Aptamers ISSN: 2514-3247 Aptamers, 2018, Vol 2, 36-44

RESEARCH REPORT

Shortlisting aptamer Candidates from HT-SELEX data

Theodore R Allnutt[1,2], Ben Wade[2], Thomas P Quinn[1,2], Mark F Richardson[1,3] and Tamsyn M Crowley[1,2,4*]

[1]Bioinformatics Core Research Group, Deakin University, Locked Bag 20000, Geelong, VIC, 3220, Australia; [2]School of Medicine, Centre for Molecular and Medical Research, Deakin University, Locked Bag 20000, Geelong, VIC 3220, Australia; [3]School of Life and Environmental Sciences, Centre for Integrative Ecology, Deakin University, Locked Bag 20000, Geelong, VIC 3220, Australia; [4]Poultry Hub Australia, University of New England, Armidale, NSW, 2351, Australia

*Correspondence to: Tamsyn Crowley, Email: tamsyn.crowley@deakin.edu.au

Received: 11 December 2017 | Revised: 21 May 2018 | Accepted: 23 May 2018 | Published: 25 May 2018

ABSTRACT

High-throughput systematic evolution of ligands by exponential enrichment (HT-SELEX) is a development of SELEX which uses next generation sequencing to analyse the aptamer selection process. Computational tools have previously been developed specifically to analyse HT-SELEX data and assist in shortlisting aptamers most likely to have high affinity to the SELEX target(s). We have exploited a published HT-SELEX data set to assess the performance of six aptamer clustering methods and four methods to rank the clusters in their ability to shortlist the highest affinity aptamers. We found that methods which attempt to take into account secondary structure, in terms of its enrichment characteristics and complexity, did not perform as well as simpler methods, which cluster based on sequence alone and rank by a measure of the final absolute enrichment. We also demonstrate that analysis methods developed for amplicon metagenomics perform well on HT-SELEX data. Importantly, we note a lack of publicly accessible HT-SELEX/validation data despite numerous studies reporting the use of this technique, which hampers extensive comparative benchmarking. We implore the community to make data public to aid methodological advances in aptamer shortlisting and benchmarking.

KEYWORDS: Aptamer selection, HT-SELEX, AptaSuite, FASTAptamer, Usearch

INTRODUCTION

The use of aptamers has increased dramatically in recent times (Zhou and Rossi, 2017). Aptamers are able to act analogously to antibodies; binding ligand molecules with high degrees of specificity and affinity (Hoinka et al, 2014). They are composed of single-stranded short nucleic acids, which fold upon themselves resulting in complex 3-D structures and produce their ligand specificity via shape recognition. Aptamers are generated using a technique known as the systematic evolution of ligands by exponential enrichment or SELEX (Tuerk and Gold, 1990). This is a sequential process that requires multiple rounds of selection and starts with a highly heterologous pool of aptamers being exposed to the target ligand, adherent aptamers are retained and non-binders are removed. SELEX results in a library of aptamers that are able to bind with specificity and strength to their target(s), these are sequenced and can then be synthesised and tested for their individual binding affinities. It is necessary to test the affinity of resultant aptamers as not all highly represented aptamers will bind the target (Tolle et al, 2014).

Traditionally, SELEX has been applied to relatively simple target ligands such as purified recombinant protein (Lakhin et al, 2013). These highly-homogenous targets generally yield numbers of aptamers in the range of one or two dozen at most. The sequencing of the final aptamer library could be facilitated using low-throughput sequencing approaches (*e.g.*, cloning and Sanger sequencing). Recently the use of

complex targets has become more common, such as the whole bacterial or eukaryotic cells (Lakhin et al, 2013). Such 'Cell-SELEX' procedures can be useful as it allows the selection of aptamers without prior knowledge of specific target ligands. The complex nature of these targets, with hundreds or thousands of cell-surface moieties, will result in many more aptamers being selected for and retained over SELEX cycles compared to simpler targets, rendering low-throughput sequencing non-viable. Instead, next generation sequencing (NGS) platforms need to be utilised in a process known as high-throughput SELEX (HT-SELEX) (Kupakuwana et al, 2011).

HT-SELEX provides deep sequencing of the SELEX aptamer pools at any or all cycles of the process. This not only supplies a larger number of candidates for testing, but also enables the analysis of the abundance of aptamer species from one cycle to the next - which can be used to reveal aptamers undergoing selection and/or containing desired structures. This process generates large data sets posing new challenges in analysis and interpretation. Hence, the overarching aim of HT-SELEX sequence analysis should be to provide a shortlist of aptamers that are most likely to be good candidates for testing target binding affinity. This is imperative as it is usually not possible to test more than a few dozen candidate aptamers due to the expense of the experiments.

Several different computational approaches have been designed specifically to assist in aptamer selection from HT-SELEX data (for a full review, see Kinghorn et al, 2017). These include, but are not limited to: AptaCluster (Hoinka et al, 2014); FASTAptamer (Alam et al, 2015); APTANI (Caroli et al, 2016) and AptaTRACE (Dao et al, 2016). Of these, AptaCluster and FASTAptamer provide methods to cluster aptamer reads in HT-SELEX data based on sequence similarity, whereas AptaTRACE and APTANI use a similar approach to identify conserved sequence motifs that define the secondary structures enriched over SELEX rounds, then build clusters that contain those motifs. In theory, the largest of these clusters would more likely contain structural features that have high affinity to the target molecule(s). Collectively we refer to these as aptamer analysis methods.

