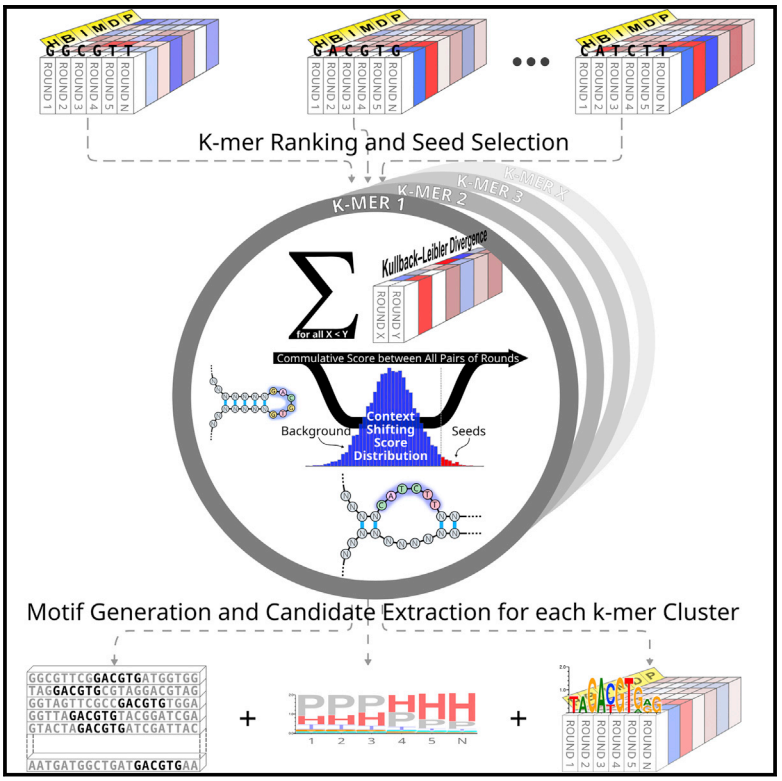
Report

Cell Systems

AptaTRACE Elucidates RNA Sequence-Structure Motifs from Selection Trends in HT-SELEX Experiments

Graphical Abstract



Highlights

- AptaTRACE detects RNA sequence motifs selected for a secondary structure context
- AptaTRACE can handle hundreds of millions of sequences harboring multiple motifs
- AptaTRACE can help to reduce the total number of selection rounds
- Experimental validation of the top motif confirmed its importance for binding

Authors

Phuong Dao, Jan Hoinka, Mayumi Takahashi, ..., Rolf Backofen, John Burnett, Teresa M. Przytycka

Correspondence

przytyck@ncbi.nlm.nih.gov

In Brief

We present AptaTRACE, a method to identify aptamer sequence motifs that undergo selection toward a particular secondary structure context during SELEX experiments. Closing a significant gap in existing methods, AptaTRACE can analyze selection pools sequenced with extremely deep coverage without restricting the search to a single motif.

Dao et al., 2016, Cell Systems *3*, 62–70
July 27, 2016 Published by Elsevier Inc.
http://dx.doi.org/10.1016/j.cels.2016.07.003

OPEN ACCESS
CellPress

Cell Systems
Focus on RECOMB
Report

AptaTRACE Elucidates RNA Sequence-Structure Motifs from Selection Trends in HT-SELEX Experiments

Phuong Dao,[1,4] Jan Hoinka,[1,4] Mayumi Takahashi,[2] Jiehua Zhou,[2] Michelle Ho,[2] Yijie Wang,[1] Fabrizio Costa,[3] John J. Rossi,[2] Rolf Backofen,[3] John Burnett,[2] and Teresa M. Przytycka[1,*]

[1]National Center of Biotechnology Information, National Library of Medicine, NIH, Bethesda, MD 20894, USA
[2]Department of Molecular and Cellular Biology, Beckman Research Institute of City of Hope, Duarte, CA 91010, USA
[3]Bioinformatics Group, Department of Computer Science, University of Freiburg, Freiburg 79110, Germany
[4]Co-first author
*Correspondence: przytyck@ncbi.nlm.nih.gov
http://dx.doi.org/10.1016/j.cels.2016.07.003

SUMMARY

Aptamers, short RNA or DNA molecules that bind distinct targets with high affinity and specificity, can be identified using high-throughput systematic evolution of ligands by exponential enrichment (HT-SELEX), but scalable analytic tools for understanding sequence-function relationships from diverse HT-SELEX data are not available. Here we present AptaTRACE, a computational approach that leverages the experimental design of the HT-SELEX protocol, RNA secondary structure, and the potential presence of many secondary motifs to identify sequence-structure motifs that show a signature of selection. We apply AptaTRACE to identify nine motifs in C-C chemokine receptor type 7 targeted by aptamers in an in vitro cell-SELEX experiment. We experimentally validate two aptamers whose binding required both sequence and structural features. AptaTRACE can identify low-abundance motifs, and we show through simulations that, because of this, it could lower HT-SELEX cost and time by reducing the number of selection cycles required.

INTRODUCTION

Aptamers are short RNA/DNA molecules capable of binding, with high affinity and specificity, a specific target molecule via sequence and structure features that are complementary to the biochemical characteristics of the target's surface. The utilization of aptamers in a multitude of biotechnological and medical sciences has recently increased dramatically. Although only 80 aptamer-related publications were added to PubMed in the year 2000, this number has since roughly doubled every 5 years, with 207 records added in 2005 alone, 565 additional inclusions in 2010, and as many as 957 new manuscripts indexed in 2014. This trend is in part attributable to the considerable diversity of possible targets, which include small organic molecules (Kim and Gu, 2013), transcription factors (Jolma et al., 2010) and other proteins or protein complexes (Berezhnoy et al., 2012), the surfaces of viruses (Binning et al., 2013), and entire cells (Daniels et al., 2003; Morris et al., 1998; Shi et al., 2013). This broad range of targets makes aptamers suitable candidates for a variety of applications ranging from molecular biosensors (Zichel et al., 2012) to drug delivery systems (Xiang et al., 2015) and antibody replacement (FDA, 2004), to just name a few.

Although the specifics vary depending on the target, aptamers are typically identified through the systematic evolution of ligands by exponential enrichment (SELEX) protocol (Ellington and Szostak, 1990). SELEX leverages the well-established paradigm of in vitro selection by repetitively enriching a pool of initially random sequences (species) with those that strongly bind a target of interest. These binders are then selected through a series of selection cycles, where each such cycle involves incubating the pool with the target; partitioning target-bound species from non-binders and removing the latter from the pool, followed by elution of the bound fraction from the target; and amplifying the remaining sequences via PCR to form the input for the subsequent round. After a target-specific number of selection cycles, the final pool is then used to extract dominating, putatively high-affinity species via traditional cloning experiments, computational analysis, and binding affinity assays. Depending on their intended application, favorable binders are often further post-processed in vitro to meet additional requirements such as improved structural stability or reducing the size of the aptamer to the relevant binding region.

