Data Collection

COSMIC Catalogue Of Somatic Mutations In Cancer

1000 Genomes Project Defining Genetic Variation in People

GENCODE

Feature Design

Multi-Alignment

Variant Effect Predictor

VeP

Intrinsic Composition

Feature Extraction

Conservation

- PhastCons Score
- PhyloP Score

VEP Annotation

- Number of Affected Sequences
- Consequences Type
- Affected Amino Acids

Genomic Context

- Distance to Gene Feature
- *K*-mer Counts
- G + C Content

Classifier Construction

Model Selection

- Random Forest
- Logistic Regression
- SVM
- GBM

Feature Selection

- Multi-Alignment
- VEP
- Genomic Context

Evaluation

Strategy

- 10-Fold CV
- Metrics
- ROC Curve

Benchmark

- CScape
- CADD