Milestone 2: Core results 01

Anders

25/1/2023

Data

One clinical trails 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 dived 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 then 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.

Respones 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 proliferaton 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 lock 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 prolif

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 challanges

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

PCA feature engineering

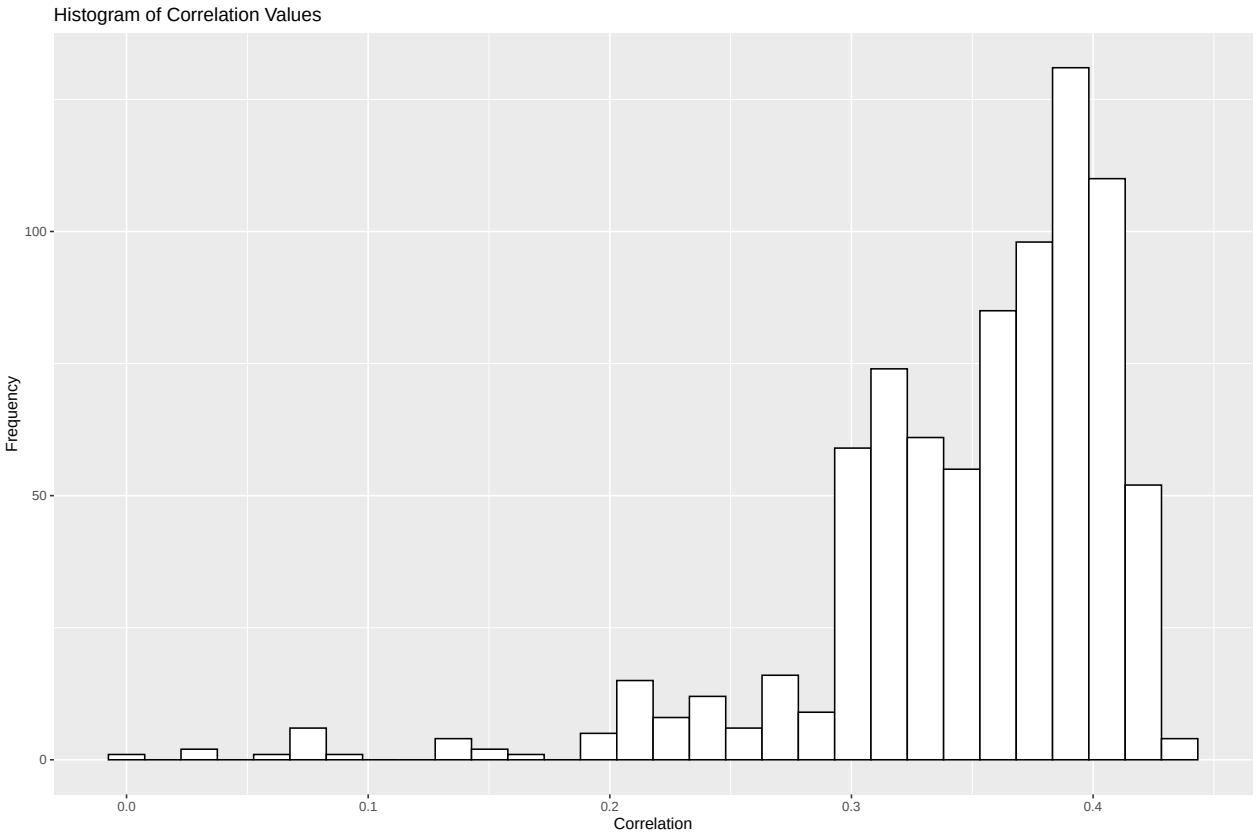
Stacking using different features in the base models

Synergistic learning

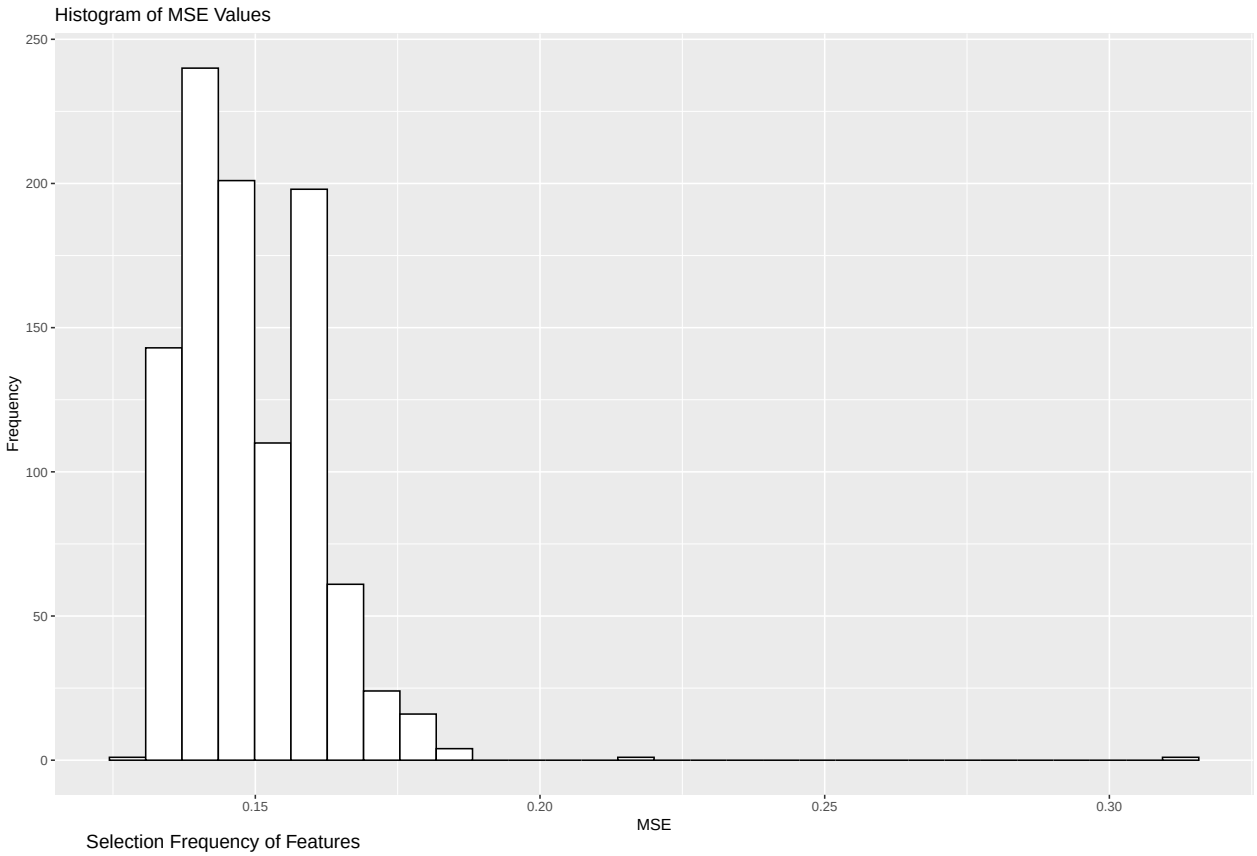
Lasso - Bootstrap

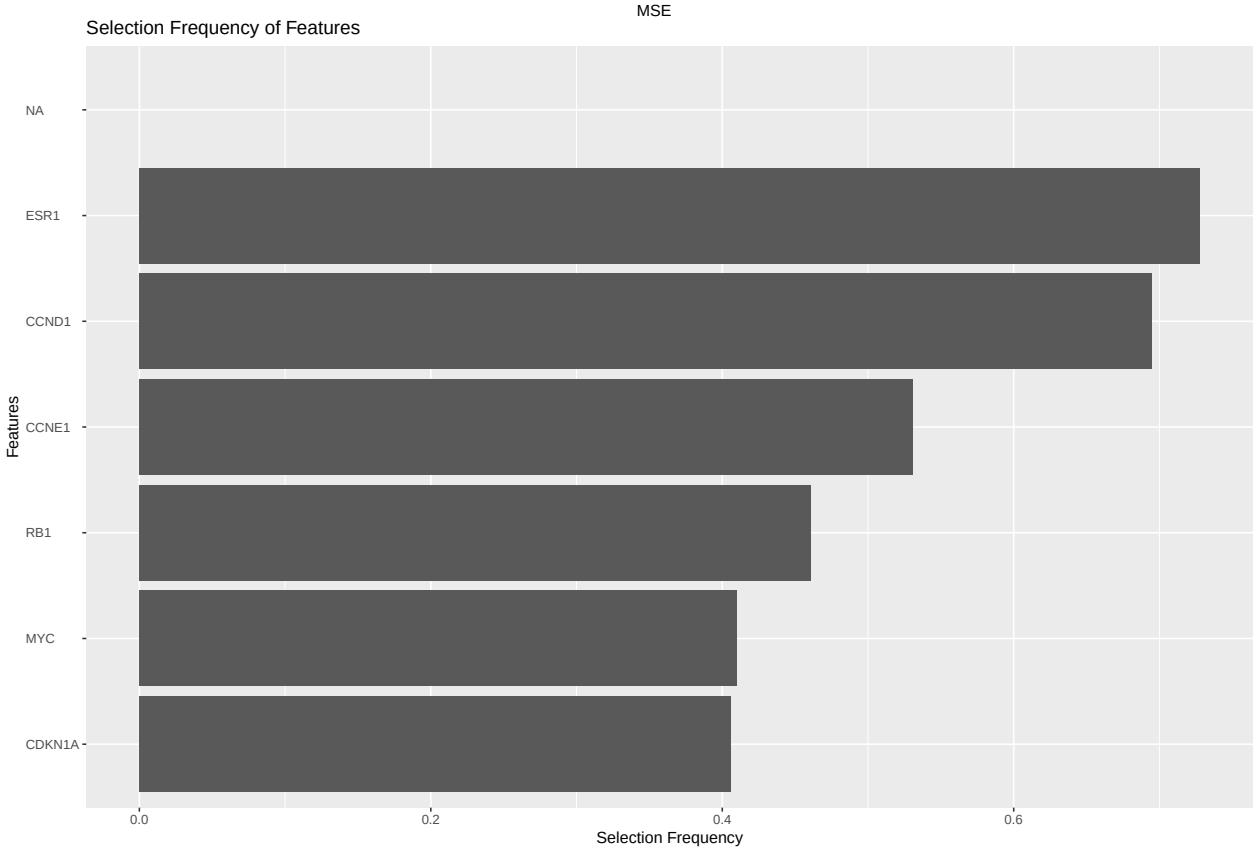
6 genes -> proliferation score

```
## 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
```


Histogram of MSE Values
Frequency
250
200
150
100
50
0
0.15
0.20
0.25
0.30
MSE

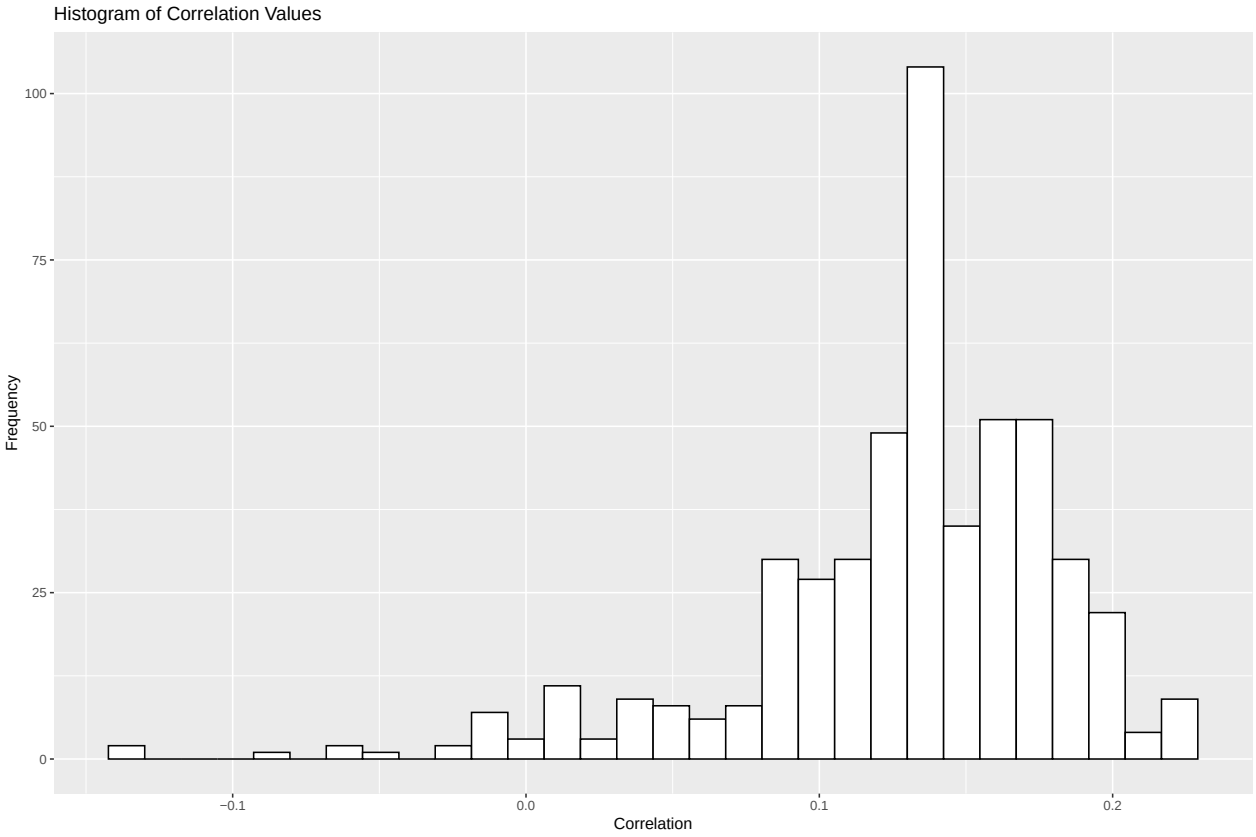

Selection Frequency of Features
Features
NA
ESR1
CCND1
CCNE1
RB1
MYC
CDKN1A
0.0
0.2
0.4
0.6
Selection Frequency

##

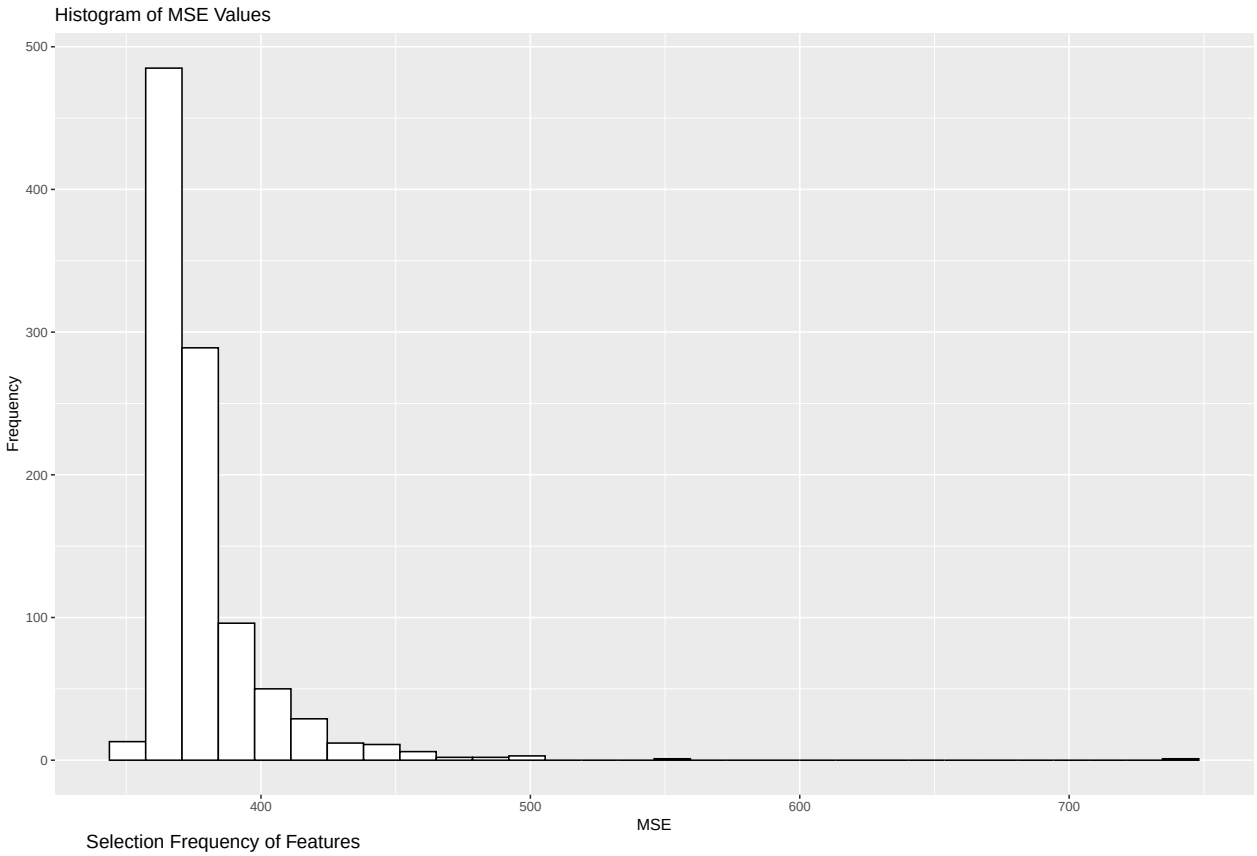
```
## Features selected 50% or more times:
## CCND1 CCNE1 ESR1
## Top 20 featrues:
##  [1] "ESR1"   "CCND1"  "CCNE1"  "RB1"    "MYC"    "CDKN1A" NA       NA
##  [9] NA       NA       NA       NA       NA       NA       NA       NA
## [17] NA       NA       NA       NA
```

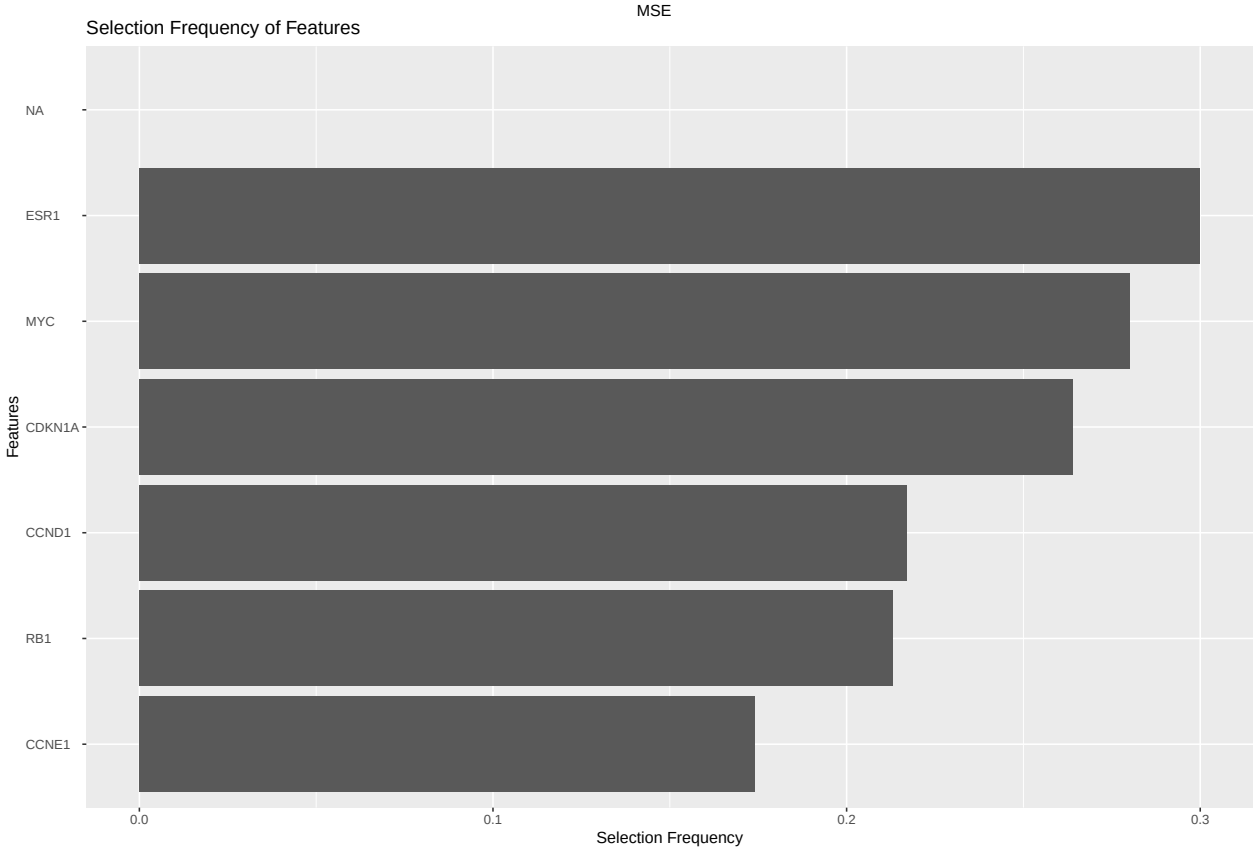
6 genes -> ROR_proliferation score

```
## 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
```


Histogram of MSE Values
Frequency
500
400
300
200
100
0
400
500
600
700
MSE

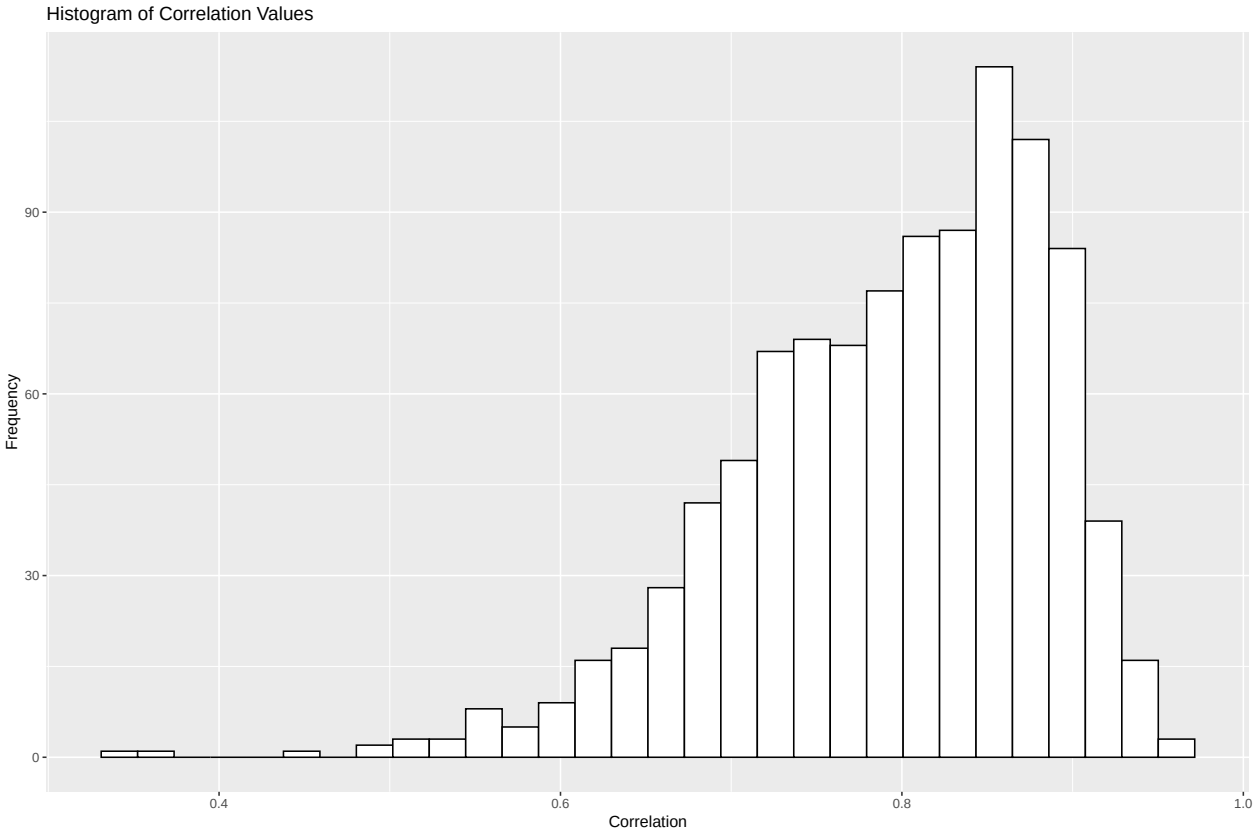

Selection Frequency of Features
Features
NA
ESR1
MYC
CDKN1A
CCND1
RB1
CCNE1
0.0
0.1
0.2
0.3
Selection Frequency

##

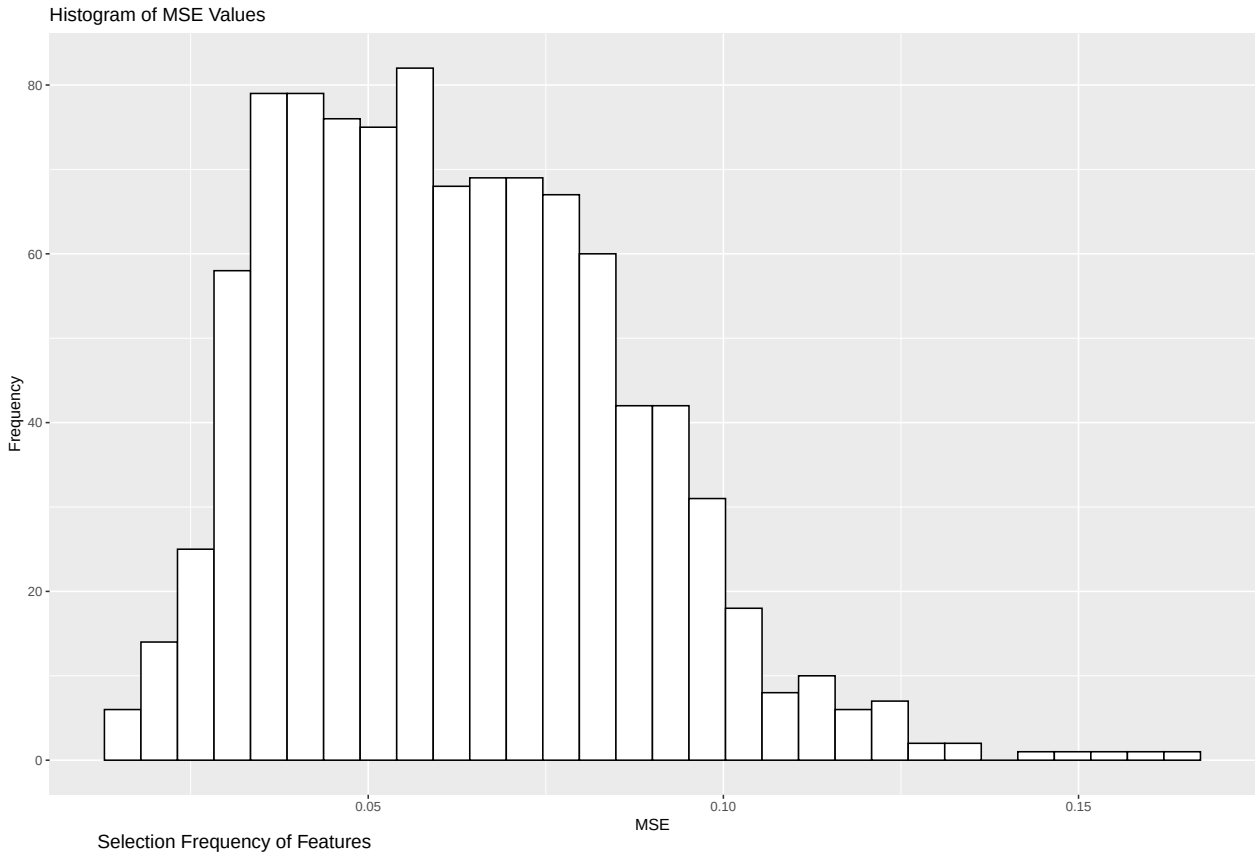
```
## Features selected 50% or more times:
##
## Top 20 featrues:
##  [1] "ESR1"   "MYC"    "CDKN1A" "CCND1"  "RB1"    "CCNE1"  NA       NA
##  [9] NA       NA       NA       NA       NA       NA       NA       NA
## [17] NA       NA       NA       NA
```

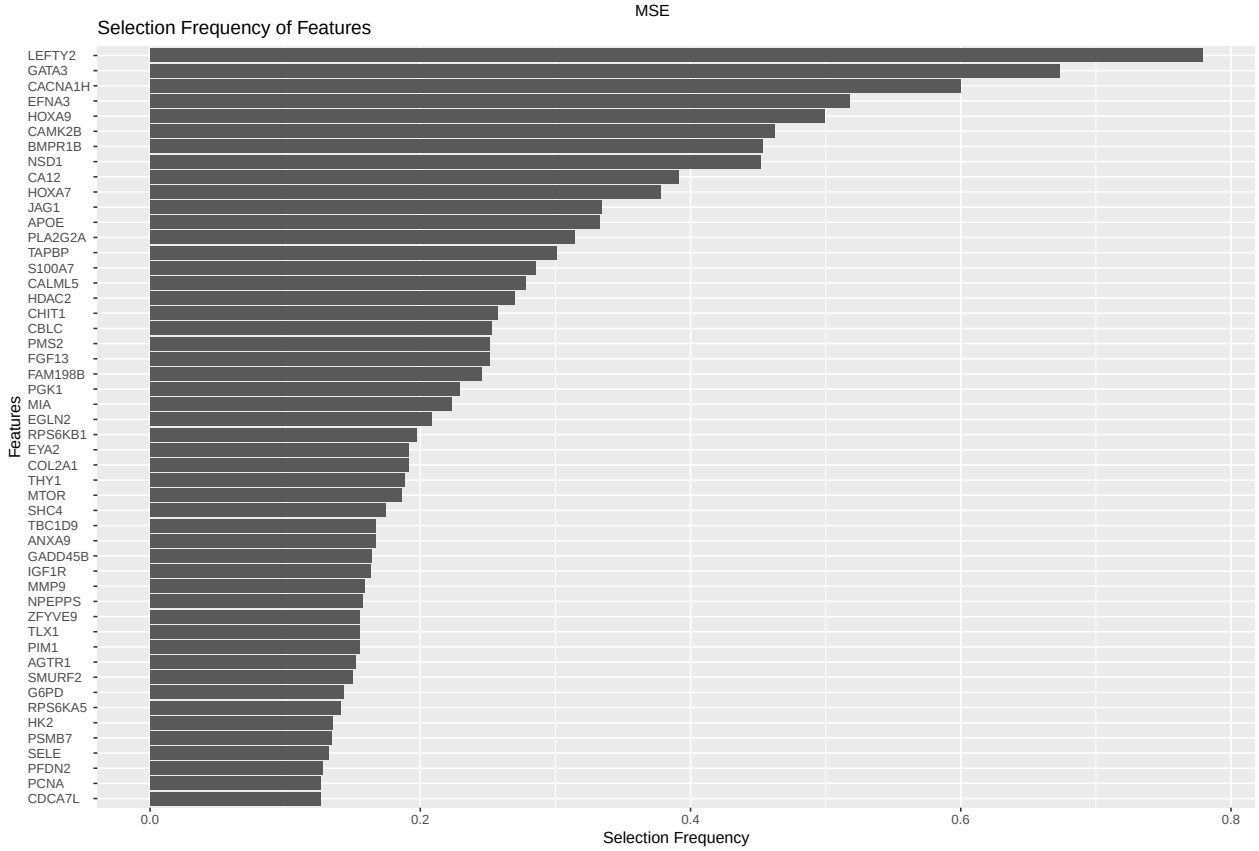
771 genes -> proliferation score

```
## 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
```




```
## MSE RESULTS
## Mean: 0.06209131
## Median: 0.0598495
## Variance: 0.0005751012
## st.dev.: 0.02398127
```


Histogram of MSE Values
Frequency
MSE
0
20
40
60
80
0.05
0.10
0.15

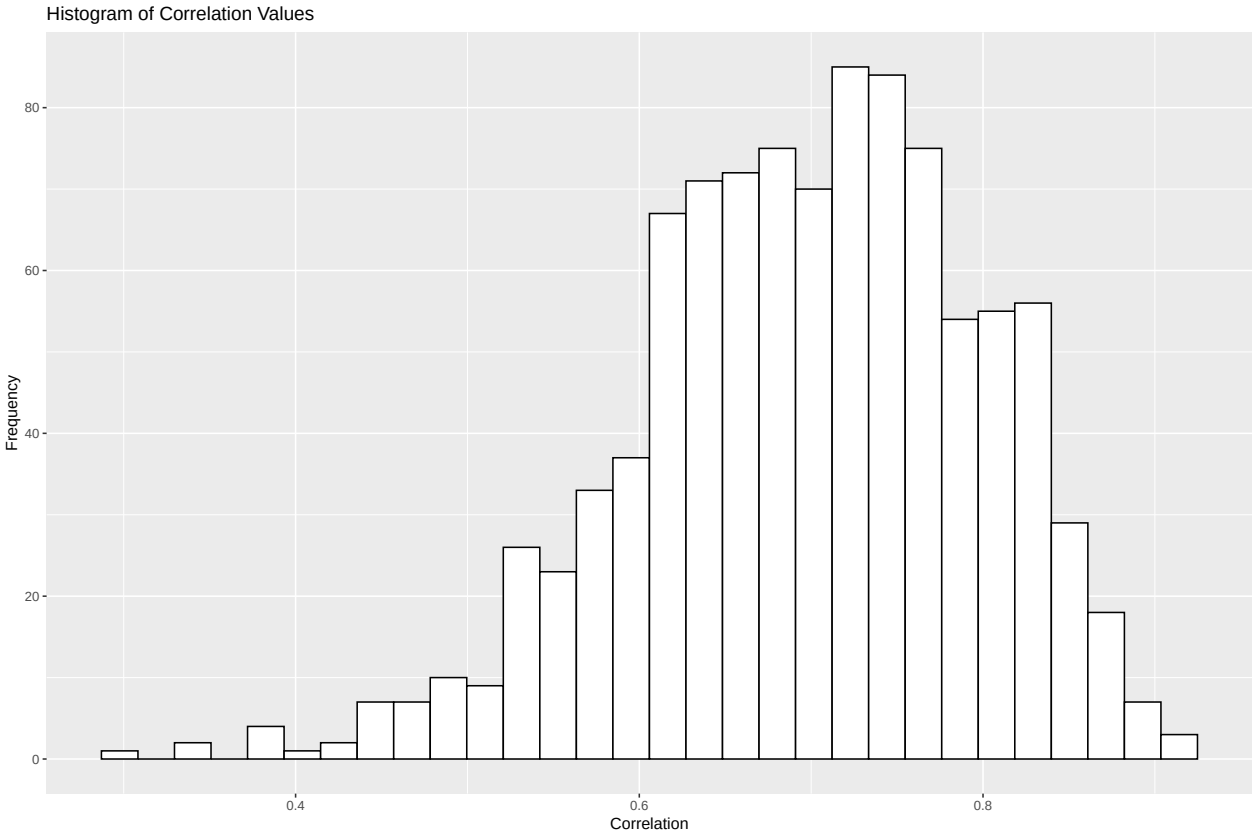

Selection Frequency of Features
Features
Selection Frequency
LEFTY2
GATA3
CACNA1H
EFNA3
HOXA9
CAMK2B
BMPR1B
NSD1
CA12
HOXA7
JAG1
APOE
PLA2G2A
TAPBP
S100A7
CALML5
HDAC2
CHIT1
CBLC
PMS2
FGF13
FAM198B
PGK1
MIA
EGLN2
RPS6KB1
EYA2
COL2A1
THY1
MTOR
SHC4
TBC1D9
ANXA9
GADD45B
IGF1R
MMP9
NPEPPS
ZFYVE9
TLX1
PIM1
AGTR1
SMURF2
G6PD
RPS6KA5
HK2
PSMB7
SELE
PFDN2
PCNA
CDCA7L
0.0
0.2
0.4
0.6
0.8

##

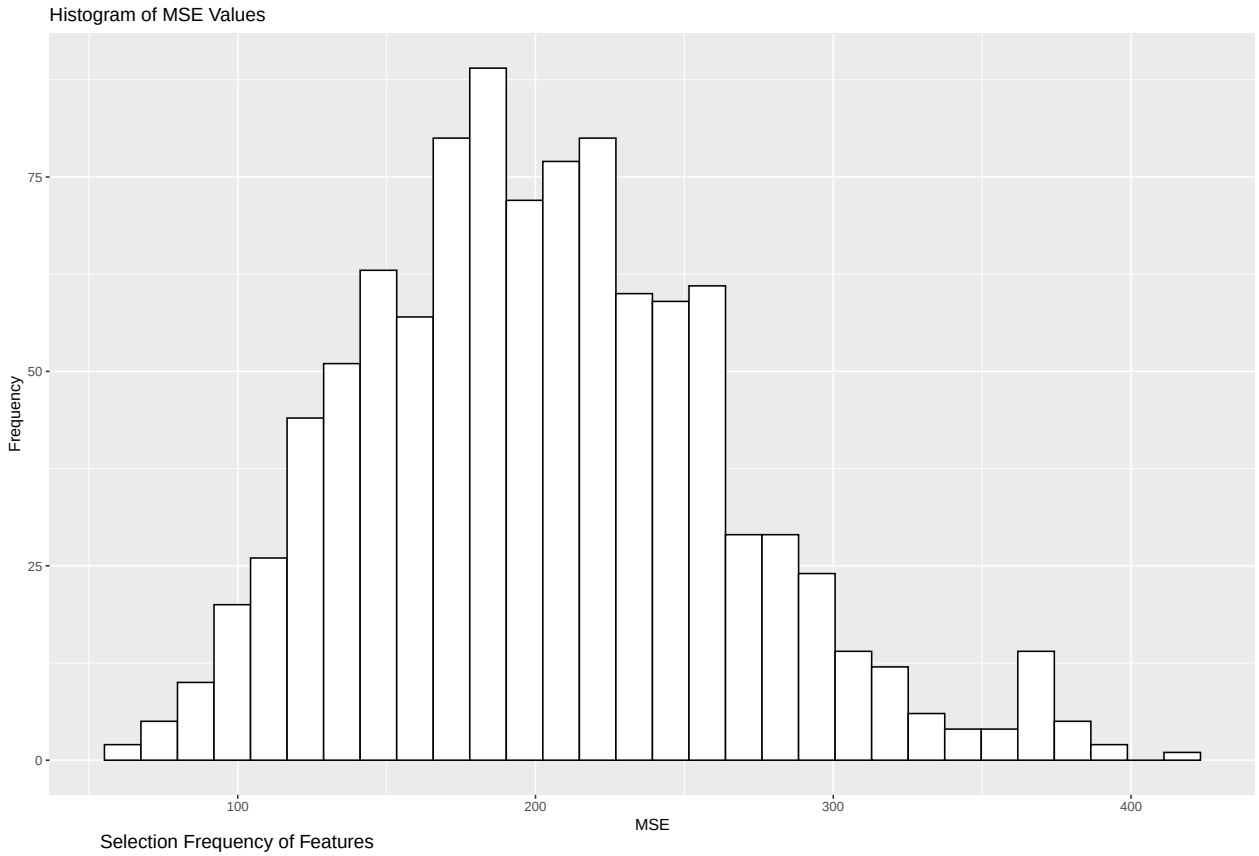
```
## Features selected 50% or more times:
## CACNA1H EFNA3 GATA3 LEFTY2
## Top 20 featrues:
##  [1] "LEFTY2"  "GATA3"   "CACNA1H" "EFNA3"   "HOXA9"   "CAMK2B"  "BMPR1B"
##  [8] "NSD1"    "CA12"    "HOXA7"   "JAG1"    "APOE"    "PLA2G2A" "TAPBP"
## [15] "S100A7"  "CALML5"  "HDAC2"   "CHIT1"   "CBLC"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

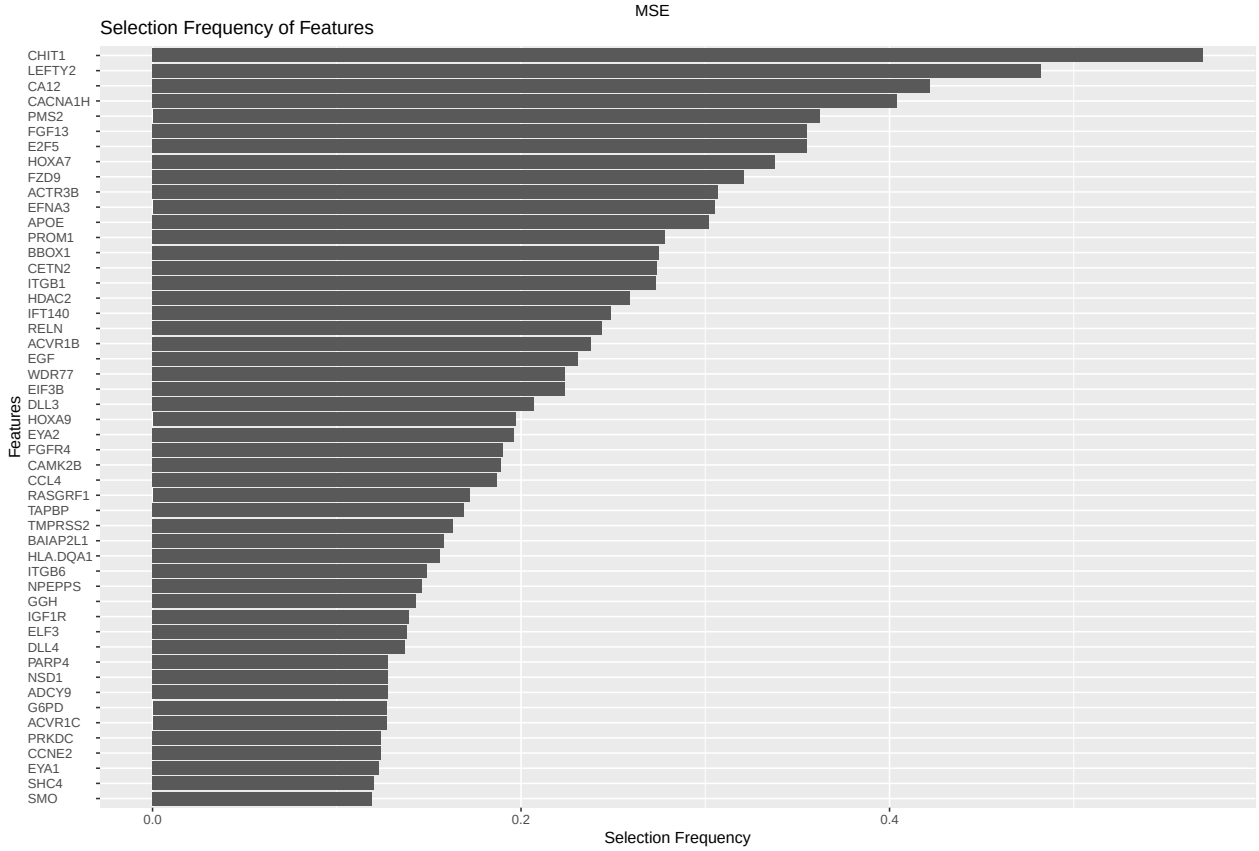
771 genes -> ROR-proliferation score