A key reason for the resurgence of interest in aptamer research relates to the utilization of affordable next-generation sequencing technologies along with traditional SELEX, referred to as high-throughput SELEX (HT-SELEX) (Jolma et al., 2010; Zhao et al., 2009). In HT-SELEX, after certain (or all) rounds of selection (including the initial pool), aptamer pools are split into two samples, the first of which serves as the starting point for the next cycle, whereas the latter is sequenced. Recently, SELEX-seq, a variation of the HT-SELEX protocol specifically designed to quantify DNA binding references for transcription factor complexes, has been introduced by Slattery et al. (2011). This protocol utilizes electrophoretic mobility shift assays to capture oligomers bound by the targets. The resulting sequencing data of both HT-SELEX and SELEX-seq, consisting of 2–50 million sequences per round, is then analyzed in silico to identify candidates that experience exponential enrichment throughout the selection (Alam et al., 2015; Hoinka et al., 2014). The massive amount of sequencing data produced by these protocols opens the opportunity for the study of many aspects of the protocol that

were either not accessible in traditional SELEX or that could be realized more accurately given hundreds of millions of data points. The development of universal methods for the analysis of HT-SELEX data is challenged by the vast diversity of selection conditions (such as temperature, salt concentration, and species-to-target ratio) and the target complexity. For example, selection against transcription factors and RNA binding molecules requires only a small number of selection rounds to produce high-quality aptamers (Jolma et al., 2010; Kupakuwana et al., 2011). On the other side of the spectrum, in the case of cell-SELEX, a variation of SELEX in which the pool is incubated with entire cells, the number of required selection cycles is significantly larger (Daniels et al., 2003). Such a target can, in general, accommodate a multitude of binding sites, each exposing different binding preferences and leading to a parallel selection toward unrelated binding motifs (Morris et al., 1998). Indeed, the discovery of aptamer binding motifs that facilitate binding to the target is one of the most challenging problems in HT-SELEX data analysis. Current motif-finding algorithms, however, have not been designed with these challenges in mind, and the need for the development of novel computational approaches that address the characteristics specific to the SELEX protocol has become highly relevant.

Traditionally, motif discovery has been defined as the problem of finding a set of common sub-sequences that are statistically enriched in a given collection of DNA, RNA, or protein sequences. To date, a large variety of computational methods in this area have been published (see Tompa et al., 2005; Zambelli et al., 2013; and Weirauch et al., 2013 for a comprehensive review). One of the first computational method for finding motifs on this type of high-throughput data is binding energy estimates using maximum likelihood (BEEML) (Zhao et al., 2009). Assuming the existence of a single binding motif, the method aims at fitting a binding energy model to the data that combines independent attributes from each position in the motif with higher-order dependencies. Another method by Jolma et al. (2010, 2013) approaches the problem by using *k*-mers to construct a position weight matrix (PWM) to infer the binding models. Similarly, Orenstein and Shamir (2015) also uses a *k*-mer approach based on frequencies from a single round of selection to identify binding motifs for transcription factor HT-SELEX data.

The search for motifs in the context of RNA molecules has to consider that binding of ssDNA and RNA molecules depends on both sequence and structure. In particular, it has been proposed that binding regions in those molecules tend to be predominantly single-stranded (Johnson and Donaldson, 2006; Schudoma et al., 2010). MEME in RNAs including secondary structures (MEMERIS) (Hiller et al., 2006) leverages this assumption by weighting nucleotides according to their likelihood of being unpaired. In contrast, RNAcontext (Kazan et al., 2010) divides the single-stranded contexts into known secondary substructures such as hairpins, bulge loops, inner loops, and stems. Consequently, RNAcontext is capable of reporting the relative preference of the structural context along with the primary structure of the potential motif. A related approach was recently proposed for combining sequence and DNA shape properties (Zhou et al., 2015). In contrast, AptaMotif (Hoinka et al., 2012) utilizes information about the structural ensemble of aptamers, obtained by enumerating of all possible structures within a user-defined energy range from the minimum free energy (MFE) structure, and applies an iterative sampling approach combined with sequence-structure alignment techniques to identify high-scoring seeds that are consequently extended to motifs over the full dataset. Subsequently, APTANI (Caroli et al., 2015) extended AptaMotif to handle larger sequence collections via a set of parameter optimizations and sampling techniques, but it also expects a high ratio of motif occurrences.

Still, none of the abovementioned methods address the full spectrum of challenges related to analyzing data from HT-SELEX selections. First, none of these approaches currently scale with the data sizes produced by modern high-throughput sequencing experiments. Next, only a few of the methods consider the existence of secondary motifs, whereas the majority operates under the assumption that only a single primary motif is present in the data. This assumption might apply to TF-SELEX, but it cannot be generalized to common-purpose HT-SELEX, where many motifs of possibly similar binding strength or optimized for additional properties, such as specificity and toxicity, must be considered. Furthermore, secondary structure information, which has proven effective in guiding the motif search to biologically relevant binding sites, is not included in most of these methods. A notable exception is RNAcontext, which can handle relatively large datasets but suffers from the single motif assumption, which cannot be easily removed. Finally, none of these approaches attempt to utilize the full scope of the information produced by modern HT-SELEX experiments, which includes sequencing data from multiple rounds of selection.

To close this gap, we have developed AptaTRACE, a method for the identification of sequence-structure motifs for HT-SELEX that utilizes the available data from all sequenced selection rounds and that is robust enough to be applicable to a broad spectrum of RNA/single-stranded DNA (ssDNA) HT-SELEX experiments independent of the target's properties. Furthermore, AptaTRACE is not limited to the detection of a single motif but capable of elucidating an arbitrary number of binding sites along with their corresponding structural preferences. Unlike previous methods, it does not rely on aptamer frequency or its derivative cycle-to-cycle enrichment. Aptamer frequency has been shown recently to be a poor predictor of aptamer affinity (Cho et al., 2010; Hoinka et al., 2014; Thiel et al., 2012), and, although cycle-to-cycle enrichment has shown a somewhat better performance, the choice of the cycles to compare is not obvious and does not always allow for extraction of sequence-structure motifs. In contrast, our method builds on tracing the dynamics of the SELEX process itself to uncover motif-induced selection trends.

