Supporting Information: Systematic Review of Aptamer Sequence Reporting in the Literature Reveals Widespread Unexplained Sequence Alterations

Alexandra A. Miller [1], Abhijit S. Rao[1], Sujana R. Nelakanti [1], Christopher Kujalowicz[1], Ted Shi[1], Ted Rodriguez[1], Andrew D. Ellington [†,‡], Gwendolyn M. Stovall[*,1,2]

[†]Institute for Molecular Biosciences, The University of Texas at Austin, Austin, Texas 78712, United States
[‡] Center for Systems and Synthetic Biology, The University of Texas at Austin, Austin, Texas 78712, United States
[1]Texas Institute for Discovery Education in Science Freshman Research Initiative, The University of Texas at Austin, Austin, Texas 78712, United States
[2]Texas Institute for Discovery Education in Science High School Research Initiative, The University of Texas at Austin, Austin, Texas 78712, United States

*Corresponding Author E-mail: gwenstovall@utexas.edu

KEYWORDS: aptamers, aptamer sequence fidelity, unexplained sequence alterations, in silico alterations

Table of Contents

Table S1. Search Terms Used for Literature Reviews. Ordered by frequency of aptamer target used in application publications (Dunn et al., 2017).

Aptamer Target	"Search within citing articles" Search Term
Human Thrombin	"aptamer" "human thrombin"
Adenosine Triphosphate (ATP)	"aptamer" "ATP"
Vascular Endothelial Growth Factor (VEGF)	"aptamer" "vegf"
Platelet Derived Growth Factor (PDGF) BB	"aptamer" "pdgf bb"
Cocaine	"aptamer" "cocaine"
Theophylline	"aptamer" "theophylline"
Lysozyme (Cox et al., 2001 and Tran et al., 2010)	"aptamer" "lysozyme"
Nucleolin	"aptamer" "nucleolin"
Immunoglobulin E (IgE)	"aptamer" "ige"
Ochratoxin A (OTA)	"aptamer" "ochratoxin a"

Table S2. Percentage Publications with Unexplained Sequence Alterations within Each Phylogeny.

Aptamer Phylogeny	Percentage Publications with Unexplained Sequence Alterations
Thrombin aptamer	13%
ATP aptamer	87%
VEGF aptamer	11%
PDGF BB aptamer	84%
Cocaine aptamer	51%
Theophylline aptamer	8%
Cox Lysozyme aptamer	93%
Tran Lysozyme aptamer	73%
Nucleolin aptamer	34%
IgE aptamer	31%
OTA aptamer	22%

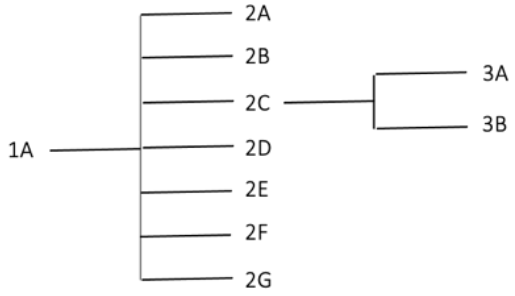
* Of note, in cases where multiple original sequences were selected and used in the literature (e.g., ochratoxin, PDGF-BB, cocaine, and IgE), these separate phylogenies were grouped together. Thus, the percent is unexplained sequence alteration found in all publications citing the original work, grouped by target.

Table S3. Total Number of Unique Sequences and Sequences with Unexplained Alterations Identified.

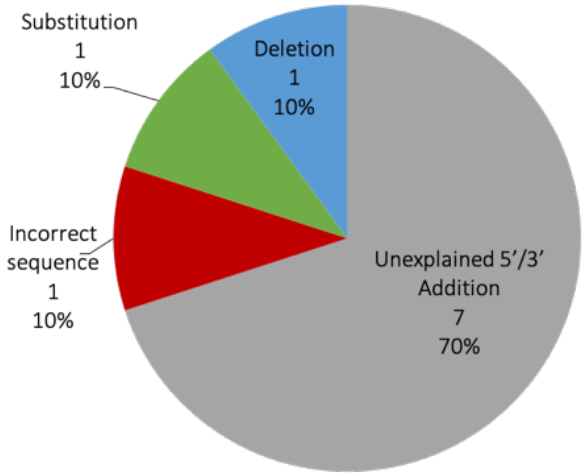
Aptamer Phylogeny	Number of Unique Sequences (IDs) Identified, Including Explained Alterations	Number of Unique Unexplained Sequence Alterations	Notes
Thrombin aptamer	10	9	One originating sequence
ATP aptamer	13	12	One originating sequence
VEGF aptamer	8	0	Seven originating clones
PDGF BB aptamer	22	19	Three originating sequences
Cocaine aptamer	30	22	Six described altered sequences and two originating clones.
Theophylline aptamer	7	6	One originating sequence
Cox Lysozyme aptamer	9	8	One original sequence examined (six clones originally).
Tran Lysozyme aptamer	5	4	One originating sequence
Nucleolin aptamer	5	3	One originating and one described sequence alteration
IgE aptamer	14	12	Two originating sequences
OTA aptamer	29	25	Four originating sequences
Total Number of Sequences	152	120 (79% of unique sequences)	41 total explained and characterized sequences, 23 originating clones total from 11 original selections

* Of note, in cases where multiple original sequences were selected and used in the literature (e.g., ochratoxin, PDGF-BB, cocaine, and IgE), these separate phylogenies were grouped together. Thus, the number of unique sequences reflects the percent unexplained sequence alteration found in all publications citing the original work, grouped by target.

A. Phylogeny of Unexplained Aptamer Sequence Alterations



B. Distribution of Apparent Error



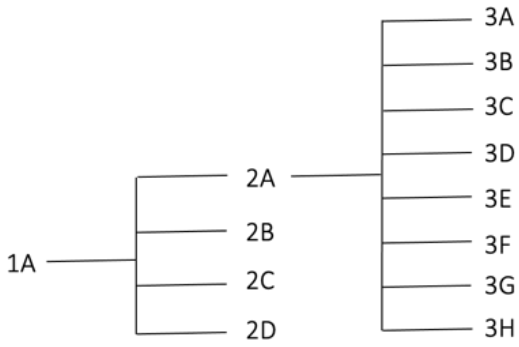
Node	*#*	*Aptamer Sequence Reported*
1A	61	GGTTGGTGTGGTTGG
2A	1	GGTGGTGGTGGTGTT
2B	1	**T**GGTTGGTGTGGTTGG**T**
2C	1	**TTTTT**GGTTGGTGTGGTTGG
3A	1	**TTTTTTTT**GGTTGGTGTGGTTGG
3B	1	**TTTTTTTTTT**GGTTGGTGTGGTTGG
2D	1	**ACTGT**GGTTGGTGTGGTTGG
2E	1	**CCATCTCCACTT**GGTTGGTGTGGTTGG
2F	1	**CCACC**GGTTGGTGTGGTTG~~G~~
2G	1	GACAGACGATGTGCTGACTACTGGTTGGTG**A**GGTTGGGTAGTCAGCACATCGTCTGTC
1B	1	GGTTGG
Total	71	

Figure S1 | Phylogeny depicting unexplained aptamer sequence alterations introduced to the 15mer DNA thrombin binding aptamer (TBA15, Bock et al., 1992). Aptamer sequences collected in December 2019.

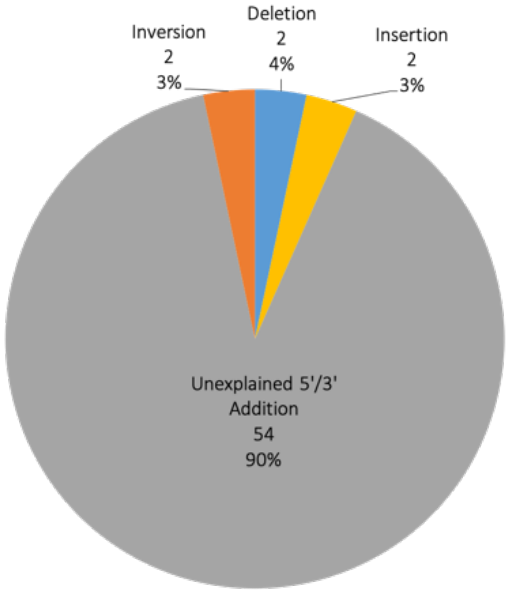
A. The nodes in the artificial phylogenetic trees represent aptamers grouped according to sequence homology and alterations. B. Unexplained aptamer sequence alterations were categorized.. In the table, unexplained insertions and additions are bolded, unexplained substitutions are bolded and underlined, unexplained deletions are struck out and justified or explained alterations are in light grey. The number (#) column indicates the number of times each sequence was reported, where multiple publications provide or use multiple clones.