```
## 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
```


Histogram of MSE Values
Frequency
0
25
50
75
100
200
300
400
MSE

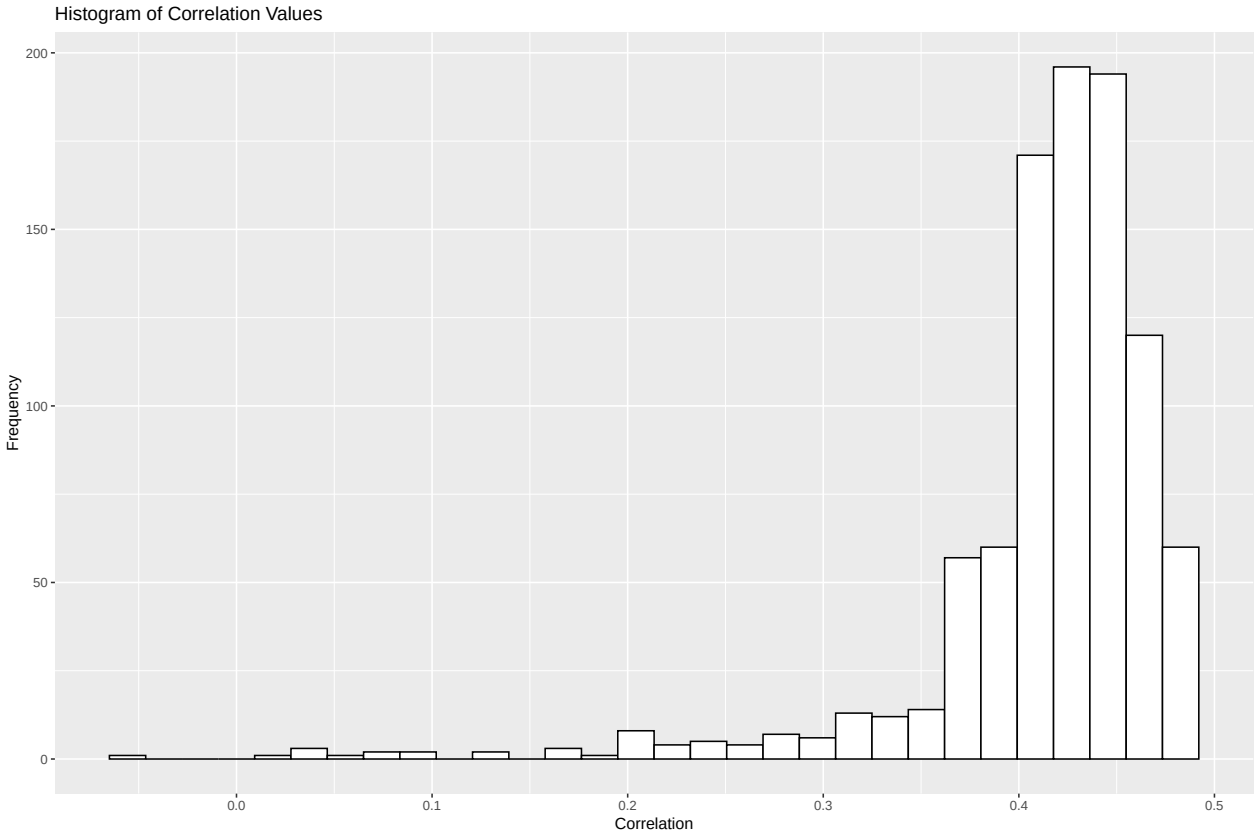

Selection Frequency of Features
Features
CHIT1
LEFTY2
CA12
CACNA1H
PMS2
FGF13
E2F5
HOXA7
FZD9
ACTR3B
EFNA3
APOE
PROM1
BBOX1
CETN2
ITGB1
HDAC2
IFT140
RELN
ACVR1B
EGF
WDR77
EIF3B
DLL3
HOXA9
EYA2
FGFR4
CAMK2B
CCL4
RASGRF1
TAPBP
TMPRSS2
BAIAP2L1
HLA.DQA1
ITGB6
NPEPPS
GGH
IGF1R
ELF3
DLL4
PARP4
NSD1
ADCY9
G6PD
ACVR1C
PRKDC
CCNE2
EYA1
SHC4
SMO
0.0
0.2
0.4
Selection Frequency

##

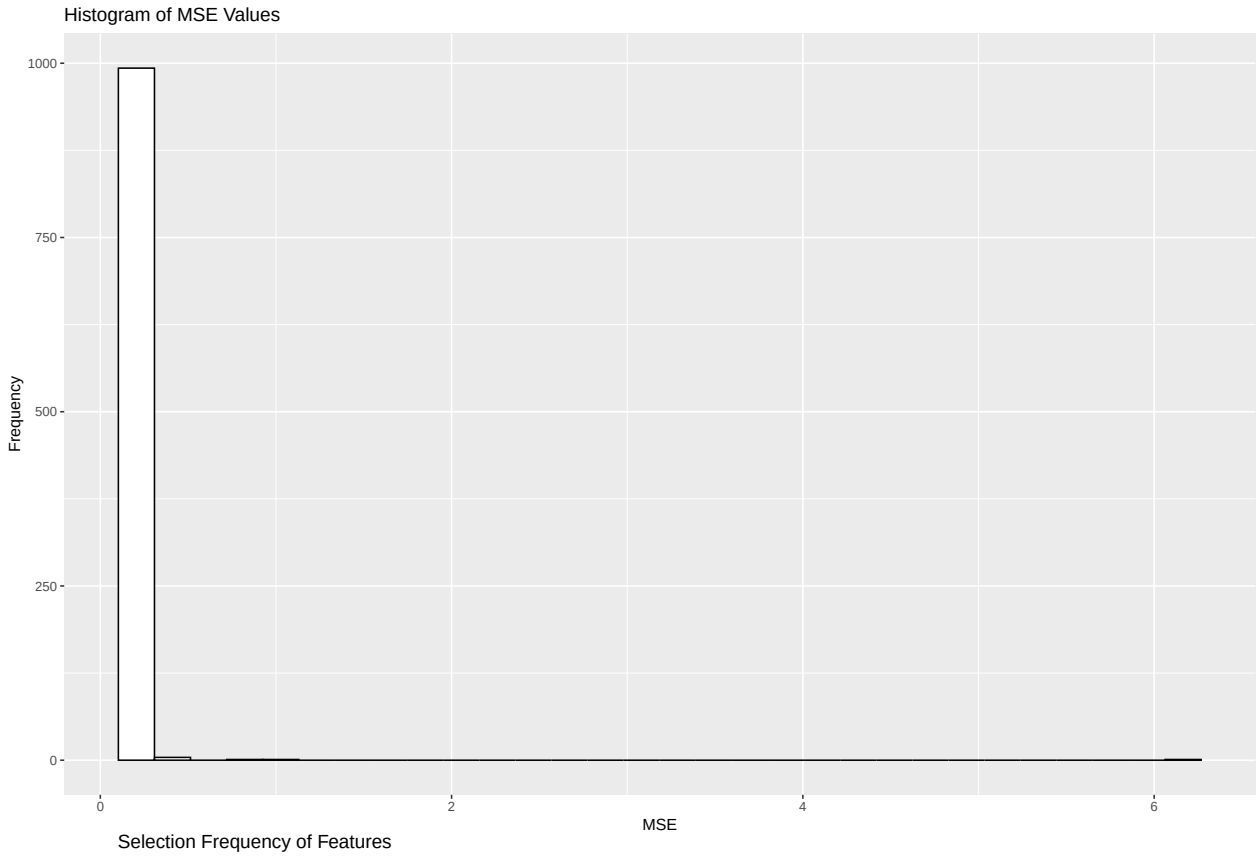
```
## Features selected 50% or more times:
## CHIT1
## Top 20 featrues:
##  [1] "CHIT1"   "LEFTY2"  "CA12"    "CACNA1H" "PMS2"    "E2F5"    "FGF13"
##  [8] "HOXA7"   "FZD9"    "ACTR3B"  "EFNA3"   "APOE"    "PROM1"   "BBOX1"
## [15] "CETN2"   "ITGB1"   "HDAC2"   "IFT140"  "RELN"    "ACVR1B"
```

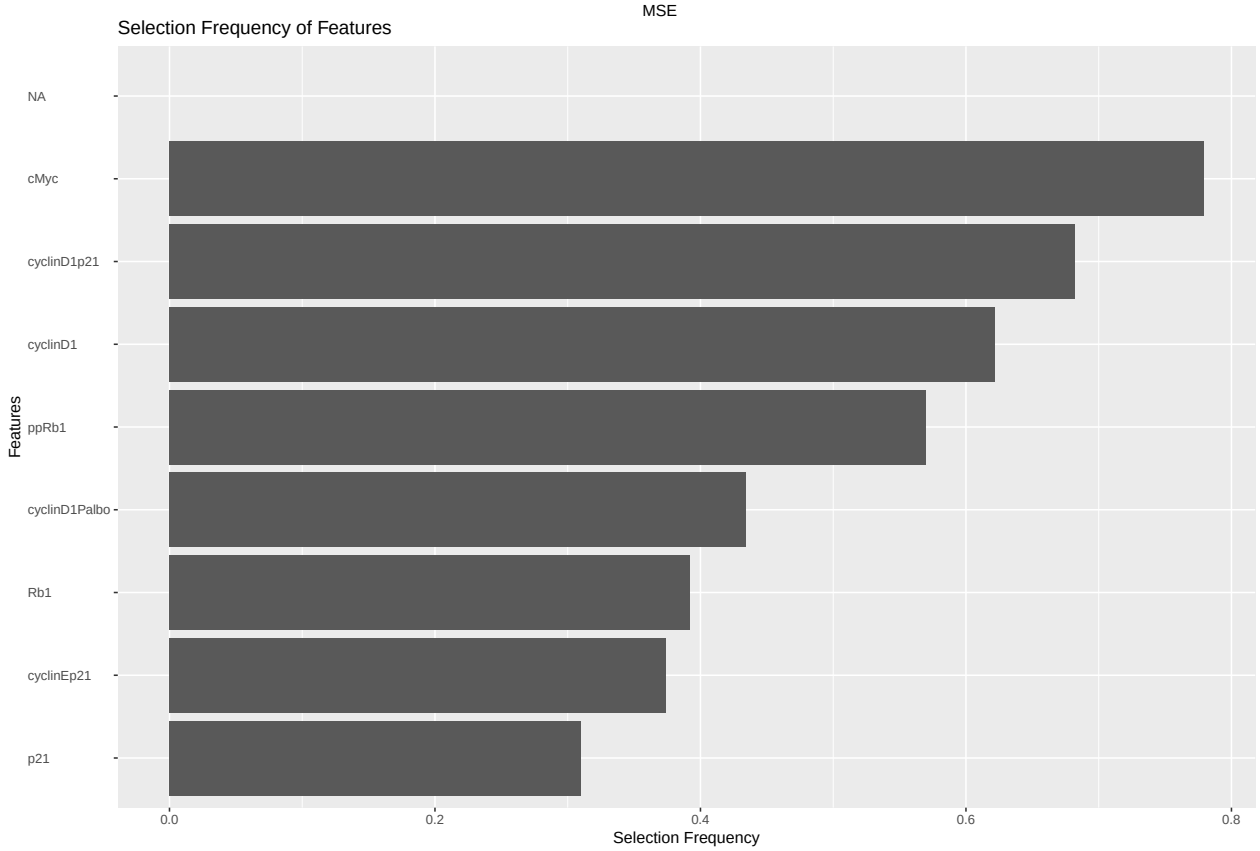
node values -> proliferation score

```
## 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
```




```
## MSE RESULTS
## Mean: 0.1479731
## Median: 0.1355805
## Variance: 0.03673469
## st.dev.: 0.191663
```


Histogram of MSE Values
Frequency
1000
750
500
250
0
0
2
4
6
MSE

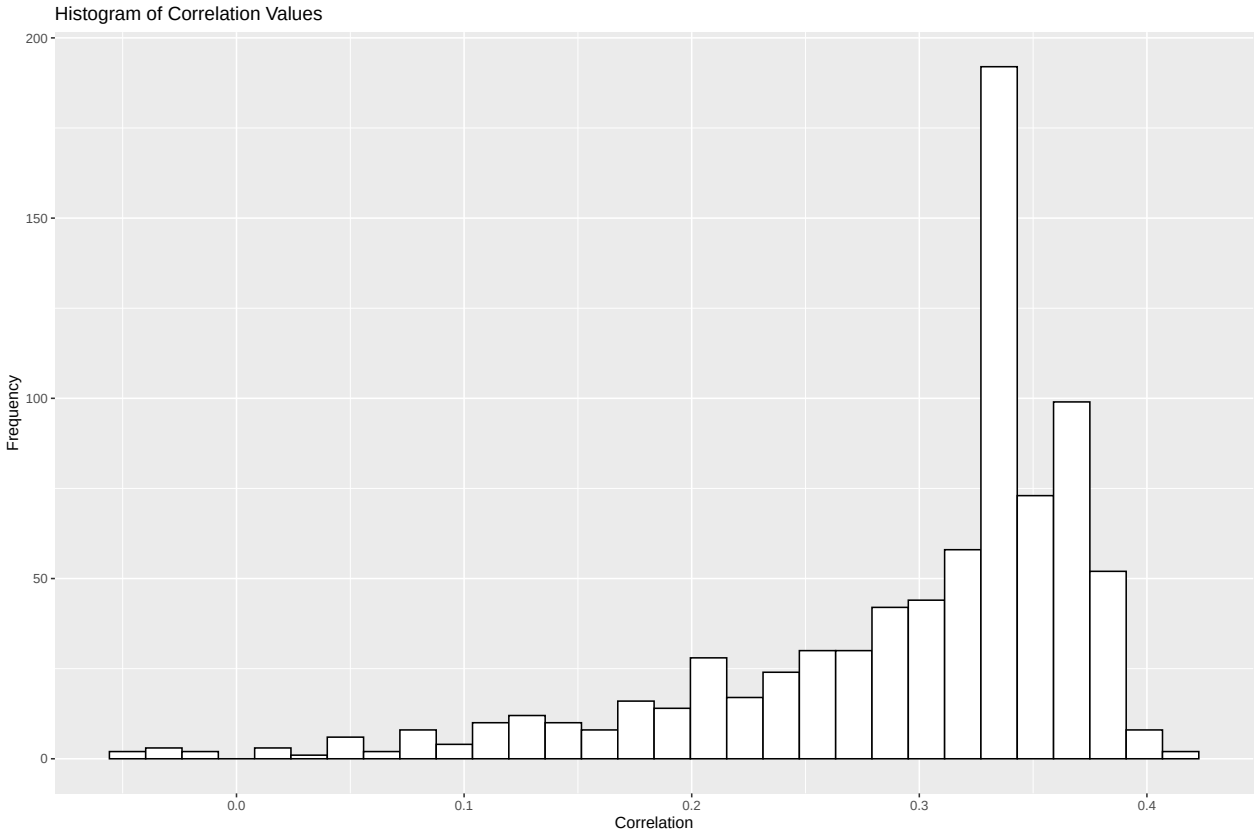

Selection Frequency of Features
Features
NA
cMyc
cyclinD1p21
cyclinD1
ppRb1
cyclinD1Palbo
Rb1
cyclinEp21
p21
0.0
0.2
0.4
0.6
0.8
Selection Frequency

##

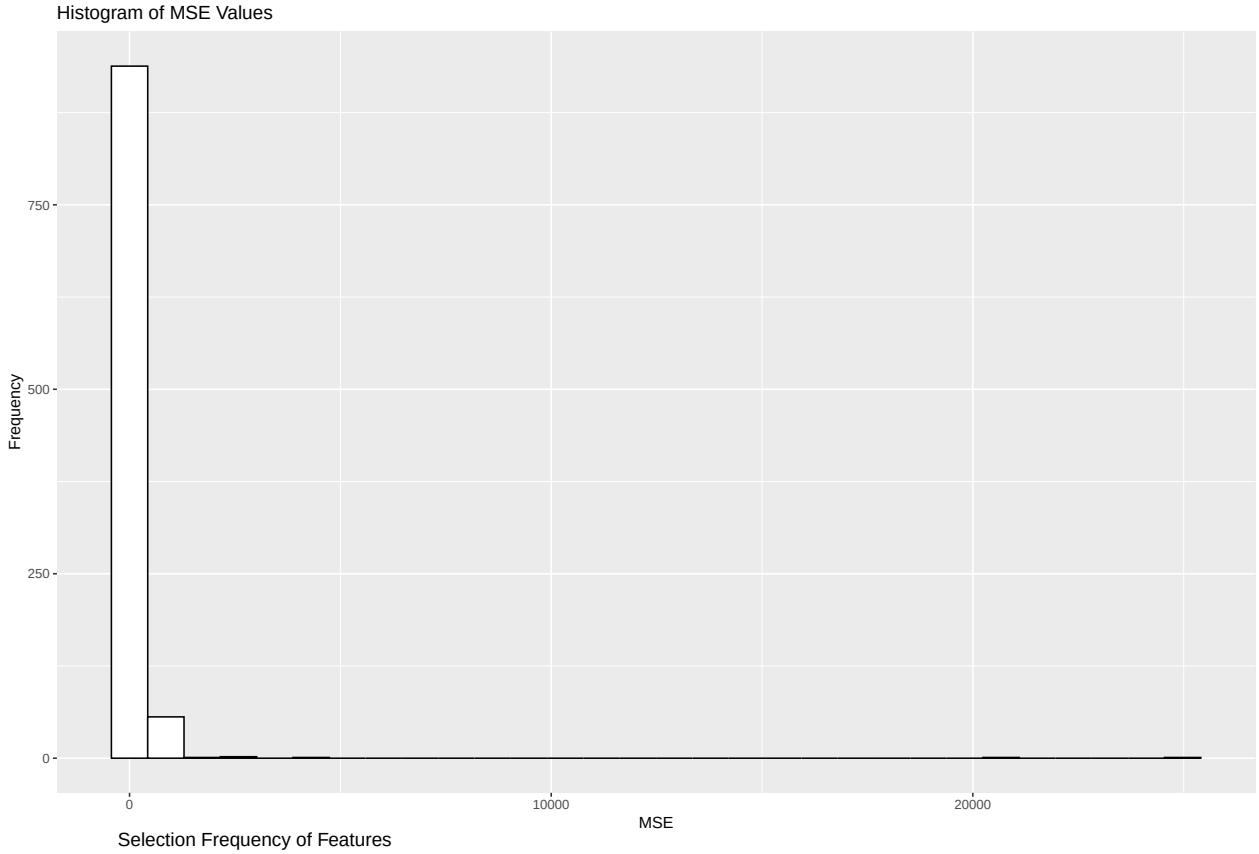
```
## Features selected 50% or more times:
## cyclinD1 cyclinD1p21 cMyc ppRb1
## Top 20 featrues:
##  [1] "cMyc"          "cyclinD1p21"   "cyclinD1"      "ppRb1"
##  [5] "cyclinD1Palbo" "Rb1"           "cyclinEp21"    "p21"
##  [9] NA              NA              NA              NA
## [13] NA      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

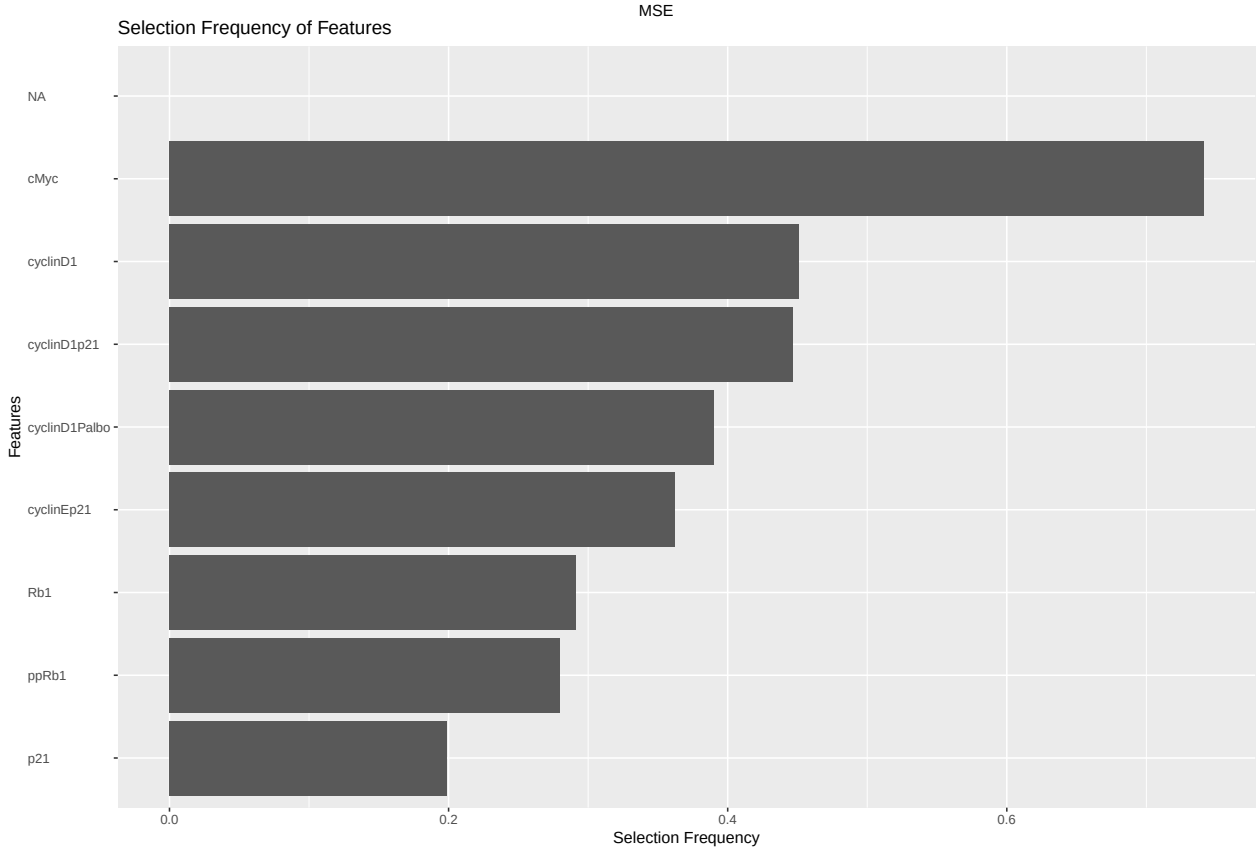
node values -> ROR-proliferation score

```
## 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
```


Histogram of MSE Values
Frequency
750
500
250
0
0
10000
20000
MSE

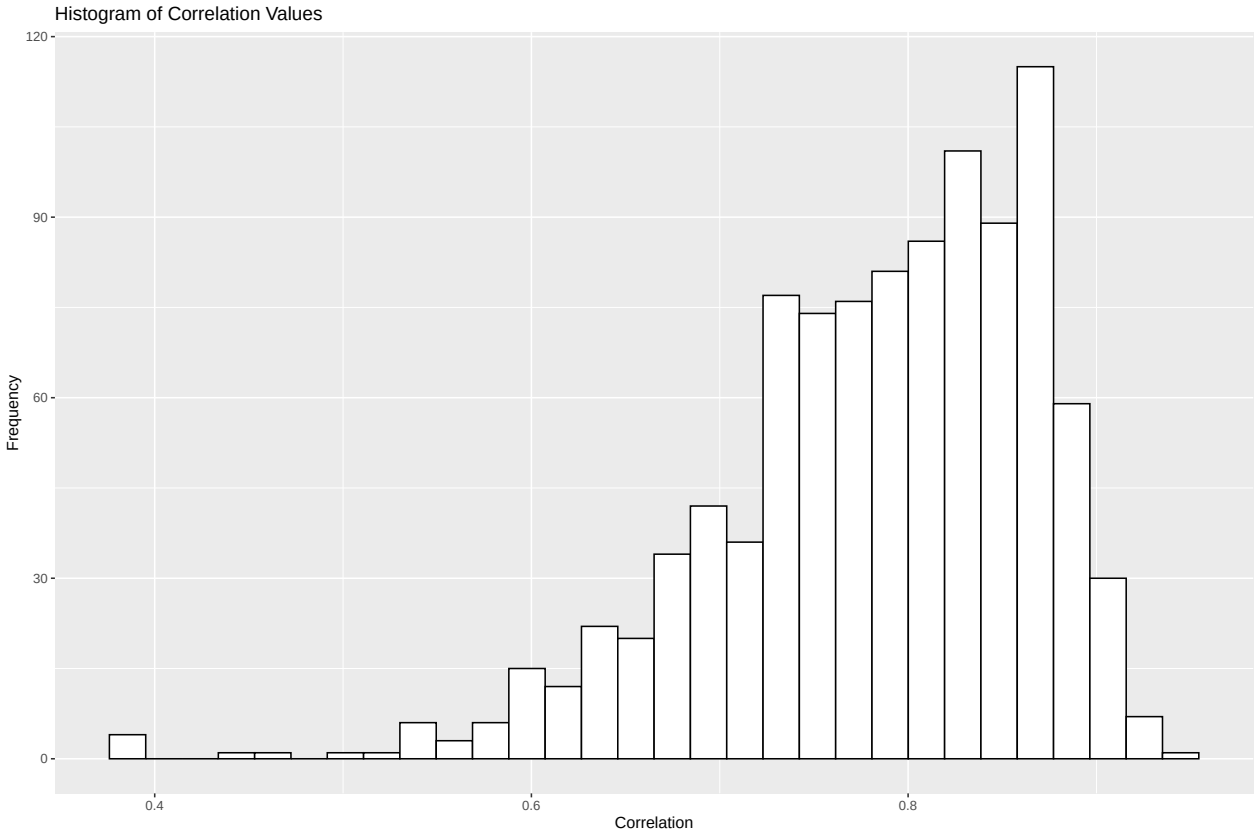

Selection Frequency of Features
Features
NA
cMyc
cyclinD1
cyclinD1p21
cyclinD1Palbo
cyclinEp21
Rb1
ppRb1
p21
0.0
0.2
0.4
0.6
Selection Frequency

##

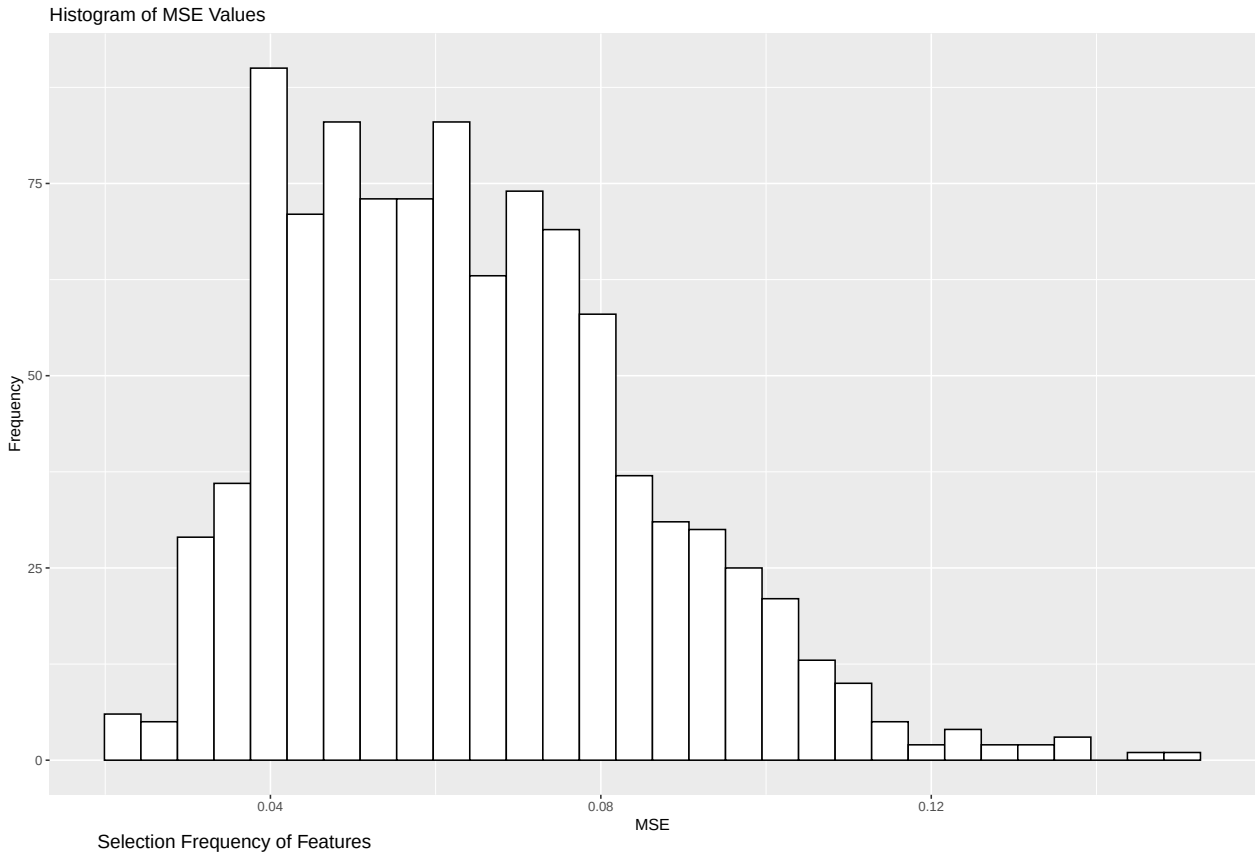
```
## Features selected 50% or more times:
## cMyc
## Top 20 featrues:
##  [1] "cMyc"          "cyclinD1"      "cyclinD1p21"   "cyclinD1Palbo"
##  [5] "cyclinEp21"    "Rb1"           "ppRb1"         "p21"
##  [9] NA              NA              NA              NA
## [13] NA              NA         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

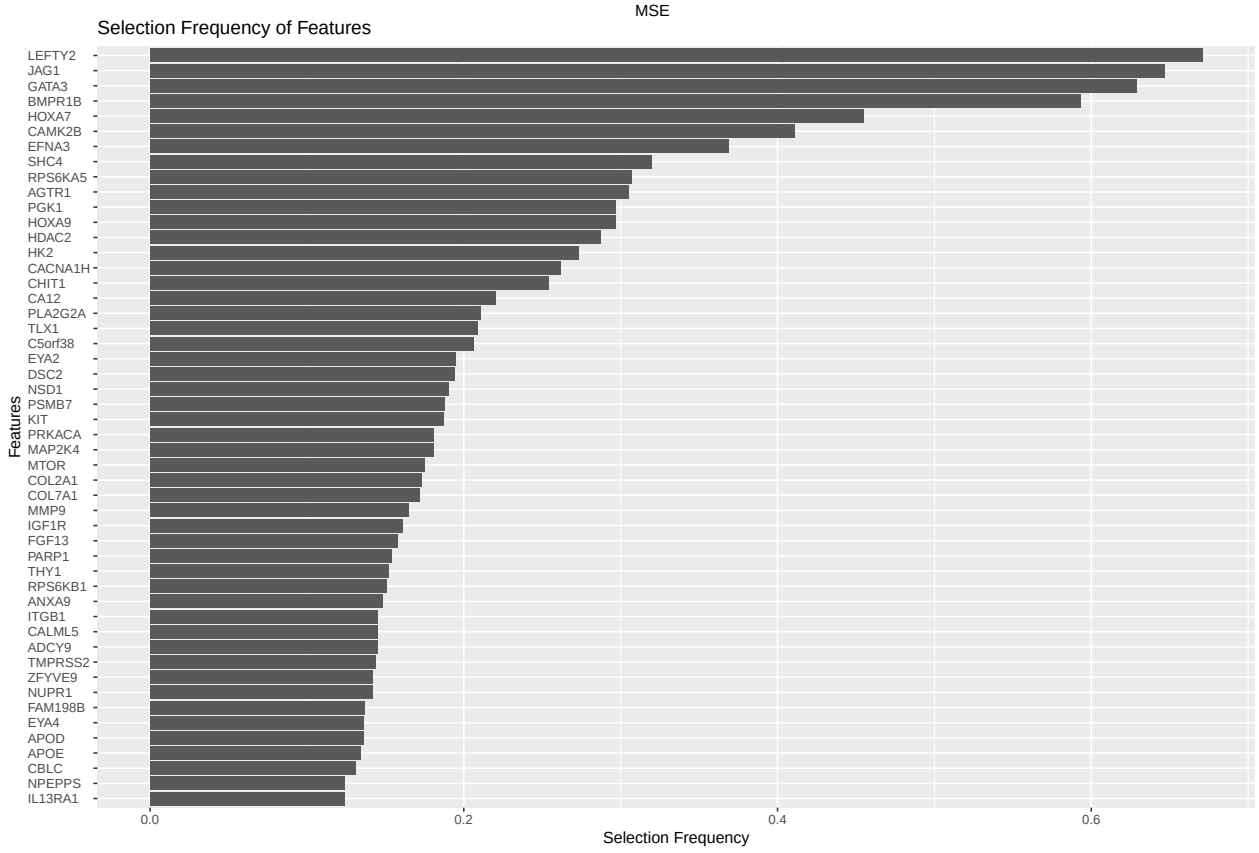
Mechanistic + Residuals -> proliferation score (additive)

```
## 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
```


Histogram of MSE Values
Frequency
75
50
25
0
0.04
0.08
0.12
MSE

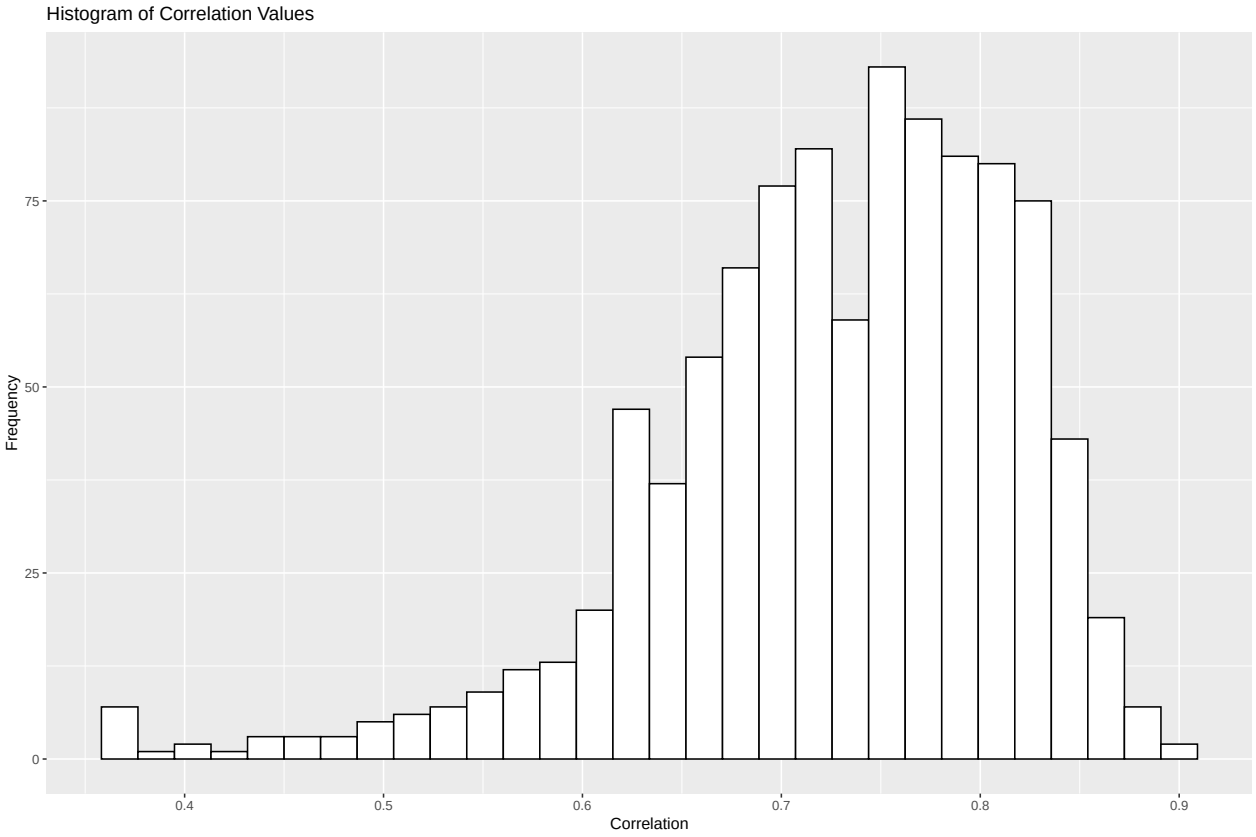

Selection Frequency of Features
Features
LEFTY2
JAG1
GATA3
BMPR1B
HOXA7
CAMK2B
EFNA3
SHC4
RPS6KA5
AGTR1
PGK1
HOXA9
HDAC2
HK2
CACNA1H
CHIT1
CA12
PLA2G2A
TLX1
C5orf38
EYA2
DSC2
NSD1
PSMB7
KIT
PRKACA
MAP2K4
MTOR
COL2A1
COL7A1
MMP9
IGF1R
FGF13
PARP1
THY1
RPS6KB1
ANXA9
ITGB1
CALML5
ADCY9
TMPRSS2
ZFYVE9
NUPR1
FAM198B
EYA4
APOD
APOE
CBLC
NPEPPS
IL13RA1
0.0
0.2
0.4
0.6
Selection Frequency

##

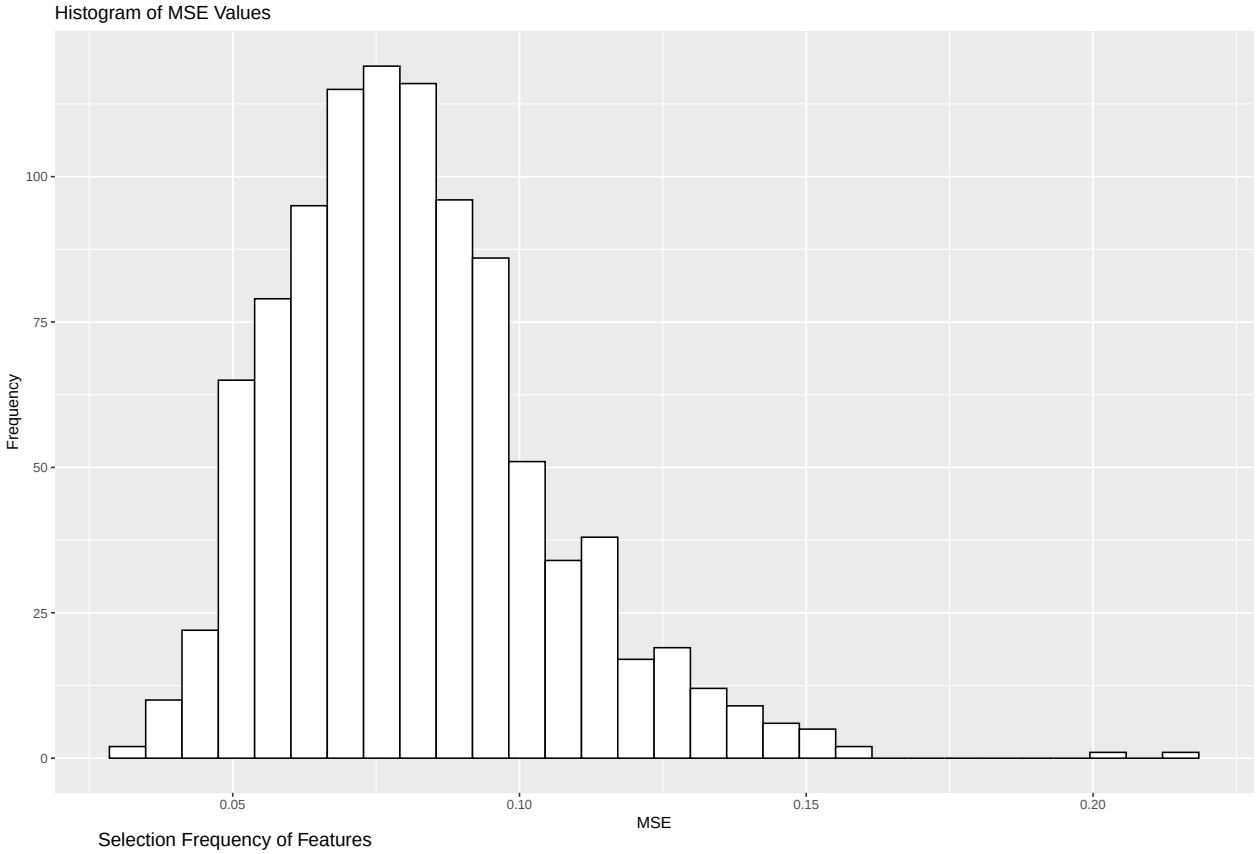
```
## Features selected 50% or more times:
## BMPR1B GATA3 JAG1 LEFTY2
## Top 20 featrues:
##  [1] "LEFTY2"  "JAG1"    "GATA3"   "BMPR1B"  "HOXA7"   "CAMK2B"  "EFNA3"
##  [8] "SHC4"    "RPS6KA5" "AGTR1"   "HOXA9"   "PGK1"    "HDAC2"   "HK2"
## [15] "CACNA1H" "CHIT1"   "CA12"    "PLA2G2A" "TLX1"    "C5orf38"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

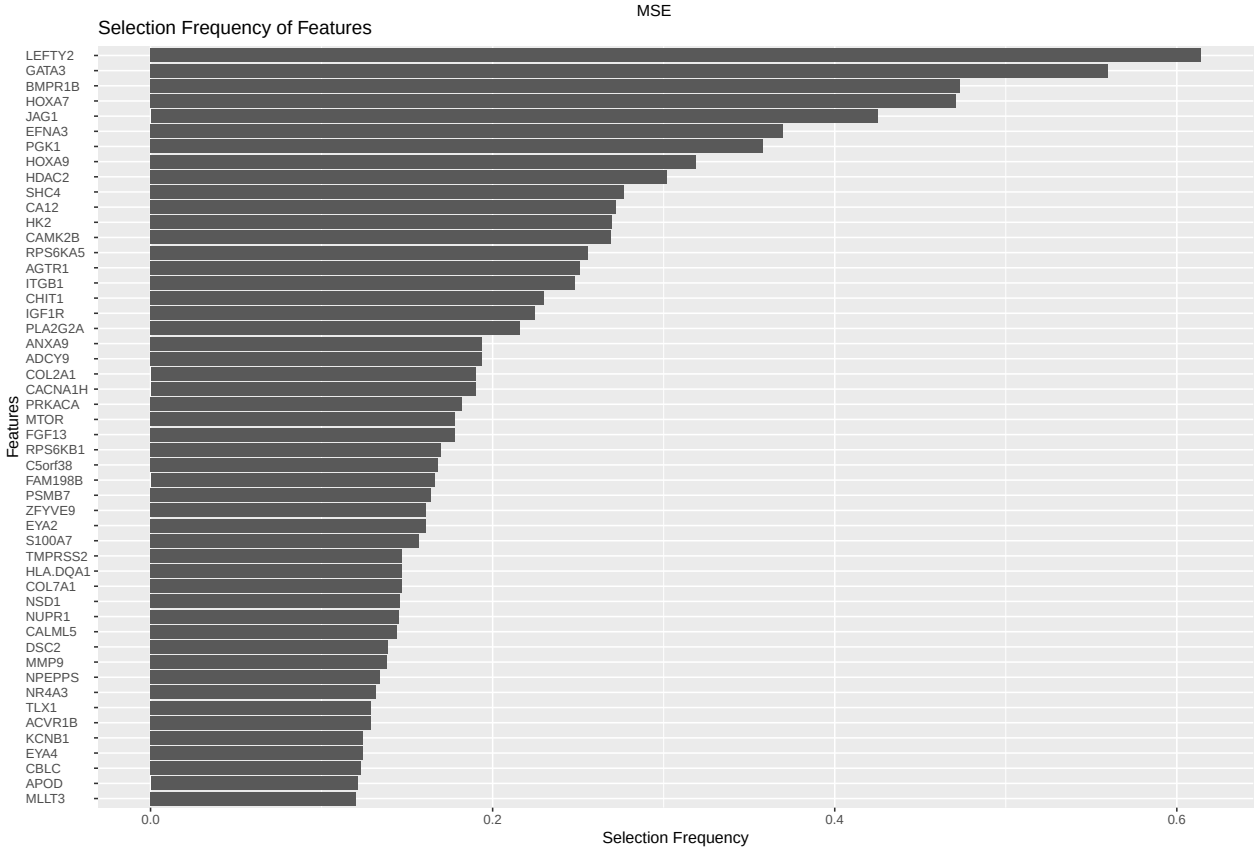
Mechanistic + Residuals -> proliferation score (multiplicative)

```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0
##
## CORRELATIONS RESULTS
## Mean: 0.7266736
## Median: 0.739984
## Variance: 0.008014481
## st.dev.: 0.08952363
```




```
## MSE RESULTS
## Mean: 0.0813415
## Median: 0.07892236
## Variance: 0.0005355891
## st.dev.: 0.0231428
```


Histogram of MSE Values
Frequency
100
75
50
25
0
0.05
0.10
0.15
0.20
MSE

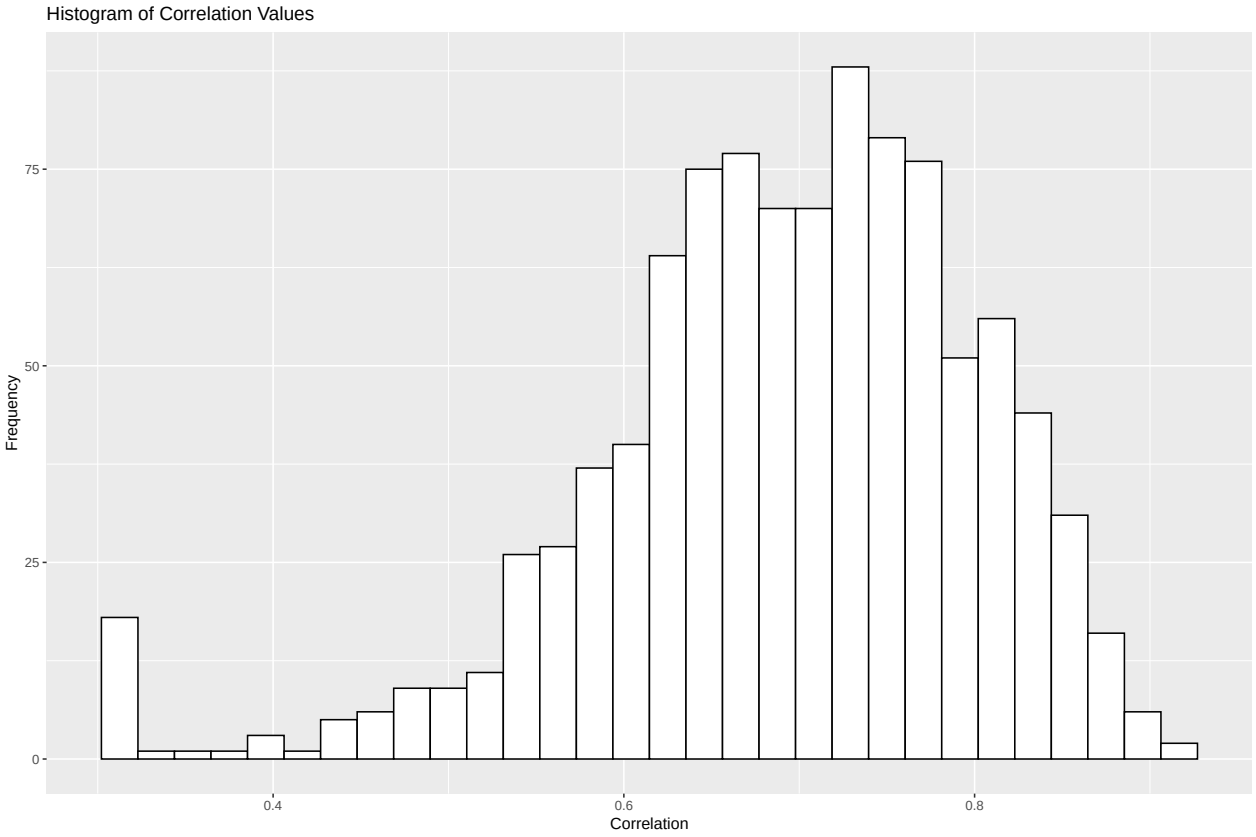

Selection Frequency of Features
Features
LEFTY2
GATA3
BMPR1B
HOXA7
JAG1
EFNA3
PGK1
HOXA9
HDAC2
SHC4
CA12
HK2
CAMK2B
RPS6KA5
AGTR1
ITGB1
CHIT1
IGF1R
PLA2G2A
ANXA9
ADCY9
COL2A1
CACNA1H
PRKACA
MTOR
FGF13
RPS6KB1
C5orf38
FAM198B
PSMB7
ZFYVE9
EYA2
S100A7
TMPRSS2
HLA.DQA1
COL7A1
NSD1
NUPR1
CALML5
DSC2
MMP9
NPEPPS
NR4A3
TLX1
ACVR1B
KCNB1
EYA4
CBLC
APOD
MLLT3
0.0
0.2
0.4
0.6
Selection Frequency

