**HERVs with fantom enhancers**

**Algorithm: Random Forest (This one has more data points because predictions were**

**generated with cross\_val\_predict instead of train\_test\_split and then predict).**

**n\_estimators = 1000**

Dataset: */dors/capra\_lab/users/abraha1/projects/transposable\_elements/data/db\_HERV/merged\_cell\_unique\_TFBS/intersect\_HERV\_TFBS\_facet\_enhancers.tsv*

Features: Transcription factors. Each transcription factor had a column in the feature matrix;

each HERV was a single row, and the value was the number of times each transcription

factor appeared within the HERV.

Results:

*/dors/capra\_lab/users/yand1/te\_ml/results/2018\_06\_15/local*

|  |  |  |
| --- | --- | --- |
|  | predicted\_negative | predicted\_positive |
| actual\_negative | 252037 | 39 |
| actual\_positive | 1410 | 1 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 0 | 0.99 | 1.00 | 1.00 | 252076 |
| 1 | 0.03 | 0.00 | 0.00 | 1411 |

**Algorithm: Random Forest with random oversampling of training data.**

**n\_estimators = 1000**

Dataset: */dors/capra\_lab/users/abraha1/projects/transposable\_elements/data/db\_HERV/merged\_cell\_unique\_TFBS/intersect\_HERV\_TFBS\_facet\_enhancers.tsv*

Features: Transcription factors. Each transcription factor had a column in the feature matrix; each HERV was a single row, and the value was the number of times each transcription factor appeared within the HERV.

Results:

*/dors/capra\_lab/users/yand1/te\_ml/results/2018\_06\_19/local/rf\_oversampled*

|  |  |  |
| --- | --- | --- |
|  | predicted\_negative | predicted\_positive |
| actual\_negative | 60031 | 2994 |
| actual\_positive | 326 | 21 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 0 | 0.99 | 0.95 | 0.97 | 63025 |
| 1 | 0.01 | 0.06 | 0.01 | 347 |

**Algorithm: Random Forest with SMOTE oversampling of training data.**

**n\_estimators = 1000**

Dataset: */dors/capra\_lab/users/abraha1/projects/transposable\_elements/data/db\_HERV/merged\_cell\_unique\_TFBS/intersect\_HERV\_TFBS\_facet\_enhancers.tsv*

Features: Transcription factors. Each transcription factor had a column in the feature matrix; each HERV was a single row, and the value was the number of times each transcription factor appeared within the HERV.

Results:

*/dors/capra\_lab/users/yand1/te\_ml/results/2018\_06\_19/local/rf\_SMOTE*

|  |  |  |
| --- | --- | --- |
|  | predicted\_negative | predicted\_positive |
| actual\_negative | 62159 | 858 |
| actual\_positive | 346 | 9 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 0 | 0.99 | 0.99 | 0.99 | 63017 |
| 1 | 0.01 | 0.03 | 0.01 | 355 |

**Algorithm: Random Forest with ADASYN oversampling of training data.**

**n\_estimators = 1000**

Dataset: */dors/capra\_lab/users/abraha1/projects/transposable\_elements/data/db\_HERV/merged\_cell\_unique\_TFBS/intersect\_HERV\_TFBS\_facet\_enhancers.tsv*

Features: Transcription factors. Each transcription factor had a column in the feature matrix; each HERV was a single row, and the value was the number of times each transcription factor appeared within the HERV.

Results:

*/dors/capra\_lab/users/yand1/te\_ml/results/2018\_06\_19/local/rf\_ADASYN*

|  |  |  |
| --- | --- | --- |
|  | predicted\_negative | predicted\_positive |
| actual\_negative | 62993 | 17 |
| actual\_positive | 362 | 0 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 0 | 0.99 | 1.00 | 1.00 | 63010 |
| 1 | 0.00 | 0.00 | 0.00 | 362 |

**Algorithm: Random Forest with random undersampling of training data.**

**n\_estimators = 1000**

Dataset: */dors/capra\_lab/users/abraha1/projects/transposable\_elements/data/db\_HERV/merged\_cell\_unique\_TFBS/intersect\_HERV\_TFBS\_facet\_enhancers.tsv*

Features: Transcription factors. Each transcription factor had a column in the feature matrix; each HERV was a single row, and the value was the number of times each transcription factor appeared within the HERV.

