# Differential binding cell-SELEX: method for identification of cell specific aptamers using high throughput sequencing

Karlis Pleiko[1]*, Liga Saulite[1], Vadims Parfejevs[1], Karlis Miculis[2], Egils Vjaters[2] and Una Riekstina[1]

[1] Faculty of Medicine, University of Latvia, Riga, LV-1004, Latvia
[2] Pauls Stradins Clinical University Hospital, Riga, LV-1002, Latvia
* To whom correspondence should be addressed. Tel: +371 28680049; Email: karlis.pleiko@lu.lv

Present Address: Karlis Pleiko, Faculty of Medicine, University of Latvia, Riga, LV-1004, Latvia

## ABSTRACT

Aptamers have evolved as a viable alternative to antibodies in recent years. High throughput sequencing (HTS) has revolutionized the aptamer research by increasing the number of reads from few using Sanger sequencing to millions of reads using HTS approach. Despite the availability and advantages of HTS compared to Sanger sequencing there are only 50 aptamer HTS sequencing samples available on public databases. HTS data for aptamer research are mostly used to compare sequence enrichment between subsequent selection cycles. This approach does not take full advantage of HTS because enrichment of sequences during selection can be due to inefficient negative selection when using live cells. Here we present differential binding cell-SELEX (systematic evolution of ligands by exponential enrichment) workflow that adapts *FASTAptamer* toolbox and bioinformatics tool *edgeR* that is mainly used for functional genomics to achieve more informative metrics about the selection process. We propose fast and practical high throughput aptamer identification method to be used with cell-SELEX technique to increase successful aptamer selection rate against whole cells. The feasibility of our approach is demonstrated by performing aptamer selection against clear cell renal cell carcinoma (ccRCC) RCC-MF cell line using RC-124 cell line from healthy kidney tissue for negative selection.

## INTRODUCTION

Aptamers are short (20 - 100 nt) oligonucleotides that, contrary to the most of other functionals nucleic acids, bind specific molecular targets owing to their folded three-dimensional (3D) structures (1). Most of the aptamers are developed for therapeutic or diagnostic purposes (2, 3). Currently several aptamer candidates are being tested in clinical trials for conditions like age-related macular degeneration (4), Duchenne muscular dystrophy (5), chronic lymphocytic leukemia (6) and others (7).

After initial description of aptamer selection method termed SELEX (systematic evolution of ligands by exponential enrichment) (8), several aptamer selection methods have been developed, among others cell-SELEX (9), where whole live cells are used. First high throughput SELEX (HT-SELEX) experiment, a variation of SELEX process that uses high throughput sequencing (HTS) methods instead of Sanger sequencing, was described by Zhao et.al in 2009 (10). Consequently adaptation of high throughput sequencing (HTS) methods for aptamer research further improved outcomes of selection procedures (11, 12).

Further on, RNA aptamer selection against active and inactive conformation of $\beta_2$ adrenoreceptor described by Kahsai et.al. employs HTS methods to characterize the fold change enrichment of particular sequences during the selection against each individual target in parallel (11). However, in case of cell-SELEX this approach might be of very limited use due to the high diversity of protein targets on cell surface that would cause enrichment of non-specifically bound sequences if no negative selection is performed.

Several research teams have developed tools to analyse HTS data from aptamer selection, notably *FASTAptamer,* a toolkit developed by Alam et.al. that can be used to track the evolutionary trajectory during the SELEX process of individual oligonucleotide sequences (12). Just recently *AptaSUITE*, a comprehensive bioinformatics framework that includes most of the previously published functionalities of different tools – data pre-processing, sequence clustering, motif identification and mutation analysis has been introduced (13).

RNA-sequencing (RNA-seq) experiments are used to quantify the differential expression of gene transcripts between samples (14). We speculated that it might be possible to adapt data analysis tools currently used for RNA-seq to be used with HT cell-SELEX experiments. During the cell-SELEX experiment, the goal to be achieved is to isolate aptamers that bind to the target cells in larger number compared to the control cells, making experimental design similar to RNA-seq analysis. Here we provide a differential binding cell-SELEX method that can be used to identify differentially abundant aptamers on the surface of target cells and negative control cells during the cell-SELEX experiments and calculate the statistical significance of these differences. Analysis includes the use of *edgeR* (15), a common tool for the analysis of RNA-seq experiments that uses negative binominal distribution to identify differentially expressed genes, *FASTAaptamer* (12) toolbox to estimate the read count, *cutadapt* (16) to remove the constant primer binding regions of aptamers and bespoke R script available for reuse. Moreover, we combine our approach with sequence enrichment analysis already used by other groups for aptamer selection to identify the most relevant sequences.

## MATERIAL AND METHODS

### Cell culturing and buffer solutions

Kidney epithelial cell line RC-124 (Cell Lines Service GmbH) established from non-tumor tissue of kidney and carbonic anhydrase 9 (CA9) positive ccRCC cell line RCC-MF (Cell Lines Service GmbH), established from the renal clear cell carcinoma pT2, N1, Mx/ GII-III (lung-metastasis) were used for cell-SELEX process as negative control and target cells accordingly. RCC-MF cells were cultured in RPMI 1640 (Gibco), RC-124 cells were cultured in McCoy's 5A medium (SigmaAldrich). Both culture media were supplemented with 10% fetal bovine serum (FBS) (Gibco), 50 U/ml penicillin and 50 µg/ml streptomycin (Gibco). Cells were propagated at $37^{o}$C, 5% $CO_2$ and 95% relative humidity.

Washing buffer was prepared by adding 0.225 g D-glucose and 0.25 ml of 1 M $MgCl_2$ to 50 ml of phosphate buffered saline (PBS) (SigmaAldrich) and filtering it through a 0.22 µM syringe filter (Corning). Binding buffer was prepared by adding 50 mg of bovine serum albumin (SigmaAldrich) and 5 mg baker's yeast tRNA (SigmaAldrich) to 50 ml of washing buffer and filtering it through a 0.22 µM syringe filter.

## Oligonucleotide library

Randomised oligonucleotide library with 40 nt and 18 nt constant primer binding regions on both sides of randomized regions (5'-ATCCAGAGTGACGCAGCA-N40-TGGACACGGTGGCTTAGT-3') was adapted from Sefah et al. (17). FAM label was attached on one primer (5′-FAM-ATCCAGAGTGACGCAGCA-3′) for flow cytometry monitoring and biotin was attached at the end of second primer for preparation of ssDNA after each cell-SELEX cycle (5′-biotin-ACTAAGCCACCGTGTCCA-3′). Oligonucleotides were ordered from Metabion or Invitrogen.

## Cell-SELEX procedure

Cell-SELEX protocol was adapted from Sefah et.al (17). Aptamer library was prepared in binding buffer at 14 µM concentration for the first selection cycle, heated at 95 $^{o}$C for 5 min and folded on ice for at least 15 min, added to fully confluent RCC-MF cells in 100 mm Petri plate (Sarstedt*)* that were washed 2 times with washing buffer before the addition of library. Initial library was applied to RCC-MF cells and incubated for 1 hour on ice with RCC-MF cells, but not with RC-124 cell in the first selection cycle. After incubation with the oligonucleotide library, cells were washed with 3 ml of washing buffer for 3 min and collected with cell scraper after adding 1 ml of DNase free water. DNase free water was used for the collection of sequences only for the first cycle, in subsequent cycles binding buffer was used to retrieve the bound sequences. After collection retrieved sequences were heated at 95 $^{o}$C for 10 min, centrifuged at 13 000 g and supernatant containing selected aptamer sequences was collected.

In subsequent selection cycles aptamer library was prepared at 500 nM concentration and incubated with negative selection cell line RC-124 beforehand. Solution containing unbound sequences was collected and applied to RCC-MF cell line after washing cells as described previously. With increasing selection cycle number several modifications were made to the selection procedure – after 4th selection cycle 60 mm plates were used instead of 100 mm plates, increasing concentration of FBS (10-20%) were added to library after folding without changing the final concentration of aptamer library, wash volume was increased to 5 ml, wash time was increased to 5 min and the number of wash times was increased to 3 after incubation.

## PCR optimization

After each selection cycle PCR optimization was performed to determine the optimal number of PCR cycles. For PCR optimization and preparative PCR cycling conditions were 12 min initial activation at

95 $^oC$, followed by repeated denaturation 30 sec at 95 $^oC$, annealing at 56.3 $^oC$ and elongation at 72 $^oC$.

### ssDNA preparation

After preparative PCR, ssDNA was acquired using agarose-streptavidin (GE Healthcare) binding to biotin labelled strand and FAM labelled ssDNA was eluted with 0.2 M NaOH (SigmaAldrich). Desalting was done using NAP5 gravity flow columns (GE Healthcare), concentration was determined measuring UV absorbance (NanoQuant Plate, M200 Pro, Tecan), and samples were concentrated using vacuum centrifugation (Eppendorf).

