7480–7494 Nucleic Acids Research, 2018, Vol. 46, No. 15 Published online 3 July 2018
doi: 10.1093/nar/gky583

Time-lapse imaging of molecular evolution by high-throughput sequencing

Nam Nguyen Quang[1,2,3], Clément Bouvier[1,2,4], Adrien Henriques[1,2,3], Benoit Lelandais[1,2,3] and Frédéric Ducongé[1,2,3,*]

[1]CEA, Fundamental Research Division (DRF), Institut of Biology François Jacob (Jacob), Molecular Imaging Research Center (MIRCen), Fontenay-aux-Roses, France, [2]Neurodegenerative Diseases Laboratory, CNRS CEA UMR 9199, Fontenay-aux-Roses, France, [3]Université Paris-Saclay, Université Paris-Sud, Orsay, France and [4]Université Pierre et Marie Curie, Paris, France

Received December 21, 2017; Revised June 12, 2018; Editorial Decision June 13, 2018; Accepted June 15, 2018

ABSTRACT

High-throughput sequencing of *in vitro* selection could artificially provide large quantities of relic sequences from known times of molecular evolution. Here, we demonstrate how it can be used to reconstruct an empirical genealogical evolutionary (EGE) tree of an aptamer family. In contrast to classical phylogenetic trees, this tree-diagram represents proliferation and extinction of sequences within a population during rounds of selection. Such information, which corresponds to their evolutionary fitness, is used to infer which sequences may have been mutated through the selection process that led to the appearance and spreading of new sequences. This approach was validated by the re-analysis of an *in vitro* selection that had previously identified an aptamer against Annexin A2. It revealed that this aptamer might be the descendant of a sequence that was more highly amplified in early rounds. It also succeeded in predicting improved variants of this aptamer and providing a means to understand the influence of selection pressure on evolution. This is the first demonstration that HTS can provide time-lapse imaging of the evolutionary pathway that is taken by a macromolecule during *in vitro* selection to evolve by successive mutations through better fitness.

INTRODUCTION

Since the 1960s, when the idea of a pre-cellular world was first introduced, Darwin's theory was extrapolated to the evolution of biomolecules outside living organisms (1). Since then, *in vitro* selection studies have demonstrated that nucleic acids and proteins can evolve in a test tube (2). Such experiments have provided fundamental information of how molecular function can arise from randomly synthesized molecules and have contributed to the establishment of the RNA-world hypothesis (3–5). Hence, *in vitro* selection has been used to study or design ribozymes with natural or novel catalytic activities (6). It has also been used to identify the sequence motifs that are recognized by several DNA- and RNA-binding proteins (7,8). The latter approach was popularized using the term SELEX (for systematic evolution of ligands by exponential enrichment) (7). SELEX has also provided new types of artificial ligands, usually called aptamers (9), which are increasingly used in various biotechnology applications. For example, they can be used to regulate gene expression (10), design biosensors (11), purify therapeutic proteins (12), discover new biomarkers (13), or the targeting of drugs, nanoparticles, or contrast agents (14). Additionally, aptamers represent a new class of drugs, with one aptamer already commercialized for the treatment of age-related macular degeneration and several others currently being evaluated in clinical trials (15).

The growing interest in *in vitro* selection has led to several improvements of the technology, but it is still based on a Darwinian evolution process (16). First, a combinatorial population with up to 10^{15} different sequences is synthesized. These sequences are then screened for their ability to bind a target or their catalytic activity. The sequences that satisfy such criteria are recovered by partition, whereas the others are removed. The 'winning' sequences are then amplified by PCR or RT-PCR and *in vitro* transcription (for DNA or RNA libraries, respectively). The repetition of selection and amplification gradually enriches the population for sequences that are adapted to the selection pressure. Additionally, these molecular species are also expected to evolve by mutations during the amplification steps. These mutations can produce new sequences that are slightly different from those of their parents. Those that contain 'beneficial' mutations for their selection have a better chance to be

[*]To whom correspondence should be addressed. Tel: +33 146 548 641; Email: frederic.duconge@cea.fr
Present address: Benoit Lelandais, Unité Imagerie et Modélisation CNRS 3691, Institut Pasteur, Paris 75015, France.

enriched in the population. Therefore, only molecules with the best-inherited traits from the starting library should survive and are expected to gradually evolve, leading to the progressive enrichment of the best-suited nucleic acid structures inside the library. However, this mechanism of molecular evolution remains largely theoretical and impossible to visualize. Accordingly, it is still not known whether the final winning sequences are already present in the original randomly synthesized library or whether they originate from variations of the original library sequences.

Until now, most aptamers have been identified after several rounds of *in vitro* selection by cloning and Sanger sequencing of approximately 100 sequences. The identification of aptamers by such low-throughput sequencing requires significant enrichment of a few sequences in the final population. However, the number of rounds and selection parameters are usually chosen arbitrarily. Indeed, *in vitro* selection is well known to be highly unpredictable due to the influence of a large number of variables (for example, the potential affinity of aptamers for the target, the amount of potential aptamer structures in the starting library, the effect of the selection parameters, etc.) (17). Additionally, better molecules are often identified after a few rounds of *in vitro* selection using a partially randomized (doped) library of aptamers or ribozymes (6,18,19). Therefore, for every *in vitro* selection, the question remains whether better sequences could have been identified with more rounds of selection or if other parameters of selection had been used.

High-throughput sequencing (HTS) technologies have been recently proposed to address this issue, enabling deeper analysis of the population during *in vitro* selection experiments (see for a review (20)). It allows, for the first time, the analysis of millions of sequences from different rounds to reveal the evolution of sequences inside the population. Such analyses have already demonstrated that the most abundant aptamer sequences in the final population are not always those with the highest affinity (21–24). They have also demonstrated that HTS can detect sequences for which their enrichment is significantly affected by a change in a selection parameter (25,26). Nevertheless, there is still no methodology for interpreting this data to understand what happens during *in vitro* selection and how this knowledge could be used to identify better compounds.

As for most big data, one of the major challenges of HTS is to transform the information contained in millions of objects (here sequencing reads) into simple, easily understandable, graphics. One way to interpret evolution is to represent it as a theoretical fitness landscape which links genotypes and the fitness for a phenotype (27,28). The peaks in such landscapes correspond to the fittest genotypes for a particular phenotype, whereas less fit sequences are down in the valleys. Jason N. Pitt and Adrian R. Ferré-D'Amaré used this concept in 2010 to analyse an *in vitro* selection experiment that was performed to select RNA ligase ribozymes (29,30). They used HTS to monitor the evolution of $\sim 10^7$ variants of a previously known RNA ligase ribozyme. This large amount of information was used to link the genotype of each variant to its activity and allowed the generation of an extraordinarily high resolution 3D graph representing an empirical fitness landscape. The peak in this graph was steep and narrow, showing that most variants of the ribozyme lose their activity. It also identified several neutral positions in which mutations had no impact. This method can be particularly efficient to analyse doped-SELEX, in which the entire library is composed of variants from a known aptamer or ribozyme. However, it is more difficult to use in a classical SELEX experiment because the library usually contains several different aptamer families. Therefore, it provides significantly fewer sequencing reads per family, allowing the analysis of very few variants per family. As a result, Jimenez *et al.* shown that the reconstructed fitness landscape from a SELEX starting with a naive library was composed of disconnected peaks formed by only a few hundred variants (31).

Another way to estimate the fitness landscape through *in vitro* selection is to study the variants that appear accidentally through mutations by polymerases. Such mutations are observable facts that are often assimilated as the raw material for evolution. Accordingly, although the mutations are random, selective pressure should promote increased enrichment of those that provide better fitness to survive. Such an evolutionary process was impossible to observe with low-throughput sequencing because mutations are rare. However, HTS can detect many variants for the most abundant sequences, with some point mutations. Some of these variants have already been identified to be better binders than the most abundant aptamer from which they were derived (24,25,32). Such variants were more highly enriched between two rounds of *in vitro* selection than the most abundant sequence. This property was used by Hoinka *et al.* to propose a score to predict the best variants (24). However, superior enrichment did not automatically correlate with better affinity, suggesting that comparing round-to-round enrichment is not sufficient to highlight beneficial mutations. Moreover, this method does not consider that evolution is a dynamic process that is expected to progress sequentially. For example, if a variant contains a mutation (x) that is beneficial, this variant will not only spread throughout the population, but will also generate other variants containing (x) and other mutations (y_n) that may increase in the population if they are neutral or provide an additional benefit. Among these new variants, those that contain additional beneficial mutations that further improve fitness should increase more rapidly in the population, and so on. Finally, such an evolutionary process must gradually lead to the extinction of other variants that were present in the population at early stages, but which are less well adapted to the selection conditions.

Evolution derived from the slight divergence of phenotypic traits, combined with the principles of selection and extinction was described by Charles Darwin in his seminal work '*On the Origin of Species by Means of Natural Selection*' (33). He used a dendrogram to represent such an evolutionary process, drawing the relationship between different hypothetical variants through successive generations. Since then, evolutionary trees based on phylogeny have been extensively used to depict evolution and most are now constructed from sequence alignments. Common methods of phylogenetic-tree construction use a metric that measures the distance between sequences to connect them by a common ancestor based on their similarity. Such reconstructions are extensively used to infer probable evolutionary pathways between species from the present to the past (34). However, phylogenetic trees are often inappropriate for depicting the evolution that occurs during *in vitro* selection because most selected sequences are different within a restricted range of mutations. Thus, most are separated by a similar Levenshtein distance and are connected to a small number of connections.

Here, we reasoned that the HTS analysis of several rounds of *in vitro* selection could artificially provide large quantities of relic sequences from known times in an evolutionary process. We use these relics, which are not commonly available in standard evolution studies, to construct a new type of phylogenetic analysis that represents an empirical genealogical evolutionary (EGE) tree. This tree-diagram is built by combining sequence alignments with historical information to infer which sequences have mutated through the selection rounds, leading to the appearance of new sequences in the population. Furthermore, it displays the frequency of each variant at each round of *in vitro* selection. This information can show the proliferation or extinction of sequences within the population, which is directly related to their evolutionary fitness. Consequently, an EGE tree can reconstruct the evolution pathways taken during an *in vitro* selection experiment to result in sequences with better fitness. This tree can be used to study the impact of the selective pressure. More importantly, it can also be used for the identification of better molecules.

We validated our approach by re-analysing a cell-SELEX experiment that we had previously carried out and that led to the identification of an anti-Annexin A2 aptamer, named ACE4 (23). Annexin A2 is an important therapeutic target that is involved in many biological processes and is overexpressed at the surface of various types of cancer cells (35). Our aim was to first evaluate whether an EGE tree can identify aptamers with a higher affinity than those previously published. We also investigated how such analysis can be used to study the effect of selection parameters and how it can replace post-SELEX optimisation, such as doped-SELEX, which is usually performed to improve aptamers (6,18,19).

