# *In silico* selection approach to develop DNA aptamers for a stem-like cell subpopulation of non-small lung cancer adenocarcinoma cell line A549

Mateja Vidic, Tina Smuc, Nika Janez, Michael Blank, Tomaz Accetto, Jan Mavri, Isis C. Nascimento, Arthur A. Nery, Henning Ulrich, Tamara T. Lah

doi: 10.2478/raon-2018-0014

## Pool amplification by PCR

**TABLE S1.** PCR conditions for constructing cell-SELEX starting library

| REAGENT | Volume |
|---|---|
| 10X buffer | 10 µl |
| $MgCl_2$ (stock 25 mM) | 10 µl |
| dNTP mix (each 10 mM) | 2 µl |
| Primer forward (100 µM) | 1 µl |
| Primer reverse (100 µM) | 1 µl |
| TaqPol (5U/ul) | 1 µl |
| betaine (stock solution=5M) | 20 µl |
| DMSO (stock solution=100%) | 5 µl |
| $ddH_2O$ | 50 µl |
| DNA template | 1 µl |

DNA was amplified under following conditions: 95°C 3 min, 95°C 45 s, 42°C 45 s, 72°C 2 min

## Restriction analysis (RFLP)

We monitored the gradual enrichment of aptamers during the selection process by flow cytometry. After seven rounds of selection, the binding ability of selected DNA pool reached a plateau, and cloning was performed to isolate individual aptamers. To monitor the evolution of the aptamer pool, the appearance of restriction sites in the population was analysed by *restriction fragment length polymorphism* (RFLP). This revealed the emergence of distinct families in the library. The PCR product of each cycle was digested with a mix of four restriction enzymes Alu I, Rsa I, Hae III and Hinf I. Following ethanol precipitation, the digested samples were loaded onto a denaturing polyacrylamide gel that was showing a reduced pattern of bands following 6 cycles of selection, suggesting that the random pool has evolved into a homogenous fraction of aptamers.

**TABLE S2.** Selected oligonucleotide candidates. The table lists the twenty most frequent oligonucleotide sequences obtained after next generation sequencing of last SELEX. In the first column are listed their names with where the first number represents number of repetitions in THc last cycle and second number frequency rating. The third column represents corresponding 5'-3' sequences (showing only the randomized region)

| NAME | NUMBER OF REPETITIONS | SEQUENCE (5'-3') |
|---|---|---|
| A786_1 | 786 | GCAGCCATTAGTTATTGTTTATTTTGATATTATA |
| A574_2 | 574 | ACACACTTTGAACTTATAACTCTATTCTCTATA |
| A452_3 | 452 | GCACTTTGGTTATTATTAACATATTATTTATATA |
| A373_4 | 373 | GTTGTTATGTACATTGTTCTACTACTTTCTATTA |
| A316_5 | 316 | GTACATTCTATCCACTACTTATTCTATGTATCTA |
| A290_6 | 290 | GCATCATTTCTACTTATTTTCTTTTGTATTGTTA |
| A288_7 | 288 | GTCGAAGTCGTTTTACCCATGTGCTATAGCCTTGAG |
| A278_8 | 278 | GTGTGTACTTATCCTATTTTGTTTACCTATATTA |
| A266_9 | 266 | GTACGGTATATCAACTTTCTATTCTTTTTACTAT |
| A254_10 | 254 | CCCTCATGTGTTGTTTTCGTATTTAAATATTATA |
| A245_11 | 245 | CCACCCCGACTCAAGGGAACCGGTCGGCCTGCCC |
| A218_12 | 218 | GGCTATGTTTGTACTGTTTATACAATATTCTATA |
| A217_13 | 217 | GTAGCAACCACATTCAATTATCTACTTTATATTA |
| A190_14 | 190 | CCACCCCGACTCAAGGGAACCGGTCGGCCTGCGC |
| A185_15 | 185 | GCAGCCCGAGCCCATCGCTCCAGGGCGGTGCCCC |
| A176_16 | 176 | GTATACATTCTACCTATATTGCACTATTCTTATT |
| A166_17 | 166 | GGTTGTTTGAGAAGTATTCCTTTTTTATTCTATA |
| A155_18 | 155 | GCCAGTTCCGGAACCCTGCGTCGGGTGCGTCCTC |
| A147_19 | 147 | GTTTAGGACTATGGTTTATTACTTTGTTTCTATA |
| A138_20 | 138 | GTACGGTATATTAACTTTCTATTCTTTTTACTAT |

## Sequencing

Sanger sequencing. The last three cycles were used for sequencing using the DNA pJET 1.2 plasmid (Fermentas Kit #K1213; Thermo Fisher Scientific Inc., Waltham, Massachusetts, USA) and *E. coli* cells competent for OneShot TOP10. 64 colonies from each pool were collected and sequenced with Next Generation approach (Ion Torrent PGM, ThermoFisher, Waltham, MA, USA).