Nine unexplained sequences were found in subsequent publications from **node 1A/1B**. **Node 2A** provided an unknown incorrect sequence that resembled the original Bock et al. (1992) sequence. In **node 2B** a thymine base was added to both the 5' and 3' end of the original 15-mer aptamer sequence, resulting in a 17-mer sequence. In **nodes 2C, 3A, 3B**, with one publication per node, each sequence alteration involved the unexplained addition of a 5' PolyT tail of varying lengths. Specifically, in **node 2C** there was a Poly5Tail, in **node 3A** there was a Poly8 Tail, and in **node 3B**, there was a Poly10 tail. **Node 2D** and **node 2E** have unexplained 5' additions. **Node 2F** contains an unexplained 5' addition as well as a deletion of a single 3' nt. Finally, **node 2G** contains described 5'/3' additions and a single substitution.

A. Phylogeny of Unexplained Aptamer Sequence Alterations



B. Distribution of Apparent Error



Node	*#*	*Aptamer Sequence Reported*
1A	9	CCTGGGGGAGTATTGCGGAGGAAGG
2A	44	**A**CCTGGGGGAGTATTGCGGAGGAAGG**T**
3A	2	**T**GGAAGGAGGCGTTATGAGGGGTCC**A**
3B	2	**AA**CCTGGGGGAGTATTGCGGAGGAAGG**T**
3C	1	**CCTCCTA**CCTGGGGGAGTATTGCGGAGGAAGG**TA**
3D	1	**TCTCTCA**CCTGGGGGAGTATTGCGGAGGAAGG**T**
3E	1	**A**CCTGGGGGAGTATTGCGGAGGAAGG**TTTT**
3F	1	**A**CCTGGGGGAGTATTGCGGAGGAAGG**TGTCACA**$(A)_{10}$
3G	1	**A**CCT**TCCT**GGGGGAGTATTGCGGAGGAAGG**T**
3H	1	**A**CCTGGGGGA~~G~~TA**A**TTGCGGAGGAAGG**T**
2B	1	CCTGGGGGA~~G~~TATTGCGGAGGAAGG
2C	4	**CA**CCTGGGGGAGTATTGCGGAGGAAGG**TT**
2D	1	**GCA**CCTGGGGGAGTATTGCGGAGGAAGG**T**
Total	69	

Figure S2 | Phylogeny depicting unexplained aptamer sequence alterations introduced to the 24mer DNA Adenosine Triphosphate (ATP) binding aptamer (DH29.36, Huizenga et al., 1994).

A. The nodes in the artificial phylogenetic trees represent aptamers grouped according to sequence homology and alterations. B. Unexplained aptamer sequence alterations were categorized. In the table, unexplained insertions and additions are bolded, unexplained substitutions are bolded and underlined, unexplained deletions are struck out and justified or explained alterations are in light grey. The number (#) column indicates the number of publications found reporting each sequence in our analysis.

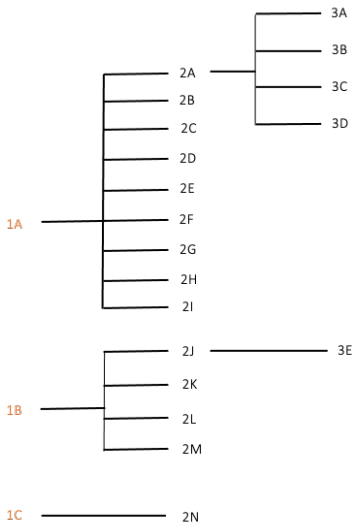
The DNA ATP binding aptamer (ABA) was discovered by Huizenga et al. (1994) and a minimized variant, or conserved sequences, was determined and folding structure was predicted (**node 1A**). Urata et al. subsequently describe three minimized variants, the 25mer described by Huizenga et al., a 23mer cutting 1 nt at the 5' and 3' ends, and one adding an A to the 5' and T to the 3'. However, all but 5 groups identified (**node 2A**) use the 27mer without citing Urata's work minimizing the variant or a justification for using this over the 25mer. **Node 3A** published the 27mer, but inverted the sequence giving the 3' to 5' sequence labeled as 5' to 3'. **Node 3B** added an additional A to the 5' end. **Node 3C** introduced an additional CCTCCT to the 5' end and an A to the 3' end without explanation of their purpose. The added A and T are presumably for stability, and the additional 5' modification was likely created for the creation of a duplex structure or a hairpin that then changes conformation upon binding to ATP making it vulnerable to Exo III degradation. **Node 3D** includes an additional TCTCTC at the 5' end of the sequence. The linker is presumably to maintain attachment to the carbon nanotube matrix when binding ATP, but this is not stated clearly. **Node 3E** includes an additional 3 T's at the 3' without justification, and **node 3F** added a stretch of nucleotides to the 3' end before the stretch of adenosines without justification. **Node 3G** includes a 4nt insertion in the center of the aptamer sequence. Finally, **node 3H** provide a sequence with a deletion and insertion. Importantly, although many of these mutations (**2A, 3B-F**) most likely served a purpose within the application, the modifications made to the original sequence were not described or cited. Notably, one paper gave the sequence with added A and T for stability in a split aptamer design and cite the sequence correctly in the text and in most placed in their figure but incorrectly cite the sequence in one of the tertiary structures, deleting a G in the middle of the stem. This was not cited as an altered sequence in the phylogeny due to the presence of correct sequence information in all other places. **Node 2B** contained a single internal deletion, and **node 2C** and **node 2D** contained unexplained 5'/3' additions.

Node	#	Aptamer Sequence Reported
1A	2	CCGGUAGUCGCAUGGCCCAUCGCGCCCGG
1B	2	(GG)[CCGG]UAGU[CGC]AUGGCCCAUC[GCG]C[CCGG]
1C	3	(GG)a[aGCUUG]AUGG[GUGAC]ACAC[GUCAU]GC[CGAGCu]
2A	1	GGAAGCUUGAUGGGUGACACACGUCAUGCCGAGCUGGCUUCC
1D	2	(G)[GAAGG]GAACC[UGC]GUCUCG[GCACCuuc]g
1E	2	(G)[GUCAACGG]UU[GA]GUCUG[UCCCGuu]c[gac]
1F	2	(G)[gcuc]a[aU]AGUUG[GAGG]CCUGU[CCUC]GCC[GU]A[GAGC]
1G	2	(GG)[aACGGUU]CU[GUGUG]UGGAC[UA]GC[CGCGGCCGuu]
1H	3	CGGAAUCAGUGAAUGCUUAUACAUCCGdT28
0	1	No sequence Provided
Total	20	

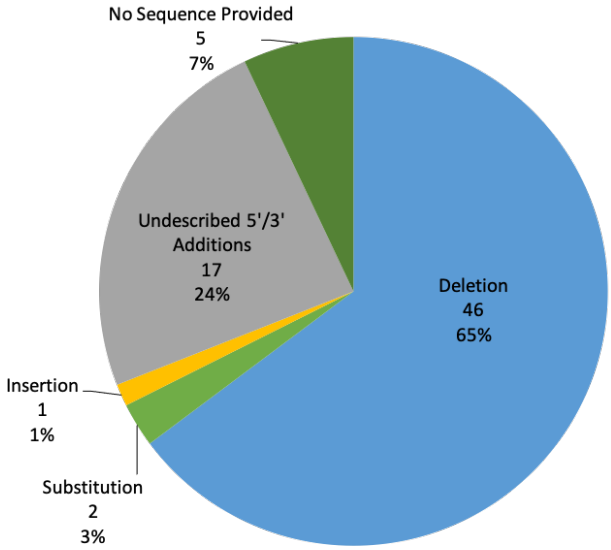
Figure S3 | Usage of the RNA vascular endothelial growth factor (VEGF) binding aptamers: family 1 (15mer), family 2 (22mer), family 3 (25mer), family 4 (19mer), family 5 (27mer), and family 6 (32mer) (Jellinek et al., 1994). Aptamer sequences collected in January 2020.

Explained additions are bolded and justified or explained alterations are in light grey. The number (#) column indicates the number of times each sequence was reported, where multiple publications provide or use multiple clones. Multiple clonal sequences were noted; for all important to note that (Gs in parentheses) represent guanines not present in the original sequences & were added to increase transcriptional efficiency, lowercase letters indicate nucleotides from the constant sequence region, [bracketed] are those nucleotides predicted to be base paired by the original authors. Search results up until 100 publications were analyzed however only 9 publications were not excluded. The remaining 91 publications were not included in accordance with our protocol because these consisted of reviews, reported novel selections, did not use the aptamer experimentally, or were written in languages other than English. Among the 9 publications, 1 publication did not report a sequence. The Jellinek et al., 1994 publication multiple clonal sequences (**1A-1G**) of the VEGF aptamer are clearly listed from six different families. In the **node 2A**, an explained 3' addition to the 44t aptamer. Additionally, **node 1H** is Macugen, which made explained chemical modifications for optimized affinity and stability.
One of publications reviewed (11%) did not provide the aptamer sequence, which is particularly problematic due to the large number of original sequences that bind VEGF (21 original aptamer sequences identified and excluded from this research alone). This phylogeny illuminates the importance of characterization of the aptamer sequences and selection based on the properties most beneficial for the application.

