High-throughput methods in aptamer discovery and analysis

Kyle H. Cole[a], **Andrej Lupták**[a,b,c,*]

[a]Department of Molecular Biology and Biochemistry, University of California, Irvine, CA, United States
[b]Department of Pharmaceutical Sciences, University of California, Irvine, CA, United States
[c]Department of Chemistry, University of California, Irvine, CA, United States
*Corresponding author: e-mail address: aluptak@uci.edu

Contents

1. Introduction 2
2. Post-selection sequence analysis 3
   2.1 Clustering 5
   2.2 Motif searching 6
   2.3 Mutagenesis 8
3. High-throughput characterization methods 10
   3.1 On-chip analysis 12
4. Conclusions 14
References 14

Abstract

Aptamers are small, functional nucleic acids that bind a variety of targets, often with high specificity and affinity. Genomic aptamers constitute the ligand-binding domains of riboswitches, whereas synthetic aptamers find applications as diagnostic and therapeutic tools, and as ligand-binding domains of regulatory RNAs in synthetic biology. Discovery and characterization of aptamers has been limited by a lack of high-throughput approaches that uncover the target-binding domains and the biochemical properties of individual sequences. With the advent of high-throughput sequencing, large-scale analysis of *in vitro* selected populations of aptamers (and catalytic nucleic acids, such as ribozymes and DNAzmes) became possible. In recent years the development of new experimental approaches and software tools has led to significant streamlining of the selection–pool analysis. This article provides an overview of post-selection data analysis and describes high–throughput methods that facilitate rapid discovery and biochemical characterization of aptamers.

Methods in Enzymology
ISSN 0076-6879
https://doi.org/10.1016/bs.mie.2019.02.009

1. Introduction

Aptamers are nucleic acids that can recognize ligands ranging from ions and small molecules, to more complex targets such as proteins and whole cells (Wu & Kwon, 2016). Because of their versatility and relative ease of isolation from high-diversity libraries, aptamers have found applications in diagnostics and drug development. A classic example of an FDA-approved therapeutic aptamer is the anti-vascular endothelial growth factor (anti-VEGF) aptamer used in the treatment of ocular vascular disease (Ng et al., 2006). Aptamer-based therapeutics compete within the niche currently filled by biologics such as monoclonal antibodies, but unlike monoclonal antibodies, they do not suffer from batch-to-batch differentiation (Shaughnessy, 2012). *In vitro* selection of nucleic-acid aptamers has gained popularity due to these applications; however, aptamer discovery is limited by the necessity of post-selection biochemical analysis (Jijakli et al., 2016). High-throughput aptamer discovery methods aim to resolve these issues through combinatorial approaches to identify ligand-targeting sequences while simultaneously determining biochemical and structurally-relevant information.

A traditional *in vitro* selection (SELEX) experiment used target molecules conjugated to solid media to separate bound molecules from highly-diverse pools of sequences. After successive rounds of selection, enriched pools were collected and cloned into *Escherichia coli* (*E. coli*) plasmids. To identify individual aptamers, the plasmids were sequenced by chain-termination methods, such as Sanger sequencing (Griffin, Toole, & Leung, 1993; Morris, Jensen, Julin, Weil, & Gold, 1998). This strategy proved laborious, because a relatively high number of individual colonies had to be sequenced to determine the distribution of selected motifs. If a specific sequence or structural motif dominated the enriched pool, the likelihood of identifying less-abundant motifs was low, due to the relatively low-throughput of this sequencing approach.

As the cost of high-throughput sequencing (HTS) decreased, deep sequencing was introduced to *in vitro* selection protocols, with the intent of improving aptamer discovery and increasing data validation through improved sequencing depth (Schütze et al., 2011). The first applications of HTS to *in vitro* selections aimed to identify new binding targets of transcription factors (Jolma et al., 2010; Slattery et al., 2011; Zhao & Stormo, 2011; Zykovich, Korf, & Segal, 2009). At the same time, Zimmerman et al.

developed a HTS method of identifying naturally-occurring RNA aptamers encoded in *E. coli* genomic DNA (Zimmermann, Gesell, Chen, Lorenz, & Schroeder, 2010), and other laboratories built methodology to select for growth-factor-specific DNA aptamers (Cho et al., 2010) and RNA inhibitors for HIV reverse transcriptase (Ditzler et al., 2013). With increased sequencing depth and computational methods, the fitness landscape of an *in vitro* selection could be visualized (Pitt & Ferré-D'Amaré, 2010), and evolutionary steps tracked between rounds of selection toward functional RNAs (Jiménez, Xulvi-Brunet, Campbell, Turk-MacLeod, & Chen, 2013; Neveu, Kim, & Benner, 2013). While these methods incorporated HTS for more expedient aptamer discovery or ribozyme sequence analysis, recent high-throughput discovery methods have combined aptamer discovery with determination of rate or equilibrium constants, as well as probing of ligand-induced conformational changes, to streamline the biochemical analysis necessary for aptamer motif definition and activity validation (Fig. 1). The objective of the following sections is to provide an overview of post-selection sequence analysis methods and high-throughput aptamer discovery methods.

2. Post-selection sequence analysis

Post-selection analysis of deep-sequenced pools can become overwhelming due to the sheer quantity of available sequences, presumably enriched for functional molecules. Software development aimed at tackling this issue has progressed significantly over the last decade, with tools intended to reduce the burden of deep sequencing data manipulation by providing analysis of sequence enrichment and sorting of aptamer sequences based on genotypic familiarity. Early high-throughput sequencing adopters used analysis packages designed for genomic sequencing data manipulation, in addition to homebrewed UNIX scripts utilizing the *sed*, *awk*, and *grep* commands (Hoon, Zhou, Janda, Brenner, & Scolnick, 2011). The availability of these custom-built scripts was limited, and scripts were typically built to answer specific experimental questions and were not intended for general enriched-pool workup. More recently, analysis of sequencing data for functional aptamers has greatly improved and the benefits of HTS have become realized; for example, for SELEX-specific constant region trimming and determination of aptamer frequencies in enriched pools (Hamada, 2018).