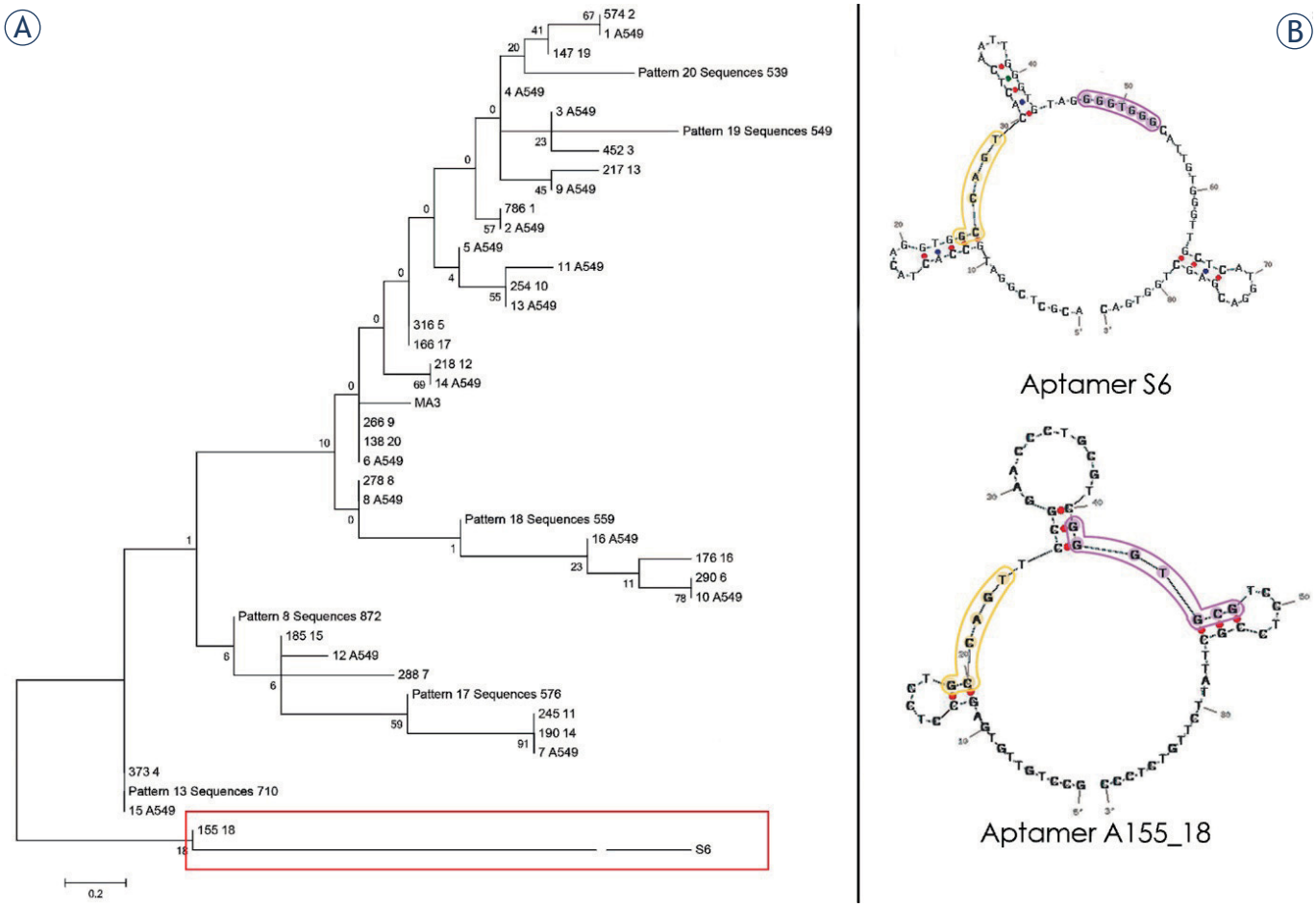


**FIGURE S1.** Dendrogram of the relationship among aptamer sequences. **(A)** Dendrogram representing evolutionary relations among selected sequences. The three was constructed with MEGA6 program and it is showing a close relation of sequences aptamers A155_18 and S6 and is presented by sharing branch. The three constructing method used maximum likelihood statistical method with 500 Bootstrap replications and Neighbour joining tree inference. **(B)** Comparison of secondary structures of A155_18 and S6. Selected structures had predicted lowest ΔG indicating their highest stability. Sharing motifs are circled in violet (GGTGG/CG) and in yellow (GCCAGT). Those motifs are pointing out the similarity between sequences and also the importance of motifs for specific binding to target cell line.

Next generation sequencing. After the preparation of the template and sequencing on Ion Torrent PGM Sequencer with Ion 316 Chip Kit v2, 2631 unique sequences were retrieved and analysed. They were grouped based on the barcode adapters, and primers were removed.

## Bioinformatics analysis

With CLC Genomics Workbench software for further processing raw data were quality-filtered by length and DNA orientation. The sense DNA sequence reads that were 34 ±1 bases long followed by a trimming of library primers resulting in 135 576 high quality reads. 72.1% reads were repeated only once and were therefore abandoned from the analysis. Number of reads with at least one copy (37 821 reads) were further narrowed down to 10% (3 782 reads). The unique sequences (2,631 reads) were aligned against the starting pool of sequences (135 576 reads) and were clustered based on 95% sequence homology.

In the second *in silico* selection procedure reads we used shell script to quality-filter by base composition followed by selecting reads having only the exact library primer sequence; all others were abandoned. The remaining reads were sorted according to copy number. The pipelined use of different programs for processing nucleotide sequences - MEME[17], MEGA6[18] and UnaFold[19] resulted in 20 oligonucleotide candidates (Table S2).

Further criteria for selection were applied: Number of repetitions, structure stability, robustness and loop maintenance.

The third *in silico* aptamer selection procedure was based on the COMPAS software (AptaIT GbmH, Planneg, Germany) with integrated bioinformatics for clustering sequences regarding the pattern in loop space region and the covariance of frequent motifs.

For construction of dendrogram (Figure 1A), the MUSCLE method[1] was used for alignment and Model with the lowest BIC (Bayesian Information Criterion) scores was computed. For our data the lowest BIC has Kimura-2 Model therefore considered to describe the substitution pattern of our sequences the best. To test the phylogeny the Bootstrap method with 500 replications and for Gap Data treatment Partial deletions used respectively. Comparison of secondary structures of sequences A155_18 and S6 with predicted lowest $\Delta G$ (Figure 1B) has shown sharing motifs.