MATERIALS AND METHODS

Reagents

Oligonucleotides were chemically synthesized by Eurogentec (Angers, France). Chemical reagents were purchased from Sigma-Aldrich (Saint-Quentin Fallavier, France) and the reagents used for molecular and cellular biology were purchased from Thermo Fisher Scientific (Villebon-sur-Yvette, France), unless otherwise specified.

Cells

MCF-7 cells derived from human breast adenocarcinoma were purchased from ATCC (Manassas, VA, USA) and grown in RPMI 1640 media supplemented with 10% FBS and 1% antibiotics, at 37°C in a 5% CO_2 atmosphere.

High-throughput sequencing

HTS was performed on a GAIIx instrument (Illumina, Little Chesterford, UK). Adapter and indexing sequences required for Illumina multiplexing sequencing were added to the DNA libraries by PCR. The PCR products were purified on a 3% agarose gel and recovered by passive elution in TE-NaCl buffer (1× Tris–EDTA, 25 mM NaCl). Each eluate was concentrated and precipitated with ethanol before resuspension in distilled water. Samples were then mixed and loaded with 5 to 10% PhiX into a flow-cell and sequenced according to the provider's instructions. HTS data were demultiplexed and recovered in FASTQ format using *bcl2fastq Conversion Software v1.8.4 (https://support.illumina.com/downloads/bcl2fastq_conversion_software_184.html)*.

Primer trimming and quality filtering of sequences.

All FASTQ files were processed using a software suite called *PATTERNITY-SEQ, developed from the aptamer platform in MIRCen (http://jacob.cea.fr/drf/ifrancoisjacob/english/Pages/Departments/MIRCen/Platforms.aspx?Type=Chapitre&numero=6)*. The different steps of the sequencing analysis are provided in Supplementary Figure S1. First, adapter and primer sequences were removed from each sequence, keeping only variable regions. Then, sequences that contained at least one base with a quality score (Q) below 30 were removed before being saved in a FASTA format. This quality score can be converted to a probability of error (P) using the formula $P = 10^{(-Q/10)}$. Thus, the recovered sequences contain bases with a potential probability of error below 0.001 (1 in 1000).

Clustering of sequences in families based on Levenshtein distance

First, the frequency of each sequence in the different libraries was calculated. Only sequences with a frequency >0.001% in at least one round were recovered to decrease the time of analysis and to remove sequences of poor interest. These sequences were then sequentially clustered in families using a Levenshtein distance of 10 (*i.e.* sequences with no more than 10 substitutions, insertions or deletions) using an approach similar to that described by Alam *et al.* (36). Clustering was performed by putting the sequences in a list sorted by their maximum frequencies (whatever the round) through SELEX. The sequence with the highest maximum frequency was used as a reference and compared to the others. All sequences meeting the distance criteria were separated from the others and defined one cluster. The algorithm repeats the same process with the remaining sequences and continues until the list is empty. All steps were performed using *PATTERNITY-SEQ*.

Correlation between the coefficient of variation (CV) and sampling size

The DNA library corresponding to round 15 of the cell-SELEX was re-sequenced to obtain a higher number of sequences from this library (11 563 613 sequences). This library was chosen because it contains sequences with heterogeneous frequencies including enriched sequences that represent up to 10% of the library. Ten samples of various sizes (100 000; 200 000; 500 000 and 1 000 000 sequences) were randomly extracted using the web-based platform *Galaxy*

Project (37). The frequencies of each sequence in the different samples were then measured. These values were further used to calculate the mean frequency and coefficient of variation (CV) for each sequence for each sample of different size.

Phylogenetic tree of the ACE4 aptamer family

All variants of the ACE4 family representing at least 0.01% of the library in one round were used to build a phylogenetic tree using the software MEGA7 (38). The tree was generated by the Maximum Parsimony method using the Subtree-Pruning-Regrafting (SPR) algorithm (39). The branch lengths were calculated using the average pathway method and are in the units of the number of changes over the whole sequence.

Empirical genealogical evolutionary (EGE) tree of the ACE4 aptamer family

The EGE tree of the ACE4 family was built using *Cytoscape* (40). For each round of selection, every variant of the ACE4 family representing at least 0.01% of the library was recovered and used to establish a node. At every round *R*, each variant was linked to a single variant of *R-1* with the closest similarity (*i.e.* Levenshtein distance) and which was the most abundant at *R-1.* Alternatively, the Levenshtein distance can be replaced by a similarity distance calculated by other methods of multi-alignment such as Clustal Omega (41) or MAFFT (42). This method is fully explained in the Results and in Supplementary Figure S2. It produces a network of interactions between variants through the rounds of selection down to a single node, which is the first variant that represents >0.01% of the population. This variant can be defined as the 'potential ancestor' of the ACE4 family. Once the network is imported into *Cytoscape*, it is first organized using the Tree layout of the software. Every node is then horizontally aligned per round of selection. Finally, the size and color scale of each node is defined by the percentage of each variant in the family at one round. The name of each variant was manually added to the dendrogram.

Evaluation of the error rate due to PCR or sequencing errors

A chemically synthesized DNA sequence of ACE4 was amplified in triplicate by 7 or 24 PCR cycles using primers elongated with illumina adapter sequences, allowing for multiplexed sequencing. The PCR was performed in a final volume of 200 μl using the same condition of the cell-SELEX. One pmole of matrix was added in a PCR mix that contained a PCR buffer with 3mM MgCl2, 5% of DMSO, 1 μM of each primer, 0.2 mM dNTP and 5 units of Taq polymerase. The buffer and the Taq polymerase were from the DyNAzyme EXT DNA Polymerase (Thermo Fisher Scientific). The cycles of PCR were 94°C for 30 s, 53°C for 1 min and 72°C for 1.5 min. The resulting libraries were sequenced and two million sequences per condition were randomly extracted using the web-based platform *Galaxy Project* (37). The positional nucleotide frequency was obtained using *BioEdit (43)*.

Preparation of radiolabeled 2′F-Py RNA aptamers

Chemically synthesized ssDNA templates were amplified by PCR before being *in vitro* transcribed in 2′F-Py RNA and purified as previously described (23,44). The sequence used as a negative control corresponded to a scrambled sequence of ACE4 (ACE4scr): 5′-GGG-AGA-UGA-UCC-GUU-GAU-GCG-AGC-ACU-ACA-ACU-GCU-GGU-CAG-CAC-UAC-UGG-GAC-GCC-AGC-UGA-CGG-CGG-AGA-AGU-CGU-CGU-UCG-UAG-GCA-GAA-UC-3′. The sequence of the ACE4 aptamer was: 5′-GGG-AGA-UGA-UCC-GUU-GAU-GCG-AGG-GAA-CGC-AAG-AAC-UGA-GGC-CAU-GAG-GCG-CCU-UCC-CUU-GCU-CAG-GAC-GCA-AGU-CGU-CGU-UCG-UAG-GCA-GAA-UC-3′. Sequences of ACE4 variants corresponded to the ACE4 aptamer, in which point mutations were introduced according to the nomenclature X*n*Y, where X is the original base, *n* is the number of its position, and Y corresponds to the mutation. If several point mutations were introduced they are separated by '/'. For example, ACE4 G33A/A44G corresponds to the sequence of ACE4 in which the guanine in position 33 and the adenine in position 44 where substituted by an adenine and a guanine, respectively. For all binding experiments, 2′F-Py RNAs were gel purified before [^{32}P] radiolabeling at the 5′ extremity, as previously described (44,45), to achieve a labelling yield of approximately 3–6 MBq/pmol of oligonucleotides.

Radioactive binding of aptamers to adherent cells

Competitive and saturation bindings were performed in triplicate using a liquid handling robot (Microlab Starlet – Hamilton, Villebon-sur-Yvette, France) and 24-well plates containing approximately 100 000 MCF-7 cells per well. For competitive binding, a radioactive ACE4 aptamer was incubated at a final concentration of 5 nM in 200 μl RPMI 1640 containing 100 μg/ml of both tRNA and Polyinosinic acid (Poly(I)). The cells were incubated at 37°C for 15 min in the presence of an equimolar concentration of an unlabelled ACE4's variant or wildtype ACE4. Unbound oligonucleotides were then removed by washing five times with 500μL RPMI 1640 and the amount of radioactive ACE4 aptamer attached to cells was counted using a MicroBeta TriLux counter (Perkin Elmer, Villebon-sur-Yvette, France). The ratio of competitive binding compared to ACE4 was calculated for each ACE4 variant by dividing the amount of radioactive ACE4 bound in the presence of wild-type ACE4 by the amount of radioactive ACE4 bound in the presence of a variant.

Saturation binding experiments were performed using a protocol extensively described elsewhere (45). Briefly, cells were incubated with radiolabelled 2′F-Py RNA aptamers at different concentrations, in 200μL RPMI 1640, containing 100 μg/ml of both tRNA and Poly(I), at 37°C for 15 min. After washing five times with 500 μl RPMI 1640, the amount of radioactive ACE4 aptamer attached to the cells was counted using a MicroBeta TriLux counter. The specific binding of aptamers was measured by subtracting the background values obtained with ACE4scr and apparent K_d values were determined by fitting the binding curves with

GraphPad Prism 6 (GraphPad Software, La Jolla, USA) using a one site-specific binding model.

Kinetic binding studies were performed using the White Ligand Tracer instrument (Ridgeview Instruments AB), which can measure the amount of radiolabelled aptamer bound to the cells in real time. One day before the binding experiments, 10^6 MCF-7 cells were seeded in a 10 cm^2 dish according to the supplier's instructions, to obtain a monolayer of cells in part of the dish. Just before binding, cells were washed with DPBS Mg^{2+}/Ca^{2+}, to remove dead cells, before adding 3 ml RPMI 1640 containing 0.1% sodium azide to avoid cellular internalization by endocytosis. The dish was then placed on the inclined support of the instrument for 10 min of background acquisition in the presence of tRNA and Poly(I) (100 μg/ml each). Radiolabelled aptamer (at a final concentration of 3nM) was added and its association with cells was measured for 30 min. The media was then replaced by fresh RPMI 1640 containing 0.1% sodium azide and the dissociation of aptamers from the cell surface was measured for 60 min. Measurements were performed using the 2 × 3 opposite positions mode and a 4 s integration time per measure. The dissociation rate constant (k_{off}) was calculated using *Trace Drawer software* (Ridgeview Instruments AB). The association rate constant (k_{on}) was calculated by dividing the k_{off} by the apparent K_{d} previously measured during the saturation binding experiments.

