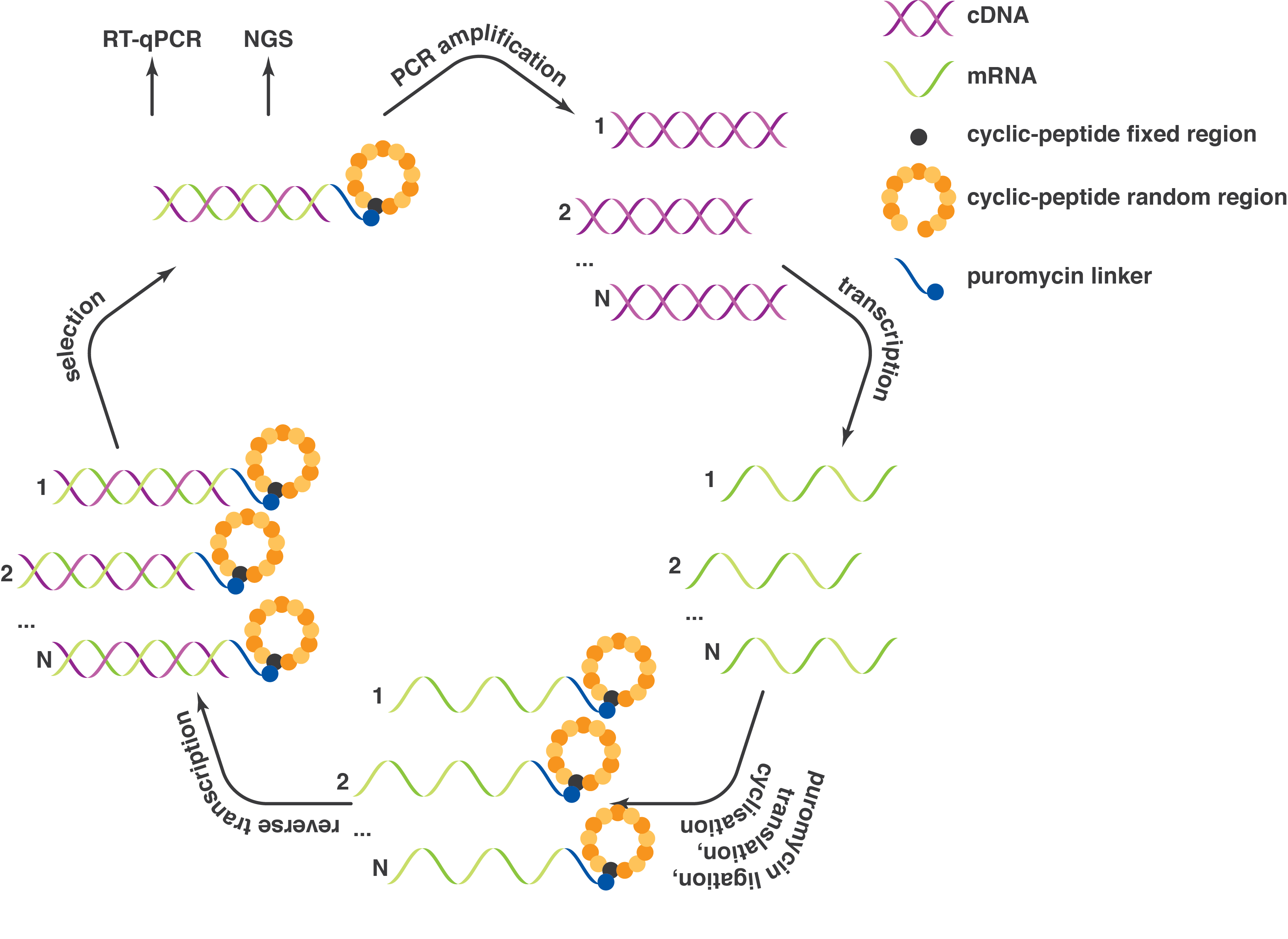
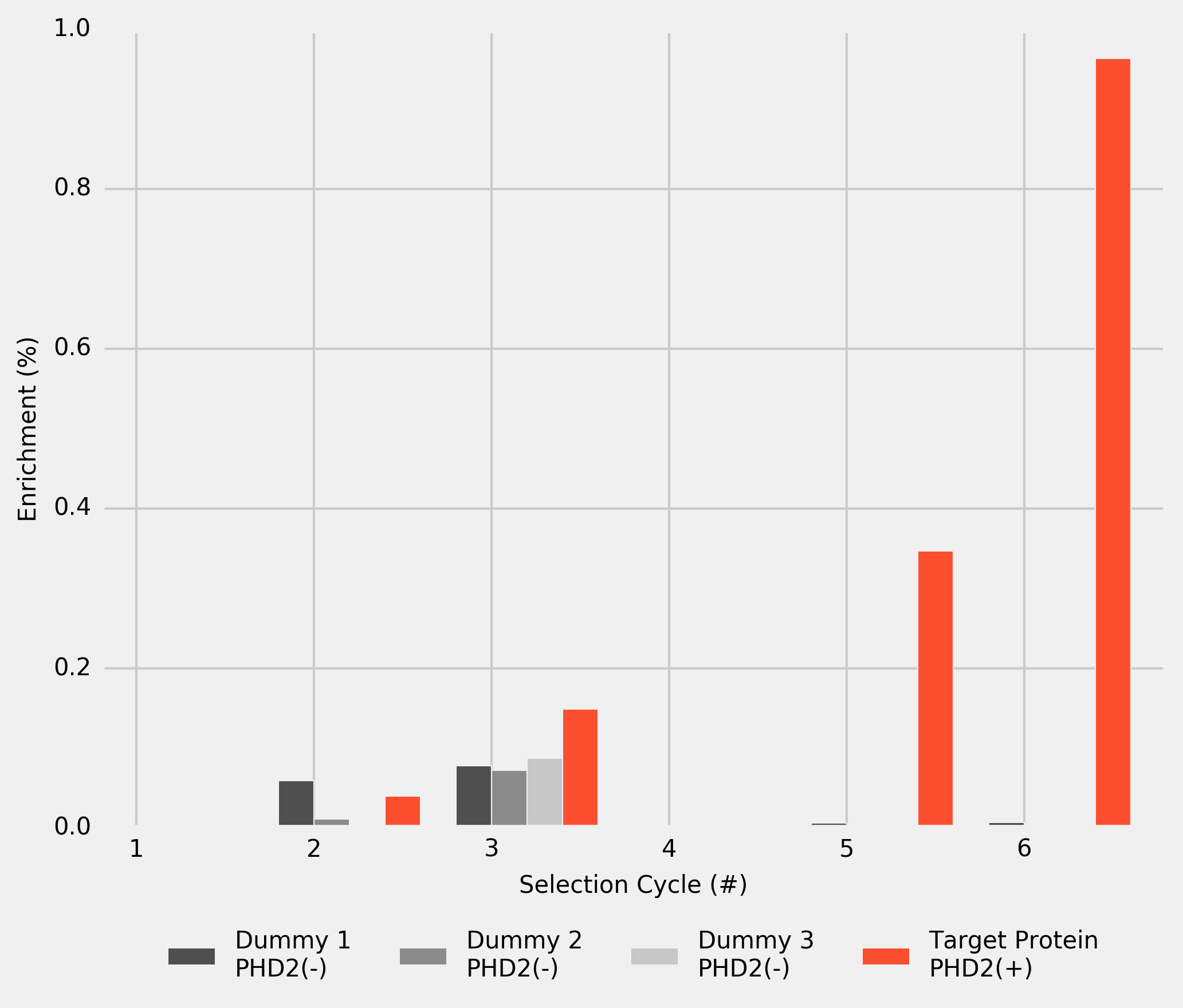
# Selection against human recombinant PHD2

