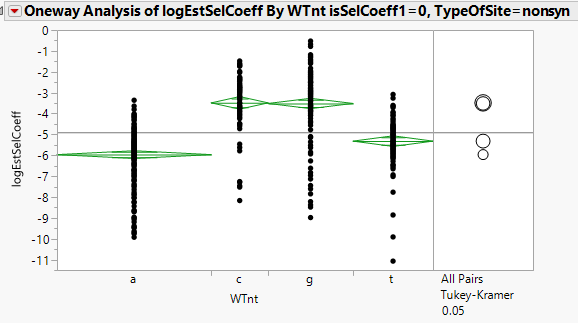
EstSelCoeff predictors

# Methods

Work described took the estimated selection coefficient for each position and tried to see the effects of several predictors on the coefficient for nonsyn mutations, in order to explain why several mutations tend to have lower values, while others tend to have higher values. I used the tests on the LOG transformation of estSelCoeff. I also ignored res, stop and syn mutations, as well as those nonsyn mutations who were absent. This makes sense as in our data absent mutations and STOP mutations get the same estSelCoeff, while premature STOP codons obviously differ from rare mutations so they should have different coefficient which is not the current case.

# Nucleotide Identity

ANOVA shows that nucleotide identity affects the selection coefficient: for C and G nucleotides, the average log(estSelCoeff) is higher by 2.5-3 orders. Further regression found that this independent variable has the highest effect on predicting logEstSelCoeff value.

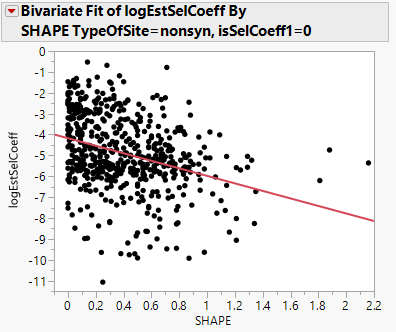


It makes sense, as A's tend to appear outside of RNA structured elements (<http://www.tandfonline.com/doi/abs/10.4161/rna.22896>), so it should be interpreted as "structural factor".

# Shape Reactivity

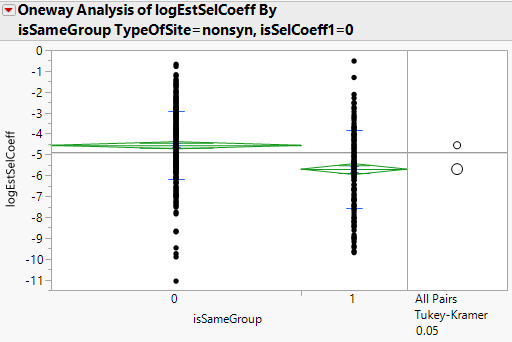
Similar to nucleotide identity, SHAPE contains information about the RNA folding of the position: positions that has lower shape score (median=0.34) tend to create secondary RNA structures. There's a statistically significant linear connection between SHAPE scores and log(EstSelCoeff) (P-value<0.0001 for both intercept and SHAPE), and the equation is:

logEstSelCoeff = -4.171505 - 1.8041432\*SHAPE



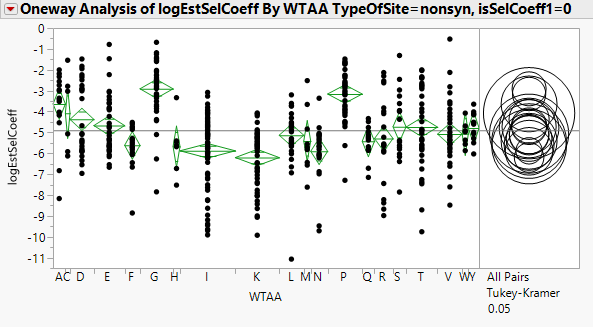
# Same AA Group

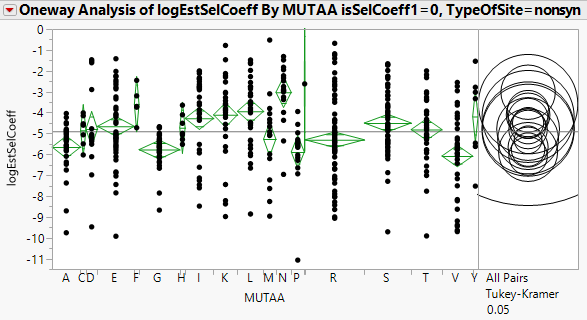
Amino acids were classified into 5 groups: Positively charged, negatively charged, uncharged, hydrophobic and special cases (Cysteine, Glycine and Proline). If the WT and Mut belonged to the same group, they received "1" otherwise "0". From the analysis it looks like the mean difference between the "1" and "0" groups is -1.1435. This difference is statistically significant (p-value<0.0001); where the target mutation is amino acid from a different group, the selection cost is higher.