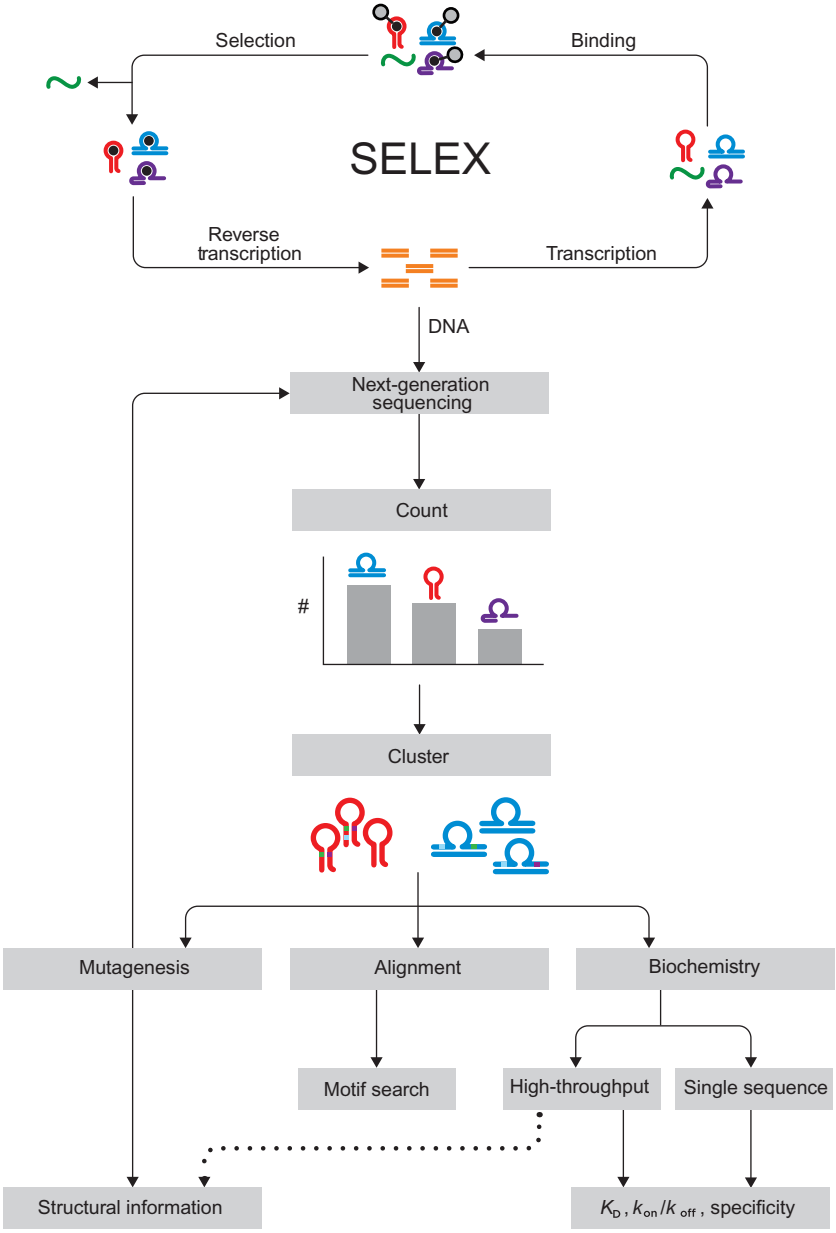

Fig. 1 Generalized workflow of *in vitro* selection (SELEX) of RNA aptamers and post-selection analysis. After several rounds of selection, the enriched pool is amplified and submitted for high-throughput sequencing. After sequencing, the reads are counted and clustered. Clustered sequences provide the basis for downstream analysis, such as mutagenesis, alignment and sequence-motif discovery, and biochemistry.

(Continued)

2.1 Clustering

In silico parsing of aptamer sequences provides the foundation of biochemical analysis by identifying sequence families enriched through selective pressure. The analysis typically begins by minimizing and sorting redundant sequencing data by identifying unique sequences and sorting them based on shared sequence characteristics. One of the initial high-throughput sequencing analysis packages specific for *in vitro* selection was Sequence Evolution With Adaptive Landscape (SEWAL), which assesses the overall fitness landscape of *in vitro* selection pools for the purpose of performing functional genotypic analysis (Pitt, Rajapakse, & Ferré-D'Amaré, 2010). SEWAL produces 3D frequency plots using a sorting algorithm to determine sequence patterns and visualizes changes in the sequencing space due to selective pressure across consecutive rounds. Identification of changes due to selective pressure can be used to predict evolutionary paths that individual sequences follow toward improved fitness.

Similarly, FASTAptamer and AptaCluster compress pool data into nonredundant sequence sets (count) and sort sequences based on similarities in genotype (clustering) (Alam, Chang, & Burke, 2015; Hoinka, Berezhnoy, Sauna, Gilboa, & Przytycka, 2014). The FASTAptamer package of scripts provides a simple, command-line data workup package that can be installed on any system capable of running Perl binaries, whereas AptaCluster requires a C++ compiler for installation, which can be difficult for inexperienced users. Both packages use counting to rank sequences based upon frequency, aiding in the identification of sequences enriched throughout the selection to assist in determining the complexity of the enriched pool.

As mentioned above, clustering sorts sequences into families of sequences which share a similar genotype. Clustering begins by separating abundant unique sequences into individual clusters. Less abundant sequences that are similar to the seed sequence of a given cluster, but vary by mutations (including insertions and deletions; indels), are then sorted. Alignments of individual clusters, using tools such as MUSCLE or MAFFT,

Fig. 1—Cont'd Mutagenesis of enriched pools, followed by reselection and next-generation sequencing, reveals functionally critical (immutable) positions, as well as structural information in the form of sequence covariation of base-paired positions. Finally, enriched sequences identified through counting and clustering can be individually tested for target binding specificity, and determination of binding constants, such as K_D and k_{on}/k_{off}. High-throughput analysis methods yield these kinetic and thermodynamic constants for many sequences simultaneously, with some methods, such as Apta-Seq (Abdelsayed et al., 2017), revealing structural information as well.

provide visualization of allowed mutations and the output of these tools is useful in determining sequence consensus in addition to potential evolutionary relationships among clustered sequences derived from different selection rounds (Ameta, Winz, Previti, & Jäschke, 2014; Edgar, 2004; Katoh, Misawa, Kuma, & Miyata, 2002). And while clustering leads to increased understanding of the genotypic variation that exists within families of aptamers, confirmation of aptamer activity relies on further biochemical analysis.

