Deep mutational scanning data analysis to reveal sequence function relationships of BRCA1

17.07.2023

Group 4: Felix Faulstich, Sophie Bosseva, Mara Rödele, Enno Müller

Table of Contents

Introduction

The paper (1) from which we got our data is about the homolgy-directed DNA repair gene BRCA1 which plays a crucial role in breast and ovarian cancer.
The researchers goal was to be able to predict the risk for developing such cancer or evaluating the risk of an existent cancer based on the DNA sequence of the BRCA1 gene.

To get the data for this the researchers created variations of the gene, which had single nucleotide variations (SNVs) in the most important domains (RING and BRCT). It's important to mention, that it was not possible for the researchers to create every possible SNV at every position. After that they screened the proteins that were created based on these SNVs for functionality and gave them function scores. From these function scores the dms_scores in our dataset were derived.

Dms values smaller than 0 mean that the protein works worse than the wildtype (>0 = better). The SNVs that lay under a dms_score of -1 get classified as bin_0. The bin (=binary) column categorises the SNVs as 0 = non functional and 1 = functional.

We chose to focus on these proteins and not on the ones that worked (some of which better than the wildtype), because the goal of this research is to evaluate cancer and cancer risk, not to figure out how humans could be genetically engineered in order to become more cancer resistant.

We were interested in the correlation of the dms_score with:

- secondary structures
- conserved regions
- and post translational modification sites.

Material and Methods

Secondary Structures

PSIPRED: Using position specific iterative (PSI-) BLAST and MSA to get the needed extra data, PSIPRED predicts secondary structures using a neural network based on various features of the amino acid sequence like the general amino acid composition, evolutionary conservation and position specific properties of the proteins sequence. (2)

Chi-Square Test: We used this statistical test to figure out, wether there is a correlation of the type of secondary structure and how many of its SNVs are classified as bin_0.

Chi-Square Formula: $\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$

O_i = Observed counts
E_i = Expected counts

Conserved Regions

BLAST: Standing for "Basic Local Alignment Search Tool", BLAST is an algorithm that uses an input sequence, also known as "query", to find matching sequences in other proteins. To do so it can use a variety of protein databases, for our analysis we used the "UniProtKB reference proteomes" database.(3) In order to find similar genes to our human BRCA1 gene, we performed a BLAST protein comparison. The results were mostly BRCA genes from other species.

Multiple Sequence Alignment: A multiple sequence alignment (MSA) takes multiple genes as input standardizes their length by aligning them based on their domains and using placeholders for missing parts. The goal is to align only homologous sequences, so parts of the protein which are derived from a common protein.(4) For our analysis, we took the first 100 sequences from the BLAST to perform a MSA, while only using one gene per species. The result was exported and can be found in our code as "fasta_file".

Results

Secondary Structures

First of all we split our data frame into its two domains RING (amino acid position (aap): 1-101) and BRCT (aap.:1631-1855). Based on the PSIPRED data for the domains, we then created the following plots to visualise which aap belonged to which secondary structure element. For this we used a scatter plot.

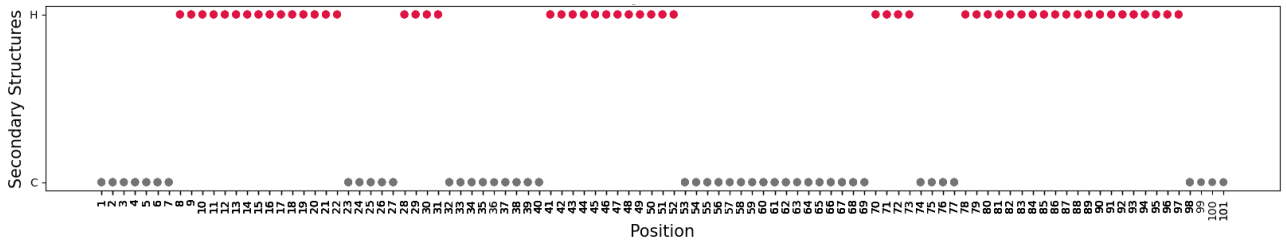

Fig. 1: Position vs Secondary Structure Element - RING (C = Coils, H = α-Helices)

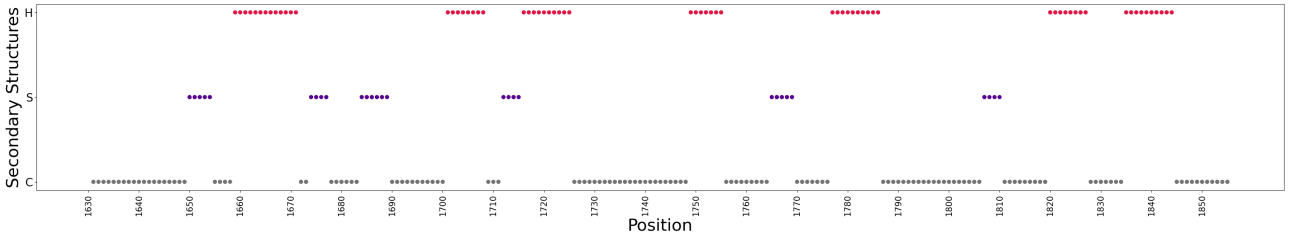

Fig. 2: Position vs. Secondary Structure Element - BRCT (C = Coils, S = β-Strands, H = α-Helices)

We also visualised the proportions of how many SNVs within a certain secondary structure cause a bin_0 dms_score with these bar plots:

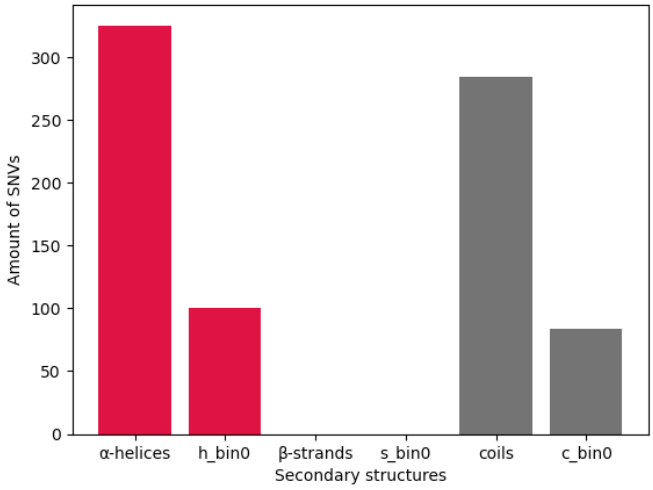

Fig. 3: SNVs per secondary structure element - RING

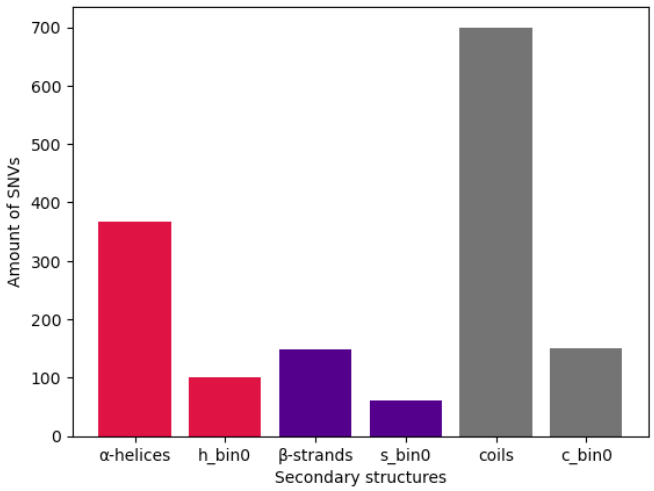

Fig. 4: SNVs per secondary structure element - BRCT

The exact ratios of a SNV that caused a bin_0 rating vs. the generell amount of SNVs in a certain secondary structure element are:

	RING	BRCT
α-Helices	30.77%	27.45%
β-Strands	–	40.54%
Coils	29.58%	21.57%

Table 1: Ratios of the amount of SNVs with a bin_0 rating for each secondary structure element

To evaluate wether the ratios we got could be used as meaningful predictions for the chances of a SNV in a certain secondary structure to cause a bin_0 score, we used the statistical test “ChiSquare” with the H0-Hypothesis being: “That there is no correlation between the type of secondary structure and how many of its SNVs have a bin_0 score”
For the RING domain this got us a P-Value of ~0.88. Since the P-value is >0.05 there is no strong indication to reject the H0 Hypothesis.
For the BRCT domain we got a P-Value of ~0.0005, which ist <0.05 so we should reject the H0-Hypothesis.

The last plots we made are for visualising the results we got from, they show every single SNV of one domain each categorised into bin_0 (red) or not bin_0 (blue) and its secondary structure type.