The target protein — biotin-tagged human recombinant PHD2 — was immobilised on M-280 streptavidin magnetic beads (see Methods). Prior to selection the activity of immobilised PHD2 was confirmed using CODD peptide as a substrate (see Methods). Original cDNA library was assembled from five libraries combined at equimolar ratios with random region varying from (NNK)8 to (NNK)12 (cDNA total length varying from 111 to 123 nucleotides, respectively). Original library diversity was 1012-1014 unique cDNA sequences, each present as 1 to 100 copies.

Selection against PHD2 was organised in cycles (see Figure\_\_\_). The cDNA library was transcribed into an mRNA library, which was *in vitro* ligated to puromycin, translated, reverse-transcribed and cyclised, thus resulting in cDNA-mRNA-peptide library. Resulting cDNA-mRNA-peptide library was incubated with immobilised PHD2 admixed at approximately 1 to 1 molar ratio. The fraction of the library recovered after incubation was used for (1) cDNA library regeneration to use as an input in the next cycle of selection, (2) real-time quantitative PCR (RT-qPCR) assay, and (3) next generation sequencing (NGS) of the library. Starting the second cycle, to eliminate the peptides interacting with magnetic beads, prior to incubation with immobilised PHD2, three consecutive dummy incubations were performed. To control the progression of selection, the cDNA library recovered after every cycle was subjected to RT-qPCR and size of the recovered cDNA library related to the input was estimated. Selection process was judged complete (1) when the fraction of cDNA library recovered after incubation with immobilised target protein exceeded that recovered after dummy incubation, and (2) when the recovery reached approximately 1% of the input (see Figure\_\_\_). Achieving this involved 6 cycles of selection.



