Supplementary Table 2. Further results from Surface Plasmon Resonance (SPR) fits. See also Table 1 and Figure 3. There is no clear relationship between concentration and reaction rate for the high-scoring sequence (G9.3415, binding-score 12.6) suggesting complex kinetics. The low-scoring sequence (C10489, binding score 5.8) only bound about as much APC as the blank control (and at a slower rate, giving negative RU values), therefore no 'on rate' could be fitted. However, an 'off-rate' can calculated for comparison with the high-scoring sequence, but should be treated with caution since all values were calculated following subtraction of the blank control.

	High-scoring sequence		Low-scoring sequence
APC concentration (nM)	Reaction rate k_{obs} (s^{-1})	'off rate' k_d (s^{-1})	'off rate' k_d (s^{-1})
214	4.5E-03	5.3E-05	1.2E-04
97	2.1E-03	4.3E-05	9.3E-05
76	1.6E-03	5.5E-05	1.1E-04
53	1.9E-03	6.1E-05	8. 9E-05
31	4.6E-04	6.4E-05	7.8E-05

Supplementary Methods

Chemicals and equipment All chemicals were used as purchased without further purification unless stated. Albumin bovine serum (BSA, molecular biology grade), sodium chloride, tween 20, sodium di-hydrogen phosphate and di-sodium hydrogen phosphate, (>99.5%) were purchased from Sigma chemical company, UK. Phycopro crosslinked allophycocyanin, APC (Europa Bioproducts Ltd) was used at a concentration 0.01 mg/ml, its characteristic excitation and emission bands are 648nm and 657nm respectively. DsRed fluorescent protein 1 was purified by affinity chromatography using a His6-tagged plasmid (pDsRed, Invitrogen). DsRed has excitation and emission bands at 560nm and 580nm, and the protein was used at a concentration of 0.01mg/ml. 20xSSC buffer was made from Tri sodium citrate, 0.3M (Fisher Scientific, UK) and sodium chloride, 0.3M pH 7.2 and diluted for use. Water was obtained from a milli Q purification system (Millipore, 18 MΩ).

Fluorescent intensities from the arrays were imaged wet, with a Genepix 4000B scanner (Axon instruments).

***In vitro* assays** Allophycocyanin (APC) is a fluorescent protein unique to cyanobacteria and red algae and a member of the phycobiliprotein family. APC is widely used as a fluorescent reagent to label biomolecules in screening assays, where it performs the role of acceptor molecule in homogeneous time-resolved fluorescence (HTRF) assays used in high throughput drug discovery. APC was chosen in these studies as it is a well established assay reagent with high quantum yield (up to 68%), fluoresces over a broad pH range, has high photostability and high water solubility 2. Moreover, the fact that APC's fluorescence is not quenched by external agents (since the fluorophore is buried in the protein interior where it is covalently bound to the protein backbone) makes APC an attractive initial protein target for aptamer design.

Optimization for the APC binding conditions was performed on two identical chips. Each chip contained 6000 sequences, repeated in duplicate, each 30 bases in length; hence each aptamer was represented on a total of 4 spots. These initial chips, termed G0 contained a wide variety of sequences. A quarter of G0 sequences were designed to form G-quadruplex structures. These were 30 bp sequences with the sequence N1-7G3-5N1-7G3-5N1-7G3-5N1-7G3-5N1-7 where the identity of N was allotted to A, T, C or G at random with equal probability and the length of each G or N run allotted at random within the 30bp length constraint. The remainder of the G0 chip comprised sequences obtained by a fully in silico evolution. The protocol for this evolution was the same as that for the main evolution described in this paper, except that the fitness assay was the score given to the sequence by the complex linear model shown in supplementary Figure 14 rather than an in vitro assay and evolution was continued for 19 generations. The model in supplementary Figure 14 derived from experiments that evolved aptamers to bind the Cy5 fluorophore (unpublished data). Arbitrary selection from the sequences in this evolution made up half of G0 sequences and the remaining quarter of G0 sequences (including all those that went into the ‘designed’ set referred to in the main text) were picked from across all generations of this evolution to have as even as possible a distribution across the range of scores obtained.

Optimized conditions are defined as no/ little APC precipitation out of solution, no/ little background fluorescence from APC arising from non-specific adsorption onto the chip’s surface and the presence of several aptamers showing some interaction with APC. During optimization a range of pH and ionic strengths was investigated. PBS experiments ranged

from a pH of 5–7.5, using 1xPBS – 10xPBS. SSC buffer pH 7.5 was also tested using both 3xSSC and 1xSSC. There was a significant increase in signal observed from several spots when both the ionic strength of the solution and pH was lowered to that of 1xPBS, pH 5.4. The chosen binding conditions were fixed for all assays and it is important to note that all the results presented, both *in vitro* and *in silico*, are valid only for these conditions.

