Epitope Fingerprinting Report

Date: February 8th, 2024

Work Package: Epitope Fingerprinting for one monoclonal antibody

Task: Epitope identification for anti-human c-myc

General comments

The c-myc antibody 9E10 was subjected to epitope fingerprinting. All available datasets fulfilled the quality criteria of sufficient sequence counts; thus, all data were perfectly suitable for statistical epitope analysis.

This antibody is a special challenge because a second sequence motif in the c-myc protein is similar to the peptide used to generate the antibody. The challenge is to prove the correct position of the epitope and the relevant neighbouring amino acids using statistical analysis. In addition, we validated our findings using known data.

This report uses and compares data from different selection procedures:

- Protein A magnetic beads (including state of the art amplification and alternatively PCR amplification to generate DNA-pools for NGS)
- Carboxy magnetic beads (EDC/NHS coupling)
- Immuno tubes

Although different sequences are selected in each experiment, all result in essentially the same results on the level of the motifs, if you look up the alignments at the end of this document.

This report summarizes data from several selection runs on the antibody as a control and earlier validation work by Nicolas Delaroque (Fraunhofer IZI, Leipzig, Germany)

Panning and NGS library statistics of previously collected datasets:

Selection	Dataset	Phagemid output (cfu min)	Sequence Count	Valid Sequence Count	Motif Count
Carboxy-Beads	ab anti-Cmyc-1pr	2.4×10^{06}	478,940	319,031	138,995
Carboxy-Beads	ab anti-Cmyc-2pr	1.3×10^{05}	599,990	405,773	132,947
proteinA Beads	c-myc-1pr	2.8×10^{05}	545,411	390,377	138,766
proteinA Beads	c-myc-2pr	2.8×10^{05}	468,843	312,593	134,622
proteinA Beads	c-myc-PCR	n.a.	572,536	414,215	138,597
Immuno tubes	it_c-myc_1pr	1.0×10^{06}	715,525	520,549	139,422
Immuno tubes	it_c-myc_2pr	1.2×10^{05}	1,061,440	750,183	125,124

Legend:

Selection: The antibody protein was chemically coupled (carboxy-beads) or directly bound (Protein A) to magnetic beads or immobilized in immuno tubes.

Dataset: Dataset identifier; '1pr' for 1st panning/selection round, '2pr' for 2nd panning/selection round, PCR: direct amplification of bound phage.

Output (cfu min): Phage rescued and titered after the selection round (often lower than real value; Typical input $4*10^{11}$ naïve resp. approximately $3*10^{09}$ preselected phagemid particles)

Seq Count: Number of sequences in illumina MiSeq™ output

Valid Count: Number of sequences matching the library's sequence design. Expected error rate is >30%.

Motif Count: Number of different 3-mer and 4-mer motifs, in theory 168,000, our threshold minimum (1st pr) 130,000. (due to library constrains (Met, Trp), this value can't go above 140,000)

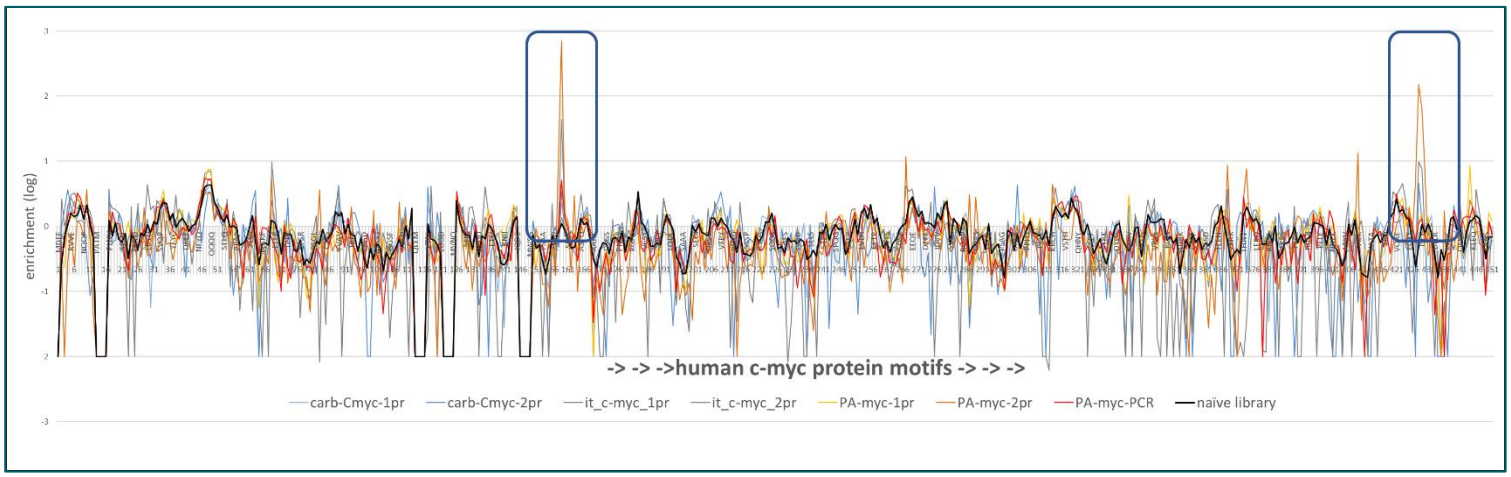
Comment/Explanations:

Experiments have been run with 10 µg mAb bound to carboxy Dynabeads with EDC/NHS coupling, directly to protein A Dynabeads or immobilized in 10 ml immuno tubes® (FisherScientific™). Phages selected from a library pool exceeding the complexity of the naïve library with the antibody on magnetic beads were used directly to infect a specially designed *E.coli* strain or eluted with a special PCR buffer. DNA pools from amplified clones resp. PCR amplification were sequenced on an Illumina MiSeq™. The number of sequences from the first round of selection was sufficient to cover the complexity of the binding peptides. Due to enrichment of single sequences in the second selection round, the number of observed motifs can be lower, and should not be less than 120,000. A lack of data for too many motifs can make statistical analysis almost impossible.

Statistical Analyses

Standard analysis based on the frequency of 4-mer motifs of the antigen sequence resulted in an overview of potential epitopes. In total, seven datasets were obtained, containing approximately 3.1 Mio sequences. One dataset had a slightly higher enrichment of a few sequences than the other (it_c-myc-2pr).

The graphic below shows the statistical enrichment of all the 4-mer peptide motifs of the c-myc protein in the datasets. The Y-axis shows the enrichment of 4-mers over the theoretically expected value defined by the library design. The scale is log10, so several motifs occur hundred to thousandfold more frequently in the NGS datasets than expected. The black line shows the measured statistical data of the naïve library's dataset from >2 Mio. sequences. The peaks in the boxes indicate motifs within the protein enriched by more than hundredfold. Such peaks could belong to positions in the protein bound by the antibody, and the sequences containing them were further analysed.



This "enrichment" curve shows that there are mainly 2 peaks for motifs:
158-KLVSE-162
and
427-KLISEEDLL-435
Both share a similar core motif: KLVSE vs KLISE. The enriched sequences will show, that the 427-KLISEEDLL-435 is likely to be the epitope of c-myc. 158-KLVSE-162 is an enriched variation caused by some similar structural features of Leu and Val. KLISE is also part of the C-terminal peptide used for generation of the antibody 9E10:
423-AEEQ**KLISEEDLL**RKRREQLKHKLEQLRNSCA-454

For each candidate motif related sequences have been retrieved from the database and aligned. The alignments were cured for sequences that were observed only once or twice. If a sequence is found at least three–five times, it is less likely to contain typical NGS sequencing errors. If multiple sequences share motifs and additional amino acids identical, or similar, to the antigen or flanking sequences (often Cys), the motif is very likely part of an epitope.

c-myc enriched motifs

This is an overview of the full-length c-myc sequence where those regions are marked in yellow, that show enrichment in the statistical analysis of 4-mer motifs in the NGS datasets.

```
>human_c-myc

1     MDFFRVVENQ QPPATMPLNV SFTNRNYDLD YDSVQPYFYC DEEENFYQQQ QQSELQPPAP
61    SEDIWKKFEL LPTPPLSPSR RSGLCSPSYV AVTPFSLRGD NDGGGGSFST ADQLEMVTEL
121   LGGDMVNQSF ICDPDDETFI KNIIIQDCMW SGFSAAAKLV SEKLASYQAA RKDSGSPNPA
181   RGHSVCSTSS LYLQDLSAAA SECIDPSVVF PYPLNDSSSP KSCASQDSSA FSPSSDSLLS
241   STESSPQGSP EPLVLHEETP PTTSSDSEEE QEDEEEIDVV SVEKRQAPGK RSESGSPSAG
301   GHSKPPHSPL VLKRCHVSTH QHNYAAPPST RKDYPAAKRV KLDSVRVLRQ ISNNRKCTSP
361   RSSDTEENVK RRTHNVLERQ RRNELKRSFF ALRDQIPELE NNEKAPKVVI LKKATAYILS
421   VQAEEQKLIS EEDLLRKRRE QLKHKLEQLR NSCA
```