2.2 Motif searching

Identification of motif sequences from enriched pools provides insight in to functional binding targets of aptamers. Methods aimed at identifying transcription factor binding sites, such as ChIP-seq, contributed to the development of DNA motif analysis software packages, such as the MEME family of tools, which identify reoccurring sequence motifs enriched throughout genomic sequencing data (Bailey et al., 2009). However, identification of motifs based on sequence alone does not necessarily coincide with target binding. BEEML, and similar methods, account for binding by utilizing a position weight matrix (PWM) to determine motifs based on energy models which determine the contribution of each base-pair to transcription factor binding (Jolma et al., 2010; Zhao & Stormo, 2011). These computational methods identify motifs based on the energetic contributions of each nucleotide position, improving the likelihood of identifying binding sequences.

RNA aptamers, on the other hand, maintain complex structures, requiring motif identification software that incorporates secondary structure prediction.

To address this issue, Backofen and coworkers developed MEME in RNAs Including secondary Structure (MEMERIS), which built on MEME by incorporating high throughput analysis capabilities in conjunction with secondary structure prediction, but is restricted to motifs within predicted single-stranded loop and bulge regions of RNA aptamer sequences (Hiller, Pudimat, Busch, & Backofen, 2006). RNAcontext improved upon the motif analysis landscape by including loops, bulges, and stems into its predictions (Kazan, Ray, Chan, Hughes, & Morris, 2010). Because RNAcontext distinguishes between these secondary structures, it can also predict the preferred conformation of the identified motif within the aptamer.

Similarly, AptaMotif and APTANI, which use a minimum free energy (MFE) approach for structure prediction in conjunction with iterative sampling and sequence alignments, have been developed to determine potential motifs across multiple input files (i.e., selection rounds) (Caroli, Taccioli, De La Fuente, Serafini, & Bicciato, 2016; Hoinka, Zotenko, Friedman, Sauna, & Przytycka, 2012). However, both of these methods are limited to identification of dominant motifs—an outcome that may not be desirable in selections for complex targets that are expected to yield multiple motifs, or in genomic SELEX, in which an exhaustive mapping of all aptamers is often preferred. AptaTrace overcomes this problem by tracing motifs through multiple rounds of selection data and testing whether these regions undergo selection toward secondary structure, such as hairpin loops or bulges (Dao et al., 2016). More recently, AptaSuite was developed as an all-encompassing software package, which includes AptaTrace and AptaCluster, and is the first open-source package designed for selection schemes to feature a graphical user interface (GUI) (Hoinka, Backofen, & Przytycka, 2018). AptaSuite can be installed on any system that runs Java Runtime Environment (JRE), providing ease-of-use to less-experienced users and removing the hassle of installing packages for adapter trimming and motif prediction.

The above-mentioned high-throughput strategies predict motifs based on the frequency of reoccurrence between rounds of selection. Alternatively, covariance model (CM) based algorithms predict secondary-structures and identify motifs based on the dependent variability of nucleotides in a given sequence (covariance). One example is CMFinder, which applies a covariance-probabilistic model to predict motifs from sequences that are dissimilar and unaligned (Yao, Weinberg, & Ruzzo, 2006). CMFinder can align and predict motifs from any input of sequences, whether *in vitro* evolved or genomic, and the output is used in the identification of sequence homology. Similarly, the Infernal package constructs a CM based on an alignment of RNAs and searches for genomic RNA homologs based on primary sequence and, most importantly, conserved-secondary structure (Nawrocki & Eddy, 2013). The output of Infernal can be applied to CMFinder for improved motif prediction in the case of genomic aptamers. In contrast, RNArobo uses a context-based motif searching algorithm for identification of novel motifs when no known homology has been previously annotated (Rampášek, Jimenez, Lupták, Vinař, & Brejová, 2016). RNArobo parses sequence searches based upon an input descriptor that defines a simplified motif map and individual structural elements; for example, a combination of secondary structure and binding-loop sequence. The methods employed by RNArobo

are expedient in determining sequences containing difficult-to-predict structural elements, such as complex pseudoknots, adding to its effectiveness in discovering genomic aptamers from *in vitro* selections.

In the case of genomic selections, sequence data require further manipulation to identify conservation and homology, in addition to aptamer *loci*. Initial sequence alignment to reference genomes can be accomplished through multiple methods. For example, Bowtie2 is beneficial for alignments of short reads obtained from selections to a reference genome (Langmead & Salzberg, 2012), while Integrated Genome Browser (IGB) provides a user interface for alignments to annotated-reference genomes, and provides additional tools including sequence pileups for visualization of consensus sequences (Freese, Norris, & Loraine, 2016). These basic alignment methods are useful when comparing HTS data to single-reference genomes but are less efficient when comparing multiple species. Packages such as the Basic Local Alignment Search Tool (BLAST) or HMMER3, allow users to query custom- or downloadable- sequence databases for more advanced conservation and homology searches (Camacho et al., 2009; Eddy, 2011). BLAST provides a powerful webserver and command-line based software package that delivers a well-rounded source for alignments, conservation, and homology, but can be nonintuitive to use. HMMER3 is an alternative homology identification package that uses a hidden Markov model (HMM) to improve upon sequence identification, with improved efficiency compared to BLAST. Information obtained from reference-genome alignments, conservation, and homology provides a basis for downstream biological characterization of genomic aptamers, such as expression analysis, *in vivo* target binding, and potential regulatory functions.

2.3 Mutagenesis

Candidate aptamer sequences that dominate a selected pool bias it toward enriched genotypes and this genotypic bias restricts possible nucleotide variation across the pool, which limits structurally-relevant information. Upon initial selection, aptamer sequences may be constrained within their given sequence space, because mutations introduced by the polymerase enzymes during the *in vitro* selection process may not be frequent enough to thoroughly sample the permissive sequence variation around a founding aptamers sequence. Mutagenesis of an enriched pool introduces such sequence variation and subsequent selection provides alternative genotypic

outcomes (Jiménez et al., 2013). Consequently, mutagenesis followed by reselection, is a useful strategy to overcome the lack of sequence diversity, allowing a more thorough characterization of the functional sequence space, in addition to potentially revealing functionally superior variants of the selected aptamers.