### Monitoring of aptamer binding by flow cytometry

Enriched aptamer pool, randomized starting library and selected lead aptamers were prepared in binding buffer at 1 µM concentrations, heated at 95 $^oC$ for 5 min and then put on ice for at least 15 min. RC-124 and RCC-MF cells were washed with PBS two times and dissociated using Versene solution (Gibco). Then 50 µL of enriched aptamer library, starting library, lead aptamers or binding buffer were added to 50 µL of cell suspension ($2.5 * 10^5$ cell per sample), followed by addition of 11 µL of FBS to each sample at final concentration 225 nM. Samples were incubated for 35 min on ice. After incubation, samples were washed two times with 500 µL of binding buffer and resuspended in 500 µL of binding buffer. Samples were passed through 40 µM cell strainer before flow cytometry analysis. Flow cytometry data were acquired using a Guava EasyCyte 8HT flow cytometer and analysed using the ExpressPro software (Merck Millipore). Flow cytometry data were analysed using FlowJo software, version 10 (FlowJo). 10’000 gated events were acquired for each sample.

### Differential binding

Aptamer pools after 4th and 11th selection cycle were prepared in binding buffer, heated and folded as described for cell-SELEX procedure at 1 ml volume with final concentration 500 nM. 500 µL were added to both RC-124 cells and RCC-MF cells grown on 60 mm plate in appropriate cell culture media up to 95% confluence. Aptamer pools were added to RC-124 and RCC-MF cells and incubated for 30 min on ice, then cells were washed two times and collected using cell scraper, heated immediately at 95$^oC$ for 10 min, centrifuged for 5 min at 13 000 g. Supernatants containing bound sequences from both cell lines were frozen at -20$^oC$. Sequencing was done to compare the differential binding profiles of enriched oligonucleotide libraries obtained from both cell lines.

### Sequencing

Samples for sequencing were prepared by performing two subsequent overlap PCRs as described in 16S metagenomic sequencing library preparation protocol (18). 1st overlap PCR used primers (5'-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG-ATCCAGAGTGACGCAGCA-3' and 5'-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG-ACTAAGCCACCGTGTCCA-3') that are complementary to constant regions of randomized oligonucleotide library with added overhang that includes Illumina platform specific sequence. Conditions for 1st overlap PCR was 12 min of initial activation, followed by 30 sec at 95 °C, 30 sec at 56.3 °C and 3 min at 72 °C. Cycle number was optimized for each sample to reduce the non-specific amplification. Afterwards, PCR products from one sample were pooled together, concentrated using DNA Clean & Concentrator (Zymo Research), run on 3% agarose gel at 110 V for 40 min and the band at 143 bp were cut out and purified using Zymoclean Gel DNA Recovery kit (Zymo Research).

2nd overlap PCR used primers that were partly complementary to previously added overhang and contained adapters to attach oligonucleotides to flow cell and i5 and i7 indexes (5'-CAAGCAGAAGACGGCATACGAGAT-[i7 index]-GTCTCGTGGGCTCGG-3' and 5'-AATGATACGGCGACCACCGAGATCTACAC-[i5 index]-TCGTCGGCAGCGTC-3'). Conditions for 2nd overhang PCR were 12 min at 95 °C, followed by 5 cycles of denaturation at 98 °C for 10 sec, annealing at 63 °C for 30 sec and elongation at 72 °C for 3 min. After PCR products from one sample were pooled together, concentrated using DNA Clean & Concentrator (Zymo Research), run on 3% agarose gel at 110 V for 45 min and the band at 212 bp were cut out and purified using Zymoclean Gel DNA Recovery kit (Zymo Research). Concentration for final products were determined using NEBNext Library Quant Kit for Illumina (New England BioLabs) by qPCR.

Sequencing was done on Illumina MiSeq platform using MiSeq 150-cycle Reagent Kit v3 in single read mode for 150-cycles. 9% of PhiX was added to the run. Sequencing was done at the Estonian Genome Center, Tartu, Estonia.

### Sequencing data analysis

Sequencing reads were filtered and demultiplexed. Constant primer binding regions were removed, sequences that are longer or shorter than 40 nt were discarded using *cutadapt* (16). Counting of recurring sequences was done using *fastaptamer-count*, matching of sequences found in replicate samples was done using *fastaptamer-enrich* (19).

Differential expression analysis tool *edgeR* (15) was further used for the analysis of sequencing data. Replicate sequencing samples (n=3) from differential binding cell-SELEX experiment after 4th and 11th selection cycles were combined and sequences with low abundance (reads per million < 2 and abundant at all in less than 2 sequencing samples) were filtered out. Normalization was performed based on reads present in each library. Differential binding was estimated using *edgeR* function for identification of significantly differentially expressed genes using following parameters: $\log_2$ fold change ($\log_2$FC) value > 2, p-value < 0.0001, adjusted for multiple comparisons using Benjamini & Hochberg (20) method.

Enrichment analysis was done separately by using all reads that came from 4th pool and 11th pool RCC-MF cell binding experiments. We calculated the mean $\log_2$ value of enrichment (mean counts per million (CPM) for sequence at 11th cycle divided by mean CPM for the same sequence at 4th cycle) for each sequence and kept the sequences that had $\log_2$FC > 6 or enrichment between the 4th and 11th cycle.

After these steps, we identified the common sequences in differential binding results and sequence enrichment results to identify most likely lead aptamer sequences. (*RNotebook* used for 4th cycle differential binding analysis and 11th cycle differential binding analysis including enrichment analysis can be found on https://github.com/KarlisPleiko/apta ).

## RESULTS

### Aptamer selection

To identify ccRCC specific aptamers initial randomized oligonucleotide library was subjected to cell-SELEX for 11 selection cycles using RCC-MF cell line as a target cell line to identify ccRCC specific aptamers and RC-124 cells as a negative control cell line to reduce the nonspecific binding. Cell specific aptamer sequence enrichment monitoring was done using flow cytometry (Guava 8HT) after 4th, 8th and 11th selection cycle. After 4th and 8th selection cycle there was a slight difference between the binding of initial randomized oligonucleotide library compared to enriched libraries. After 11th selection cycle we observed binding of enriched library to more than >95% of cells. However, the observed binding was nonspecific and selected aptamer sequences were binding to both RC-124 (Fig. 1a) and RCC-MF (Fig. 1b) cell lines.

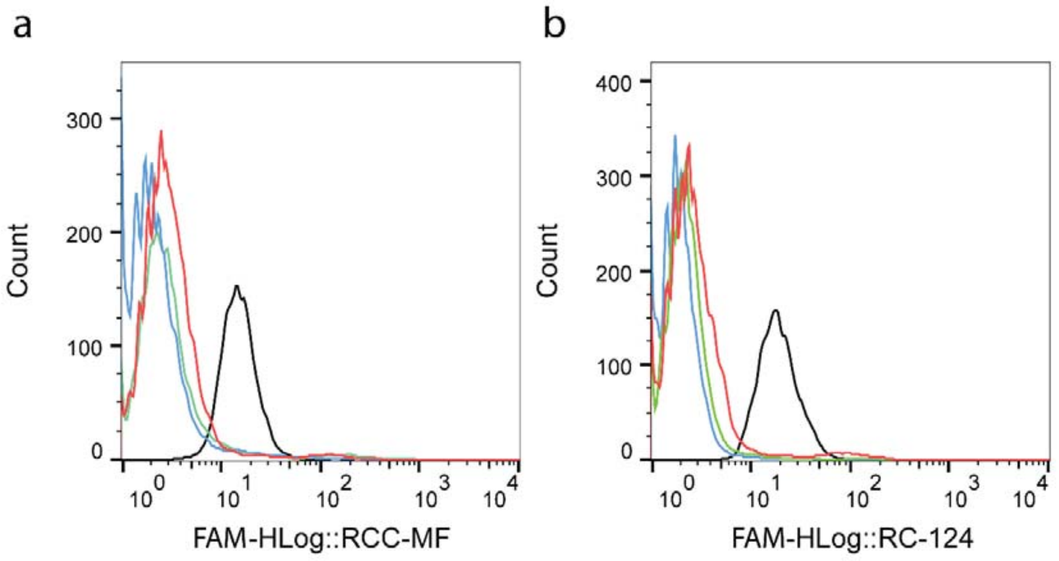


Fig. 1. Flow cytometry plots demonstrate fluorescence intensity changes of enriched libraries during the cell-SELEX procedure. Monitoring binding sequences' enrichment during cell-SELEX to negative control cells RC-124 (a) and target cells RCC-MF (b). Blue – randomized oligonucleotide library, green – 4th cycle, red – 8th cycle, black – 11th cycle.

