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 X numbers of "signature genes"; which are thought to represent functional unities with respect to cancer biology... overview of signatures..

Additionally, both data-set contians clinical data as... (how to model in...)

Respones used

Proliferation score

Risk of relapse score (ROR)

...

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 to see if some is superior

Evaluation of models

Two levels of evaluation is scheduled:

1. Relative comparison of 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)

Models

Lasso

Post Lasso

Ridge

Elastic Net

Boosting with stumps as base learner

- mboost

- xgboost

Feature selecting ensemble model

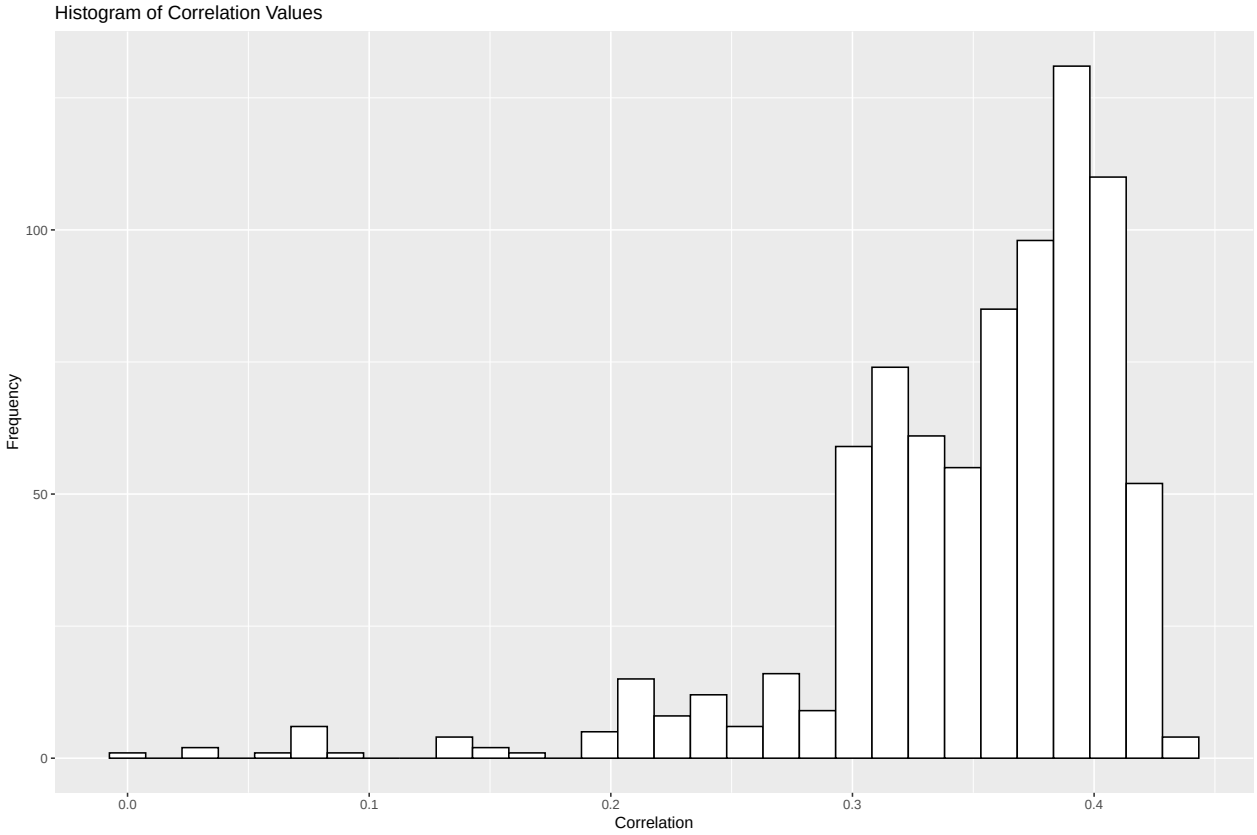
RESULTS

Lasso

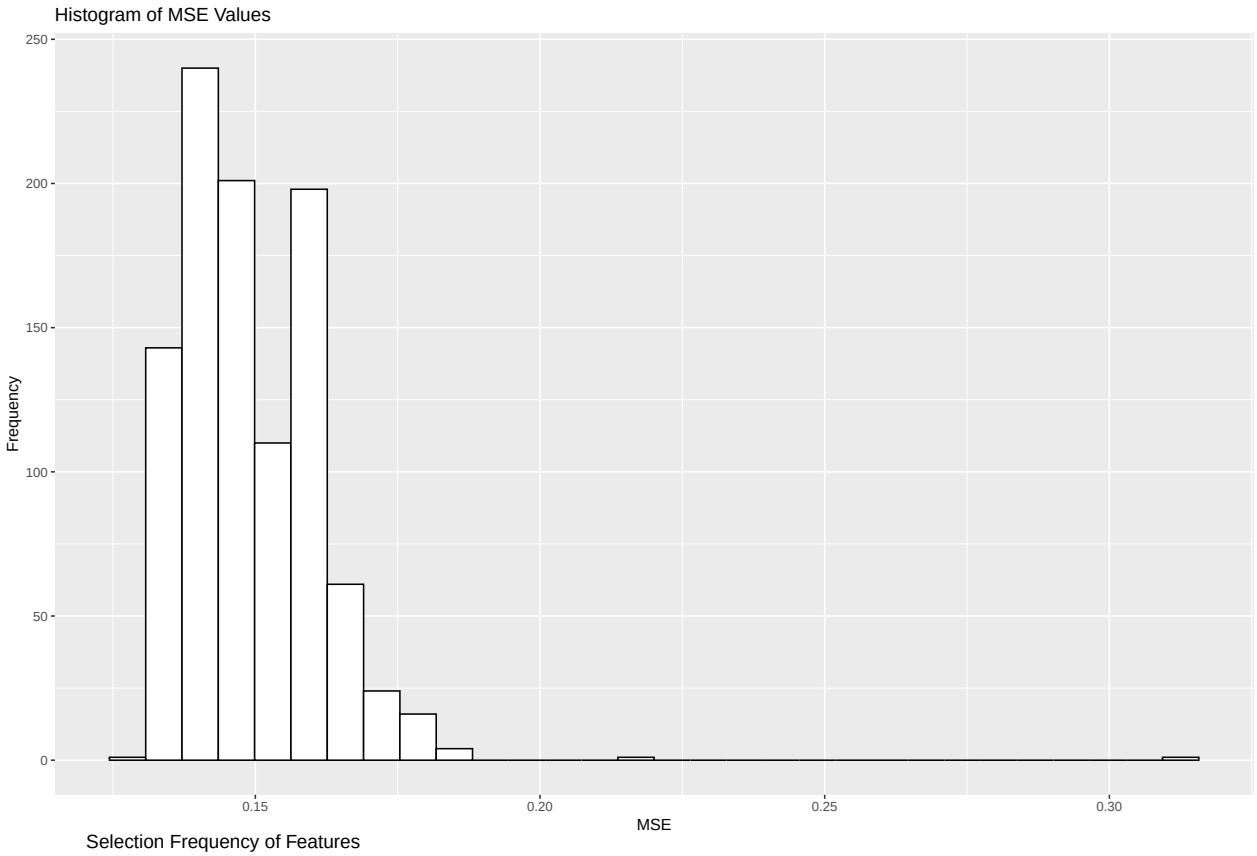
Bootstrap

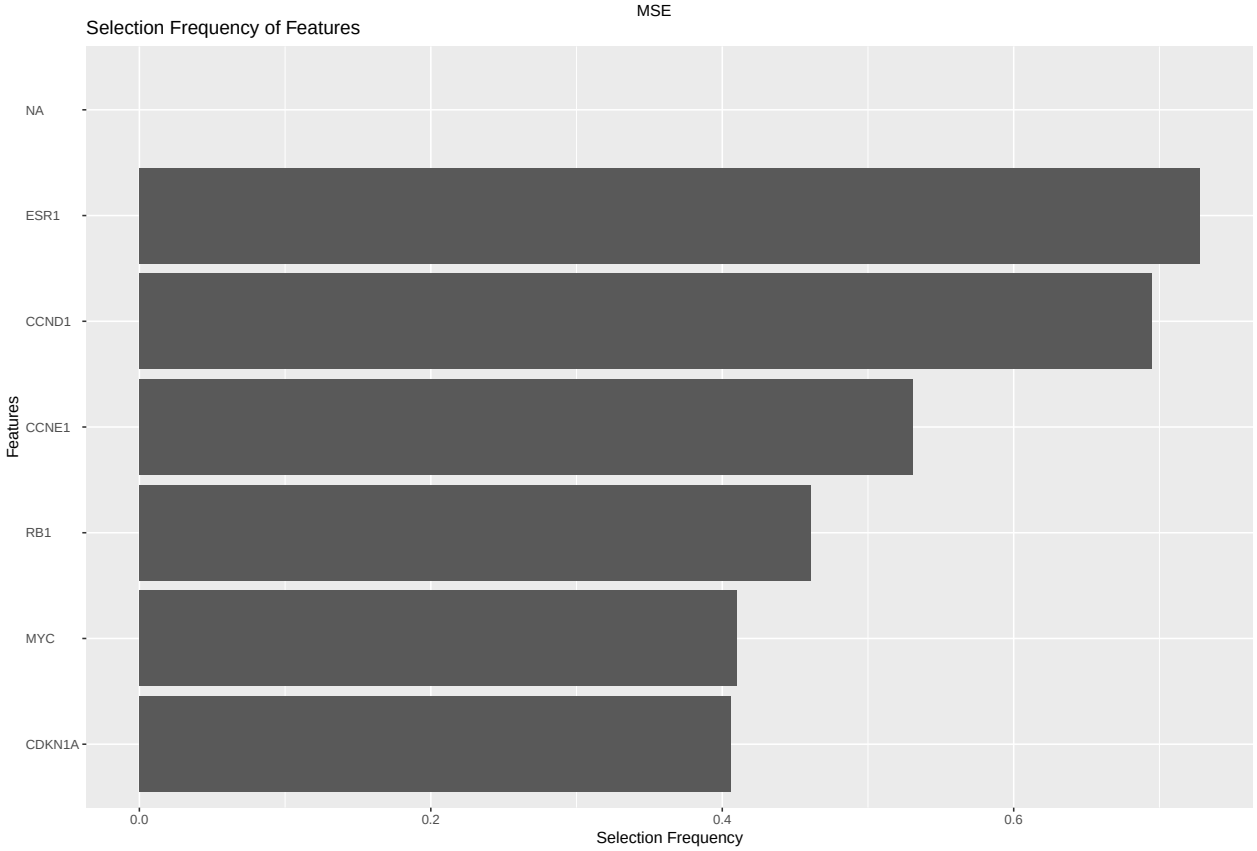
6 genes -> proliferation score

```
## 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

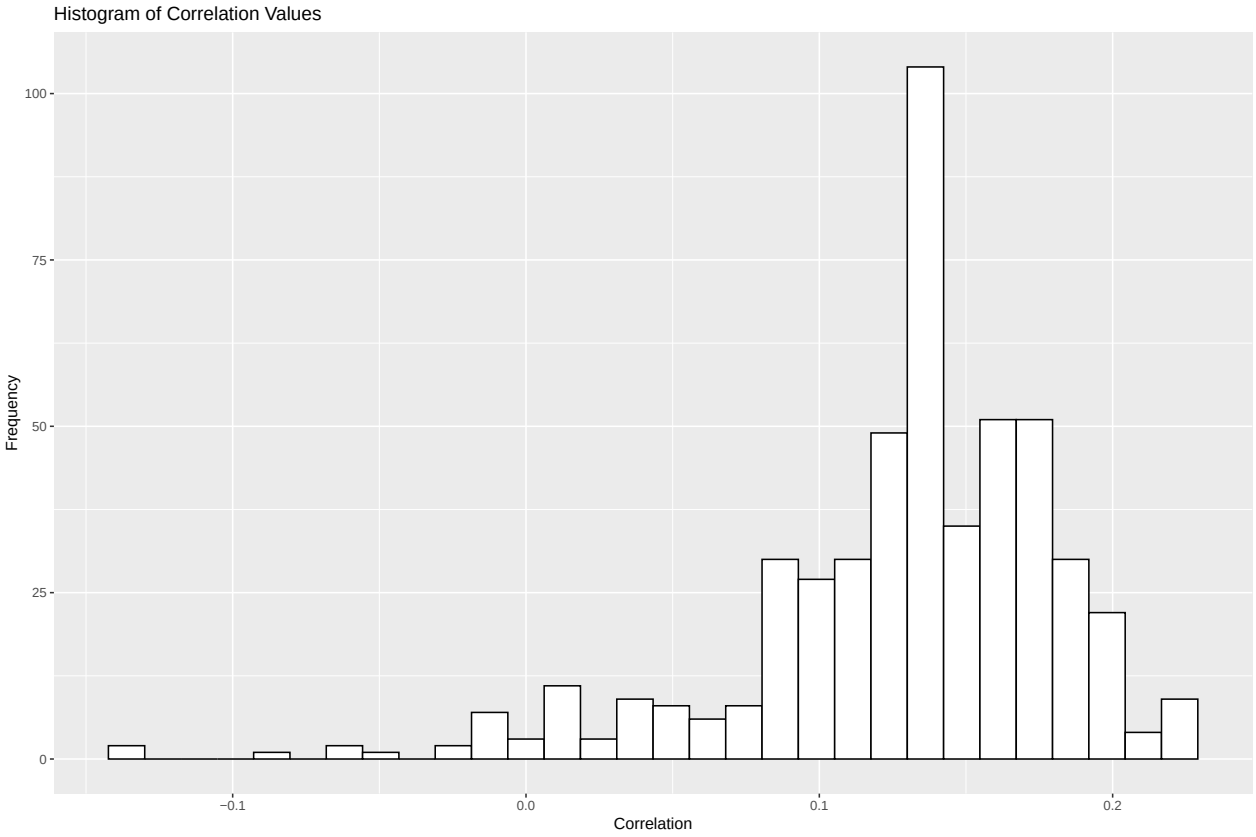

Selection Frequency of Features
Features
NA
ESR1
CCND1
CCNE1
RB1
MYC
CDKN1A
0.0
0.2
0.4
0.6
Selection Frequency

##

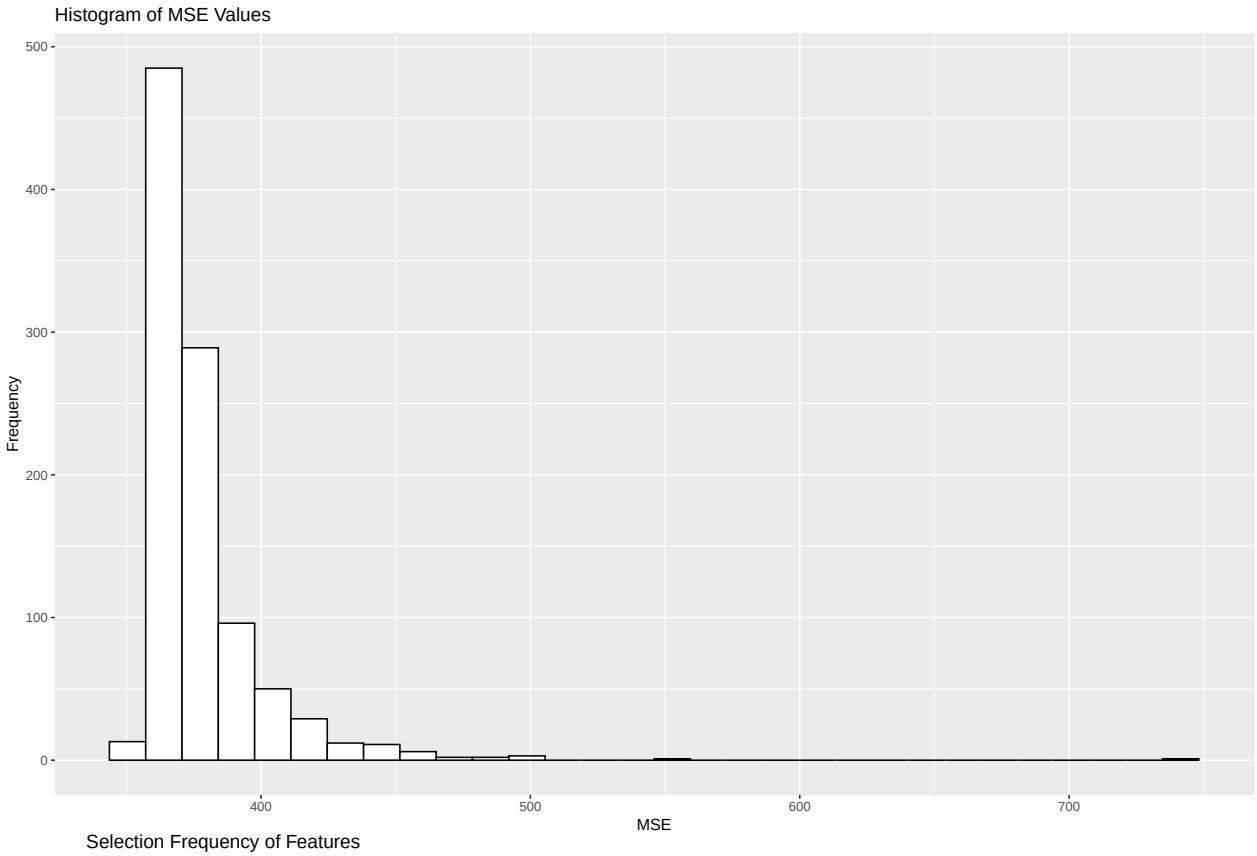
```
## Features selected 50% or more times:
## CCND1 CCNE1 ESR1
## Top 20 featrues:
##  [1] "ESR1"   "CCND1"  "CCNE1"  "RB1"    "MYC"    "CDKN1A" NA       NA
##  [9] NA       NA       NA       NA       NA       NA       NA       NA
## [17] NA       NA       NA       NA
```

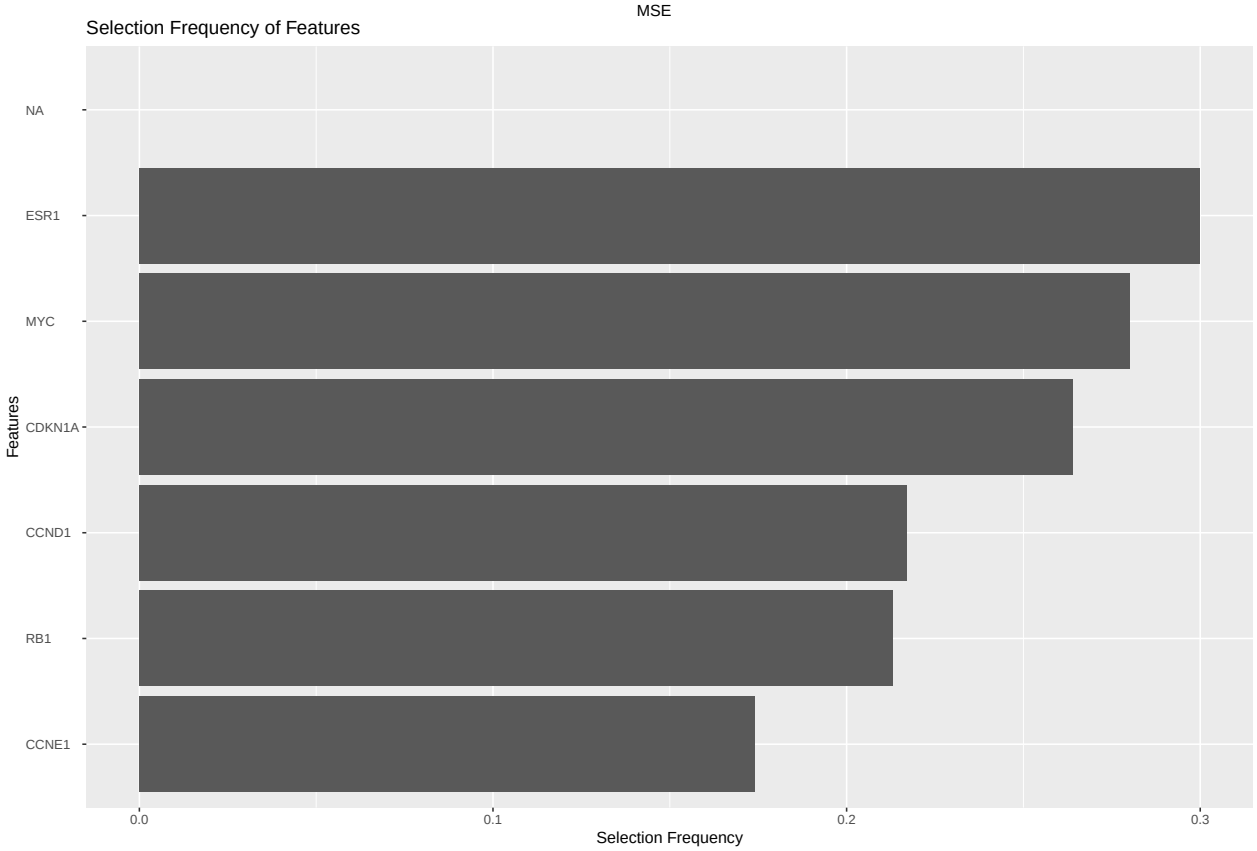
6 genes -> ROR_proliferation score