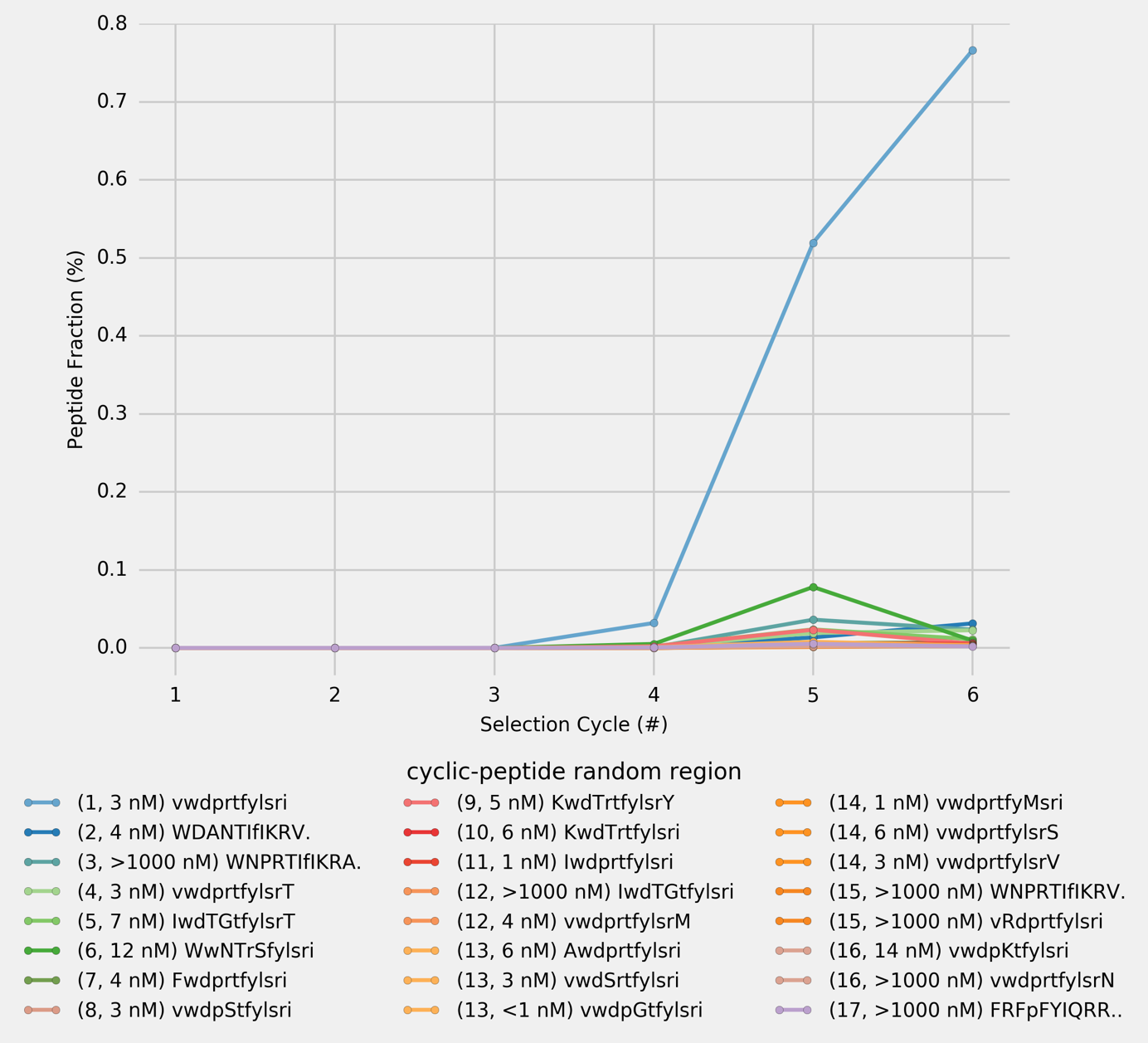
Figure\_\_\_. Schematic representation of a selection cycle. Selection cycle begins with cDNA library, which is transcribed into mRNA library, and subsequently ligated to puromycin, translated, cyclised and reverse-transcribed *in vitro*, thus resulting in cDNA-mRNA-peptide library. This library is incubated with immobilised target protein. The fraction of the library recovered after incubation is used for (1) cDNA library regeneration to use as an input in the next cycle of selection, (2) RT-qPCR assay, and (3) next generation sequencing (NGS) of the library.



Figure\_\_\_. cDNA-library recovery after dummy incubations 1 to 3 (M-280 streptavidin beads without PHD2), and after incubation with PHD2 immobilised on M-280 streptavidin magnetic beads. The results are based on the RT-qPCR assay and normalised by the cDNA input in every cycle.

# Analysis of Selection Results

After each cycle, a sample cDNA library was subject to NGS. The resulting cDNAs were *in silico* translated into peptides and, to estimate the enrichment of individual peptides, the instances of unique peptides were counted at each cycle. After the final cycle, 206 unique peptide sequences were identified. These peptides were ranked based on their enrichment in final cycle (rank 1 being the most enriched peptide). 24 top-ranked peptides were selected and their enrichment was plotted for every selection cycle (see Figure\_\_\_).

