable[$row][$col] == $fromAbove) {
$$a1Ref[$row - 1] = -1;
$as1 = $$s1Ref[$row - 1] . $as1;
$as2 = '-' . $as2;
$row--;
}
elsif($traceTable[$row][$col] == $fromLeft) {
$$a2Ref[$col - 1] = -1;
$as1 = '-' . $as1;
$as2 = $$s2Ref[$col - 1] . $as2;
$col--;
}
}
# test output
#	print "Score table:\n";
#	for(my $row = 0; $row <= $len1; $row++) {
#		for(my $col = 0; $col <= $len2; $col++) {
#			print "\t$scoreTable[$row][$col]";
#		}
#		print "\n";
#	}
#
```

```
#	print "Trace table:\n";
#	for(my $row = 0; $row <= $len1; $row++) {
#		for(my $col = 0; $col <= $len2; $col++) {
#			print "\t$traceTable[$row][$col]";
#		}
#		print "\n";
#	}
$$alignRef = $as1 . "\n" . $as2 . "\n";
return $scoreTable[$len1][$len2];
}
sub SAME_BASE {
my $base1 = shift;
my $base2 = shift;
$base1 =~ s/$base1/\U$base1\E/;
$base2 =~ s/$base2/\U$base2\E/;
if($base1 eq $base2) {
return 1;
}
if($base1 eq 'N') {
return 1;
}
if($base2 eq 'N') {
return 1;
}
return 0;
}
sub PRNMATCH {
my $FH = shift;
my $hIdx = shift;
my $rh = shift;
for(my $mIdx = 0; $mIdx <= $#{$hIdx}; $mIdx++) {
my $rIdx = ${$hIdx}[$mIdx];
if($rIdx == -1) {
print $FH "-";
}
else {
print $FH "${$rh}[$rIdx]";
}
}
}
```

Appendix 3

DNA Sequences (All sequences written 5' → 3'). Indexes are underlined.

Indexed Adapter 1/ Adapter 1 Complement (Indices 1-8 of 16 were synthesized and listed below)

Adapter 1 (Index 1), (39 bases)
ACACTCTTTCCCTACACGACGCTCTTCCGATCT**ATCGTA**

Adapter 1 Complement (Index 1), (51 bases)
GCATGCGCGTGT**TACGAT**AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT

Adapter 1 (Index 2), (39 bases)
ACACTCTTTCCCTACACGACGCTCTTCCGATCT**GATAGC**

Adapter 1 Complement (Index 2), (51 bases)
GCATGCGCGTGT**GCTATC**AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT

Adapter 1 (Index 3), (39 bases)
ACACTCTTTCCCTACACGACGCTCTTCCGATCT**CGATCT**

Adapter 1 Complement (Index 3), (51 bases)
GCATGCGCGTGT**AGATCG**AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT

Adapter 1 (Index 4), (39 bases)
ACACTCTTTCCCTACACGACGCTCTTCCGATCT**TCGCTA**

Adapter 1 Complement (Index 4), (51 bases)
GCATGCGCGTGT**TAGCGA**AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT

Adapter 1 (Index 5), (39 bases)
ACACTCTTTCCCTACACGACGCTCTTCCGATCT**TACGTC**

Adapter 1 Complement (Index 5), (51 bases)
GCATGCGCGTGT**GACGTA**AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT

Adapter 1 (Index 6), (39 bases)
ACACTCTTTCCCTACACGACGCTCTTCCGATCT**AGTAGT**

Adapter 1 Complement (Index 6), (51 bases)
GCATGCGCGTGT**ACTACT**AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT

Adapter 1 (Index 7), (39 bases)
ACACTCTTTCCCTACACGACGCTCTTCCGATCT**GCATGA**

Adapter 1 Complement (Index 7), (51 bases)
GCATGCGCGTGT**TCATGC**AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT

Adapter 1 (Index 8), (39 bases)
ACACTCTTTCCCTACACGACGCTCTTCCGATCT**CTGCTG**

Adapter 1 Complement (Index 8), (51 bases)
GCATGCGCGTGT**CAGCAG**AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT

Modified Adapter 2 (33bases)
Phos/GACAGAGGTCAGTCGTATGCCGTCTTCTGCTTG

Adapter 2 Complement (45 bases)
CAAGCAGAAGACGGCATACGACTGACCTCTGTCTGTGGCGCATGC

Appendix 4

DNA Sequences (All sequences written 5' → 3'). Indexes are underlined.

Oligonucleotide sequences for TruSeq™ RNA and DNA Sample Prep Kits:

TruSeq Universal Adapter (58 bases)
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT

TruSeq Adapter Index 1 (63 bases)
GATCGGAAGAGCACACGTCTGAACTCCAGTCAC**ATCACG**ATCTCGTATGCCGTCTTCTGCTTG

TruSeq Adapter, Index 2 (63 bases)
GATCGGAAGAGCACACGTCTGAACTCCAGTCAC**CGATGT**ATCTCGTATGCCGTCTTCTGCTTG

TruSeq Adapter, Index 3 (63 bases)
GATCGGAAGAGCACACGTCTGAACTCCAGTCAC**TTAGGC**ATCTCGTATGCCGTCTTCTGCTTG

TruSeq Adapter, Index 4 (63 bases)
GATCGGAAGAGCACACGTCTGAACTCCAGTCAC**TGACCA**ATCTCGTATGCCGTCTTCTGCTTG

TruSeq Adapter, Index 5 (63 bases)
GATCGGAAGAGCACACGTCTGAACTCCAGTCAC**ACAGTG**ATCTCGTATGCCGTCTTCTGCTTG