```
## 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

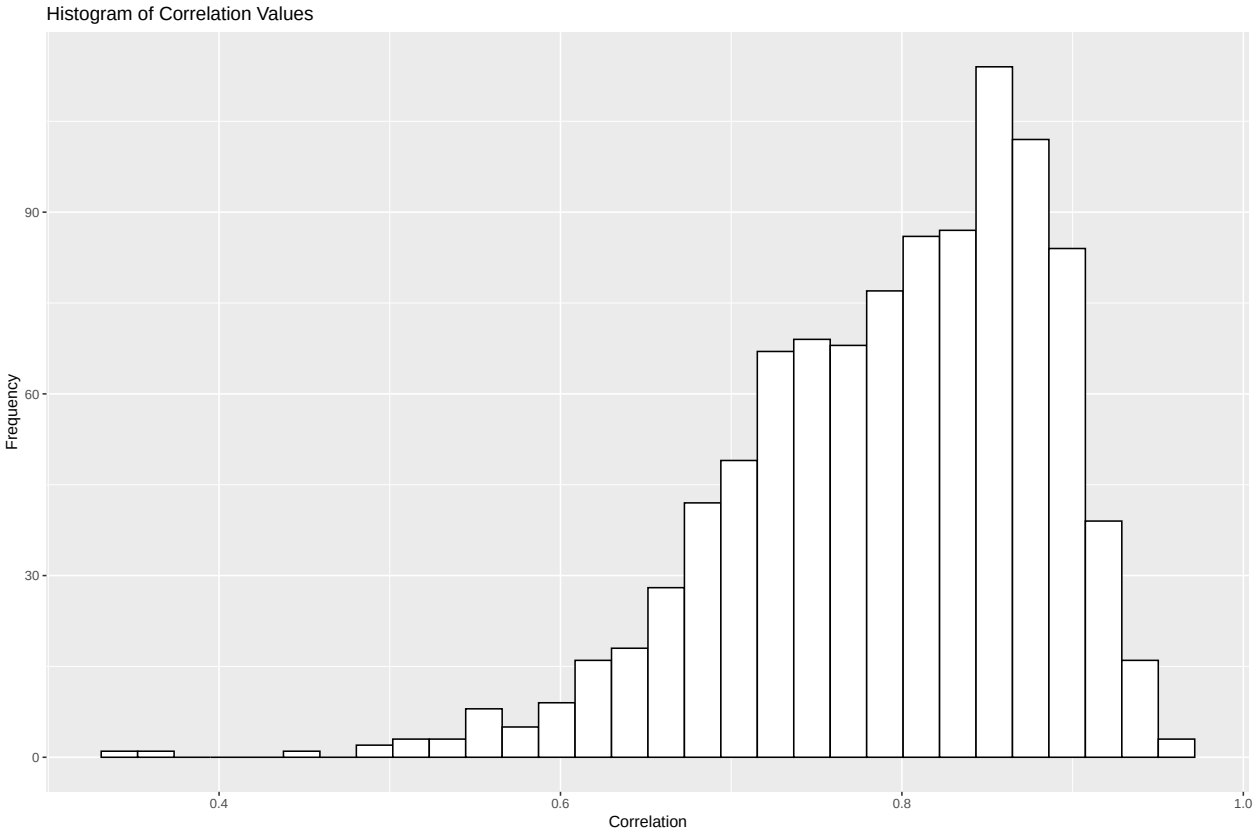

Selection Frequency of Features
Features
NA
ESR1
MYC
CDKN1A
CCND1
RB1
CCNE1
0.0
0.1
0.2
0.3
Selection Frequency

##

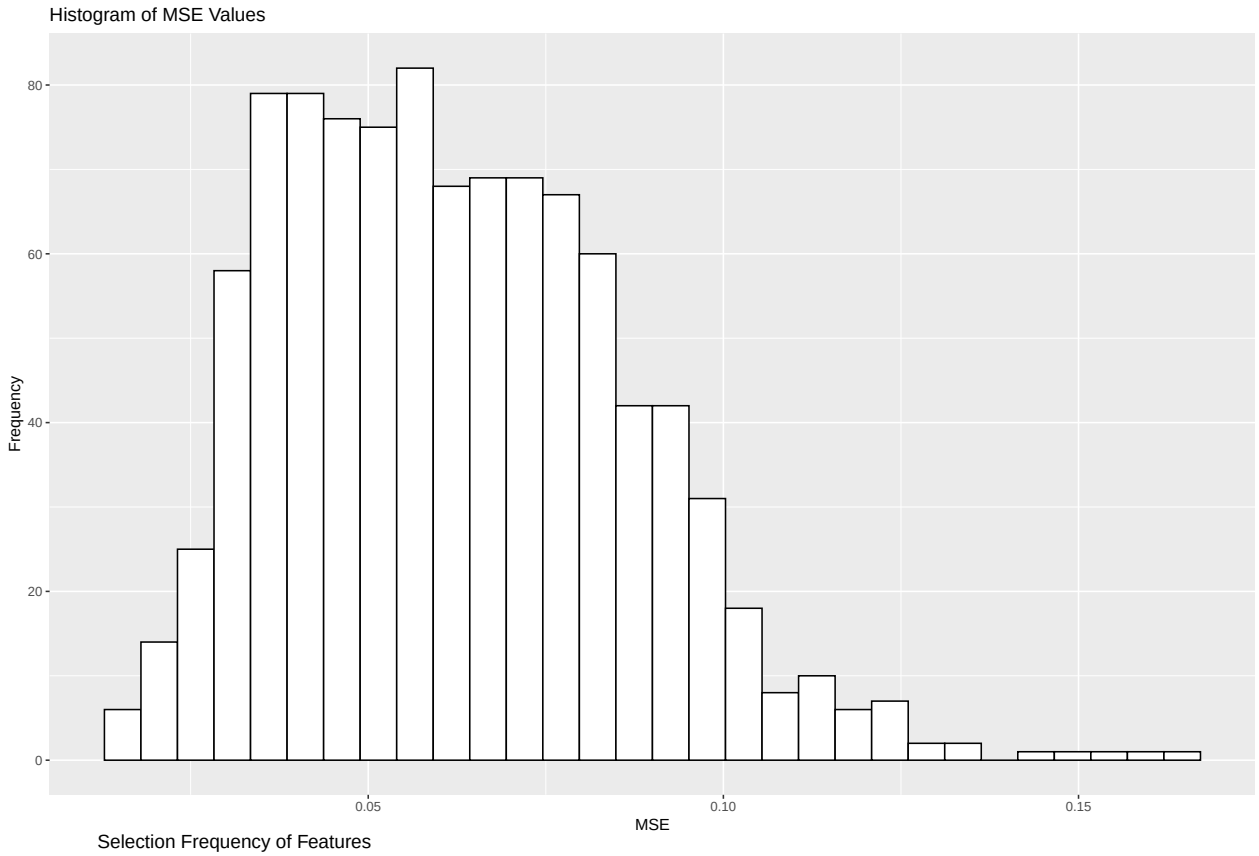
```
## Features selected 50% or more times:
##
## Top 20 featrues:
##  [1] "ESR1"   "MYC"    "CDKN1A" "CCND1"  "RB1"    "CCNE1"  NA       NA
##  [9] NA       NA       NA       NA       NA       NA       NA       NA
## [17] NA       NA       NA       NA
```

