# mRNA-Display Mechanism And Its Implications For The Results Analysis And Interpretation

# Summary

mRNA display is a very potent technique which enables serendipitous discovery of functional peptide-ligands for target enzymes. To date, however, the mechanism of this technique remains vaguely understood, which may result in biased experimental design and the interpretation of the results. This paper provides a detailed analysis of a peptide-ligand selection using mRNA display, and suggests a thermodynamic model, which can be used to develop and optimise mRNA display, provides and algorithm aimed at reducing the bias of mRNA display introduced in the selection process. It also suggests some potential improvements to the experimental design and analysis routine.

# Keywords

mRNA display, PHD2, RaPID system, thermodynamic model, PCR bias

# Introduction

DNA-encoded peptide libraries opened a new way to search for peptide ligands with useful properties — drug-like compounds{Goldflam:2015ju}, protein inhibitors, protein activators, or co-crystallisation chaperons{Hipolito:2014bg}. Selection of peptide ligands from such libraries is based on iterative algorithm: (1) incubation of a DNA-tagged peptide library with a target protein, (2) recovery of DNA-tagged ligands, and (3) regeneration of a less diverse library from the recovered DNA tags{Goldflam:2015ju}. Selection iterations are repeated until library diversity reduces sufficiently to make its analysis possible. Identified peptides are subjected to verification of their complex formation with the target protein and identification of their potentially useful properties. Selection is termed ‘successful’ if amongst thus recovered peptides, some demonstrate high affinity towards the target protein (kD < 10-7 M).

Based on such selection algorithm, in the last three decades, phage display and mRNA display were developed. These are of particular interest because they enable selection from highly diverse DNA-encoded peptide libraries (approximately 109 for phage display to 1012 for mRNA display). Successful selections using display techniques produced a number of useful peptide ligands. Unfortunately, many selections are failing to produce a ligand for a target protein{Menendez:2005ck}{Vodnik:2011iy}. This has most commonly been attributed (1) to interference of the target protein with the phage, cDNA or mRNA part of a ligand or (2) to individual cDNA sequences amplification efficiencies.

A peptide-ligand recovery is thought to be regulated by its ability to form a complex with a target protein; and for the sake of the analysis, ligand recovery is taken directly proportional to its affinity towards target protein (*i.e.* ligand-target complex KD value). Thus, it is often concluded that peptides recovered in higher proportion have higher affinity towards target protein. However, without verification of the mechanism of selection this conclusion may result in less enriched peptides, yet having lower KD, being disregarded.

Recovery of structurally similar peptide ligands is often viewed as evidence for a particular consensus sequence being favoured in the selection (*i.e.* a complex between target protein and such consensus peptide ligand having lower KD). Often based on the recovered peptide sequences, a consensus-sequence is constructed or conserved motif is identified, and such sequences are further tested for useful properties{Takahashi:2003ik}. However, if the structurally similar peptides recovered in a selection are genetically related (*i.e.* are mutants of one prevalent peptide), rather than accidentally present in the original library, such approach would produce a heavily biased consensus sequence or conserved motif, which may not have high affinity towards the target protein.

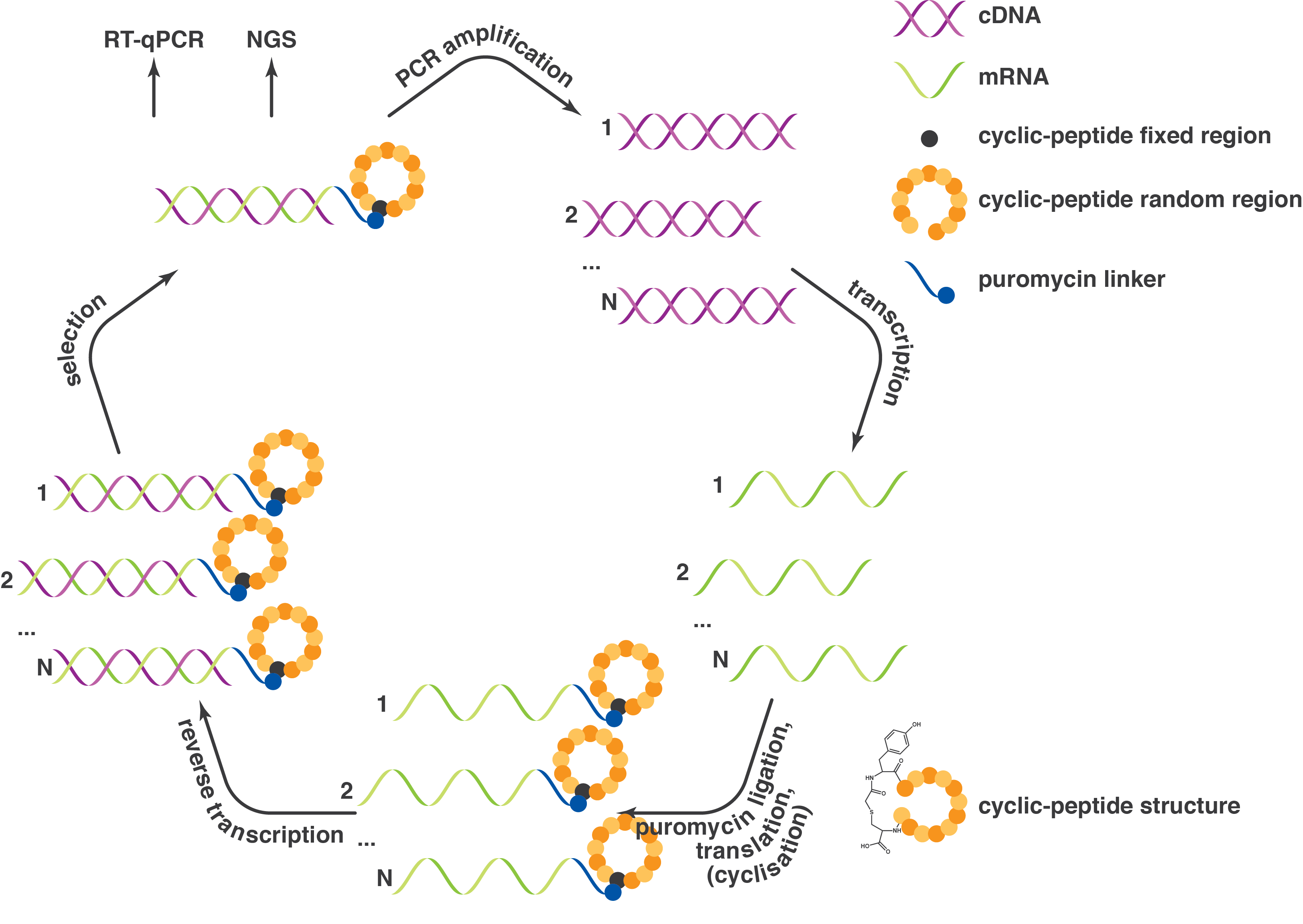
The analysis of display results relies on two assumptions: (1) the recovery of a peptide ligand in the process of selection is directed by and is proportional to its affinity towards a protein target, and (2) the amplification of DNA tags does not affect the outcome of selection process. To date, the former assumption has been invalidated for both phage displays{Derda:2011fy} and mRNA displays{JalaliYazdi:2016fg}. However, it remains unclear to which extent the results are biased by the DNA-tag amplification, and how the results of the selection are to be interpreted.