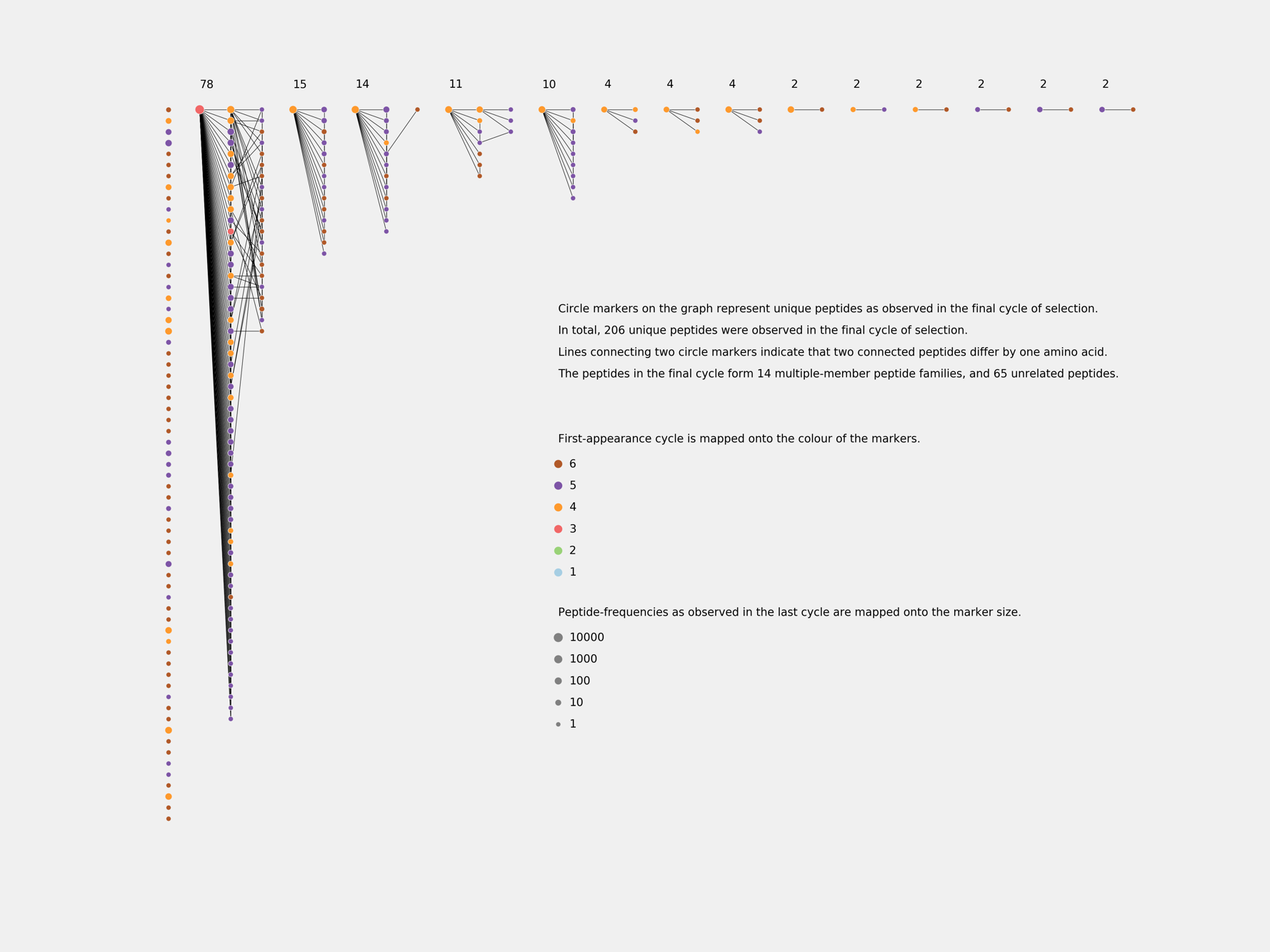


Figure\_\_\_. Peptide enrichment at the end of each cycle for the 24 most frequent unique peptides as observed after the final cycle of selection (Cycle # 6). The peptide sequences are formatted so that amino acids in each position, if different from the top peptide (vwdprtfylsri), are capitalised; if the peptide sequence is shorter, the missing amino acid is indicated by the ‘.’ at the end of the sequence. Peptides rank and their KD are provided in the brackets after the random-region peptide sequence.

As selection progresses the library is being enriched in one top peptide (vwdprtfylsri, KD = 3 nM), which constitutes over 75% of the library in the final cycle. To confirm that the top peptide has the highest affinity towards target protein, and to verify a correlation between peptide enrichment after the final selection cycle and their KD values, top 24 peptides were chemically synthesised and KD values of their complexes with PHD2 were measured using surface plasmon resonance (SPR; see Figure\_\_\_). The KD values of the top 24 peptides cover the range from below 1 nM to over 1000 nM. Importantly, several of the less enriched peptides were found to have higher affinity towards the target protein than the top peptide. No correlation was found between peptide enrichment and KD. Notably, 16 out of top 24 peptides differ from the top peptide and one another by no more than two amino acids, with KD values differ by approximately 3 orders of magnitude. Noteworthy, peptide ranked 3 has KD higher than 1000 nM and is followed in ranking by peptides with KD in the lower nM region. These findings indicate, that a peptide enrichment is not exclusively regulated by its affinity towards target protein *(i.e.* is not primarily driven by ligand’s phenotype, but results from a more complex selection mechanism).