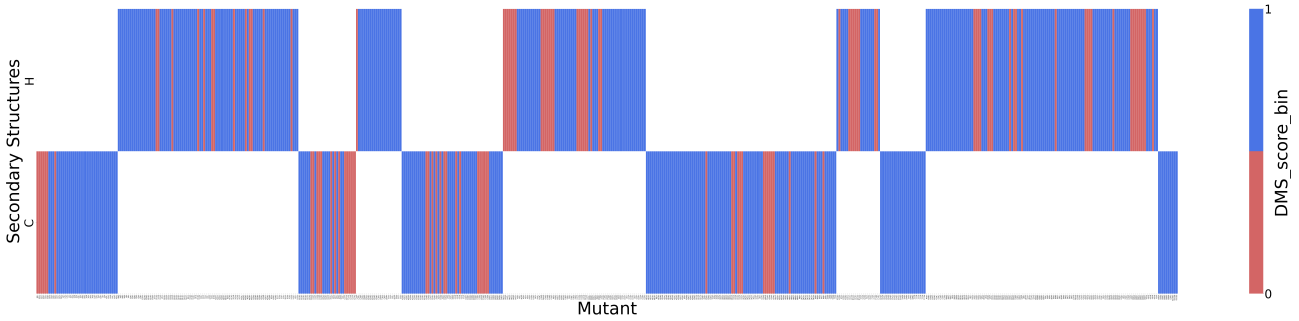

Fig. 5: Visualisation for RING

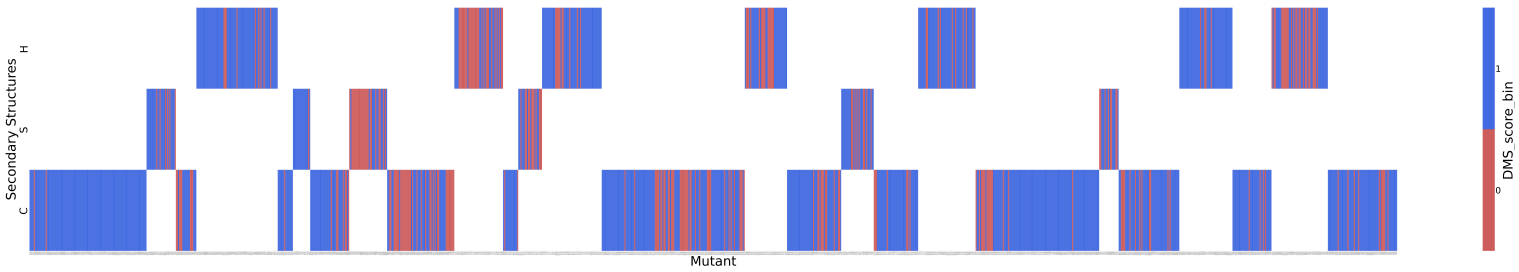

Fig. 6: Visualisation for BRCT

Conserved Regions

To determine the level of conservation present in our protein, a BLAST analysis was performed in order to find proteins with a similar structure. The first 100 genes from this BLAST analysis were further used for a multiple sequence alignment. The data from the MSA was the input for the creation of a sequence logo in order to see which amino acids are prevalent at a given position.

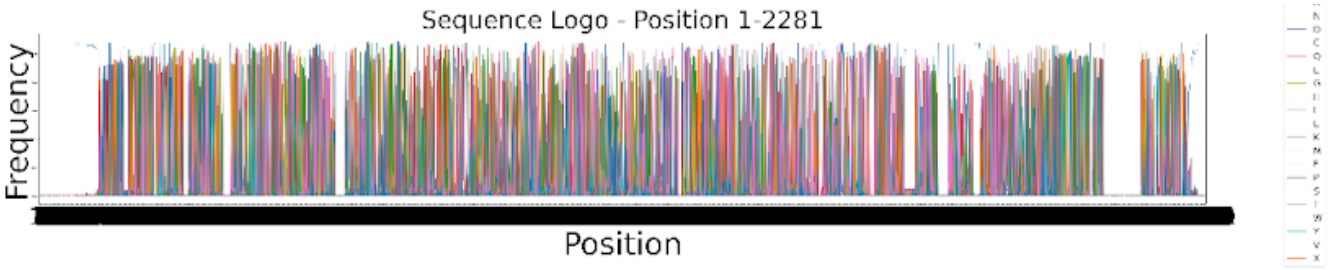

Fig. 7: Sequence logo of the whole MSA

In this form, which shows the whole sequence of the MSA, we can see that the peaks of the graphs seem to be really high, which indicates a high level of conservation at those spots. However, we can not see if the high levels of conservation are consistent throughout the whole gene. In order to determine whether this is the case, we split the sequence into smaller blocks, more specifically into the aforementioned RING and BRCT domains, which allows for a clearer representation of amino acid variance at a given spot.

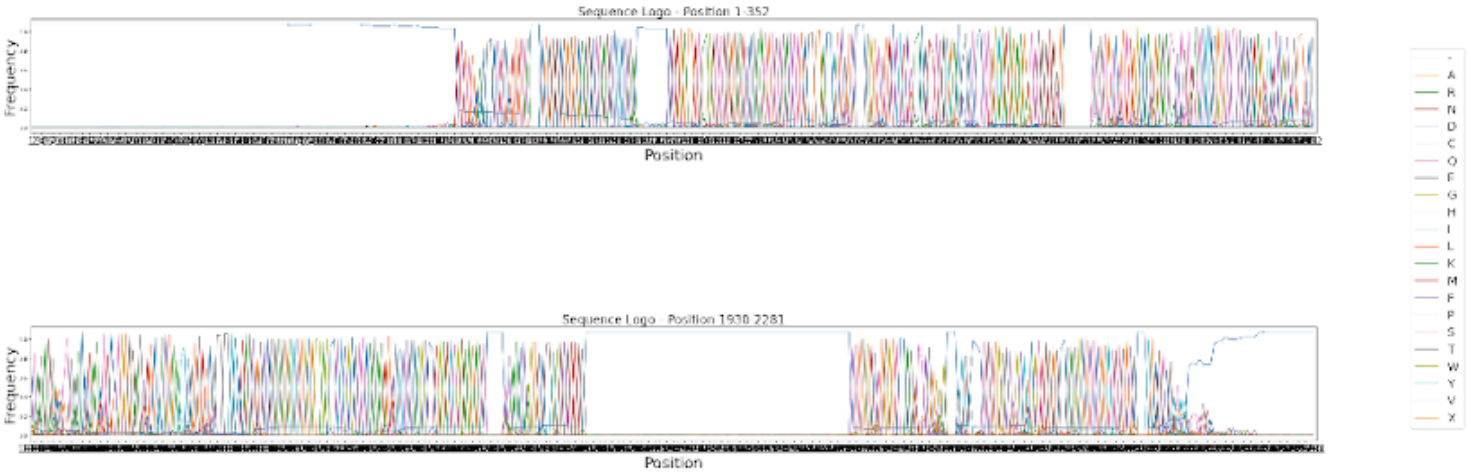

Fig. 8: Sequence logo of the RING (top) and BRCT (bottom) domain

In this form, the data is much more readable and allows for a more accurate analysis of amino acid frequency. When looking at this plot it is apparent that most of the positions have a really high frequency (upwards of 0.8), which indicates that at those positions there is little to no variance between the 100 analyzed amino acids. However there are some regions where some variance seems to be present.