Mutagenesis of an enriched pool can be accomplished through mutagenic PCR or error-prone PCR methods (EP-PCR). Mutagenic PCR is facilitated by nucleoside analogs, such as 8-oxo-dG and dPTP, to introduce templated mismatches. 8-oxo-dG and dPTP can be utilized by *Taq* polymerase allowing for introduction of the analogs into standard PCR reaction mixtures in excess, and the rate at which mutations are introduced can be tuned by the ratio of the analogs to standard dNTPs (Zaccolo, Williams, Brown, & Gherardi, 1996). EP-PCR takes advantage of the low fidelity of *Taq* polymerase to introduce mutations through doubling of the pool (Wilson & Keefe, 2001). PCR is inherently error-prone and contributes to the genotypic variation of *in vitro* selection; however, by increasing the concentration of Mg^{2+} and introducing Mn^{2+} metal ions, the error rate of *Taq* polymerase increases from 0.02% to 0.066% per nucleotide position when coupled with disproportionate dNTP concentrations (Cadwell & Joyce, 1992).

Sequence covariation is a sought-after outcome of mutagenic selection and provides benefit to motif and secondary structure prediction. Due to the lack of sequence diversity of selected pools, covariation of sequences that form the secondary structure elements of functional nucleic acids tends be low. As mentioned previously, software such as CMFinder and Infernal utilize covariance models for motif and homology prediction (Nawrocki & Eddy, 2013; Yao et al., 2006). Each tool utilizes covariance in secondary structure prediction strategies by identifying nucleotide positions which mutate dependently. Compensatory evolution of nucleotide positions is most likely the result of structural motifs which rely on the interaction of the two positions to maintain functionality (Parsch, Braverman, & Stephan, 2000). Secondary-structure prediction based on covariation data is accomplished with software such as ViennaRNA's RNAalifold, which can take several hundred aligned sequences as input through the webserver and command-line versions (Bernhart, Hofacker, Will, Gruber, & Stadler, 2008). Sequence covariation can be also revealed through alignment of homologous sequences using resources like the BLAST-like alignment tool (BLAT) (Kent, 2002), but mutagenic selection has the potential to provide more diversity, which is ideal for structural and motif prediction.

3. High-throughput characterization methods

While deep sequencing has increased the diversity of potential aptamers from selections by aiding in the identification of less abundant sequences, validation relies on biochemical confirmation of ligand binding. Given the potentially high number of functional sequences in an enriched pool, binding assays of aptamers to immobilized ligands can prove tedious, unless the pool is dominated by a small number of motifs (Sassanfar & Szostak, 1993). Moreover, traditional binding assays do not provide insights into conformational dynamics upon ligand binding or the location of binding domains. High-throughput methods of aptamer discovery aim to address these issues through combinatorial measurements of binding kinetics or thermodynamics, as well as structural probing, followed by HTS. These methods intend to alleviate the strain of *in vitro* and *in silico* selection workup by identifying ligand-associated motifs and determining binding or rate constants.

Sequence enrichment does not necessarily equate to stronger or more specific binding; therefore, identification of high-affinity binders in a pool traditionally requires binding or rate constant measurements of individual sequences. To increase the experimental throughput of these measurements, one approach characterizes of functional RNAs by probing their dynamic properties against a titration of ligand. The technique, Apta-Seq, utilizes a combinatorial, multiplexed approach focused on RNA chemical modification and deep HTS to evaluate structural interaction of aptamers with their target ligand (Abdelsayed et al., 2017). Apta-Seq utilizes selective 2′-hydroxyl acylation with primer extension (SHAPE) to determine whether a 2′-hydroxyl is solvent-accessible to allow acylation by a SHAPE reagent, such as 2-(azidomethyl)nicotinic acid acyl imidazole (NAI-N_3), under variable ligand concentrations (Merino, Wilkinson, Coughlan, & Weeks, 2005; Spitale et al., 2015). Apta-Seq is based on an established workflow known as SHAPE-seq to map HTS data of reverse transcriptase (RT) stops caused by acylation of 2′-hydroxyls of pool RNAs, leading to truncated cDNAs with heterogenous 3′ ends (Lucks et al., 2011). The heterogenous 3′ ends of the cDNAs are resolved by introducing an RT primer with a 5′ overhang and self-ligating the cDNA by CircLigase (Lucigen). Together with the RT primer sequence, the introduced 5′ overhang can be subsequently used for priming a PCR amplification reaction, and submitted to HTS. RT stop sites are mapped using ShapeFinder,

a package used to quantitatively map RT stops at nucleotide resolution and when a ligand titration is used, the stops uncover RNA conformational changes due ligand binding (Vasa, Guex, Wilkinson, Weeks, & Giddings, 2008). When this method was applied to a human-genomic *in vitro* selection for ATP-binders, it revealed three novel human-genome-derived ATP-binding RNA aptamers, and confirmed the existence of two previously identified aptamers (Abdelsayed et al., 2017; Vu et al., 2012). These results demonstrated the robustness of the method to perform structural analysis in a high-throughput pipeline and allowed determination of dissociation rate constants, K_Ds, for selected pools while eliminating single-clone biochemical assessment. This general workflow provides a unique framework which can be modified for all nucleic acids, including DNA and XNA aptamers, if base-specific modifications such as dimethyl sulfate (DMS), 1-ethyl-3-(3-dimethylaminopropyl) carbodiimide (EDC), or nicotinoyl azide (NAz) probing are introduced (Feng et al., 2018; Kwok, Ding, Tang, Assmann, & Bevilacqua, 2013; Mitchell et al., 2018; Wang, Sexton, Culligan, & Simon, 2018; Zinshteyn et al., 2018).