During further selection and process optimization by changing incubation time, library concentration, FBS concentration and temperature complete specificity against RCC-MF cell line was not achieved up to 11th cycle. We calculated approximate dissociation constant ($K_d$) values of the whole enriched library after 11th selection cycles by adding increasing concentrations to RC-124 cells (Fig. 2a) and RCC-MF cells (Fig. 2b) and measuring the green fluorescence increase after the incubation on ice using flow cytometry. Based on these measurements we plotted the geometrical mean fluorescence to determine approximate $K_d$ values to control cells RC-124 (Fig. 2c) and target cells RCC-MF (Fig. 2d). $K_d$ values observed were almost identical, 189 nM for RC-124 and 169 nM for RCC-MF cells.

We concluded that complete specificity against ccRCC cells is not achieved. However, low $K_d$ value measured for enriched library after 11th pool suggested that it might be possible that the library includes also some ccRCC cell specific aptamers. To explore the differences that might exist within the library, we developed differential binding cell-SELEX approach.

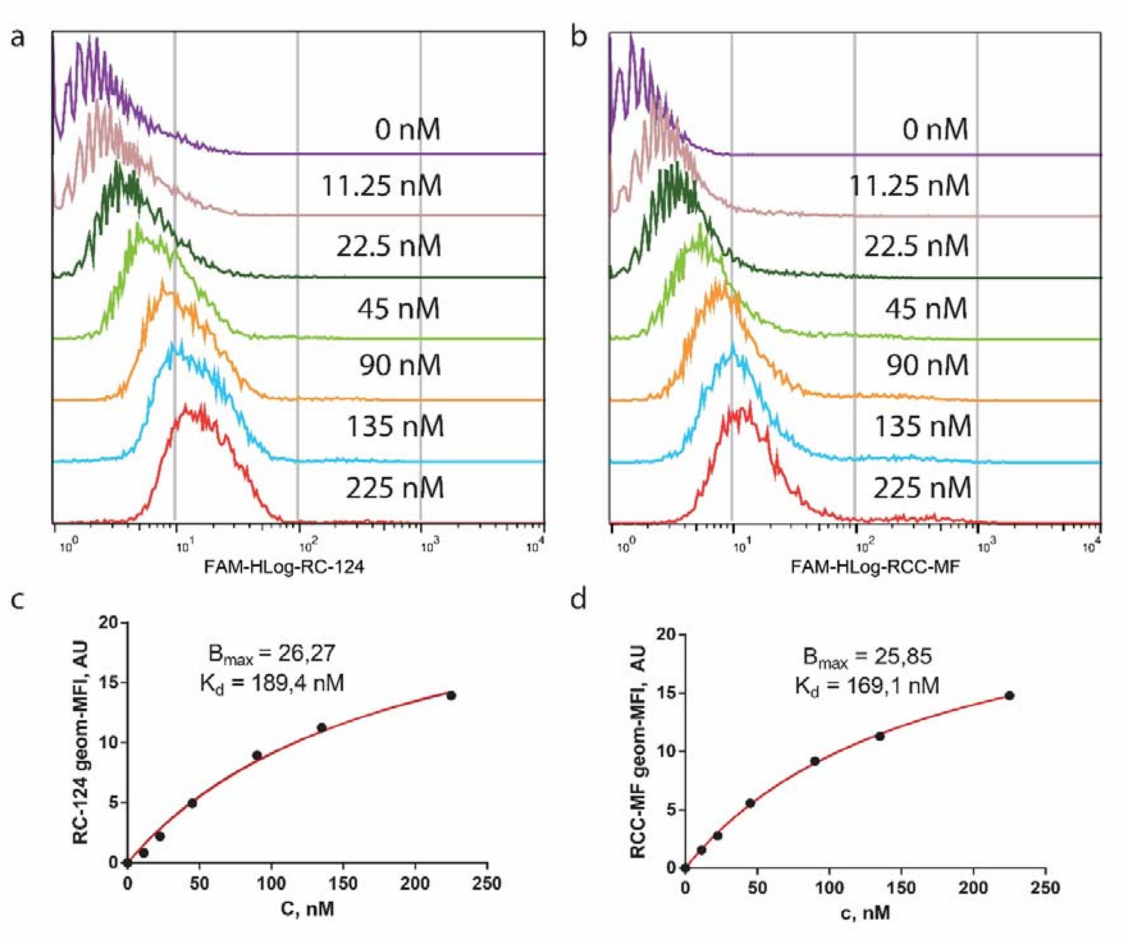


Fig. 2. Aptamer binding $K_d$ calculations after 11th selection cycle. Aptamer binding measurements by flow cytometry using 11th pool enriched library at different concentrations on control cells RC-124 (a) and clear cell carcinoma cells RCC-MF (b). $K_d$ value determination using geometrical mean fluorescence intensity during the same experiment for RC-124 cells (c) and RCC-MF cells (d).

### Differential binding cell-SELEX

Differential binding cell-SELEX process (Fig.3) was performed after 4th and 11th selection cycles. After incubation with identically split aptamer libraries and retrieval of bound sequences to both RC-124 and RCC-MF, we performed two subsequent overlap PCR reactions and confirmed that both constructs after 1st overhang PCR and 2nd overhang PCR are of expected size (Fig.4). Quantification of final libraries were done using NEBNext Library Quant Kit (New England BioLabs) to quantify only sequences that have flow cell adapters attached to them (Table 1).

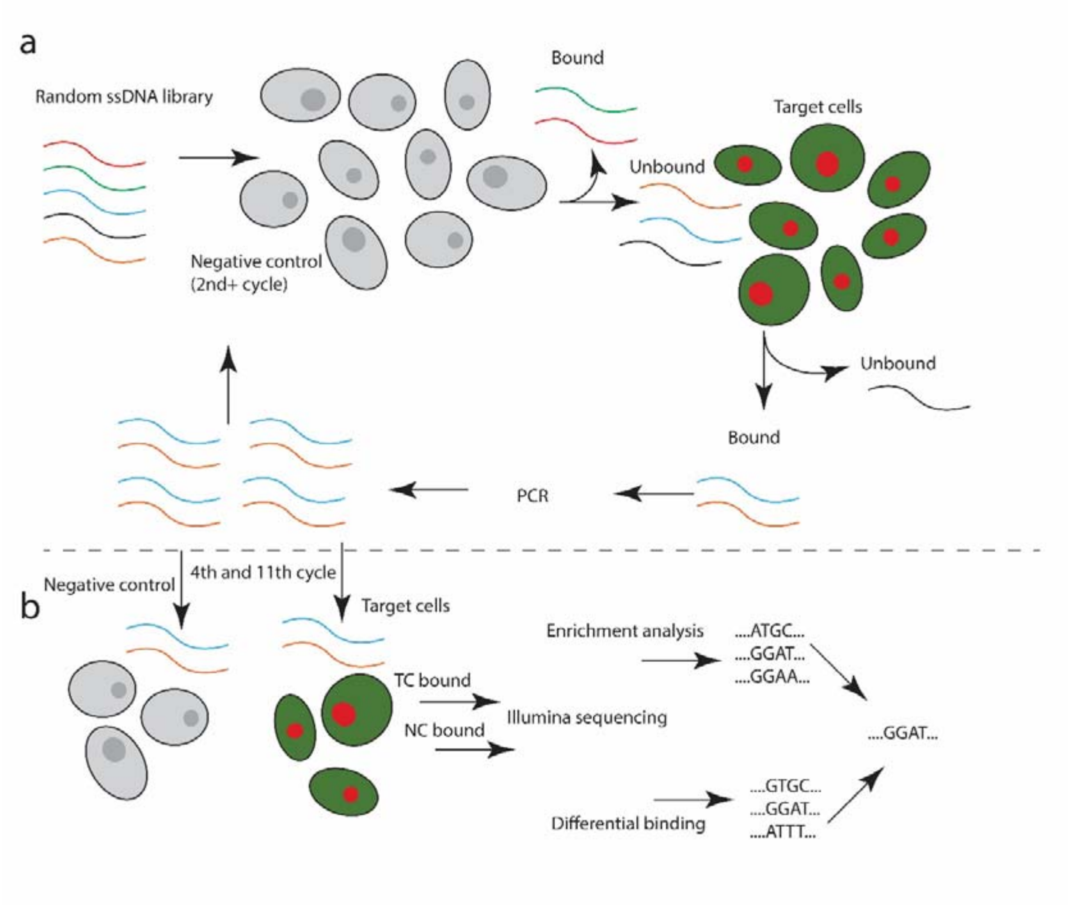

a
Random ssDNA library
Negative control
(2nd+ cycle)
Bound
Unbound
Target cells
Unbound
Bound
PCR
b
Negative control
4th and 11th cycle
Target cells
TC bound
NC bound
Illumina sequencing
Enrichment analysis
....ATGC...
....GGAT...
....GGAA...
....GGAT...
Differential binding
....GTGC...
....GGAT...
....ATTT...