Replicate sequences within each chip showed good correlation (on average ρ=0.90), and excellent correlation between the pair of chips was also observed (ρ=0.94). The sequences from G0 were ranked in order of Fluorescent intensity, a high intensity representing a good binder to APC. A selection of 500 sequences was taken representing, as evenly as possible, the entire range of binding intensities observed. These were labelled as control sequences, and used for normalization in all generations as described below and selection in G1.

The remaining chips have all sequences present in duplicate, and each generation of probes is synthesized on at least two chips, each with a unique random assignment of sequences and control aptamers to spatial positions.

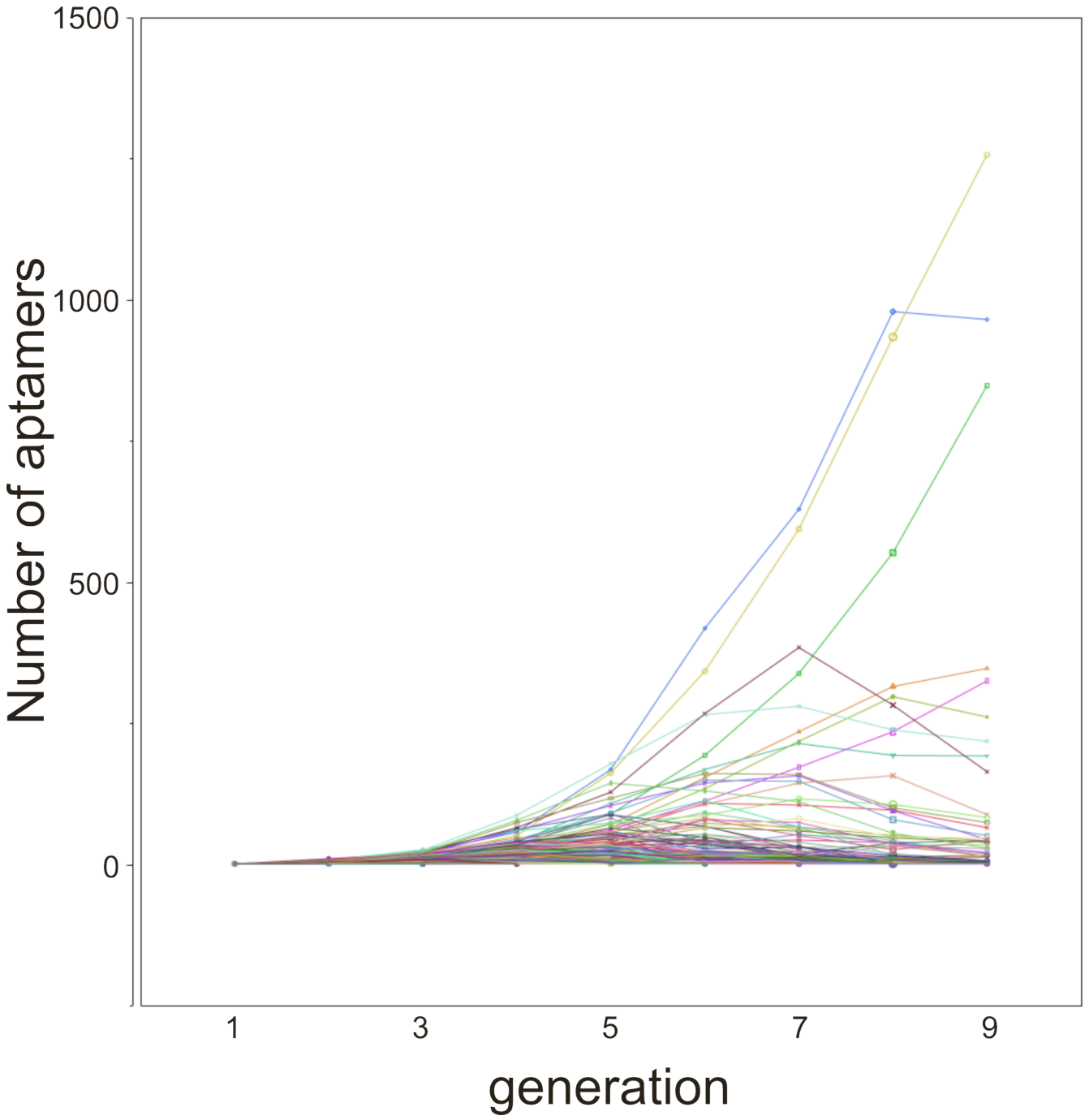
Multidimensional scaling Figure 5B is a two dimensional representation of the 50,000 x 50,000 distance matrix of the data points generated during the evolution of aptamers to APC. The generation of such a low-dimensional representation of a distance matrix is known as multidimensional scaling.