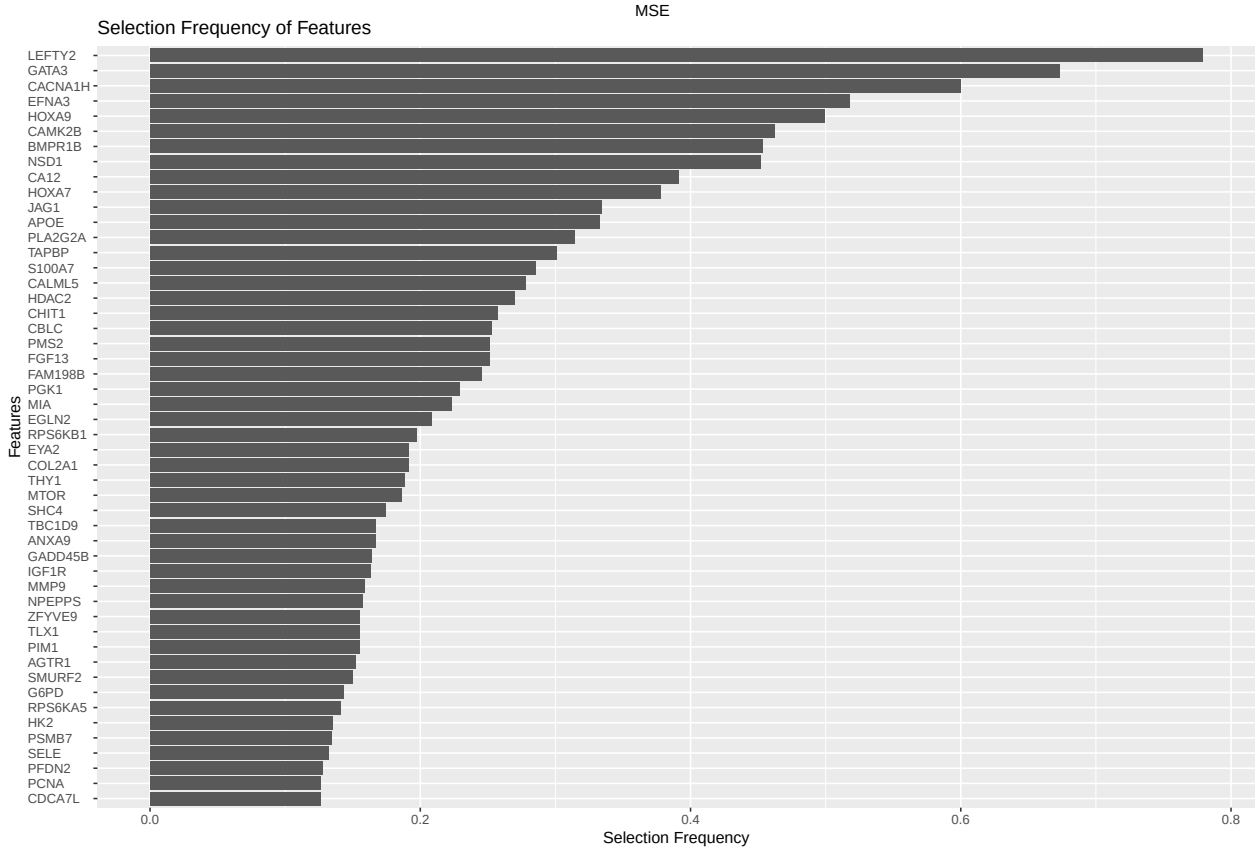
771 genes -> proliferation score

```
## 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

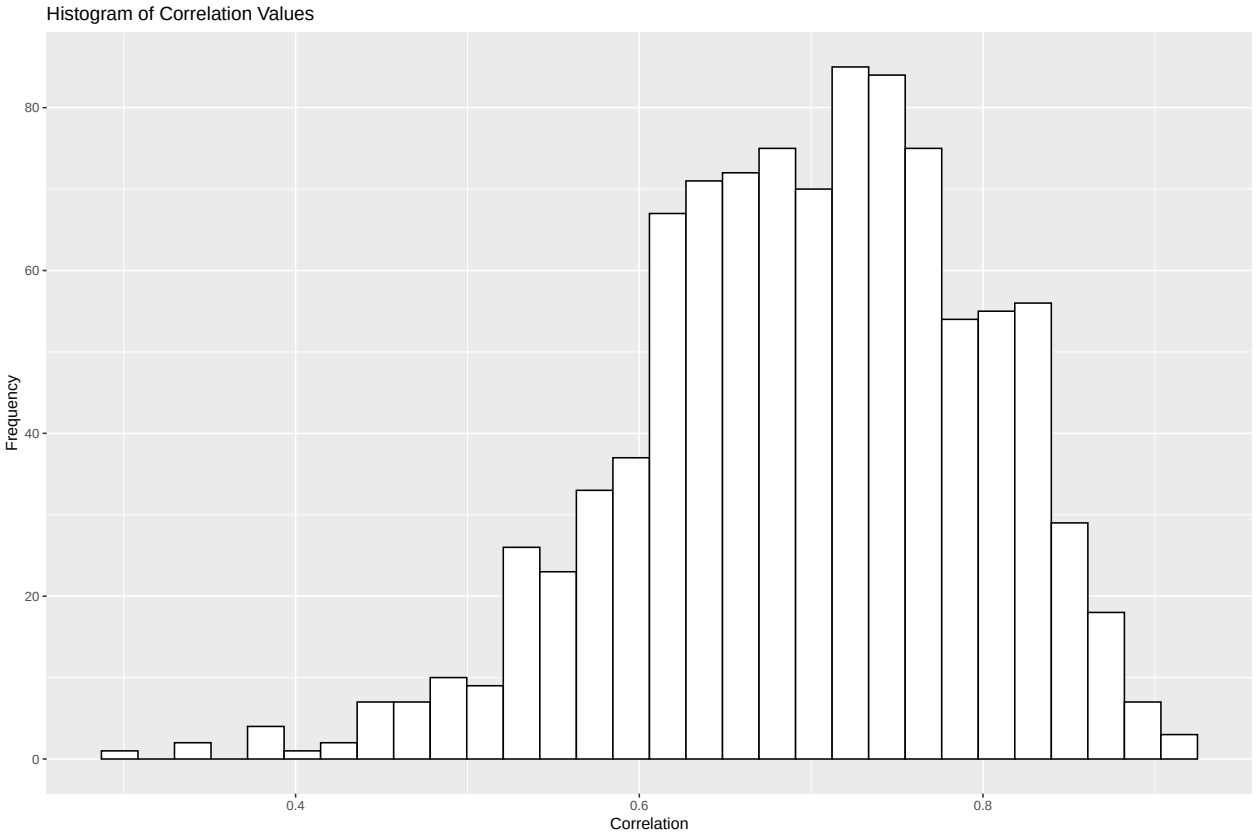

Selection Frequency of Features
Features
Selection Frequency
LEFTY2
GATA3
CACNA1H
EFNA3
HOXA9
CAMK2B
BMPR1B
NSD1
CA12
HOXA7
JAG1
APOE
PLA2G2A
TAPBP
S100A7
CALML5
HDAC2
CHIT1
CBLC
PMS2
FGF13
FAM198B
PGK1
MIA
EGLN2
RPS6KB1
EYA2
COL2A1
THY1
MTOR
SHC4
TBC1D9
ANXA9
GADD45B
IGF1R
MMP9
NPEPPS
ZFYVE9
TLX1
PIM1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

##

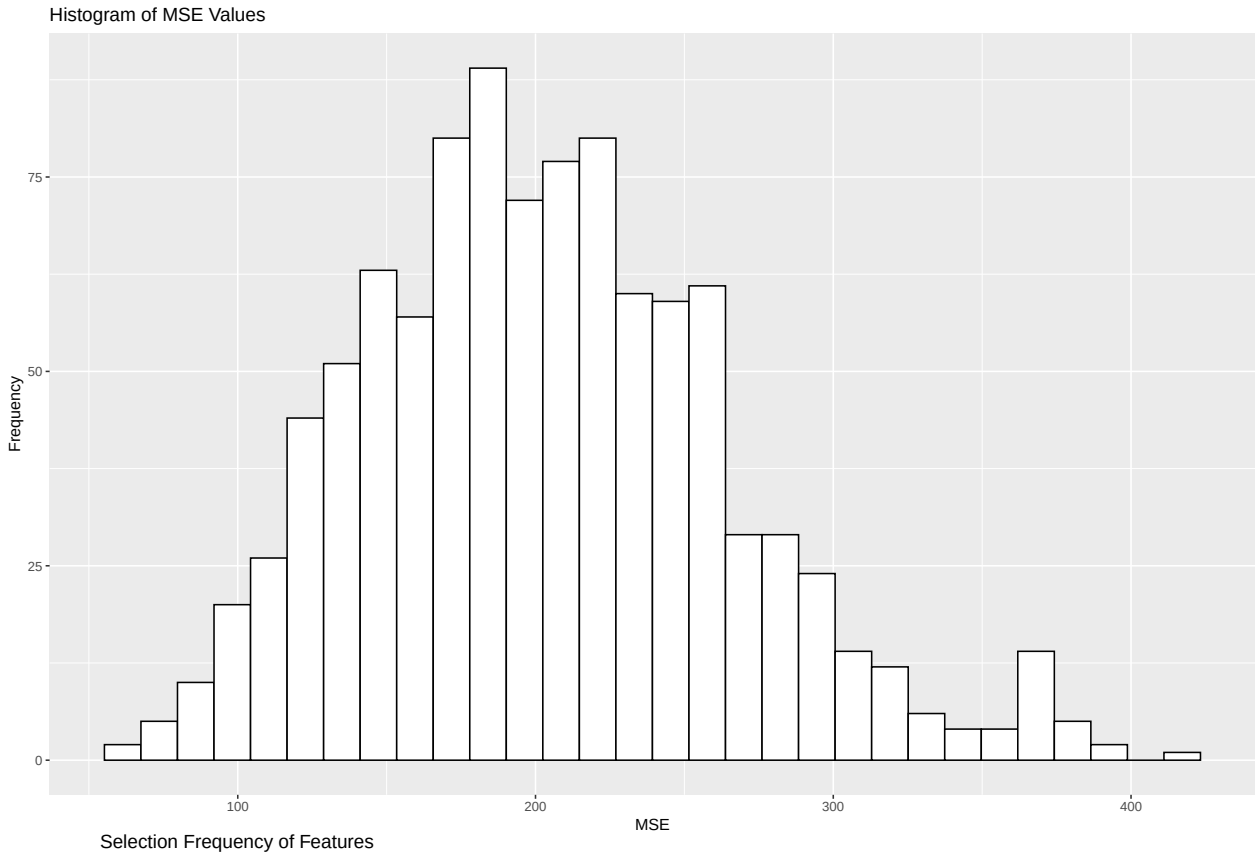
```
## Features selected 50% or more times:
## CACNA1H EFNA3 GATA3 LEFTY2
## Top 20 featrues:
##  [1] "LEFTY2"  "GATA3"   "CACNA1H" "EFNA3"   "HOXA9"   "CAMK2B"  "BMPR1B"
##  [8] "NSD1"    "CA12"    "HOXA7"   "JAG1"    "APOE"    "PLA2G2A" "TAPBP"
## [15] "S100A7"  "CALML5"  "HDAC2"   "CHIT1"   "CBLC"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

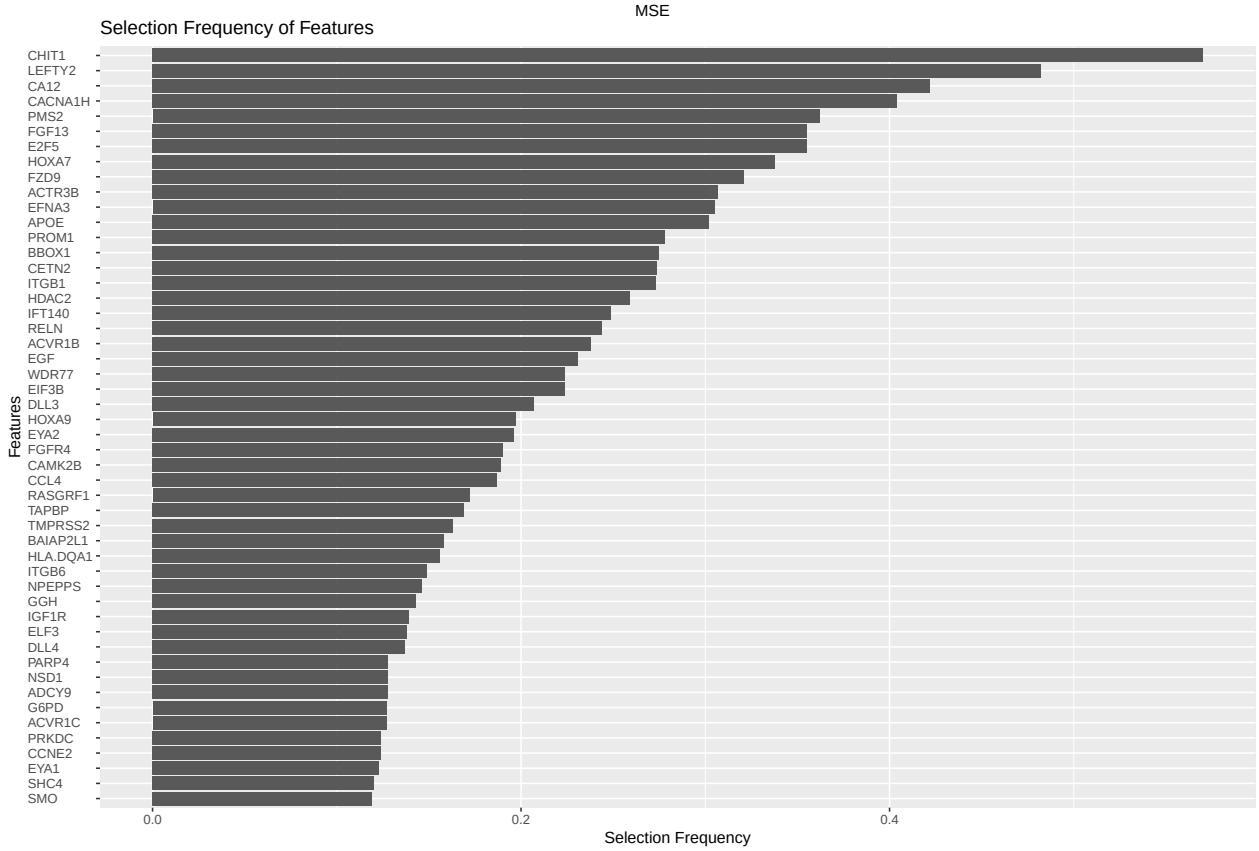
771 genes -> ROR-proliferation score

```
## 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

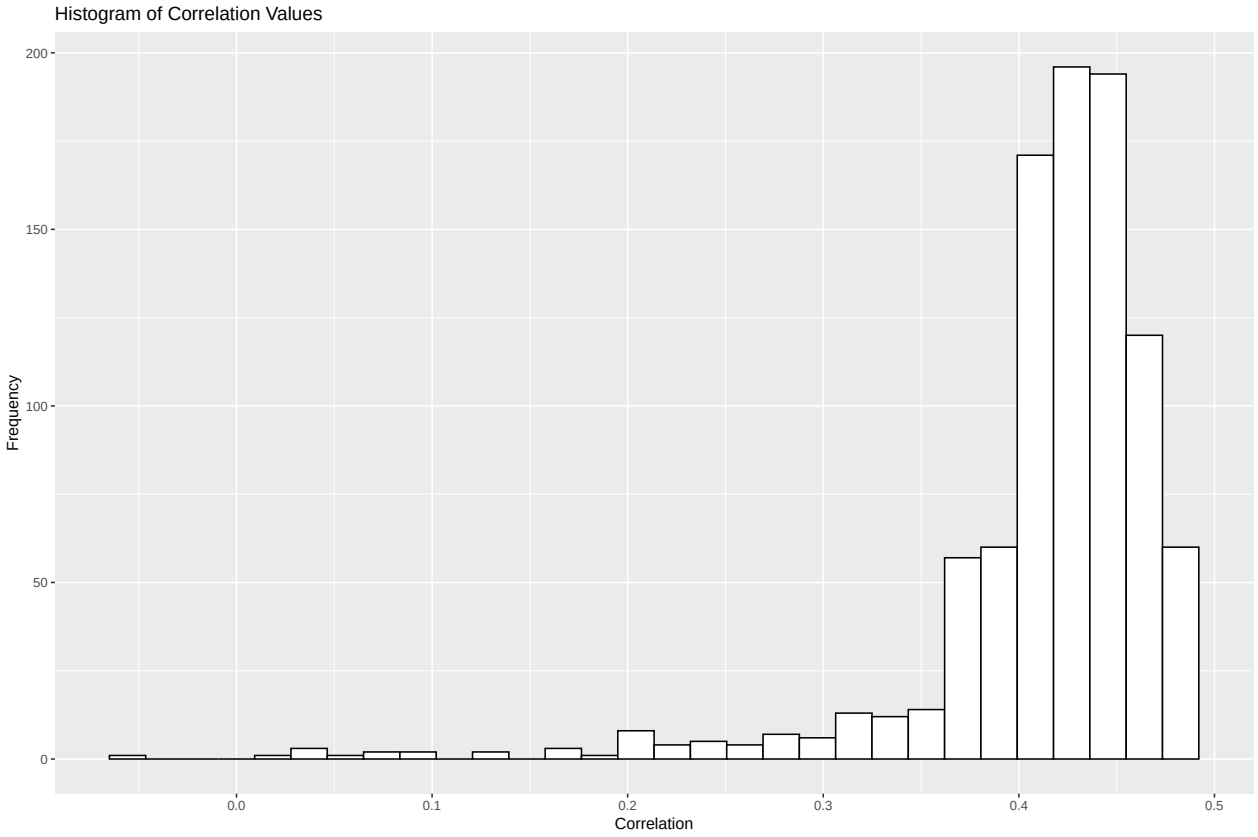

Selection Frequency of Features
Features
CHIT1
LEFTY2
CA12
CACNA1H
PMS2
FGF13
E2F5
HOXA7
FZD9
ACTR3B
EFNA3
APOE
PROM1
BBOX1
CETN2
ITGB1
HDAC2
IFT140
RELN
ACVR1B
EGF
WDR77
EIF3B
DLL3
HOXA9
EYA2
FGFR4
CAMK2B
CCL4
RASGRF1
TAPBP
TMPRSS2
BAIAP2L1
HLA.DQA1
ITGB6
NPEPPS
GGH
IGF1R
ELF3
DLL4
PARP4
NSD1
ADCY9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

##

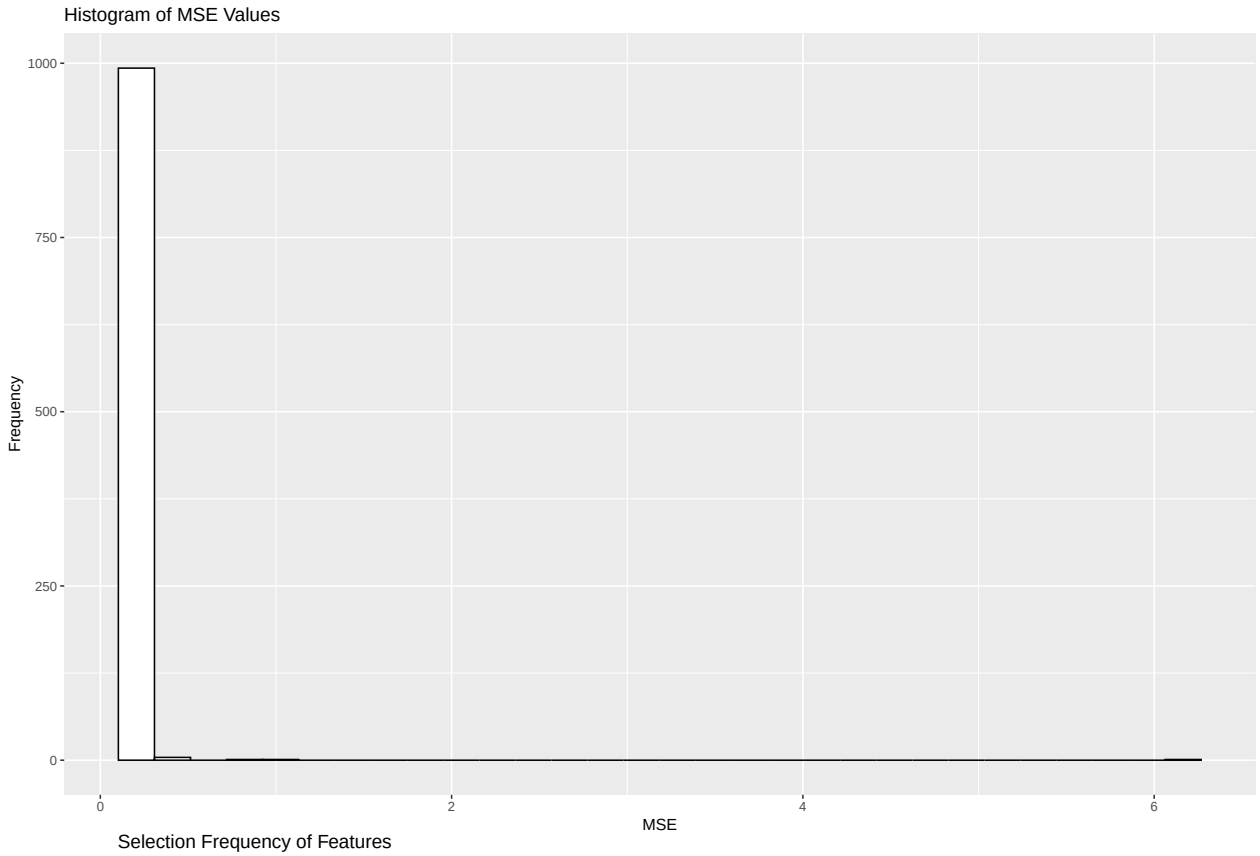
```
## Features selected 50% or more times:
## CHIT1
## Top 20 featrues:
##  [1] "CHIT1"   "LEFTY2"  "CA12"    "CACNA1H" "PMS2"    "E2F5"    "FGF13"
##  [8] "HOXA7"   "FZD9"    "ACTR3B"  "EFNA3"   "APOE"    "PROM1"   "BBOX1"
## [15] "CETN2"   "ITGB1"   "HDAC2"   "IFT140"  "RELN"    "ACVR1B"
```

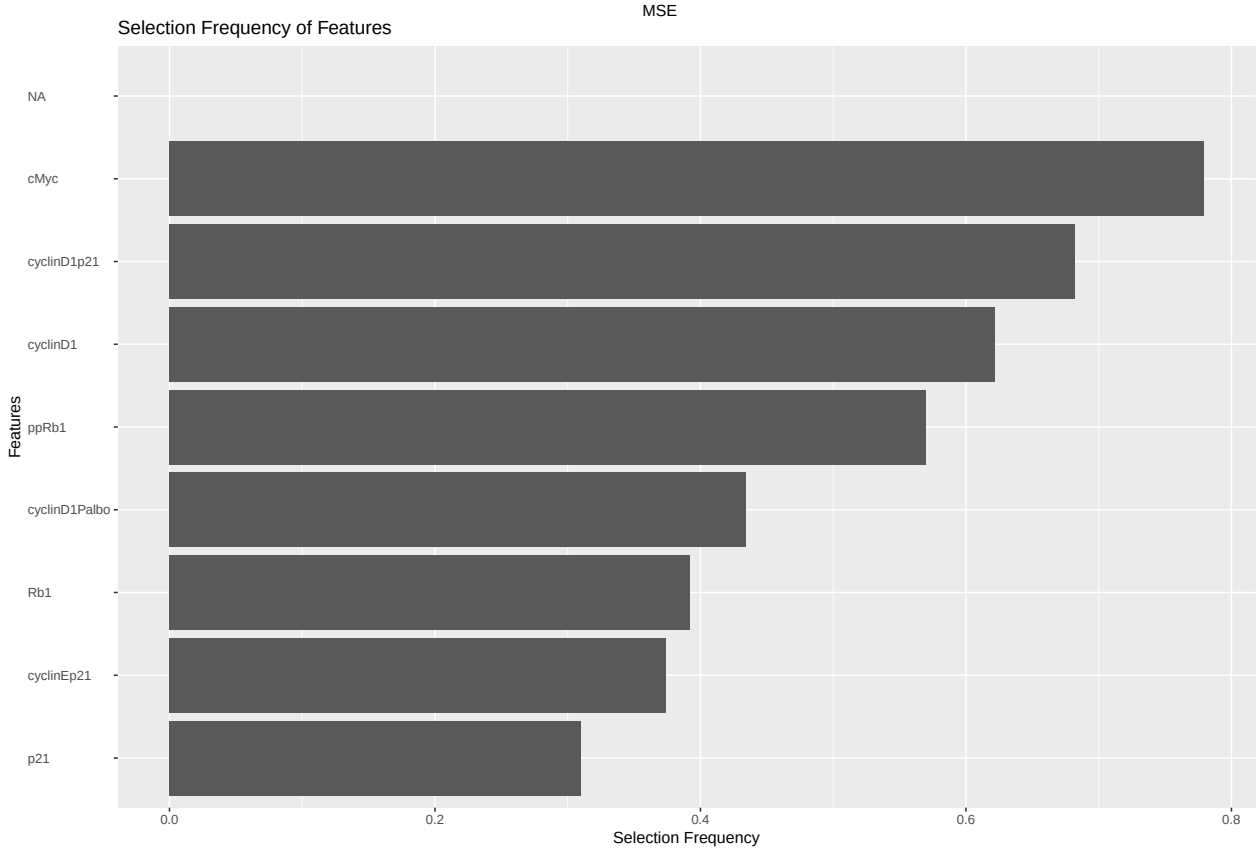
node values -> proliferation score

