

Milestone 2: Core results 01

Anders

25/1/2023

Data

One clinical trials on breast cancer (advanced HR+/HER2- and HER2-E breast cancer) using two different drug combination; and a cohorts study (here used as test data set).

Both data set have mRNA expression of 771 genes at baseline (prior to treatment). This genes are specifically selected based on their potential roles in breast cancer pathology:

The gene set is divided into 25 sets of “signature genes”; which are thought to represent functional unities with respect to cancer biology. Often signaling pathways. Furthermore, 8 immune cells are represented with specific genes. These sets are substantially smaller than the signature genes; which I presume leads to some issue in modeling (as for clinical data too - see next sentence).

Additionally, both data-set contains clinical data; which up to now is not used in any models. If included they maybe should have a higher weight or be implemented differently from a sole gene. Maybe in a stacked ensemble model as signature.

Responses used

Proliferation score

A score based on expression level of some of the genes. Range: -1.1366 to 0.8511 ### Risk of relapse score (ROR) A score based on expression level of some of the genes. Range: -8.035678 to 75.13174 ### Risk of relapse score with proliferation score (ROR-Prolif) A combined score based on expression level of more genes. Range: 1 to 97 (1-100).

The two scores involving ROR also have categorical variants containing: low, medium, high

Progression free survival (PFS)

This is the outcome used in the clinical cohort. Here i have used correlation with the scores described above. Spearman can maybe be used. Another approach is to use the above scores to divide the patient in to two groups and see if the two groups show clearly separable PFS over time (basically look at the graphs). The differentiation into two groups is done base on best values from a ROC curve.

Trail

Two treatments which differ with respect to drug combination - Target: ribociclib and endocrine therapy (letrozole) - Chemotherapy: doxorubicin, cyclophosphamide and paclitaxel. approx. 50 patients in each group. Endpoints: proliferation score, ROR score, combined ROR and prolifer

Cohort

The primary objective of this study is to compare two cdk4/6 targeted drugs (Palbociclib, n=36; Abemaciclib, n=3 in combination with endocrine therapy (tamoxifen, fulvestrant or aromatase inhibitors, I think?)

Endpoints: progression free survival (months), OS?, and status of the two former (dont know what that means)

Major goal

1. Find best model to predict outcome of cancer treatment with genetic profile as predictive features
2. Features selection in order to understand cancer biology

Major challenges

Preliminary experiments (on trail 1) showed instability in prediction and feature selection between bootstrap samples of Lasso. I believe this is a classical problem of high-dim data?

Approch

Test all thinkable models in a search for superior models

Evaluation of models

Two levels of evaluation is considered:

1. Relative comparison of the different models

1000 bootstrap models are fitted and then evaluated on the original sample. This gives a relative comparison of the various models with respect to data very similar to the given data set. In addition to Correlation and MSE, frequency of selected features is compared.

2. Expected outcome of future patients

3 strategies are considered:

1. Repeated cross-validations (200 rep, 5-fold)
2. Bootstrap models with 0.632 (or 0.632?) adjustment (Not done)
3. Use the cohort as test data-set (Challenge: This trail have different responses)

RESULTS

Results of individual modles:

Models tested

Lasso

Post Lasso

Residuals

Ridge

Elastic Net

Boosting with stumps as base learner

- mboost

- xgboost

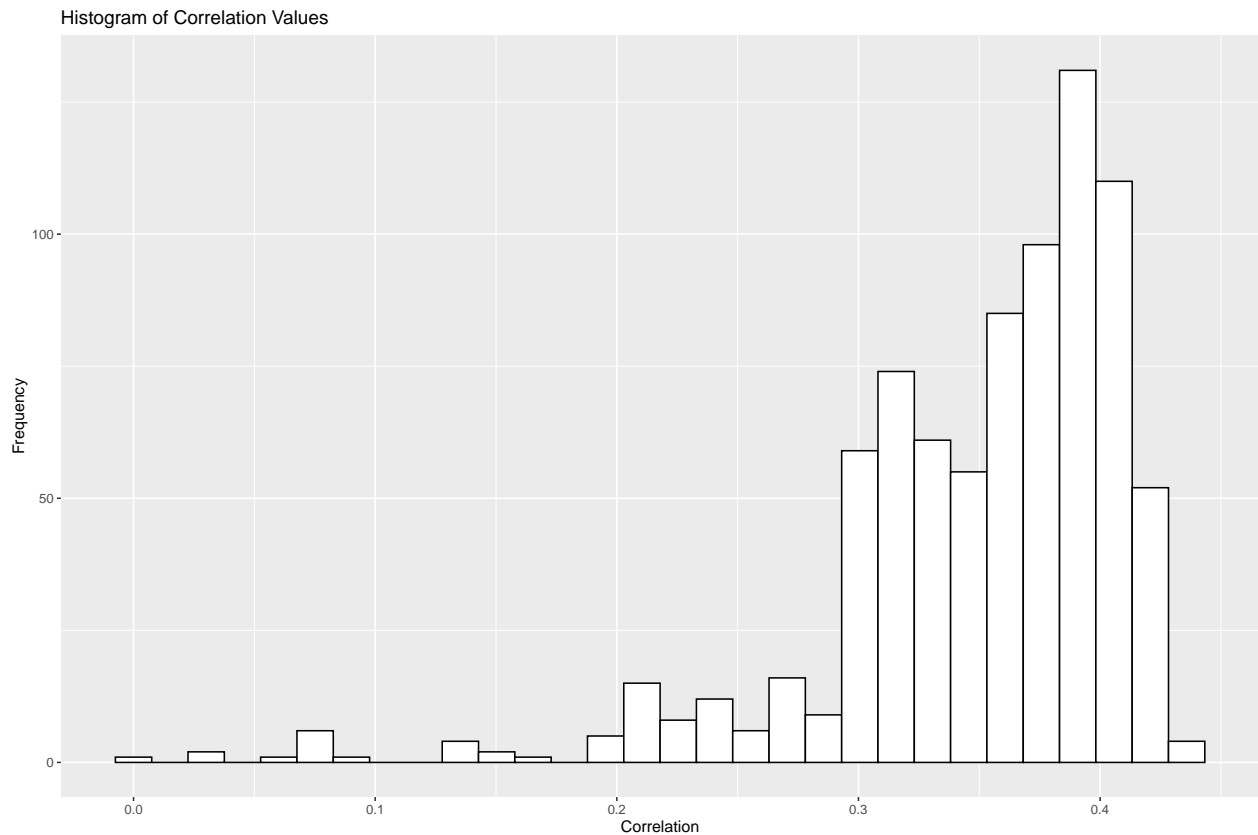
Stacking using different features in the base models

Synergistic learning

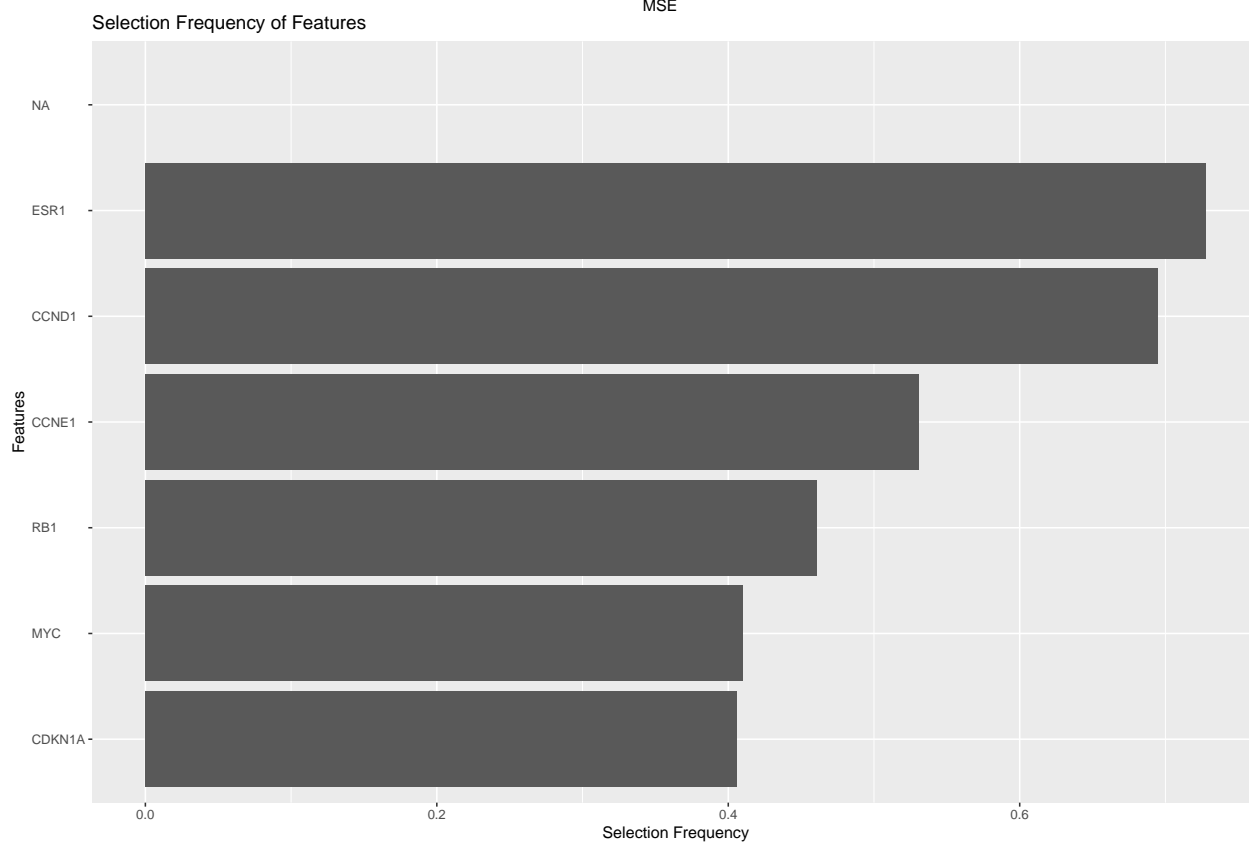
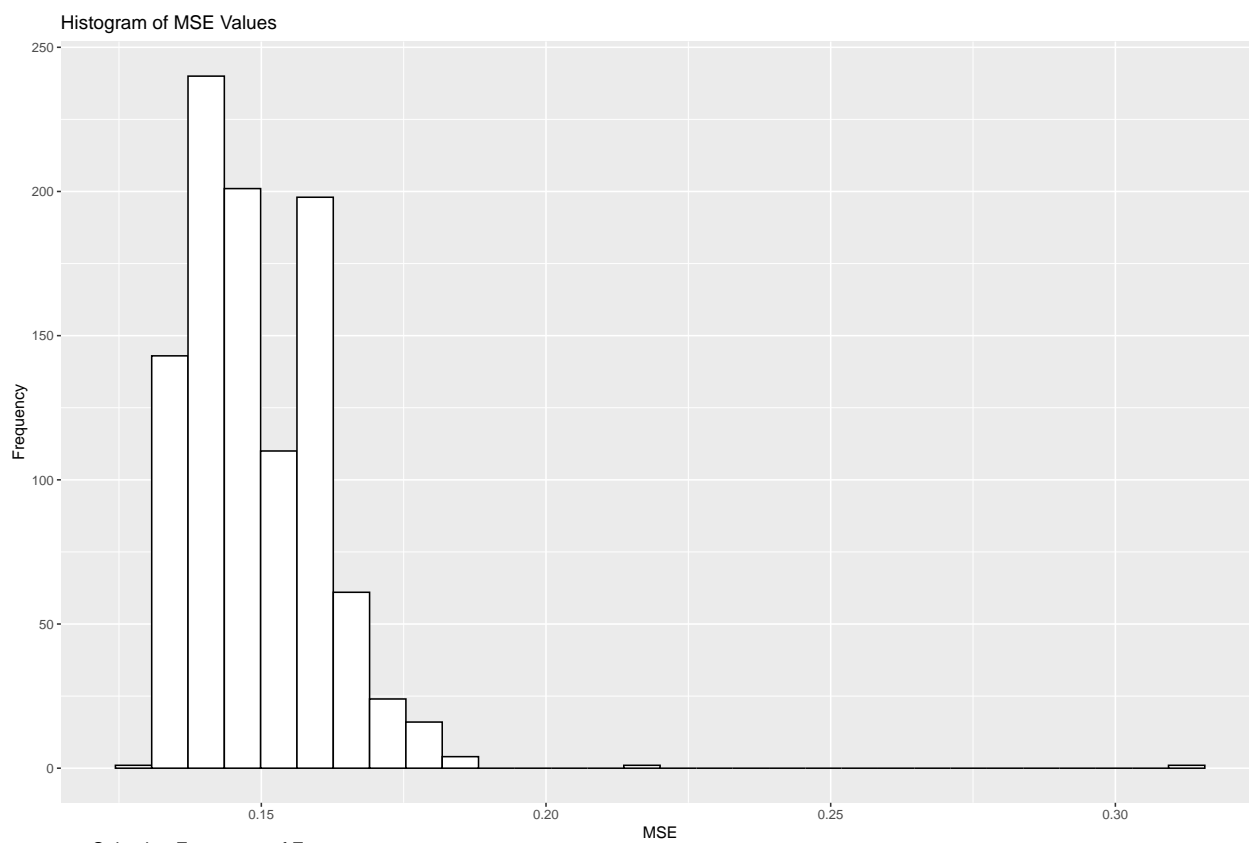
Lasso - Bootstrap

6 genes -> proliferation score

```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0.182
##
## CORRELATIONS RESULTS
## Mean: 0.3498379
## Median: 0.3672243
## Variance: 0.003984261
## st.dev.: 0.063121
```



```
## MSE RESULTS
## Mean: 0.1492302
## Median: 0.1469228
## Variance: 0.0001550247
## st.dev.: 0.01245089
```

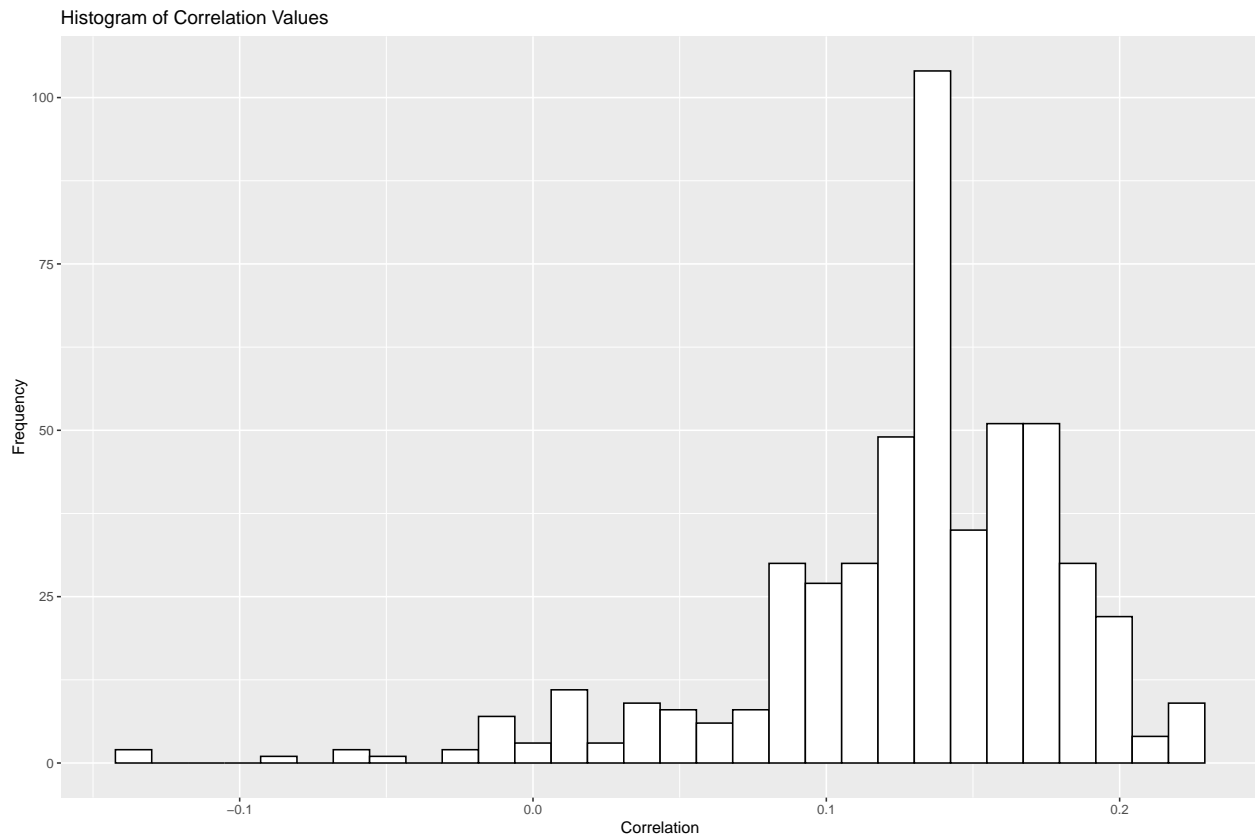


##

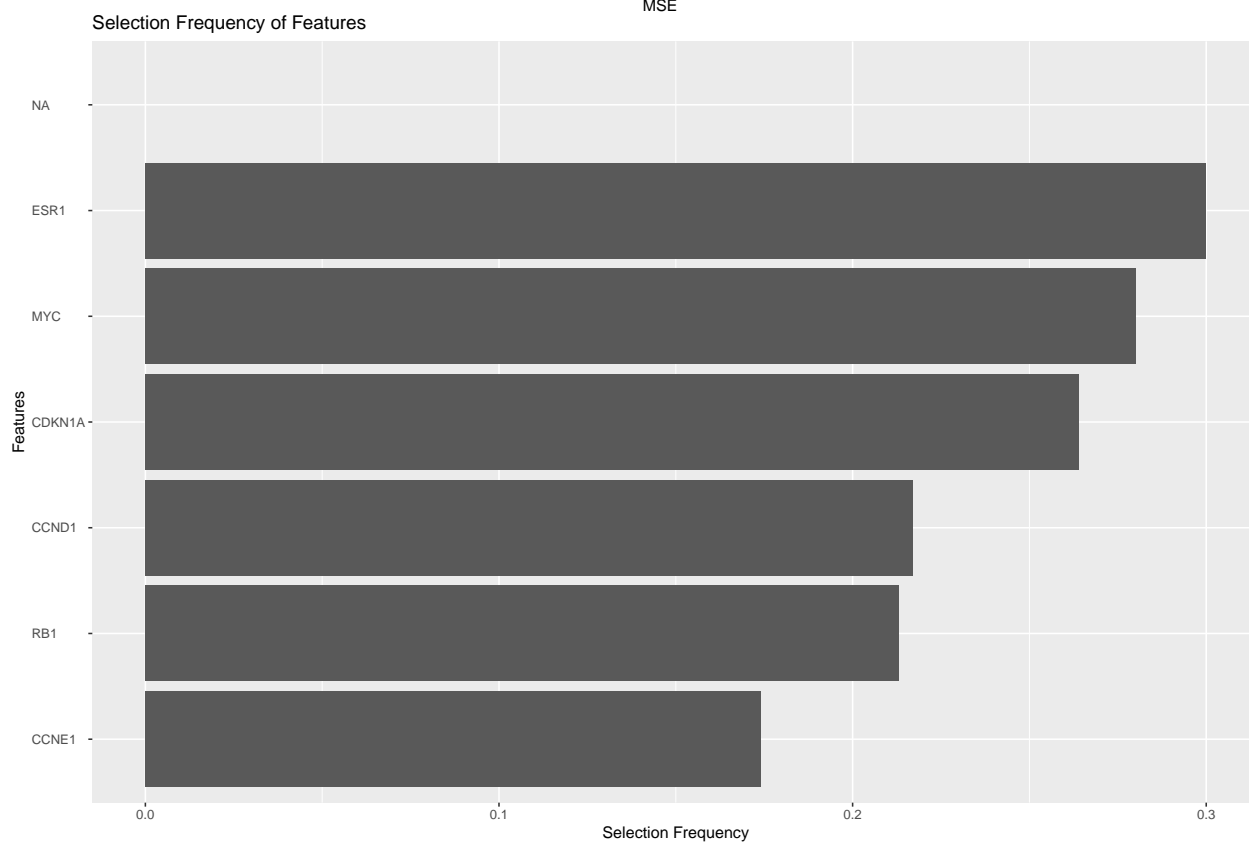
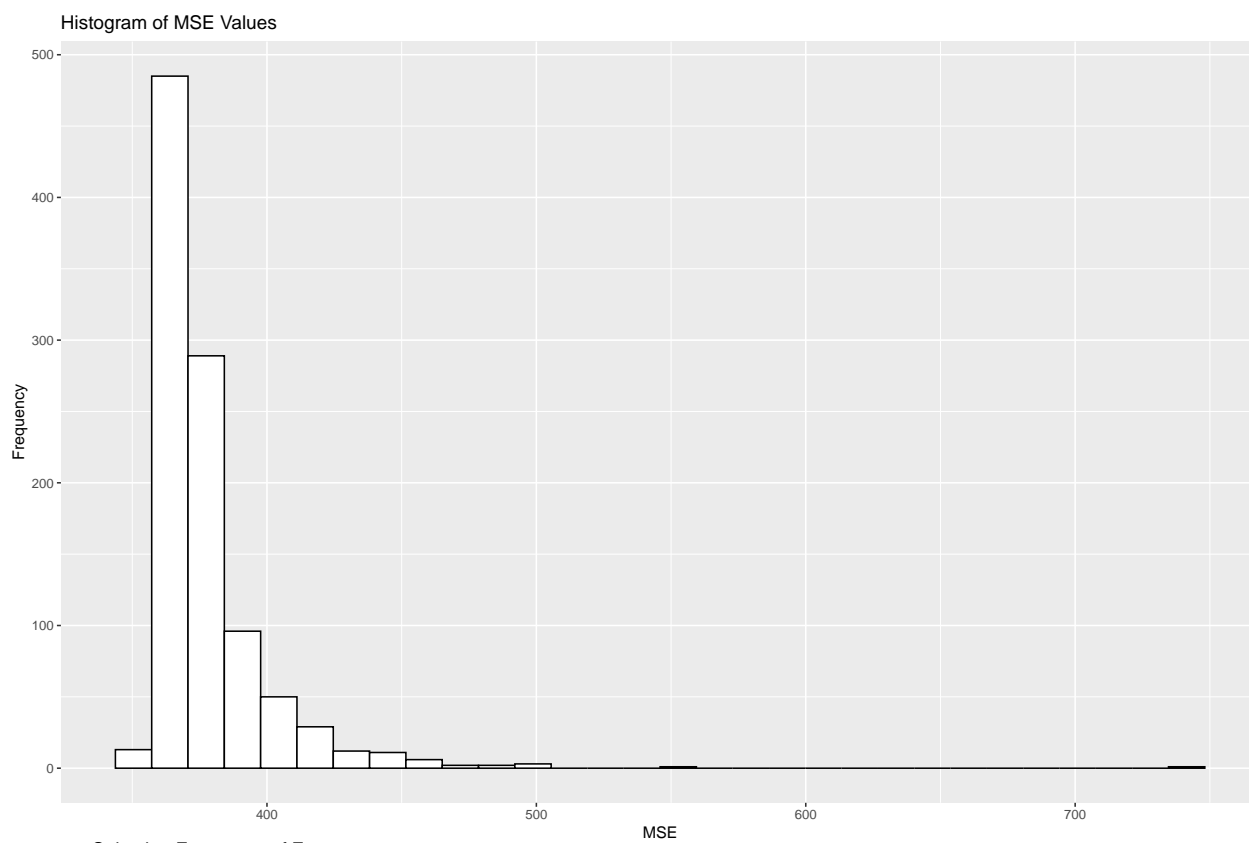
```
## Features selected 50% or more times:
## CCND1 CCNE1 ESR1
## Top 20 featruess:
## [1] "ESR1" "CCND1" "CCNE1" "RB1" "MYC" "CDKN1A" NA NA
## [9] NA NA NA NA NA NA NA NA
## [17] NA NA NA NA
```