Classical multidimensional scaling requires operations on the NxN distance matrix which is effectively intractable when working with such large quantities of data. To maintain scalability, the algorithm used here (1) does not reproduce the full distance matrix but first calculates the locations of group of landmark points (group of sequences from the first generation) in the coordinate space. The coordinates of the remaining sequences within the dataset are then triangulated based on these fixed points (2). The Shepard diagram in supplementary Figure 11 displays the goodness of fit of the projection to the real sequence distances. It is clear that, while there is a reasonable correlation (ρ=0.50, N=1x10^5) between true distances and 2D distances as used in Figure 5B, all points that are distant in Figure 5B are distant in reality, but points that are close in Figure 5B may not be close in reality.

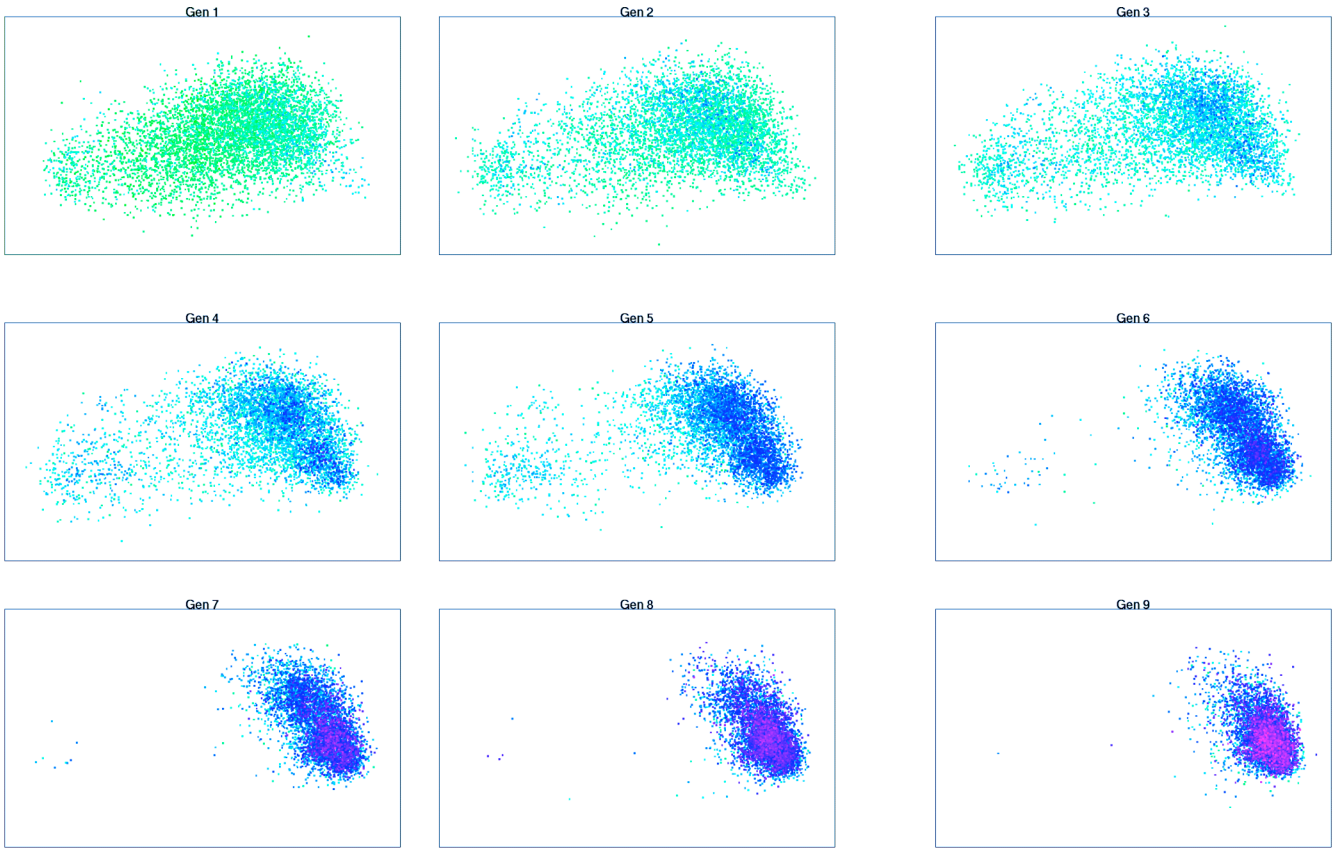
Distances were generated using a Needleman-Wunsch pairwise global sequence alignment (3). The Needleman-Wunsch algorithm was implemented with a set mismatch penalty of 1 and a gap penalty of 1; there was no special weighting for mismatch for transitions/transversions or gap extensions. Sequence similarity scores were converted to distances by subtraction from 30 (the sequence length). These same distances were used for the sequence-distance vs. score-distance plot (Figure 6).

