# Introduction

The emergence of genetically-encoded peptide libraries has facilitated the discovery and further development of targeted peptide ligands — such as drug-like compounds{Goldflam:2015ju}, protein inhibitors, protein activators, or co-crystallisation chaperons{Hipolito:2014bg} - against proteins of biomedical interest. Genetically encoded peptide display library is a library of compounds which combine phenotype (peptide sequence) and genotype (mRNA sequence). In the last three decades, various genetically encoded peptide display technologies have been developed, including phage display, ribosome display, and mRNA display. All of these libraries allow for selection of protein ligands from highly diverse peptide libraries, using a iterative selection algorithm: (1) incubation of encoded peptide library with a target protein, (2) recovery of protein-binding DNA-tagged ligands, and (3) regeneration of a less diverse library from the recovered DNA tags{Goldflam:2015ju}. Selection iterations are carried out until some peptide sequences enrich and the library diversity reduces sufficiently to make its analysis possible. Identified peptides sequences are synthesised and their binding properties verified against their target protein. Selection is termed ‘successful’ if amongst recovered peptides, some demonstrate high affinity towards the target protein (e.g. kD < 10-7 M).

These binding peptides are of particular interest because they are selected from highly diverse genetically-encoded peptide libraries (approximately 109 for phage display to 1012 for mRNA display) and successful hits generally display significant selectivity and binding potency. However, not all selection produce ligands for a target protein{Menendez:2005ck}{Vodnik:2011iy}. Most commonly attributed factors include (1) enrichment of off-target binding peptides (‘target unrelated peptides’, (2) interference of the target protein with the phage, cDNA/mRNA tags, or (3) to individual cDNA sequences amplification efficiencies.

A peptide-ligand recovery in display technologies is thought to be regulated by its ability to form a complex with a target protein; thus, it is often assumed that the ligand recovery is directly proportional to its affinity towards target protein (*i.e.* ligand-target complex KD value). Thus, it is often conceived that peptide sequences recovered in higher proportion (i.e. greater enrichment) have higher affinity towards target protein. However, several studies have demonstrated that the rank order of a sequence is not necessarily predictive of its binding affinity (Ref Jalali-Yazdi etal), and 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). Consensus-sequences or conserved motifs can be identified from recovered sequenced data, and such sequences can be further tested for useful properties{Takahashi:2003ik}. However, if the structurally similar peptides are genetically related (*i.e.* are mutants of one prevalent peptide), rather than present in the original library, such analysis would produce a heavily biased consensus sequence or conserved motif that may be misleading, and lack affinity towards the target protein.

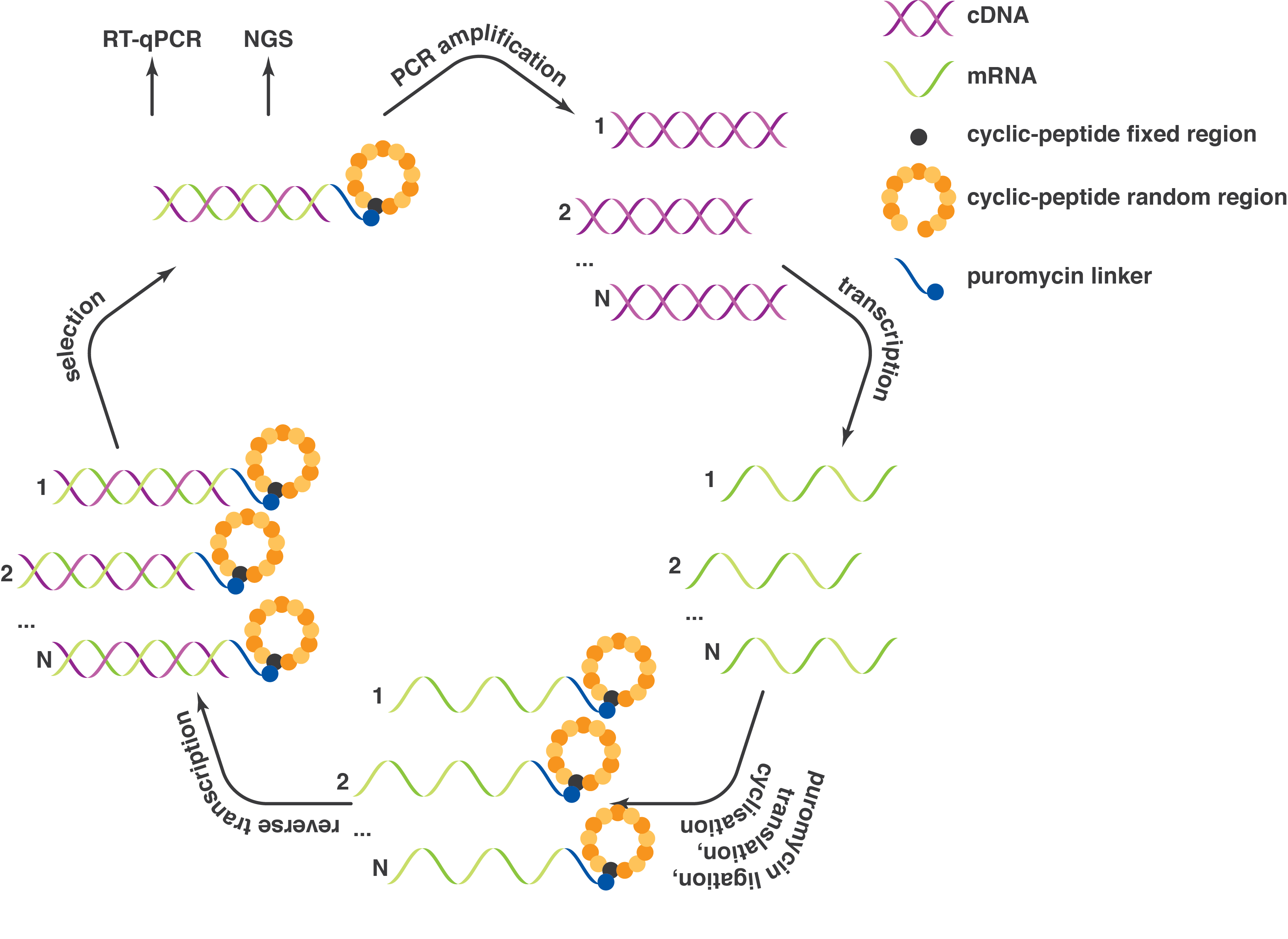
The analysis of display results is majorly based on two assumptions: (1), the recovery of a peptide ligand in the process of selection is directed by and is proportional to its affinity towards a protein target, and (2) the DNA tags do not affect the outcome of selection process. To date, these assumptions have not been validated.