TruSeq Adapter, Index 6 (63 bases)
GATCGGAAGAGCACACGTCTGAACTCCAGTCAC**GCCAAT**ATCTCGTATGCCGTCTTCTGCTTG

TruSeq Adapter, Index 7 (63 bases)
GATCGGAAGAGCACACGTCTGAACTCCAGTCAC**CAGATC**ATCTCGTATGCCGTCTTCTGCTTG

TruSeq Adapter, Index 8 (63 bases)
GATCGGAAGAGCACACGTCTGAACTCCAGTCAC**ACTTGA**ATCTCGTATGCCGTCTTCTGCTTG

TruSeq Adapter, Index 9 (63 bases)
GATCGGAAGAGCACACGTCTGAACTCCAGTCAC**GATCAG**ATCTCGTATGCCGTCTTCTGCTTG

TruSeq Adapter, Index 10 (63 bases)
GATCGGAAGAGCACACGTCTGAACTCCAGTCAC**TAGCTT**ATCTCGTATGCCGTCTTCTGCTTG

TruSeq Adapter, Index 11 (63 bases)
GATCGGAAGAGCACACGTCTGAACTCCAGTCAC**GGCTAC**ATCTCGTATGCCGTCTTCTGCTTG

TruSeq Adapter, Index 12 (63 bases)
GATCGGAAGAGCACACGTCTGAACTCCAGTCAC**CTTGTA**ATCTCGTATGCCGTCTTCTGCTTG

m=15 Adapter Library (61 bases)
Phos/ACGACGCTCTTCCGATCT –m15- A**ATGCCTGG**GAGCACACGTCTGAACTCC

CAP1 Primer (20 bases)
ACACGACGCTCTTCCGATCT

CAP2 Primer (25 bases)
GTGACTGGAGTTCAGACGTGTGCTA

CAP1 Short (18 bases)
ACGACGCTCTTCCGATCT

CAP2 Short (17 bases)
AGACGGCATACGAGCTC

P1 Forward Primer (57 bases)
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT

P2.1 Reverse Primer (55 bases)
CAAGCAGAAGACGGCATACGAGAT**CGTGAT**GTGACTGGAGTTCAGACGTGTGCTC

P2.2 Reverse Primer (55 bases)
CAAGCAGAAGACGGCATACGAGAT**ACATCG**GTGACTGGAGTTCAGACGTGTGCTC

P2.3 Reverse Primer (55 bases)
CAAGCAGAAGACGGCATACGAGAT**GCCTAA**GTGACTGGAGTTC AGACGTGTGCTC

P2.4 Reverse Primer (55 bases)
CAAGCAGAAGACGGCATACGAGAT**TGGTCA**GTGACTGGAGTTC AGACGTGTGCTC

P2.5 Reverse Primer (55 bases)
CAAGCAGAAGACGGCATACGAGAT**CACTGT**GTGACTGGAGTTC AGACGTGTGCTC

P2.6 Reverse Primer (55 bases)
CAAGCAGAAGACGGCATACGAGAT**ATTGGC**GTGACTGGAGTTC AGACGTGTGCTC

P2.7 Reverse Primer (55 bases)
CAAGCAGAAGACGGCATACGAGAT**GATCTG**GTGACTGGAGTTC AGACGTGTGCTC

P2.8 Reverse Primer (55 bases)
CAAGCAGAAGACGGCATACGAGAT**TCAAGT**GTGACTGGAGTTC AGACGTGTGCTC

P2.9 Reverse Primer (55 bases)
CAAGCAGAAGACGGCATACGAGAT**CTGATC**GTGACTGGAGTTC AGACGTGTGCTC

P2.10 Reverse Primer (55 bases)
CAAGCAGAAGACGGCATACGAGAT**AAGCTA**GTGACTGGAGTTC AGACGTGTGCTC

P2.11 Reverse Primer (55 bases)
CAAGCAGAAGACGGCATACGAGAT**GTAGCC**GTGACTGGAGTTC AGACGTGTGCTC

P2.12 Reverse Primer (55 bases)
CAAGCAGAAGACGGCATACGAGAT**TACAAG**GTGACTGGAGTTC AGACGTGTGCTC

Read 1 Primer (33 bases)
ACACTCTTTCCCTACACGACGCTCTTCCGATCT

Custom Read 2 Primer (25 bases)
GTGACTGGAGTTCAGACGTGTGCTC

Custom Index Read Primer (25 bases)
GAGCACACGTCTGAACTCCAGTCAC

Illumina TruSeq Flow Cell Oligo (P5), (39 bases)
TTTTTTTTTTAATGATACGGCGACCACCGAGAUCTACAC

Illumina TruSeq Flow Cell Oligo (P7), (34 bases)
TTTTTTTTTTCAAGCAGAAGACGGCATACGAGAT

Top 20 Sequences, 60:1, m=15 adapter library bottom-bottom band

Rank	Sequence	Count	Identifier
1	GGTGGTTGTTGTGGT	927	TBAsc
2	GGTGGTTGTTGTGAT	7	TBAsc
3	GGTGGTTATTGTGGT	6	TBAsc
4	AGTGGTTGTTGTGGT	6	TBAsc
5	GGTGGTTGTTATGGT	5	TBAsc
6	GGTGGTTGTTGAGGT	4	TBAsc
7	GCGCAGTGCCACATC	3	
8	CGATGACTGACGCGT	3	
9	GGTGGTTGTTGTGCT	3	TBAsc
10	GATGGTTGTTGTGGT	3	TBAsc
11	TGTGGTTGTTGTGGT	3	TBAsc
12	GGTGGTTGTCGTGGT	3	TBAsc
13	NGTGGTTGTTGTGGT	3	TBAsc
14	GGTGGTTGTTCTGGT	2	TBAsc
15	GTGCGACCGCACCCC	2	
16	CGCCTCTTCTGATGC	2	
17	ACACGGCCCGCCCGG	2	
18	CGACCTCCGTCCAGC	2	
19	CGCGCCCAGTCCTCC	2	
20	CTGCTTCGCTCAGGC	2	

Top 20 Sequences, 60:1, m=15 adapter library bottom-top band

Rank	Sequence	Count	Identifier
1	GGTGGTTGTTGTGGT	534	TBAsc
2	GGTGATTGTTGTGGT	6	TBAsc
3	GGTAGTTGTTGTGGT	6	TBAsc
4	GGTGGTTGTTGTAGT	6	TBAsc
5	GGTGGTTGTTATGGT	5	TBAsc
6	TTACGATGTCCGACT	3	
7	AGTGGTTGTTGTGGT	3	
8	CGGTTCGCGATGGAC	3	
9	GATGGTTGTTGTGGT	3	
10	GCGCCCGATCGTGAT	3	
11	GGTGGTTGTTGCGGT	3	TBAsc
12	GGTGGCTGTTGTGGT	3	TBAsc
13	GCACCAGCCGTTACC	3	
14	CTTGATTGCTTGGTG	2	
15	TCTGCCGGGCCCAAC	2	
16	GGCCAGGCCCACTGC	2	
17	GGCGTGACACGGTCG	2	
18	CCACTGCTTGTTATC	2	
19	GCGATGCTTACTCGC	2	
20	ATTTCTCCATGGGTC	2	

Appendix 5

DNA Sequences (All sequences written 5' → 3'). Indexes are underlined.

