**Variant library – binding affinity report**

**(Target: Homo sapiens clone HE1-34 circulating B cell antibody heavy chain variable region mRNA)**

1. **Distribution of reads**

Analysis was performed on the numbers of read counts of the enriched sequences. Distributions of input and output sequence reads are shown below in the respective bar charts.

Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated

After removing those sequences with 0 reads, we observe the following distributions.

A normal distribution, with median of 13 reads is observed in the input read count data. A left-skewed distribution with a median of 30 is observed in the output read count data.

Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated

Sequences with 0 reads in their output (that have a positive number of reads in their input) would suggest that binding did not occur, this is to be expected. It is surprising to see the considerable number of sequences with an input read count of 0. It would be worth investigating at which stage in the pipeline this issue occurred (design, PCR, NGS, filtering).

We also see a number of sequences with no reads in the input but a positive number of reads in the output. It would also be worth investigating this further.

1. **DNA sequence design**

Total enriched sequences 13046

Sequences that did not match design 1484 (11.4%)

Sequences that were longer than design 206 (1.58%)

A picture containing chart

Description automatically generated

1484 (11.4%) of sequences did not match the design. Of these, 206 were longer than the original design. 112 were 1 nucleotide longer, 63 were 2 nucleotides longer and 31 were 3 nucleotides longer.

1. **Protein binding affinity**

To assess protein binding affinity, the following metric was used:

**Affinity = Read count in output library / Read count in input library**

This ratio gives the fold change after the test for binding affinity. After removing sequences with 0 reads from the data, affinities range between 0.056 and 394. 99% of the data have an affinity score of less than 16.5. The 50 sequences with the highest affinity scores range between 17.75 and 394. Interestingly, the highest scoring sequence deviated from our design. In order to verify this, it would be good to repeat the experiment for this sequence, to eliminate the possibility of error e.g. sample mix-up. It is possible that this is a spurious result as it has an affinity five time higher than the second highest sequence.

The three proteins with the highest affinity are:

Sequence 42, affinity 394.0

CAGGTGCAGCTGGTGGAGTCTGGGGGAGGCTTGGTCAAGCCTGGAGGGTCCCTGAGACTCTCCTGTGCAGCCTCTGGATTCTCCCCCGCCTGCTACTACATGAGCTGGATCCGCCAGGCTCCAGGGAAGGGGCTGGAGTGGGTTTCATACATTAGTGTCTTGGTCTCCACCATATACTACGCAGACTCTGTGAAGGGCCGATTCACCATCTCCAGGGACAACGCCAAGAACTCACTGTATCTGCAAATGAACAGCCTGAGAGCCGAGGACACGGCCGTGTATTACTGTGCGAGAGA

Sequence 1803, affinity 78.2

CAGGTGCAGCTGGTGGAGTCTGGGGGAGGCTTGGTCAAGCCTGGAGGGTCCCTGAGACTCTCCTGTGCAGCCTCTGGATTCCGCGCGGCCTACTACTACATGAGCTGGATCCGCCAGGCTCCAGGGAAGGGGCTGGAGTGGGTTTCATACATTAGTGGCTAGGAGTTCACCATATACTACGCAGACTCTGTGAAGGGCCGATTCACCATCTCCAGGGACAACGCCAAGAACTCACTGTATCTGCAAATGAACAGCCTGAGAGCCGAGGACACGGCCGTGTATTACTGTGCGAGAGA

Sequence 6713, affinity 52.0

CAGGTGCAGCTGGTGGAGTCTGGGGGAGGCTTGGTCAAGCCTGGAGGGTCCCTGAGACTCTCCTGTGCAGCCTCTGGATTCGTCGACTGCTTCTACTACATGAGCTGGATCCGCCAGGCTCCAGGGAAGGGGCTGGAGTGGGTTTCATACATTAGTGTCGGCTCCTGCACCATATACTACGCAGACTCTGTGAAGGGCCGATTCACCATCTCCAGGGACAACGCCAAGAACTCACTGTATCTGCAAATGAACAGCCTGAGAGCCGAGGACACGGCCGTGTATTACTGTGCGAGAGA

1. **Next steps**

Going forward, it would be interesting to investigate the subset of sequences that do not match the design, to identify whether particular positions or regions have a higher prevalence of change and whether these errors are produced in a random or more systematic way. It would also be worth investigating this in the respective amino acid sequence, to investigate whether secondary or tertiary structures could be affecting the output at the PCR or NGS stages.

I aim to investigate further into the sequences with the highest affinity scores, these representing our best candidate proteins. Correlations between nucleotide position and affinity would be of interest.