Doped cell-SELEX

A partially randomized DNA library was synthesized based on the sequence of the ACE4 aptamer 5′-GGG-AGA-TGA-TCC-GTT-GAT-GCG-AGg-gaa-cgc-aag-aac-tga-ggc-cat-gag-gcg-cct-tcc-ctt-gct-cag-gac-gcA-AGT-CGT-CGT-TCG-TAG-GCA-GAA-TC-3′ (bases in lowercase correspond to the ACE4 aptamer in which point mutations were introduced at a rate of 7.5% with an equal mixture of all three other bases; bases in uppercase are identical to the ACE4 aptamer and correspond to primer binding sites used during amplification).

Doped-SELEX was performed as the cell-SELEX previously described for the identification of the ACE4 aptamer (23,44). Briefly, the doped DNA library was amplified by 12 cycles of PCR and *in vitro* transcribed using 2′F-Pyrimidines (Trilink Biotechnologies, San Diego, CA, USA) and a mutant form of T7 RNA polymerase (T7Y639F, kind gift of R. Souza). After treatment with DNAse I and PAGE purification, 25 pmol 2′F-Py RNA library containing approximately 10^{12} sequences was heated at 85°C for 5 min, snap-cooled on ice for 5 min, and allowed to warm to 37°C. Then, 2′F-Py RNAs were incubated 10 min at 37°C with 1.65 × 10^6 adherent MCF-7 cells in 500μl RPMI 1640 containing 5 μg yeast tRNA. The cells were washed five times with 5 ml RPMI 1640 to remove unbound sequences (the last wash lasted 5min). Finally, bound oligonucleotides were recovered using the NucleoSpin RNA II RNA extraction kit (Macherey-Nagel, Hoerdt, France). Then, 2′F-Py RNAs were reverse transcribed with Superscript II RT before re-amplification by PCR and *in vitro* transcription. Four rounds of doped cell-SELEX were performed. An aliquot from each PCR was analysed by two independent high-throughput sequencing runs.

Analysis of the mutational landscape from the doped cell-SELEX

Galaxy Project was used to randomly select 800 000 sequences from the starting doped library and each round of selection. The positional nucleotide frequency of each library was obtained using *BioEdit* (43). For each position *p*, a normalized enrichment ratio (*RN*) between rounds 4 and 0 (starting doped library) was calculated for each base *b* as described by the formula:

$$RN_{p,b} = \frac{r_{p,b,4/0}}{\sum_{b\epsilon B} r_{p,b,4/0}}$$

where $B = \{A, U, C, G\}$ and

$$r_{p,b,4/0} = \frac{P_{p,b,\,R=4}}{P_{p,b,\,R=0}}$$

where $P_{p,b,R=4}$ is the percentage of base *b* at position *p* at round 4 and $P_{p,b,R=0}$ is the percentage of base *b* at position p at round 0.

These *RN*s have been previously used by Carothers *et al.* to analyse a doped-SELEX experiment (46). They were added, using the software *VARNA* (47), as a color scale to the predicted secondary structure of ACE4 built by *MFold* (48).

The percentage of each sequence *x* and their Levenshtein distance to the wildtype ACE4 were obtained by *PATTERNITY-SEQ* to reconstruct a 2D mutational landscape of the ACE4 aptamer. Only sequences with a frequency >0.001% in one round were recovered. The relative enrichment factor (*E*) between rounds 4 (R4) and 0 (R0) relative to that of the wildtype ACE4 was then calculated for each sequence *x* by the formula:

$$E = \frac{\frac{P_{x,\,R=4}}{P_{x,\,R=0}}}{\frac{P_{ACE4,\,R=4}}{P_{ACE4,\,R=0}}}$$

where $P_{x,R=4}$ is the percentage of the sequence *x* at round 4, $P_{x,R=0}$ the percentage of the sequence *x* at round 0, $P_{ACE4,R=4}$ the percentage of the wildtype ACE4 sequence at round 4 and $P_{ACE4,R=0}$ the percentage of the wildtype ACE4 sequence at round 0. Finally, a 2D mutational fitness landscape was drawn by GraphPad Prism 6, plotting *E* on the y-axis and the Levenshtein distance on the x-axis for each sequence.

RESULTS

HTS sequencing reveals the Darwinian evolution that occurs during *in vitro* selection

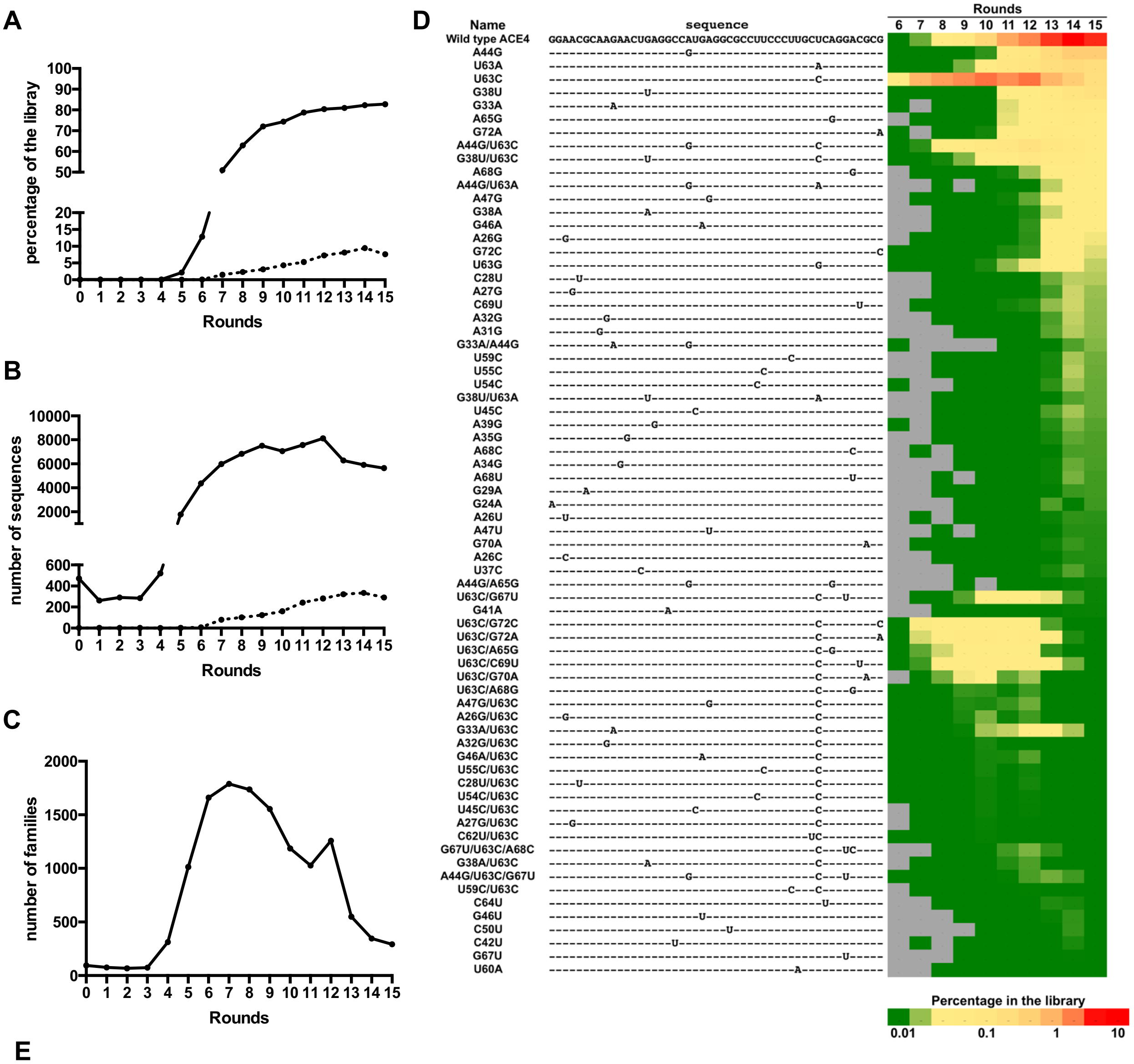
We reanalysed fifteen rounds of a cell-SELEX, that we previously published, to investigate the type of evolution that can be measured during *in vitro* selection using HTS (23). One aptamer, called ACE4, was judged to be particularly attractive during this SELEX because it targets Annexin A2, a protein that is overexpressed at the surface of several types of cancer cells (35). We conducted a more in-depth analysis by examining approximately 15 million sequences, which corresponds to approximately 500 000 to 2 000 000 sequences per round, to better study the evolution of sequences during this SELEX.

We observed that sequences with a shorter random region were enriched during this SELEX (Supplementary Table S1). We decided to recover all the sequences with a random region between 47 and 52 nucleotides to allow the detection of sequences with deletions or insertions. All the sequences were clustered in families using a Levenshtein distance of 10. Thus, every family was composed of similar sequences with no more than 10 substitutions, insertions, or deletions (Supplementary Table S2). The sequences of most families were separated by a maximum Levenshtein distance of four. We monitored the sequences that were enriched during the selection, by analyzing those that could be detected in at least one round at a frequency >0.001% in the library (i.e. 10 copies per million sequences, Supplementary Table S3). We detected a few hundred sequences with a frequency >0.001% in the library until the round 5; but they collectively represented less than 0.1% of the population (solid lines in Figure 1A and B and Supplementary Table S3). Their number increases up to 8652 sequences from round 6 to round 12 before slowly decreasing down to 6680 sequences by round 15 (Figure 1B). Simultaneously, their total prevalence in the pool increased exponentially and represented more than half the pool by round 7 and 83% by round 15 (Figure 1A). The fact that the number of sequences >0.001% decreased after round 12 while their prevalence increased in the library demonstrates that some sequences started to disappear from the population due to greater amplification of others. Such extinction could be clearly seen for instance for the sequences of the ACE22 and ACE105 families, which were the most amplified families at round 7, but which progressively disappeared from the library thereafter (Supplementary Tables S2 and S3). This correlated with an increase in the number of families to 1737 at round 8 before a continuous decrease to 296 at round 15 (Figure 1C). This demonstrates that there is predominant amplification of a few families relative to the others from round 8. Moreover, some of these families contained an increasing number of variants and, consequently, the number of sequences >0.001% decreased less rapidly than the number of families. As an example of such an evolutionary process, the frequency of the family that contained the ACE4 aptamer steadily increased up to 9.5% at round 14 before decreasing slightly to 7.7% at round 15 (dotted line in Figure 1A).