Adapters for ligating un-flanked libraries:

Adapter 1 (58 bases)
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT

Adapter 1 Complement, Full Length (64 bases)
AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT

Adapter 1 Complement, Short, n=1 (34 bases)
NAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT

Adapter 1 Complement, Short, n=2 (35 bases)
NNAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT

Adapter 1 Complement, Short, n=3 (36 bases)
NNNAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT

Adapter 1 Complement, Short, n=4 (37 bases)
NNNNAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT

Adapter 2, Index 1 (64 bases)
Phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCAC**ATCACG**ATCTCGTATGCCGTCTTCTGCTTG

Adapter 2 Complement, Index 1, n=1 (3'-Amine), (65 bases)
CAAGCAGAAGACGGCATACGAGATCGTGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCN/3AmMO/

Adapter 2 Complement, Index 1, n=1 (3'-Amine), (66 bases)
CAAGCAGAAGACGGCATACGAGATCGTGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCNN

Adapter 2 Complement, Index 1, n=1 (3'-Amine), (67 bases)
CAAGCAGAAGACGGCATACGAGATCGTGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCNNN

Adapter 2 Complement, Index 1, n=1 (3'-Amine), (68 bases)
CAAGCAGAAGACGGCATACGAGATCGTGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCNNNN

Click ligation adapters and libraries

m=15 DNA Click Library (5'-azide, 3'-alkyne)
TNNNNNNNNNNNNNNC

Click TBA (5'-azide, 3'-alkyne)
TGGTTGGTGTGGTTGGC

Adapter 1(3'-alkyne) (59 bases)
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTC

Adapter 1 Complement, A/ n=3 splint (38 bases)
5NNNAGAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT

Adapter 1 Complement, Fixed splint (38 bases)
5ACCAGAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT

Adapter 2, Index 1 (5'-azide)
/GATCGGAAGAGCACACGTCTGAACTCCAGTCAC**ATCACG**ATCTCGTATGCCGTCTTCTGCTTG

Adapter 2 Complement, G/n=3 splint (68 bases)
CAAGCAGAAGACGGCATACGAGAT**CGTGAT**GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCAGNN
N

Adapter 2 Complement, Fixed splint (68 bases)
CAAGCAGAAGACGGCATACGAGAT**CGTGAT**GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCAGCC
A

Adapters for ligating tailed libraries:

Adapter 1 (58 bases)
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT

Adapter 1 Complement, Short, Fixed splint (47 bases)
GTGTAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT

Adapter 1 Complement, Short, Fixed/n=14 splint (41 bases)
NNNNGTGTAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT

Adapter 2, Index 1 (64 bases)
Phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCAC**ATCACG**ATCTCGTATGCCGTCTTCTGCTTG

Adapter 2 Complement, Index , Fixed splint (3'-Amine), (68 bases)
CAAGCAGAAGACGGCATACGAGAT**CGTGAT**GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGTG/3AmMO/

Adapter 2 Complement, Index 1, Fixed/n=4 splint (3'-Amine), (72 bases)
CAAGCAGAAGACGGCATACGAGAT**CGTGAT**GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGTGNNNN/3AmMO/

Appendix 6

DNA Sequences (All sequences written 5' → 3').

Pyrrolo-C oligos, underlined base indicates Pyrrolo-C →C substitution:

m=15 tailed library, external Py-C (24 bases)
pACACNNNNNNNNNNNNNN**C**ACA

m=15 tailed library, internal PyC (24 bases)
pACACNNNNNNN**C**NNNNNNNCACA

Tailed TBA-PyC (24 bases)
pACACGGTTGGTGTGGTTGG**C**ACA

Tailed TBAsc-PyC (24 bases)
pACACGGTGGTTGTTGTGGT**C**ACA

44-mer library/TBA (44 bases)
pCANNNNN(**C**N)$_4$GGTTGGTGTGGTTGG(N)$_{12}$GC

References

1. Famulok, M., J.S. Hartig, and G. Mayer, *Functional Aptamers and Aptazymes in Biotechnology, Diagnostics, and Therapy.* Chemical Reviews, 2007. **107**(9): p. 3715-3743.

2. Potyrailo, R.A., et al., *Adapting selected nucleic acid ligands (aptamers) to biosensors.* Anal. Chem., 1998. **70**: p. 3419.

3. Hamaguchi, N., A. Ellington, and M. Stanton, *Aptamer Beacons for the Direct Detection of Proteins.* Analytical Biochemistry, 2001. **294**(2): p. 126-131.

4. DeCiantis, C.L., et al., *A Nucleic Acid Switch Triggered by the HIV-1 Nucleocapsid Protein†.* Biochemistry, 2007. **46**(32): p. 9164-9173.

5. Cheng, A.K.H., D. Sen, and H.-Z. Yu, *Design and testing of aptamer-based electrochemical biosensors for proteins and small molecules.* Bioelectrochemistry, 2009. **77**(1): p. 1-12.

6. Vallée-Bélisle, A. and K.W. Plaxco, *Structure-switching biosensors: inspired by Nature.* Current opinion in structural biology, 2010. **20**(4): p. 518-526.

7. Yin, J., et al., *Label-Free and Turn-on Aptamer Strategy for Cancer Cells Detection Based on a DNA–Silver Nanocluster Fluorescence upon Recognition-Induced Hybridization.* Analytical Chemistry, 2013. **85**(24): p. 12011-12019.

8. Zhu, Y., et al., *Building an aptamer/graphene oxide FRET biosensor for one-step detection of bisphenol A.* ACS Appl Mater Interfaces, 2015. **7**(14): p. 7492-6.

9. Contreras Jimenez, G., et al., *Aptamer-based label-free impedimetric biosensor for detection of progesterone.* Anal Chem, 2015. **87**(2): p. 1075-82.

10. Zhou, W., et al., *Aptamer-based biosensors for biomedical diagnostics.* Analyst, 2014. **139**(11): p. 2627-40.

11. Chen, F., et al., *Aptamer from whole-bacterium SELEX as new therapeutic reagent against virulent Mycobacterium tuberculosis.* Biochemical and Biophysical Research Communications, 2007. **357**(3): p. 743-748.

12. Ng, E.W.M. and A.P. Adamis, *Anti-VEGF Aptamer (Pegaptanib) Therapy for Ocular Vascular Diseases.* Annals of the New York Academy of Sciences, 2006. **1082**(1): p. 151-171.

13. Keefe, A.D. and R.G. Schaub, *Aptamers as candidate therapeutics for cardiovascular indications.* Curr. Opin. Pharmacol., 2008. **8**: p. 147-152.

14. Keefe, A.D., S. Pai, and A. Ellington, *Aptamers as therapeutics.* Nat Rev Drug Discov, 2010. **9**(7): p. 537-550.

15. Thiel, K.W. and P.H. Giangrande, *Therapeutic Applications of DNA and RNA Aptamers.* Oligonucleotides, 2009. **19**(3): p. 209-222.

16. Pastor, F., et al., *Induction of tumour immunity by targeted inhibition of nonsense-mediated mRNA decay.* Nature, 2010. **465**(7295): p. 227-30.

17. Zhou, J., et al., *Dual functional RNA nanoparticles containing phi29 motor pRNA and anti-gp120 aptamer for cell-type specific delivery and HIV-1 inhibition.* Methods, 2011. **54**(2): p. 284-94.