This paper provides detailed analysis of a successful selection carried out against prolyl-hydroxylase enzyme isoform 2 (PHD2), an important oxygen sensing enzymes in mammals (Ref). Recombinant biotinylated human PHD2 was used as target for selection against a cyclic-peptide library using *Ra*ndom non-standard *P*eptide *I*ntegrated *D*iscovery) RaPID system. RaPID builds on mRNA display technology (mRNA/cDNA-peptide ligation via puromycin), and integrates the flexible in vitro translation (FIT) of an mRNA library and the cyclisation of translated peptides using the non-proteinogenic amino acids incorporated into the peptide chain to allow generation of a diverse (>1012) macrocyclic peptide display library. Our work reveals that the cDNA tags can affect the selection results, and we describe a robust approach to assess the selection-results to reduce this bias in analysis. We suggest a thermodynamic model that informs the extent to which the ligand-to-target affinity defines the selection outcome, and highlights potential limitations of mRNA display and potential ways of reduce these biases for improved selection outcome.

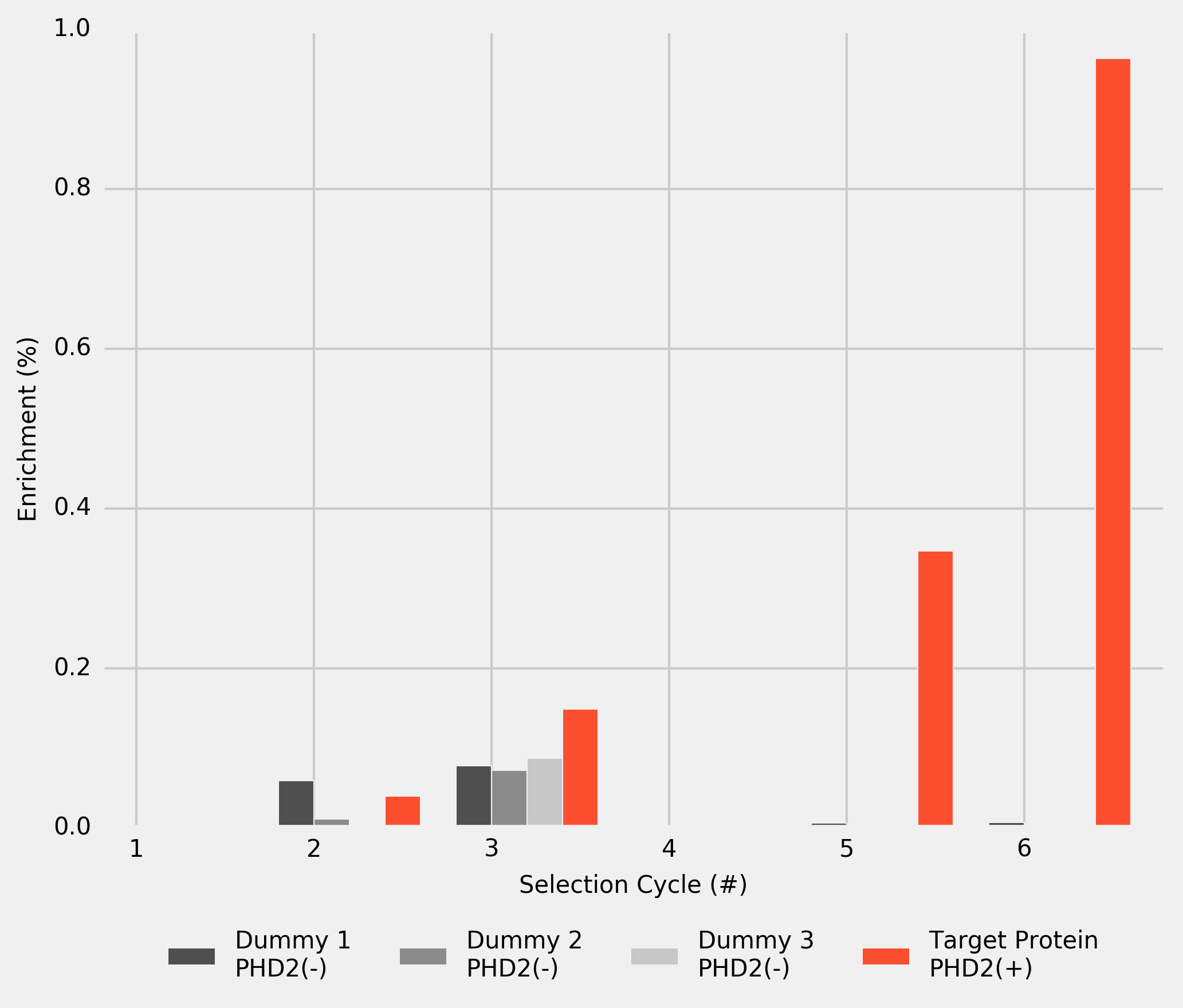
# Selection against human recombinant PHD2

