Assignment 3: Analysing the Fitness of GFP variants

Abdullah Faqih Al Mubarok - vpx267

October 25, 2023

1 Task 1: quality control and translation

1. How many unique barcodes are found? How many unique DNA variant sequences?

   Answer:

   Before cleanup, there are 68,039 barcodes and 58,359 unique DNA variants.

   After cleanup, ther are 65,679 barcodes and 56,029 unique DNA variants.

2. How many unique protein sequences after cleanup?

   Answer:

   There are 51,716 unique protein sequences after cleanup.

3. Determine is the most common protein sequence that is not wild-type, and report the mutation(s) found in this sequence. Keep in mind that a protein sequence can be encoded by several different barcodes.

   Answer:

   The moost non-WT common protein sequence which has 59 uniqueBarcodes:

   SKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGK
   LPVPWPTLVTTLSYGVQCFSRYPDHMKQHDFLKSAMPEGYVQERTIFFKD
   DGNYKTRAEVKFEGDTLVNRIELKGIDFKEDGNILGHKLEYNYNSHNVYIM
   ADKQKNGIKVNFKIRHNIEDGSVQLADHYQQNTPIGDGPVLLPDNHYLSTQ
   SALSKDPNEKRDHMVLLEFVTAAGITHGMDELYK*

   List of the mutation(s) of the most common protein sequence (non-WT):

   - F83L

2 Task 2: protein-level variants

1. Are the deviations you observe beyond what you expect based on the experimental error? Submit plot and discussion as part of your hand-in.

   Answer:

   It was anticipated that there would be a positive and linear correlation between the average score of all variants and the scores of individual variants. As seen from

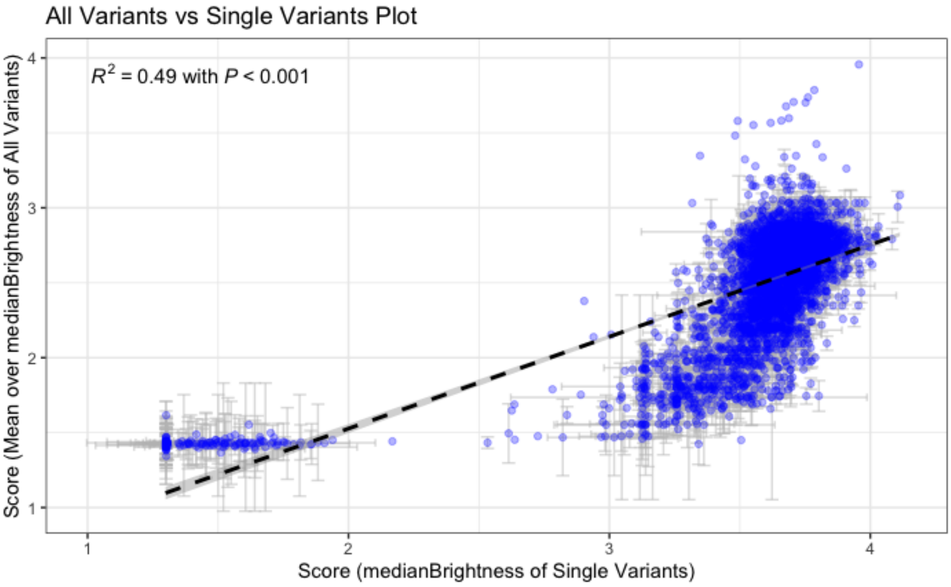


Figure 1: All Variants vs Single Variants Plot

figure 1, the two seem to be positively correlated. However, the variance of score over all variants are poorly explained by the score over single variants with simple linear additive model ($R^2 = 0.49$, $pval < 0.001$). This might be due to the presence of multiple mutations which may give rise to epistasis, which can manifest as either positive or negative interactions. Moreover, The epistasis effects seem to not appear if the variant scores were averaged only for the variants with single mutation (figure 2).

