Assignment 3

Mahdi

October 19, 2020

Task 1

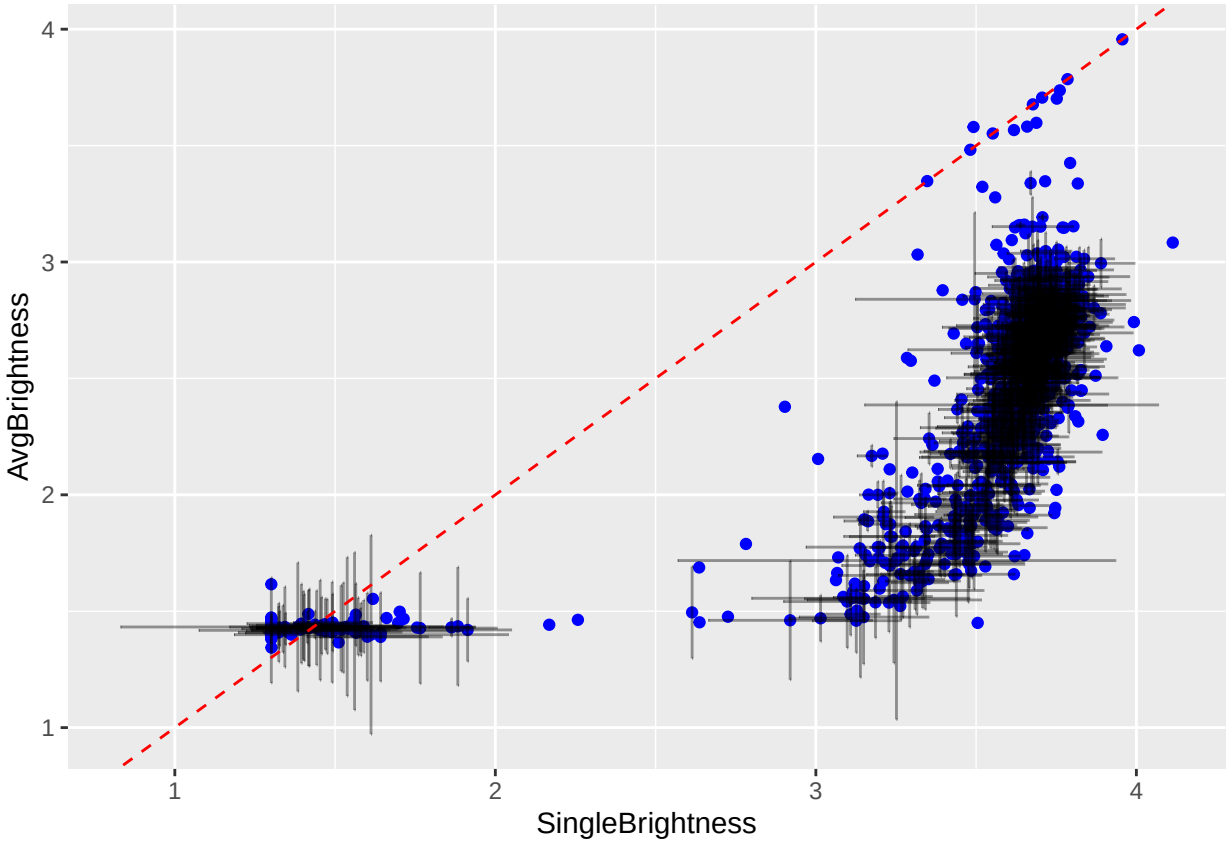
56029 unique barcodes are found.
51716 unique protein sequences are found.
The most common sequence:
SKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKLPVPWPTLVTTLSYGVQ
CFSRYPDHMKQHDFFKSAMPEGYVQERTIFFKDDGNYKTRAEVKFEGDTLVNRIELKGID-
FKEDGNIL GHKLEYNYNSHNVYIMADKQKNGIKVNLKIRHNIEDGSVQLADHYQQNTPIGDG-
PVLLPDNHYLSTQS ALSKDPNEKRDHMVLLEFVTAAGITHGMDELYK*

Task 2

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

```
## [1] "sequence"         "uniqueBarcodes"   "medianBrightness" "stdErr"
```

```
## Warning: Removed 382 rows containing missing values (geom_errorbarh).
```