Selection against the target protein — biotin-tagged human recombinant PHD2 — was performed as described previously {Hayashi:2012bb} using 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 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. The resulting cDNA-mRNA-peptide library was incubated with immobilised PHD2 at approximately 1 to 1 molar ratio, taking into account approximate translation efficiency. The fraction of the library recovered after incubation was then used to (1) generate the input cDNA library for the next cycle of selection, (2) real-time quantitative PCR (RT-qPCR) assay to calculate the recovery yield, and for (3) next generation sequencing (NGS) of the library. From second cycle forward, to eliminate the peptides interacting with magnetic beads, pre-screening was carried out three times against biotin-loaded beads (negative selection, ‘dummy’) prior to incubation with immobilised PHD2. To monitor the progression of selection, the cDNA library recovered after every cycle was subjected to RT-qPCR and fraction size of the recovered cDNA library relative to the input was estimated. Selection process was completed when the fraction of cDNA library recovered after incubation with immobilised target protein exceeded that recovered from the negative selection, and when the recovery reached approximately 1% of the input (see Figure\_\_\_). Six cycles of selection were carried out for PHD2.



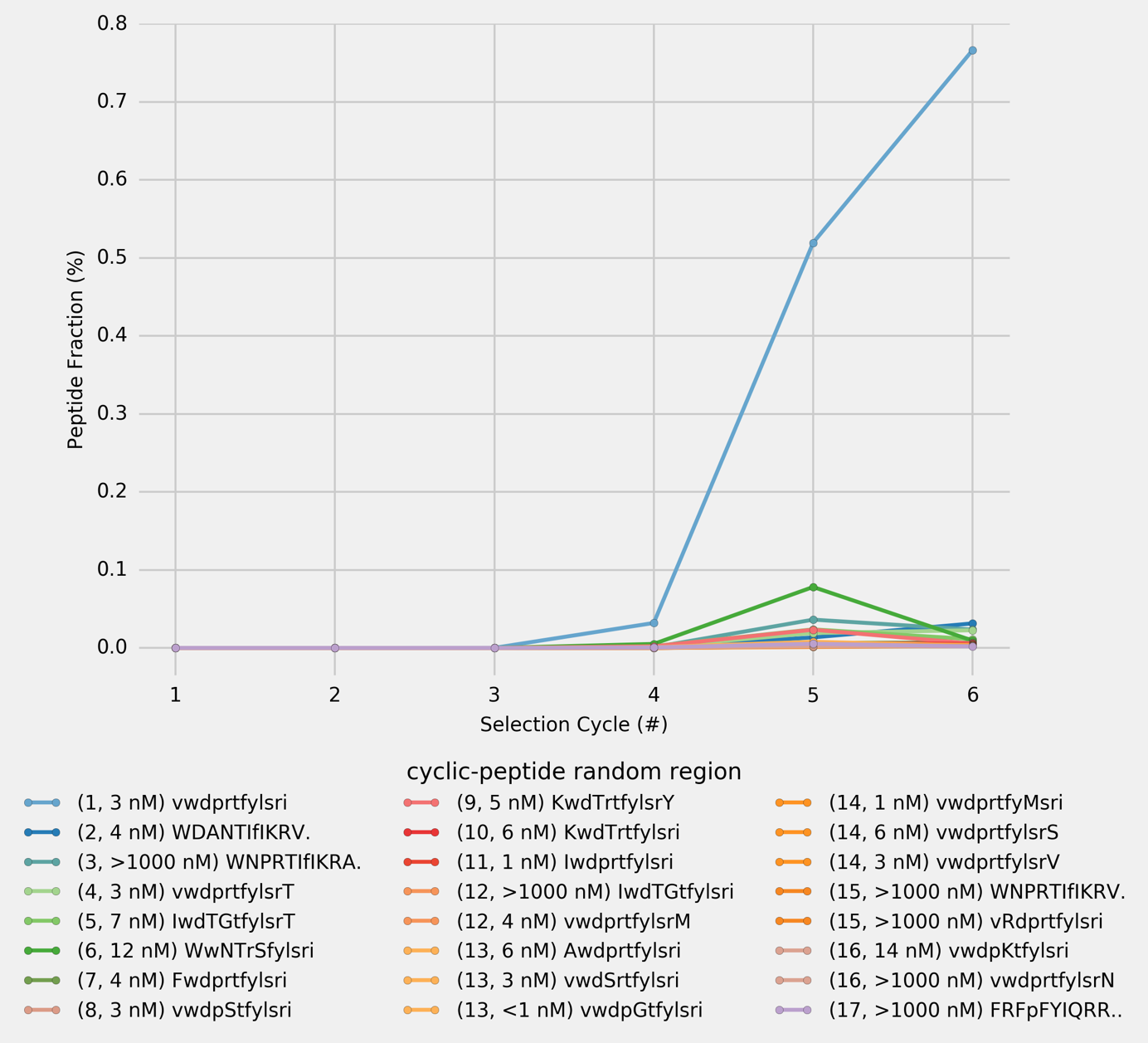
*Figure\_\_\_. Schematic representation of a selection cycle. Selection cycle begins with cDNA library, which is transcribed into mRNA library, and subsequently ligated to puromycin, translated, cyclised and reverse-transcribed in vitro, thus resulting in cDNA-mRNA-peptide library. This library is incubated 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 output cDNA library was subject to NGS. The resulting sequences were *in silico* translated into peptides and, and the instances of unique peptides were counted at each cycle to estimate the enrichment of individual peptides. 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). Twenty four top-ranked peptides were selected and their enrichment was plotted for every selection cycle (see Figure\_\_\_).