We tested AptaTRACE on sequencing data obtained from realistically simulating SELEX over ten rounds of selection with known binding motifs and then applied it to an in vitro cell-SELEX experiment over nine selection cycles (40 million sequences per cycle). In both cases, our method was successful in extracting highly significant sequence-structure motifs while scaling well with the 10-fold increase in data size. We verified the biological relevance of the top motif by a series of mutation studies in which either the primary or secondary structure of the motif was removed from a candidate aptamer. In both cases, we observed a significant decrease in binding affinity compared with the wild-type. Our results furthermore indicate that the vast majority of motifs are residing in, and are selected for, single-stranded

regions, consistent with previous reports regarding RNA target binding (Johnson and Donaldson, 2006). In addition, we observed that, with sufficient sequencing depth, these motifs can be detected by AptaTRACE relatively early during the selection. Therefore, the ability of AptaTRACE to handle very large input sets opens the possibility of reducing the required number of selection rounds, which are typically expensive to perform in terms of time and cost. AptaTRACE is available for download at http://www.ncbi.nlm.nih.gov/CBBresearch/Przytycka/index.cgi#aptatools.

RESULTS

We start with a high-level outline of the method and refer the reader to Section A in the Supplemental Experimental Procedures for a detailed description of AptaTRACE, implementation, and runtime. Next we use simulated data, produced with an extended version of our AptaSim program (Hoinka et al., 2015) developed for this study, to compare the performance of AptaTRACE with other methods that can handle similar data sizes or incorporate secondary structure into their models. Finally, we show our results of applying AptaTRACE to an in vitro selection consisting of high-throughput data from nine rounds of cell-SELEX (Takahashi et al., BioProject: PRJNA321551).

Top-Level Description of AptaTRACE

Our method builds on accepted assumptions regarding the general HT-SELEX procedure. First we assume that the affinity and specificity of aptamers are mainly attributed to a combination of localized sequence and structural features that exhibit complementary biochemical properties to a target's binding site. Given a large number of molecules in the initial pool, it is expected that such binding motifs are embedded in multiple distinct aptamers. Consequently, during the selection process, aptamers containing these highly target-affine sequence-structure motifs will become enriched compared with target non-specific sequences. Notably, under these assumptions, aptamers that contain only the sequence motif without the appropriate structural context are either not enriched at all or enriched to a much lower degree. The second critical assumption we make is the existence of a multitude of sequence-structure binding motifs that either compete for the same binding site or are binding to different surface regions of the target (Morris et al., 1998; Zichel et al., 2012).

Leveraging the above properties of the SELEX protocol, AptaTRACE detects sequence-structure motifs by identifying sequence motifs that undergo selection toward a particular secondary structure context. Specifically, we expect that, in the initial pool, the structural contexts of each *k*-mer are distributed according to a background distribution that can be determined from the data. However, for sequence motifs involved in binding, in later selection cycles, this distribution becomes biased toward the structural context favored by the binding interaction with the target site. Consequently, AptaTRACE aims at identifying sequence motifs whose tendency of residing in a hairpin, bulge loop, inner loop, multiple loop, or dangling end or of being paired converges to a specific structural context throughout the selection. To achieve this, for each sequenced pool, we compute the distribution of the structural contexts of all possible *k*-mers (all possible nucleotides sequences of length *k*) in all aptamers. Figures 1A–1D provide a schematic of this procedure.

Next we use the relative entropy (Kullback-Leibler [KL] divergence) to estimate, for every *k*-mer, the change in the distribution of its secondary structure contexts (K-context distribution for short) between any cycle to a later cycle (Figures 1E and 1F). The sum of these KL-divergence scores over all pairs of selection cycles defines the context shifting score for a given *k*-mer. The context shifting score is thus an estimate of the selection toward the preferred structure(s). Complementing the context shifting score is the K-context trace, which summarizes the dynamics of the changes in the K-context distribution over consecutive selection cycles.

To assess the statistical significance of these context shifting scores, we additionally compute a null distribution consisting of context shifting scores derived from *k*-mers of all low-affinity aptamers in the selection. This background is used to determine a p value for the structural shift for each *k*-mer (Figure 1G). Predicted motifs are then constructed by aggregating overlapping *k*-mers under the restriction that the structural preferences in the overlapped region are consistent (Figure 1H). Finally, position-specific weight matrices of these motifs, specifically their sequence logos, along with their motif context traces (the average K-context traces of the *k*-mers used in the PWM construction) and the corresponding aptamers in which these occur, are reported to the user (Figure 1I).

Results on Simulated Data

To validate our approach, we applied AptaTRACE to a dataset generated by means of in silico SELEX. To this end, we used an extension to our AptaSim program (Hoinka et al., 2015) allowing for implanting specific sequence-structure motifs into the initial pool. We generated a dataset of 4 million sequences per round containing five motifs (denoted here as motifs a–e), 5–8 nt in length and located predominantly in unpaired regions. Note that the motifs' primary structures also occur randomly in the background aptamers, albeit in arbitrary structural contexts, and that the motifs are hence not over-represented in the initial pool. Each motif was initially present in 100 different target-affine aptamer species and consequently selected for over ten rounds of SELEX. A complete description of the simulation and the parameters used during in silico SELEX are available in Section B in the Supplemental Experimental Procedures.

We applied AptaTRACE as well as discriminative regular expression motif elicitation (DREME) and RNAcontext to the dataset to compare their capability of extracting these motifs. Notably, RNAcontext was not capable of handling 4 million sequences in a reasonable time frame, prompting us to sample the 10,000 most frequent and least frequent sequences of the last selection cycle as input. The full scope of parameters used for these methods during the comparison is detailed in Section D of the Supplemental Experimental Procedures. The results of the comparison are summarized in Figure 2A. Notably, RNAcontext did not return any of the implanted motifs and is hence not represented in the figure and was excluded from further comparisons. AptaTRACE was applied to the full dataset as well as to the last selection cycle only to facilitate a comparison with DREME. Although DREME failed to identify the low-affinity motifs d and e, AptaTRACE was able to recover all motifs in