##

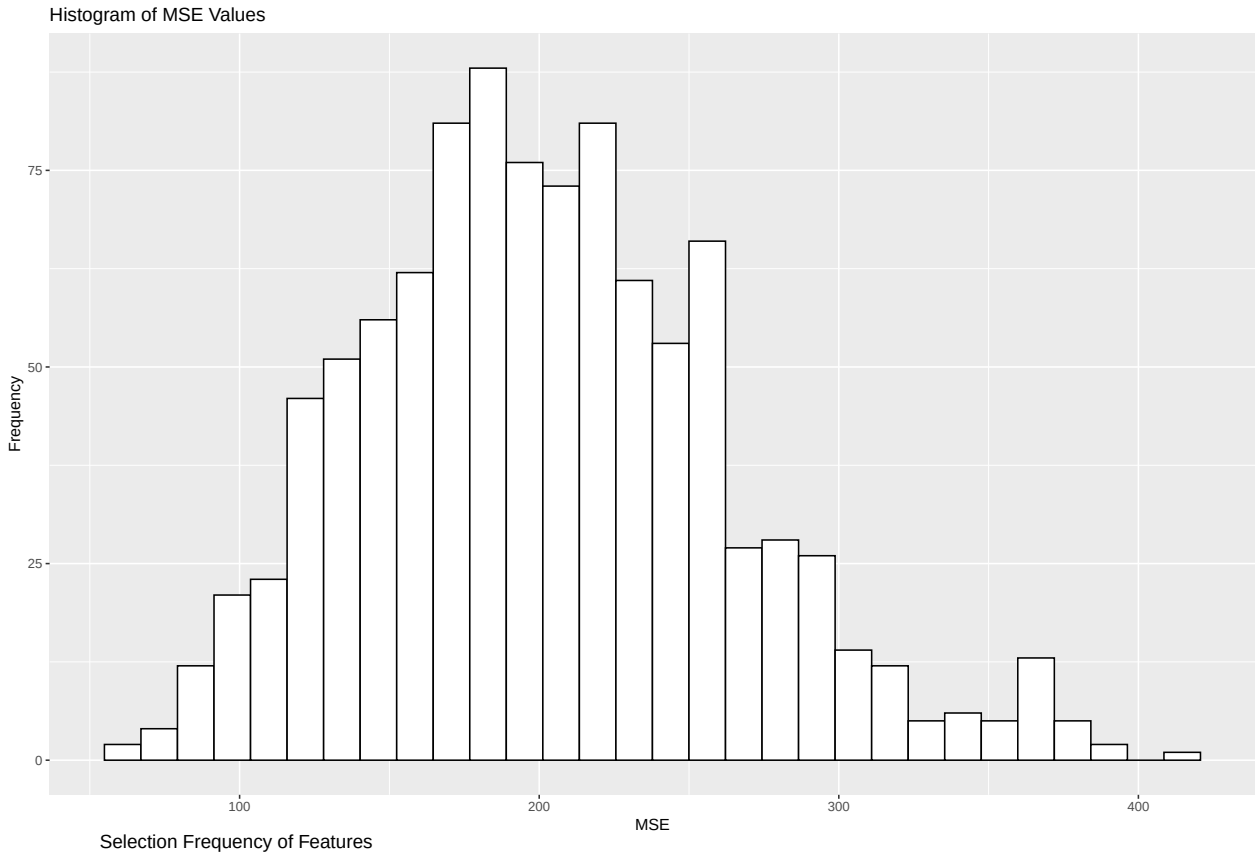
```
## Features selected 50% or more times:
## GATA3 LEFTY2
## Top 20 featrues:
##  [1] "LEFTY2"  "GATA3"   "BMPR1B"  "HOXA7"   "JAG1"    "EFNA3"   "PGK1"
##  [8] "HOXA9"   "HDAC2"   "SHC4"    "CA12"    "HK2"     "CAMK2B"  "RPS6KA5"
## [15] "AGTR1"   "ITGB1"   "CHIT1"   "IGF1R"   "PLA2G2A" "ADCY9"
```

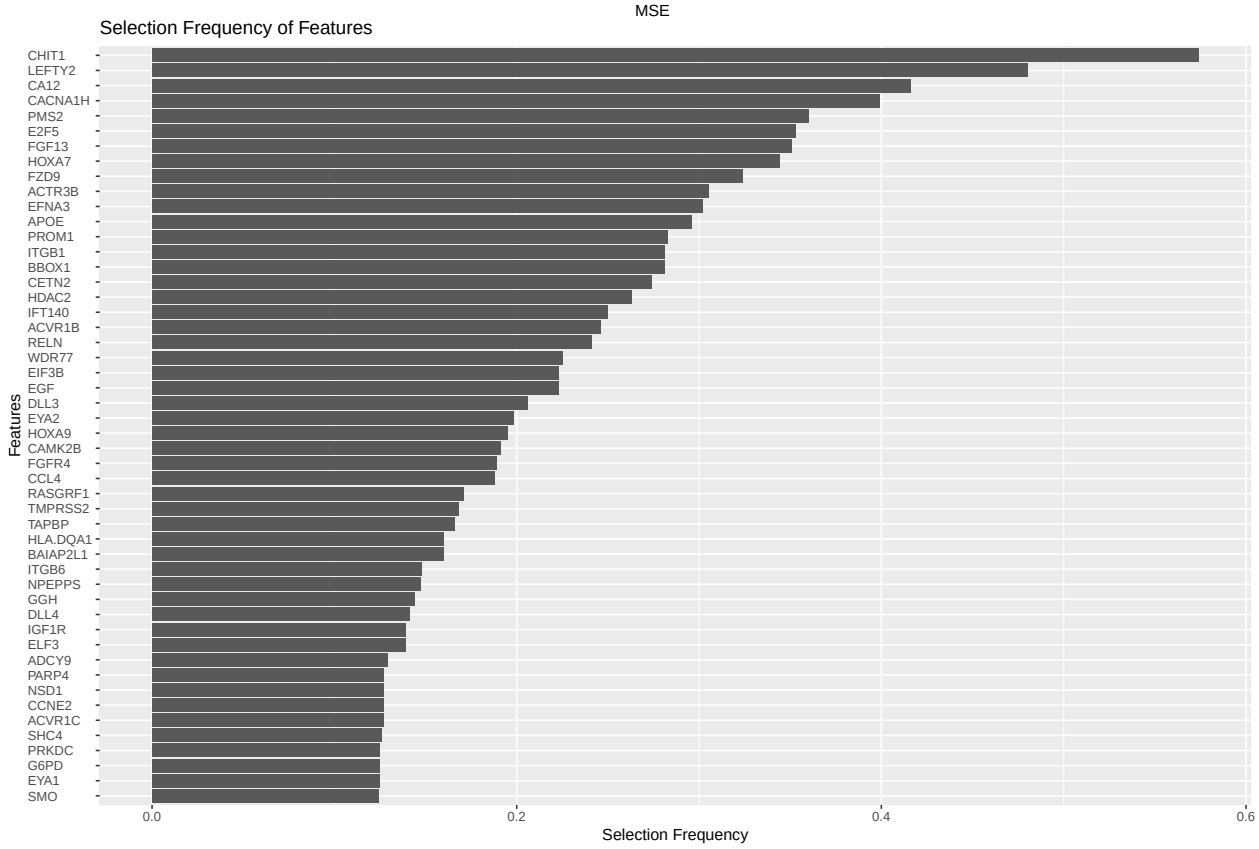
Mechnaistic + Residuals -> ROR-proliferation score (additive)

```
## 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
Frequency
0
25
50
75
100
200
300
400
MSE

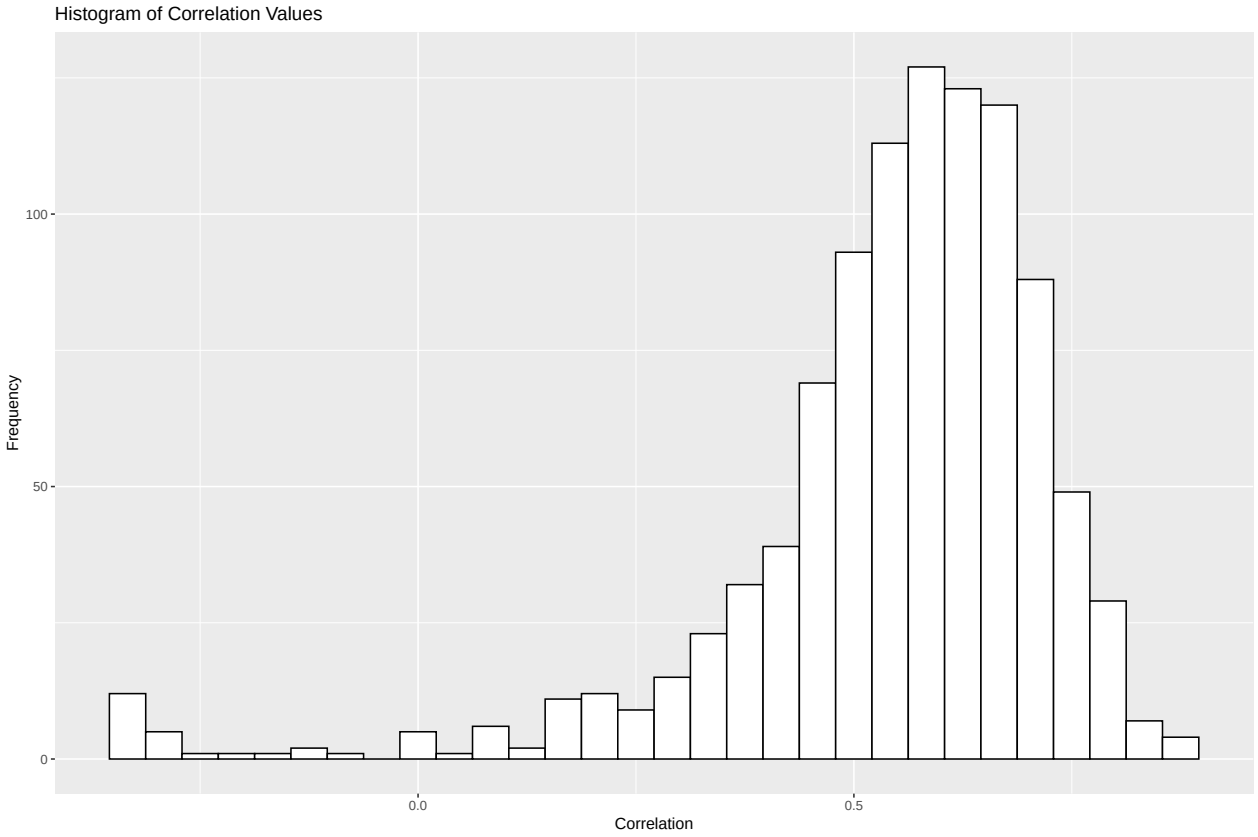

Selection Frequency of Features
Features
CHIT1
LEFTY2
CA12
CACNA1H
PMS2
E2F5
FGF13
HOXA7
FZD9
ACTR3B
EFNA3
APOE
PROM1
ITGB1
BBOX1
CETN2
HDAC2
IFT140
ACVR1B
RELN
WDR77
EIF3B
EGF
DLL3
EYA2
HOXA9
CAMK2B
FGFR4
CCL4
RASGRF1
TMPRSS2
TAPBP
HLA.DQA1
BAIAP2L1
ITGB6
NPEPPS
GGH
DLL4
IGF1R
ELF3
ADCY9
PARP4
NSD1
CCNE2
ACVR1C
SHC4
PRKDC
G6PD
EYA1
SMO
0.0
0.2
0.4
0.6
Selection Frequency

##

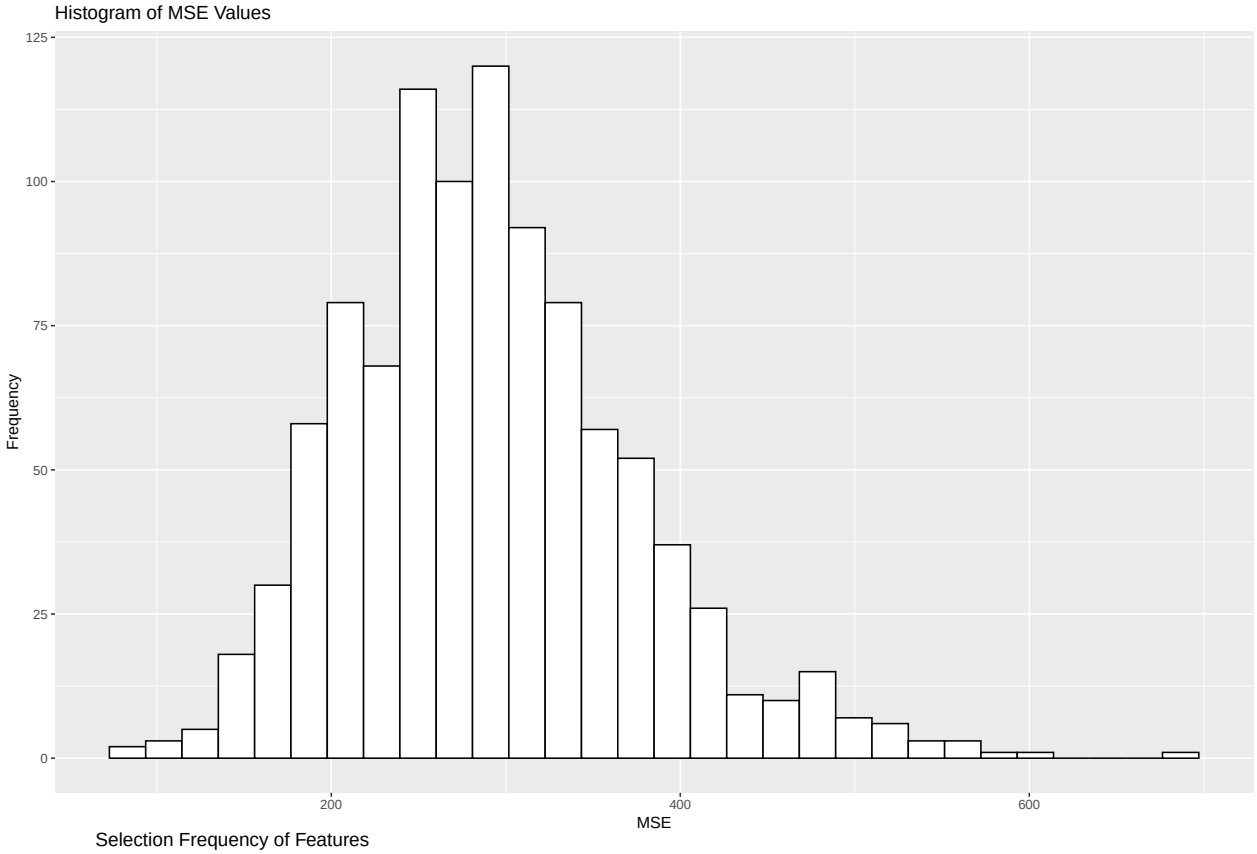
```
## Features selected 50% or more times:
## CHIT1
## Top 20 featrues:
##  [1] "CHIT1"   "LEFTY2"  "CA12"    "CACNA1H" "PMS2"    "E2F5"    "FGF13"
##  [8] "HOXA7"   "FZD9"    "ACTR3B"  "EFNA3"   "APOE"    "PROM1"   "BBOX1"
## [15] "ITGB1"   "CETN2"   "HDAC2"   "IFT140"  "ACVR1B"  "RELN"
```

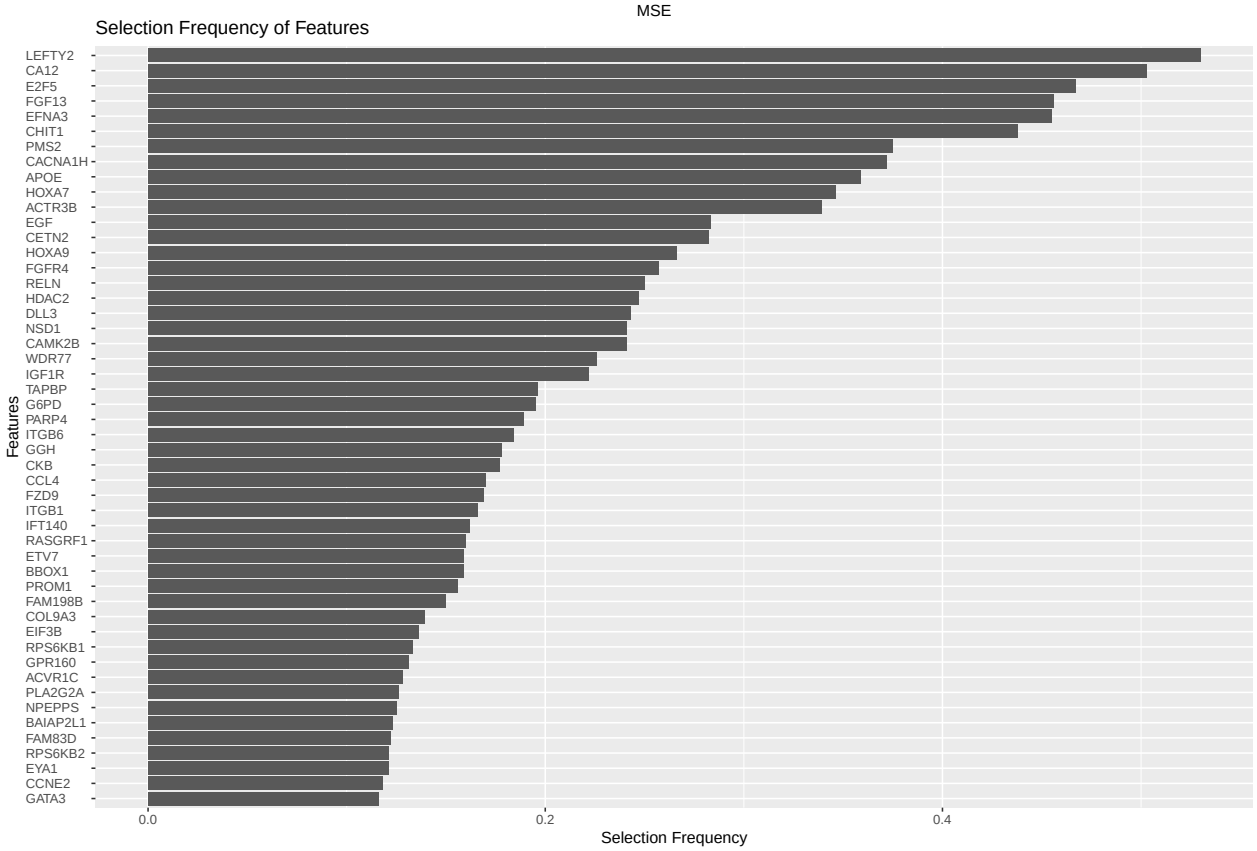
Mechnaistic + Residuals -> ROR-proliferation score (multiplicative)

```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0
##
## CORRELATIONS RESULTS
## Mean: 0.5427757
## Median: 0.5790305
## Variance: 0.03640698
## st.dev.: 0.1908061
```




```
## MSE RESULTS
## Mean: 291.0186
## Median: 284.2399
## Variance: 6854.567
## st.dev.: 82.79231
```


Histogram of MSE Values
Frequency
MSE
0
25
50
75
100
125
200
400
600


Selection Frequency of Features
Features
Selection Frequency
LEFTY2
CA12
E2F5
FGF13
EFNA3
CHIT1
PMS2
CACNA1H
APOE
HOXA7
ACTR3B
EGF
CETN2
HOXA9
FGFR4
RELN
HDAC2
DLL3
NSD1
CAMK2B
WDR77
IGF1R
TAPBP
G6PD
PARP4
ITGB6
GGH
CKB
CCL4
FZD9
ITGB1
IFT140
RASGRF1
ETV7
BBOX1
PROM1
FAM198B
COL9A3
EIF3B
RPS6KB1
GPR160
ACVR1C
PLA2G2A
NPEPPS
BAIAP2L1
FAM83D
RPS6KB2
EYA1
CCNE2
GATA3
0.0
0.2
0.4

##

```
## Features selected 50% or more times:
## CA12 LEFTY2
## Top 20 featrues:
##  [1] "LEFTY2"  "CA12"    "E2F5"    "FGF13"   "EFNA3"   "CHIT1"   "PMS2"
##  [8] "CACNA1H" "APOE"    "HOXA7"   "ACTR3B"  "EGF"     "CETN2"   "HOXA9"
## [15] "FGFR4"   "RELN"    "HDAC2"   "DLL3"    "CAMK2B"  "NSD1"
```

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	0.7266736	0.0895236	0.0813415	0.0231428

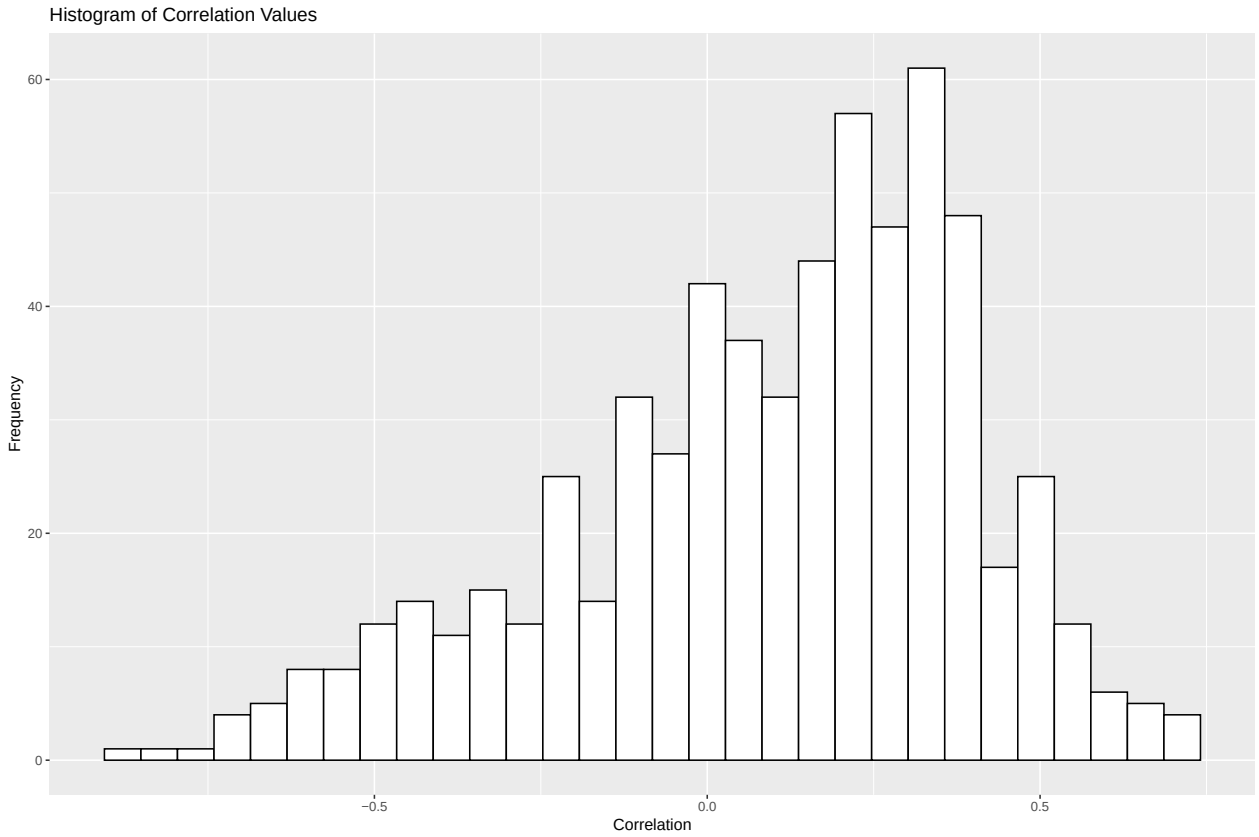
Summery results: lasso ROR+proliferation score (bootstrap)

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	0.5427757	0.1908061	291.0186	82.79231

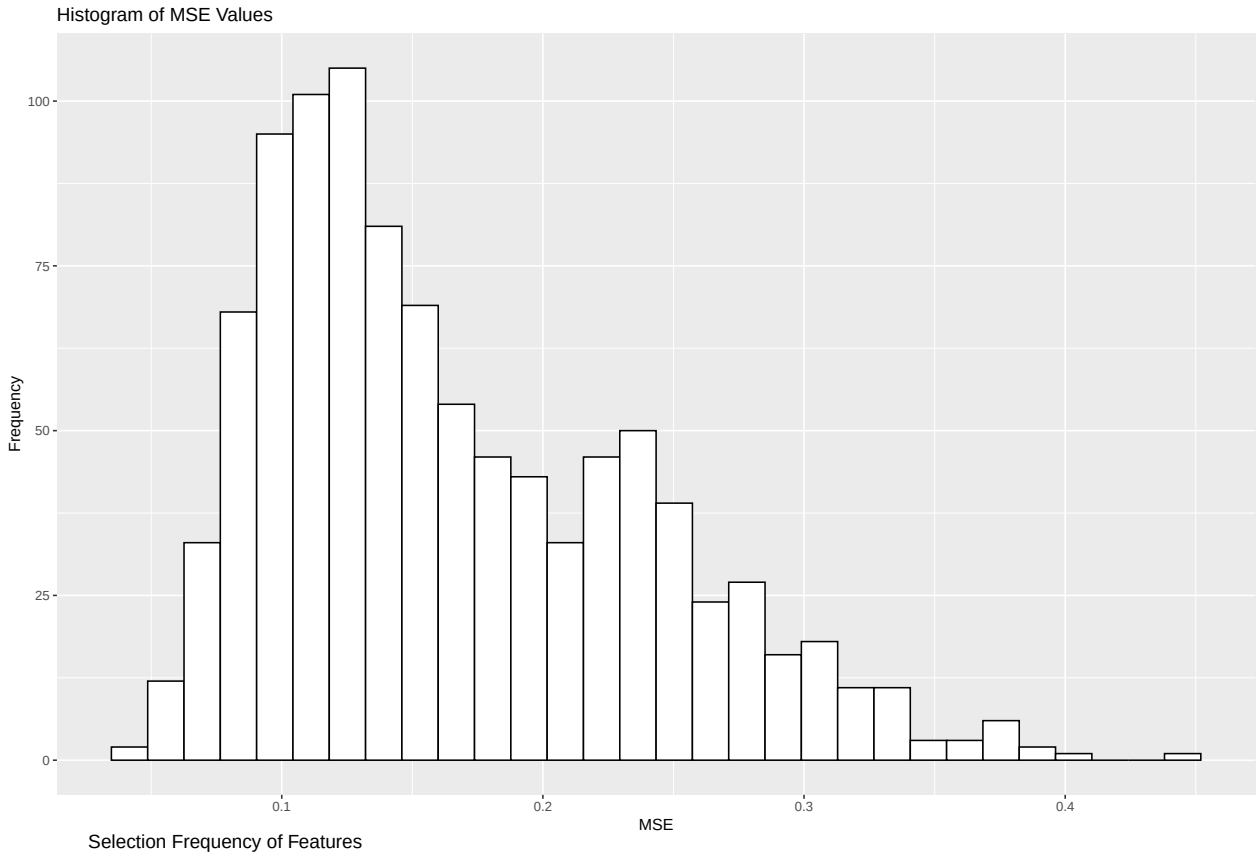
Lasso - Repeated cross-validation

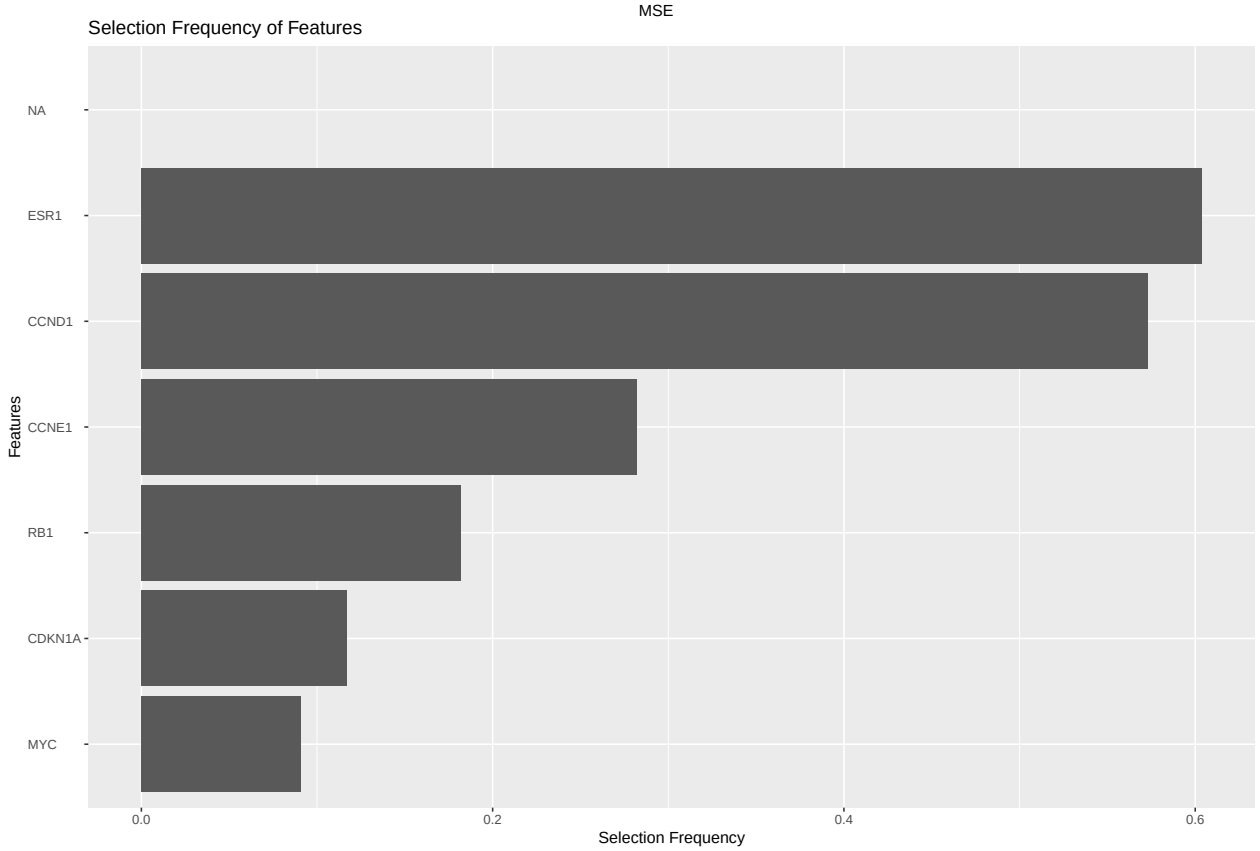
200 repeats of five fold cross-validation ### 6 genes -> proliferation score

```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0.373
##
## CORRELATIONS RESULTS
## Mean: 0.09196628
## Median: 0.1502755
## Variance: 0.09413904
## st.dev.: 0.3068209
```




```
## MSE RESULTS
## Mean: 0.1655931
## Median: 0.1468293
## Variance: 0.005160684
## st.dev.: 0.0718379
```


Histogram of MSE Values
Frequency
100
75
50
25
0
0.1
0.2
0.3
0.4
MSE

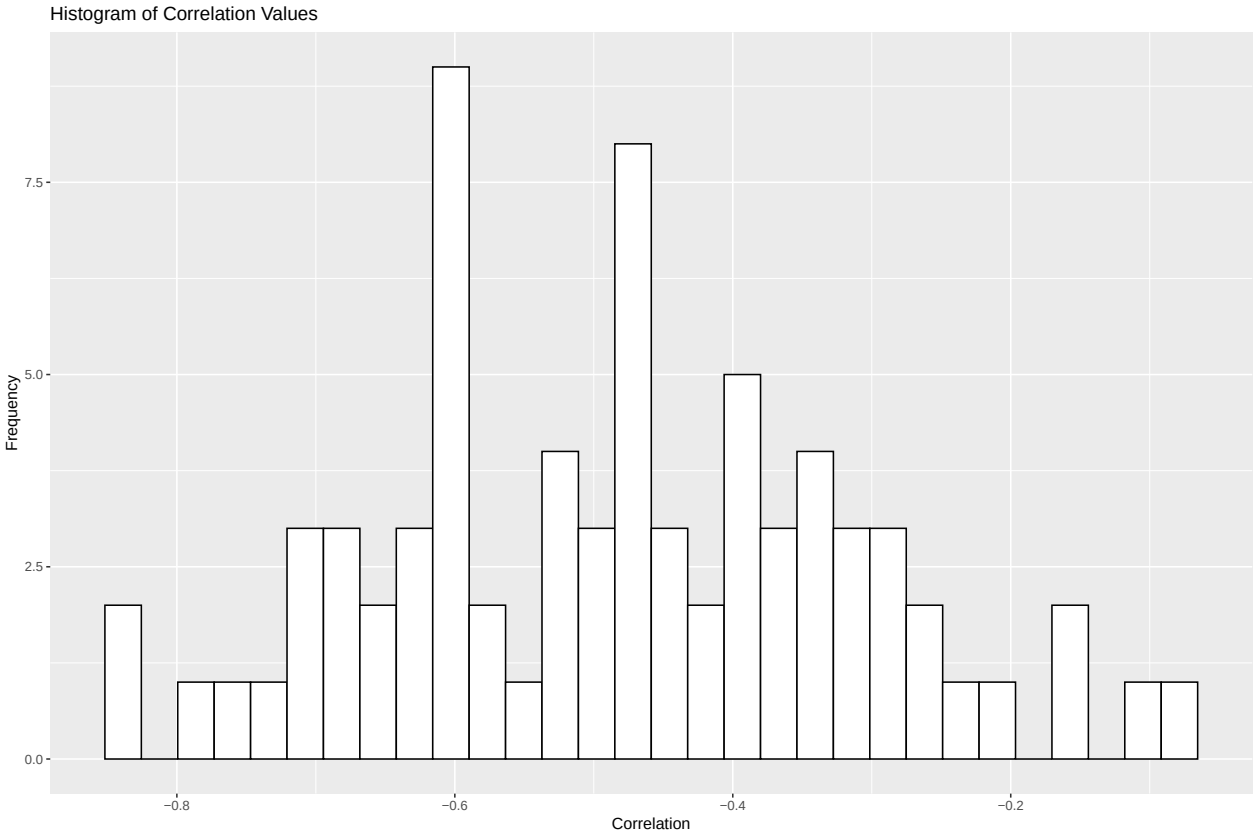

Selection Frequency of Features
Features
NA
ESR1
CCND1
CCNE1
RB1
CDKN1A
MYC
0.0
0.2
0.4
0.6
Selection Frequency

##

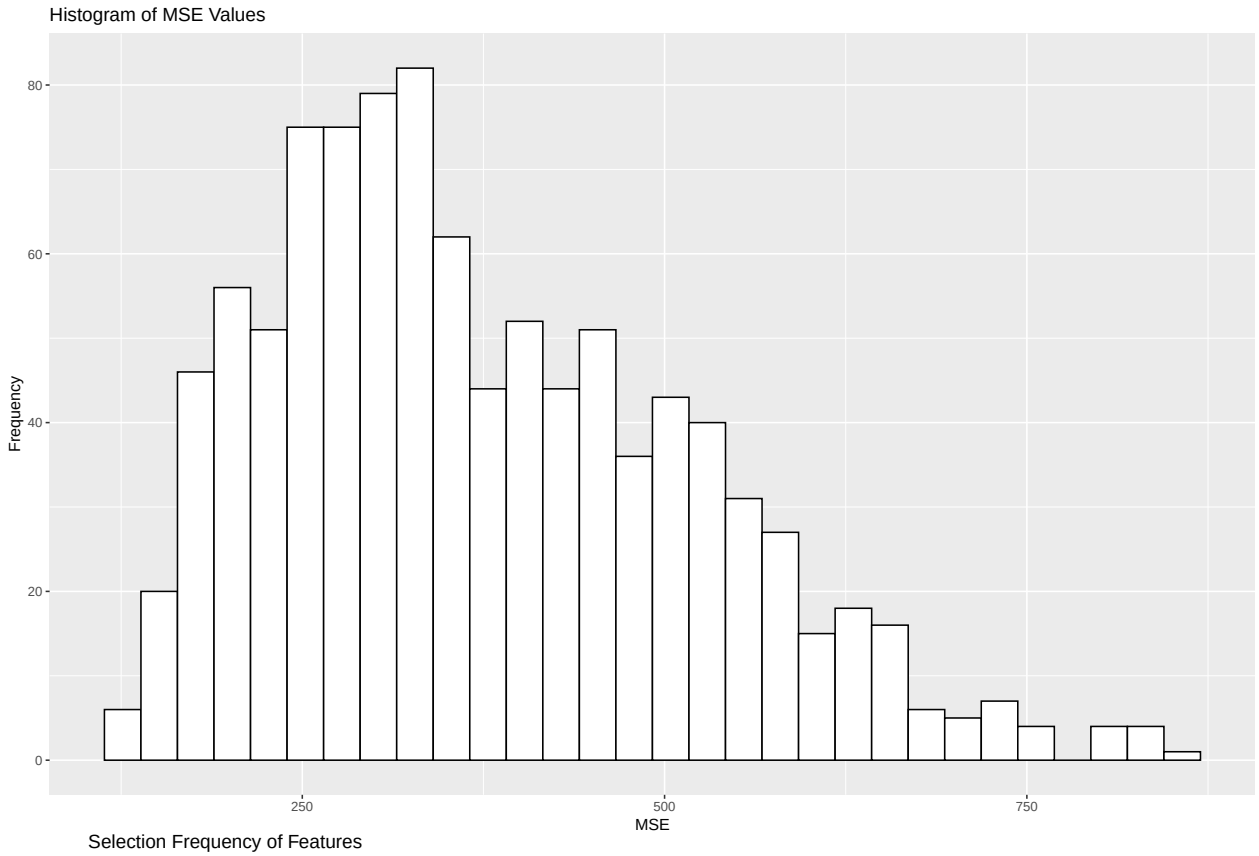
```
## Features selected 50% or more times:
## CCND1 ESR1
## Top 20 featrues:
##  [1] "ESR1"   "CCND1"  "CCNE1"  "RB1"    "CDKN1A" "MYC"    NA       NA
##  [9] NA       NA       NA       NA       NA       NA       NA       NA
## [17] NA       NA       NA       NA
```

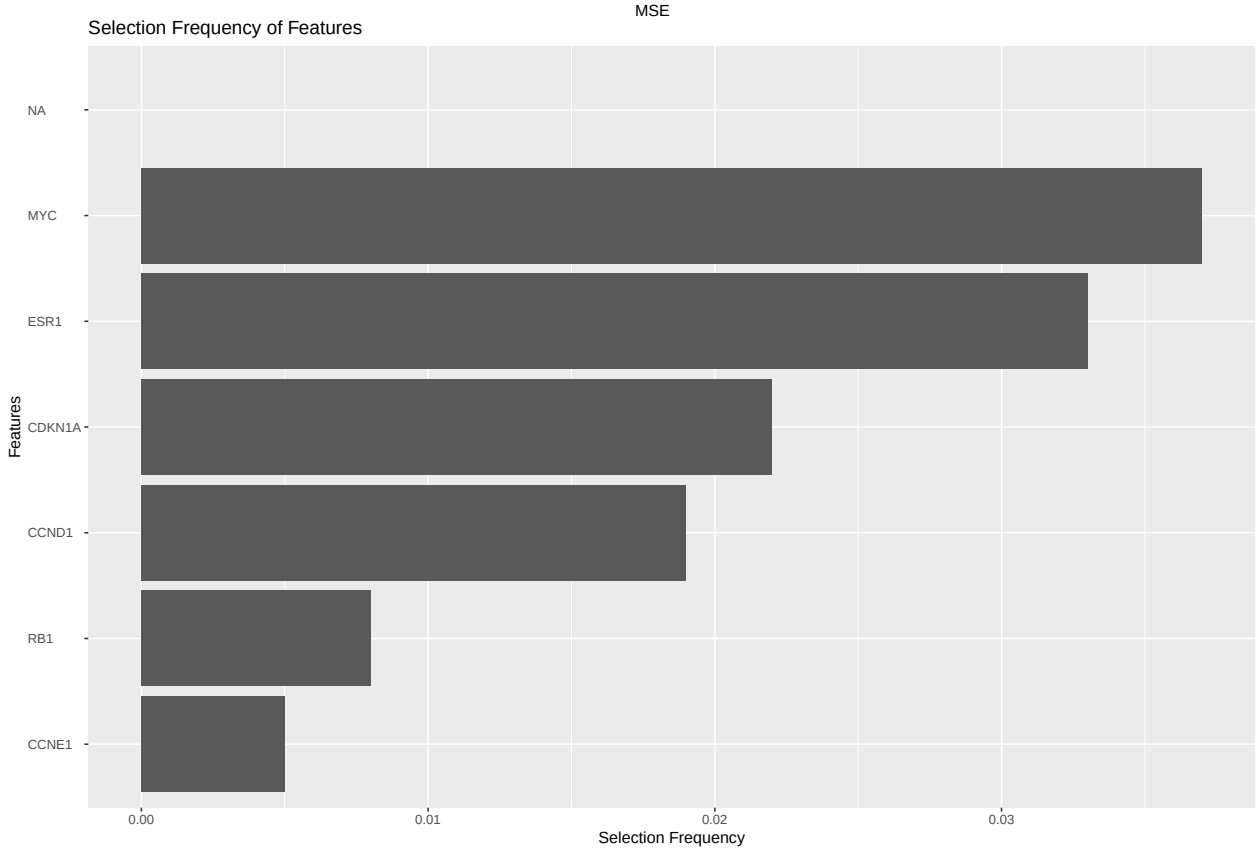
6 genes -> ROR_proliferation score

```
## 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
```


Histogram of MSE Values
Frequency
0
20
40
60
80
250
500
750
MSE

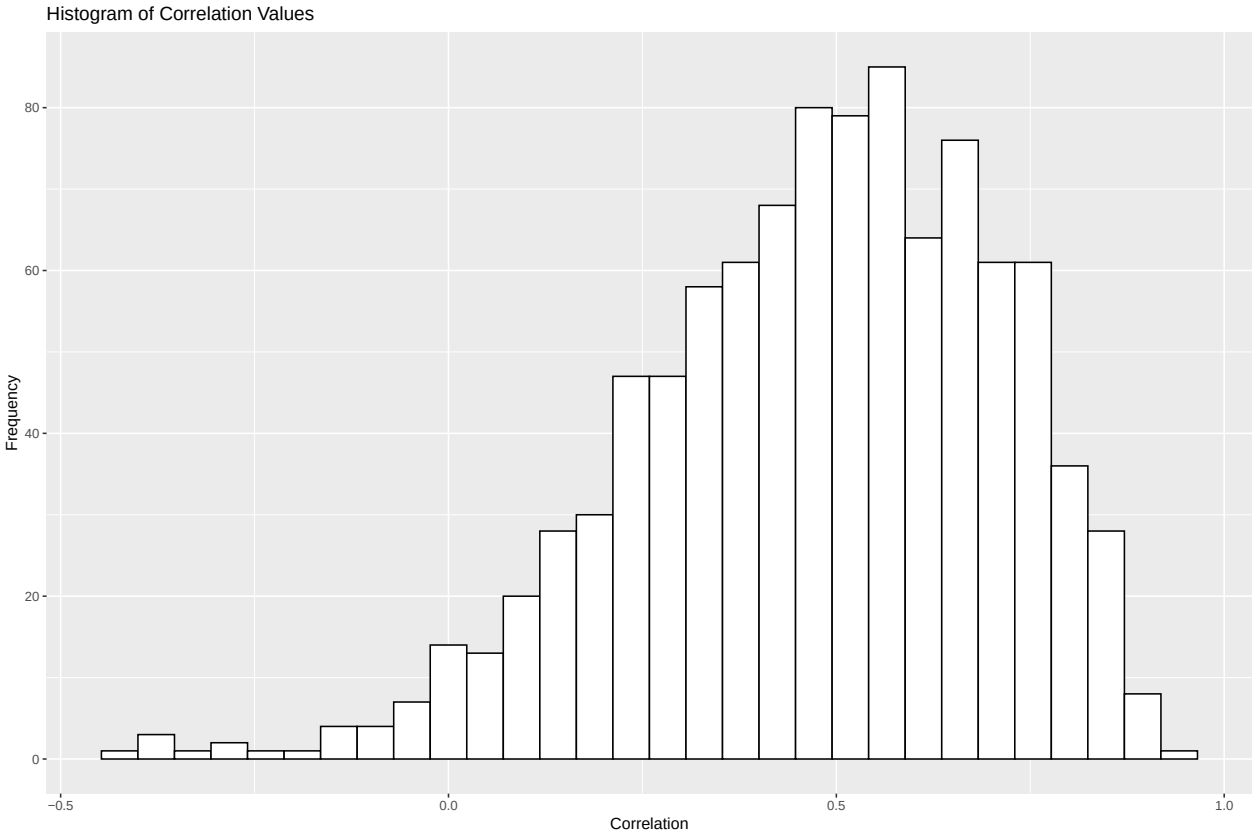

Selection Frequency of Features
Features
NA
MYC
ESR1
CDKN1A
CCND1
RB1
CCNE1
0.00
0.01
0.02
0.03
Selection Frequency

##

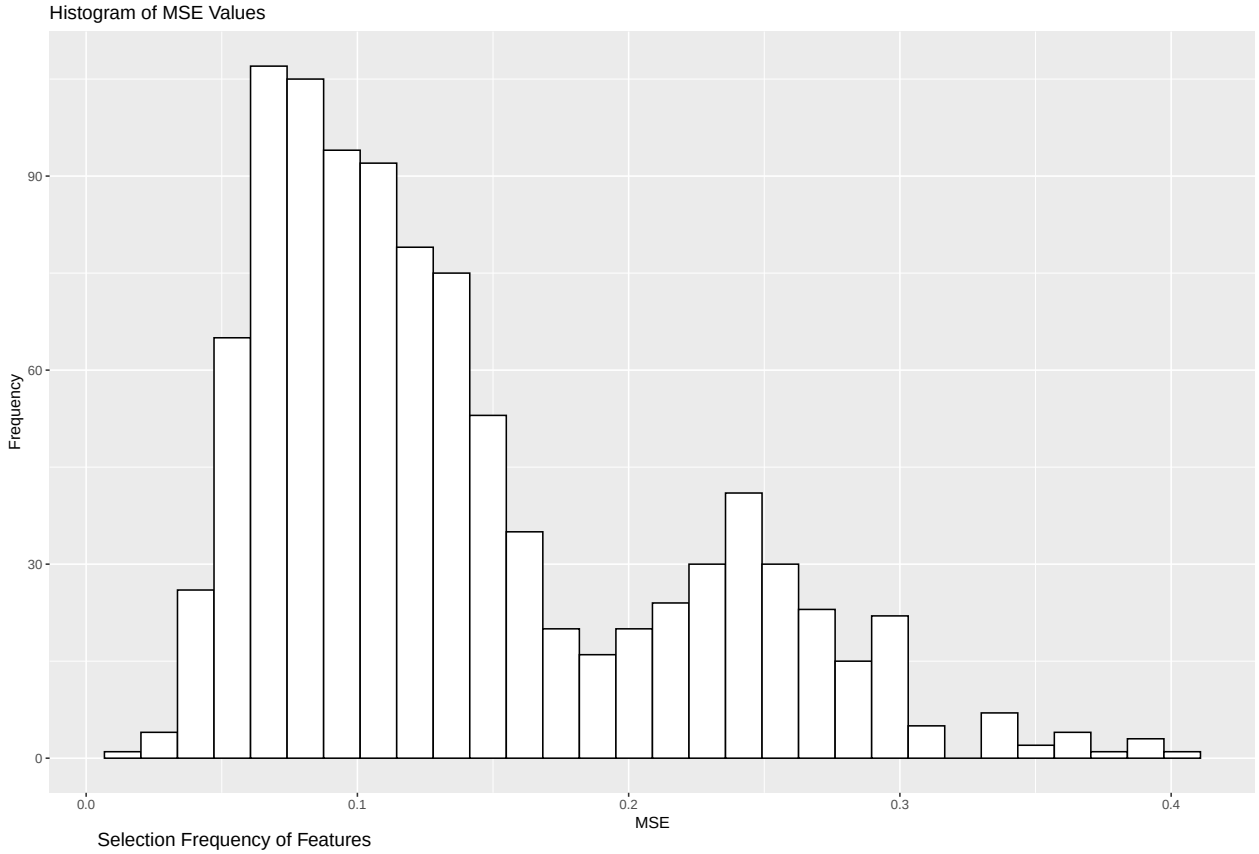
```
## Features selected 50% or more times:
##
## Top 20 featrues:
##  [1] "MYC"    "ESR1"   "CDKN1A" "CCND1"  "RB1"    "CCNE1"  NA       NA
##  [9] NA       NA       NA       NA       NA       NA       NA       NA
## [17] NA       NA       NA       NA
```

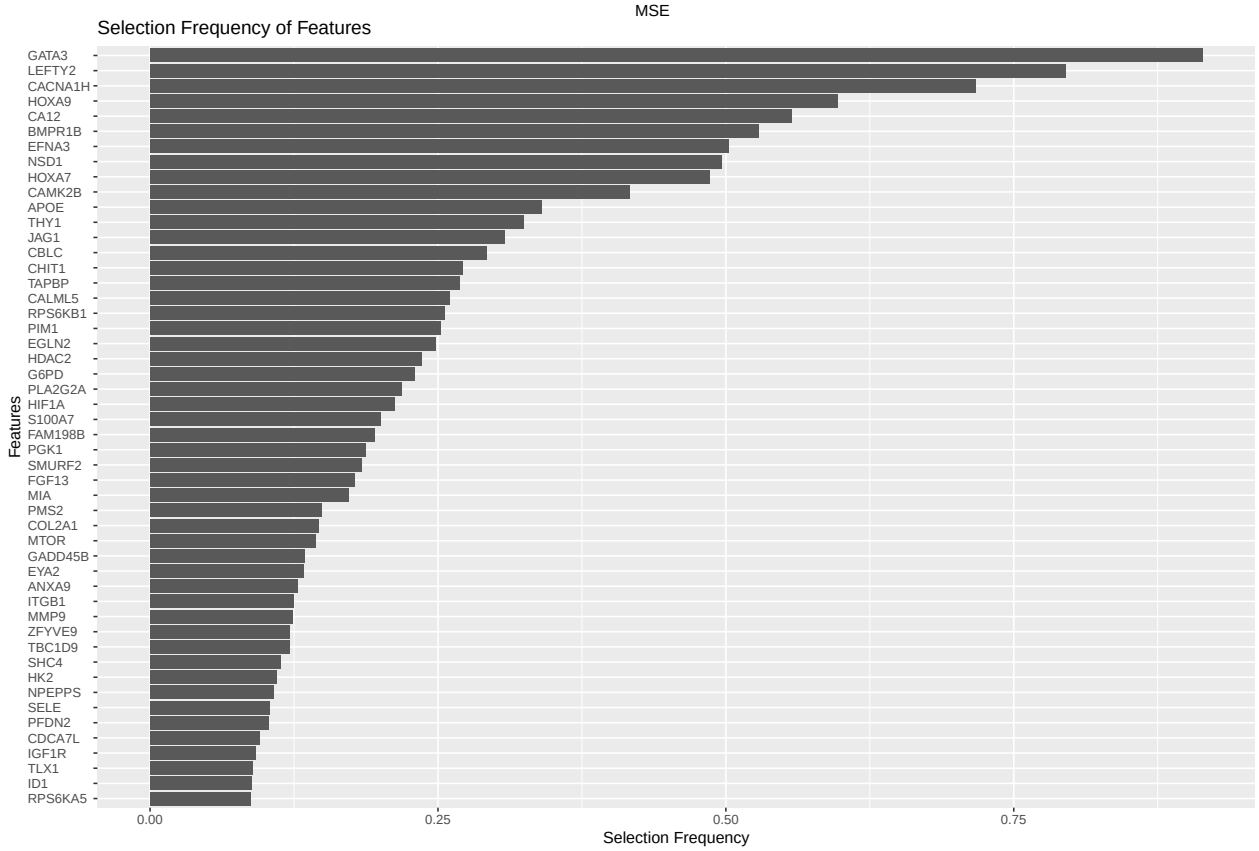
771 genes -> proliferation score