# mRNA-Display Phylogenetics

After the final cycle of selection, library sequencing produced 5601 reads, with 505 unique DNA sequences encoding for 206 unique peptide sequences. Many of these peptides are structurally similar. To verify whether structurally similar peptides found in the final cycle of selection have their origin in the original library or result from mutation of some originally present sequences, a graph of the unique peptide sequences found after the final cycle of selection was constructed. Two unique DNA sequences were joined together if these two sequences differ by one nucleotide. Thus, the unique cDNA sequences were arranged into a set of disjoint groups. cDNA sequences in each group were translated into peptide sequences *in silico* and the unique peptide sequences within such groups were joined together if two peptides differ by one amino acid. The resulting graph was visualised as a scatterplot, where each circle marker represents a unique peptide sequence and each line connecting any two circles indicates that two connected peptide sequences differ by one amino acid (see Figure\_\_\_). For detailed analysis see Supplementary\_\_\_ showing peptide sequences and their characteristics.



Figure\_\_\_. Phylogenetic trees of unique peptides observed in the final cycle of selection (Cycle # 6) against PHD2 using cyclic-peptide library, based on cDNA (NNK)8-12 random library.

Peptides which appear earlier in selection are more frequent, and are encoded by a greater number of DNA mutants, some of which are not to be found in the original NNK library. Given the constant rate of cDNA mutation in the process of PCR amplification, it is conceivable that the later appearing DNA sequences have fewer DNA mutants. These results suggest, that in a pair of peptides distinguishable by only one amino acid, the more frequent and earlier appearing peptide is likely to be genetically antecedent to the lest frequent later appearing peptides, thus, evincing genetic rather than accidental structural relatedness of the peptides.

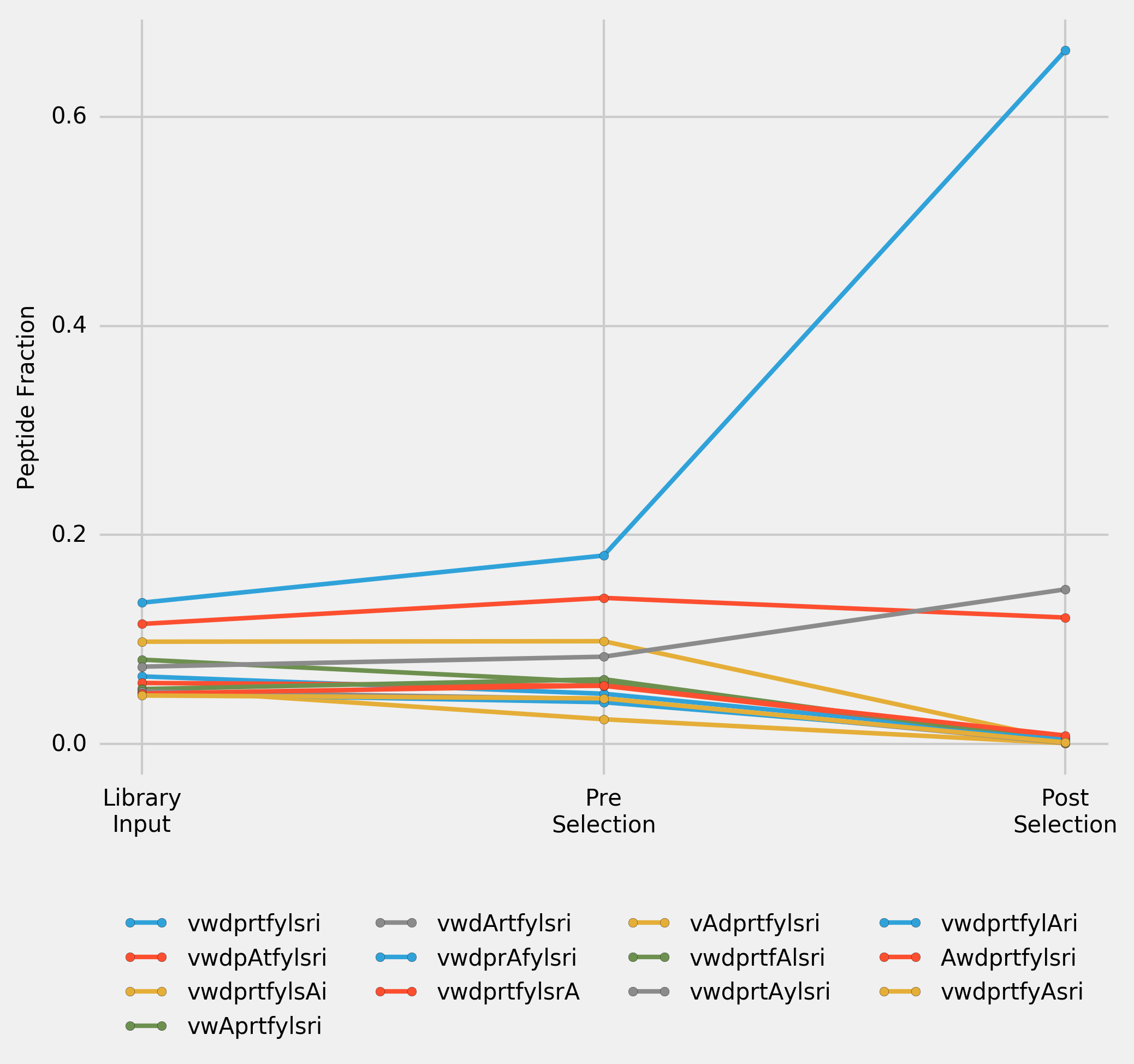
Of the 505 unique cDNA sequences found in final cycle, 361 (71%) have mutations beyond original NNK library. Crucially, some mutant peptides persist through selection and are highly enriched despite their low affinity towards target peptide (KD >1000 nM). Thus, secondary diversity, resulting from mutation of original sequences, does not reflect the affinity of either individual peptides or a consensus peptide towards target protein, and may bias the analysis in favour of the most prolific peptide family constituting 78 out of 206 (35%) unique peptide sequences. The top peptide is encoded by 82 cDNA mutants, of which 78 have mutations are not in the original NNK library and are evidently extrinsic to the original library.

