Milestone 2: Core results 01

Anders

25/1/2023

Data

One clinical trails on breast cancer (advanced HR+/HER2-) using two different drug combination

The 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). In the domain knowledge part at the end 5 signature gene sets are use. Make sense to add more...

Additionally, the 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 in study -ish

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 combined score based on expression level of genes and some clinical findings. Range: -8.035678 to 75.13174

(only used in combination with proliferation score as described bellow)

Risk of relapse score with proliferaton score (ROR_Prolif)

A combined score of the two above. Range: 1 to 97 (1-100)

The two scores involving ROR also have categorical variants containing: low, medium, high (but not used...)

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

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. Correlation, MSE and 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

Features tested:

6 genes

771 genes

node values of mech model

residuals of mech model

Responses tested:

proliferation score

combined score including proliferation and ROR

Models tested:

Lasso

Ridge

Elastic Net

Boosting with stumps as base learner

(- mboost)

- xgboost

PCA on subsets of genes

Stacking using different features in the base models

Sparse group lasso (not done)

Iterative learning (ongoing)

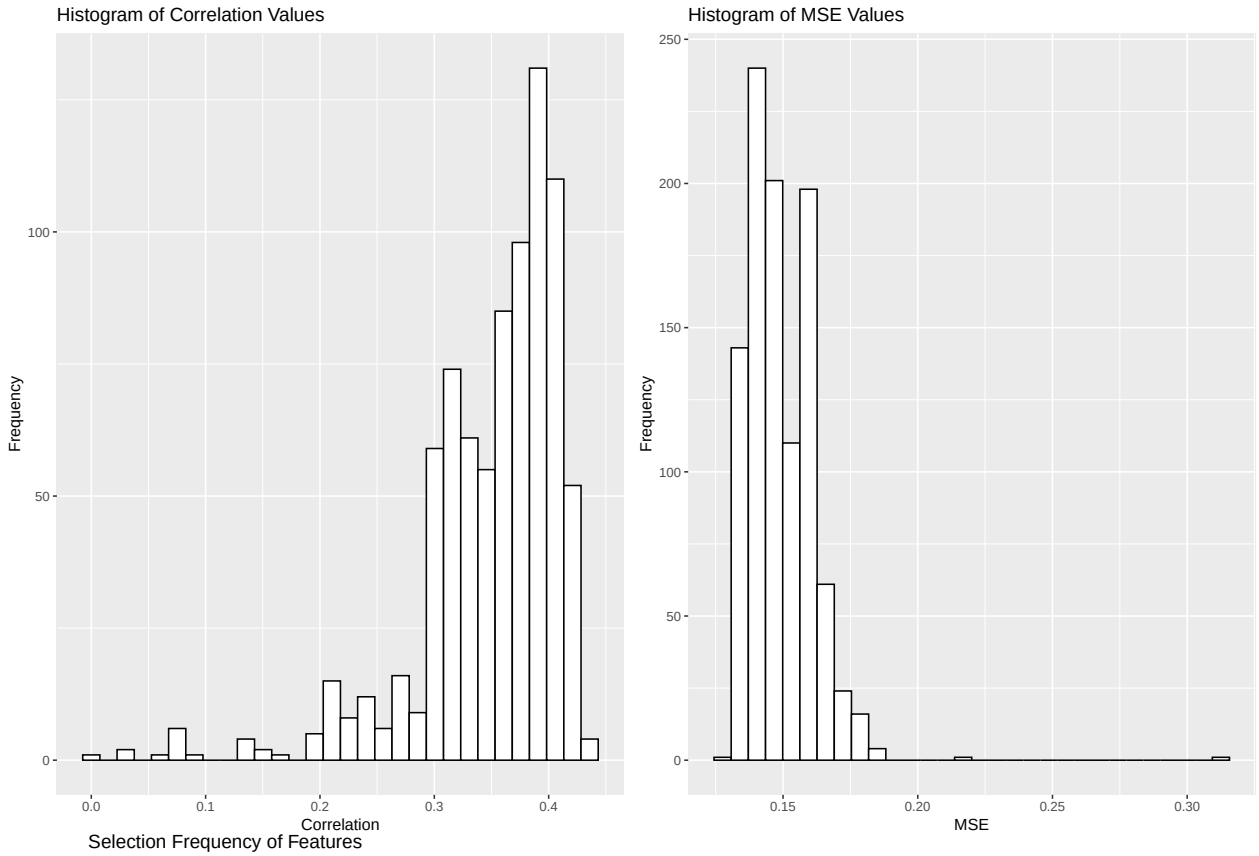
Post Lasso (not done)

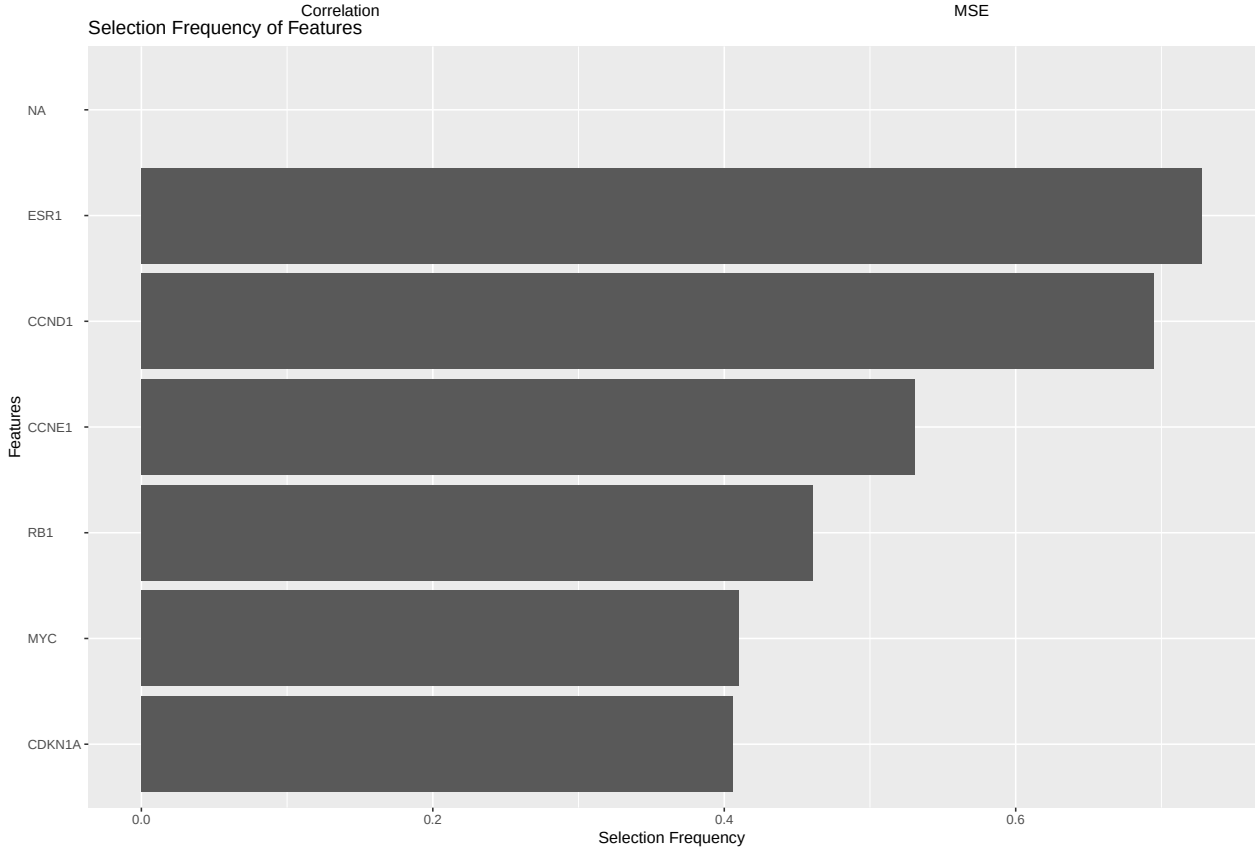
Results of individual modles:

Lasso - Bootstrap

6 genes -> proliferation score (lasso - bootstrap)

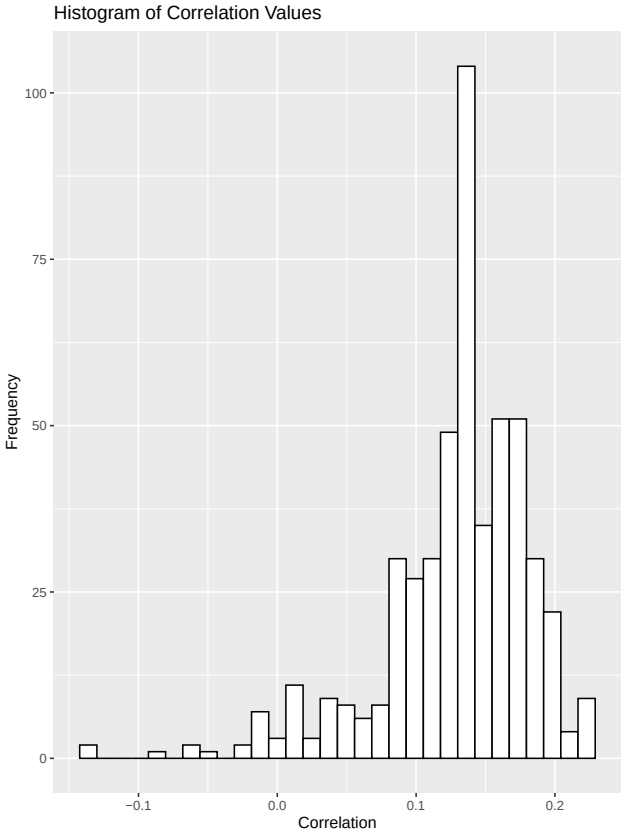
```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0.182
##
## CORRELATIONS RESULTS
## Mean: 0.3498379
## Median: 0.3672243
## st.dev.: 0.063121
##
## MSE RESULTS
## Mean: 0.1492302
## Median: 0.1469228
## st.dev.: 0.01245089
##
## Features selected 50% or more times:
## 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
```

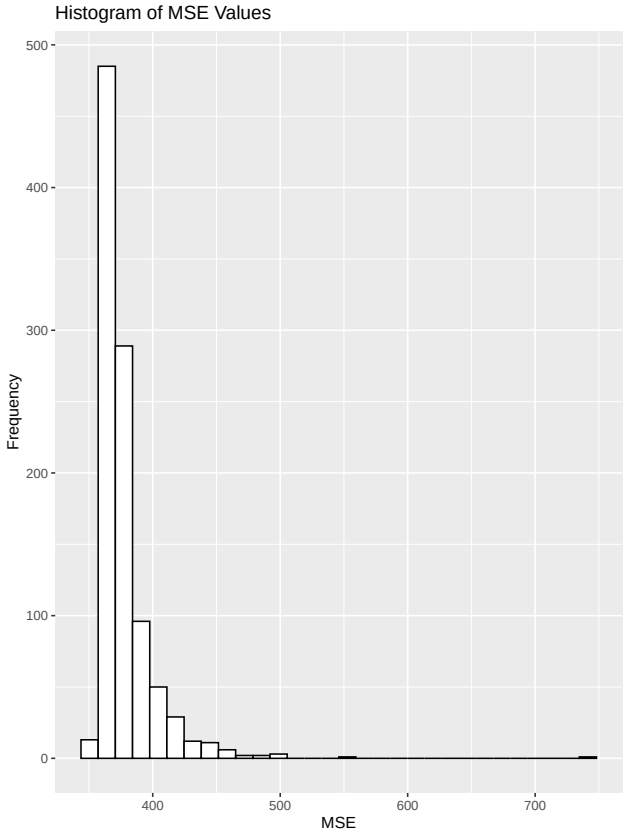

Histogram of Correlation Values
Frequency
Correlation
0
50
100
0.0
0.1
0.2
0.3
0.4
Histogram of MSE Values
Frequency
MSE
0
50
100
150
200
250
0.15
0.20
0.25
0.30

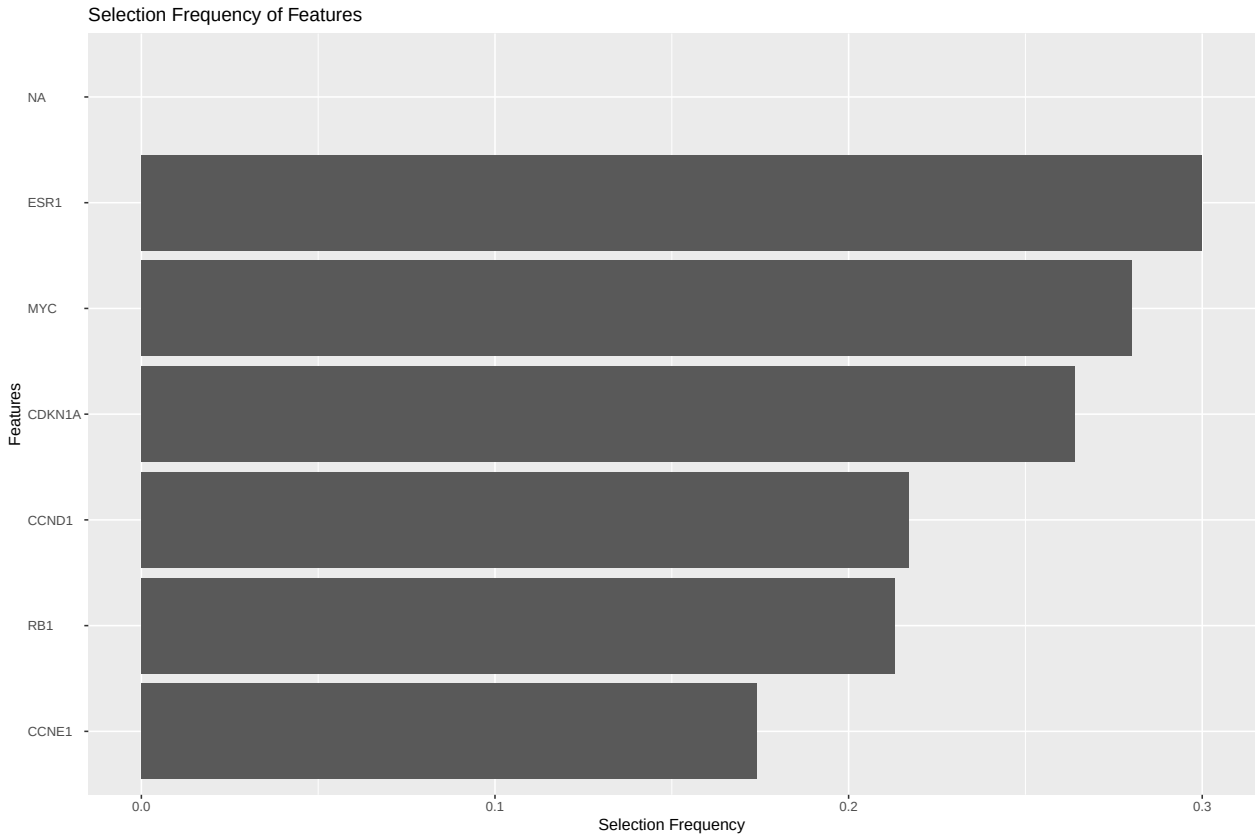

Selection Frequency of Features
Features
Selection Frequency
NA
ESR1
CCND1
CCNE1
RB1
MYC
CDKN1A
0.0
0.2
0.4
0.6

6 genes -> ROR_proliferation score (lasso - bootstrap)

```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0.495
##
## CORRELATIONS RESULTS
## Mean: 0.1282306
## Median: 0.1311715
## st.dev.: 0.05341282
##
## MSE RESULTS
## Mean: 378.094
## Median: 370.7201
## st.dev.: 23.44024
##
## Features selected 50% or more times:
## Non selected that many 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
```