6 genes -> ROR_proliferation score

```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0.495
##
## CORRELATIONS RESULTS
## Mean: 0.1282306
## Median: 0.1311715
## Variance: 0.00285293
## st.dev.: 0.05341282
```



```
## MSE RESULTS
## Mean: 378.094
## Median: 370.7201
## Variance: 549.4448
## st.dev.: 23.44024
```



##

```
## Features selected 50% or more times:
```

```
##
```

```
## Top 20 featrues:
```

```
## [1] "ESR1"  "MYC"   "CDKN1A" "CCND1" "RB1"   "CCNE1" NA      NA
## [9] NA      NA      NA      NA      NA      NA      NA      NA
## [17] NA      NA      NA      NA
```

```
771 genes -> proliferation score
```

```
## number of models fitted: 1000
```

```
## Fraction of model fits with no selected genes: 0.002
```

```
##
```

```
## CORRELATIONS RESULTS
```

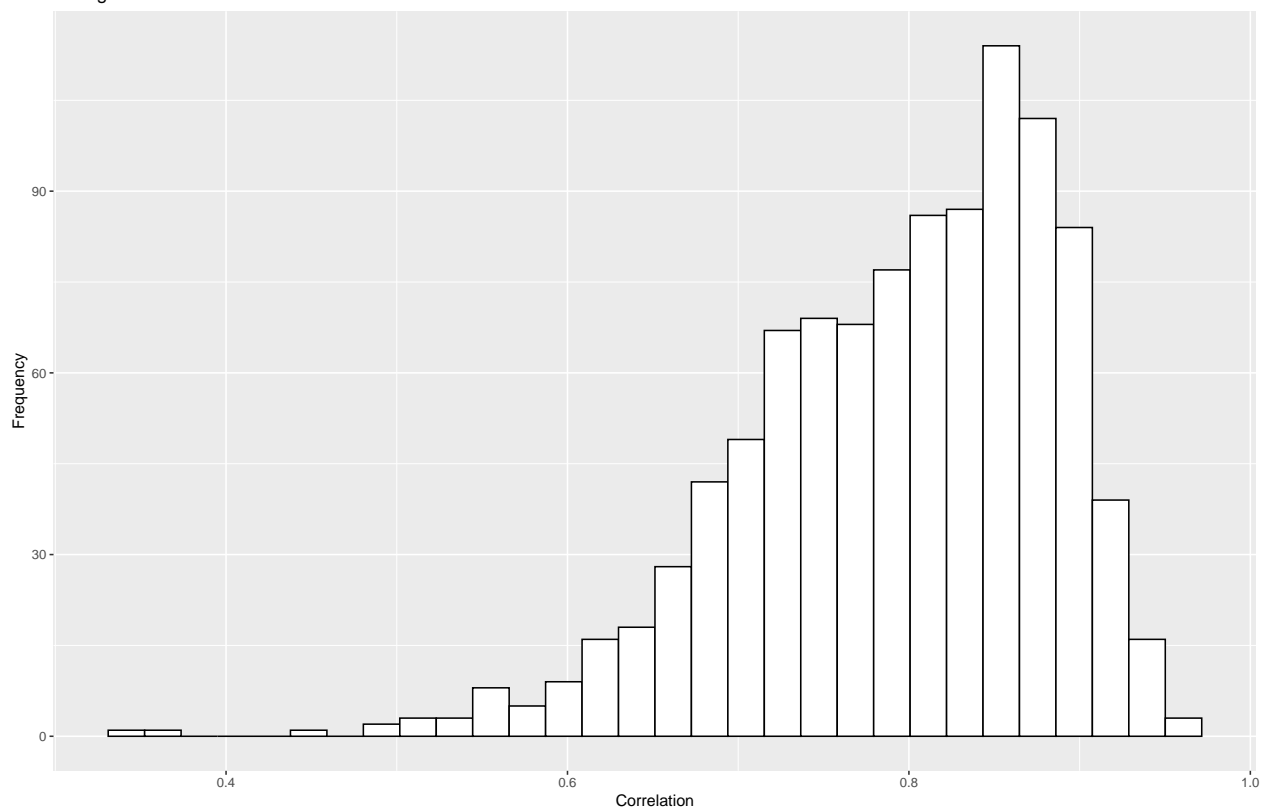
```
## Mean: 0.7941413
```

```
## Median: 0.8101886
```

```
## Variance: 0.008119272
```

```
## st.dev.: 0.090107
```

Histogram of Correlation Values



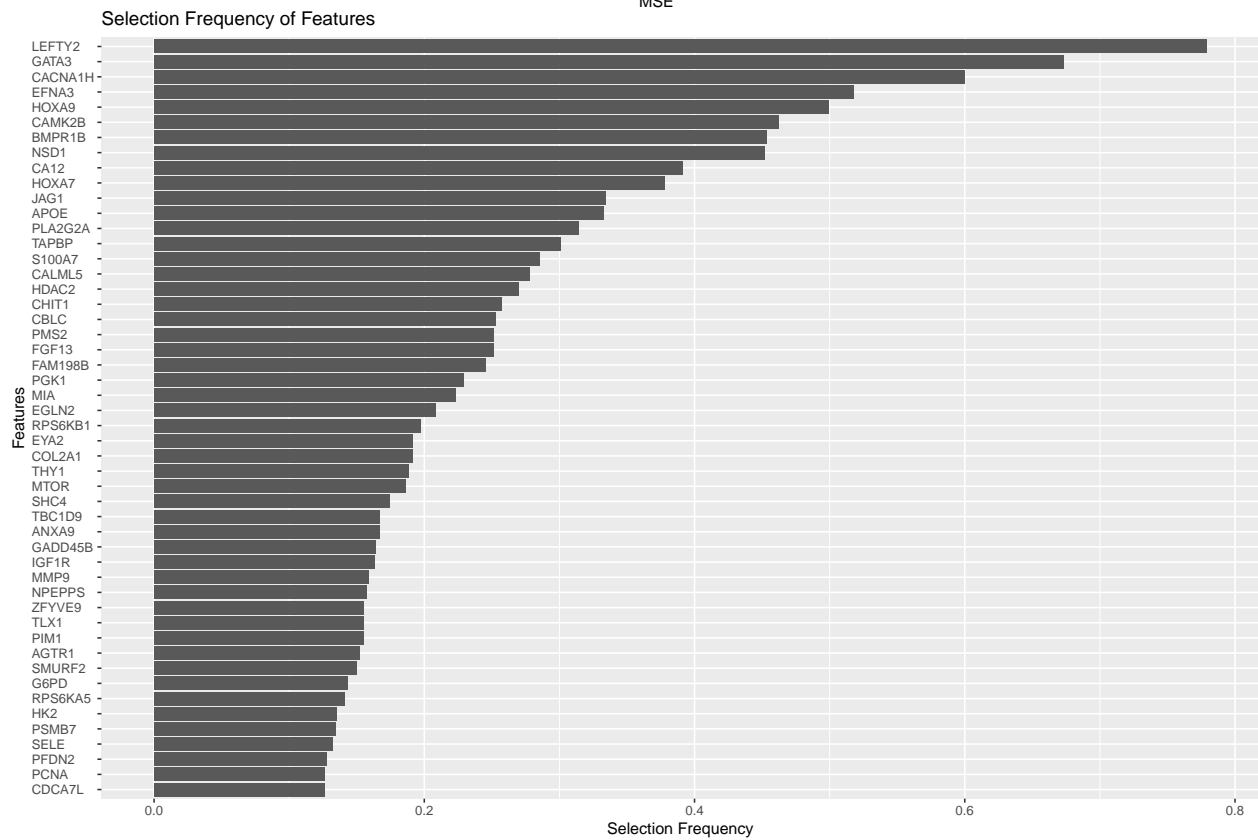
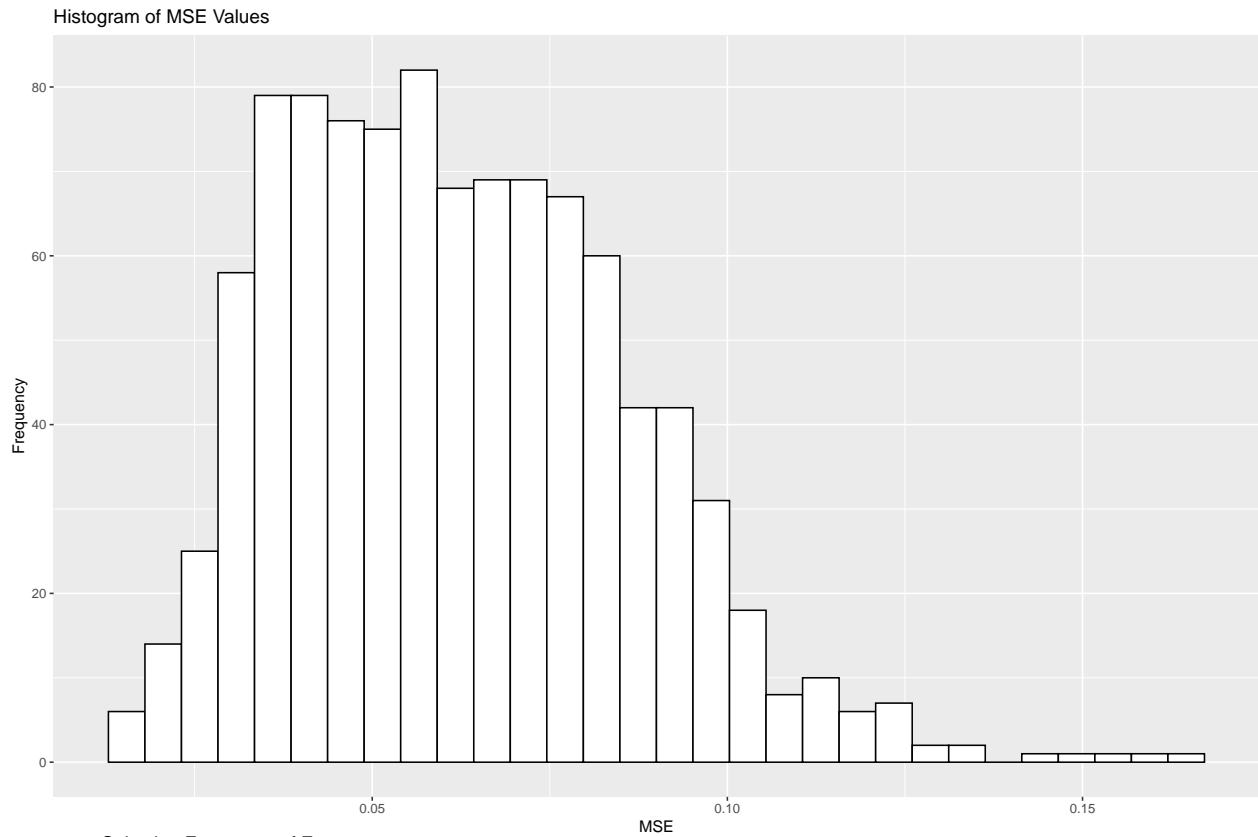
```
## MSE RESULTS
```

```
## Mean: 0.06209131
```

```
## Median: 0.0598495
```

```
## Variance: 0.0005751012
```

```
## st.dev.: 0.02398127
```

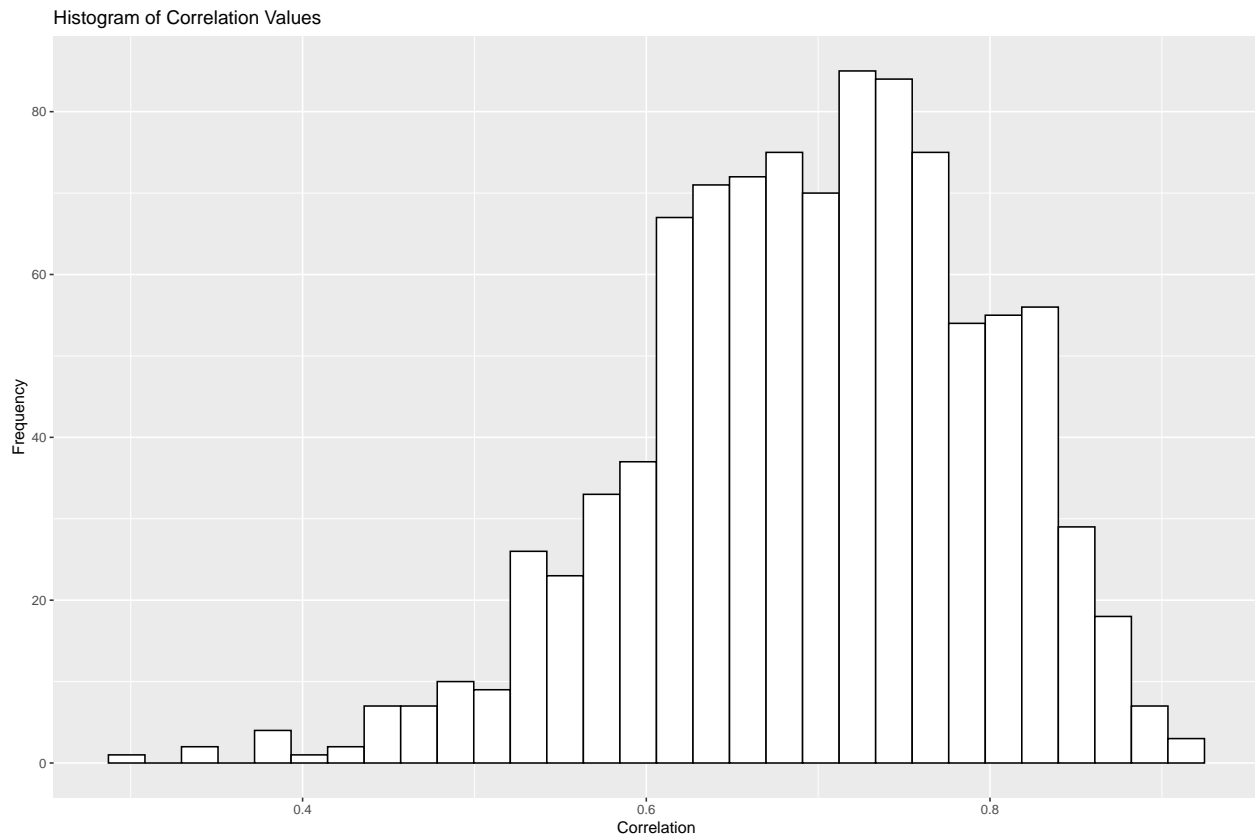


##

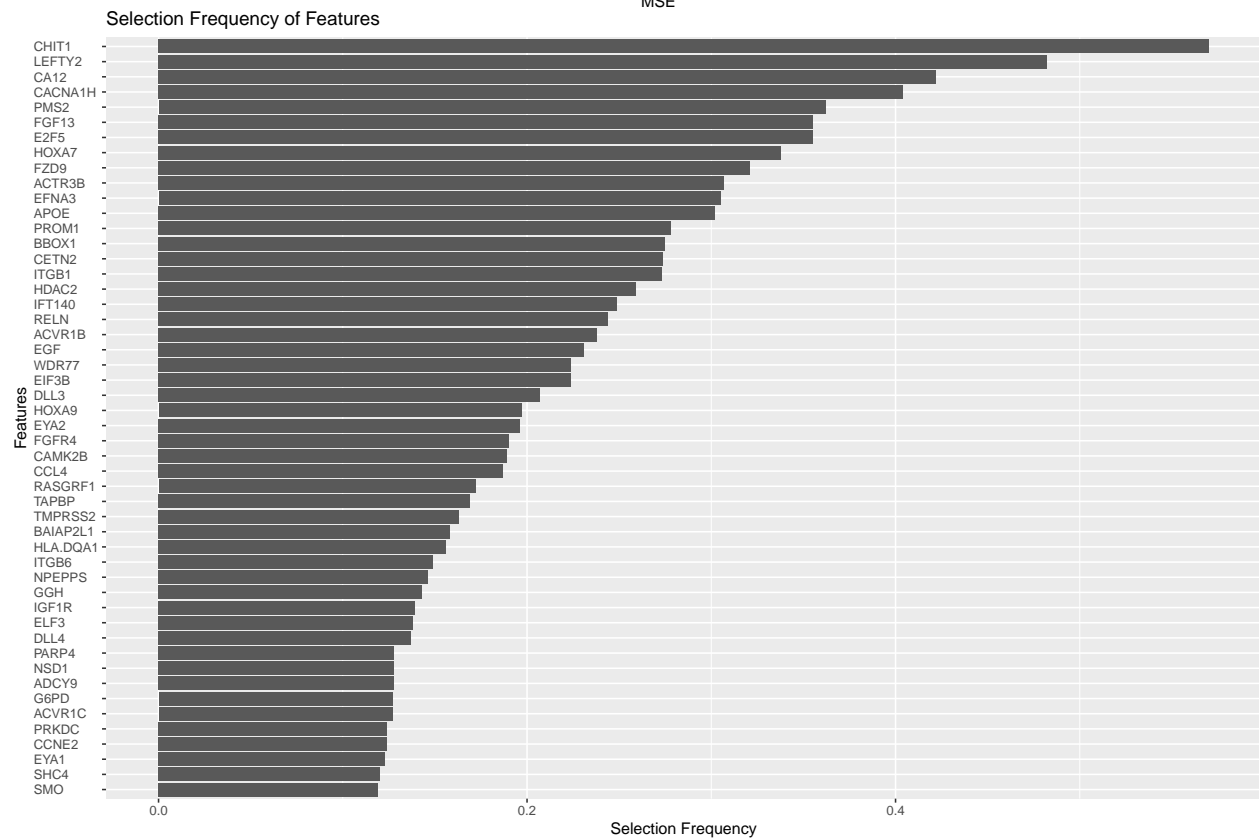
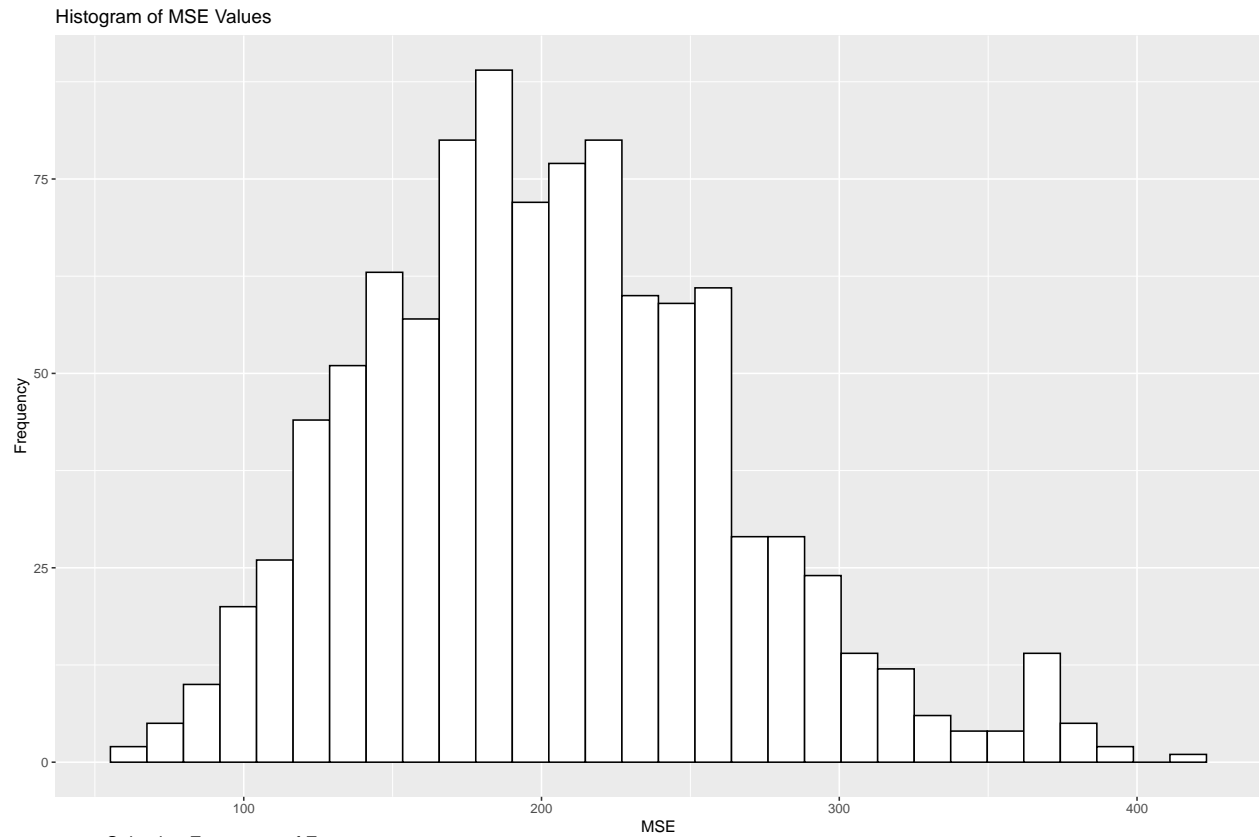

```
## Features selected 50% or more times:
## CACNA1H EFNA3 GATA3 LEFTY2
## Top 20 featrues:
## [1] "LEFTY2" "GATA3" "CACNA1H" "EFNA3" "HOXA9" "CAMK2B" "BMPR1B"
## [8] "NSD1" "CA12" "HOXA7" "JAG1" "APOE" "PLA2G2A" "TAPBP"
## [15] "S100A7" "CALML5" "HDAC2" "CHIT1" "CBLC" "FGF13"
```