The next thing we did was using the sequence logo to create a consensus sequence, which we compared to the sequence of our BRCA1 gene.

Fig. 9: Sequence comparison of the whole consensus sequence and human BRCA1

Because of the size of the BRCA1 gene, the plot of the whole sequence can not really be used, so we split up into domains again, using the same cutting points as before.

Fig. 10: Sequence comparison of RING (top) and BRCT (bottom)

The Sequence in the middle is the consensus sequence and if there are differences between the consensus sequence and the sequence it is compared to, the sequence splits up displaying the amino acid of the consensus sequence in green on top and the amino acid of the compared gene in red below of the consensus sequence. There are differences in about 9% of positions, but they are scattered almost equally across the whole gene, so there are no larger parts which have a completely different sequence.

Post-translational Modifications

In this study, we originally aimed to investigate patterns between DMS scores and specific positions within our dataset. To achieve this, we created a secondary dataset exclusively comprising of mutations with a bin_score of 0, focusing on mutations associated with cancer. Our initial analysis explored mutations categorized into polar, non-polar, positively charged, and negatively charged amino acids. We examined both newly introduced mutations within these groups and mutations originating from the aforementioned categories. We observed that in certain positions, all possible mutations led to the introduction of either a polar or non-polar amino acid, all of which exhibited particularly unfavorable DMS scores. However, the occurrence of such positions was a rarity, leading us to exclude them from further analysis.

Building upon our findings, we further investigated mutations involving the introduction or removal of post translational modification amino acids. Specifically, we constructed heatmaps for amino acids associated with phosphorylation (S, Y, T)(Fig. 12), acetylation (K)(Fig. 13), and glycosylation (S, T, N)(Fig. 11). We extensively analyzed mutations leading to these specific amino acids as well as mutations originating from amino acids capable of phosphorylation, glycosylation, or acetylation. Notably, we discovered that all mutations originating from these amino acids consistently resulted in the production of proteins with particularly low DMS score. Moreover, these mutations were found to exclusively occupy specific positions, allowing us to identify a total of 27 positions strongly associated with post translational modifications.

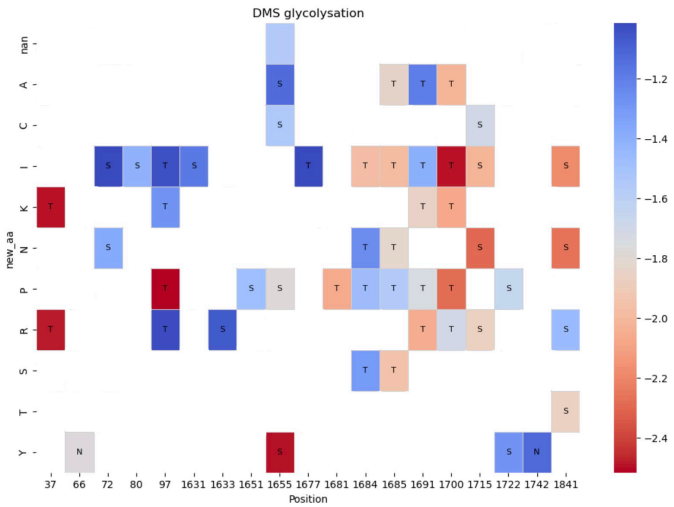

Fig. 11: The glycosylation heatmap shows the DMS scores of mutations originating from S, T, N (annotated on the heatmap), to new amino acids (y-axis). The position is represented on the x-axis. Only unfit proteins are taken into account.

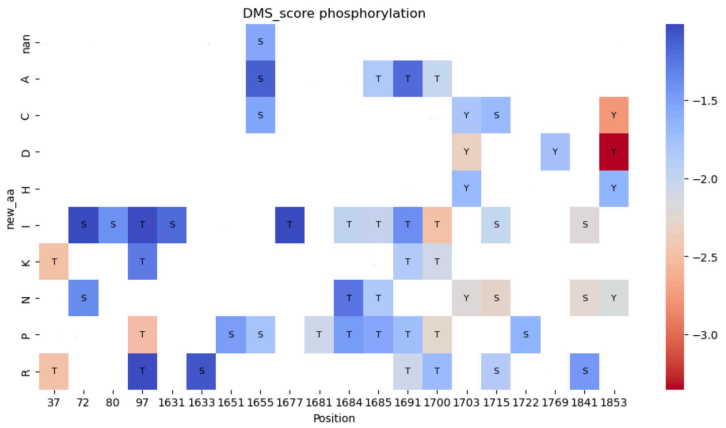

Fig.12: The phosphorylation heatmap shows the DMS scores of mutations originating from S, T, Y (annotated on the heatmap), to new amino acids (y-axis). The position is represented on the x-axis. Only unfit proteins are taken into account.