This paper provides detailed analysis of a successful selection against human recombinant PHD2 from a cyclic-peptide library performed using RaPID system, a version of mRNA display based on (1) the flexible in vitro translation (FIT) of an mRNA library, (2) mRNA-to-peptide ligation using puromycin, and (3) peptide cyclisation using the non-proteinogeinc amino acids incorporated into the peptide chain. This work verifies the effect of the cDNA tags on the selection results and introduces a robust approach to the selection-results analysis. In particular, this paper suggests a thermodynamic mechanism of selection that informs on to which extent the selection process is defined by the ligand-to-target affinity, highlights potential limitations of mRNA display and indicates potential ways of protocol improvements.

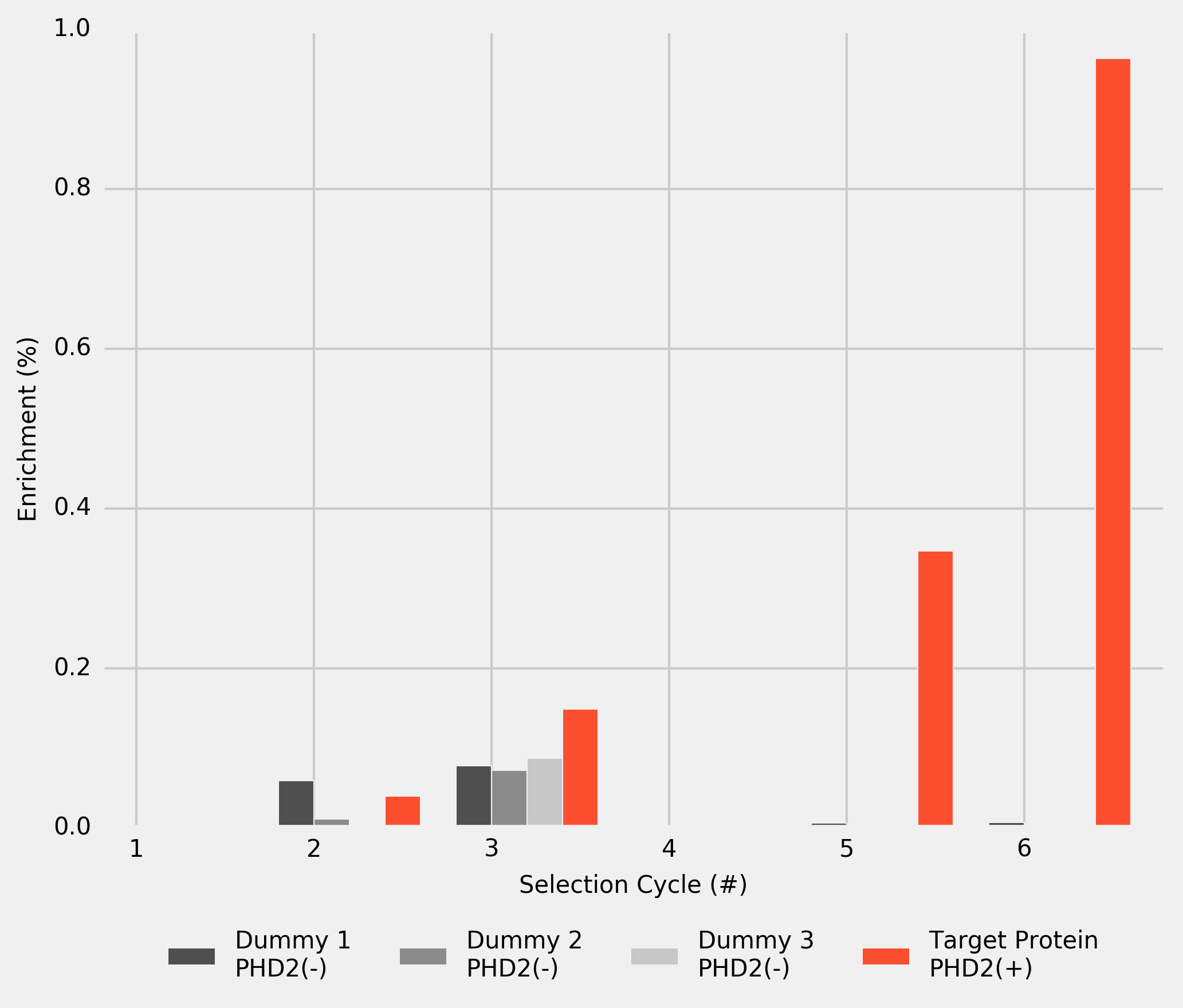
# Selection against human recombinant PHD2