A. Phylogeny of Unexplained Aptamer Sequence Alterations



B. Distribution of Apparent Error



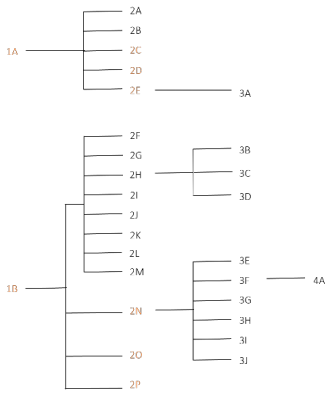
Node	*#*	*Aptamer Sequence Reported*
1A	6	CACAGGCTACGGCACGTAGAGCATCACCATGATCCTGTG
2A	27	~~CA~~CAGGCTACGGCACGTAGAGCATCACCATGATCCTG~~TG~~
3A	2	~~CA~~CAGGCTACGGCACGTAGAGCATCACCATGATCCTG~~TG~~**TTTTT**
3B	1	~~CA~~CAGGCTACGGCACGTAGAGCATCACCATGATCCTG~~TG~~**TTTTTT**
3C	1	~~CA~~CAGGCTACGGCACGTAGAGCATCACCATGATCCTG~~TG~~**AAAAAAAAAAAAAAAAAAAAAAAA**
3D	3	**TTTTTTTTTT**~~CA~~CAGGCTACGGCACGTAGAGCATCACCATGATCCTG~~TG~~
2B	3	CACAGGCTACGGCACGTAGAGCATCACCATGATCCTGTG**T**
2C	6	~~CA~~CAGGCTACGGCACGTAGAGCATCACCATGATCCTG~~TG~~
2D	1	~~CA~~CAGGCTACGGCACGTAGAGCATCACCATGATCCTGTG**ACTAACCTG** (G_4)
2E	1	~~CA~~CAGGCTACGGCACGTAGAGCATCACCATGATCCTGT~~G~~
2F	1	C**T**CAGGCTACGGCACGTAGAGCATCACCATGATCCTG**A**G
2G	1	CACAGGCTACGGCACGTAGAGCATCACCATGATCCTGTG**TT**
2H	1	CACAGGCTACGGCACGTAGAGCATC**ATC**ACCATGATCCTGTG**T**
2I	1	~~CAC~~AGGCTACGGCACGTAGAGCATCACCATGATCCT~~GTG~~
1B	5	TACTCAGGGCACTGCAAGCAATTGTGGTCCCAATGGGCTGAGTA
2J	1	TACTCAGGGCACTGCAAGCAATTGTGGTCCCAATGGGCTGAGTA**T**
3E	2	**AAAAAAAAAA**TACTCAGGGCACTGCAAGCAATTGTGGTCCCAATGGGCTGAGTA**T**
2K	1	~~T~~ACTCAGGGCACTGCAAGCAATTGTGGTCCCAATGGGCTGAGTA
2L	1	TACTCAGGGCACTGCAAGCAATTGTGGTCCCAATGGGCTGAGTA**TTTTTT**
2M	1	TACTCAGG~~G~~CACTGCAAGCAATTGTGGTCCCAATGGGCTGAGTA
1C	2	TGGGAGGGCGCGTTCTTCGTGGTTACTTTTAGTCCCG
2N	1	**A**GGGCG~~GG~~CG~~CG~~TTCTTCGTGGTTACTTTTAGTCCCG
0	5	No sequence provided
Total	74	

Figure S4 | Phylogenies depicting unexplained aptamer sequence alterations introduced to the three DNA platelet-derived growth factor (PDGF-BB) binding aptamers: 36t a 36mer (node 1A), 41t a 41mer (1B), and 20t a 20mer (1C), Green and Jellinek, 1996.

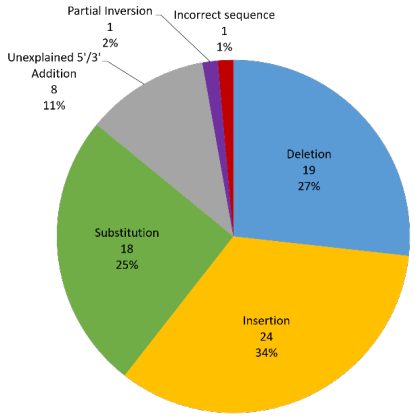
A. The nodes in the artificial phylogenetic trees represent aptamers grouped according to sequence homology and alterations. B. Unexplained aptamer sequence alterations were categorized. In the table, unexplained insertions and additions are bolded, unexplained substitutions are bolded and underlined, unexplained deletions are struck out and justified or explained alterations are in light grey. The number (#) column indicates the number of times each sequence was reported, where multiple publications provide or use multiple clones.

The above phylogeny documents the alteration to the PDGF ssDNA aptamer sequences originally identified by Green and Jellinek 1996 using the SELEX methodology. Green and Jellinek identified three different minimized variant sequences that exhibited binding to PDGF BB (**1A, 1B, 1C**). **Node 2A** reported a sequence that truncated the original sequence by two nucleotides on either end, and was the most commonly-reported sequence (39%). Sequences stemming from this node (**3A-D**), employed various unjustified 5' or 3' extensions to the 2A sequence. **Node 2B** contains an unexplained 3' T addition, **node 2C** deletes the first nt and last 2 nucleotides, **node 2D** deletes the first nucleotide and makes an unexplained 3' addition, **node 2E** deletes one nucleotide from both the 5' and 3' end, **node 2F** makes two unexplained substitutions, **node 2G** contains an unexplained 3' TT, **node 2H** inserts an additional ATC in the sequence and makes a 3' T addition, and **node 2I** deletes 3nt from both the 5' and 3' ends. Sequence alterations made to **sequence 1B** largely consisted of unexplained 5' or 3' additions (**nodes 2J, 3E, and 2L**). **Node 2K** deleted a 5' nucleotide and **node 2M** deleted a G in the core binding sequence. Of note, **node 2N** reports a sequence that included deletions of nucleotides within the region. While this group was able to generate results from this sequence alteration, target binding characteristics could have changed with the introduction of these altered sequences.