This family contained an increasing number of variants for which the frequency in the pool was >0.001% (dotted line in Figure 1B, Supplementary Tables S3 and S4). One of these variants (named U63C because the uracil in position 63 of ACE4 is replaced by a cytosine) was predominantly amplified in the family between rounds 6 and 10, with an increase in frequency from 0.07% to 2.8% (Figure 1D and Supplementary Table S4). The percentage was then stable for two rounds before continuously decreasing to 0.3% by round 15. This decrease was concomitant with the enrichment of other variants, including ACE4, which was the most abundant variant of the family by round 15, where it represented ∼4% of the total library and 63% of the family (Figure 1D). Overall, these observations suggest that ACE4 was a variant of an ancestral sequence (U63C), which was predominantly enriched in the population by round 6. It also suggests that this ancestral sequence was less well adapted than ACE4 to the selection pressure that increased progressively (Figure 1E). In conclusion, accordingly to previous studies (24,25,32), HTS analysis of the population from every round of our cell-SELEX clearly depicts several evolutionary patterns described by Darwin's theory, including the amplification, extinction, and divergence of sequence species inside families (33).

The frequencies of variants detected in the ACE4 family are higher than the mutation rate that may result from sequencing errors or mutations introduced by PCR

It is well known that sequencing errors can occur, even though the quality of HTS has been considerably improved during past few years (49). In addition, some positions within a sequence may be more prone to mutations introduced by the polymerases than others. We wished to know whether the ACE4 variants come from mutations during PCR or were due to sequencing errors. Thus, we amplified a chemically synthesized sequence of the ACE4 aptamer for 7 or 24 PCR cycles under the same conditions of our SELEX, but using primers that were elongated with adapter sequences for multiplexing HTS. We then analysed two million sequences per PCR products by HTS in triplicate. The number of variants dramatically increased between 7 and 24 PCR cycles and, consequently, the non-mutated ACE4 sequences decreased from 96% to 83%, respectively (data not shown). This suggests that the mutations were mainly due to polymerase errors rather than sequencing errors. This was further confirmed by comparing the mutation frequency per position. Indeed, the primer sequence, which is not enzymatically synthesized during PCR but elongated, was seven times less mutated after 24 PCR cycles than the sequence synthesized by Taq polymerase (Figure 2A). The polymerase mutated twice as many adenines and uracils as guanines and cytosines. Furthermore, transitions were favoured over transversions, regardless of the position (Figure 2B). Thus, a purine was more prone to be mutated in another purine, whereas a pyrimidine had a higher chance of being mutated to another pyrimidine. This result further confirms that observed mutations are predominantly introduced by Taq polymerase, as recent studies demonstrated that Illumina's technologies introduce many more A to C and G to T substitutions (50). Thus, the Taq polymerase that was used during our cell-SELEX can generate a high number of variants, although the mutations are not entirely random. Nevertheless, the highest frequency of mutation was around 0.01 (Figure 2A). As a result, an ACE4 variant generated by PCR or sequencing errors should represent no more than 0.0005% in the library because the ACE4 aptamer was detected at a maximum frequency around 5%. This mutation frequency is too low to explain the high enrichment of particular variants observed during the cell-SELEX, which likely correlates with their evolutionary fitness.



E

Rounds	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Millions of cells in counter-selection	2 x 10	2 x 10	2 x 10	2 x 10	2 x 5	2 x 5	2 x 5	2 x 5	2 x 5	2 x 5	2 x 5	2 x 5	2 x 5	5	5
Millions of cells in selection	10	10	10	10	5	5	5	5	5	5	5	5	5	5	5
Inversion of counter-selection and selection														+	+
Concentration of library (μM)	2	1	1	1	1	1	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
Incubation time (min)	30	30	25	25	25	20	20	15	15	10	10	10	10	10	10
Number of washes	2	3	3	3	4	4	5	5	5	5	5	5	5	5	5
Particular condition of washing									5min for the last wash	5min for the last wash	5min for the last wash	5min for the last wash	5min for the 2 last washes	5min for the 2 last washes	5min for the 2 last washes
Addition of tRNA during incubation (100 μg/mL)											+	+	+	+	+

Figure 1. Evolution of the library and ACE4 family during the cell-SELEX. (**A**) Percentage of the library that is composed of sequences with a frequency >0.001% (solid line). Percentage of the ACE4 family inside the library (dotted line). (**B**) Number of sequences with a frequency > 0.001% (solid line). Number of sequences inside the ACE4 family with a frequency >0.001% (dotted line). (**C**) Number of families that are composed of sequences higher than 0.001% of the library. (**D**) The heatmap shows the evolution of the 70 most amplified sequences of the ACE4 family from round 6 to round 15. Frequencies at 0.01% or less are in darkest green; frequencies at 0 are in grey. The sequences corresponding to the random region are presented, highlighting their mutations relative to the wildtype ACE4 aptamer at the top. The sequences were named according to their mutations relative to ACE4. (**E**) Selection pressure used during cell-SELEX (23). The grey highlighting corresponds to the conditions that changed between two rounds.

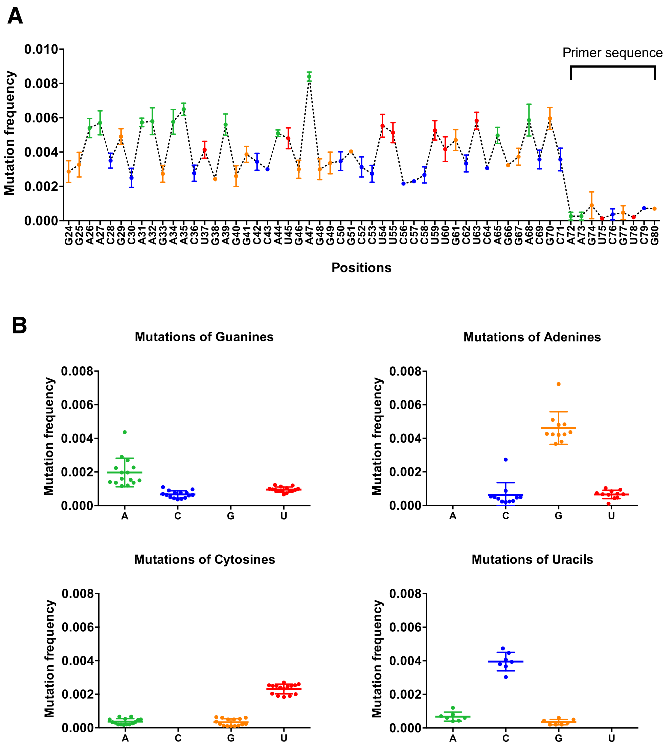


Figure 2. Mutation frequency of the ACE4 aptamer after 24 cycles of PCR. The mutation frequencies were calculated from 2 000 000 sequences obtained after 24 cycles of PCR ($n = 3$). (**A**) Mean mutation frequency for each position of the ACE4 aptamer. The mutation frequencies in a part of the primer sequence (positions 72–80) revealed errors that are strictly due to sequencing since this region is not synthesized by the Taq polymerase. These errors were seven times less frequent than those in the region elongated by the Taq polymerase during PCR (positions 24–71). The mutation yield appeared to be approximately constant for each base, regardless of the position. However, the mutation rate was approximately two-fold higher for adenines (green) and uracils (red) than for guanines (yellow) and cytosines (blue). (**B**) Mutations for each base of the ACE4 aptamer elongated by Taq polymerase. Transitions were favoured over transversions for each base. For example, purines were much more frequently mutated to another purine, and pyrimidines much more frequently to another pyrimidine.

An empirical genealogical evolutionary (EGE) tree can highlight the evolutionary pathway used during cell-SELEX towards the ACE4 aptamer

Phylogenetic analyses based on sequence alignment are often performed to build family relationships between different variants of aptamers. However, these methods provide little information on the possible evolution of aptamers that may have occurred during the successive rounds of *in vitro* selection. Accordingly, such analysis of the ACE4 family only reveals two groups of variants, those containing a U63C mutation with respect to ACE4 and those without (Figure 3A). It does not provide any information about the evolution of these sequences during the rounds of SELEX nor predict possible better variants. Nevertheless, the data provided by HTS analysis of the population during the selection rounds represent a high number of relic sequences from different times of the evolution. We reasoned that this information could be used in addition to sequence alignment analysis to reconstruct an empirical genealogical evolutionary (EGE) tree of an aptamer family during *in vitro* selection.

For this purpose, we postulated that a threshold needs to be arbitrarily set to define at which round a variant appeared in the population at a frequency that indicates undeniable enrichment. We defined this threshold by performing several population analyses to estimate the effect of sampling size on the coefficient of variation (CV) of the measured frequencies. As expected, the CV decreases when the frequency increases and the CV for a given frequency decreases when the sampling size increases (Supplementary Figure S3). Based on this analysis, we chose a frequency of 0.01% as the threshold because we analysed at least 500 000 sequences per round, which should have guaranteed measurements with a CV below 20%. Among the 476 different variants of ACE4 aptamer that were detected, only 70 could be detected in at least one round at a frequency above this threshold. We then used two rules to reconstruct the most probable family relationship between these variants. Rule 1: every variant should be the descendant of a 'potential parent' that is one of the several variants detected above the threshold in the previous round. Rule 2: this potential parent is defined as the most highly similar and most highly abundant variant in the previous round (see Supplementary Figure S2 for an illustration of the method). These rules were used to link the variants through the rounds of selection down to a 'potential ancestor'. These connections were further used to draw an EGE tree that represents a time tree of evolution in which the time unit is one round of selection (Figure 3B). For every round, the percentage of each variant in the family was added to the tree as a node using both colour and size codes. Thus, this tree also provides an indication of the evolutionary fitness of the variants showing the enrichment and decrease of every variant inside a family. To the best of our knowledge, this is the first dendrogram that can simultaneously represent both the fitness and family Relationship of variants during evolution.