Identification of structural motifs is critical for the determination of ligand binding domains, assisting in the minimization of aptamer sequences derived from a given pool. Previously, natural RNA aptamers, the ligand-binding domains of riboswitches, have mostly been discovered via computational investigation for sequence and structural homology (Barrick & Breaker, 2007). However, this approach limits the discovery of novel, naturally occuring aptamers to those which share consensus with characterized sequences or sequence models, such as those modeled with CMFinder. To alleviate this limitation, Parallel Analysis of RNA Conformations Exposed to Ligand Binding (PARCEL) was developed as a high-throughput method of identifying RNA-ligand interactions in whole transcriptomes (Tapsin et al., 2018). The PARCEL workflow uses three methods of RNA-ligand interaction assessment in parallel, followed by next-generation sequencing. RNA-footprinting with double-strand specific RNase V1 and single-strand specific S1 nuclease is performed in parallel in the presence or absence of a target metabolite. Additionally, the protocol applies partial hydrolysis of transcripts, in-line probing, to determine transcript flexibility (Regulski & Breaker, 2008; Soukup & Breaker, 1999). The methodology relies on the assumption that the target ligand provides protection to the RNA when bound. Transcripts treated in the presence or absence of metabolite are then sequenced and mapped to a reference genome. Reads are normalized

to ligand conditions and while PARCEL utilizes nuclease cleavage and in-line probing, both methods could be applied in parallel with acylation in Apta-Seq to provide additional structural insight and K_Ds.

A different approach to determining the strength of aptamer-ligand interactions is based on the measurement of binding rate constants. High-throughput sequencing kinetics (HTSK), is a an efficient method for the determination of k_{on} and k_{off} rate constants for pools of mRNA-peptide fusions (Jalali-Yazdi, Lai, Takahashi, & Roberts, 2016; Roberts & Szostak, 1997). HTSK provides k_{on}/k_{off} rates for the majority of sequences in a given pool through a time-point analysis protocol. To determine the association rate constants, k_{on}, an enriched pool is added to ligands immobilized on magnetic beads and at various time points, bead fractions are removed, washed, and the isolated sequences are amplified and prepared for next-generation sequencing. The fractional composition of a given sequence can be determined by measuring the frequency of that sequence at a given time point, and the amount of bound aptamer can be determined by multiplying the fractional composition by the total counts of the bound pool at each time point. Dissociation rate constants, k_{off}, are determined following the on-rate time points by washing the beads in the presence of excess ligand and collecting fractions at various time-points. Fractional composition of sequences is determined similarly to the on-rates and K_Ds can be calculated from the rate constants, assuming a simple binding model.

RNA Bind-n-Seq is a similar approach to HTSK, used to identify RNA-protein interactions by combining high-throughput *in vitro* selection of RNA with rate constant measurements (Lambert et al., 2014). A randomized pool of RNA is incubated with varying concentrations of streptavidin-tagged RNA binding proteins (RBPs), which can be pulled down using biotinylated beads, allowing the bound RNA to be eluted and sequenced. Association and dissociation rate constants can be determined from these sequenced data, assuming complete pull-down of the tagged RBPs; however, this method is limited to proteins or target molecules that can be tagged for affinity-binding assays.

3.1 On-chip analysis

The Illumina sequencing platform has been utilized for deep sequencing of whole genomes, single-cell RNA-seq, and for understanding epigenetics through bisulphite sequencing, among many other applications

(Lippert et al., 2017; Raine, Manlig, Wahlberg, Syvänen, & Nordlund, 2017; Ziegenhain et al., 2017). After sequencing, the DNA-displayed chips can be used for further experiments that utilize the strong fluorescent signal, good physical separation, and known sequences of the DNA polonies generated by the sequencing method. The Illumina platform's versatility for aptamer discovery was first illustrated with high-throughput sequencing fluorescent ligand interaction profiling (HITS-FLIP) (Nutiu et al., 2011). HITS-FLIP quantitatively measures DNA-protein affinity through changes in fluorescent ligand associations of millions of DNA clusters bound to the Illumina flowcell. On/off rates can be determined by varying the protein concentrations introduced across the flowcell and measuring subsequent changes in fluorescence.

Similarly, the chip-hybridized associated-mapping platform (CHAMP) takes advantage of used MiSeq chips with linked DNA clusters to determine DNA-protein interactions (Jung et al., 2017). The DNA clusters are reamplified to remove associated fluorescent nucleotides and a fluorescent-oligonucleotide probe is hybridized to the clusters and used as a reference marker for the DNA sequences. Fluorescent proteins are incubated in varying concentrations, imaged, and associations between the binding experiment and the sequences of individual colonies are made to the DNA sequence through imaging and software analysis. In another example of this approach, two groups built on HITS-FLIP for high-throughput RNA aptamer discovery (Buenrostro et al., 2014; Tome et al., 2014). Both methods use diverse single-stranded DNA libraries which contain T7 promoters to initiate transcription on the sequencer flow cell, and employ an approach to halt transcription with the intent on maintaining the association of the transcript with the DNA cluster. Fluorescently-labeled proteins that bind the halted transcripts can be detected through fluorescent imaging of the flowcell. This approach allows identification of RNA aptamers that bind the protein ligand through the sequence of the associated DNA. Additionally, genotypic variations can be directly compared, decreasing the need for downstream mutagenic analysis and thus providing covariation information for structure prediction. These methods are currently limited to ligands which can be fluorescently labeled, such as target proteins labeled with fluorescent dyes or fused to fluorescent proteins. On the other hand, many targets of SELEX are modified for bead-binding during *in vitro* selection experiments, and this chemistry can often be exploited for flourescent labeling allowing for optical detection of binding on HTS chips.

4. Conclusions

Increased interest in nucleic-acid aptamers has, in recent years, led to the development of improved high-throughput methods of aptamer discovery. High-throughput methods that incorporate biochemical assessment of individual sequences decrease the time requirement of aptamer identification and increases the likelihood of uncovering larger diversity of functional sequences. Current methods have been used to uncover human ATP-binding RNA aptamers, analyze MS2 aptamers, and identify CRISPR-Cas complex specificity (Abdelsayed et al., 2017; Buenrostro et al., 2014; Jung et al., 2017). These techniques are valuable in the advancement of *in vitro* selection and aid in understanding of the evolution of functional RNAs from random sequence space. Continued improvement to these methods will support the growth of aptamer-based drug development and should fill the shortcomings in current therapeutic biologics.

References

Abdelsayed, M. M., Ho, B. T., Vu, M. M. K., Polanco, J., Spitale, R. C., & Lupták, A. (2017). Multiplex aptamer discovery through apta-seq and its application to ATP aptamers derived from human-genomic SELEX. *ACS Chemical Biology*, *12*(8), 2149–2156. https://doi.org/10.1021/acschembio.7b00001.