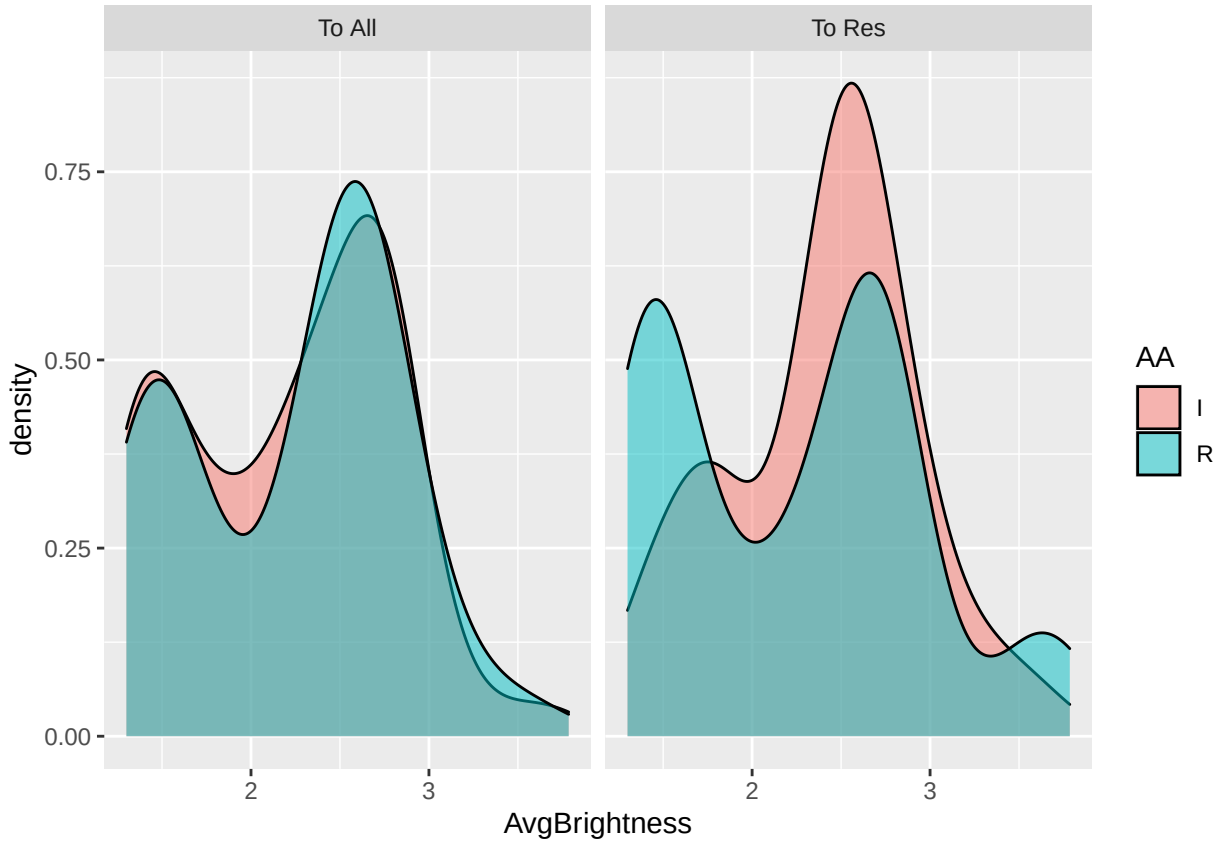


Since we know the majority of mutations lead to a decrease in fitness, we expect the brightness across sequences with multiple mutation to be smaller than the brightness for sequences with a single mutation. For example, Mutation A results in high fitness so the Brightness for sequences with single mutations (only Mutation A) will be high. But many sequences may have Mutation A along with additional mutations which reduce the fitness and hence brightness, so when taking the average of all those sequences, the brightness will be lower.

From the plot we can see two main clusters, one where both brightness across sequences and the brightness for single mutation sequences are high and one where both are low. However, the single mutation brightness is lower than the brightness across sequences for the majority of mutations, which is as expected.

Many points have large error values for both brightness values, while some points do not have any error bars because the sequeunce had a unique Barcode of 1 and thus no error value was recorded.