Selection against the target protein — biotin-tagged human recombinant PHD2 — was performed according to the previously published protocol{Hayashi:2012bb} with minor modifications. 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 estimated 1012-1014 unique cDNA sequences, each present at 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 (using Taq DNA polymerase) 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 from the second cycle, to eliminate the peptides interacting with magnetic beads, prior to incubation with immobilised PHD2, three consecutive dummy incubations (without PHD2) were performed. To control the progression of selection, the cDNA library recovered after every cycle was subjected to RT-qPCR and quantity of the recovered cDNA library relative 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.



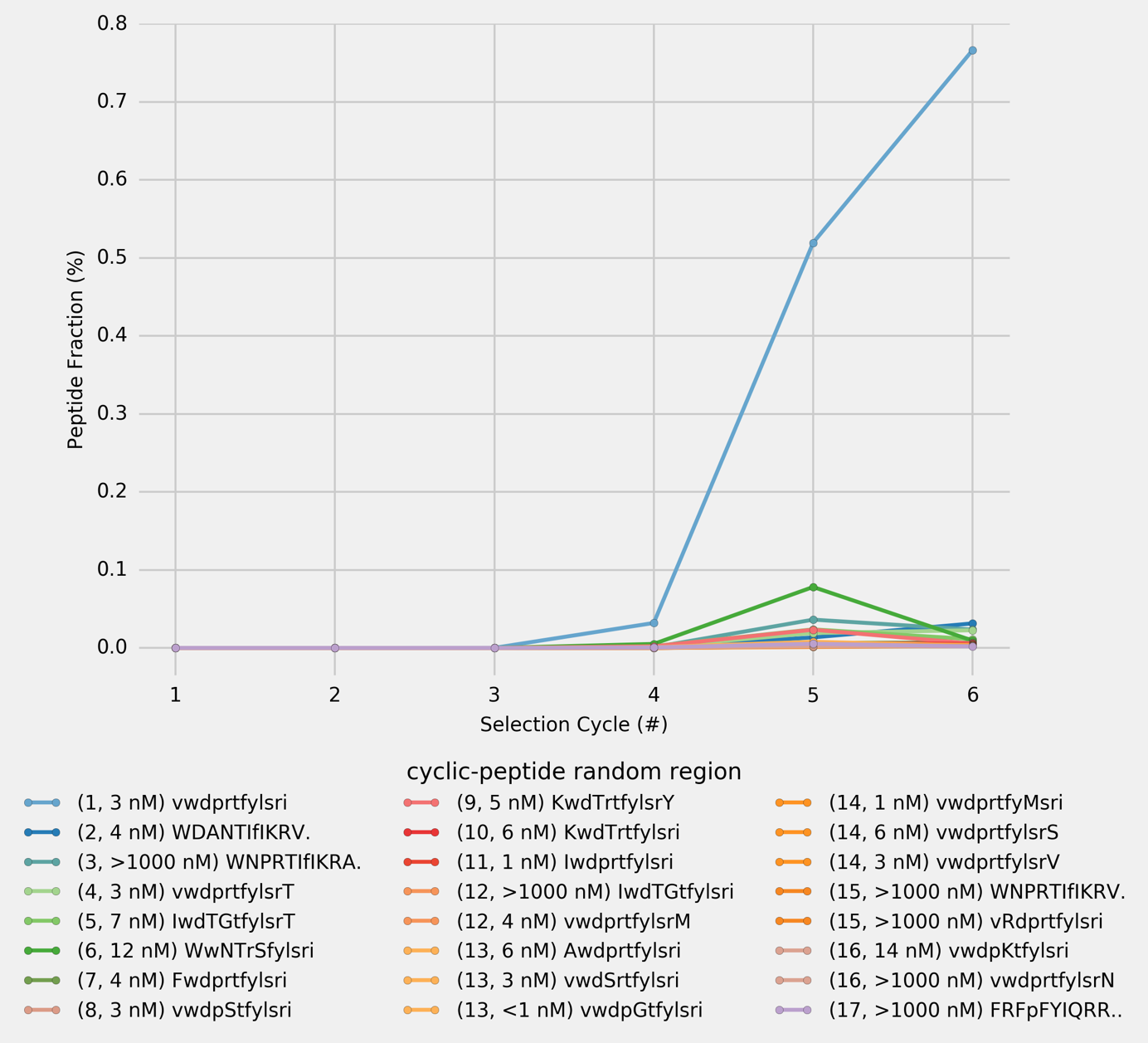
Figure\_\_\_. Schematic representation of a selection cycle. Selection cycle begins with cDNA library, which is transcribed into mRNA library, and subsequently ligated to puromycin, translated, cyclised and reverse-transcribed *in vitro*, thus resulting in cDNA-mRNA-peptide library. This library is incubated with immobilised target protein. The fraction of the library recovered after incubation is used for (1) cDNA library amplification using Taq DNA polymerase (thus regenerated library serves 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 of cDNA library was subjected to NGS. The resulting cDNAs were *in silico* translated into peptides and, to estimate the frequency 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 are ranked based on their frequency in final cycle (rank 1 being the most enriched peptide). 24 top-ranked peptides were selected and their frequency was plotted for every selection cycle (see Figure\_\_\_).