```
## 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
```




```
## MSE RESULTS
## Mean: 0.1376002
## Median: 0.1154157
## Variance: 0.005670929
## st.dev.: 0.07530557
```


Histogram of MSE Values
Frequency
90
60
30
0
0.0
0.1
0.2
0.3
0.4
MSE

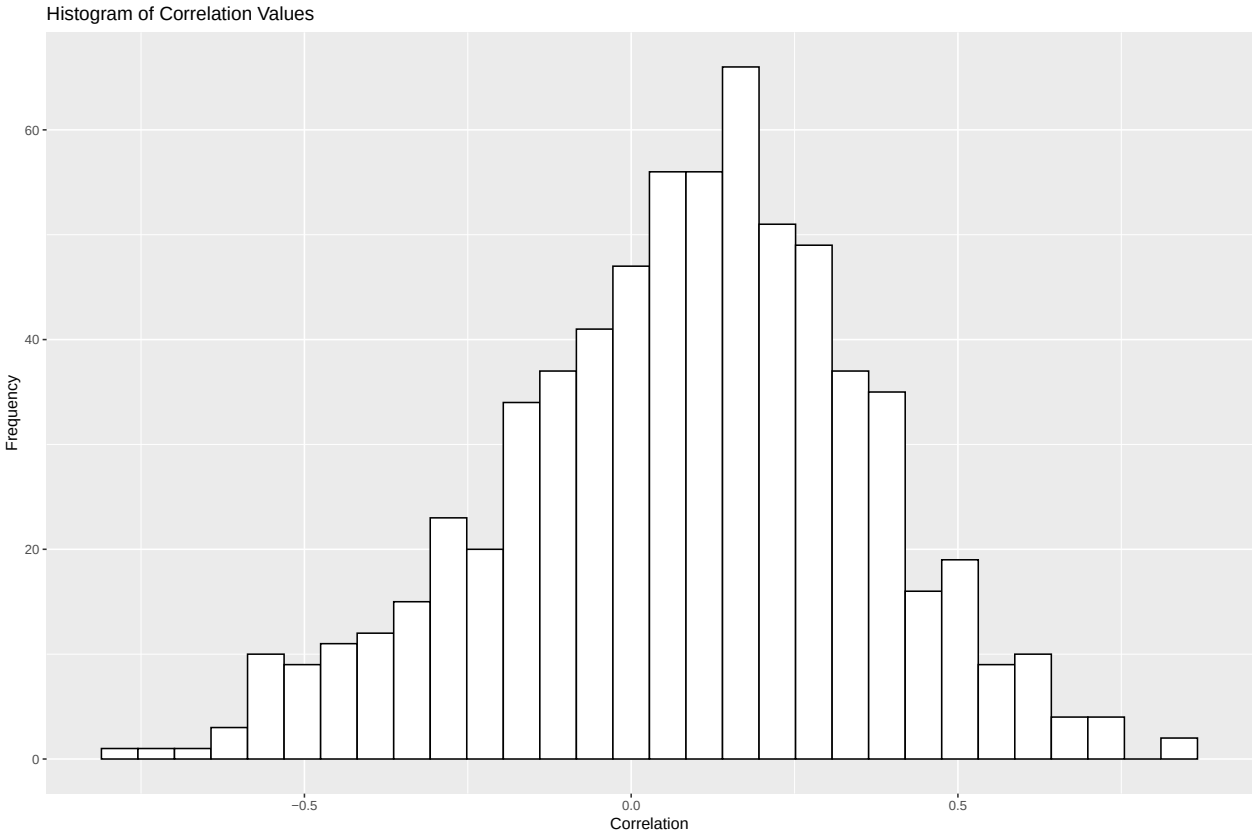

Selection Frequency of Features
Features
GATA3
LEFTY2
CACNA1H
HOXA9
CA12
BMPR1B
EFNA3
NSD1
HOXA7
CAMK2B
APOE
THY1
JAG1
CBLC
CHIT1
TAPBP
CALML5
RPS6KB1
PIM1
EGLN2
HDAC2
G6PD
PLA2G2A
HIF1A
S100A7
FAM198B
PGK1
SMURF2
FGF13
MIA
PMS2
COL2A1
MTOR
GADD45B
EYA2
ANXA9
ITGB1
MMP9
ZFYVE9
TBC1D9
SHC4
HK2
NPEPPS
SELE
PFDN2
CDCA7L
IGF1R
TLX1
ID1
RPS6KA5
0.00
0.25
0.50
0.75
Selection Frequency

##

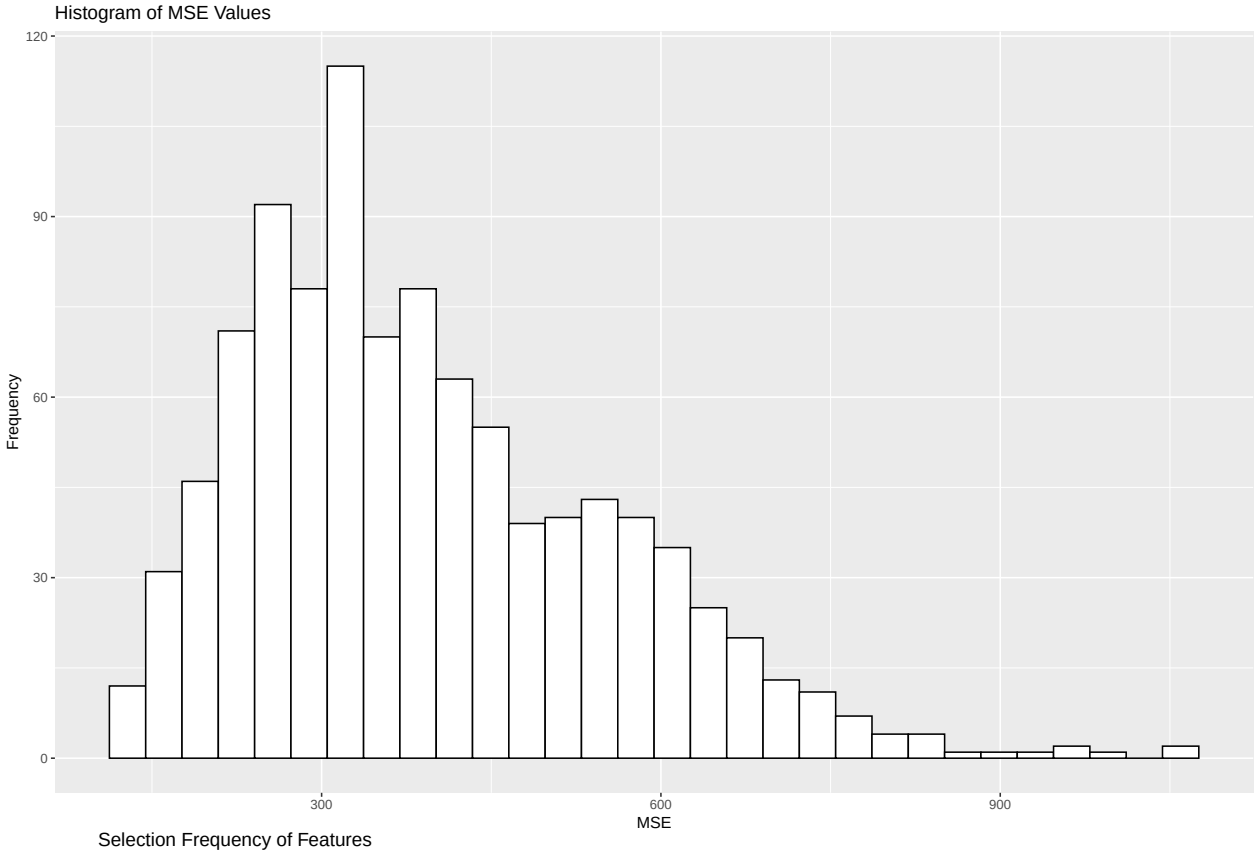
```
## Features selected 50% or more times:
## BMPR1B CA12 CACNA1H EFNA3 GATA3 HOXA9 LEFTY2
## Top 20 featrues:
##  [1] "GATA3"   "LEFTY2"  "CACNA1H" "HOXA9"   "CA12"    "BMPR1B"  "EFNA3"
##  [8] "NSD1"    "HOXA7"   "CAMK2B"  "APOE"    "THY1"    "JAG1"    "CBLC"
## [15] "CHIT1"   "TAPBP"   "CALML5" 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

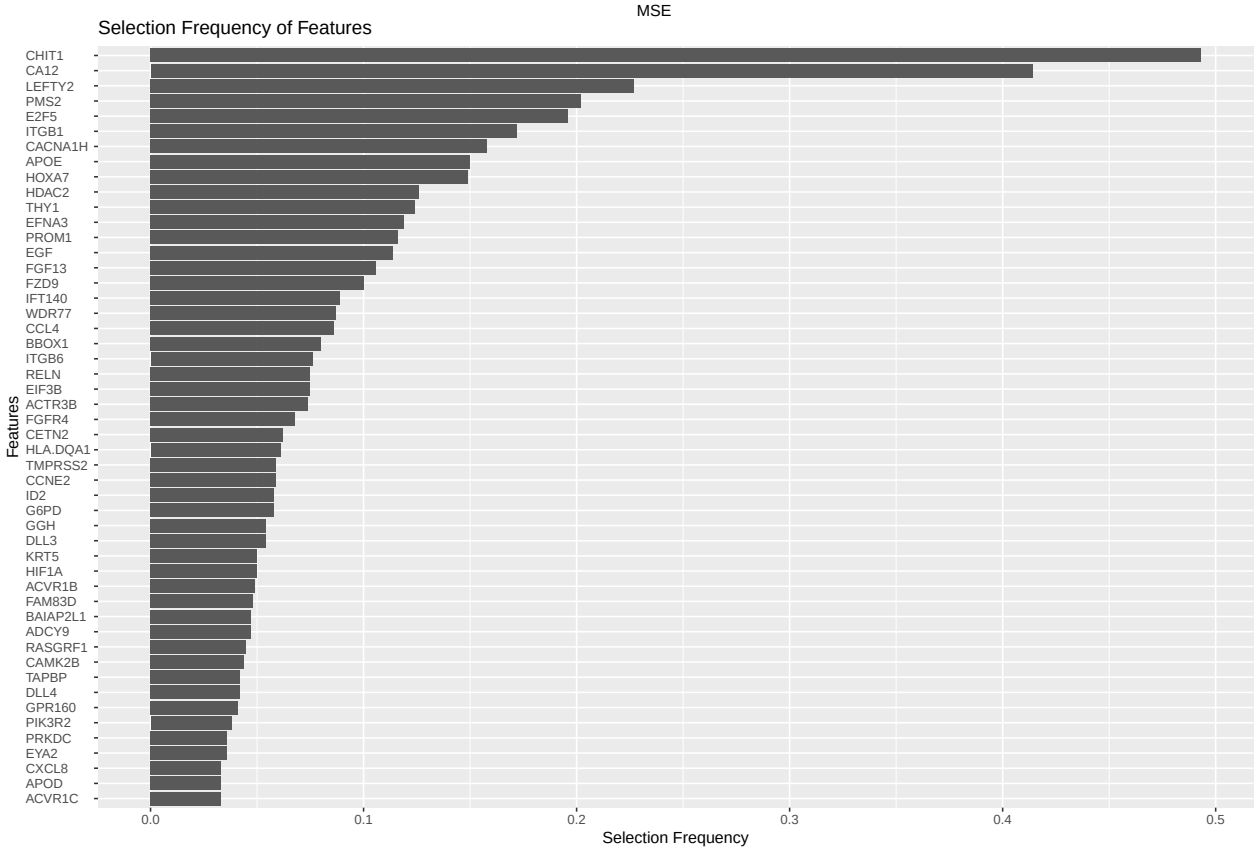
771 genes -> ROR-proliferation score

```
## 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
```


Histogram of MSE Values
Frequency
120
90
60
30
0
300
600
900
MSE

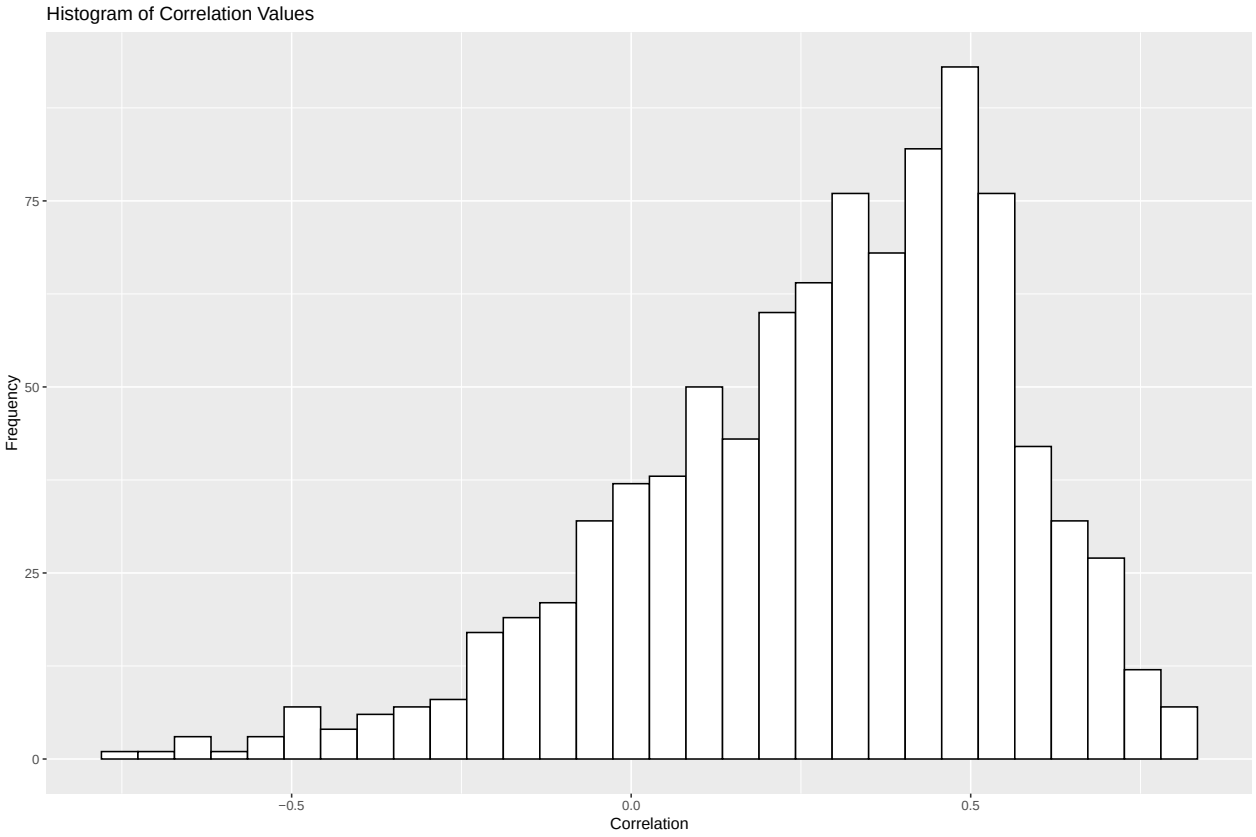

Selection Frequency of Features
Features
CHIT1
CA12
LEFTY2
PMS2
E2F5
ITGB1
CACNA1H
APOE
HOXA7
HDAC2
THY1
EFNA3
PROM1
EGF
FGF13
FZD9
IFT140
WDR77
CCL4
BBOX1
ITGB6
RELN
EIF3B
ACTR3B
FGFR4
CETN2
HLA.DQA1
TMPRSS2
CCNE2
ID2
G6PD
GGH
DLL3
KRT5
HIF1A
ACVR1B
FAM83D
BAIAP2L1
ADCY9
RASGRF1
CAMK2B
TAPBP
DLL4
GPR160
PIK3R2
PRKDC
EYA2
CXCL8
APOD
ACVR1C
0.0
0.1
0.2
0.3
0.4
0.5
Selection Frequency

##

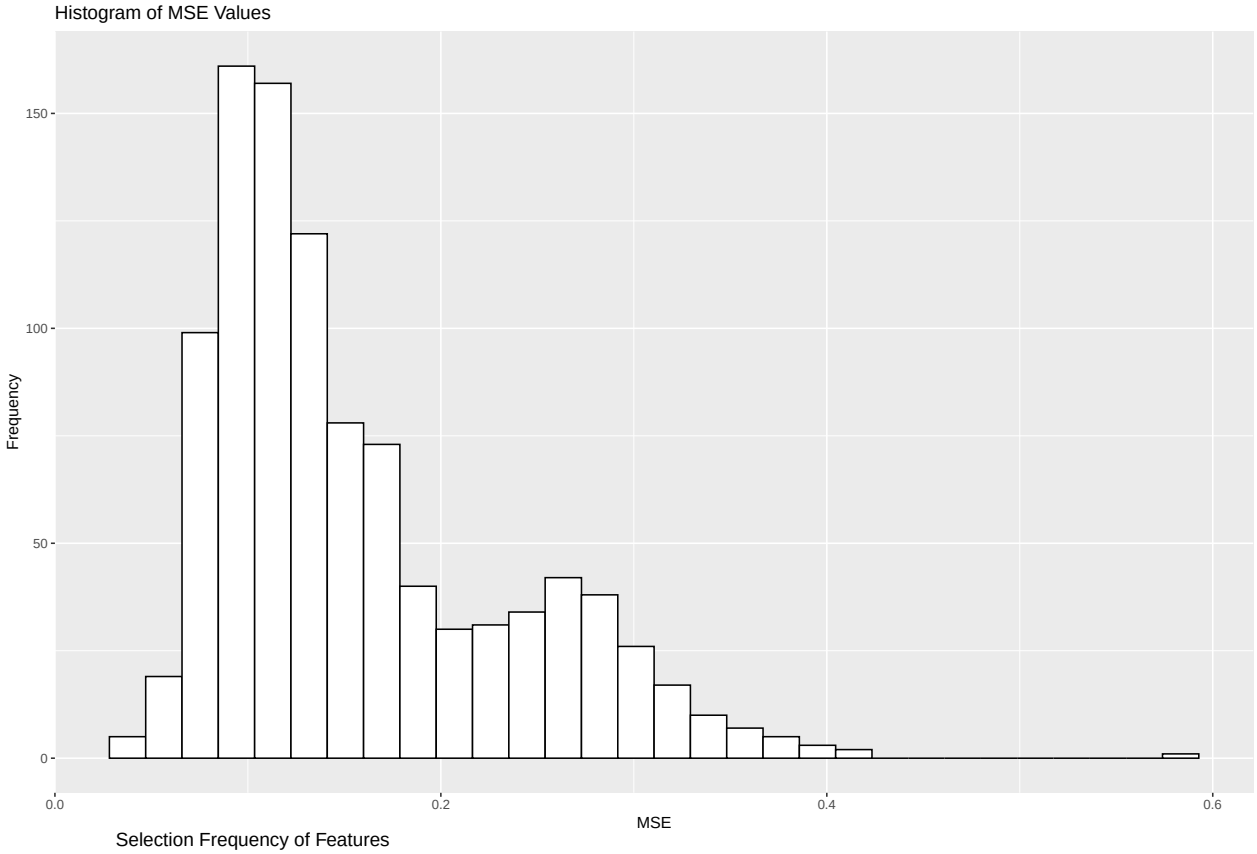
```
## Features selected 50% or more times:
##
## Top 20 featrues:
##  [1] "CHIT1"   "CA12"    "LEFTY2"  "PMS2"    "E2F5"    "ITGB1"   "CACNA1H"
##  [8] "APOE"    "HOXA7"   "HDAC2"   "THY1"    "EFNA3"   "PROM1"   "EGF"
## [15] "FGF13"   "FZD9"    "IFT140"  "WDR77"   "CCL4"    "BBOX1"
```

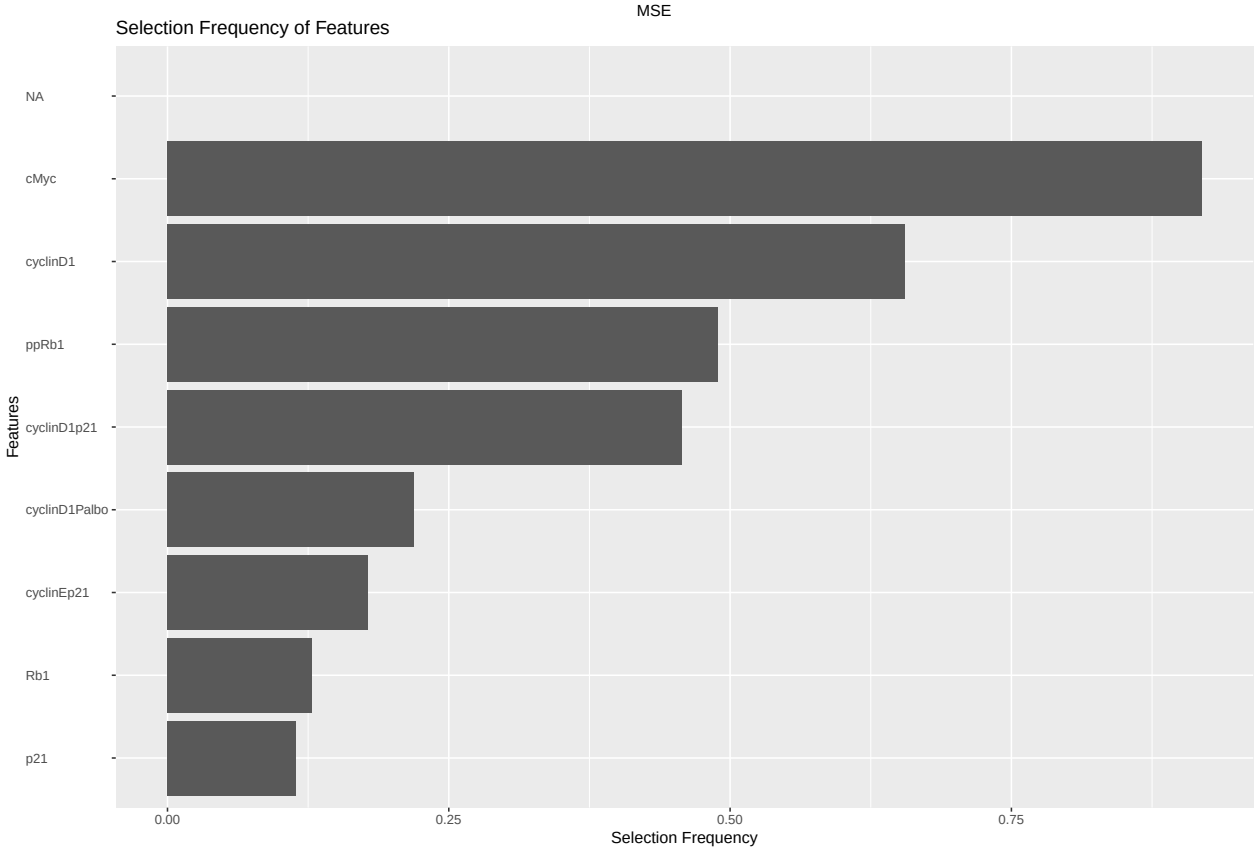
node values -> proliferation score

```
## 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
```




```
## MSE RESULTS
## Mean: 0.1560308
## Median: 0.1314678
## Variance: 0.005819908
## st.dev.: 0.07628832
```


Histogram of MSE Values
Frequency
150
100
50
0
0.0
0.2
0.4
0.6
MSE

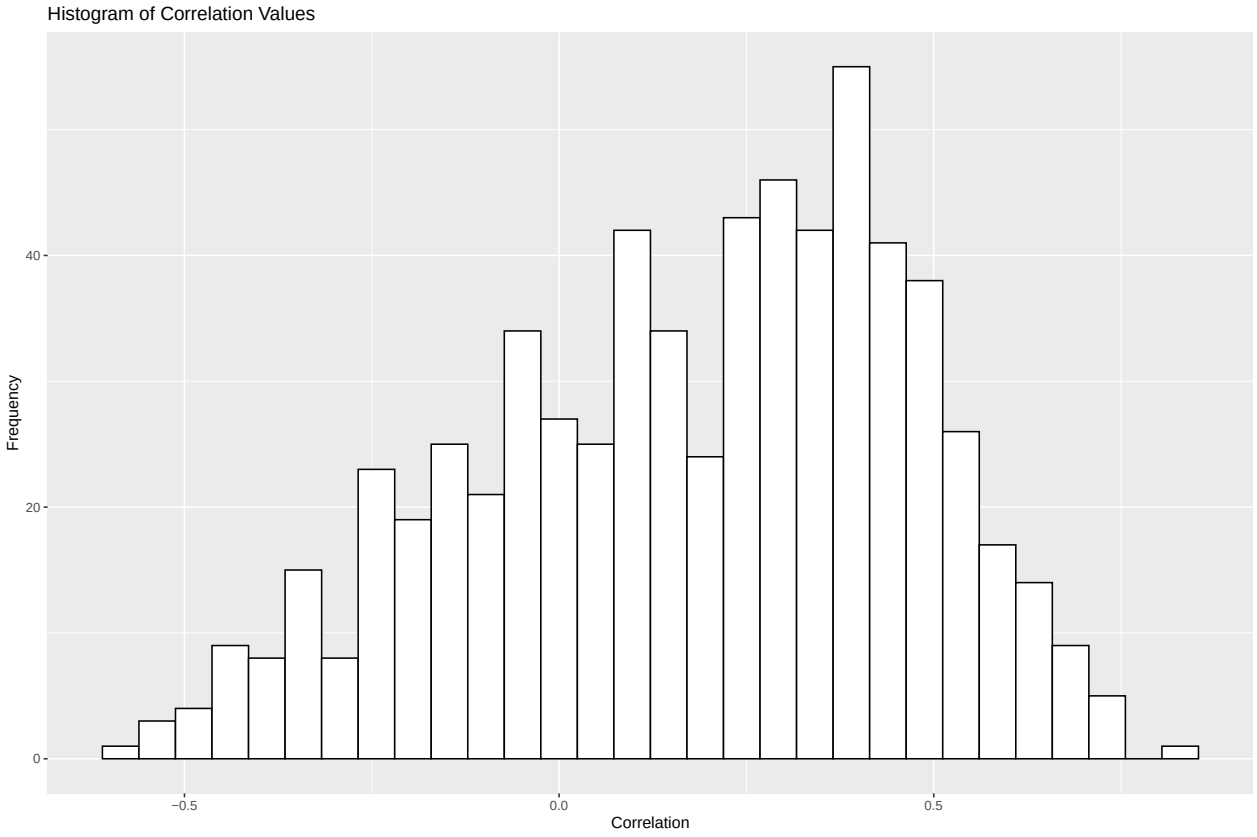

Selection Frequency of Features
Features
NA
cMyc
cyclinD1
ppRb1
cyclinD1p21
cyclinD1Palbo
cyclinEp21
Rb1
p21
0.00
0.25
0.50
0.75
Selection Frequency

##

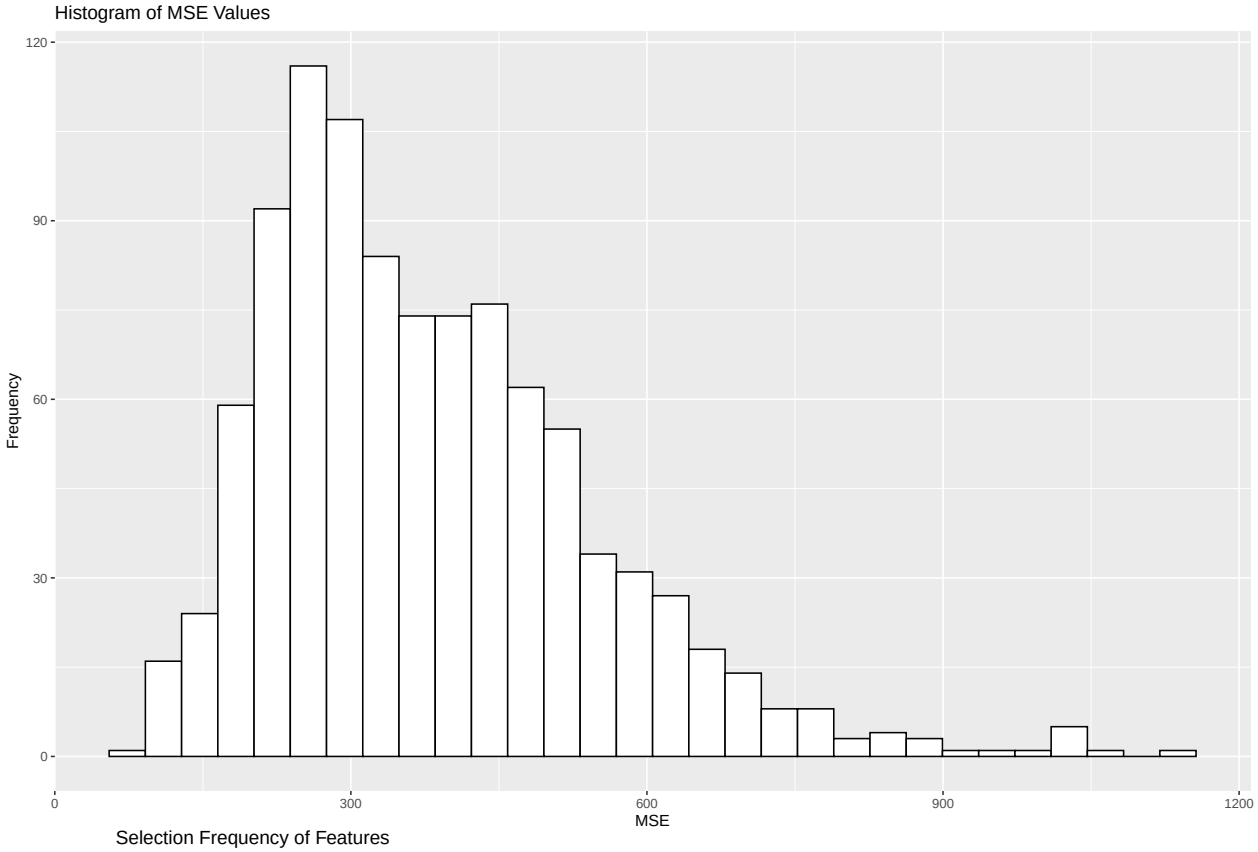
```
## Features selected 50% or more times:
## cyclinD1 cMyc
## Top 20 featrues:
##  [1] "cMyc"          "cyclinD1"      "ppRb1"         "cyclinD1p21"
##  [5] "cyclinD1Palbo" "cyclinEp21"    "Rb1"           "p21"
##  [9] NA              NA              NA              NA
## [13] NA              NA  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

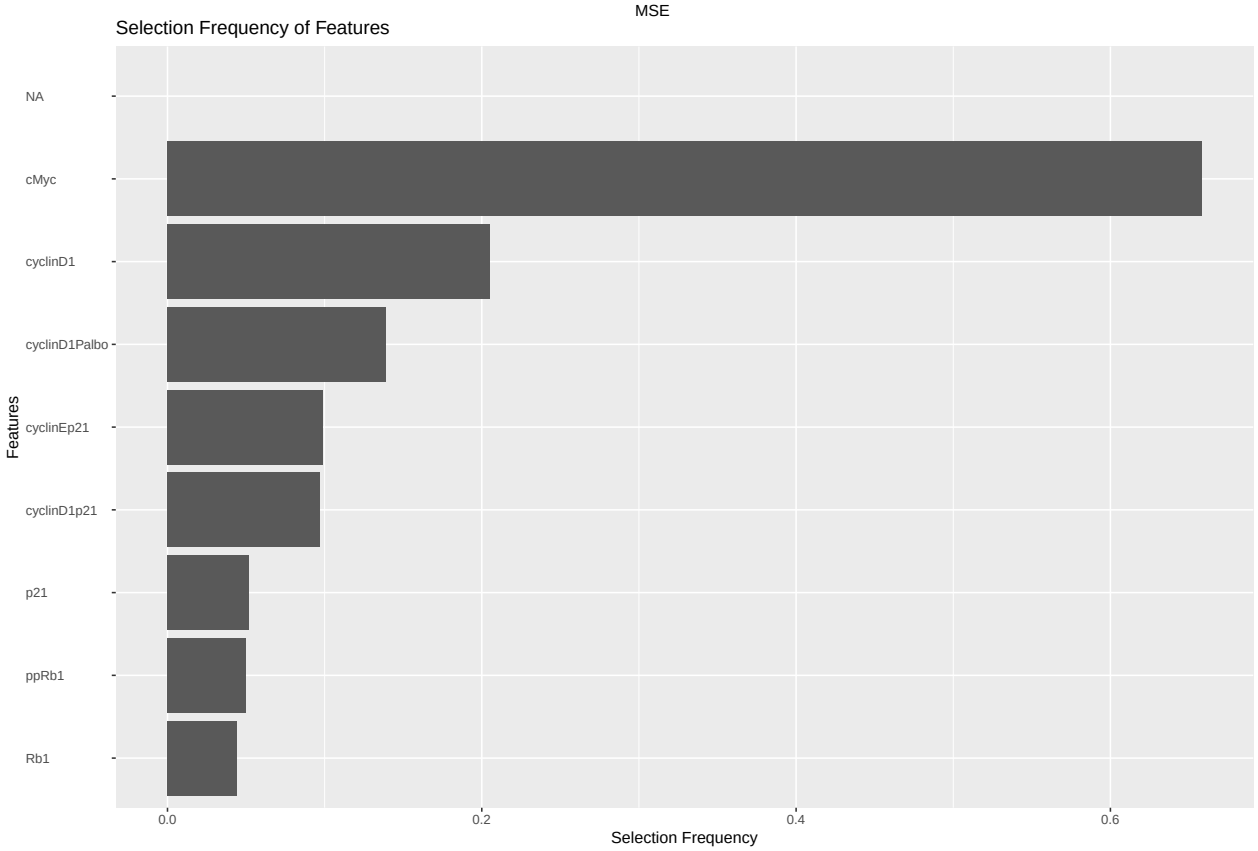
node values -> ROR-proliferation score

```
## 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
```


Histogram of MSE Values
Frequency
MSE
0
30
60
90
120
0
300
600
900
1200

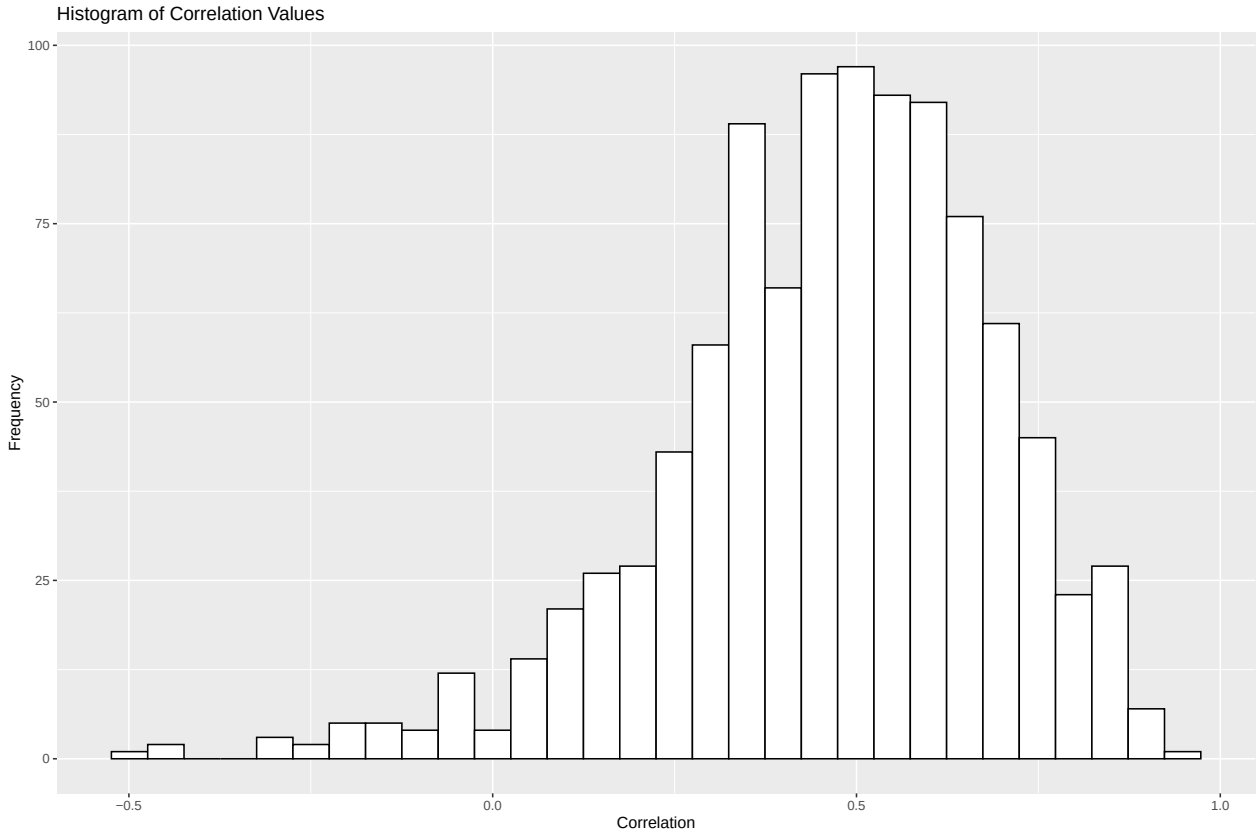

Selection Frequency of Features
Features
Selection Frequency
NA
cMyc
cyclinD1
cyclinD1Palbo
cyclinEp21
cyclinD1p21
p21
ppRb1
Rb1
0.0
0.2
0.4
0.6

##

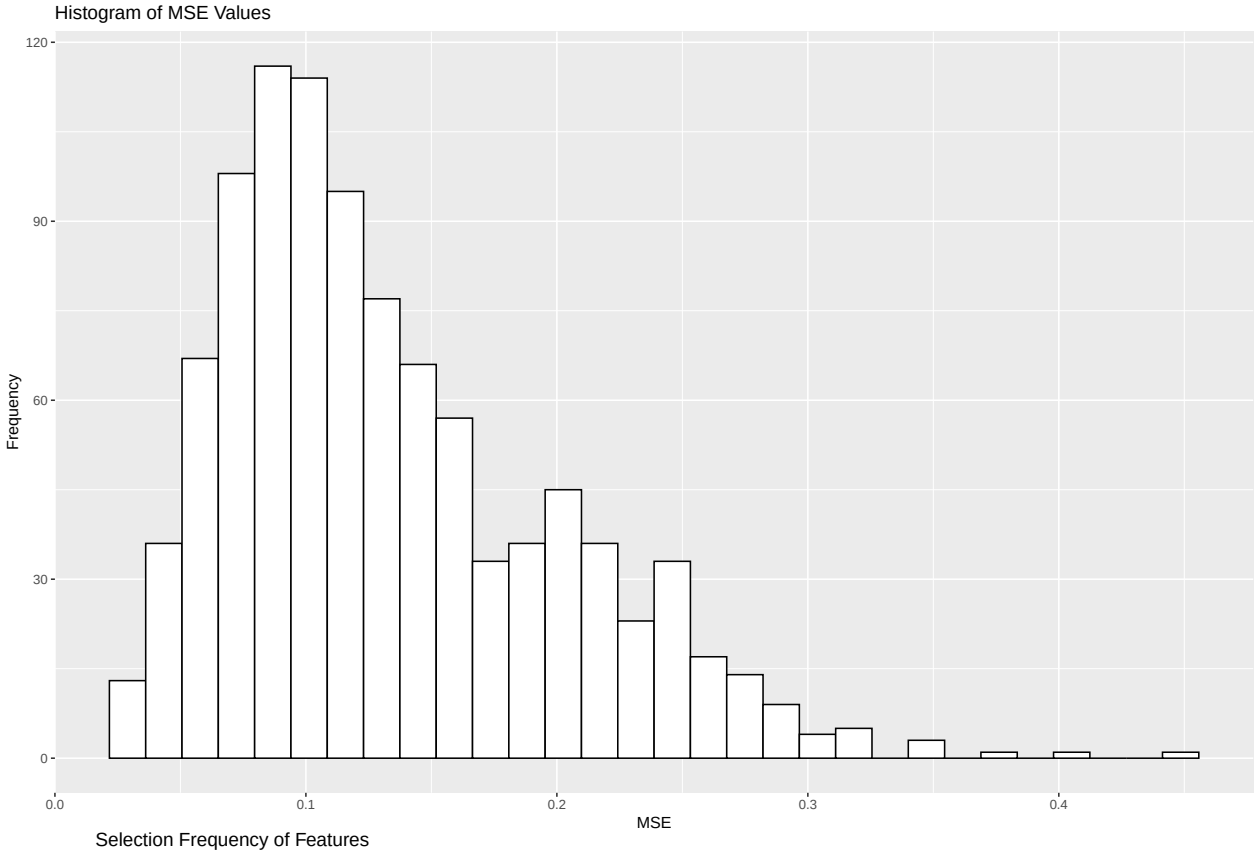
```
## Features selected 50% or more times:
## cMyc
## Top 20 featrues:
##  [1] "cMyc"          "cyclinD1"      "cyclinD1Palbo" "cyclinEp21"
##  [5] "cyclinD1p21"   "p21"           "ppRb1"         "Rb1"
##  [9] NA              NA              NA              NA
## [13] NA              NA            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