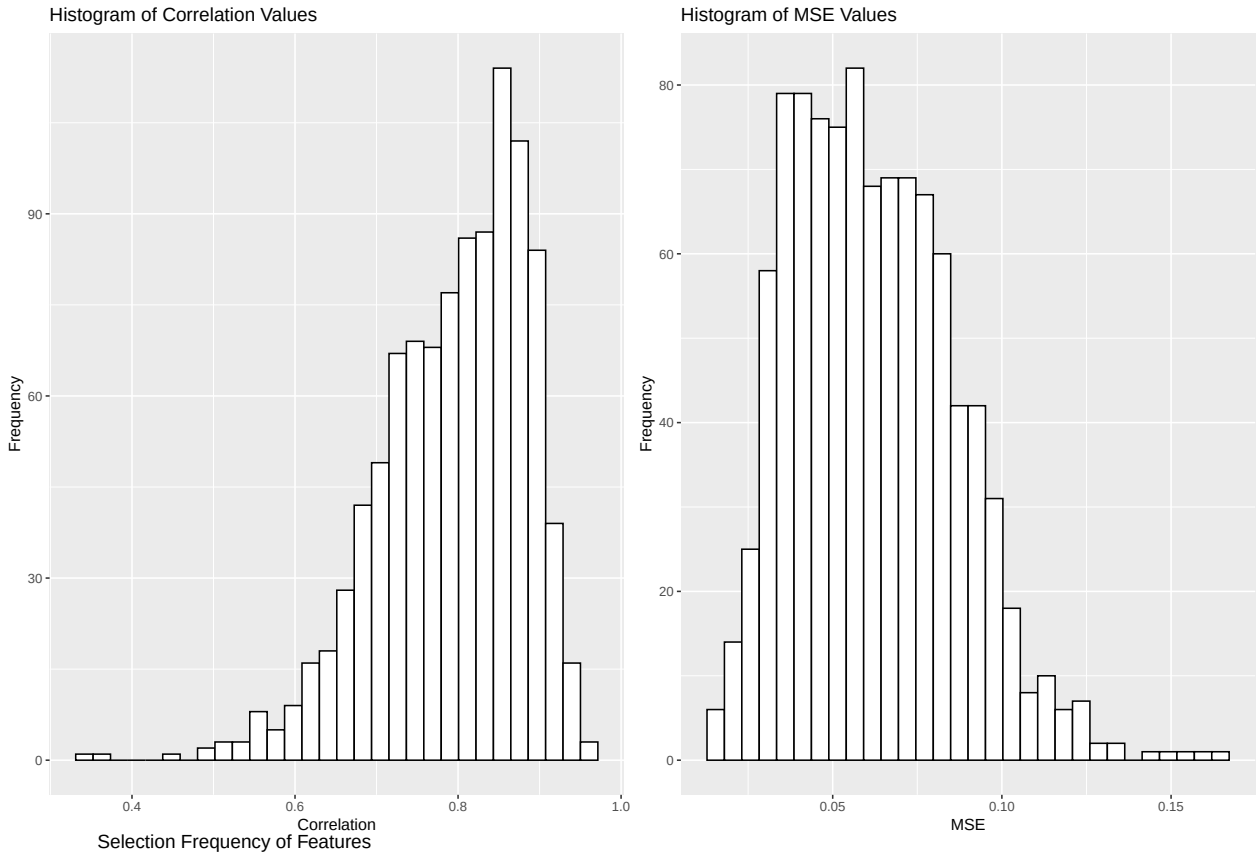




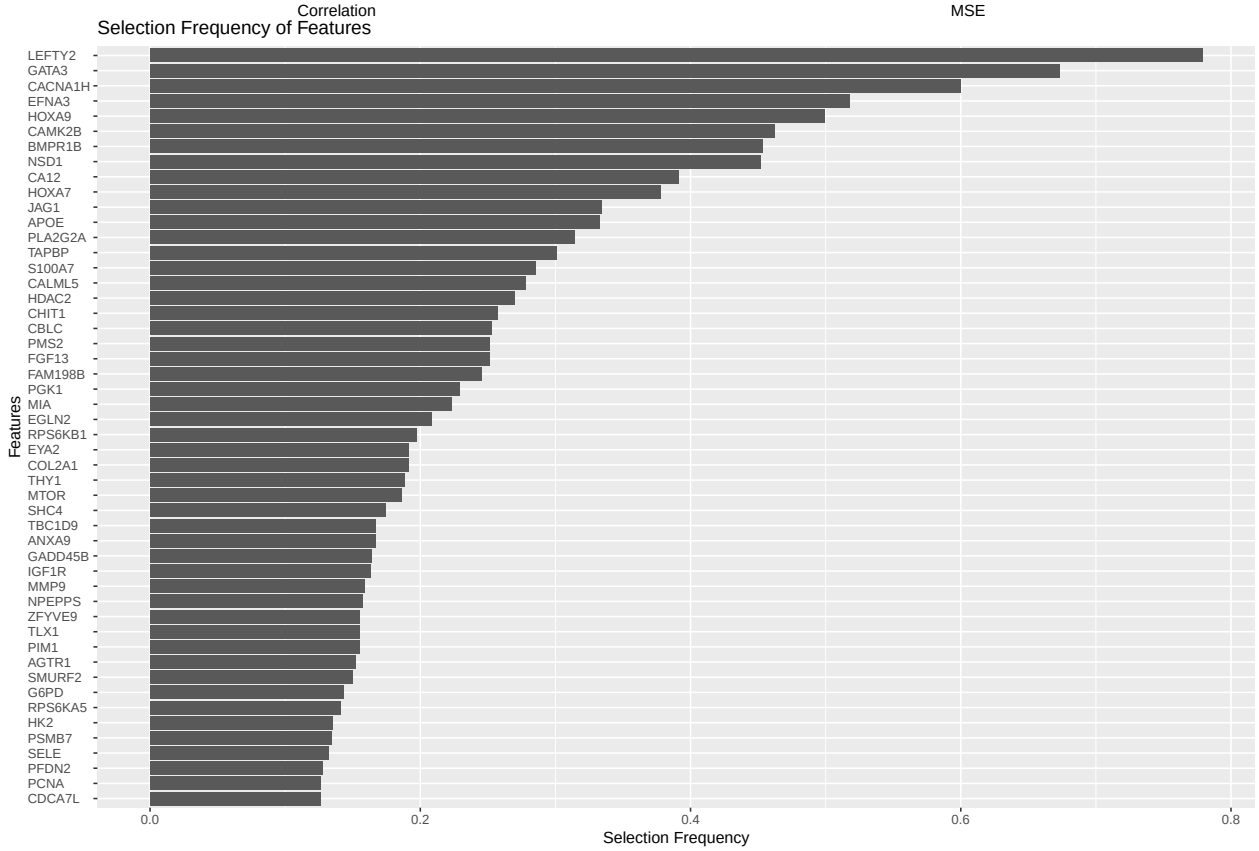


771 genes -> proliferation score (lasso - bootstrap)

```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0.002
##
## CORRELATIONS RESULTS
## Mean: 0.7941413
## Median: 0.8101886
## st.dev.: 0.090107
##
## MSE RESULTS
## Mean: 0.06209131
## Median: 0.0598495
## st.dev.: 0.02398127
##
## Features selected 50% or more times:
## CACNA1H EFNA3 GATA3 LEFTY2
##
## Top 20 featrues:
##  [1] "LEFTY2"  "GATA3"   "CACNA1H" "EFNA3"   "HOXA9"   "CAMK2B"  "BMPR1B"
##  [8] "NSD1"    "CA12"    "HOXA7"   "JAG1"    "APOE"    "PLA2G2A" "TAPBP"
## [15] "S100A7"  "CALML5"  "HDAC2"   "CHIT1"   "CBLC"    "FGF13"
```

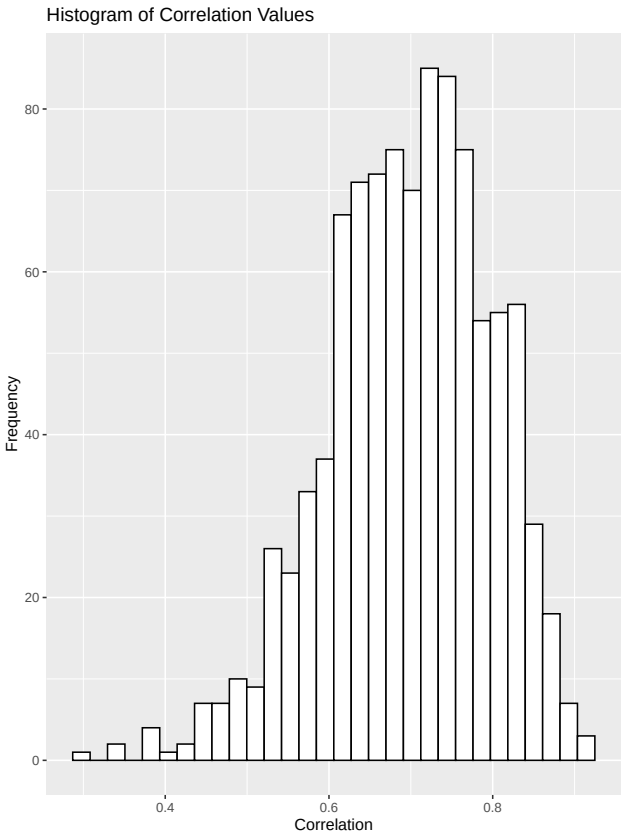

Histogram of Correlation Values
Frequency
Correlation
0
30
60
90
0.4
0.6
0.8
1.0

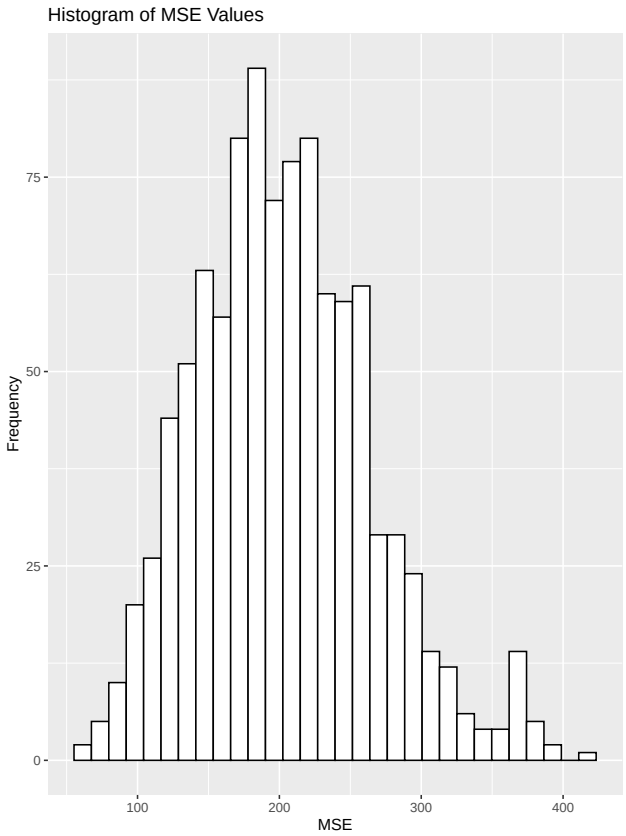
Histogram of MSE Values
Frequency
MSE
0
20
40
60
80
0.05
0.10
0.15

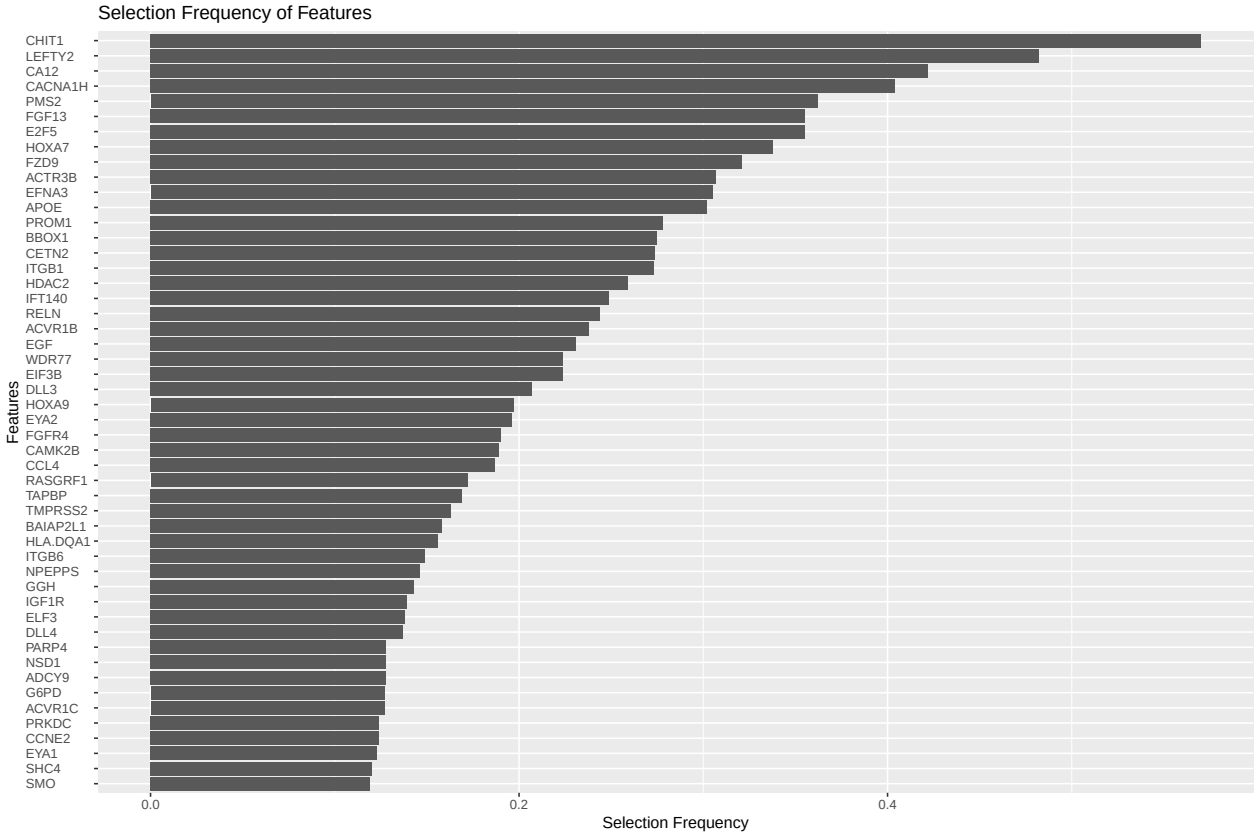

Selection Frequency of Features
Features
Selection Frequency
LEFTY2
GATA3
CACNA1H
EFNA3
HOXA9
CAMK2B
BMPR1B
NSD1
CA12
HOXA7
JAG1
APOE
PLA2G2A
TAPBP
S100A7
CALML5
HDAC2
CHIT1
CBLC
PMS2
FGF13
FAM198B
PGK1
MIA
EGLN2
RPS6KB1
EYA2
COL2A1
THY1
MTOR
SHC4
TBC1D9
ANXA9
GADD45B
IGF1R
MMP9
NPEPPS
ZFYVE9
TLX1
PIM1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

771 genes -> ROR-proliferation score (lasso - bootstrap)

```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0.017
##
## CORRELATIONS RESULTS
## Mean: 0.6968101
## Median: 0.7035889
## st.dev.: 0.09950598
##
## MSE RESULTS
## Mean: 203.408
## Median: 198.455
## st.dev.: 61.34872
##
## Features selected 50% or more times:
## CHIT1
##
## Top 20 featrues:
##  [1] "CHIT1"   "LEFTY2"  "CA12"    "CACNA1H" "PMS2"    "E2F5"    "FGF13"
##  [8] "HOXA7"   "FZD9"    "ACTR3B"  "EFNA3"   "APOE"    "PROM1"   "BBOX1"
## [15] "CETN2"   "ITGB1"   "HDAC2"   "IFT140"  "RELN"    "ACVR1B"
```