References

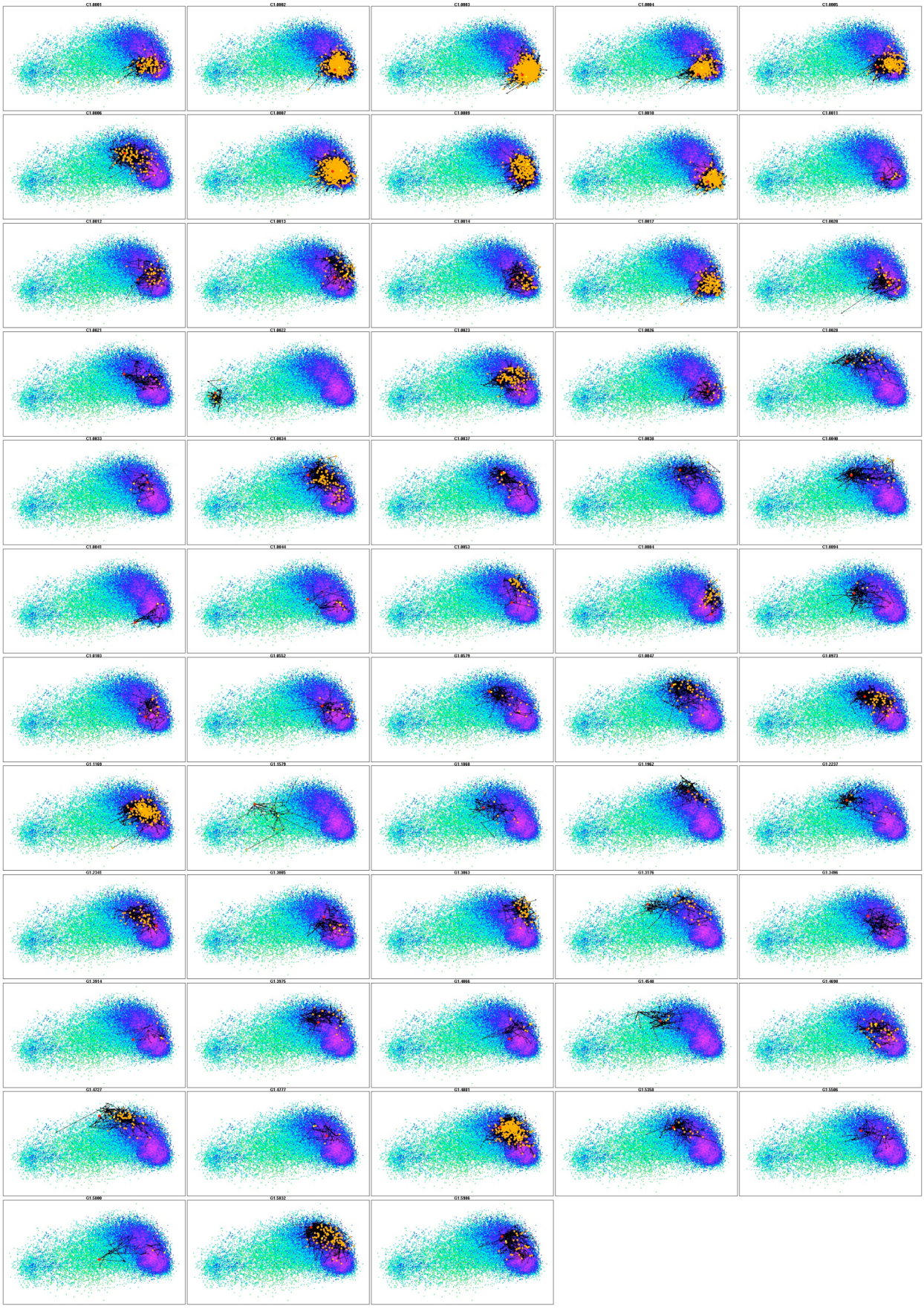
1. Tenenbaum, J. B., de Silva, V. & Langford, J. C. (2000) A global geometric framework for nonlinear dimensionality reduction. *Science* 290, 2319-23.
2. de Silva, V. & Tenenbaum, J. B. (2003) Global versus local methods in nonlinear dimensionality reduction. *Advances in Neural Information Processing Systems* 15, 705-712.
3. Needleman, S. B. & Wunsch, C. D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* 48, 443-53.