18. Sun, H., et al., *Oligonucleotide Aptamers: New Tools for Targeted Cancer Therapy.* Mol Ther Nucleic Acids, 2014. **3**: p. e182.

19. Gnanum, A.J., et al., *Development of aptamers specific for potential diagnostic targets in B. pseudomallei.* Transactions of the Royal Society of Tropical Medicine and Hygiene, 2008. **102**(0 1): p. S55-S57.

20. Gold, L., et al., *Aptamer-Based Multiplexed Proteomic Technology for Biomarker Discovery.* PLoS ONE, 2010. **5**(12): p. e15004.

21. Kirby, R., et al., *Aptamer-Based Sensor Arrays for the Detection and Quantitation of Proteins.* Analytical Chemistry, 2004. **76**(14): p. 4066-4075.

22. Sefah, K., et al., *Nucleic acid aptamers for biosensors and bio-analytical applications.* Analyst, 2009. **134**(9): p. 1765-1775.

23. Wan, Y., et al., *Surface-immobilized aptamers for cancer cell isolation and microscopic cytology.* Cancer Res, 2010. **70**(22): p. 9371-80.

24. Ozalp, V.C., et al., *Aptamers: molecular tools for medical diagnosis.* Curr Top Med Chem, 2015. **15**(12): p. 1125-37.

25. Wang, K., et al., *Highly sensitive and specific colorimetric detection of cancer cells via dual-aptamer target binding strategy.* Biosensors and Bioelectronics, 2015. **73**: p. 1-6.

26. Silverman, S.K., *Functional Nucleic Acids for Sensing and Other Analytical Applications*, ed. Y. Lu, Li, Y. 2007: Springer.

27. Fitzwater, T. and B. Polisky, *[17] A SELEX primer*, in *Methods in Enzymology*, N.A. John, Editor. 1996, Academic Press. p. 275-301.

28. Jayasena, S.D., *Aptamers: An Emerging Class of Molecules That Rival Antibodies in Diagnostics.* Clin. Chem., 1999. **45**: p. 1628.

29. Schultz, R.G. and S.M. Gryaznov, *Oligo-2'-fluoro-2'-deoxynucleotide N3'-->P5' phosphoramidates: synthesis and properties.* Nucleic Acids Research, 1996. **24**(15): p. 2966-2973.

30. Burmeister, P.E., et al., *Direct in vitro selection of a 2'-O-methyl aptamer to VEGF*. Chem Biol, 2005. **12**(1): p. 25-33.

31. Klussman, S., et al., *Mirror-image RNA that binds D-adenosine.* Nature Biotech., 1996. **14**: p. 1112-1115.

32. Nolte, A., et al., *Mirror-design of L-oligonucleotide ligands binding to L-arginine.* Nat Biotechnol, 1996. **14**(9): p. 1116-9.

33. Kupakuwana, G.V., et al., *Acyclic Identification of Aptamers for Human alpha-Thrombin Using Over-Represented Libraries and Deep Sequencing.* PLoS ONE, 2011. **6**(5): p. e19395.

34. **Song, K., Lee, S. and Ban, C.**, *Aptamers and Their Biological Applications*. 2011: Sensors. p. 612-631.

35. Cullen, B.R. and W.C. Greene, *Regulatory pathways governing HIV-1 replication.* Cell, 1989. **58**(3): p. 423-426.

36. Marciniak, R.A., M.A. Garcia-Blanco, and P.A. Sharp, *Identification and characterization of a HeLa nuclear protein that specifically binds to the trans-activation-response (TAR) element of human immunodeficiency virus.* Proceedings of the National Academy of Sciences of the United States of America, 1990. **87**(9): p. 3624-3628.

37. Hayashi, T., et al., *RNA packaging signal of human immunodeficiency virus type 1.* Virology, 1992. **188**: p. 590.

38. O'Malley, R.P., et al., *A mechanism for the control of protein synthesis by adenovirus VA RNAI.* Cell, 1986. **44**(3): p. 391-400.

39. Tuerk, C. and L. Gold, *Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase.* Science, 1990. **249**(4968): p. 505-10.

40. Ellington, A.D. and J.W. Szostak, *In vitro selection of RNA molecules that bind specific ligands.* Nature, 1990. **346**: p. 818-822.

41. Gold, L., et al., *Diversity of oligonucleotide functions.* Annual Review of Biochemistry, 1995. **64**: p. 763-797.

42. Marshall, K.A. and A.D. Ellington, *In vitro selection of RNA aptamers.* Methods Enzymol, 2000. **318**: p. 193-214.

43. Lozupone, C., et al., *Selection of the simplest RNA that binds isoleucine.* RNA, 2003. **9**(11): p. 1315-1322.

44. Legiewicz, M., et al., *Size, constant sequences, and optimal selection.* RNA, 2005. **11**(11): p. 1701-1709.

45. Connell, G.J., M. Illangesekare, and M. Yarus, *Three small ribooligonucleotides with specific arginine sites.* Biochemistry, 1993. **32**(21): p. 5497-502.

46. Ferreira, C.S., C.S. Matthews, and S. Missailidis, *DNA aptamers that bind to MUC1 tumour marker: design and characterization of MUC1-binding single-stranded DNA aptamers.* Tumour Biol., 2006. **27**: p. 289-301.

47. Hale, S.P. and P. Schimmel, *Protein synthesis editing by a DNA aptamer.* Proc. Natl Acad. Sci. USA, 1996. **93**: p. 2755-2758.

48. Missailidis, S., et al., *Selection of aptamers with high affinity and high specificity against C595, an anti-MUC1 IgG3 monoclonal antibody, for antibody targeting.* Journal of Immunological Methods, 2005. **296**(1–2): p. 45-62.

49. Conrad, R., et al., *Isozyme-specific inhibition of protein kinase C by RNA aptamers.* J Biol Chem, 1994. **269**(51): p. 32051-4.

50. Ellington, A.D. and J.W. Szostak, *Selection in vitro of single-stranded DNA molecules that fold into specific ligand-binding structures.* Nature, 1992. **355**(6363): p. 850-2.

51. Kumar, P.K., et al., *Isolation of RNA aptamers specific to the NS3 protein of hepatitis C virus from a pool of completely random RNA.* Virology, 1997. **237**(2): p. 270-82.

52. Sassanfar, M. and J.W. Szostak, *An RNA motif that binds ATP.* Nature, 1993. **364**: p. 550-553.

53. Yamamoto, R., T. Baba, and P.K. Kumar, *Molecular beacon aptamer fluoresces in the presence of Tat protein of HIV-1.* Genes Cells, 2000. **5**: p. 389.