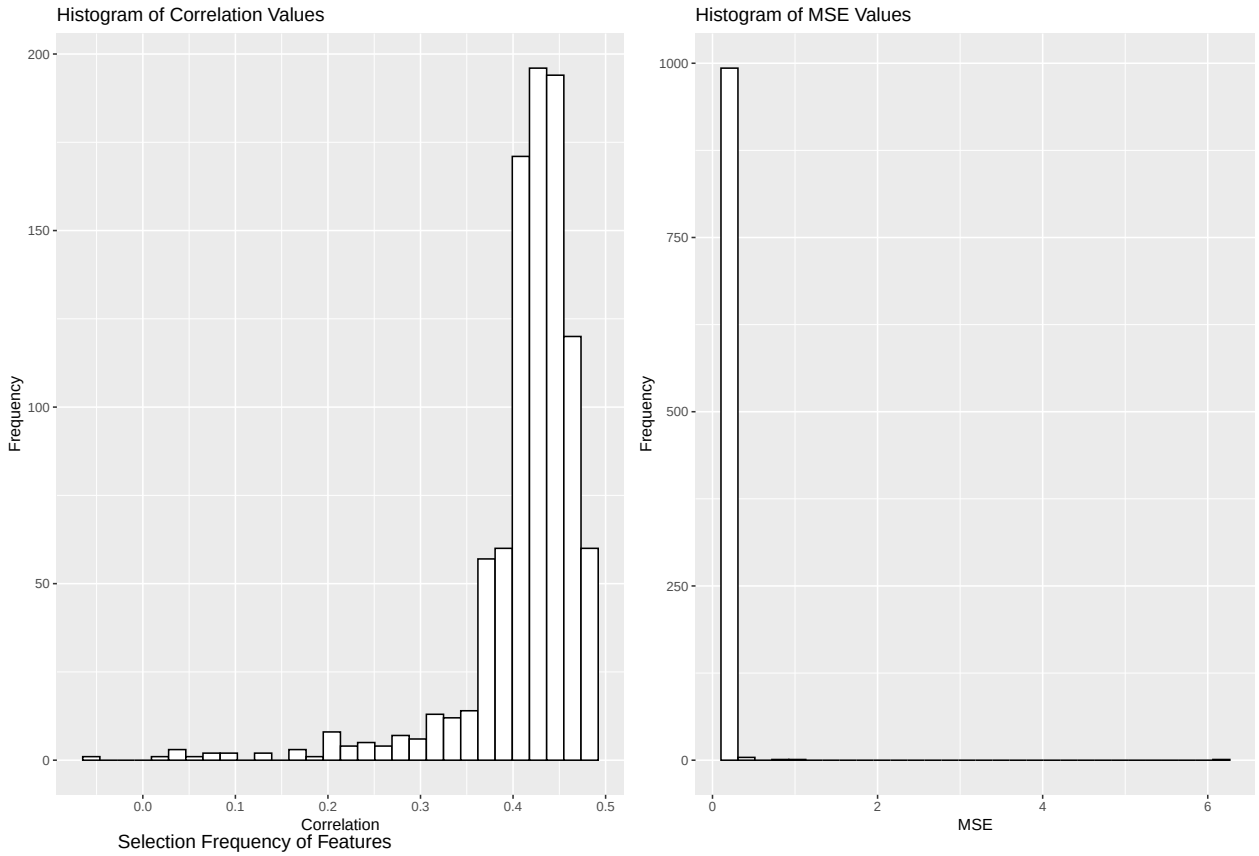




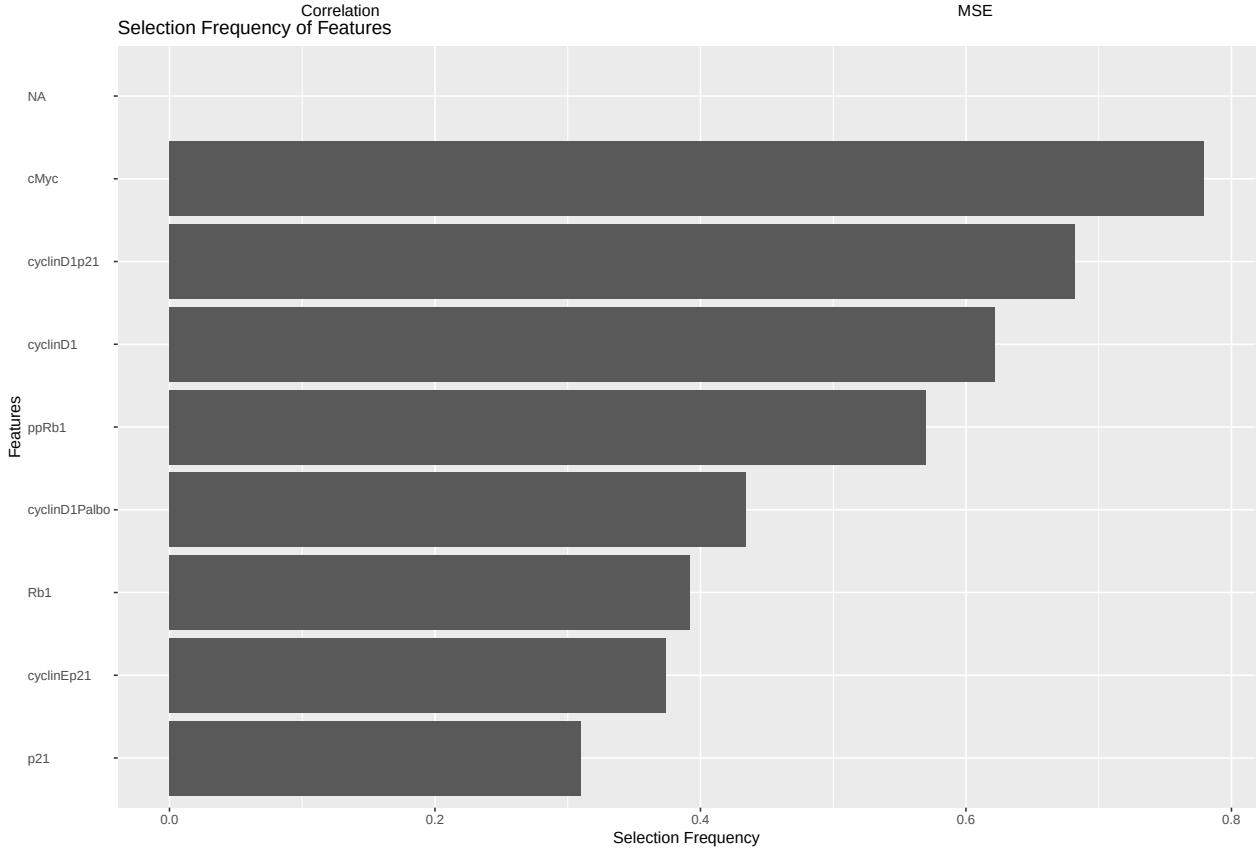


node values -> proliferation score (lasso - bootstrap)

```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0.053
##
## CORRELATIONS RESULTS
## Mean: 0.414496
## Median: 0.4275114
## st.dev.: 0.06427223
##
## MSE RESULTS
## Mean: 0.1479731
## Median: 0.1355805
## st.dev.: 0.191663
##
## Features selected 50% or more times:
## cyclinD1 cyclinD1p21 cMyc ppRb1
##
## Top 20 featrues:
##  [1] "cMyc"          "cyclinD1p21"   "cyclinD1"      "ppRb1"
##  [5] "cyclinD1Palbo" "Rb1"           "cyclinEp21"    "p21"
##  [9] NA              NA              NA              NA
## [13] NA              NA              NA              NA
## [17] NA              NA              NA              NA
```

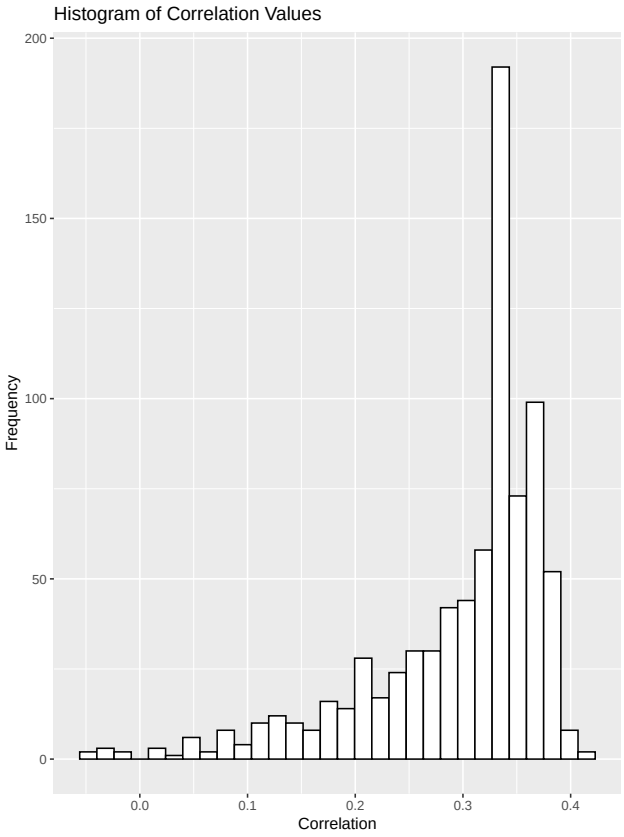

Histogram of Correlation Values
Frequency
Correlation
0
50
100
150
200
0.0
0.1
0.2
0.3
0.4
0.5

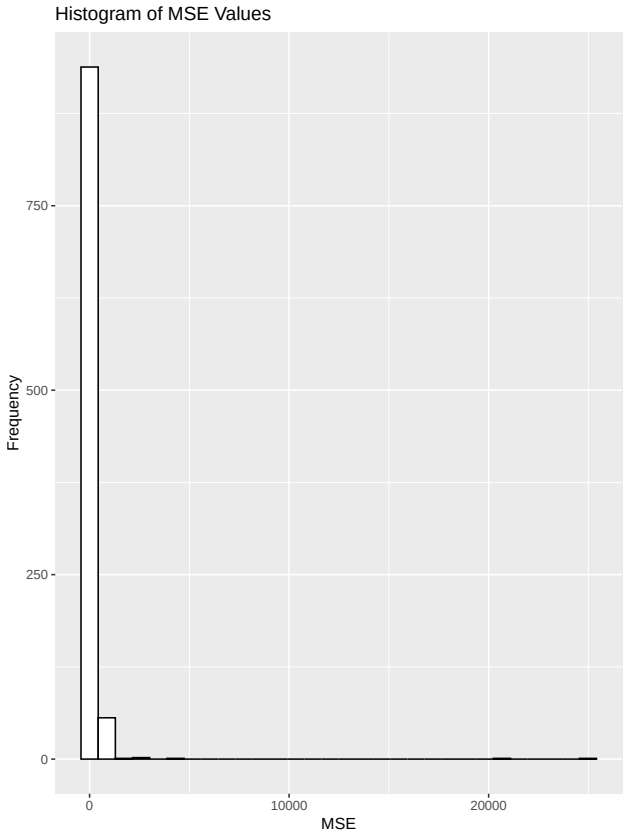
Histogram of MSE Values
Frequency
MSE
0
250
500
750
1000
0
2
4
6

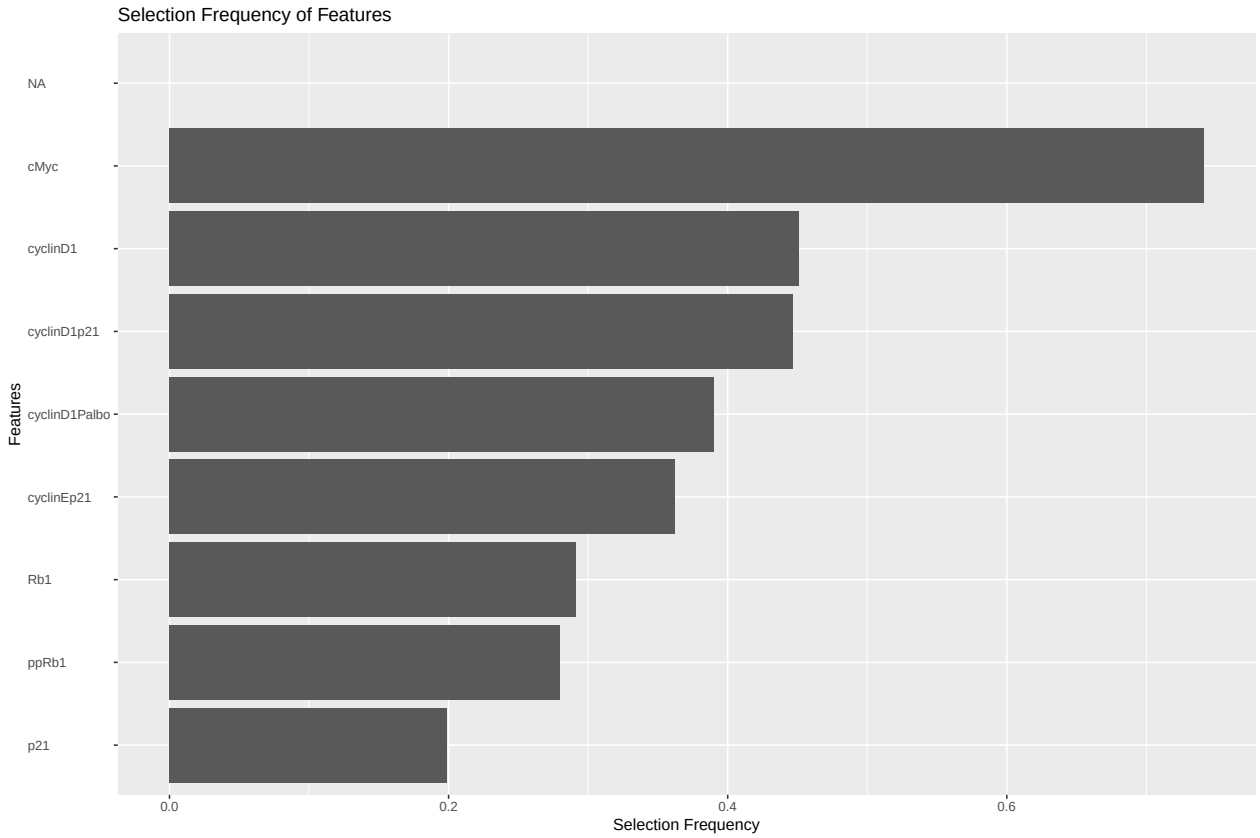

Selection Frequency of Features
Features
Selection Frequency
NA
cMyc
cyclinD1p21
cyclinD1
ppRb1
cyclinD1Palbo
Rb1
cyclinEp21
p21
0.0
0.2
0.4
0.6
0.8

node values -> ROR-proliferation score (lasso - bootstrap)

```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0.2
##
## CORRELATIONS RESULTS
## Mean: 0.2964169
## Median: 0.3317433
## st.dev.: 0.08306956
##
## MSE RESULTS
## Mean: 417.6667
## Median: 353.8176
## st.dev.: 1027.062
##
## Features selected 50% or more times:
## cMyc
##
## Top 20 featrues:
##  [1] "cMyc"          "cyclinD1"      "cyclinD1p21"   "cyclinD1Palbo"
##  [5] "cyclinEp21"    "Rb1"           "ppRb1"         "p21"
##  [9] NA              NA              NA              NA
## [13] NA              NA              NA              NA
## [17] NA              NA              NA              NA
```

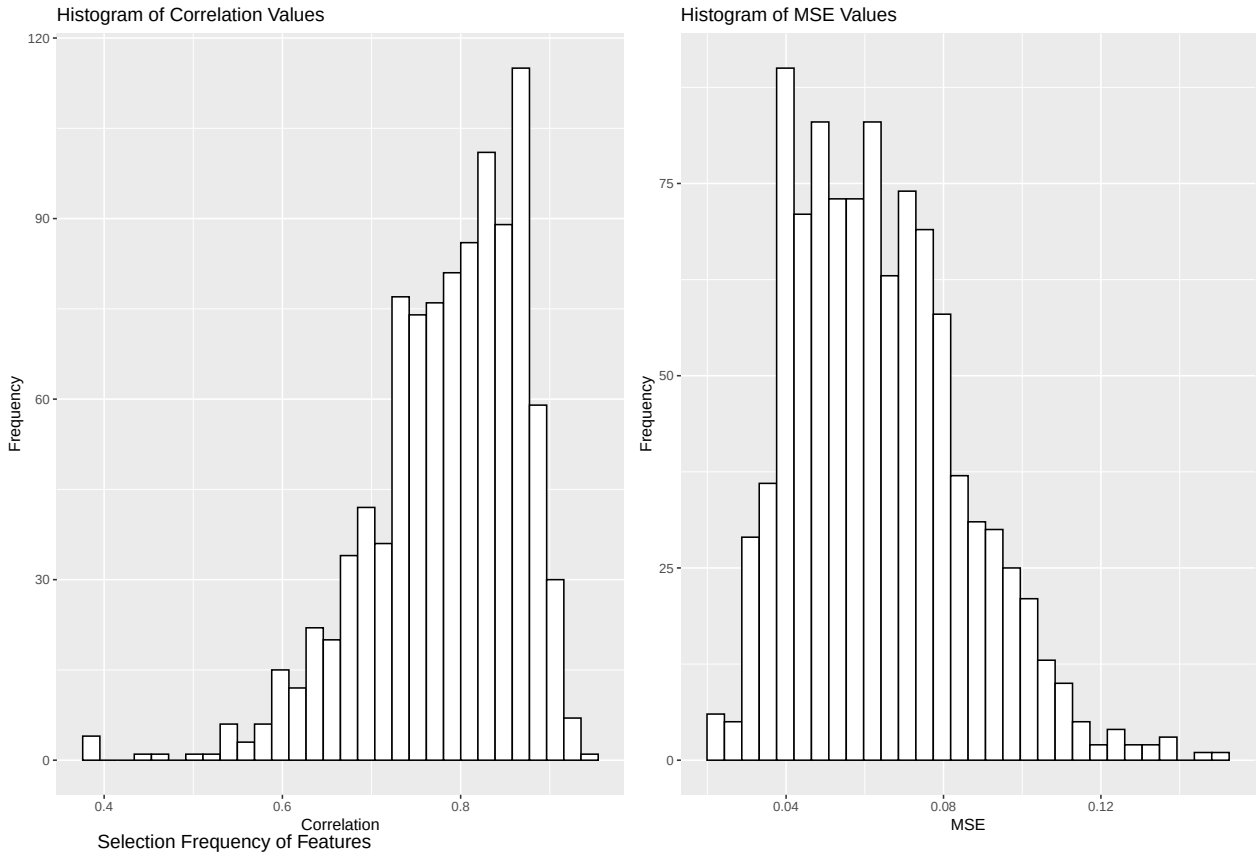


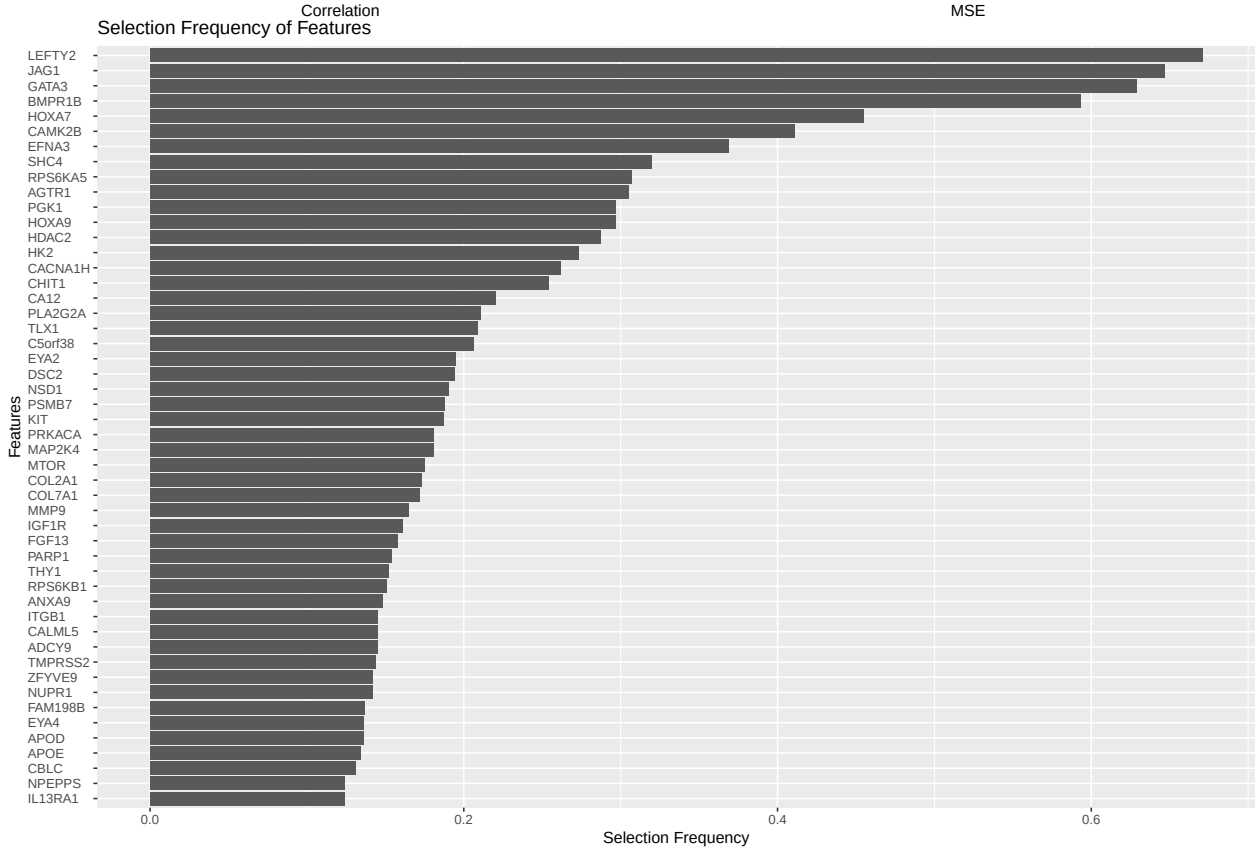




Mechanistic + Residuals (additive) -> proliferation score (lasso - bootstrap)