Figure 3. Differential binding cell-SELEX workflow combines (a) cell-SELEX workflow with (b) additional differential binding and data analysis steps to estimate the relative number of aptamer sequences within the pool that bind to each type of cells.

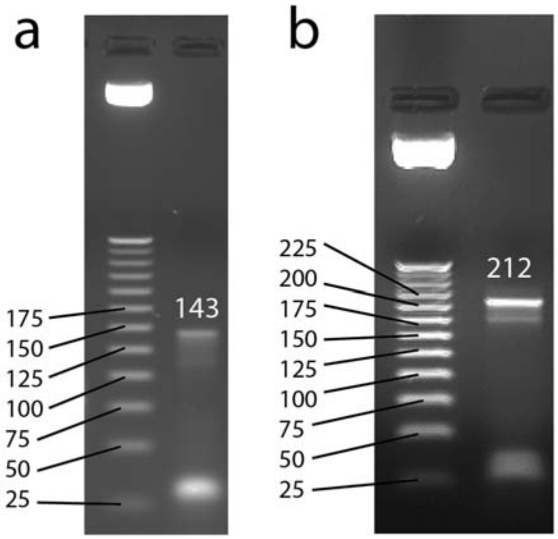


Figure 4. Gel images of aptamers after adding *Illumina* sequencing specific adapters and indexes. Aptamers after (a) 1st overhang PCR product with a length of 143 bp and (b) 2nd overhang PCR product with a length of 212 bp.

Table 1. Concentration determined by qPCR before sequencing and sequencing reads per sample for sequenced aptamer libraries.

| **Sample No** | **Samle name** | **Concentration (nM)** | **Reads** |
|---|---|---|---|
| 1 | RCC-MF P4_1 | 142.6 | 520`534 |
| 2 | RCC-MF P4_2 | 90.95 | 781`654 |
| 3 | RCC-MF P4_3 | 185.8 | 1`130`509 |
| 4 | RC-124 P4_1 | 76.23 | 169`024 |
| 5 | RC-124 P4_2 | 67.06 | 619`635 |
| 6 | RC-124 P4_3 | 82.44 | 548`781 |
| 7 | RCC-MF P11_1 | 15.22 | 277`070 |
| 8 | RCC-MF P11_2 | 11.99 | 498`937 |
| 9 | RCC-MF P11_3 | 5.23 | 326`402 |
| 10 | RC-124 P11_1 | 15.58 | 742`255 |
| 11 | RC-124 P11_2 | 28.48 | 1`142`819 |
| 12 | RC-124 P11_3 | 12.88 | 730`489 |

Sequencing data confirms the successful differential binding cell-SELEX experiments based on developed protocols.

**Data analysis for differential binding cell-SELEX**

Sequencing was done after 4th and 11th selection cycles. Reads per sample after initial quality filtration, adapter and constant primer binding region removal and length filtration (40 nt) varied from 169`024 to 1`142`856 (Table 1).

Combining all replicates from both samples after data clean-up we identified 3’627’938 unique sequences within 4th selection cycle experiment and 503’107 unique sequences in 11th selection cycle experiment. After filtering the reads by *edgeR* to remove sequences that had lower count per million (CPM) than two per sample and that were present in less than two replicates, we were left with 1’015 unique sequences for 4th cycle aptamers and 35’859 sequences for 11th cycle aptamers.

For differential binding data analysis (Fig. 5) we further used selected sequences to run *edgeR* package, a statistical analysis software that is used to estimate differential expression from RNA-seq data. Resulting data were adjusted for multiple comparisons using built-in Benjamini-Hochberg approach and filtered by removing all sequences that have $\log_2$ fold change (logFC) values less than two or adjusted p-value was higher than 0.0001.

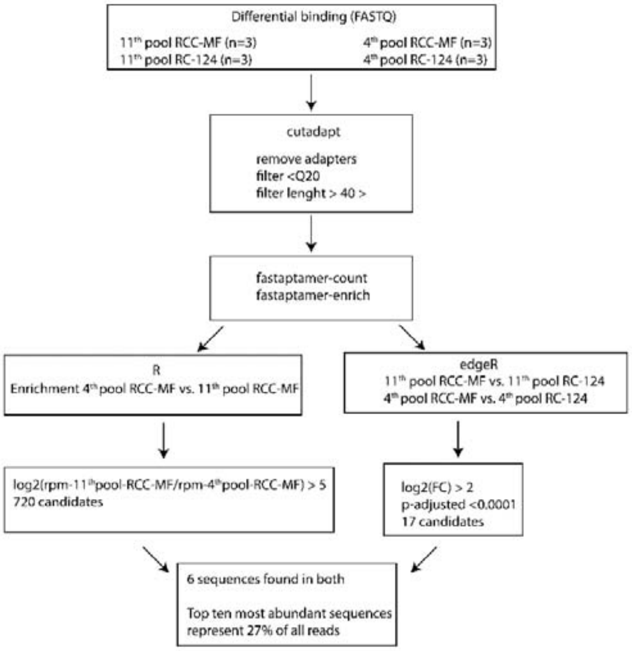

Differential binding (FASTQ)
11th pool RCC-MF (n=3)
11th pool RC-124 (n=3)
4th pool RCC-MF (n=3)
4th pool RC-124 (n=3)
cutadapt
remove adapters
filter <Q20
filter lenght > 40 >
fastaptamer-count
fastaptamer-enrich
R
Enrichment 4th pool RCC-MF vs. 11th pool RCC-MF
edgeR
11th pool RCC-MF vs. 11th pool RC-124
4th pool RCC-MF vs. 4th pool RC-124
log2(rpm-11thpool-RCC-MF/rpm-4thpool-RCC-MF) > 5
720 candidates
log2(FC) > 2
p-adjusted <0.0001
17 candidates
6 sequences found in both
Top ten most abundant sequences represent 27% of all reads

Figure 5. Data analysis pipeline for differential binding cell-SELEX data processing. After trimming using *cutadapt, FASTAptamer* tools *fastaptamer-count* and *fastaptamer-enrich* were used to count the reads for each sequence. Enrichment analysis was done using *R* and *tidyverse* package to identify 720 sequences with enrichment $\log_2 > 5$. *edgeR* was used to perform differential binding analysis

resulting in 17 candidate sequences. Matching the sequences resulted in six aptamer candidates that are represented in both analyses.

Comparing differential binding datasets using 4th selection cycle enriched library, we were unable to identify any statistically significantly differentially bound sequences based on count per million (CPM) of each sequence and fold change (FC) comparison between two cell lines (Fig. 6a). Most of the sequences bound from 4th cycle enriched library had a low abundance. However, analysis of 11th selection enriched library discovered 195 statistically significant differentially bound sequences according to the same criteria as described for the first experiment (multiple comparison adjusted p-value $< 0.0001$, $\log_2$(CPM) > abs(2)) (Fig. 6b). 178 sequences had $\log_2$(CPM) < -2 compared to 17 sequences that had $\log_2$(CPM) > 2 (Supplementary Table 1), indicating that more cell type specific sequences were identified for control RC-124 cells than for target RCC-MF cells (Fig. 6c).

Enrichment analysis identified 720 unique sequences that have $\log_2$(meanCPM@11th cycle/meanCPM@4th cycle) > 5 or the sequences is enriched in CPM terms 32 times from 4th to 11th cycle (Supplementary Table 2). We further combined differential binding results that resulted in 17 unique sequences with 720 from enrichment analysis. This identified only 6 sequences that were present in both datasets (Supplementary Table 3) as the most likely candidates to specifically target ccRCC cells (if $\log_2$ cut off values is decreased to 5, it is possible to identify 6 sequences that can be found in both differential binding analysis and in enrichment analysis results). We also ordered all unique sequences that were present in 11th pool by CPM and calculated the $\log_2$ enrichment value between 4th and 11th cycle (Supplementary Table 4). $\text{Log}_2$ enrichment values for top 10 most abundant sequences ranged from 4.7 to 6.2 and seven out of 10 sequences had value above 5. This means that these sequences are also included in enrichment analysis results. These 10 most abundant sequences contribute to approximately 27% of all sequencing reads from 11th pool. However, none of the top 10 most abundant sequences passed the statistical significance threshold or FC threshold in differential binding analysis.