54. Li, Y., C.R. Geyer, and D. Sen, *Recognition of anionic porphyrins by DNA aptamers.* Biochemistry, 1996. **35**(21): p. 6911-22.

55. Osborne, S.E. and A.D. Ellington, *Nucleic Acid Selection and the Challenge of Combinatorial Chemistry.* Chemical Reviews, 1997. **97**(2): p. 349-370.

56. Knight, R. and M. Yarus, *Analyzing partially randomized nucleic acid pools: straight dope on doping.* Nucleic Acids Research, 2003. **31**(6): p. e30-e30.

57. Bock, L.C., et al., *Selection of single-stranded DNA molecules that bind and inhibit human thrombin.* Nature, 1992. **355**(6360): p. 564-566.

58. Kato, T., et al., *In vitro selection of DNA aptamers which bind to cholic acid.* Biochim Biophys Acta, 2000. **1493**(1-2): p. 12-8.

59. Vianini, E., M. Palumbo, and B. Gatto, *In vitro selection of DNA aptamers that bind L-tyrosinamide.* Bioorg Med Chem, 2001. **9**(10): p. 2543-8.

60. Daniels, D.A., et al., *A tenascin-C aptamer identified by tumor cell SELEX: Systematic evolution of ligands by exponential enrichment.* Proceedings of the National Academy of Sciences, 2003. **100**(26): p. 15416-15421.

61. Tuerk, C. and L. Gold, *Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase.* Science, 1990. **249**: p. 505-510.

62. Gopinath, S.C., *Methods developed for SELEX.* Anal Bioanal Chem, 2007. **387**(1): p. 171-82.

63. Wang, C., et al., *In vitro selection of high-affinity DNA aptamers for streptavidin.* Acta Biochimica et Biophysica Sinica, 2009. **41**(4): p. 335-340.

64. Hianik, T., et al., *Influence of ionic strength, pH and aptamer configuration for binding affinity to thrombin.* Bioelectrochemistry, 2007. **70**(1): p. 127-33.

65. Russo Krauss, I., et al., *High-resolution structures of two complexes between thrombin and thrombin-binding aptamer shed light on the role of cations in the aptamer inhibitory activity.* Nucleic Acids Research, 2012. **40**(16): p. 8119-8128.

66. Lorenz, C., F. von Pelchrzim, and R. Schroeder, *Genomic systematic evolution of ligands by exponential enrichment (Genomic SELEX) for the identification of protein-binding RNAs independent of their expression levels.* Nat. Protocols, 2006. **1**(5): p. 2204-2212.

67. Allen, P., S. Worland, and L. Gold, *Isolation of high-affinity RNA ligands to HIV-1 integrase from a random pool.* Virology, 1995. **209**: p. 327-336.

68. Mendonsa, S.D. and M.T. Bowser, *In vitro evolution of functional DNA using capillary electrophoresis.* J Am Chem Soc, 2004. **126**(1): p. 20-1.

69. Mendonsa, S.D. and M.T. Bowser, *In vitro selection of high-affinity DNA ligands for human IgE using capillary electrophoresis.* Anal Chem, 2004. **76**(18): p. 5387-92.

70. Mosing, R.K., S.D. Mendonsa, and M.T. Bowser, *Capillary electrophoresis-SELEX selection of aptamers with affinity for HIV-1 reverse transcriptase.* Anal Chem, 2005. **77**(19): p. 6107-12.

71. Berezovski, M., et al., *Nonequilibrium capillary electrophoresis of equilibrium mixtures: a universal tool for development of aptamers.* J Am Chem Soc, 2005. **127**(9): p. 3165-71.

72. Drabovich, A.P., et al., *Selection of smart aptamers by methods of kinetic capillary electrophoresis.* Anal Chem, 2006. **78**(9): p. 3171-8.

73. Tran, D.T., et al., *Selection and characterization of DNA aptamers for egg white lysozyme.* Molecules, 2010. **15**(3): p. 1127-40.

74. Rose, C.M., et al., *Capillary electrophoretic development of aptamers for a glycosylated VEGF peptide fragment.* Analyst, 2010. **135**(11): p. 2945-51.

75. Yang, J. and M.T. Bowser, *Capillary electrophoresis-SELEX selection of catalytic DNA aptamers for a small-molecule porphyrin target.* Anal Chem, 2013. **85**(3): p. 1525-30.

76. Mendonsa, S.D. and M.T. Bowser, *In vitro selection of aptamers with affinity for neuropeptide Y using capillary electrophoresis.* J Am Chem Soc, 2005. **127**(26): p. 9382-3.

77. Ozer, A., J.M. Pagano, and J.T. Lis, *New Technologies Provide Quantum Changes in the Scale, Speed, and Success of SELEX Methods and Aptamer Characterization.* Mol Ther Nucleic Acids, 2014. **3**: p. e183.

78. Berezovski, M., et al., *Non-SELEX selection of aptamers.* J Am Chem Soc, 2006. **128**(5): p. 1410-1.

79. Tok, J., et al., *Selection of aptamers for signal transduction proteins by capillary electrophoresis.* ELECTROPHORESIS, 2010. **31**(12): p. 2055-2062.

80. Ashley, J., K. Ji, and S.F.Y. Li, *Selection of bovine catalase aptamers using non-SELEX.* ELECTROPHORESIS, 2012. **33**(17): p. 2783-2789.

81. Oh, S.S., et al., *Generation of highly specific aptamers via micromagnetic selection.* Anal Chem, 2009. **81**(13): p. 5490-5.

82. Lou, X., et al., *Micromagnetic selection of aptamers in microfluidic channels.* Proc Natl Acad Sci U S A, 2009. **106**(9): p. 2989-94.

83. Cho, M., et al., *Quantitative selection of DNA aptamers through microfluidic selection and high-throughput sequencing.* Proc Natl Acad Sci U S A, 2010. **107**(35): p. 15373-8.

84. Ahn, J.Y., et al., *A sol-gel-based microfluidics system enhances the efficiency of RNA aptamer selection.* Oligonucleotides, 2011. **21**(2): p. 93-100.

85. Birch, C.M., et al., *Identification of malaria parasite-infected red blood cell surface aptamers by inertial microfluidic SELEX (I-SELEX).* Sci. Rep., 2015. **5**.

86. Cox, J.C., P. Rudolph, and A.D. Ellington, *Automated RNA selection.* Biotechnol. Prog., 1998. **14**: p. 845-850.

87. Cox, J.C. and A.D. Ellington, *Automated selection of anti-protein aptamers.* Bioorg. Med. Chem., 2001. **9**: p. 2525-2531.

88. Cox, J.C. and A.D. Ellington, *Automated selection of anti-protein aptamers.* Bioorg Med Chem, 2001. **9**(10): p. 2525-31.