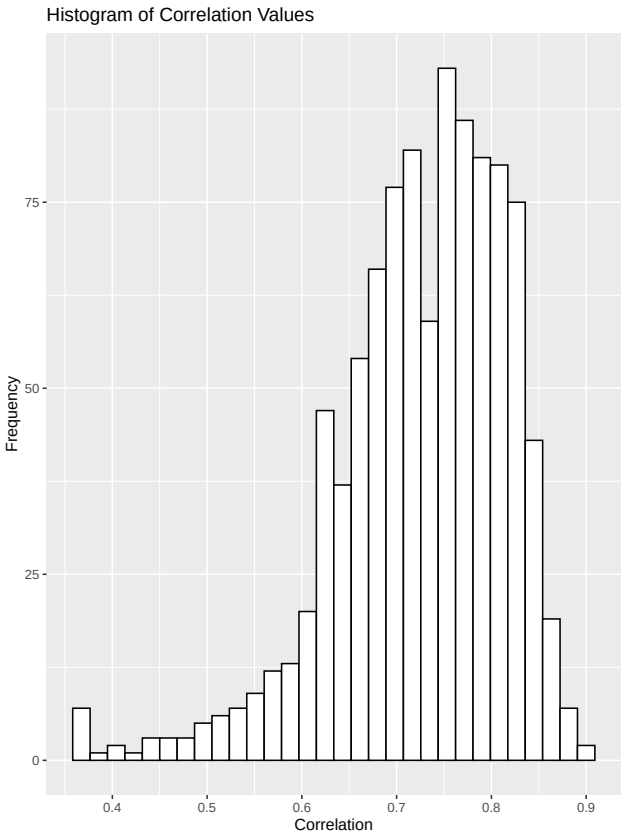
```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0
##
## CORRELATIONS RESULTS
## Mean: 0.7835808
## Median: 0.7962129
## st.dev.: 0.08543355
##
## MSE RESULTS
## Mean: 0.06384841
## Median: 0.06161072
## st.dev.: 0.0213956
##
## Features selected 50% or more times:
## BMPR1B GATA3 JAG1 LEFTY2
##
## Top 20 featrues:
##  [1] "LEFTY2"  "JAG1"    "GATA3"   "BMPR1B"  "HOXA7"   "CAMK2B"  "EFNA3"
##  [8] "SHC4"    "RPS6KA5" "AGTR1"   "HOXA9"   "PGK1"    "HDAC2"   "HK2"
## [15] "CACNA1H" "CHIT1"   "CA12"    "PLA2G2A" "TLX1"    "C5orf38"
```

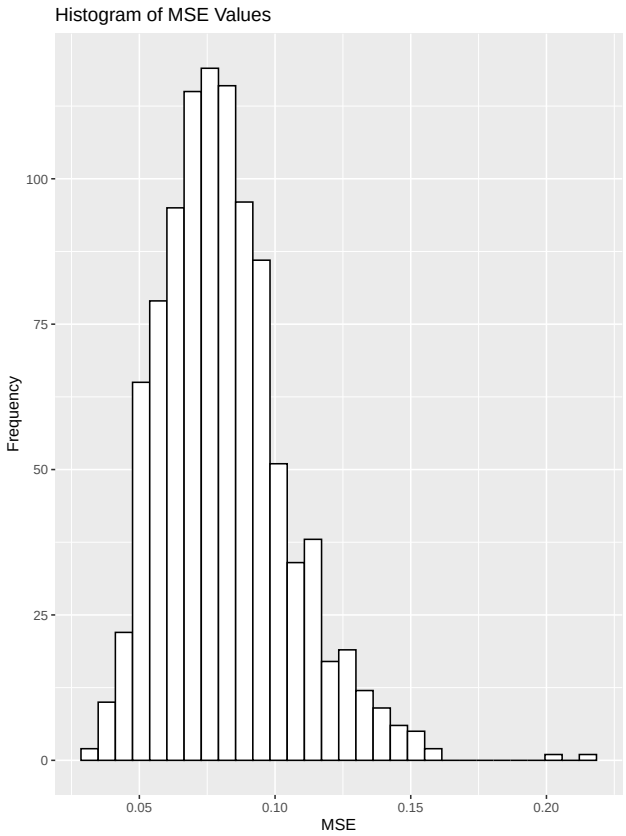

Histogram of Correlation Values
Frequency
Correlation
Histogram of MSE Values
Frequency
MSE

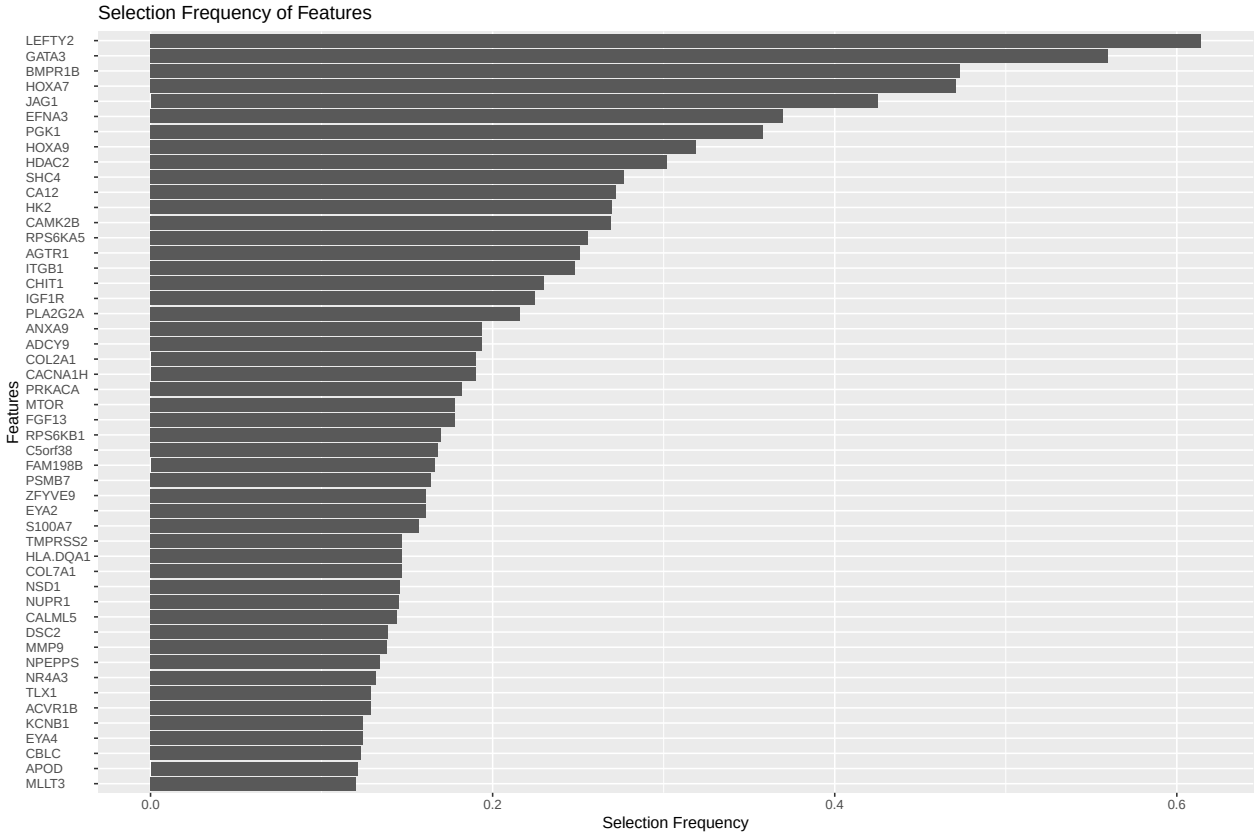

Selection Frequency of Features
Features
LEFTY2
JAG1
GATA3
BMPR1B
HOXA7
CAMK2B
EFNA3
SHC4
RPS6KA5
AGTR1
PGK1
HOXA9
HDAC2
HK2
CACNA1H
CHIT1
CA12
PLA2G2A
TLX1
C5orf38
EYA2
DSC2
NSD1
PSMB7
KIT
PRKACA
MAP2K4
MTOR
COL2A1
COL7A1
MMP9
IGF1R
FGF13
PARP1
THY1
RPS6KB1
ANXA9
ITGB1
CALML5
ADCY9
TMPRSS2
ZFYVE9
NUPR1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

Mechanistic + Residuals (multiplicative) -> proliferation score (lasso - bootstrap)

```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0
##
## CORRELATIONS RESULTS
## Mean: 0.7266736
## Median: 0.739984
## st.dev.: 0.08952363
##
## MSE RESULTS
## Mean: 0.0813415
## Median: 0.07892236
## st.dev.: 0.0231428
##
## 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"
```

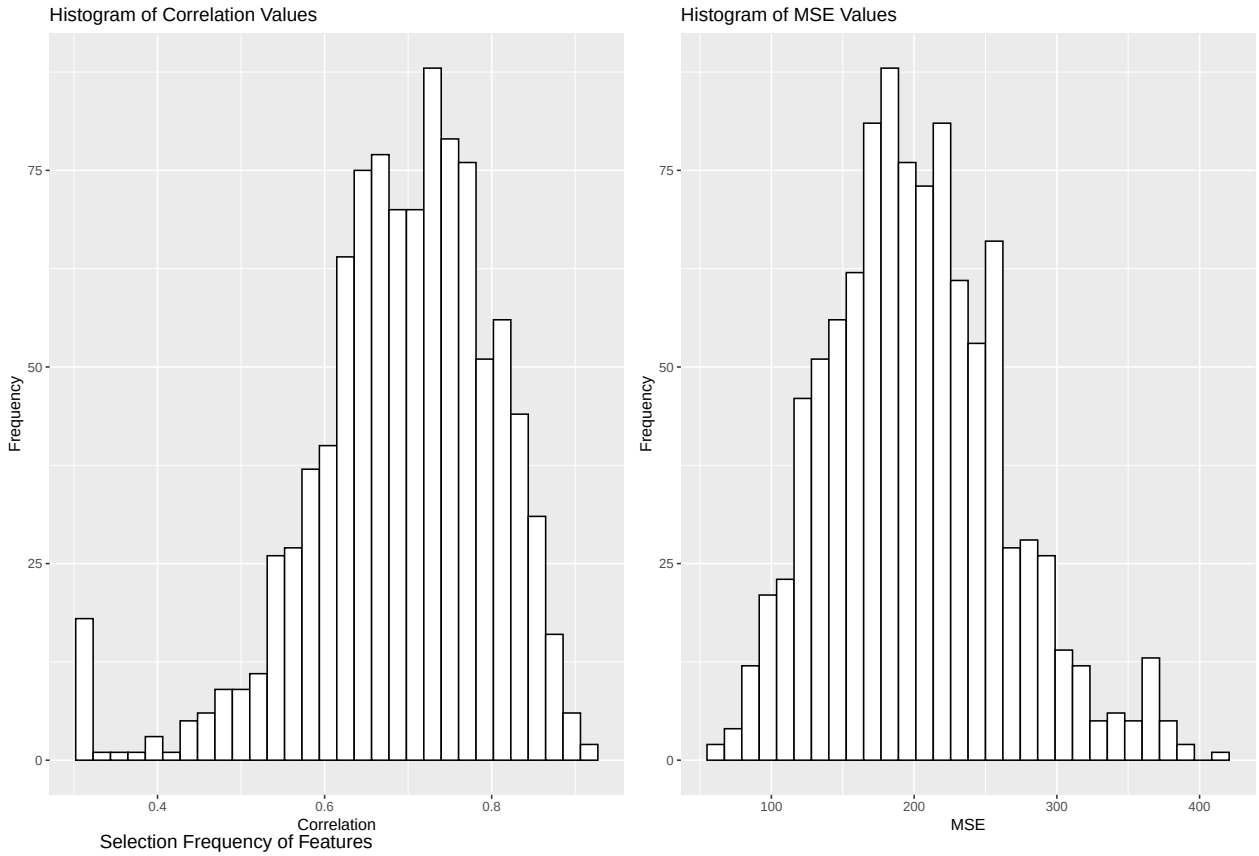




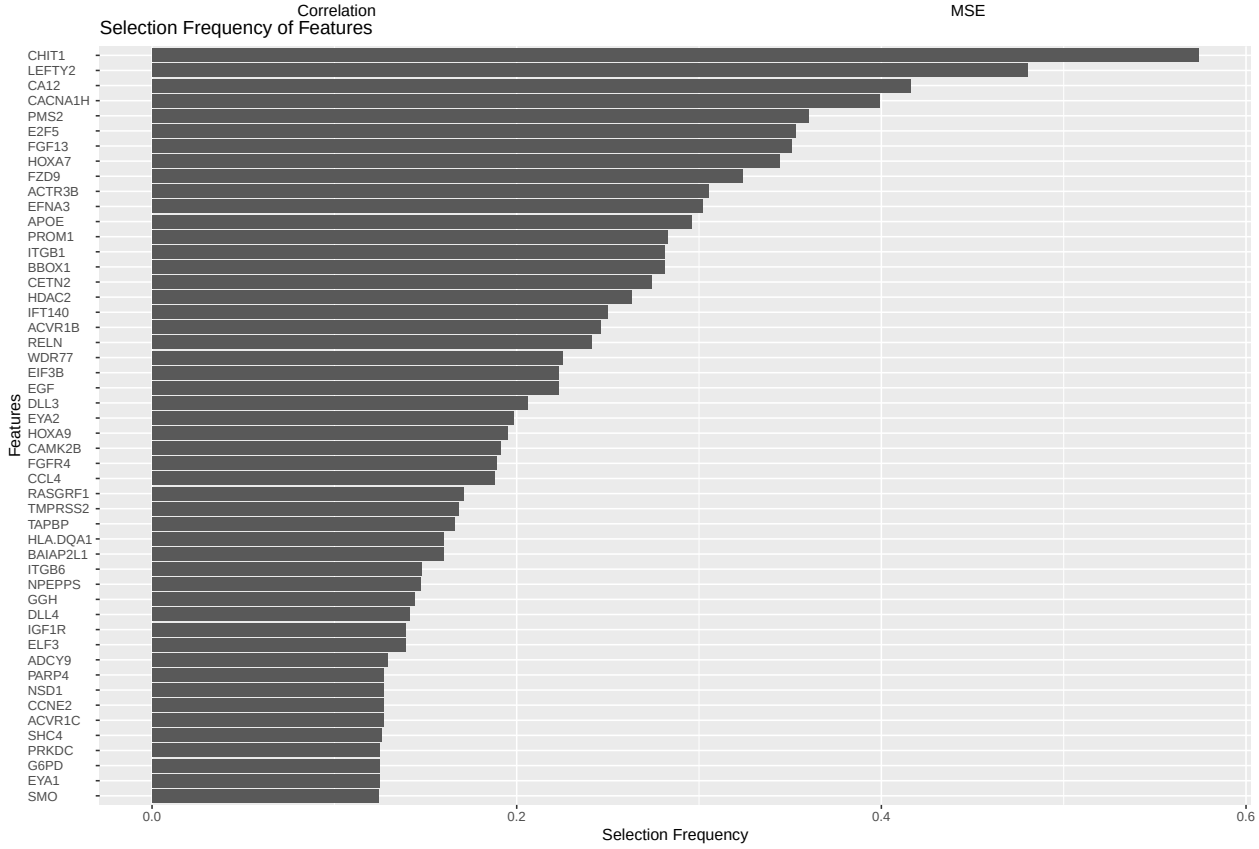


Mechnaistic + Residuals (additive) -> ROR-proliferation score (lasso - bootstrap)

```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0
##
## CORRELATIONS RESULTS
## Mean: 0.6925953
## Median: 0.7037142
## st.dev.: 0.1092315
##
## MSE RESULTS
## Mean: 202.3235
## Median: 197.4845
## st.dev.: 61.04236
##
## Features selected 50% or more times:
## CHIT1
##
## Top 20 featrues:
##  [1] "CHIT1"   "LEFTY2"  "CA12"    "CACNA1H" "PMS2"    "E2F5"    "FGF13"
##  [8] "HOXA7"   "FZD9"    "ACTR3B"  "EFNA3"   "APOE"    "PROM1"   "BBOX1"
## [15] "ITGB1"   "CETN2"   "HDAC2"   "IFT140"  "ACVR1B"  "RELN"
```