```
## 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

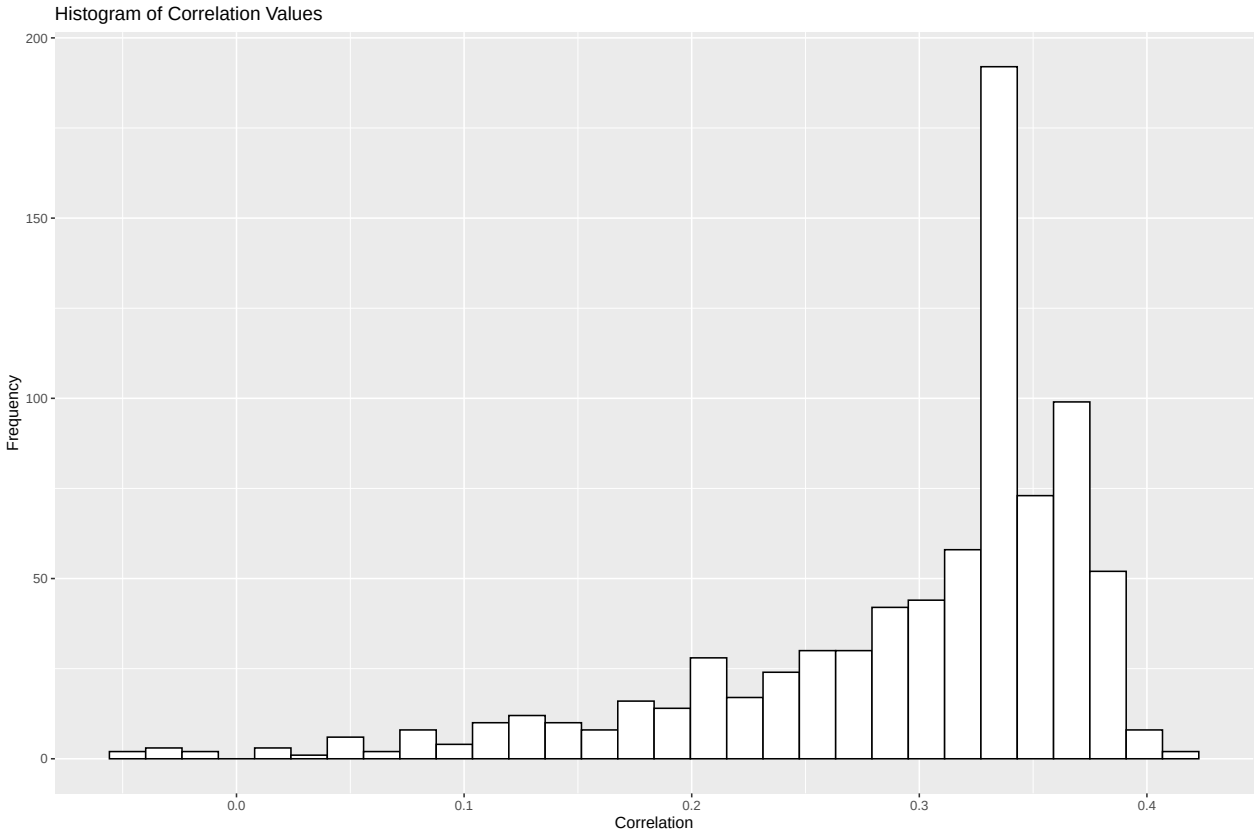

Selection Frequency of Features
Features
NA
cMyc
cyclinD1p21
cyclinD1
ppRb1
cyclinD1Palbo
Rb1
cyclinEp21
p21
0.0
0.2
0.4
0.6
0.8
Selection Frequency

##

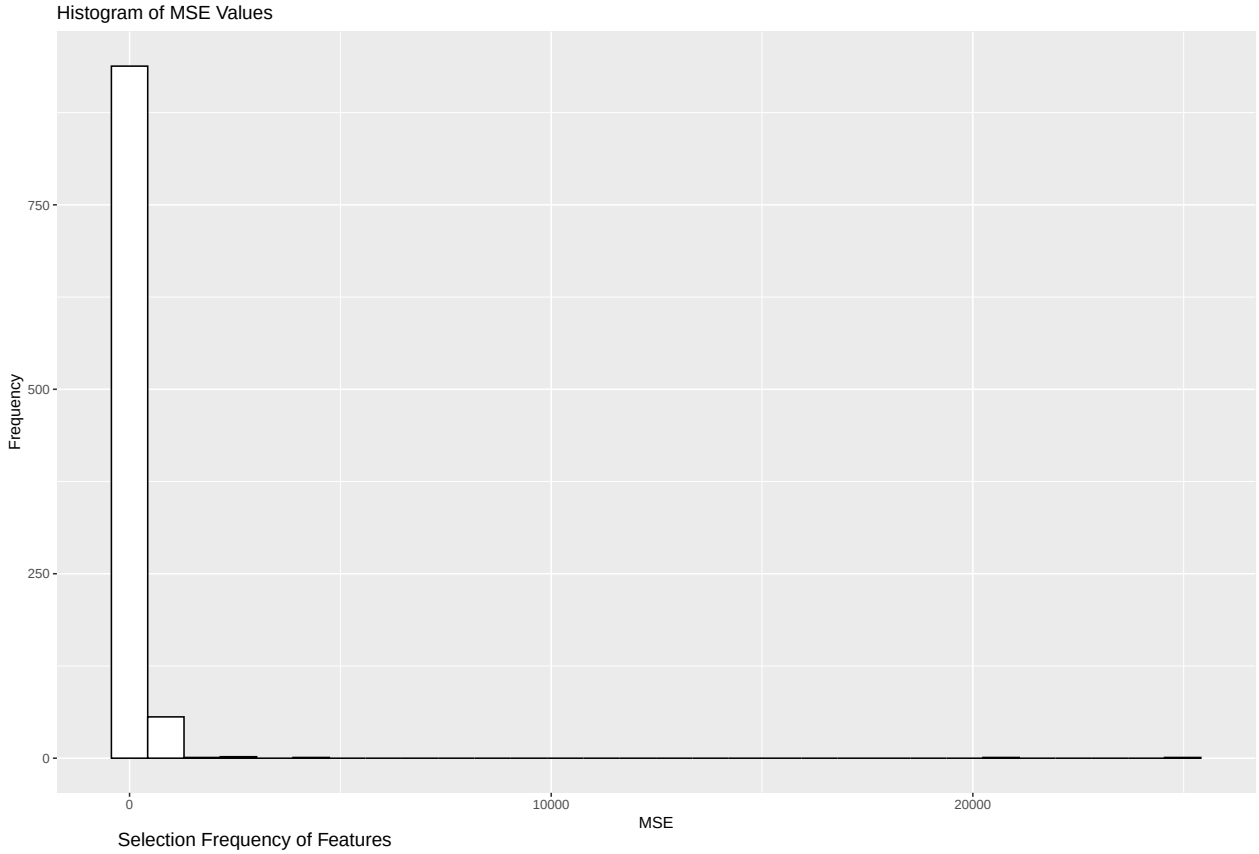
```
## Features selected 50% or more times:
## cyclinD1 cyclinD1p21 cMyc ppRb1
## Top 20 featrues:
##  [1] "cMyc"          "cyclinD1p21"   "cyclinD1"      "ppRb1"
##  [5] "cyclinD1Palbo" "Rb1"           "cyclinEp21"    "p21"
##  [9] NA              NA              NA              NA
## [13] NA      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

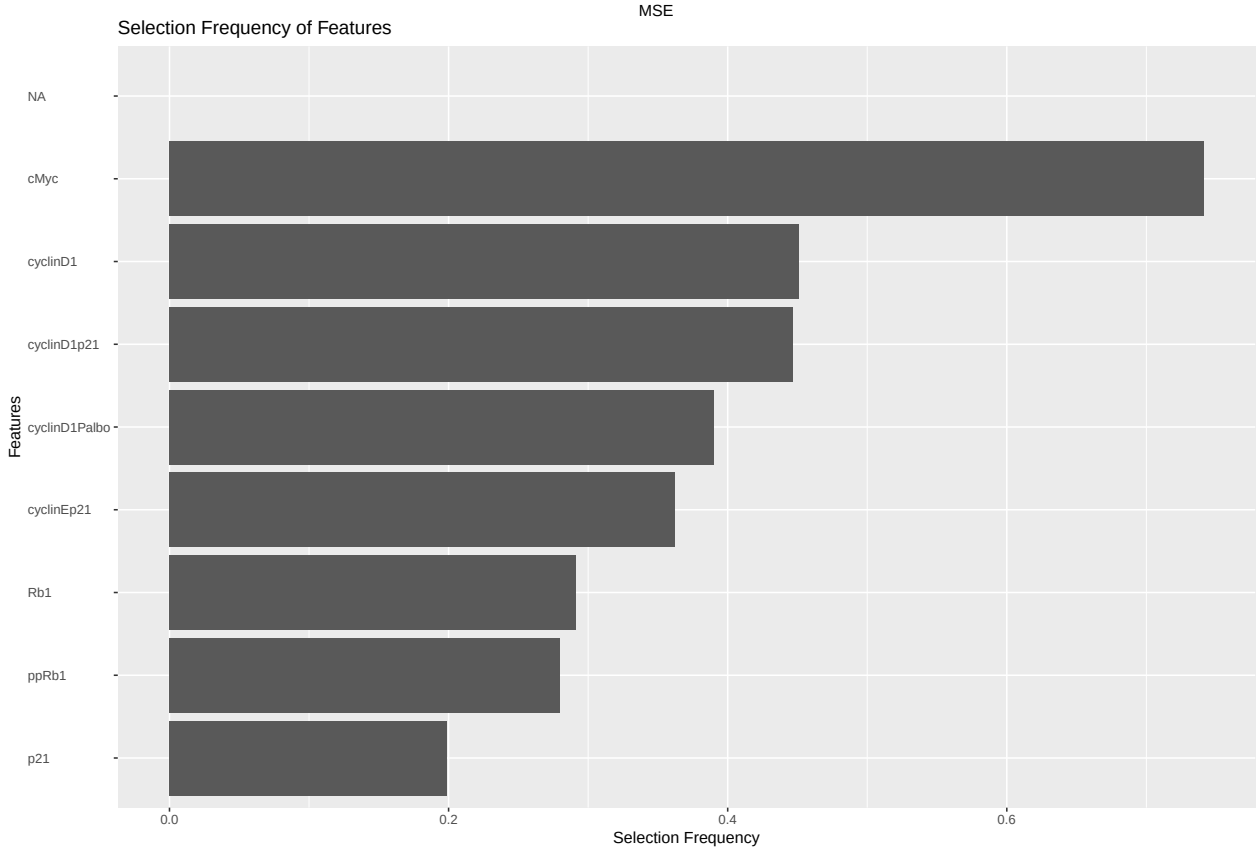
node values -> ROR-proliferation score

