A reproducible approach for the use of aptamer libraries for the identification of Aptamarkers for brain amyloid deposition based on plasma analysis.

Cathal Meehan[1], Soizic Lecocq[1] and Gregory Penner[1]

(1) NeoNeuro SAS, Villejuif Bio Park, 1 Mail du Professeur Georges Mathé, 94800, Villejuif, France

Corresponding Author:

Gregory Penner

NeoNeuro SAS

Villejuif Biopark

1 mail du Georges Mathé

94800 Villejuif, France

gpenner@neoneuro-aptamers.com

+1 519 835-8677

Keywords:

Alzheimer's disease, biomarkers, Aptamarkers, Neomer library

20

21 **Classification:**

22 Biological Sciences, Neuroscience

Abstract

An approach for the agnostic identification and validation of aptamers for the prediction of a medical state from plasma analysis is presented in application to a key risk factor for Alzheimer's disease, brain amyloid deposition. This method involved the use of a newly designed aptamer library with sixteen random nucleotides interspersed with fixed sequences called a Neomer library. The Neomer library approach enables the direct application of the same starting library on multiple plasma samples, without the requirement for pre-enrichment associated with the traditional approach. Eight aptamers were identified as a result of the selection process and screened across 390 plasma samples from the same 400 individuals by qPCR assay. Results from both processes were analysed using multiple machine learning algorithms from the Scikit-learn package along with clinical variables including cognitive status, age and sex to create predictive models. An Extra Trees Classifier model provided the highest predictive power. The Neomer approach resulted in a sensitivity of 0.74, specificity of 0.76, and AUC of 0.78. The only clinical variable that remained included in the model was age. We conclude that the Neomer approach represents a clear improvement for the agnostic identification of aptamers (Aptamarkers) that bind to unknown biomarkers of a medical state.

Introduction

The diagnosis of complex diseases, including dementia, psychiatric disorders, auto-immune disorders, and gastro-intestinal diseases, remains a significant challenge when relying on blood-based approaches. Various "omics" approaches, such as genomics, transcriptomics, proteomics

44 and metabolomics have encountered limitations in addressing this issue. Firstly, genomics and
45 transcriptomics-based strategies struggle to provide accurate predictions for complex diseases
46 due to the intricate nature of genetic risk factors. The complexity arises from the involvement of
47 numerous genetic variants, each with a modest individual effect, having positive or negative
48 correlations with the condition, making it challenging to pinpoint causative factors or establish
49 predictive models. Additionally, transcriptomics fails to correlate gene expression from these
50 individual genetic factors to disease onset or progression due to technical variation introduced
51 during sample preparation, library construction, and sequencing[1]. Both of these approaches are
52 limited by the gap between genetic variation, transcript abundance and protein abundance as a
53 function of translation. Neither of these approaches provide information about molecular
54 complexes or post-translational modifications.

55 Further downstream, proteomics and metabolomics, pose another set of challenges. These
56 platforms confront substantial variability, both within individuals due to dynamic fluctuations in
57 protein levels or metabolite concentrations over time, and across different subjects (inter-
58 individual) due to genetic diversity, lifestyle, or environmental factors[2]. Proteomics and
59 metabolomics are also both limited in their ability to detect complexes and to quantify the
60 relative abundance of molecules. Uncharacterized forms of both metabolites and proteins, as
61 well as protein complexes, further limit the ability of these approaches to form predictive models
62 based on current scientific literature. Proteins and metabolites in life often exist in complexes,
63 many proteins exist and function as members of non-covalently bound homo or hetero-mers of
64 varying composition, while metabolites due to their inherent hydrophobicity often exist in the
65 blood stream as non-covalent complexes with protein. The existence of complexes is not directly
66 measured by either proteomics or metabolomics, it is at best inferred.

There is clearly a need for an agnostic approach to biomarker identification that is not constrained by these limitations. The concept of using aptamer libraries for the agnostic discovery of biomarkers for predicting disease states has been attempted by ourselves[3,4] and others but was also limited by the following constraints.

1.) The need to immobilize target material in order to parse aptamers that bind from aptamers that do not bind.
2.) The need for reiterative selection cycles in order to enrich aptamer sequences from single unique copies in the naïve library. The reiterative selection process implicitly introduces PCR bias into the selection process and is a cause for variation between samples.
3.) The inability to apply the same set of sequences in a naïve library to different biofluid or tissue samples.

In this paper we will describe how we have combined innovative approaches that overcome these constraints with an aptamer library application for the diagnosis of brain amyloid deposition (a risk factor for Alzheimer's disease) in blood.

We have overcome the need to immobilize targets in order to separate aptamers in a library that bind from those that do not with an approach called FRELEX[5]. This approach relies on immobilized antisense oligonucleotides on gold nanoparticles to compete with targets in blood for the binding of all aptamers in a library. A proportion of each aptamer sequence that does not bind to a target will hybridize to the immobilized antisense and can be removed with centrifugation. That portion of any aptamer sequence that binds to a target will remain in the supernatant and is recovered and submitted for next generation sequencing (NGS) analysis. The more any given aptamer is bound to targets in selection the higher it's relative frequency will be in NGS analysis.

90 We overcame the latter two constraints listed above, the need for reiterative selection and the
91 inability to apply the same aptamer sequences to different biological samples by reinventing
92 aptamer selection. A key constraint with the traditional SELEX method is the contiguous nature
93 of the random region[6]. This implicitly restricts structural diversity resulting in a need to design
94 libraries with large numbers of random nucleotides. The standard SELEX library design involves
95 40 contiguous nucleotides (nt) in the random region. This represents a total of 1.2E24 possible
96 sequences. In practical terms the upper limit in regard to the number of sequences that can
97 practically be applied to a sample is 1E15 to 1E16. This means that it is not practical to apply
98 aptamer libraries in selection that contain multiple copies of each possible sequence. It is also not
99 possible to apply the same library of sequences to different targets. To overcome this constraint
100 we designed an aptamer library with 16 random nt interspersed with fixed sequences, as
101 described in Figure 1. The key to this design is that there is minimal capacity for hybridization
102 among the fixed sequences. With 16 random nt the total number of possible sequences is 4.29E9.
103 This approach enables the use of 4E12 sequences such that there is an average of 1,000 copies of
104 each possible sequence in the naïve library, and the same naïve library sequences can be applied
105 to different biological samples. We refer to this novel design of an aptamer library as a neomer
106 library.