# Most Abundant Peptide Alanine Scan

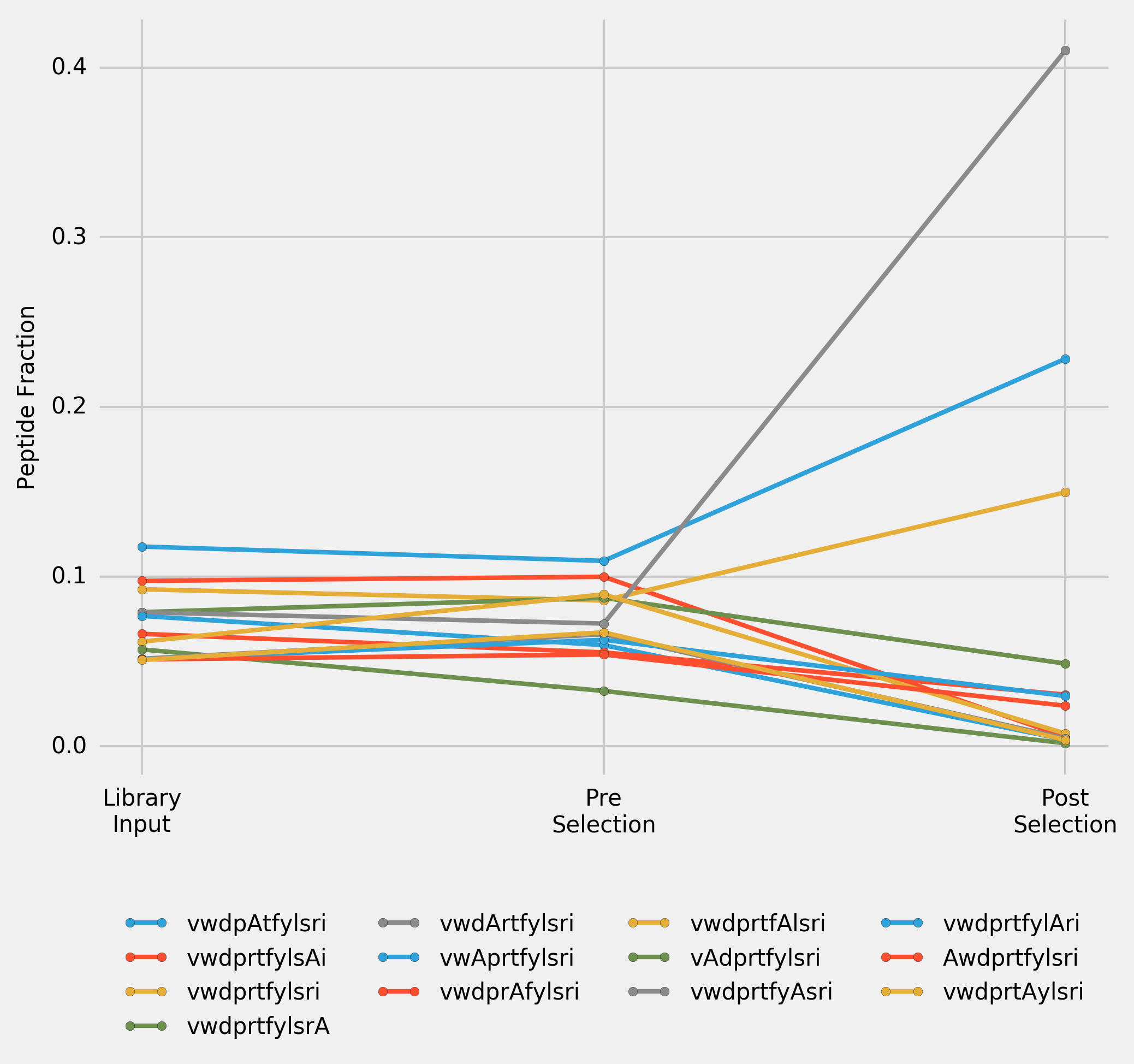
To provide further insight into the mechanism of the selection, one-cycle selections were carried out with two cDNA libraries comprised of 13 sequences. In both libraries (1) and (2) 12 cDNAs — encoding for the alanine-scan peptides of the most frequent peptide in original selection — were designed based on the most frequent E. coli codons. Last cDNA — encoding for the most frequent peptide in original selection — was either (1) the original cDNA sequence recovered from selection (Figure\_\_\_), or (2) one designed based on the most frequent E. coli codons (Figure\_\_\_). Both selections have been performed with and without dummy incubations, each selection was repeated twice and all provided essentially identical results.