Alam, K. K., Chang, J. L., & Burke, D. H. (2015). FASTAptamer: A bioinformatic toolkit for high-throughput sequence analysis of combinatorial selections. *Molecular Therapy Nucleic Acids*, *4*(3), e230. https://doi.org/10.1038/mtna.2015.4.

Ameta, S., Winz, M.-L., Previti, C., & Jäschke, A. (2014). Next-generation sequencing reveals how RNA catalysts evolve from random space. *Nucleic Acids Research*, *42*(2), 1303–1310. https://doi.org/10.1093/nar/gkt949.

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., et al. (2009). MEME suite: Tools for motif discovery and searching. *Nucleic Acids Research*, *37*(1), W202–W208. https://doi.org/10.1093/nar/gkp335.

Barrick, J. E., & Breaker, R. R. (2007). The distributions, mechanisms, and structures of metabolite-binding riboswitches. *Genome Biology*, *8*(11), R239. https://doi.org/10.1186/gb-2007-8-11-r239.

Bernhart, S. H., Hofacker, I. L., Will, S., Gruber, A. R., & Stadler, P. F. (2008). RNAalifold: Improved consensus structure prediction for RNA alignments. *BMC Bioinformatics*, *9*(1), 474. https://doi.org/10.1186/1471-2105-9-474.

Buenrostro, J. D., Araya, C. L., Chircus, L. M., Layton, C. J., Chang, H. Y., Snyder, M. P., et al. (2014). Quantitative analysis of RNA-protein interactions on a massively parallel array reveals biophysical and evolutionary landscapes. *Nature Biotechnology*, *32*, 562. https://doi.org/10.1038/nbt.2880.

Cadwell, R. C., & Joyce, G. F. (1992). Randomization of genes by PCR mutagenesis. *PCR Methods and Applications*, *2*(1), 28–33. https://doi.org/10.1101/gr.2.1.28.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, *10*, 421. https://doi.org/10.1186/1471-2105-10-421.

Caroli, J., Taccioli, C., De La Fuente, A., Serafini, P., & Bicciato, S. (2016). APTANI: A computational tool to select aptamers through sequence-structure motif analysis of HT-SELEX data. *Bioinformatics*, *32*(2), 161–164. https://doi.org/10.1093/bioinformatics/btv545.

Cho, M., Xiao, Y., Nie, J., Stewart, R., Csordas, A. T., Oh, S. S., et al. (2010). Quantitative selection of DNA aptamers through microfluidic selection and high-throughput sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(35), 15373–15378. https://doi.org/10.1073/pnas.1009331107.

Dao, P., Hoinka, J., Takahashi, M., Zhou, J., Ho, M., Wang, Y., et al. (2016). AptaTRACE elucidates RNA sequence-structure motifs from selection trends in HT-SELEX experiments. *Cell Systems*, *3*(1), 62–70. https://doi.org/10.1016/j.cels.2016.07.003.

Ditzler, M. A., Lange, M. J., Bose, D., Bottoms, C. A., Virkler, K. F., Sawyer, A. W., et al. (2013). High-throughput sequence analysis reveals structural diversity and improved potency among RNA inhibitors of HIV reverse transcriptase. *Nucleic Acids Research*, *41*(3), 1873–1884. https://doi.org/10.1093/nar/gks1190.

Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLoS Computational Biology*, 7(10), e1002195. https://doi.org/10.1371/journal.pcbi.1002195.

Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797. https://doi.org/10.1093/nar/gkh340.

Feng, C., Chan, D., Joseph, J., Muuronen, M., Coldren, W. H., Dai, N., et al. (2018). Light-activated chemical probing of nucleobase solvent accessibility inside cells. *Nature Chemical Biology*, *14*, 276. https://doi.org/10.1038/nchembio.2548.

Freese, N. H., Norris, D. C., & Loraine, A. E. (2016). Integrated genome browser: Visual analytics platform for genomics. *Bioinformatics*, *32*(14), 2089–2095. https://doi.org/10.1093/bioinformatics/btw069.

Griffin, L. C., Toole, J. J., & Leung, L. L. K. (1993). The discovery and characterization of a novel nucleotide-based thrombin inhibitor. *Gene*, *137*(1), 25–31. https://doi.org/10.1016/0378-1119(93)90247-Z.

Hamada, M. (2018). In silico approaches to RNA aptamer design. *Biochimie*, *145*, 8–14. https://doi.org/10.1016/j.biochi.2017.10.005.

Hiller, M., Pudimat, R., Busch, A., & Backofen, R. (2006). Using RNA secondary structures to guide sequence motif finding towards single-stranded regions. *Nucleic Acids Research*, *34*(17), e117. https://doi.org/10.1093/nar/gkl544.

Hoinka, J., Backofen, R., & Przytycka, T. M. (2018). AptaSUITE: A full-featured bioinformatics framework for the comprehensive analysis of aptamers from HT-SELEX experiments. *Molecular Therapy—Nucleic Acids*, *11*, 515–517. https://doi.org/10.1016/j.omtn.2018.04.006.

Hoinka, J., Berezhnoy, A., Sauna, Z. E., Gilboa, E., & Przytycka, T. M. (2014). AptaCluster—A method to cluster HT-SELEX aptamer pools and lessons from its application. *Research in Computational Molecular Biology*, *8394*, 115–128. https://doi.org/10.1007/978-3-319-05269-4_9.

Hoinka, J., Zotenko, E., Friedman, A., Sauna, Z. E., & Przytycka, T. M. (2012). Identification of sequence-structure RNA binding motifs for SELEX-derived aptamers. *Bioinformatics*, *28*(12), i215–i223. https://doi.org/10.1093/bioinformatics/bts210.

Hoon, S., Zhou, B., Janda, K. D., Brenner, S., & Scolnick, J. (2011). Aptamer selection by high-throughput sequencing and informatic analysis. *BioTechniques*, *51*(6), 413–416. https://doi.org/10.2144/000113786.

Jalali-Yazdi, F., Lai, L. H., Takahashi, T. T., & Roberts, R. W. (2016). High-throughput measurement of binding kinetics by mRNA display and next-generation sequencing. *Angewandte Chemie*, *55*(12), 4007–4010. https://doi.org/10.1002/anie.201600077.

