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:

A comparison of two approaches 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. One method involved the use of a traditional aptamer library with a contiguous random region of 40 nucleotides (SELEX), the other 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. For the SELEX approach, 24 aptamers were identified as a result of the selection process and screened across plasma samples from 400 individuals using a qPCR assay. For the Neomer 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 analysis. Results from both processes were analyzed with the Scikit-Learn machine learning library along with clinical variables including cognitive status, age and sex to create predictive models. In both cases 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, while the SELEX approach resulted in a sensitivity of 0.55, a specificity of 0.77, and AUC of 0.75. The only clinical variable that remained included in the Neomer model was age, while both age and cognitive status (MCI versus CN) were included in the SELEX model. 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:

Alzheimer's disease is a neurodegenerative process that is thought to be caused by three-step process[1]. The first step is the accumulation of amyloid plaques in the brain, followed by Tau tangles within neurons and finally neuronal death. As such, the presence of high brain amyloid is considered by clinicians in combination with memory tests in their definition of the disease in patients presenting memory symptoms. The use of positron emission tomography (PET) scans or the evaluation of Abeta 42 to Abeta 40 peptide ratios in cerebrospinal fluid[2] are the current gold standards for brain amyloid diagnosis. Unfortunately, both of these diagnostic procedures are invasive and expensive. The need for a simple blood test as a first line screen has been suggested as desirable by several sources[3, 4].

Aptamer selection has traditionally been performed by variations of a method called SELEX (Systematic Evolution of Ligands by EXponential enrichment)[5]. With a SELEX based aptamer development approach, it is necessary to use a random region of 40 nucleotides (nt) in order to create a library of sequences with sufficient structural diversity. This need arises given the contiguous nature of the bases and implicit physical constraints thereof, and the potential for folding back of random sequences on the defined primer recognition sequences. With 40 random nucleotides the number of possible sequences is $4^{40} = 1.2 \text{ X } 10^{24}$. This number of sequences would have a mass greater than a kilogram and thus be difficult to work with effectively in a laboratory setting. As a result, it is common practice to sub-sample the synthesized library and start with a pool of $1.0 \text{ X } 10^{15}$ sequences. This represents only $8.27 \text{ X } 10^{-10}$ coverage of the original library. There are two implications of this extremely low level of sampling of all possible sequences.

1.) It is unlikely that any given sequence will be sampled more than once, thus the average copy number of each unique sequence in the library will be one.
2.) It is unlikely that the same sequences will be sampled in different samplings of the original library. Thus, each starting sample of the library for selection will be comprised of entirely different sequences.

As such SELEX is not a reproducible method.

It has been our intent to develop a method for the use of aptamer libraries to identify unknown biomarkers in biofluids derived from individuals that differ for the presence of a medical state. To enable agnostic selection of aptamers against unknown targets we utilized a method (FRELEX) that we had invented for the selection of small molecules that obviates the need for target molecule immobilization[6]. The ability to partition aptamers that bind to a target from those that do not, enables agnostic selection for binding without a need for knowledge of the targets that are bound. The application of the FRELEX selection process to aptamer selection in plasma, results in the enrichment of aptamers that bind to all possible targets in biological fluid samples. To identify aptamers that bind to targets (biomarkers) that differ between biological fluids as a function of a medical state requires assaying the proportion of individual aptamer bound across samples derived from multiple individuals that differ for the medical state.

We have previously described the use of the FRELEX selection method with aptamer libraries to agnostically identify aptamers (Aptamarkers) that bind to unknown molecules in plasma (biomarkers) that are predictive of a disease[7,8]. In this study we extend the use of FRELEX to a larger data set and compare performance of this approach for predicting brain amyloid deposition based on blood analysis to a novel approach with a reproducible aptamer library.

The AIBL cohort[9] kindly provided us with samples from 400 individuals for this study.

To overcome the problem of different sequences in every aliquot of a naïve SELEX library we first enriched a single SELEX library with positive selection against a pool of samples with high brain amyloid and counter selection against a pool of samples with low brain amyloid. After ten rounds of selection on these pools aptamers that bind to biomarkers that differ between the pools should be enriched in copy number, and aptamers that do not bind to biomarkers should be eliminated, or at least be present at very low copy number. This enriched library was then applied to individual samples to further characterize aptamers that bound to targets that differed between samples as a function of the medical state in question. This approach fulfilled the necessary assumption that the library being applied to each sample was reasonably similar prior to selection.

A total of 24 aptamers were identified as being maximally predictive of brain amyloid based on t tests across high and low brain amyloid samples. These aptamers were applied to plasma of 400 individuals (300 training, 100 testing) and their level of binding to unknown targets was quantified by qPCR analysis. The assignment of samples to the training or testing sets was balanced for a defined set of clinical variables. The qPCR data from all 24 Aptamarkers along with the clinical variables age, sex and cognitive status (mild cognitive impairment (MCI) versus cognitively normal (CN)) was analyzed with the use of the scikit-learn machine learning library This resulted in the identification of an Extra Trees Classifier model for the best accuracy on the training set. This model resulted in prediction values with the test set of samples of 55% sensitivity, 77% specificity, an overall accuracy of 67% and an area under the response curve of 0.75.

111 We were able to develop a model with a sensitivity of 95%, specificity of 83%, and an AUROC
112 of 0.954 for prediction when we limited analysis to the 69 MCI individuals within the 400
113 samples analyzed. The remaining 331 were cognitively normal.

114 This approach suffered from certain drawbacks, including;

115 1.) Potential for bias in the original pool selection by individuals with large differences in
116 target abundance unrelated to the medical state.
117 2.) A substantial reduction in the number of unique sequences being assayed against
118 individual samples due to the pool selection enrichment process. In effect, only
119 approximately 1,000 sequences had a sufficiently high copy number in the enriched
120 library to be useful in the analysis across individual samples.
121 3.) Potential for selection process bias (primarily PCR amplifiability) as a result of the
122 multiple rounds of pool selection.

123 To improve the predictive power arising from the SELEX approach from this study we applied a
124 novel aptamer library design and selection process that we have termed the Neomer platform. A
125 Neomer library is designed such that there are sixteen random nucleotides interspersed by fixed
126 known sequences. The library was designed such that there was minimal potential for
127 hybridization among the fixed sequences. This means that structural diversity in the library is
128 driven primarily by the identity of the random sequences. The interspersing of the random
129 sequences with fixed sequences reduces the physical constraint on structural diversity that is
130 implicit with contiguous random nucleotides.

131 Given that there are sixteen random nucleotides the library consists of 4^{16} possible sequences or
132 approximately 4.29 billion (4.29×10^9). We applied an average of 6,000 copies of each sequence

133 in this library to twenty plasma samples from the AIBL cohort that were characterized for brain
134 amyloid deposition based on positron emission tomography (PET) scans. Ten samples with high
135 brain amyloid and ten samples with low brain amyloid were used in a single round of selection.
136 Prior to NGS analysis each aptamer sequence was amplified and divided into two modules
137 through the use of a restriction enzyme to a site in the middle of the sequence. Each module of
138 the selected aptamer was processed for NGS analysis separately. Each module contained eight
139 random nucleotides thus there was 65,536 possible unique sequences per module. The frequency
140 of each of these unique sequences was determined based on module specific NGS analysis. The
141 frequency of each possible full length selected aptamer sequence was determined based on the
142 multiplication of the two module sequence frequencies with an outer product matrix.

143 This approach enabled the characterization of the effect of each individual sample selection on
144 the frequency of each of the 4.29 billion aptamer sequences in the library.

145 Four sequences were selected as maximally discriminative for high amyloid prediction (Ham)
146 and four sequences selected as maximally discriminative for low amyloid prediction (Lam).
147 These aptamers are now referred to as Aptamarkers. We designed specific primer sequences for
148 each aptamer such that we could amplify them individually from a pool. The pool of eight
149 Aptamarkers were applied to 10 uL of plasma from 390 samples (AIBL cohort) in the presence
150 of antisense oligonucleotides (Figure 2) with the capacity to hybridize to them. The antisense
151 oligonucleotides were immobilized on gold nanoparticles and that portion of the Aptamarker
152 pool that hybridized to them was removed with a centrifugation step. The remaining supernatant
153 was used as a template for qPCR analysis.

154 **Figure 2:** Schematic overview of Neomer library analysis.

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

159

160 Prior to processing the samples as described, we maintained the labeling of 291 of the samples
161 from the 300 used for SELEX analysis as training samples and 99 of the 100 used for SELEX
162 analysis as testing samples. The reduction in the number of samples was due to insufficient
163 material remaining of these samples. The qPCR results in association with clinical variables were
164 applied to Scikit-Learn machine learning library on the training sample alone. Those models that
165 provided the best areas under the ROC on the training samples were then tested for predictive
166 capacity on the test set.