Results:

*/dors/capra\_lab/users/yand1/te\_ml/results/2018\_06\_19/local/rf\_undersampled*

|  |  |  |
| --- | --- | --- |
|  | predicted\_negative | predicted\_positive |
| actual\_negative | 46367 | 16666 |
| actual\_positive | 94 | 245 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 0 | 1.00 | 0.74 | 0.85 | 63033 |
| 1 | 0.01 | 0.72 | 0.03 | 339 |

**Algorithm: Support Vector Classifier with rbf kernel and C = 10**

Dataset: */dors/capra\_lab/users/abraha1/projects/transposable\_elements/data/db\_HERV/merged\_cell\_unique\_TFBS/intersect\_HERV\_TFBS\_facet\_enhancers.tsv*

Features: Transcription factors. Each transcription factor had a column in the feature matrix; each HERV was a single row, and the value was the number of times each transcription factor appeared within the HERV.

Results:

*/dors/capra\_lab/users/yand1/te\_ml/results/2018\_06\_22\_svc\_rbf/accre*

|  |  |  |
| --- | --- | --- |
|  | predicted\_negative | predicted\_positive |
| actual\_negative | 62987 | 0 |
| actual\_positive | 385 | 0 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 0 | 0.99 | 1.00 | 1.00 | 62987 |
| 1 | 0.00 | 0.00 | 0.00 | 385 |

**Algorithm: Support Vector Classifier with rbf kernel and C = 1000**

Dataset: */dors/capra\_lab/users/abraha1/projects/transposable\_elements/data/db\_HERV/merged\_cell\_unique\_TFBS/intersect\_HERV\_TFBS\_facet\_enhancers.tsv*

Features: Transcription factors. Each transcription factor had a column in the feature matrix; each HERV was a single row, and the value was the number of times each transcription factor appeared within the HERV.

Results:

*/dors/capra\_lab/users/yand1/te\_ml/results/2018\_06\_21\_svm\_grid*

|  |  |  |
| --- | --- | --- |
|  | predicted\_negative | predicted\_positive |
| actual\_negative | 62965 | 56 |
| actual\_positive | 336 | 15 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 0 | 0.99 | 1.00 | 1.00 | 63021 |
| 1 | 0.21 | 0.04 | 0.07 | 351 |

**HERVs with chromhmm enhancers**

**Algorithm: Random Forest with n\_estimators = 100**

Dataset: */dors/capra\_lab/users/yand1/te\_ml/data/2018\_06\_29\_kmers\_enhancers\_intersect/reformatted\_hervs\_kmers\_enhancers\_intersect.tsv*

Features: 6-mers. Each 6-mer had a column with the number of times that 6-mer appeared within the HERV, with HERVs in rows. Column to predict contained 1 or 0 for if the HERV overlapped an enhancer.

Results:

*/dors/capra\_lab/users/yand1/te\_ml/results/2018\_07\_02\_kmers\_pca\_rf*

|  |  |  |
| --- | --- | --- |
|  | predicted\_negative | predicted\_positive |
| actual\_negative | 29707 | 6128 |
| actual\_positive | 20054 | 7950 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 0 | 0.60 | 0.83 | 0.69 | 35835 |
| 1 | 0.56 | 0.28 | 0.38 | 28004 |

**Algorithm: Gradient Boosting with decision trees**

Dataset: */dors/capra\_lab/users/yand1/te\_ml/data/2018\_06\_29\_kmers\_enhancers\_intersect/reformatted\_hervs\_kmers\_enhancers\_intersect.tsv*

Features: 6-mers. Each 6-mer had a column with the number of times that 6-mer appeared within the HERV, with HERVs in rows. Column to predict contained 1 or 0 for if the HERV overlapped an enhancer.