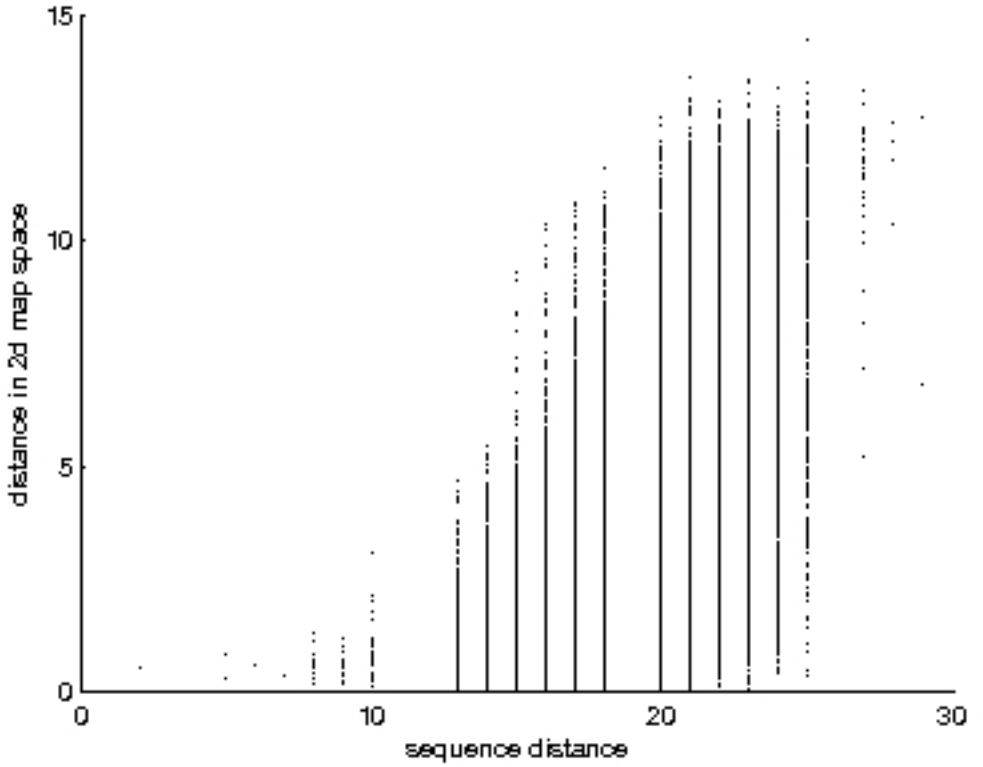
Supplementary Figure 8. The dynamics of all 6000 lineages over the course of the evolution. Each lineage starts as a single sequence in generation 1, the total population being 5500 sequences in all subsequent generations. The dynamics are complex, 1376 lineages rise then fall in frequency, of which 54 lineages subsequently rise again.



Supplementary Figure 9. The same as Figure 5B except that points are plotted separately for each generation.

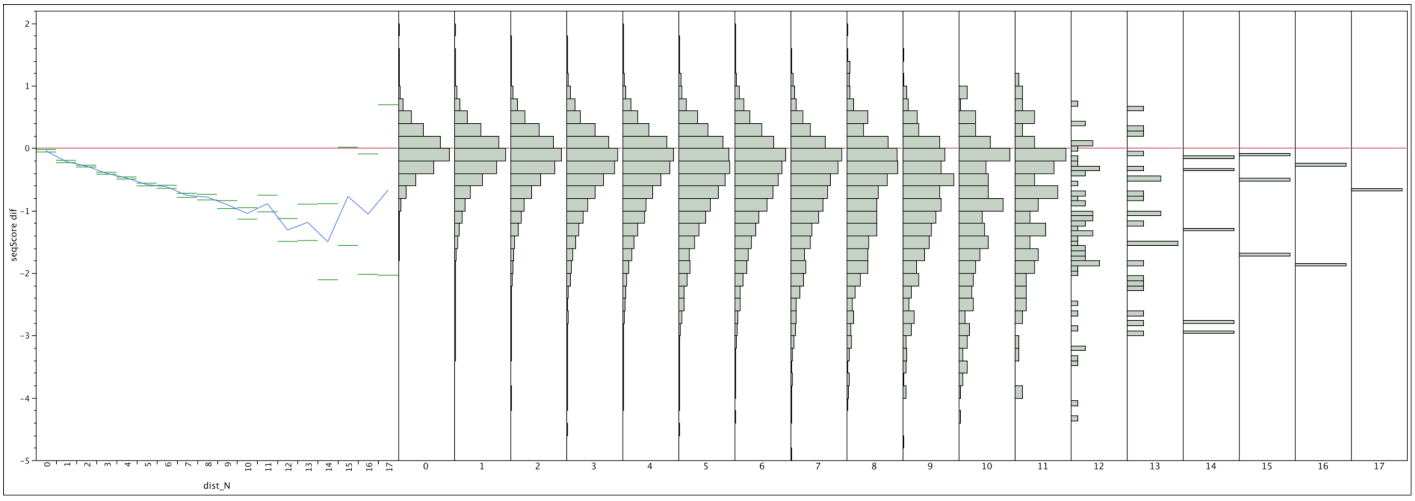


Supplementary Figure 10. Evolutionary paths of the 58 lineages with probes in generation 9. All the lineages in Figure 5A are shown in the same way as the example lineages in Figure 5B. Colours and symbols as in Figure 5B.