EPITOPIC

Epitope in detail

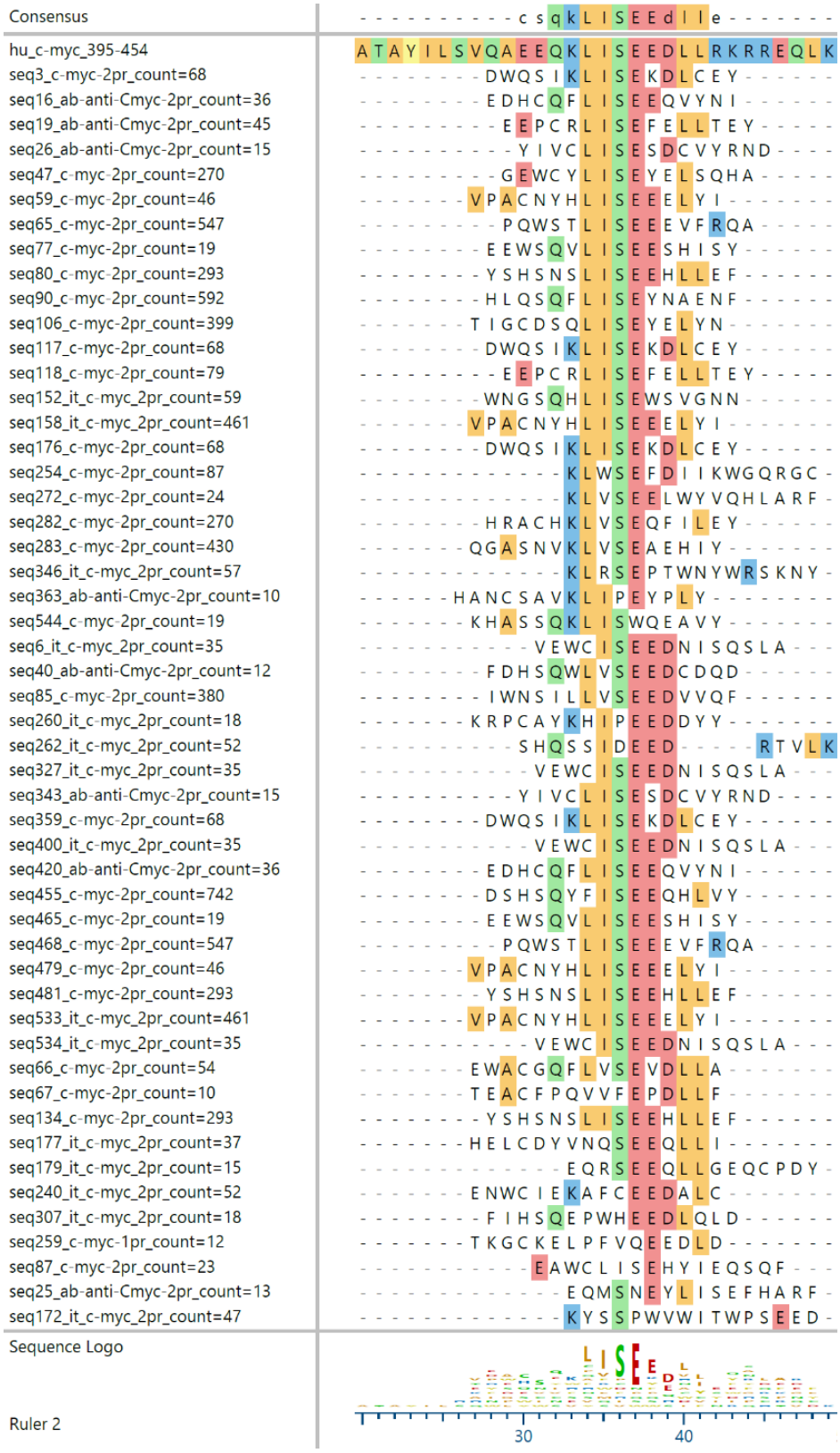
Alignment of all sequences that contain KLISE, ISEED, or EEDLL (out of 427-KLISEEDLL-435) and 1 amino acid variations of these (a*aaa, aa*aa and aaa*a).