771 genes -> ROR-proliferation score

```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0.017
##
## CORRELATIONS RESULTS
## Mean: 0.6968101
## Median: 0.7035889
## Variance: 0.009901439
## st.dev.: 0.09950598
```



```
## MSE RESULTS
## Mean: 203.408
## Median: 198.455
## Variance: 3763.666
## st.dev.: 61.34872
```



##

```
## Features selected 50% or more times:
```

```
## CHIT1
```

```
## Top 20 featrues:
```

```
## [1] "CHIT1"    "LEFTY2"   "CA12"     "CACNA1H"  "PMS2"     "E2F5"     "FGF13"
## [8] "HOXA7"    "FZD9"     "ACTR3B"   "EFNA3"    "APOE"     "PROM1"    "BBOX1"
## [15] "CETN2"    "ITGB1"    "HDAC2"    "IFT140"   "RELN"     "ACVR1B"
```

```
node values -> proliferation score
```

```
## number of models fitted: 1000
```

```
## Fraction of model fits with no selected genes: 0.053
```

```
##
```

```
## CORRELATIONS RESULTS
```

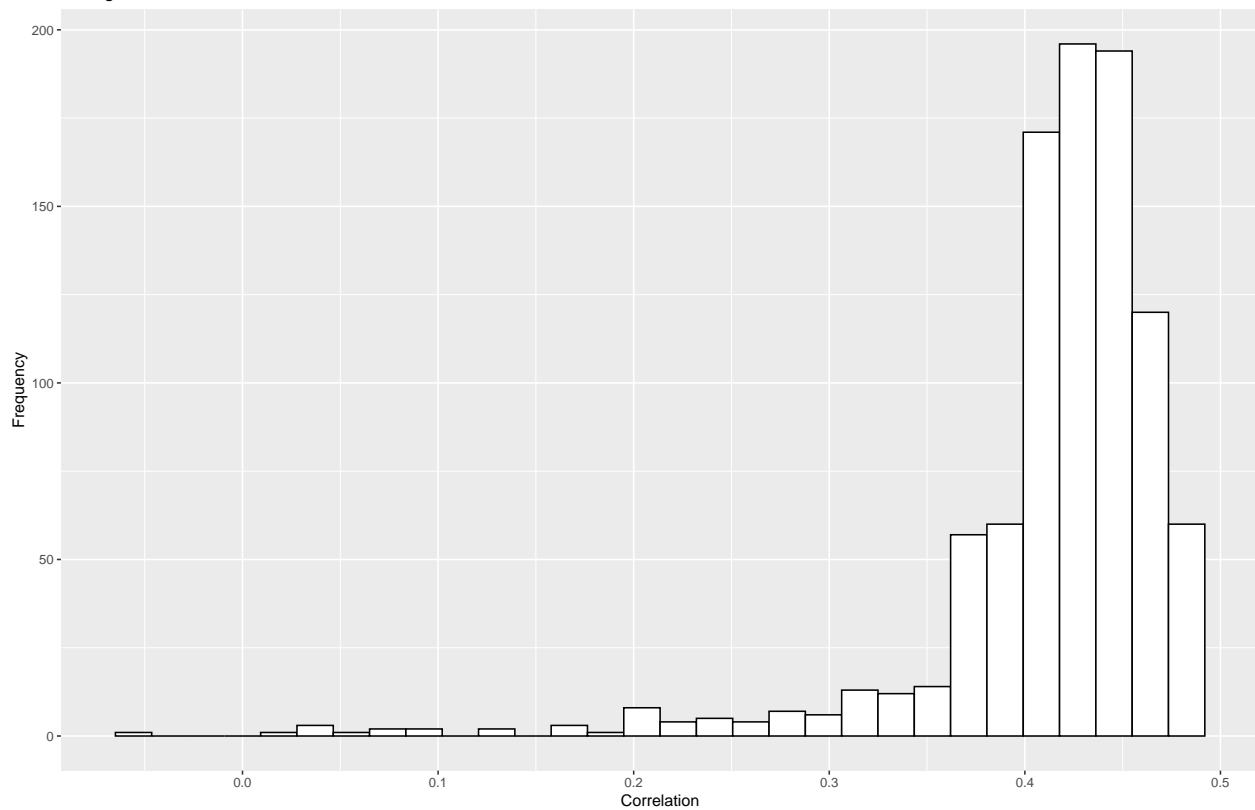
```
## Mean: 0.414496
```

```
## Median: 0.4275114
```

```
## Variance: 0.00413092
```

```
## st.dev.: 0.06427223
```

```
Histogram of Correlation Values
```



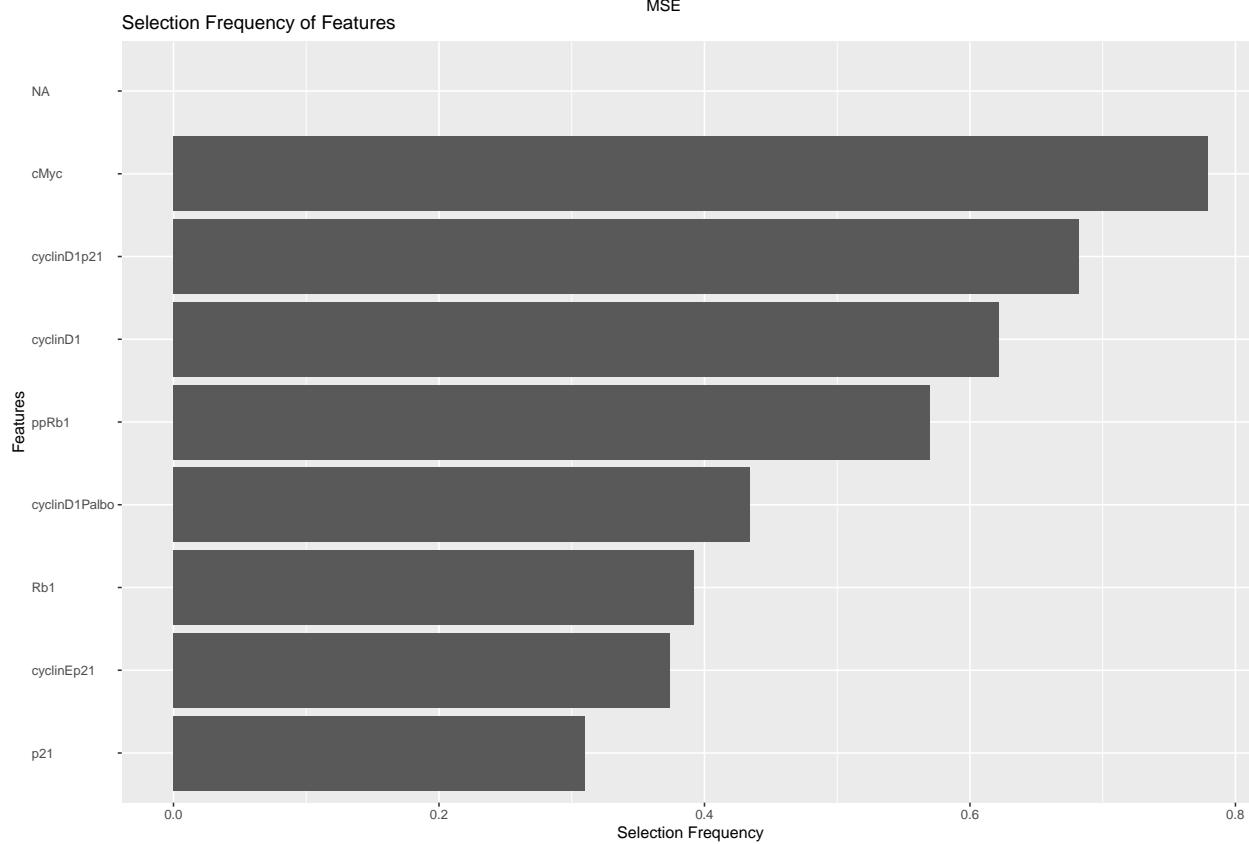
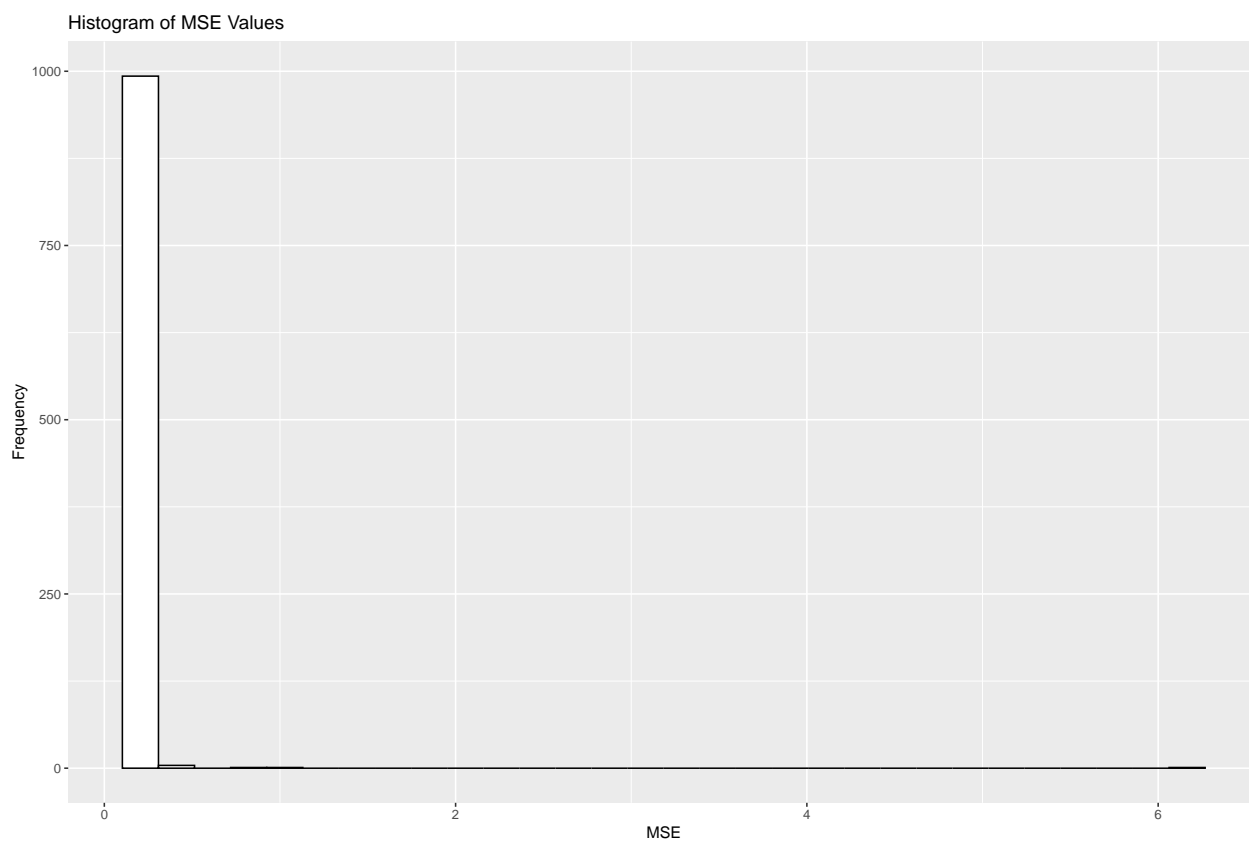
```
## MSE RESULTS
```

```
## Mean: 0.1479731
```

```
## Median: 0.1355805
```

```
## Variance: 0.03673469
```

```
## st.dev.: 0.191663
```

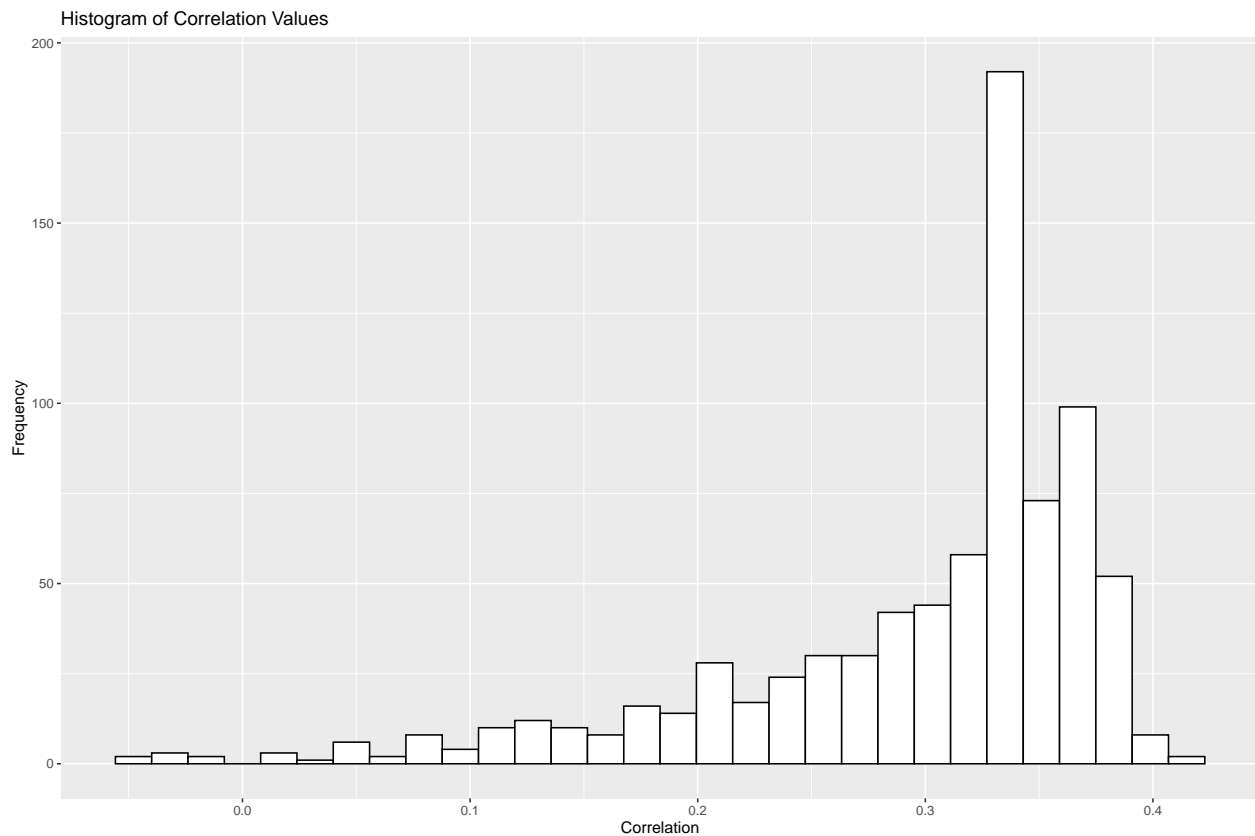


##

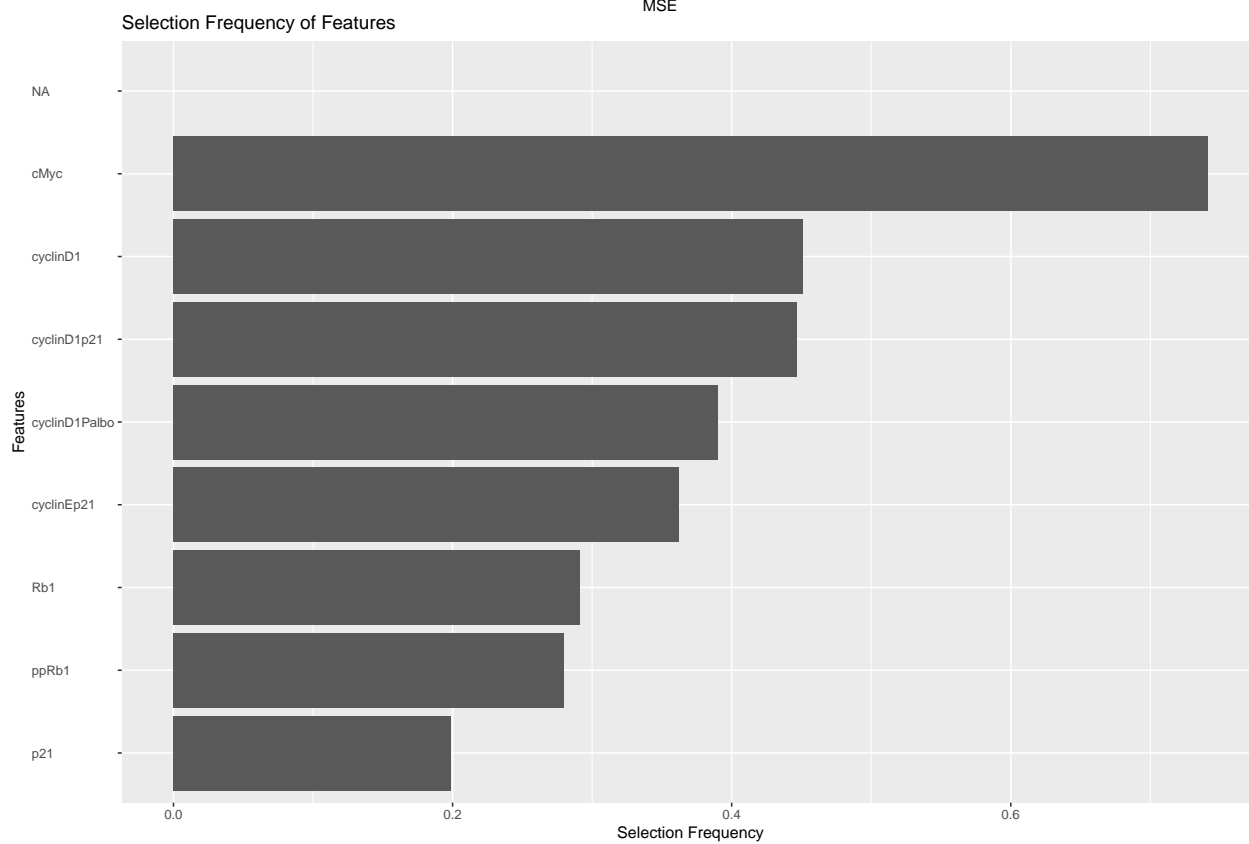
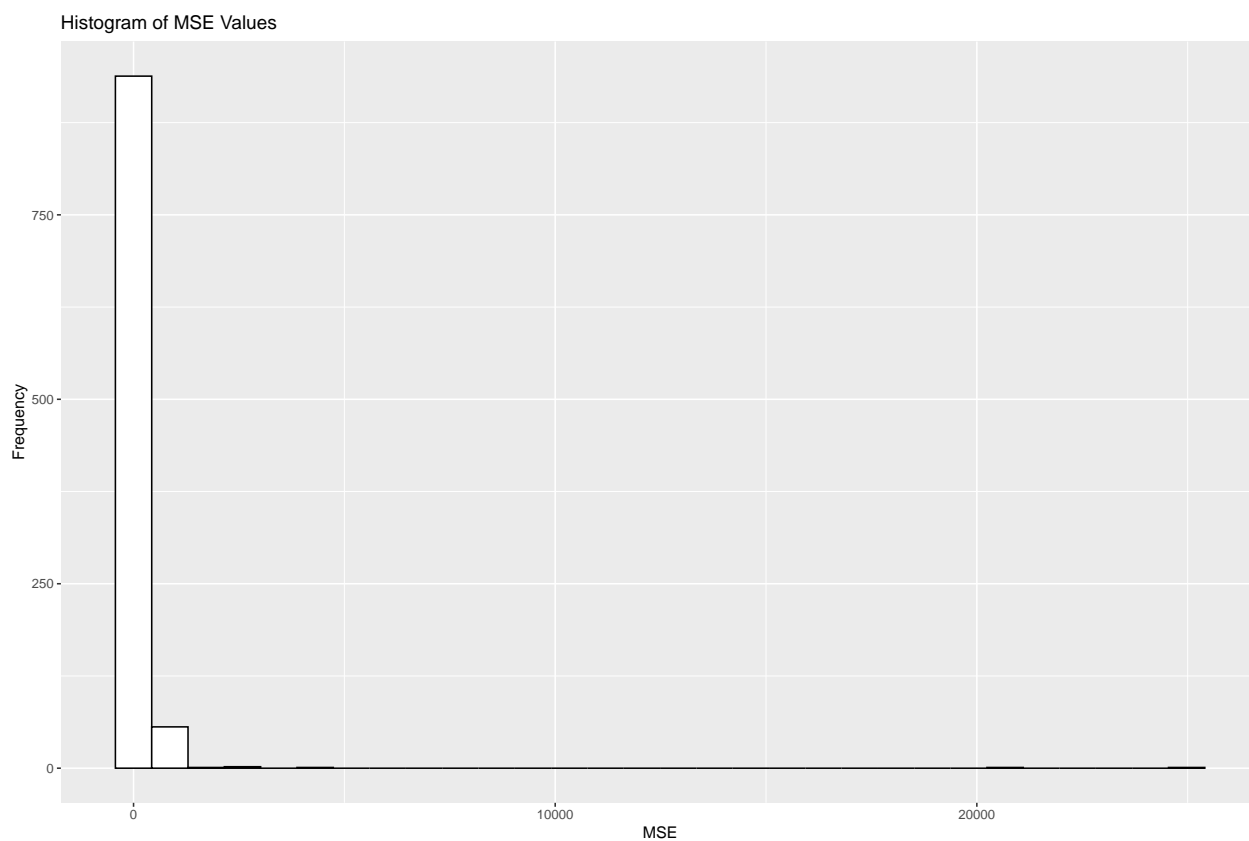
```
## Features selected 50% or more times:
## cyclinD1 cyclinD1p21 cMyc ppRb1
## Top 20 featrues:
## [1] "cMyc"          "cyclinD1p21"  "cyclinD1"     "ppRb1"
## [5] "cyclinD1Palbo" "Rb1"          "cyclinEp21"   "p21"
## [9] NA              NA              NA              NA
## [13] NA              NA              NA              NA
## [17] NA              NA              NA              NA
```

node values -> ROR-proliferation score

```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0.2
##
## CORRELATIONS RESULTS
## Mean: 0.2964169
## Median: 0.3317433
## Variance: 0.006900552
## st.dev.: 0.08306956
```



```
## MSE RESULTS
## Mean: 417.6667
## Median: 353.8176
## Variance: 1054857
## st.dev.: 1027.062
```