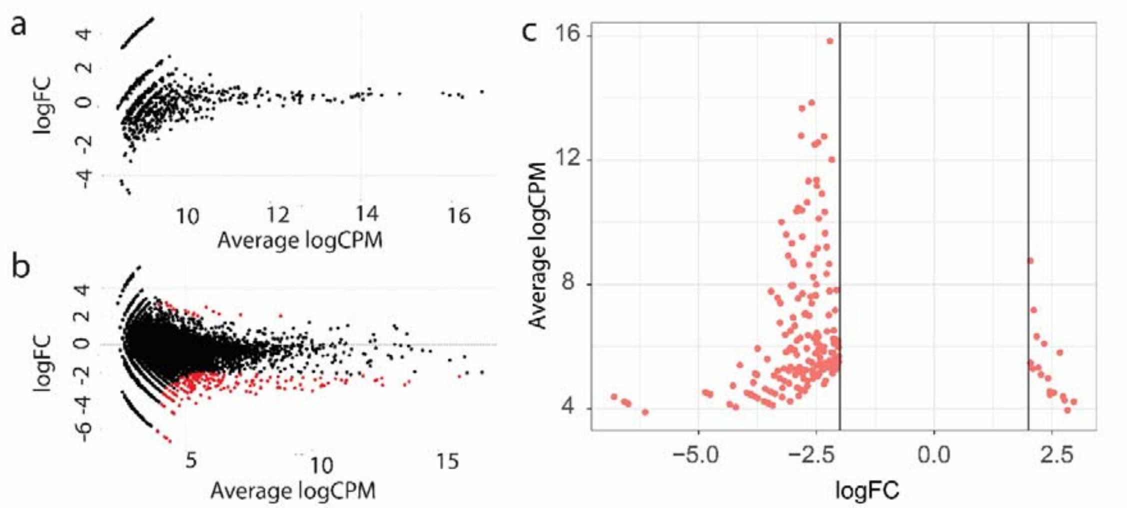


Figure 6. Differential binding cell-SELEX results at 4th cycle (a) and 11th cycle (b) of selection. Negative logFC value indicates increased binding to control cells RC-124, positive logFC value indicates increased binding to ccRCC cells RCC-MF, red dots indicate that these results are statistically significant according to adjusted p-value < 0.0001 using *edgeR* and have logFC > 2 in absolute numbers. All results that fulfil these criteria can be seen in (c).

Differential binding results confirm that it is possible to use *edgeR* within our pipeline to identify most likely candidate molecules for further testing.

**Lead aptamer testing**

For lead aptamer testing using flow cytometry we chose four sequences identified by combining differential binding cell-SELEX analysis and enrichment analysis. We also tested 15 other sequences that were identified by enrichment analysis alone. We compared the binding results as median fluorescence intensities (MFI) from each sample. To remove the background, data were corrected by subtracting MFI from sample that was incubated with randomized starting library with each cell type from MFI from samples that were incubated with lead sequences ($MFI_{lead\text{-}sequence}$-$MFI_{random\text{-}library}$). Corrected MFIs were compared using 1-way ANOVA with Sidak’s multiple comparison test to determine if our identified sequences altogether bind stronger to RCC-MF cells than to RC-124 cells (Fig. 7).

Sequence No. 14 that was identified using combined differential binding analysis with enrichment analysis showed the largest MFI difference between RC-124 and RCC-MF cells. The differences in MFI results were expressed as $\log_2$ values from $MFI_{RCC\text{-}MF}/MFI_{RC\text{-}124}$ and using this approach sequence No. 14 had the second highest $\log_2$ value of 1.66 (Supplementary Table 5).

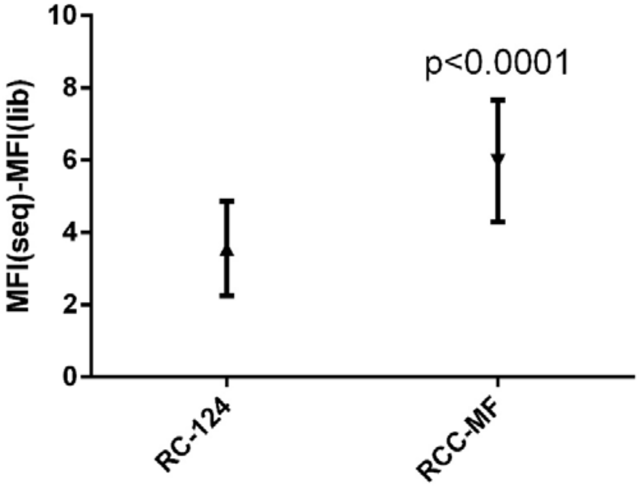


Figure 7. Comparison of median fluorescence intensities from lead sequences binding to RC-124 and RCC-MF cells identified by enrichment analysis and differential binding. Error bars represent SD, p-value determined by 1-way ANOVA followed by Sidak's multiple comparisons test.

## DISCUSSION

Recent review on aptamer discovery mentions that there are 141 entries of aptamer selection against cells as of 2017. For comparison, proteins as targets have 584 entries and small molecules have 234 research entries (1). This is not surprising considering the advanced technological procedure involved in cell-SELEX process compared to protein or small molecule SELEX. Several methods have been developed in recent years to improve the success rate of cell-SELEX, for example, HT-SELEX (10), FACS-SELEX (21) and cell-internalization SELEX (22). HTS adaptation for aptamer sequencing has been described as one of the most fundamental changes to aptamer selection technology (23).

The main goal achieved in this research is the development of differential binding cell-SELEX method. This method can identify cell type specific aptamer sequences from cell-SELEX selection pools that would not be selected by other cell-SELEX methods and thus would remain overlooked by the investigators.

Currently the most often used analysis for aptamer finding using HTS data includes enrichment analysis, which means comparison of abundance of one particular sequence at the beginning of the SELEX procedure to the abundance of the same sequence after SELEX procedure. Enrichment analysis can identify a large number of oligonucleotides with very similar $\log_2$ enrichment values as can be seen by our results (Supplementary Table 2). However, it is rarely useful for cell-SELEX because of the high possibility to enrich non-specific sequences. Using enrichment analysis with cut-off value of $\log_2 > 5$, we identified 720 sequences to be further tested. However, when the same sequencing dataset was submitted for differential binding analysis using *edgeR,* we identified 17 sequences that were more abundant on the surface of RCC-MF cells than on RC-124 cells.

Remarkably, seven out of 10 most abundant aptamer sequences after cell-SELEX process were enriched above the set cut-off value $\log_2 > 5$. However, none of them appeared in differential binding results because they did not pass the statistical significance test. These observations are in line with previous statements that the most abundant aptamer sequences are not necessary the best binders (24). This proves the value of differential binding approach for excluding the non-specifically enriched sequences during cell-SELEX procedure.

Differential binding cell-SELEX uses *edgeR* to compare how all sequences that can be found in final enriched aptamer library interact with control and target cells and estimate the statistical significance of these differences. There are several bioinformatics tools available to analyse statistical significance of differential expression for RNA-seq data (15, 25, 26). To our knowledge, so far none of these tools have been applied for estimation of differentially bound aptamers on the cell surface. *edgeR* was chosen because it is compatible with the existing data analysis workflows in *R* (27). Combination of enrichment analysis and differential binding approach provides an algorithm to choose target sequences for further analysis.

Using combined analysis approach, we identified six aptamer sequences. We confirmed by flow cytometry that aptamers identified by differential binding approach had a higher selectivity towards RCC cells compared to sequences that were identified by enrichment analysis (Supplementary Table 5).

Altogether, we demonstrate a combined analysis pipeline that can be used to identify hit aptamers from low binding specificity aptamer libraries after cell-SELEX experiments. We propose fast and practical high throughput aptamer identification method to be used with cell-SELEX technique to increase successful aptamer selection rate against whole cells.

Higher number of sequencing reads during differential binding cell-SELEX could even further increase the likelihood to identify low abundance, but differentially bound sequences specific to cells of interest.

Cell-SELEX design described in this research uses commercially available human cells RCC-MF and RC-124 both as a target and negative control. We are first to use these cell lines for aptamer selection using cell-SELEX approach. However, it could be more suitable to use patient-matched primary cells isolated from tumour site and adjacent healthy kidney tissue within few passages after isolation while cells are most likely to represent the diversity found in clinical settings (28).

Differential binding cell-SELEX method developed here can be used to accelerate aptamer selection based on HTS analysis. Additional information from differential binding cell-SELEX reduces the time needed to identify aptamers. This can lead to broader use of cell-SELEX technique not only to identify aptamers against cell lines, but also against primary cells isolated from patient samples. Further research is needed to determine the usability of the developed method for identification of aptamers binding to patient-derived samples.