Thousands of similar sequences share more than the initial motif used to identify the sequences in the dataset. The alignment on the right shows only those sequences containing these motifs with at least 5 amino acid identities to the antigen, and occuring at least 20 times in the individual dataset. In total, there were appr. 1,500 unique sequences out of 12,195 sequences containing one motif.

In this case, besides the enriched and conserved core motif KLISE, more amino acids further to the C-terminus can fit identically to the antigen while still showing a high frequency. Many sequences feature six or even 7 amino acid identities with the antigen near the core motif.

Based on the alignment and considering more or less conserved residues the epitope would be best described as

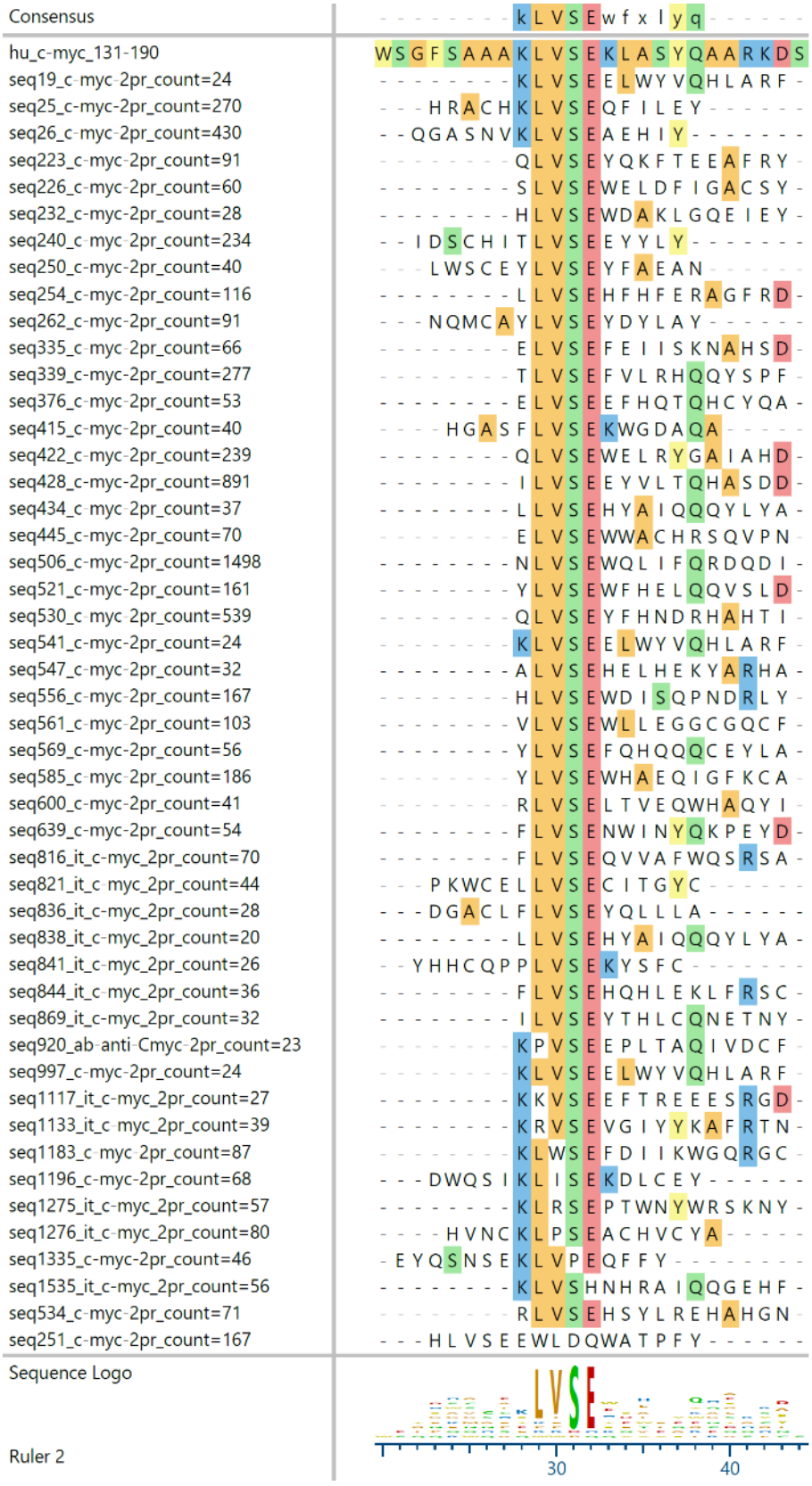
KLISEedLl



For this report we also aligned sequences containing the enriched motif 158-KLVSE-162 allowing one position being variable as above. Displayed are the 50 most frequent sequences of all data sets, which are sharing at least five amino acid identities with the antigen and occurring with a frequency of at least 30.