167 The application of the Neomer libraries with eight Aptamarkers resulted in an Extra Trees
168 Classifier model with a sensitivity of 74%, specificity of 76%, overall accuracy of 75% and an
169 area under the ROC curve of 0.78, with a model containing all eight Aptamarkers as variables
170 and age.

171 The application of the Neomer library approach to Aptamarker development successfully
172 overcame the constraints previously encountered with a SELEX approach. The predictive power
173 of the Aptamarkers derived from the Neomer approach alone was higher than that achieved with
174 the Aptamarkers derived with the SELEX approach. The predictive power of the
175 Neomer/Aptamarkers also did not include cognitive status as a clinical variable, indicating that
176 this was not a significant criteria for this test. A key advantage of the method as described is the

agnostic nature of the analysis. This same approach can be applied to develop Aptamarkers for any disease. The method can also be applied to detect differences in blood related to response to treatment, or predisposition to side effects.

Materials and Methods:

SELEX/FRELEX selection and predictive model building.

In the introduction, we described the use of the SELEX method for Aptamarker development. We refer to this as SELEX even though we use a FRELEX approach because the basis for the selection process is identical. FRELEX is a means of performing SELEX that does not require immobilization of the target molecules.

Population Sample

The human samples used in this study were obtained from the Australian Imaging, Biomarker & Lifestyle Flagship Study of Ageing (AIBL) cohort, which can be accessed at https://aibl.csiro.au/. The recruitment process for the cohort, which included assessments of neuropsychological functioning, lifestyle factors, and mood, has been described in detail previously[9]. In summary, a total of 1,166 participants with a baseline age of over 60 years were initially recruited. However, 54 participants were later excluded due to comorbid disorders or withdrawal of consent. To ensure accurate and consistent diagnoses among participants, a clinical review panel utilized the international criteria for Alzheimer's disease (AD) diagnosis by the National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer's Disease and Related Disorders Association (NINCDS-ARDA), as well as the

198 symptomatic predementia phase criteria for mild cognitive impairment (MCI) diagnosis. In this
199 particular investigation, the focus is on 400 individuals (200 male and 200 female), of whom
200 16% had mild cognitive impairment and 84% were cognitively normal, who completed the full
201 study assessment. Machine learning techniques were applied to 300 of these samples, while the
202 remaining 100 samples were used for evaluation of the predictive power of the machine learning
203 algorithms. The AIBL study received ethics approval from the institutional ethics committees of
204 Austin Health, St. Vincent's Health, Hollywood Private Hospital, and Edith Cowan University.
205 Prior to their inclusion in the study, all volunteers provided written informed consent. None of
206 the authors of this paper had access to or will have access to the identity of the patients enrolled
207 in the study, only coded numbers for sets of clinical data were received.

208 **Selection of Aptamarkers**

209 Depletion was performed with the Albumin/IgG Removal Kit (Pierce) according to the
210 manufacturer's recommendations on the samples used for selection. Aliquots (10 µL) of each of
211 the pooled samples were placed in spin columns containing 80 µL of gel slurry. Each column
212 was capped, vortexed, and incubated for 30 minutes at room temperature on an orbital shaker.
213 The column was then centrifuged at 1,000 × g for 2 minutes, and the filtrate was retained and
214 used subset analysis.

215 NeoNeuro SAS (France) was kindly permitted to use the FRELEX selection platform (patent
216 application: WO 2017035666 A1) for this study by NeoVentures Biotechnology Inc. (Canada)
217 for the selection of aptamer libraries. FRELEX requires the preparation of an immobilization
218 field consisting of a gold chip surface (approximately 1 cm^2) that was coated with random eight
219 nucleotide DNA oligonucleotides (immobilized through a thiol bond). The thiolated random

220 oligos was reduced with TCEP following IDT specifications. Reduced random oligos were then
221 conjugated to the gold surface.

222 The library was composed as follows.

223 5' TGTGCTATCGTTCGACTCTATCC (N40) GCTCATCGACTGTACCC 3'

224 This was amplified with the use of a forward primer, 5' TGTGCTATCGTTCGACTCTATCC 3'
225 and a reverse primer, 5' GGGTACAGTCGATGAGC 3'. For the selection of aptamer libraries.
226 FRELEX requires the preparation of an immobilization field consisting of a gold chip coated
227 with thiolated random 8 base pair DNA oligonucleotides. The 8-mer thiolated random
228 oligonucleotides were dissolved in 50 µL of phosphate buffer saline (PBS) (8.0 mM Na2PO4,
229 1.4 mM KH2PO4, 136 mM NaCl, 2.7 mM KCl, pH 7.4) at a concentration of 10 µM. This
230 solution was incubated at room temperature (RT) for 1 hour on gold surface chip, dimensions 7 x
231 10 x 0.3 mm (Xantec, Germany). The chip was then air-dried and 50 µL of a solution containing
232 thiol terminated polyethylene glycol (SH-PEG) molecules and incubated for 30 min at RT with
233 gentle shaking. This step blocks any remaining gold surface that is not covered with 8mers. SH-
234 PEG was subsequently added a second time for 16 hours. After that, the SH-PEG solution was
235 removed from the chip and the functionalized gold chip surface was washed with de-ionised
236 water and air-dried.

237 In the first step of FRELEX employed in this study, 10^{16} sequences from the random aptamer
238 library described previously were snap cooled by heating the library to 95 °C for 10 min
239 followed by immediate immersion in ice bath. These single stranded (ss) DNA sequences were
240 incubated with the functionalized immobilization field (gold chip with 8mers) in 50 µL of
241 Selection Buffer (10 mM Tris, 120 mM NaCl, 5 mM MgCl2, 5 mM KCl) for 30 min at RT. The

242 remaining solution was removed and discarded. This removes sequences that have too much
243 secondary structure to enable binding to the 8mers on the surface. The immobilization field was
244 washed twice with 50 µL of 10X TE buffer and the remaining bound oligonucleotides were
245 eluted and recovered with two incubations of 15 min each 1 mL of Selection Buffer at 95 °C.
246 These elutions were pooled and purified using the GeneJET PCR Purification Kit (Thermo
247 Fisher Scientific, Germany) as described by the manufacturer and eluted with 25 µL of de-
248 ionised water.

249 This aptamer library selected for capacity to bind to 8mers was then used for positive selection
250 with a pool of blood serum from positive amyloid in a total volume of 50 µL 1X Selection
251 Buffer. This solution was incubated for 15 min then incubated with the immobilization field for
252 15 minutes at RT. The remaining solution was recovered carefully and stored in a fresh tube. The
253 immobilization field was washed twice with 50 µL of 1X selection buffer with each wash being
254 collected and pooled with the solution removed in the first step. This solution contains sequences
255 that did not bind to the immobilization, presumably because they are bound to some moieties
256 within the blood serum instead. This pooled solution was purified as described for the phase I
257 step, eluted in 400 µL and subjected to PCR amplification for an appropriate number of cycles to
258 create a clear band of approximately 5 ng of amplified DNA.

259 After the first round of selection, PCR was used to amplify the selected ssDNA into double
260 stranded DNA (dsDNA) for an appropriate number of cycles to create a clear band of
261 approximately 5 ng of amplified DNA. All PCR procedures were carried out according to
262 standard molecular biology protocols and under the following conditions: 95°C for 5 min, 4
263 cycles at 95°C for 10 s, 52°C for 15s and 72°C for 30 sec, then x cycles at 95°C for 10 s, 68°C
264 for 15s and 72°C for 30 sec, followed by a final extension at 72°C for 5 min. Then the dsDNA is

265 used as a template for in vitro transcription to obtain RNA. The RNA was treated with DNase I,
266 purified with RNeasy Minelute cleanup (Qiagen). It was then used as the template in an RT to
267 obtain cDNA. The cDNA-RNA mixture was treated with RNase H according to manufacturer
268 instructions. The cDNA was purified and used as template for the next round of selection.

269 From the second selection round, selection were performed in the same manner with the
270 exception that a pool of blood serum from healthy patients was added in the negative selection
271 phase (phase I of FRELEX), where we select for aptamers that exhibit the capacity to bind to the
272 immobilization field. The inclusion of healthy blood serum enables the selection against
273 aptamers that bind to targets that are present within this material. This creates a contrast in which
274 we are effectively selecting aptamers that bind to epitopes that are enriched in the late human AD
275 blood pool. This approach was maintained for a total of ten rounds of selection.

276 After 10 rounds of selection the enriched library was applied to each of the individual blood
277 samples used for selection for a further three rounds of selection. Each of these three rounds of
278 selection as well as the enriched library from selection round 10 on the pooled samples were
279 prepared and analyzed by next generation sequencing.