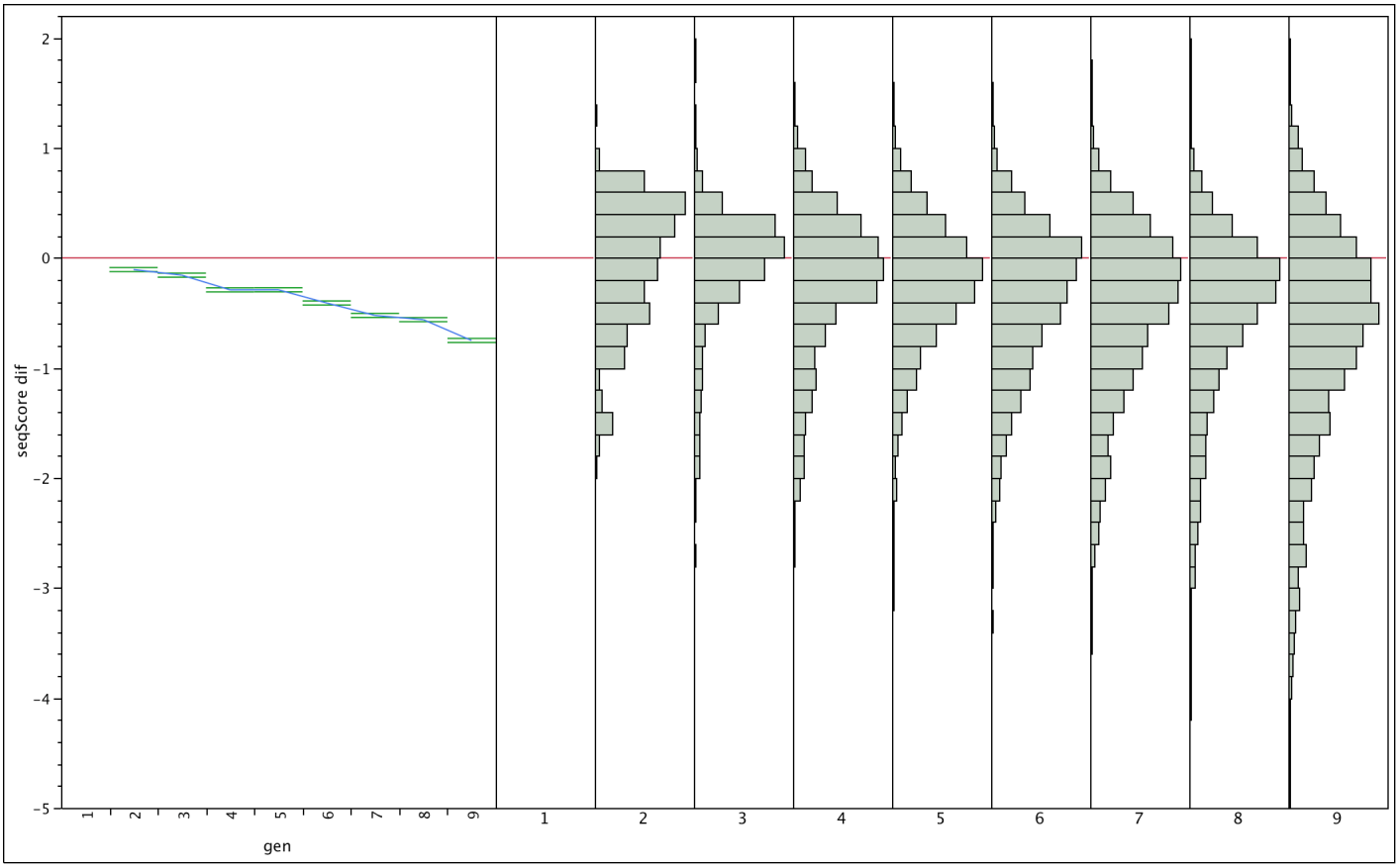
Supplementary Figure 11. Shepard diagram. Displays distance between points in 2-dimensional space (as plotted in Figure 5B) against sequence distance. The plot is based on a sample of 100,000 distances.

