**Table S3 Feature Selection**

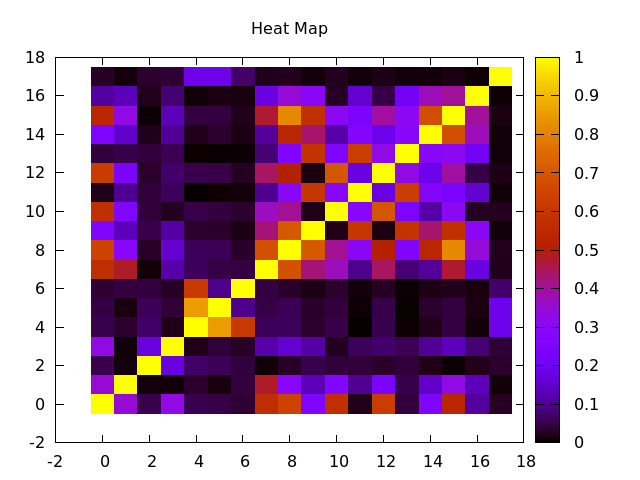
**1. Four different feature selection methods**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Feature/Method | STEP-Wise Logistic Regression  (p-value, smallest the better) | Lasso Logistic Regression | Randomized Logistic Regression  (Feature importance, the larger, the better) | Random Forest  (Feature importance, the larger, the better) |
| Folding Energy | <2e-16 | C | 1.0 | 0.241 |
| Seed match | 9.24e-12 | C | 1.0 | 0.076 |
| Accessibility | 1.30e-15 | C | 1.0 | 0.06 |
| AU content | <2e-16 | C | 1.0 | 0.053 |
| Stem conservation | 2.73e-14 | C | 0.47 | 0.057 |
| Flanking conservation | X | C | 0.765 | 0.057 |
| Difference between stem and flanking conservation | X | X | 0 | 0.028 |
| m/e motif | 3.19e-07 | C | 0.5 | 0.083 |
| The total number of paired positions | 7.13e-13 | X | 0.65 | 0.034 |
| The length of the target mRNA region | 1.14e-10 | C | 1 | 0.062 |
| The length of the largest consecutive pairs | <2e-16 | C | 1 | 0.094 |
| The position of the largest consecutive pairs relative to miRNA 5’ | 4.46e-15 | X | 0.815 | 0.011 |
| The length of the largest consecutive pairs allowing 2 mismatches | X | X | 0.015 | 0.015 |
| The position of the largest consecutive pairs allowing 2 mismatches | X | X | 0.04 | 0.028 |
| The number of paired positions at the miRNA 3’ end | 7.72e-09 | C | 0.965 | 0.024 |
| The total number of paired positions in the seed region and the miRNA 3’ end | 0.000602 | X | 0.005 | 0.019 |
| The difference between the number of paired positions in the seed region and that in the miRNA 3’ end | <2e-16 | C | 1 | 0.054 |
| Exon preference | X | X | 0 | 0.0005 |

**Note**: In the above table, C: Chosen,X: Not Chosen; **Criteria for feature selection: a)** STEP-Wise p-value <0.05 **b)** Lasso Chosen **c)**Randomized logistic feature importance/Random Forest feature importance > smallest of the conventional feature (Energy, Seed, Accessibility (including AU), Conservation). **d)** If any feature was selected by at least 2 methods, it will be chosen.

**2. Correlation analysis of listed 18 features**

The following is the Heatmap of pair-wise Pearson Correlation of the 18 features (id:0-17).



(1) Feature 6 -(Difference between STEM and Flanking conservation) is highly correlated to Feature 4 (STEM conservation) with correlation coefficient 0.61 .

(2) Feature 12- (The length of largest consecutive pairs allowing 2 mismatches) is highly correlated with feature 10 (The length of largest consecutive pairs) with correlation coefficient 0.70.

(3) Feature 13 (The position of largest consecutive pairs allowing 2 mismatches) is highly correlated with feature 11 (The position of largest consecutive pairs) with correlation coefficient 0.62.

(4) Feature 15 (The total number of paired positions in the seed region and the miRNA 3’ end) is highly correlated with feature 14 (Number of paired positions in the miRNA 3’ end) with correlation coefficient 0.67 and feature 16 (The difference between the number of paired positions in the seed region and that in the miRNA 3’ end) with correlation coefficient 0.40.

(5) Feature 17 (exon preference) is not correlated to any of the listed features.

The correlation analysis tells that there’s redundancy existed in these 18 features. However, we can’t tell directly what features should we choose based on the correlation analysis. For example, feature 0 (energy) is moderately correlated with seed (1) with correlation coefficient 0.35. Both energy and seed are traditional features and widely used by many existed tools (miRanda, TargetScan, etc.). Therefore, we utilized 4 different widely used feature selection methods. There are 13 out of 18 features were selected based on the 4 different feature selection methods.

The correlation analysis supports the feature selection methods. Features 4, 5, 6 are both conservation-related features and we also know that feature 6 is highly correlated with feature 4 and 5. Therefore, feature 6 could be redundant as it correlates to same type of features and this feature 6 was discarded by our feature selection analysis. We have similar results for the features 12,13,15. They are all highly correlated with the same type of features (10, 11, 14/16) and are discarded by the feature selection analysis. Feature 17 is not correlated (even not weakly correlated) with any of the listed features and this feature is not selected by any of the 4 feature selection methods. This implies that this feature might be not informative at all.