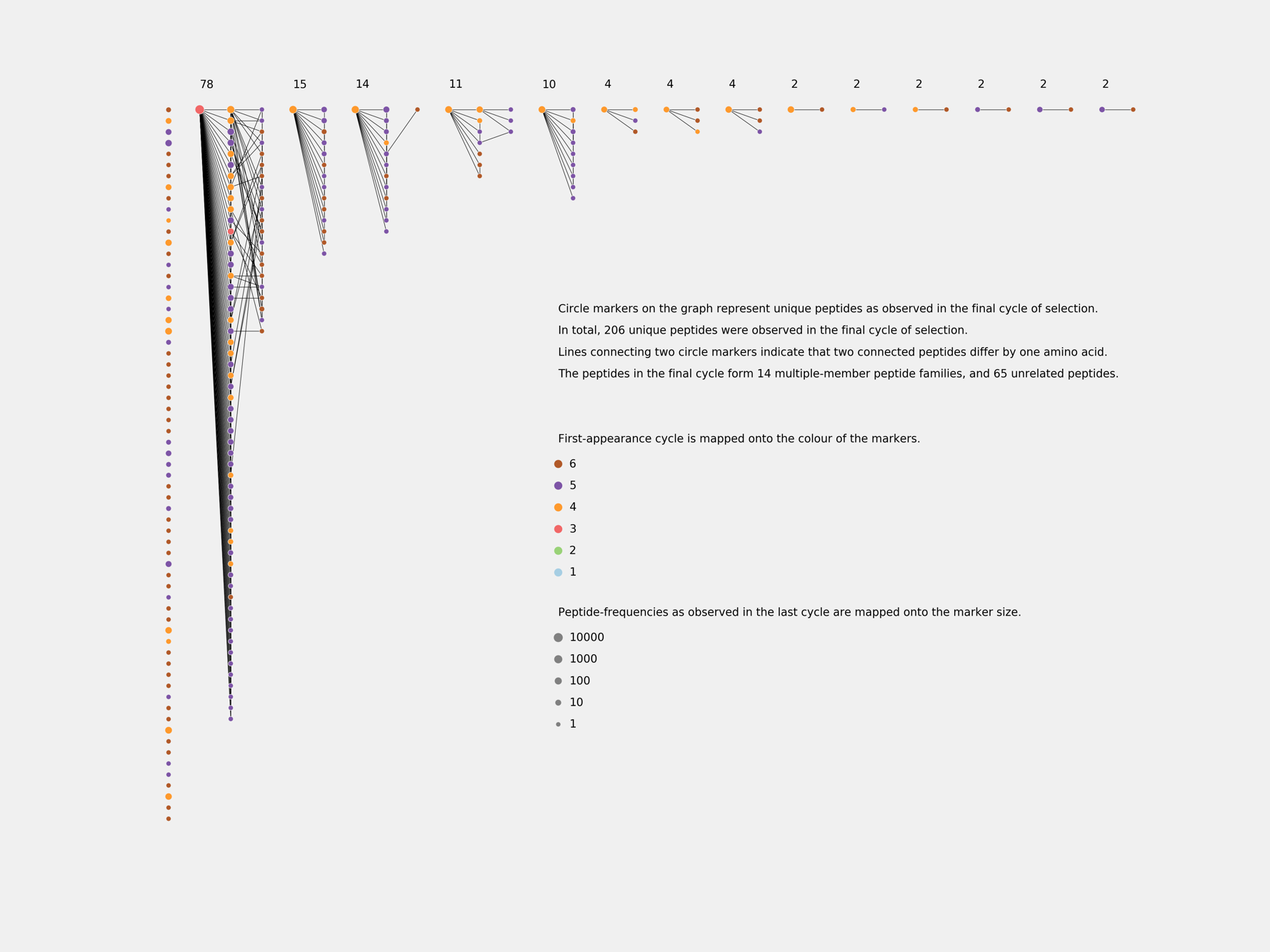


Figure\_\_\_. Peptide frequency at the end of each cycle for the 24 most frequent unique peptides as observed after the final cycle of selection (Cycle # 6). The peptide sequences are formatted so that amino acids in each position, if different from *the top-ranking peptid*e (with random region vwdprtfylsri), 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 peptide (with random region vwdprtfylsri, and KD = 3 nM; referred to as *the top-ranking peptide* throughout the paper), which constitutes over 75% of the library after the 6th cycle. To confirm that the top-ranking peptide has the highest affinity towards the target protein, and to verify a correlation between peptide frequency 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 even higher affinity towards the target protein than the top-ranking peptide. No correlation was found between peptide frequency and KD. Notably, 16 out of top 24 peptides differ from the top-ranking 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 the frequency of a peptide is not exclusively regulated by its affinity towards a 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 in earlier selection cycles 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 peptide, thus, evincing genetic rather than accidental structural relatedness of the two peptides.