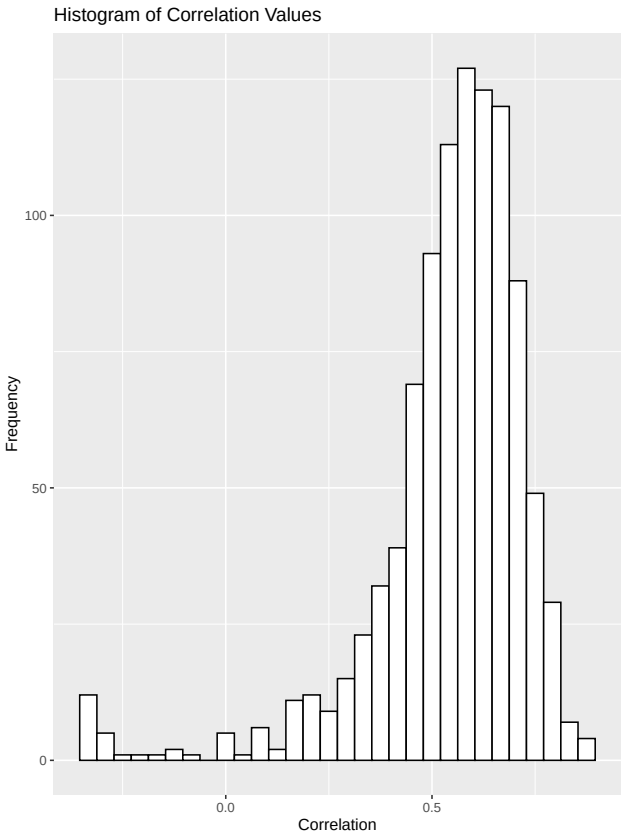

Histogram of Correlation Values
Frequency
Correlation
0
25
50
75
0.4
0.6
0.8

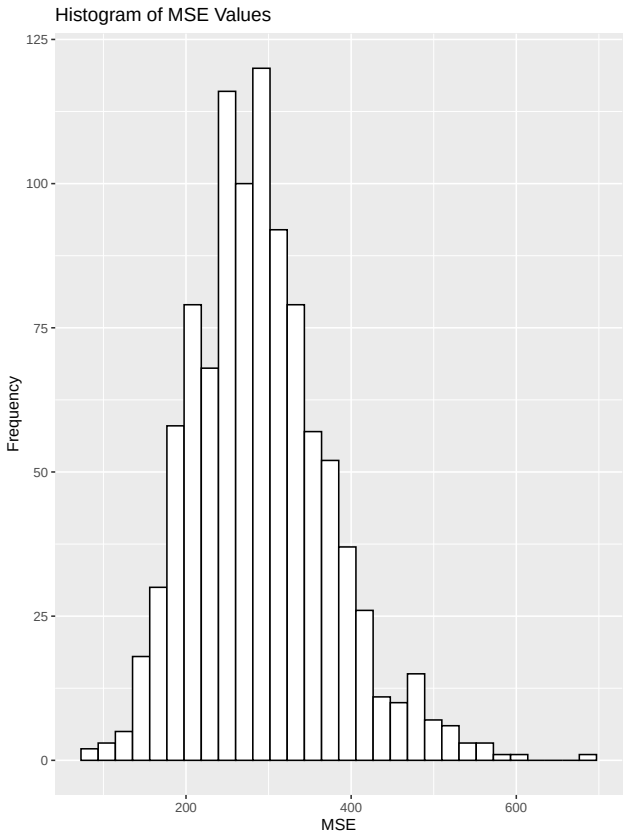
Histogram of MSE Values
Frequency
MSE
0
25
50
75
100
200
300
400

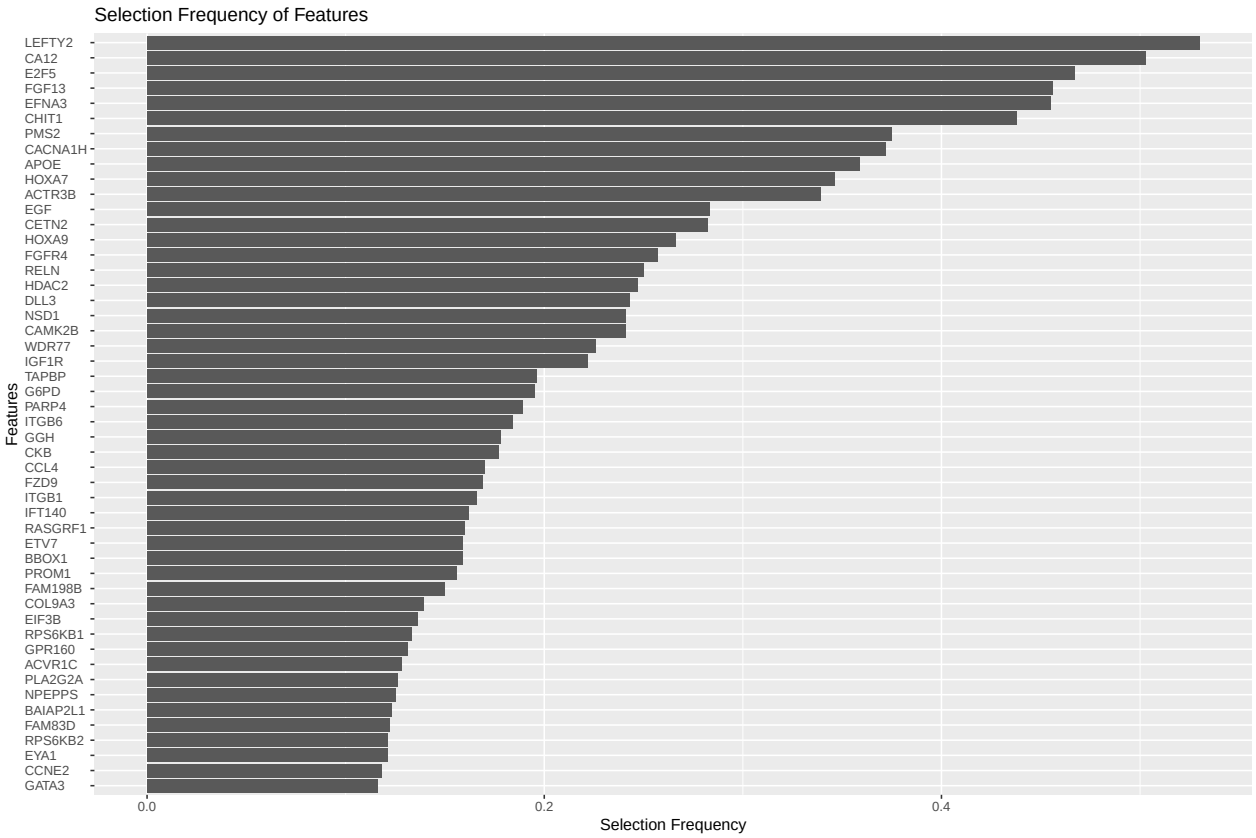

Selection Frequency of Features
Features
Selection Frequency
CHIT1
LEFTY2
CA12
CACNA1H
PMS2
E2F5
FGF13
HOXA7
FZD9
ACTR3B
EFNA3
APOE
PROM1
ITGB1
BBOX1
CETN2
HDAC2
IFT140
ACVR1B
RELN
WDR77
EIF3B
EGF
DLL3
EYA2
HOXA9
CAMK2B
FGFR4
CCL4
RASGRF1
TMPRSS2
TAPBP
HLA.DQA1
BAIAP2L1
ITGB6
NPEPPS
GGH
DLL4
IGF1R
ELF3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

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

```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0
##
## CORRELATIONS RESULTS
## Mean: 0.5427757
## Median: 0.5790305
## st.dev.: 0.1908061
##
## MSE RESULTS
## Mean: 291.0186
## Median: 284.2399
## st.dev.: 82.79231
##
## 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)

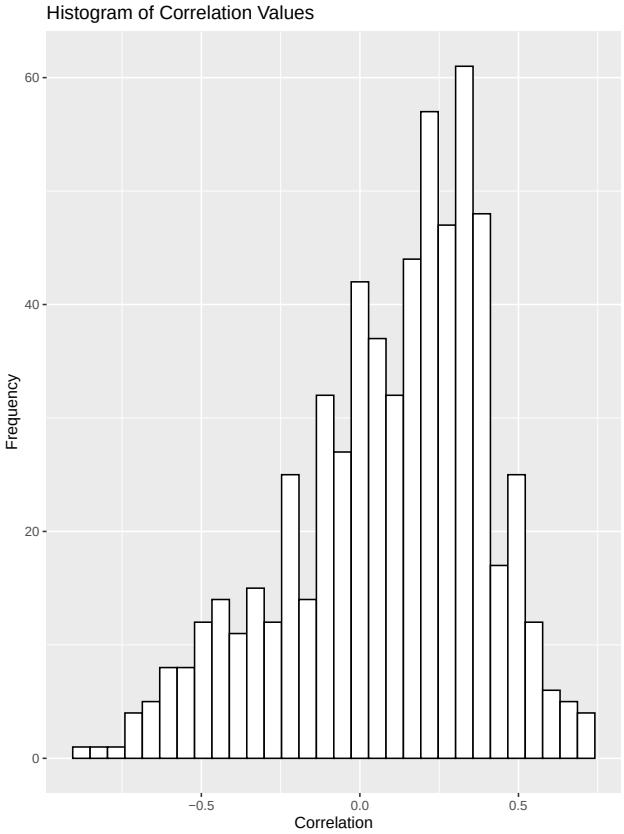
Model	cor_mean	sd_cor	MSE_mean	MSE_sd
lasso 6 genes	0.1282306	0.0534128	378.0940	23.44024
lasso 771 genes	0.6968101	0.0995060	203.4080	61.34872
Nodes	0.2964169	0.0830696	417.6667	1027.06231
Residual additive	0.6925953	0.1092315	202.3235	61.04236
Residual multiplicative	0.5427757	0.1908061	291.0186	82.79231

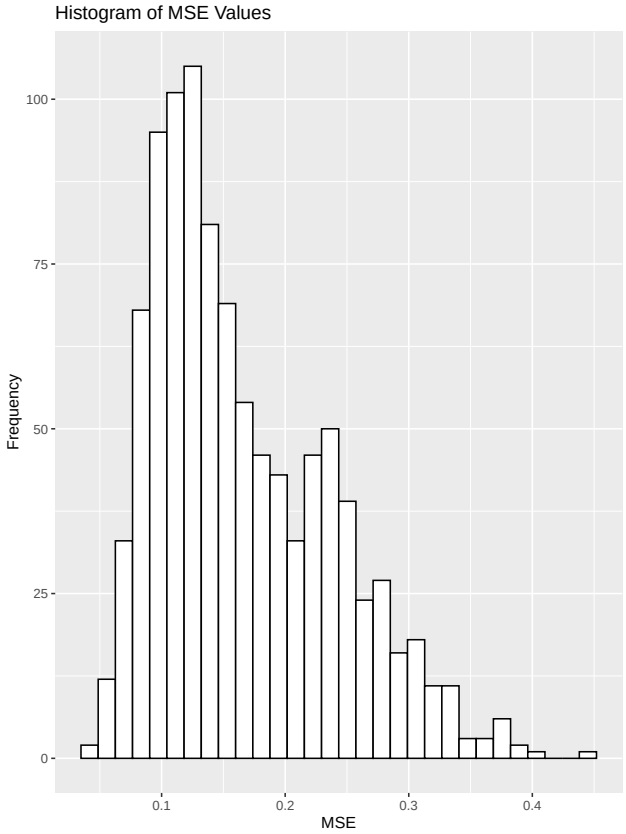
Lasso - Repeated cross-validation

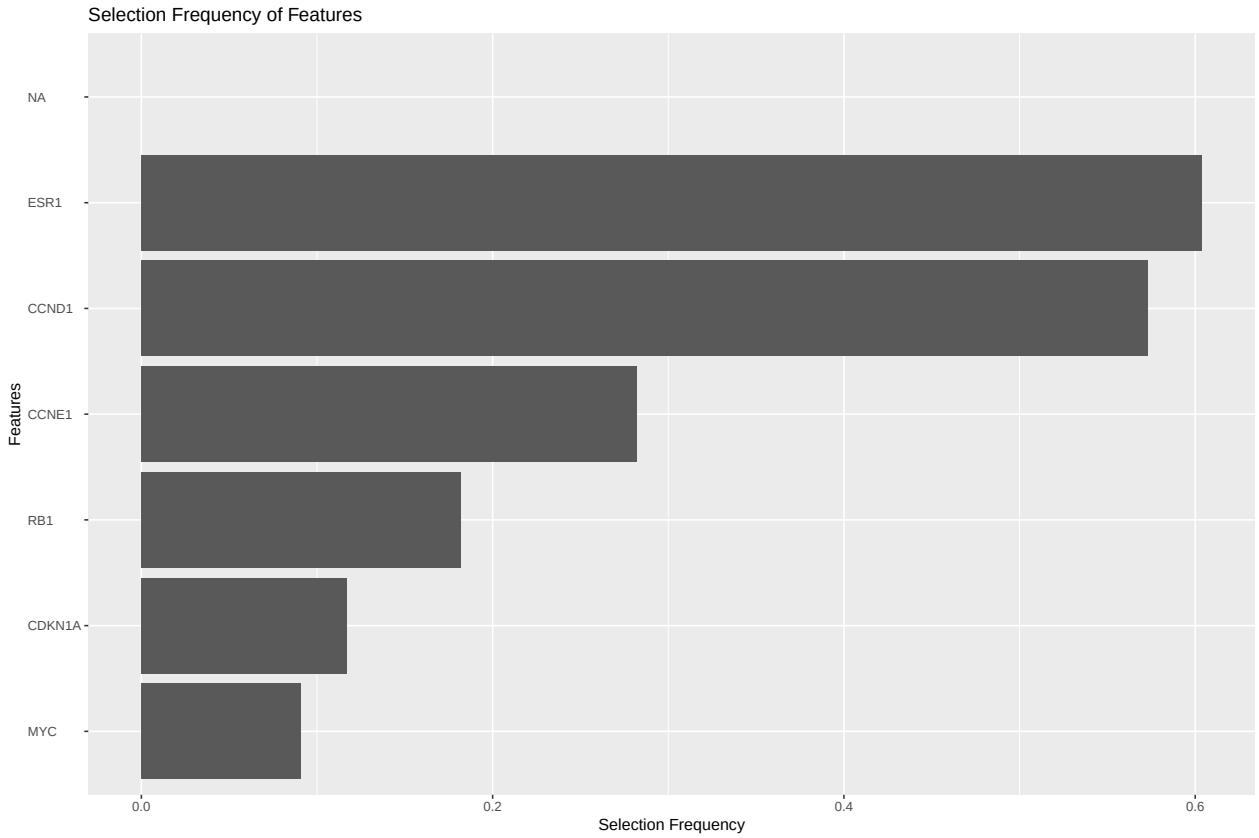
200 repeats of five fold cross-validation

6 genes -> proliferation score (lasso - repeated cross-validation)