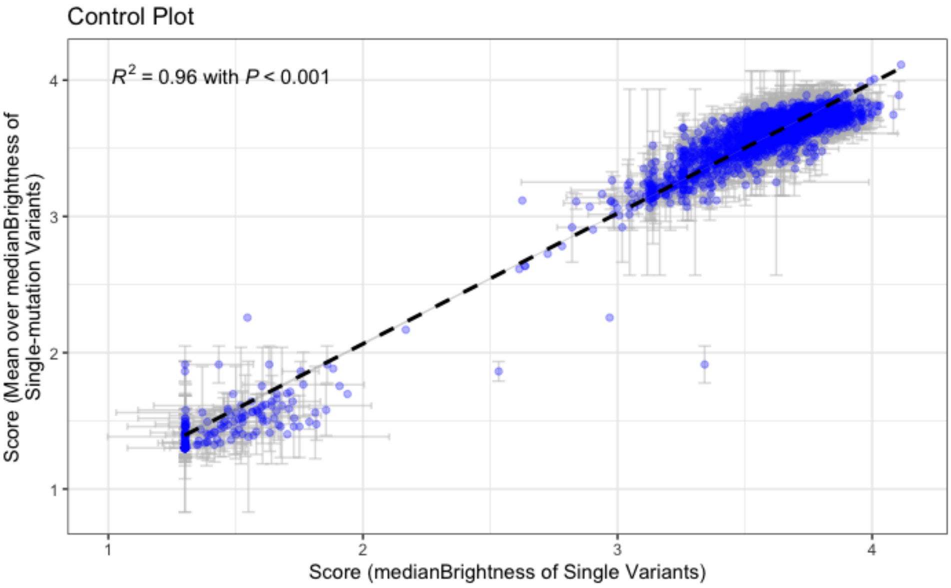


Figure 2: All Single Variants vs Single Variants Plot

2. remove all data that was only observed once, and repeat the comparison of averages

over all sequences vs. single mutations. Include that plot. How does the pattern change? What's your hypothesis for why we see this?

Answer:

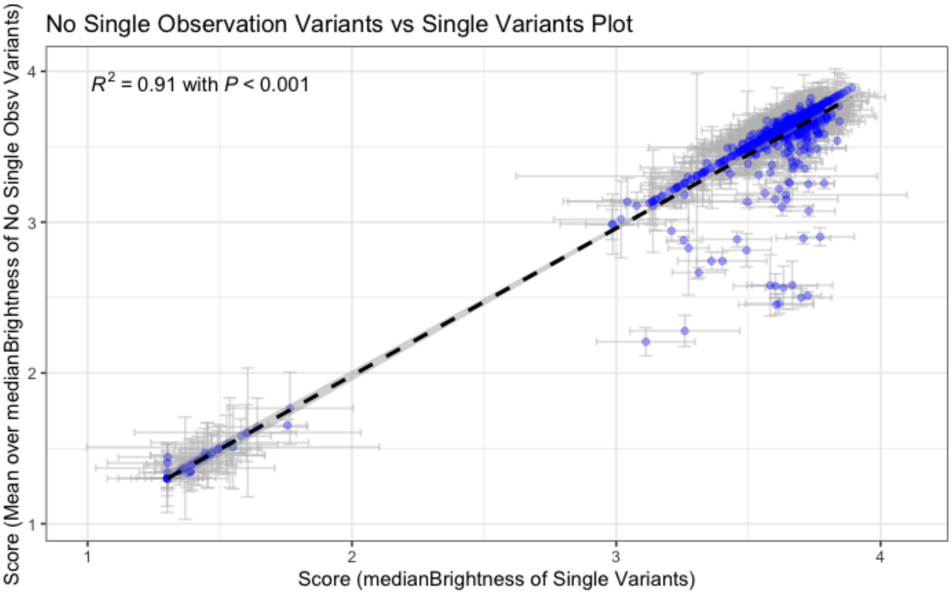


Figure 3: No Single Observation Variants vs Single Variants Plot

After removing the observations that possess only one uniqueBarcode, by fitting a linear additive model as above, the R^2 changes to be 0.91 ($pval < 0.001$) which indicates that the simple additive model explains more variance than before exclusion (91% vs 49%). This observation implies that the presence of a single unique barcode in the variants may result in an increased mutation rate as a consequence of sequencing errors. Additionally, in the absence of sequencing errors, these mutations may potentially induce structural instability in the protein, leading to impaired folding ability and consequent impact on both the abundance (uniqueBarcode) and score.