# AA Identity

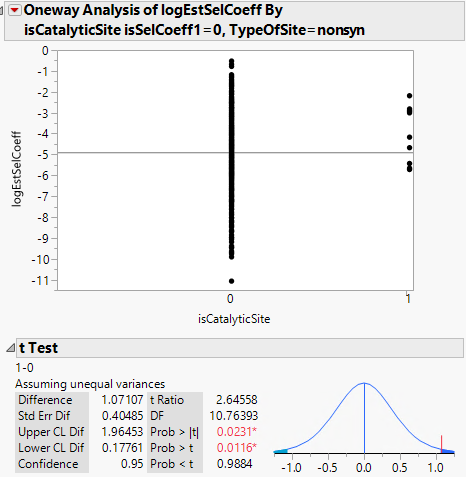
Both WT and Mutation AA identity showed statistical significances. There are some AAs that when are WT or Mutation show different selection penalty. This is not a balanced ANOVA design but ANOVA is robust to most assumptions violations.





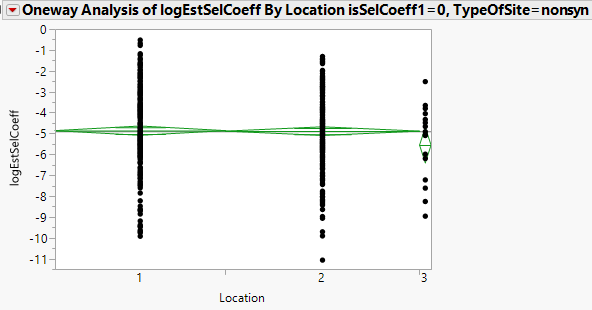
# Catalytic Site

There are a total of 6 catalytic sites in the data: 3 in PROT and 3 in RT. They sum to a total of 11 nonsyn mutations, and there's a statistical significance for having higher logEstSelCoeff for catalytic sites (p-value=0.0116 when assuming equal variances between the groups). The difference itself is 1.07107.



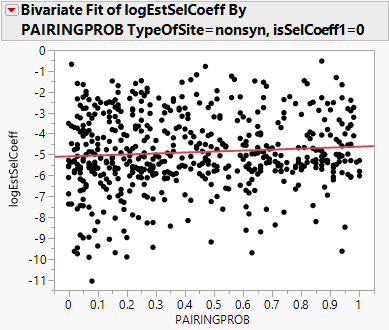
# Codon Position

Using codon positions (1, 2 and 3) to distinguish between higher and lower selection failed; looks like there's no statistical significance in this path.



# Pairing Probabilities

The SHAPE paper also introduces pairing probabilities. By themselves they don't show statistical significance to the logEstSelCoeff values.



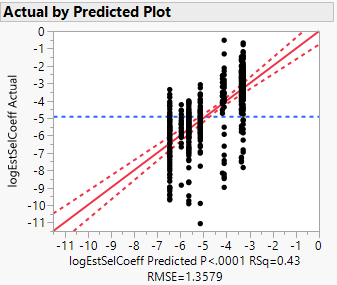
# Transformations on ExpMutRate and PairingProb

Doing transformations on expected mutation rate lost their significance when nucleotide identity or same AA group are introduced into the model.

# Regression for LogEstSelCoeff

Doing linear regression of the logEstSelCoeff results with R2=0.43 with as little as 2 categorical variables: nucleotide identity and same AA group:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| RSquare | | | | 0.429278 | |
| RSquare Adj | | | | 0.42512 | |
| Root Mean Square Error | | | | 1.357862 | |
| Mean of Response | | | | -4.89635 | |
| Observations (or Sum Wgts) | | | | 554 | |
| **Term** | |  | | | **Estimate** | | **Std Error** | | **t Ratio** | | **Prob>|t|** | |
| Intercept | |  | | | -4.763012 | | 0.068581 | | -69.45 | | <.0001\* | |
| WTnt[a] | |  | | | -1.281535 | | 0.09028 | | -14.20 | | <.0001\* | |
| WTnt[c] | |  | | | 0.9949523 | | 0.122275 | | 8.14 | | <.0001\* | |
| WTnt[g] | |  | | | 1.0701179 | | 0.105629 | | 10.13 | | <.0001\* | |
| isSameGroup[0] | |  | | | 0.4113995 | | 0.064447 | | 6.38 | | <.0001\* | |
| **Source** | **Nparm** | | **DF** | | **Sum of Squares** | | | **F Ratio** | | **Prob > F** | |  | |
| WTnt | 3 | | 3 | | 609.88035 | | | 110.2584 | | <.0001\* | |  | |
| isSameGroup | 1 | | 1 | | 75.13266 | | | 40.7490 | | <.0001\* | |  | |



Adding the mutation AA improves prediction to R2=0.471:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| RSquare | | | 0.471909 | |
| RSquare Adj | | | 0.450029 | |
| Root Mean Square Error | | | 1.328118 | |
| Mean of Response | | | -4.89635 | |
| Observations (or Sum Wgts) | | | 554 | |
| **Source** | **Nparm** | **DF** | | **Sum of Squares** | | **F Ratio** | **Prob > F** |  |
| WTnt | 3 | 3 | | 358.61737 | | 67.7698 | <.0001\* |  |
| isSameGroup | 1 | 1 | | 44.85561 | | 25.4298 | <.0001\* |  |
| MUTAA | 18 | 18 | | 75.61082 | | 2.3814 | 0.0012\* |  |