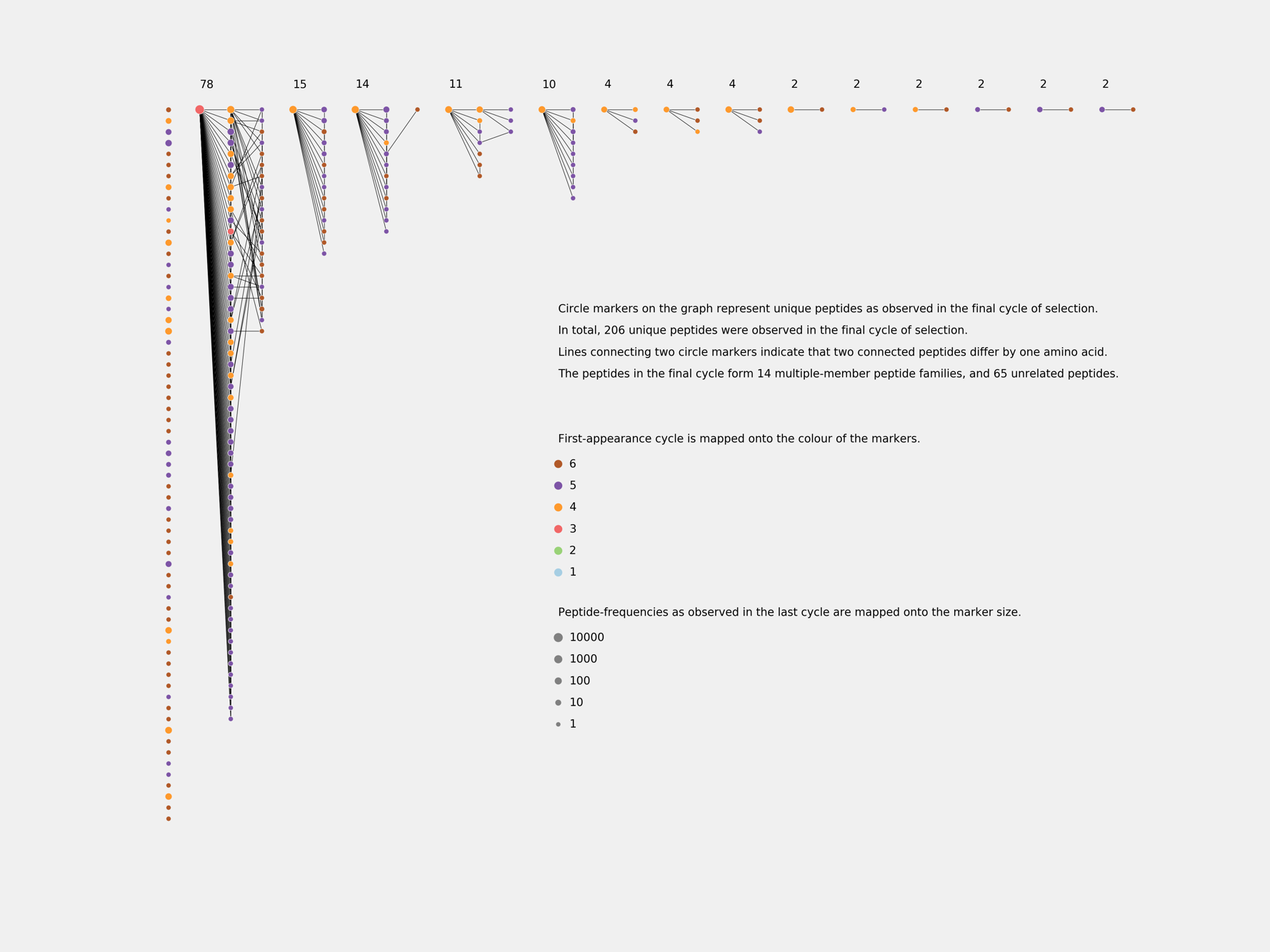


*Figure\_\_\_. Population analysis and enrichment of 24 top-ranking unique peptides (from cycle # 6) monitored over the RaPID selection cycles. The peptide sequences are formatted so that amino acids in each position, if different from the top peptide (vwdprtfylsri), are capitalised; if the peptide sequence 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, enrichment of the library by specific peptide sequences is observed. In particular, a single peptide (vwdprtfylsri, KD = 3 nM) constitutes over 75% of the library in the final cycle (Figure\_). The top ranking 24 peptides from the final cycle were chemically synthesised and KD values against PHD2 were measured using surface plasmon resonance (SPR; see Figure\_\_\_). Interestingly, the KD values of the top 24 peptides covered a wide range of binding affinities, from below 1 nM to over 1000 nM. Importantly, several of the low abundancy peptides were found to have higher affinities for PHD2 than the top ranking peptides. 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, but KD values differed by approximately 3 orders of magnitude. Intriguingly, the 3rd ranked peptide had KD > 1000 nM, whereas the 4th ranked peptide showed KD = 3nM. 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 (that satisfy the parameters), 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 they differ by one nucleotide. Thus, the unique cDNA sequences were arranged into a set of disjointed 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.*