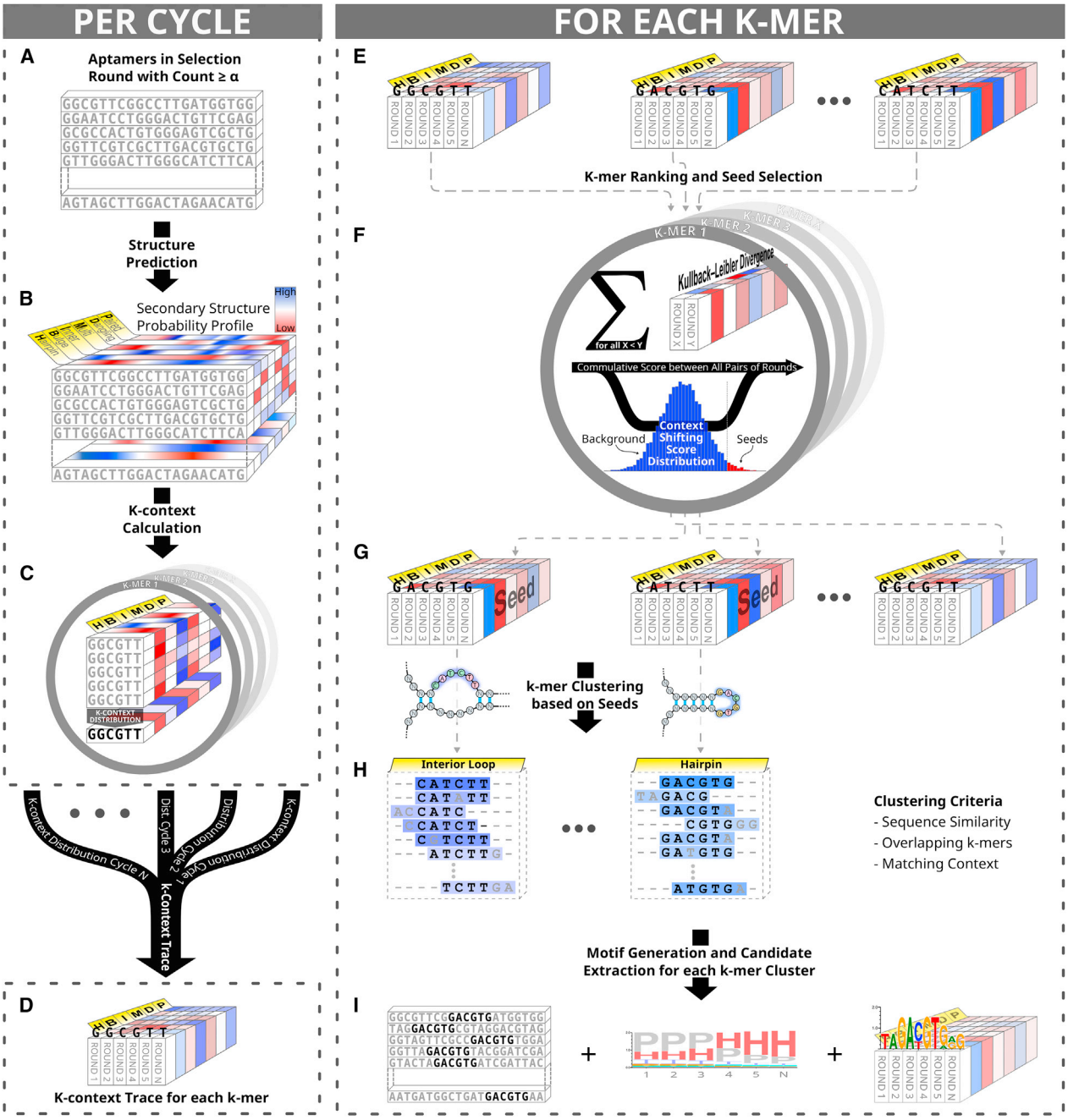


Figure 1. Schematic Overview of the AptaTRACE Method

(A) For each cycle, all sequences with a frequency above a user-defined threshold α are selected as input.

(B) Computation of secondary structure probability profiles for each aptamer using SFOLD. For each nucleotide, the profile describes the probability of residing in a hairpin, bulge loop, inner loop, multiple loop, or dangling end or of being paired.

(C) K-context and K-context distribution calculation for each *k*-mer.

(D) Generation of the K-context trace for each *k*-mer.

(E–G) *k*-mer ranking and statistical significance estimation. Given any two selection cycles, the relative entropy (KL-divergence) is used to estimate the change in the distribution of its K-context distribution. The sum of these KL-divergence scores over all pairs of selection cycles defines the context shifting score for a given *k*-mer. To assess the statistical significance of these context shifting scores, a null distribution is computed, consisting of context shifting scores derived from *k*-mers of all low-affinity aptamers in the selection (frequency $\leq \alpha$). This background is used to determine a p value for the structural shift for each *k*-mer. Top scoring *k*-mers are selected as seeds.

(H) Predicted motifs are constructed by aggregating *k*-mers overlapping with the seed under the restriction that the structural preferences in the overlapped region are consistent.

(I) Position-specific weight matrices representing these motifs, along with their K-context traces, and corresponding aptamers are reported to the user.