Submit the plots for all 4 distributions as part of your handin

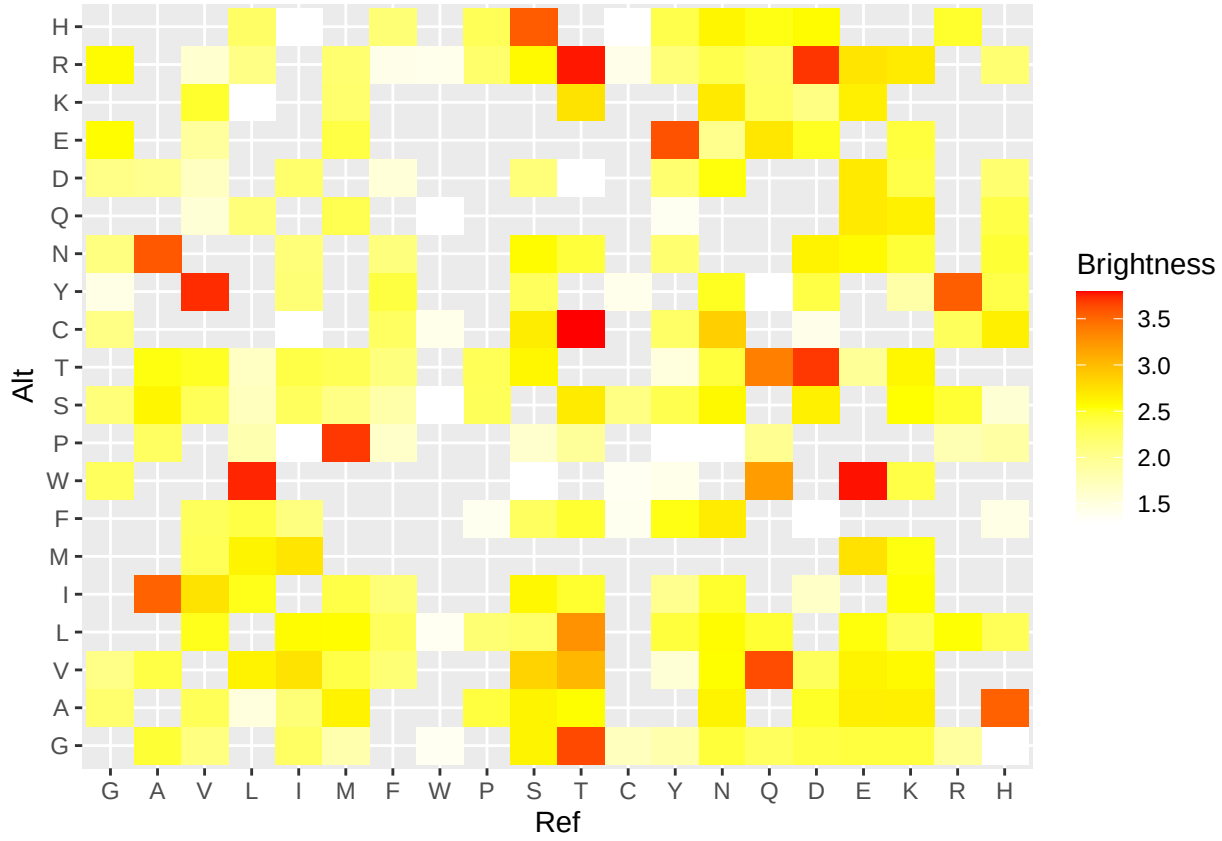


For a synonymous mutation, we expect the brightness to remain high because the sequence is unchanged. For a missense mutation, we expect the brightness to be comparitively high if the mutation doesn't significantly alter the properties of the amino acid. If the property is altered, then the brightness is expected to be low due to decreased fitness. The location of the mutation also plays a role as mutations in some locations will not affect the fitness much while other locations will have a large effect.

Isoleucine is a hydrophobic amino acid while Arginine is a hydrophilic amino acid. From the plot on the left, we can see that multiple peaks form. Each peak corresponds to the chosen residue being converted to either a residue of the opposite group, which results in a peak with low brightness/fitness or a residue in the same group, which results in a peak with higher brightness and thus fitness. For example, the peak with a lower brightness for Isoleucine could correspond to a mutation to a hydrophilic residue and then the peak with a higher brightness could correspond to a mutation to a hydrophobic residue.

For the plot on the right, we can see three peaks when mutating all amino acids to arginine. The large peak with a low brightness is probably caused by a mutation from hydrophobic amino acids. This may be because mutating a hydrophobic amino acid to a hydrophilic amino acid in the core of the protein may have destabilized it. Similarly, the peaks with medium and high brightness may have been caused by mutations from hydrophilic amino acids which would not reduce brightness much, or mutations at certain locations that strongly increased the fitness. Mutation to Isoleucine resulted in most sequences having a medium brightness probably because the neither the core nor the outer regions were strongly affected by this change.