We conclude that differential binding cell-SELEX method can be used to characterise not only enrichment of sequences between cycles, but also differential binding of aptamer sequences to the target and control cells used for SELEX experiment. We demonstrate the feasibility of our approach by performing aptamer selection against ccRCC cell line RCC-MF using RC-124 cell line from healthy kidney tissue for negative selection.

## AVAILABILITY

*edgeR* is available as *Bioconductor* package (http://bioconductor.org/packages/edgeR/ ),

*FASTAptamer* was downloaded from github (https://github.com/FASTAptamer/FASTAptamer ),

*cutadapt* was installed using *Bioconda* (29)

RNotebooks for data analysis using *tidyverse* (30) are available here:
https://github.com/KarlisPleiko/apta

## ACCESSION NUMBERS

Sequencing data are available at SRA under accession number PRJEB28411.

## SUPPLEMENTARY DATA

Supplementary Table 1. List of 17 aptamer variable sequences identified using differential binding.

Supplementary Table 2. List of 720 aptamer variable sequences identified using enrichment analysis.

Supplementary Table 3. List of 6 aptamer variable sequences identified using combined enrichment analysis and differential binding.

Supplementary Table 4. List of 10 most abundant aptamer variable sequences.

Supplementary Table 5. List of lead aptamer sequences tested by flow cytometry.

## ACKNOWLEDGEMENT

K.P., U.R., E.V. conceived and designed the project. K.P., L.S., V.P., K.M. carried out the experiments. K.P., U.R. wrote the paper. All read and approved the final manuscript.

## FUNDING

This work was supported by University of Latvia Foundation [grant number 2182].

## CONFLICT OF INTEREST

None declared

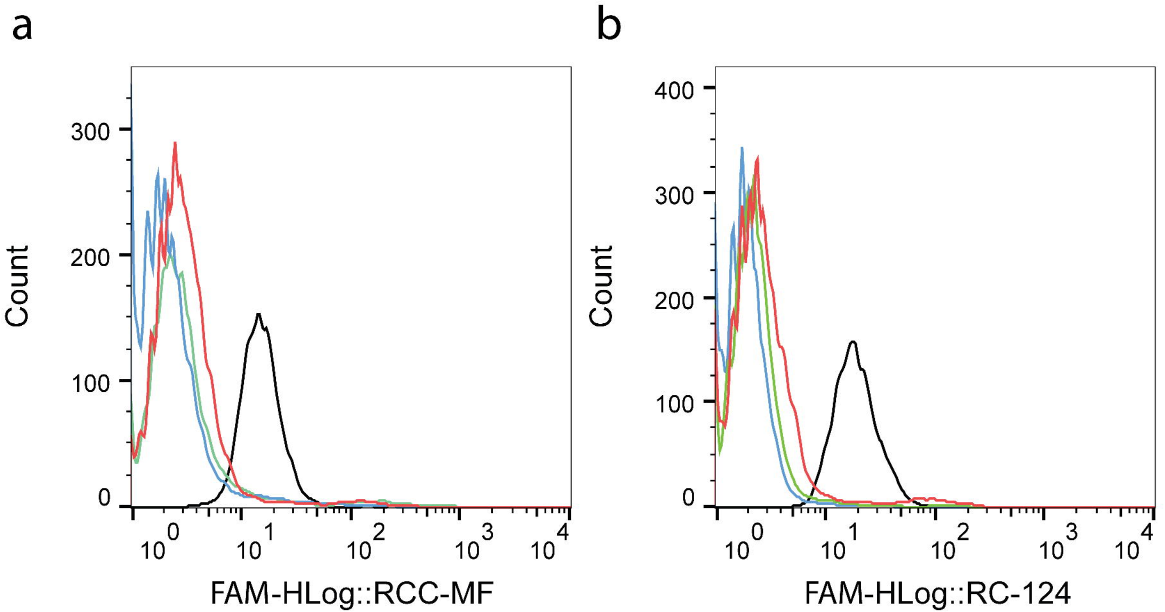
## REFERENCES

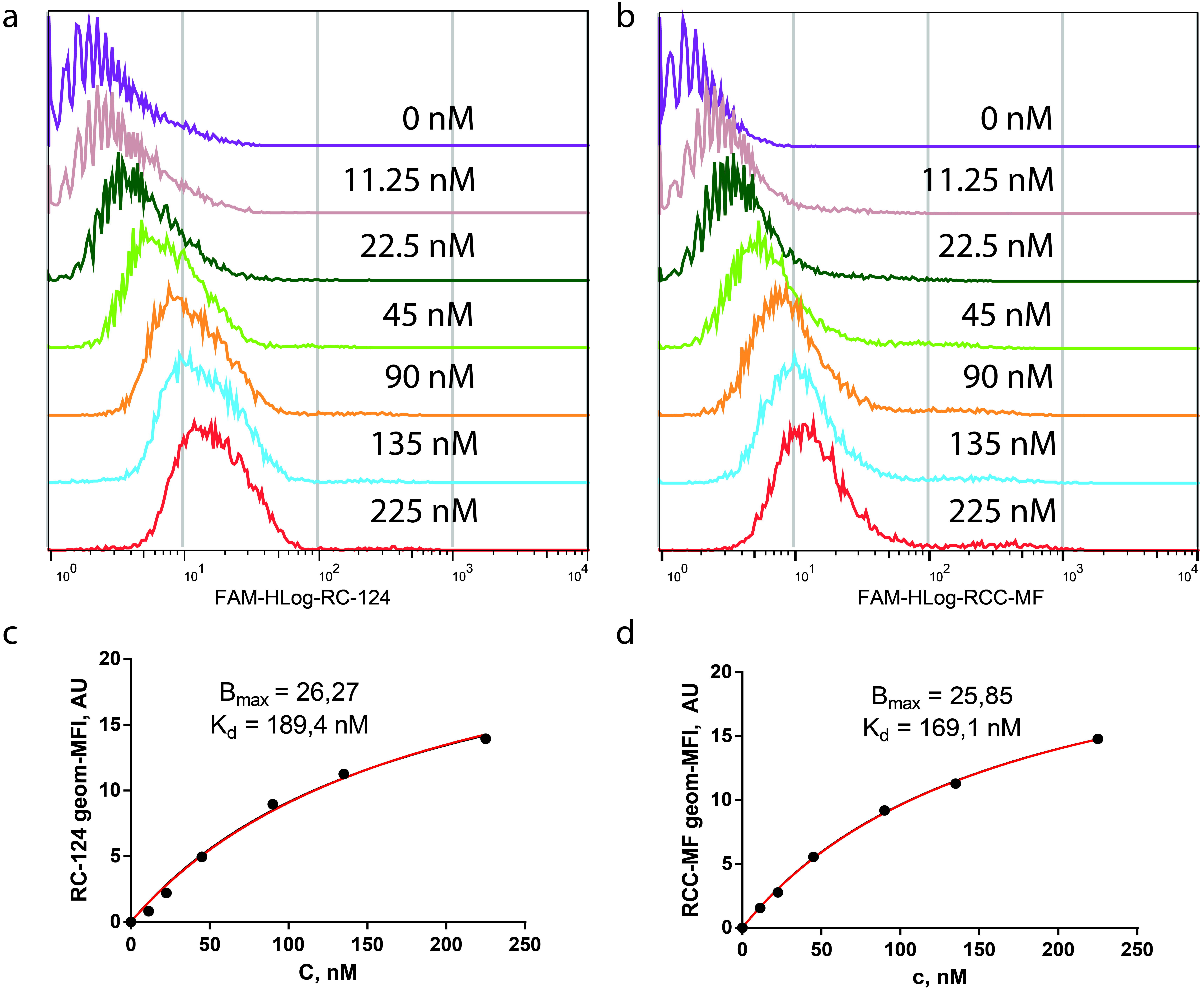
1. Dunn,M.R., Jimenez,R.M. and Chaput,J.C. (2017) Analysis of aptamer discovery and technology. *Nat. Rev. Chem.*, **1**, 0076.
2. Pereira,R.L., Nascimento,I.C., Santos,A.P., Ogusuku,I.E.Y., Lameu,C., Mayer,G. and Ulrich,H. (2018) Aptamers: novelty tools for cancer biology. *Oncotarget*, **9**, 26934–26953.
3. Zhou,J. and Rossi,J. (2017) Aptamers as targeted therapeutics: Current potential and challenges. *Nat. Rev. Drug Discov.*, **16**, 181–202.
4. A Safety and Efficacy Study of E10030 (Anti-PDGF Pegylated Aptamer) Plus Lucentis for Neovascular Age-Related Macular Degeneration (NCT01089517).
5. A Phase II Open-label, Multicenter Extension Study to Assess the Long-term Safety and Efficacy of Vamorolone in Boys with Duchenne Muscular Dystrophy (DMD)(EudraCT No: 2016-004263-38).
6. NOX-A12 in Combination With Bendamustine and Rituximab in Relapsed Chronic Lymphocytic Leukemia (CLL)(NCT01486797).
7. Kaur,H., Bruno,J.G., Kumar,A. and Sharma,T.K. (2018) Aptamers in the Therapeutics and Diagnostics Pipelines. *Theranostics*, **8**, 4016–4032.
8. Ellington,A.D. and Szostak,J.W. (1990) In vitro selection of RNA molecules that bind specific ligands. *Nature*, **346**, 818.
9. Hicke,B.J., Marion,C., Chang,Y.F., Gould,T., Lynott,C.K., Parma,D., Schmidt,P.G. and Warren,S. (2001) Tenascin-C Aptamers Are Generated Using Tumor Cells and Purified Protein. *J. Biol. Chem.*, **276**, 48644–48654.
10. Zhao,Y., Granas,D. and Stormo,G.D. (2009) Inferring binding energies from selected binding sites. *PLoS Comput. Biol.*, **5**.
11. Kahsai,A.W., Wisler,J.W., Lee,J., Ahn,S., Cahill,T.J., Dennison,S.M., Staus,D.P., Thomsen,A.R.B., Anasti,K.M., Pani,B., *et al.* (2016) Conformationally selective RNA aptamers allosterically modulate the β 2-Adrenoceptor. *Nat. Chem. Biol.*, **12**, 709–716.
12. Alam,K.K., Chang,J.L. and Burke,D.H. (2015) FASTAptamer: A bioinformatic toolkit for high-throughput sequence analysis of combinatorial selections. *Mol. Ther. - Nucleic Acids*, **4**, 1–10.
13. Hoinka,J., Backofen,R. and Przytycka,T.M. (2018) AptaSUITE: A Full-Featured Bioinformatics Framework for the Comprehensive Analysis of Aptamers from HT-SELEX Experiments. *Mol. Ther. - Nucleic Acids*, **11**, 515–517.
14. Werner,T. (2010) Next generation sequencing in functional genomics. *Brief. Bioinform.*, **11**, 499–511.
15. Robinson,M.D., McCarthy,D.J. and Smyth,G.K. (2010) {edgeR}: a {Bioconductor} package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.
16. Martin,M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, **17**, 10.
17. Sefah,K., Shangguan,D., Xiong,X., O'Donoghue,M.B. and Tan,W. (2010) Development of DNA