```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0.373
##
## CORRELATIONS RESULTS
## Mean: 0.09196628
## Median: 0.1502755
## st.dev.: 0.3068209
##
## MSE RESULTS
## Mean: 0.1655931
## Median: 0.1468293
## st.dev.: 0.0718379
##
## 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
```

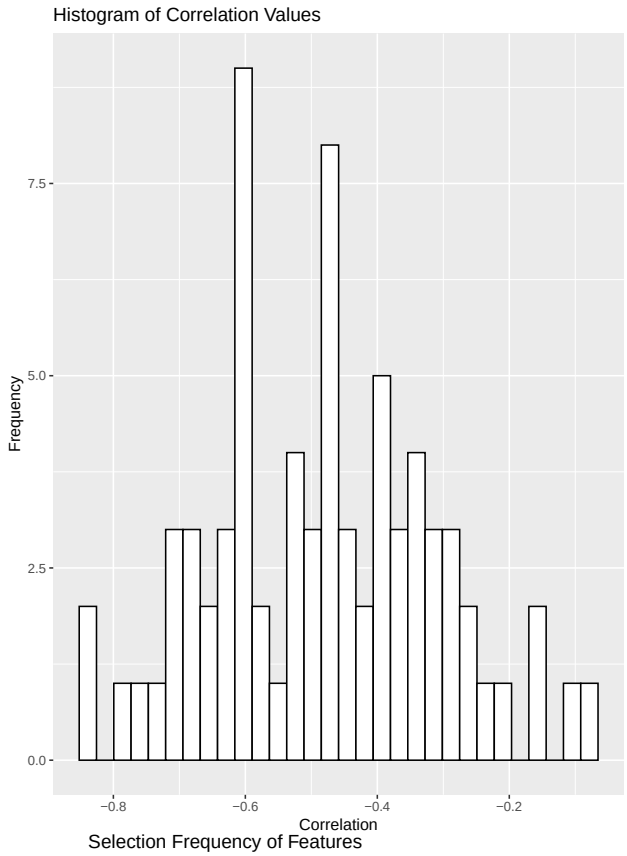


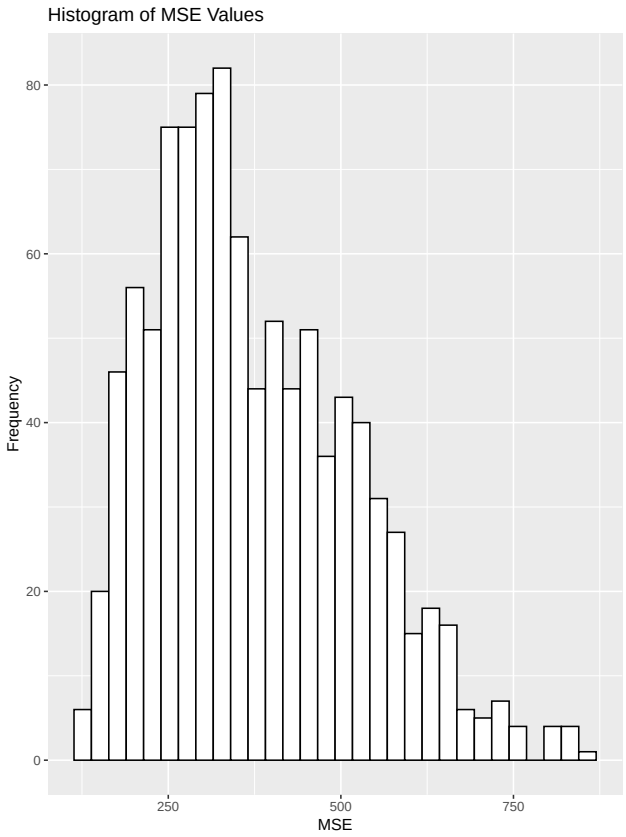


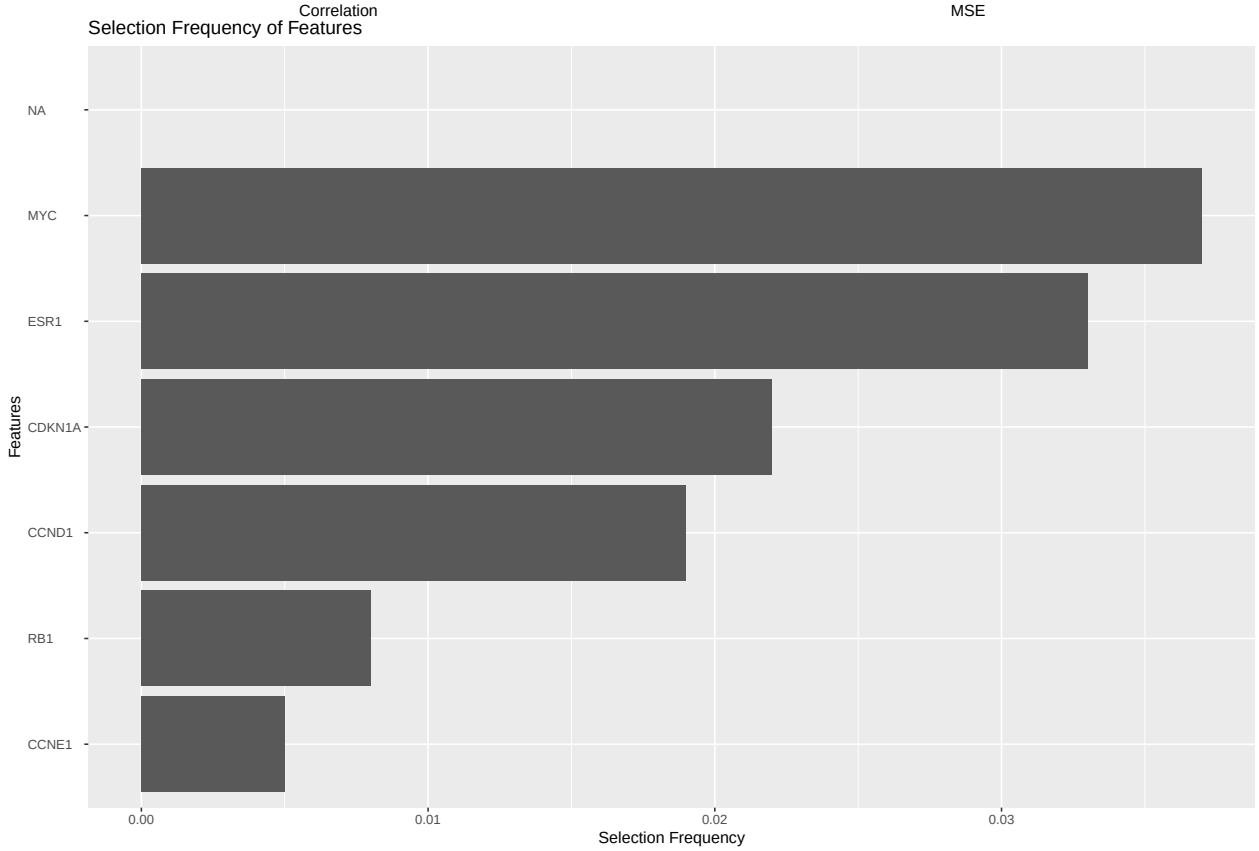


6 genes -> ROR_proliferation score (lasso - repeated cross-validation)

```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0.926
##
## CORRELATIONS RESULTS
## Mean: -0.4822298
## Median: -0.4810641
## st.dev.: 0.1724985
##
## MSE RESULTS
## Mean: 374.1519
## Median: 343.2105
## st.dev.: 144.1559
##
## Features selected 50% or more times:
## Non selected that many 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
```

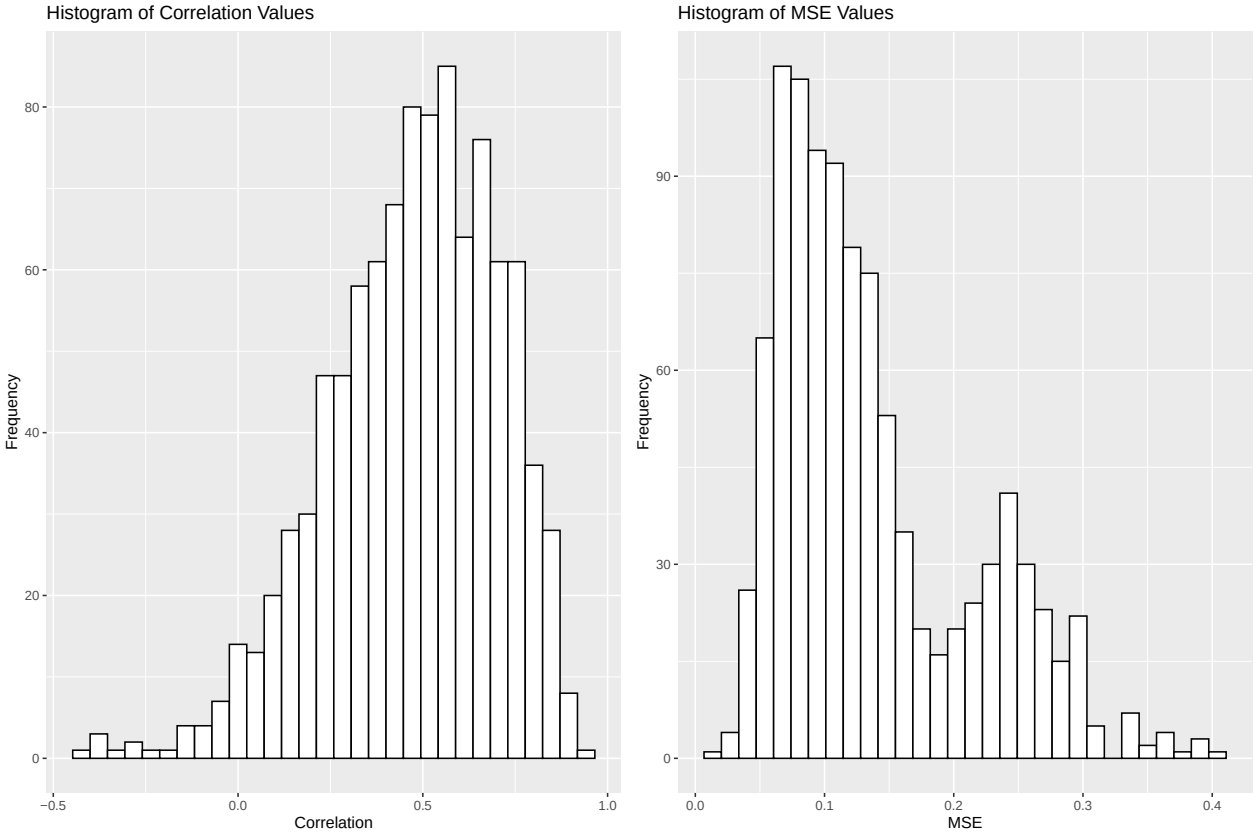

Histogram of Correlation Values
Frequency
Correlation
0.0
2.5
5.0
7.5
−0.8
−0.6
−0.4
−0.2

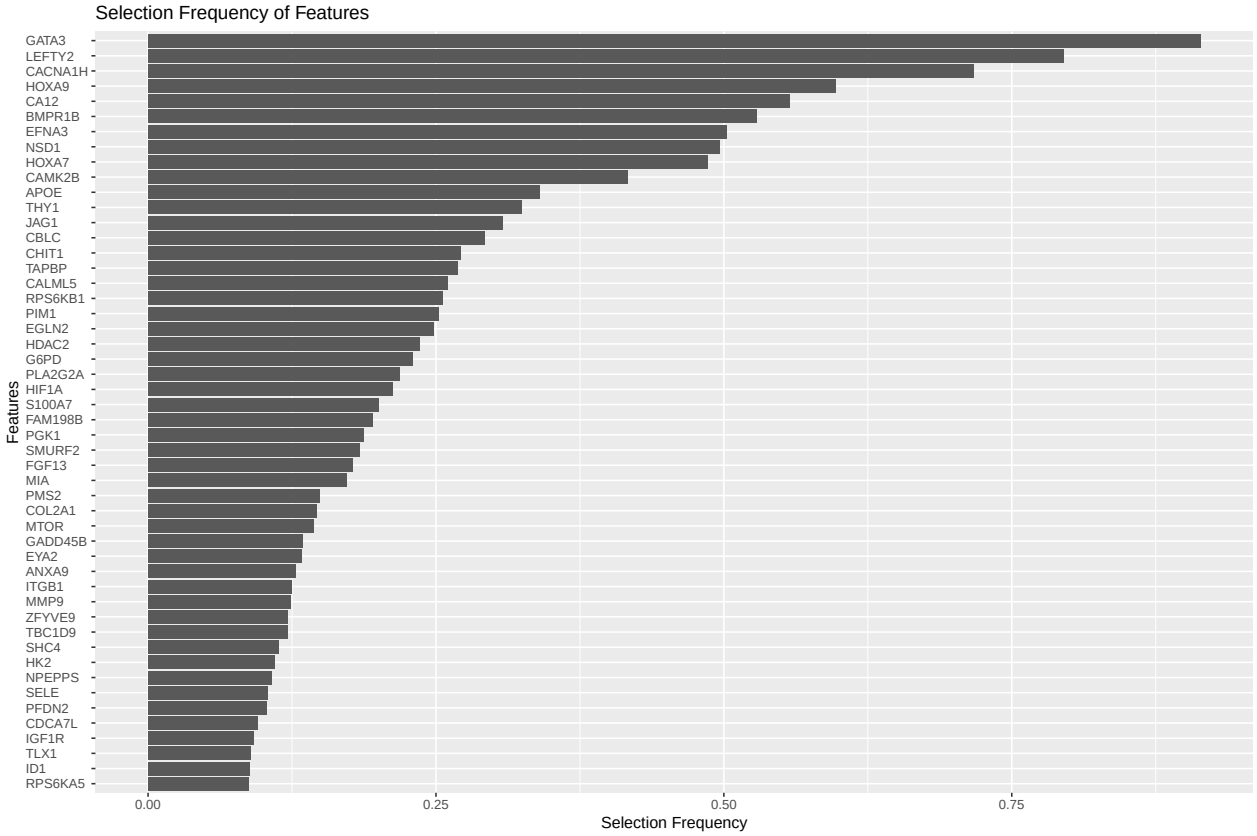

Histogram of MSE Values
Frequency
MSE
0
20
40
60
80
250
500
750


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

771 genes -> proliferation score (lasso - repeated cross-validation)

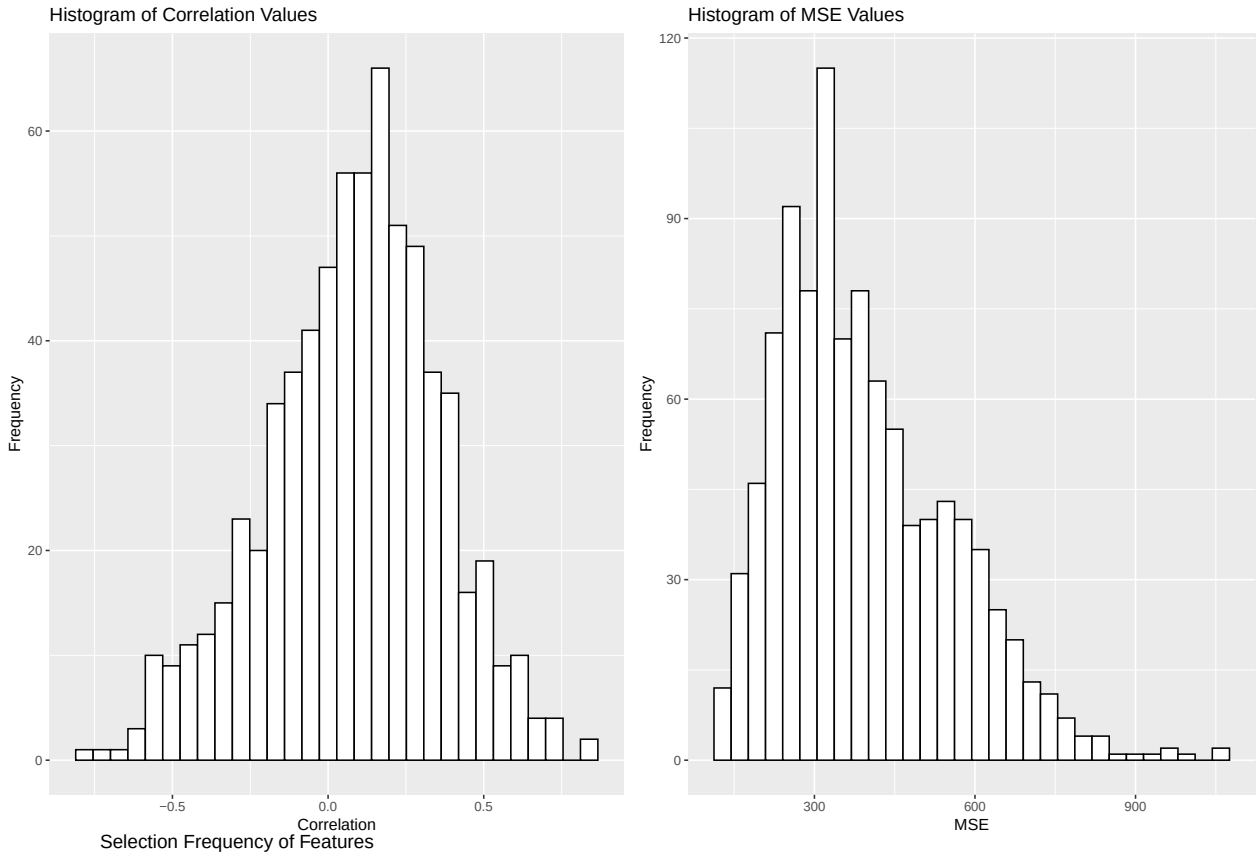
```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0.011
##
## CORRELATIONS RESULTS
## Mean: 0.4737037
## Median: 0.4959203
## st.dev.: 0.2310209
##
## MSE RESULTS
## Mean: 0.1376002
## Median: 0.1154157
## st.dev.: 0.07530557
##
## Features selected 50% or more times:
## 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"
```

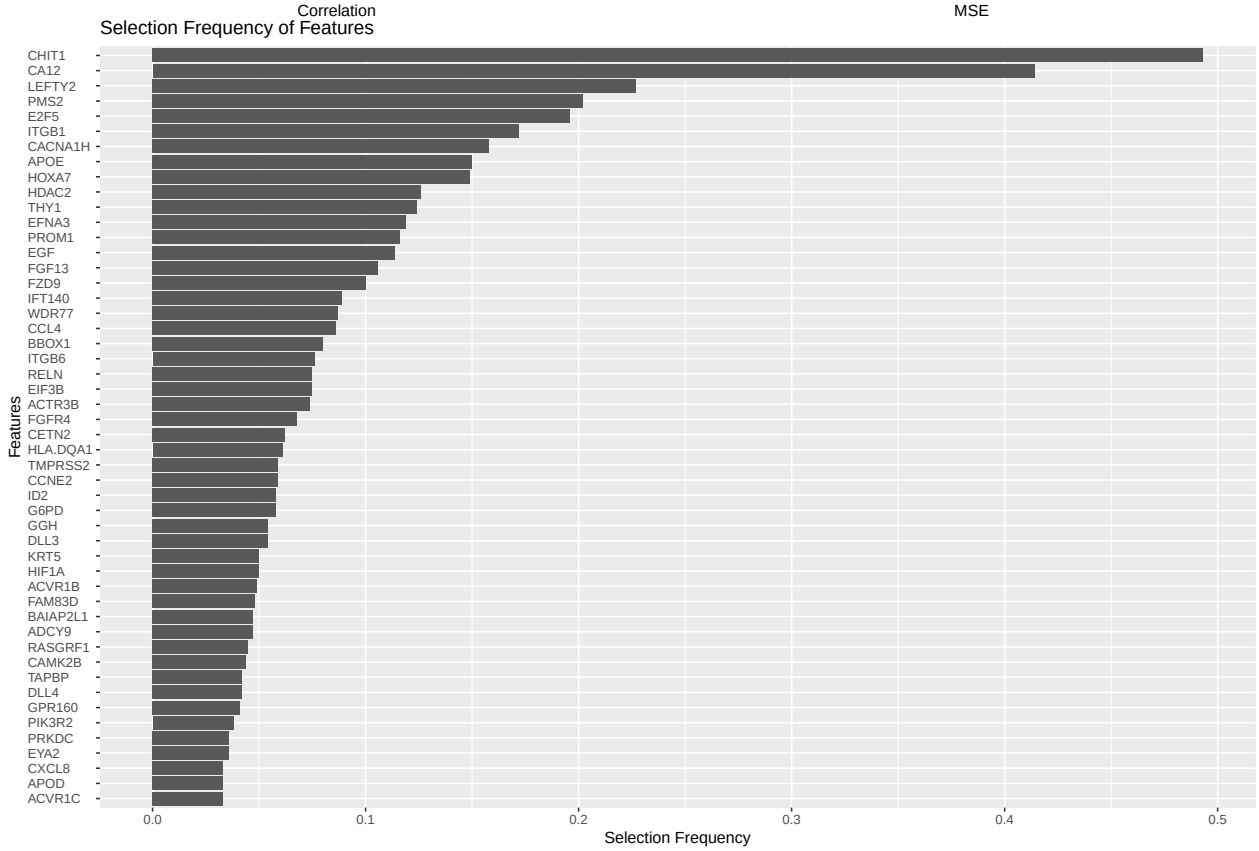




771 genes -> ROR-proliferation score (lasso - repeated cross-validation)

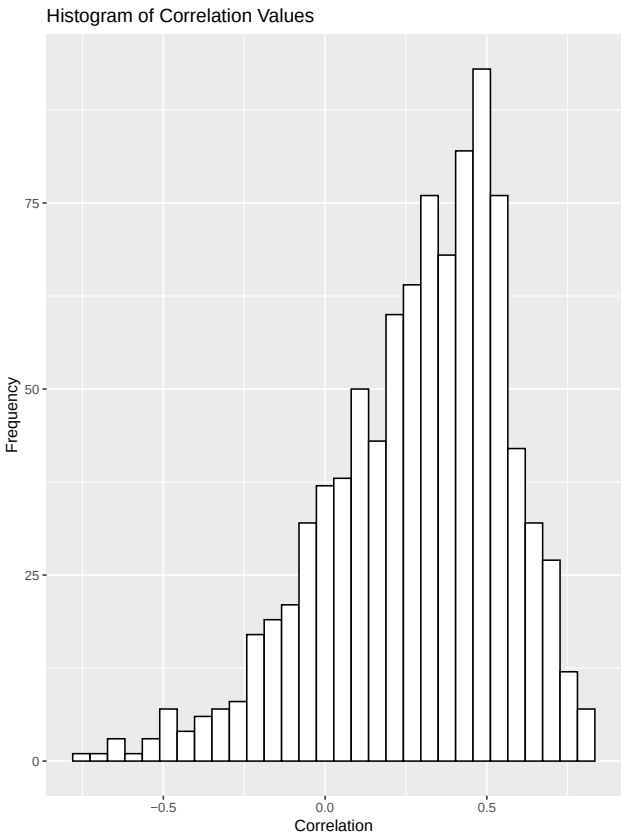
```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0.321
##
## CORRELATIONS RESULTS
## Mean: 0.08062366
## Median: 0.1014264
## st.dev.: 0.2767153
##
## MSE RESULTS
## Mean: 393.8069
## Median: 360.5105
## st.dev.: 159.6451
##
## Features selected 50% or more times:
## Non selected that many 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"
```

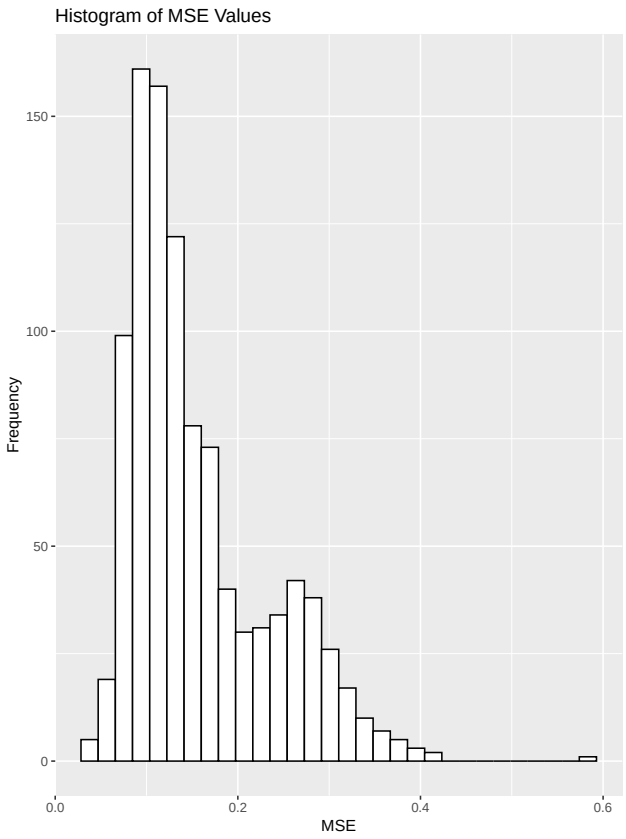

Histogram of Correlation Values
Frequency
Correlation
Histogram of MSE Values
Frequency
MSE

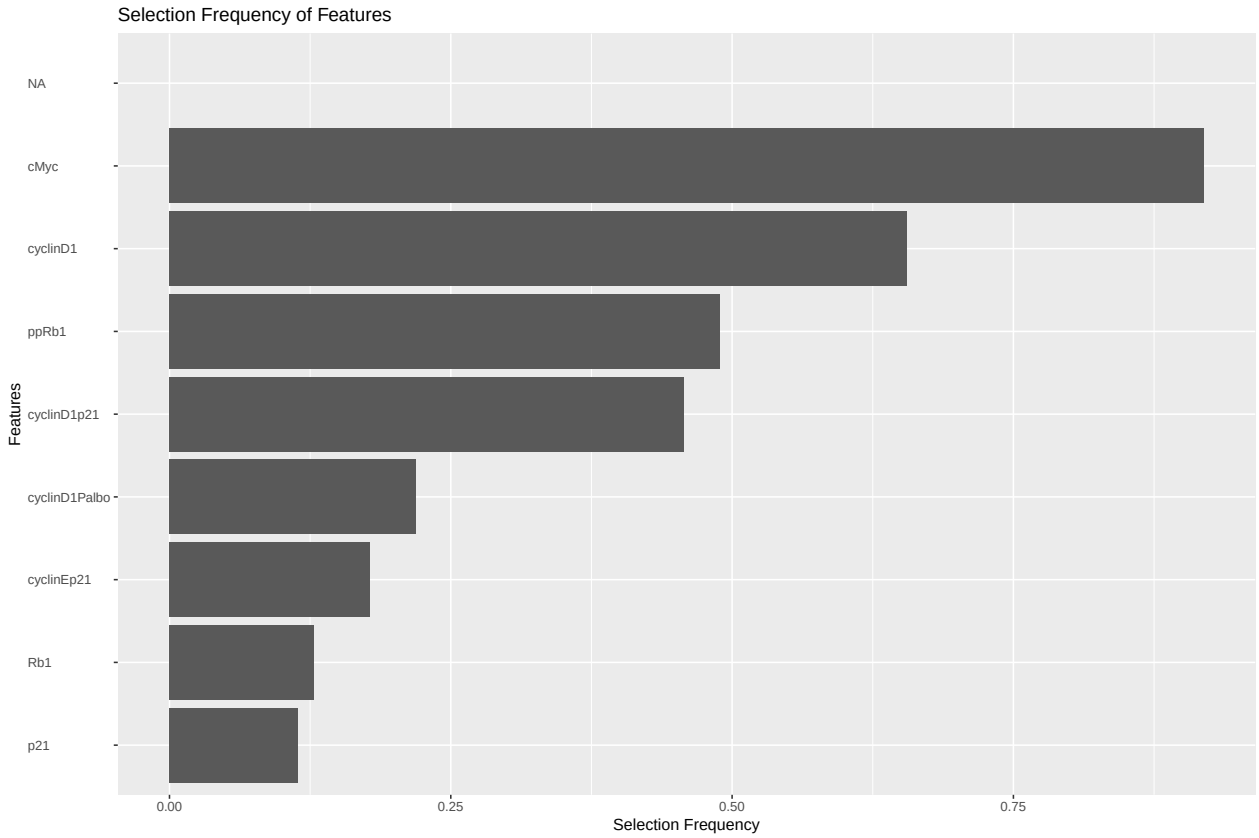

Selection Frequency of Features
Features
CHIT1
CA12
LEFTY2
PMS2
E2F5
ITGB1
CACNA1H
APOE
HOXA7
HDAC2
THY1
EFNA3
PROM1
EGF
FGF13
FZD9
IFT140
WDR77
CCL4
BBOX1
ITGB6
RELN
EIF3B
ACTR3B
FGFR4
CETN2
HLA.DQA1
TMPRSS2
CCNE2
ID2
G6PD
GGH
DLL3
KRT5
HIF1A
ACVR1B
FAM83D
BAIAP2L1
ADCY9
RASGRF1
CAMK2B
TAPBP
DLL4
GPR160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

node values -> proliferation score (lasso - repeated cross-validation)