Using this diagram, we observed that a 'potential ancestor' of the ACE4 family appeared at round 6 and was not the ACE4 aptamer previously identified as the most abundant variant after 15 rounds of cell-SELEX. This ancestor has a cytosine at position 63 instead of the uracil found in ACE4. Until round 10, this ancestor was the principal representative of the family, although it generated numerous variants, including the ACE4 aptamer, which was first detected at round 7. The frequency of several of these variants increased in the population between rounds 7 and 9. However, the frequency of the ancestor and all its variants containing a cytosine in position 63 decreased dramatically from round 10. Simultaneously, the frequency of the ACE4 aptamer steadily increased and generated new variants with a uracil in position 63, whereas the ancestor produced almost no new variants. This result demonstrates the validity of our model as the appearance of variants correlated with the abundance of their potential parents. It also demonstrates that there was an important shift in the evolution of the family after round 9. Furthermore, although the ACE4 aptamer was the most abundant sequence of the family at round 15, the EGE tree shows that the frequency of three of its variants started to increase in the family during the last three rounds (shown by arrows in Figure 3B). This suggests

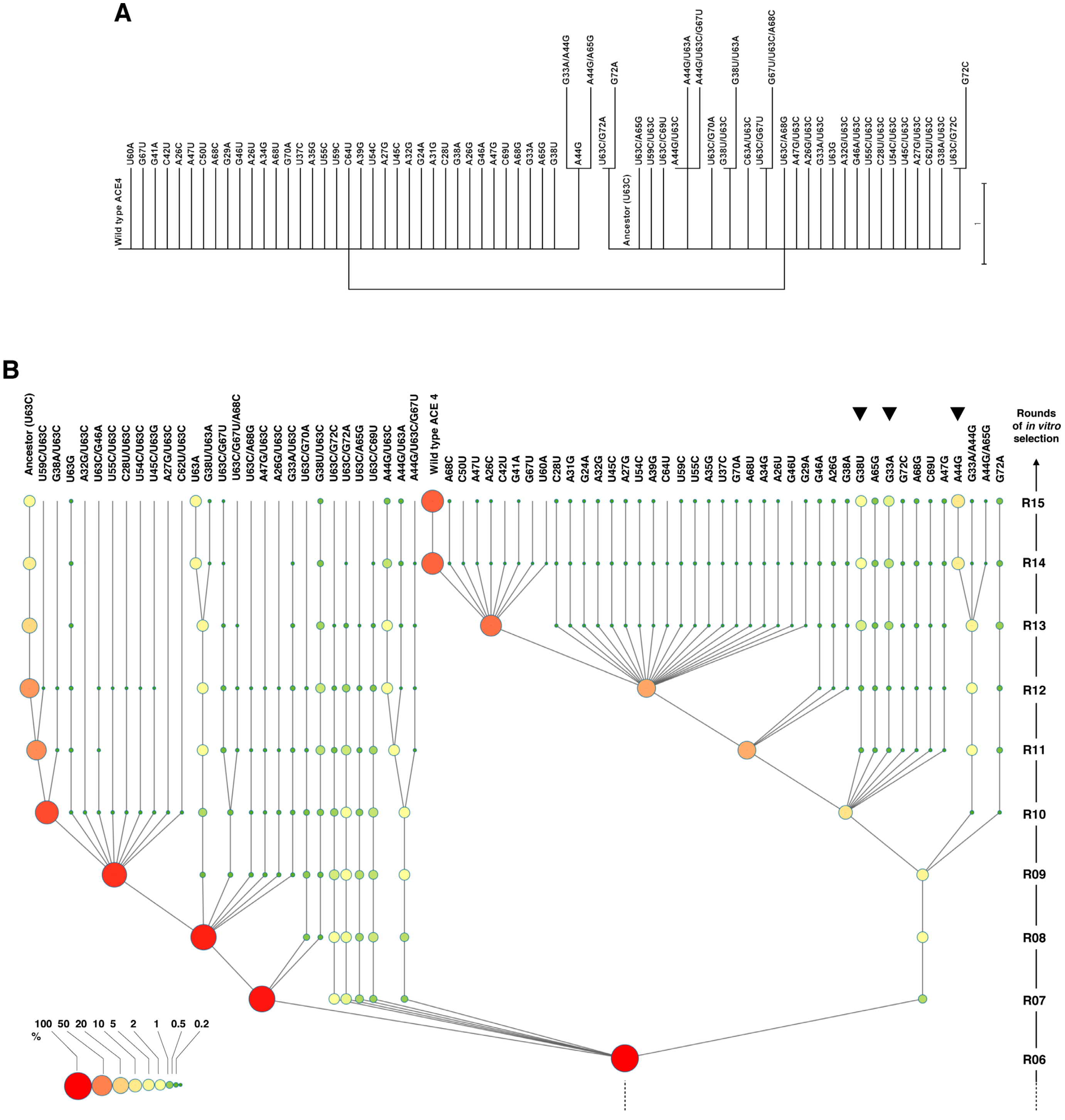


Figure 3. Classical phylogenetic tree and empirical genealogical evolutionary (EGE) tree of the ACE4 family. Every variant of ACE4 that was detected in the library with a frequency that is > 0.01% was used to build evolutionary trees. The name of each variant is shown at the top of each tree. (**A**) Classical phylogenetic tree built using the Maximum Parsimony method. (**B**) The EGE tree was drawn using Cytoscape (40). Nodes with a gradient of different size and colour present the percentage of each variant in the family at each round. These nodes are aligned horizontally for each round, indicated at the right of the tree. These nodes are connected through their potential parents in the previous rounds by lines. The identification of the potential parents is inferred based on the methodology described in Supplementary Figure S2. This dendrogram highlights different evolutionary profiles for groups of variants. It should reflect their fitness to survive the selection pressure during rounds of selection, summarized in Figure 1E. The left branch, which corresponds to the evolution of a potential ancestor sequence (U63C), contains nodes that reflect a high percentage in the family until round 9. All the variants that appear between rounds 7 and 10 are linked to this branch, meaning that they should have been created by mutations of this sequence. After round 9, the frequency of one of the mutants, the ACE4 aptamer with a uracil in position 63, increases within the family and starts to generate new variants through further mutations (branches on the right). Simultaneously, the frequency of variants with a cytosine in position 63 decreases within the family (branches on the left). Despite over-representation of the wildtype ACE4, some of its variants (A44G, G33A and G38U) start to increase within the family from round 13 (black arrows). This suggests that these new variants could be better binders than ACE4 and possibly become dominant within the family and generate other variants if further rounds were performed.

that these variants (containing either the mutation A44G, G33A, or G38U relative to ACE4) may be better ligands than the ACE4 aptamer.

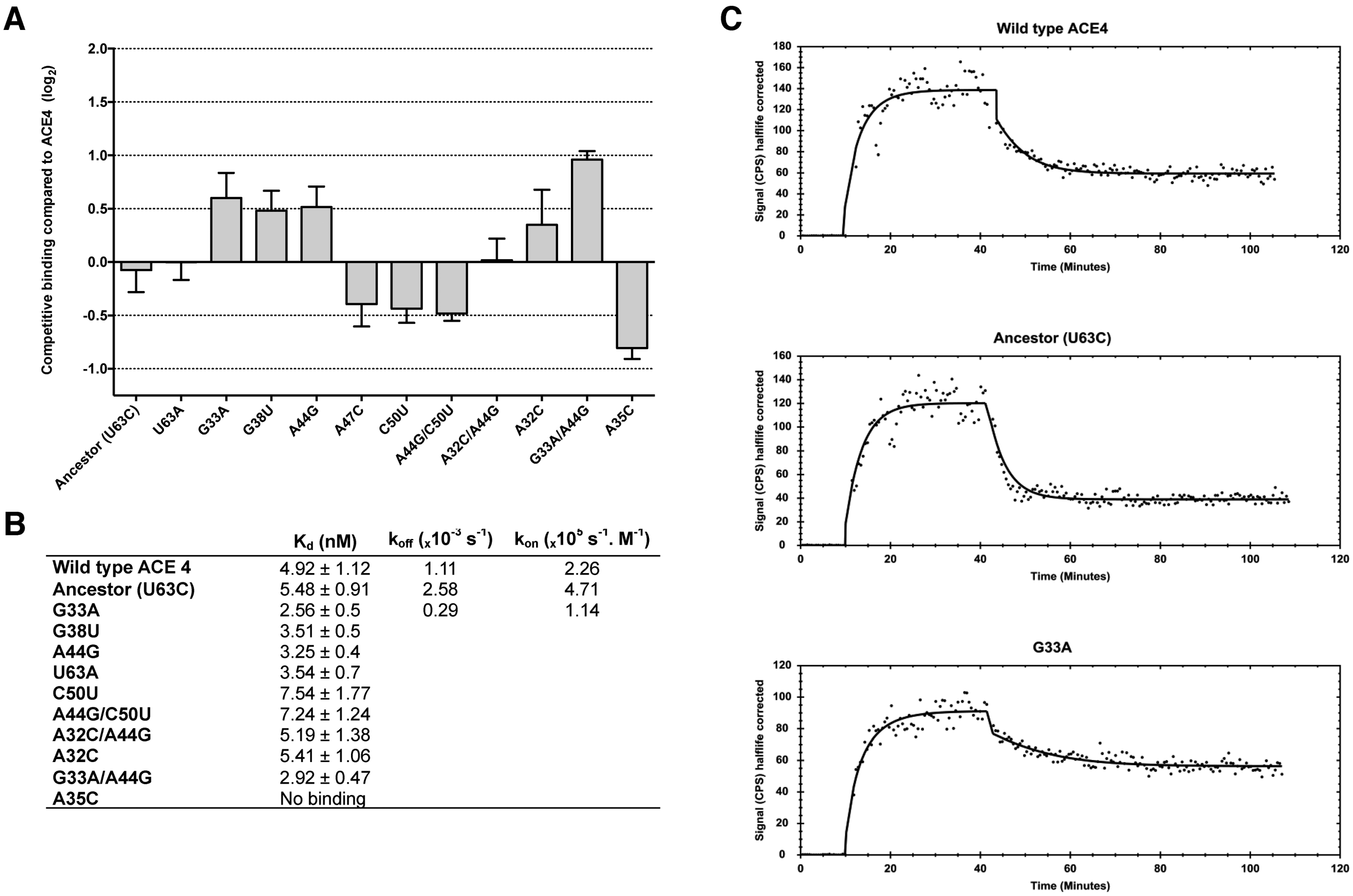
The shift of evolution is correlated with the selection of variants with slower dissociation rate

We evaluated five variants by competitive binding assays, using ACE4 as a reference, to determine whether the evolution observed with the EGE tree correlated with the amplification of better ligands (Figure 4A). As anticipated, the three variants (A44G, G33A and G38U), which demonstrated better enrichment than ACE4 during the last three rounds, showed a 2-fold better competitive binding. Furthermore, variant U63A, which was highly stable in the family during selection, demonstrated the same binding as ACE4. However, variant U63C, which is the ancestor that progressively decreased during selection relative to ACE4, showed an unexpectedly similar binding affinity to that of ACE4. We also performed binding experiments on cells to measure the apparent dissociation constant K_d of these variants (Figure 4B and Supplementary Figure S4). These studies confirmed the previous results, as the ancestor and the U63A variant had the same affinity as ACE4, whereas the three other variants had an approximately two-fold lower K_d than ACE4.