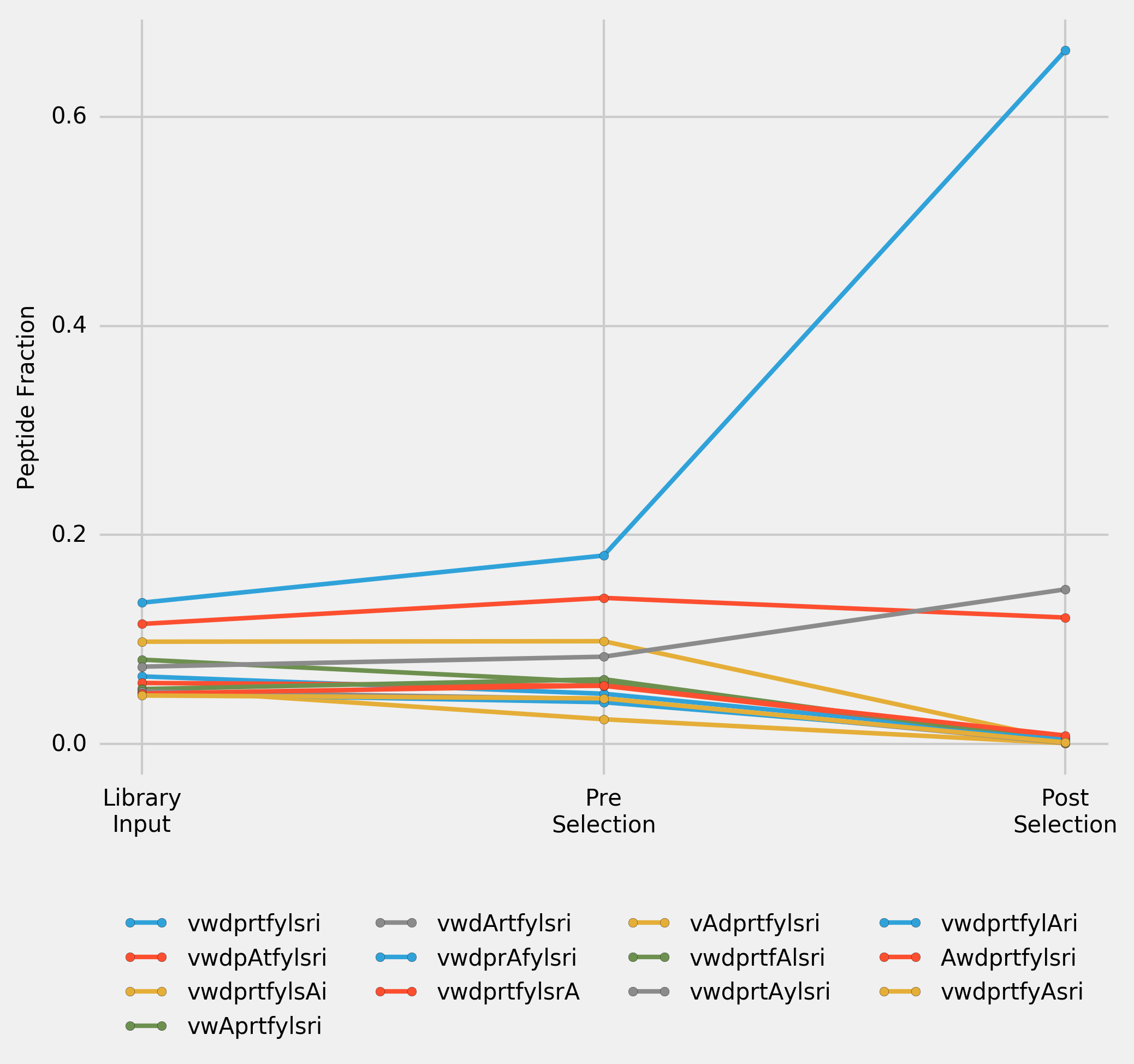
Of the 505 unique cDNA sequences found in final cycle, 361 (71%) have mutations beyond original NNK library. Crucially, some mutant peptides persist through selection and are highly enriched despite their low affinity towards target peptide (KD >1000 nM). Thus, secondary diversity, resulting from mutation 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-ranking peptide is encoded by 82 cDNA mutants, of which 78 have mutations extrinsic to the original NNK library and are evidently introduced later in the selection process.

# Most Abundant Peptide Alanine Scan

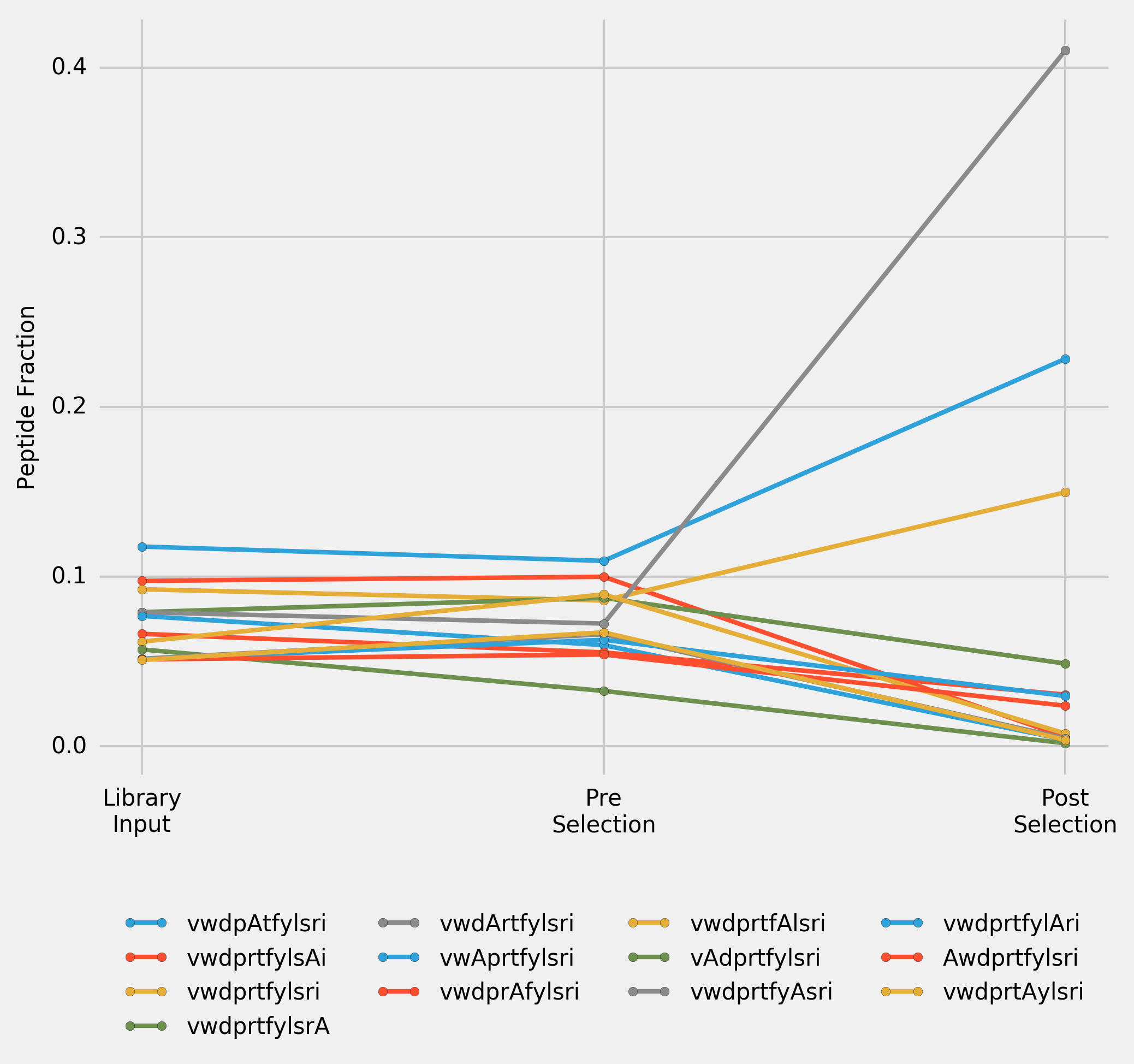
To provide further insight into the mechanism of the selection, one-cycle selections were carried out with two cDNA libraries comprised of 13 sequences. In both libraries (1) and (2) 12 cDNAs — encoding for the alanine-scan peptides of the most frequent peptide in original selection — were designed based on the most common 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 frequency 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 frequency 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\_\_\_).