89. Bryant, K.F., et al., *Binding of herpes simplex virus-1 US11 to specific RNA sequences.* Nucleic Acids Res, 2005. **33**(19): p. 6090-100.

90. Cox, J.C., et al., *Automated acquisition of aptamer sequences.* Comb Chem High Throughput Screen, 2002. **5**(4): p. 289-99.

91. Eulberg, D., *Development of an automated in vitro selection protocol to obtain RNA-based aptamers: identification of a biostable substance P antagonist.* Nucleic Acids Res., 2005. **33**: p. e45.

92. Cox, J.C., *Automated selection of aptamers against protein targets translated in vitro: from gene to aptamer.* Nucleic Acids Res., 2002. **30**: p. e108.

93. Hünniger, T., et al., *Just in Time-Selection: A Rapid Semiautomated SELEX of DNA Aptamers Using Magnetic Separation and BEAMing.* Analytical Chemistry, 2014. **86**(21): p. 10940-10947.

94. Blackwell, T.K., *[41] Selection of protein binding sites from random nucleic acid sequences*, in *Methods in Enzymology*, I.M.V. Peter K. Vogt, Editor. 1995, Academic Press. p. 604-618.

95. Tsai, R.Y. and R.R. Reed, *Identification of DNA recognition sequences and protein interaction domains of the multiple-Zn-finger protein Roaz.* Mol Cell Biol, 1998. **18**(11): p. 6447-56.

96. Goodman, S.D., et al., *In Vitro Selection of Integration Host Factor Binding Sites.* Journal of Bacteriology, 1999. **181**(10): p. 3246-3255.

97. Yao, W., K. Adelman, and J.A. Bruenn, *In vitro selection of packaging sites in a double-stranded RNA virus.* J Virol, 1997. **71**(3): p. 2157-62.

98. Stoltenburg, R., N. Nikolaus, and B. Strehlitz, *Capture-SELEX: Selection of DNA Aptamers for AMinoglycoside Antibiotics*. 2012: Journal of Analytical Methods in Chemistry. p. 14.

99. Spiga, F.M., P. Maietta, and C. Guiducci, *More DNA–Aptamers for Small Drugs: A Capture–SELEX Coupled with Surface Plasmon Resonance and High-Throughput Sequencing.* ACS Combinatorial Science, 2015. **17**(5): p. 326-333.

100. Golden, M.C., et al., *Diagnostic potential of PhotoSELEX-evolved ssDNA aptamers.* Journal of Biotechnology, 2000. **81**(2–3): p. 167-178.

101. Jensen, K.B., et al., *Using in vitro selection to direct the covalent attachment of human immunodeficiency virus type 1 Rev protein to high-affinity RNA ligands.* Proceedings of the National Academy of Sciences of the United States of America, 1995. **92**(26): p. 12220-12224.

102. Eaton, B.E., *The joys of in vitro selection: chemically dressing oligonucleotides to satiate protein targets.* Curr Opin Chem Biol, 1997. **1**(1): p. 10-6.

103. Rohloff, J.C., et al., *Nucleic Acid Ligands With Protein-like Side Chains: Modified Aptamers and Their Use as Diagnostic and Therapeutic Agents.* Mol Ther Nucleic Acids, 2014. **3**: p. e201.

104. Maasch, C., et al., *Polyetheylenimine-polyplexes of Spiegelmer NOX-A50 directed against intracellular high mobility group protein A1 (HMGA1) reduce tumor growth in vivo.* J Biol Chem, 2010. **285**(51): p. 40012-8.

105. Purschke, W.G., *A DNA Spiegelmer to staphylococcal enterotoxin B.* Nucleic Acids Res., 2003. **31**: p. 3027-3032.

106. Purschke, W.G., *An L-RNA-based aquaretic agent that inhibits vasopressin in vivo.* Proc. Natl Acad. Sci. USA, 2006. **103**: p. 5173-5178.

107. Vater, A., *Short bioactive Spiegelmers to migraine-associated calcitonin gene-related peptide rapidly identified by a novel approach: tailored-SELEX.* Nucleic Acids Res., 2003. **31**: p. e130.

108. Vater, A. and S. Klussmann, *Turning mirror-image oligonucleotides into drugs: the evolution of Spiegelmer® therapeutics.* Drug Discovery Today, 2015. **20**(1): p. 147-155.

109. Burmeister, P.E., *2'-Deoxy purine, 2'-O-methyl pyrimidine (dRmY) aptamers as candidate therapeutics.* Oligonucleotides, 2006. **16**: p. 337-351.

110. Richardson, F.C., et al., *Polymerization of 2'-fluoro- and 2'-O-methyl-dNTPs by human DNA polymerase alpha, polymerase gamma, and primase.* Biochem Pharmacol, 2000. **59**(9): p. 1045-52.

111. Smith, C.M. and J.A. Steitz, *Sno storm in the nucleolus: new roles for myriad small RNPs.* Cell, 1997. **89**(5): p. 669-72.

112. Lebruska, L.L. and L.J. Maher, 3rd, *Selection and characterization of an RNA decoy for transcription factor NF-kappa B.* Biochemistry, 1999. **38**(10): p. 3168-74.

113. Pieken, W.A., et al., *Kinetic characterization of ribonuclease-resistant 2'-modified hammerhead ribozymes.* Science, 1991. **253**(5017): p. 314-317.

114. Sekiya, S., et al., *Characterization and application of a novel RNA aptamer against the mouse prion protein.* J Biochem, 2006. **139**(3): p. 383-90.

115. Rhodes, A., et al., *The generation and characterization of antagonist RNA aptamers to human oncostatin M.* J Biol Chem, 2000. **275**(37): p. 28555-61.

116. Padilla, R. and R. Sousa, *Efficient synthesis of nucleic acids heavily modified with non-canonical ribose 2'-groups using a mutantT7 RNA polymerase (RNAP).* Nucleic Acids Research, 1999. **27**(6): p. 1561-1563.

117. Padilla, R. and R. Sousa, *A Y639F/H784A T7 RNA polymerase double mutant displays superior properties for synthesizing RNAs with non-canonical NTPs.* Nucleic Acids Research, 2002. **30**(24): p. e138-e138.

118. Chelliserrykattil, J. and A.D. Ellington, *Evolution of a T7 RNA polymerase variant that transcribes 2[prime]-O-methyl RNA.* Nat Biotech, 2004. **22**(9): p. 1155-1160.

119. Rhie, A., et al., *Characterization of 2'-fluoro-RNA aptamers that bind preferentially to disease-associated conformations of prion protein and inhibit conversion.* J Biol Chem, 2003. **278**(41): p. 39697-705.