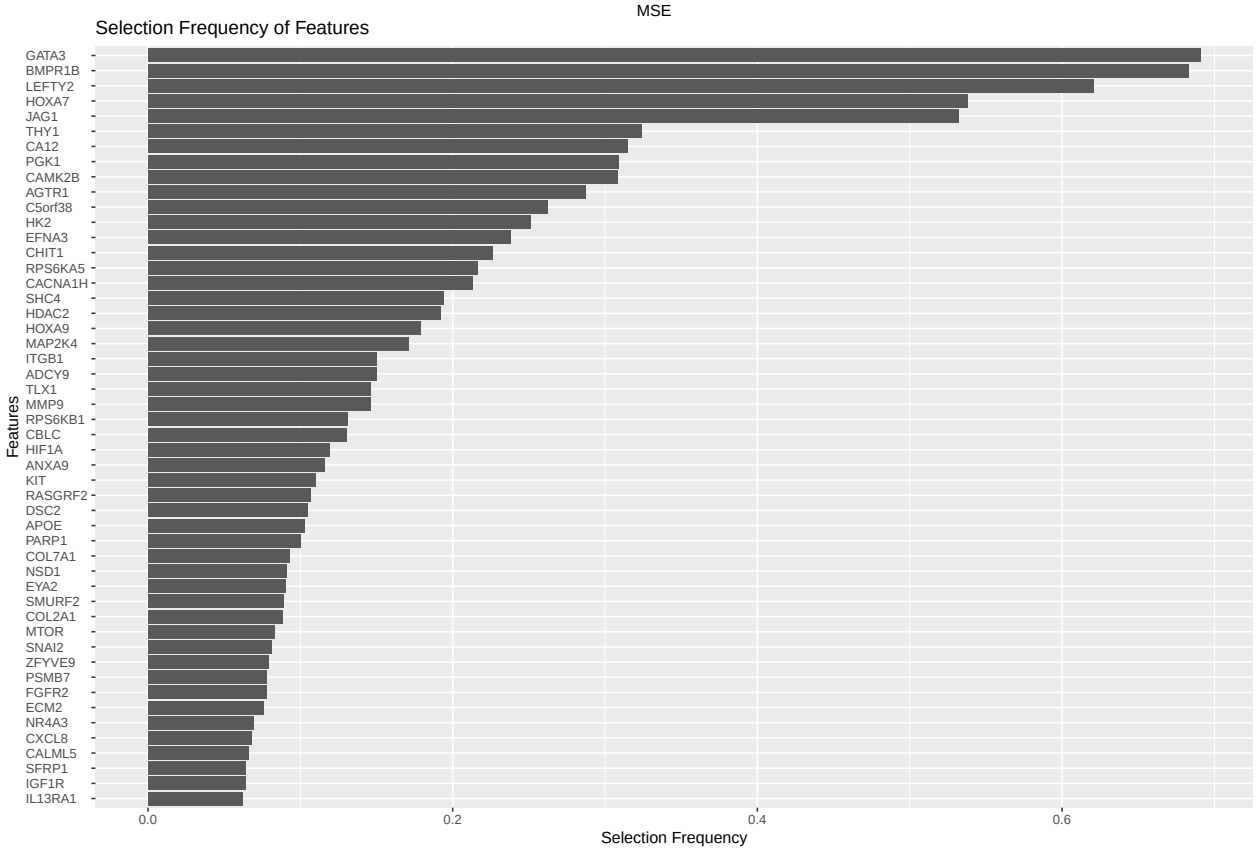
Mechanistic + Residuals -> proliferation score (additive)

```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0
##
## CORRELATIONS RESULTS
## Mean: 0.4633095
## Median: 0.4870052
## Variance: 0.04959996
## st.dev.: 0.2227105
```




```
## MSE RESULTS
## Mean: 0.1331785
## Median: 0.1164927
## Variance: 0.004278014
## st.dev.: 0.06540653
```


Histogram of MSE Values
Frequency
MSE
0
30
60
90
120
0.0
0.1
0.2
0.3
0.4

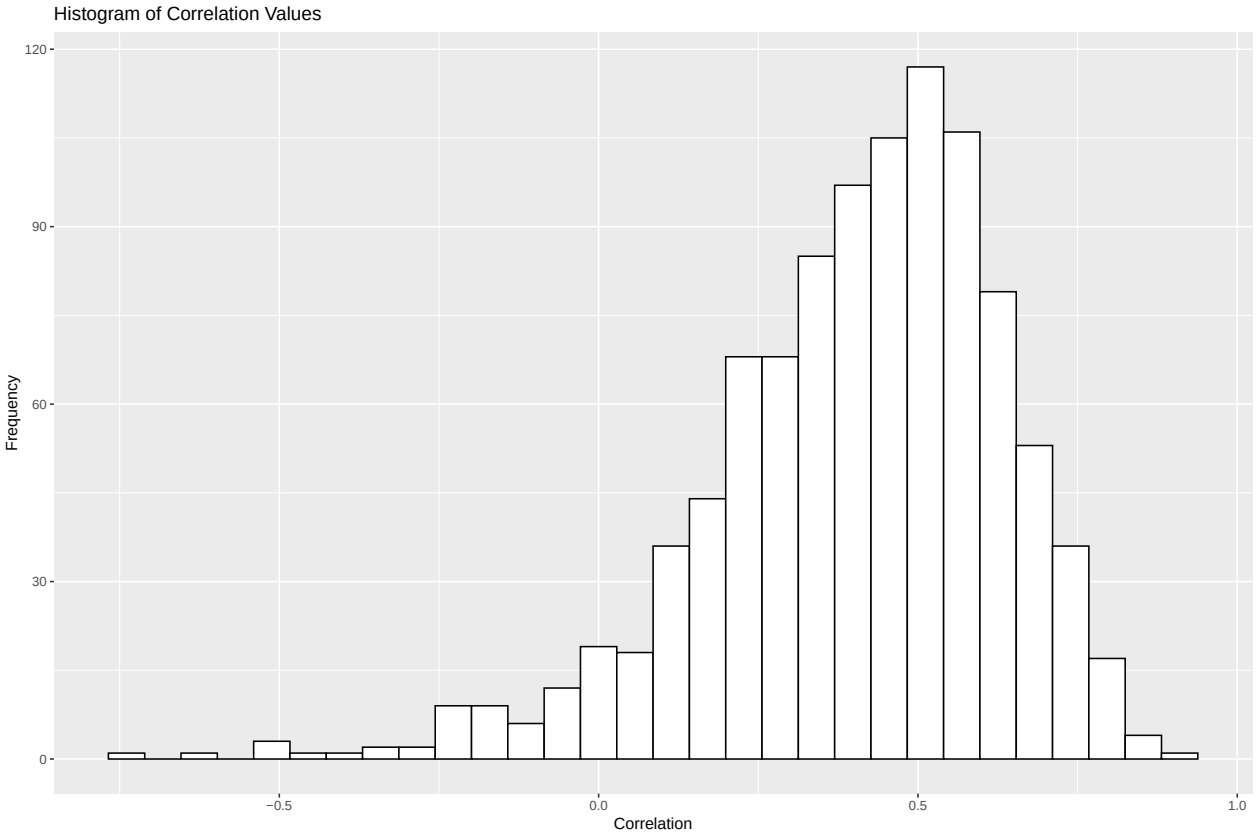

Selection Frequency of Features
Features
Selection Frequency
GATA3
BMPR1B
LEFTY2
HOXA7
JAG1
THY1
CA12
PGK1
CAMK2B
AGTR1
C5orf38
HK2
EFNA3
CHIT1
RPS6KA5
CACNA1H
SHC4
HDAC2
HOXA9
MAP2K4
ITGB1
ADCY9
TLX1
MMP9
RPS6KB1
CBLC
HIF1A
ANXA9
KIT
RASGRF2
DSC2
APOE
PARP1
COL7A1
NSD1
EYA2
SMURF2
COL2A1
MTOR
SNAI2
ZFYVE9
PSMB7
FGFR2
ECM2
NR4A3
CXCL8
CALML5
SFRP1
IGF1R
IL13RA1
0.0
0.2
0.4
0.6

##

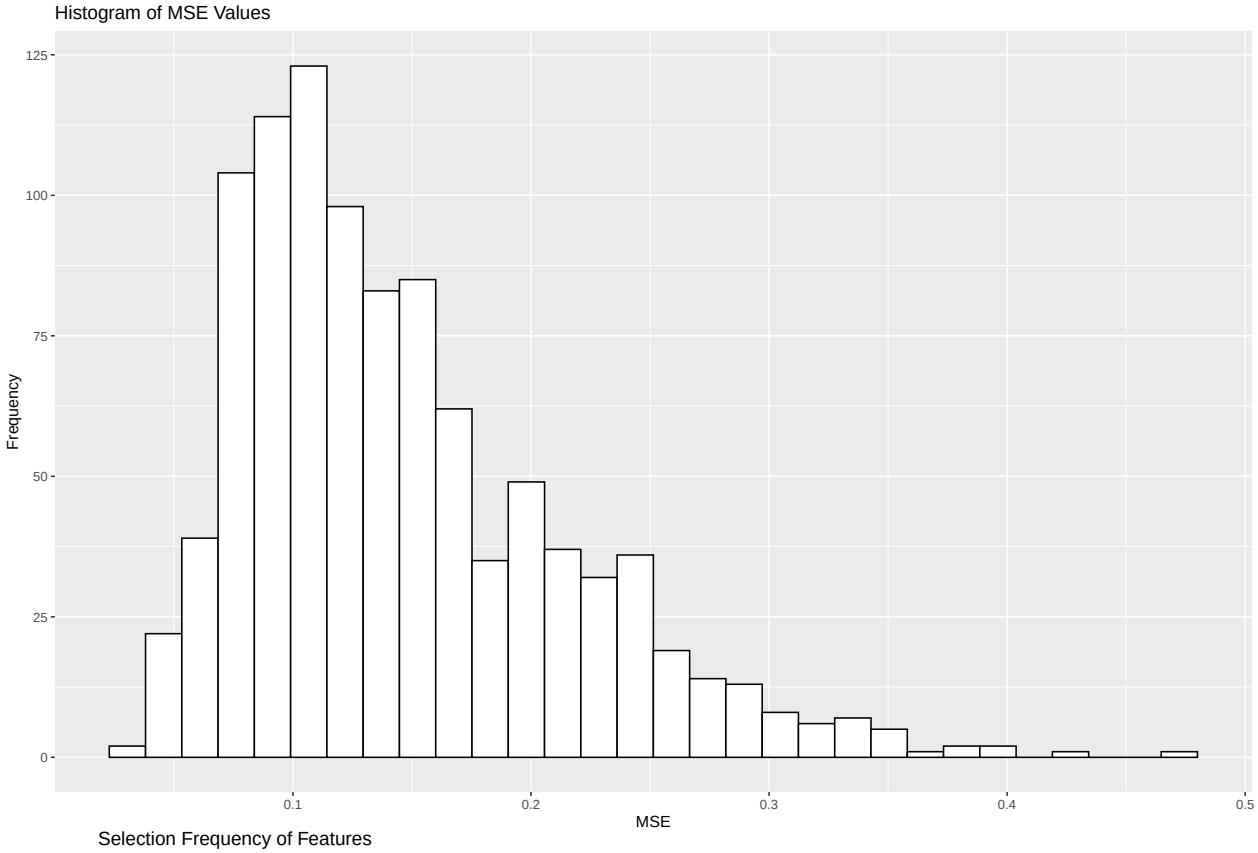
```
## Features selected 50% or more times:
## BMPR1B GATA3 HOXA7 JAG1 LEFTY2
## Top 20 featrues:
##  [1] "GATA3"   "BMPR1B"  "LEFTY2"  "HOXA7"   "JAG1"    "THY1"    "CA12"
##  [8] "PGK1"    "CAMK2B"  "AGTR1"   "C5orf38" "HK2"     "EFNA3"   "CHIT1"
## [15] "RPS6KA5" "CACNA1H" "SHC4"    "HDAC2"   "HOXA9"   "MAP2K4"
```

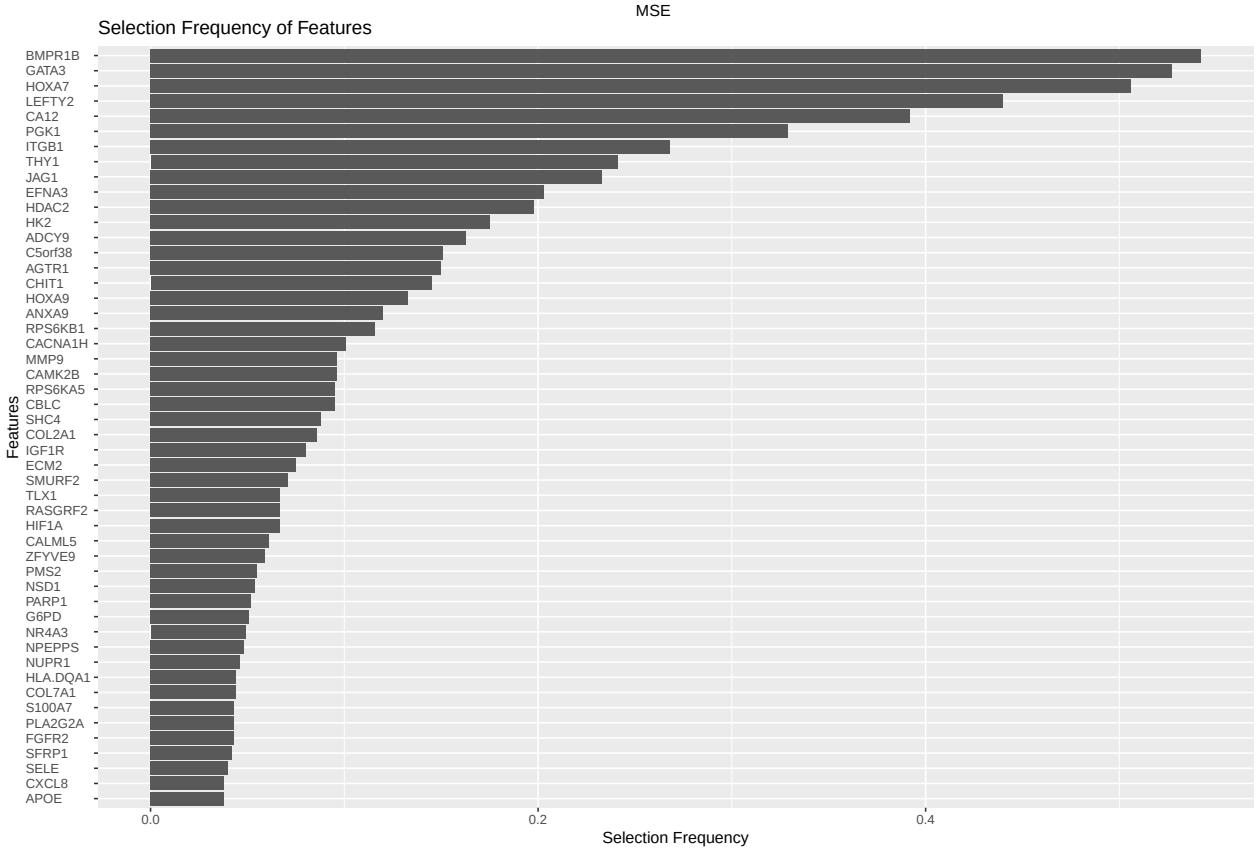
Mechanistic + Residuals -> proliferation score (multiplicative)

```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0
##
## CORRELATIONS RESULTS
## Mean: 0.4028471
## Median: 0.437445
## Variance: 0.05302116
## st.dev.: 0.2302632
```




```
## MSE RESULTS
## Mean: 0.1455819
## Median: 0.1286019
## Variance: 0.004632394
## st.dev.: 0.06806169
```


Histogram of MSE Values
Frequency
125
100
75
50
25
0
0.1
0.2
0.3
0.4
0.5
MSE

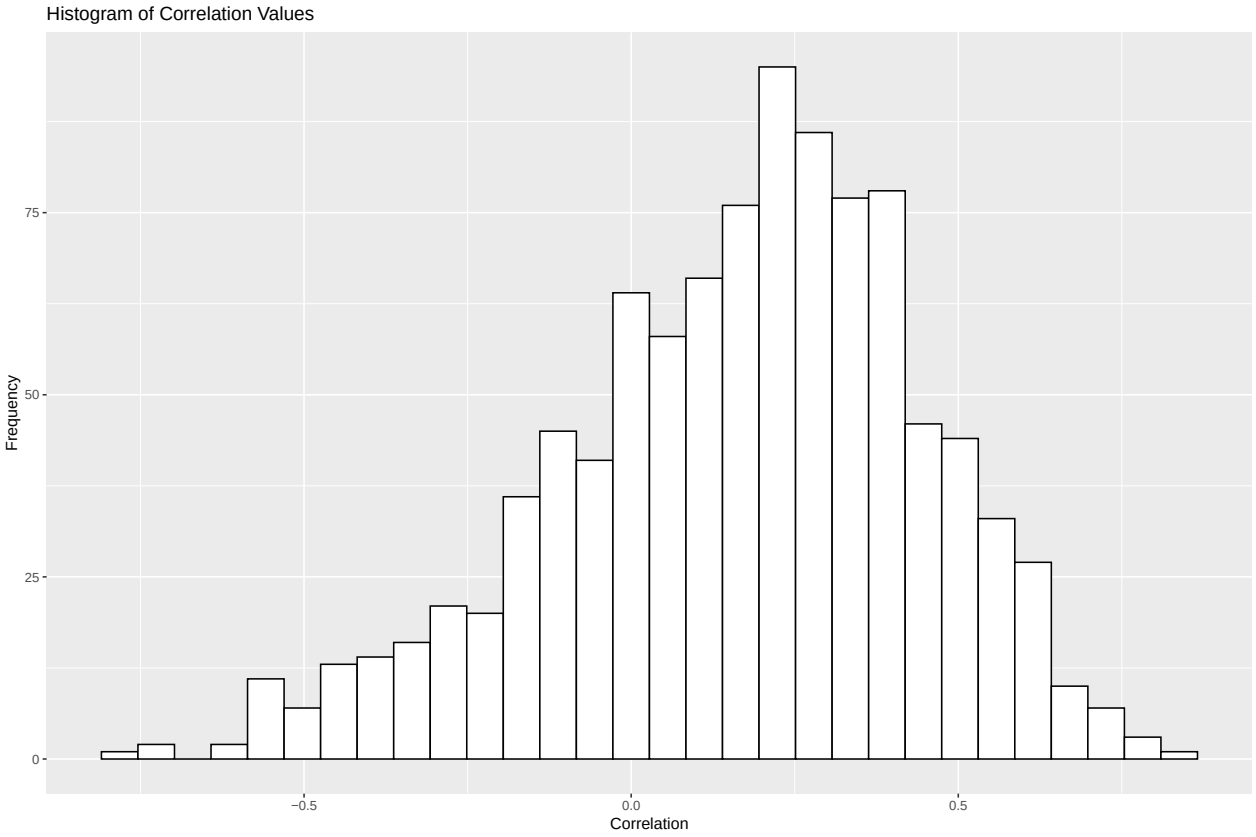

Selection Frequency of Features
Features
BMPR1B
GATA3
HOXA7
LEFTY2
CA12
PGK1
ITGB1
THY1
JAG1
EFNA3
HDAC2
HK2
ADCY9
C5orf38
AGTR1
CHIT1
HOXA9
ANXA9
RPS6KB1
CACNA1H
MMP9
CAMK2B
RPS6KA5
CBLC
SHC4
COL2A1
IGF1R
ECM2
SMURF2
TLX1
RASGRF2
HIF1A
CALML5
ZFYVE9
PMS2
NSD1
PARP1
G6PD
NR4A3
NPEPPS
NUPR1
HLA.DQA1
COL7A1
S100A7
PLA2G2A
FGFR2
SFRP1
SELE
CXCL8
APOE
0.0
0.2
0.4
Selection Frequency

##

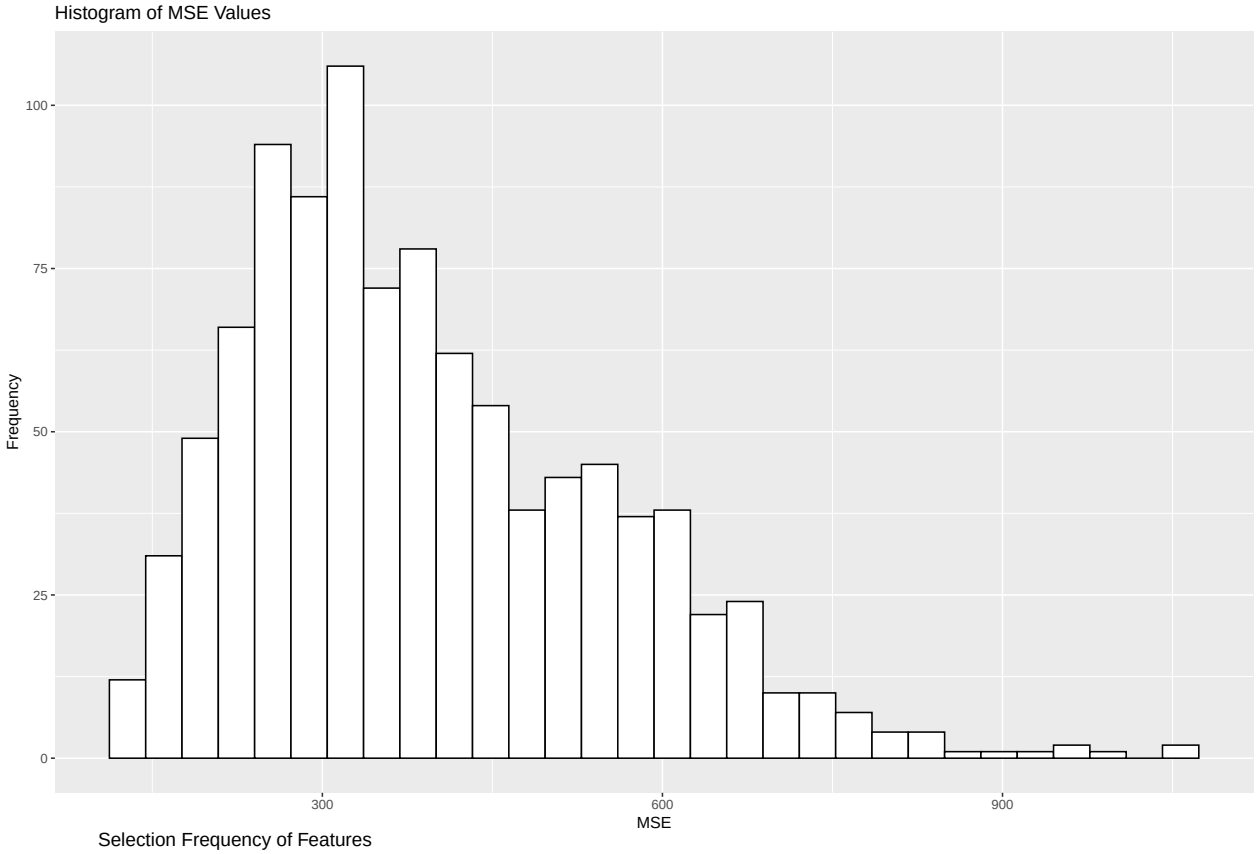
```
## Features selected 50% or more times:
## BMPR1B GATA3 HOXA7
## Top 20 featrues:
##  [1] "BMPR1B"  "GATA3"   "HOXA7"   "LEFTY2"  "CA12"    "PGK1"    "ITGB1"
##  [8] "THY1"    "JAG1"    "EFNA3"   "HDAC2"   "HK2"     "ADCY9"   "C5orf38"
## [15] "AGTR1"   "CHIT1"   "HOXA9"   "ANXA9"   "RPS6KB1" "CACNA1H"
```

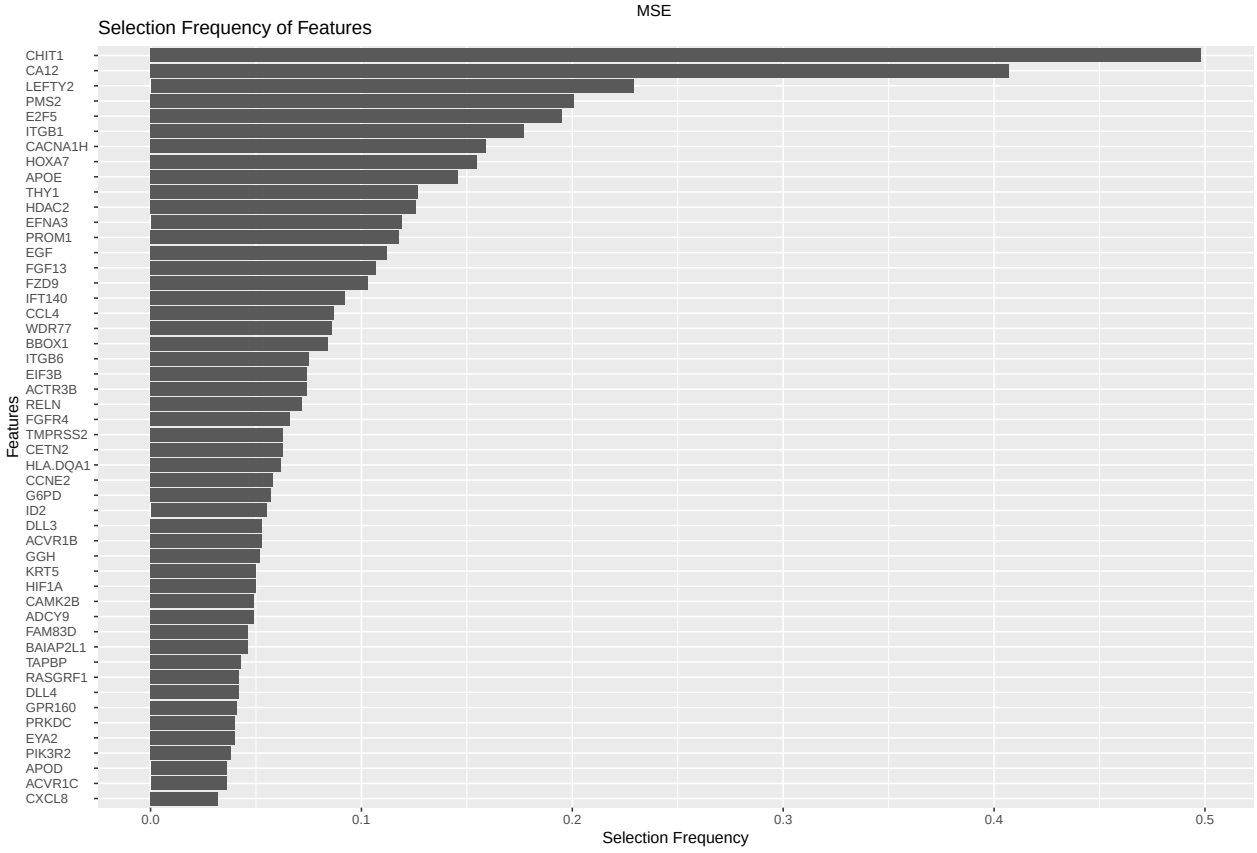
Mechnaistic + Residuals -> ROR-proliferation score (additive)

```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0
##
## CORRELATIONS RESULTS
## Mean: 0.16425
## Median: 0.198308
## Variance: 0.07775369
## st.dev.: 0.2788435
```




```
## MSE RESULTS
## Mean: 392.5436
## Median: 360.3419
## Variance: 25240.73
## st.dev.: 158.8733
```


Histogram of MSE Values
Frequency
100
75
50
25
0
300
600
900
MSE

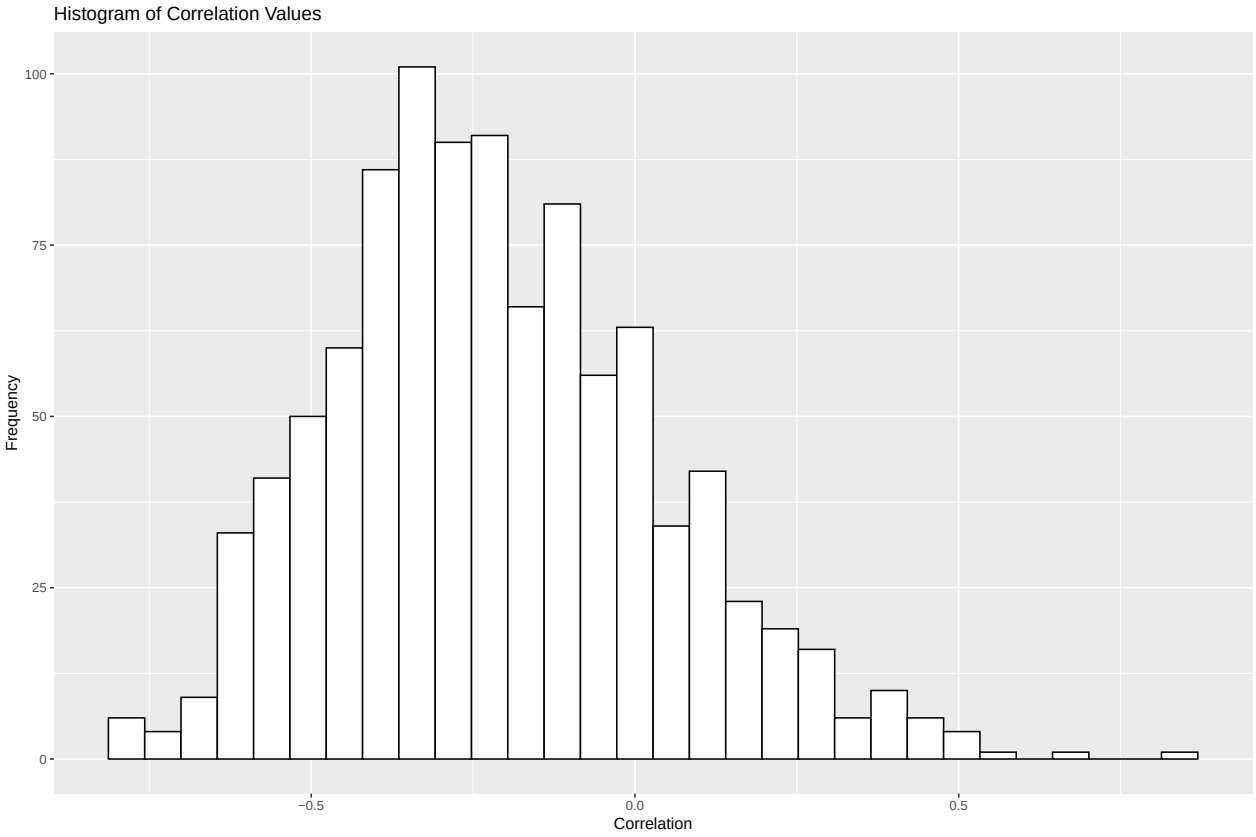

Selection Frequency of Features
Features
CHIT1
CA12
LEFTY2
PMS2
E2F5
ITGB1
CACNA1H
HOXA7
APOE
THY1
HDAC2
EFNA3
PROM1
EGF
FGF13
FZD9
IFT140
CCL4
WDR77
BBOX1
ITGB6
EIF3B
ACTR3B
RELN
FGFR4
TMPRSS2
CETN2
HLA.DQA1
CCNE2
G6PD
ID2
DLL3
ACVR1B
GGH
KRT5
HIF1A
CAMK2B
ADCY9
FAM83D
BAIAP2L1
TAPBP
RASGRF1
DLL4
GPR160
PRKDC
EYA2
PIK3R2
APOD
ACVR1C
CXCL8
0.0
0.1
0.2
0.3
0.4
0.5
Selection Frequency

##

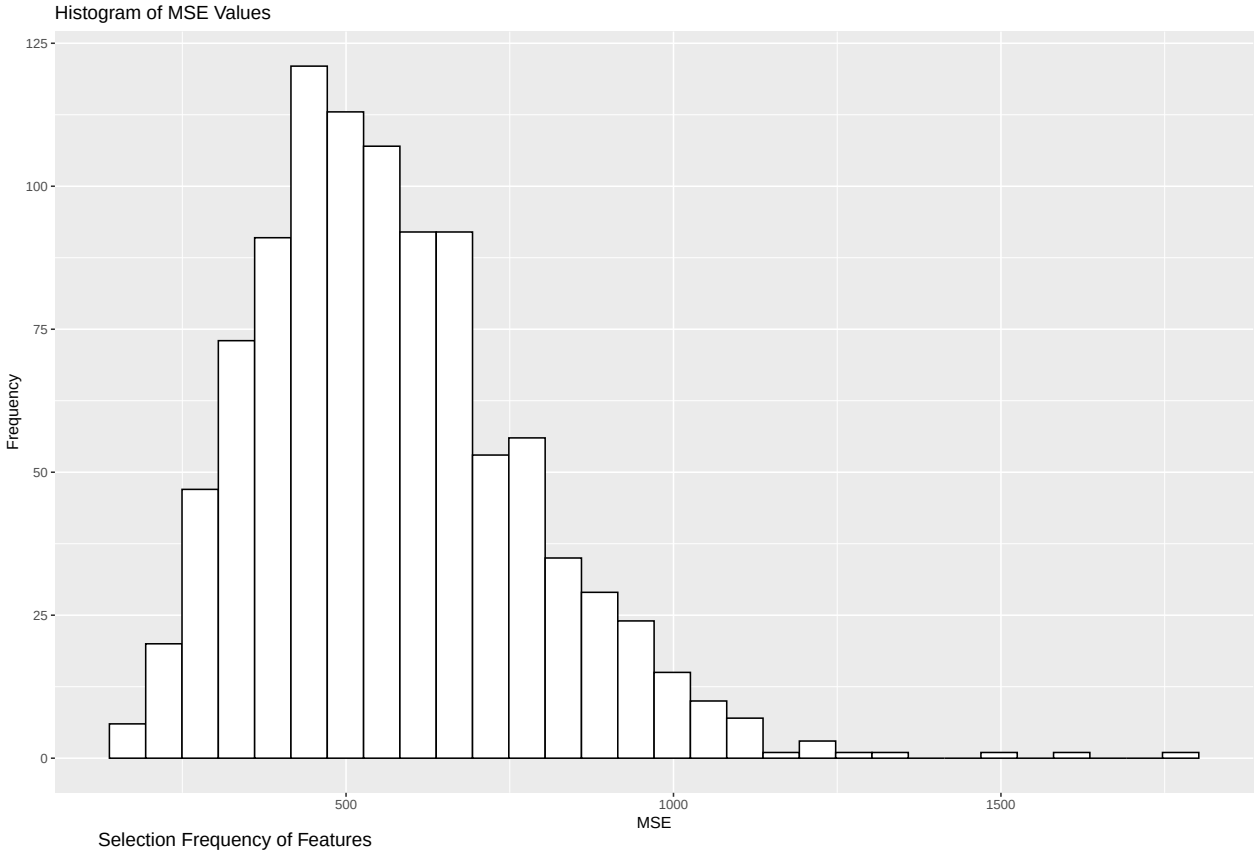
```
## Features selected 50% or more times:
##
## Top 20 featrues:
##  [1] "CHIT1"   "CA12"    "LEFTY2"  "PMS2"    "E2F5"    "ITGB1"   "CACNA1H"
##  [8] "HOXA7"   "APOE"    "THY1"    "HDAC2"   "EFNA3"   "PROM1"   "EGF"
## [15] "FGF13"   "FZD9"    "IFT140"  "CCL4"    "WDR77"   "BBOX1"
```

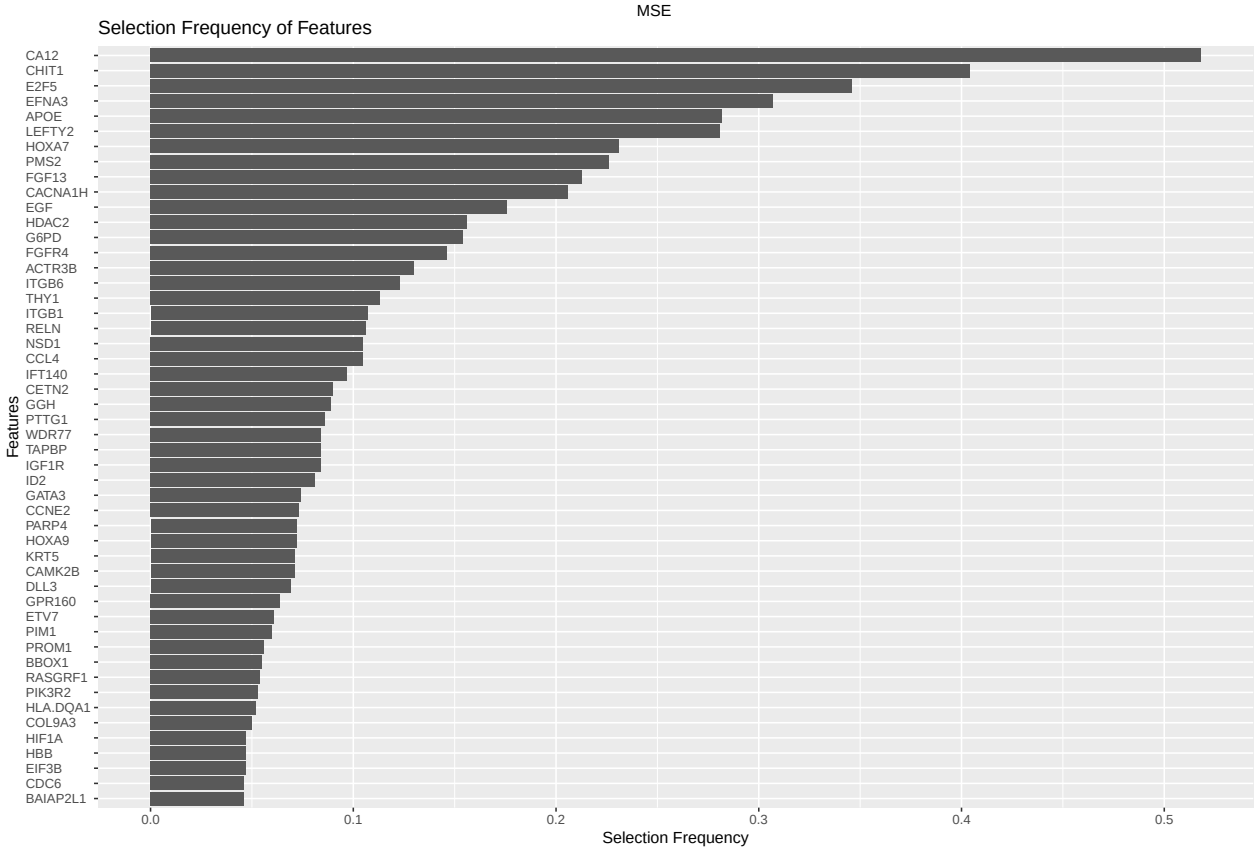
Mechnaistic + Residuals -> ROR-proliferation score (multiplicative)

```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0
##
## CORRELATIONS RESULTS
## Mean: -0.2145253
## Median: -0.2415987
## Variance: 0.06311304
## st.dev.: 0.2512231
```




```
## MSE RESULTS
## Mean: 568.8063
## Median: 541.2023
## Variance: 43510.24
## st.dev.: 208.5911
```


Histogram of MSE Values
Frequency
125
100
75
50
25
0
500
1000
1500
MSE


Selection Frequency of Features
Features
CA12
CHIT1
E2F5
EFNA3
APOE
LEFTY2
HOXA7
PMS2
FGF13
CACNA1H
EGF
HDAC2
G6PD
FGFR4
ACTR3B
ITGB6
THY1
ITGB1
RELN
NSD1
CCL4
IFT140
CETN2
GGH
PTTG1
WDR77
TAPBP
IGF1R
ID2
GATA3
CCNE2
PARP4
HOXA9
KRT5
CAMK2B
DLL3
GPR160
ETV7
PIM1
PROM1
BBOX1
RASGRF1
PIK3R2
HLA.DQA1
COL9A3
HIF1A
HBB
EIF3B
CDC6
BAIAP2L1
0.0
0.1
0.2
0.3
0.4
0.5
Selection Frequency

##

```
## Features selected 50% or more times:
## CA12
## Top 20 featrues:
##  [1] "CA12"    "CHIT1"   "E2F5"    "EFNA3"   "APOE"    "LEFTY2"  "HOXA7"
##  [8] "PMS2"    "FGF13"   "CACNA1H" "EGF"     "HDAC2"   "G6PD"    "FGFR4"
## [15] "ACTR3B"  "ITGB6"   "THY1"    "ITGB1"   "RELN"    "CCL4"
```

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

Model	cor_mean	sd_cor	MSE_mean	MSE_sd
lasso 6 genes	0.0919663	0.3068209	0.1655931	0.0718379
lasso 771 genes	0.4737037	0.2310209	0.0620913	0.0753056
Nodes	0.2842257	0.2768458	0.1560308	0.0762883
Residual additive	0.4633095	0.2227105	0.1331785	0.0654065
Residual multiplicative	0.4028471	0.2302632	0.1455819	0.0680617

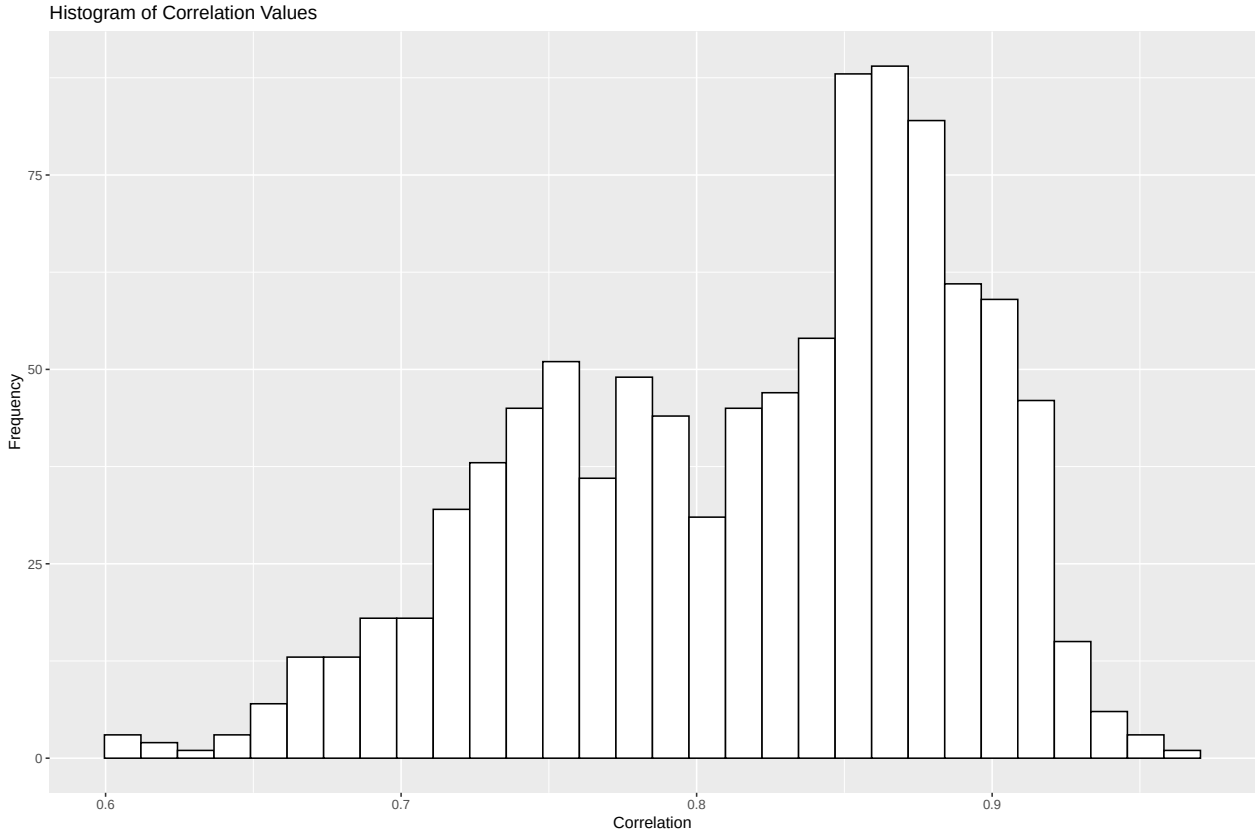
Summery results: lasso ROR+proliferation score (repeated cross-validation)

Model	cor_mean	sd_cor	MSE_mean	MSE_sd
lasso 6 genes	-0.4822298	0.1724985	374.1519	144.1559
lasso 771 genes	0.0806237	0.2767153	393.8069	159.6451
Nodes	0.1806504	0.2854892	380.1157	164.5869
Residual additive	0.1642500	0.2788435	392.5436	158.8733
Residual multiplicative	-0.2145253	0.2512231	568.8063	208.5911

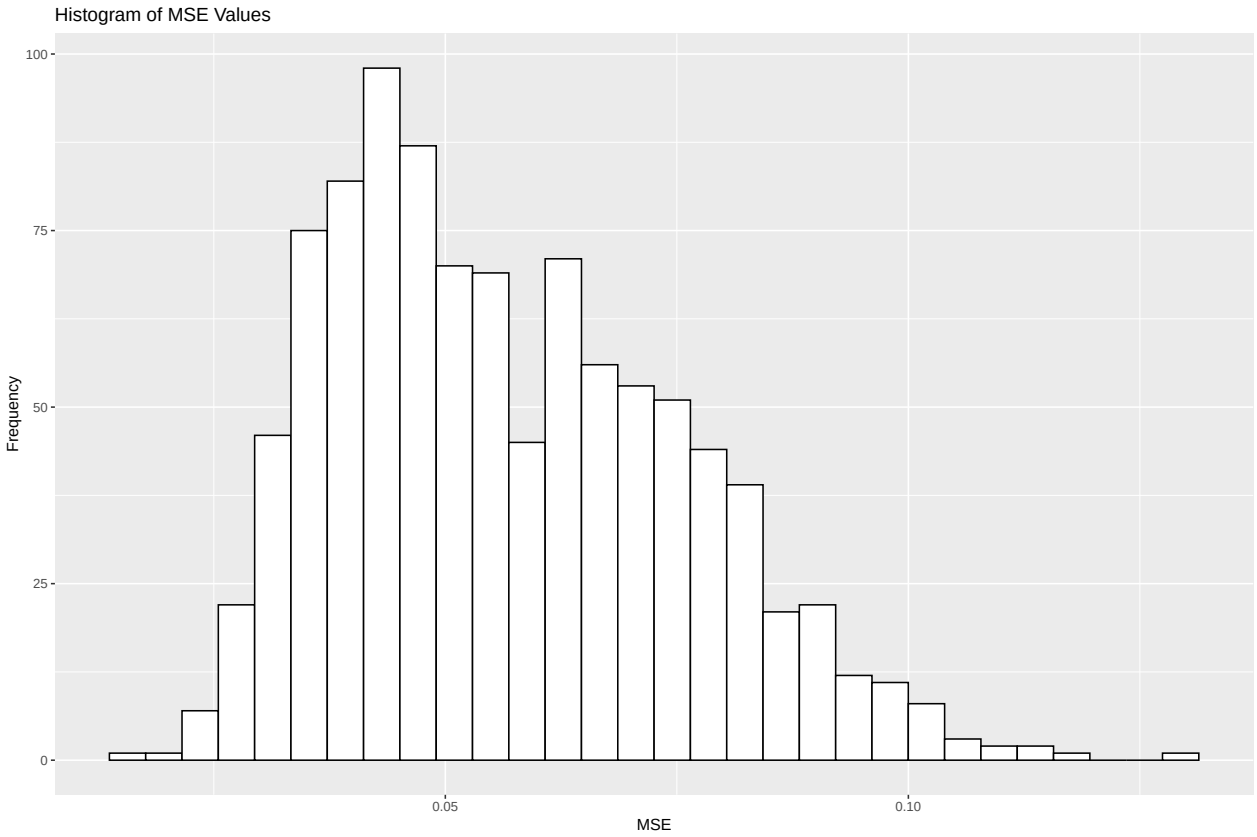
Ridge bootstrap

771 genes -> proliferation score

```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0
##
## CORRELATIONS RESULTS
## Mean: 0.8189568
## Median: 0.8355063
## Variance: 0.005044925
## st.dev.: 0.07102764
```