```
## 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

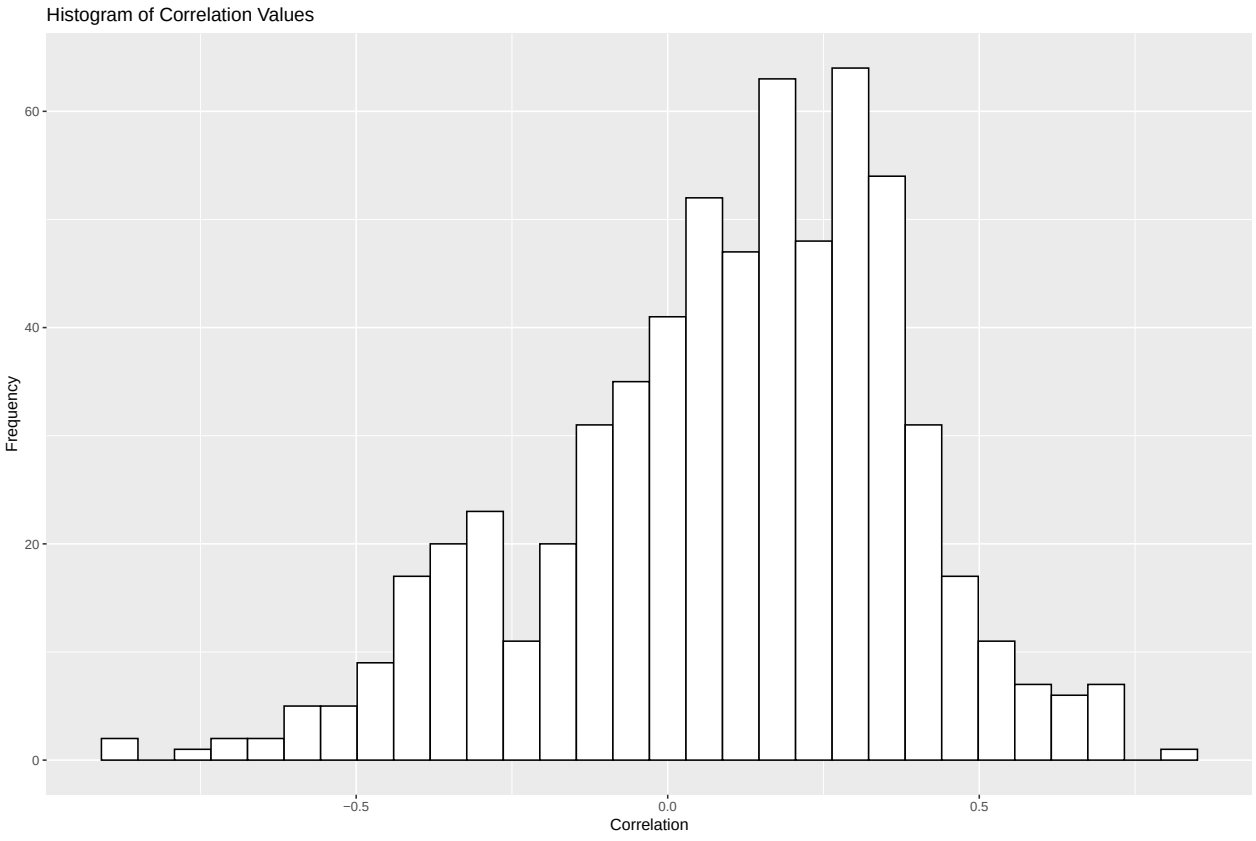
##

```
## 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
```

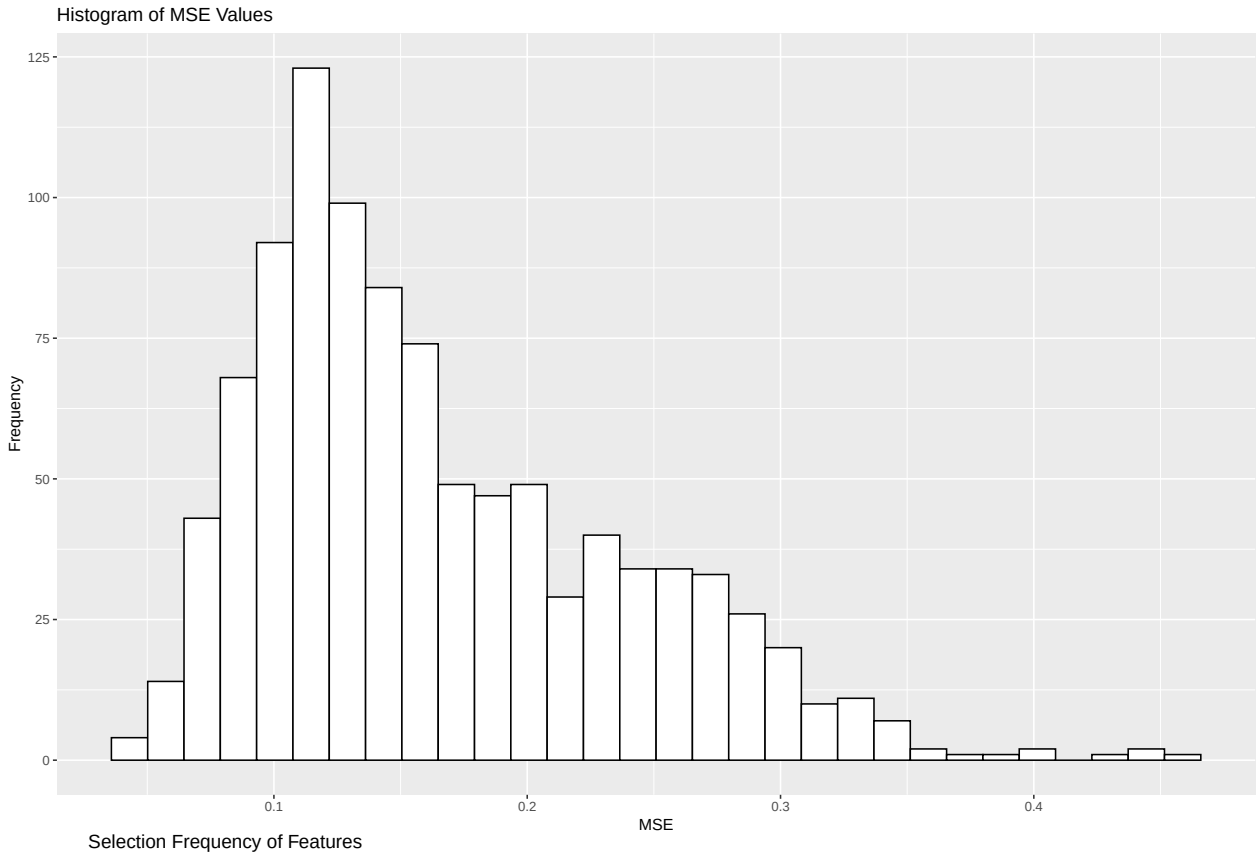
Repeated cross-validation

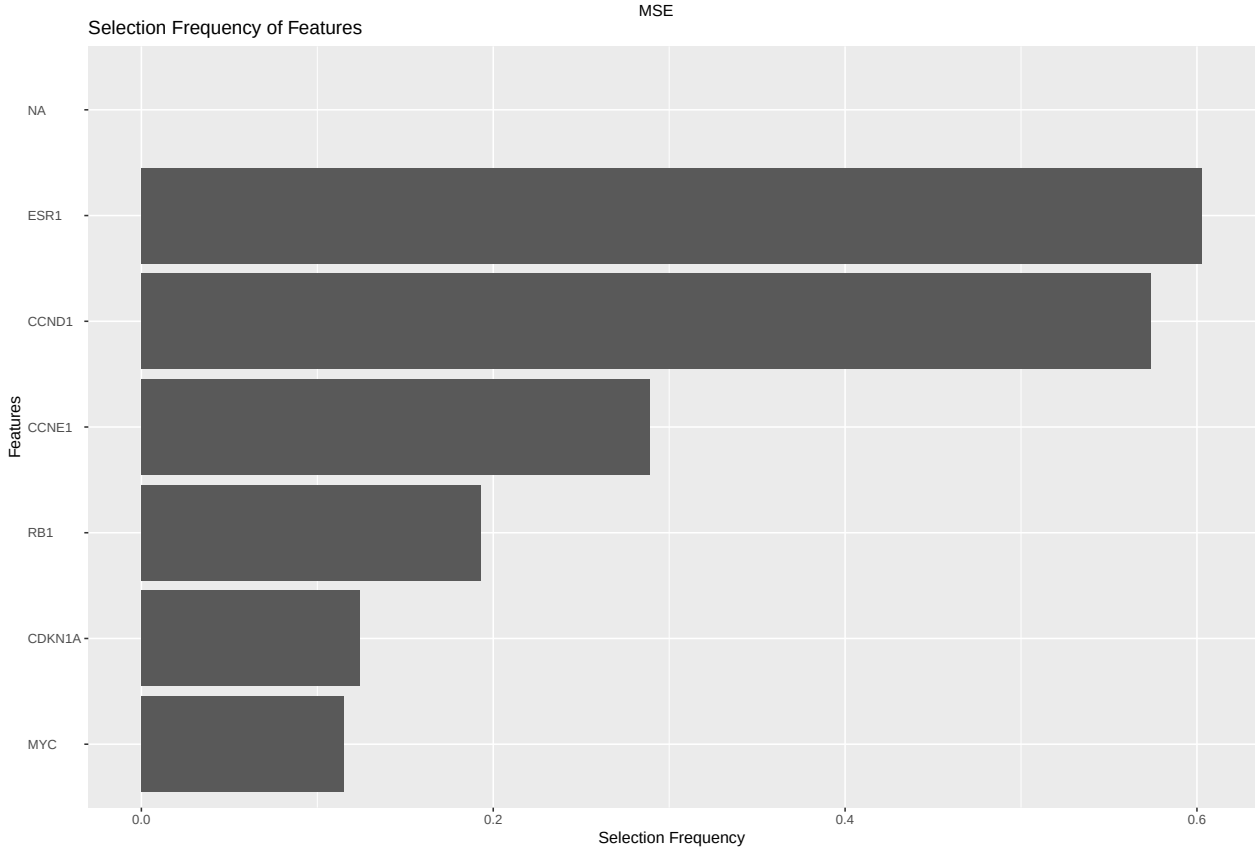
6 genes -> proliferation score

```
## 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
```


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

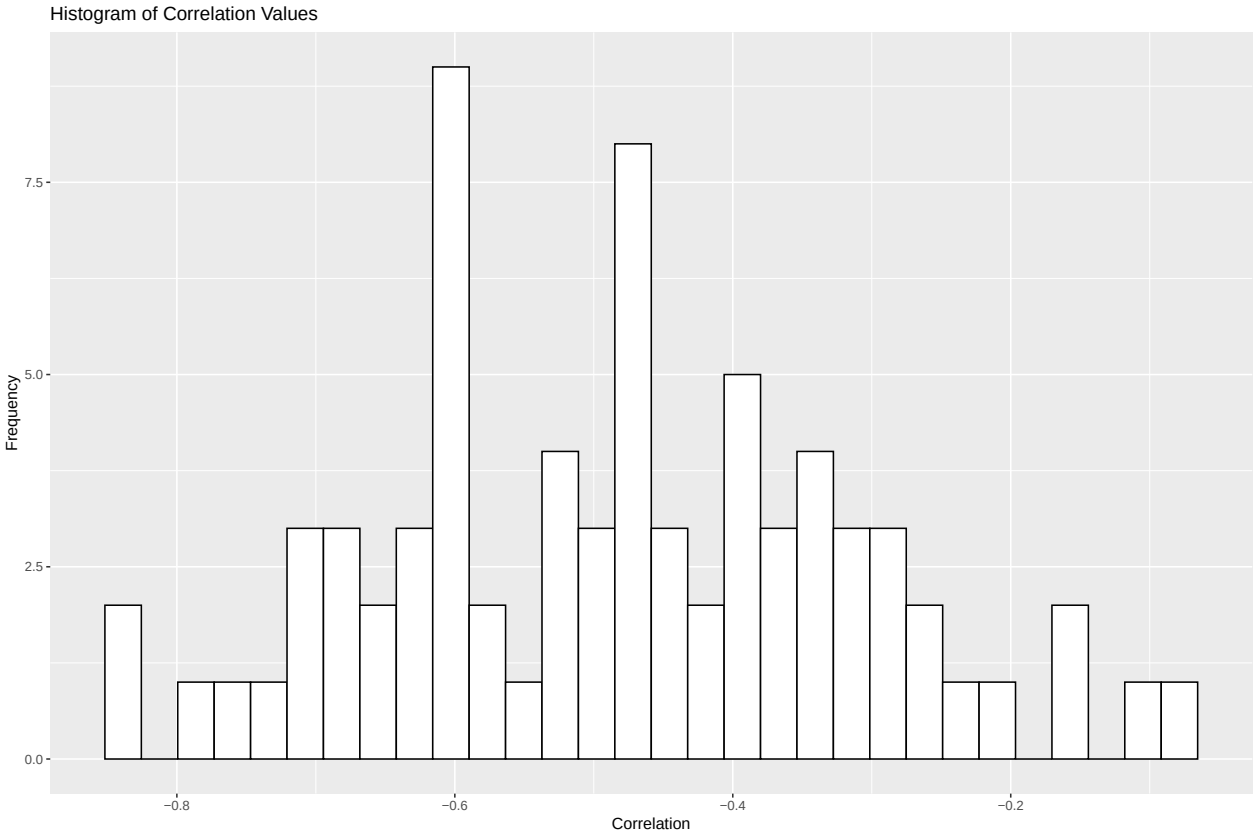

Selection Frequency of Features
Features
NA
ESR1
CCND1
CCNE1
RB1
CDKN1A
MYC
0.0
0.2
0.4
0.6
Selection Frequency

##

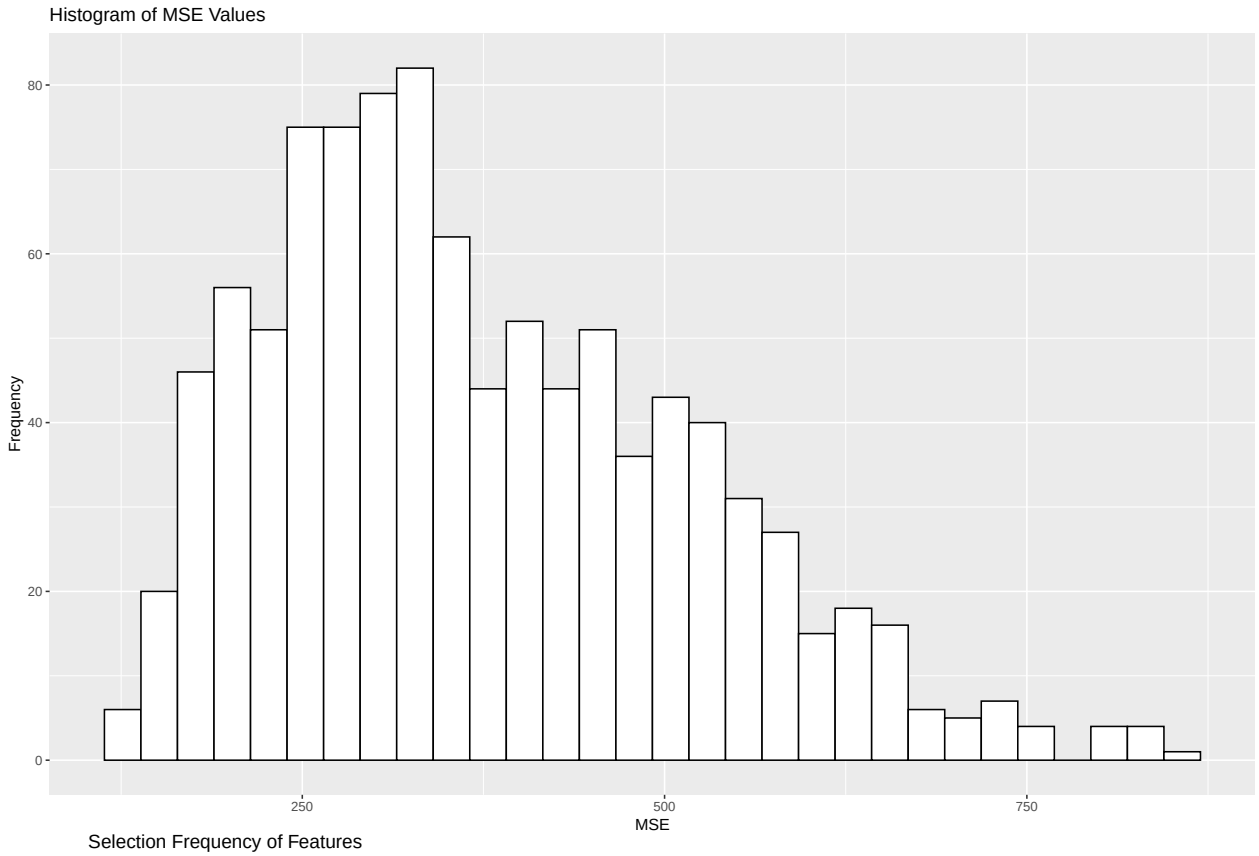
```
## Features selected 50% or more times:
## CCND1 ESR1
## Top 20 featrues:
##  [1] "ESR1"   "CCND1"  "CCNE1"  "RB1"    "CDKN1A" "MYC"    NA       NA
##  [9] NA       NA       NA       NA       NA       NA       NA       NA
## [17] NA       NA       NA       NA
```