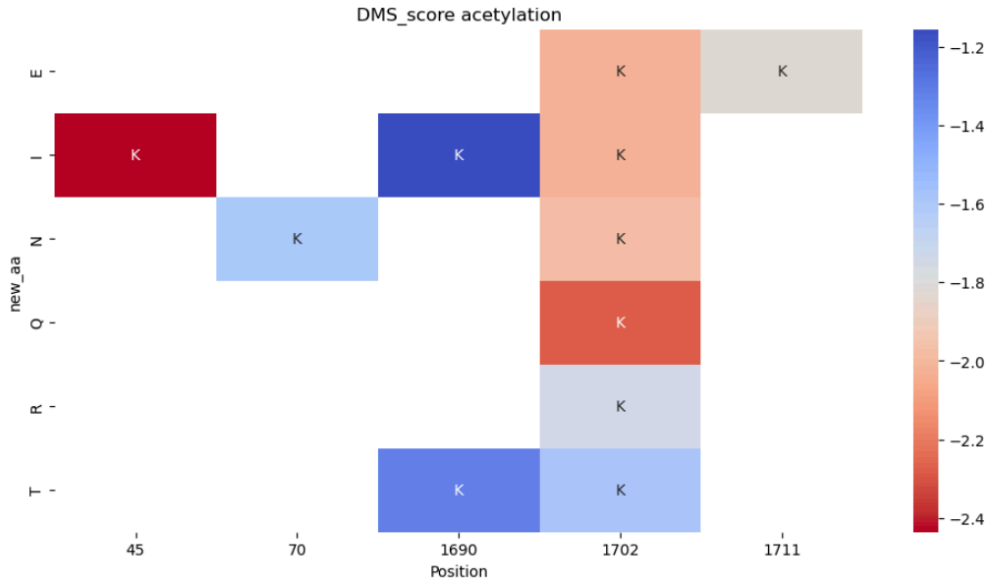

Fig. 13: The acetylation heatmap shows the DMS scores of mutations originating from K (annotated on the heatmap), to new amino acids (y-axis). The position is represented on the x-axis. Only unfit proteins are taken into account.

Subsequently, we focused our attention on mutations with a bin_score of 1, again focusing on those leading to the loss or gain of an amino acid capable of executing post translational modifications. We curated a new dataset exclusively comprising of mutations with a bin_score of 1 and conducted further investigations. Assuming a DMS score of 0 for the wild-type BRCA1 sequence, we aimed to interpret our results within this context. We observed that mutations originating from phosphorylation (Fig.14) or glycosylation (Fig.15) amino acids almost always exhibited DMS scores below 0, with a few exceptions. Similarly, mutations leading to these amino acids (Fig.16) also predominantly had DMS scores below 0.

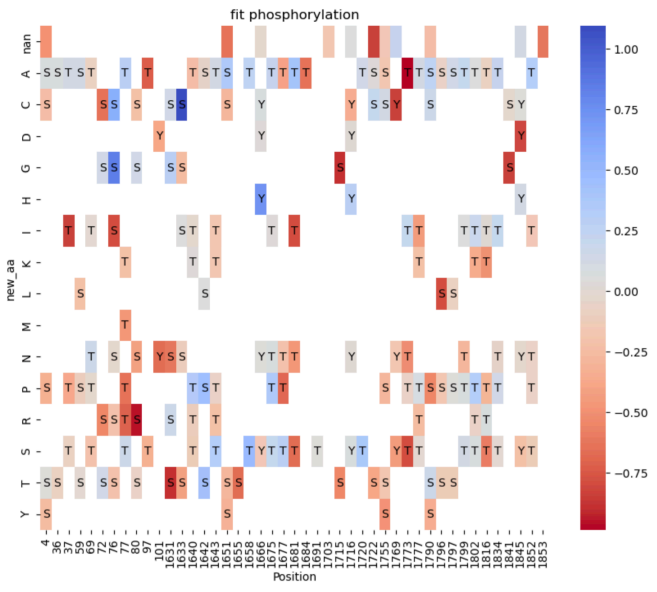

Fig. 14: The phosphorylation fit heatmap shows t he DMS scores of mutations originating from S, T, Y (annotated on the heatmap), to new amino acids (y-axis). The position is represented on the x-axis. Only fit proteins are taken into account.

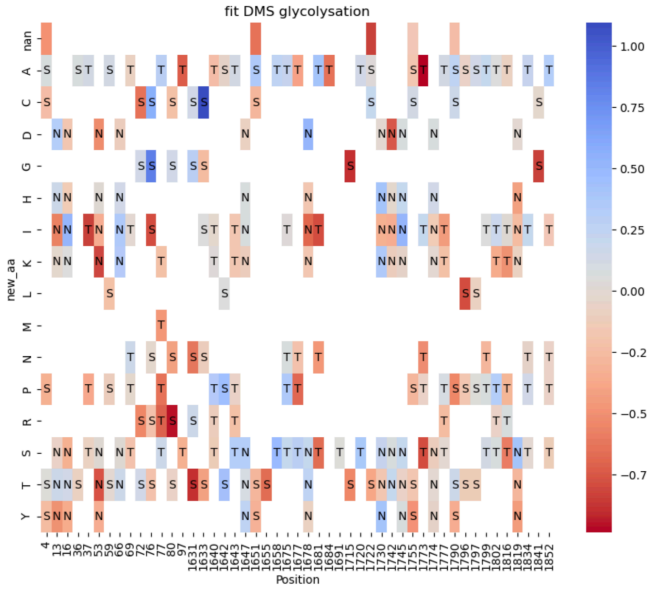

Fig. 15: The glycosylation fit heatmap shows the DMS scores of mutations originating from S, T, N (annotated on the heatmap), to new amino acids (y-axis). The position is represented on the x-axis. Only fit proteins are taken into account.