280 The following sequences were chosen from this analysis for validation with qPCR analysis.

281 **Table 1:** Sequences chosen for validation with qPCR analysis

Sequence Name	Sequence Aptamer	Sequence Rvs Primer corresponding	Concentration

Aibl021	TGT GCT ATC GTT CGA CTC TAT CCC CGA AAA TCC GAT AAT CCT ACT AAC CTA GGG ACA CAC GAG GCT CAT CGA CTG TAC CC	CCT CGT GTG TCC CTA GGT TA	6,6 pM
Aibl038	TGT GCT ATC GTT CGA CTC TAT CCT CAA ACT AGA ACC GCT TAA AAG CTT CCT TGA ATA AAC CTA GCT CAT CGA CTG TAC CC	GGA AGC TTT TAA GCG GTT CTA GT	6,6 pM
Aibl042	TGT GCT ATC GTT CGA CTC TAT CCG CAA ATC ATT CAA AGA GGA AAG AGC CAA CAT GAG ATC TGT GCT CAT CGA CTG TAC CC	ACA GAT CTC ATG TTG GCT CTT	6,6 pM
Aibl110	TGT GCT ATC GTT CGA CTC TAT CCG TGG ACG TAC ACG CAG ATG GAG TAA GAC CCA TGT TCT TCA GCT CAT CGA CTG TAC CC	GCT GAA GAA CAT GGG TCT TAC T	6,6 pM
Aibl132	TGT GCT ATC GTT CGA CTC TAT CCA AAT CCA TAG ATA GAA CTT ATC GAA ACG ACC AAG GCT TAC GCT CAT CGA CTG TAC CC	TAA GCC TTG GTC GTT TCG AT	6,6 pM
Aibl143	TGT GCT ATC GTT CGA CTC TAT CCC GCG AAA ACC ACG ACA AAT CAA ACA TAG ACA GAA AAC CGA GCT CAT CGA CTG TAC CC	ATG AGC TCG GTT TTC TGT CT	6,6 pM
Aibl160	TGT GCT ATC GTT CGA CTC TAT CCA ACA ACG TAA CAT ATT TGT TCT TCT TAC CGA AAA TCA TCC GCT CAT CGA CTG TAC CC	GAG CGG ATG ATT TTC GGT AAG A	6,6 pM

Aibl314	TGT GCT ATC GTT CGA CTC TAT CCG GCA TTA AAC GTC ACC AGA ACT CAG AAG CAG AAC AAA CGG GCT CAT CGA CTG TAC CC	CGT TTG TTC TGC TTC TGA GTT	6,6 pM
Aibl315	TGT GCT ATC GTT CGA CTC TAT CCG CTA ATC GAT TGG CGA TGC TAT TAC TGA TAT TCT CCG TAC GCT CAT CGA CTG TAC CC	GAG CGT ACG GAG AAT ATC AGT AAT	6,6 pM
Aibl465	TGT GCT ATC GTT CGA CTC TAT CCA GTA AGG TGT TGT TGA AGA CAA GCA GAT TCC ACA TTA CCT GCT CAT CGA CTG TAC CC	TGT GGA ATC TGC TTG TCT TCA	6,6 pM
Aibl597	TGT GCT ATC GTT CGA CTC TAT CCC GCT GCA ACG CAA AAG GGA AAG TGA ACC CGA ACA ACA CCG GCT CAT CGA CTG TAC CC	GGT GTT GTT CGG GTT CAC T	6,6 pM
Aibl607	TGT GCT ATC GTT CGA CTC TAT CCT TAG TTT CAC TAC CCC AGC CCT ATG TTA TAT TTA GAA TTA GCT CAT CGA CTG TAC CC	TAA CAT AGG GCT GGG GTA GT	6,6 pM
Aibl724	TGT GCT ATC GTT CGA CTC TAT CCT ATA AAC GAA CGT TTG GGA TAC AAA CCT GAA CAC ATC GCC GCT CAT CGA CTG TAC CC	CGG CGA TGT GTT CAG GTT	6,6 pM
Aibl746	TGT GCT ATC GTT CGA CTC TAT CCC TCA TCA GGA AAA ATC AGA ACT GAG CGA GTG ATC TCG AAT GCT CAT CGA CTG TAC CC	GAG ATC ACT CGC TCA GTT CT	6,6 pM

282

283 The same forward primer was used for each Aptamer: 5' TGTGCTATCGTTCGACTTACC 3'

284

285 **qPCR analysis of Aptamarkers**

286 The 14 aptamers listed in Table 1 were pooled at a concentration of 6.6 pM each. This mixture
287 was heated for 10 min at 95°C and snap cooled on ice for 5 min. The cooled solution was then
288 applied with selection buffer (20 mM Tris, 5 mM MgCl2, 5 mM KCl, 120 mM NaCl) on 10 µL
289 blood samples from high or low amyloid patients and incubated for 30 min at RT then applied for
290 another 15 min on a gold chip previously coated with random 8 mer sequences. The supernatant
291 was taken off the chip and cleaned with a GeneJET PCR Purification Kit (Thermo Fisher
292 Scientific, Germany) as described by the manufacturer and eluted with 35 µL of de-ionised
293 water. This process was repeated two more times, except for the last elution of 150 µL.

294 These aptamers were analyzed through real-time PCR (RT-PCR) using corresponding primers
295 (Table 1). The primers were designed to limit complementarity to each other in order to decrease
296 nonspecific amplification of self-dimerizing primers.

297 qPCR amplification was quantified through the use of a SYBR green master mix according to
298 the manufacturer's instructions (Bio-Rad). Each 20 µl reaction mix contained 5 µl of extracted
299 DNA and 250 nM PCR primers. The reaction was initially subjected to 95°C for 5 minutes
300 followed by 35 cycles of amplification (95°C for 10 seconds, 60°C for 15 seconds and 72°C for
301 30 seconds).

302 # Data analysis:

303 Each sample was analyzed in two well qPCR replicates. Following qPCR analysis, the data was
304 reviewed according to quality control standards with a coefficient of variance of 6 between
305 replicates, only data passing this standard was retained. The raw fluorescence data was analyzed
306 to identify the point at which fluorescence increased to six times the base level. The base level
307 was determined in the first five cycles. The point between the PCR cycles was determined by
308 linear regression analysis. This provided a Cq value. An efficiency value (E) for each sample was
309 determined based on identification of the linear portion of the amplification curve above the Cq
310 value.

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

314 The Cq and E values were averaged across replicates and the relative amount of template (RT)
315 was then determined with the following formula.

316 $$\frac{1}{E^{Cq}}$$

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

319 **Neomer selection and predictive modelling**

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

322

323 5'

324 CCAGATACAGACNNGAGGNNNGAATNNNAACCATCGGCGCCAACANNNCATTCNNN

325 CAGANNCAGTAGACAGC 3' (IDT, Belgium).

326

327 A prediction of the secondary structure formed by this library is provided in Figure 1 based on

328 the use of the RNAfold program[10] as visualized with Forna structure software[11].

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

330

331 **Selection of Aptamarkers**

332 Depletion of albumin and IgG was not performed with the Neomer library.

333 FRELEX was again used for selection with the Neomer library with the following modifications.

334 Instead of a gold chip we coated 40 nm Gold Nanoparticules (GNPs – Cytodiagnostics) with a

335 specific thiolated antisense DNA (/5ThioMC6-

336 D/AAAACAAAGTCTACTTGTTGGTTCTGTAT). This antisense was designed to have

337 complementary binding to fixed sequences in the Neomer library. The thiolated antisense was

338 reduced with TCEP following IDT specifications. Reduced antisense was then conjugated with

339 GNPs following supplier specifications and dissolved in 200 µL 0,1X of phosphate buffer saline

340 (PBS) (8.0 mM Na2PO4, 1.4 mM KH2PO4, 136 mM NaCl, 2.7 mM KCl, pH 7.4). This

341 concentration of the antisense in solution on the GNPs was measured with a NanoDrop

342 instrument in order to ensure that an antisense concentration of 0.24 µM was used for each

343 sample.

344 In the first step of FRELEX employed with the Neomer library, 7.15E-12 moles of library from
345 the random aptamer library described previously were snap cooled by heating the library to
346 95 °C for 10 min followed by immediate immersion in ice bath. These single stranded (ss) DNA
347 sequences were incubated with the individual plasma sample in 50 µL of Selection Buffer (10
348 mM Tris, 120 mM NaCl, 5 mM MgCl2, 5 mM KCl) for 15 min at RT. This solution was then
349 applied to the functionalized GNPs for 15 min at RT. The solution was then centrifuged for 20
350 min at 3500 g. Supernatant was removed and purified using the Oligonucleotide Clean-up
351 protocol of the Monarch PCR & DNA Clean-up kit, as described by NEB and eluted with 400 µL
352 of de-ionised water.

353 The overall process used for Neomer library selection is provided as a schematic in Figure 2.

354 **Figure 2:** Schematic overview of Neomer library analysis.

355 **Legend for Figure 2:** A, processing of library for NGS analysis. B, Outer product matrix of all
356 4.29 billion possible sequences. C, statistical analysis of aptamer frequencies across samples.
357 The distribution of the fixed sequences and random sequences is more complex than shown here
358 but is clearly described in the sequence provided elsewhere.