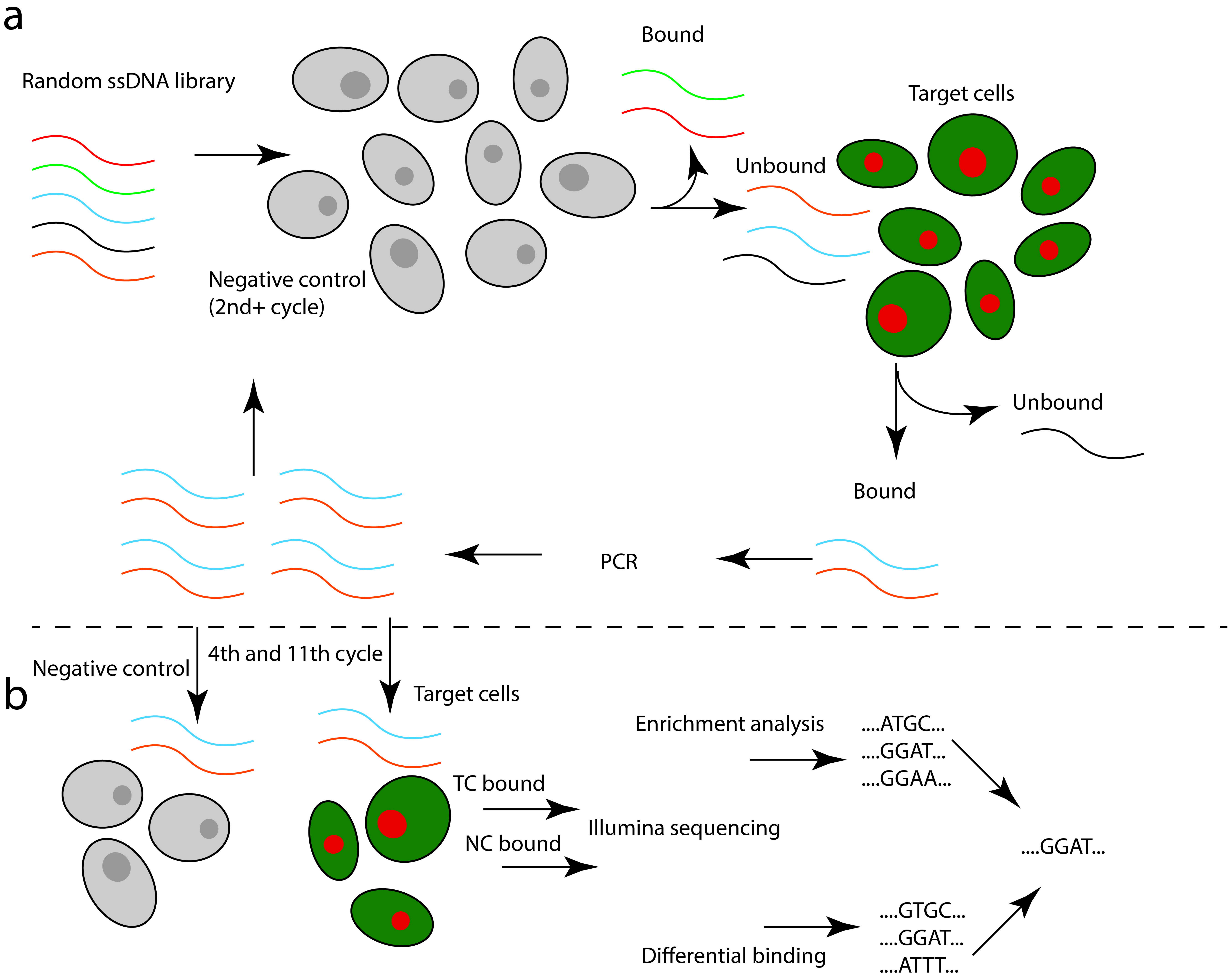
aptamers using cell-selex. *Nat. Protoc.*, **5**, 1169–1185.

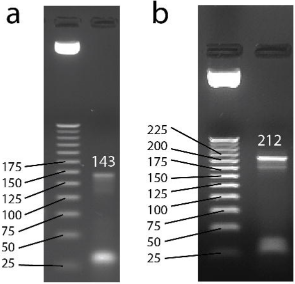
18. Illumina (2013) 16S Metagenomic Sequencing Library Preparation.
19. Alam,K.K., Chang,J.L. and Burke,D.H. (2015) FASTAptamer: A bioinformatic toolkit for high-throughput sequence analysis of combinatorial selections. *Mol. Ther. - Nucleic Acids*, **4**, 1–10.
20. Benjamini,Y. and Hochberg,Y. (1995) Controlling The False Discovery Rate - A Practical And Powerful Approach To Multiple Testing. *J. R. Stat. Soc., Ser. B*, **57**, 289–300.
21. Mayer,G., Ahmed,M.S.L., Dolf,A., Endl,E., Knolle,P.A. and Famulok,M. (2010) Fluorescence-activated cell sorting for aptamer SELEX with cell mixtures. *Nat. Protoc.*, **5**, 1993–2004.
22. Thiel,W.H., Thiel,K.W., Flenker,K.S., Bair,T., Dupuy,A.J., McNamara,J.O., Miller,F.J. and Giangrande,P.H. (2015) Cell-Internalization SELEX: Method for Identifying Cell-Internalizing RNA Aptamers for Delivering siRNAs to Target Cells. In *Methods in molecular biology (Clifton, N.J.)*.Vol. 1218, pp. 187–199.
23. Ozer,A., Pagano,J.M. and Lis,J.T. (2014) New technologies provide quantum changes in the scale, speed, and success of SELEX methods and aptamer characterization. *Mol. Ther. - Nucleic Acids*, **3**, 1–18.
24. Hoinka,J., Berezhnoy,A., Dao,P., Sauna,Z.E., Gilboa,E. and Przytycka,T.M. (2015) Large scale analysis of the mutational landscape in HT-SELEX improves aptamer discovery. *Nucleic Acids Res.*, **43**, 5699–5707.
25. Trapnell,C., Roberts,A., Goff,L., Pertea,G., Kim,D., Kelley,D.R., Pimentel,H., Salzberg,S.L., Rinn,J.L. and Pachter,L. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.*, **7**, 562–578.
26. Ritchie,M.E., Phipson,B., Wu,D., Hu,Y., Law,C.W., Shi,W. and Smyth,G.K. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, **43**, e47–e47.
27. Team,R.C. (2018) R: A Language and Environment for Statistical Computing.
28. Lobo,N.C., Gedye,C., Apostoli,A.J., Brown,K.R., Paterson,J., Stickle,N., Robinette,M., Fleshner,N., Hamilton,R.J., Kulkarni,G., *et al.* (2016) Efficient generation of patient-matched malignant and normal primary cell cultures from clear cell renal cell carcinoma patients: Clinically relevant models for research and personalized medicine. *BMC Cancer*, **16**, 1–15.
29. Grüning,B., Dale,R., Sjödin,A., Chapman,B.A., Rowe,J., Tomkins-Tinch,C.H., Valieris,R., Köster,J. and Bioconda Team (2018) Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nat. Methods*, **15**, 475–476.
30. Wickham,H. (2017) tidyverse: Easily Install and Load the ‘Tidyverse’.