Generally, following sequence quality control, the first step in HT-SELEX sequence analysis is clustering the sequences into groups. As this is a frequent procedure in bioinformatics, it is reasonable to suggest that other bioinformatic, non-aptamer specific, analysis tools that cluster sequences may be useful. In particular, we suggest that the algorithms developed for amplicon metagenomics (or collectively amplicon analysis methods), *e.g.*, Uclust and Unoise, both parts of the Usearch package (Edgar, 2016) are particularly relevant. This is because HT-SELEX data has similarities to metagenomic amplicons in that reads are generally of a discrete and narrow size range and are produced by competitive PCR reactions, giving rise to similar types and rates of errors. Furthermore, in later SELEX cycles, sequences form populations of varying sizes with a log-normal frequency distribution, as would be expected for metagenomic amplicon populations (Paulson et al, 2013).

Among clustering and structural analysis there is the need for methods to be easily applied, computationally efficient and scalable to the potentially vast numbers of reads in a given HT-SELEX sequence data set. Despite the availability of several aptamer-specific analysis methods, to date no useful comparison of such programs, with respect to their ability to rank sequenced aptamers with known target affinities, has been conducted. Here, we present a comparison of shortlisting methods in terms of speed and accuracy, using the freely available programs: AptaCluster, FASTAptamer, the Usearch programs: Uclust and Unoise, and a secondary structure clustering method developed for this study. We then assess the accuracy of different metrics to rank the order of clusters, such that high-affinity aptamers are most likely to be shortlisted. We note that there is a paucity of publicly accessible HT-SELEX sequence data and associated aptamer affinity data. This hampers efforts to conduct extensive comparative benchmarking, while also limiting the development of aptamer analysis best practices. Our results provide an important first step in the discourse regarding how best to analyse HT-SELEX data, whilst also highlighting the differences between some existing analysis methods. We conclude that the amplicon clustering methods are applicable to shortlisting aptamers from HT-SELEX data, but further investigation into their generalisability is warranted.

METHODS

Data acquisition

Figure 1 shows an overview of the analysis process used. To compare the available aptamer selection methods, we required a data set with both NGS data and measurements of an aptamers binding affinity to the target molecule (*i.e.*, dissociation constant, K_d values). Twenty studies with HT-SELEX sequencing data are available on the NCBI short read archive (SRA), as of 20/3/2018. However, only one data set provides significant numbers of K_d data enabling its use as a benchmarking dataset – that published in Hoinka et al, 2015; and Levay et al, 2015 (available via the NCBI SRA; BioProject PRJNA315881; Runs SRR3279660 and SRR3279661). The total number of reads per round was: round 2 = 2,644,594; round 3 = 1,298,160; round 4 = 1,958,875; and round 5 = 5,746,926. Associated K_d data is available for 33 of the most abundant aptamers selected in Hoinka et al, 2015 and Levay et al, 2015, hereafter these are referred to as the affinity-scored aptamers (Table 1).

Aptamer selection approaches

Following NGS quality control (the effects of which we do not examine in this paper), HT-SELEX sequence analysis has two steps to produce a shortlist of candidate aptamers for validation: first - cluster together sequences based on similarity and count the size of clusters; and second - rank the clusters by using a metric that attempts to place aptamers with desired binding properties at the top of the list (Figure 1), we used six different methods to do this. Note that a recent aptamer analysis package, APTANI (Caroli et al, 2016), was not included in this study because it would not operate on our Linux system due to technical problems despite repeated attempts.