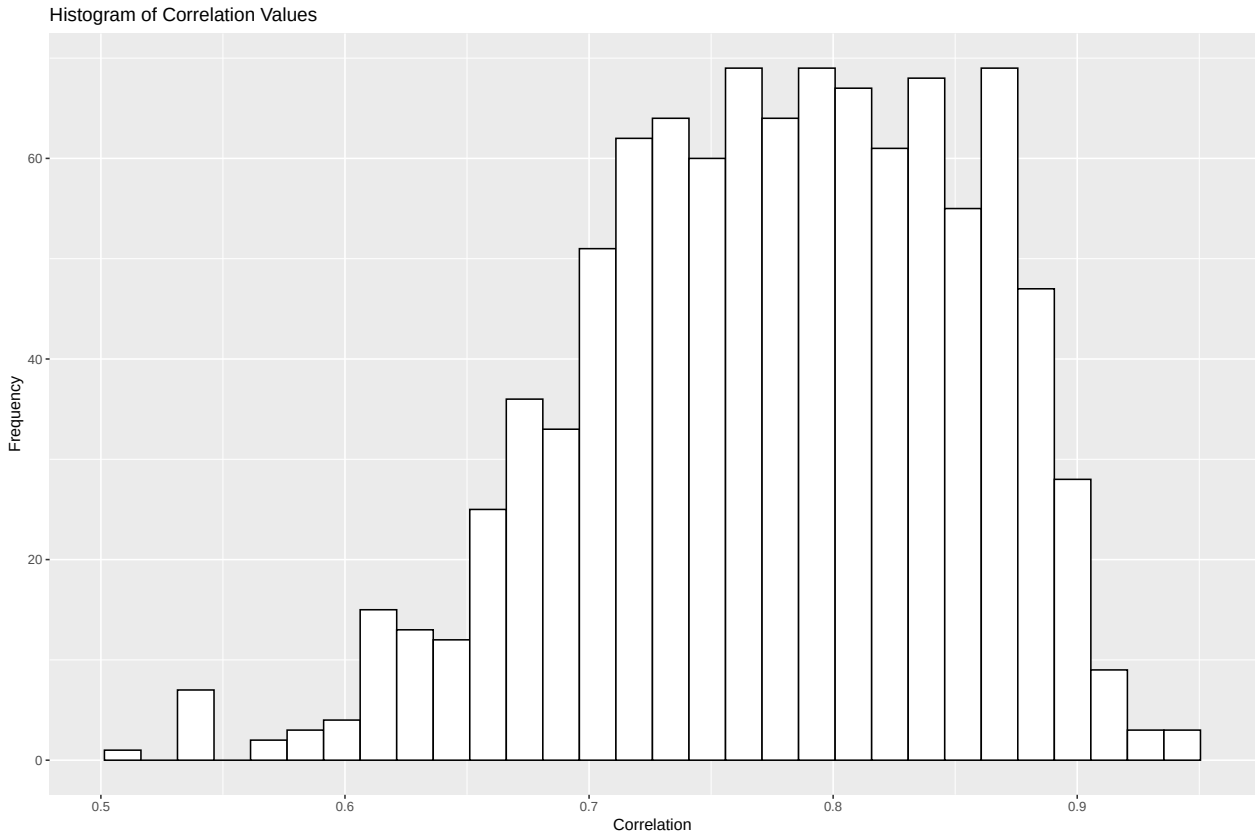



```
## MSE RESULTS
## Mean: 0.05665241
## Median: 0.0532761
## Variance: 0.0003488416
## st.dev.: 0.0186773
```

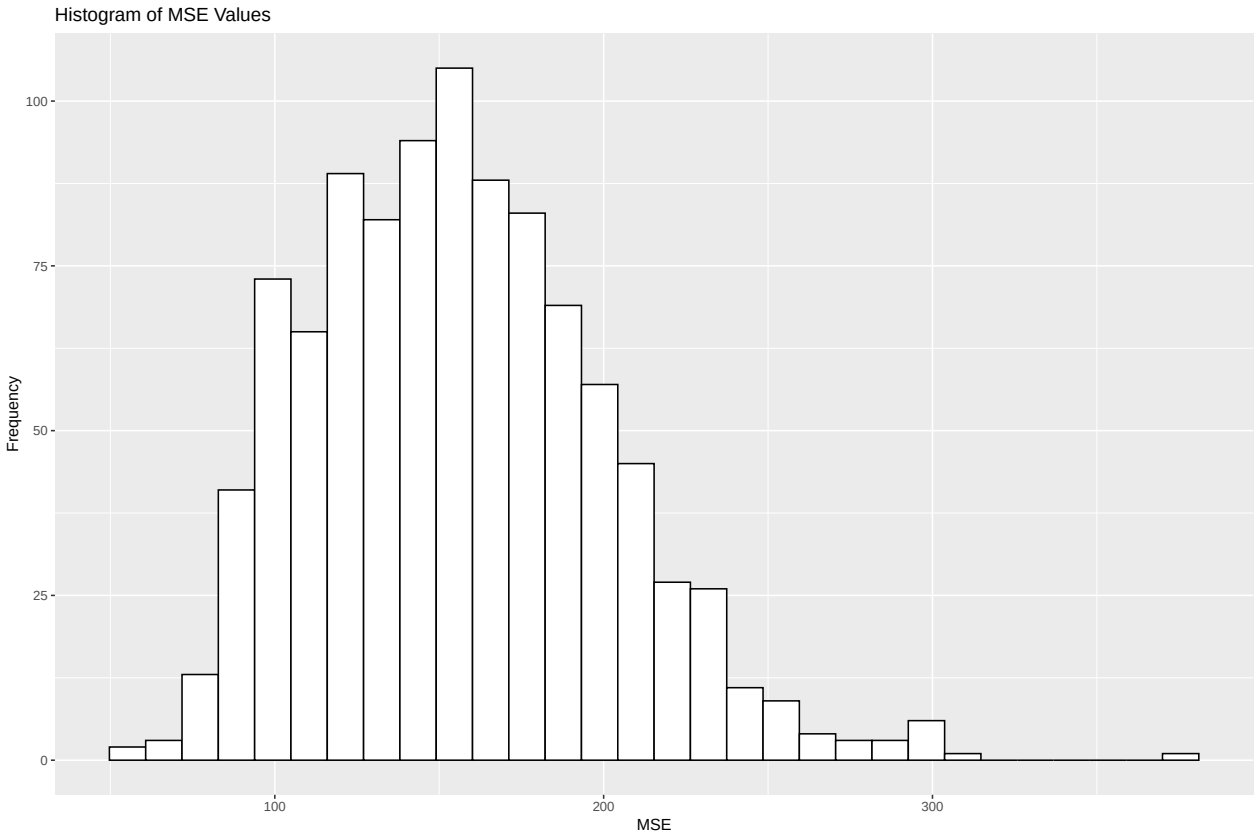


771 genes -> ROR-proliferation score

```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0
##
## CORRELATIONS RESULTS
## Mean: 0.7761924
## Median: 0.7811637
## Variance: 0.006001423
## st.dev.: 0.07746885
```



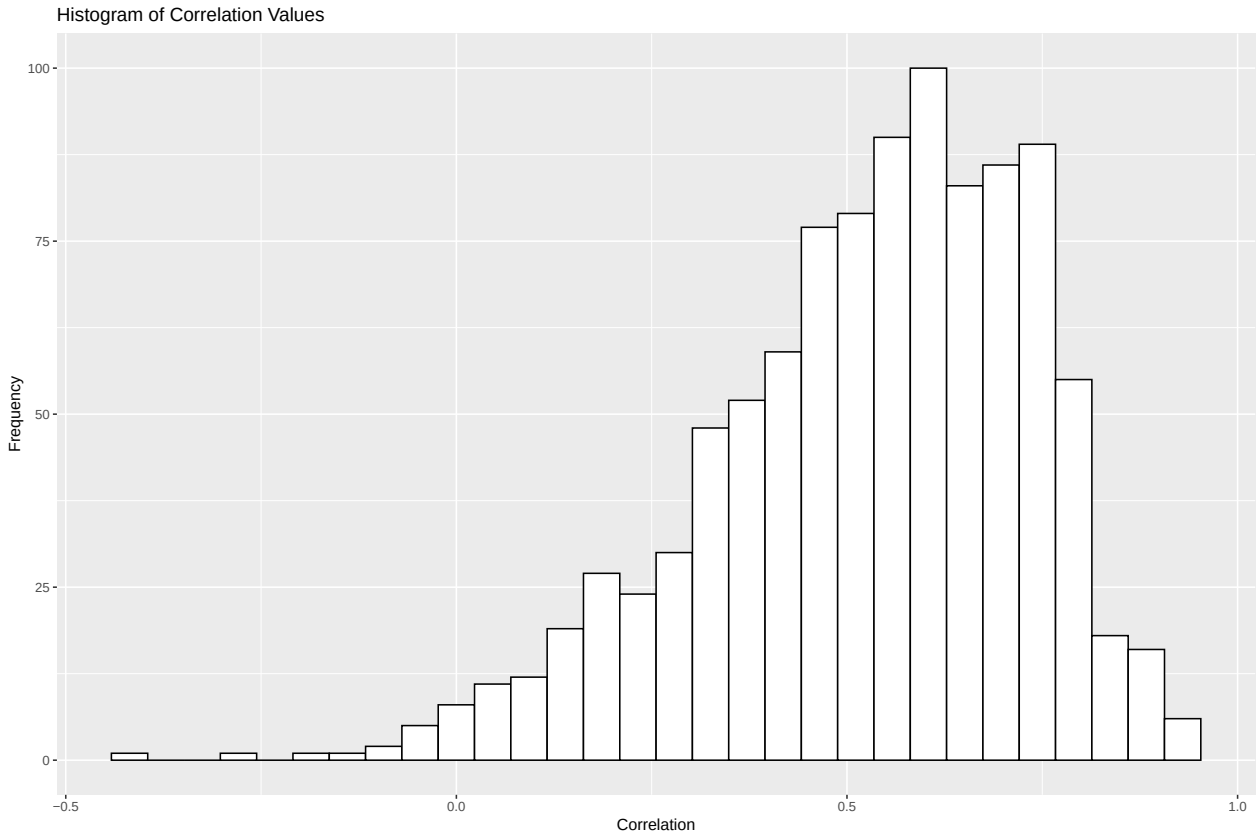

```
## MSE RESULTS
## Mean: 156.065
## Median: 154.0679
## Variance: 1952.025
## st.dev.: 44.18173
```



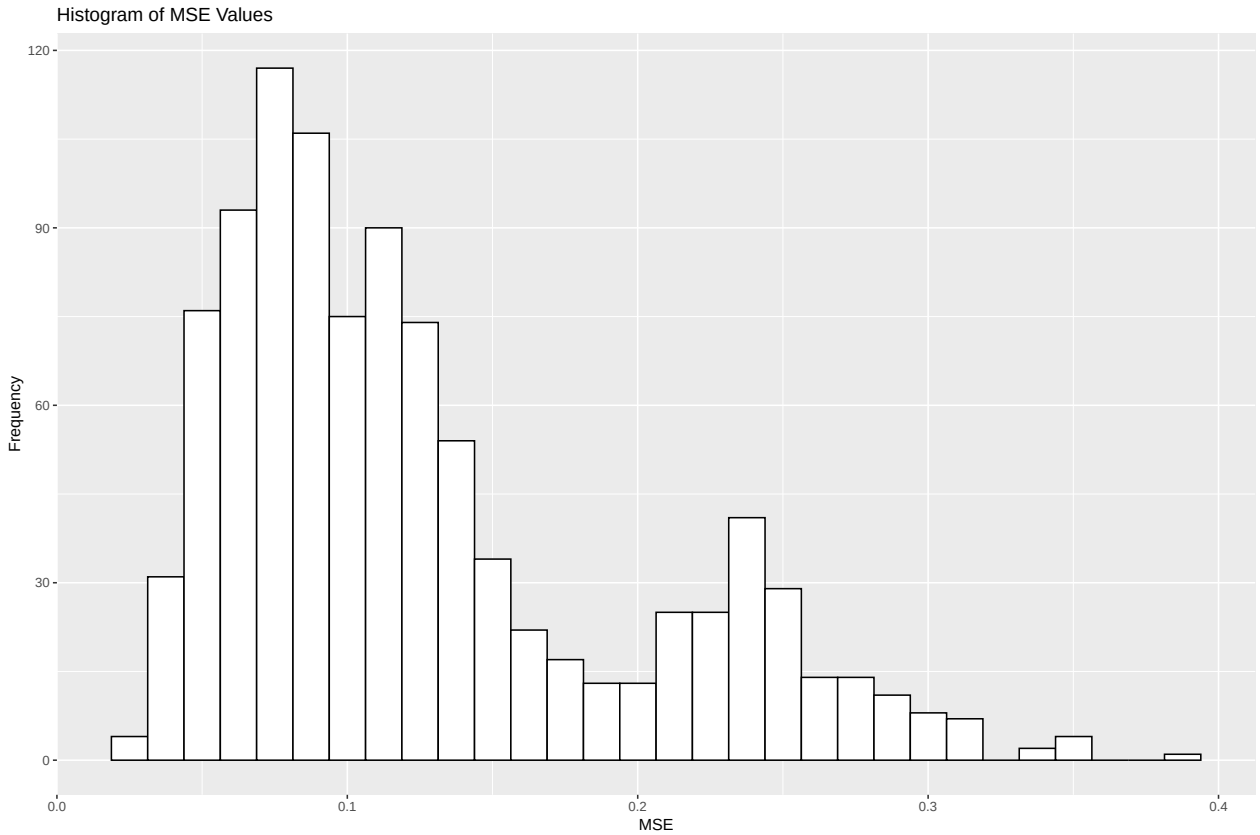
Ridge repeated cross-validation

771 genes -> proliferation score

```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0
##
## CORRELATIONS RESULTS
## Mean: 0.5268342
## Median: 0.5562175
## Variance: 0.04291266
## st.dev.: 0.2071537
```

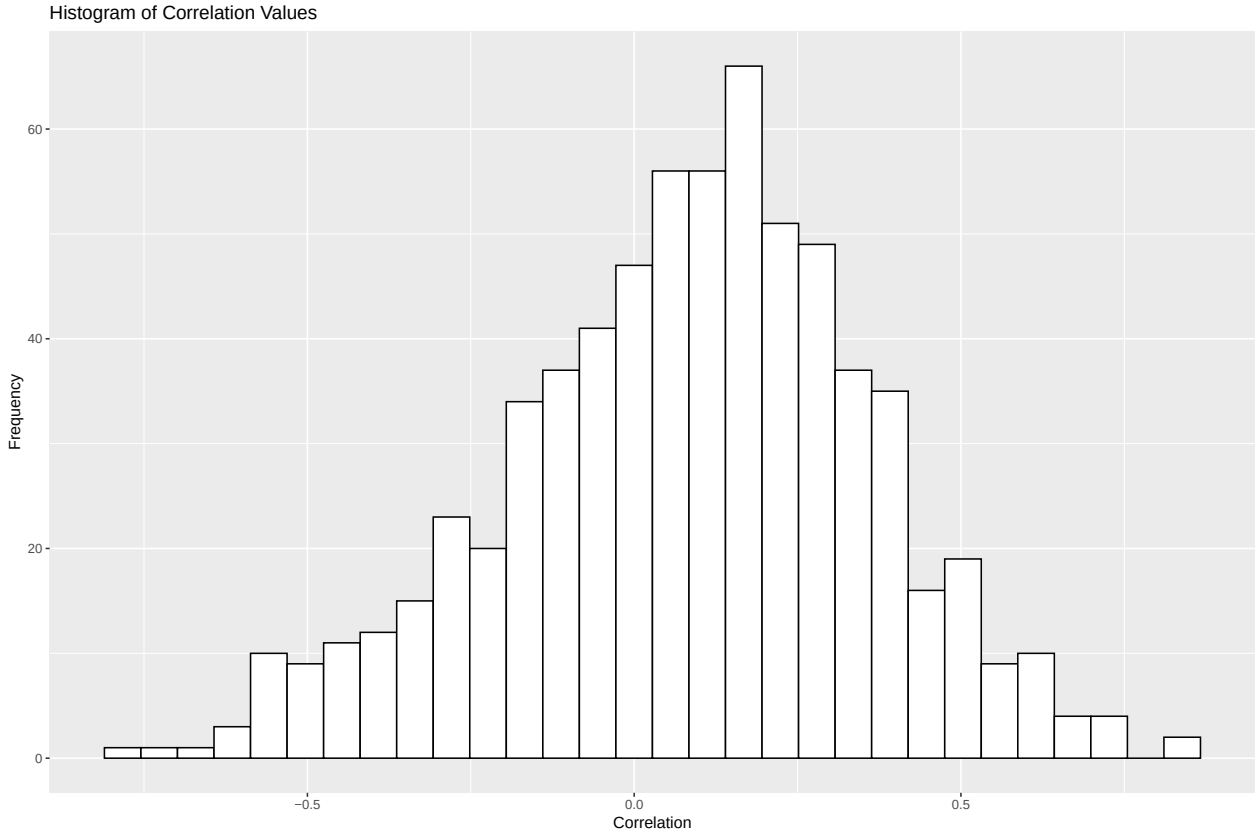



```
## MSE RESULTS
## Mean: 0.1256548
## Median: 0.1059589
## Variance: 0.004890454
## st.dev.: 0.06993178
```

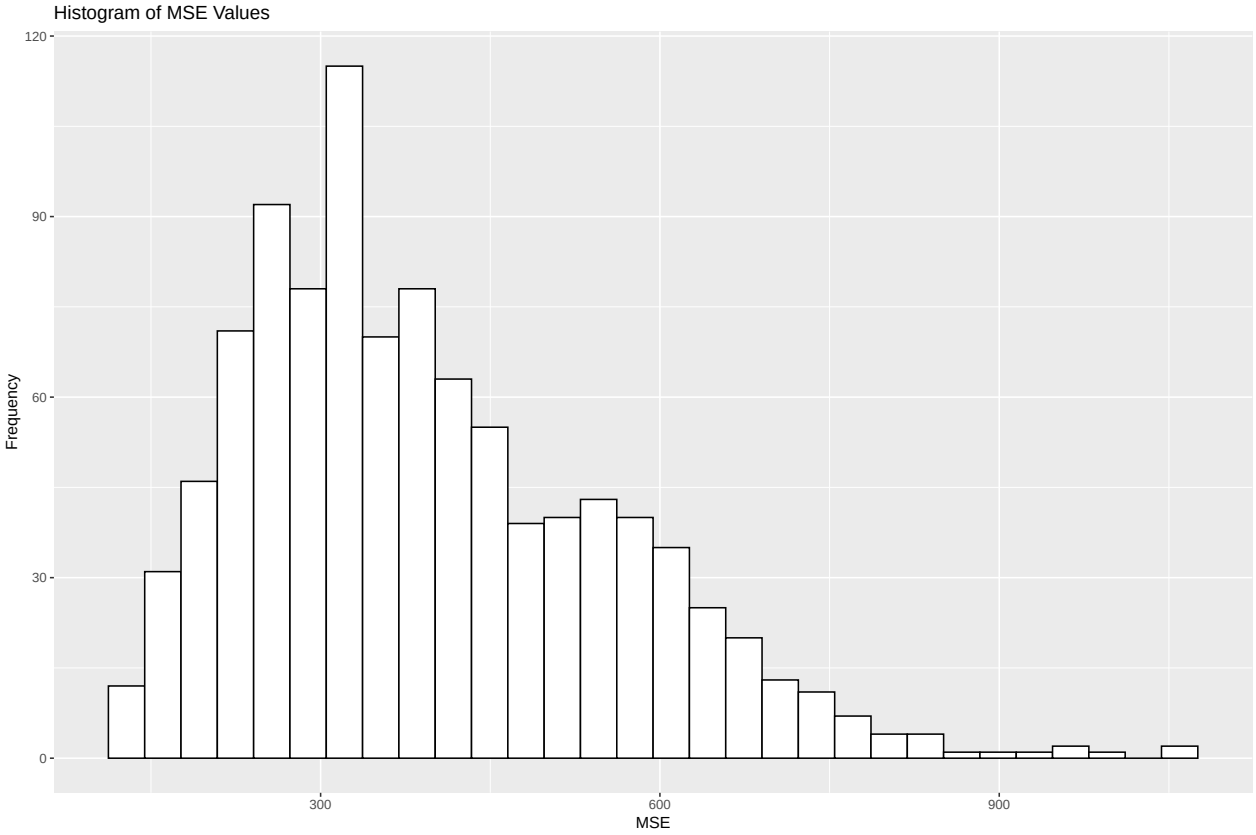


771 genes -> ROR-proliferation score

```
## 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
```



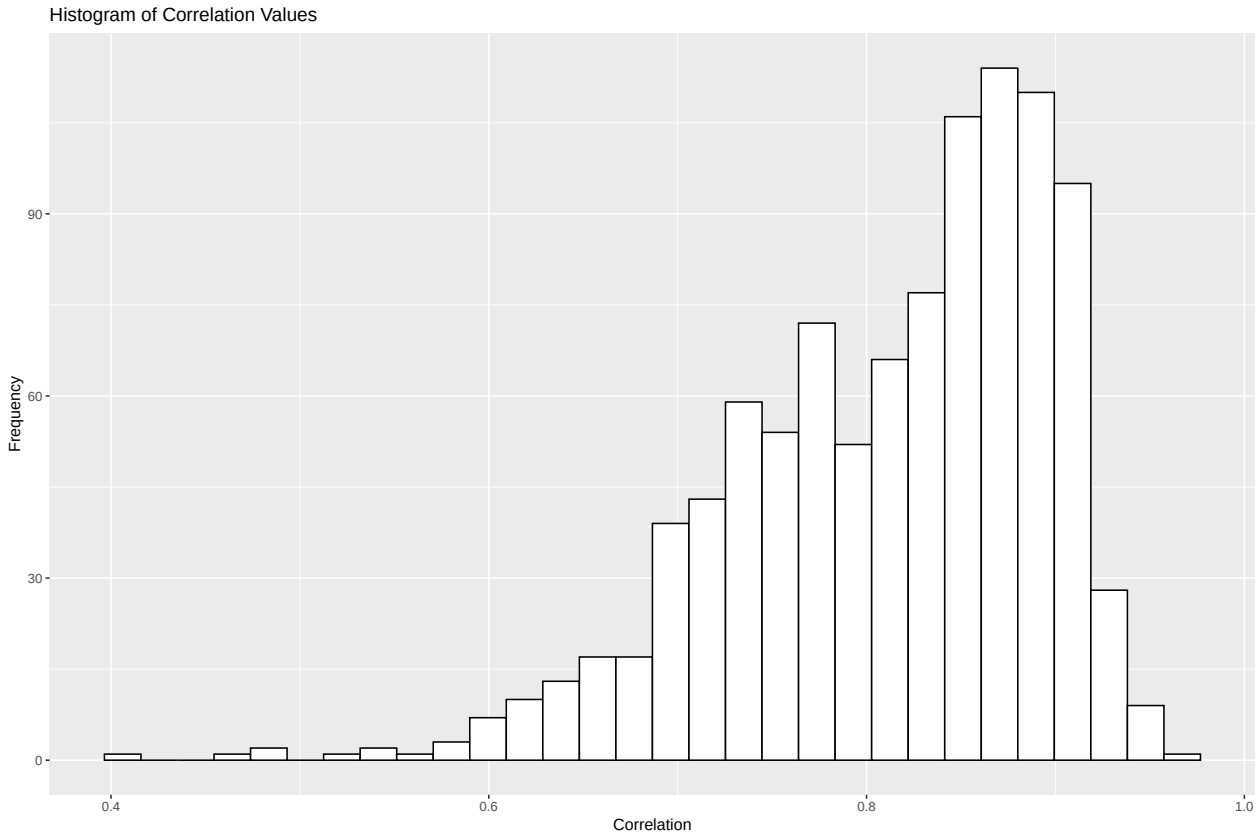
Summery results: ridge 771 genes bootstrap and repeated cross-validation

Model	cor_mean	sd_cor	MSE_mean	MSE_sd
prolif boot	0.8189568	0.0710276	0.0566524	0.0186773
ROR-prolif boot	0.7761924	0.0774688	156.0649552	44.1817261
prolif rep cross-val	0.5268342	0.2071537	0.1256548	0.0699318
ROR-prolif rep cross-val	0.0806237	0.2767153	393.8068910	159.6450765

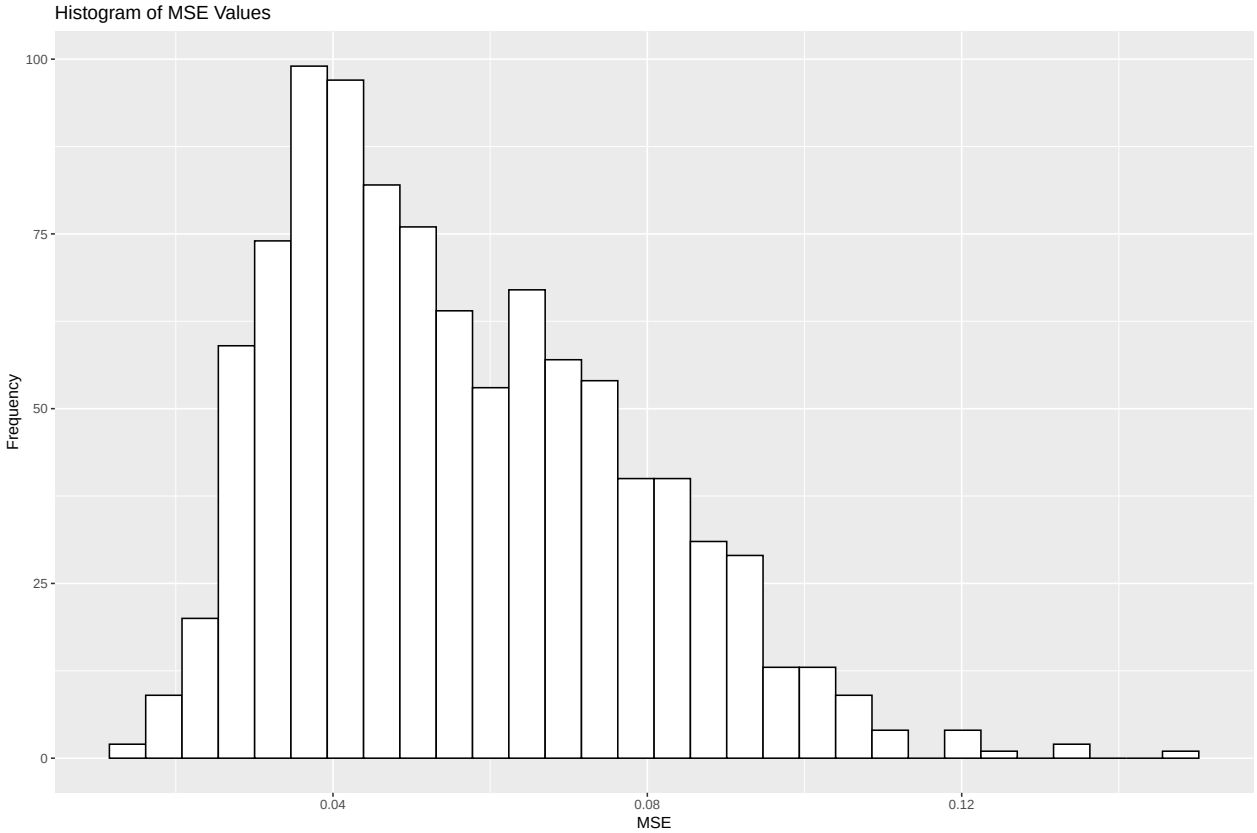
Elastic Net - bootstrap

771 genes -> proliferation score

```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0
##
## CORRELATIONS RESULTS
## Mean: 0.8125591
## Median: 0.8324809
## Variance: 0.007038791
## st.dev.: 0.0838975
```

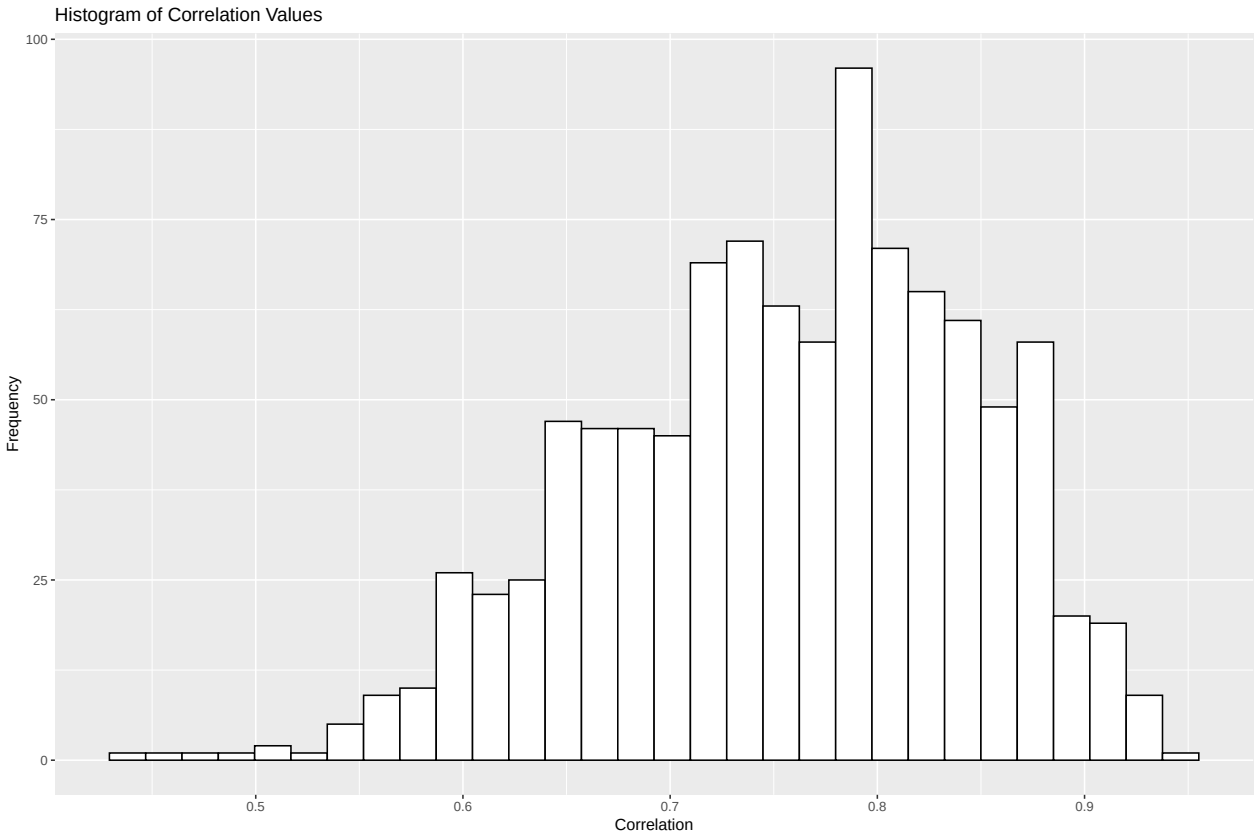



```
## MSE RESULTS
## Mean: 0.0558955
## Median: 0.05152763
## Variance: 0.0004710269
## st.dev.: 0.02170316
```

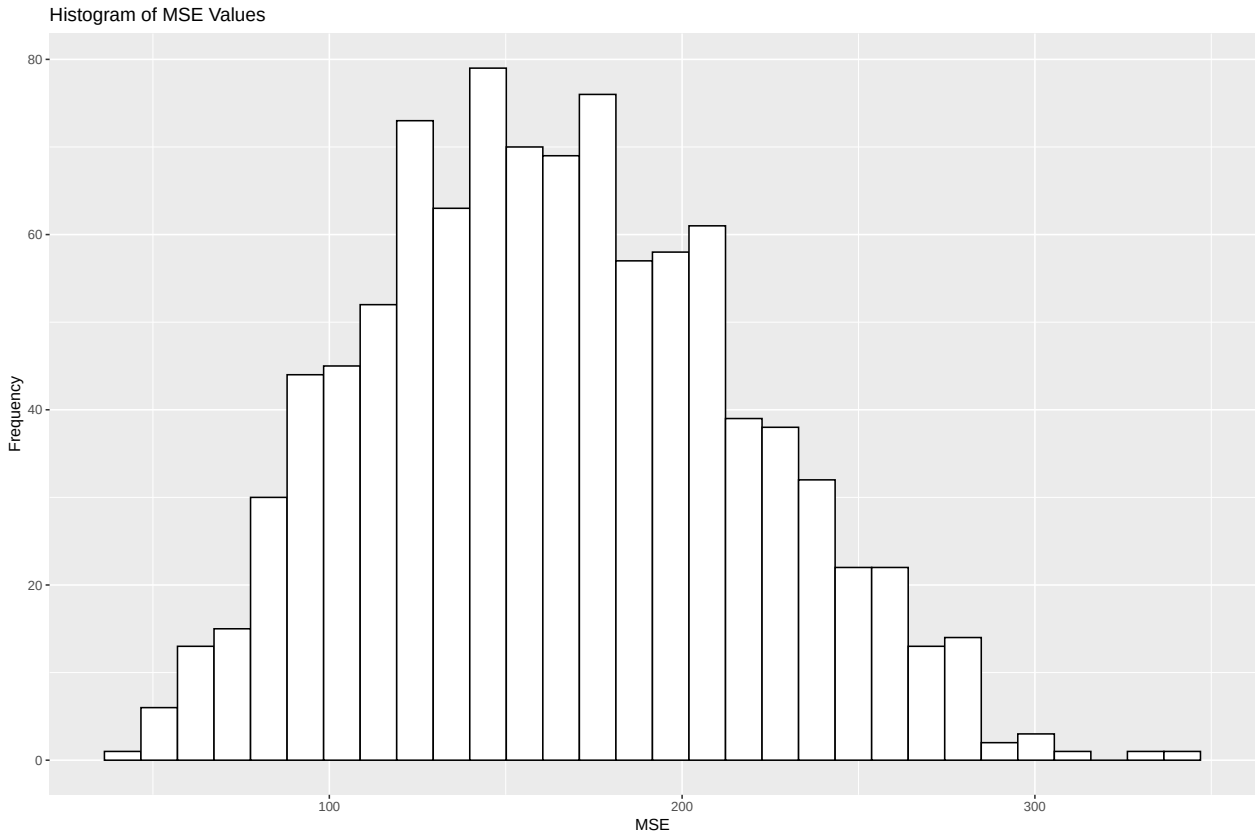


771 genes -> ROR-proliferation score

```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0
##
## CORRELATIONS RESULTS
## Mean: 0.7565123
## Median: 0.7644687
## Variance: 0.007887981
## st.dev.: 0.08881431
```



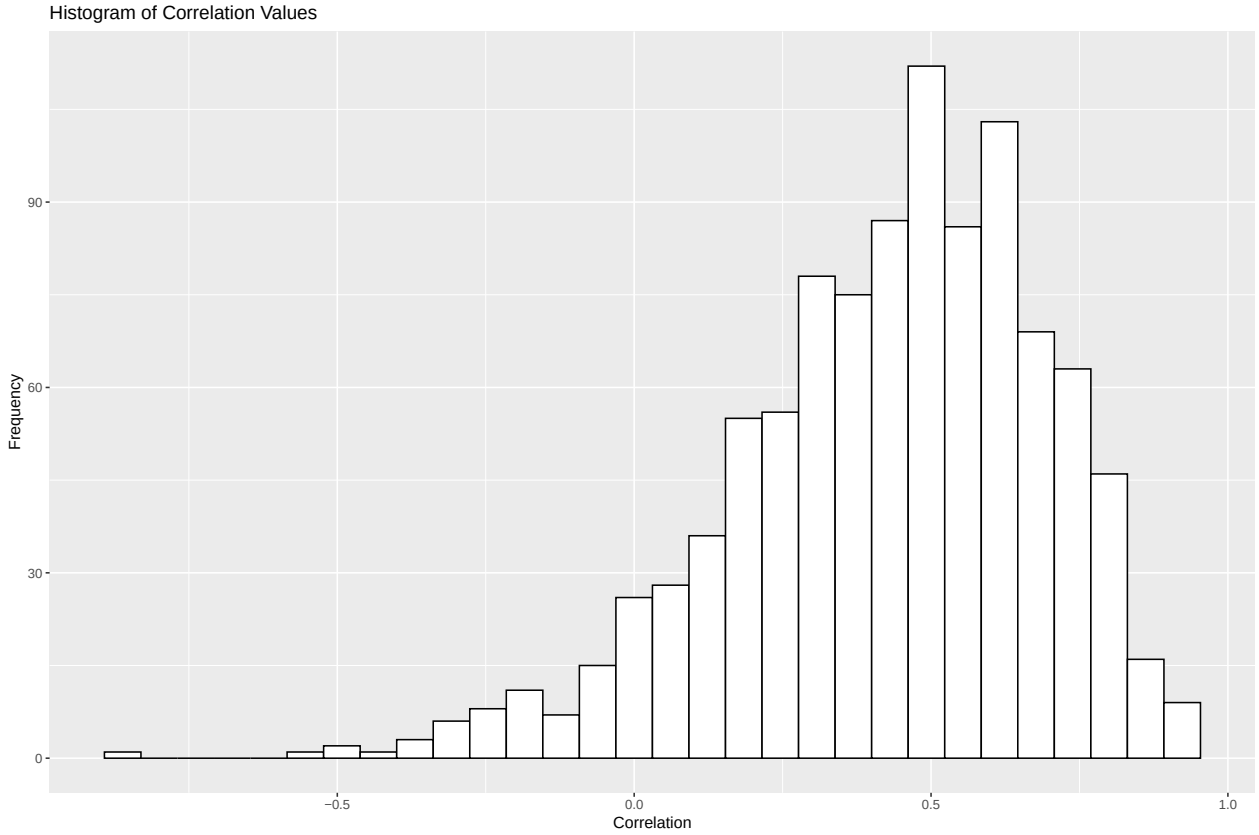

```
## MSE RESULTS
## Mean: 164.4116
## Median: 162.2643
## Variance: 2785.22
## st.dev.: 52.77518
```



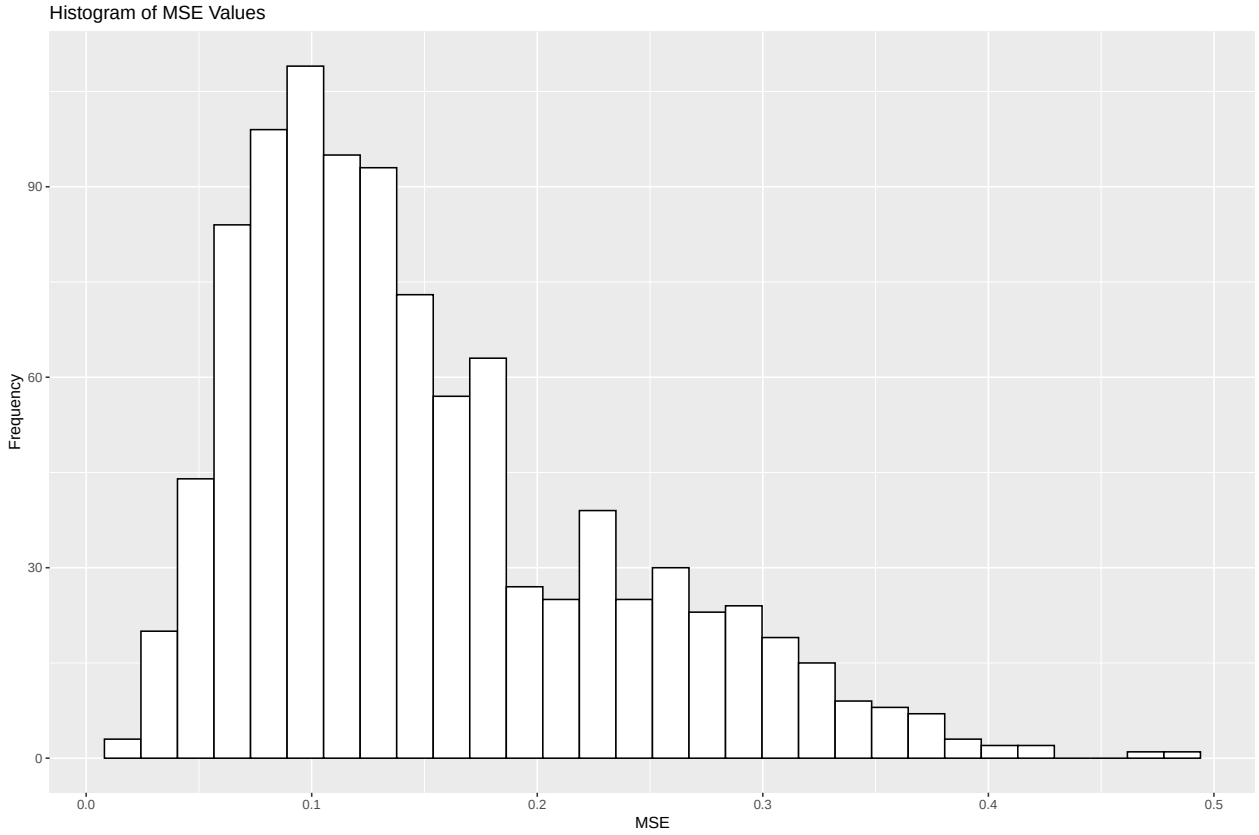
Elastic Net: cross-validation

771 genes -> proliferation score

```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0
##
## CORRELATIONS RESULTS
## Mean: 0.427189
## Median: 0.4644365
## Variance: 0.06953924
## st.dev.: 0.2637029
```

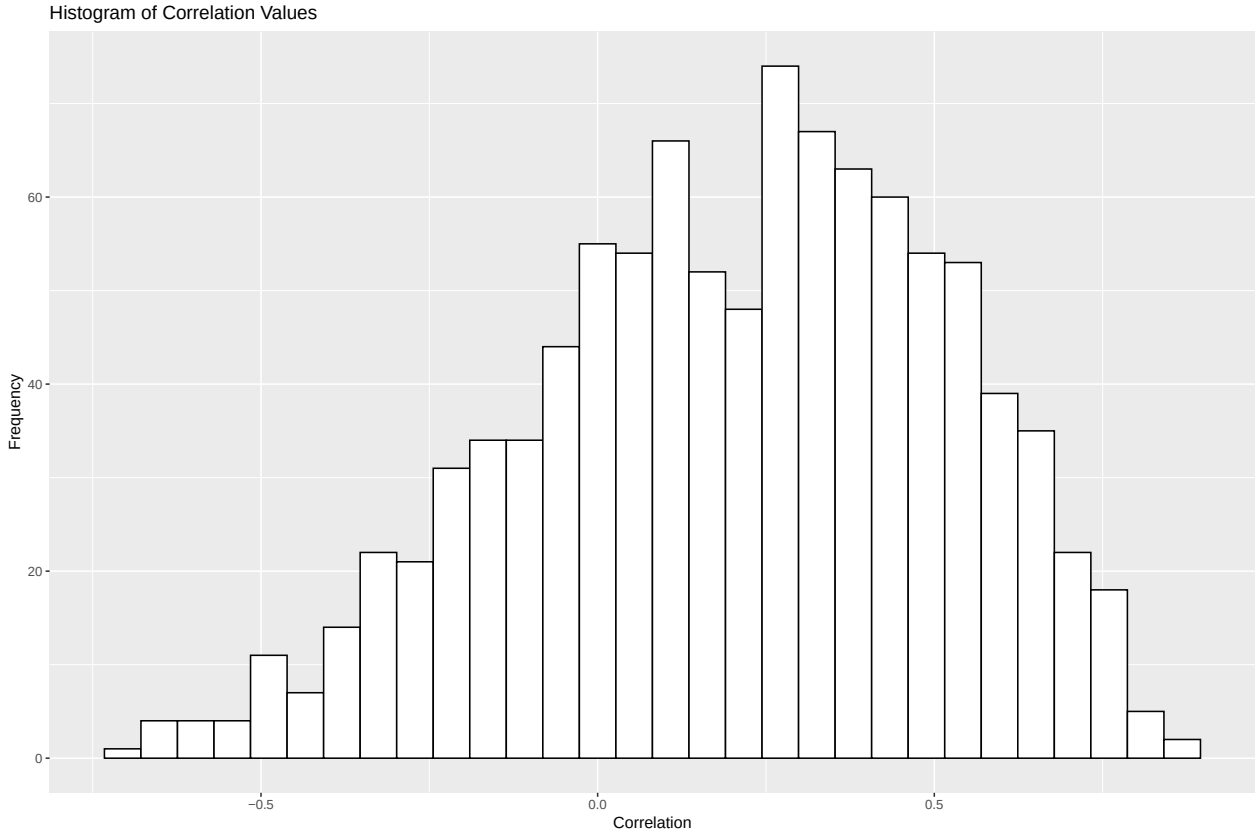



```
## MSE RESULTS
## Mean: 0.1501025
## Median: 0.1284895
## Variance: 0.006763425
## st.dev.: 0.08224004
```