A. Phylogeny of Unexplained Sequence Alterations



B. Distribution of Apparent Error

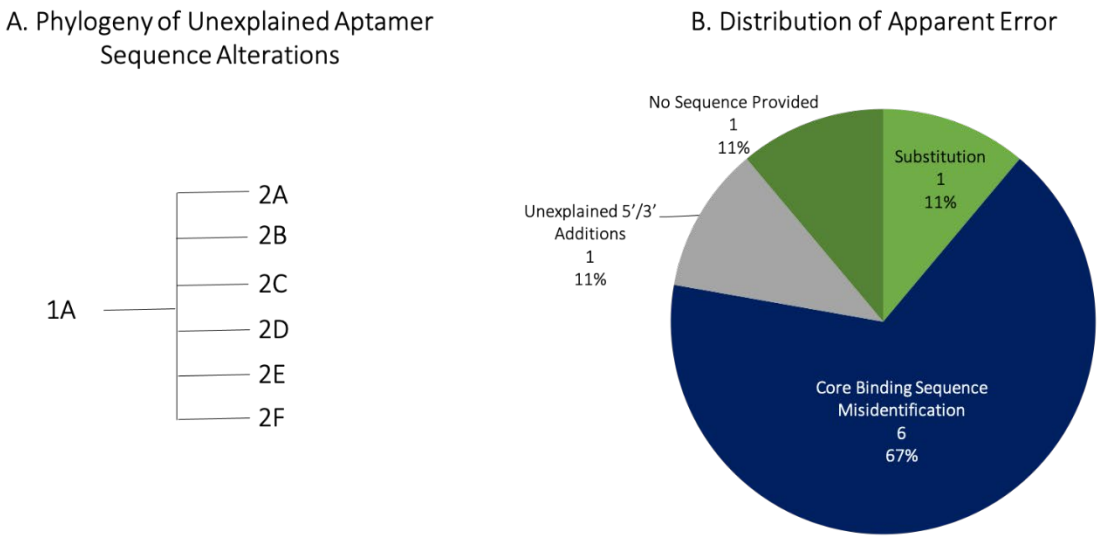


Node	#	Aptamer Sequence Reported
1A	9	GACAAGGAAAATCCTTCAATGAAGTGGGTC
2A	1	GACAAGGA**T**AAATCCTTCAATGAAGTGGGTCACTCATCTGTGA
2B	1	**A**GACAAGGAAAATCCTTCAATGAAGTGGGTC**TC**
2C	7	**A**GACAAGGAAAATCCTTCAATGAAGTGGGTC**G**
2D	7	**GGC**GACAAGGAAAATCCTTCAA**C**GAAGTGGGTC**GCC**
2E	2	GACAAGGAAAATCCTTCAA**C**GAAGTGGGTC
3A	1	GACAAG~~GAAAA~~**GCGACGTCGC**T**T**CTTCAA**C**GAAGTGGGTC
1B	6	GGGAGACAAGGAAAATCCTTCAATGAAGTGGGTCGACA
2F	1	GGGAGACAAGGAAA~~A~~TCCTTCAATGAAGTGGGTCGACA
2G	1	GGGAG**T**CAAG~~G~~AA**C**, 5' ACAGC**A**GGGTGAAGTAACTTC**T**T**G** 3'
2H	4	GGGAG**T**CAAG~~G~~AA**CGAATTCGT**TC~~G~~TTCAATGAAGTGGG**A**CGACA
3B	1	GGGAG**T**CAAG~~G~~AA**CGAATTCGT**TC~~G~~TTCAATGAAGTGGG**A**CGACA**TTTCTTCTTCTTTTCATGTCGTCCGTTT**
3C	1	**TTTTT**GGGAG**T**CAAG~~G~~AA**CGAATTCGT**TC~~G~~TTCAATGAAGTGGG**A**CGACA
3D	1	GGGAG**T**CAAG~~G~~AA**CGAATTCGT**TC**G**TTC**TT**CAATGAAGTGGG**A**CGACA
2I	3	GGGAG**T**CAAG~~G~~AA**CGT**TCTTCAATGAAGTGGG**A**CGACA
2J	1	**C**GG**C**GACAAG~~G~~AA**GCGACGTCGC**T**T**CTTCAA**C**GAAGTGGGTCGAC**G**
2K	1	GGGAG**T**CAAG~~G~~AA~~AA~~**CTTTG**T**T**CTTCAAT~~GA~~AGTGGG**A**CGACA
2L	1	GGGAG**T**CAAG~~G~~AAAA**CCGG**T**T**CTTCAATGAAGTGG~~TG~~GACA
2M	1	GGGAG**T**CAAG~~G~~AA**CA**AA**GT**TCCTTCAATGAAG**T****GT**GGG**A**CGACA
2N	2	GGGAGACAAGGAAAATCCTTCAATGAAGTGGGTC**TCCC**
3E	2	**GG**GGGAGACAAGGAAAATCCTTCAATGAAGTGGGTCTCCC
3F	3	GGGAGACAAGGA**T**AAATCCTTCAATGAAGTGGGTCTCCC
4A	1	GGGAGACAAGGA**T**AAATCCTTCAATGAAGTGGGTC~~TCCC~~
3G	1	**ACTCATCTGTGAATCTC**GGGAGACAAGGA**T**AAATCCTTCAATGAAGTGGGTCTCCC
3H	1	**CCATA**GGGAGACAAGGA**T**AAATCCTTCAATGAAGTGGGTCTCCC
3I	1	TTTTTTTACTCATGGGAGACAAGGA**TT**AATCCTTCAATGAAGTGGGTCTCCC
3J	1	**ATCTC**GGGA**C**GACAAGGA**T****TT**ATCCT~~T~~CAATGAAGTGGGTC**G**TCCC
2O	2	GGGAGACAAGGAAAATCCTTCAA**C**GAAGTGGGTC**TCCC**
2P	1	~~G~~GGAGACAAGGAAAATCCTTCAATGAAGTGGGTCGAC~~A~~
3	1	AGCAACCTCAAACAGACACCATGG
Total	66	

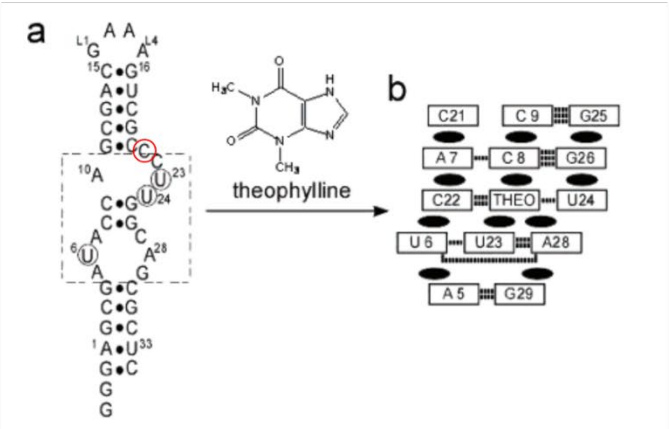
Figure S5 | Phylogenies depicting unexplained aptamer sequence alterations introduced to the two DNA cocaine binding aptamers: MNS4.1 a 30mer (node 1B) and MNS7.9/MSN6 a 38mer (1A), Stojanovic et al., 2001.

A. The nodes in the artificial phylogenetic trees represent aptamers grouped according to sequence homology and alterations. B. Unexplained aptamer sequence alterations were categorized. In the table, unexplained insertions and additions are bolded, unexplained substitutions are bolded and underlined, unexplained deletions are struck out and justified or explained alterations are in light grey. The number (#) column indicates the number of times each sequence was reported, where multiple publications provide or use multiple clones. Orange nodes indicate well described and characterized sequence alterations.

The above phylogeny follows the evolution of the MSN-4.1 (**node 1B**) and MNS-7.9/MN6 (**node 1A**) ssDNA cocaine aptamer sequences since their discovery in 2001 by Stojanovic, et al. The MSN-7.9 aptamer was constructed by truncating the MSN-4.1 aptamer and is described "with a shortened S_1 stem [and found] that it retained significant, albeit reduced, affinity for cocaine." They further modify MSN-7.9 with a 5'-fluorescein and 3' dabcyl to mitigate effects of decreased binding. A 32-mer (**node 2C**) with single nucleotide additions to both the 5' and 3' ends of the original MN6 sequence was well described and characterized. Further, two sequence variants were described **(nodes 2D and 2E)** that include a single nucleotide substitution in stem-loop 3, shifting a noncanonical G-T base pairing to G-C. In addition to this substitution, the MN4 sequence (**node 1C**) features three-nucleotide additions to both the 5' and 3' ends, while the MN19 sequence (**2E**) does not share this change. The original 2001 Stojanovic publication included an additional sequence, MNS 4.1 (**node 1B**), which included a four-nucleotide addition to both the 5' and 3' ends. **Node 2B** is a well described a 38-mer in which the final four nucleotides on the 3' end are replaced with TCCC for stabilization of the stem. **Node 2C** contains this same substitution in addition to a single nucleotide substitution for added stability. Finally, the MN1 sequence (**node 2D**) featured the deletion of a single 5' and 3' nucleotide. All other changes noted in the phylogeny above made changes to the original MN6 or MNS 4.1 sequences without sufficient explanation of their modifications. The **node 2G** used a split aptamer design, inverting one fragment, but not the other in addition to a deletion and five undescribed substitutions. **Node 3** provided an unknown sequence, citing the Stojanovic et al., 2001 sequence. Upon review of the included publications for this analysis, it was noted that the downstream applications for the cocaine aptamer were almost exclusively sensor-based and diagnostic in nature, primarily using in-vitro methodology.



C. Core Binding Sequence Misidentification in Node 2D



Node	*#*	*Aptamer Sequence Reported*
1A	103	(Nx) AUACCA (Nx) CCUUGG (C/A) AG (Nx)
2A	1	(AAGUG) **C**UACCA (GCAUCGUCUUGAUGC) CCUUGGCAG (CACUUCA)
2B	1	$(CH_2)_6$ CCUUGGAAG *(CC)* 5′ *(GG)* AUACCA **$(CH_2)_6$–HS**
2C	1	(GGC) **G**AUACCA**G** (CCGAAAGG) **C**CCUUGGCAG**C** (GUC)
2D	1	(Nx) ~~A~~UACCA (GCCGAAAGGCC) CUUGGC~~A~~G (Nx)
2E	3	*(Nx)* **G**AUACCA *(Nx)* CCUUGGCAG**C** *(Nx)*
2F	1	(GGGAGCG) AUACCA (GCGACGAAAGUCGC) ~~c~~CUUGGCAG (CGCUC)
0	1	No sequence provided
Total	112	

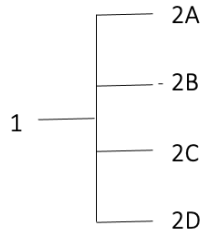
Figure S6 | Phylogeny depicting unexplained aptamer sequence alterations introduced to the RNA anti-theophylline aptamer (the conserved theophylline binding region, Jenison et al., 1994).