107 The next issue to be solved was how to characterize the frequency of all 4.29E9 sequences cost-
108 effectively in NGS analysis. Our solution was to include a restriction site in the middle of the
109 library sequence. The presence of this restriction site drives the only occurrence within the fixed
110 sequences of self-hybridization. After selection, the library is amplified and then this restriction
111 site is used to cut the library into two modules. Each module contains eight random nucleotides
112 representing a total of 65,536 possible sequences. Each module is prepared separately for NGS

113 analysis and in a single NGS run it is possible to determine the frequency of each of all possible
114 sequences.

115 The frequency of each of the original 4.29E9 possible sequences is obtained for each selection by
116 multiplying the frequencies of each of the possible 65,536 sequences in one module by the
117 frequencies in the other module through the use of an outer product matrix. This mathematical
118 approach is synonymous to the concept of a Punnett square used in genetics.

119 We refer to the aptamers within these libraries used for agnostic selection as Aptamarkers. We
120 applied this neomer approach to identify Aptamarkers for brain amyloid deposition in plasma
121 with 390 individuals from the Australian Imaging, Biomarker and Lifestyle (AIBL) cohort[7,8].

122 Alzheimer's disease is a neurodegenerative process that is thought to be caused by a three-step
123 process[9]. The first step is the accumulation of amyloid plaques in the brain, followed by Tau
124 tangles within neurons and finally neuronal death. The presence of high brain amyloid is
125 considered by clinicians in combination with memory tests in their definition of the disease in
126 patients presenting memory symptoms. The use of positron emission tomography (PET) scans or
127 the evaluation of Abeta 42 to Abeta 40 peptide ratios in cerebrospinal fluid[10] are the current gold
128 standards for brain amyloid diagnosis. Unfortunately, both diagnostic procedures are invasive
129 and expensive. The need for a simple blood test as a first line screen has been suggested as
130 desirable by several sources[11, 12].

131 We applied a neomer library to plasma samples from twenty individuals (ten high brain amyloid,
132 ten low brain amyloid) balanced for other clinical variables and characterized the frequencies of
133 all 4.29E9 sequences across all samples by NGS analysis. We identified four Aptamarkers that
134 were preferentially enriched on blood samples from individuals with high brain amyloid

135 deposition (HAM) and four sequences that were preferentially enriched on blood samples from
136 four individuals with low brain amyloid (LAM). We designed specific primer sequences for each
137 Aptamarker such that we could amplify them individually from a pool. The pool of eight
138 Aptamarkers were applied to 10 uL of plasma from 390 samples (AIBL cohort) in the presence
139 of antisense oligonucleotides (Figure 2) with the capacity to hybridize to them. The antisense
140 oligonucleotides were immobilized on gold nanoparticles and that portion of the Aptamarker
141 pool that hybridized to them was removed with a centrifugation step. The remaining supernatant
142 was used as a template for qPCR analysis.

143 Prior to processing the samples we predefined 291 of the samples as training samples and 99 as
144 test samples. Machine learning (ML) analysis of the qPCR results and clinical variables was
145 carried out using a variety of different algorithms on the training samples. ML models that were
146 assessed for sensitivity, specificity and area under the receiving operator curve (AUROC) and
147 were tested for predictive capacity on the test samples.

148 ML analysis of the Neomer libraries with eight Aptamarkers and age resulted in an Extra Trees
149 Classifier model with a sensitivity of 74%, specificity of 76%, overall accuracy of 75% and an
150 area under the AUROC curve of 0.78. The predictive power of the Neomer/Aptamarkers did not
151 include cognitive status or sex as clinical variables, indicating that this was not a significant
152 criteria for this test. A key advantage of the method as described is the agnostic nature of the
153 analysis. This same approach can be applied to develop Aptamarkers for any disease. The
154 method can also be applied to detect differences in blood related to response to treatment, or
155 predisposition to side effects.

156

Materials and Methods

Neomer selection and predictive modelling

The ssDNA Neomer library used for selection was composed of a 16 random interspersed by fixed sequences.

5' CCAGATACAGACNNGAGGNNNGAATNNNAACCATCGGCGCCAACANNNCATTCNNN CAGANNCAGTAGACAGC 3' (IDT, Belgium).

This template was selected to limit the contiguous nature of traditional aptamer libraries and to maximise the structural diversity of the aptamers. A prediction of the secondary structure formed by this library is provided in Figure 1 based on the use of the RNAfold program[13] as visualized with Forna structure software[14].

Figure 1: Predicted secondary structure of fixed sequences in Neomer library.

We also conjugated single stranded (ss) DNA antisense sequences (/5ThioMC6-D/AAAACAAAGTCTACTTGTTGGTTCTGTAT) with homology to fixed sequences in the neomer library on 40 nm Gold Nanoparticles (GNPs – Cytodiagnostics). This antisense was designed to have complementary binding to fixed sequences in the Neomer library. The thiolated antisense was reduced with TCEP following IDT specifications. Reduced antisense was then conjugated with GNPs following supplier specifications and dissolved in 200 µL 0,1X of phosphate buffer saline (PBS) (8.0 mM Na2PO4, 1.4 mM KH2PO4, 136 mM NaCl, 2.7 mM

179 KCl, pH 7.4). This concentration of the antisense in solution on the GNPs was measured with a
180 NanoDrop instrument in order to ensure that an antisense concentration of 0.24 μM was used for
181 each sample.