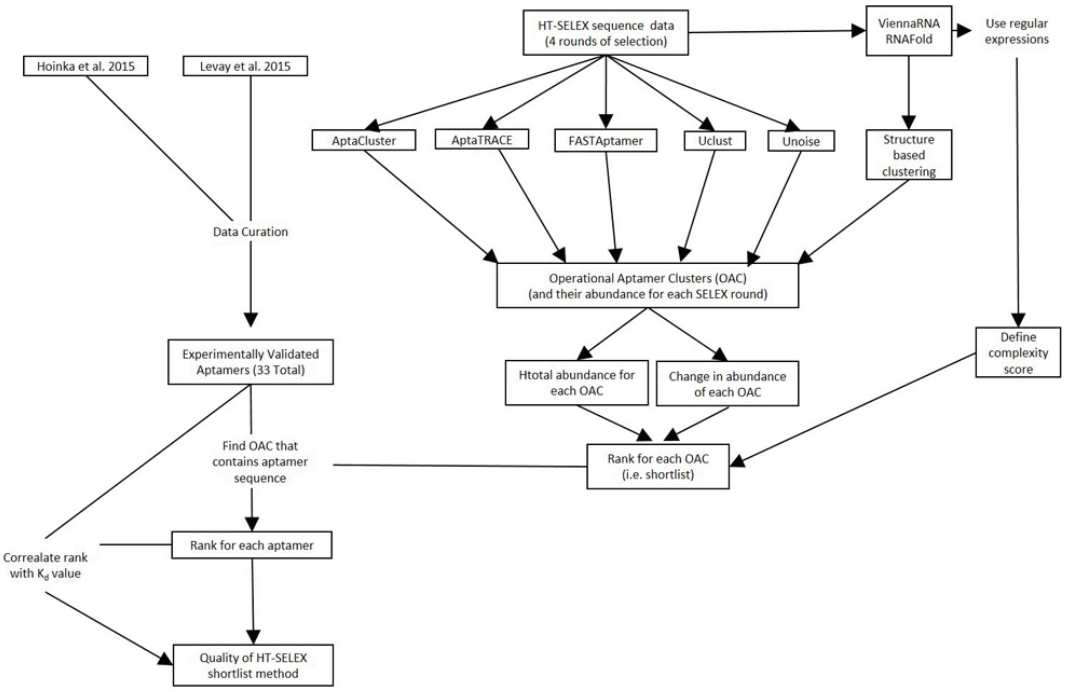


Figure 1. Diagram showing the analysis process.

CLUSTERING METHODS

We used three aptamer specific clustering methods: FASTAptamer (v1.03), AptaCluster and AptaTRACE (both implemented by Aptasuite v0.4.5), and two amplicon clustering methods: Uclust and Unoise (implemented by USEARCH v10.0.240_i86linux64, Edgar, 2010; Edgar, 2016). We used default parameters for all methods, apart from those specifics described below. We used Uclust to cluster sequences and generate counts without error correction and used Unoise to do the same but with error correction. For Unoise, HT-SELEX reads were mapped to the 'Zotu' representative sequences using USEARCH's '-otutab' command to obtain counts. For all methods, identity thresholds of 97% and 90% were investigated, for the top 500 clusters. For FASTAptamer clustering, the options '-c 500' and '-f 100' were used to limit the total number of clusters to 500 and remove reads with fewer than 100 identical copies from the clustering process. All other clustering programs used default options and the total number of clusters was trimmed to the 500 largest.

Structure analysis

For structure-based clustering, we first used RNAfold (v2.4.1) (Lorenz et al, 2011) to generate secondary structures for reads (a single, minimum free entropy structure - using a single possible structure greatly decreased computation time), then translated the resulting "dot and bracket" notation (see Figure 2) into a pseudo-sequence with three possible symbols: G, C, and A, where 'A' stood for no structure present ('dots' in RNAfold notation); and 'G' and 'C' stood for paired forward and reverse bases respectively (brackets in RNAfold notation). The pseudo-sequences were then clustered as normal by Uclust at a clustering identity threshold of 97%.

Performance of aptamer selection approaches

For each clustering method, we tested five different metrics for ranking sequence clusters: counts (total abundance across all SELEX cycles); enrichment gradient (gradient of the abundances from 2nd to 5th cycle); absolute enrichment (between final (5th) and 4th cycle, expressed as the absolute difference in counts); proportional enrichment (between final (5th) and 4th cycle, expressed as the proportional change); and percentage of hairpin tips (HTP) - a measure of secondary structure complexity as detailed below and in the Supplementary material.

We explored whether secondary structure alone could predict binding affinity. To do this, we devised measures to represent the complexity of the secondary structures of the 33 affinity-scored aptamers. By applying regular expressions to the RNAfold "dot and bracket" notation (described above), we characterised four motifs (illustrated in Figure 2). For each structural property, we calculated the number of times they occur, the average size of the motif(s), and the total size of the motif(s). We then correlated these measures with the K_d values. We applied a standardisation of the best performing measure (HTP, or the percent of the molecule containing hairpin tips) to the top 500 largest clusters, effectively re-ranking the top 500 most abundant clusters by this complexity metric. The RNAFold program has the potential to produce multiple structures for a given sequence depending on parameters such as temperature, as such we explored the impact of altering temperature on the 33 affinity tested aptamers to temperatures of 21, 26, 31 and 37 degrees. Results (shown in supplementary Table 1) showed, broadly speaking, little difference in the correlation between any of the predicted structural features (Figure 2) and binding affinity. As such default parameters were subsequently used.

To assess the performance of combined clustering and ranking methods, the affinity-scored aptamers with K_d

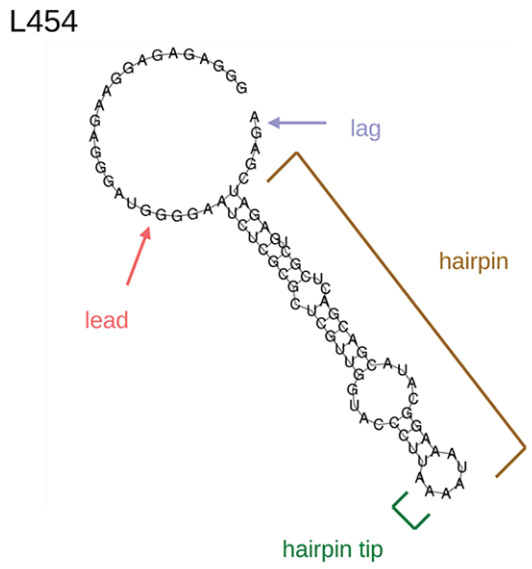
Table 1. Comparison of aptamer sequence clustering and ranking methods to binding data (K_d, nM). For each method's clustered and sorted data, 'count' columns show the rank position of each aptamer in the top 500 largest clusters by cluster size; and 'Enr.' shows the rank position of each aptamer when the clusters are sorted by their change in enrichment between 4th and 5th rounds of SELEX. If an aptamer was not found then a rank of 500 was used. The 'Top 10 correct' value is the number of strong binding aptamers (K_d < 100 nM) that were ranked in the top 10 (therefore higher value is better, range = 0–10). rs = Spearman's correlation; r = Pearson's correlation. Analysis time is shown in hours.