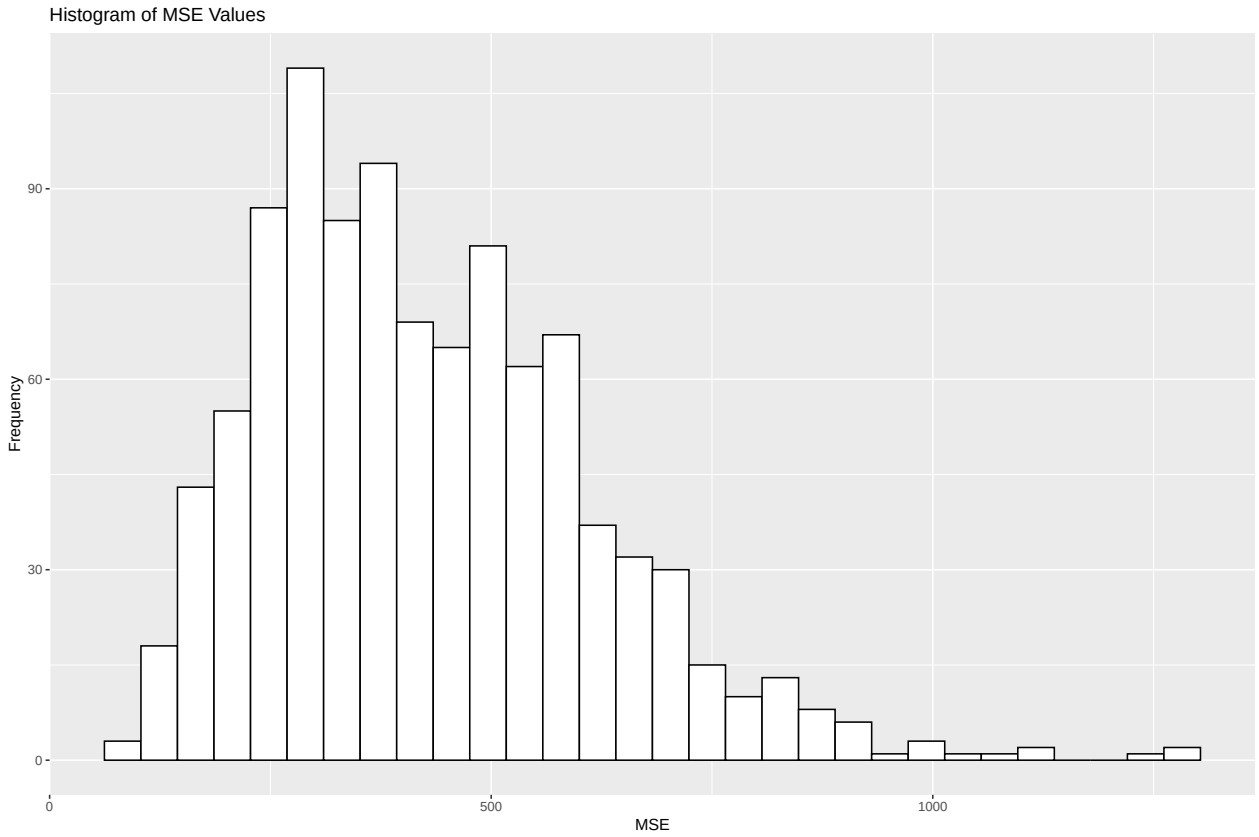


771 genes -> ROR-proliferation score

```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0.002
##
## CORRELATIONS RESULTS
## Mean: 0.2049341
## Median: 0.226514
## Variance: 0.0968435
## st.dev.: 0.3111969
```




```
## MSE RESULTS
## Mean: 427.3517
## Median: 396.2622
## Variance: 34687.75
## st.dev.: 186.2465
```



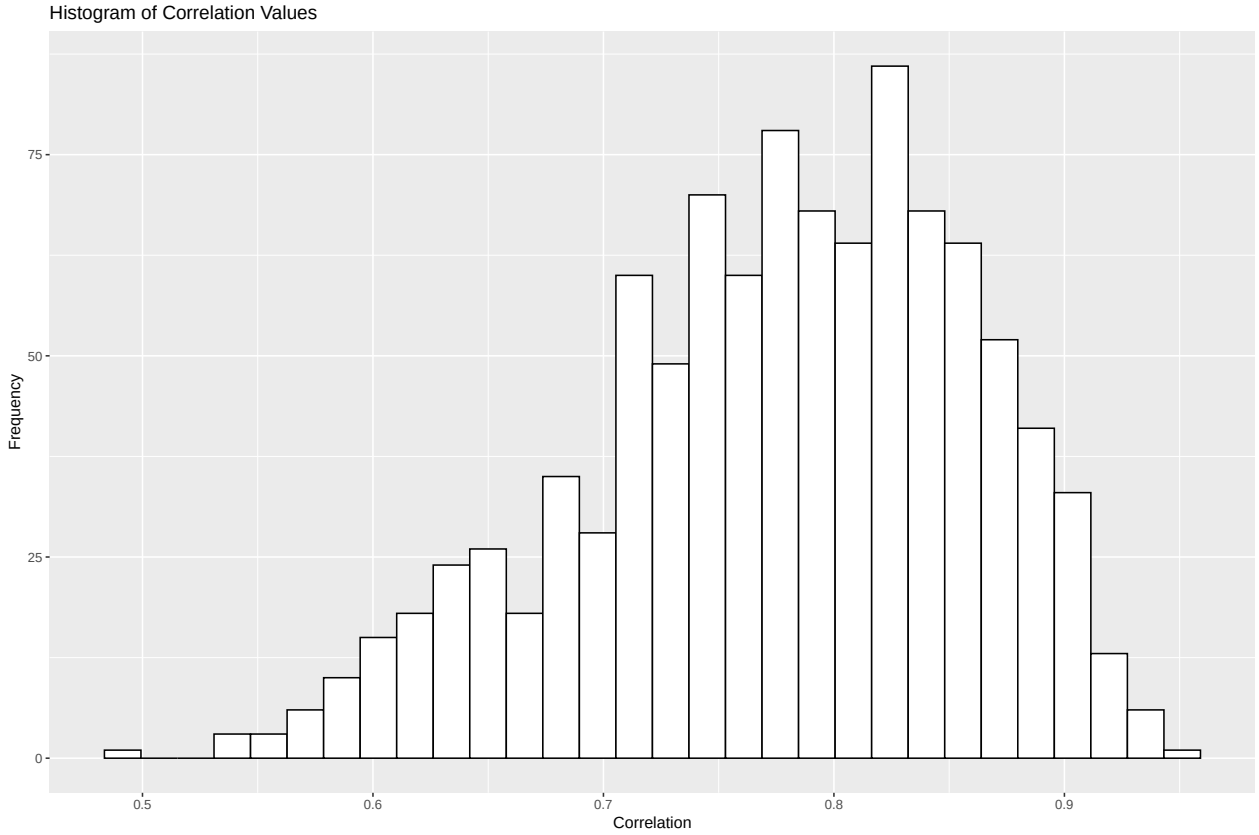
Summery results: elastic net 771 genes bootstrap and repeated cross-validation

Model	cor_mean	sd_cor	MSE_mean	MSE_sd
prolif boot	0.8125591	0.0838975	0.0558955	0.0217032
ROR-prolif boot	0.7565123	0.0888143	164.4116160	52.7751849
prolif rep cross-val	0.4271890	0.2637029	0.1501025	0.0822400
ROR-prolif rep cross-val	0.2049341	0.3111969	427.3517412	186.2464725

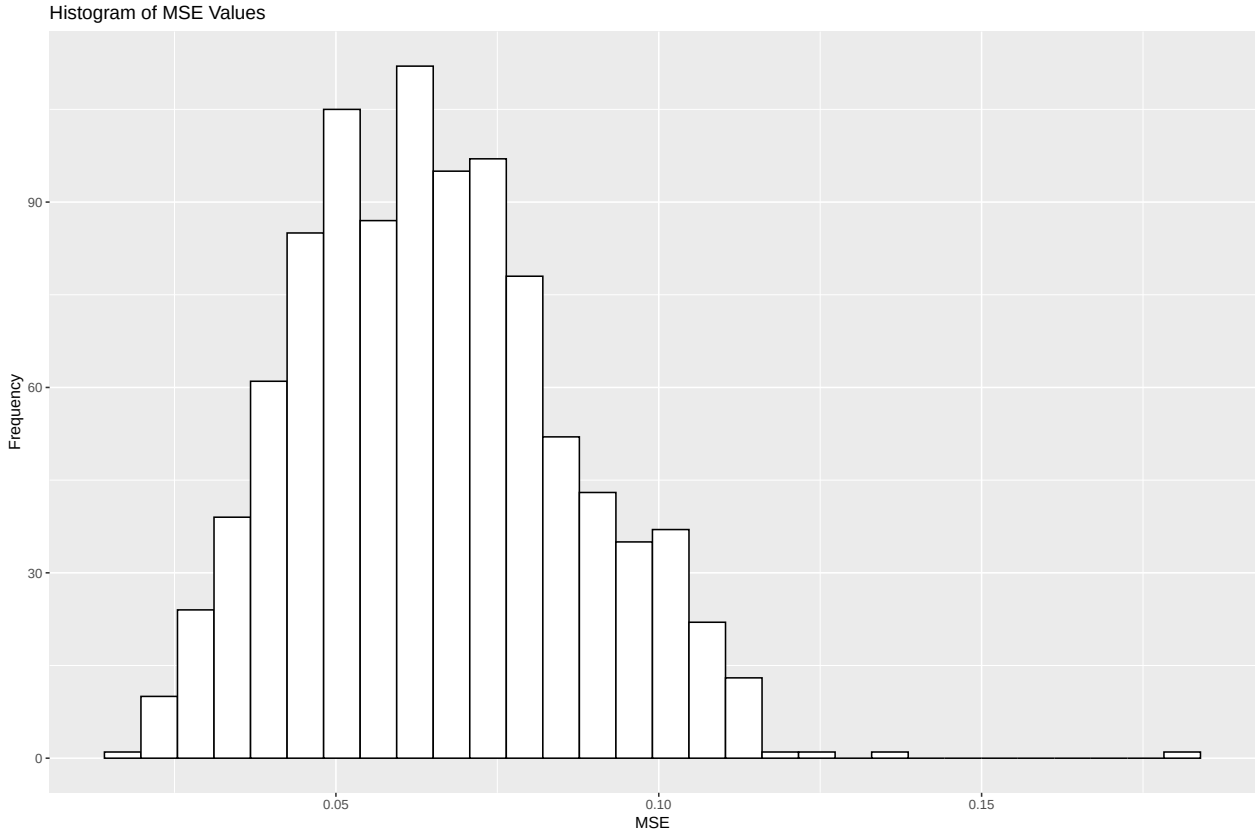
Boosting with stumps as base learner - bootstrap

771 genes -> proliferation score

```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0
##
## CORRELATIONS RESULTS
## Mean: 0.7760479
## Median: 0.7841718
## Variance: 0.006853002
## st.dev.: 0.08278286
```

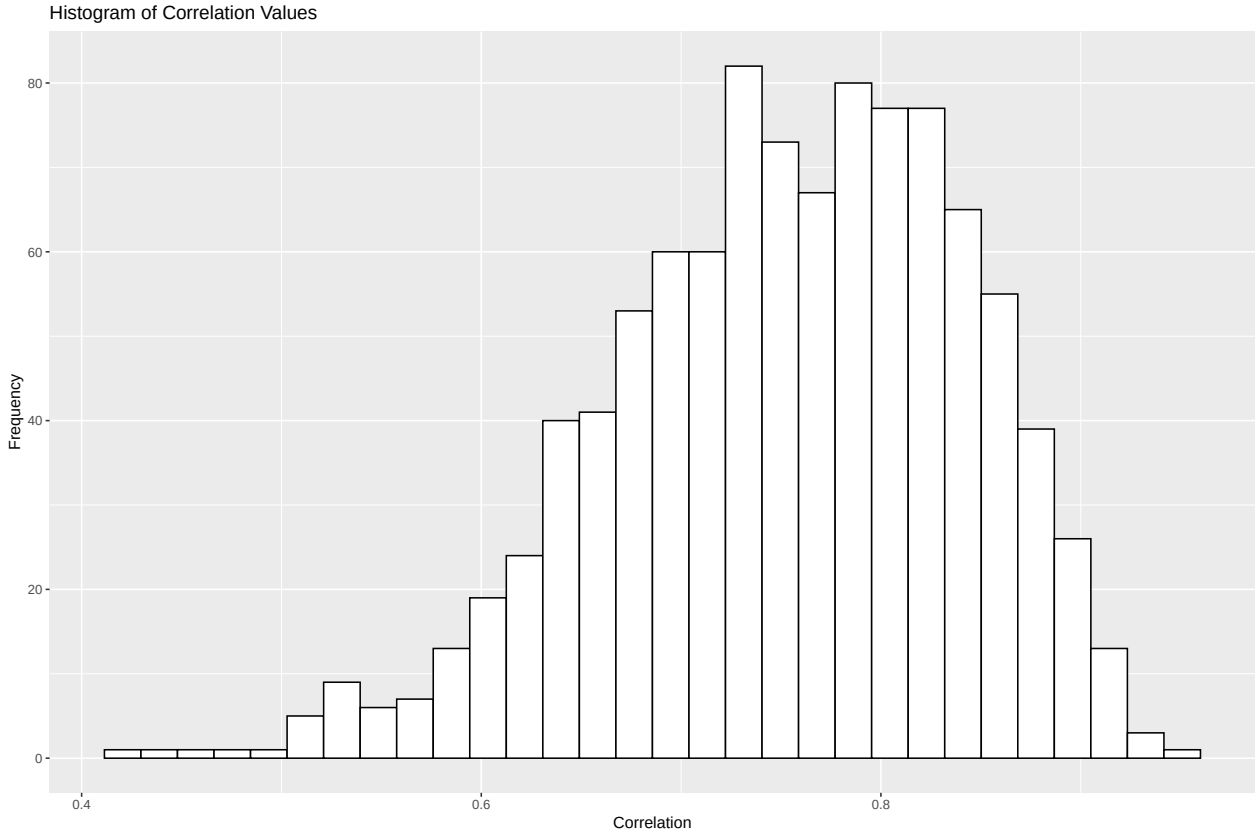



```
## MSE RESULTS
## Mean: 0.06537103
## Median: 0.06397191
## Variance: 0.0004410638
## st.dev.: 0.02100152
```

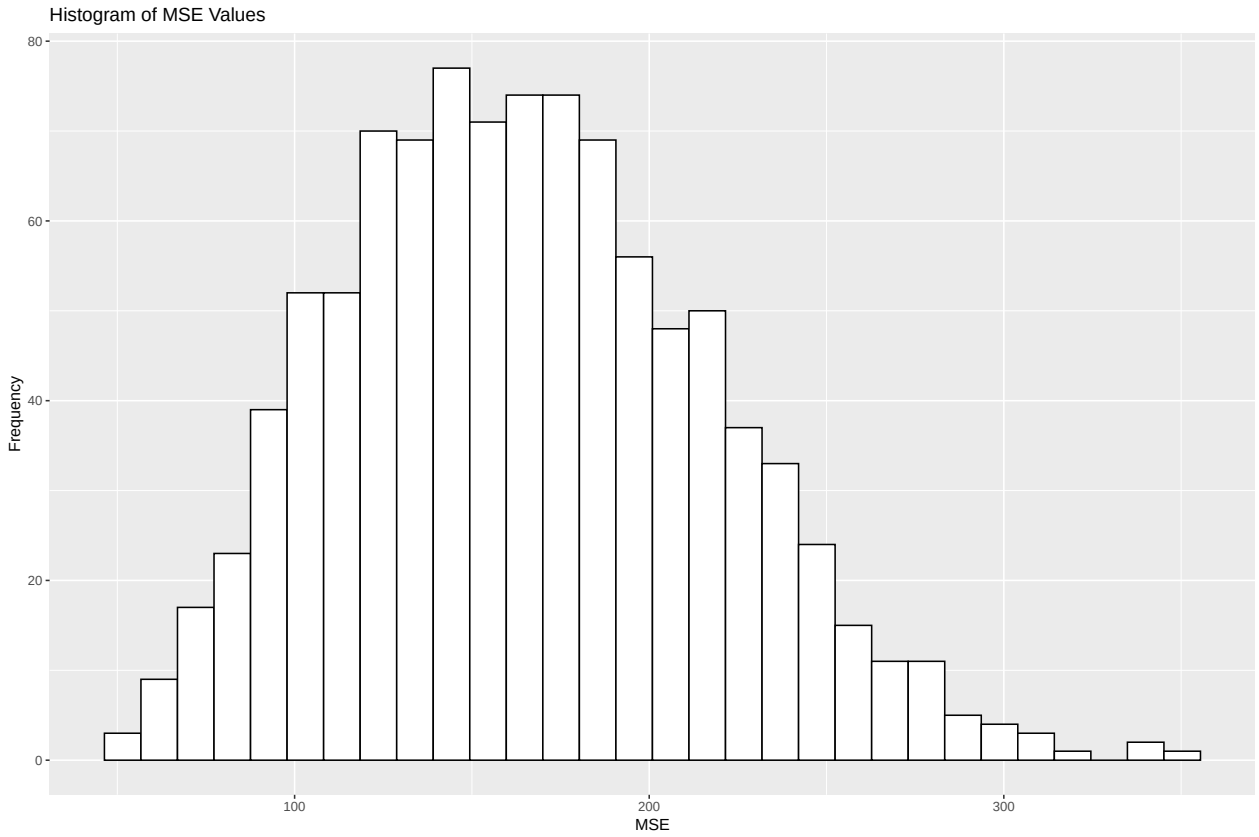


771 genes -> ROR-proliferation score

```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0
##
## CORRELATIONS RESULTS
## Mean: 0.7530515
## Median: 0.7593775
## Variance: 0.007786865
## st.dev.: 0.08824321
```



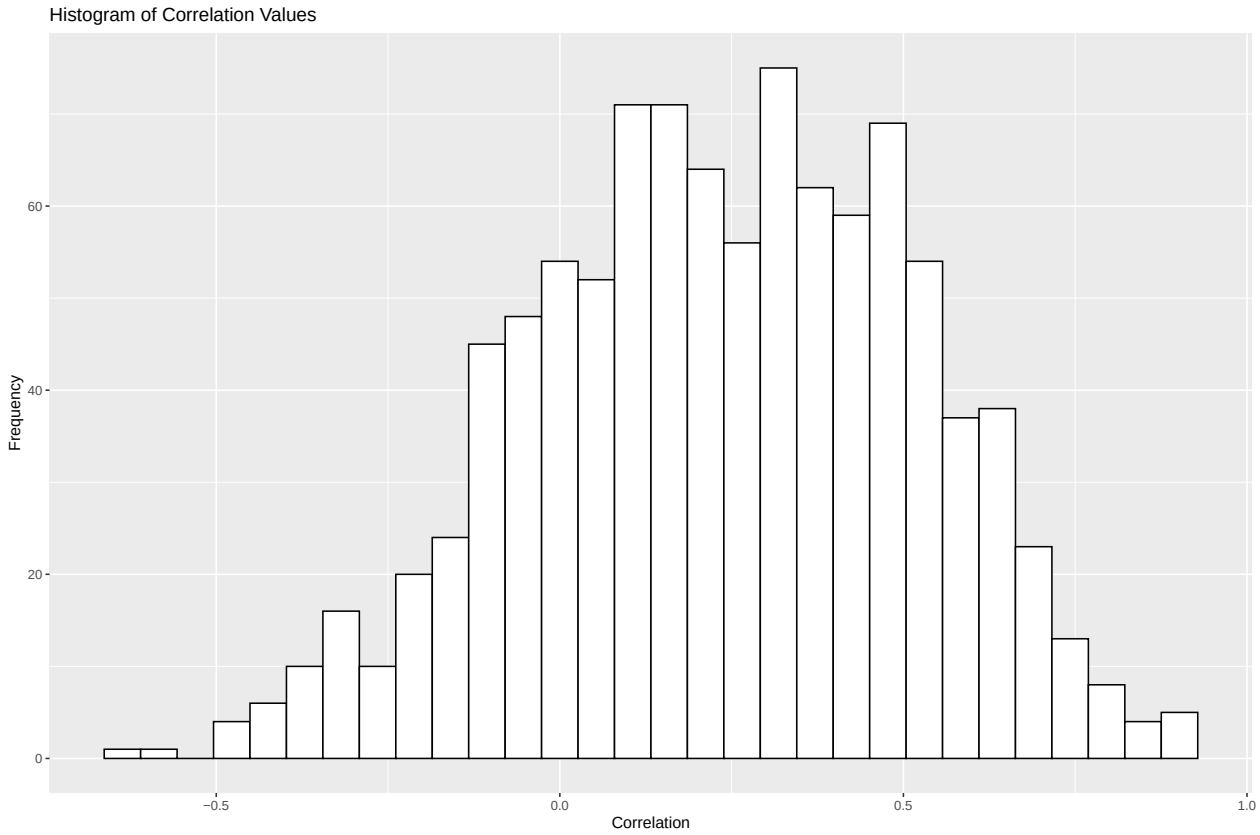

```
## MSE RESULTS
## Mean: 165.145
## Median: 162.2361
## Variance: 2690.091
## st.dev.: 51.86608
```



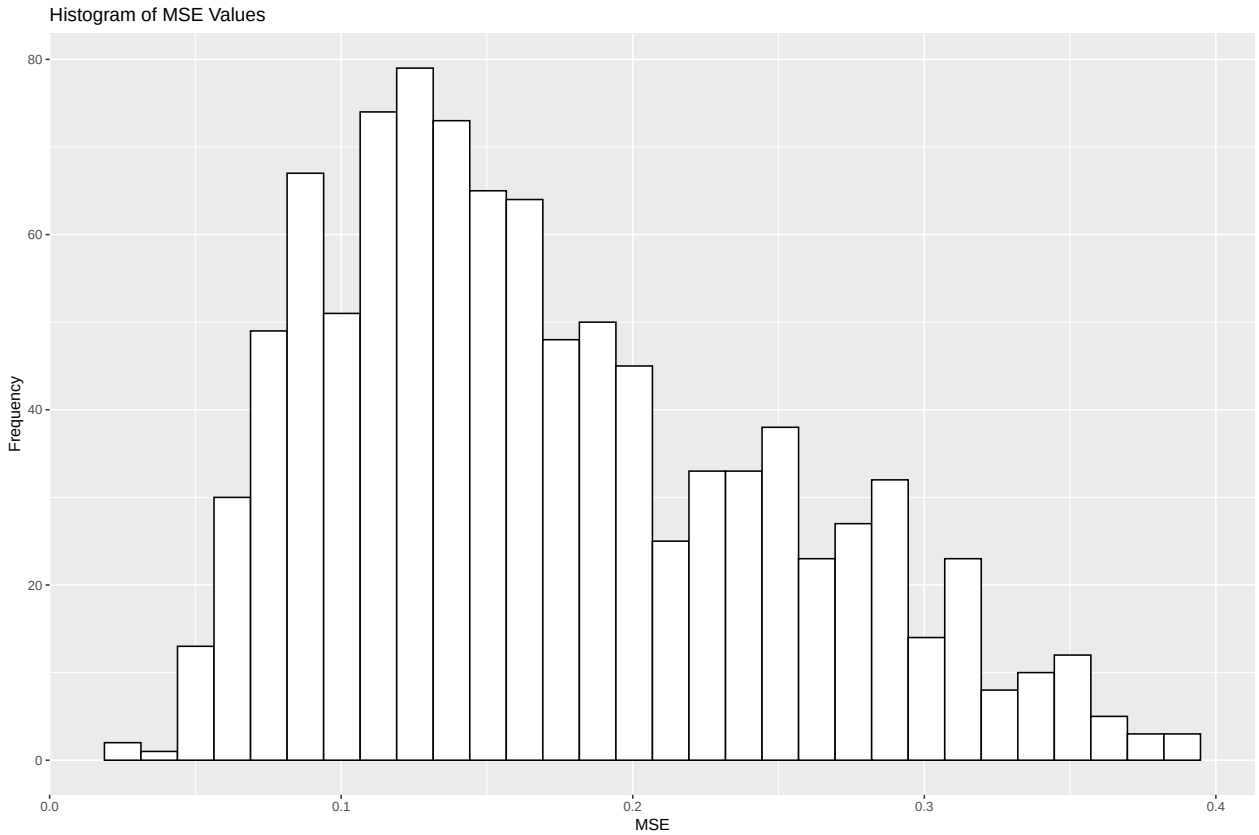
Boosting with stumps as base learner cross-validation

771 genes -> proliferation score

```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0
##
## CORRELATIONS RESULTS
## Mean: 0.2364594
## Median: 0.2426694
## Variance: 0.07812256
## st.dev.: 0.2795041
```

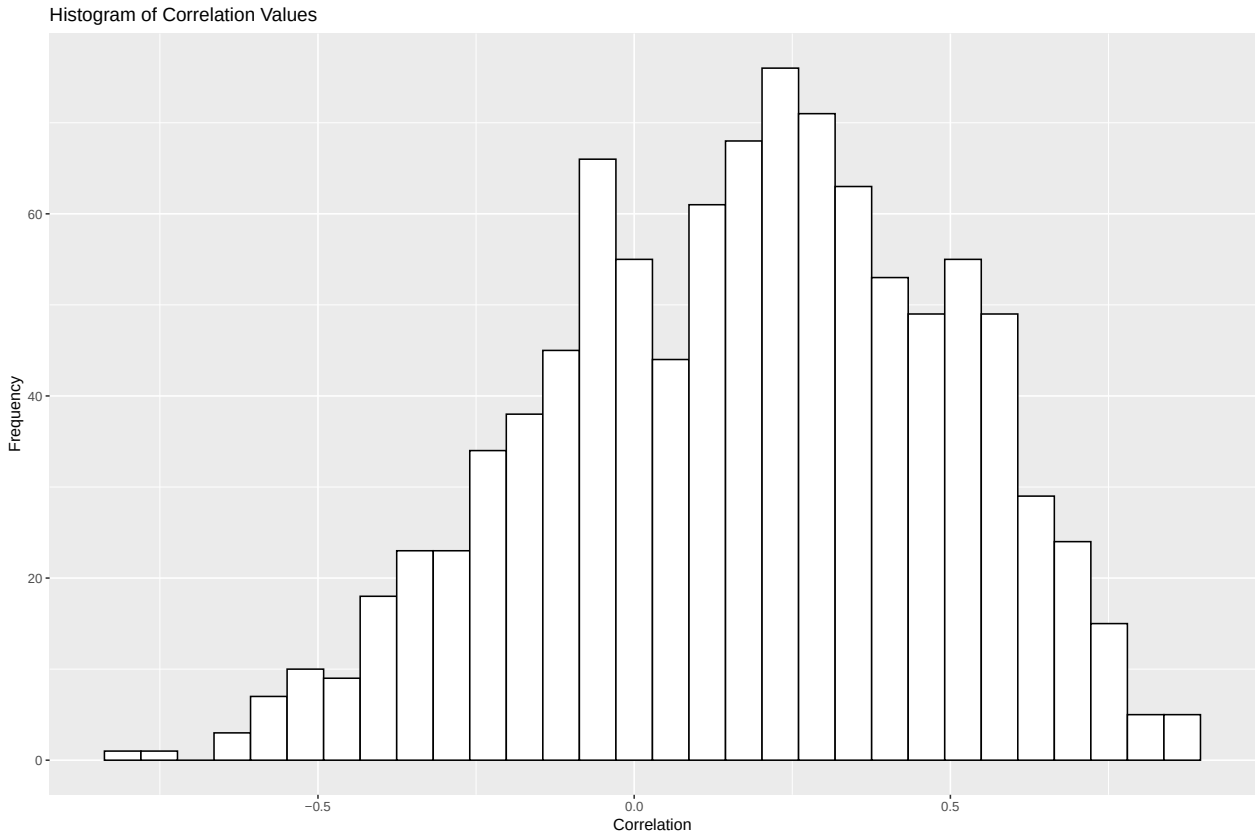



```
## MSE RESULTS
## Mean: 0.1712618
## Median: 0.1554505
## Variance: 0.005837176
## st.dev.: 0.07640141
```

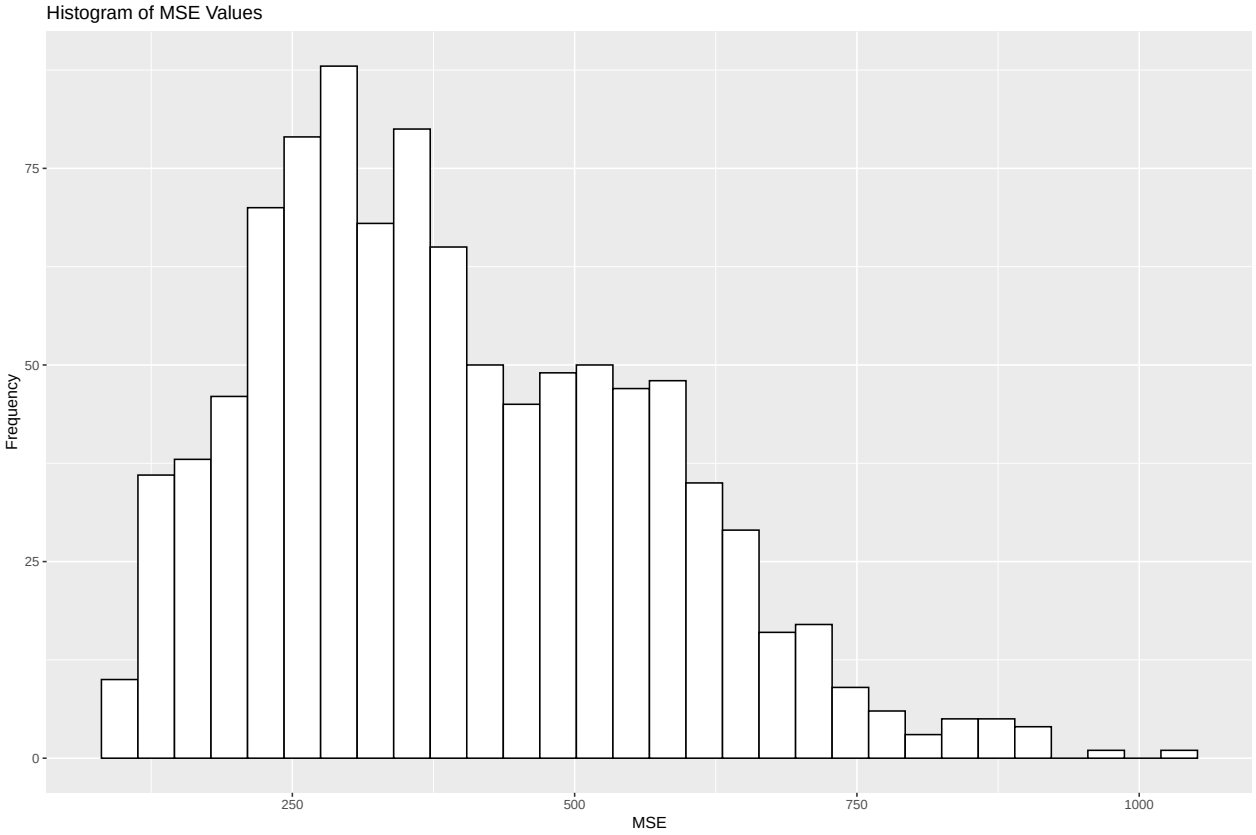


771 genes -> ROR-proliferation score

```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0
##
## CORRELATIONS RESULTS
## Mean: 0.1744792
## Median: 0.1948469
## Variance: 0.09932263
## st.dev.: 0.3151549
```




```
## MSE RESULTS
## Mean: 394.2634
## Median: 366.3266
## Variance: 29461.83
## st.dev.: 171.6445
```



Summery results: Boosting with stumps 771 genes bootstrap and repeated cross-validation

Model	cor_mean	sd_cor	MSE_mean	MSE_sd
prolif boot	0.7760479	0.0827829	0.0653710	0.0210015
ROR-prolif boot	0.7530515	0.0882432	165.1450271	51.8660843
prolif rep cross-val	0.2364594	0.2795041	0.1712618	0.0764014
ROR-prolif rep cross-val	0.1744792	0.3151549	394.2634498	171.6444924

Post Lasso

not done

Summery of all results

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	0.7266736	0.0895236	0.0813415	0.0231428

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	0.5427757	0.1908061	291.0186	82.79231

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

Model	cor_mean	sd_cor	MSE_mean	MSE_sd
lasso 6 genes	0.0919663	0.3068209	0.1655931	0.0718379
lasso 771 genes	0.4737037	0.2310209	0.0620913	0.0753056
Nodes	0.2842257	0.2768458	0.1560308	0.0762883
Residual additive	0.4633095	0.2227105	0.1331785	0.0654065
Residual multiplicative	0.4028471	0.2302632	0.1455819	0.0680617

Summery results: lasso ROR+proliferation score (repeated cross-validation)

Model	cor_mean	sd_cor	MSE_mean	MSE_sd
lasso 6 genes	-0.4822298	0.1724985	374.1519	144.1559
lasso 771 genes	0.0806237	0.2767153	393.8069	159.6451
Nodes	0.1806504	0.2854892	380.1157	164.5869
Residual additive	0.1642500	0.2788435	392.5436	158.8733
Residual multiplicative	-0.2145253	0.2512231	568.8063	208.5911

Summery results: ridge 771 genes bootstrap and repeated cross-validation

Model	cor_mean	sd_cor	MSE_mean	MSE_sd
prolif boot	0.8189568	0.0710276	0.0566524	0.0186773
ROR-prolif boot	0.7761924	0.0774688	156.0649552	44.1817261
prolif rep cross-val	0.5268342	0.2071537	0.1256548	0.0699318
ROR-prolif rep cross-val	0.0806237	0.2767153	393.8068910	159.6450765

Summery results: elastic net 771 genes bootstrap and repeated cross-validation

Model	cor_mean	sd_cor	MSE_mean	MSE_sd
prolif boot	0.8125591	0.0838975	0.0558955	0.0217032
ROR-prolif boot	0.7565123	0.0888143	164.4116160	52.7751849
prolif rep cross-val	0.4271890	0.2637029	0.1501025	0.0822400
ROR-prolif rep cross-val	0.2049341	0.3111969	427.3517412	186.2464725

Summery results: Boosting with stumps 771 genes bootstrap and repeated cross-validation

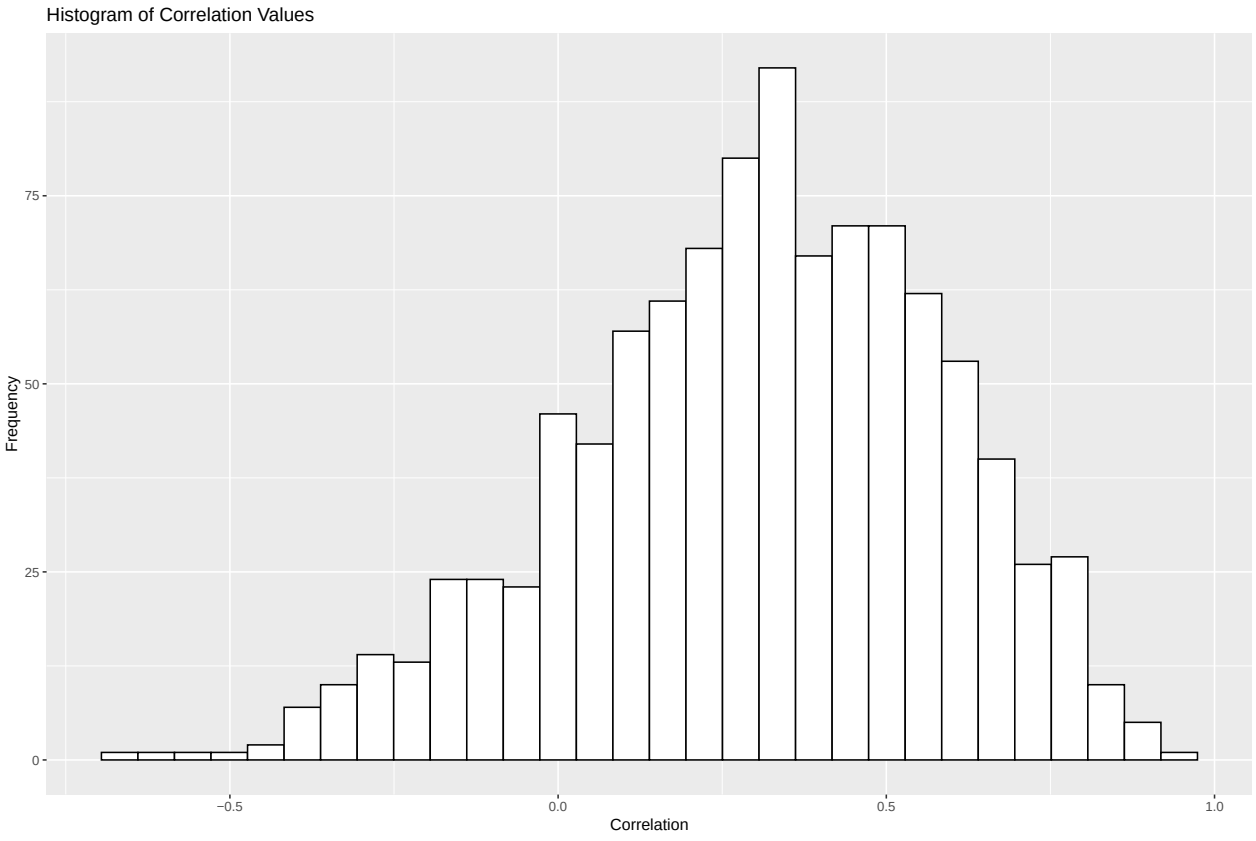
Model	cor_mean	sd_cor	MSE_mean	MSE_sd
prolif boot	0.7760479	0.0827829	0.0653710	0.0210015
ROR-prolif boot	0.7530515	0.0882432	165.1450271	51.8660843
prolif rep cross-val	0.2364594	0.2795041	0.1712618	0.0764014
ROR-prolif rep cross-val	0.1744792	0.3151549	394.2634498	171.6444924

START USING DOMAIN KNOWLEDGE

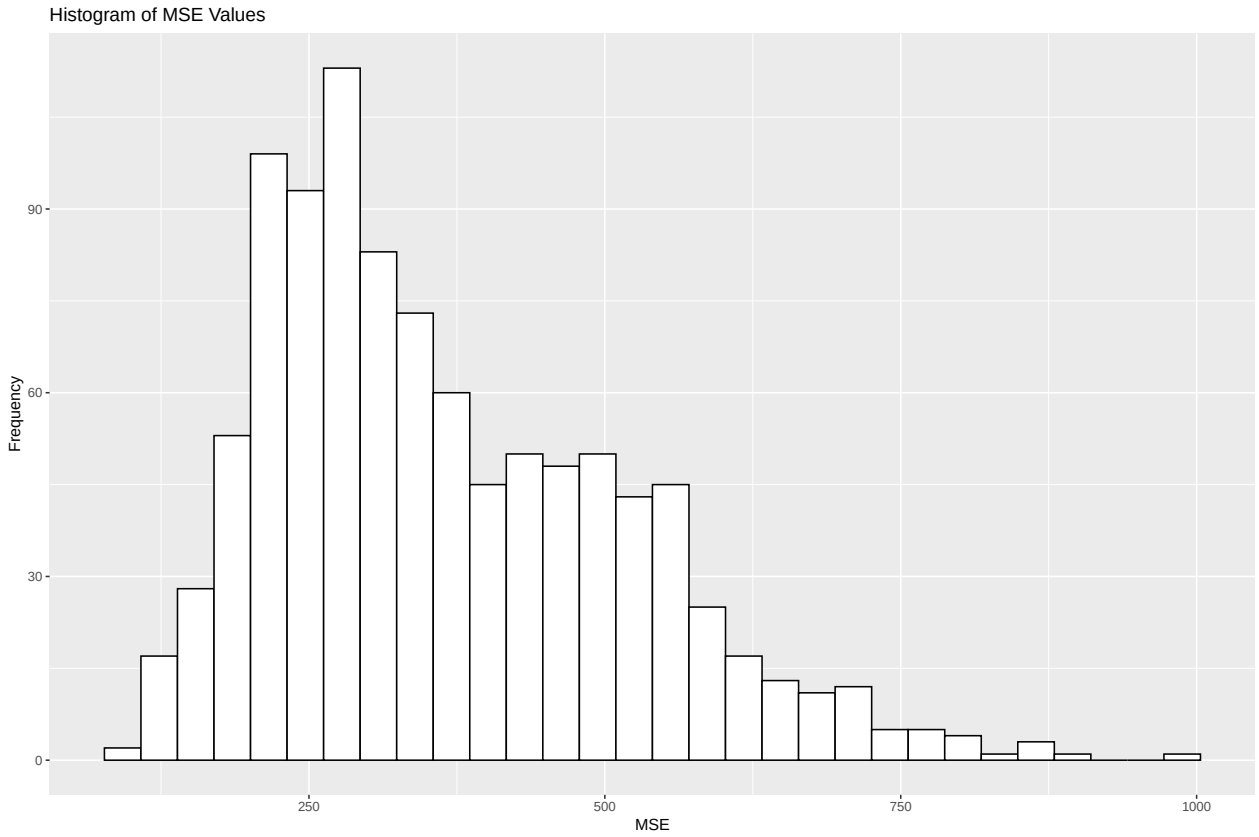
PCA on signature gene sets, cross-validation

Ridge: 771 genes -> ROR-proliferation score

```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0
##
## CORRELATIONS RESULTS
## Mean: 0.3028091
## Median: 0.319312
## Variance: 0.07725244
## st.dev.: 0.2779432
```

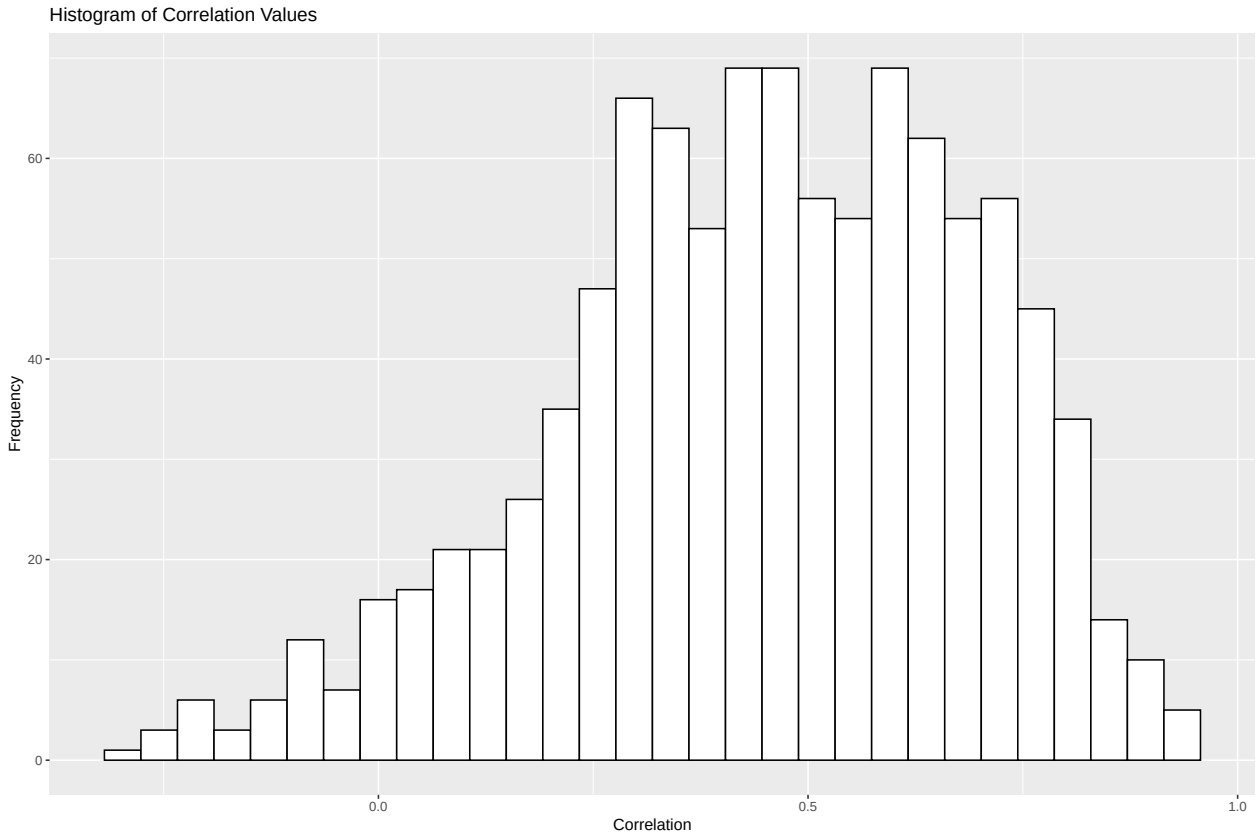



```
## MSE RESULTS
## Mean: 364.5733
## Median: 327.0747
## Variance: 22807.28
## st.dev.: 151.0208
```

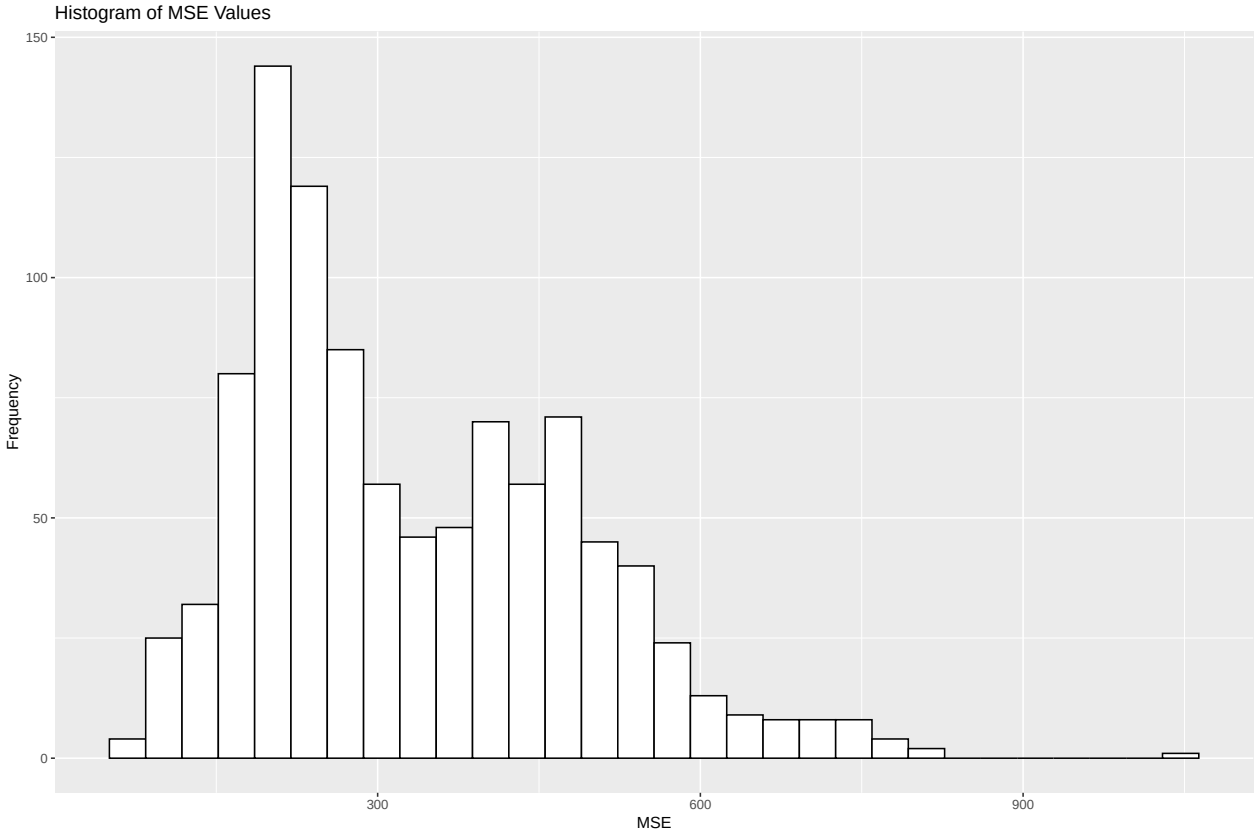


Ridge: 771 genes -> ROR-proliferation score + interactions between PCs

```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0
##
## CORRELATIONS RESULTS
## Mean: 0.4505837
## Median: 0.463892
## Variance: 0.05781415
## st.dev.: 0.2404457
```