A. The nodes in the artificial phylogenetic trees represent aptamers grouped according to sequence homology and alterations. B. Unexplained aptamer sequence alterations were categorized. C. The core binding sequence misattribution designated as node 2D is shown. From the original Jenison et al. 1994 paper, the conserved/consensus core binding sequences were identified as *region 1* 5′-AUACCA-3′ and *region 2* 5′-CCUUGG(C/A)AG-3′. In contrast, a figure from the Kawai et al. 2005 paper (reprinted above in Figure S6C) identified "the structural interactions in the binding site," which did not include the first 5′-C in region 2 (noted by our group with a red circle). This nucleotide omission is not discussed in the text. This cited figure is reprinted in part with permission from Kawai, R., Kimoto, M., Ikeda, S., Mitsui, T., Endo, M., Yokoyama, S., & Hirao, I. (2005). Site-specific fluorescent labeling of RNA molecules by specific transcription using unnatural base pairs. *Journal of the American Chemical Society*, *127*(49), 17286-17295. Copyright 2004 American Chemical Society. In the lower table, unexplained insertions and additions are bolded, unexplained substitutions are bolded and underlined, unexplained deletions are struck out

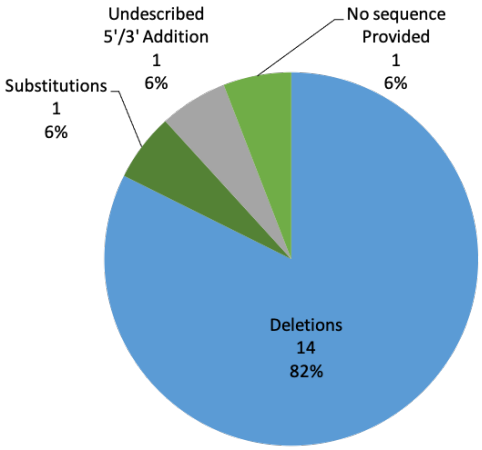
and justified or explained alterations are in light grey. The number (#) column indicates the number of publications found reporting each sequence in our analysis.

Jenison et al. 1994 selected an anti-theophylline aptamer and minimized a 14nt conserved theophylline binding region with a semiconserved C27 which can be either C or A (**node 1A**). Zimmerman et al., 1997, 1998 and 2000 later further elucidated the binding characteristic of the anti-theophylline aptamer. In the binding complex, two base triplets surround the target (U6-U23-A28 and A7-C8-G26), only forming stably in the presence of theophylline (i.e., not in the presence of caffeine, **Figure part C**). **Node 2A** made an undescribed substitution of a C to A in region 1. **Node 2B** uses a split aptamer design but contains an unexplained 3' addition. Finally, in **nodes 2C-2F** the authors include a misidentification of the core binding sequence, as described above (see C). Because this aptamer sequence misattribution does not change the effectiveness of the aptamer in these cases, we do not categorize them as unexplained sequence alterations, but rather a separate category similar to unexplained 5'/3' additions. Briefly, **node 2C** boxes an addition nucleotide on the 5' and 3' ends of both regions, **node 2D** deletes one nt from the 5' end of region 1 and one from the 3' end of region 2, **node 2E** adds a G to the 5' of region 1 and a C to the 3' of region 2, and **node 2F** contains a single addition on the 5' end of region 2.

A. Phylogeny of Unexplained Aptamer Sequence Alterations



B. Distribution of Apparent Error



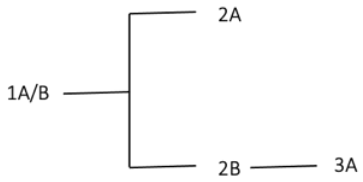
Node	*#*	*Aptamer Sequence Reported*
1	6	AGCAGCACAGAGGTCAGATGGCAGCTAAGCAGGCGGCTCACAAAACCATTCGCATGCGGCCCTATGCGTGCTACCGTGAA
2A	1	AGCAGCACAGAGGTCAGATGGCAG**G**TAAGCAGGCGGCTCACAAAACCATTCGCATGCGGCCCTATGCGTGCTACCGTGAA
2B	12	~~AGCAGCACAGAGGTCAGATG~~GCAGCTAAGCAGGCGGCTCACAAAACCATTCGCATGCGGC~~CCTATGCGTGCTACCGTGAA~~
2C	1	**CGAACAACGCTAAAATTC**~~AGCAGCACAGAGGTCAGATG~~GCAGCTAAGCAGGCGGCTCACAAAACCATTCGCATGCGGCCCTATGCGTGCTACCGTGAA**ATAGGAGACGGGCAACAGCTGATCCTGATGG**
2D	1	SH-(CH2)6- AGCAGCACAGAGGTCAGATG~~GCAG~~CTAAGCAGGCGGCTCACAAAACCATTCGCATGCGGCCCTATGCGTGCTACCGTGAA
0	1	No sequence Provided
Total	22	

Figure S7 | Phylogeny depicting unexplained aptamer sequence alterations introduced to the 80mer DNA lysozyme binding aptamer (Apta1, Tran et al., 2010).

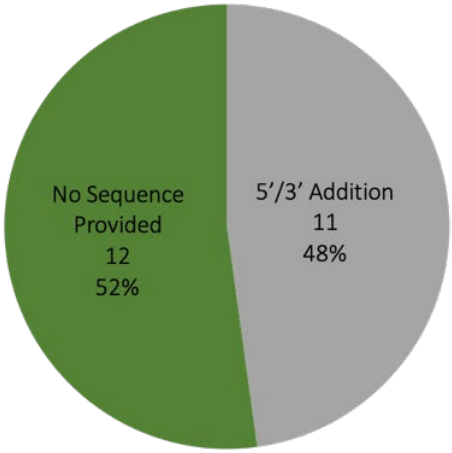
A. The nodes in the artificial phylogenetic trees represent aptamers grouped according to sequence homology and alterations. B. Unexplained aptamer sequence alterations were categorized. In the table, unexplained insertions and additions are bolded, unexplained substitutions are bolded and underlined, unexplained deletions are struck out and justified or explained alterations are in light grey. The number (#) column indicates the number of publications found reporting each sequence in our analysis.

Tran et al. selected a new ssDNA (**node 1**) aptamer in 2010 that showed higher binding affinity than the aptamer previously described by Cox et al. Since then, it has been used in a variety of aptasensors. **Node 2A** used the Tran aptamer publishing the sequence with the primers included, indicating that the first and last 22 nt were a part of the primer, with a single nt mutation with no apparent or articulated purpose. Several subsequent publications, (**node 2B**) publish the sequence without the primer without citing a minimized variant of the aptamer. **Node 2C** uses the Tran aptamer with novel primers and the deletion of the forward primer. Finally, **node 2D** deleted 4nt after the forward primer. Additionally, 1 publication presumably uses the aptamer developed by Tran et al. but does not provide the specific sequence, although several aptamers exist for the same target.

A. Phylogeny of Unexplained Aptamer Sequence Alterations



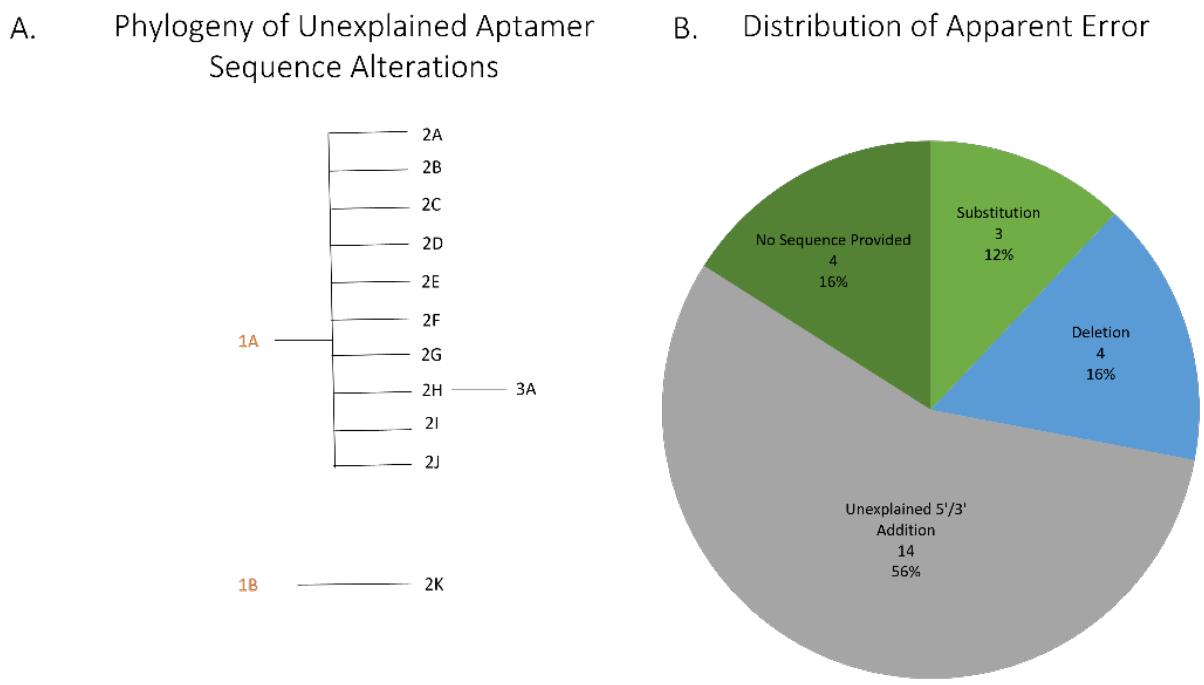
B. Distribution of Apparent Error



Node	*#*	*Aptamer Sequence Reported*
1A	6	TTTGGTGGTGGTGGTTGTGGTGGTGGTGG
1B	39	~~TTT~~GGTGGTGGTGGTTGTGGTGGTGGTGG
2A	7	**TT**GGTGGTGGTGGTTGTGGTGGTGGTGG
2B	1	GGTGGTGGTGGTTGTGGTGGTGGTGG**T**
3A	3	GGTGGTGGTGGTTGTGGTGGTGGTGG**TTT**
0	12	No sequence provided
Total	68	