3 Task 3: summary matrix

1. Average the brightness data across all variants in a 20x20 matrix showing the wild-type and target amino acids, as we did in the exercises in class. Submit a plot of the matrix (see e.g. ex. 3) as part of your homework assignment

   Answer:

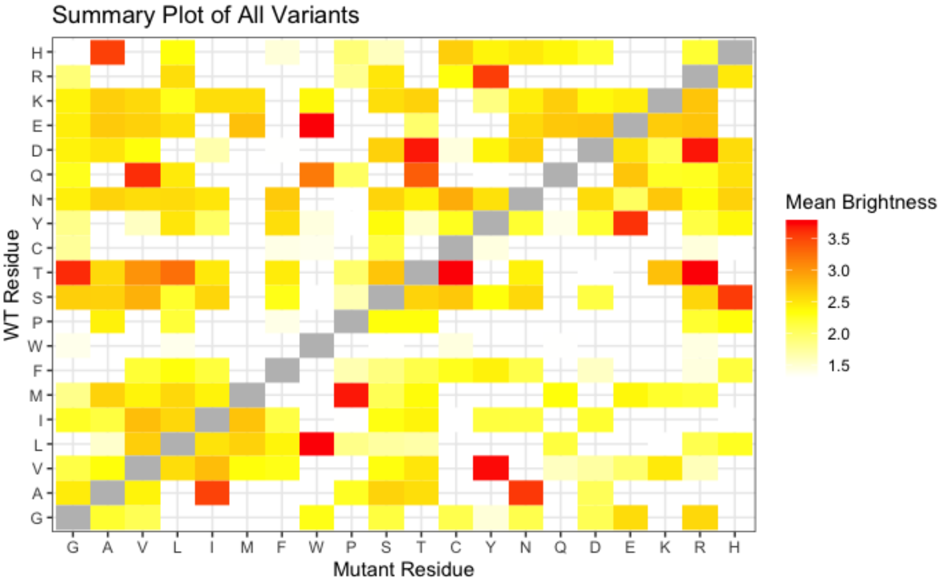


Figure 4: Summary Plot of All Variants

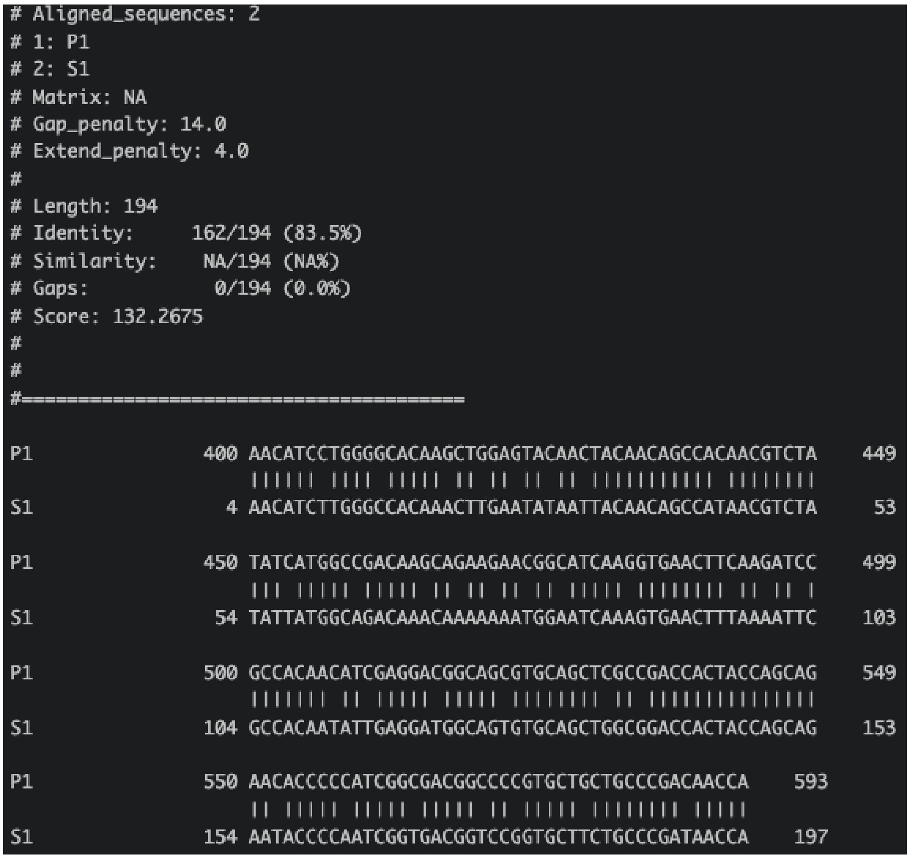
4 Task 4: compare to the other GFP mutagenesis dataset

1. Compare it to the nativeDNA included above (e.g. by pairwise sequence alignment), then translate both sequences to protein and compare those. Write a short paragraph describing what you observe. is local or global alignment more suitable here?