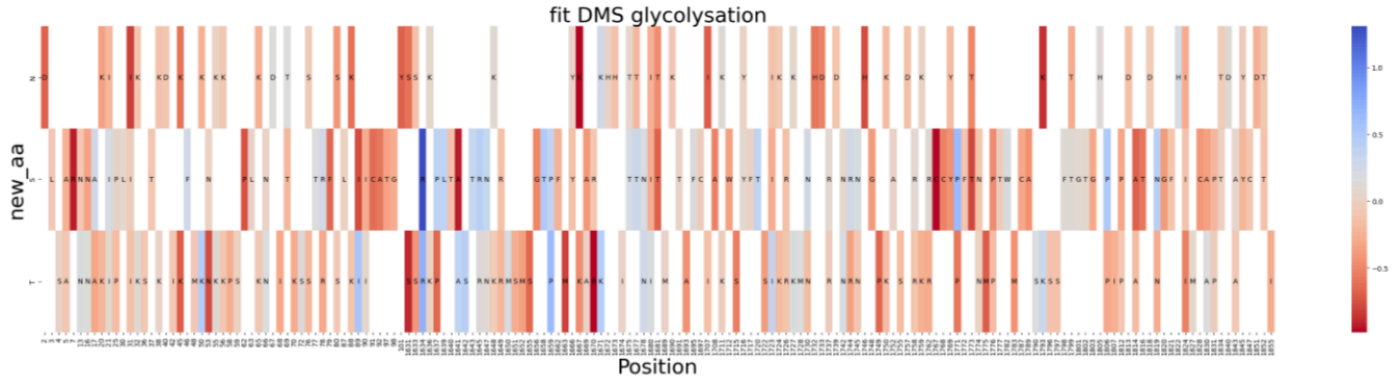

Fig. 15: The glycolysation fit heatmap shows the DMS scores of mutations leading to S, T, N, (y-axis). Annotated on the heatmap are the precursor amino acids. The position is represented on the x-axis. Only fit proteins are taken into account.

In contrast, mutations originating from (Fig.18) or leading to lysine (Fig.17), which is responsible for acetylation, displayed a less straightforward pattern. These mutations exhibited a mix of positive and negative DMS scores, indicating a lack of immediately identifiable pattern. Interestingly, a significant number of mutations resulting in the loss of lysine had DMS scores above 0.

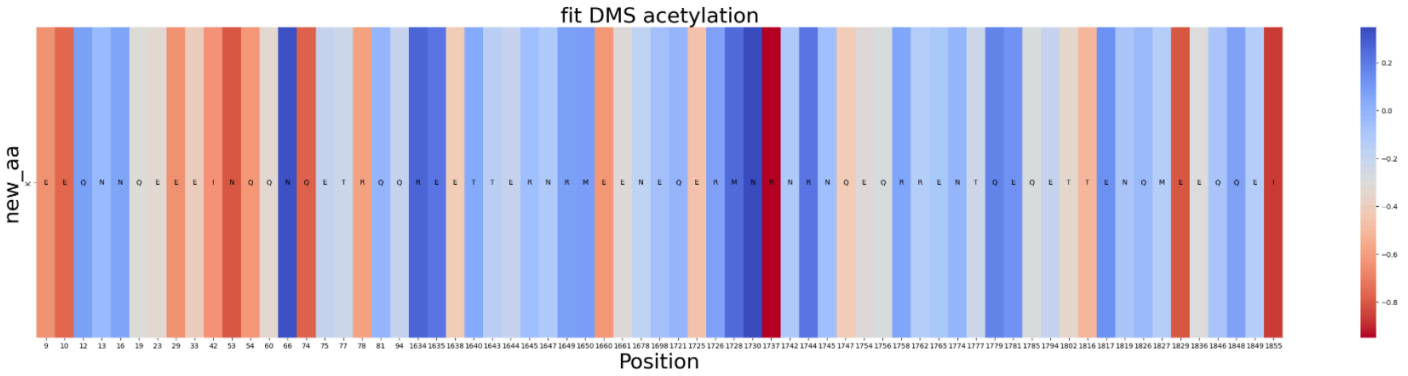

Fig. 17: The acetylation fit heatmap shows the DMS scores of mutations leading to K, with the precursor amino acids annotated on the heatmap. The position is represented on the x-axis. Only fit proteins are taken into account.

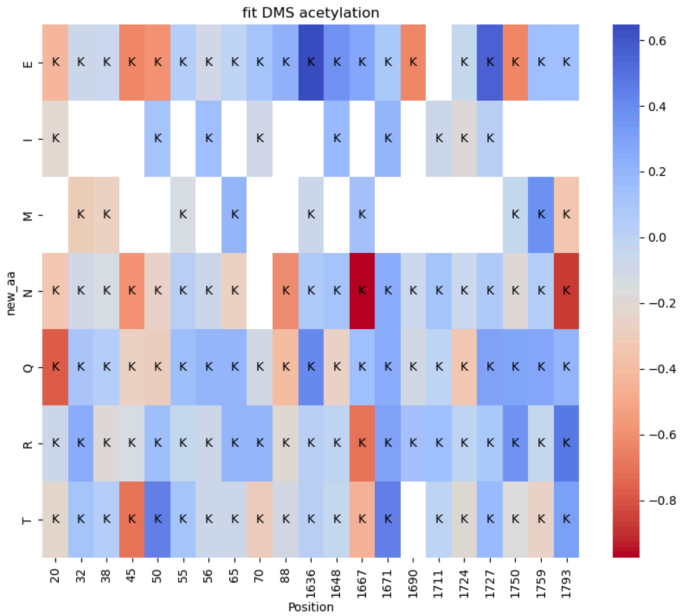

Fig. 18. The acetylation fit heatmap shows the DMS scores of mutations originating from K (annotated on the heatmap), to new amino acids (y-axis). The position is represented on the x-axis. Only fit proteins are taken into account.