Figure S8 | Phylogeny depicting unexplained aptamer sequence alterations introduced to the 26mer DNA nucleolin binding aptamer (AS1411/AGRO100, Bates et al., 1999).

A. The nodes in the artificial phylogenetic trees represent aptamers grouped according to sequence homology and alterations. B. Unexplained aptamer sequence alterations were categorized. In the table, unexplained insertions and additions are bolded, unexplained substitutions are bolded and underlined, unexplained deletions are struck out and justified or explained alterations are in light grey. The number (#) column indicates the number of publications found reporting each sequence in our analysis.
The phylogeny constructed traces alterations in the AS1411 ssDNA aptamer sequence since its discovery in 1999 by Bates et al. by screening anti-proliferative activity among antisense oligonucleotides rather than the SELEX method typical in aptamer development. One notable early modification made to the GRO29A sequence (**Node 1A**) removed the 5' 3T cap and 3'aminoalkyl because they were found to be unnecessary for serum stability and protection from nuclease activity resulting in the AS1411/AGRO100 sequence (**Node 1B**). **Node 2A** includes an undescribed 5' 2T addition. **Node 2B** states that they use a 26mer but provide a 27nt sequence with an unexplained added T at the 3' end. Similarly, **node 3A** adds three 3' Ts. Further, 12 of the 50 publications reviewed do not report the aptamer sequence.

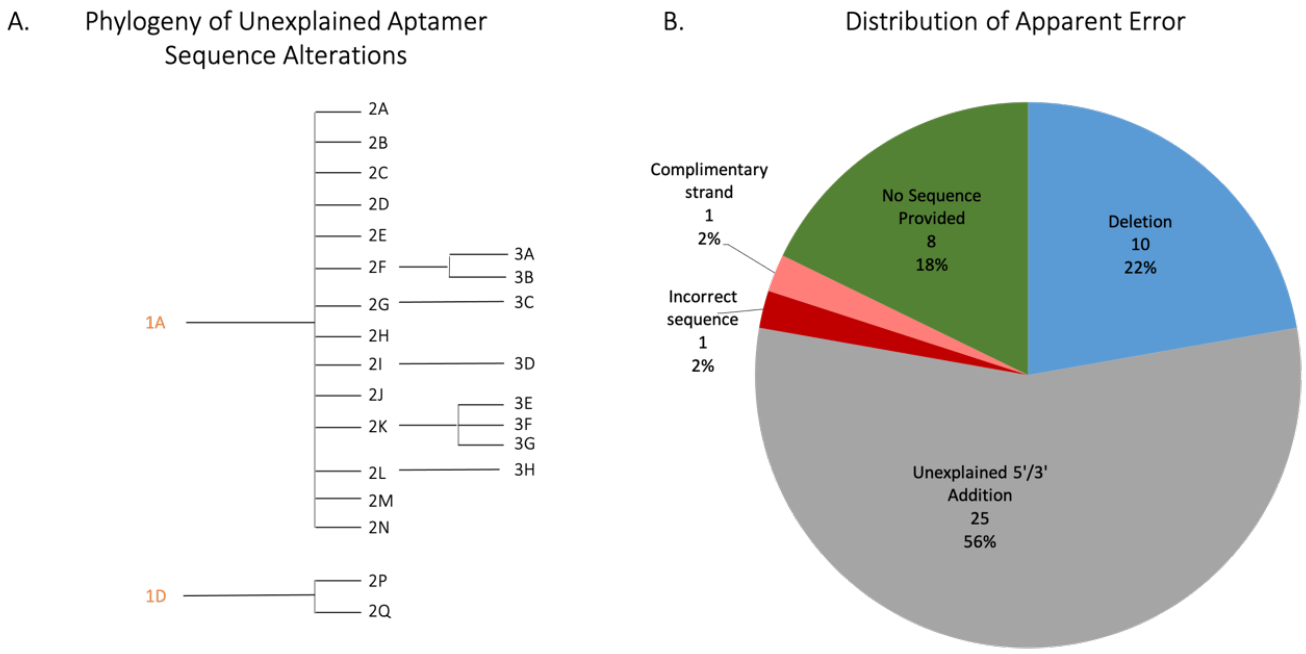


Node	*#*	*Aptamer Sequence Reported*
1A	43	GGGGCACGTTTATCCGTCCCTCCTAGTGGCGTGCCCC
2A	9	**GCGC**GGGGCACGTTTATCCGTCCCTCCTAGTGGCGTGCCCC**GCGC**
2B	1	GGGGCACGTTTATCCGTCCC~~TCCT~~AGTGGCGTGCCCC
2C	1	**ATCTAC**GGGGCACGTTTATCCGTCCCTCCTAGTGGCGTGCCCC
2D	2	~~GGGG~~TCACGTTTATCCGTCCCTCCTAGTGGCGTG**A**~~CCCC~~
2E	1	**TTTTT**GGGGCACGTTTATCCGTCCCTCCTAGTGGCGTGCCCC
2F	1	**TTTTTTTTTT**GGGGCACGTTTATCCGTCCCTCCTAGTGGCGTGCCCC
2G	1	**AAAAA**GGGGCACGTTTATCCGTCCCTCCTAGTGGCGTGCCCC
2H	3	**AAAAAAAAAAAAAAA**GGGGCACGTTTATCCGTCCCTCCTAGTGGCGTGCCCC
3A	1	**AAAAAAAAAAAAAAA**GGGGCACGTTTATCCGTCCCTCCTAGTGGCGTGCCC**G**
2I	1	~~GGG~~GCACGTTTATCCGTCCCTCCTAGTGGCGTGC~~CCC~~
2J	1	GGGGCACGTTTATCCGTCCCTCCTAGTGGCGTGCCCC**A**
1B	1	TTTATCCGTTCCTCTTAGTGG
2K	2	TTTATCCGTTCCTC**C**TAGTGG
0	4	No sequence provided
Total	72	

Figure S9 | Phylogenies depicting unexplained aptamer sequence alterations introduced to the two DNA Immunoglobulin E (IgE) binding aptamers: D17.4 a 37mer (node 1A) and "consensus sequence" a 21mer (2A) Wiegand et al., 1996.

A. The nodes in the artificial phylogenetic trees represent aptamers grouped according to sequence homology and alterations. B. Unexplained aptamer sequence alterations were categorized. In the table, unexplained insertions and additions are bolded, unexplained substitutions are bolded and underlined, unexplained deletions are struck out and justified or explained alterations are in light grey. The number (#) column indicates the number of times each sequence was reported, where multiple publications provide or use multiple clones.

The IgE binding aptamer, clone D17.4, was discovered by Wiegand et al. (1996) (**ID 1A**). Stjovanic et al., also described a DNA IgE consensus sequence was discovered as well (**node 1B**). In the Wiegand et al. (1996) publication, a secondary structure was proposed with four added G-C pairs, two to each 5' and 3' end.to promote extra stability of the sequence. This addition of G-C pair nucleotides is present in a subsequent publication in which the aptamer is now referred to as D17.4ext, and this publication clearly reports that the 8-nucleotide expansion was added to promote the stability of the secondary structure (**node 2A**). Subsequent publications reported the same sequence with the G-C pair addition described; however, two of the publications failed to report the purpose this change served in their application. In **Node 2B** 3 nucleotides C-C-T were deleted from the middle of the sequence, but the purpose of this change was not explained nor was a secondary structure prediction made in either publication. Several groups made unexplained 5' or 3' modifications to the sequence (**nodes 2C, 2E, 2F, 2G, 2H,** and **2J**) noted here. In **Node 2D** the first four GGGG nucleotides were deleted from the 5' end and the last four CCCC nucleotides were deleted from the 3' end, a T nucleotide was added to the beginning of the 5' end and an A nucleotide was added to the end of the 3' end. **Node 3A** contains the same unexplained 5' poly A as node 1I, but also contains a 3' substitution. In **node 2I**, a three-nucleotide deletion was made to both the 5' and 3' end. Interestingly, although 6 nucleotides were deleted from the sequence, the aptamer was still referred to as a 37-mer sequence even though the length was transformed to a 31-mer sequence. When tracking the changes of the secondary sequence presented in node 2A, two publications made a substitution that converted a thymine base into a cytosine base (**node 2K**).