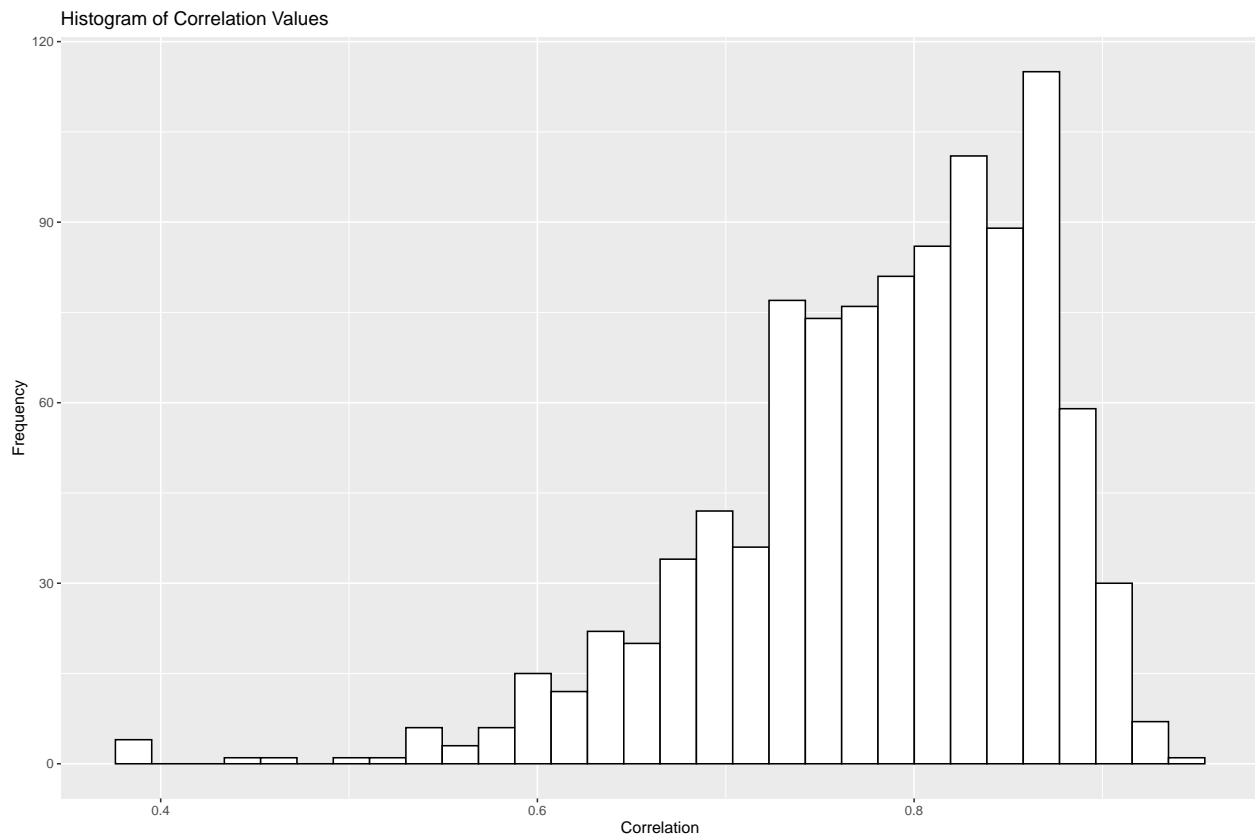


##

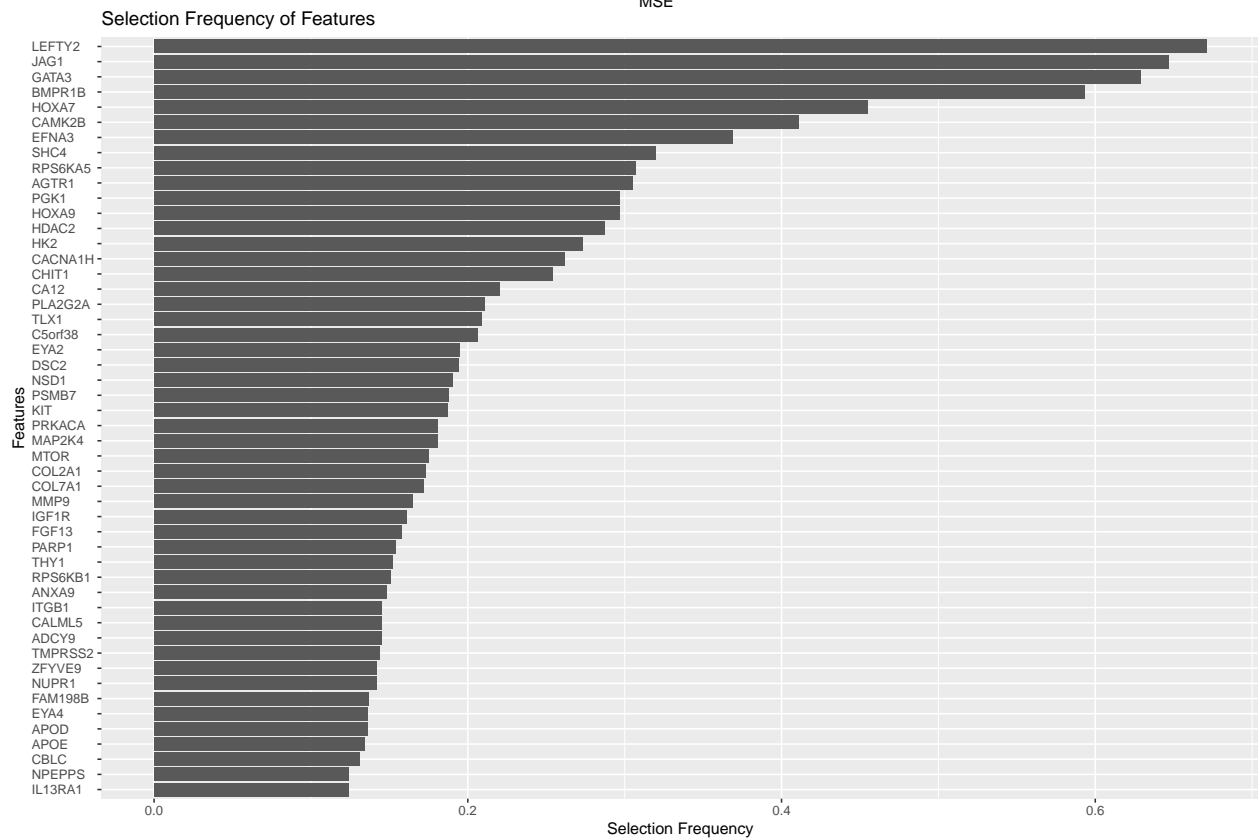
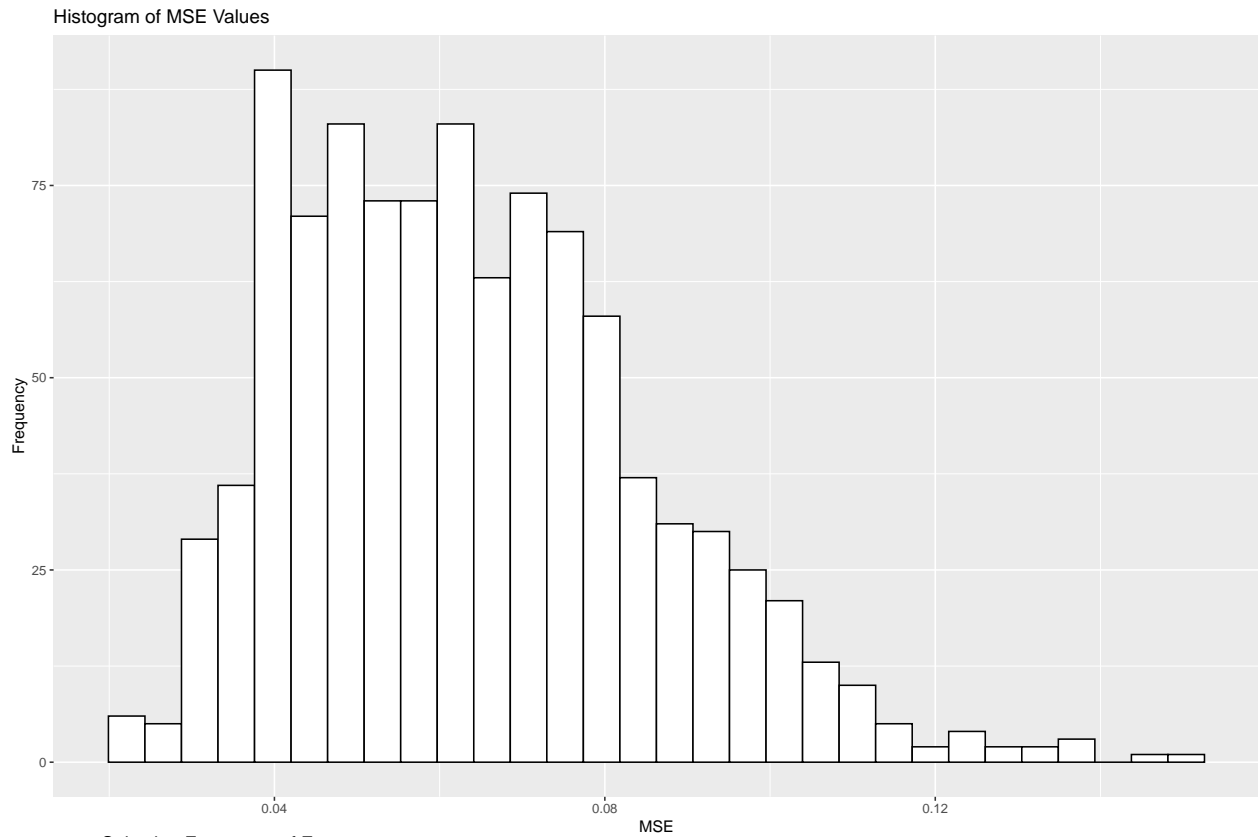
```
## Features selected 50% or more times:
## cMyc
## Top 20 featrues:
## [1] "cMyc"          "cyclinD1"      "cyclinD1p21"   "cyclinD1Palbo"
## [5] "cyclinEp21"    "Rb1"           "ppRb1"         "p21"
## [9] NA              NA              NA              NA
## [13] NA             NA              NA              NA
## [17] NA             NA              NA              NA
```

Mechanistic + Residuals -> proliferation score (additive)

```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0
##
## CORRELATIONS RESULTS
## Mean: 0.7835808
## Median: 0.7962129
## Variance: 0.007298891
## st.dev.: 0.08543355
```



```
## MSE RESULTS
## Mean: 0.06384841
## Median: 0.06161072
## Variance: 0.0004577716
## st.dev.: 0.0213956
```

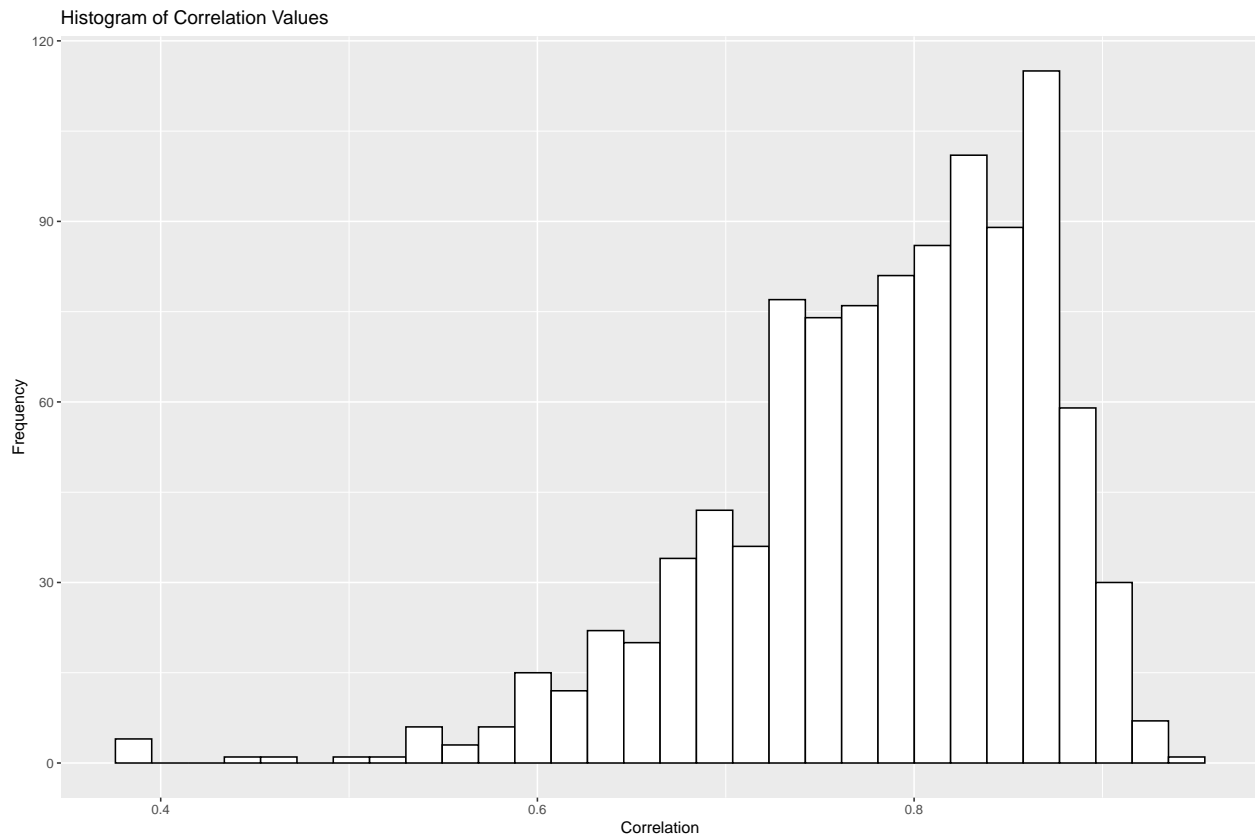


##

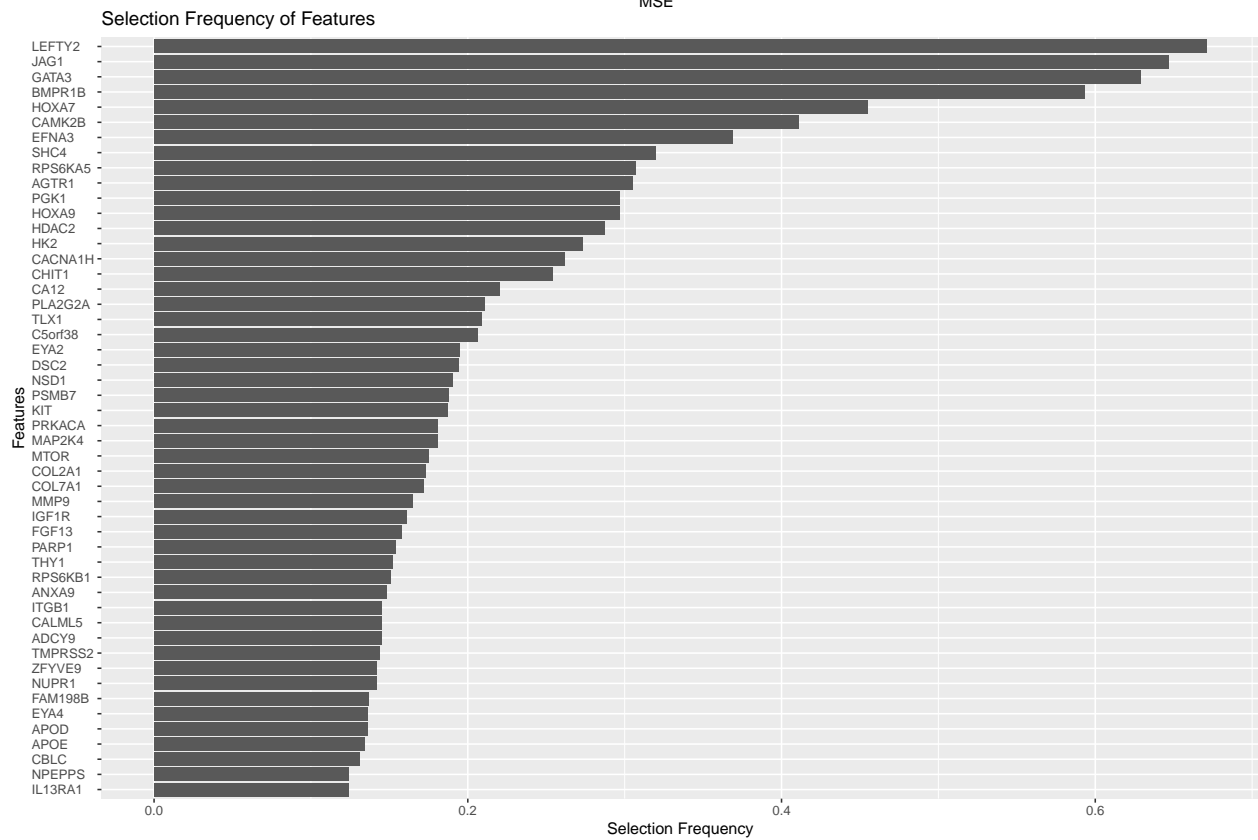
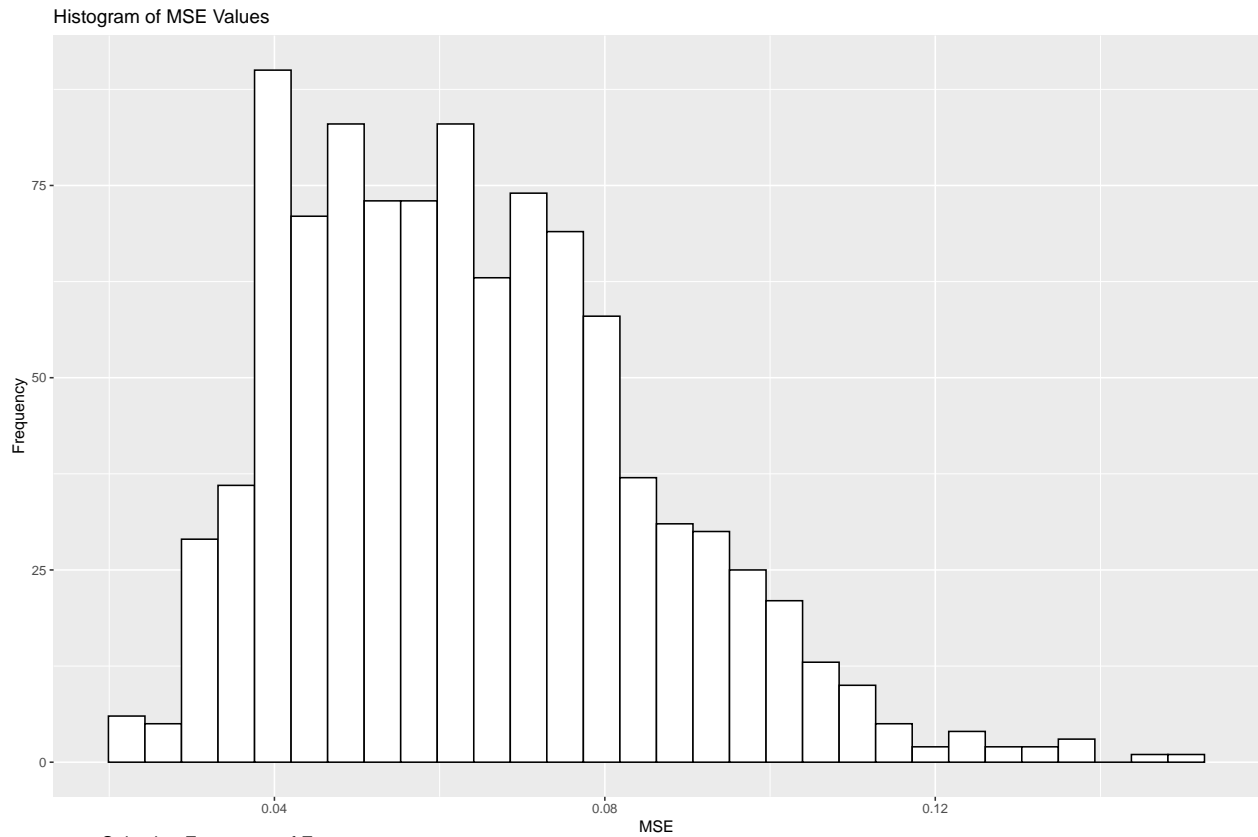

```
## Features selected 50% or more times:
## BMPR1B GATA3 JAG1 LEFTY2
## Top 20 featrues:
## [1] "LEFTY2" "JAG1" "GATA3" "BMPR1B" "HOXA7" "CAMK2B" "EFNA3"
## [8] "SHC4" "RPS6KA5" "AGTR1" "HOXA9" "PGK1" "HDAC2" "HK2"
## [15] "CACNA1H" "CHIT1" "CA12" "PLA2G2A" "TLX1" "C5orf38"
```

Mechanistic + Residuals -> proliferation score (multiplicative)

```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0
##
## CORRELATIONS RESULTS
## Mean: 0.7835808
## Median: 0.7962129
## Variance: 0.007298891
## st.dev.: 0.08543355
```



```
## MSE RESULTS
## Mean: 0.06384841
## Median: 0.06161072
## Variance: 0.0004577716
## st.dev.: 0.0213956
```