   Answer:

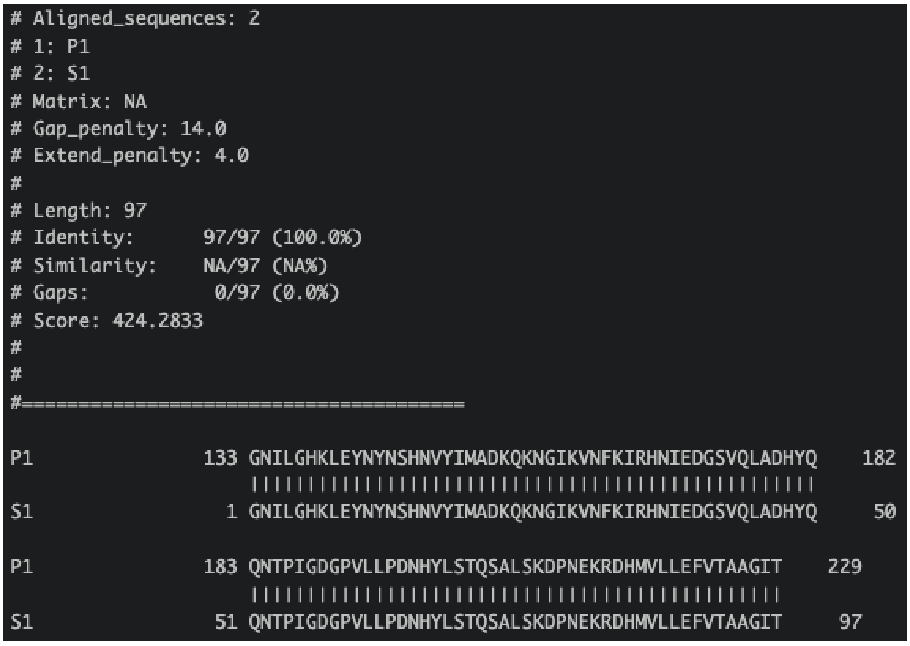
   Since the length of the nativeDNA from the exercise is half of nativeDNA from the assignment, It is more suitable to look for the alignment between subsequences of them and look for possible same domain. Thus, we use the local alignment.

   Based on the figure 5 and 6, we might conclude that the nativeAA from the exercise is a subsequence of the nativeAA from the assignment. In addition, despite the unmatched bases on the DNA pairwise alignment, the AA alignment is perfectly matched. We could also see that the exercise nativeAA aligned starts at index 133 until 229 of assignment NativeAA.



```
# Aligned_sequences: 2
# 1: P1
# 2: S1
# Matrix: NA
# Gap_penalty: 14.0
# Extend_penalty: 4.0
#
# Length: 194
# Identity:     162/194 (83.5%)
# Similarity:    NA/194 (NA%)
# Gaps:           0/194 (0.0%)
# Score: 132.2675
#
#
#=======================================

P1               400 AACATCCTGGGGCACAAGCTGGAGTACAACTACAACAGCCACAACGTCTA    449
                     |||||| |||| ||||| || || || || ||||||||||| ||||||||
S1                 4 AACATCTTGGGCCACAAACTTGAATATAATTACAACAGCCATAACGTCTA     53

P1               450 TATCATGGCCGACAAGCAGAAGAACGGCATCAAGGTGAACTTCAAGATCC    499
                     ||| ||||| ||||| || || || || ||||| |||||||| || || |
S1                54 TATTATGGCAGACAAACAAAAAAATGGAATCAAAGTGAACTTTAAAATTC    103

P1               500 GCCACAACATCGAGGACGGCAGCGTGCAGCTCGCCGACCACTACCAGCAG    549
                     ||||||| || ||||| ||||| |||||||| || |||||||||||||||
S1               104 GCCACAATATTGAGGATGGCAGTGTGCAGCTGGCGGACCACTACCAGCAG    153

P1               550 AACACCCCCATCGGCGACGGCCCCGTGCTGCTGCCCGACAACCA    593
                     || ||||| ||||| ||||| || ||||| |||||||| |||||
S1               154 AATACCCCAATCGGTGACGGTCCGGTGCTTCTGCCCGATAACCA    197
```