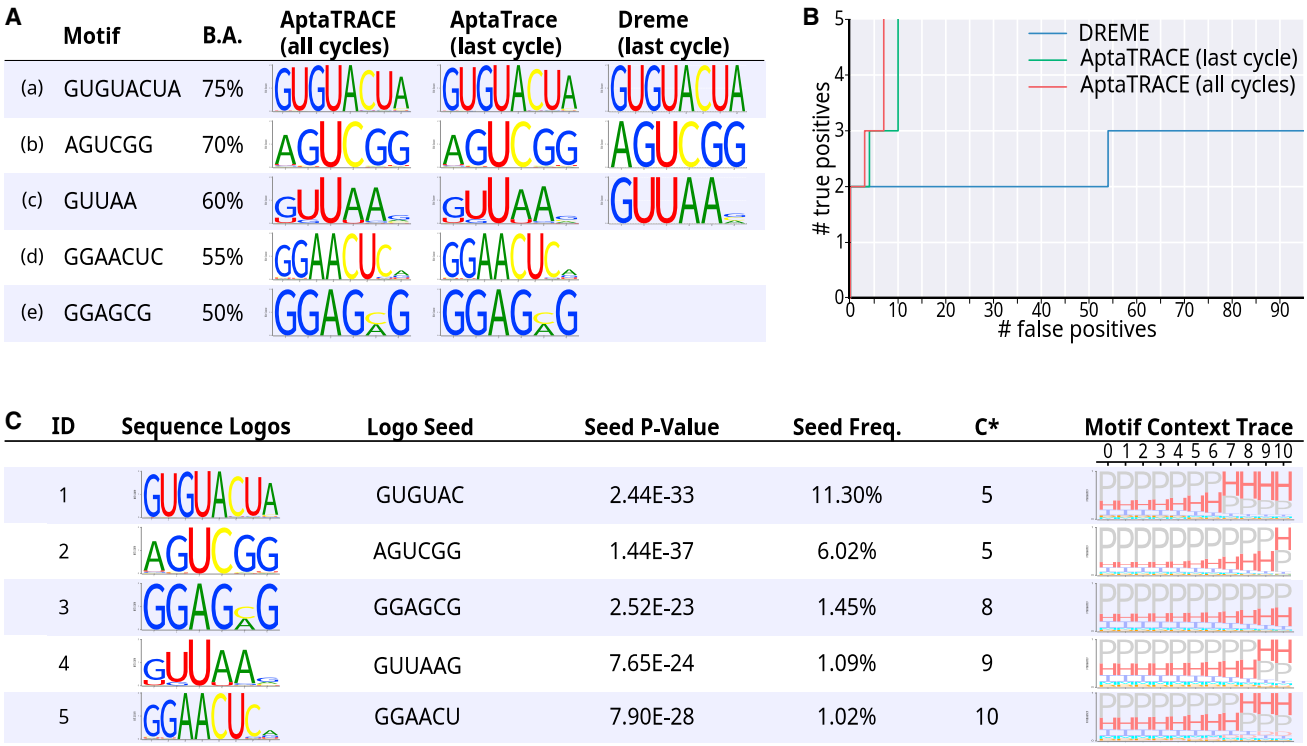


Figure 2. Comparison between AptaTRACE and Other Methods Using Simulated Data

(A) Comparison of AptaTRACE against other methods based on simulated data. AptaTRACE was applied to the entire dataset and, for comparison with DREME, to the last selection cycle only. Shown in the first two columns are the implanted motifs and their binding affinity used throughout the selection (B.A.). The output PWMs produced by the tested methods that correspond to the implanted motifs are displayed in the remaining columns.

(B) Plot depicting the number of false positive motifs reported by DREME (blue line) and AptaTRACE on the x axis against the number of true positives recovered on the y axis. For AptaTRACE, we utilized only the initial pool and the last cycle as input (green line) and the full simulated dataset (red line). The former yielded 15 motifs. When applying our method onto the full dataset, a total of 12 motifs were identified.

(C) Sequence-structure motifs identified by AptaTRACE from virtual SELEX given all ten selection cycles, including the initial pool, as input. Shown here are the identified sequence logos, the *k*-mer that scored highest in significance used for construction of each motif (seed) and its p value, the abundance of seed of the motif in the final selection round (Frequency), the first cycle at which the motif was detected (C^*), as well as the motif context trace throughout the selection from the initial pool to round 10.

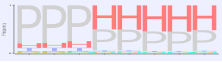
both test scenarios. In addition, AptaTRACE exhibits, by a large margin, the lowest false discovery rate compared with DREME (Figure 2B). We note that DREME took approximately 10 days to complete, whereas AptaTRACE only required a total of 12 hr for both computing the secondary structure profiles and for identifying the sequence-structure motifs. A more detailed summary of the sequence logos extracted by our approach on the full dataset, including their motif context traces and statistical significance, is available in Figure 2C. Interestingly, a visual inspection of the motif context trace (last column, Figure 2C) points to the possibility of capturing most of these motifs at earlier cycles. Indeed, computing the selection round in which a motif was first detected by AptaTRACE (column C^*, Figure 2C), confirmed this expectation.

Results on Cell-SELEX Data

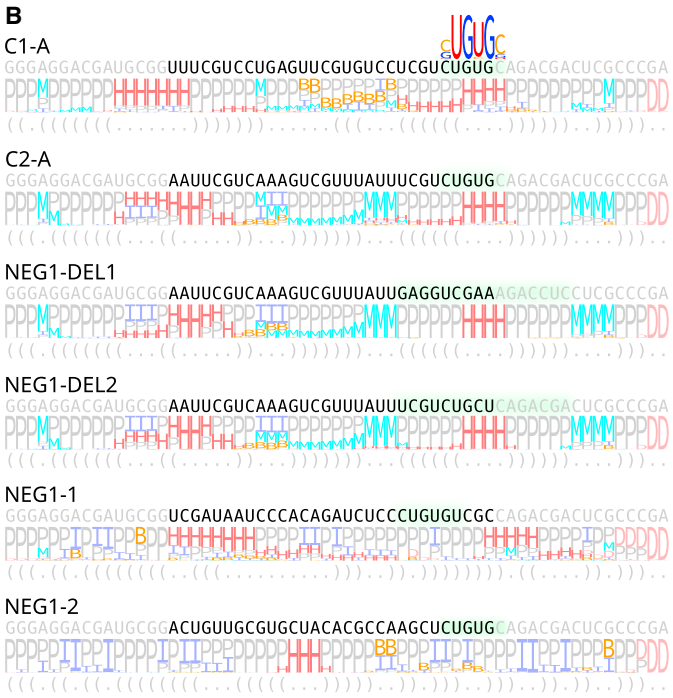
Next we applied AptaTRACE to the results of an in vitro cell-SELEX experiment targeting the C-C chemokine receptor type 7 (CCR7), where the initial pool as well as seven of nine selection rounds have been sequenced, averaging 40 million aptamers per cycle (see Experimental Procedures for a detailed description of the experimental procedure). AptaTRACE was able to successfully extract a total of nine motifs (Figure 3A).

The context trace of these motifs hints toward two properties of the selection process. First, a clear selection toward single-stranded regions for every extracted motif can be observed. It has always been postulated that ssDNA/RNA binding motifs are predominantly located in loop regions (Schudoma et al., 2010). Indeed, this assumption was leveraged by MEMERIS (Hiller et al., 2006) by imposing priors, directing the motif search toward single-stranded regions. In the case of AptaTRACE, no prior assumption of this type was made. The fact that, despite a lack of such priors, motifs detected by AptaTRACE conform with the expected properties of RNA sequence-structure binding sites supports their relevance for binding. Next, the trend of the structural preferences of these motifs emerges relatively early during the selection process, indicating that, in conjunction with our method, the identification of biologically relevant binding sites in general-purpose HT-SELEX data might be possible with fewer selection cycles.