Node	*#*	*Aptamer Sequence Reported*
1A	122	GATCGGGTGTGGGTGGCGTAAAGGGAGCATCGGACA
2A	2	GATCGGGTGTGGGTGGCGTAAAGGGAGCATCGGACA**GGCGAAGACAGGTGCTTAGT**
2B	1	GGAGGACGAAGCGGAACCGGGTGTGGGTGCCTTGATCCAGGGAGTCTCAGAAGACACGCCCGACA
2C	1	TGTCCGATGCTCCCTTTACGCCACCCACACCCGATC
2D	1	TTCGATCGGGTGTGGGTGGCGTAAAGGGAGCATCGGA~~CA~~AAAA
2E	1	**GCTGAGTCTGAGTC**GATCGGGTGTGGGTGGCGTAAAGGGAGCATCGGACA
2F	4	**GCATCT**GATCGGGTGTGGGTGGCGTAAAGG~~GAGCATCGGACA~~
3A	1	AAGGGTAGGGCGGGTTGGGTAAATAAAAA**GCATCT**GATCGGGTGTGGGTGGCGTAAAGG~~GAGCATCGGACA~~**AAA**
3B	1	AAGGGTAGGGCGGGTTGGGTAAAT**GCATCT**GATCGGGTGTGGG~~T~~GGCGTAAAGG~~GAGCATCGGACA~~**AAA**
2G	1	**TTTTT**GATCGGGTGTGGGTGGCGTAAAGGGAGCATCGGACA
2H	1	**CCTGGGAGGGAGGGAGG**GATCGGGTGTGGGTGGCGTAAAGGGAGCATCGGACA**CCCGATCCC**
3C	1	**CTGGGAGGGAGGGAGG**GATCGGGTGTGGGTGGCGTAAAGGGAGCATCGGACA**CCCGATCCC**
2I	1	GATCGGGTGTGGGTGGCGTAAAGGGAGCATCGGACA**AAAAAAAAA**
3D	1	GATCGGGTGTGGGTGGCGTAAAGGGAGCATCGGACA**AAAAAAAAAAAAAAAAAA**
2J	1	**TGGGTAGGGCGGGTTGGGAAA**GATCGGGTGTGGGTGGCGTAAAGGGAGCATCGGACA
2K	1	**AAA**GATCGGGTGTGGGTGGCGTAAAGGGAGCATCGGACA
3E	2	**AAAAAA**GATCGGGTGTGGGTGGCGTAAAGGGAGCATCGGACA
3F	2	**AAAAAAAAAAAA**GATCGGGTGTGGGTGGCGTAAAGGGAGCATCGGACA
3G	1	**AAAAAAAAAAAAAAAAAA**GATCGGGTGTGGGTGGCGTAAAGGGAGCATCGGACA
2L	1	GATCGGGTGTGGGTGGCGTAAAGGGAGCATCGGACA**CGCCACCCACAC**
3H	1	GATCGGGTGTGGGTGGCGTAAAGGGAGCATCGGACA**CGCCACCCACACA**
2M	1	**TGAG**GATCGGGTGTGGGTGGCGTAAAGGGAGCATCGGACA
2N	1	CTGGGAGGGAGGGAGG~~GATCGGGTGTGGGTGGCGTAAAGGGAGCATCGGACA~~CCCGATCCC
1B	3	TGGTGGCTGTAGGTCAGCATCTGATCGGGTGTGGGTGGCGTAAAGGGAGCATCGGACAACG
1C	3	GATCGGGTGTGGGTGGCGTAAAGGGAGCATC
1D	7	GATCGGGTGTGGGTGGCGTAAAGGGAGCATCGG
2P	1	GATCGGGTGTGGGTGGCGTAAAGG~~G~~AGCATCGG
2Q	1	**TTTTTTTTTT**GATCGGGTGTGGGTGGCGTAAAGGGAGCATCGG
1E	1	~~TGGTGGCTGTAGGTCAG~~GGGGTGAAACGGGTCCCG~~CGTTGTCCGATGCTC~~
0	8	No sequence provided
Total	174	

Figure S10 | Phylogenies depicting unexplained aptamer sequence alterations introduced to the three DNA Ochratoxin A (OTA) binding aptamers: 1.12 a 61mer (node 1B), 1.12.5 a 36mer (1A), and 1.12.8 33mer (1D), Cruz-Aguado et al., 2008.

A. The nodes in the artificial phylogenetic trees represent aptamers grouped according to sequence homology and alterations. B. Unexplained aptamer sequence alterations were categorized. In the table, unexplained insertions and additions are bolded, unexplained substitutions are bolded and underlined, unexplained deletions are struck out and justified or explained alterations are in light grey. The number (#) column indicates the number of times each sequence was reported, where multiple publications provide or use multiple clones.

Orange IDs are the clones described by Cruz-Aguado (2008), **node 1A** was 1.12.5 (36mer), **node 1B** was 1.12 (61mer), **node 1C** was 1.12.12 (31mer), **node 1D** was 1.12.8 (33mer), and **node 1E** was the random region of 1.13 with the primers deleted (notably, this is not a described or characterized altered sequence. **Node 2B** cites the Cruz-Aguado aptamer but provide the sequence of a different OTA aptamer described in 2010. **Node 2C** provided the complimentary strand to the aptamer. Several sequences identified included 5'/3' additions that were not explained in the text (**nodes 2A, 2D, 2E, 2G, 2H, 3C, 2I, 3D, 2J, 2K, 3E-G, 2L, 3H, 2M, and 2Q**). **Nodes 2F, 3A,** and **3B** contain the same 5' addition and 3' deletion**.** However, **node 3A** contains an additional 3' addition and **node 3B** contains that addition with a single nt deletion. **Nodes 2N** and **2D** include explained 5'/3' additions as well as unexplained internal deletions. Finally, **node 2P** contains a single nt deletion.

Table S4. Consolidated Aptamer Sequence Reporting Data.

ALL PHYLOGENIES					
Target Name	# publications reviewed	Number of Publications Reporting Unexlpained Altered Sequences	Number of Publications Reporting Unaltered Sequences	Number of Unique Unexlpained Altered Sequences Identified	Number of Unaltered Sequences Identified
lysozyme (Cox)	61	57	4	58	4
lysozyme (Tran)	22	16	6	16	6
nucleolin	68	23	45	23	45
ATP	69	60	9	60	9
IgE	71	22	49	22	50
theophylline	112	9	103	9	103
ochratoxin A	171	38	133	39	135
thrombin	70	9	61	9	62
VEGF	9	1	8	1	17
PDGF BB	70	59	11	61	13
cocaine	57	29	28	30	35
Total Pubs	780	323	457	328	479
Percentage in all publications		41.41%	58.59%	40.64%	59.36%
MEDIAN SPLIT DATA					
Phylogenies with Greater than 34% Total Undescribed Sequence Alterations					
lysozyme (Cox)	61	57	4	58	4
lysozyme (Tran)	22	16	6	16	6
ATP	69	60	9	60	9
PDGF BB	70	59	11	61	13
cocaine	57	29	28	30	35
Total Pubs	279	221	58	225	67
Percentage in all publications		79.21%	20.79%	77.05%	22.95%
Phylogenies with Less than or equal to 34% Undescribed Sequence Alterations					
nucleolin	68	23	45	23	45
IgE	71	22	49	22	50
theophylline	112	9	103	9	103
VEGF	9	1	8	1	17
ochratoxin A	171	38	133	39	135
thrombin	70	9	61	9	62
Total Pubs	501	102	399	103	412
Percentage in all publications		20.36%	79.64%	20.00%	80.00%

The findings for the individual phylogenies as well as composite data are reported here. In several cases, individual publications reported and used several sequences. Thus, we identify both the total number of publications that report altered sequences (left grouping of data) as well as the total number of altered sequences and unaltered sequences identified (right grouping of data). To further categorize the types of apparent *in silico* sequence alterations that are perpetuated in the literature, we used a "median split" to divide the dataset into two groups: "low"/few apparent mutations and "high"/more apparent mutations (as described in the manuscript, data under the "Median Split Data" Heading). The median percentage of unexplained sequence alterations within each phylogeny was 34%, and this value was used for the division into groups, shown above.

