**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:

predicted\_negative predicted\_positive

actual\_negative 252037 39

actual\_positive 1410 1

precision recall f1-score

0.03 0.00 0.00

**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:

predicted\_negative predicted\_positive

actual\_negative 60031 2994

actual\_positive 326 21

precision recall f1-score

0.01 0.06 0.01

**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:

predicted\_negative predicted\_positive

actual\_negative 62159 858

actual\_positive 346 9

precision recall f1-score

0.01 0.03 0.01

**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:

predicted\_negative predicted\_positive

actual\_negative 62993 17

actual\_positive 362 0

precision recall f1-score

0.01 0.03 0.01

**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:

predicted\_negative predicted\_positive

actual\_negative 46367 16666

actual\_positive 94 245

precision recall f1-score

0.01 0.72 0.03

**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:

predicted\_negative predicted\_positive

actual\_negative 62987 0

actual\_positive 385 0

precision recall f1-score

0.00 0.00 0.00

**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:

predicted\_negative predicted\_positive

actual\_negative 62965 56

actual\_positive 336 15

precision recall f1-score

0.21 0.04 0.07

**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:

predicted\_negative predicted\_positive

actual\_negative 29707 6128

actual\_positive 20054 7950

precision recall f1-score

0.56 0.28 0.38

**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:

predicted\_negative predicted\_positive

actual\_negative 33043 2544

actual\_positive 24093 4159

precision recall f1-score

0.62 0.15 0.24

**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:

predicted\_negative predicted\_positive

actual\_negative 715 288

actual\_positive 596 299

precision recall f1-score

0.51 0.33 0.40

**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):

predicted\_negative predicted\_positive

actual\_negative 45944 17309

actual\_positive 9881 53610

precision recall f1-score

0.76 0.84 0.80