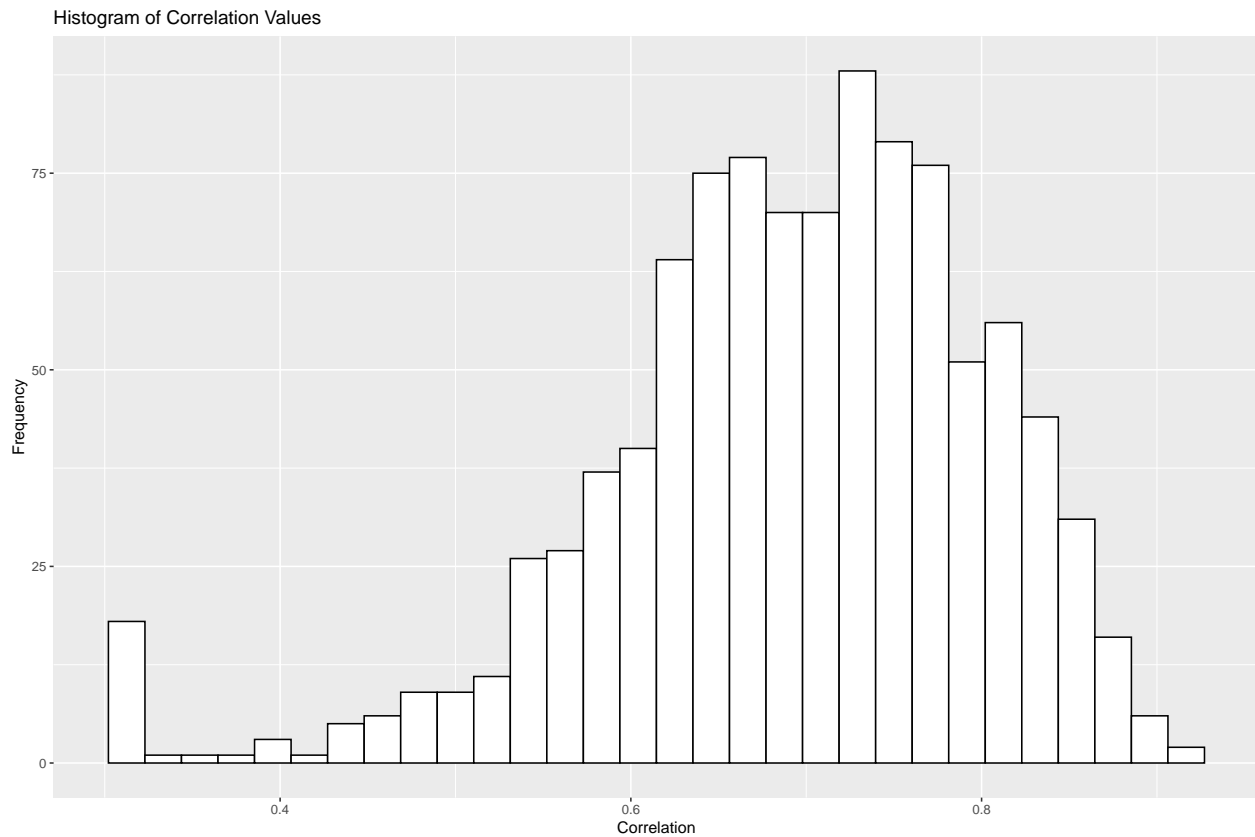


##

```
## Features selected 50% or more times:
## BMPR1B GATA3 JAG1 LEFTY2
## Top 20 featrues:
## [1] "LEFTY2" "JAG1" "GATA3" "BMPR1B" "HOXA7" "CAMK2B" "EFNA3"
## [8] "SHC4" "RPS6KA5" "AGTR1" "HOXA9" "PGK1" "HDAC2" "HK2"
## [15] "CACNA1H" "CHIT1" "CA12" "PLA2G2A" "TLX1" "C5orf38"
```

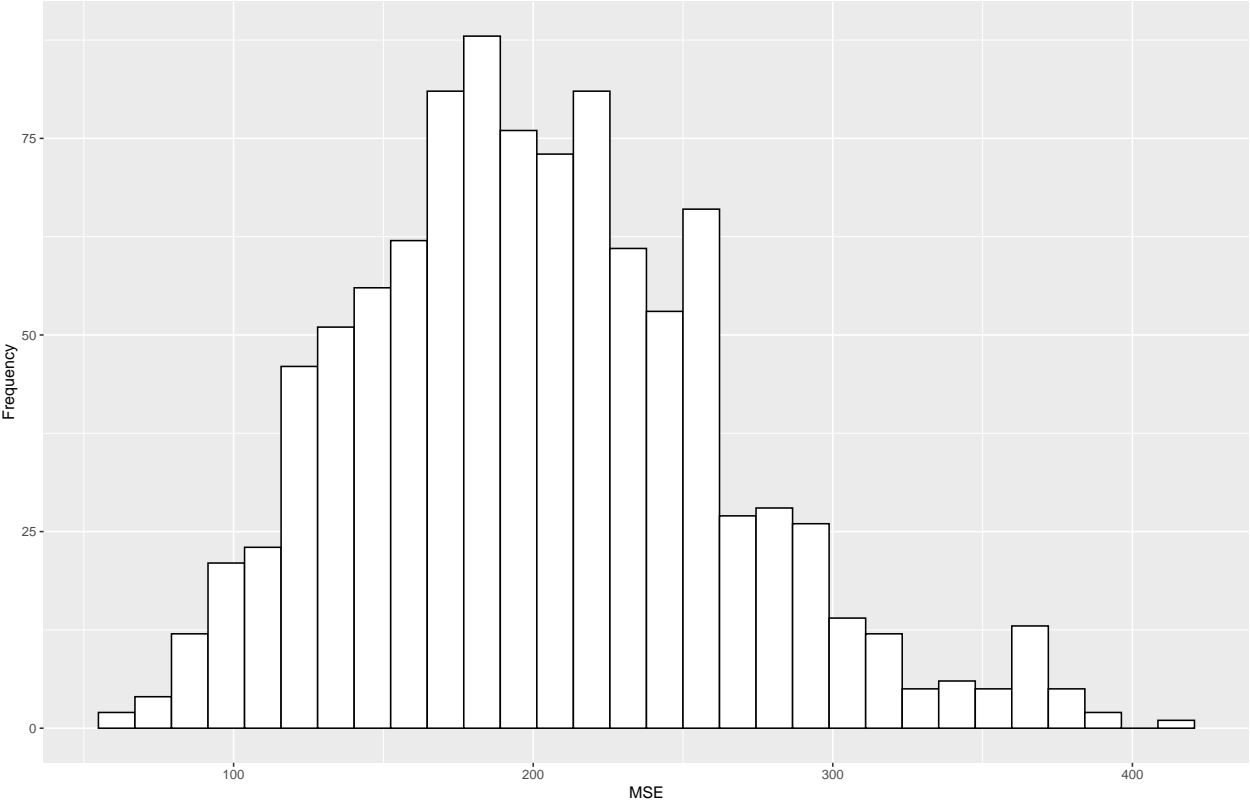
Mechnaistic + Residuals -> ROR-proliferation score (additive)

```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0
##
## CORRELATIONS RESULTS
## Mean: 0.6925953
## Median: 0.7037142
## Variance: 0.01193153
## st.dev.: 0.1092315
```

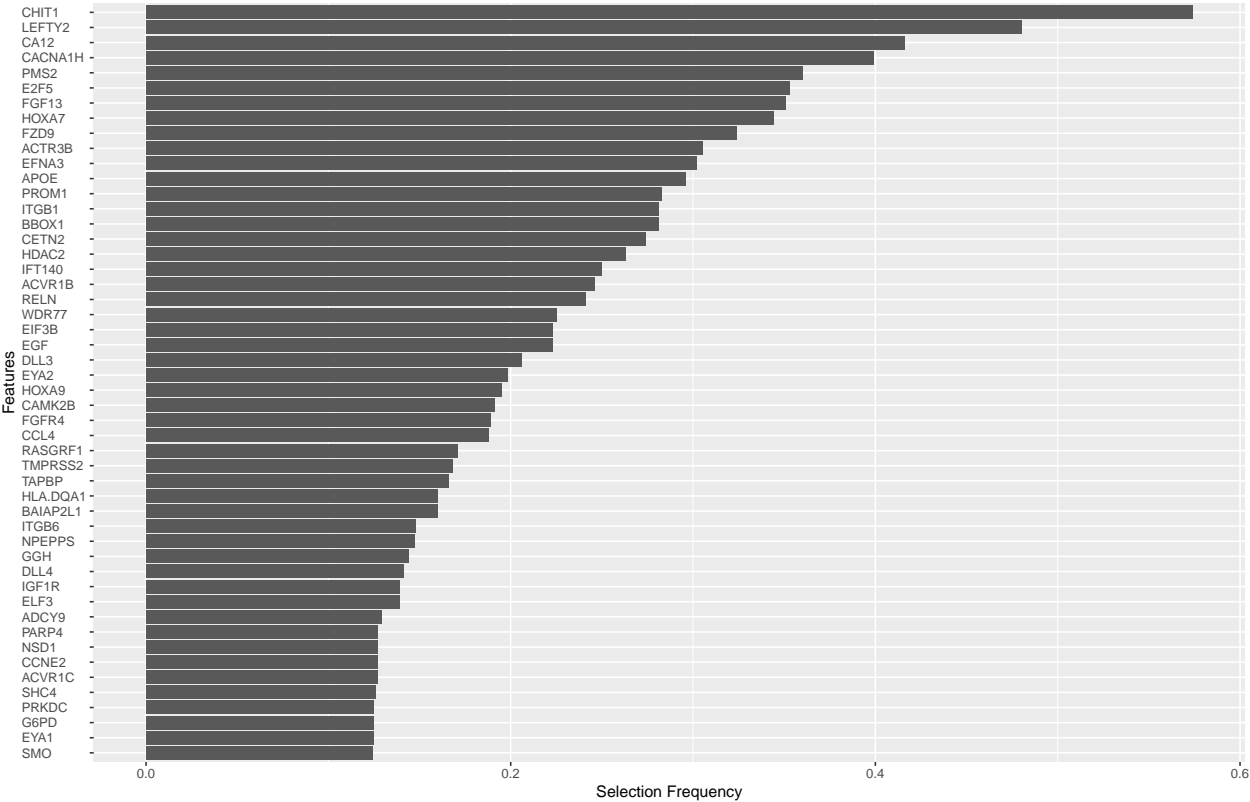


```
## MSE RESULTS
## Mean: 202.3235
## Median: 197.4845
## Variance: 3726.169
## st.dev.: 61.04236
```

Histogram of MSE Values



Selection Frequency of Features



##

```

## Features selected 50% or more times:
## CHIT1
## Top 20 featrues:
## [1] "CHIT1"    "LEFTY2"   "CA12"     "CACNA1H"  "PMS2"     "E2F5"     "FGF13"
## [8] "HOXA7"    "FZD9"     "ACTR3B"   "EFNA3"    "APOE"     "PROM1"    "BBOX1"
## [15] "ITGB1"    "CETN2"    "HDAC2"    "IFT140"   "ACVR1B"   "RELN"

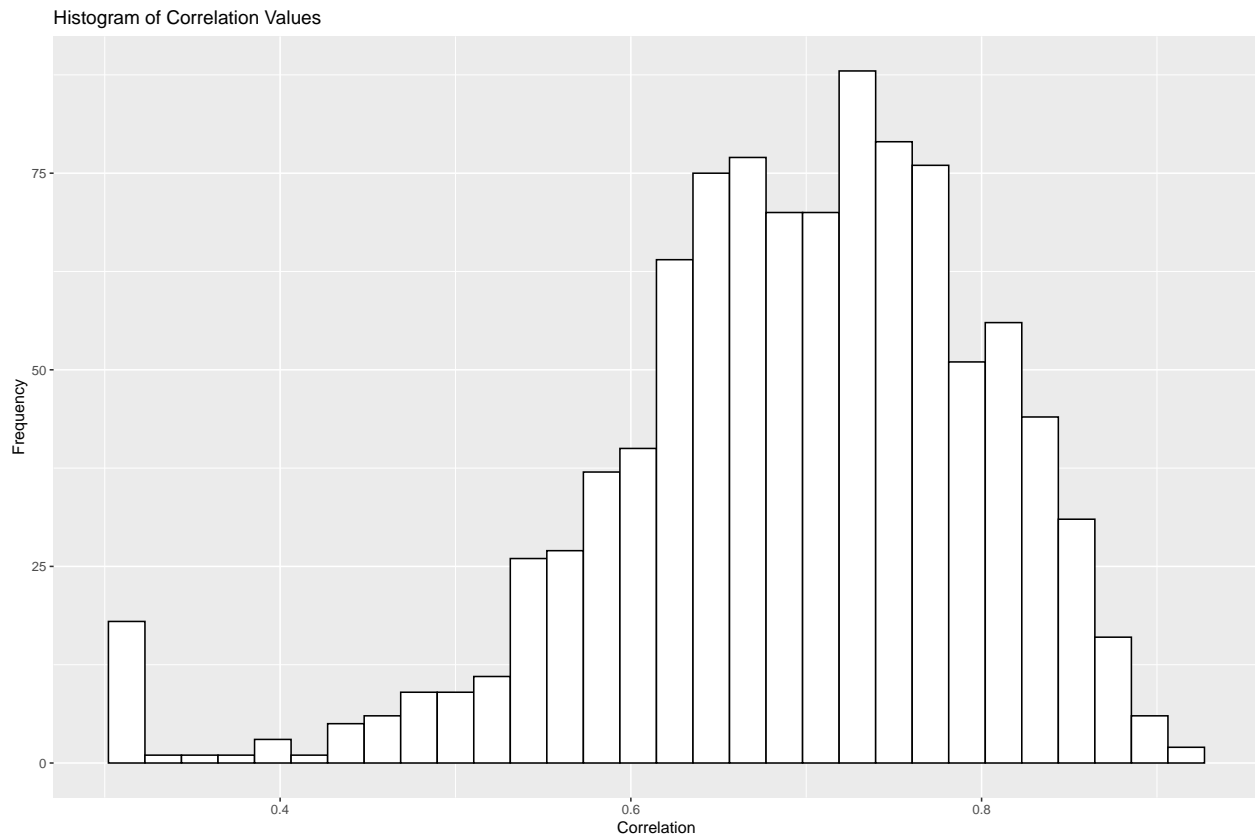
```

Mechnaistic + Residuals -> ROR-proliferation score (multiplicative)

```

## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0
##
## CORRELATIONS RESULTS
## Mean: 0.6925953
## Median: 0.7037142
## Variance: 0.01193153
## st.dev.: 0.1092315

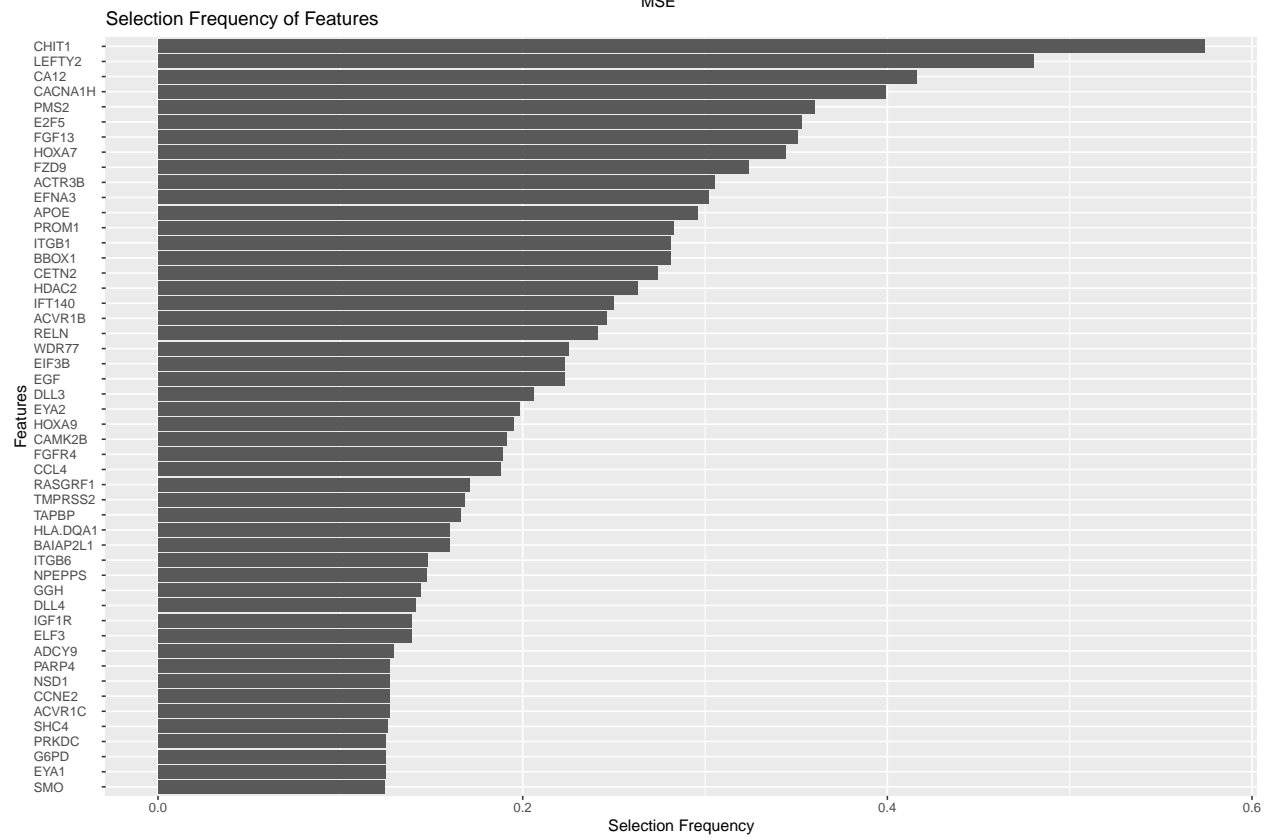
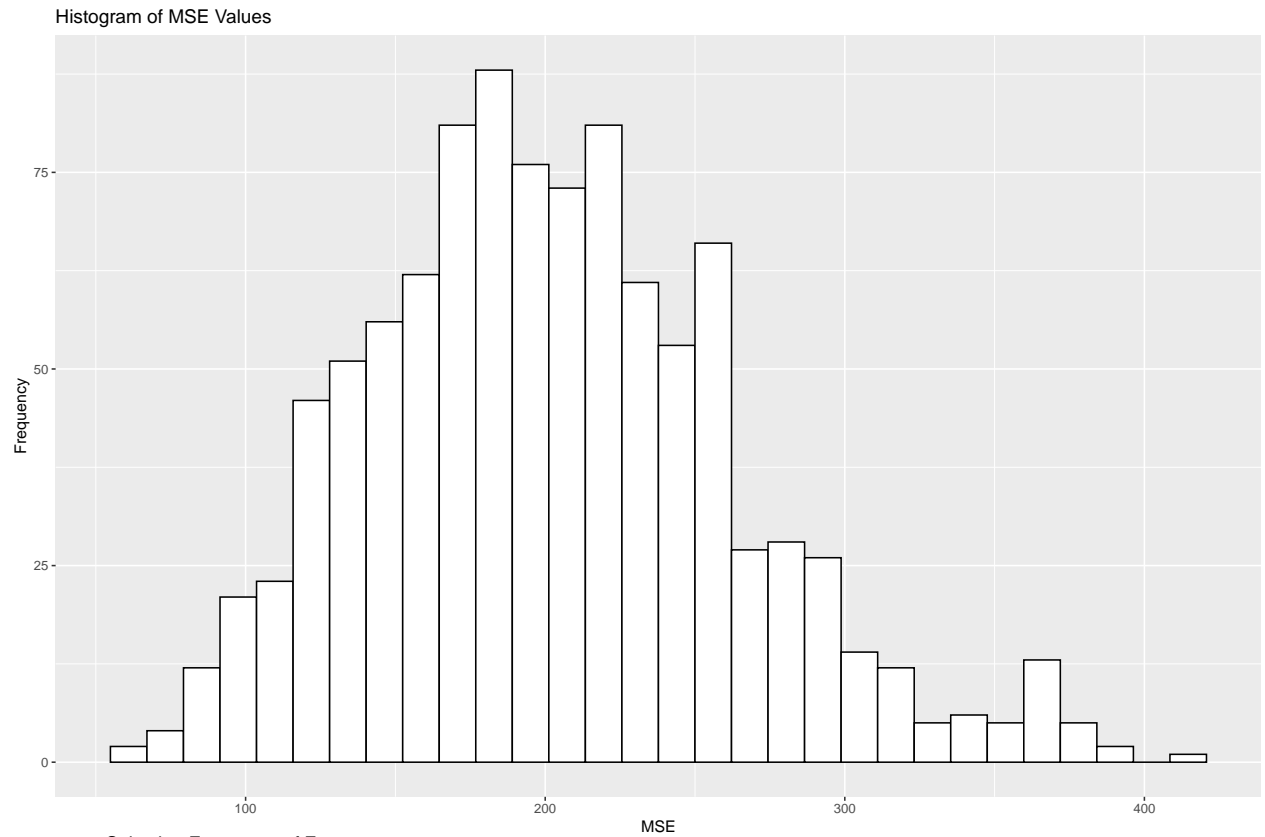
```



```

## MSE RESULTS
## Mean: 202.3235
## Median: 197.4845
## Variance: 3726.169
## st.dev.: 61.04236

```



##

```
## Features selected 50% or more times:
## CHIT1
## Top 20 featrues:
## [1] "CHIT1" "LEFTY2" "CA12" "CACNA1H" "PMS2" "E2F5" "FGF13"
## [8] "HOXA7" "FZD9" "ACTR3B" "EFNA3" "APOE" "PROM1" "BBOX1"
## [15] "ITGB1" "CETN2" "HDAC2" "IFT140" "ACVR1B" "RELN"
```

Summery results: lasso proliferation score (bootstrap)