182 In the first step of FRELEX employed in this study, 2.57E13 sequences from the neomer aptamer
183 library were snap cooled by heating the library to 95 °C for 10 min followed by immediate
184 immersion in ice bath. These libraries were then incubated with individual plasma sample in 50
185 μL of Selection Buffer (10 mM Tris, 120 mM NaCl, 5 mM MgCl2, 5 mM KCl) for 15 min at RT.
186 This solution was then applied to the functionalized GNPs for 15 min at RT. The solution was
187 centrifuged for 20 min at 3500 g. The supernatant was removed and purified using the
188 Oligonucleotide Clean-up protocol of the Monarch PCR & DNA Clean-up kit, as described by
189 NEB and eluted with 400 μL of de-ionised water.

190 **Figure 2:** Schematic representation of the application of neomer aptamers to plasma samples.

191 After this selection, PCR amplification was used to double strand the library (dsDNA) for an
192 appropriate number of cycles to create a clear band of approximately 5 ng of amplified DNA. All
193 PCR procedures were carried out according to standard molecular biology protocols and under
194 the following conditions: 95°C for 5 min, 4 cycles at 95°C for 10 s, 35°C for 15s, 72°C for 30
195 sec, and then remaining cycles at 95°C for 10 s, 64°C for 15s and 72°C for 30 sec, followed by a
196 final extension at 72°C for 5 min. Table 2 provides the primer sequences used for all steps
197 described here.

198 **Table 1:** Sequences of primers used to process Neomer libraries for next generation sequencing
199 analysis.

Full Library Fwd	5’ CAA ATA CGT ATG AGG TCG CTC GTT CCC AGA TAC AGA C 3’
Full library rvs	5’ TAA TAC GAC TCA CTA TAG GGA TAA TGC TGT CTA CTG 3’
NGS-1A Fwd	5’ CCC TAC ACG ACG CTC TTC CGA TCT ATC ACG CAA ATA CGT ATG AGG TCG CTC GTT C 3’
NGS-1A Rvs	5’ GGT CAG ACG TGT GCT CTT CCG ATC GGG GCG CCG ATG GTT 3’
NGS-1B Fwd	5’ CCC TAC ACG ACG CTC TTC CGA TCT ATC ACG GCG CCA ACA 3’
NGS-1B Rbs	5’ GGT CAG ACG TGT GCT CTT CCG ATC GGG TAA TAC GAC TCA CTA TAG GGA TAA TGC TGT CTA CTG 3’

The underlined sequence in Table 1 is a hex code that was used for library parsing from the fasta file. The sequence shown is only one example, we used 21 different hex codes for this analysis. We used the same hex code for module A and module B in all cases.

It is important to maintain the average expected copy of each sequence at a level above 100 copies throughout the NGS preparation procedure. To ensure this we performed the Oligonucleotide Clean-up protocol of the Monarch PCR & DNA Clean-up kit (NEB) on 10 PCR reaction tubes. The 40 µL volume of elution was used as a template for restriction with KasI enzyme (NEB) at 37°C for 60 min followed by 20 min at 65°C to obtain two distinct modules (A and B). The two modules were treated in the same way, but separately for Next Generation Sequencing (NGS) preparation. The restricted products were diluted 10X prior to next step.

211 The amplified product from NGS preparation was purified using the Monarch PCR & DNA
212 Clean-Up kit prior to be sequenced. In addition, we processed a buffer sample containing no
213 blood in each analysis, as a further test.

214 All libraries were sequenced using Illumina NovaSeq at the TCAG facility (Hospital for Sick
215 Children, Toronto, Canada). Fastq files obtained upon sequencing were converted into a Fasta
216 format, and the copy number of each of the possible 65,536 sequences in each of Module A and
217 Module B were characterized using a proprietary Python script that we have developed for this
218 application. The frequency of each Module sequence was determined by dividing copy number
219 by the total number of reads for that Module. The frequency of each of the 4.29 billion sequences
220 in the original selected library was then estimated through an outer product calculation of the
221 Module A and Module B frequencies.

222 The average frequency for each of these 4.29 billion sequences was determined for the ten high
223 amyloid samples and the ten low amyloid samples. The standard deviation of each of these
224 averages was also determined. The average of the low amyloid samples for each sequence was
225 subtracted from the average for the same sequence across the high amyloid samples. This value
226 was divided by the average of the standard deviation from both sets of samples to generate a Z
227 score. The top 10,000 sequences in terms of Z score were identified.

228 We then identified the top four sequences within these 10,000 based on the separation of all the
229 ten high samples from all the ten low samples.

230 We also ran this analysis by subtracting the average of each sequence with the high amyloid
231 samples from the average of the same sequence averaged across the low amyloid samples. The

top 10,000 sequences based on Z scores was also identified as low amyloid sequences and the top four identified based on separation of all ten low samples from all ten high samples.

The sequences of these eight Aptamarkers is provided in Table 2. Specific reverse primers were designed for each of these sequences and tested to ensure that in conjunction with the same forward primer (GTT CCC AGA TAC AGA C) in each case they specifically amplified the desired sequence.

Table 2: Sequences of Aptamarkers identified by the Neomer process and specific reverse primer sequences for their amplification.

Name of the aptamer	Aptamer sequence	Sequence of the specific reverse primer associated
HAM_2753	CCA GAT ACA GAC TCG AGG ACT GAA TCG GAA CCA TCG GCG CCA ACA AGA CAT TCA TAC AGA AAC AGT AGA CAG C	TTT CTG TAT GAA TGT CT
HAM_6700	CCA GAT ACA GAC TCG AGG TAA GAA TGG GAA CCA TCG GCG CCA ACA CAG CAT TCA ATC AGA GAC AGT AGA CAG C	TCT CTG ATT GAA TGC TG
HAM_6968	CCA GAT ACA GAC CAG AGG CGT GAA TTT CAA CCA TCG GCG CCA ACA CCC CAT TCC TGC AGA TAC AGT AGA CAG C	TAT CTG CAG GAA TGG GG