Task 3



Task 4

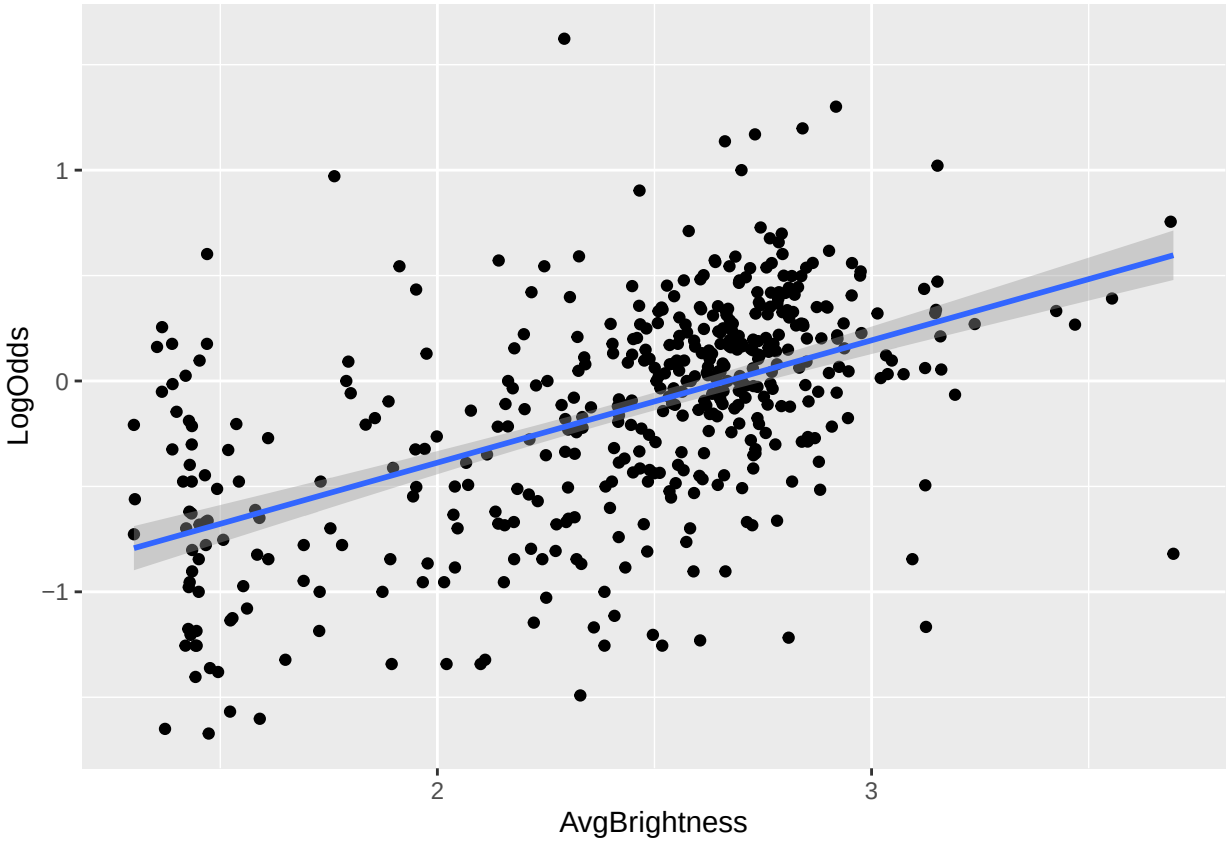
4.1

The DNA from the exercise aligns to the DNA from the assignment at position 397. The Protein from the exercise aligns to the Protein from the assignment at position 133. The protein sequences align perfectly except for the last amino acid in the exercise protein which is replaced by a stop codon. In contrast, the DNA sequence has many mismatches throughout the alignment. This is due to the redundancy of the genetic code which allows different codons to code for the same amino acid.

4.2

465 variants are found in both datasets
1346 variants are found in only the Sarkisyan dataset
59 variants are found only in the dataset we used in class

4.3



A logOdds value greater than zero tells us a mutation is more prevalent in the bright dataset which means it has a higher fitness. Conversely, a logsodds value less than zero tells us the mutation has low brightness. Since the two sequences align perfectly, we would expect a high fitness in one to indicate a high fitness in the other, which would mean we expect a high brightness across sequences to correspond to a high log odds ratio. We observe a positive correlation between the two values, which is what we expect. Although the correlation is not very strong which may be explained by a variety of causes such as loss of useful data when filtering the exercise dataset, the two proteins being differently affected by the mutations due to their different size, noise in the data, etc.