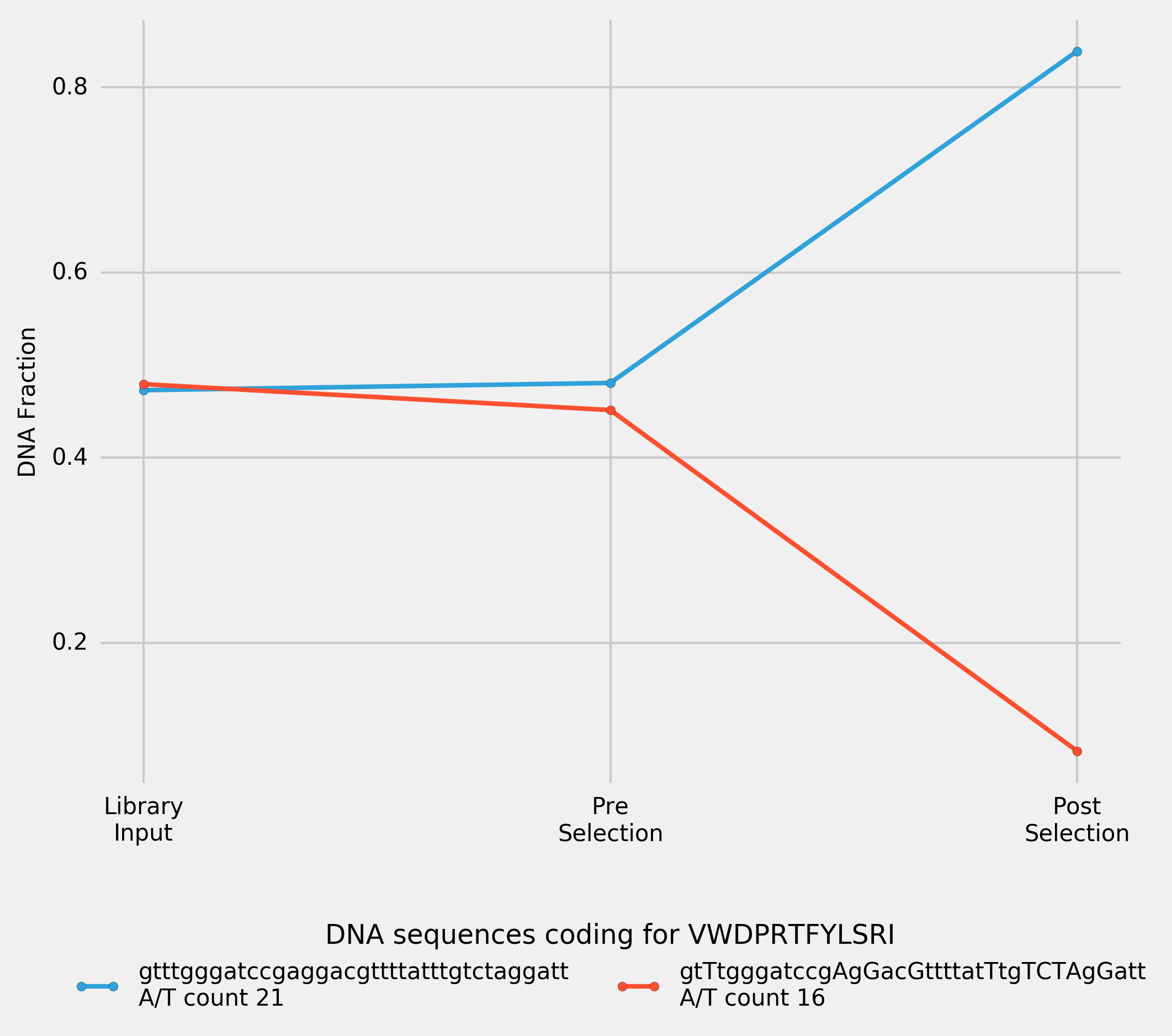
Figure\_\_\_. Peptide frequency over one cycle of selection for a library (1) of 13 cDNAs. 12 cDNAs — encoding for the alanine-scan peptides of the top-ranking peptide in the original selection — were designed based on the most frequent E. coli codons, last cDNA — encoding for the top-raning peptide 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 frequency over one cycle of selection for a library of 13 cDNAs. 12 cDNAs (encoding for the alanine-scan peptides of the top-ranking peptide in the original selection) and cDNA (encoding for the top-ranking 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