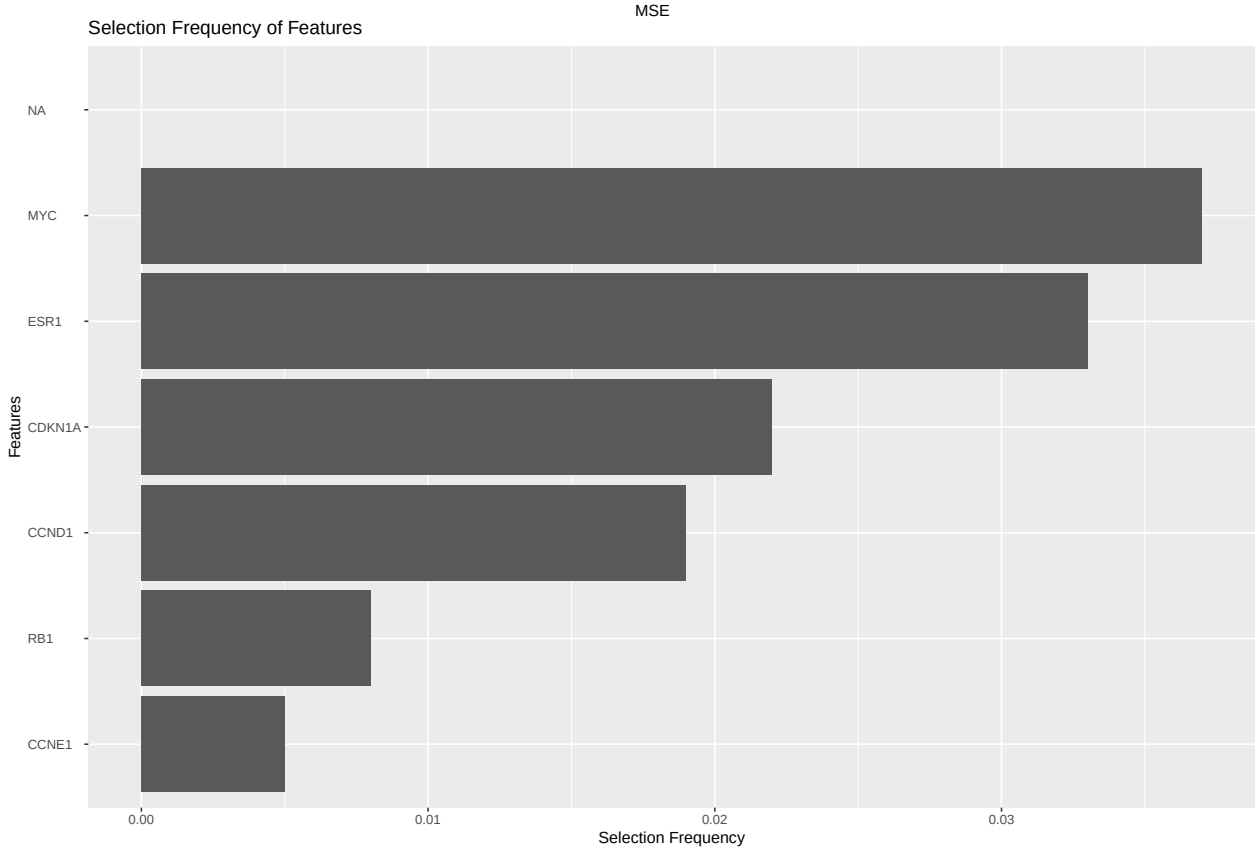
6 genes -> ROR_proliferation score

```
## 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

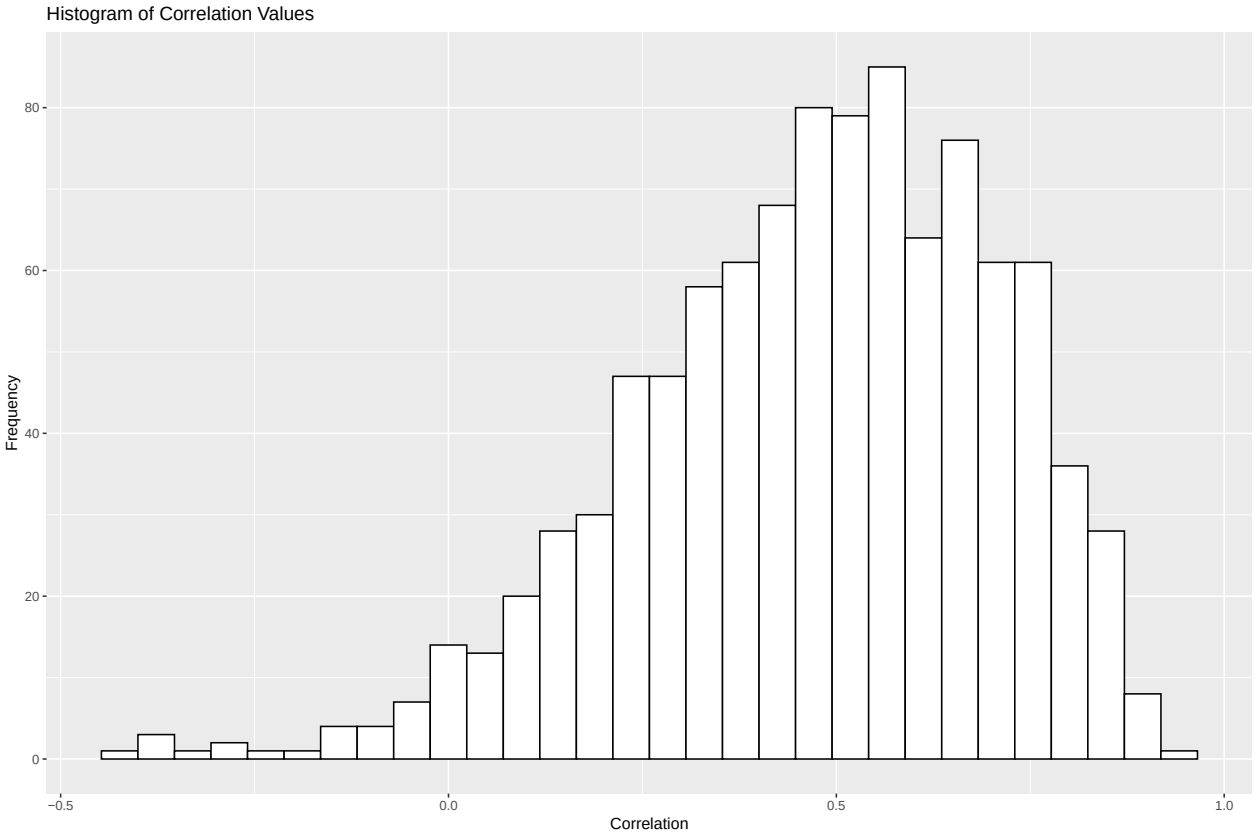

Selection Frequency of Features
Features
NA
MYC
ESR1
CDKN1A
CCND1
RB1
CCNE1
0.00
0.01
0.02
0.03
Selection Frequency

##

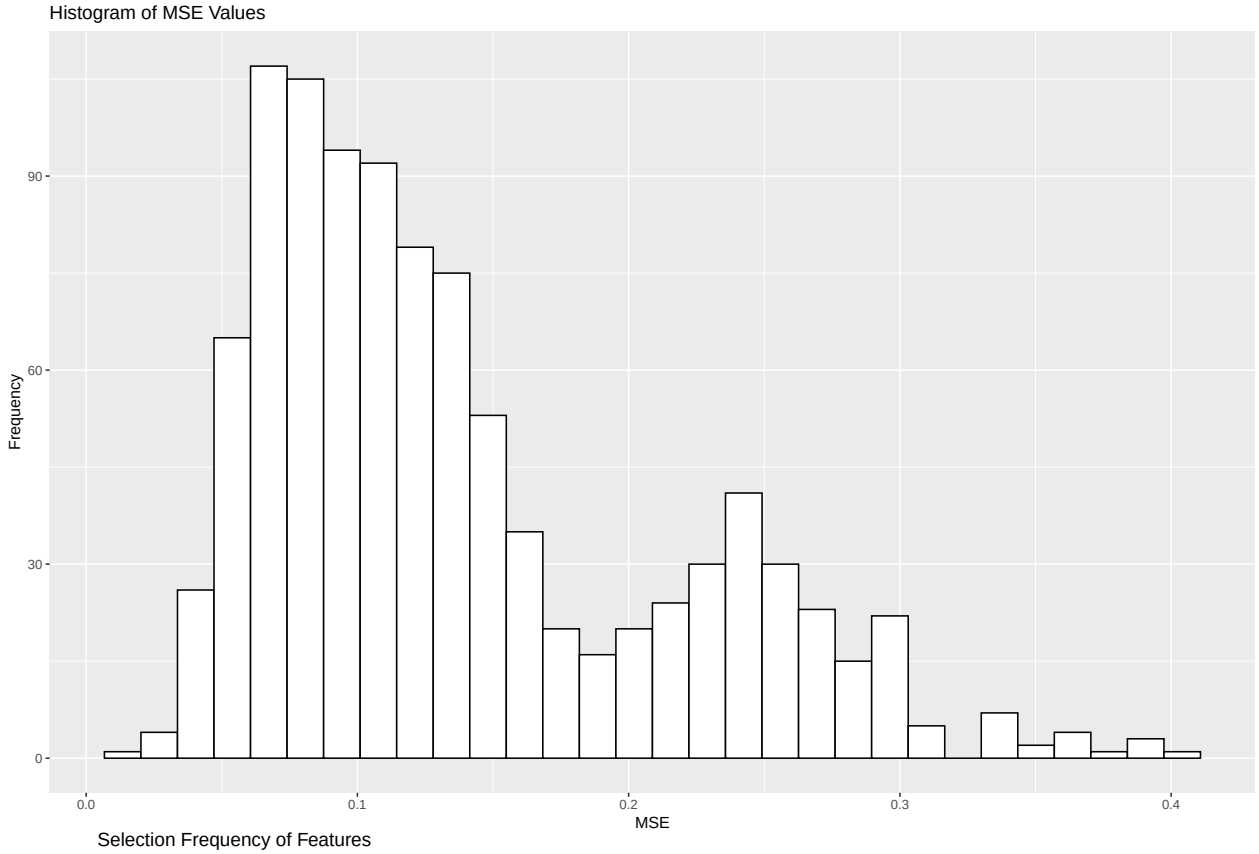
```
## Features selected 50% or more times:
##
## Top 20 featrues:
##  [1] "MYC"    "ESR1"   "CDKN1A" "CCND1"  "RB1"    "CCNE1"  NA       NA
##  [9] NA       NA       NA       NA       NA       NA       NA       NA
## [17] NA       NA       NA       NA
```

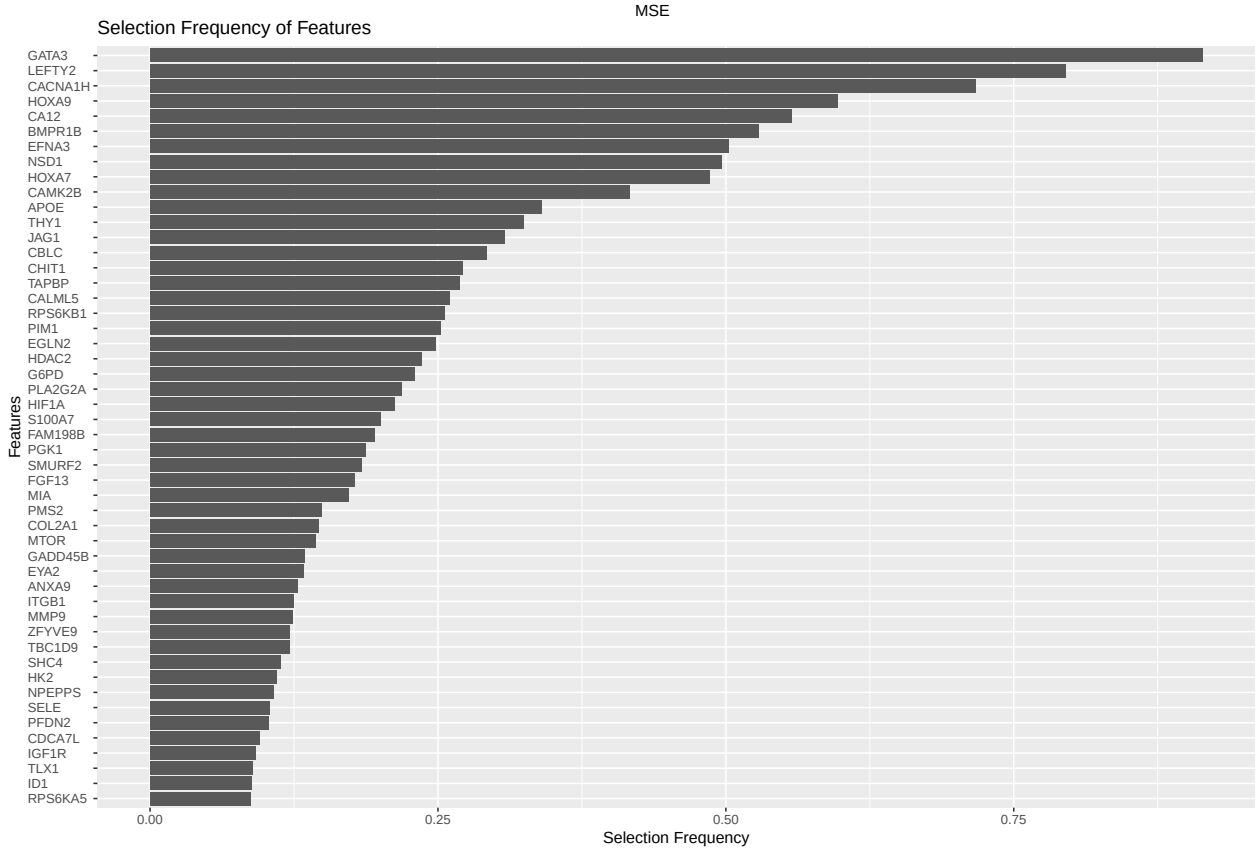
771 genes -> proliferation score