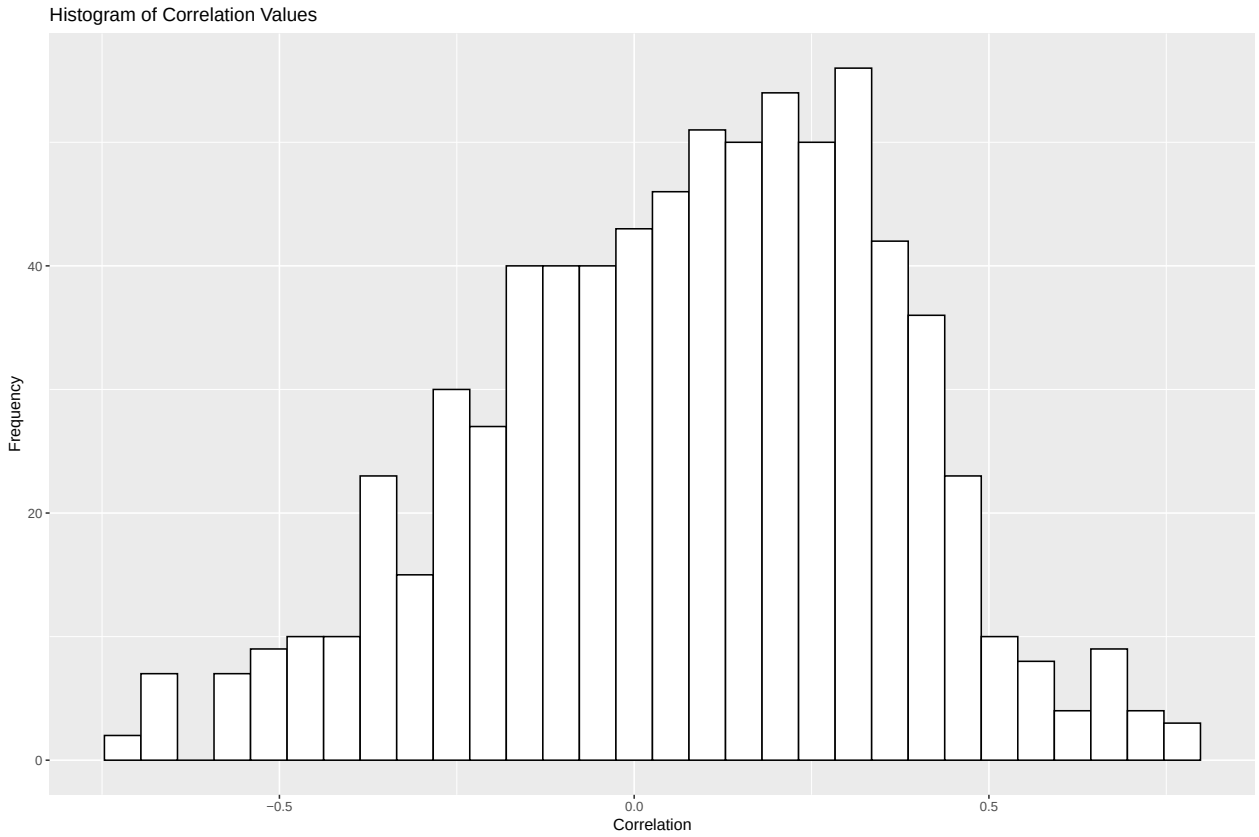



```
## MSE RESULTS
## Mean: 333.3975
## Median: 291.1335
## Variance: 23000.06
## st.dev.: 151.6577
```

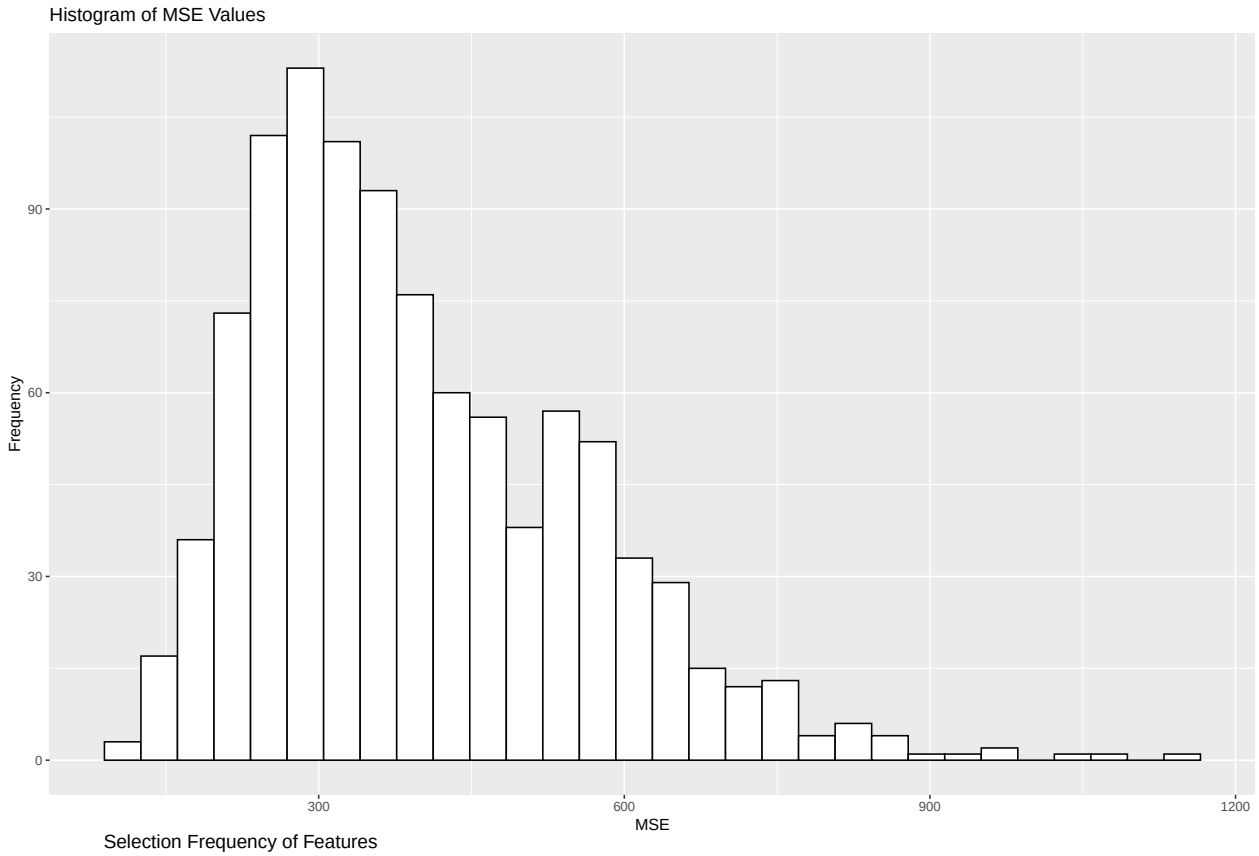


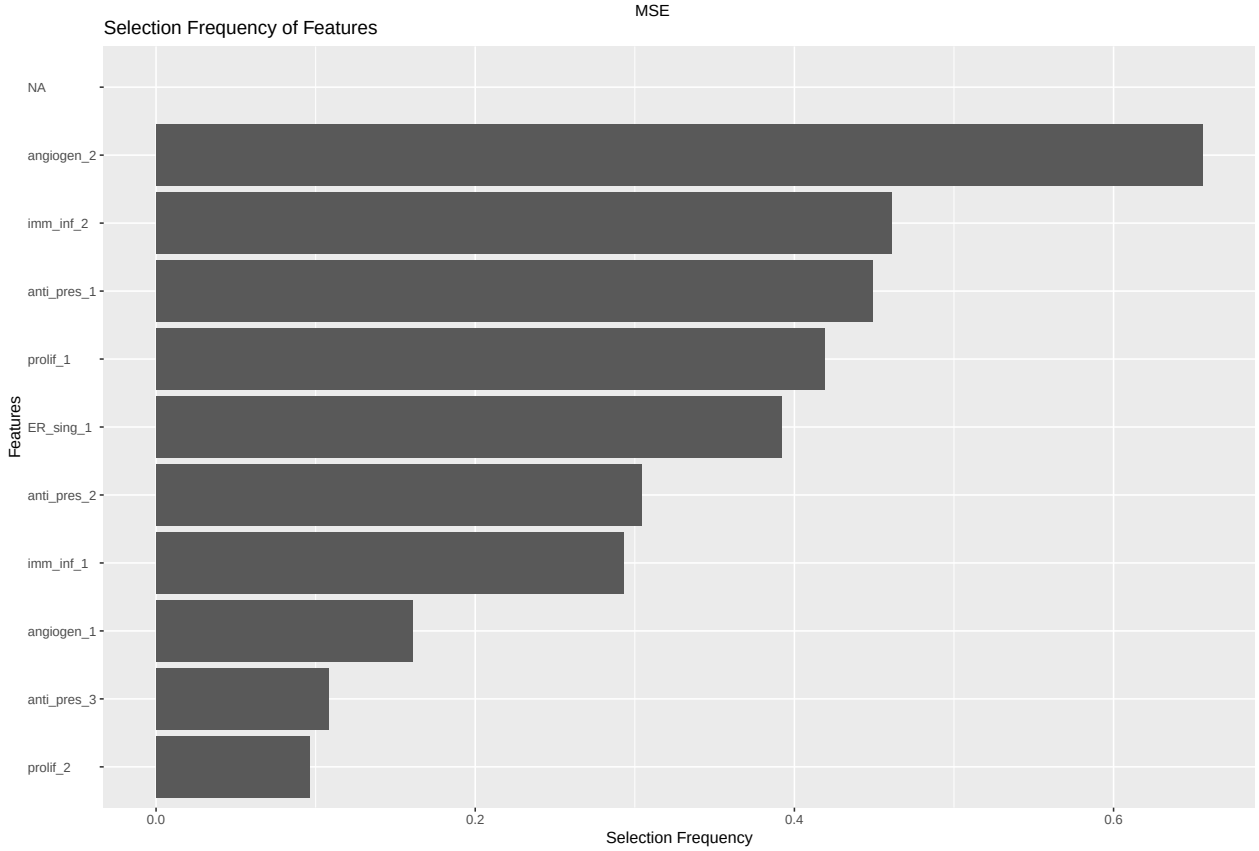
Lasso: 771 genes -> ROR-proliferation score

```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0.251
##
## CORRELATIONS RESULTS
## Mean: 0.07849613
## Median: 0.1028317
## Variance: 0.08158919
## st.dev.: 0.2856382
```




```
## MSE RESULTS
## Mean: 396.4902
## Median: 361.1201
## Variance: 25641.71
## st.dev.: 160.1303
```


Histogram of MSE Values
Frequency
90
60
30
0
300
600
900
1200
MSE

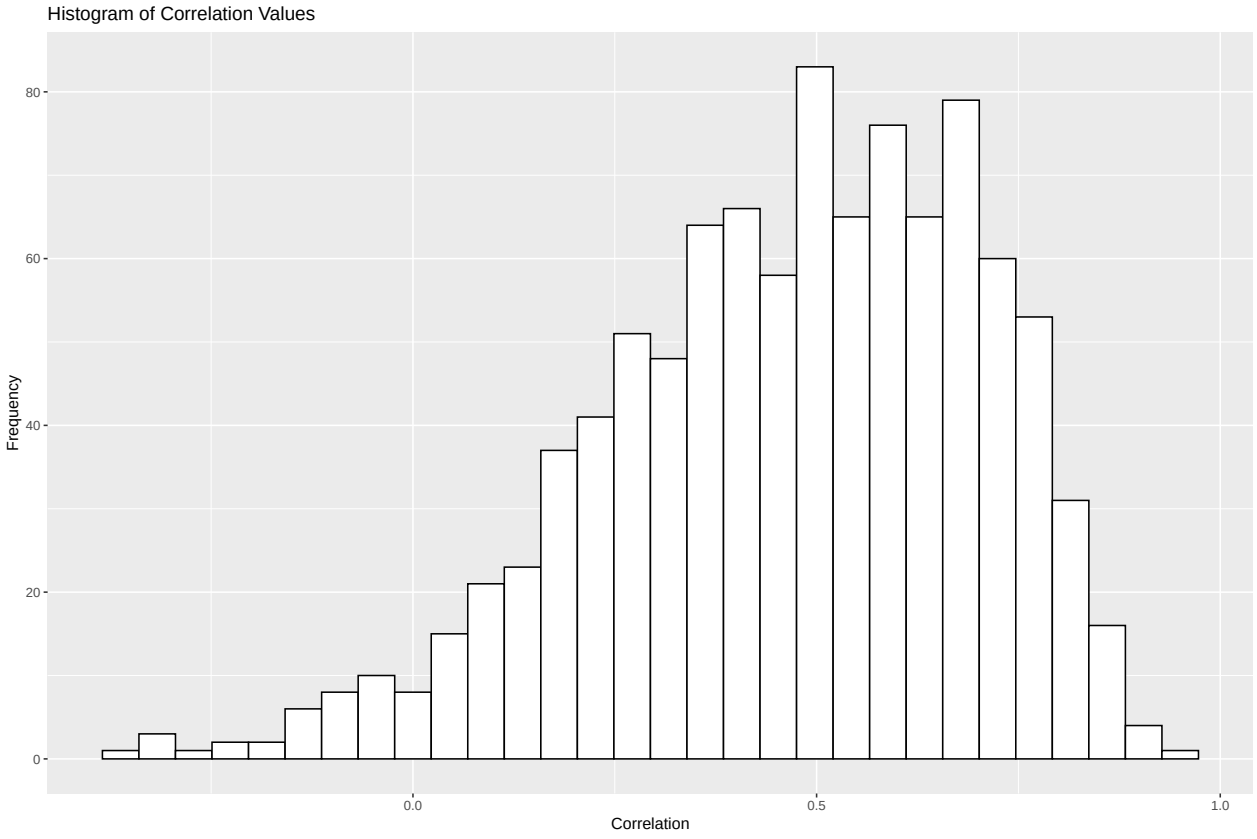

Selection Frequency of Features
Features
NA
angiogen_2
imm_inf_2
anti_pres_1
prolif_1
ER_sing_1
anti_pres_2
imm_inf_1
angiogen_1
anti_pres_3
prolif_2
0.0
0.2
0.4
0.6
Selection Frequency

##

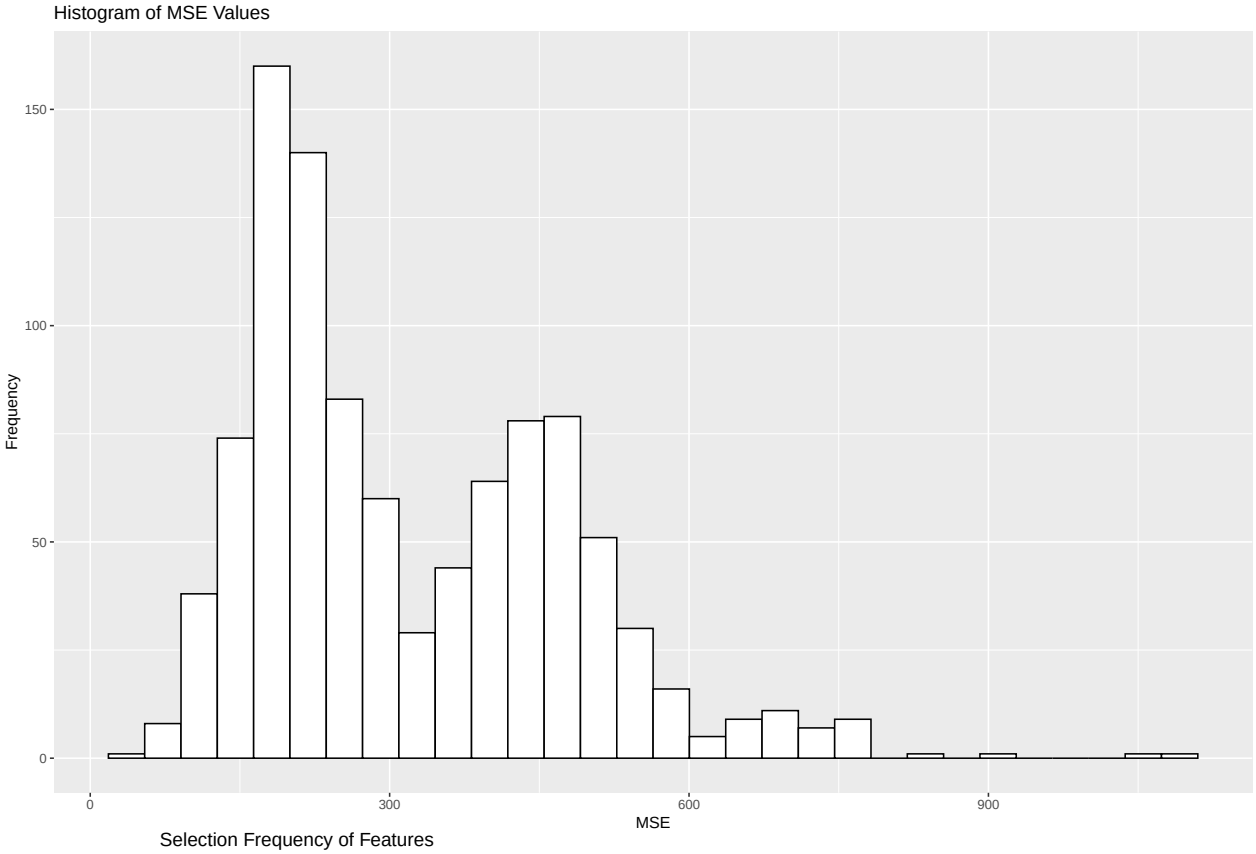
```
## Features selected 50% or more times:
## angiogen_2
## Top 20 featrues:
##  [1] "angiogen_2"  "imm_inf_2"   "anti_pres_1" "prolif_1"    "ER_sing_1"
##  [6] "anti_pres_2" "imm_inf_1"   "angiogen_1"  "anti_pres_3" "prolif_2"
## [11] NA            NA            NA            NA            NA
## [16] NA            NA            NA            NA            NA
```

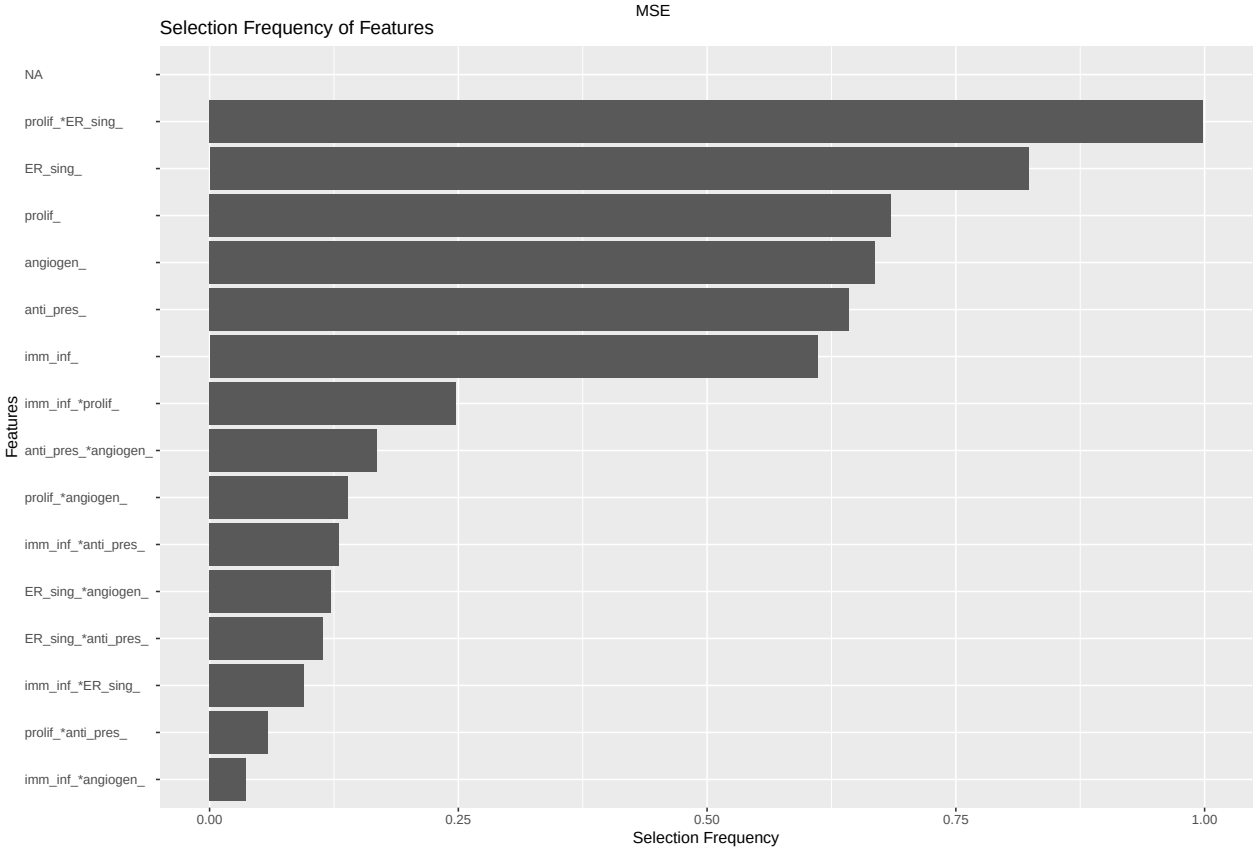
Lasso: 771 genes -> ROR-proliferation score + interactions between PCs

```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0.002
##
## CORRELATIONS RESULTS
## Mean: 0.4687056
## Median: 0.4972697
## Variance: 0.05426225
## st.dev.: 0.2329426
```




```
## MSE RESULTS
## Mean: 321.0086
## Median: 265.9525
## Variance: 24834.46
## st.dev.: 157.5895
```


Histogram of MSE Values
Frequency
150
100
50
0
0
300
600
900
MSE

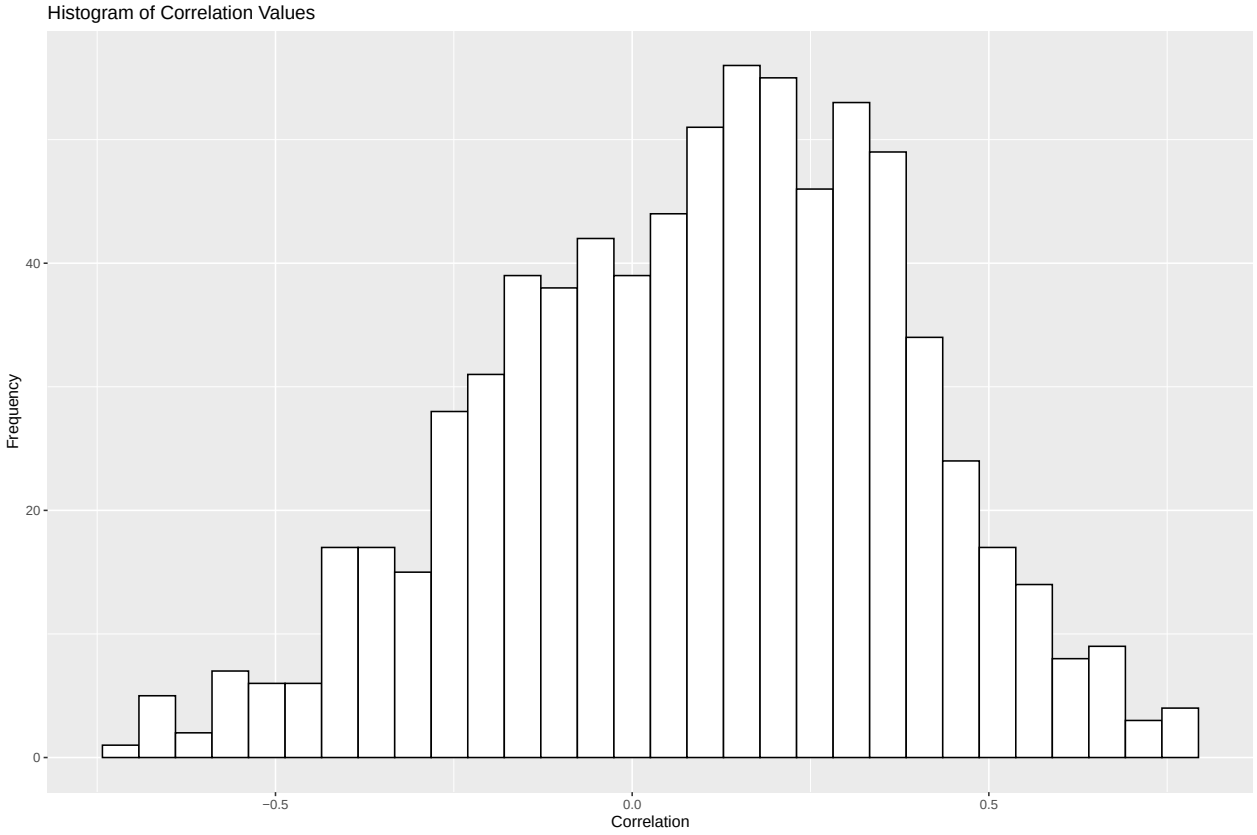

Selection Frequency of Features
Features
NA
prolif_*ER_sing_
ER_sing_
prolif_
angiogen_
anti_pres_
imm_inf_
imm_inf_*prolif_
anti_pres_*angiogen_
prolif_*angiogen_
imm_inf_*anti_pres_
ER_sing_*angiogen_
ER_sing_*anti_pres_
imm_inf_*ER_sing_
prolif_*anti_pres_
imm_inf_*angiogen_
0.00
0.25
0.50
0.75
1.00
Selection Frequency

##

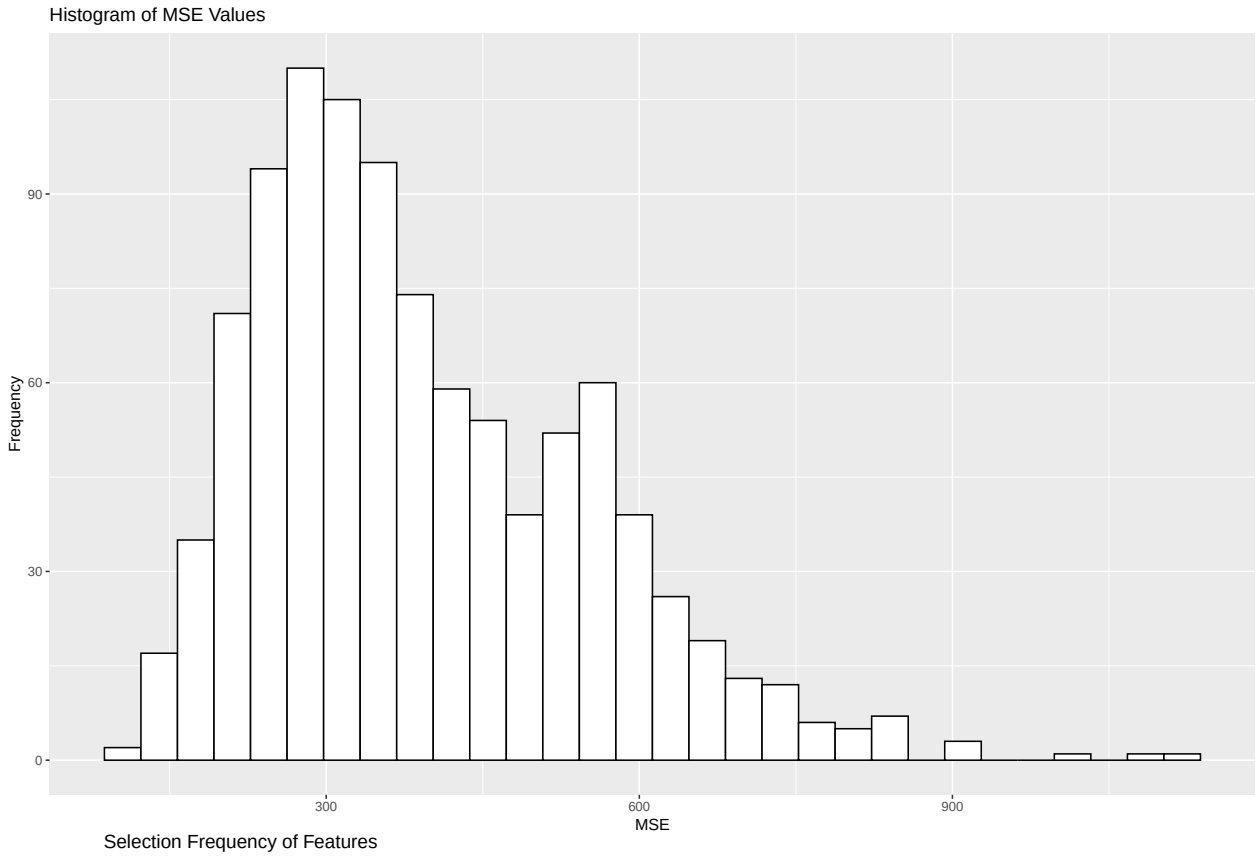
```
## Features selected 50% or more times:
## imm_inf_ prolif_ ER_sing_ anti_pres_ angiogen_ prolif_*ER_sing_
## Top 20 featrues:
##  [1] "prolif_*ER_sing_"     "ER_sing_"             "prolif_"
##  [4] "angiogen_"            "anti_pres_"           "imm_inf_"
##  [7] "imm_inf_*prolif_"     "anti_pres_*angiogen_" "prolif_*angiogen_"
## [10] "imm_inf_*anti_pres_"  "ER_sing_*angiogen_"   "ER_sing_*anti_pres_"
## [13] "imm_inf_*ER_sing_"    "prolif_*anti_pres_"   "imm_inf_*angiogen_"
## [16] NA                     NA                     NA
## [19] NA                     NA
```

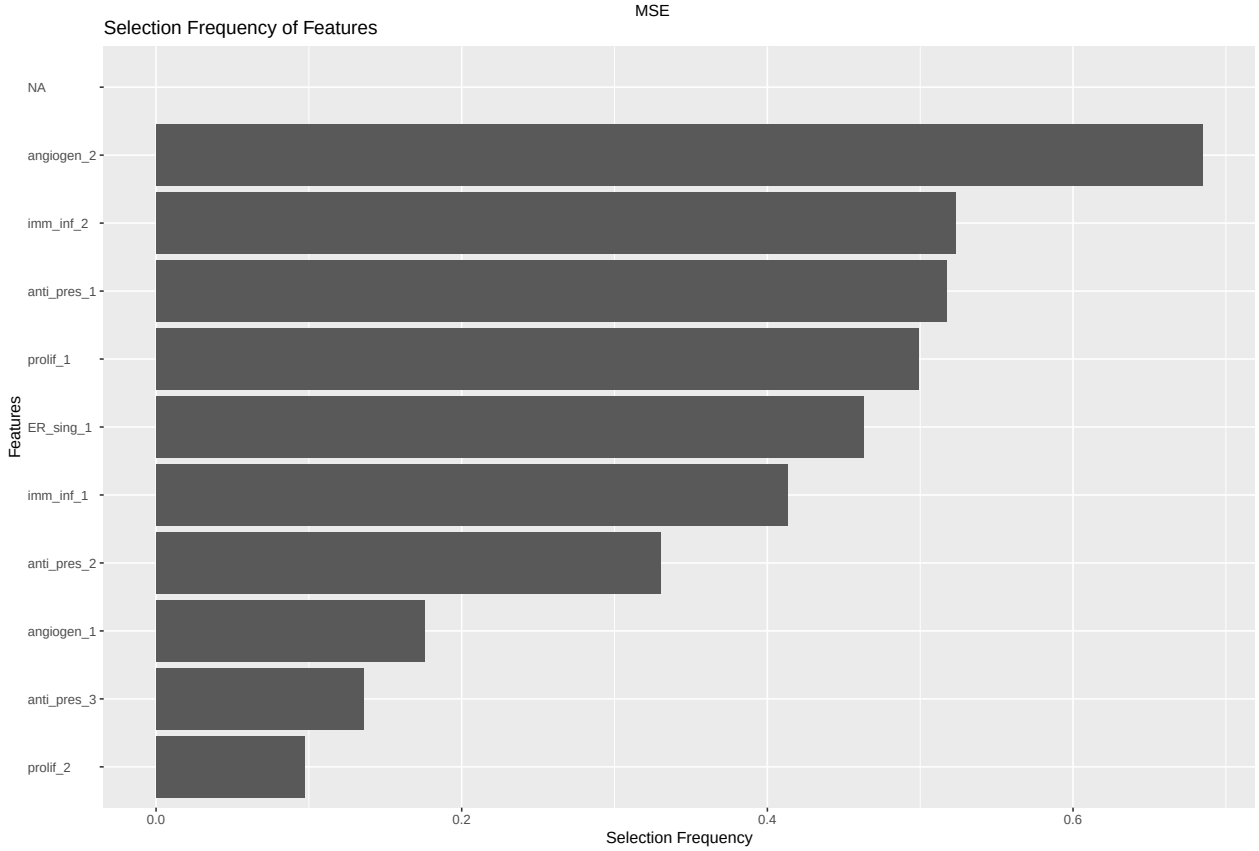
ElasticNet: 771 genes -> ROR-proliferation score

```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0.24
##
## CORRELATIONS RESULTS
## Mean: 0.09606929
## Median: 0.1167623
## Variance: 0.08091802
## st.dev.: 0.2844609
```




```
## MSE RESULTS
## Mean: 392.3862
## Median: 358.5637
## Variance: 24990.96
## st.dev.: 158.0853
```


Histogram of MSE Values
Frequency
90
60
30
0
300
600
900
MSE

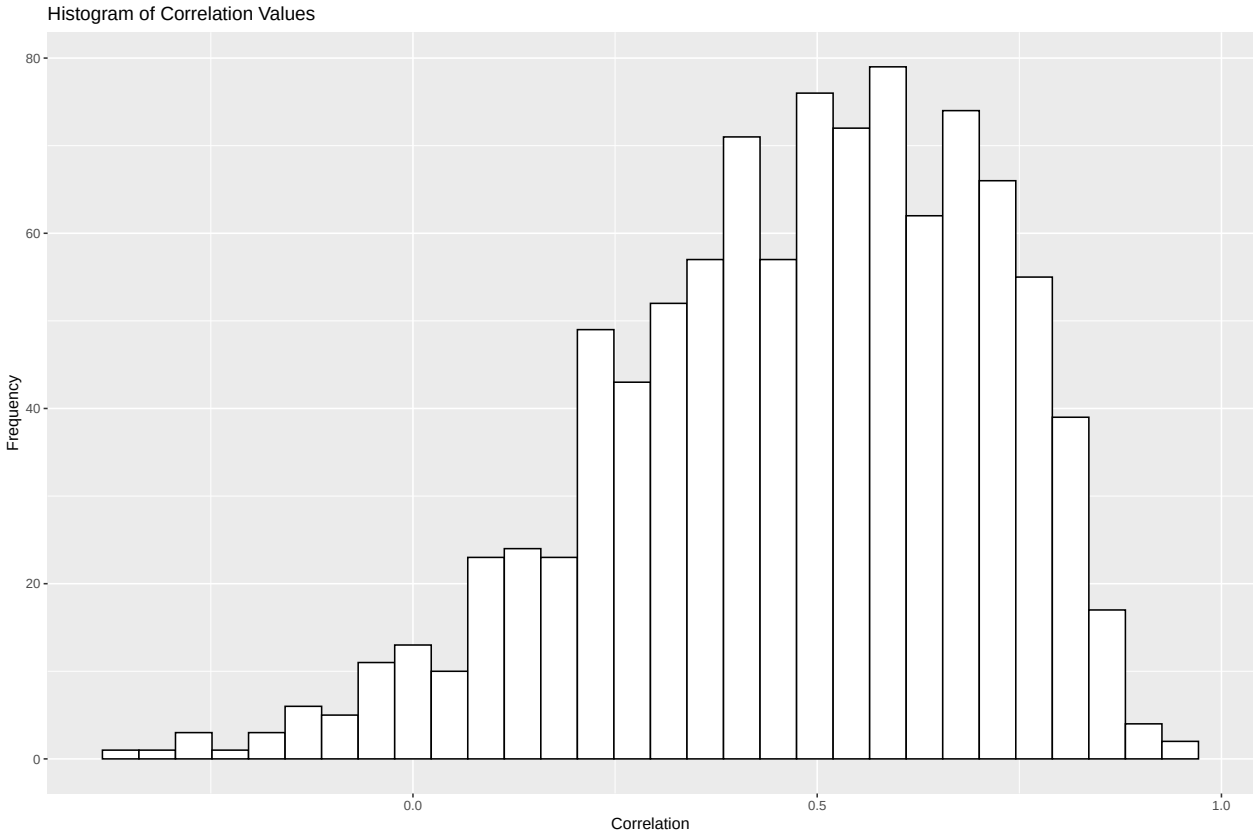

Selection Frequency of Features
Features
NA
angiogen_2
imm_inf_2
anti_pres_1
prolif_1
ER_sing_1
imm_inf_1
anti_pres_2
angiogen_1
anti_pres_3
prolif_2
0.0
0.2
0.4
0.6
Selection Frequency

##

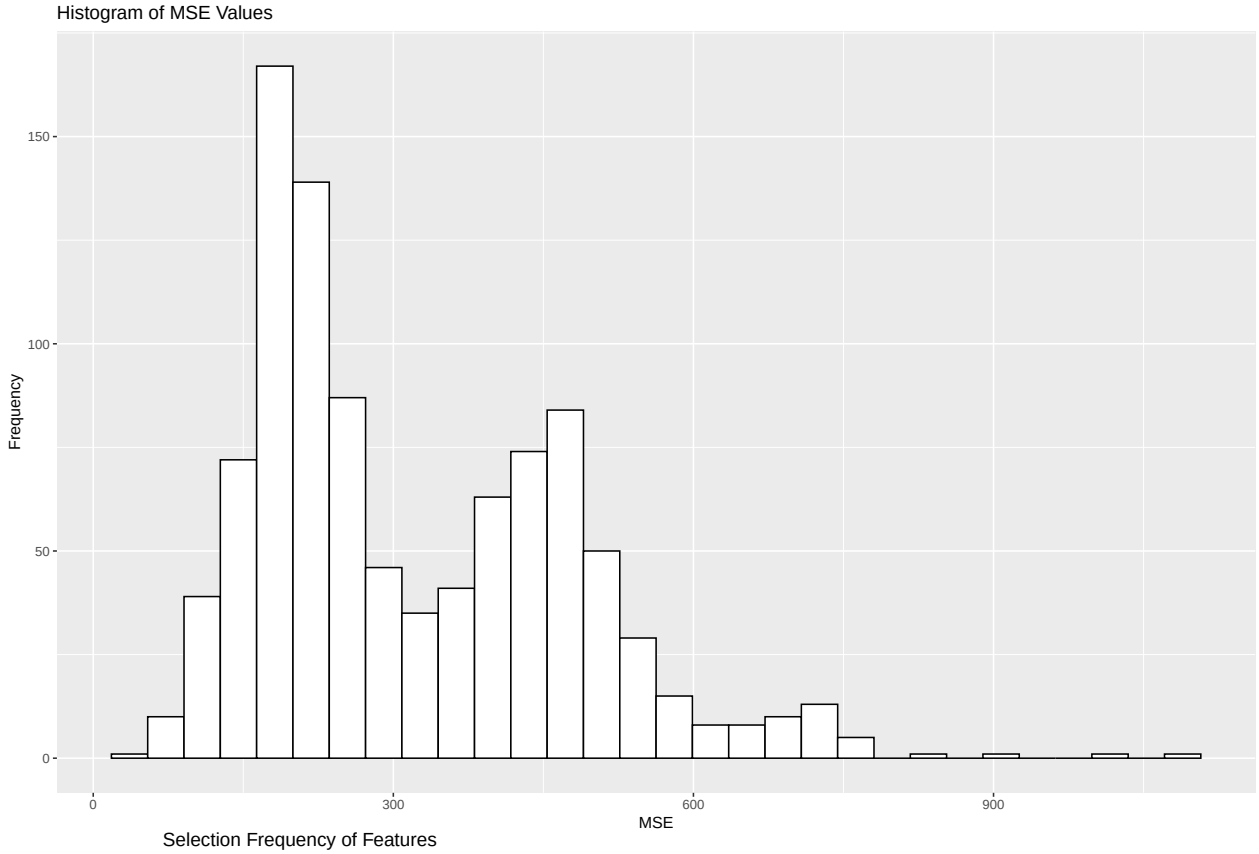
```
## Features selected 50% or more times:
## imm_inf_2 anti_pres_1 angiogen_2
## Top 20 featrues:
##  [1] "angiogen_2"  "imm_inf_2"   "anti_pres_1" "prolif_1"    "ER_sing_1"
##  [6] "imm_inf_1"   "anti_pres_2" "angiogen_1"  "anti_pres_3" "prolif_2"
## [11] NA            NA            NA            NA            NA
## [16] NA            NA            NA            NA            NA
```

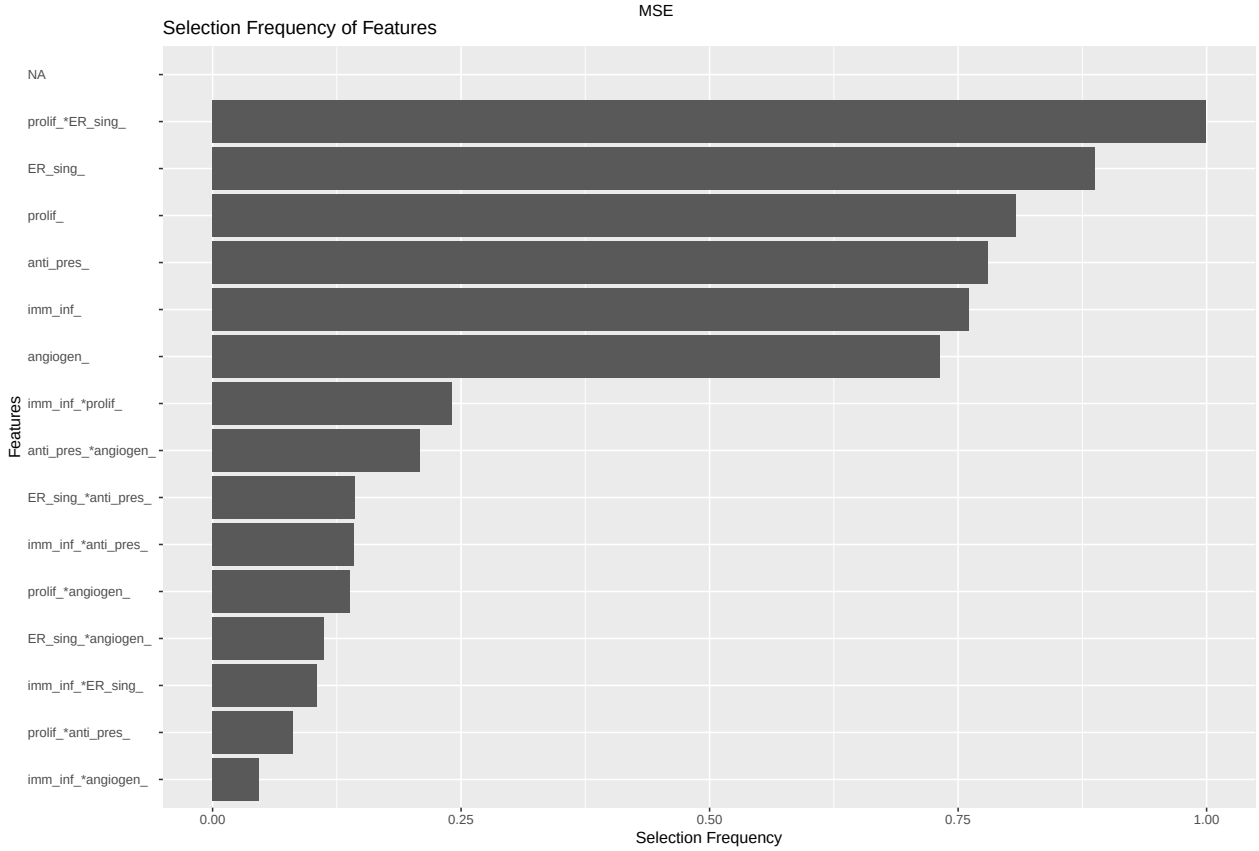
ElasticNet: 771 genes -> ROR-proliferation score + interactions between PCs

```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0.001
##
## CORRELATIONS RESULTS
## Mean: 0.4749342
## Median: 0.5041917
## Variance: 0.05452847
## st.dev.: 0.2335133
```




```
## MSE RESULTS
## Mean: 319.7269
## Median: 263.5692
## Variance: 24916.52
## st.dev.: 157.8497
```


Histogram of MSE Values
Frequency
150
100
50
0
0
300
600
900
MSE


Selection Frequency of Features
Features
NA
prolif_*ER_sing_
ER_sing_
prolif_
anti_pres_
imm_inf_
angiogen_
imm_inf_*prolif_
anti_pres_*angiogen_
ER_sing_*anti_pres_
imm_inf_*anti_pres_
prolif_*angiogen_
ER_sing_*angiogen_
imm_inf_*ER_sing_
prolif_*anti_pres_
imm_inf_*angiogen_
0.00
0.25
0.50
0.75
1.00
Selection Frequency

##

```
## Features selected 50% or more times:
## imm_inf_ prolif_ ER_sing_ anti_pres_ angiogen_ prolif_*ER_sing_
## Top 20 featrues:
##  [1] "prolif_*ER_sing_"     "ER_sing_"             "prolif_"
##  [4] "anti_pres_"           "imm_inf_"             "angiogen_"
##  [7] "imm_inf_*prolif_"     "anti_pres_*angiogen_" "ER_sing_*anti_pres_"
## [10] "imm_inf_*anti_pres_"  "prolif_*angiogen_"    "ER_sing_*angiogen_"
## [13] "imm_inf_*ER_sing_"    "prolif_*anti_pres_"   "imm_inf_*angiogen_"
## [16] NA                     NA                     NA
## [19] NA                     NA
```

Summery results: Boosting with stumps 771 genes bootstrap and repeated cross-validation

Model	cor_mean	sd_cor	MSE_mean	MSE_sd
prolif boot	0.7760479	0.0827829	0.0653710	0.0210015
ROR-prolif boot	0.7530515	0.0882432	165.1450271	51.8660843
prolif rep cross-val	0.2364594	0.2795041	0.1712618	0.0764014
ROR-prolif rep cross-val	0.1744792	0.3151549	394.2634498	171.6444924