A



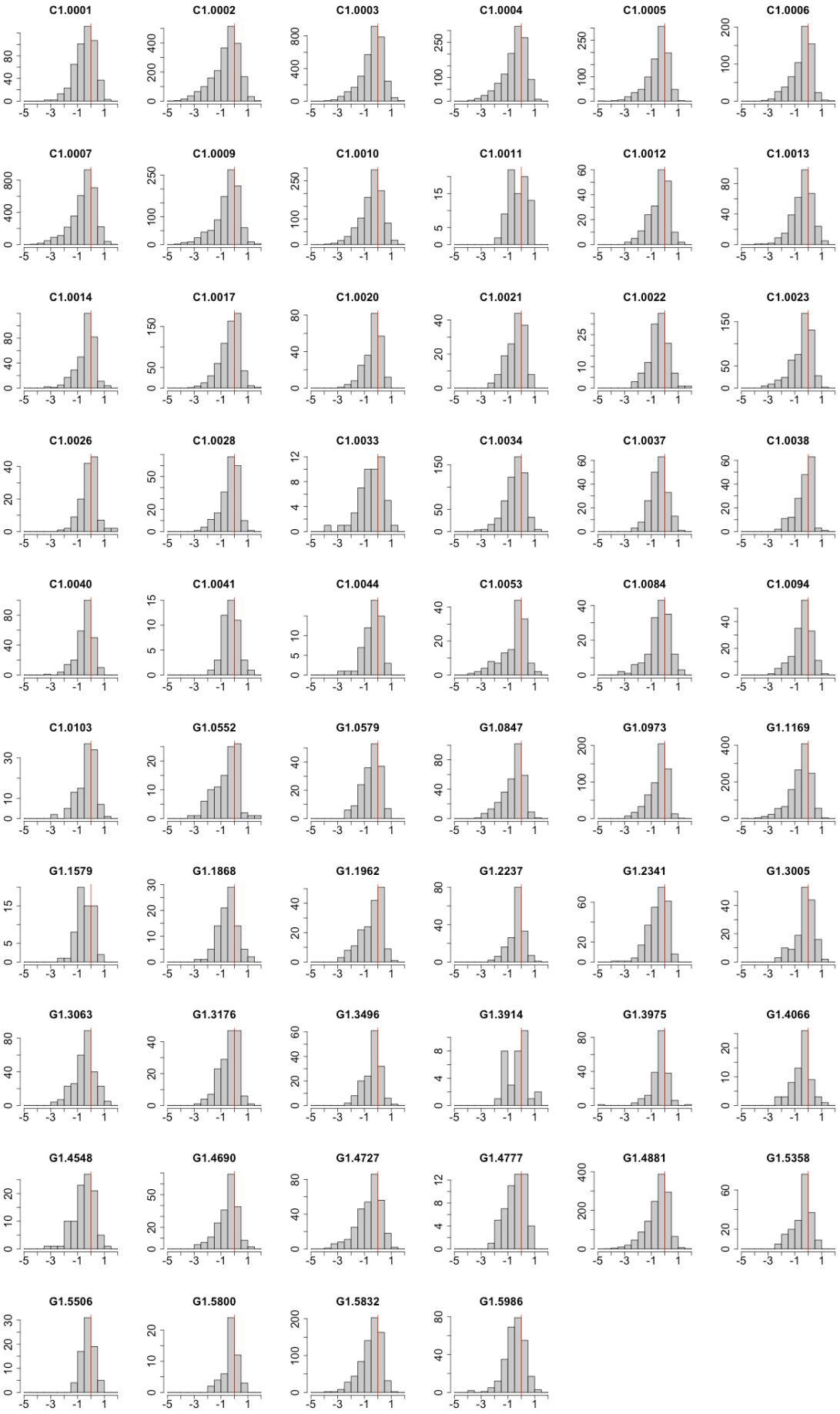
Step size

B



Generation

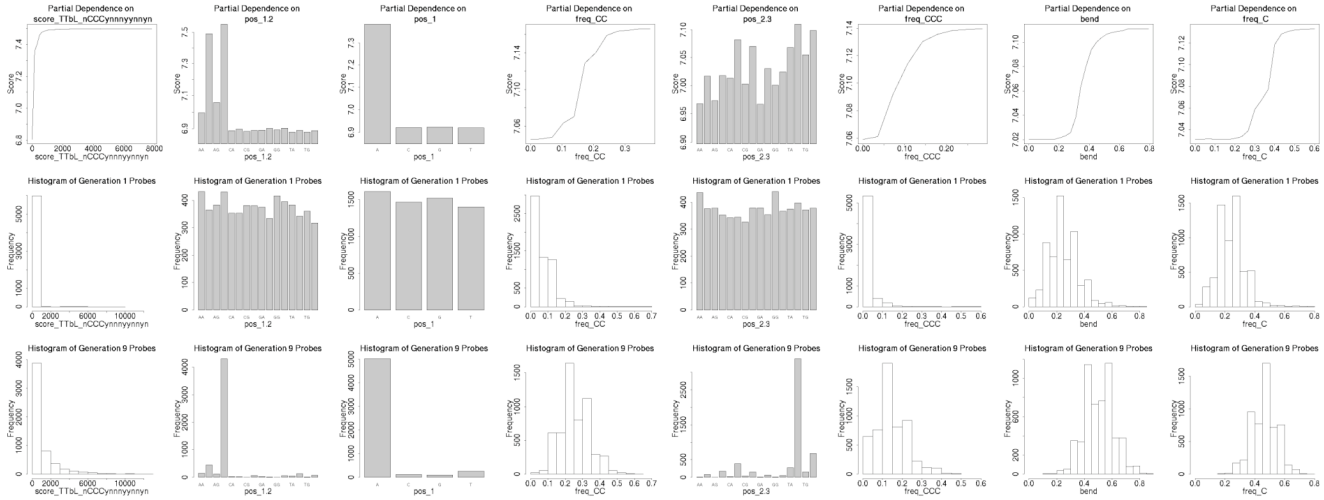
c



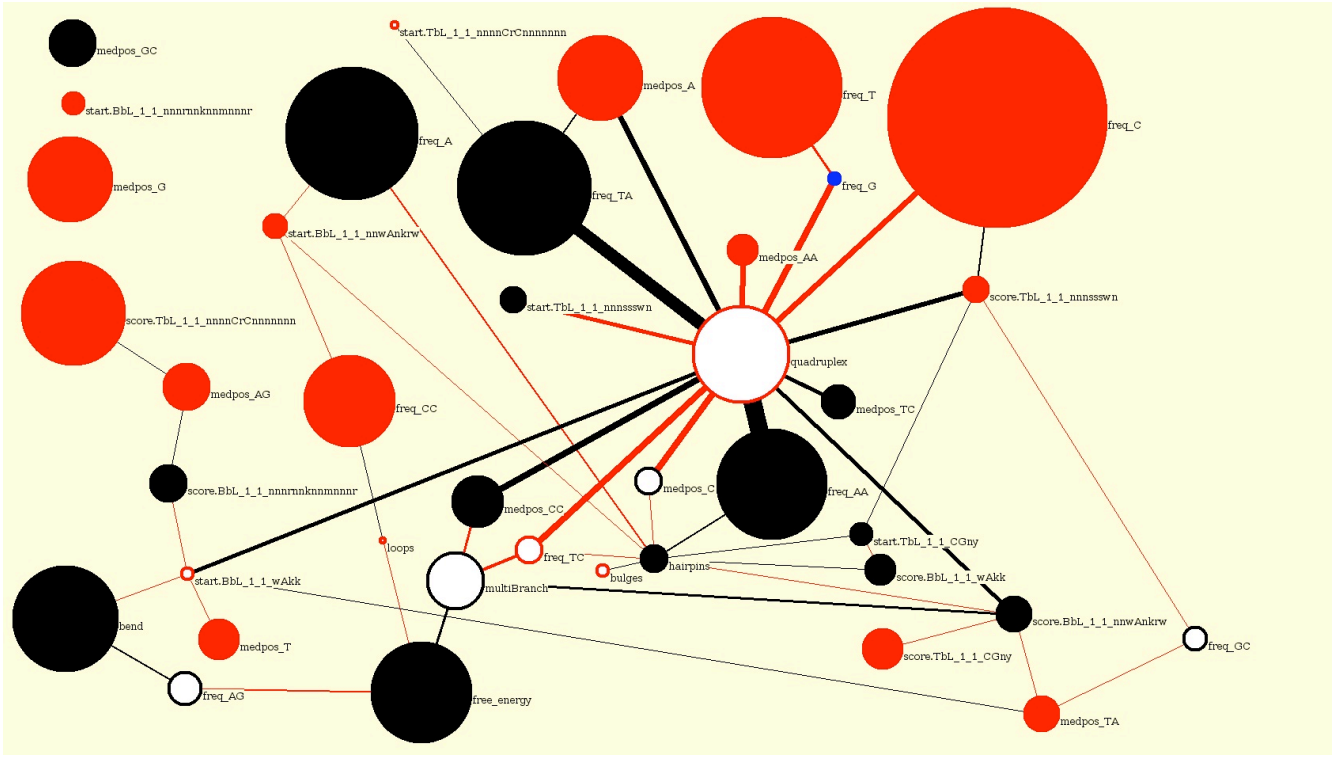
Frequency

Score change

Supplementary Figure 12. Score changes due to individual mutation events. All the values are the difference in score between one aptamer and its immediate ancestor in the previous generation, positive values indicate beneficial mutation events, negative deleterious, in each case, zero (a precisely neutral mutation) is indicated by a red line. (**A**) All score changes divided according to the step size- each step is made up of point mutations (step size of 1), an indels (step size of 2). The left-hand side gives the mean and 95% confidence interval of the score difference at each step size, the right-hand side the full distribution. The step size of 0 indicates sequences that are un-mutated between generations. (**B**) As in (**A**) except that the score changes are divided by generation, generation 1 being absent since these were the starting sequences without ancestors. (**C**) Histograms of score differences within the 58 individual lineages that are represented in the final generation.



Supplementary Figure 13. Effects of the most important explanatory variables in the Random Forest model as realised in the test set. For each variable three graphs are given- the partial dependence on the output of the Random Forest on the level of the given variable, the distribution of the variable in G1 and the distribution of the variable in G9. Lines and open bars correspond to continuous variables, shaded bars to discrete variables. Note the differing x scales for the continuous variable plots.



Supplementary Figure 14. Linear model used for designed probes. A complex form of sequence design was achieved by evolving sequences *in silico* to maximise the score predicted by the linear model shown. The size of circles indicates the absolute size of the linear coefficient applied to the given variable. Lines indicate 2-way interactions, the thicker the line the larger the interaction applied. Variable names are constructed in the same way as in Table S2. Red circles and lines indicate positive coefficients, black negative (the blue circle has no coefficient since, having included the frequencies of A, C and T, the frequency of G is already taken into account). White centres indicate that an effect was non-significant when the model was constructed.