			Uclust 97%		Unoise 97%		Structure 97%*		AptaCluster 3 bp max. diff.		Fastaptamer 97%		AptaTrace
Aptamer ID	**Kd**	**rank**	**count**	**Enr.**	**count**	**Enr.**	**count**	**Enr.**	**count**	**Enr.**	**Count**	**Enr.**	**na**
L462	2	1	15	11	14	10	22	24	18	15	18	15	15
L464	4	2	22	38	21	35	30	101	28	214	28	106	30
L455	4	3	31	15	30	14	54	37	38	27	38	25	500
L454	8	4	16	9	15	8	20	15	16	11	15	11	500
H33	10	5	55	20	50	18	115	70	148	99	120	73	112
L463	12	6	30	41	27	38	47	110	34	186	34	100	38
H4	18	7	11	7	10	6	10	7	11	7	11	7	500
H12	20	8	18	13	17	12	23	21	19	17	19	17	16
H22	20	8	33	31	32	27	50	109	43	146	42	89	50
H30	25	9	48	19	44	17	104	62	115	74	98	59	85
H0	25	9	1	1	1	1	1	1	1	1	1	1	1
L465	25	9	25	549	24	443	33	418	29	481	30	304	48
L418	35	10	14	12	13	11	25	25	15	16	16	16	500
L413	40	11	19	111	18	130	26	152	23	273	23	118	22
H6	50	12	8	480	7	499	8	257	8	408	8	233	7
H3	60	13	5	6	5	5	5	6	6	6	6	6	5
H2	65	14	4	3	4	3	4	4	4	4	4	4	4
H8	80	15	12	8	11	7	12	11	12	8	12	8	8
L420	80	15	28	21	28	19	42	81	36	92	36	68	34
H40	120	16	66	24	61	21	146	90	186	139	141	87	34
H1	120	16	3	2	3	2	3	2	2	2	2	2	146
L412	120	16	32	547	31	491	56	474	42	467	41	298	2
H14	123	17	13	430	12	458	21	178	17	284	17	121	500
H16	375	18	17	469	16	457	19	496	20	487	20	309	19
H7	375	18	10	14	9	13	9	30	9	21	9	21	26
H9	375	18	9	521	8	500	11	366	10	446	10	243	9
H20	375	18	20	474	19	493	28	427	24	486	24	308	12
L409	500	19	26	372	29	422	55	398	44	439	43	235	32
H26	500	19	29	455	26	497	40	495	37	492	37	265	122
L417	500	19	27	71	25	64	41	113	31	285	31	122	88
H5	500	19	7	285	6	498	7	299	7	429	7	215	35
H15	500	19	21	35	20	33	29	128	25	147	25	90	6
H24	500	19	23	488	22	496	34	454	27	491	27	312	25
r_s			–0.13	0.49	–0.12	0.54	–0.12	0.52	–0.10	0.52	–0.10	0.52	–0.31
r			–0.07	0.47	–0.06	0.55	–0.10	0.60	–0.12	0.60	–0.12	0.60	–0.27
Top 10 correct			4	6	5	7	5	4	4	5	4	5	5
Analysis Time			75		3		2		1		1		11

*Structure clustering was performed on one million subsampled reads

values were sorted from best to worst binding affinity (lower the K_d the better the binding) and this list was compared to the ranking of the top 500 largest clusters from each of the six shortlisting methods. The rank of each affinity-scored aptamer in each shortlisting methods' clusters was established by using a Python search script that enabled up to five bp mismatches (no more than two were observed in practice). We then assessed performance of each shortlisting method based on their rank of affinity-scored aptamer compared to the K_d using both Spearman's rank correlation, r_s, and Pearson's correlation r and the number of 'good' binders (K_d < 100) observed in the top 10 (of the 500) of each shortlisting methods aptamer clusters (henceforward referred to as the 'top10' measure).

To ensure the ability to re-run this analysis in its entirety, alll scripts and analysis pipelines used are available via the GitHub repository: https://github.com/bioinformatics-deakin/htselex2017.



GGGAGAGAGGAAGAGGGAUGGGGAAUCUCGCGCUCGUUGGUACCCUUAAAAUAAAGGCAUACGACGACUCGCUGAGAUCGAGA
.....((((((((.((((((....((((......))))....))))))..)))).))))).....

Figure 2. A 2D visualisation of the aptamer L454, as rendered by the ViennaTools command line RNAfold software (using default parameters). The diagram shows the four secondary motifs used to procedurally characterise aptamers by secondary structure; lead, lag, hairpin and hairpin tip. The lead is the series of unpaired bases at the 5' end of the transcript shown in pink. The lag is the series of unpaired bases at the 3' end of the transcript in blue. The hairpin is the series of bases that occur between the first closed pair. The hairpin tip is the series of unbound bases between the last closed pair. Below, the primary sequence and the secondary sequence (in dot-bracket notation) for L454 is provided.