HAM_8505	CCA GAT ACA GAC AGG AGG GCG GAA TTC CAA CCA TCG GCG CCA ACA TTT CAT TCT AAC AGA CAC AGT AGA CAG C	TGT CTG TTA GAA TGA AA
C-LAM_1	CCA GAT ACA GAC TAG AGG TAT GAA TAG AAA CCA TCG GCG CCA ACA ATT CAT TCA TTC AGA TGC AGT AGA CAG C	CAT CTG AAT GAA TGA AT
C-LAM_168	CCA GAT ACA GAC AAG AGG ACC GAA TGT CAA CCA TCG GCG CCA ACA TTA CAT TCA TTC AGA TTC AGT AGA CAG C	AAT CTG AAT GAA TGT AA
C-LAM_262	CCA GAT ACA GAC CAG AGG TTT GAA TTC CAA CCA TCG GCG CCA ACA TTA CAT TCC TCC AGA ATC AGT AGA CAG C	ATT CTG GAG GAA TGT AA
C-LAM_2709	CCA GAT ACA GAC TTG AGG TTC GAA TCT CAA CCA TCG GCG CCA ACA TGA CAT TCC TTC AGA TTC AGT AGA CAG C	AAT CTG AAG GAA TGT CA

Screening of Aptamarkers

Each sample was analyzed in two well qPCR replicates. Following qPCR analysis, the data was reviewed according to quality control standards with a coefficient of variance of 6 between replicates, only data passing this standard was retained. The raw fluorescence data was analyzed to identify the point at which fluorescence increased to six times the base level. The base level

was determined in the first five cycles. The point between the PCR cycles was determined by linear regression analysis. This provided a Cq value. An efficiency value (E) for each sample was determined based on identification of the linear portion of the amplification curve above the Cq value.

The average of all Cq and E values within a given qPCR experiment were calculated. This average was divided by a fixed value to define a normalization factor for each qPCR run. Both the Cq and E values were divided by this normalization factor by qPCR set.

The Cq and E values were averaged across replicates and the relative amount of template (RT) was then determined with the following formula - 1/E^Cq

Each individual RT value was divided by the sum of all eight RT values per sample to define a proportion of the RT value within each sample (RTp). This value was used for model building.

Dataset Preprocessing

The dataset model comprised 390 samples, with variables for all 8 aptamers and the same 3 clinical variables. Both were preprocessed for zero-mean scaling, to make them suitable for ML analyses.

Feature Selection Using Default Parameters

For each classification algorithm, a model with default parameters was applied to calculate the importance of individual features. The algorithms from the scitkit-learn package[15] and others were used include Logistic Regression, Decision Trees, Random Forests, Gradient Boosting, k-Nearest Neighbors, Support Vector Machines, Gaussian Naive Bayes, Multi-Layer Perceptron,

XGBoost[16], Ridge Classifier, and ExtraTrees. Based on the importance scores obtained, features were ranked. In a sequential approach, the least significant features were excluded, thus generating reduced training datasets.

Hyperparameter Tuning

For each iteration of feature selection, hyperparameter tuning was conducted. A predefined dictionary of hyperparameter grids was set for each algorithm under evaluation. The GridSearchCV technique, complemented by a 5-fold cross-validation, was used to pinpoint the optimal hyperparameters for each model, with accuracy serving as the optimization criterion.

Model Evaluation

Models were subsequently trained on the reduced datasets obtained post-feature selection. During each model iteration, evaluative metrics such as accuracy, sensitivity, specificity, and AUROC were gauged.

Results

The ssDNA Neomer library used in our selection was a 16-base random sequence interspersed with fixed sequences. This library was designed to optimise secondary structure diversity through the placement of regions of random sequence and by the lack of capacity of the fixed sequences to hybridize with each other (Figure 1). This means that structural diversity is entirely driven by the random nucleotides.

285 During NGS preparation, it was crucial to maintain an average expected sequence copy above
286 100. The amplified products for NGS were purified before sequencing. A buffer sample was
287 incorporated in each analysis as a control. Following sequencing, a proprietary Python script was
288 employed to determine the frequency of every sequence in Modules A and B. The frequency of
289 each of the 4.29 billion sequences in the selected library was deduced through an outer product
290 calculation.

291 From the analysis, we discerned the average frequency for all sequences across ten high and low
292 amyloid samples. Z scores were then derived by considering the difference in averages and
293 standard deviations in these sample sets. The top 10,000 sequences based on Z scores were
294 identified and further refined to the top four sequences offering the best separation across sample
295 sets when performing contrasts between high and low brain amyloid, and low and high brain
296 amyloid deposition (HAM and LAM aptamers).

297 **Figure 3:** The distribution of Z and fold values for a) the high amyloid versus low amyloid
298 contrast and b) the low amyloid versus high amyloid contrast. The fold values are simply the
299 difference between the high and low average minus unity. The low versus high graph is presented
300 with the opposite sign of the high versus low amyloid contrast.

301 It is interesting to note that the fold differences were higher with the high versus low amyloid
302 contrast as compared to the low versus high contrast. The higher fold value for the high amyloid
303 aptamers may indicate that these will be more predictive than the low amyloid ones.

304 **Figure 4:** Comparative analysis illustrating the stratification of samples utilizing the top 4 of
305 high-amyloid and low-amyloid binding aptamers. Each point corresponds to an individual
306 sample.

307 In the qPCR analysis of Aptamarkers, we established a strict quality threshold with a coefficient
308 of variance of 6 between replicates. Using the fluorescence data, we computed the Cq and
309 efficiency values (E) for each sample, subsequently determining a normalization factor. Relative
310 amounts of template (RT) were calculated, and from these, proportions of RT values per sample
311 (RTp) were derived, forming the basis for model building. We used all eight Aptamarkers as
312 variables and the clinical variables sex, clinical status (mild cognitive impairment (MCI) or
313 cognitively normal (CN)), and age. We did not include ApoE genetic status in our primary model
314 based on the understanding that this variable is not commonly applied in clinical practice in
315 Europe.

316 The dataset was preprocessed to zero-mean scaling, catering to ML applications. Feature
317 selection employed default parameters from various algorithms, sequentially excluding less
318 significant features. With each iteration, hyperparameter tuning was performed using the
319 GridSearchCV technique and a 5-fold cross-validation. The models, trained on the reduced
320 datasets, were evaluated based on multiple performance metrics. Table 4 provides a summary of
321 all algorithms that returned perfect predictive results on the test set.