The fact that ACE4 and the ancestor had the same affinity was unanticipated, because the frequency of variants with a cytosine in position 63, such as the ancestor, decreased from round 9 relative to those containing a uracil, such as ACE4 (Figure 3B). Several selection parameters varied during this cell-SELEX to progressively improve the stringency of the selection (Figure 1E). The shift in evolution from round 9 correlated with a more stringent washing of the cells (5 min), which was introduced from round 9 to favour the enrichment of aptamers with a slow dissociation rate. Therefore, we compared the binding kinetic of the ancestor, the wildtype ACE4 aptamer and the G33A variant, which demonstrated the best affinity. We measured the interaction of radioactive aptamers with living cells in real-time using *LigandTracer* (Figure 4B and C). Although the three aptamers had similar K_d, their interaction kinetics were markedly different. The ancestor displayed a faster dissociation rate from cells, with a dissociation rate constant (k_{off}) approximately two- and 10-fold higher than those of ACE4 and the G33A variant, respectively. Based on these k_{off} values and the K_ds that were previously calculated from saturation binding experiments, we estimated the association rate constant (k_{on}) of the aptamers. The ancestor displayed a k_{on} approximately 2- and 4-fold higher than those of ACE4 and the G33A variant, respectively (Figure 4B). Overall, these results suggest that our cell-SELEX conditions favored the enrichment of aptamers with slower dissociation rates from round 9, although they also had slower association rates. In contrast, aptamers with faster association rates may have been favored in early rounds.

Doped cell-SELEX provides a better fitness landscape of the ACE4 aptamer but does not generate better variants

The EGE tree reveals the evolutionary pathway that was taken during the cell-SELEX experiment to generate better variants. However, the question remains whether the mutations introduced by PCR are sufficient to fully explore the fitness landscape of the aptamer, i.e. whether the same variants would have been identified if the *in vitro* selection had been carried out several times. We addressed this question by performing four rounds of cell-SELEX using a library of doped ACE4 aptamers that contain 92% of the original residues and 2.7% of each other residue at each position. This method, called 'doped-SELEX', is usually performed post-selection for the identification of better variants by exploring the fitness landscape of aptamers (6,18,19,29). The starting doped library and the library from each round were analysed by HTS. ACE4 represented only 1.8% of the starting doped library, whereas most of the library contained ACE4 variants with one, two, or tree mutations (Supplementary Table S5). As expected, variants with more mutations were found at a lower frequency in the library. For example, each variant with one mutation was present at a frequency of approximately 0.06%, whereas each variant with two mutations was at a frequency of approximately 0.003%. However, three variants were unexpectedly overrepresented in the starting library, two with one mutation (A35C or A47G) and one with both mutations (each variant represented 1.8, 1.6, and 1.4% of the library, respectively). This bias was therefore observed for the base frequency per position. Indeed, the mutations A35C and A47G were present at the same frequency as their corresponding wildtype bases (Supplementary Figure S5). The frequency of the other mutations was ∼50–100 times lower than their corresponding wildtype base. Mutations of uracils were also two-times less frequent than mutations of the other bases. This information on the bias of the starting doped library was used to normalize the evolution of the base frequencies during selection using a method previously described by Carothers *et al.* (46). It clearly highlighted an increase in several base frequencies at various positions, mostly in the 5′ region (Figure 5A). Such increases should reveal the bases that play a key role in the binding of the aptamer. In contrast, the base frequencies for some positions, mostly in 3′ region, were mostly unchanged, suggesting that they are not crucial for the aptamer. When this information was added to a predicted structure of ACE4 it revealed that some predicted G-C base pairs (positions 24–57, 25–56, 29–52, 30–51 and 42–49), as well as some bases in a predicted bulge (A34, A35, C36, A39 and G40), appear to be crucial and do not tolerate mutation (Figure 5B). In contrast, the frequencies of five bases in a bulge and a loop increased much more than their original base, suggesting that these mutations (A32C, G33A, G38U, A44G and A47C) may be beneficial for the aptamer. Three of these mutations included the three beneficial mutations that were previously identified by the EGE tree. Two other mutations (A26G and C50U) were also enriched, although they were predicted to be engaged in base pairing. Accordingly, most of the 1,124 variants that displayed a better enrichment than ACE4 contained at least one of these mutations (Supplementary Table S5). We studied several variants that presented the highest amplification during the doped cell-SELEX and contain the mutation (A32C, A47C and C50U) (Figure 6 and Supplementary Table S5). Unexpectedly, competitive binding revealed that the variants A47C, A44G/C50U, and A32C/A44G did



	K_d (nM)	k_{off} (x10^{-3} s^{-1})	k_{on} (x10^5 s^{-1}. M^{-1})
Wild type ACE 4	4.92 ± 1.12	1.11	2.26
Ancestor (U63C)	5.48 ± 0.91	2.58	4.71
G33A	2.56 ± 0.5	0.29	1.14
G38U	3.51 ± 0.5		
A44G	3.25 ± 0.4		
U63A	3.54 ± 0.7		
C50U	7.54 ± 1.77		
A44G/C50U	7.24 ± 1.24		
A32C/A44G	5.19 ± 1.38		
A32C	5.41 ± 1.06		
G33A/A44G	2.92 ± 0.47		
A35C	No binding		

Figure 4. Binding of ACE4 variants to MCF7 cells overexpressing Annexin A2. (**A**) Competitive binding assays of ACE4 variants compared to wildtype ACE4. [^{32}P]-radiolabeled ACE4 aptamer was incubated in the presence of an unlabelled ACE4 variant or wildtype ACE4. The competitive binding ratio was calculated by dividing the amount of radioactive ACE4 bound in the presence of wildtype ACE4 by the amount of radioactive ACE4 bound in the presence of a variant (n = 3). The results are presented in log 2: the variants with higher competitive binding than ACE4 have positive values and those with lower competitive binding have negative values. (**B**) Equilibrium and kinetics constants for the binding of ACE4 variants to cells. Apparent K_ds were measured by saturation binding experiments (n = 3, see Supplementary Figure S3). Apparent dissociation rate constants (k_{off}) were measured by fitting the dissociation of aptamers from the cell surface measured by the Ligand Tracer instrument in (**C**) The apparent association rate constant (k_{on}) was calculated by dividing the apparent k_{off} by the apparent K_d previously measured during saturation binding experiments. (**C**) Association and dissociation of radioactive wildtype ACE4 and two variants (U63C and G33A) measured in real time by the Ligand Tracer instrument. Association and the dissociation were measured for 30 and 60 min, respectively. The fit of the dissociation allows calculation of the k_{off} presented in (**B**).

not have a higher affinity than ACE4, whereas they were much more enriched (Figure 4A and B). Only the A32C variant provides a slightly higher affinity than ACE4. However, its competitive binding was lower than that of the variants previously identified by the EGE tree. In contrast, the G33A/A44G variant, which also appeared in the two last rounds of the tree, had a much higher affinity. This result suggests that the EGE tree may be as effective as a doped-SELEX in identifying mutations that provide the highest affinity.

DISCUSSION

For more than half a century, *in vitro* selection has generated a high number of new peptides, proteins, and nucleic acids some of which are even used as medicines (1,2,4). However, the evolutionary processes engaged during this approach have remained largely unknown, although the rise and fall of isotypes during a selection was already observed since the first original experiments (7,51). Until recently, the population was only sequenced at the end of the experiment. In addition, a hundred or fewer sequences were usually analysed by Sanger sequencing, which provided a very low resolution snapshot of the population. HTS provides an opportunity to improve the resolution of this analysis, in an unprecedented way, by investigating millions of sequences. Most importantly, it also offers the ability to perform time-lapse analysis of the population during selection. Therefore, it can provide access, for the first time, to something only dreamed by evolutionary biologists: a large quantity of relics from precise moments of evolution. We used relic sequences from various rounds of *in vitro* selection to reconstruct an empirical genealogical evolutionary (EGE) tree that represents the evolution of several anti-annexin A2 aptamer variants. Our method only analyses sequences that are sampled above a certain threshold. As such, we hypothesize that the observed species in later rounds are more likely to arise from species whose frequencies are higher than this

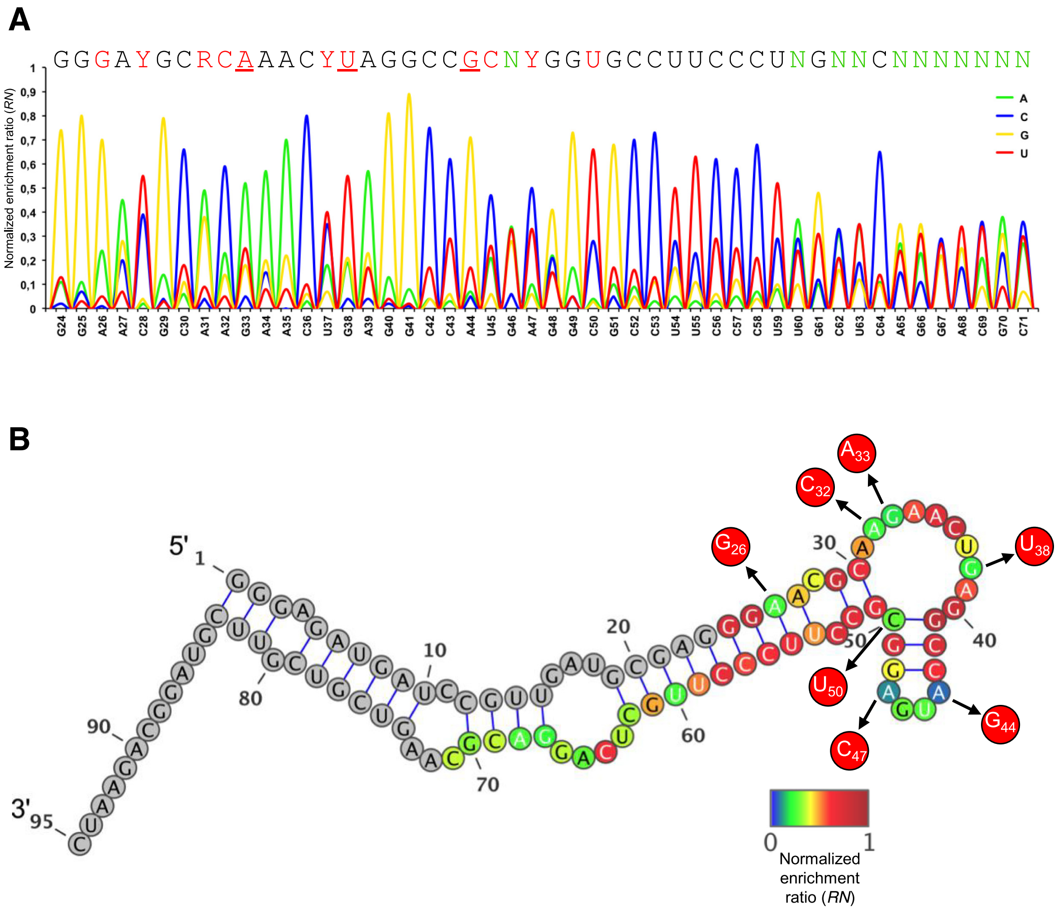


Figure 5. Evolution of nucleotide frequencies per position during doped-SELEX. (**A**) Normalized enrichment ratio (*RN*) for each base per position after four rounds of doped SELEX. *RN*s are presented as an artificial sequencing chromatogram. It emphasizes important positions of the aptamer for which some bases are predominantly amplified, suggesting that they play an important role in the interaction of the aptamer with its target. Each base that is predominantly amplified per position was used to compose the sequence at the top. Bases that were different from those of wildtype ACE4 are shown in red. Y designates pyrimidines, R purines and N (in green) positions for which it is difficult to observe the enrichment of a single base. (**B**) Structural prediction of the ACE4 aptamer showing bases that were predominantly amplified after four rounds of doped-SELEX. *RN*s were added as a colour scale to the predicted secondary structure of ACE4. Arrows indicate mutated bases that are more highly amplified than the original bases.