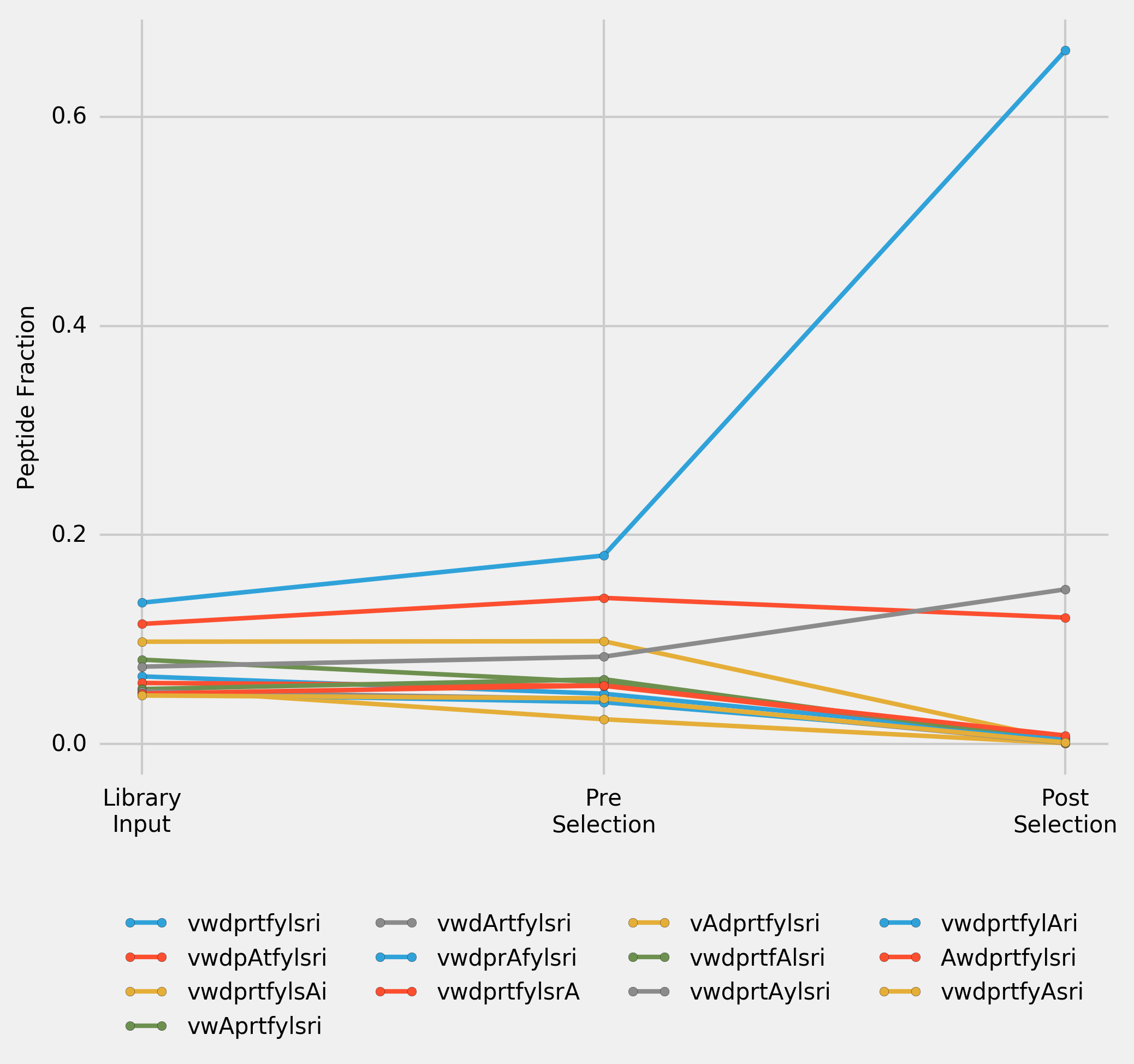
Of the 505 unique cDNA sequences found in final cycle, 361 (71%) have mutations beyond the original NNK library. Crucially, some mutant peptides persist through selection and are highly enriched despite their low affinity towards target peptide (KD >1000 nM). Thus, secondary diversity, resulting from 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