359

360 After this selection, PCR amplification was used to double strand the library (dsDNA) for an
361 appropriate number of cycles to create a clear band of approximately 5 ng of amplified DNA. All
362 PCR procedures were carried out according to standard molecular biology protocols and under
363 the following conditions: 95°C for 5 min, 4 cycles at 95°C for 10 s, 35°C for 15s, 72°C for 30
364 sec, and then remaining cycles at 95°C for 10 s, 64°C for 15s and 72°C for 30 sec, followed by a

365 final extension at 72°C for 5 min. Table 2 provides the primer sequences used for all steps
366 described here.

367 **Table 2:** Sequences of primers used to process Neomer libraries for next generation sequencing
368 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 <u>ATC ACG</u> 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 <u>ATC ACG</u> 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'

369

370 The underlined sequence in Table 2 is a hex code used for library parsing from the fastq file. The
371 sequence shown is only one example, we used 21 different hex codes for this analysis. We used
372 the same hex code for module A and module B in all cases.

373 It is important to maintain the average expected copy of each sequence at a level above 100
374 copies throughout the NGS preparation procedure. To ensure this we performed the
375 Oligonucleotide Clean-up protocol of the Monarch PCR & DNA Clean-up kit (NEB) on 10 PCR

376 reaction tubes. The 40 µL volume of elution was used as a template for restriction with KasI
377 enzyme (NEB) at 37°C for 60 min followed by 20 min at 65°C to obtain two distinct modules (A
378 and B). The two modules were treated in the same way, but separately for Next Generation
379 Sequencing (NGS) preparation. The restricted products were diluted 10X prior to next step.

380 The amplified product from NGS preparation was purified using the Monarch PCR & DNA
381 Clean-Up kit prior to be sequenced.

382 In addition, we processed a buffer sample containing no blood in each analysis, as a further test.

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

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

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

We also ran this analysis by subtracting the average of each sequence with the high amyloid 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 3. 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 3: 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

409

410 **Screening of Aptamarkers**

For the Neomer Aptamarkers screening and data analysis was performed in the same manner as described for the SELEX/FRELEX approach.

Results and Discussion:

The frequency of the top 1,000 SELEX aptamers following selection on the pools is provided in Figure 3. This represents the number of reads of each sequence divided by the total number of reads of the library.

Figure 3: Frequency of top 1,000 SELEX sequences after pool selection.

The 1,000 sequences characterized in Figure 3, were analyzed for variation in their response to three rounds of selection on the twenty individual samples that made up the pools (Figure 4). Fourteen aptamers that exhibited the strongest deviation from a regression line drawn through the data within Figure 4 were selected as Aptamarkers. These sequences are provided in Table 1.

Figure 4: Comparison of average frequencies of high and low amyloid individual samples after three rounds of selection with the enriched library.

The top machine learning algorithms applied to the qPCR results of these 14 Aptamarkers across the 400 samples is provided in Table 4.

Table 4: Top performing models from machine learning training on 300 samples.

Algorithm	Excluded	Accuracy	Sensitivity	Specificity	AUROC
DecisionTreeClassifier		1	1	1	1
DecisionTreeClassifier	Clinical Classification	1	1	1	1
DecisionTreeClassifier	Abl465, Clinical Classification	1	1	1	1
DecisionTreeClassifier	Abl160, Abl465, 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	Abl042, Clinical Classification, Sex	1	1	1	1
GradientBoostingClassifier		1	1	1	1
XGBClassifier	Abl597, Abl021, Abl160	1	1	1	1
ExtraTreesClassifier		0.976667	0.95	1	1
ExtraTreesClassifier	Sex	0.98	0.957143	1	1
ExtraTreesClassifier	Clinical Classification, Sex	0.993333	0.985714	1	1
ExtraTreesClassifier	Abl042, Clinical Classification, Sex	1	1	1	1
BaggingClassifier	Sex	1	1	1	1

430

Based on this analysis the algorithm Extra Trees Classifier with the clinical variable, sex, excluded was applied to the test samples as this most closely conformed with the model used for the Neomer Aptamarker analysis. The other models led to similar results on the test samples. The use of this model resulted in a test sensitivity of 0.53, specificity of 0.92, and an overall accuracy of 0.74. The ROC curve for this model on these 400 samples is provided in Figure 5, and the contribution of each variable to predictive power is provided in Figure 6.

Figure 5: Analysis of predictive power of SELEX Aptamarkers on 400 samples - Receiver-operator curve

Figure 6: Analysis of predictive power of SELEX Aptamarkers on 400

Each of the aptamers contributes a small amount to the predictive power. Clinical classification and age contribute more to prediction than any Aptamarker alone. However, the majority of the weight of the prediction is based on the combined contribution of all of the Aptamarkers.

Neomer selection:

The Z values versus fold values for the top 10,000 aptamers from the Neomer selection are are presented in Figure 7 a and b for the high versus low and low versus high amyloid contrasts respectively.

Figure 7: The distribution of Z and fold values for a) the high amyloid versus low amyloid contrast and b) the low amyloid versus high amyloid contrast. The fold values are simply the

difference between the high and low average minus unity. The low versus high graph is presented with the opposite sign of the high versus low amyloid contrast.

It is interesting to note that the fold differences were higher with the high versus low amyloid contrast as compared to the low versus high contrast. The Z values have very low p values given that these contrasts are based on a comparison between 10 samples for each phenotype. The higher fold value for the high amyloid aptamers may indicate that these will be more predictive than the low amyloid ones.

Figure 8 provides the separation of each of the samples for the top four high amyloid and top four low amyloid aptamers.

Figure 8: Separation of high and low samples by the four best Aptamarkers for a) high amyloid and b) low amyloid.

Each of the spots in Figure 5 corresponds to an individual sample. Aptamers were also selected such that the result on buffer alone (no plasma) is aligned with the non-targeted phenotype.

For machine learning analysis we used the scikit program in Python with all eight Aptamarkers as variables and the clinical variables sex, clinical status (mild cognitive impairment (MCI) or cognitively normal (CN)), and age. We did not include ApoE genetic status in our primary model based on the understanding that this variable is not commonly applied in clinical practice in Europe. Table 5 provides a summary of all algorithms that returned perfect predictive results on the test set.

471 **Table 5:** Top models based on machine learning analysis of training set samples.

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 390 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 AUC-ROC 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 AUC-ROC 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 9 provides a description of the relative weight of each of the variables in the model.

Figure 9: Weight attributable to each variable in 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 10.

Figure 10: Receiver operator response

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

495 **Conclusions:**

496 The application of the Neomer library approach improves the predictive capacity of the
497 Aptamarkers identified over those developed by the SELEX method It is clear that the
498 Aptamarker approach clearly results in predictive capacity for brain amyloid deposition from
499 plasma analysis.