120. Ruckman, J., et al., *2'-Fluoropyrimidine RNA-based aptamers to the 165-amino acid form of vascular endothelial growth factor (VEGF165). Inhibition of receptor binding and VEGF-induced vascular permeability through interactions requiring the exon 7-encoded domain.* J Biol Chem, 1998. **273**(32): p. 20556-67.

121. White, R.R., et al., *A nuclease-resistant RNA aptamer specifically inhibits angiopoietin-1-mediated Tie2 activation and function.* Angiogenesis, 2008. **11**(4): p. 395-401.

122. Miyakawa, S., et al., *Structural and molecular basis for hyperspecificity of RNA aptamer to human immunoglobulin G.* RNA, 2008. **14**(6): p. 1154-1163.

123. Rusconi, C.P., et al., *RNA aptamers as reversible antagonists of coagulation factor IXa.* Nature, 2002. **419**(6902): p. 90-94.

124. Pan, W., P. Xin, and G.A. Clawson, *Minimal primer and primer-free SELEX protocols for selection of aptamers from random DNA libraries.* Biotechniques, 2008. **44**(3): p. 351-60.

125. Pan, W. and G.A. Clawson, *The Shorter the Better: Reducing Fixed Primer Regions of Oligonucleotide Libraries for Aptamer Selection.* Molecules (Basel, Switzerland), 2009. **14**(4): p. 10.3390/molecules14041353.

126. Pan, W., et al., *Primer-Free Aptamer Selection Using A Random DNA Library.* Journal of Visualized Experiments : JoVE, 2010(41): p. 2039.

127. Cowperthwaite, M.C. and A.D. Ellington, *Bioinformatic Analysis of the Contribution of Primer Sequences to Aptamer Structures.* Journal of molecular evolution, 2008. **67**(1): p. 95-102.

128. Hesselberth, J.R., et al., *In vitro selection of RNA molecules that inhibit the activity of ricin A-chain.* J Biol Chem, 2000. **275**(7): p. 4937-42.

129. Shtatland, T., et al., *Interactions of escherichia coli RNA with bacteriophage MS2 coat protein: genomic SELEX.* Nucleic Acids Res, 2000. **28**(21): p. E93.

130. Ng, E.W., et al., *Pegaptanib, a targeted anti-VEGF aptamer for ocular vascular disease.* Nat Rev Drug Discov, 2006. **5**(2): p. 123-32.

131. Rusconi, C.P., et al., *Antidote-mediated control of an anticoagulant aptamer in vivo.* Nat Biotechnol, 2004. **22**(11): p. 1423-8.

132. Eikelboom, J.W., S.L. Zelenkofske, and C.P. Rusconi, *Coagulation factor IXa as a target for treatment and prophylaxis of venous thromboembolism.* Arterioscler Thromb Vasc Biol, 2010. **30**(3): p. 382-7.

133. Vavalle, J.P. and M.G. Cohen, *The REG1 anticoagulation system: a novel actively controlled factor IX inhibitor using RNA aptamer technology for treatment of acute coronary syndrome.* Future Cardiol, 2012. **8**(3): p. 371-82.

134. Green, L.S., *Inhibitory DNA ligands to platelet-derived growth factor B-chain.* Biochemistry, 1996. **35**: p. 14413-14424.

135. Diener, J.L., et al., *Inhibition of von Willebrand factor-mediated platelet activation and thrombosis by the anti-von Willebrand factor A1-domain aptamer ARC1779.* J Thromb Haemost, 2009. **7**(7): p. 1155-62.

136. Cosmi, B., *ARC-1779, a PEGylated aptamer antagonist of von Willebrand factor for potential use as an anticoagulant or antithrombotic agent.* Curr Opin Mol Ther, 2009. **11**(3): p. 322-8.

137. Jilma, B., et al., *A randomised pilot trial of the anti-von Willebrand factor aptamer ARC1779 in patients with type 2b von Willebrand disease.* Thromb Haemost, 2010. **104**(3): p. 563-70.

138. Gilbert, J.C., *First-in-human evaluation of anti von Willebrand factor therapeutic aptamer ARC1779 in healthy volunteers.* Circulation, 2007. **116**: p. 2678-2686.

139. Ireson, C.R. and L.R. Kelland, *Discovery and development of anticancer aptamers.* Mol Cancer Ther, 2006. **5**(12): p. 2957-62.

140. Jenison, R.D., *Oligonucleotide inhibitors of P-selectin-dependent neutrophil-platelet adhesion.* Antisense Nucleic Acid Drug Dev., 1998. **8**: p. 265-279.

141. Vance, S.A. and M.G. Sandros, *Zeptomole Detection of C-Reactive Protein in Serum by a Nanoparticle Amplified Surface Plasmon Resonance Imaging Aptasensor.* Sci. Rep., 2014. **4**.

142. Levy-Nissenbaum, E., et al., *Nanotechnology and aptamers: applications in drug delivery.* Trends Biotechnol, 2008. **26**(8): p. 442-9.

143. Ostroff, R.M., et al., *Unlocking Biomarker Discovery: Large Scale Application of Aptamer Proteomic Technology for Early Detection of Lung Cancer.* PLoS ONE, 2010. **5**(12): p. e15003.

144. Ostroff, R.M., et al., *Early detection of malignant pleural mesothelioma in asbestos-exposed individuals with a noninvasive proteomics-based surveillance tool.* PLoS One, 2012. **7**(10): p. e46091.

145. Kiddle, S.J., et al., *Candidate blood proteome markers of Alzheimer's disease onset and progression: a systematic review and replication study.* J Alzheimers Dis, 2014. **38**(3): p. 515-31.

146. Bruno, J.G., *Predicting the Uncertain Future of Aptamer-Based Diagnostics and Therapeutics.* Molecules, 2015. **20**(4): p. 6866-87.

147. Schütze, T., et al., *Probing the SELEX Process with Next-Generation Sequencing.* PLoS ONE, 2011. **6**(12): p. e29604.

148. Kupakuwana, G., *High Throughput Screening of Aptamers*, in *Structural Biology, Biochemistry and Biophyics*. 2011, Syracuse University. p. 236.

149. Wang, K.Y., et al., *A DNA aptamer which binds to and inhibits thrombin exhibits a new structural motif for DNA.* Biochemistry, 1993. **32**(8): p. 1899-904.

150. Kenan, D.J., D.E. Tsai, and J.D. Keene, *Exploring molecular diversity with combinatorial shape libraries.* Trends Biochem Sci, 1994. **19**(2): p. 57-64.

151. Padmanabhan, K., et al., *The structure of alpha-thrombin inhibited by a 15-mer single-stranded DNA aptamer.* J Biol Chem, 1993. **268**(24): p. 17651-4.

152. Padmanabhan, K. and A. Tulinsky, *An ambiguous structure of a DNA 15-mer thrombin complex.* Acta Crystallogr D Biol Crystallogr, 1996. **52**(Pt 2): p. 272-82.