Model	cor_mean	sd_cor	MSE_mean	MSE_sd
lasso 6 genes	0.3498379	0.0631210	0.1492302	0.0124509
lasso 771 genes	0.7941413	0.0901070	0.0620913	0.0239813
Nodes	0.4144960	0.0642722	0.1479731	0.1916630
Residual additive	0.7835808	0.0854335	0.0638484	0.0213956
Residual multiplicative	0.7835808	0.0854335	0.0638484	0.0213956

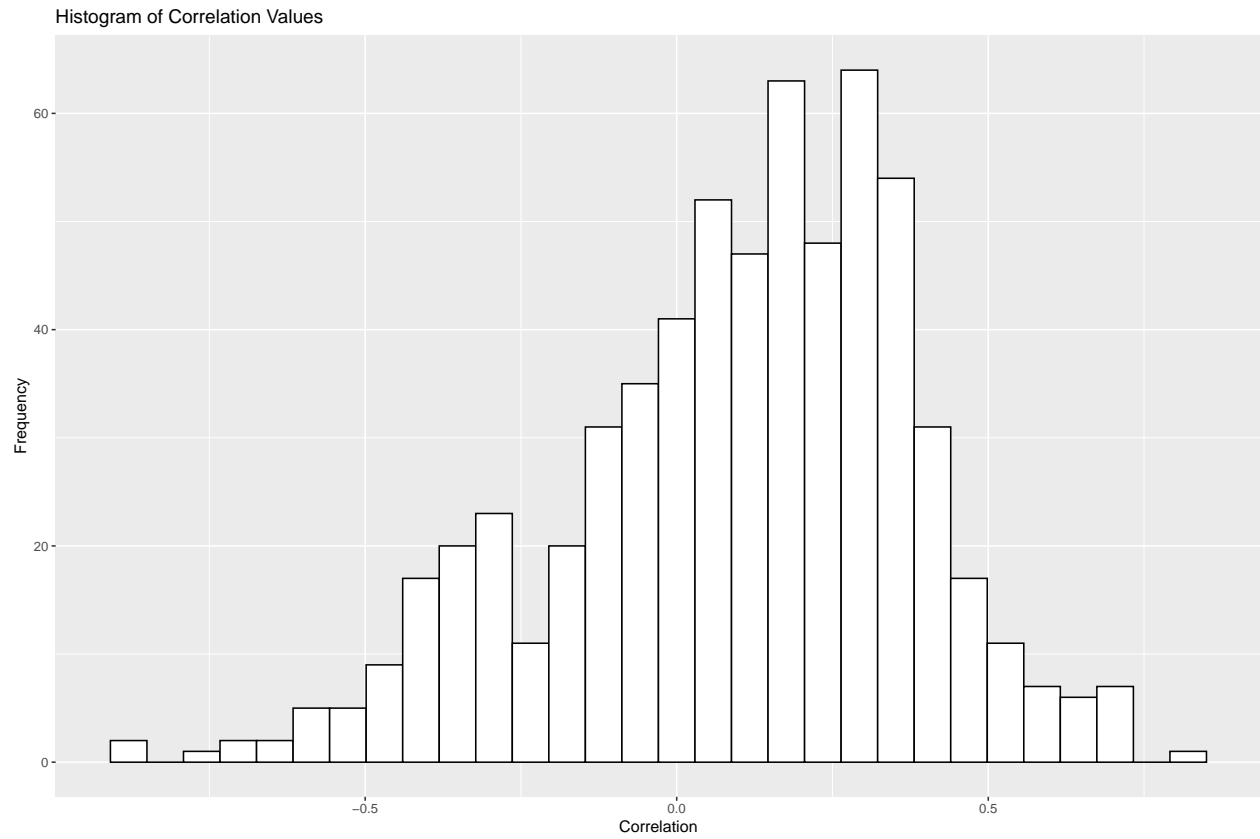
Summery results: lasso ROR+proliferation score (bootstrap)

Model	cor_mean	sd_cor	MSE_mean	MSE_sd
lasso 6 genes	0.1282306	0.0534128	378.0940	23.44024
lasso 771 genes	0.6968101	0.0995060	203.4080	61.34872
Nodes	0.2964169	0.0830696	417.6667	1027.06231
Residual additive	0.6925953	0.1092315	202.3235	61.04236
Residual multiplicative	0.6925953	0.1092315	202.3235	61.04236

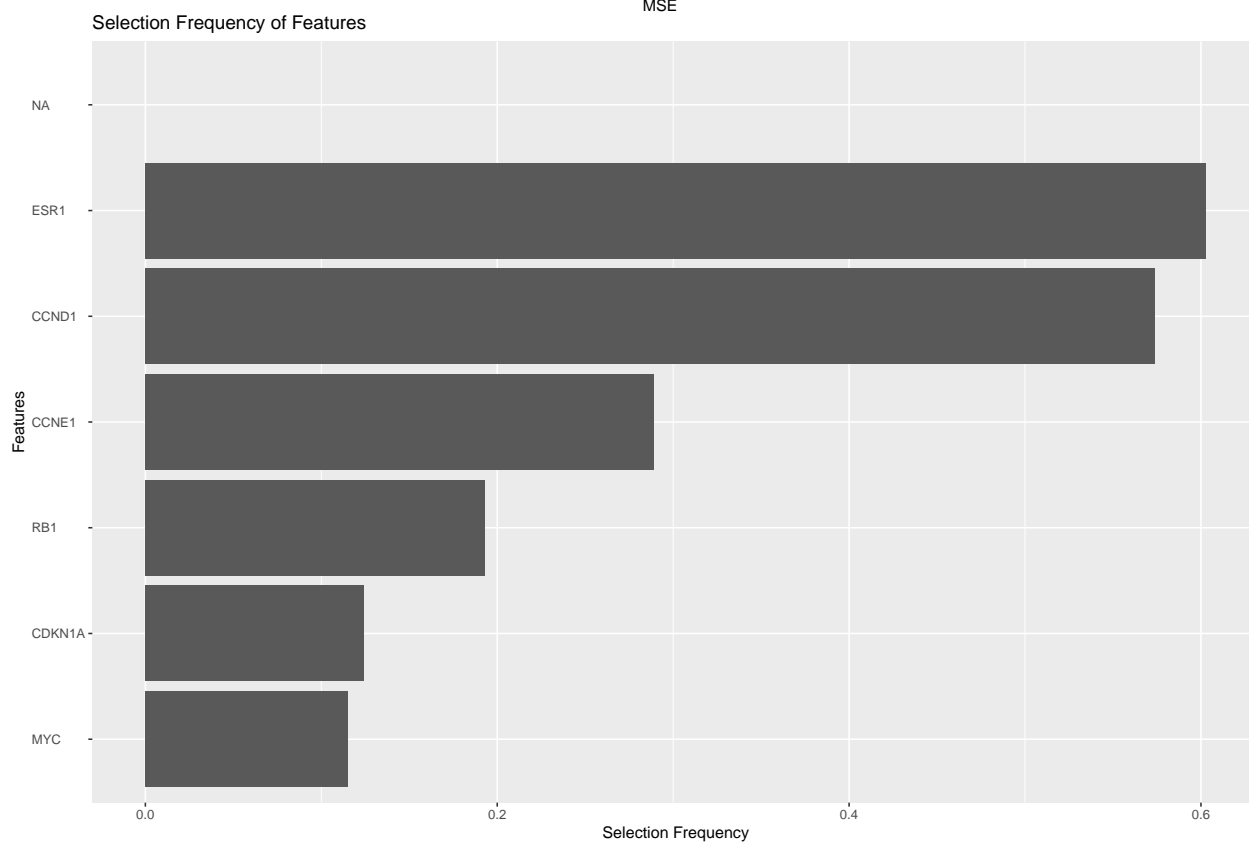
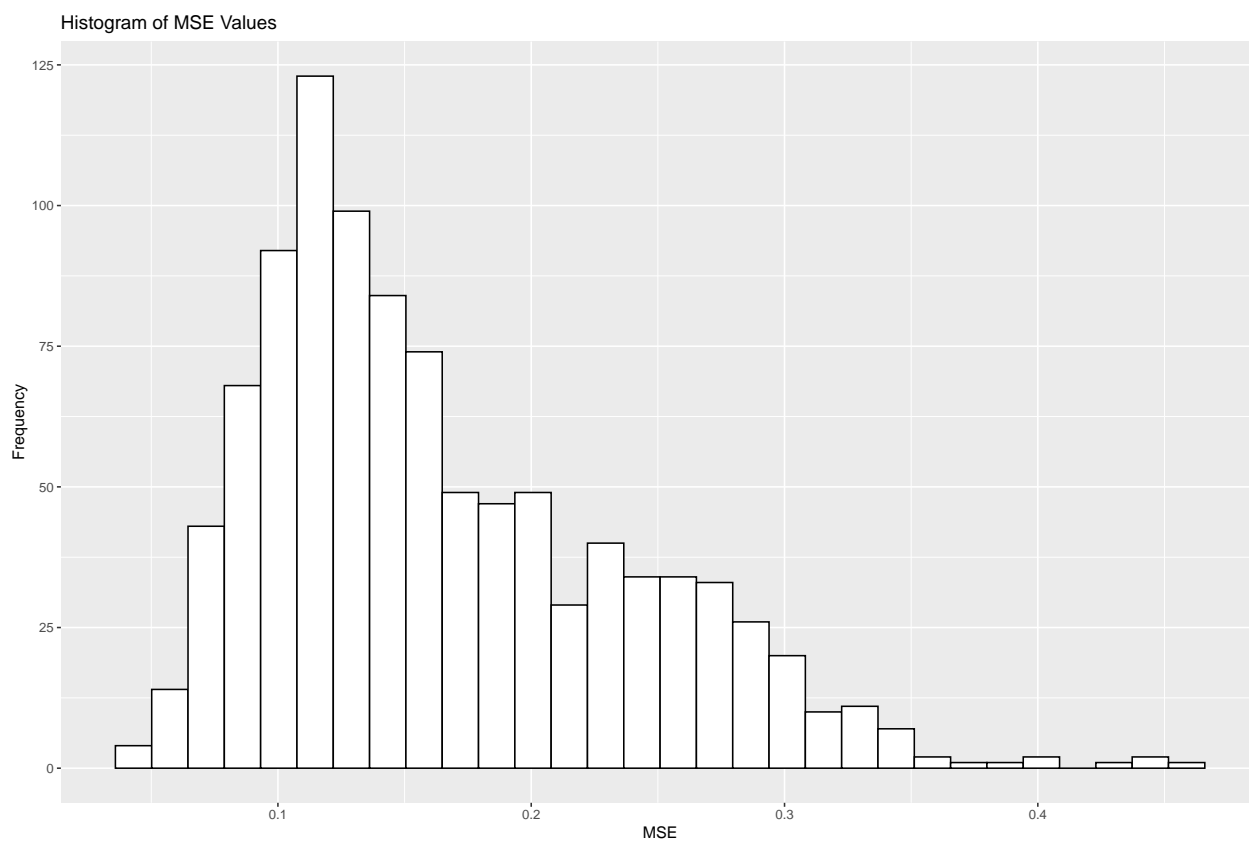
Lasso - Repeated cross-validation

200 repeats of five fold cross-validation ### 6 genes -> proliferation score

```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0.368
##
## CORRELATIONS RESULTS
## Mean: 0.09734364
## Median: 0.1428286
## Variance: 0.0805456
## st.dev.: 0.2838056
```



```
## MSE RESULTS
## Mean: 0.1650201
## Median: 0.146494
## Variance: 0.005185708
## st.dev.: 0.07201186
```

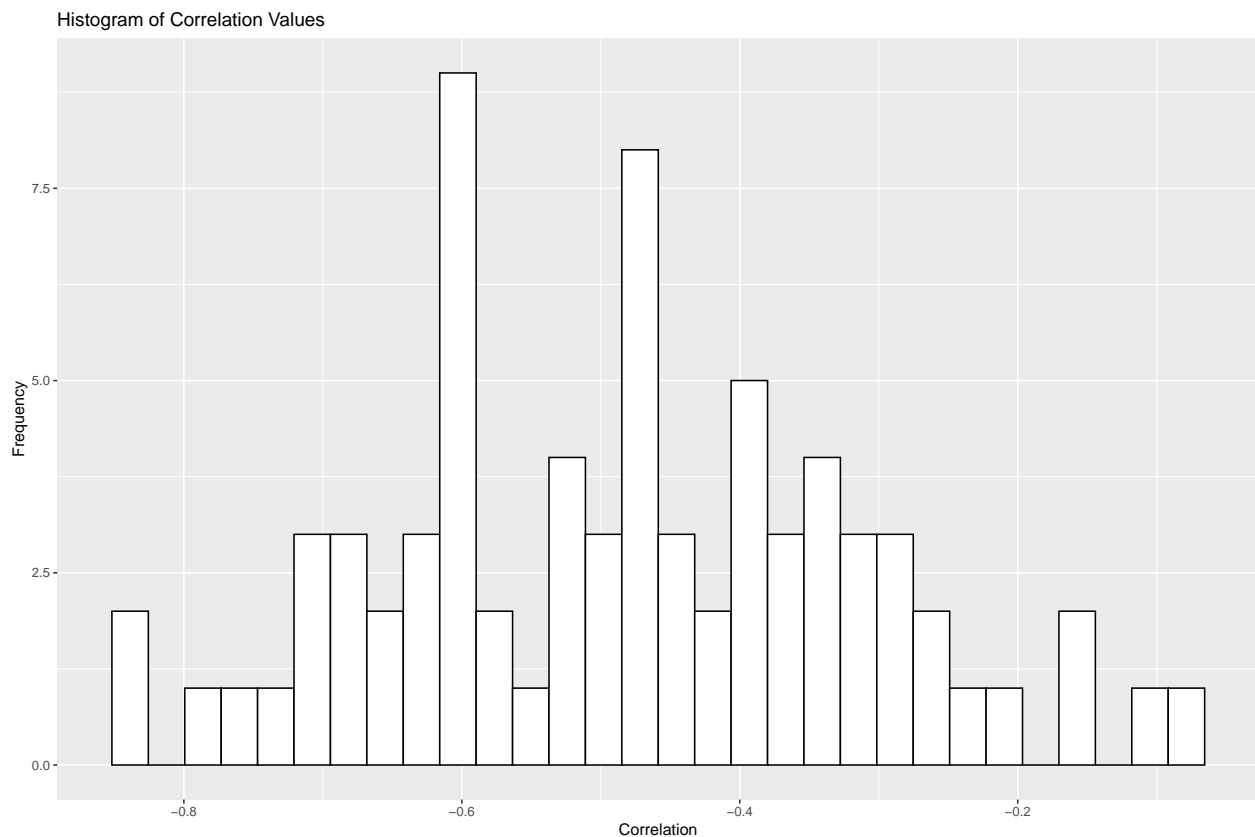



##

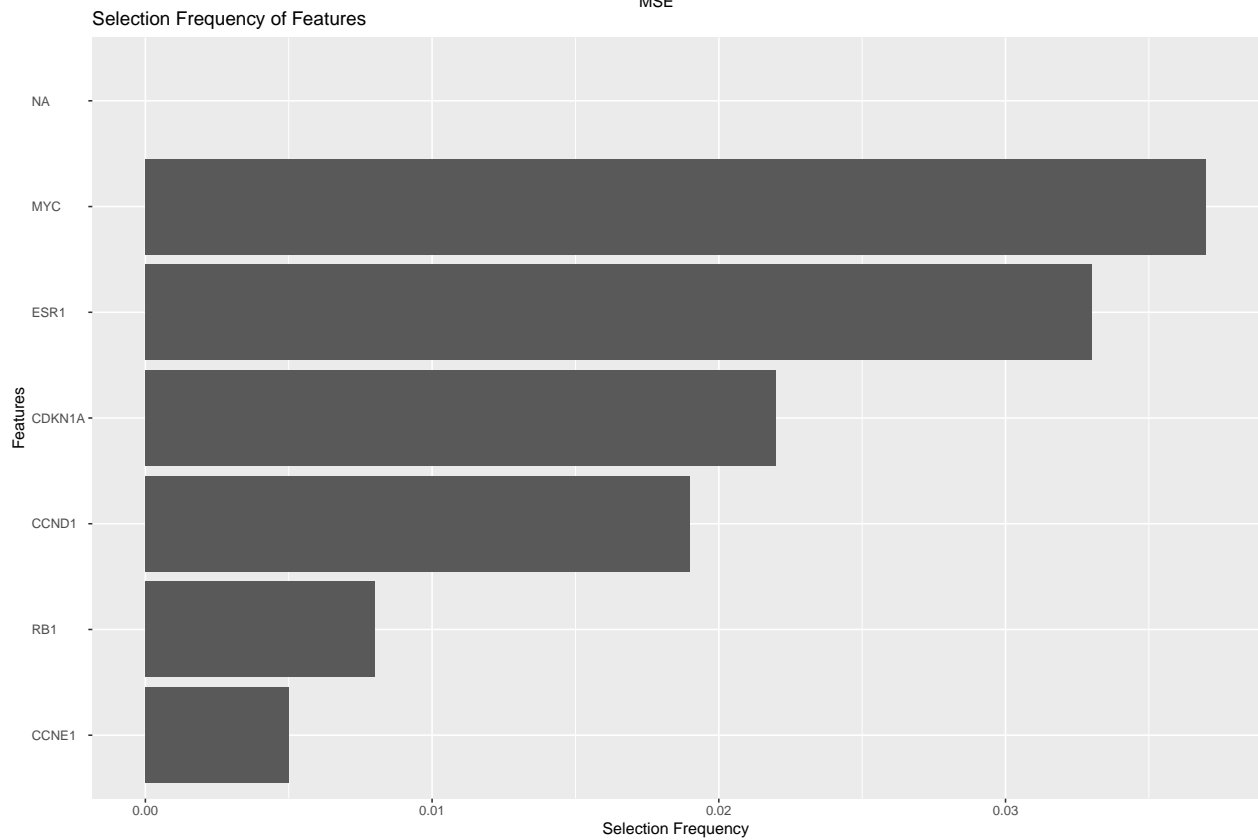
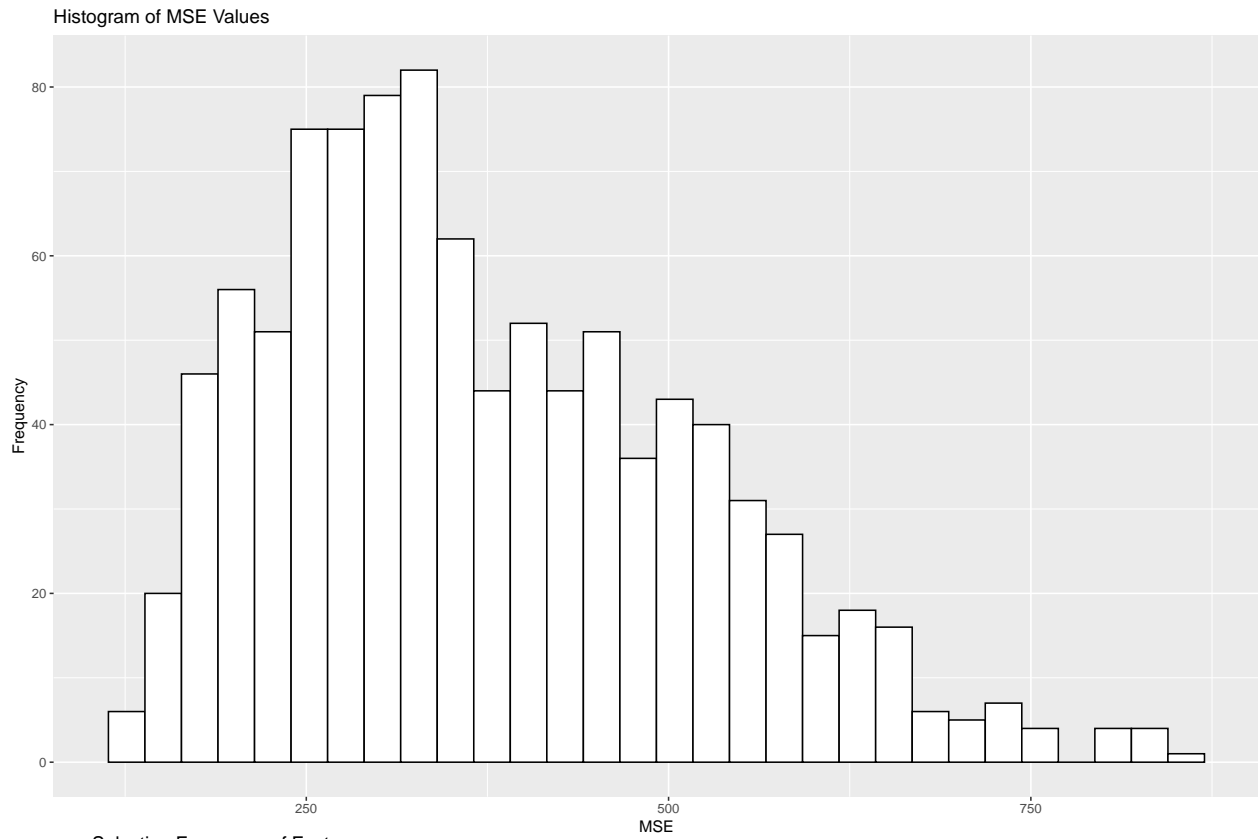
```
## Features selected 50% or more times:
## CCND1 ESR1
## Top 20 featrues:
## [1] "ESR1" "CCND1" "CCNE1" "RB1" "CDKN1A" "MYC" NA NA
## [9] NA NA NA NA NA NA NA NA
## [17] NA NA NA NA
```

6 genes -> ROR_proliferation score

```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0.926
##
## CORRELATIONS RESULTS
## Mean: -0.4822298
## Median: -0.4810641
## Variance: 0.02975572
## st.dev.: 0.1724985
```



```
## MSE RESULTS
## Mean: 374.1519
## Median: 343.2105
## Variance: 20780.92
## st.dev.: 144.1559
```



##

```
## Features selected 50% or more times:
```

```
##
```

```
## Top 20 features:
```

```
## [1] "MYC"      "ESR1"      "CDKN1A"    "CCND1"     "RB1"       "CCNE1"     NA         NA
## [9] NA         NA          NA          NA          NA          NA          NA         NA
## [17] NA         NA          NA          NA
```

```
771 genes -> proliferation score
```

```
## number of models fitted: 1000
```

```
## Fraction of model fits with no selected genes: 0.011
```

```
##
```

```
## CORRELATIONS RESULTS
```