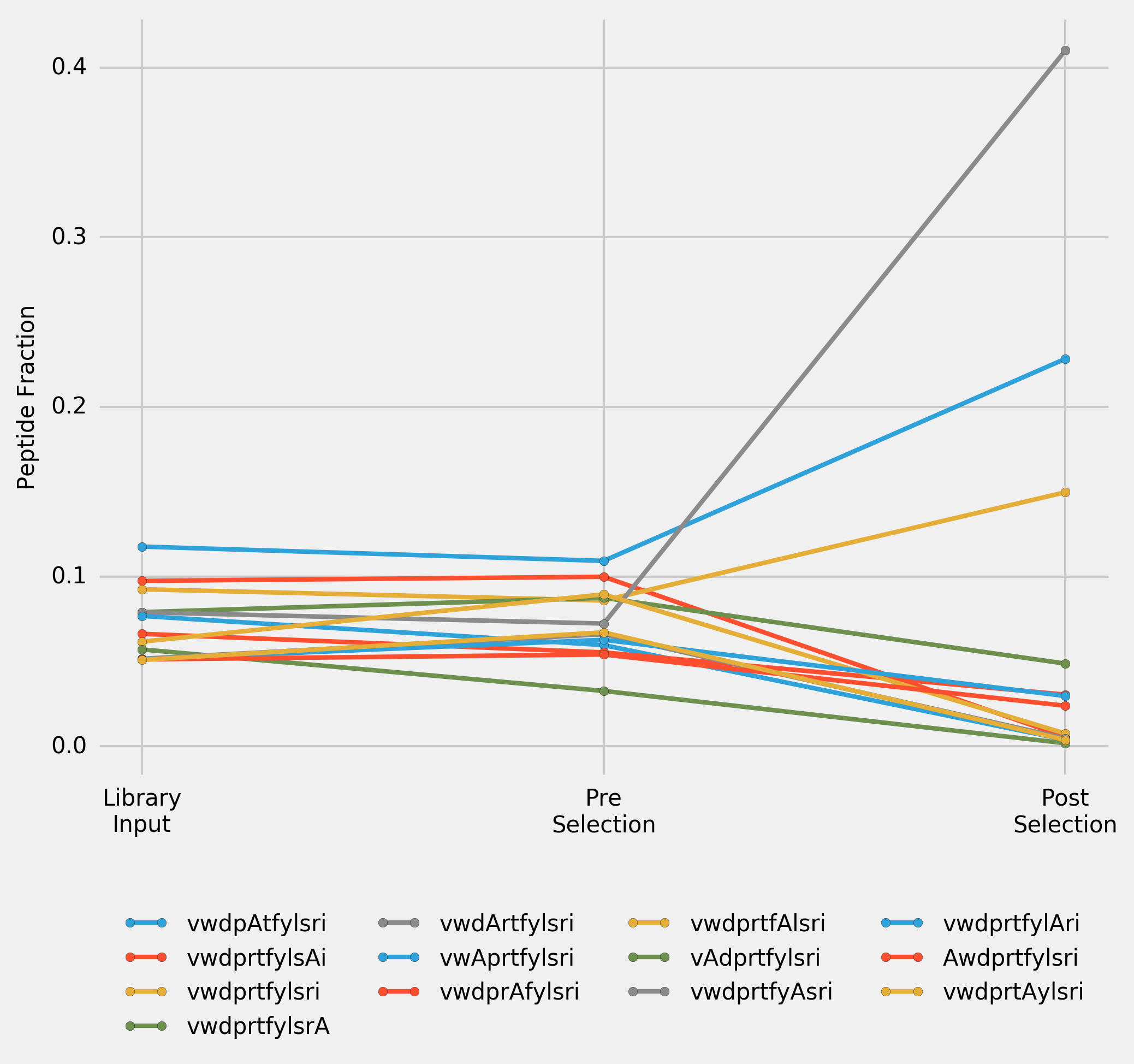
To provide further insight into the mechanism of the selection, one-cycle selections were carried out with two cDNA libraries comprised of 13 sequences. In both libraries (1) and (2) 12 cDNAs — encoding for the alanine-scan peptides of the most frequent peptide in original selection — were designed based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