cDNAs based on the most frequent E. coli codons differ from one another by no more than 3 nucleotides and A/T count of their random region varies between 13 and 16. The original cDNA sequence recovered from selection differs from the other cDNAs by up to 12 nucleotides, with A/T count equal 21.

Single-round selection using library (1) produced the same top peptide (enriched over 60%) as the original selection from random library. In selection using library (2) the same peptide was enriched only to 15% and ranked 3. Minimising the differences in cDNA not only affects the enrichment of individual peptides, but also changes the ranks of peptides, and previously top peptide sequence becomes third. Changing the genotype of only one peptide affects not only the enrichment of this peptide, but also the outcome of the whole selection. KD values for chemically synthesised alanine-scan peptides were assayed using SPR, and compared with results of such selections (Table\_\_\_).



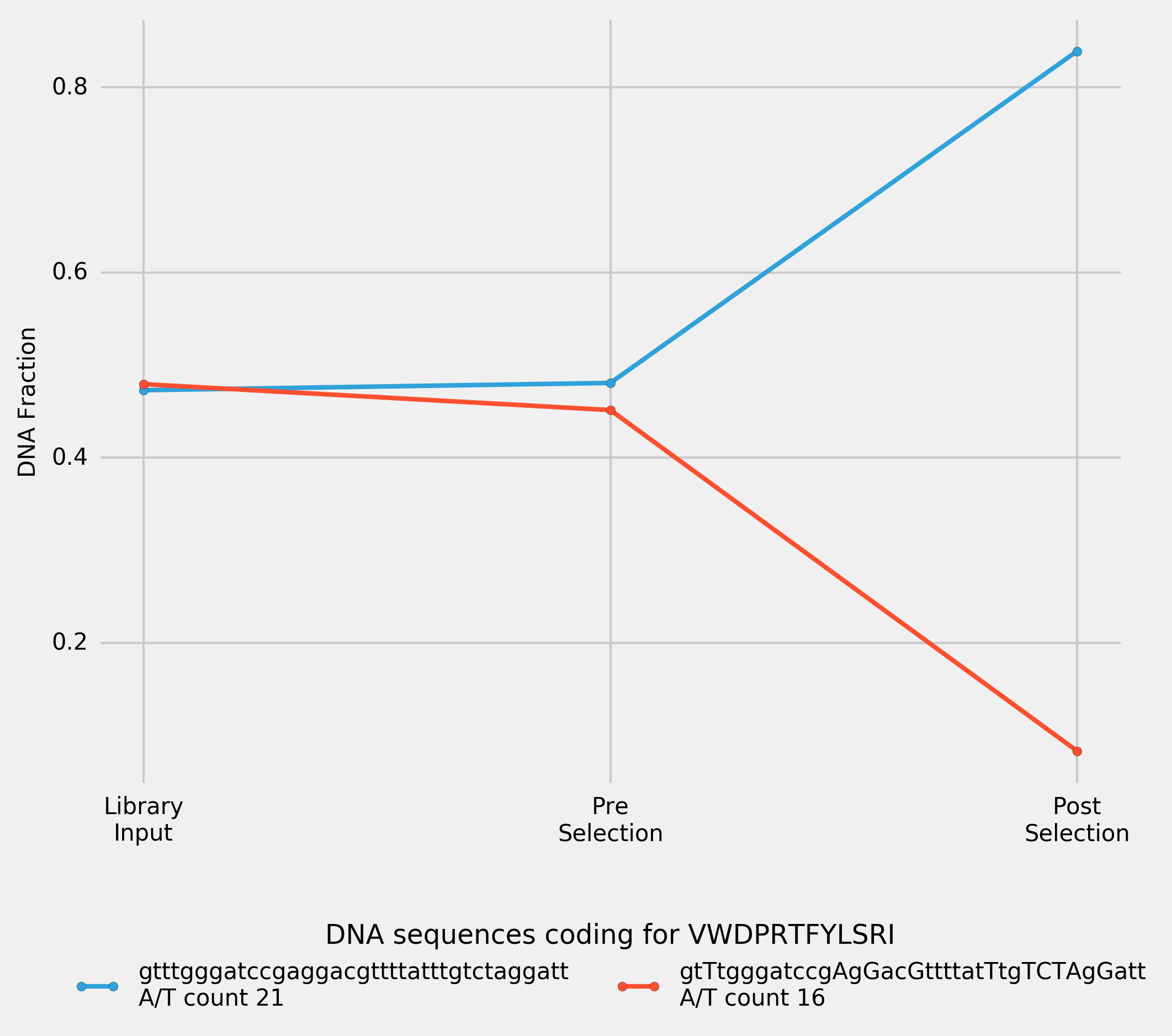
Figure\_\_\_. Peptide enrichment over one cycle of selection for a library (1) of 13 cDNAs. 12 cDNAs — encoding for the alanine-scan peptides of top peptide in the original selection — were designed based on the most frequent E. coli codons, last cDNA — encoding for the top peptide in original selection — is the original cDNA sequence recovered from selection. Input Library was transcribed, puromycine ligated, translated and reverse transcribed to convert it into Pre-Selection Library. Post-Selection Library was the library recovered after incubation of Pre-Selection Library with immobilised PHD2. Incorporated alanines in the peptide sequences are highlighted using capitalisation.