322

323 **Table 4:** Top models based on analysis of training set samples using a neomer approach.

Algorithm	Variables Excluded	Accuracy	Sensitivity	Specificity	AUROC
DecisionTreeClassifier	Rvs HAM6700, Clinical Classification, Sex	1	1	1	1
ExtraTreesClassifier	Clinical Classification	1	1	1	1
ExtraTreesClassifier	Sex, Clinical Classification	1	1	1	1
ExtraTreesClassifier	Rvs c-LAM262, Sex, Clinical Classification	1	1	1	1
RandomForestClassifier		1	1	1	1
RandomForestClassifier	Sex	1	1	1	1
RandomForestClassifier	Clinical Classification, Sex	1	1	1	1
RandomForestClassifier	Rvs HAM6700, Clinical Classification, Sex	1	1	1	1
XGBClassifier		1	1	1	1
XGBClassifier	Sex	1	1	1	1
XGBClassifier	Rvs HAM6700, Sex	1	1	1	1
XGBClassifier	Rvs c-LAM262, Rvs HAM6700, Sex	1	1	1	1

The ML models that exhibited the highest fit with the observed data are all similar in nature basically being based on multiple decision trees. We chose the Extra Trees Classifier model for application to the test set of data given that this model performed consistently well on subsets of the data prior to the completion of the 391 samples.

This model performed best on the test set when clinical classification and sex were not included with a sensitivity of 0.74, specificity of 0.76, overall accuracy of 0.75 and AUROC of 0.78.

The inclusion of ApoE status resulted in an increase in sensitivity to 0.78, a reduction in specificity to 0.72 and a maintenance of overall accuracy of 0.75. The AUROC was also unchanged at 0.78. We interpret this result as meaning that the Aptamarkers are implicitly providing the predictive power of the ApoE allelic status. Figure 5 provides a description of the relative weight of each of the variables in the model.

Figure 5: Feature importance attributable to each variable in the predictive model.

It is intriguing that the Aptamarkers all exhibit relatively similar predictive capacity with each individual Aptamarker contributing about half the value of age as a variable. The sum of all the Aptamarker values represents more than 75% of the predictive capacity of the model. The receiver operator response analysis is provided in Figure 6.

Figure 6: Receiver operator response

The area under the curve of the receiver operator AUROC analysis provides a description of the expected robustness of the test. In this case, that value is 0.78.

Conclusions

The application of the Neomer library approach addresses many limitations of the traditional SELEX method and results in the identification of Aptamarkers with high predictive capacity. The Aptamarker approach clearly results in high predictive capacity for brain amyloid deposition from plasma analysis.

The agnostic nature of this approach means that this should be applicable to any disease.

364 # List of Abbreviations

365 AIBL - Australian imaging, biomarker and lifestyle

366 AUROC – Area under receiver operator curve

367 CN – Cognitively normal

368 E – Efficiency value

369 GNPs - Gold nanoparticles

370 LAM - Low brain amyloid deposition

371 HAM - High brain amyloid deposition

372 IDT - Integrated DNA technologies

373 MCI – Mild cognitive impairment

374 ML – Machine learning

375 NGS – Next generation sequencing analysis

376 NT - Nucleotide

377 PCR – Polymerase chain reaction

378 PET - Positron emission tomography

379 qPCR – Quantitative polymerase chain reaction

380 RT – Relative amount of template

381 SELEX - Systematic evolution of ligands by exponential enrichment

Acknowledgments

Ethics

All data utilized in this study have been anonymized at the source. We did not directly collect this information; therefore, the personal data we accessed cannot be used to identify specific individuals.

Availability of Data

The data underlying the figures and tables presented in this study are available at the following DOI: X .

Competing Interests

The research was funded by NeoVentures Biotechnology Ltd., NeoNeuro and the Alzheimer's Drug Discovery Foundation. The authors, including Cathal Meehan, Soizic Lecocq, and Gregory Penner, are paid employees of NeoNeuro.

Funding

The research was funded by NeoVentures Biotechnology Ltd., NeoNeuro and the Alzheimer's Drug Discovery Foundation.

Author's contributions

Cathal Meehan: Responsible for designing the propriety python script, statistical methods, and machine learning analysis. Assisted in writing and revising the manuscript.

Soizic Lecocq: Led the blood sample management, selection, qPCR analysis, aptamer design, and antisense design. Assisted in writing and revising the manuscript.

Gregory Penner: Conceived and designed the experiments, coordinating research in both wet and dry lab activities. Assisted in writing and revising the manuscript.

All authors were involved in reviewing the manuscript and approved the final version.

References

1. McIntyre LM, Lopiano KK, Morse AM, Amin V, Oberg AL, Young LJ, Nuzhdin SV. RNA-seq: technical variability and sampling. BMC genomics. 2011 Dec;12:1-3.
2. Perng W, Aslibekyan S. Find the needle in the haystack, then find it again: replication and validation in the 'omics era. Metabolites. 2020 Jul 12;10(7):286.

419 3. Lecocq, S., Spinella, K., Dubois, B., Lista, S., Hampel, H. and Penner, G., 2018.
420 Aptamers as biomarkers for neurological disorders. Proof of concept in transgenic
421 mice. *PLoS One*, *13*(1), p.e0190212.
422 4. Penner G, Lecocq S, Chopin A, Vedoya X, Lista S, Vergallo A, Cavedo E, Lejeune FX,
423 Dubois B, Hampel H, INSIGHT-preAD study group. Aptamarker prediction of brain
424 amyloid-β status in cognitively normal individuals at risk for Alzheimer's disease. Plos
425 one. 2021 Jan 4;16(1):e0243902.
426 5. Penner G, inventor; NeoVentures Biotechnology Inc, assignee. Method for the selection
427 of aptamers for unbound targets. United States patent US 10,415,034. 2019 Sep 17.
428 6. Tuerk C, Gold L. Systematic evolution of ligands by exponential enrichment: RNA
429 ligands to bacteriophage T4 DNA polymerase. Science, 1990; 249:505-510.
430 7. Ellis KA, Bush AI, Darby D, De Fazio D, Foster J, Hudson P, Lautenschlager NT, Lenzo
431 N, Martins RN, Maruff P, Masters C. The Australian Imaging, Biomarkers and Lifestyle
432 (AIBL) study of aging: methodology and baseline characteristics of 1112 individuals
433 recruited for a longitudinal study of Alzheimer's disease. International psychogeriatrics.
434 2009 Aug;21(4):672-87.
435 8. Collie A, Maruff P, Darby DG, Masters C, Currie J. The Melbourne aging study. In Mild
436 Cognitive Impairment 2013 May 13 (pp. 98-118). Psychology Press.
437 9. Jack Jr CR, Bennett DA, Blennow K, Carrillo MC, Dunn B, Haeberlein SB, Holtzman
438 DM, Jagust W, Jessen F, Karlawish J, Liu E. NIA-AA research framework: toward a
439 biological definition of Alzheimer's disease. Alzheimer's & Dementia. 2018 Apr
440 1;14(4):535-62.