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

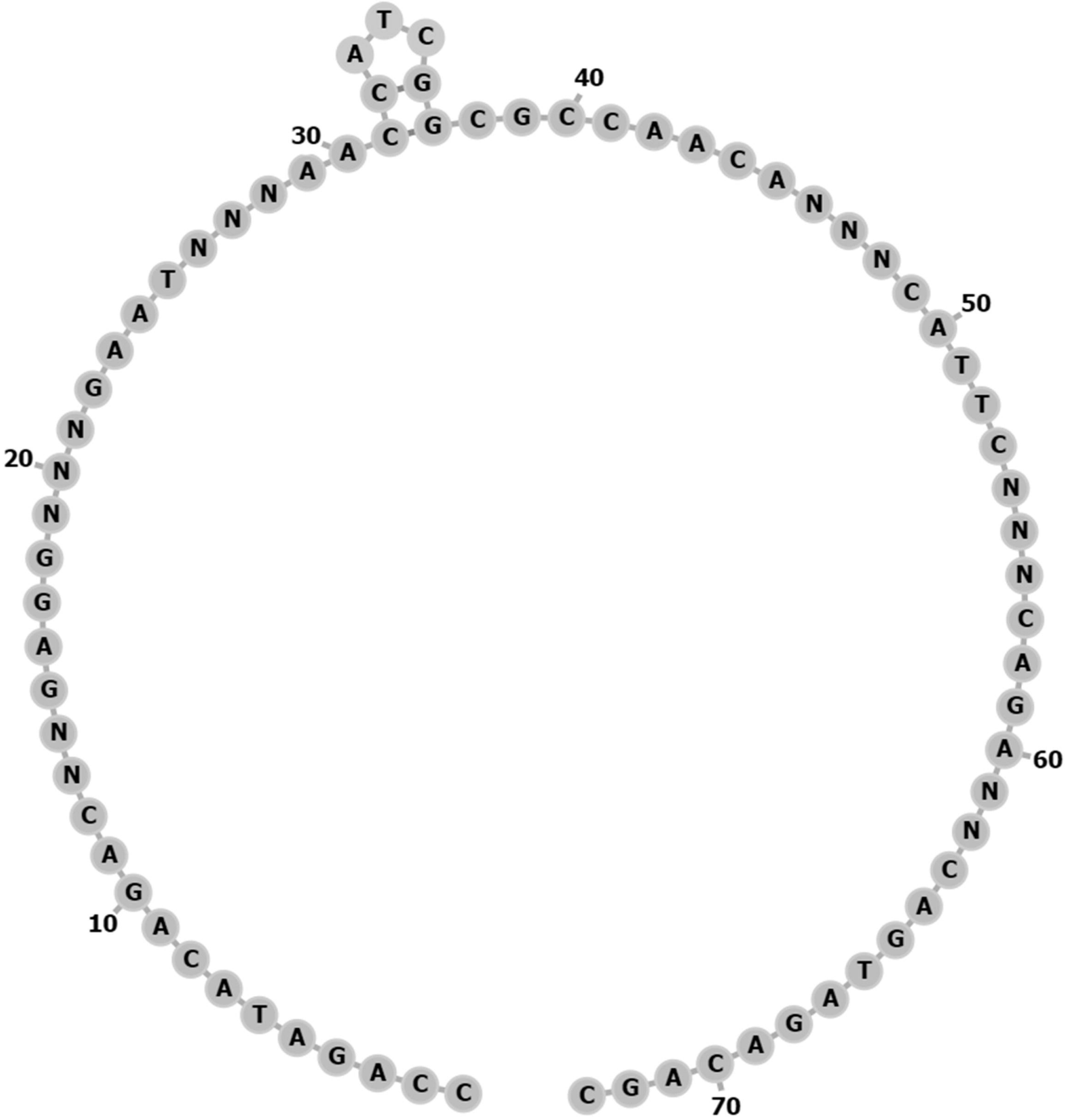
501

References:

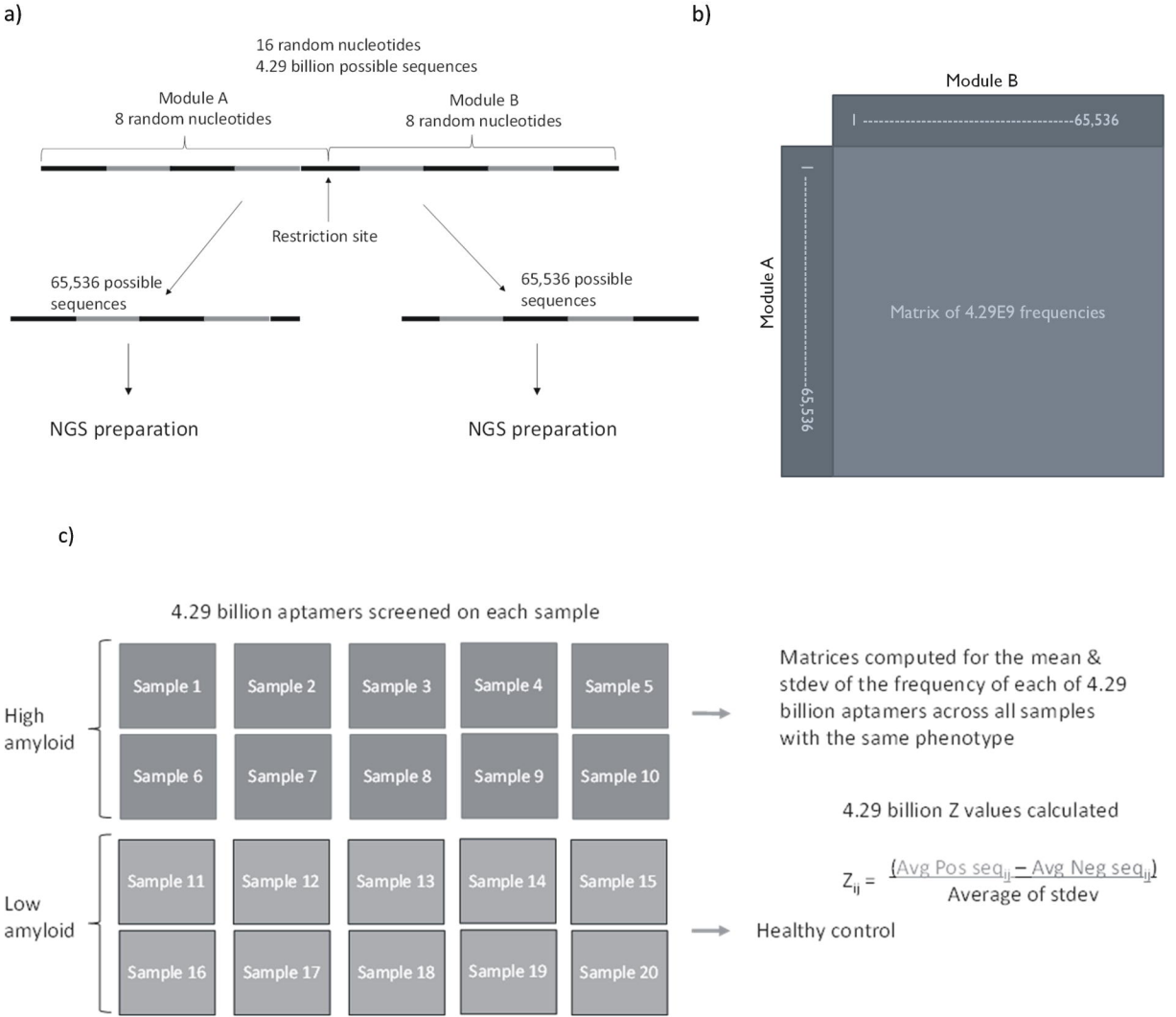
1. Jack CR Jr, Bennett DA, Blennow K, Carrillo MC, Dunn B, Haeberlin SB, et al. "NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease" Alzheimer's & dementia: The Journal of the Alzheimer's Association vol. 14,4 (2018): 535–562. pmid:29653606
2. Reimand J, Groot C, Teunissen CE, Windhorst AD, Boellaard R, Barkhof F, et al. Why Is Amyloid-β PET Requested After Performing CSF Biomarkers? J Alzheimers Dis. 2020;73(2):559-569. doi: 10.3233/JAD-190836. PMID: 31796674; PMCID: PMC7081099.
3. Li Y, Schindler SE, Bollinger JG, Ovod V, Mawuenyega KG, Weiner MW, et al. Validation of plasma amyloid-β 42/40 for detecting Alzheimer disease amyloid plaques. *Neurology*. 2022;98(7):e688-e699. doi: 10.1212/WNL.0000000000013211.
4. Zetterberg H, Schott JM. Blood biomarkers for Alzheimer's disease and related disorders. Acta Neurol Scand. 2022;146(1):51-55.
5. Tuerk C, Gold L. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. Science, 1990; 249:505–510.
6. Penner G, Method For The Selection Of Aptamers For Unbound Target, WO 2017/035666A1, US 10,415,304 B2, 2020
7. Penner G, Lecocq S, Chopin A, Vedoya X, Lista S, Vergallo A, et al. (2021) Aptamarker prediction of brain amyloid-β status in cognitively normal individuals at risk for Alzheimer's disease. PLoS ONE 16(1): e0243902. https://doi.org/10.1371/journal.pone.0243902,