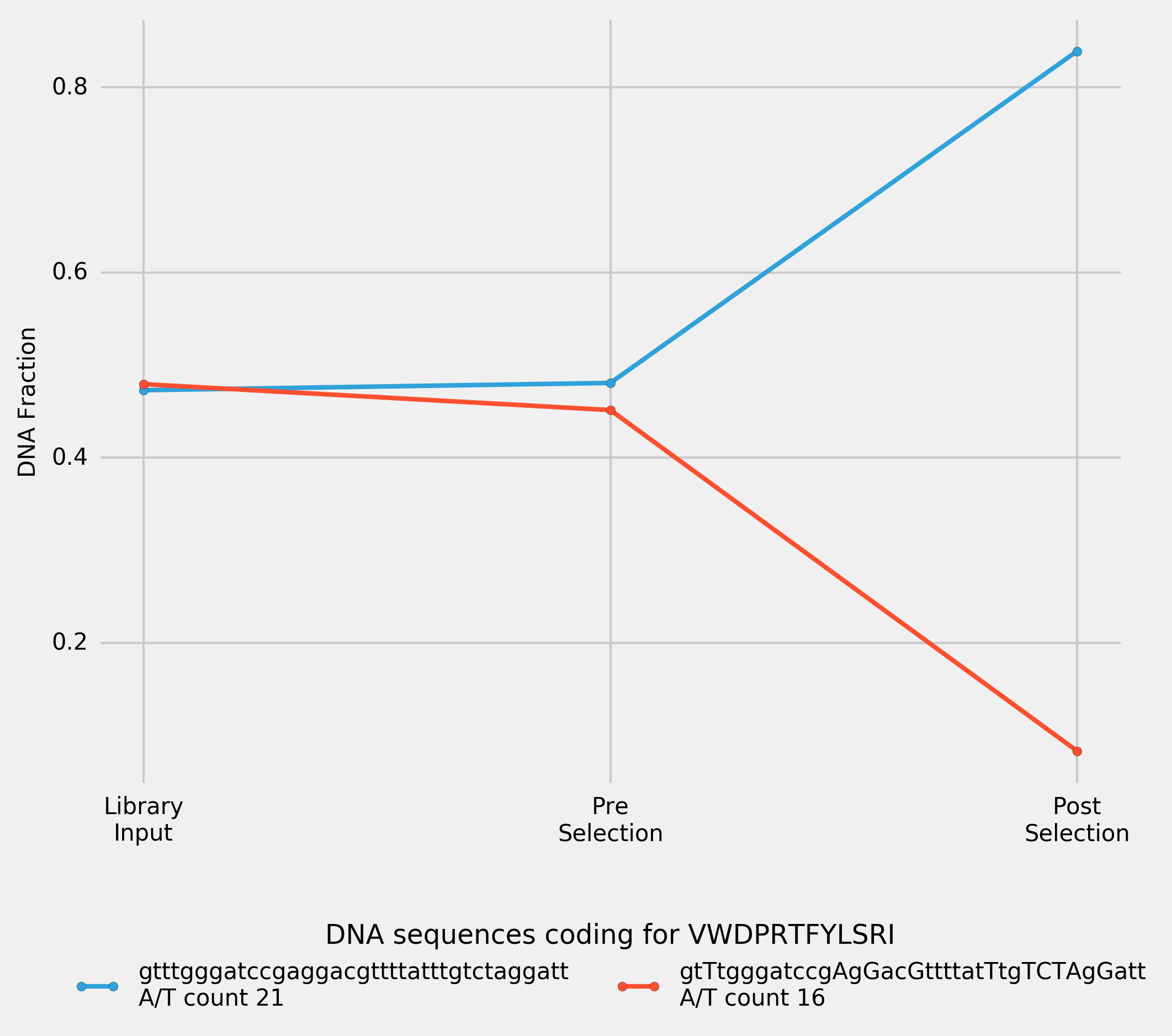
*Figure\_\_\_. Peptide enrichment over one cycle of selection for a library (1) of 13 cDNAs. 12 cDNAs — encoding for the alanine-scan peptides of top peptide in the original selection — were designed based on the most frequent E. coli codons, last cDNA — encoding for the top peptide in original selection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

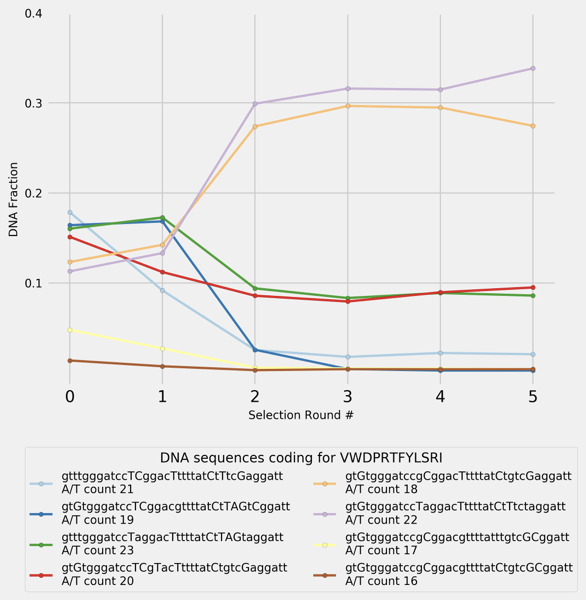
To directly compare the impact of cDNA on the selection outcome one-cycle selection was performed with a library comprised of two cDNAs encoding for the most frequent peptide in original selection: (1) the original cDNA sequence recovered from selection, and (2) one designed based on the most frequent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

# Thermodynamic model for ligand enrichment

While there have been recent efforts to study the individual affinities of libraries (REF - Jalali-Yazdi), currently existing experimental methods are limited in monitoring target binding affinities from peptide libraries with diversities >109 unique ligands. To better understand to which extent the complex formation between a ligand and a target protein affects the selection outcome in the first cycle of selection, a thermodynamic model was used.

Thermodynamic ligand-target 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, characterised by total concentration of all ligands, and the average KD of a complex with a target protein.