Table S5. Consolidated Aptamer Sequence Alteration Data.

ALL PHYLOGENIES										
Target Name	No Sequence Provided	Deletion	Insertion	Substitution	Inversion	Unexplained 5'/3' Addition	Complimentary Strand	Reported Incorrect Sequence	Core Binding Sequence Misattribution	Partial Inversion
lysozyme (Cox)	1	56	2	2	0	0	0	0	0	0
lysozyme (Tran)	1	14	0	1	0	1	0	0	0	0
nucleolin	12	0	0	0	0	11	0	0	0	0
ATP	0	2	2	0	2	54	0	0	0	0
IgE	4	4	0	3	0	14	0	0	0	0
theophylline	1	0	0	1	0	1	0	0	6	0
ochratoxin A	8	9	0	0	0	25	1	1	0	0
thrombin	0	1	0	1	0	7	0	1	0	0
VEGF	1	0	0	0	0	0	0	0	0	0
PDGF BB	5	46	1	2	0	17	0	0	0	0
cocaine	0	19	24	18	0	8	0	1	0	1
Total Number of Sequence Alteration Identified	33	151	29	28	2	138	1	3	6	1
Percentage of all errors found (i.e., x% of all alterations found were deletions)	8.42%	38.52%	7.40%	7.14%	0.51%	35.20%	0.26%	0.77%	1.53%	0.26%
MEDIAN SPLIT DATA (TOTAL)										
Phylogenies with Greater than 34% Total Undescribed Sequence Alterations										
lysozyme (Cox)	1	56	2	2	0	0	0	0	0	0
lysozyme (Tran)	1	13	0	1	0	1	0	0	0	0
ATP	0	2	2	0	2	54	0	0	0	0
PDGF BB	5	46	1	2	0	17	0	0	0	0
cocaine	0	19	24	18	0	8	0	1	0	1
Total Number of Sequence Alteration Identified	7	136	29	23	2	80	0	1	0	1
Percentage of all errors found (i.e., x% of all alterations found were deletions)	2.51%	48.75%	10.39%	8.24%	0.72%	28.67%	0.00%	0.36%	0.00%	0.36%
Phylogenies with Less than or equal to 34% Undescribed Sequence Alterations										
nucleolin	12	0	0	0	0	11	0	0	0	0
IgE	4	4	0	3	0	14	0	0	0	0
theophylline	1	0	0	1	0	1	0	0	6	0
VEGF	1	0	0	0	0	0	0	0	0	0
ochratoxin A	8	9	0	0	0	25	1	1	0	0
thrombin	0	1	0	1	0	7	0	1	0	0
Total Number of Sequence Alteration Identified	26	14	0	5	0	58	1	2	6	0
Percentage of all errors found (i.e., x% of all alterations found were deletions)	23.21%	12.50%	0.00%	4.46%	0.00%	51.79%	0.89%	1.79%	5.36%	0.00%

To better understand the types of unexplained sequence alterations reported in the literature, we categorized each altered sequence (i.e., deletion, insertion, substation, etc.). As stated in Table S4, in several cases, individual publications reported and used several sequences. Further, several sequences contained multiple types of sequence alteration (i.e., one altered sequence containing both a deletion and substitution). For this analysis, we tallied all instances of that sequence alteration type. Thus, several sequences were "counted" more than once. Notably, in two cases a known incorrect sequence (i.e., a different aptamer for a different target) was reported. We also categorized publications that did not provide a sequence in the manuscript or the supplement as containing a sequence alteration because their work is not reproducible on its own. Finally, any addition to the 5'/3' end of the sequence that were not described in the text and did not appear to serve an obvious purpose in the application, were considered unexplained 5'/3' additions. This includes single nucleotide additions, that in many cases likely served no purpose (e.g., node 2B of the PDGF-BB phylogeny, Figure S4) and are more likely to be copy-edit mistakes.
To further categorize the types of apparent *in silico* sequence alterations that are perpetuated in the literature, we used a "median split" to divide the dataset into two groups: "low"/few apparent mutations and "high"/more apparent mutations (as described in the manuscript, data under the "Median Split Data" Heading). The median percentage of unexplained sequence alterations within each phylogeny was 34%, and this value was used for the division into groups, shown above. We hoped this would help to elucidate what sequence alteration types were common in phylogenies that contained a greater proportion of mutation, and thus what mutation types are more likely to result in more frequent reporting of altered sequences.
Interestingly, there was a larger contribution of deletions (48%) and insertions (10%) in the high-level of unexplained sequence alteration group relative to the low-level of unexplained sequence alteration group, which had 12% deletions and no insertions. In contrast, the low-level unexplained sequence alteration group contained a higher contribution of unexplained 5'/3' modifications (51%) compared to the high sequence alteration group (28%) as well as a greater proportion of publications that did not report a sequence (23%) compared to the high sequence alteration group (2%). This observation generally suggests that deletions and insertions lead to subsequent reporting of sequence alterations in the literature, while 5'/3' modifications that may have unique, researcher-specific purpose do not. Further, while publications that do not report a sequence are irreproducible, they are unlikely to beget subsequent sequence alterations.

References

1. Dunn, M. R., Jimenez, R. M., & Chaput, J. C. (2017). Analysis of aptamer discovery and technology. *Nature Reviews Chemistry*, *1*(10), 1-16.
2. Cruz-Aguado, J. A., & Penner, G. (2008). Determination of ochratoxin A with a DNA aptamer. *Journal of agricultural and food chemistry*, *56*(22), 10456-10461.
3. Wiegand, T. W., Williams, P. B., Dreskin, S. C., Jouvin, M. H., Kinet, J. P., & Tasset, D. (1996). High-affinity oligonucleotide ligands to human IgE inhibit binding to Fc epsilon receptor I. *The journal of immunology*, *157*(1), 221-230.
4. Bates, P. J., Kahlon, J. B., Thomas, S. D., Trent, J. O., & Miller, D. M. (1999). Antiproliferative activity of G-rich oligonucleotides correlates with protein binding. *Journal of Biological Chemistry*, *274*(37), 26369-26377.
5. Tran, D. T., Janssen, K. P., Pollet, J., Lammertyn, E., Anné, J., Van Schepdael, A., & Lammertyn, J. (2010). Selection and characterization of DNA aptamers for egg white lysozyme. *Molecules*, *15*(3), 1127-1140.
6. Jenison, R. D., Gill, S. C., Pardi, A., & Polisky, B. (1994). High-resolution molecular discrimination by RNA. *Science*, *263*(5152), 1425-1429.
7. Zimmermann, G. R., WICK, C. L., SHIELDS, T. P., JENISON, R. D., & PARDI, A. (2000). Molecular interactions and metal binding in the theophylline-binding core of an RNA aptamer. *Rna*, *6*(5), 659-667.
8. Zimmermann, G. R., Jenison, R. D., Wick, C. L., Simorre, J. P., & Pardi, A. (1997). Interlocking structural motifs mediate molecular discrimination by a theophylline-binding RNA. *Nature structural biology*, *4*(8), 644-649.
9. Zimmermann, G. R., Shields, T. P., Jenison, R. D., Wick, C. L., & Pardi, A. (1998). A semiconserved residue inhibits complex formation by stabilizing interactions in the free state of a theophylline-binding RNA. *Biochemistry*, *37*(25), 9186-9192.
10. Kawai, R., Kimoto, M., Ikeda, S., Mitsui, T., Endo, M., Yokoyama, S., & Hirao, I. (2005). Site-specific fluorescent labeling of RNA molecules by specific transcription using unnatural base pairs. *Journal of the American Chemical Society*, *127*(49), 17286-17295.
11. Stojanovic, M. N., De Prada, P., & Landry, D. W. (2001). Aptamer-based folding fluorescent sensor for cocaine. *Journal of the American Chemical Society*, *123*(21), 4928-4931.
12. Green, L. S., Jellinek, D., Jenison, R., Östman, A., Heldin, C. H., & Janjic, N. (1996). Inhibitory DNA ligands to platelet-derived growth factor B-chain. *Biochemistry*, *35*(45), 14413-14424.
13. Jellinek, D., Green, L. S., Bell, C., & Janjic, N. (1994). Inhibition of receptor binding by high-affinity RNA ligands to vascular endothelial growth factor. *Biochemistry*, *33*(34), 10450-10456.
14. Huizenga, D. E., & Szostak, J. W. (1995). A DNA aptamer that binds adenosine and ATP. *Biochemistry*, *34*(2), 656-665.
15. Bock, L. C., Griffin, L. C., Latham, J. A., Vermaas, E. H., & Toole, J. J. (1992). Selection of single-stranded DNA molecules that bind and inhibit human thrombin. *nature*, *355*(6360), 564-566.