A

ID	Sequence Logo	Logo Seed	Seed P-Value	Seed Freq.	Motif Freq.	Motif Contect Trace (0 1 3 5 6 7 8 9)
1		CUGUG	3.43E-03	15.40%	33.00%	
2		CGCUG	2.47E-03	6.57%	11.30%	
3		UGCGC	2.59E-04	5.05%	11.13%	
4		UAUUG	2.51E-03	4.83%	12.24%	
5		CUGGC	9.59E-04	4.82%	11.87%	
6		UGGUG	2.24E-03	4.67%	5.33%	
7		CUGCA	1.29E-03	4.54%	4.54%	
8		GCGUG	3.60E-03	3.46%	3.46%	
9		UGCCG	2.09E-04	1.37%	2.56%	

B



C

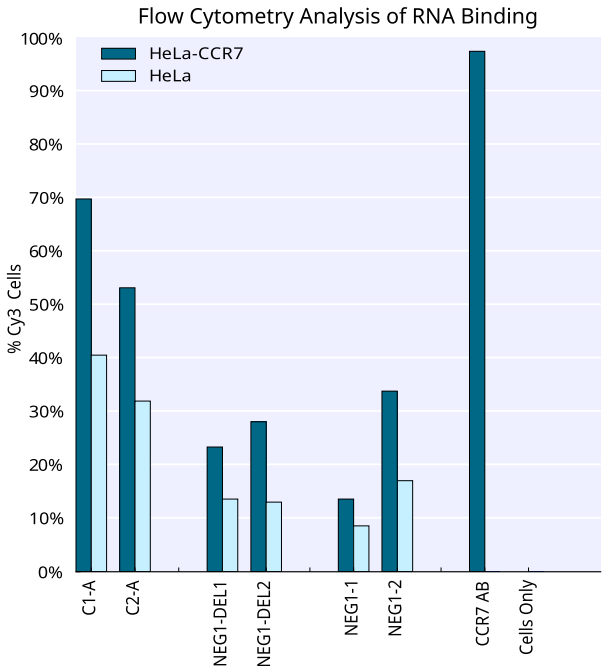


Figure 3. Analysis of Motifs Identified by AptaTRACE in Cell-SELEX Experiments

(A) All sequence-structure motifs as identified by AptaTRACE on cell-SELEX data. The sequence logo as well as the most frequent *k*-mer constituting the logo (Logo Seed) and its p value are depicted for each motif together with its seed frequency. The motif context trace for the sequenced cycles (0, 1, 3, 5, 6, 7, 8, and 9) is shown in the last column. Experimental validation of sequence-structure motifs identified by AptaTRACE.

(B) The selected aptamers, their secondary structure probability profiles, as well as their MFE structure. For each species, the primer regions of the sequence are colored gray, whereas the 30-nt-long randomized region is presented in black. The secondary structure probability profile reflects, for each nucleotide, the probability of residing in a hairpin, inner loop, bulge loop, or multiple loop or of being paired. The position of the motif on the aptamer (as shown in the sequence logo) is highlighted in green, and the primary structure region in which the motif was removed is highlighted in violet.

(C) Flow cytometry analysis of the sequences as shown in (B). Depicted are the percentages of aptamer-bound cells for target-expressing HeLa cells (dark blue) and target-lacking HeLa cells (light blue) compared with a high-affinity antibody (CCR7 AB) and HeLA cells with no aptamers for control (Cells Only).

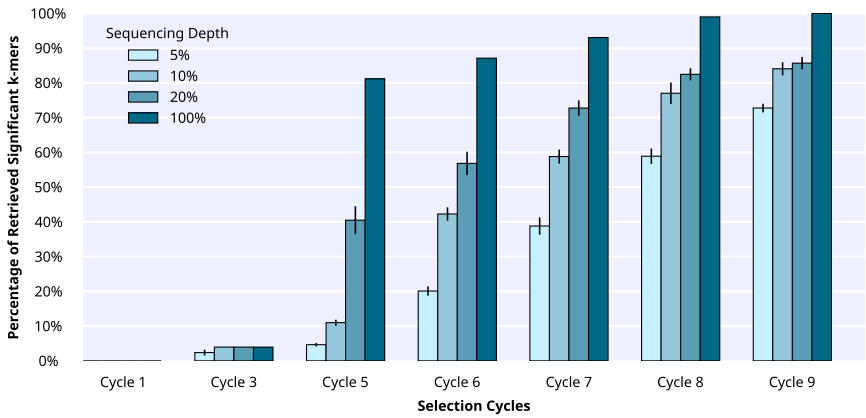


Figure 4. Trade-Off between the Number of Selection Cycles and Sequencing Depth

Shown is the percentage of significant *k*-mers identified by AptaTRACE as a function of the number of cycles and sequencing depth. Each bar corresponds to the application of our approach onto a reduced dataset containing all selection cycles up to round *x* while utilizing a random subset of *y*% reads of each cycle. The height of each column stands for the percentage of the number of retrieved significant *k*-mers compared with running AptaTRACE on the full cell-SELEX data. The SDs correspond to the sampling effects caused by repeating each experiment 20 times.

Experimental Validation

We further substantiated our findings in vitro by performing a number of flow cytometry-based binding assays using chemokine receptor-expressing HeLa cells and HeLa cells in which the receptor is not expressed (Experimental Procedures). Using the most prevalent sequence-structure motif as a reference, we selected two highly enriched aptamers, denoted C1-A and C2-A (Figure 3B), that contain this motif in a hairpin located at the far 3′ end of the randomized region. To verify that both sequence and structure are responsible for the binding interaction with the target, we additionally engineered four control experiments based on C2-A in which we either preserved the secondary structure of the aptamer but replaced the primary structure of the motif with an arbitrary sequence not related to any motifs identified by AptaTRACE (NEG1-DEL1 and NEG1-DEL2) or selected aptamers from the pool which retained the primary structure of the motif but in which the secondary structure is contained within a paired region of the species (NEG1-1 and NEG1-2).

Our binding assays show that, although aptamer C1-A exhibits the highest affinity to the target, C2-A shows substantially greater specificity (Figure 3C). The control experiments demonstrate that eliminating either sequence or structure from the motif results in a significant decrease of binding ability to the target, strengthening our argument that AptaTRACE is capable of identifying biologically relevant sequence-structure motifs from complex HT-SELEX data. Notably, removing the secondary structure component from the motif resulted in the largest drop in affinity compared with replacing the primary structure only, further validating the correlation between motifs located in unpaired regions and expected binding affinity.

Trade-Off between the Number of Selection Cycles and Sequencing Depth

The ability of AptaTRACE to analyze very large datasets opens the possibility of reducing the number of selection cycles by increasing the sequencing depth. Such a reduction in number of cycles is desirable for two main reasons. First, with current technology, the cost savings from reducing the number of cycles outweigh the added cost because of deeper sequencing. Next, a decrease in the amount of selection rounds allows to reduce the number of potential artifacts that can accumulate during this multi-step procedure. Because the sequencing depth of our cell-SELEX experiment significantly exceeds current practices, we were able to explore the relationship between sequencing depth, the number of required selection cycles, and the number of identified motifs by AptaTRACE. For this purpose, we performed a series of sampling tests on the original data. Specifically, we iteratively reduced the number of selection cycles down to the first round and randomly selected 5%, 10%, 20%, and 100% of the sequences in each round. We then utilized AptaTRACE on the scaled-down datasets and computed the ratio between the number of identified seed *k*-mers and significant *k*-mers compared with running AptaTRACE on the full dataset.