Figure 5: Pairwise Alignment of Exercise nativeDNA and Assignment nativeDNA



```
# Aligned_sequences: 2
# 1: P1
# 2: S1
# Matrix: NA
# Gap_penalty: 14.0
# Extend_penalty: 4.0
#
# Length: 97
# Identity:      97/97 (100.0%)
# Similarity:    NA/97 (NA%)
# Gaps:           0/97 (0.0%)
# Score: 424.2833
#
#
#=======================================

P1               133 GNILGHKLEYNYNSHNVYIMADKQKNGIKVNFKIRHNIEDGSVQLADHYQ    182
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
S1                 1 GNILGHKLEYNYNSHNVYIMADKQKNGIKVNFKIRHNIEDGSVQLADHYQ     50

P1               183 QNTPIGDGPVLLPDNHYLSTQSALSKDPNEKRDHMVLLEFVTAAGIT    229
                     |||||||||||||||||||||||||||||||||||||||||||||||
S1                51 QNTPIGDGPVLLPDNHYLSTQSALSKDPNEKRDHMVLLEFVTAAGIT     97
```

Figure 6: Pairwise Alignment of Exercise nativeAA and Assignment nativeAA

2. How many variants are observed in both datasets? only observed in the Sarkisyan dataset? only observed in the dataset we worked with in class?

Answer:

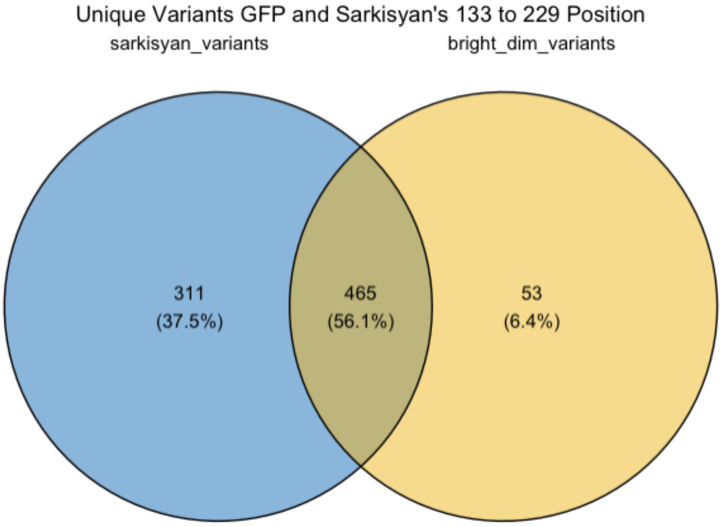


Figure 7: Venn Diagram of Unique Variants of GFP and Sarkisyan's 133 to 229 Position

From the figure 7, we can see that there are 465 variants observed in both datasets, 53 variants only observed in the dataset we worked with in class, and 311 variants only observed in the Sarkisyan dataset.

3. For the variants found in both datasets, create a scatterplot to compare their averaged medianBrightness (see task 2) vs. log(bright/dim) ratio. Briefly describe what trends you observe, and whether those are what you would expect.

   Answer:

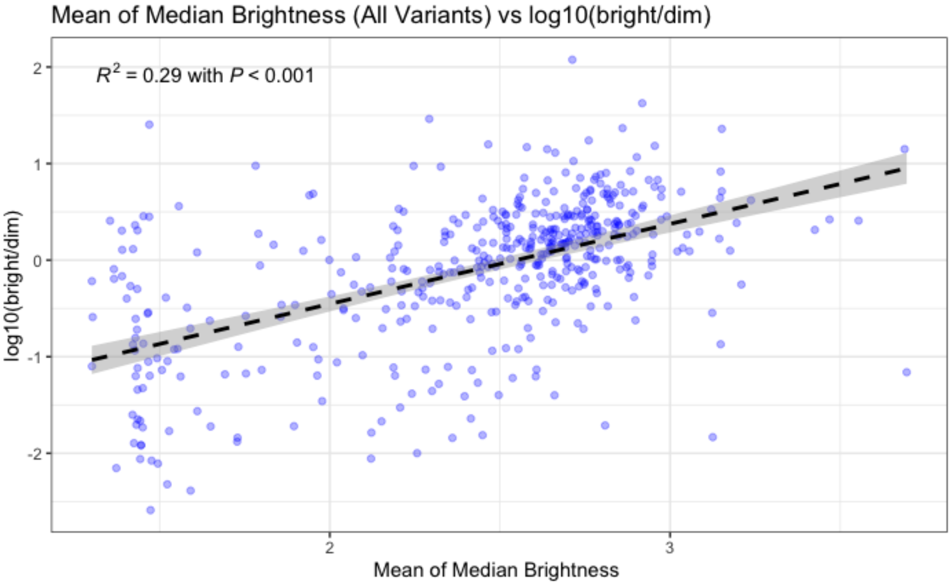


Figure 8: Mean medianBrigthness from Sarkisyan vs log10(bright/dim) from Exercise