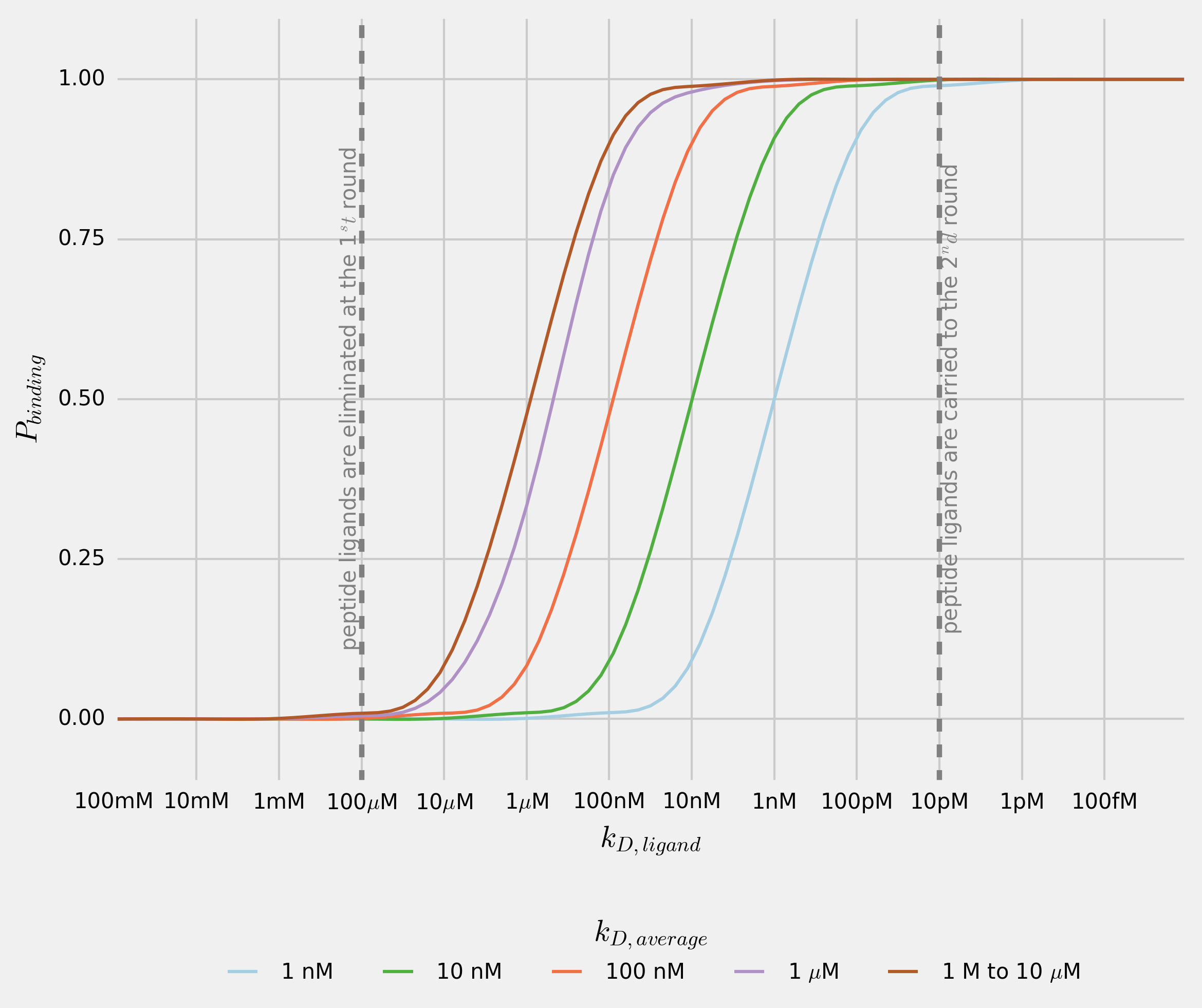
Thus, the amount of the unbound target protein ([T]) can be expressed as follows:

[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. Each ligand is represented by 1 to 100 copies in the first cycle, and can be eliminated completely from the selection if it does not bind to a target.



*Figure\_\_\_. Probability of a ligand (n) being bound to a target protein is dependent on the dissociation constant (KD,n) of the ligand-target complex. 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 during incubation with cDNA-mRNA-peptide fusion library 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 with KD exceeding 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 enrichment depends on the amplification efficiency, rather than their affinity towards the target protein.

# Discussion

In the process of selection, the enrichment of cDNA-mRNA-peptide ligands in the library can be analysed – in the simplest terms - by two distinct properties of each ligand: Firstly, the affinity of the peptide ligand-target protein complex (characterised by KD of ligand-target complex - phenotype), and secondly the PCR amplification of the cDNA tag (characterised by cDNA amplification efficiency – ‘genotype’). 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. In this study based on PHD2 selection as an example, we have demonstrated that, while the top ranking sequences that dominate the final cycle of selection can be efficient binders (KD in nM region), enrichment in certain ligands (as exemplified by peptide rank 4, KD > 1000 nM, or the top ranked peptide sequence with different DNA sequences) can also be driven by PCR amplification rather than ligand-target affinity, *i.e.* the process is overall driven by the genotype rather than phenotype of the cDNA-mRNA-peptide ligands.

Based on the results of this study, several suggestions can be made to improvement the design of selection process. While it is not possible to completely eliminate the effect of PCR bias introduced at the library amplification stage, high fidelity and high GC DNA polymerases such as Phusion, Q5 and AccuPrime can be used to minimise PCR bias. 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} targeted motif can be defined more specifically; to ensure the recovery of the ligand after the first cycle of selection, the ‘warhead’ of KD at 1 nM or lower will ensure advantaged selection, as demonstrated in the binding probably model (Figure \_).

It has been previously suggested that in the first round less ‘stringent’ conditions are applied to 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. The higher the temperature at which the peptide-ligands library is incubated with a target protein, the lower the average KD value of the recovered library, and the more likely the 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 selections may benefit from being carried out at higher temperature from the first cycle.

To minimise the undesirable interference between cDNA-mRNA tag and the target protein, short random strands of DNA can be introduced into the blocking solution.[Ref on papers with salmon sperm DNA used in the selection]

If the total-ligand to target ratio is reduced to 1:10000, ligands with KD > 10 nM can most likely eliminated from selection, thus ensuring the recovery of peptide ligands with KD below 10 nM.

Selection results can be heavily biased by the rapidly amplifying sequences and their mutants. Peptide sequences present in the original library that amplify at a lower rate yet having high affinity towards target protein, are thus often disregarded. An alternative phylogenetic approach to analysing the selection-results presented in this study may provide new opportunities to identify such high affinity / low PCR amplification efficiency ligands present in the original library. Uncovering such ligands will increase the rate of successful ligand discoveries using display technologies.