The results suggest that AptaTRACE is capable of identifying motif signals from as early as selection cycle 3 and that, with only five rounds, the vast majority of motifs (80%) can be recovered (Figure 4). These findings therefore strongly indicate the possibility of trading off additional (and expensive) selection cycles in favor of deeper (and more economic) high-throughput sequencing, even when analyzing complex landscapes as those generated by cell-SELEX experiments.

DISCUSSION

Unlike in traditional SELEX, where only a handful of potential binders are retrieved and exhaustively tested experimentally, HT-SELEX returns a massive amount of sequencing data sampled from some or all selection rounds. These data consequently serve as the basis for the challenging task of identifying suitable binding candidates and for deriving their sequence-structure properties that are key for binding affinity and specificity. Except for the special case of TF-binding aptamers, no previous tool addressing this task existed. Several potential factors during any stage of the selection contribute to the complexity of developing efficient approaches for the identification of sequence-structure binding motifs from HT-SELEX sequencing data. They include, but are not limited to, polymerase amplification biases, sequencing biases, contamination of foreign sequences, and non-specific binding. These factors prompted aptamer experts to consider cycle-to-cycle enrichment instead of frequency counts as a predictor for binding affinity. Although cycle-to-cycle enrichment did increase the predictive power of these methods, it cannot bypass problems related to amplification bias nor can it identify aptamer properties that drive binding affinity and specificity. In contrast,

AptaTRACE is specifically designed to identify sequence-structure binding motifs in HT-SELEX data and is thus suitable to predict the features behind binding affinity and specificity.

An important feature of AptaTRACE is that, rather than using quantitative information, it directly leverages the experimental design of the SELEX protocol and identifies motifs that are under selection through appropriately composed scoring functions. By focusing on local motifs that are selected for, AptaTRACE bypasses global biases such as the PCR bias, which is typically related to more universal sequence properties such as the CG content. In addition, because AptaTRACE measures selection toward a sequence-structure motif by its shift in the distribution of the structural context and not based on abundance, it can uncover statistically significant motifs that are selected for, even when these only form a small fraction of the pool. This is an important property that can ultimately help to shorten the number of cycles required for selection and thus to reduce the overall cost of the procedure. Indeed, our results have confirmed that, with deep enough sequencing, only a limited number of selection cycles might be required for exhaustively elucidating sequence-structure motifs in HT-SELEX data. In addition, our analysis also shows that the dynamics of K-context traces is not the same for all motifs. Although most trends essentially stabilize at a relatively early cycle, some continue to grow. We hypothesize that this type of information can aid the identification of the most promising binders. AptaTRACE is therefore not only a powerful method to detect emerging sequence-structure motifs but also a flexible tool that can be readily adopted to interrogate such selection dynamics.

EXPERIMENTAL PROCEDURES

Cell-based SELEX was performed using an RNA library containing a randomized 30-nt region flanked by fixed primer sequences. Cell-based selection was performed as described previously (Kim and Gu 2013) by employing open PCR for DNA amplification during each selection round. Positive selection was performed on HeLa cells transduced with a bicistronic lentiviral vector expressing the target surface receptor and GFP, whereas unmodified HeLa cells, which lack expression of the target receptor, were used for negative selection. High throughput sequencing (HTS) was performed on the positive selection at rounds 0, 1, 3, 5, 6, 7, 8, and 9.

Binding assays were performed using standard flow cytometry. We used both chemokine receptor-expressing HeLa cells and HeLa cells in which the receptor was not expressed for the analysis of aptamer binding (see the Flow Cytometry Analysis of Cell Surface Binding section in the Supplemental Experimental Procedures for details).

AptaTRACE is available for download at http://www.ncbi.nlm.nih.gov/CBBresearch/Przytycka/index.cgi#aptatools and as Data S1.

ACCESSION NUMBER

The accession number for the data reported in this paper is BioProject: PRJNA321551.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures and one data file and can be found with this article online at http://dx.doi.org/10.1016/j.cels.2016.07.003.

AUTHOR CONTRIBUTIONS

J.H., P.D., R.B., and T.M.P. conceived the study. M.T. and J.Z performed the cell-SELEX experiment, and M.H performed the flow cytometry experiment. P.D. and J.H. performed data processing and computational analysis. T.M.P., R.B., and J.B. supervised the study. Y.W., F.C., and J.J.R. provided critical feedback. J.H., P.D., and T.M.P. wrote the manuscript.

ACKNOWLEDGMENTS

This work was supported by the Intramural Research Program of the NIH, National Library of Medicine (J.H., P.D., and T.M.P.). Funding for the open access charge was provided by the NIH. This work utilized the computational resources of the NIH HPC Biowulf cluster. J.J.R., J.Z., and M.T. have patent disclosures entitled "CCR7 and CD2 RNA Aptamer-Functionalized Conjugates to Target and Activate HIV Latently Infected Lymphocytes," City of Hope, filed in January 2016. An early version of this paper was submitted to and peer-reviewed at the 2016 Annual International Conference on Research in Computational Molecular Biology (RECOMB). The manuscript was revised and then independently further reviewed at *Cell Systems*.

Received: May 14, 2016
Revised: June 24, 2016
Accepted: July 1, 2016
Published: July 27, 2016

REFERENCES

Alam, K.K., Chang, J.L., and Burke, D.H. (2015). FASTAptamer: A Bioinformatic Toolkit for High-throughput Sequence Analysis of Combinatorial Selections. Mol. Ther. Nucleic Acids *4*, e230.

Berezhnoy, A., Stewart, C.A., Mcnamara, J.O., 2nd, Thiel, W., Giangrande, P., Trinchieri, G., and Gilboa, E. (2012). Isolation and optimization of murine IL-10 receptor blocking oligonucleotide aptamers using high-throughput sequencing. Mol. Ther. *20*, 1242–1250.

Binning, J.M., Wang, T., Luthra, P., Shabman, R.S., Borek, D.M., Liu, G., Xu, W., Leung, D.W., Basler, C.F., and Amarasinghe, G.K. (2013). Development of RNA aptamers targeting Ebola virus VP35. Biochemistry *52*, 8406–8419.

Caroli, J., Taccioli, C., De La Fuente, A., Serafini, P., and Bicciato, S. (2015). APTANI: a computational tool to select aptamers through sequence-structure motif analysis of HT-SELEX data. Bioinformatics *32*, 161–164.