```
## 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

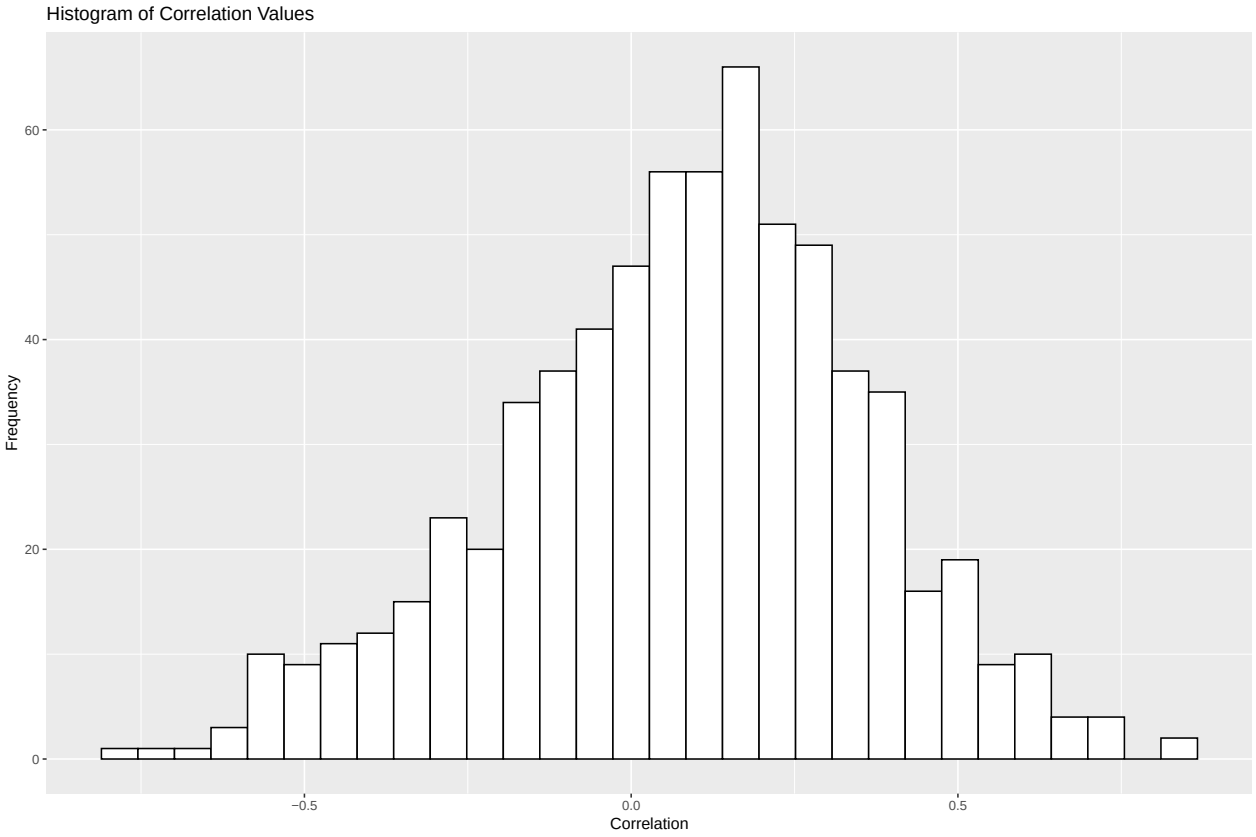

Selection Frequency of Features
Features
GATA3
LEFTY2
CACNA1H
HOXA9
CA12
BMPR1B
EFNA3
NSD1
HOXA7
CAMK2B
APOE
THY1
JAG1
CBLC
CHIT1
TAPBP
CALML5
RPS6KB1
PIM1
EGLN2
HDAC2
G6PD
PLA2G2A
HIF1A
S100A7
FAM198B
PGK1
SMURF2
FGF13
MIA
PMS2
COL2A1
MTOR
GADD45B
EYA2
ANXA9
ITGB1
MMP9
ZFYVE9
TBC1D9
SHC4
HK2
NPEPPS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

##

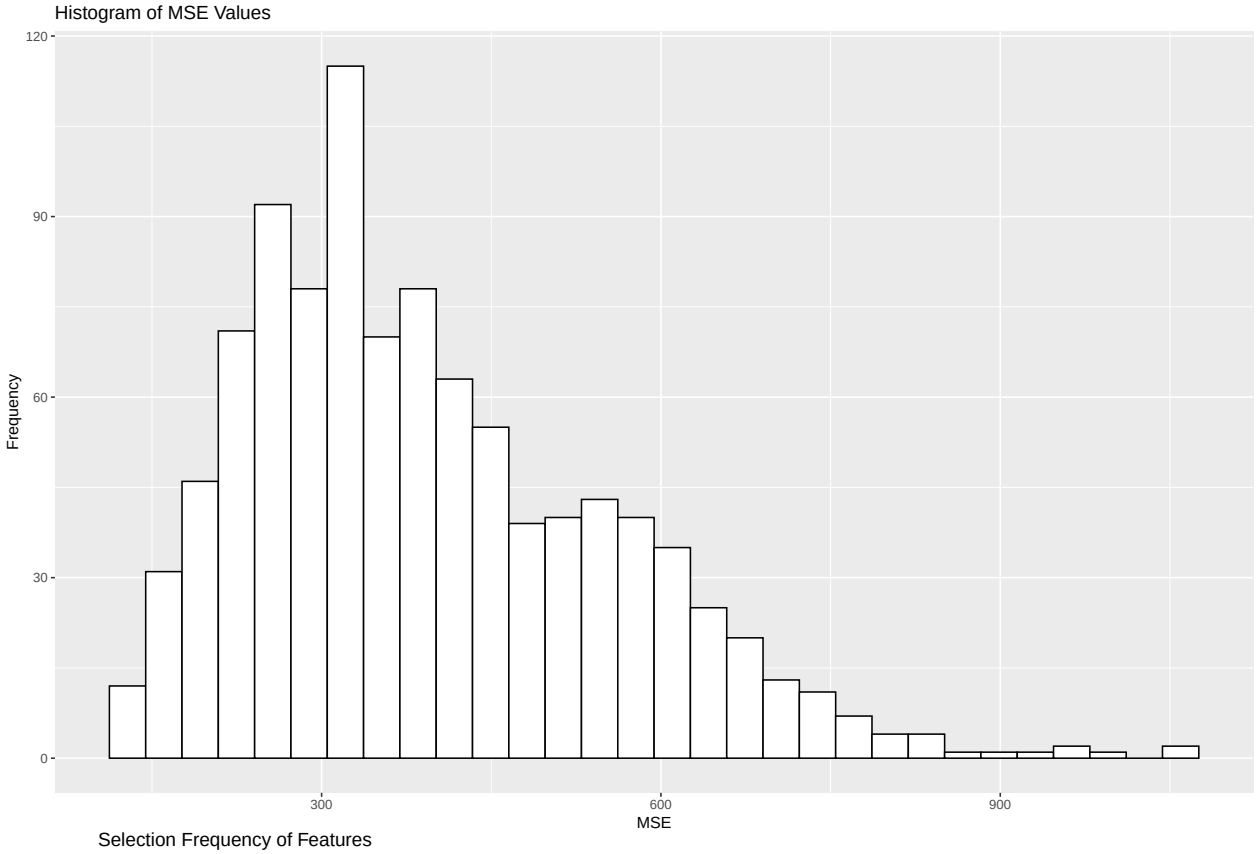
```
## Features selected 50% or more times:
## BMPR1B CA12 CACNA1H EFNA3 GATA3 HOXA9 LEFTY2
## Top 20 featrues:
##  [1] "GATA3"   "LEFTY2"  "CACNA1H" "HOXA9"   "CA12"    "BMPR1B"  "EFNA3"
##  [8] "NSD1"    "HOXA7"   "CAMK2B"  "APOE"    "THY1"    "JAG1"    "CBLC"
## [15] "CHIT1"   "TAPBP"   "CALML5" 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

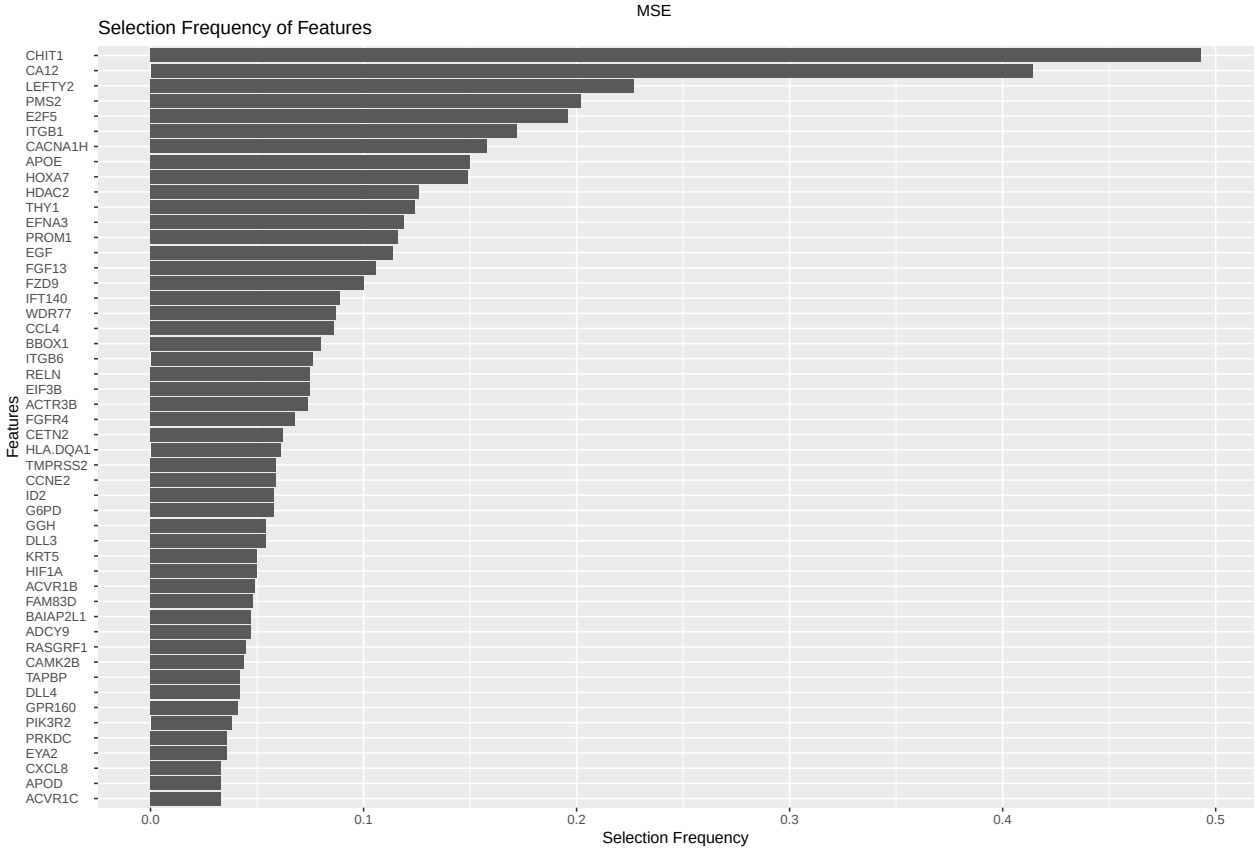
771 genes -> ROR-proliferation score

```
## 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

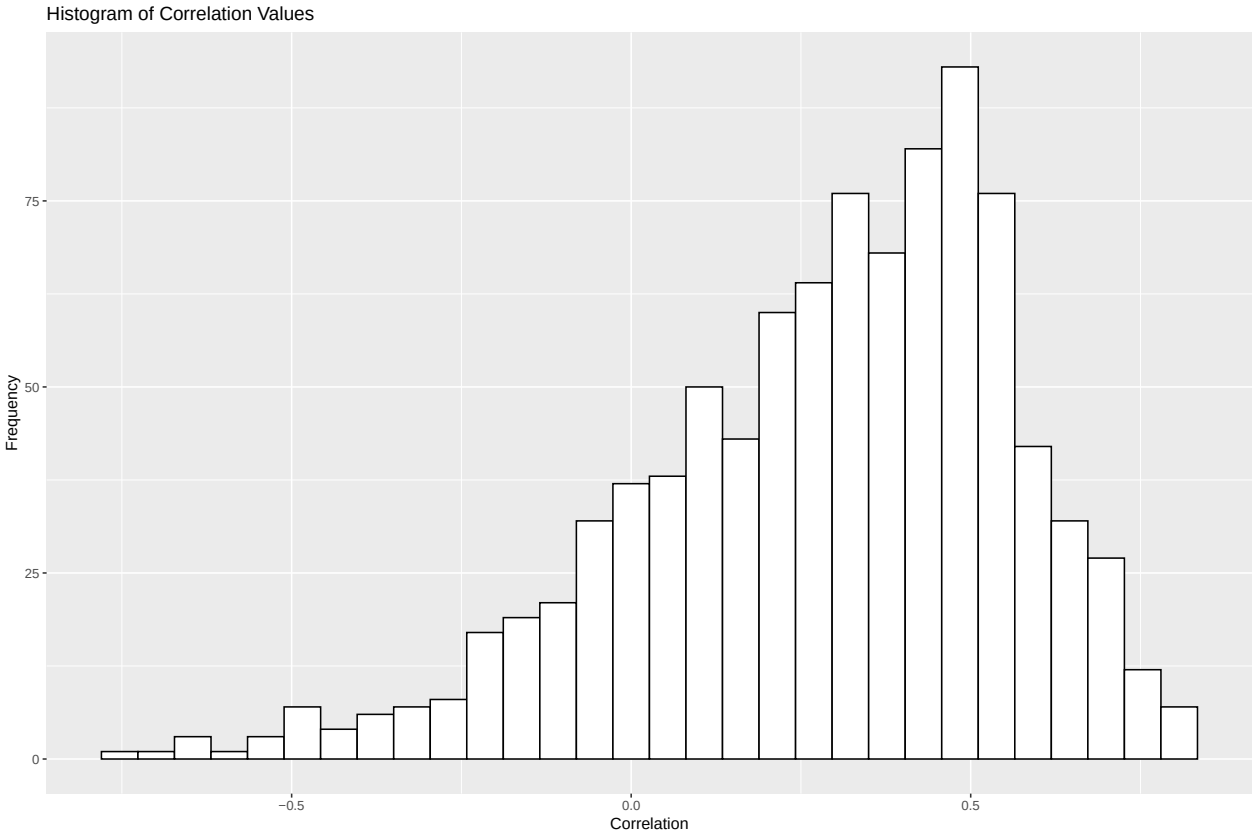

Selection Frequency of Features
Features
CHIT1
CA12
LEFTY2
PMS2
E2F5
ITGB1
CACNA1H
APOE
HOXA7
HDAC2
THY1
EFNA3
PROM1
EGF
FGF13
FZD9
IFT140
WDR77
CCL4
BBOX1
ITGB6
RELN
EIF3B
ACTR3B
FGFR4
CETN2
HLA.DQA1
TMPRSS2
CCNE2
ID2
G6PD
GGH
DLL3
KRT5
HIF1A
ACVR1B
FAM83D
BAIAP2L1
ADCY9
RASGRF1
CAMK2B
TAPBP
DLL4
GPR160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