10. Reimand, J., Groot, C., Teunissen, C.E., Windhorst, A.D., Boellaard, R., Barkhof, F., Nazarenko, S., van der Flier, W.M., van Berckel, B.N.M., Scheltens, P. and Ossenkoppele, R., 2019. Why Is Amyloid-β PET Requested After Performing Cerebrospinal Fluid Biomarkers?. Journal of Alzheimer's Disease.
11. Li, Y., Schindler, S.E., Bollinger, J.G., Ovod, V., Mawuenyega, K.G., Weiner, M.W., Shaw, L.M., Masters, C.L., Fowler, C.J., Trojanowski, J.Q. and Korecka, M., 2022. Validation of plasma amyloid-β 42/40 for detecting Alzheimer disease amyloid plaques. Neurology, 98(7), pp.e688-e699.
12. Zetterberg H, Schott JM. Blood biomarkers for Alzheimer's disease and related disorders. Acta Neurol Scand. 2022;146(1):51-55.
13. Gruber AR, Lorenz R, Bernhart SH, Neuböck R, Hofacker IL. The vienna RNA websuite. Nucleic acids research. 2008 Apr 19;36(suppl_2):W70-4.
14. Kerpedjiev P, Hammer S, Hofacker IL. Forna (force-directed RNA): Simple and effective online RNA secondary structure diagrams. Bioinformatics. 2015 Oct 15;31(20):3377-9.
15. Kramer O. Scikit-learn. Machine learning for evolution strategies. 2016:45-53.
16. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining 2016 Aug 13 (pp. 785-794).

Note: Patent applications have been submitted globally describing both the Neomer application as well as the Aptamarker approach.

Figure 1: Predicted secondary structure of fixed sequences in Neomer library.

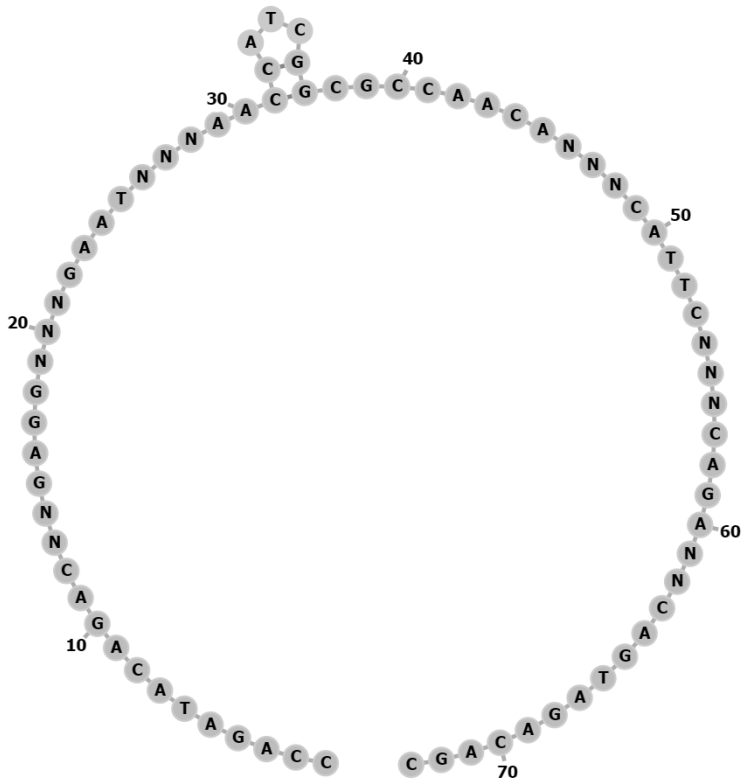


Figure 2: Schematic overview of Neomer library analysis.

Legend: A, processing of library for NGS analysis. B, Outer product matrix of all 4.29 billion possible sequences. C, statistical analysis of aptamer frequencies across samples. The distribution of the fixed sequences and random sequences is more complex than shown here but is clearly described in the sequence provided elsewhere.