```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0.063
##
## CORRELATIONS RESULTS
## Mean: 0.2842257
## Median: 0.3249779
## st.dev.: 0.2768458
##
## MSE RESULTS
## Mean: 0.1560308
## Median: 0.1314678
## st.dev.: 0.07628832
##
## Features selected 50% or more times:
## cyclinD1 cMyc
##
## Top 20 featrues:
##  [1] "cMyc"          "cyclinD1"      "ppRb1"         "cyclinD1p21"
##  [5] "cyclinD1Palbo" "cyclinEp21"    "Rb1"           "p21"
##  [9] NA              NA              NA              NA
## [13] NA              NA              NA              NA
## [17] NA              NA              NA              NA
```

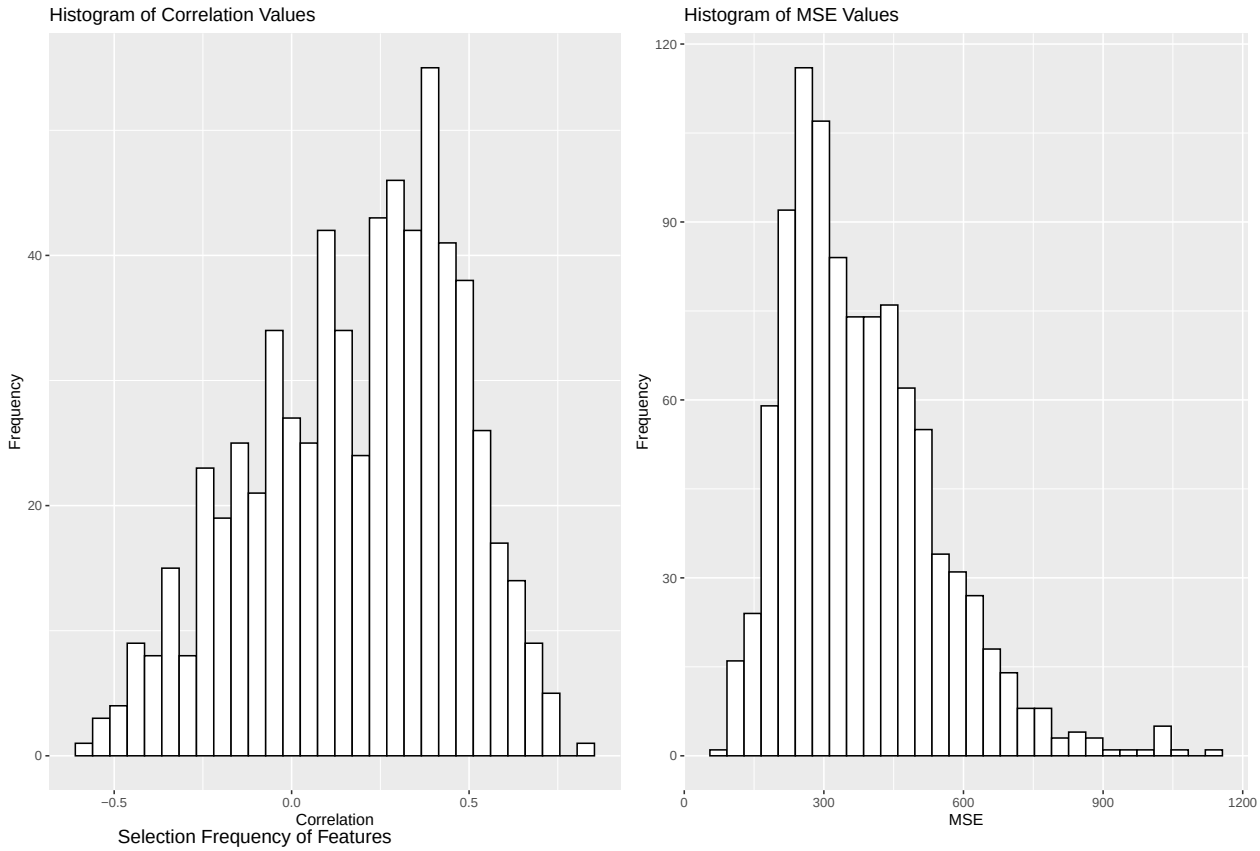


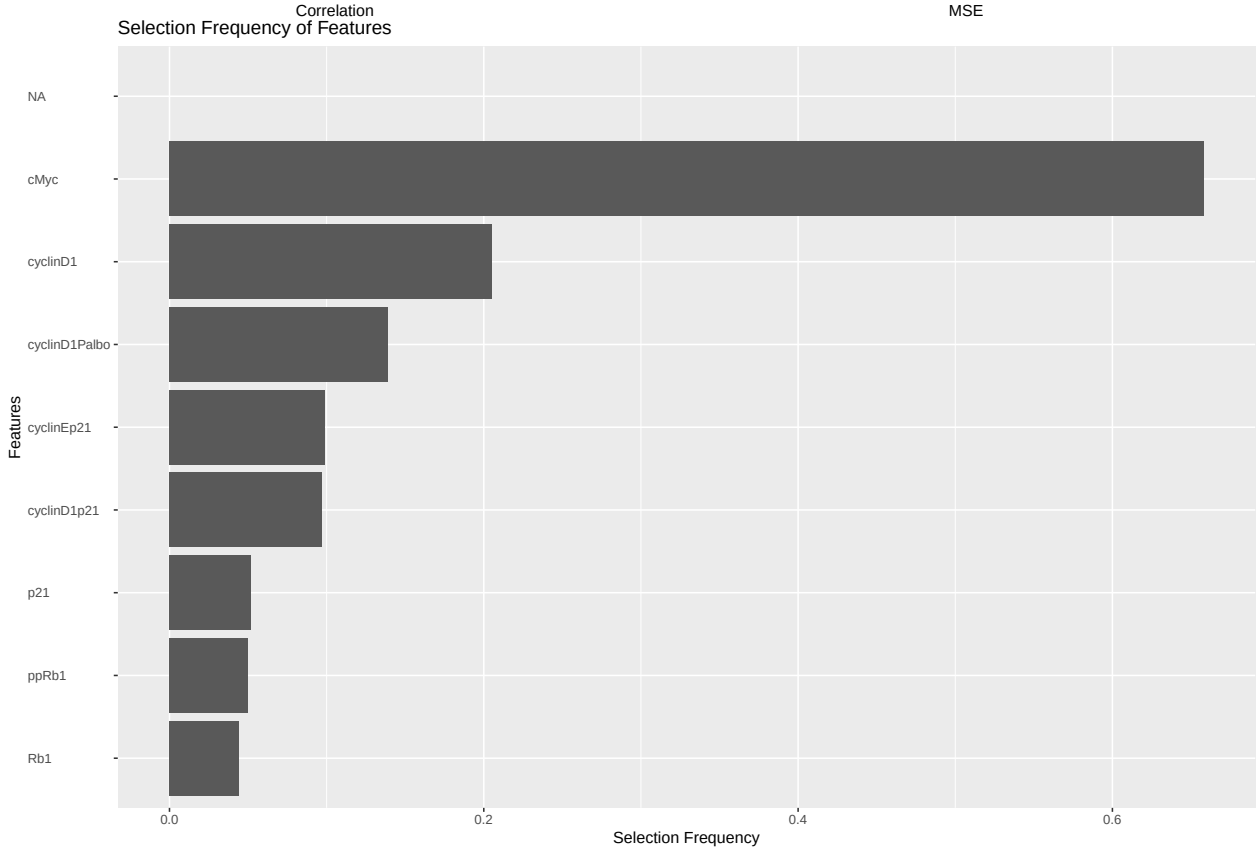




node values -> ROR-proliferation score (lasso - repeated cross-validation)

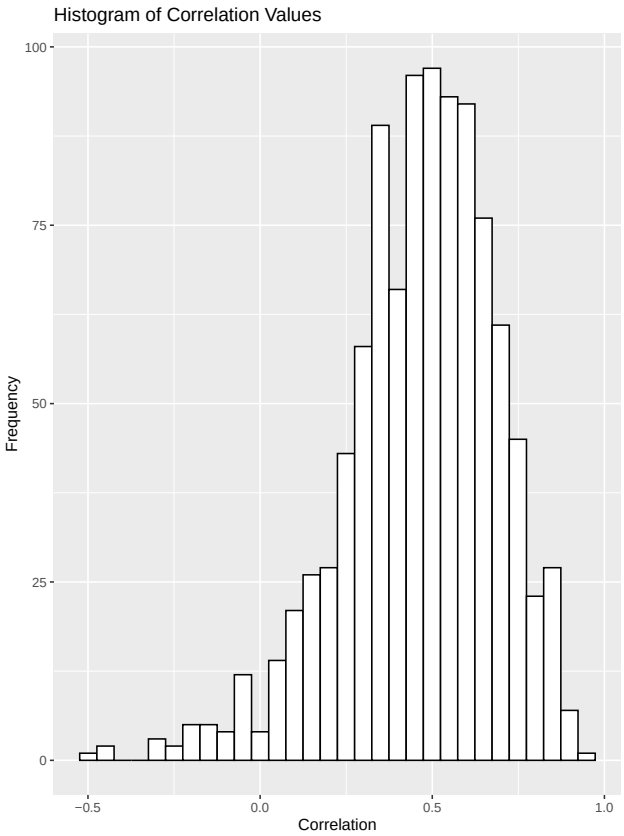
```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0.341
##
## CORRELATIONS RESULTS
## Mean: 0.1806504
## Median: 0.2237481
## st.dev.: 0.2854892
##
## MSE RESULTS
## Mean: 380.1157
## Median: 349.3312
## st.dev.: 164.5869
##
## Features selected 50% or more times:
## cMyc
##
## Top 20 featrues:
##  [1] "cMyc"          "cyclinD1"      "cyclinD1Palbo" "cyclinEp21"
##  [5] "cyclinD1p21"   "p21"           "ppRb1"         "Rb1"
##  [9] NA              NA              NA              NA
## [13] NA              NA              NA              NA
## [17] NA              NA              NA              NA
```

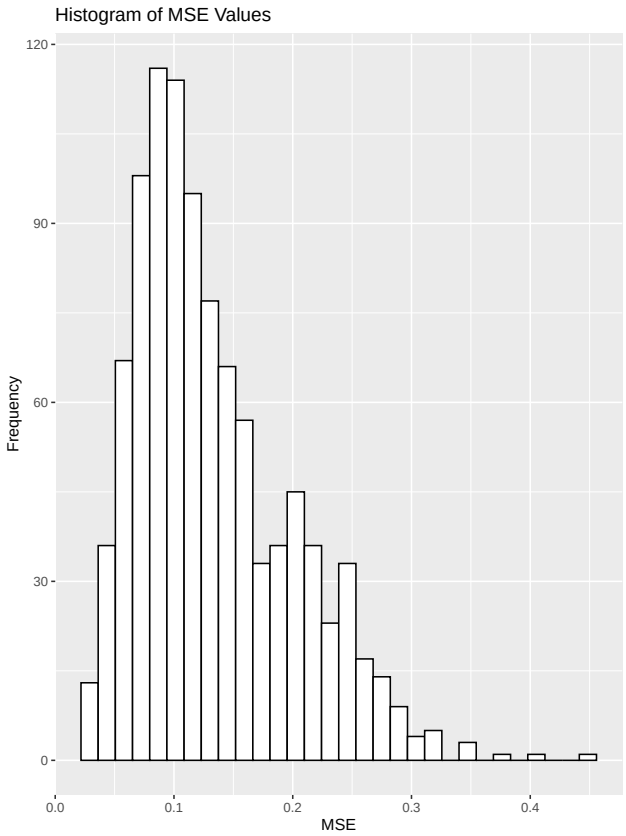

Histogram of Correlation Values
Frequency
Correlation
Histogram of MSE Values
Frequency
MSE

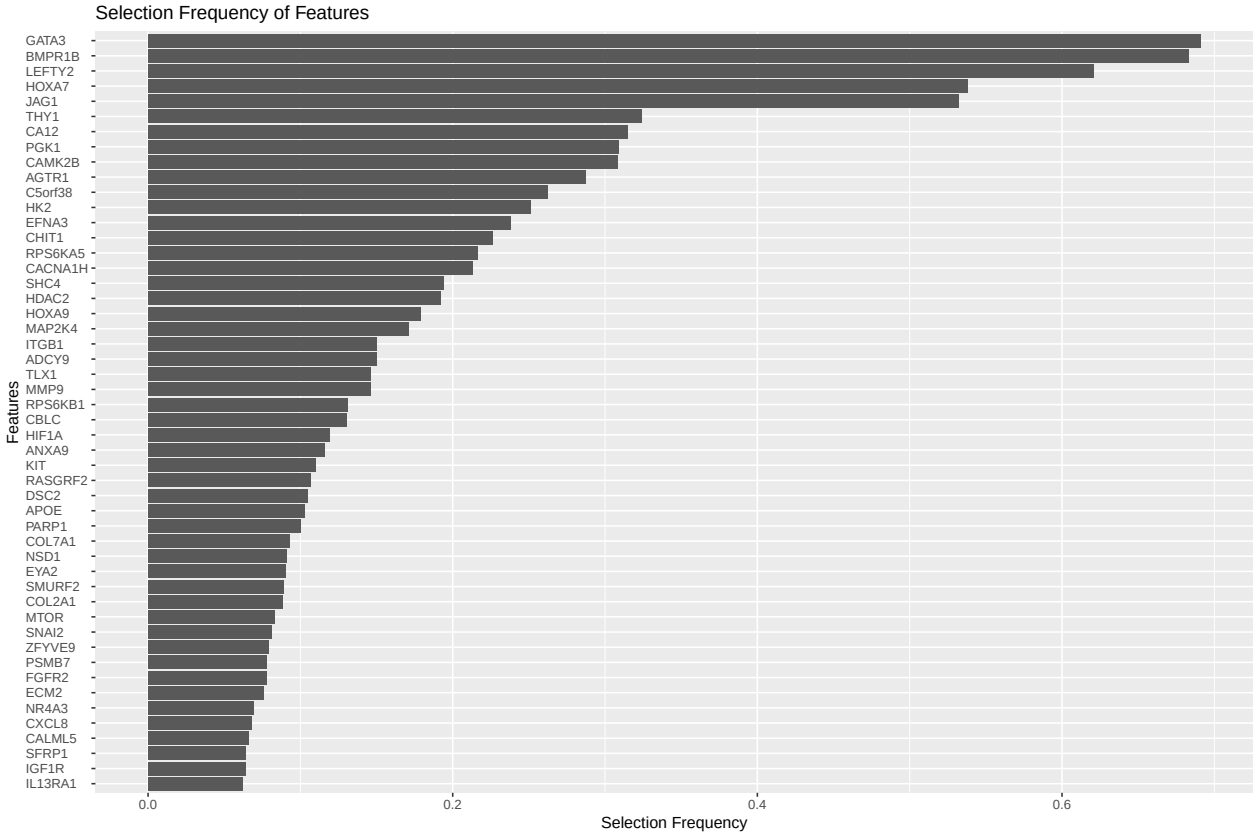

Selection Frequency of Features
Features
NA
cMyc
cyclinD1
cyclinD1Palbo
cyclinEp21
cyclinD1p21
p21
ppRb1
Rb1
Selection Frequency

Mechanistic + Residuals (additive) -> proliferation score (lasso - repeated cross-validation)