Cho, M., Xiao, Y., Nie, J., Stewart, R.T., Csordas, A.T., Oh, S.S., Thomson, J.A., and Soh, H.T. (2010). Quantitative selection of DNA aptamers through microfluidic selection and high-throughput sequencing. Proc. Natl. Acad. Sci. U.S.A *107*, 15373–15378.

Daniels, D.A., Chen, H., Hicke, B.J., Swiderek, K.M., and Gold, L. (2003). A tenascin-C aptamer identified by tumor cell SELEX: Systematic evolution of ligands by exponential enrichment. Proc. Natl. Acad. Sci. U.S.A *100*, 15416–15421.

Ellington, A.D., and Szostak, J.W. (1990). In vitro selection of RNA molecules that bind specific ligands. Nature *346*, 818–822.

FDA (2004). FDA Approves New Drug Treatment for Age-Related Macular Degeneration. http://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/2004/ucm108385.htm. Accessed December 20, 2004.

Hiller, M., Pudimat, R., Busch, A., and Backofen, R. (2006). Using RNA secondary structures to guide sequence motif finding towards single-stranded regions. Nucleic Acids Res. *34*, e117.

Hoinka, J., Zotenko, E., Friedman, A., Sauna, Z.E., and Przytycka, T.M. (2012). Identification of sequence-structure RNA binding motifs for SELEX-derived aptamers. Bioinformatics *28*, i215–i223.

Hoinka, J., Berezhnoy, A., Sauna, Z.E., Gilboa, E., and Przytycka, T.M. (2014). AptaCluster - A Method to Cluster HT-SELEX Aptamer Pools and Lessons from its Application. Res. Comput. Mol. Biol. *8394*, 115–128.

Hoinka, J., Berezhnoy, A., Dao, P., Sauna, Z.E., Gilboa, E., and Przytycka, T.M. (2015). Large scale analysis of the mutational landscape in HT-SELEX improves aptamer discovery. Nucleic Acids Res. *43*, 5699–5707. http://paperpile.com/b/nC8l3T/KYH4.

Johnson, P.E., and Donaldson, L.W. (2006). RNA recognition by the Vts1p SAM domain. Nat. Struct. Mol. Biol. *13*, 177–178.

Jolma, A., Kivioja, T., Toivonen, J., Cheng, L., Wei, G., Enge, M., Taipale, M., Vaquerizas, J.M., Yan, J., Sillanpää, M.J., et al. (2010). Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. Genome Res. *20*, 861–873.

Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K.R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G., et al. (2013). DNA-binding specificities of human transcription factors. Cell *152*, 327–339.

Kazan, H., Ray, D., Chan, E.T., Hughes, T.A., and Morris, Q. (2010). RNAcontext: a new method for learning the sequence and structure binding preferences of RNA-binding proteins. PLOS Comput. Biol. *6*, e1000832.

Kim, Y.S., and Gu, M.B. (2013). Advances in aptamer screening and small molecule aptasensors. Adv. Biochem. Eng. Biotechnol. *140*, 29–67.

Kupakuwana, G.V., Crill, J.E., 2nd, McPike, M.P., and Borer, P.N. (2011). Acyclic identification of aptamers for human alpha-thrombin using over-represented libraries and deep sequencing. PLoS ONE *6*, e19395.

Morris, K.N., Jensen, K.B., Julin, C.M., Weil, M., and Gold, L. (1998). High affinity ligands from in vitro selection: complex targets. Proc. Natl. Acad. Sci. USA *95*, 2902–2907.

Orenstein, Y., and Shamir, R. (2015). HTS-IBIS: fast and accurate inference of binding site motifs from HT-SELEX data. bioRxiv. http://dx.doi.org/10.1101/022277.

Schudoma, C., May, P., Nikiforova, V., and Walther, D. (2010). Sequence-structure relationships in RNA loops: establishing the basis for loop homology modeling. Nucleic Acids Res. *38*, 970–980.

Shi, H., Cui, W., He, X., Guo, Q., Wang, K., Ye, X., and Tang, J. (2013). Whole cell-SELEX aptamers for highly specific fluorescence molecular imaging of carcinomas in vivo. PLoS ONE *8*, e70476.

Slattery, M., Riley, T., Liu, P., Abe, N., Gomez-Alcala, P., Dror, I., Zhou, T., Rohs, R., Honig, B., Bussemaker, H.J., and Mann, R.S. (2011). Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. Cell *147*, 1270–1282.

Thiel, W.H., Bair, T., Peek, A.S., Liu, X., Dassie, J., Stockdale, K.R., Behlke, M.A., Miller, F.J., Jr., and Giangrande, P.H. (2012). Rapid identification of cell-specific, internalizing RNA aptamers with bioinformatics analyses of a cell-based aptamer selection. PLoS ONE *7*, e43836.

Tompa, M., Li, N., Bailey, T.L., Church, G.M., De Moor, B., Eskin, E., Favorov, A.V., Frith, M.C., Fu, Y., Kent, W.J., et al. (2005). Assessing computational tools for the discovery of transcription factor binding sites. Nat. Biotechnol. *23*, 137–144.

Weirauch, M.T., Cote, A., Norel, R., Annala, M., Zhao, Y., Riley, T.R., Saez-Rodriguez, J., Cokelaer, T., Vedenko, A., Talukder, S., et al.; DREAM5 Consortium (2013). Evaluation of methods for modeling transcription factor sequence specificity. Nat. Biotechnol. *31*, 126–134.

Xiang, D., Shigdar, S., Qiao, G., Wang, T., Kouzani, A.Z., Zhou, S.F., Kong, L., Li, Y., Pu, C., and Duan, W. (2015). Nucleic acid aptamer-guided cancer therapeutics and diagnostics: the next generation of cancer medicine. Theranostics *5*, 23–42.

Zambelli, F., Pesole, G., and Pavesi, G. (2013). Motif discovery and transcription factor binding sites before and after the next-generation sequencing era. Brief. Bioinform. *14*, 225–237.

Zhao, Y., Granas, D., and Stormo, G.D. (2009). Inferring binding energies from selected binding sites. PLoS Comput. Biol. *5*, e1000590.

Zhou, T., Shen, N., Yang, L., Abe, N., Horton, J., Mann, R.S., Bussemaker, H.J., Gordân, R., and Rohs, R. (2015). Quantitative modeling of transcription factor binding specificities using DNA shape. Proc. Natl. Acad. Sci. USA *112*, 4654–4659.

Zichel, R., Chearwae, W., Pandey, G.S., Golding, B., and Sauna, Z.E. (2012). Aptamers as a sensitive tool to detect subtle modifications in therapeutic proteins. PLoS ONE *7*, e31948.