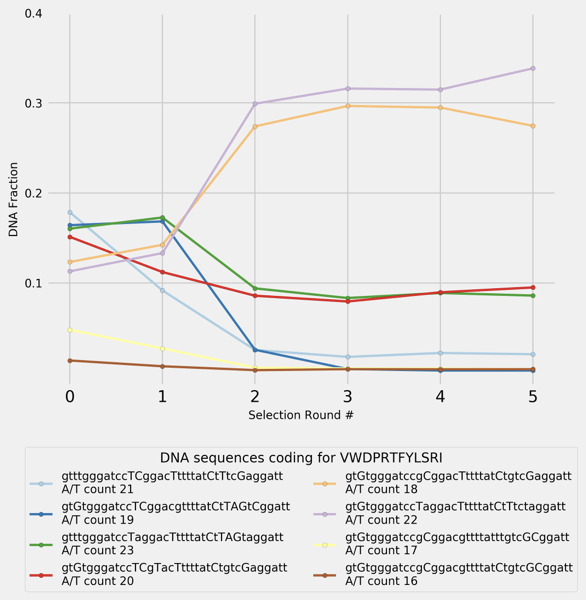
To directly compare the impact of cDNA on the selection outcome, one-cycle selection was performed with a library comprised of two cDNAs encoding for the most frequent peptide in original selection: (1) the original cDNA sequence recovered from selection, and (2) one designed based on the most frequent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



Figure\_\_\_. Peptide frequency over one cycle of selection for a library of 2 cDNAs. Both cDNAs are encoding for the top-ranking 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-ranking 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 frequency 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 frequency. 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.

# Discussion

In the process of selection, cDNA-mRNA-peptide ligands in the library undergo two distinct types of reactions, controlled by different parts of the fusion ligand: (1) complex formation between the ligand and the target protein (controlled by the peptide part of a ligand, and is characterised by KD of ligand-target complex), and (2) PCR amplification (controlled by the cDNA part of a ligand, and characterised by cDNA amplification efficiency). The former follows linear law and is performed approximately 10 times during selection, while the latter follows exponential law and is performed approximately 150 times during selection. Close examination of the selection process suggests that the frequency in certain ligands is primarily driven by PCR amplification rather than complex formation with target protein, *i.e.* the process is overall driven by the genotype rather than phenotype of the cDNA-mRNA-peptide ligands. It remains clear however, that ligand-target binding does have influence on the selection.

Unfortunately, currently existing experimental methods do not allow one to closely monitor selection from the libraries with diversities 109 to 1012 unique ligands. Instead, to better understand to which extent the complex formation between a ligand and a target protein affects the selection outcome, a thermodynamic model was developed.

Thermodynamic ligand-target (L-T) complex formation-dissociation equilibrium for a given ligand (n) is described by equations:

LT ⇌ Ln + T

kD,n = [Ln]・[T]/[LnT]

[LnT] = [Ln]・[T]/kD,n

The probability of a target protein being bound (PT,b) is described by equation:

Pb,T = [LT]/([T] +[LT]) = ([L]/kD)/(1 + [L]/kD)

Where L is an average ligand, charachterised by total concentration of all ligands, and the average KD of a complex with a target protein.

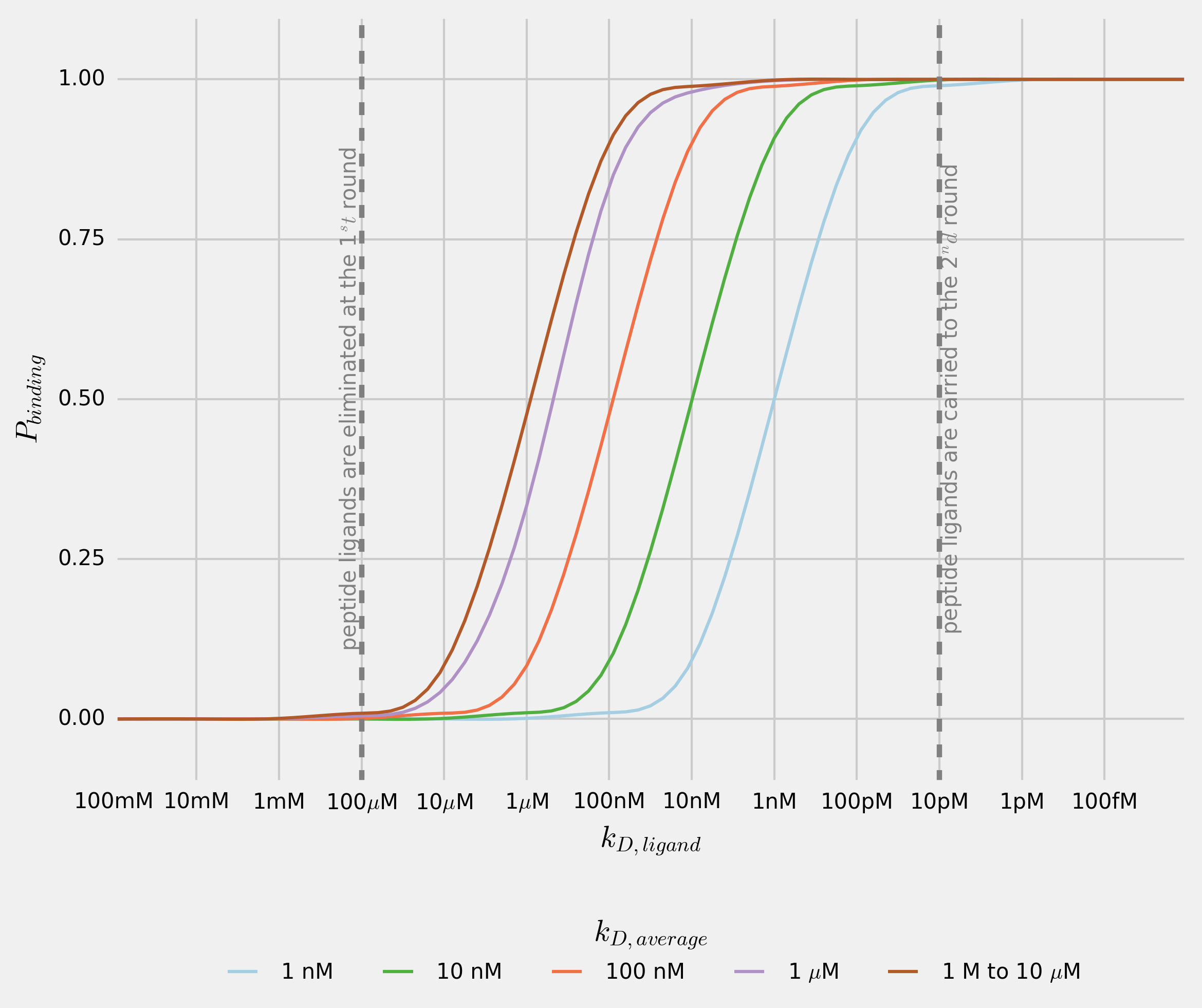
Thus, the amount of the unbound target protein ([T]) can be expressed as follows:

[T] = (1 - P(b,T))・T

The probability of a given ligand being bound (Pn,b) to a target protein is described by equation:

Pb,n = [LnT]/([Ln] +[LnT]) = ([T]/KD,n)/(1 + [T]/KD,n).

Figure\_\_ shows the probability of a ligand being bound to a target protein depending on a given ligand KD, and the average KD of total ligand population. Primary diversity of the mRNA library (as opposed to secondary diversity resulting from PCR-induced mutations) is to a great extent regulated by the first cycle of selection, at which point each ligand is represented by 1 to 100 copies and can be eliminated completely from the selection, if not bound to a ligand.



Figure\_\_\_. Probability of a ligand (n) being bound to a target protein dependence on a ligand-target complex dissociation constant (KD,n). Every line shows the probability of a ligand (n) being bound to a target protein for total-ligand-population average KD in the range 1 nM to 10 µM.

The concentration of a target protein in during incubation with cDNA-mRNA-peptide fusion libraty is approximately 10-6 M, and the total library concentration is about 10-6 M (total-ligand to target ratio 1:1). Given that peptide ligands with 1 µM KD values are found after selection, total library average KD can be safely assumed to be at 10 µM or higher. It can be seen that a ligand which complex with the target protein KD exceeds 100 µM, has close to 0 probability of being bound to the target protein. In the first cycle of selection, when only 1 to 100 copies of a given ligand are present in the mixture, such ligand is likely to be eliminated completely from the selection. Ligands with KD lower than 100 µM are likely to be recovered after the first round of selection and amplified, from which point their frequency depends on the amplification efficiency, rather than their affinity towards the target protein.

Based on the results of this study and suggested thermodynamic model of selection, several suggestions for the improvement of selection process can be made. While it is not possible to completely eliminate the effect of PCR bias, introduces at the library amplification stage, it may be possible to minimise it using high fidelity and high GC DNA polymerases such as Phusion, Q5 and AccuPrime. To reduce by-product formation during the PCR amplification, the protocol may benefit from using emulsion PCR.

Previously suggested introduction of a ‘warhead’{Morimoto:2012cp} can be defined more specifically: to ensure the recovery of the ligand after the first cycle of selection, the ‘warhead’ has to ensure KD at 1 nM or lower.

It has been previously suggested that in the first round less ‘stringent’ conditions are applied 1to the selection (*e.g.* fewer washes, and lower temperature){Pande:2010jx}. Temperature regulation of selection affects complex formation between the ligands and the target proteins in predictable manner: the higher is the temperature at which the peptide-ligands library is incubated with a target protein the lower is the average KD value, and the more likely it is that he peptides bound will have lower KD values. Selection at 37ºC (as opposed to commonly applied 4ºC) was previously used to regulate the selection pressure at later cycles of selection{Hipolito:2013gm}. Thermodynamic model proposed in this work suggests, that selection may benefit from being carried out at higher temperature from the first cycle.

To minimise the undesirable interference between cDNA-mRNA part of the ligand and target protein, short random strands of DNA can be introduced into the blocking solution.

If the total-ligand to target ratio is reduced to 1:10000 it is possible to ensure that ligands form complexes with a target protein at KD higher than 10 nM are eliminated from selection, thus ensuring the recovery of peptide ligands with KD below 10 nM.

Due to the PCR bias, peptide sequences present in the original library and amplifying at a lower rate yet having high affinity towards target protein, are often disregarded and taken out of consideration. Instead, selection results are heavily biased by the rapidly amplifying sequences and their mutants. While it may be impossible to completely eliminate the amplification bias or alter the amplification efficiencies of ligands, the alternative approach to the selection-results analyses offered in this work allows one to identify the peptide sequences present in the original library, yet amplifying at a lower rate due to the PCR bias. It is hoped that the analysis suggested in this paper may uncover such ligands, thus increasing the rate of successful ligand discoveries.