Jijakli, K., Khraiwesh, B., Fu, W., Luo, L., Alzahmi, A., Koussa, J., et al. (2016). The in vitro selection world. *Methods*, *106*, 3–13. https://doi.org/10.1016/j.ymeth.2016.06.003.

Jiménez, J. I., Xulvi-Brunet, R., Campbell, G. W., Turk-MacLeod, R., & Chen, I. A. (2013). Comprehensive experimental fitness landscape and evolutionary network for small RNA. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(37), 14984–14989. https://doi.org/10.1073/pnas.1307604110.

Jolma, A., Kivioja, T., Toivonen, J., Cheng, L., Wei, G., Enge, M., et al. (2010). Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Research*, *20*(6), 861–873. https://doi.org/10.1101/gr.100552.109.

Jung, C., Hawkins, J. A., Jones, S. K., Jr., Xiao, Y., Rybarski, J. R., Dillard, K. E., et al. (2017). Massively parallel biophysical analysis of CRISPR-cas complexes on next generation sequencing chips. *Cell*, *170*(1), 35–47.e13. https://doi.org/10.1016/j.cell.2017.05.044.

Katoh, K., Misawa, K., Kuma, K.-i., & Miyata, T. (2002). MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, *30*(14), 3059–3066. https://doi.org/10.1093/nar/gkf436.

Kazan, H., Ray, D., Chan, E. T., Hughes, T. R., & Morris, Q. (2010). RNAcontext: A new method for learning the sequence and structure binding preferences of RNA-binding proteins. *PLoS Computational Biology*, *6*(7), e1000832. https://doi.org/10.1371/journal.pcbi.1000832.

Kent, W. J. (2002). BLAT—The blast-like alignment tool. *Genome Research*, *12*(4), 656–664. https://doi.org/10.1101/gr.229202.

Kwok, C. K., Ding, Y., Tang, Y., Assmann, S. M., & Bevilacqua, P. C. (2013). Determination of in vivo RNA structure in low-abundance transcripts. *Nature Communications, 4*, 2971. https://doi.org/10.1038/ncomms3971.

Lambert, N., Robertson, A., Jangi, M., McGeary, S., Sharp, P. A., & Burge, C. B. (2014). RNA bind-n-seq: Quantitative assessment of the sequence and structural binding specificity of RNA binding proteins. *Molecular Cell*, *54*(5), 887–900. https://doi.org/10.1016/j.molcel.2014.04.016.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*, 357. https://doi.org/10.1038/nmeth.1923.

Lippert, C., Sabatini, R., Maher, M. C., Kang, E. Y., Lee, S., Arikan, O., et al. (2017). Identification of individuals by trait prediction using whole-genome sequencing data. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(38), 10166–10171. https://doi.org/10.1073/pnas.1711125114.

Lucks, J. B., Mortimer, S. A., Trapnell, C., Luo, S., Aviran, S., Schroth, G. P., et al. (2011). Multiplexed RNA structure characterization with selective 2′-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). *Proceedings of the National Academy of Sciences of the United States of America*, *108*(27), 11063–11068. https://doi.org/10.1073/pnas.1106501108.

Merino, E. J., Wilkinson, K. A., Coughlan, J. L., & Weeks, K. M. (2005). RNA structure analysis at single nucleotide resolution by selective 2'-hydroxyl acylation and primer extension (SHAPE). *Journal of the American Chemical Society*, *127*(12), 4223–4231. https://doi.org/10.1021/ja043822v.

Mitchell, D., Renda, A. J., Douds, C. A., Babitzke, P., Assmann, S. M., & Bevilacqua, P. C. (2018). In vivo RNA structural probing of uracil and guanine base pairing by 1-ethyl-3-(3-dimethylaminopropyl)carbodiimide (EDC). *RNA*, *25*, 147–157. https://doi.org/10.1261/rna.067868.118. Advanced Online Publication.

Morris, K. N., Jensen, K. B., Julin, C. M., Weil, M., & Gold, L. (1998). High affinity ligands from in vitro selection: Complex targets. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(6), 2902–2907. https://doi.org/10.1073/pnas.95.6.2902.

Nawrocki, E. P., & Eddy, S. R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, *29*(22), 2933–2935. https://doi.org/10.1093/bioinformatics/btt509.

Neveu, M., Kim, H.-J., & Benner, S. A. (2013). The "strong" RNA world hypothesis: Fifty years old. *Astrobiology*, *13*(4), 391–403. https://doi.org/10.1089/ast.2012.0868.

Ng, E. W. M., Shima, D. T., Calias, P., Cunningham, E. T., Jr., Guyer, D. R., & Adamis, A. P. (2006). Pegaptanib, a targeted anti-VEGF aptamer for ocular vascular disease. *Nature Reviews Drug Discovery*, *5*, 123. https://doi.org/10.1038/nrd1955.

Nutiu, R., Friedman, R. C., Luo, S., Khrebtukova, I., Silva, D., Li, R., et al. (2011). Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument. *Nature Biotechnology*, *29*, 659. https://doi.org/10.1038/nbt.1882.

Parsch, J., Braverman, J. M., & Stephan, W. (2000). Comparative sequence analysis and patterns of covariation in RNA secondary structures. *Genetics*, *154*(2), 909–921.

Pitt, J. N., & Ferré-D'Amaré, A. R. (2010). Rapid construction of empirical RNA fitness landscapes. *Science*, *330*(6002), 376–379. https://doi.org/10.1126/science.1192001.

Pitt, J. N., Rajapakse, I., & Ferré-D'Amaré, A. R. (2010). SEWAL: An open-source platform for next-generation sequence analysis and visualization. *Nucleic Acids Research*, *38*(22), 7908–7915. https://doi.org/10.1093/nar/gkq661.

Raine, A., Manlig, E., Wahlberg, P., Syvänen, A.-C., & Nordlund, J. (2017). SPlinted ligation adapter tagging (SPLAT), a novel library preparation method for whole genome bisulphite sequencing. *Nucleic Acids Research*, *45*(6), e36. https://doi.org/10.1093/nar/gkw1110.

Rampášek, L., Jimenez, R. M., Lupták, A., Vinař, T., & Brejová, B. (2016). RNA motif search with data-driven element ordering. *BMC Bioinformatics*, *17*(1), 216. https://doi.org/10.1186/s12859-016-1074-x.