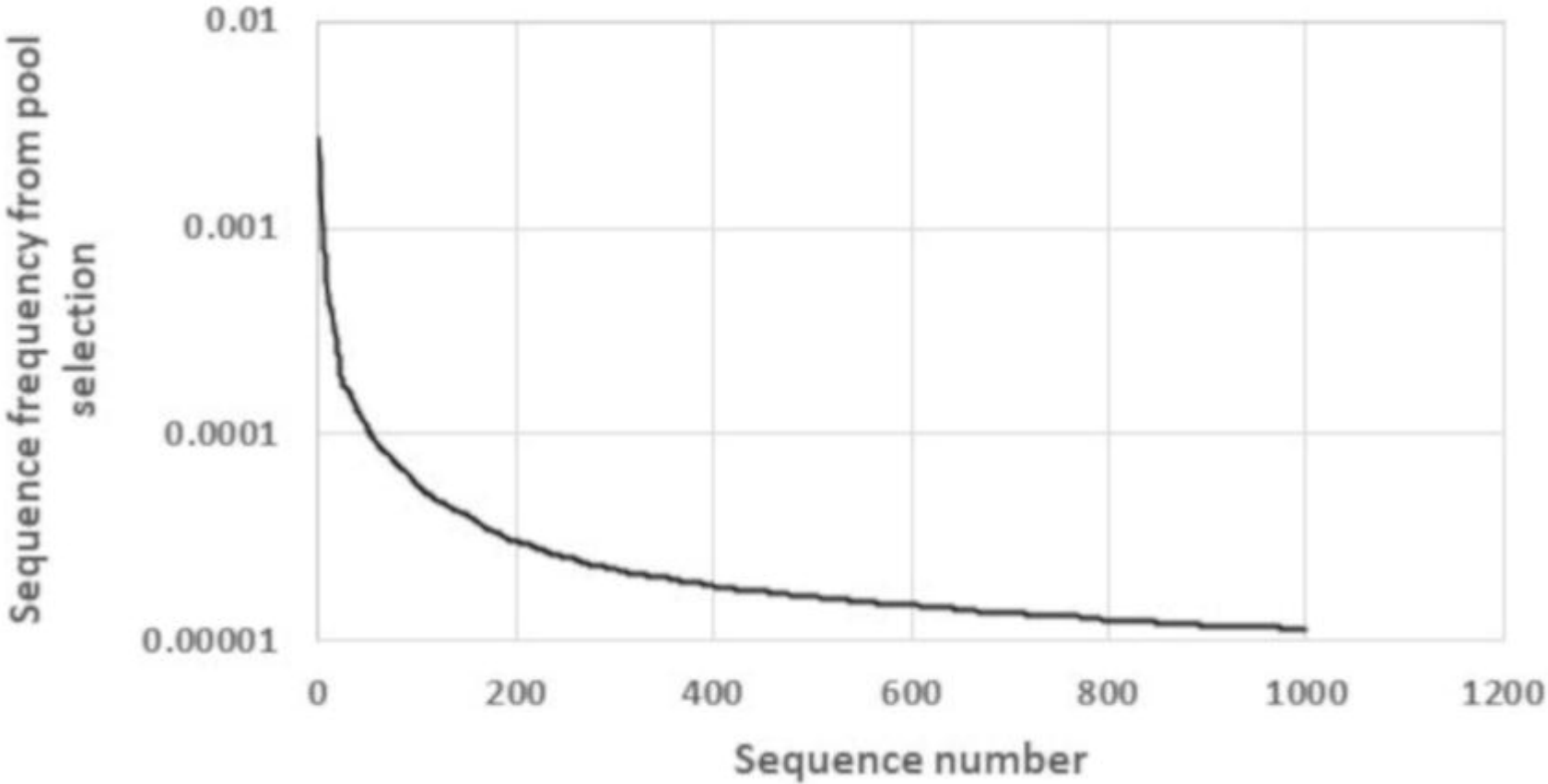
8. Lecocq S, Spinella K, Dubois B, Lista S, Hampel H, Penner G (2018) Aptamers as biomarkers for neurological disorders. Proof of concept in transgenic mice. PLoS ONE 13(1): e0190212. https://doi.org/10.1371/journal.pone.0190212).
9. Collie A, Maruff P, Darby DG, Masters C, Currie J. The Melbourne Aging Study. In: Mild Cognitive Impairment: International Perspectives. Edited by Tuokko HA, Hultsch DF. New York: Taylor & Francis; 2006: 93-113.
10. Gruber AR, Lorenz R, Bernhart SH, Neuböck R, Hofacker IL. The Vienna RNA Websuite. Nucleic Acids Research, Volume 36, Issue suppl_2, 1 July 2008, Pages W70-W74, DOI: 10.1093/nar/gkn188.
11. Kerpedjiev P, Hammer S, Hofacker IL (2015). Forna (force-directed RNA): Simple and effective online RNA secondary structure diagrams. Bioinformatics 31(20):3377-9.

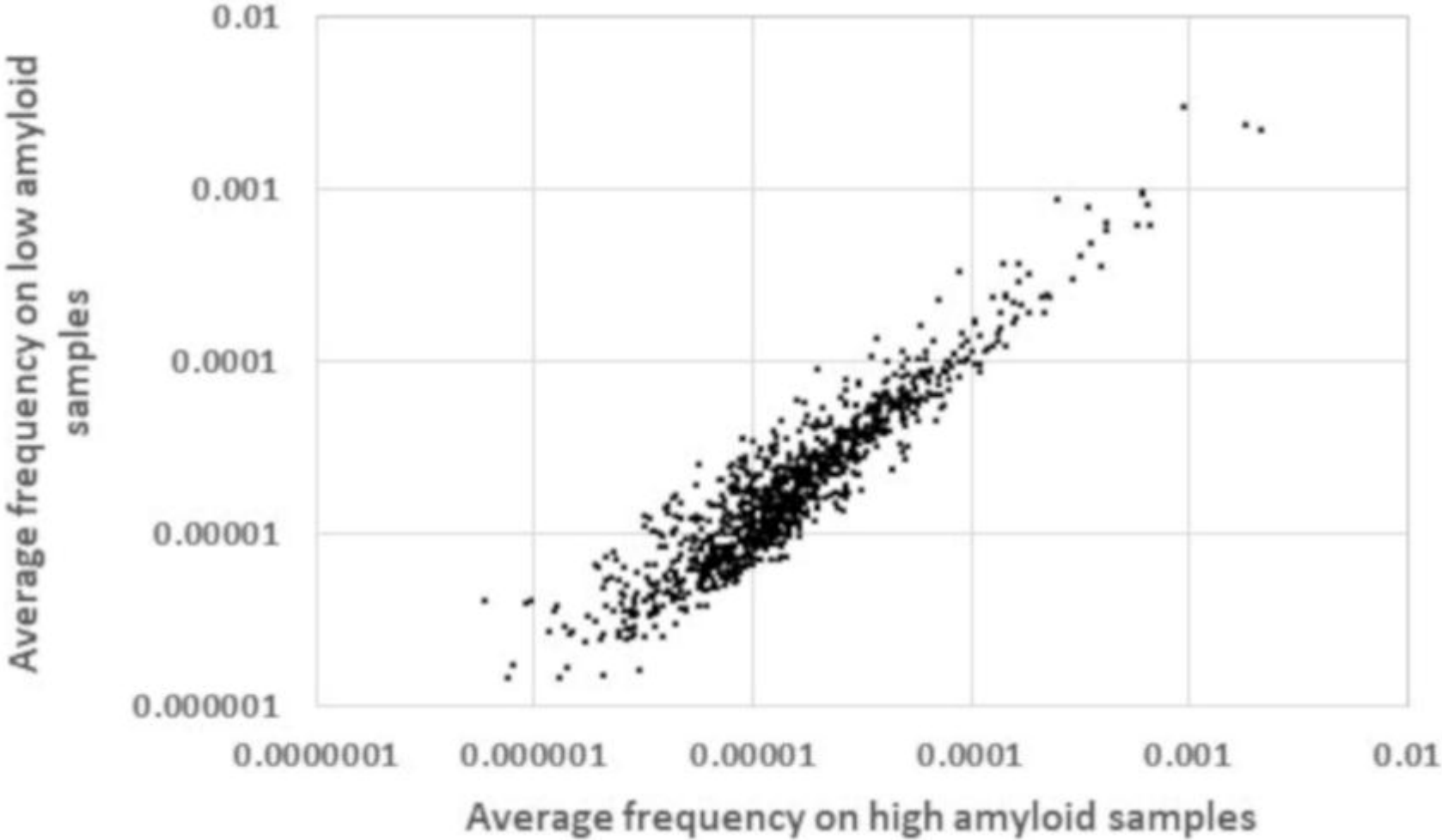
Note: Patent applications have been submitted globally describing both the Neomer application as well as the Aptamarker approach.