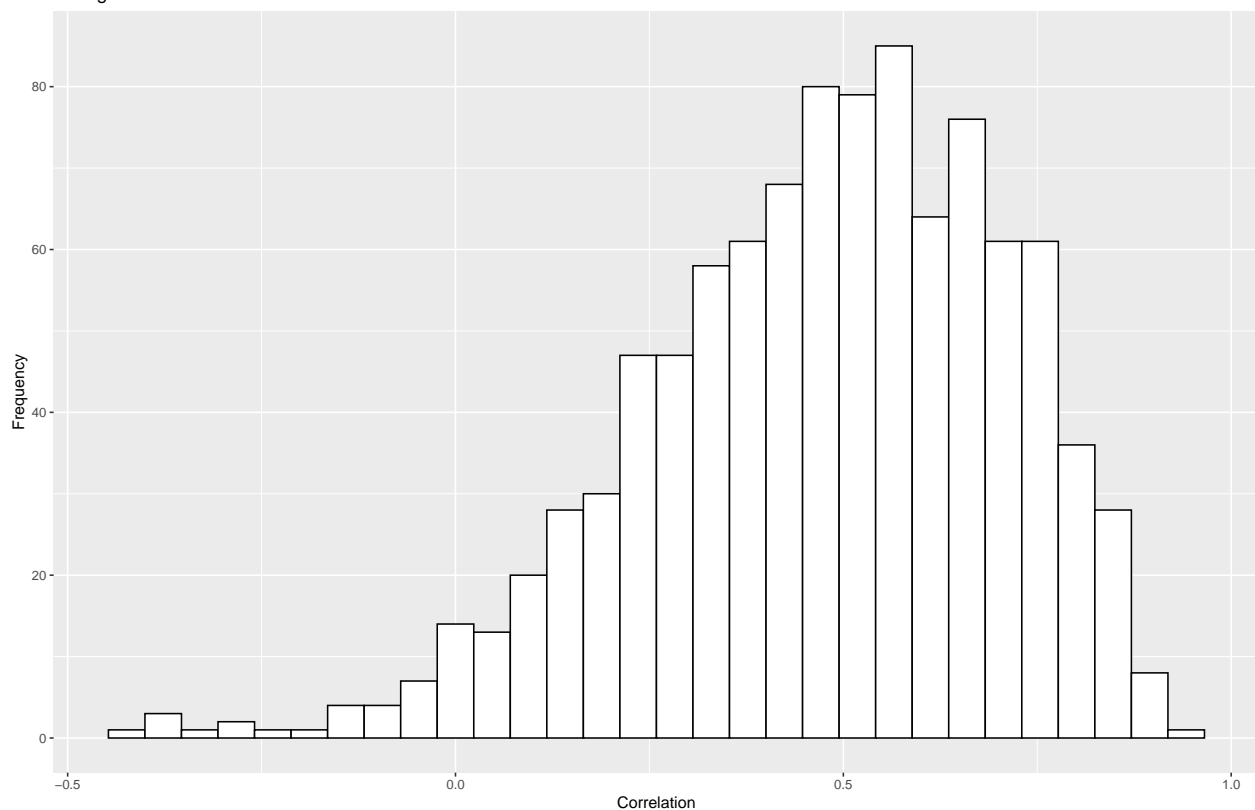
```
## Mean: 0.4737037
```

```
## Median: 0.4959203
```

```
## Variance: 0.05337068
```

```
## st.dev.: 0.2310209
```

```
Histogram of Correlation Values
```



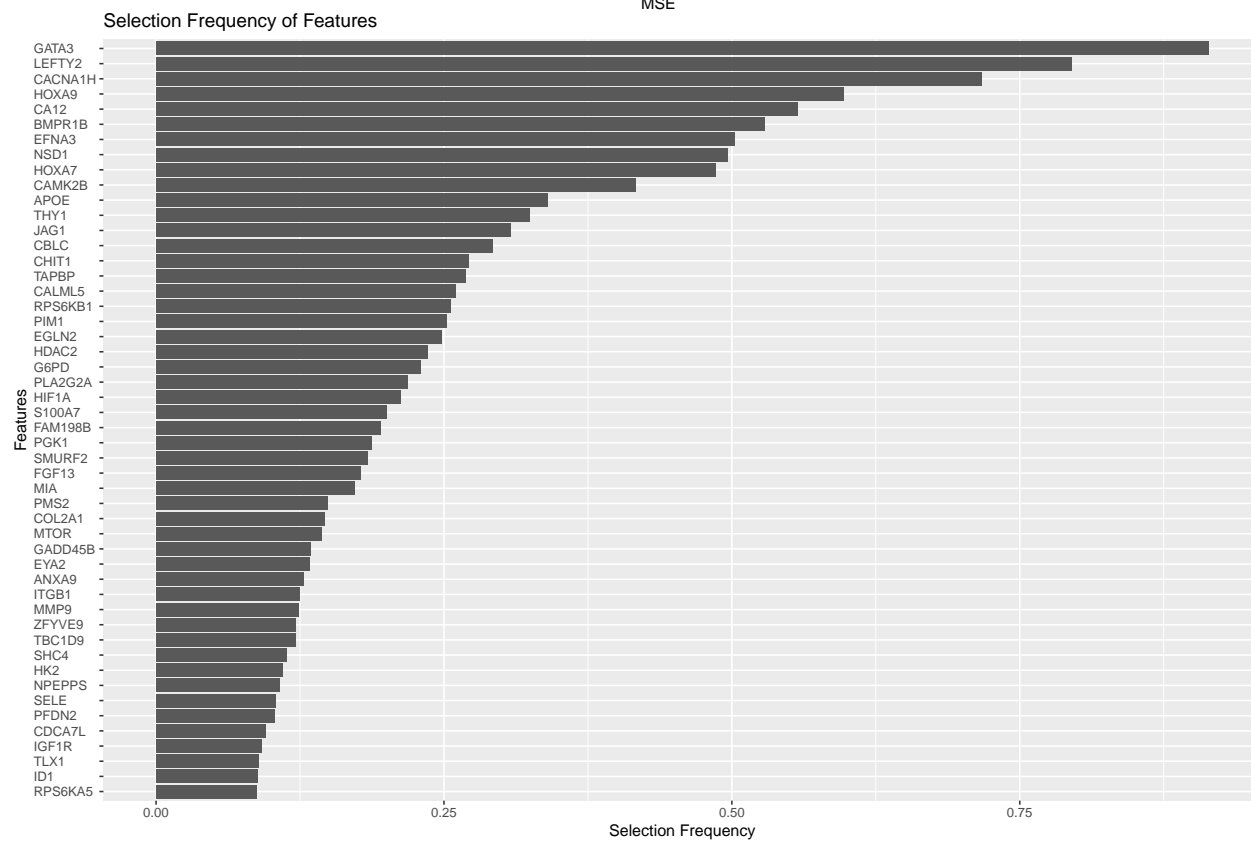
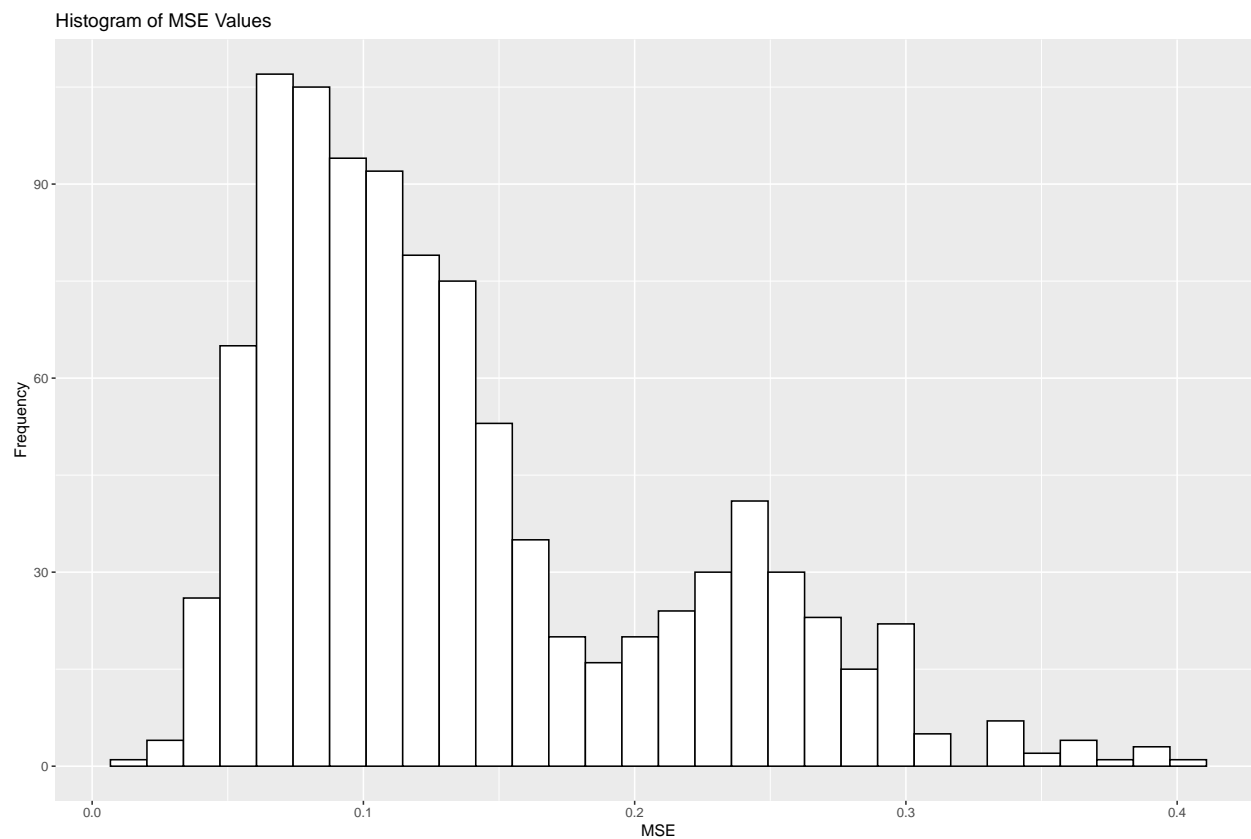
```
## MSE RESULTS
```

```
## Mean: 0.1376002
```

```
## Median: 0.1154157
```

```
## Variance: 0.005670929
```

```
## st.dev.: 0.07530557
```

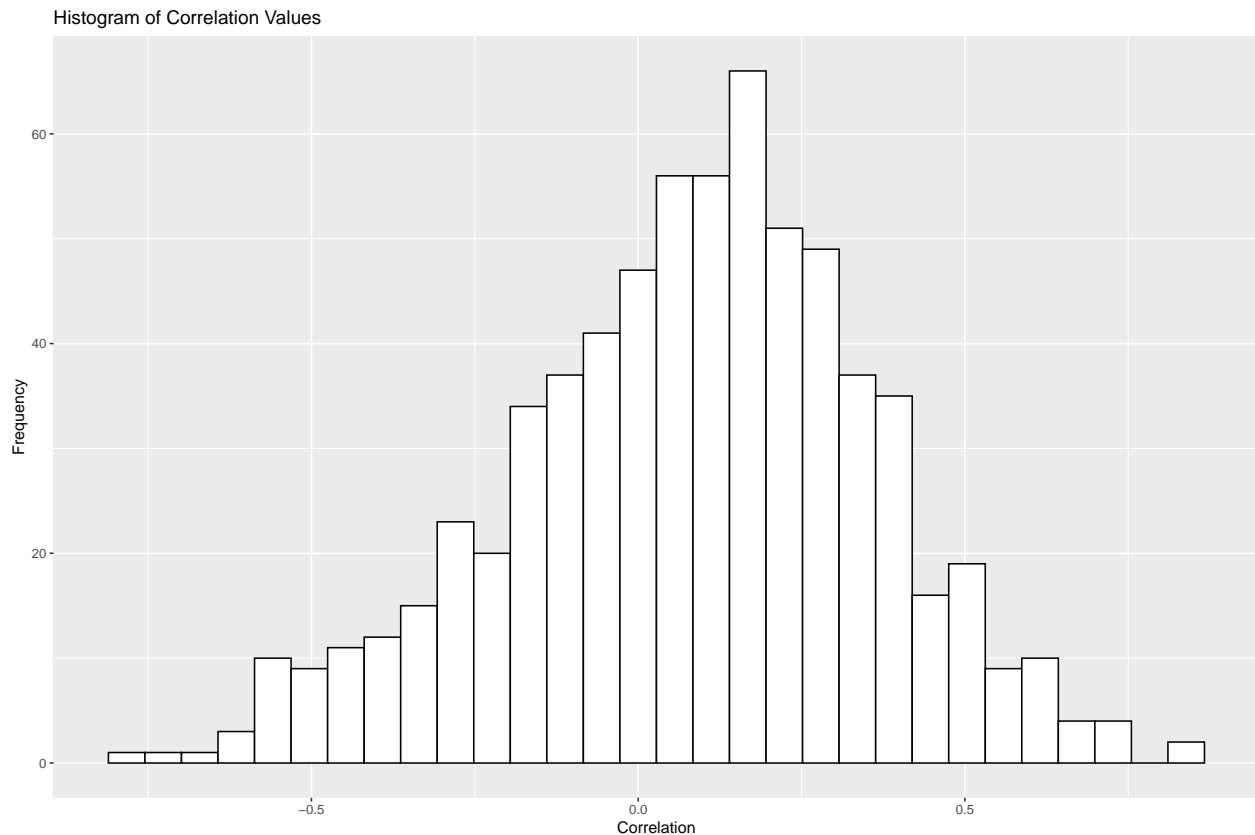


##

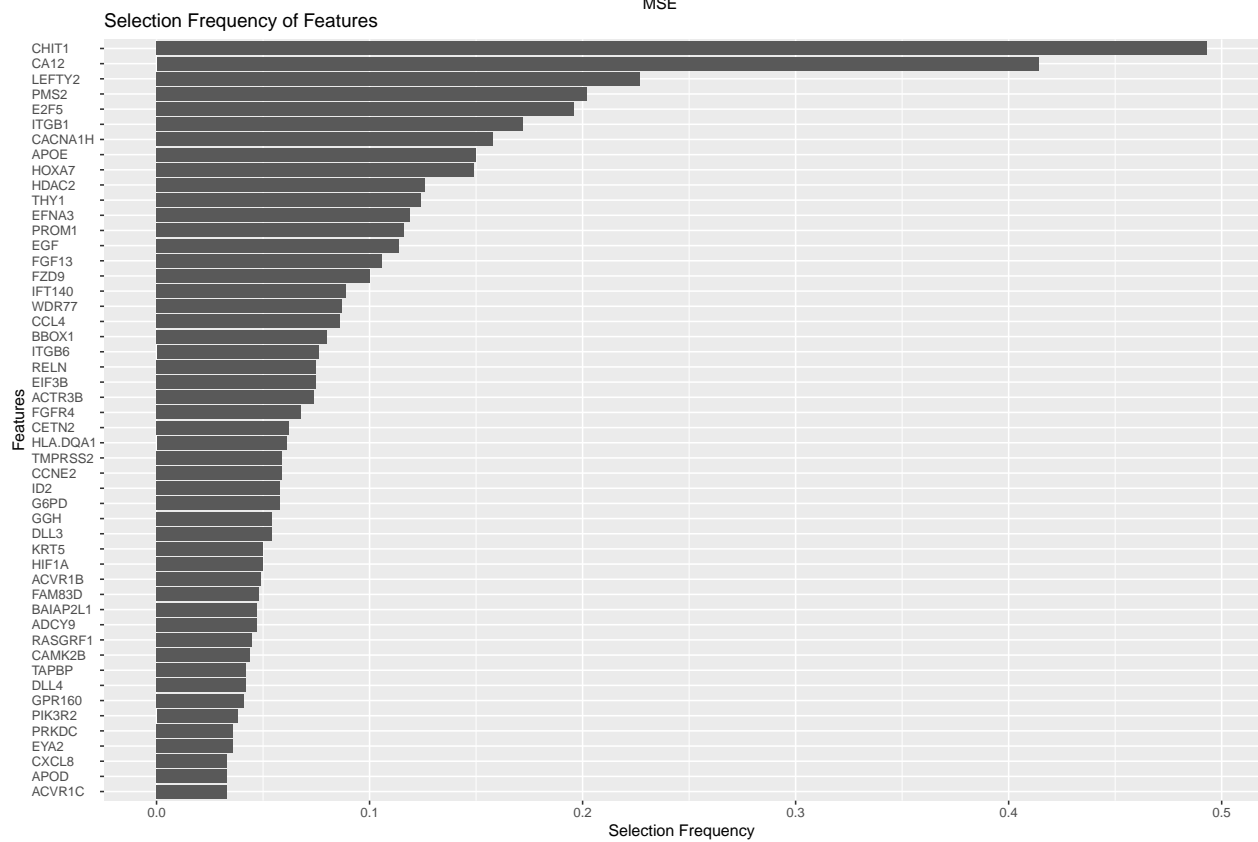
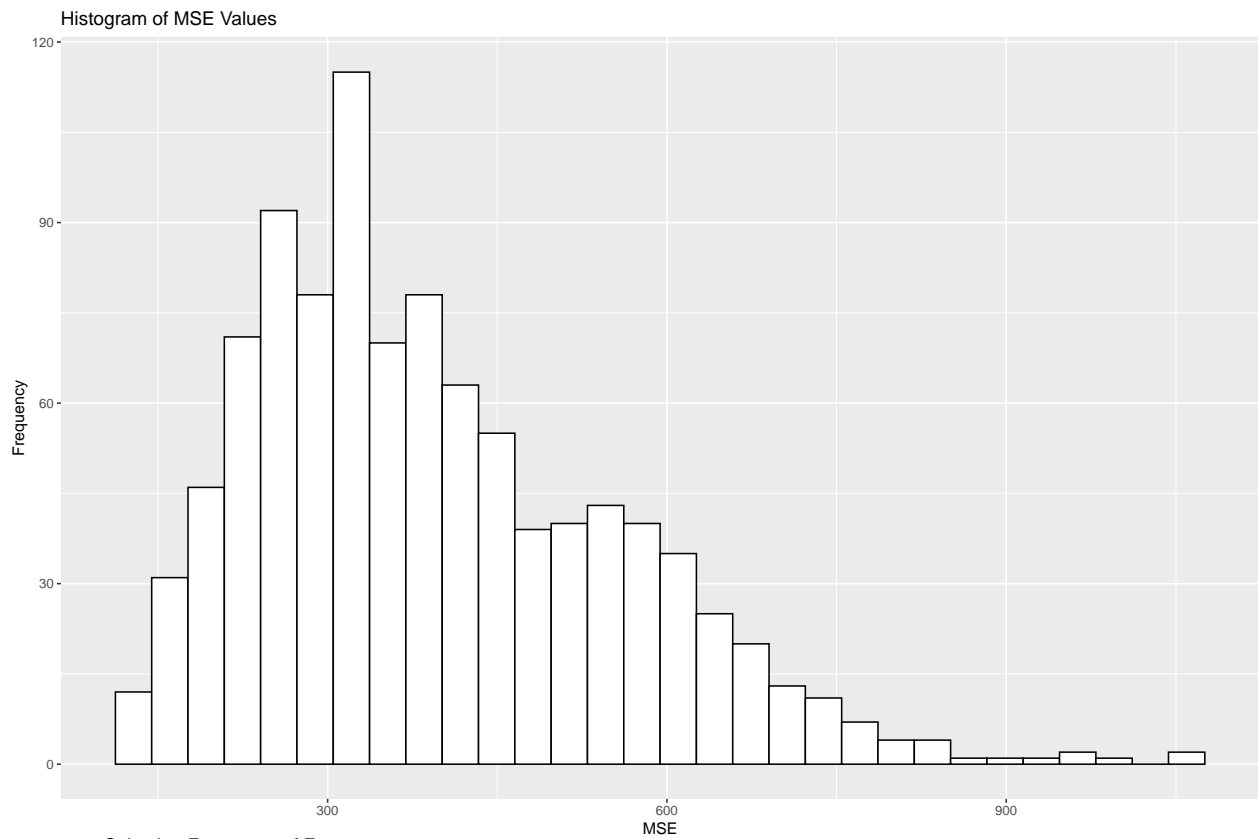
```
## Features selected 50% or more times:
## BMPR1B CA12 CACNA1H EFNA3 GATA3 HOXA9 LEFTY2
## Top 20 featruess:
## [1] "GATA3" "LEFTY2" "CACNA1H" "HOXA9" "CA12" "BMPR1B" "EFNA3"
## [8] "NSD1" "HOXA7" "CAMK2B" "APOE" "THY1" "JAG1" "CBLC"
## [15] "CHIT1" "TAPBP" "CALML5" "RPS6KB1" "PIM1" "EGLN2"
```

771 genes -> ROR-proliferation score

```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0.321
##
## CORRELATIONS RESULTS
## Mean: 0.08062366
## Median: 0.1014264
## Variance: 0.07657135
## st.dev.: 0.2767153
```



```
## MSE RESULTS
## Mean: 393.8069
## Median: 360.5105
## Variance: 25486.55
## st.dev.: 159.6451
```



##

```
## Features selected 50% or more times:
```

```
##
```

```
## Top 20 featrues:
```

```
## [1] "CHIT1" "CA12" "LEFTY2" "PMS2" "E2F5" "ITGB1" "CACNA1H"
```

```
## [8] "APOE" "HOXA7" "HDAC2" "THY1" "EFNA3" "PROM1" "EGF"
```

```
## [15] "FGF13" "FZD9" "IFT140" "WDR77" "CCL4" "BBOX1"
```

```
node values -> proliferation score
```

```
## number of models fitted: 1000
```

```
## Fraction of model fits with no selected genes: 0.063
```

```
##
```

```
## CORRELATIONS RESULTS
```