Results:

*/dors/capra\_lab/users/yand1/te\_ml/results/2018\_07\_02\_kmers\_pca\_rf*

|  |  |  |
| --- | --- | --- |
|  | predicted\_negative | predicted\_positive |
| actual\_negative | 33043 | 2544 |
| actual\_positive | 24093 | 4159 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 0 | 0.58 | 0.93 | 0.71 | 35587 |
| 1 | 0.62 | 0.15 | 0.24 | 28252 |

**MLT1Ks (subset of HERVs) with chromhmm enhancers**

**Algorithm: Random Forest with n\_estimators = 100**

Dataset: */dors/capra\_lab/users/yand1/te\_ml/data/2018\_07\_05\_subset\_hervs/ mlt1k\_kmers\_features\_matrix.tsv*

Features: 6-mers. Each 6-mer had a column with the number of times that 6-mer appeared within the MLT1K, with MLT1Ks in rows. Column to predict contained 1 or 0 for if the MLT1K overlapped an enhancer.

Results:

*/dors/capra\_lab/users/yand1/te\_ml/results/2018\_07\_05\_subset\_hervs*

|  |  |  |
| --- | --- | --- |
|  | predicted\_negative | predicted\_positive |
| actual\_negative | 715 | 288 |
| actual\_positive | 596 | 299 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 0 | 0.55 | 0.71 | 0.62 | 1003 |
| 1 | 0.51 | 0.33 | 0.40 | 895 |

**HERVs vs shuffled parts of human genome**

**Algorithm: Random Forest with n\_estimators = 100**

Dataset: */dors/capra\_lab/users/yand1/te\_ml/data/2018\_07\_06\_genome\_shuffle/all\_pairs\_features\_matrix.tsv*

Features: 6-mers. Each 6-mer had a column with the number of times that 6-mer appeared within the DNA segment, with DNA segments in rows. Column to predict contained 1 or 0 for if the segment of DNA was a HERV or not.

Results (positives are HERVs, negatives are shuffled parts of human genome):

*/dors/capra\_lab/users/yand1/te\_ml/results/2018\_07\_09\_shuffle\_ml*

|  |  |  |
| --- | --- | --- |
|  | predicted\_negative | predicted\_positive |
| actual\_negative | 45944 | 17309 |
| actual\_positive | 9881 | 53610 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 0 | 0.82 | 0.73 | 0.77 | 63253 |
| 1 | 0.76 | 0.84 | 0.84 | 63491 |

**All transposable elements with chromhmm enhancers**

**Algorithm: Random Forest with default parameters**

Dataset: */dors/capra\_lab/users/yand1/te\_ml/data/2018\_06\_25\_chromehmm\_ml\_input*

Features: Transcription factors. Each transcription factor had a column in the feature matrix; each transposable element was a single row, and the value was the number of times each transcription factor appeared within the transposable element.

Results:

*/dors/capra\_lab/users/yand1/te\_ml/results/2018\_07\_18\_tf\_chromhmm*

0 represents transposable element that does not overlap with enhancer; 1 represents overlap.

|  |  |  |
| --- | --- | --- |
|  | predicted\_0 | predicted\_1 |
| actual\_0 | 917424 | 39380 |
| actual\_1 | 248936 | 20343 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 0 | 0.79 | 0.96 | 0.86 | 956804 |
| 1 | 0.34 | 0.08 | 0.12 | 269279 |

**Algorithm: Random Forest with class weights {0: 1, 1: 10000000}**

Dataset: */dors/capra\_lab/users/yand1/te\_ml/data/2018\_06\_25\_chromehmm\_ml\_input*

Features: Transcription factors. Each transcription factor had a column in the feature matrix; each transposable element was a single row, and the value was the number of times each transcription factor appeared within the transposable element.

Results:

*/dors/capra\_lab/users/yand1/te\_ml/results/2018\_07\_19\_tf\_chromhmm*

0 represents transposable element that does not overlap with enhancer; 1 represents overlap.

|  |  |  |
| --- | --- | --- |
|  | predicted\_0 | predicted\_1 |
| actual\_0 | 805169 | 151194 |
| actual\_1 | 209616 | 60104 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 0 | 0.79 | 0.84 | 0.82 | 956363 |
| 1 | 0.28 | 0.22 | 0.25 | 269720 |

ROC AUC: 0.578

**Algorithm: Random Forest with class weights = “balanced”**

Dataset: */dors/capra\_lab/users/yand1/te\_ml/data/2018\_06\_25\_chromehmm\_ml\_input*

Features: Transcription factors. Each transcription factor had a column in the feature matrix; each transposable element was a single row, and the value was the number of times each transcription factor appeared within the transposable element.