153. Macaya, R.F., et al., *Thrombin-binding DNA aptamer forms a unimolecular quadruplex structure in solution.* Proc Natl Acad Sci U S A, 1993. **90**(8): p. 3745-9.

154. Kelly, J.A., J. Feigon, and T.O. Yeates, *Reconciliation of the X-ray and NMR structures of the thrombin-binding aptamer d(GGTTGGTGTGGTTGG).* J Mol Biol, 1996. **256**(3): p. 417-22.

155. Griffin, L.C., *In vivo anticoagulant properties of a novel nucleotide-based thrombin inhibitor and demonstration of regional anticoagulation in extracorporeal circuits.* Blood, 1993. **81**: p. 3271-3276.

156. Schwienhorst, A., *Direct thrombin inhibitors – a survey of recent developments.* Cellular and Molecular Life Sciences CMLS, 2006. **63**(23): p. 2773-2791.

157. Wagner-Whyte, J., et al., *Discovery of a potent, direct thrombin inhibiting aptamer.* 2007: Journal of Thrombosis nad Haemostasis. p. P-S-067.

158. *KAPA Library Quanitification Technical Guide*. 2014, KAPA Biosystems.

159. *Illumina Sequencing Technology*. 2010, Illumina.

160. Nilsson, B., M.K. Horne, 3rd, and H.R. Gralnick, *The carbohydrate of human thrombin: structural analysis of glycoprotein oligosaccharides by mass spectrometry.* Arch Biochem Biophys, 1983. **224**(1): p. 127-33.

161. Shinsky, S.A., et al., *Biochemical Reconstitution and Phylogenetic Comparison of Human SET1 Family Core Complexes Involved in Histone Methylation.* Journal of Biological Chemistry, 2015.

162. Shinsky, S.A., et al., *A non-active site SET domain surface crucial for the interaction of MLL1 and the RbBP5-ASH2L heterodimer within MLL family core complexes.* Journal of molecular biology, 2014. **426**(12): p. 2283-2299.

163. Patel, A., et al., *On the mechanism of multiple lysine methylation by the human mixed lineage leukemia protein-1 (MLL1) core complex.* J Biol Chem, 2009. **284**(36): p. 24242-56.

164. Sefah, K., et al., *Development of DNA aptamers using Cell-SELEX*. Nat Protoc, 2010. **5**(6): p. 1169-85.

165. Ruano, G. and K.K. Kidd, *Modeling of heteroduplex formation during PCR from mixtures of DNA templates.* Genome Research, 1992. **2**(2): p. 112-116.

166. Dabney, J. and M. Meyer, *Length and GC-biases during sequencing library amplification: a comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries.* Biotechniques, 2012. **52**(2): p. 87-94.

167. El-Sagheer, A.H. and T. Brown, *Click Nucleic Acid Ligation: Applications in Biology and Nanotechnology.* Accounts of Chemical Research, 2012. **45**(8): p. 1258-1267.

168. Kocalka, P., A.H. El-Sagheer, and T. Brown, *Rapid and efficient DNA strand cross-linking by click chemistry.* Chembiochem, 2008. **9**(8): p. 1280-5.

169. El-Sagheer, A.H., et al., *Biocompatible artificial DNA linker that is read through by DNA polymerases and is functional in Escherichia coli.* Proceedings of the National Academy of Sciences, 2011. **108**(28): p. 11338-11343.

170. *Protocol: Click-Chemistry Labeling of Oligonucleotides and DNA*. 2015, Lumiprobe.

171. *Con A Sepharose 4B Instruction Manual.* 2011, GE Healthcare.

172. Brodolin, K., in *DNA-protein Interactions: A Practical Approach*. 2000, Oxford University Press.

173. Lu, K., et al., *Structural characterization of formaldehyde-induced cross-links between amino acids and deoxynucleosides and their oligomers.* J Am Chem Soc, 2010. **132**(10): p. 3388-99.

174. Barker, S., M. Weinfeld, and D. Murray, *DNA–protein crosslinks: their induction, repair, and biological consequences.* Mutation Research/Reviews in Mutation Research, 2005. **589**(2): p. 111-135.

175. Niranjanakumari, S., et al., *Reversible cross-linking combined with immunoprecipitation to study RNA-protein interactions in vivo.* Methods, 2002. **26**(2): p. 182-90.

176. Robertson, G., et al., *Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing.* Nat Meth, 2007. **4**(8): p. 651-657.

177. Research, G., *Glen Report: Pyrrolo-C - a fluorescent nucleoside base analogue that codes efficiently as C*. 2003.

178. Vincent, A. and K. Scherrer, *A rapid and sensitive method for detection of proteins in polyacrylamide SDS gels: staining with ethidium bromide.* Mol Biol Rep, 1979. **5**(4): p. 209-14.

179. Boggy, G.J. and P.J. Woolf, *A Mechanistic Model of PCR for Accurate Quantification of Quantitative PCR Data.* PLoS ONE, 2010. **5**(8): p. e12355.

180. SantaLucia, J. and G.J. Boggy. *Counting PCR: A new method to obtain absolute DNA copy number without a standard curve*. DNASoftware.

181. *Octet RED96 System: Superior Quantitation and Kinetics Performance with Increased Cost Efficiency*. 2010, ForteBio.

182. *Streptavidin (SA) Biosensors: For Kinetic Analysis, Screening, and Quantitation of Most Proteins*. 2009: ForteBio.

183. *Dip and Read Amine Reactive Second-Generation (AR2G) Biosensors*. 2011, ForteBio.

184. Zaccolo, M., et al., *An approach to random mutagenesis of DNA using mixtures of triphosphate derivatives of nucleoside analogues.* J Mol Biol, 1996. **255**(4): p. 589-603.

185. Petrie, K.L. and G.F. Joyce, *Deep sequencing analysis of mutations resulting from the incorporation of dNTP analogs.* Nucleic Acids Research, 2010. **38**(22): p. 8095-8104.

186. Nutiu, R., et al., *Direct visualization of DNA affinity landscapes using a high-throughput sequencing instrument.* Nature biotechnology, 2011. **29**(7): p. 659-664.

Bibliographic Information

NAME OF AUTHOR: Caitlin M. Miller

PLACE OF BIRTH: Newport News, Virginia, USA

DATE OF BIRTH: January 23, 1988

DEGRESS AWARDED:

Bachelor of Science in Chemistry, 2009, Seton Hall University, South Orange, NJ, USA

Master of Philosophy in Chemistry, 2011, Syracuse University, Syracuse, NY, USA

AWARDS AND HONORS:

University Fellowship, Syracuse University, 2009 – 2012

Clare Booth Luce Scholar, Seton Hall University, 2006 - 2009