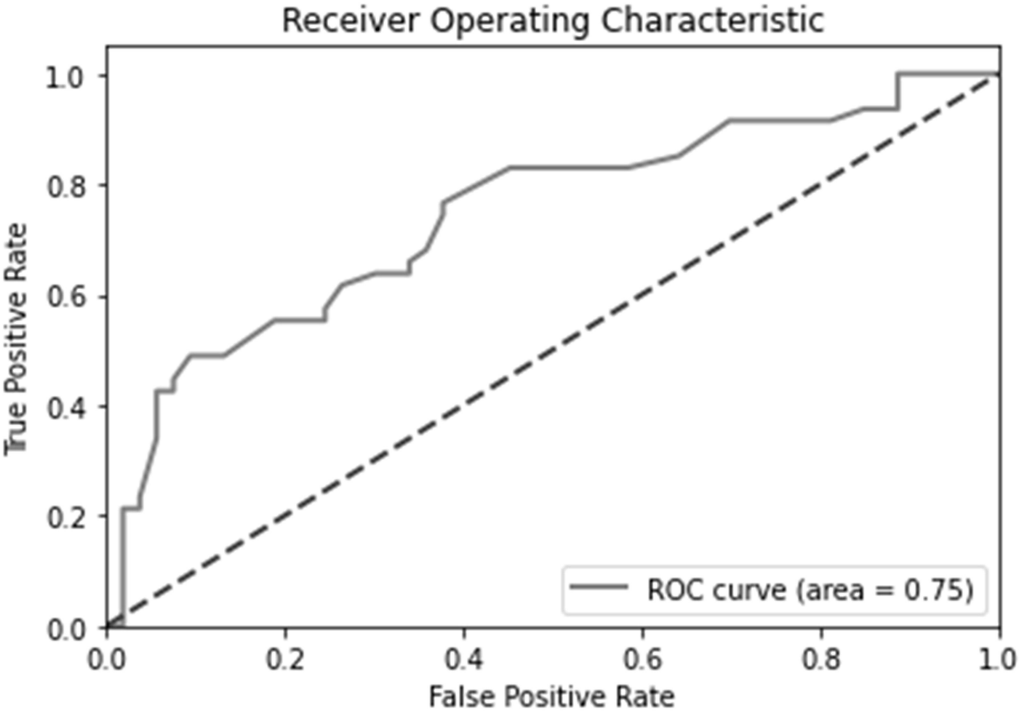


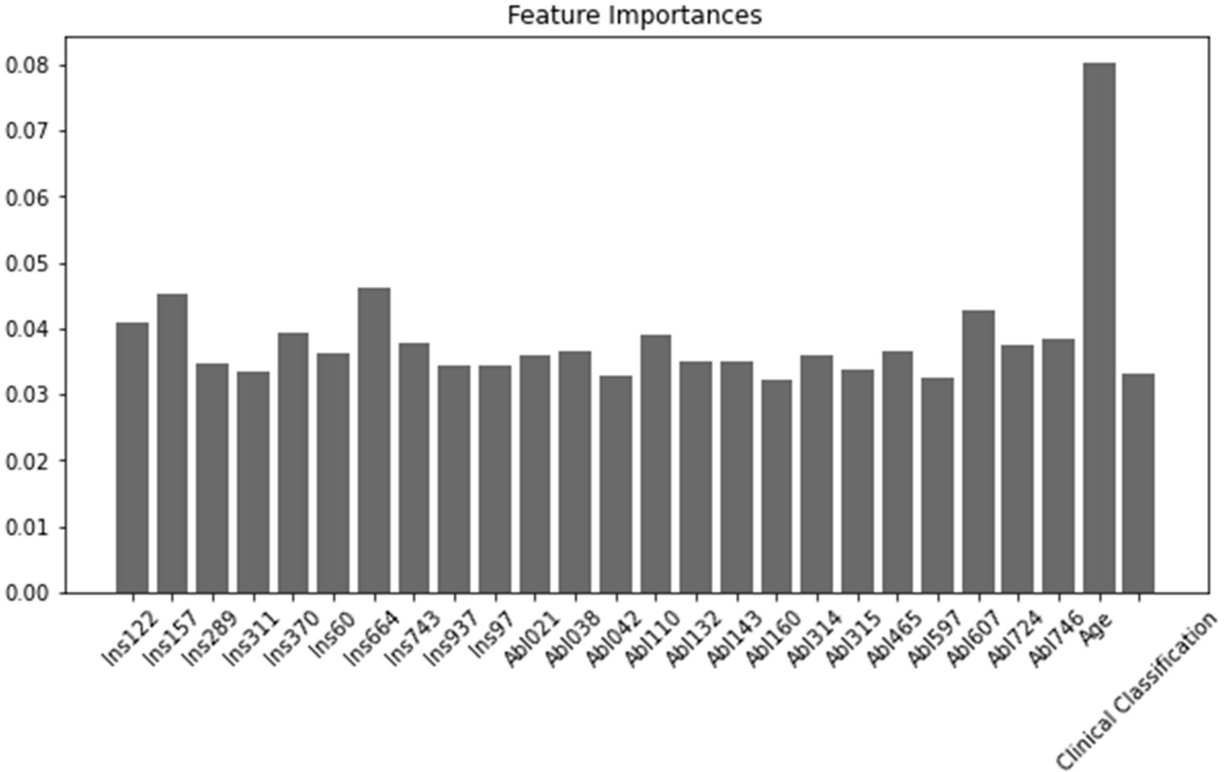
40
30
50
20
60
10
70


a)
16 random nucleotides
4.29 billion possible sequences
Module A
8 random nucleotides
Module B
8 random nucleotides
Restriction site
65,536 possible sequences
65,536 possible sequences
NGS preparation
NGS preparation
b)
Module B
1
65,536
Module A
1
65,536
Matrix of 4.29E9 frequencies
c)
4.29 billion aptamers screened on each sample
High amyloid
Sample 1
Sample 2
Sample 3
Sample 4
Sample 5
Sample 6
Sample 7
Sample 8
Sample 9
Sample 10
Low amyloid
Sample 11
Sample 12
Sample 13
Sample 14
Sample 15
Sample 16
Sample 17
Sample 18
Sample 19
Sample 20
Matrices computed for the mean & stdev of the frequency of each of 4.29 billion aptamers across all samples with the same phenotype
4.29 billion Z values calculated
$Z_{ij} = \frac{(\text{Avg Pos seq}_{ij} - \text{Avg Neg seq}_{ij})}{\text{Average of stdev}}$
Healthy control

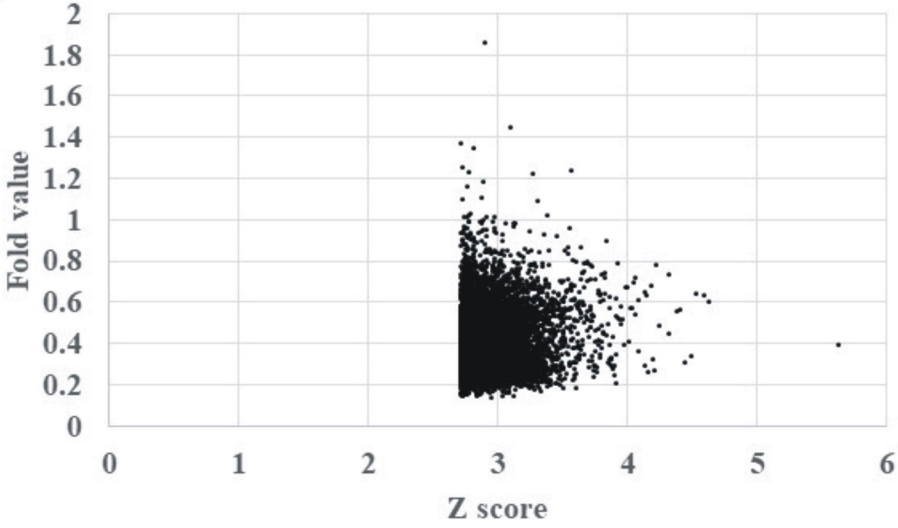

0.01
0.001
0.0001
0.00001
Sequence frequency from pool selection
0
200
400
600
800
1000
1200
Sequence number


0.01
0.001
0.0001
0.00001
0.000001
Average frequency on low amyloid samples
0.0000001
0.000001
0.00001
0.0001
0.001
0.01
Average frequency on high amyloid samples

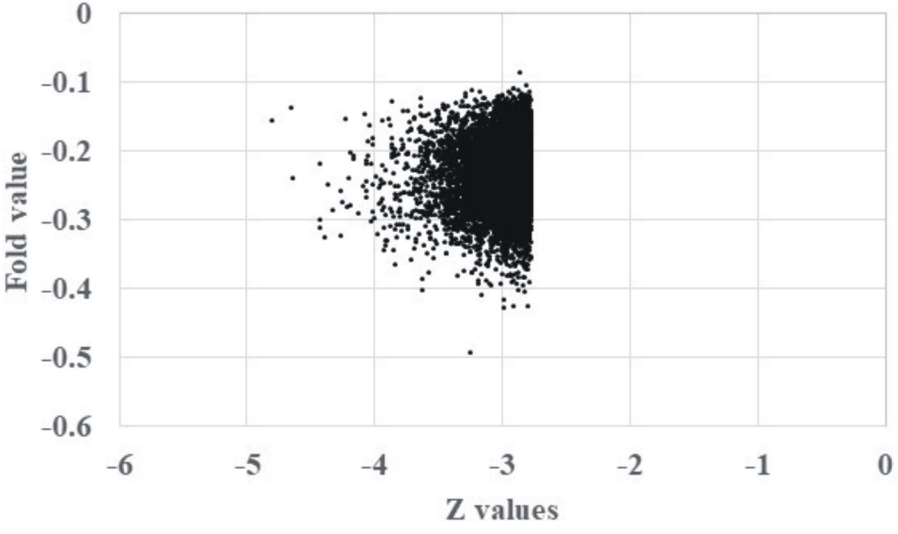

Receiver Operating Characteristic
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.75)


Feature Importances
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
Ins122
Ins157
Ins289
Ins311
Ins370
Ins60
Ins664
Ins743
Ins937
Ins97
Abl021
Abl038
Abl042
Abl110
Abl132
Abl143
Abl160
Abl314
Abl315
Abl465
Abl597
Abl607
Abl724
Abl746
Age
Clinical Classification