Since the mutations occurred within the same domain, I expected that the mean of medianBrightness vs log(bright/dim) ratio plot should be positively correlated.

Figure 8 could describe that the two seem to be positively correlated. However, using the linear additive model, the mean of medianBrightness could only explain 29% of the log(bright/dim) variance ($pval < 0.001$). The low R^2 might be due to the differences in the experimental methods and/or conditions.

5 Task 5: integrating MAVE assays of stability and activity

1. Create a scatter plot of all the variants described by the two MAVEs.

   Answer:

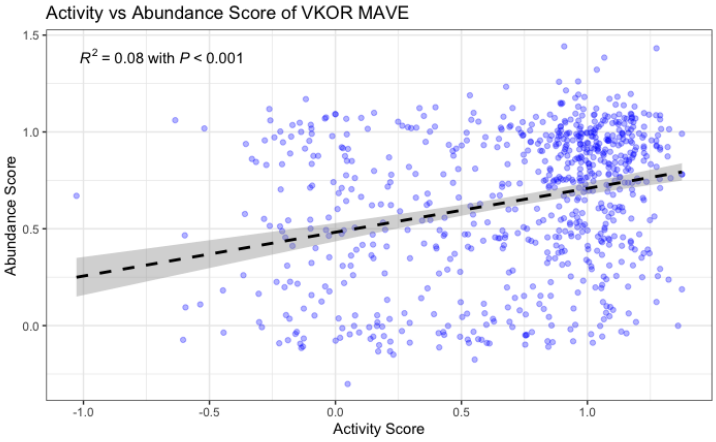


Figure 9: Activity vs Abundance VKOR

2. Create a scatter plot of only the variants listed in gnomAD. Label the variants that have annotation in the "Clinical significance" field.

   Answer:

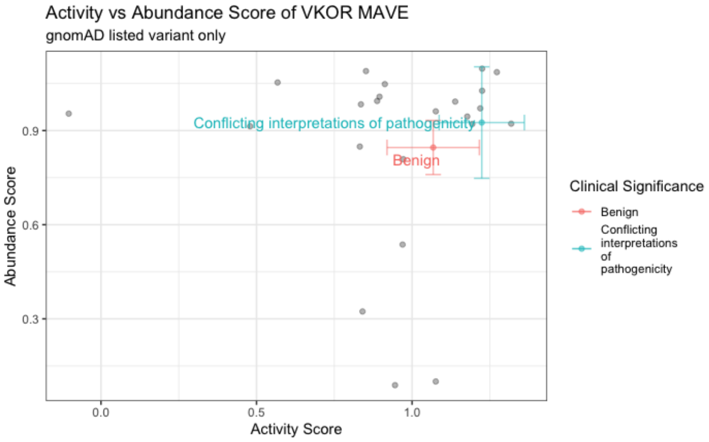


Figure 10: Activity vs Abundance VKOR gnomAD Variants

3. How many variants have an abundance score below 0.5?

   Answer:

   Number of intersection of the MAVEs and gnomAD variants which have abundance score below 0.5: 3

4. How many variants have an activity score below 0.4 and would thus be categorised as inactive?

   Answer:

   Number of intersection of the MAVEs and gnomAD variants which have activity score below 0.4: 1

5. How many variants are in the intersection of those two categories, so, low abundance and low activity according to the assays?

   Answer:

   Number of intersection of the MAVEs and gnomAD variants which have low abundance and activity score: 0

6. There are variants with conflicting interpretations in ClinVar. Can we make a better estimate regarding their effects after having seen the assay data?

   Answer:

   Since the conflicting interpretation exhibits a high level of both activity and abundance scores, and furthermore cannot be distinguished from the benign label, it is reasonable to infer that this variant does not have any influence on gene function and does not contribute to disease progression (benign).