RESULTS AND DISCUSSION

HT-SELEX analysis approaches primarily aim to inform the selection of high affinity and target-specific candidate aptamers before experimental validation and downstream application (Kinghorn et al, 2017). The validation of candidates is still essential because not all candidate aptamers will have the desired affinity or specificity, despite potentially being highly represented in the HT-SELEX libraries. Validation cannot be avoided, is time consuming and generally requires the modification of the synthesised aptamers with various functional groups, greatly increasing cost. Accordingly, analysis methods most likely to shortlist functional aptamers are of great benefit and importance.

We compared the shortlisting accuracy and speed of HT-SELEX sequence data clustering and ranking methods, including analysis methods and a structural clustering in order to find those that provide the highest likelihood of identifying high affinity aptamers for subsequent validation. It was our intention to undertake this analysis using multiple HT-SELEX datasets, but to do this it required both sequence and associated binding affinity data. To our dismay we only located one such dataset. Most published studies describing HT-SELEX data available on the SRA database (20 datasets) reported only a handful (< 10) of aptamer binding affinities if any at all. We also encountered studies within the literature that utilised HT-SELEX but made no mention if the sequencing data had been deposited. The lack of publicly deposited data and/or lack of reported binding affinities represents a hitherto unrecognised problem within the field that could be addressed with a dedicated online repository for HT-SELEX, containing data such as binding affinities and NGS reads. The adoption and use of a dedicated repository could, in our opinion, represent best-practice for the field.

An ideal aptamer can bind to a target with high affinity to form aptamer-target complexes. These complexes happen as a result of the conformations assumed by the macromolecules, and the way in which these conformations interact in real space. Therefore, one would expect that the primary sequence of an aptamer alone may not sufficiently explain whether a sequence preferentially binds to a target. While the analysis of HT-SELEX makes use of the abundances of unique sequence clusters (*i.e.*, after each round of SELEX) to prioritise high affinity aptamers, it follows that the inclusion of higher-order structural information could augment the interpretation of these results. To explore this possibility, we used regular expressions to procedurally characterise aptamer complexity, then ranked aptamers independent of any HT-SELEX results.

The Usearch program provides a clustering algorithm, Uclust, that is fast, accurate and precise that could provide an alternative to AptaCluster or FASTAptamer. Unoise identifies Illumina sequencing (or PCR) errors in reads and corrects them, producing a list of 'Zotus' (zero radius operational taxonomic units). The program also implements PCR chimera (heteroduplex) removal, which may alleviate some of the potential errors that occur in the SELEX procedure. Following de-noising, reads are then mapped on to the set of Zotus with a predefined similarity threshold, *e.g.*, 97%, to provide counts of each Zotu over SELEX cycles. This mapping process is considerably faster than clustering algorithms and provides a means to count the abundance of all available sequences in reference to a relatively shortlist of Zotus, while also correcting potential errors. However,

these corrected errors may contain random PCR substitutions that may or may not contribute to improved aptamer binding - but if these errors do contribute to such, we would expect them to persist in later cycles and grow into separate Zotus that will still be detected by Unoise.

Performance of complexity measures

Table 2 shows an example of the correlation of 8 measures of complexity with K_d values. Only percentage of the total length that were hairpin tips (HTP) showed moderate correlation with aptamer rank. This would suggest that the more free nucleotides there are on the exposed end of the loop-structure (or the more loop-structures there are in total), the less affinity the aptamer has to its target. Given this moderate correlation it was decided to additionally apply the HTP measure to the subsequent clustering methods to investigate its predictive potential in regards to selecting 'strong binding' aptamers with this dataset.

Performance of clustering and ranking methods

To score the performance of each clustering method and ranking combination we chose the 'top10' as our primary measure, combined with both Spearman's rank correlation compared to the K_d rank (r_s), and Pearson's correlation compared to the K_d value. Table 1 shows the rank position of each affinity-scored aptamer in the top 500 largest clusters for each of the six clustering methods and the two best ranking methods: total count and absolute enrichment (count difference between final two SELEX cycles) (full results including all other ranking methods can be found in supplementary Table 2). The best performance (according to our primary measure) was shown by Unoise (clustered

Table 2. Complexity scores used to rank aptamer clusters applied to the published aptamers with K_d values using default settings for the RNAfold program.