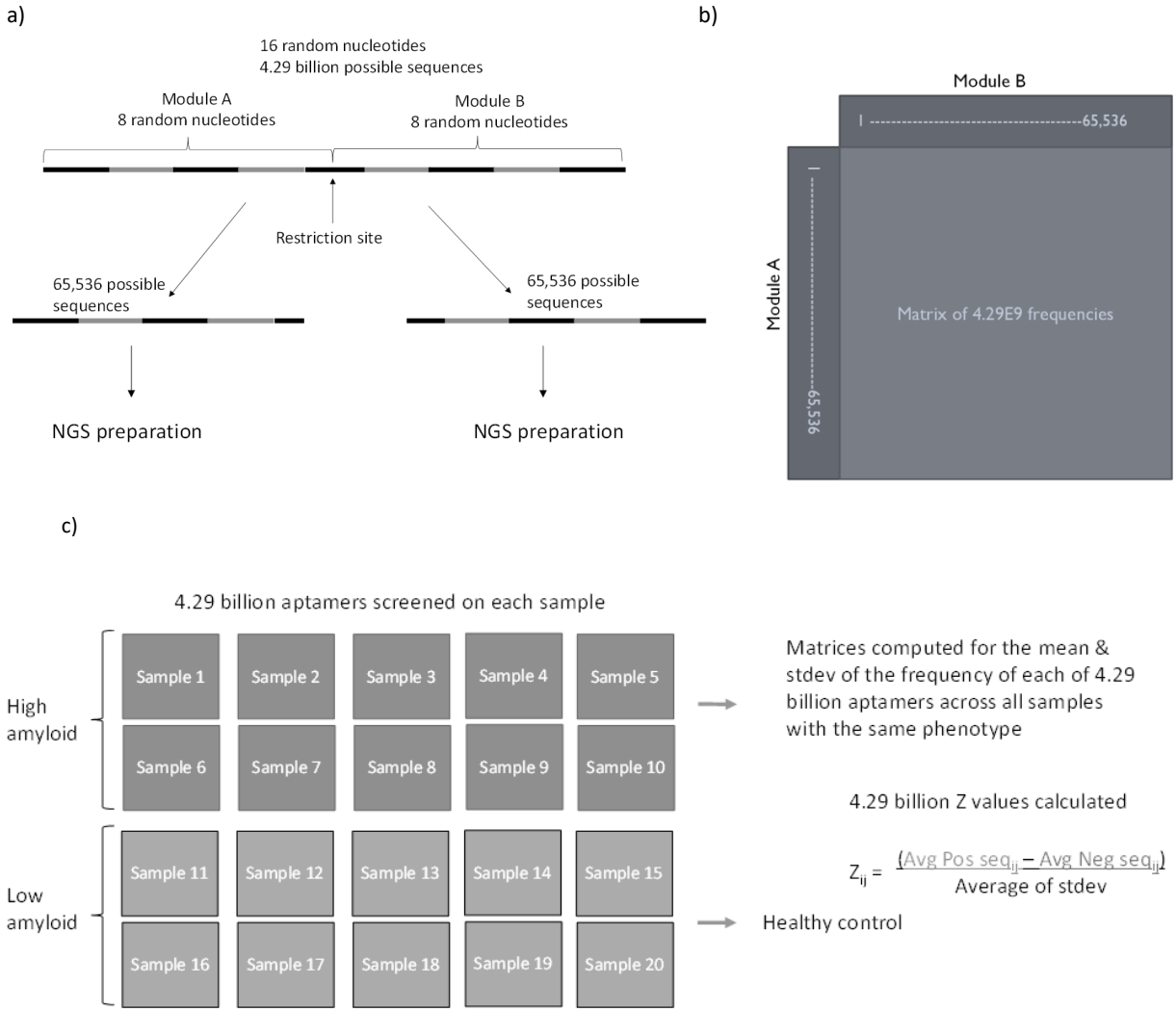
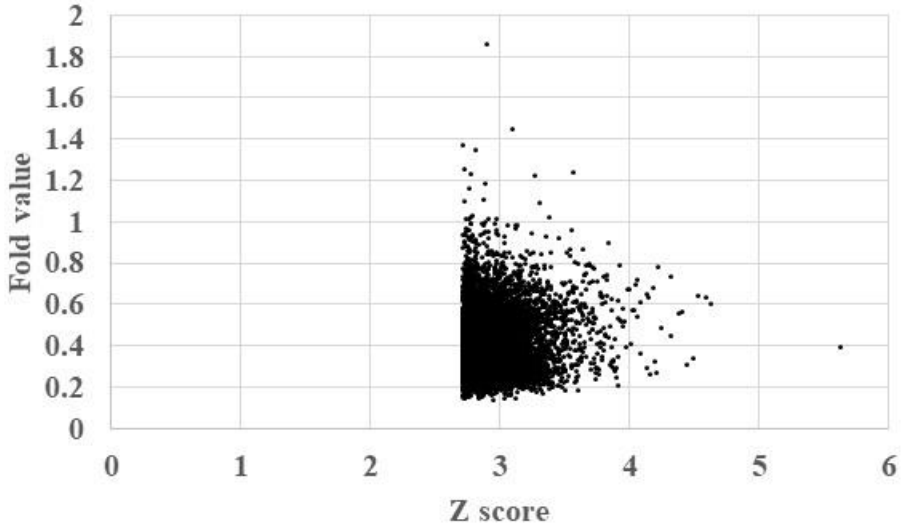


Figure 3: a and b for the high versus low and low versus high amyloid contrasts respectively.

A(



b)

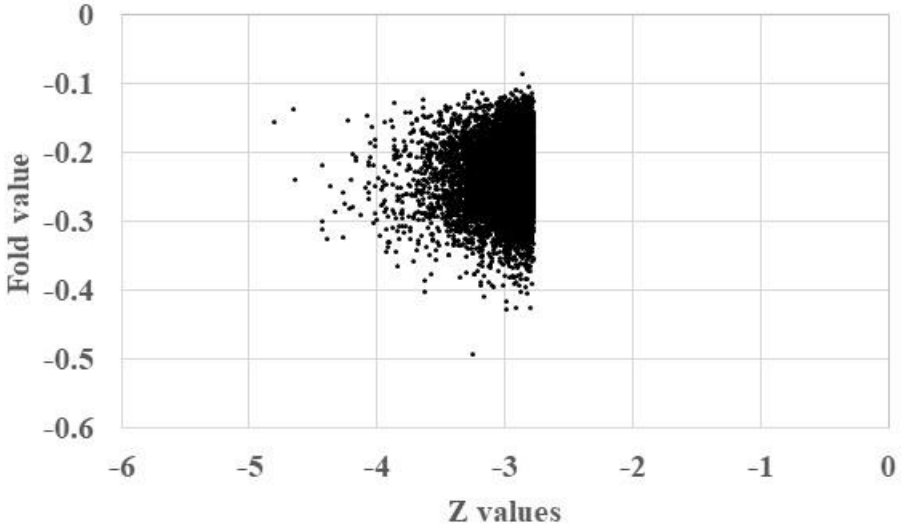


Figure 4: Separation of high and low samples by the four best Aptamarkers for a) high amyloid and b) low amyloid.