Discussion

The ratios we got in Table 1 for the distribution of bin_0 scores for each secondary structure element in the RING domain are with 30.77% for α-Helices and 29.58% for β-Strands quite even. These distributions can also be observed in Fig. 5. The same can not be said for the BRCT domain, here the distributions (α-Helices: 27.45%; β-Strands: 40.54%; Coils: 21.57%) show a clear tendency suggesting that a mutation in a β-strand has the highest risk of making the resulting protein non functional. This can also be seen very clearly in the visualisation for this domain (Fig. 6), here the amount of red (=bin_0) is proportionally the highest in the β-strands.

To figure out whether these results were statistically relevant and could maybe be used to predict the effect of a SNV, we used the chi-square test method. We chose this method because it is suitable for our data since it works with independent, categorical data and big sample sizes (unlike the Fisher Exact Test which is used for small sample sizes).

The H0-Hypothesis for the test was “That there is no correlation between the type of secondary structure and how many of its SNVs have a bin_0 score”. For the RING domain this gave us a p-value of ~0.88 (>0.05) suggesting to not reject the H0-Hypothesis. For BRCT we got a p-value of ~0.0005 (<0.05) suggesting to reject the H0-Hypothesis.

Thus the result of our secondary structure analysis is, that in future prediction methods of the BRCA1 genes functionality the type of secondary structure a mutation is irrelevant for the RING domain, but could be a valuable factor if the mutation is located in the BRCT domain.

It would be interesting to conduct further tests and experiments in order to figure out, why the secondary structure element of a SNV is more important for BRCT than in RING.

Next up we performed conservation analysis, which showed the following data:

81.1% of our gene hit an amino acid frequency of 70%

68.6% hit an amino acid frequency of 80%

42.7% hit an amino acid frequency of 90%

25.2% hit an amino acid frequency of 95%

This shows that there is at least some level of conservation present, but further investigation in the form of a statistical test is needed to determine how significant it really is. Additionally, to set this into a better context, a similar analysis should be performed on a MSA of a certified highly conserved gene, for example a histone. While keeping those limitations in mind, this result is in line with our expectations, as the overall pretty low DMS-scores of our dataset led us to believe that considerable levels of conservation are present, as a highly conserved protein is not expected to be tolerant regarding mutations. The majority of regions with visibly low conservation levels were outside of our dataset, which only covers the RING and BRCA domain and are thus of little importance for this study. However, they could be interesting when working with a dataset that covers the entirety of the human BRCA1 gene. To further examine the accuracy of this thesis it would also be sensible to run the analysis with another MSA consisting of other, preferably more, sequences.

To get more information about the conservation levels of our specific gene, a sequence comparison with a consensus sequence created from the “fasta_file” was performed. This showed that about 90% of the sequences were identical, further backing up the thesis of considerable levels of conservation being present. It also has to be noted that the differences between human BRCA1 and the consensus sequence occur scattered all over the sequence, with the largest group of differing positions being 4. That implies that even though about 10% of our gene is not identical with the consensus sequence, there are no regions of the protein which are completely non homologous. This again is evidence for conservation, because some amino acid differences in positions of potentially little evolutionary relevance seem to be less altering in protein function than

the change of bigger, more coherent regions. This is of course only a thesis and would need to get thoroughly tested in order to be of relevance.

Next, we conducted an analysis to investigate the patterns between DMS scores and specific amino acids within our dataset. We focused on mutations that might be associated with cancer by creating a secondary dataset comprising mutations with a bin_score of 0.

Our initial analysis examined mutations categorized into polar, non-polar, positively charged, and negatively charged amino acids. We observed that in certain positions, all possible mutations resulted in the introduction of either a polar or non-polar amino acid, which consistently exhibited unfavorable DMS scores. However, given the rare occurrence of such positions, we excluded them from further analysis, suggesting their minimal impact on the overall patterns observed.

To delve deeper into the functional consequences of mutations, we focused on amino acids, which could be post translationally modified. We constructed heatmaps to visualize the impact of mutations associated with phosphorylation (S, Y, T), acetylation (K), and glycosylation (S, T, N) on their DMS scores. Our analysis revealed a consistent trend: mutations originating from or leading to these specific amino acids consistently led to the production of proteins with low DMS scores. This observation highlights the functional importance of post translational modifications in protein structure and stability. Moreover, we identified specific positions strongly associated with post translational modifications, totaling 27 positions. The loss or gain of modifiable amino acids in these positions resulted in the production of unfit proteins, underscoring their critical role in maintaining protein functionality.

Furthermore, we investigated mutations with a bin_score of 1, which only include fit proteins. By focusing on mutations leading to the loss or gain of amino acids capable of being modified, we aimed to interpret their impact on DMS scores compared to the wild-type BRCA1 sequence. We assumed that the wild-type has a DMS score 0. Intriguingly, mutations introducing phosphorylation or glycosylation amino acids consistently, with few exceptions, exhibited DMS scores below 0, indicating a detrimental effect on protein functionality. Similarly, mutations originating from these amino acids and introducing proteins not capable of post transitional modifications had DMS scores below 0, reaffirming the importance of these amino acids. Despite being classified as fit mutations, they exhibited lower DMS scores compared to the wild-type BRCA1 sequence. This means that all mutations, fit and unfit, that might be involved in post translational modification, lead to a sequence that performs worse than the wild type, suggesting their potential functional implications in disease development.