Results:

*/dors/capra\_lab/users/yand1/te\_ml/results/2018\_07\_20\_tf\_chromhmm*

0 represents transposable element that does not overlap with enhancer; 1 represents overlap.

|  |  |  |
| --- | --- | --- |
|  | predicted\_0 | predicted\_1 |
| actual\_0 | 880173 | 77203 |
| actual\_1 | 235286 | 33421 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 0 | 0.79 | 0.92 | 0.85 | 957376 |
| 1 | 0.30 | 0.12 | 0.18 | 268707 |

ROC AUC: 0.573

**HERVs, chromhmm enhancers, HERV-enhancer intersect, and random set length matched to HERVs**

**Algorithm: Random Forest with max\_depth = 25**

Dataset: */dors/capra\_lab/users/yand1/te\_ml/data/2018\_07\_27\_sampled\_hervs\_enhancers\_random\_rf/* *sampled\_hervs\_enhancers\_shuffled\_features\_matrix.tsv*

(stratified sample of 10,000 instances)

Features: 6-mers. Each 6-mer had a column with the number of times that 6-mer appeared within the DNA segment, with DNA segments in rows. Column to predict label for HERV, enhancer, HERV-enhancer intersect, or random set.

Results: */dors/capra\_lab/users/yand1/te\_ml/results/2018\_07\_27\_sampled\_hervs\_enhancers\_random\_rf/rf\_10000*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | predicted enhancer only | predicted herv enhancer intersect | predicted herv only | predicted shuffled genome |
| actual enhancer only | 707 | 17 | 34 | 199 |
| actual herv enhancer intersect | 105 | 36 | 65 | 170 |
| actual herv only | 125 | 29 | 91 | 176 |
| actual shuffled genome | 215 | 26 | 53 | 452 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1 | support |
| enhancer | 0.61 | 0.74 | 0.67 | 957 |
| herv-enhancer | 0.33 | 0.1 | 0.15 | 376 |
| herv | 0.37 | 0.22 | 0.27 | 421 |
| random | 0.45 | 0.61 | 0.52 | 746 |
| average | 0.48 | 0.51 | 0.48 | 2500 |

**Algorithm: Random Forest with max\_depth = 25**

Dataset: */dors/capra\_lab/users/yand1/te\_ml/data/2018\_07\_27\_sampled\_hervs\_enhancers\_random\_rf/* *sampled\_100000\_hervs\_enhancers\_shuffled\_features\_matrix.tsv*

(stratified sample of 100,000 instances)

Features: 6-mers. Each 6-mer had a column with the number of times that 6-mer appeared within the DNA segment, with DNA segments in rows. Column to predict label for HERV, enhancer, HERV-enhancer intersect, or random set.

Results: */dors/capra\_lab/users/yand1/te\_ml/results/2018\_07\_27\_sampled\_hervs\_enhancers\_random\_rf/rf\_100000*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | predicted enhancer only | predicted herv enhancer intersect | predicted herv only | predicted shuffled genome |
| actual enhancer only | 7915 | 104 | 211 | 1404 |
| actual herv enhancer intersect | 751 | 383 | 660 | 1573 |
| actual herv only | 1012 | 390 | 1079 | 1868 |
| actual shuffled genome | 1732 | 293 | 524 | 5101 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1 | support |
| enhancer | 0.69 | 0.82 | 0.75 | 9634 |
| herv-enhancer | 0.44 | 0.25 | 0.32 | 3367 |
| herv | 0.44 | 0.25 | 0.32 | 4349 |
| random | 0.51 | 0.67 | 0.58 | 7650 |
| average | 0.54 | 0.58 | 0.55 | 25000 |