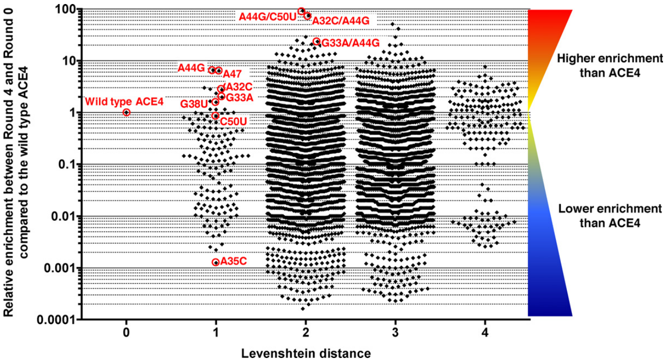


Figure 6. Mutational landscape of the ACE4 aptamer based on doped-SELEX. The relative enrichment of 5872 ACE4 variants relative to that of wildtype ACE4 was measured after four rounds of doped-SELEX (presented in log 10 in the y-axis). These variants contained up to four mutations relative to ACE4 (x-axis). Most variants were less enriched than ACE4, but 1124 variants were more enriched (relative enrichment superior to 1). Variants evaluated in the binding experiments are encircled in red.

limit of detection in prior rounds. We believe that this inference used by the EGE tree may be better for the analysis of *in vitro* molecular evolution processes than classical phylogenetic trees because it uses historical information on the abundance of each variant throughout the process, in addition to sequence alignment, to determine the family relationships between variants. In contrast, most classical phylogenetic analysis simply use sequence homology to infer family relationships. Furthermore, the EGE tree can also show the increase or extinction of individual sequences inside the population. Thus, it can reveal the evolutionary pathways taken by macromolecules by successive mutations to evolve towards variants with the best fitness to survive the selection pressure. It is impossible to obtain such information using a classical phylogenetic tree built from multiple sequence alignments (see Figure 3A).

Our study was inspired by the diagram published by Charles Darwin in 1859 to explain his theory of evolution (33). His diagram resembles a genealogy tree in which different hypothetical species with slight variations are depicted at different times of evolution and connected to their parents up to a hypothetical ancestor. Darwin used this diagram to show that all variations may not appear simultaneously, but rather by successive variations. Furthermore, he supposed that only the variations that are profitable would be preserved and generate new variations. Moreover, the improved variants may take the place of earlier, and less well adapted variants, leading to their extinction. All these patterns of evolution described for the evolution of animal species were directly observed at a molecular scale for the evolution of anti-annexin A2 aptamers. An ancestor of the aptamer was highly abundant in the earlier rounds of selection and generates several variants. One variant, the ACE4 aptamer, was much better adapted to the selection pressure and enriched in the population in the further rounds of *in vitro* selection. This led to the progressive extinction of the ancestor and other variants. At the same time, ACE4 generated several new variants, some of which were better suited to withstand the selection pressure that was increased experimentally by increasing the time of washing. Accordingly, the EGE tree showed that the uracil in position 63, in addition to the mutations G33A, A44G, or G38U, may provide variants with slower dissociation rates, which was confirmed by binding experiments. These mutations all concerned bases that are predicted to be in single-strand loops, suggesting that these loops may be important for the binding of the aptamer.

Previously, Hoinka et al proposed a method, called 'Aptamut', to highlight the best variants of an aptamer family calculating a Log score (24). We compared our results with this method (Supplementary Tables S6 and S7). One of the main differences is that Aptamut used a Locality sensitive hashing approach to cluster sequences. As a result, we observed that the clustering was never rigorously identical when we use this method several times. It identified about 690,000 different families while our approach reproducibly identified 1,887 different families. More importantly, Aptamut clustered ACE4 variants in several different families, which complicated the analysis. In addition, it takes in account sequences whose frequency is less than 0.001% whereas we demonstrated that they are subject to high measurement errors (Supplementary Figure S3). Focusing on the most abundant ACE4 family identified by Aptamut (Supplementary Table S6), several sequences with the highest Log score matched the wild-type ACE4 sequence with mutation in the primer regions. Some sequences also contained the G33A, A44G and G38U mutations. However, they have been diluted with several sequences containing mutations that were not enriched during the doped-SELEX, which raises questions about their relevance for improving binding (i. e. the mutations C57G, C64G, G66U, G67U, G67A, A68G or A68C). These results suggest that identifying the best variants could be easier and more reproducible with the EGE tree.

The evolution of the ACE4 aptamer reveals that *in vitro* selection is far more than just an improved screening method. Hence, several mathematical models have been developed to understand the possible evolution of a population during *in vitro* selection. Most of the mathematical models consider that the starting population contains the aptamer sequences and they do not take into account mutations (17). Based on this assumption, it has been proposed that aptamers could be selected in only one round of selection using a very stringent protocol. Although such selection has been successfully used on occasion, it usually leads to aptamers of lower affinity than those identified by classical SELEX. Our study suggests that the starting population may not contain the aptamer sequence that will be identified after several rounds of *in vitro* selection. Instead, the population may contain some sequences close to the final sequence, but with lower affinity. Therefore, several rounds of selection are likely necessary to select and mutate these sequences to obtain better aptamers in the population. This hypothesis may explain why selection is sometimes more or less rapid depending on whether the population contains sequences more or less close to the aptamer structure with the best fitness.

Favouring and controlling the mutation rate during the amplification steps seems therefore crucial for *in vitro* selection. HTS analysis is a powerful technique to investigate this parameter. Under our conditions, the mean mutation rate per PCR cycle was estimated by HTS to be ∼$1.7 \pm 0.6 \times 10^{-4}$ (an error every 5,882 incorporated nucleotides), consistent with the mutation rate calculated for Taq polymerases that are not improved to increase fidelity (52). After 24 cycles of PCR amplification of ACE4, only 83% of sequences were unmutated and 16% contained one mutation (approximately 0.1% for each sequence). Therefore, a variant with one mutation has a high chance of being generated during each round of *in vitro* selection, and PCR may be considered to generate a library of doped ACE4 aptamers that contains ∼99.6% of the original residues and 0.13% of each other residue at each position. In contrast, our starting library for the doped cell-SELEX contained many more variations and only 1.8% sequences corresponded to the wildtype sequence. The frequency of sequences with only one mutation was ∼0.05% for each and more than half sequences contained two to four mutations. This result could explain why the doped cell-SELEX did not succeed in generating better variants.

Our SELEX used the natural mutation rate of polymerases. However, mutagenic PCR has long been used during *in vitro* selection to explore the sequence space around species that survive the early selection cycles (6,32,53). We believe that the EGE tree could be particularly useful for analysing such experiments and evaluating how the mutation rate can improve SELEX. Indeed, Pressman *et al.* have already used HTS to study the *in vitro* selection of a ribozyme that was conducted using mutagenic PCR (32). They showed that, for half the identified families, sequences containing a few mutations, usually corresponding to a single nucleotide mutation, replaced the sequence with the highest abundance in the previous rounds. By focusing on four families, they demonstrated that three of these variants, which displayed the highest abundance in the last round, exhibited a higher activity than the sequences previously identified by cloning and Sanger sequencing analysis. We built an EGE tree for all these families that confirmed that these better ribozymes have a high enrichment in their respective family. However, the EGE tree allowed to identify beneficial mutations and to predict better variants that were not identified during this previous study (Supplementary Figure S6).

Finally, a chain of mountains, for which the peaks represent sequences with the best fitness, is often used to illustrate the theoretical fitness landscape of nucleic acids or proteins. Such a landscape is difficult to access because it would be necessary to know the fitness of all possible sequences. Nevertheless, an EGE tree can reconstruct the pathways that have been taken during *in vitro* selection to climb to the top of these mountains. Here, we demonstrated how it can improve the discovery of aptamers and allow a better understanding of the effect of selection pressures. However, it may be useful for other molecular evolution studies, such as those concerning ribozymes or proteins.

DATA AVAILABILITY

The method to build an EGE tree can be executed by following the instructions found on GitHub platform (https://github.com/AptaFred/EGE_tree). All fastq files are available on the European Nucleotide Archive (ENA) at http://www.ebi.ac.uk/ena/data/view/PRJEB22637.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

This work has benefited from the facilities and expertise of the high throughput sequencing core facility of I2BC (Centre de Recherche de Gif – http://www.i2bc-saclay.fr/). We also thank Noémie Ménétrot and Benoit Jego for technical assistance during the doped SELEX, as well as Larry Gold for kind suggestions to improve the manuscript. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

FUNDING

'Agence Nationale pour la Recherche' under the frame of EuroNanoMed (project META); Investissement d'Avenir [ANR-11-INBS-0011]—NeurATRIS: A Translational Research Infrastructure for Biotherapies in Neurosciences. N.N.Q. was supported by a Grant from LFB biotechnologies. Funding for open access charge: Laboratory budget.

Conflict of interest statement. A patent application was filed by one of the author (F.D.) covering the anti-Annexin A2 aptamers described in this paper and their use in diagnostic and therapeutic applications. This does not alter the authors' adherence to all *Nucleic Acids Research* policies on sharing data and materials.