In contrast, mutations involving lysine, which is a potential target for acetylation, displayed a more intricate pattern. These mutations exhibited a mix of positive and negative DMS scores, lacking an obviously identifiable pattern. Interestingly, a substantial number of mutations resulting in the loss or gain of lysine had DMS scores above 0. This observation implies that the mutated protein may exhibit improved performance compared to the wild-type BRCA1, suggesting potential compensatory mechanisms or alternative functional pathways.

Overall, our analyses uncovered distinct patterns in DMS scores concerning different positions and amino acid modifications. The indicated correlation between mutations introducing phosphorylation and glycosylation amino acids and low DMS scores emphasizes the importance of these modifications in protein functionality. In contrast, lysine mutations demonstrated a more complex relationship, warranting further investigation. These findings contribute to a deeper understanding of the functional consequences of specific mutations associated with post translational modifications and their implications in cancer development.

Future studies should focus on validating our findings and exploring the underlying molecular mechanisms to provide valuable insights into BRCA1-related diseases and potential therapeutic interventions.

References

(1) Accurate classification of BRCA1 variants with saturation genome editing - Gregory M. Findlay1, riza M. Daza1, Beth Martin1, Melissa D. Zhang1, Anh P. leith1, Molly Gasperini1, Joseph D. Janizek1, Xingfan Huang1, lea M. Starita1,2* & Jay Shendure1,2,3* https://www.nature.com/articles/s41586-018-0461-z

(2) Jones, D. T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. 292, 195-202

(3) Korf, Ian, Mark Yandell, and Joseph Bedell. *Blast*. " O'Reilly Media, Inc.", 2003.

(4) Edgar, Robert C., and Serafim Batzoglou. "Multiple sequence alignment." *Current opinion in structural biology* 16.3 (2006): 368-373.

Appendix

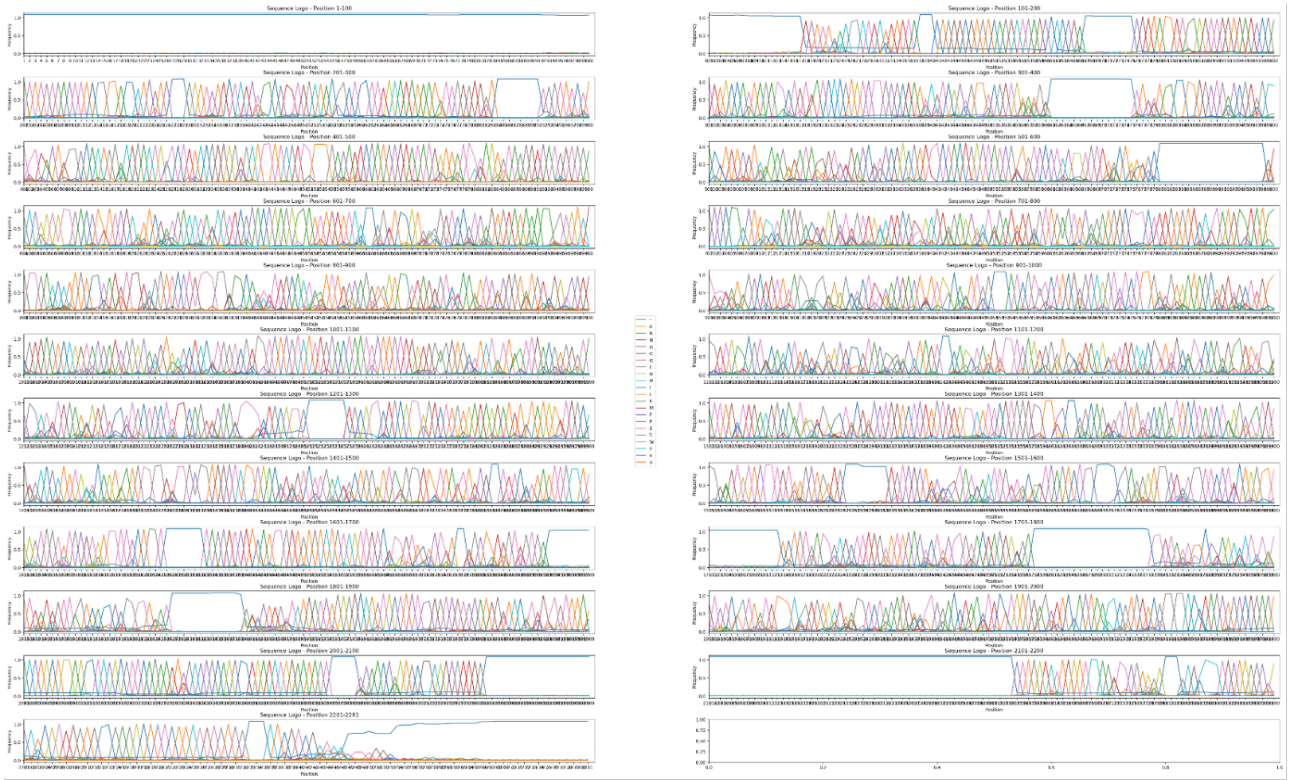

Fig. 19: Sequence logo split into blocks of 100 amino acids

Fig. 20: Sequence comparison split into blocks of 100 amino acids