Figure\_\_\_. Peptide enrichment over one cycle of selection for a library of 13 cDNAs. 12 cDNAs (encoding for the alanine-scan peptides of top peptide in the original selection) and cDNA (encoding for the top peptide in original selection) were designed based on the most frequent E. coli codons. Input Library was transcribed, puromycine ligated, translated and reverse transcribed to convert it into Pre-Selection Library. Post-Selection Library was the library recovered after incubation of Pre-Selection Library with immobilised PHD2. Incorporated alanines in the peptide sequences are highlighted using capitalisation.

# Direct Comparison of the Original Sequence

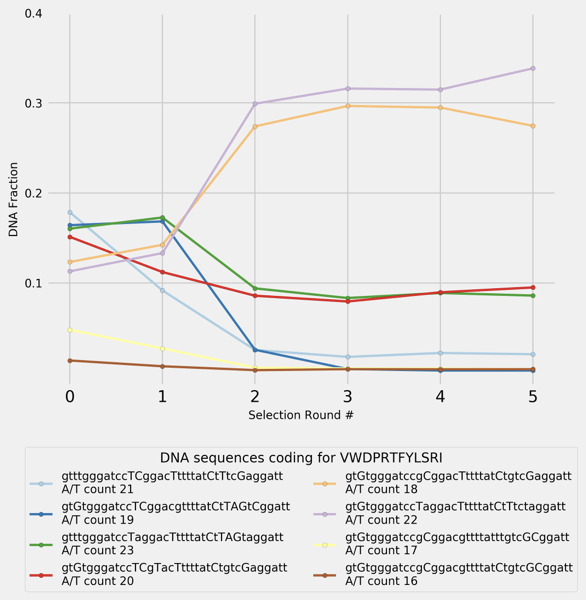
To directly compare the impact of cDNA on the selection outcome one-cycle selection was performed with a library comprised of two cDNAs encoding for the most frequent peptide in original selection: (1) the original cDNA sequence recovered from selection, and (2) one designed based on the most frequent E. coli codons.



Figure\_\_\_. Peptide enrichment over one cycle of selection for a library of 2 cDNAs. Both cDNAs are encoding for the top peptide in the original selection (vwdprtfylsri): one is the original cDNA sequence recovered after selection, another is designed based on the most frequent E. coli codons. Nucleotides which differ the original sequence are highlighted using capitalisation. Input Library was transcribed, puromycine ligated, translated and reverse transcribed to convert it into Pre-Selection Library. Post-Selection Library was the library recovered after incubation of Pre-Selection Library with immobilised PHD2.

# PCR Amplification Efficiency

To verify whether PCR amplification efficiency is A/T content dependent, 8 (NNK)12 sequences with A/T content varying from 16 to 23 were picked at random and used for the mock-selection test, in which the equimolar library was PCR-amplified for 120 cycles, and the samples were collected for the NGS sequencing after every 20 cycles of PCR (Figure\_\_\_).



Figure\_\_\_. PCR-amplification of 8 randomly selected (NNK)12 cDNAs encoding for the top peptide in the original selection (vwdprtfylsri). Every 20 cycles of PCR, part of the sample was used for the following amplification, and the rest was collected for NGS sequencing. 0-point shows the composition of the original input library.

The enrichment in two sequences was achieved by performing PCR amplification alone (without selection) for as few as 40 cycles, for a very simple library. For a more complex library, it would take more cycles of PCR to produce this level of enrichment. It appears that A/T content does not have significant influence on the PCR-amplification efficiency, and some more complex mechanism, such as forming intramolecular folds, may be involved.