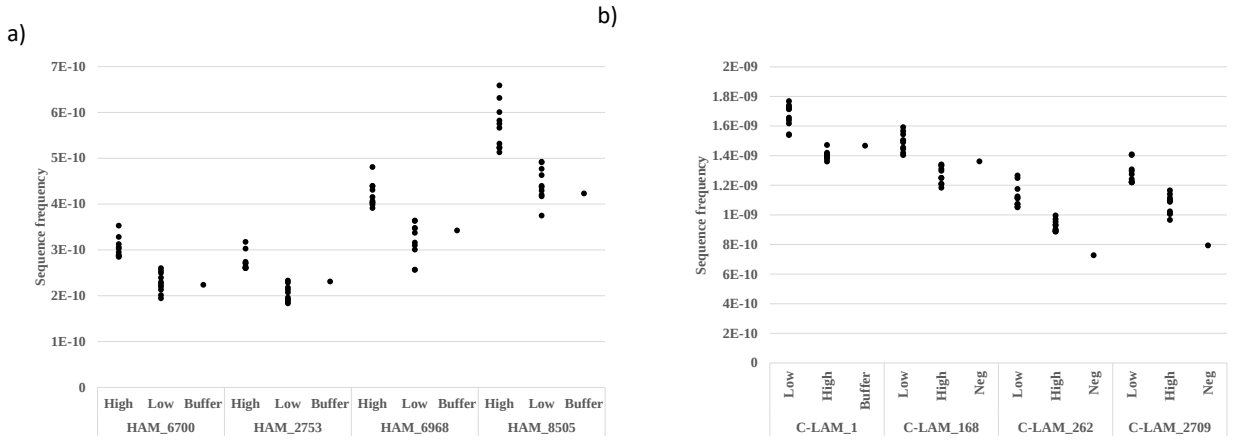


Figure 5: Weight attributable to each variable in predictive model.

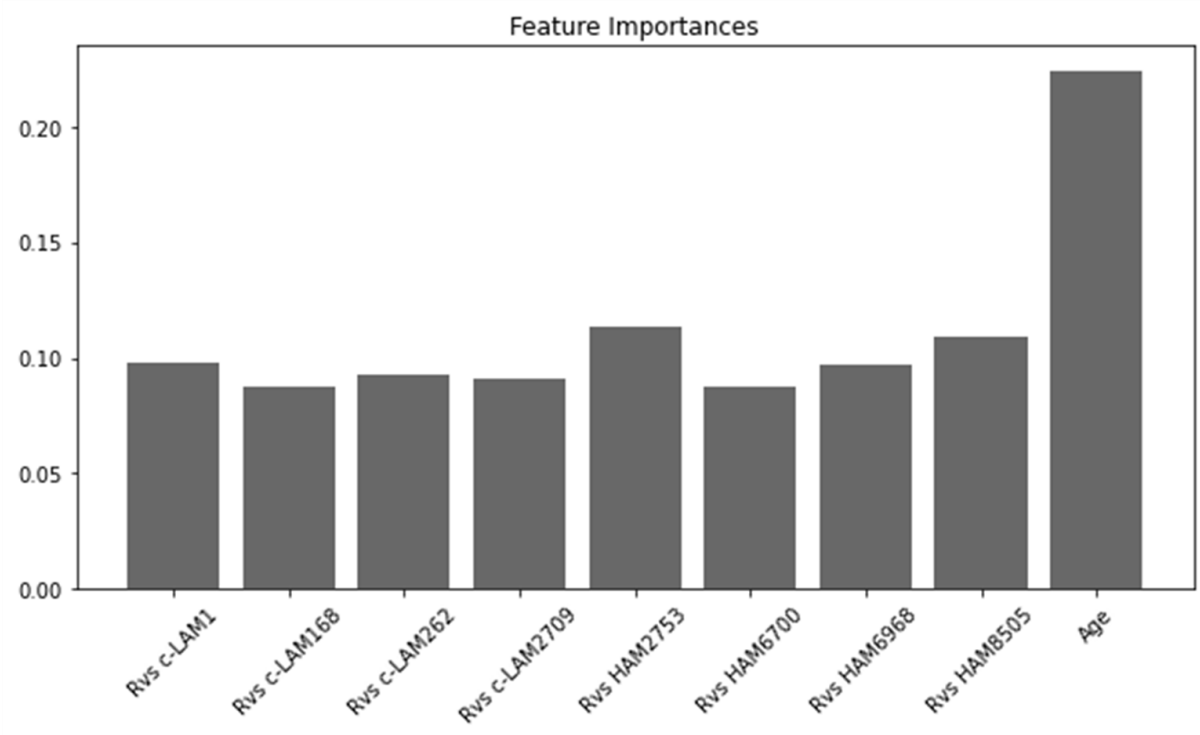
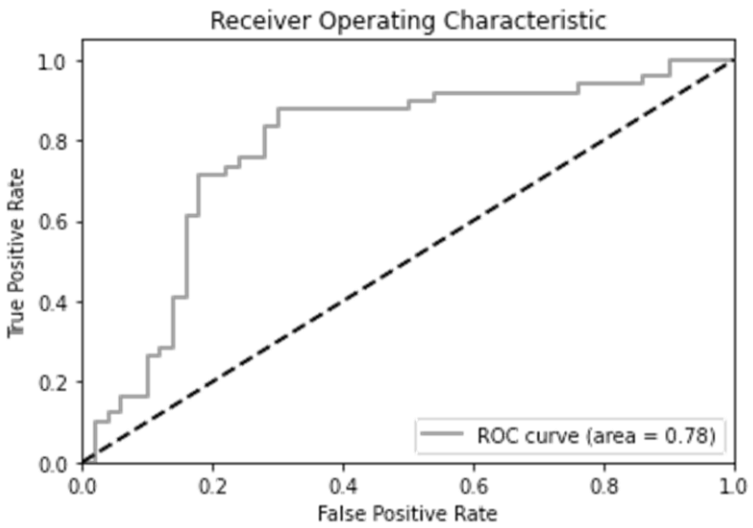



Figure 6: Receiver operator response