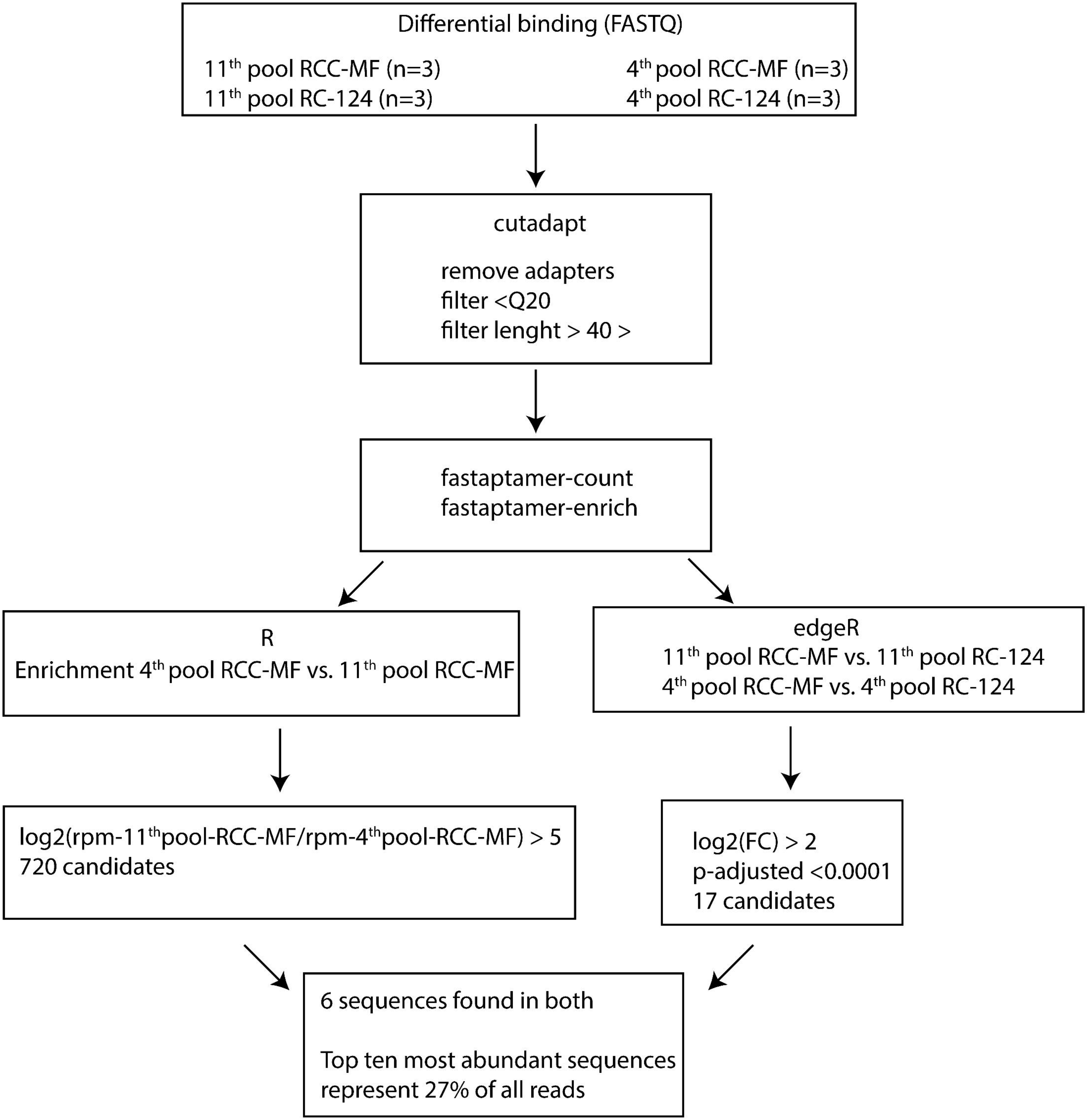
## TABLE AND FIGURES LEGENDS


a
Count
300
200
100
0
$10^0$
$10^1$
$10^2$
$10^3$
$10^4$
FAM-HLog::RCC-MF
b
Count
400
300
200
100
0
$10^0$
$10^1$
$10^2$
$10^3$
$10^4$
FAM-HLog::RC-124

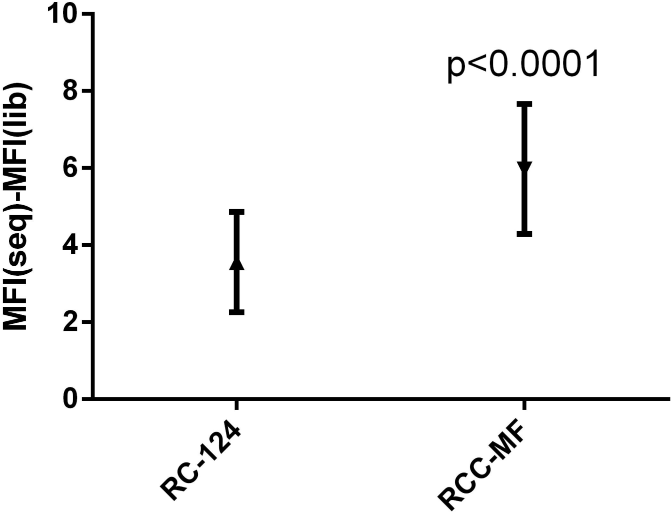

a
0 nM
11.25 nM
22.5 nM
45 nM
90 nM
135 nM
225 nM
$10^0$
$10^1$
$10^2$
$10^3$
$10^4$
FAM-HLog-RC-124
b
0 nM
11.25 nM
22.5 nM
45 nM
90 nM
135 nM
225 nM
$10^0$
$10^1$
$10^2$
$10^3$
$10^4$
FAM-HLog-RCC-MF
c
RC-124 geom-MFI, AU
$B_{max}$ = 26,27
$K_d$ = 189,4 nM
0
5
10
15
20
0
50
100
150
200
250
C, nM
d
RCC-MF geom-MFI, AU
$B_{max}$ = 25,85
$K_d$ = 169,1 nM
0
5
10
15
20
0
50
100
150
200
250
c, nM


a
Random ssDNA library
Negative control
(2nd+ cycle)
Bound
Unbound
Target cells
Unbound
Bound
PCR
b
Negative control
4th and 11th cycle
Target cells
TC bound
NC bound
Illumina sequencing
Enrichment analysis
....ATGC...
....GGAT...
....GGAA...
Differential binding
....GTGC...
....GGAT...
....ATTT...
....GGAT...


a
175
150
125
100
75
50
25
143
b
225
200
175
150
125
100
75
50
25
212


Differential binding (FASTQ)
11th pool RCC-MF (n=3)
11th pool RC-124 (n=3)
4th pool RCC-MF (n=3)
4th pool RC-124 (n=3)
cutadapt
remove adapters
filter <Q20
filter lenght > 40 >
fastaptamer-count
fastaptamer-enrich
R
Enrichment 4th pool RCC-MF vs. 11th pool RCC-MF
edgeR
11th pool RCC-MF vs. 11th pool RC-124
4th pool RCC-MF vs. 4th pool RC-124
log2(rpm-11thpool-RCC-MF/rpm-4thpool-RCC-MF) > 5
720 candidates
log2(FC) > 2
p-adjusted <0.0001
17 candidates
6 sequences found in both
Top ten most abundant sequences represent 27% of all reads

a
logFC
4
2
0
-2
-4
10
12
14
16
Average logCPM
b
logFC
4
2
0
-2
-4
-6
5
10
15
Average logCPM
c
Average logCPM
16
12
8
4
−5.0
−2.5
0.0
2.5
logFC


MFI(seq)-MFI(lib)
10
8
6
4
2
0
p<0.0001
RC-124
RCC-MF