Only the core motif KLVSE was enriched, whereas the other one to two additional C-terminal amino acids partially fit the antigen sequence. However, the frequency of identities is hardly above the statistical variations and is likely to be only by chance.

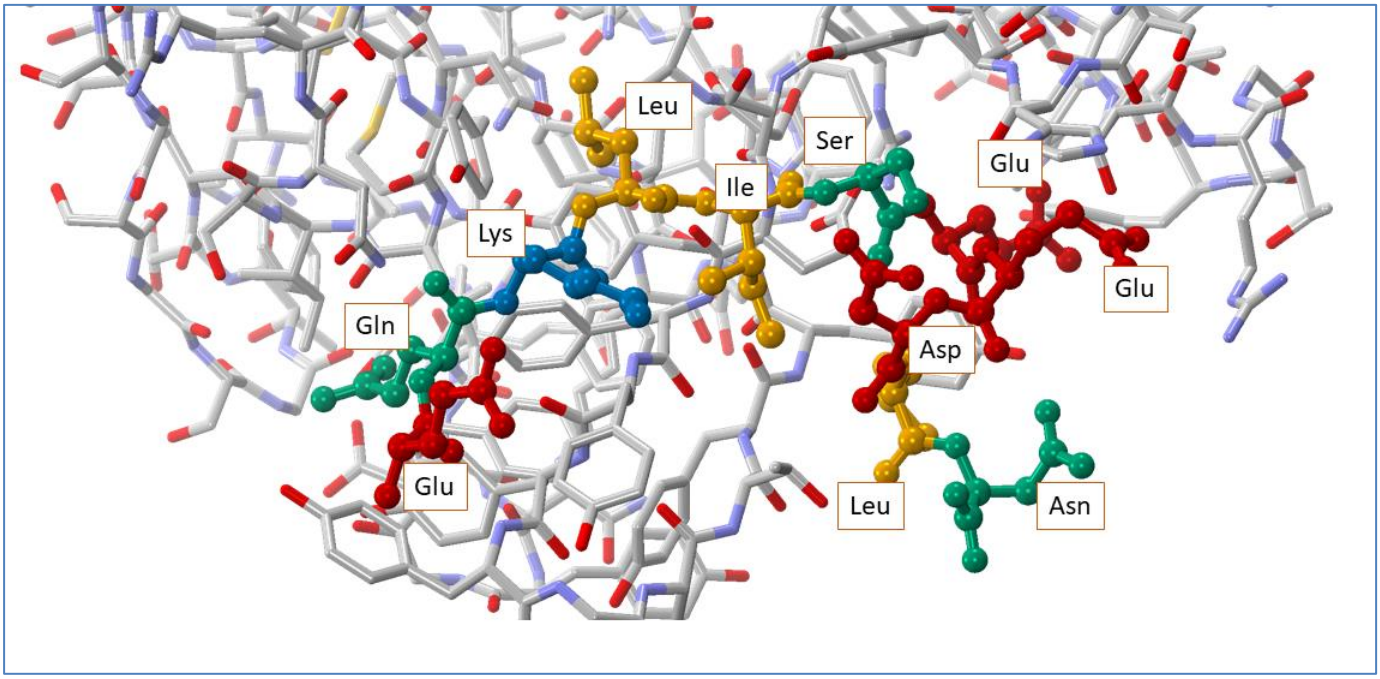
However, 1,538 different sequences were found 48,267 times, and the KLVSE motif is more likely to bind the antibody in the same way as the KLISE motif-containing sequences. Several sequences from this alignment are also found in the previous alignment.



Structure

The picture below shows the core peptide of the epitope co-crystallized with the Fab of 9E10 accessible as PDB structure 2or9.

It can be seen that the central Ile is pointing away from the antibody. Therefore, in this position, it does not matter whether Val or Ile is present. That explains the enormous enrichment of the KLVSE motif.



The motif **K**L**ISE**ed**L**l reads from the left to bottom right (though the peptide binding the protein used for the structure is actually EQKLISEEDLN).

Separately available for download

In this section epitopic usually provides information about additional supporting graphs and sequence data.

Any additional data available on request!