Aptamer ID	K_d (nM)	Lead Length (bp)	Lag Length (bp)	Hairpin Tips Total	Hairpin Tips Length Mean (bp)	Hairpin Tips Total (%)	Hairpins Total	Hairpins Length Mean (bp)	Hairpins Total (%)
L462	2	1	3	2	3.50	8	2	36.50	88
L464	4	17	3	2	5.50	13	2	13.00	31
L455	4	0	3	2	8.50	20	2	37.50	90
L454	8	24	5	1	6.00	7	1	52.00	63
H33	10	4	3	2	4.00	10	2	26.00	63
L463	12	2	1	3	9.67	35	3	23.67	86
H4	18	0	3	2	3.50	8	2	37.00	89
H12	20	21	4	1	5.00	6	1	56.00	67
H22	20	18	2	1	3.00	4	1	61.00	73
H30	25	3	10	1	6.00	7	1	68.00	82
H0	25	14	3	4	4.00	19	4	14.00	67
L465	25	2	1	3	10.33	37	3	23.67	86
L418	35	3	10	1	5.00	6	1	68.00	82
L413	40	17	1	3	9.33	34	3	16.33	59
H6	50	0	3	2	9.50	23	2	37.00	89
H3	60	23	2	1	3.00	4	1	56.00	67
H2	65	0	3	2	7.00	17	2	37.00	89
H8	80	16	4	1	14.00	17	1	61.00	73
L420	80	2	1	3	4.67	17	3	23.67	86
H40	120	1	1	2	7.50	18	2	37.50	90
H1	120	0	1	3	5.67	20	3	25.00	90
L412	120	1	11	2	4.00	10	2	28.00	67
H14	123	2	1	3	6.33	23	3	23.67	86
H16	375	2	1	3	6.33	23	3	23.67	86
H7	375	8	0	2	6.00	14	2	30.00	72
H9	375	9	2	2	5.50	13	2	28.50	69
H20	375	2	1	3	11.00	40	3	23.67	86
L409	500	0	11	1	7.00	8	1	70.00	84
H26	500	2	1	3	14.33	52	3	23.67	86
L417	500	0	9	2	13.00	31	2	33.00	80
H5	500	2	1	3	10.33	37	3	23.67	86
H15	500	2	1	3	11.67	42	3	23.67	86
H24	500	6	3	3	4.00	14	3	19.33	70
r_s	–0.22		–0.31	0.30	0.39	0.41	0.30	–0.15	0.14
r	–0.31		–0.02	0.27	0.42	0.47	0.27	–0.21	0.19

at 97% identity) combined with absolute enrichment ranking (top10 = 7; r_s = 0.54; r = 0.55). The second best was Uclust (clustered at 97% identity), also combined with absolute enrichment ranking of aptamers (top10 = 6; r_s = 0.49; r = 0.47). Several other methods (such as AptaCluster, AptaTRACE and FASTAptamer) also ranked four or five of the strong binders in their top 10 candidates, indicating that there was little difference between these selection methods. Our results are congruent with another study that noted the enrichment method produced the best results for ranking clusters (Hoinka et al, 2015), although the latter used proportional enrichment, rather than absolute counts.

We found that proportional enrichment was inferior to absolute enrichment scores - being more susceptible to ranking clusters with low absolute numbers (supplementary Table 2). This finding is further supported when we look at the number of 'weak binders' (K_d > 100 nM) that ranked within the top 10 - effectively a measure of false positives. Across all clustering methods, when total counts were used, between three and four 'weak binders' were present among the top 10 clusters (supplementary Table 2). Importantly, in all examples, this number was reduced when the absolute enrichment ranking method was used, with only the same single weak binder being assigned to the top 10 in all cases (H1). Not only does this support the use of absolute enrichment ranking for aptamer selection but it further underlines the importance of sequencing multiple rounds of SELEX selection, not just the terminal round.

Of the other ranking methods used to score clustering approaches; total count, proportional enrichment, and percentage of hairpin tips (HTP) performed less well (supplementary Table 2). HTP and proportional enrichment were consistently the poorest performing ranking methods (both only ranking one strong binder in the top ten of the Unoise analysis). Clustering based on secondary structure did not perform better than that based on sequence: AptaTRACE and our own structure clustering methods achieved five of the 'strong' binding aptamers in their top 10 and r_s coefficients of: AptaTRACE, r_s = –0.31; structure, absolute enrichment, r_s = 0.52; structure, counts, r_s = -0.12. Clusters based on secondary structure contained negligible sequence variation (supplementary data; Table 3). Average sequence divergence of the 500 largest clusters was 1.5% (s.d. = 3.4%). Five sequence variation outliers were present with divergences ranging from 32% to 35% but their largest cluster size was 114 and none of them contained the affinity-scored aptamers.

We found that AptaTRACE was among the lowest performing methods at shortlisting aptamers with this dataset. It took approximately four times longer to run than the Unoise method (see below) and although it managed to rank five of the 'strong' binders in its top 10, it also failed to rank four of the 'strong' within its top 500 and had the poorest r_s coefficient (-0.27, enrichment) of any method. AptaTRACE does not produce data for each cycle, but rather identifies clusters that contain structural motifs that undergo positive selection over each SELEX cycle (Dao et al, 2016) and reports the count observed in the last cycle where the cluster was observed. This means that the absolute enrichment ranking could not be applied to AptaTRACE to rank the clusters - which may have improved the shortlist it provided. Uclust performed moderately well, however, it is only practical to run on sub-sampled data due to the required CPU time on the full data set (75hr).

Clustering by predicted secondary structure alone did not improve the aptamer shortlist. We expected that aptamers with very similar structures could arise from different sequences during the SELEX process – akin to convergent evolution. However, we observed that only the smallest clusters contained appreciable sequence variation - meaning that in this experiment at least, secondary structure selection remained coupled to sequence selection. This could be a consequence of the initial sequence library not containing enough sequence variants of each potential structure. If so, this negates the utility of structural clustering to rank aptamers. Since the global secondary structure of an aptamer can drastically change with a single base change, we included the primers in structure prediction on the grounds that these primers are present in the sequence during SELEX.