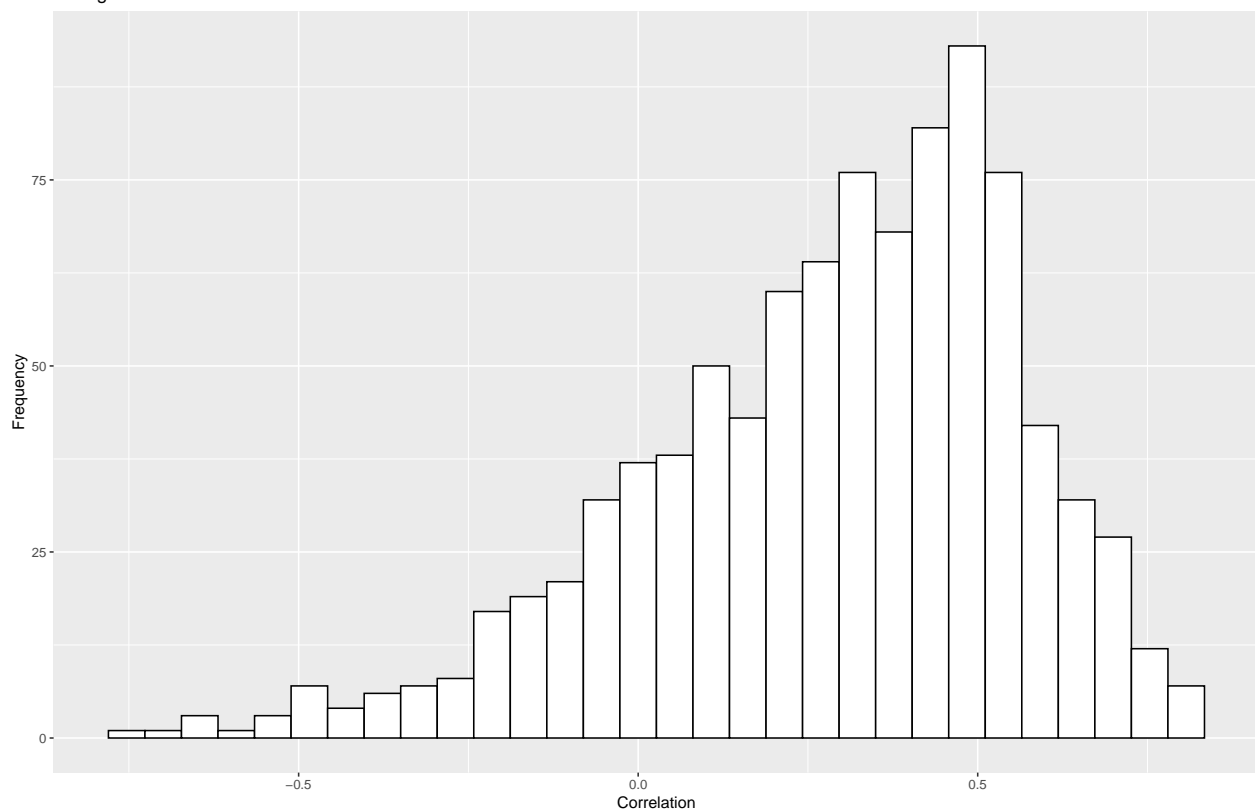
```
## Mean: 0.2842257
```

```
## Median: 0.3249779
```

```
## Variance: 0.07664357
```

```
## st.dev.: 0.2768458
```

```
Histogram of Correlation Values
```



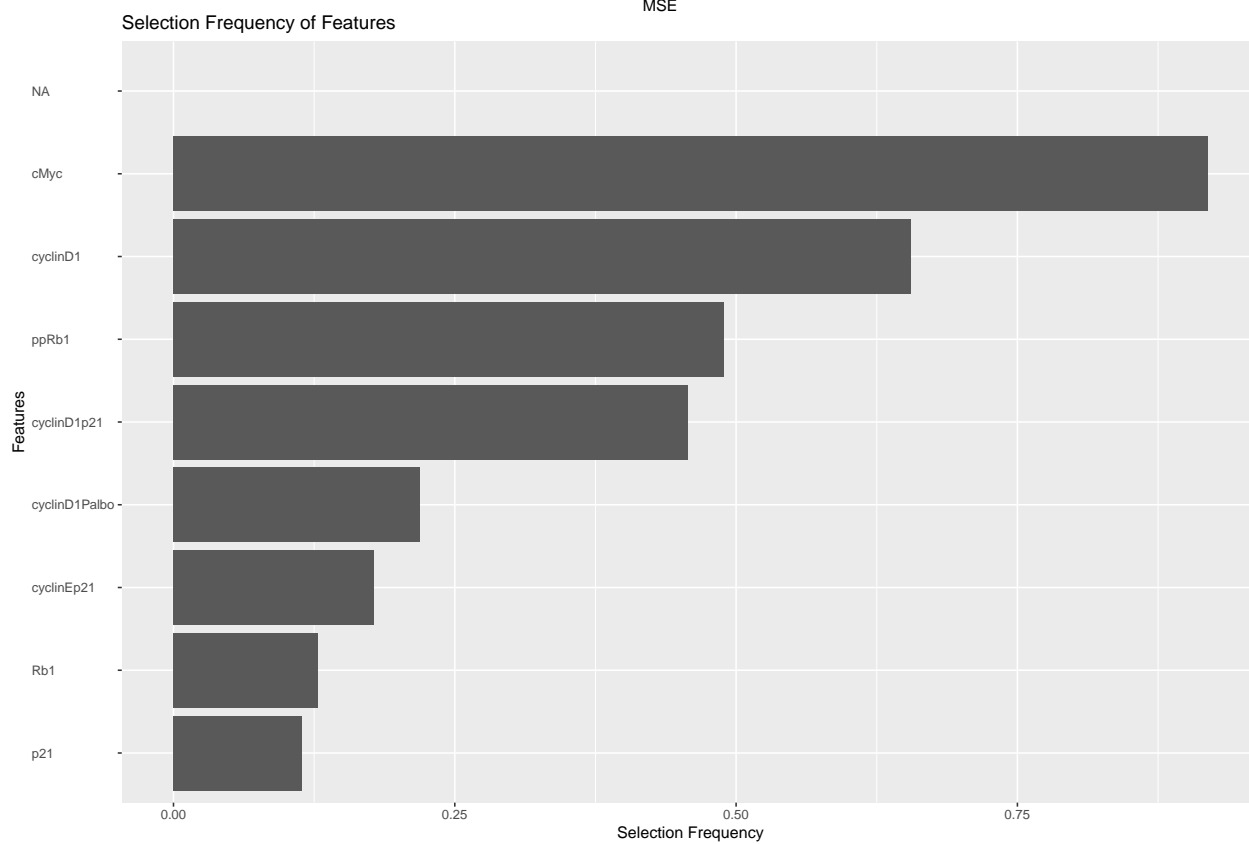
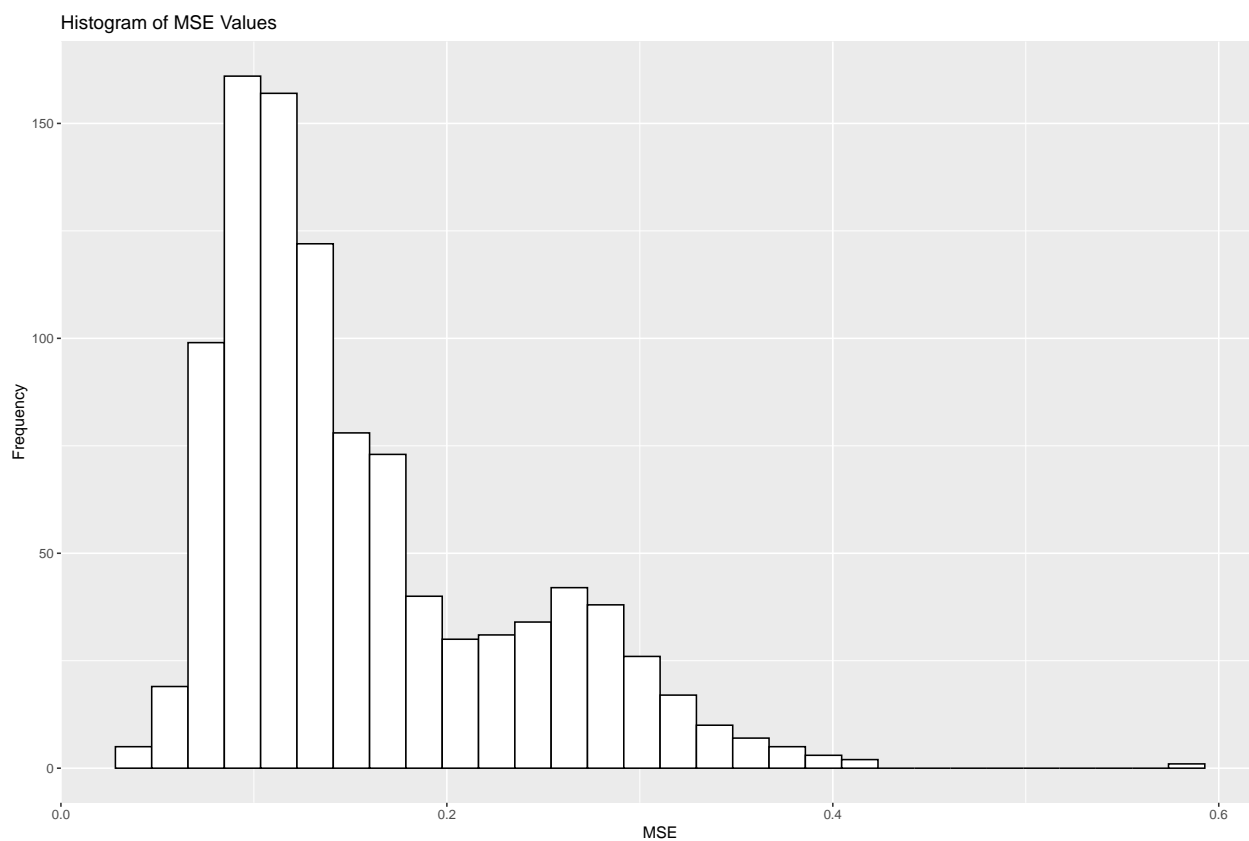
```
## MSE RESULTS
```

```
## Mean: 0.1560308
```

```
## Median: 0.1314678
```

```
## Variance: 0.005819908
```

```
## st.dev.: 0.07628832
```

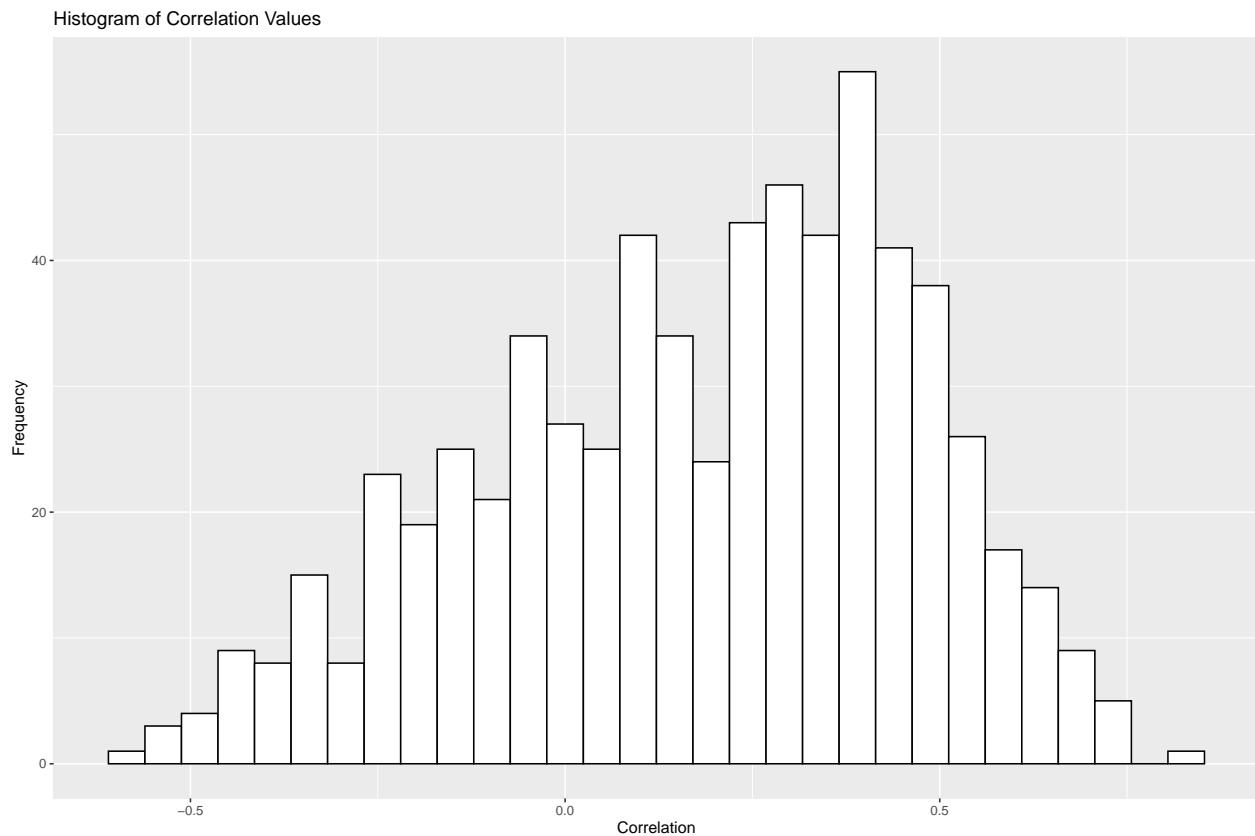



##

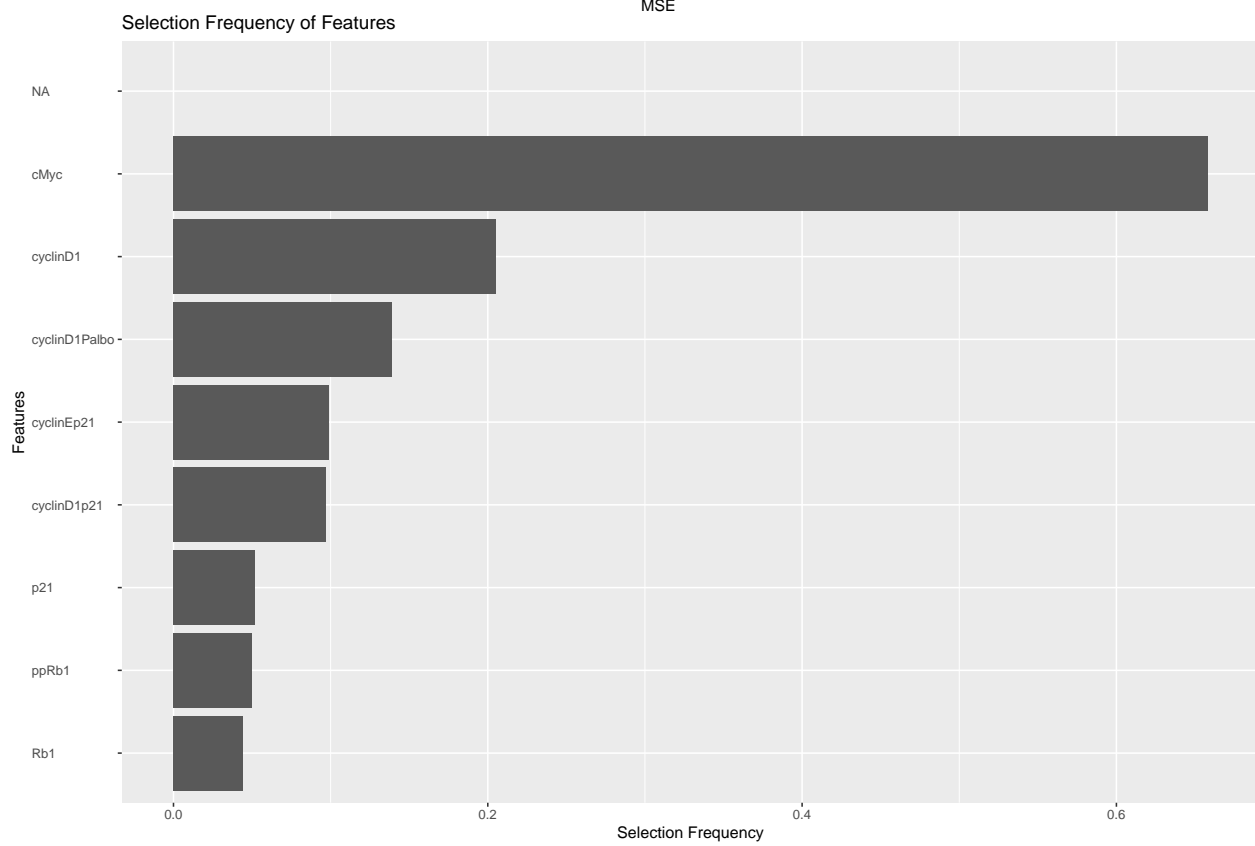
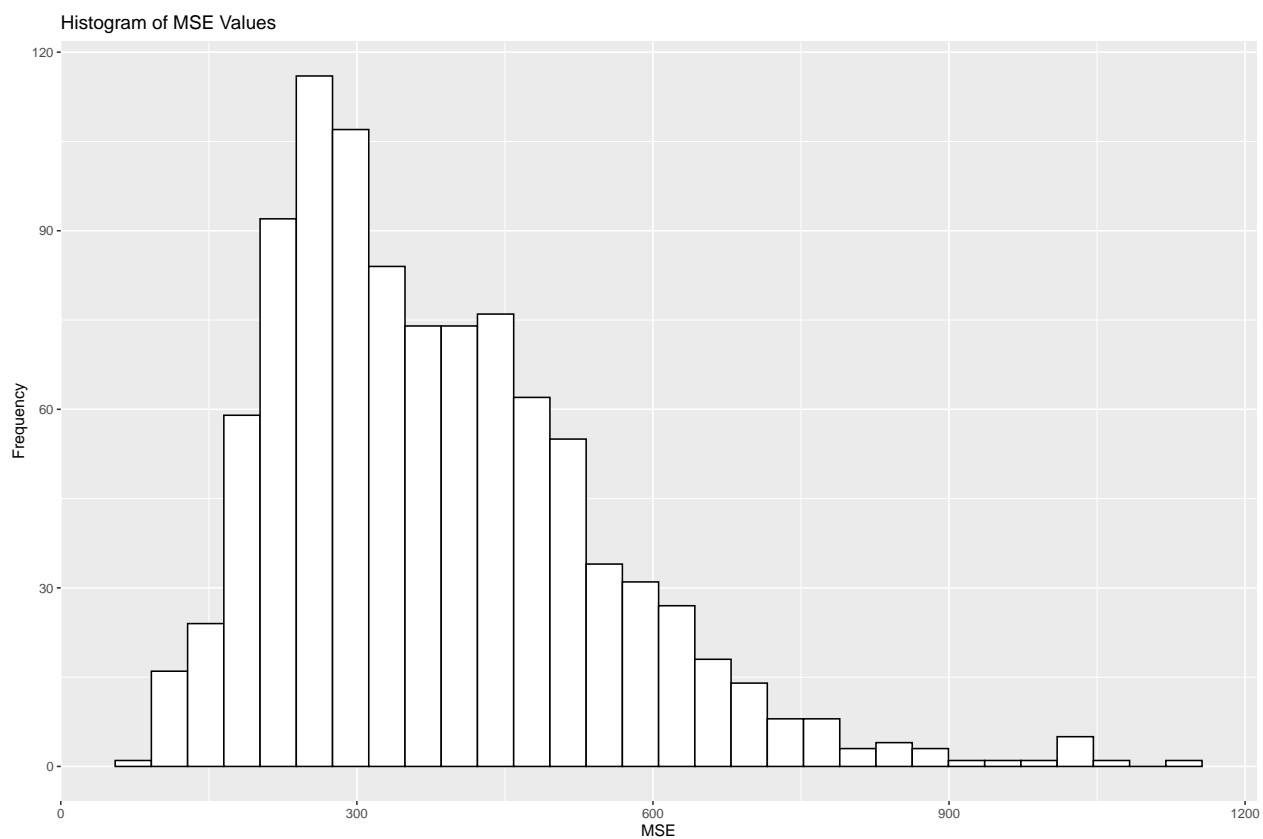
```
## Features selected 50% or more times:
## cyclinD1 cMyc
## Top 20 featrues:
## [1] "cMyc"          "cyclinD1"      "ppRb1"         "cyclinD1p21"
## [5] "cyclinD1Palbo" "cyclinEp21"    "Rb1"           "p21"
## [9] NA              NA              NA              NA
## [13] NA             NA              NA              NA
## [17] NA             NA              NA              NA
```

node values -> ROR-proliferation score

```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0.341
##
## CORRELATIONS RESULTS
## Mean: 0.1806504
## Median: 0.2237481
## Variance: 0.08150408
## st.dev.: 0.2854892
```



```
## MSE RESULTS
## Mean: 380.1157
## Median: 349.3312
## Variance: 27088.84
## st.dev.: 164.5869
```



##

```
## Features selected 50% or more times:
## cMyc
## Top 20 featrues:
## [1] "cMyc"          "cyclinD1"      "cyclinD1Palbo" "cyclinEp21"
## [5] "cyclinD1p21"   "p21"           "ppRb1"          "Rb1"
## [9] NA              NA              NA              NA
## [13] NA             NA              NA              NA
## [17] NA             NA              NA              NA
```

Summery results: lasso proliferation score (Repeated cross-validation)

Model	cor_mean	sd_cor	MSE_mean	MSE_sd
lasso 6 genes	0.3498379	0.0631210	0.1492302	0.0124509
lasso 771 genes	0.7941413	0.0901070	0.0620913	0.0239813
Nodes	0.4144960	0.0642722	0.1479731	0.1916630
Residual additive	0.7835808	0.0854335	0.0638484	0.0213956
Residual multiplicative	0.7835808	0.0854335	0.0638484	0.0213956

Ridge

Elastic Net

Boosting with stumps as base learner

Post Lasso

Summery of results

Lasso proliferation score (bootstrap)

Model	cor_mean	sd_cor	MSE_mean	MSE_sd
lasso 6 genes	0.3498379	0.0631210	0.1492302	0.0124509
lasso 771 genes	0.7941413	0.0901070	0.0620913	0.0239813
Nodes	0.4144960	0.0642722	0.1479731	0.1916630
Residual additive	0.7835808	0.0854335	0.0638484	0.0213956
Residual multiplicative	0.7835808	0.0854335	0.0638484	0.0213956