Regulski, E. E., & Breaker, R. R. (2008). In-line probing analysis of riboswitches. In J. Wilusz (Ed.), *Post-transcriptional gene regulation* (pp. 53–67). Totowa, NJ: Humana Press. https://doi.org/10.1007/978-1-59745-033-1_4.

Roberts, R. W., & Szostak, J. W. (1997). RNA-peptide fusions for the in vitro selection of peptides and proteins. *Proceedings of the National Academy of Sciences of the United States of America*, *94*(23), 12297–12302. https://doi.org/10.1073/pnas.94.23.12297.

Sassanfar, M., & Szostak, J. W. (1993). An RNA motif that binds ATP. *Nature*, *364*, 550. https://doi.org/10.1038/364550a0.

Schütze, T., Wilhelm, B., Greiner, N., Braun, H., Peter, F., Mörl, M., et al. (2011). Probing the SELEX process with next-generation sequencing. *PLoS One*, *6*(12), e29604. https://doi.org/10.1371/journal.pone.0029604.

Shaughnessy, A. F. (2012). Monoclonal antibodies: Magic bullets with a hefty price tag. *BMJ345*, e8346. https://doi.org/10.1136/bmj.e8346.

Slattery, M., Riley, T., Liu, P., Abe, N., Gomez-Alcala, P., Dror, I., et al. (2011). Cofactor binding evokes latent differences in DNA binding specificity between hox proteins. *Cell*, *147*(6), 1270–1282. https://doi.org/10.1016/j.cell.2011.10.053.

Soukup, G. A., & Breaker, R. R. (1999). Relationship between internucleotide linkage geometry and the stability of RNA. *RNA*, *5*(10), 1308–1325. https://doi.org/10.1017/S1355838299990891.

Spitale, R. C., Flynn, R. A., Zhang, Q. C., Crisalli, P., Lee, B., Jung, J.-W., et al. (2015). Structural imprints in vivo decode RNA regulatory mechanisms. *Nature*, *519*(7544), 486–490. https://doi.org/10.1038/nature14263.

Tapsin, S., Sun, M., Shen, Y., Zhang, H., Lim, X. N., Susanto, T. T., et al. (2018). Genome-wide identification of natural RNA aptamers in prokaryotes and eukaryotes. *Nature Communications*, *9*(1), 1289. https://doi.org/10.1038/s41467-018-03675-1.

Tome, J. M., Ozer, A., Pagano, J. M., Gheba, D., Schroth, G. P., & Lis, J. T. (2014). Comprehensive analysis of RNA-protein interactions by high-throughput sequencing-RNA affinity profiling. *Nature Methods*, *11*(6), 683–688. https://doi.org/10.1038/nmeth.2970.

Vasa, S. M., Guex, N., Wilkinson, K. A., Weeks, K. M., & Giddings, M. C. (2008). ShapeFinder: A software system for high-throughput quantitative analysis of nucleic acid reactivity information resolved by capillary electrophoresis. *RNA*, *14*(10), 1979–1990. https://doi.org/10.1261/rna.1166808.

Vu, M. M. K., Jameson, N. E., Masuda, S. J., Lin, D., Larralde-Ridaura, R., & Lupták, A. (2012). Convergent evolution of adenosine aptamers spanning bacterial, human, and random sequences revealed by structure-based bioinformatics and genomic SELEX. *Chemistry & Biology*, *19*(10), 1247–1254. https://doi.org/10.1016/j.chembiol.2012.08.010.

Wang, P. Y., Sexton, A. N., Culligan, W. J., & Simon, M. D. (2018). Carbodiimide reagents for the chemical probing of RNA structure in cells. *RNA*, *25*, 135–146. https://doi.org/10.1261/rna.067561.118. Advanced Online Publication.

Wilson, D. S., & Keefe, A. D. (2001). Random mutagenesis by PCR. *Current Protocols in Molecular Biology*, *51*(1), 8.3.1–8.3.9. https://doi.org/10.1002/0471142727.mb0803s51.

Wu, Y. X., & Kwon, Y. J. (2016). Aptamers: The "evolution" of SELEX. *Methods*, *106*, 21–28. https://doi.org/10.1016/j.ymeth.2016.04.020.

Yao, Z., Weinberg, Z., & Ruzzo, W. L. (2006). CMfinder—A covariance model based RNA motif finding algorithm. *Bioinformatics*, *22*(4), 445–452. https://doi.org/10.1093/bioinformatics/btk008.

Zaccolo, M., Williams, D. M., Brown, D. M., & Gherardi, E. (1996). An approach to random mutagenesis of DNA using mixtures of triphosphate derivatives of nucleoside analogues. *Journal of Molecular Biology*, *255*(4), 589–603. https://doi.org/10.1006/jmbi.1996.0049.

Zhao, Y., & Stormo, G. D. (2011). Quantitative analysis demonstrates most transcription factors require only simple models of specificity. *Nature Biotechnology*, *29*(6), 480–483. https://doi.org/10.1038/nbt.1893.

Ziegenhain, C., Vieth, B., Parekh, S., Reinius, B., Guillaumet-Adkins, A., Smets, M., et al. (2017). Comparative analysis of single-cell RNA sequencing methods. *Molecular Cell*, *65*(4), 631–643. https://doi.org/10.1016/j.molcel.2017.01.023.

Zimmermann, B., Gesell, T., Chen, D., Lorenz, C., & Schroeder, R. (2010). Monitoring genomic sequences during SELEX using high-throughput sequencing: Neutral SELEX. *PLoS One*, *5*(2), e9169. https://doi.org/10.1371/journal.pone.0009169.

Zinshteyn, B., Chan, D., England, W., Feng, C., Green, R., & Spitale, R. C. (2018). Assaying RNA structure with LASER-seq. *Nucleic Acids Research*, *47*, 43–55. gky1172. https://doi.org/10.1093/nar/gky1172.

Zykovich, A., Korf, I., & Segal, D. J. (2009). Bind-n-seq: High-throughput analysis of in vitro protein–DNA interactions using massively parallel sequencing. *Nucleic Acids Research*, *37*(22), e151. https://doi.org/10.1093/nar/gkp802.