Additional considerations

Clustering approaches are computationally demanding for large data sets like those produced in HT-SELEX (Coleman et al, 2000). As such, we compared the computational demand (measured as CPU hours) for all of the five different methods, with both the full data set (11,648,555 reads) and a subsampled set containing one million reads. The computational demand of clustering methods varied considerably, however, note that the CPU time required by ranking methods was negligible (seconds) compared to the clustering step. With a 12 CPU (Intel Xeon E5–2680 v3), 2.5 GHz computer with 132 GB RAM, Uclust (97% identity) took 75hr to complete clustering of the full data set, and 1hr 27min to process the subsampled data set. Aptasuite (which generated AptaCluster and AptaTRACE results) took one hour to generate AptaCluster results on the full dataset, and 11hr on the full data set to generate AptaTRACE results (1hr 30min on the subsampled set). Structural clustering was not attempted on the full data set, due to the length of time required for its Uclust step (75hr), and took 1hr 54min for secondary structure prediction and clustering on one million subsampled reads. Unoise (de-noising and Zotu mapping) took 2hr 54min on the full data set (subsampling not deemed necessary due performance speed). Aptasuite therefore demonstrated that its clustering algorithm was 6.8x faster than Uclust. However, when Uclust was applied to AptaCluster clusters, new clusters were found: using one million of the top AptaCluster representative sequences 75 further sequences were added to the largest cluster and a further 15896 clusters were given at least one new member; demonstrating that AptaCluster may trade some precision for speed. All time comparisons are shown in Table 1. Uclust, although it performed well, proved impractically slow to cluster the complete data set compared to AptaCluster, which was the fastest clustering algorithm. However, Unoise, including the mapping of all reads to Zotus, was sufficiently fast to be practical to use on the full data set. We limited the number of CPUs employed to a reasonable level (12) that would be commonly available to researchers with access to cluster computers. However,

Unoise and AptaCluster could be run on a desktop PC with e.g. four CPUs within a reasonable length of time (nine and three hours respectively) on this size of data set. The FASTAptamer program performed quickly (1hr), but only following the addition of limiting parameters to reduce the number of input de-replicated reads and total number of clusters.

There are additional considerations that must be taken into account when using FASTAptamer and AptaCluster. These programs cluster sequences within the individual rounds and not across all rounds (as done by Uclust and Unoise). This has the effect that some clusters will be present in one round but not another, which may lead to the loss of potential aptamers in the ranking step because it prevents the measurement of enrichment between rounds - although this is only likely with smaller clusters that are less likely to contain good binders' sequence. This behaviour is the likely reason why the performance of FASTAptamer and AptaCluster were nearly identical.

Subsampling the full sequence data set to one million reads had little apparent effect on the performance of each method of selecting aptamers compared to using the full data set. For example Uclust (97% identity), with all data, produced top10 = 6 and $r_s = 0.49$, compared to top10 = 5 and $r_s = 0.48$ when one million reads were sampled (supplementary Table 3). It appears subsampling to one million reads did not greatly affect the observed aptamer ranking, it is possible that it could increase the error in enrichment calculation in smaller clusters or be a consequence of the dataset on which we performed the analysis. Therefore, we would suggest careful exploration of subsampling on a case by case basis.

It is important to acknowledge a limitation of this study - the number of validated aptamers available for benchmarking. An idealised data set would have data on the binding properties of all aptamer candidates, however, this is unfeasible due to the high cost of performing such experiments. Hence, when applying our top10 measure, in all cases, some of those top ten were uncharacterised aptamers. It is beyond the scope of this study to validate all of these additional aptamer candidates, so it is difficult to determine if they constituted strong or weak binders. As such, some caution must be taken when interpreting these results. However, using this data set, we can state that Unoise and Uclust performed at least as well as AptaCluster, FASTAptamer and AptaTRACE in predicting and ranking the aptamers for which K_d values were available. Whilst caution must be taken extrapolating the results of this study broadly we see no reason why the use of analysis tools such as Uclust and Unoise on other HT-SELEX datasets would not be similarly useful.

In future, it would be greatly beneficial for the aptamer researching community to access all such available data from a dedicated online repository, including metadata such as SELEX conditions. Although at present it is difficult to validate very large numbers of aptamers from HT-SELEX, it is likely that solid-state affinity testing methods, such as surface plasmon resonance (Rubio et al, 2016), will enable high-throughput validation - which will greatly assist future analytical methods, our understanding of the SELEX process, and the more rapid and efficient development of aptamers.