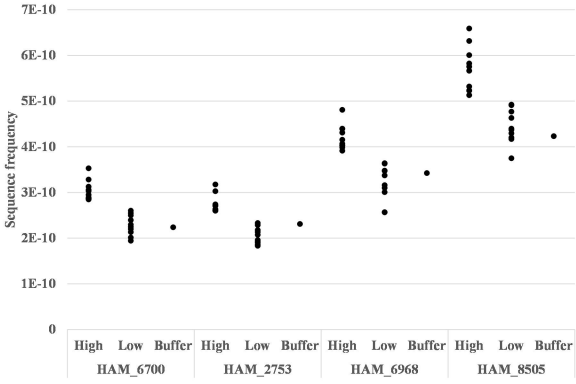
a)


Fold value
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
0
1
2
3
4
5
6
Z score

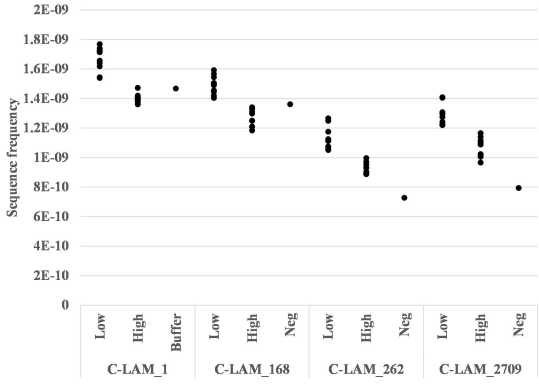
b)

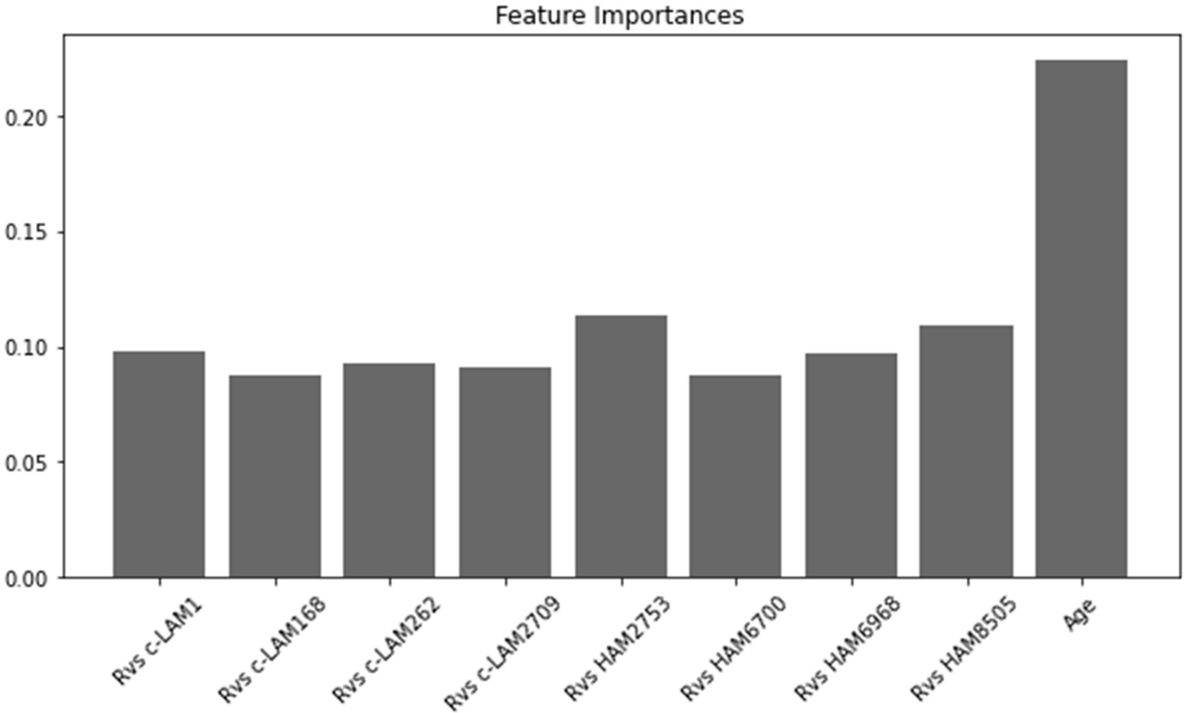

Fold value
0
-0.1
-0.2
-0.3
-0.4
-0.5
-0.6
-6
-5
-4
-3
-2
-1
0
Z values

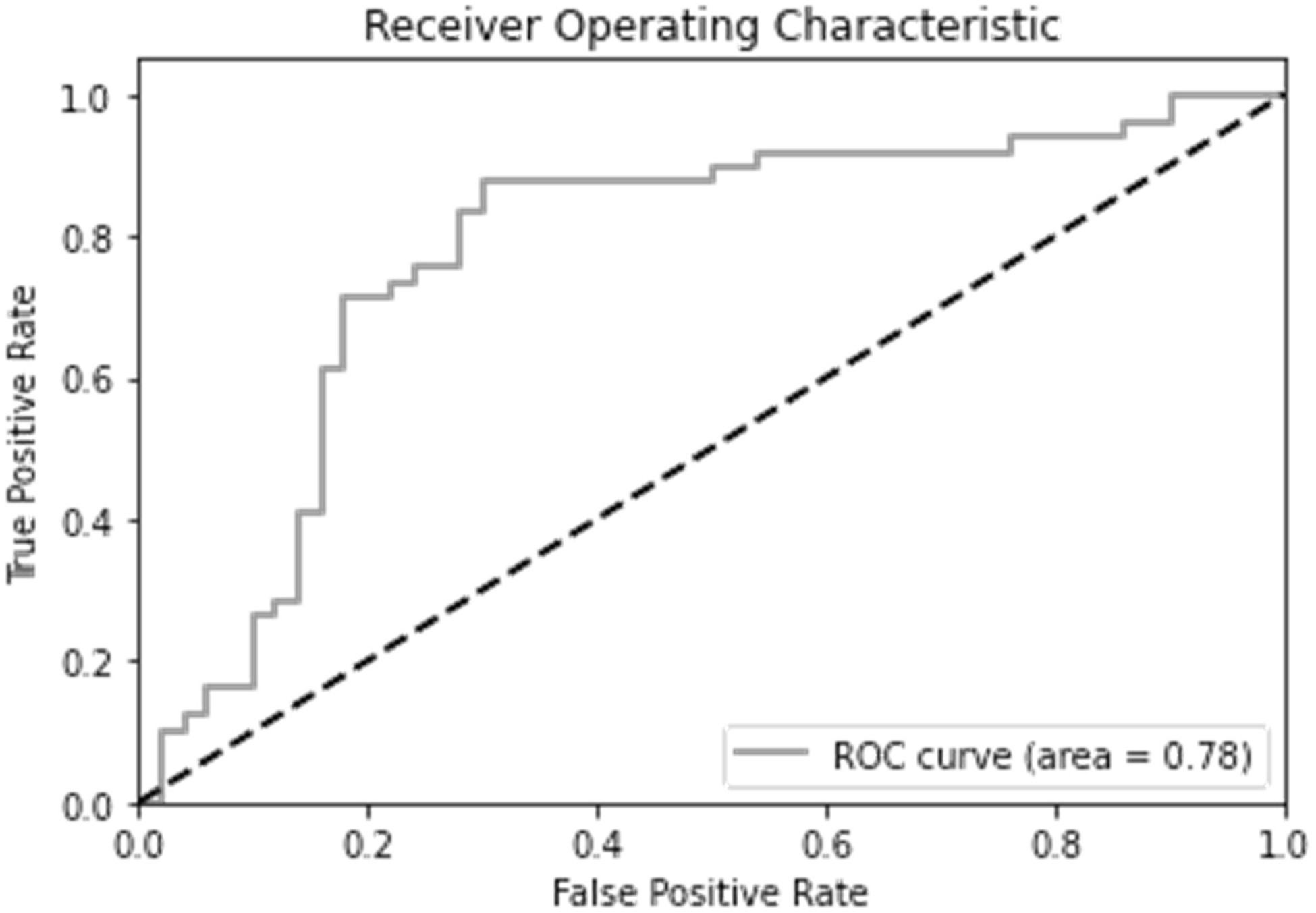
a)


7E-10
6E-10
5E-10
4E-10
3E-10
2E-10
1E-10
0
Sequence frequency
High Low Buffer High Low Buffer High Low Buffer High Low Buffer
HAM_6700
HAM_2753
HAM_6968
HAM_8505

b)


2E-09
1.8E-09
1.6E-09
1.4E-09
1.2E-09
1E-09
8E-10
6E-10
4E-10
2E-10
0
Sequence frequency
Low High Buffer Low High Neg Low High Neg Low High Neg
C-LAM_1
C-LAM_168
C-LAM_262
C-LAM_2709


Feature Importances
0.20
0.15
0.10
0.05
0.00
Rvs c-LAM1
Rvs c-LAM168
Rvs c-LAM262
Rvs c-LAM2709
Rvs HAM2753
Rvs HAM6700
Rvs HAM6968
Rvs HAM8505
Age


Receiver Operating Characteristic
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.78)