REFERENCES

1. Mills,D.R., Peterson,R.L. and Spiegelman,S. (1967) An extracellular Darwinian experiment with a self-duplicating nucleic acid molecule. *Proc. Natl. Acad. Sci. U.S.A.*, **58**, 217–224.
2. Jijakli,K., Khraiwesh,B., Fu,W., Luo,L., Alzahmi,A., Koussa,J., Chaiboonchoe,A., Kirmizialtin,S., Yen,L. and Salehi-Ashtiani,K. (2016) The in vitro selection world. *Methods*, **106**, 1–13.
3. Robertson,M.P., Joyce and,G.F. (2012) The origins of the RNA World. *Cold Spring Harb. Perspect. Biol.*, **4**, a003608.
4. Cech,T.R. (2012) The RNA worlds in context. *Cold Spring Harb. Perspect. Biol.*, **4**, a006742.
5. Neveu,M., Kim,H.J. and Benner,S.A. (2013) The "strong" RNA world hypothesis: fifty years old. *Astrobiology*, **13**, 391–403.
6. Beaudry,A.A. and Joyce,G.F. (1992) Directed evolution of an RNA enzyme. *Science*, **257**, 635–641.
7. Tuerk,C. and Gold,L. (1990) Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science*, **249**, 505–510.
8. Jolma,A., Yan,J., Whitington,T., Toivonen,J., Nitta,K.R., Rastas,P., Morgunova,E., Enge,M., Taipale,M., Wei,G. *et al.* (2013) DNA-binding specificities of human transcription factors. *Cell*, **152**, 327–339.
9. Ellington,A.D. and Szostak,J.W. (1990) In vitro selection of RNA molecules that bind specific ligands. *Nature*, **346**, 818–822.
10. Berens,C., Groher,F. and Suess,B. (2015) RNA aptamers as genetic control devices: the potential of riboswitches as synthetic elements for regulating gene expression. *Biotechnol. J.*, **10**, 246–257.
11. Pfeiffer,F. and Mayer,G. (2016) Selection and biosensor application of aptamers for small molecules. *Front. Chem.*, **4**, 25.
12. Forier,C., Boschetti,E., Ouhammouch,M., Cibiel,A., Ducong e,F., Nogre,M., Tellier,M., Bataille,D., Bihoreau,N., Santambien,P. *et al.* (2017) DNA aptamer affinity ligands for highly selective purification of human plasma-related proteins from multiple sources. *J. Chromatogr. A*, **1489**, 39–50.
13. Berezovski,M.V., Lechmann,M., Musheev,M.U., Mak,T.W. and Krylov,S.N. (2008) Aptamer-Facilitated biomarker discovery (AptaBiD). *J. Am. Chem. Soc.*, **130**, 9137–9143.
14. Cibiel,A., Pestourie,C. and Duconge,F. (2012) In vivo uses of aptamers selected against cell surface biomarkers for therapy and molecular imaging. *Biochimie*, **94**, 1595–1606.
15. Zhou,J. and Rossi,J. (2016) Aptamers as targeted therapeutics: current potential and challenges. *Nat. Rev. Drug Discov.*, **16**, 440.
16. Cibiel,A., Dupont,D.M. and Duconge,F. (2011) Methods to identify aptamers against cell surface biomarkers. *Pharmaceuticals*, **4**, 1216–1235.
17. Spill,F., Weinstein,Z.B., Irani Shemirani,A., Ho,N., Desai,D. and Zaman,M.H. (2016) Controlling uncertainty in aptamer selection. *Proc. Natl. Acad. Sci. U.S.A.*, **113**, 12076–12081.
18. Bartel,D.P., Zapp,M.L., Green,M.R. and Szostak,J.W. (1991) HIV-1 rev regulation involves recognition of non-Watson-Crick base pairs in viral RNA. *Cell*, **67**, 529–536.
19. Autour,A., Westhof,E. and Ryckelynck,M. (2016) iSpinach: a fluorogenic RNA aptamer optimized for in vitro applications. *Nucleic Acids Res.*, **44**, 2491–2500.
20. Nguyen Quang,N., Perret,G. and Ducong e,F. (2016) Applications of High-Throughput sequencing for in vitro selection and characterization of aptamers. *Pharmaceuticals (Basel)*, **9**, E76.
21. Cho,M., Xiao,Y., Nie,J., Stewart,R., Csordas,A.T., Oh,S.S., Thomson,J.A. and Soh,H.T. (2010) Quantitative selection of DNA aptamers through microfluidic selection and high-throughput sequencing. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 15373–15378.
22. Schutze,T., Wilhelm,B., Greiner,N., Braun,H., Peter,F., Morl,M., Erdmann,V.A., Lehrach,H., Konthur,Z., Menger,M. *et al.* (2011) Probing the SELEX process with next-generation sequencing. *PLoS ONE*, **6**, e29604.
23. Cibiel,A., Quang,N.N., Gombert,K., Theze,B., Garofalakis,A. and Duconge,F. (2014) From ugly duckling to swan: unexpected identification from cell-SELEX of an anti-Annexin A2 aptamer targeting tumors. *PLoS One*, **9**, e87002.
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. Ditzler,M.A., Lange,M.J., Bose,D., Bottoms,C.A., Virkler,K.F., Sawyer,A.W., Whatley,A.S., Spollen,W., Givan,S.A. and Burke,D.H. (2013) High-throughput sequence analysis reveals structural diversity and improved potency among RNA inhibitors of HIV reverse transcriptase. *Nucleic Acids Res.*, **41**, 1873–1884.
26. Dupont,D.M., Larsen,N., Jensen,J.K., Andreasen,P.A. and Kjems,J. (2015) Characterisation of aptamer-target interactions by branched selection and high-throughput sequencing of SELEX pools. *Nucleic Acids Res.*, **43**, e139.
27. Wright,S. (1932) The roles of mutation, inbreeding, crossbreeding and selection in evolution. *Proc. Natl. Acad. Sci. U.S.A.*, **1**, 356–366.

28. Maynard Smith,J. (1970) Natural selection and the concept of a protein space. *Nature*, **225**, 563–564.
29. Pitt,J.N. and Ferré-D'Amaré,A.R. (2010) Rapid construction of empirical RNA fitness landscapes. *Science*, **330**, 376–379.
30. Pitt,J.N., Rajapakse,I. and Ferre-D'Amare,A.R. (2010) SEWAL: an open-source platform for next-generation sequence analysis and visualization. *Nucleic Acids Res.*, **38**, 7908–7915.
31. Jimenez,J.I., Xulvi-Brunet,R., Campbell,G.W., Turk-MacLeod,R. and Chen,I.A. (2013) Comprehensive experimental fitness landscape and evolutionary network for small RNA. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 14984–14989.
32. Pressman,A., Moretti,J.E., Campbell,G.W., Muller,U.F. and Chen,I.A. (2017) Analysis of in vitro evolution reveals the underlying distribution of catalytic activity among random sequences. *Nucleic Acids Res.*, **45**, 8167–8179.
33. Darwin,C. (1859) *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. John Murray, London.
34. Morlon,H. (2014) Phylogenetic approaches for studying diversification. *Ecol. Lett.*, **17**, 508–525.
35. Bharadwaj,A., Bydoun,M., Holloway,R. and Waisman,D. (2013) Annexin A2 heterotetramer: structure and function. *Int. J. Mol. Sci.*, **14**, 6259–6305.
36. Alam,K.K., Chang,J.L. and Burke,D.H. (2015) FASTAptamer: A bioinformatic toolkit for High-throughput sequence analysis of combinatorial selections. *Mol. Ther. Nucleic Acids*, **4**, e230.
37. Afgan,E., Baker,D., van den Beek,M., Blankenberg,D., Bouvier,D., Cech,M., Chilton,J., Clements,D., Coraor,N., Eberhard,C. *et al.* (2016) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res.*, **44**, W3–W10.
38. Kumar,S., Stecher,G. and Tamura,K. (2016) MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.*, **33**, 1870–1874.
39. Nei,M. and Kumar,S. (2000) *Molecular Evolution and Phylogenetics*. Oxford University Press, Inc., NY.
40. Shannon,P., Markiel,A., Ozier,O., Baliga,N.S., Wang,J.T., Ramage,D., Amin,N., Schwikowski,B. and Ideker,T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.
41. Sievers,F., Wilm,A., Dineen,D., Gibson,T.J., Karplus,K., Li,W., Lopez,R., McWilliam,H., Remmert,M., Soding,J. *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, **7**, 539.
42. Katoh,K., Kuma,K., Miyata,T. and Toh,H. (2005) Improvement in the accuracy of multiple sequence alignment program MAFFT. *Genome Inform. Int. Conf. Genome Informatics*, **16**, 22–33.
43. Hall,T.A. (1999) BioEdit: A User-Friendly biological sequence alignment editor and analysis program for Windows 95/98/NT,. *Nucleic Acids Symp. Ser.*, **41**, 95–98.
44. Quang,N.N., Miodek,A., Cibiel,A. and Duconge,F. (2017) Selection of aptamers against whole living cells: from cell-SELEX to identification of biomarkers. *Methods Mol. Biol.*, **1575**, 253–272.
45. Quang,N.N., Pestourie,C., Cibiel,A. and Ducongé,F. (2016) How to measure the affinity of aptamers for membrane proteins expressed on the surface of living adherent cells. *Methods*, **97**, 35–43.
46. Carothers,J.M., Oestreich,S.C., Davis,J.H. and Szostak,J.W. (2004) Informational complexity and functional activity of RNA structures. *J. Am. Chem. Soc.*, **126**, 5130–5137.
47. Darty,K., Denise,A. and Ponty,Y. (2009) VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics*, **25**, 1974–1975.
48. Zuker,M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, **31**, 3406–3415.
49. Nakamura,K., Oshima,T., Morimoto,T., Ikeda,S., Yoshikawa,H., Shiwa,Y., Ishikawa,S., Linak,M.C., Hirai,A., Takahashi,H. *et al.* (2011) Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res.*, **39**, e90.
50. Schirmer,M., Ijaz,U.Z., D'Amore,R., Hall,N., Sloan,W.T. and Quince,C. (2015) Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Res.*, **43**, e37.
51. Bartel,D.P. and Szostak,J.W. (1993) Isolation of new ribozymes from a large pool of random sequences [see comment]. *Science*, **261**, 1411–1418.
52. Lee,D.F., Lu,J., Chang,S., Loparo,J.J. and Xie,X.S. (2016) Mapping DNA polymerase errors by single-molecule sequencing. *Nucleic Acids Res.*, **44**, e118.
53. Joyce,G.F. and Inoue,T. (1989) A novel technique for the rapid preparation of mutant RNAs. *Nucleic Acids Res.*, **17**, 711–722.