CONCLUSIONS

We found that the use of two analytical pipelines, Unoise and Uclust, originally designed to analyse amplicon metagenomic data, performed at least as well for shortlisting HT-SELEX data than a number of aptamer-specific pipelines. In particular, Unoise produced promising results and did so in a short timeframe comparable to the most rapid of the aptamer-specific methods tested here. Methods that used the secondary structure to cluster and/or rank aptamers (APTAtrace, structure clustering, complexity score ranking) did not perform as well in this study. This suggests that it may not be worth the greater computing effort that these methods require. However, given that our comparison is based on the only currently available validation HT-SELEX set, we are unable to provide conclusive recommendations for the choice of one shortlisting method over any other. Yet, we provide an important start to the discourse on method suitability and clearly demonstrate that Usearch clustering methods are applicable to shortlisting aptamers from HT-SELEX data, but further investigation into their generalisability is warranted. It is important to note that the proportional enrichment ranking, as previously suggested by Hoinka et al (2015), performed consistently poorly (albeit on the single dataset analysed here). We recommend the use of the consistently high performing metric, absolute enrichment between the last two SELEX cycles, instead. The lack of multiple HT-SELEX experiments with validation binding affinities (for both strong and weak binders) that are publically available greatly hampers community efforts to conduct extensive comparative benchmarking, limiting the development of aptamer analysis best practices. We advocate for, and implore the community to make such data public to aid methodological advances in aptamer shortlisting.

ACKNOWLEDGEMENTS

We thank Jan Hoinka for providing de-multiplexed and merged reads from the SRA data.

COMPETING INTERESTS

None declared.

LIST OF ABBREVIATIONS

bp: **B**ase **P**air
CPU: **C**entral **P**rocessing **U**nit
GB: **G**iga**b**ytes
GHz: **G**iga**h**ert**z**
HTP: **H**airpin **T**ips **P**ercent
HT-SELEX: High-throughput systematic evolution of ligands by exponential enrichment
K_d: **D**issociation **C**onstant
NGS: **N**ext **G**eneration **S**equencing
r: **P**earson's **C**orrelation **C**oefficient
r_s: **S**pearman's **R**ank **C**orrelation **C**oefficient
RAM: **R**andom **A**ccess **M**emory
SRA: **S**hort **R**ead **A**rchive
ZOTU: **Z**ero-radius **O**perational **T**axonomic **U**nit

REFERENCES

Alam KK, Chang JLandBurke DH. 2015. FASTAptamer: A Bioinformatic Toolkit for High-throughput Sequence Analysis of Combinatorial Selections. Mol Ther Nucleic Acids, 4, e230.

Caroli J, Taccioli C, De La Fuente A, Serafini P and Bicciato S. 2016. APTANI: a computational tool to select aptamers through sequence-structure motif analysis of HT-SELEX data. Bioinformatics, 32, 161–164.

Coleman DAandWoodruff DL. 2000. Cluster Analysis for Large Datasets: An Effective Algorithm for Maximizing the Mixture Likelihood. J Comput Graph Stats, 9, 672–688.

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 Syst, 3, 62–70.

Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics, 26, 2460–2461.

Edgar CE. 2016. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. Biorxiv, 081257.

Hoinka J, Berezhnoy A, Dao P, Sauna ZE, Gilboa E and Przytycka TM. 2015. Large scale analysis of the mutational landscape in HT-SELEX improves aptamer discovery. Nucleic Acids Res, 43, 5699–5707.

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

Rubio JM, Svobodova M, Mairal T and O'Sullivan CK. 2016. Surface plasmon resonance imaging (SPRi) for analysis of DNA aptamer:beta-conglutin interactions. Methods, 97, 20–26.

Kinghorn A, Fraser L, Lang S, Shiu S and Tanner J. 2017. Aptamer Bioinformatics. Int J Mol Sci, 18, 2516.

Kupakuwana GV, Crill JE, 2nd, McPike MPandBorer PN. 2011. Acyclic identification of aptamers for human alpha-thrombin using over-represented libraries and deep sequencing. PLoS One, 6, e19395.

Lakhin AV, Tarantul VZ and Gening LV. 2013. Aptamers: Problems, Solutions and Prospects. Acta Naturae, 5, 34–43.

Levay A, Brenneman R, Hoinka J, et al. 2015. Identifying high-affinity aptamer ligands with defined cross-reactivity using high-throughput guided systematic evolution of ligands by exponential enrichment. Nucleic Acids Res, 43, e82.

Lorenz R, Bernhart SH, Honer Zu Siederdissen C, et al. 2011. ViennaRNA Package 2.0. Algorithms Mol Biol, 6, 26.

Paulson JN, Stine OC, Bravo HC and Pop M. 2013. Differential abundance analysis for microbial marker-gene surveys. Nature Methods, 10, 1200.

Schirmer M, Ijaz UZ, D'Amore R, Hall N, Sloan WT and Quince C. 2015. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. Nucleic Acids Research, 43, e37-e37.

Takahashi M, Wu X, Ho M, Chomchan P, Rossi JJ, Burnett JC, et al. 2016. High throughput sequencing analysis of RNA libraries reveals the influences of initial library and PCR methods on SELEX efficiency. Sci Rep, 6, 33697.

Tolle F, Wilke J, Wengel J and Mayer G. 2014. By-Product Formation in Repetitive PCR Amplification of DNA Libraries during SELEX. PLoS One, 9, e114693.

Tsao S-M, Lai J-C, Horng H-E, Liu T-C and Hong C-Y. 2017. Generation of Aptamers from A Primer-Free Randomized ssDNA Library Using Magnetic-Assisted Rapid Aptamer Selection. Sci Rep, 7, 45478.

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

Zhou J and Rossi J. 2016. Aptamers as targeted therapeutics: current potential and challenges. Nat Rev Drug Discov, 16, 181.