##

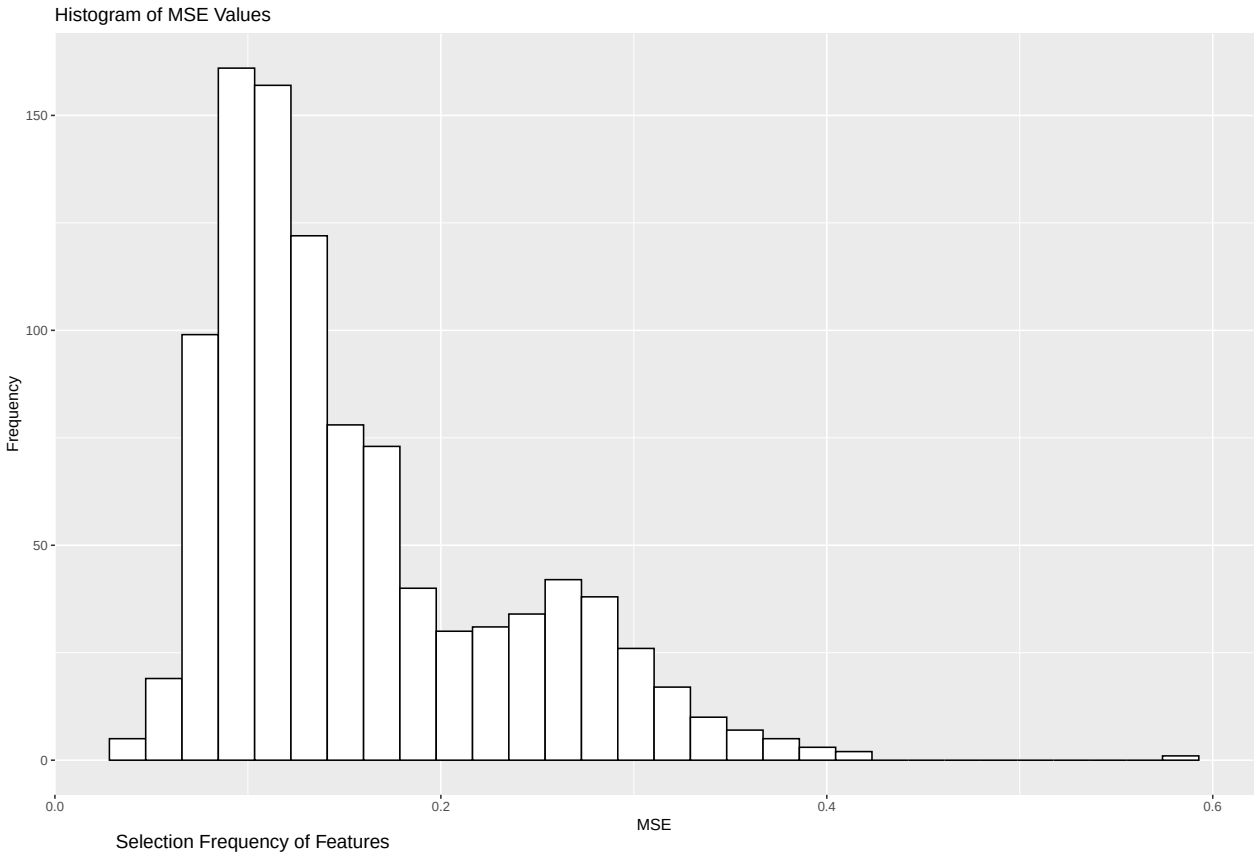
```
## Features selected 50% or more times:
##
## Top 20 featrues:
##  [1] "CHIT1"   "CA12"    "LEFTY2"  "PMS2"    "E2F5"    "ITGB1"   "CACNA1H"
##  [8] "APOE"    "HOXA7"   "HDAC2"   "THY1"    "EFNA3"   "PROM1"   "EGF"
## [15] "FGF13"   "FZD9"    "IFT140"  "WDR77"   "CCL4"    "BBOX1"
```

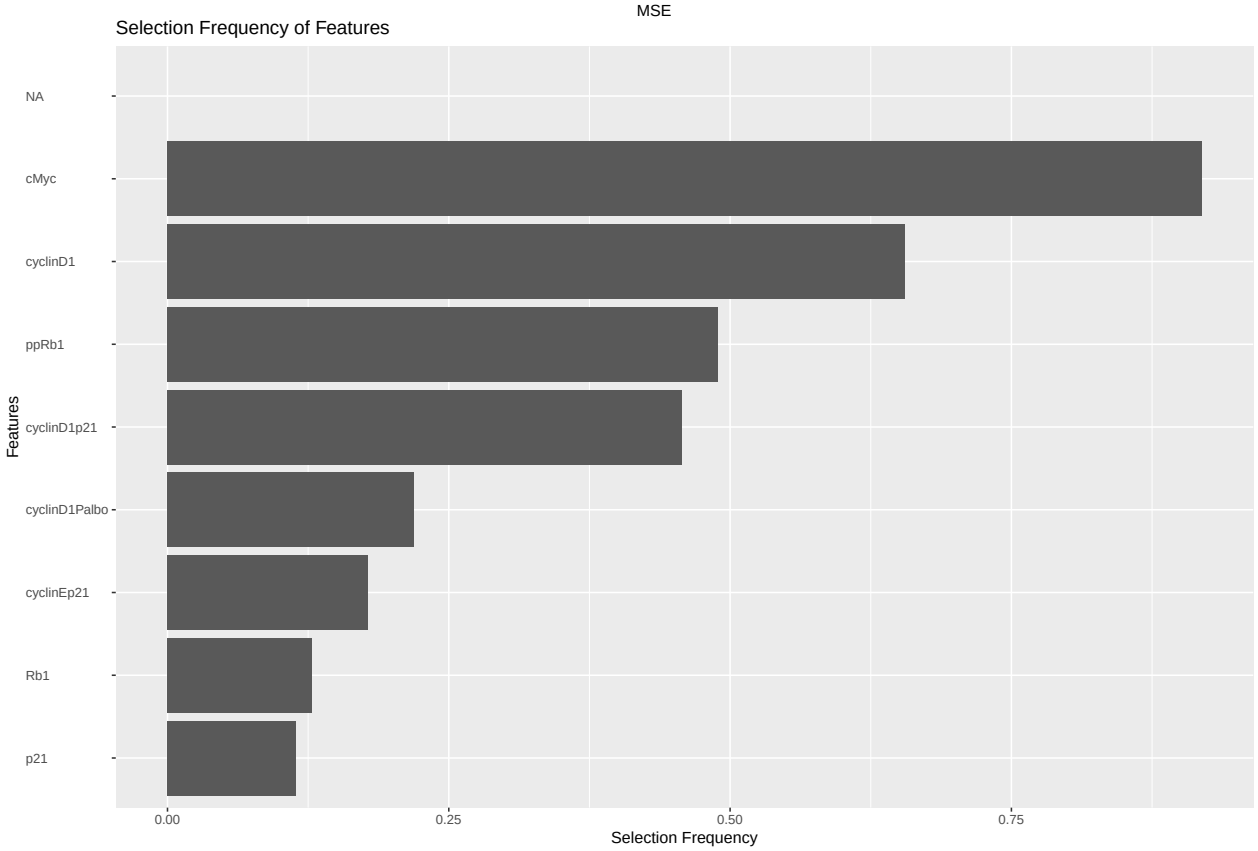
node values -> proliferation score

```
## 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

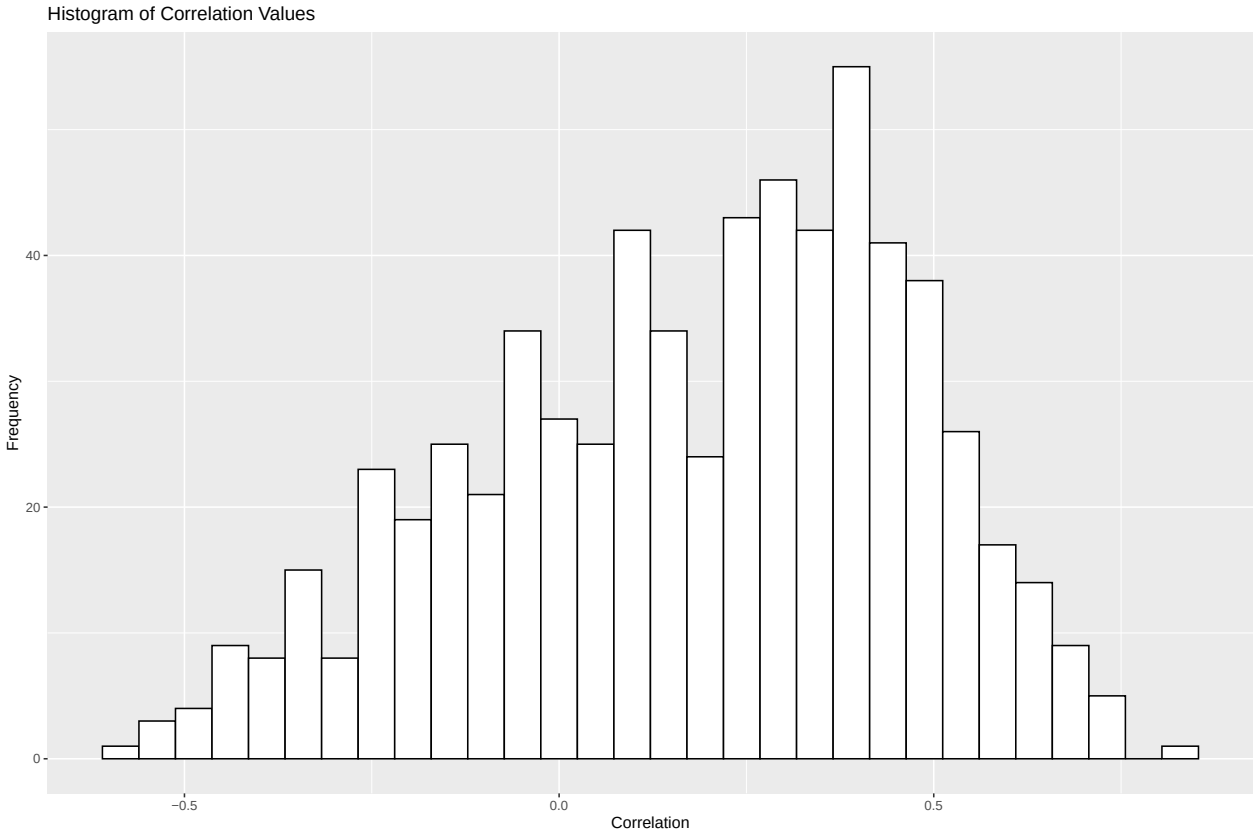

Selection Frequency of Features
Features
NA
cMyc
cyclinD1
ppRb1
cyclinD1p21
cyclinD1Palbo
cyclinEp21
Rb1
p21
0.00
0.25
0.50
0.75
Selection Frequency

##

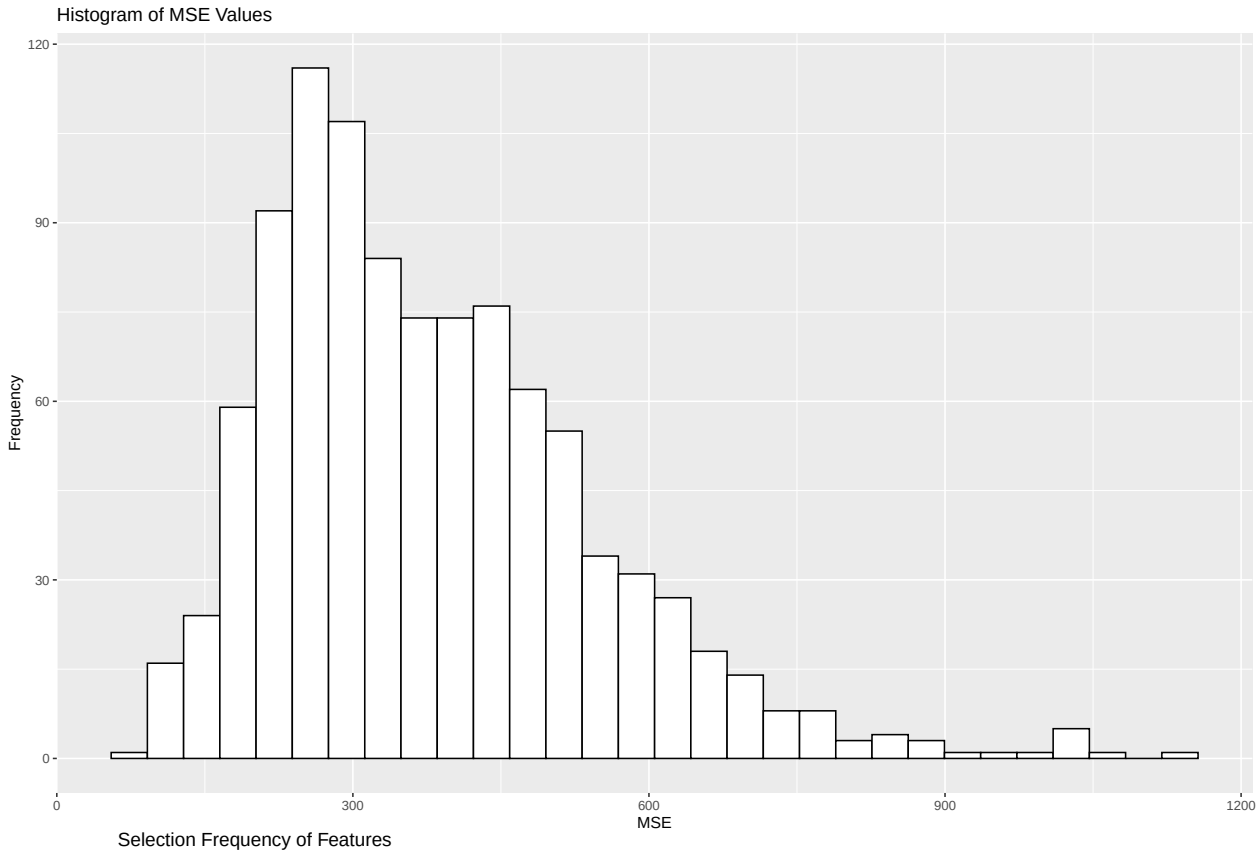
```
## Features selected 50% or more times:
## cyclinD1 cMyc
## Top 20 featrues:
##  [1] "cMyc"          "cyclinD1"      "ppRb1"         "cyclinD1p21"
##  [5] "cyclinD1Palbo" "cyclinEp21"    "Rb1"           "p21"
##  [9] NA              NA              NA              NA
## [13] NA              NA  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

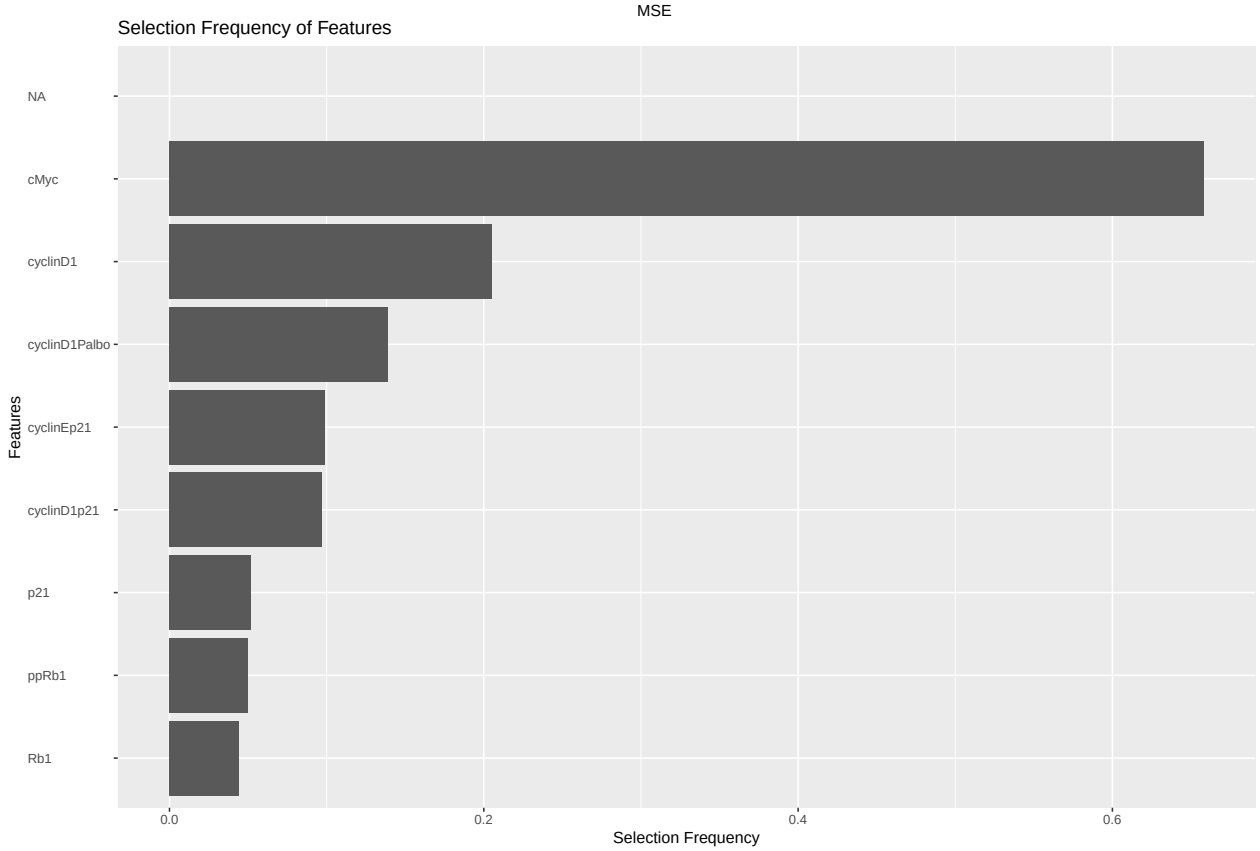
node values -> ROR-proliferation score

```
## 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
```

Ridge

Elastic Net

Boosting with stumps as base learner

Post Lasso