```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0
##
## CORRELATIONS RESULTS
## Mean: 0.4633095
## Median: 0.4870052
## st.dev.: 0.2227105
##
## MSE RESULTS
## Mean: 0.1331785
## Median: 0.1164927
## st.dev.: 0.06540653
##
## 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"
```

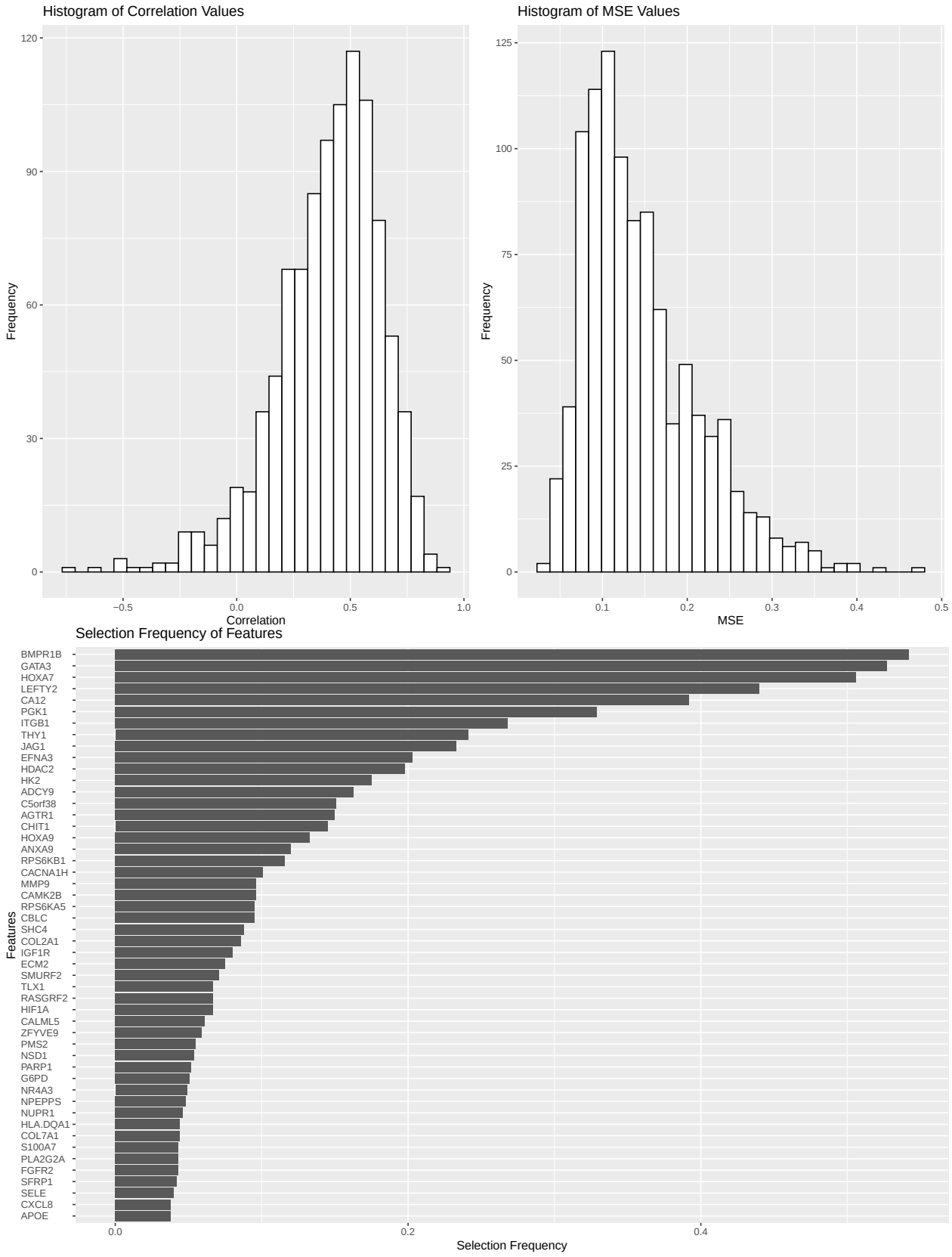






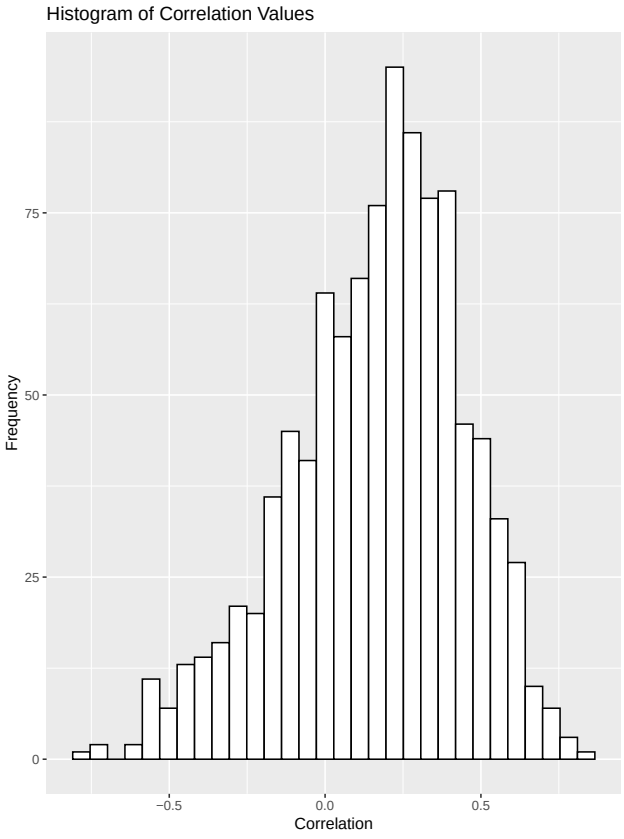
Mechanistic + Residuals (multiplicative) -> proliferation score (lasso - repeated cross-validation)

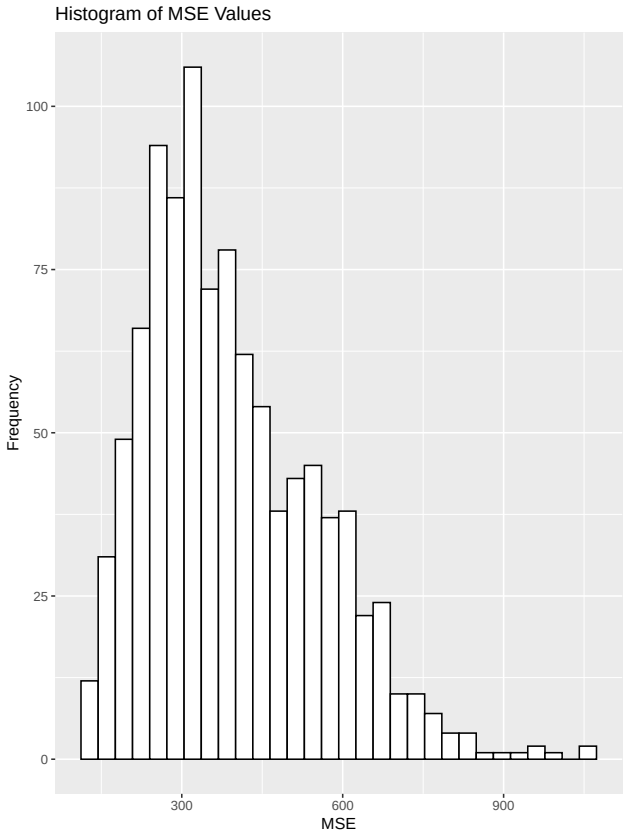
```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0
##
## CORRELATIONS RESULTS
## Mean: 0.4028471
## Median: 0.437445
## st.dev.: 0.2302632
##
## MSE RESULTS
## Mean: 0.1455819
## Median: 0.1286019
## st.dev.: 0.06806169
##
## Features selected 50% or more times:
## 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"
```

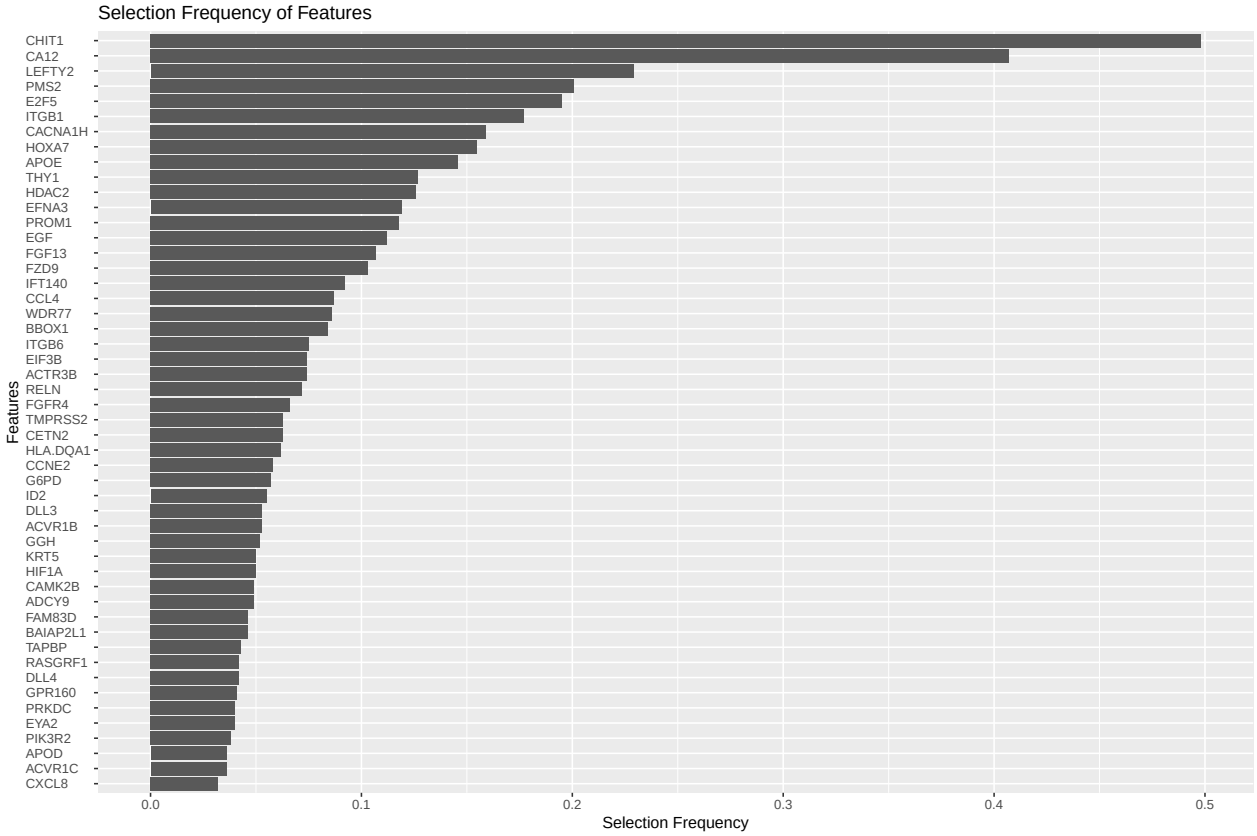

Histogram of Correlation Values
Frequency
Correlation
Histogram of MSE Values
Frequency
MSE
Selection Frequency of Features
Features
BMPR1B
GATA3
HOXA7
LEFTY2
CA12
PGK1
ITGB1
THY1
JAG1
EFNA3
HDAC2
HK2
ADCY9
C5orf38
AGTR1
CHIT1
HOXA9
ANXA9
RPS6KB1
CACNA1H
MMP9
CAMK2B
RPS6KA5
CBLC
SHC4
COL2A1
IGF1R
ECM2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

Mechnaistic + Residuals (additive) -> ROR-proliferation score (lasso - repeated cross-validation)

```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0
##
## CORRELATIONS RESULTS
## Mean: 0.16425
## Median: 0.198308
## st.dev.: 0.2788435
##
## MSE RESULTS
## Mean: 392.5436
## Median: 360.3419
## st.dev.: 158.8733
##
## Features selected 50% or more times:
## Non selected that many 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"
```

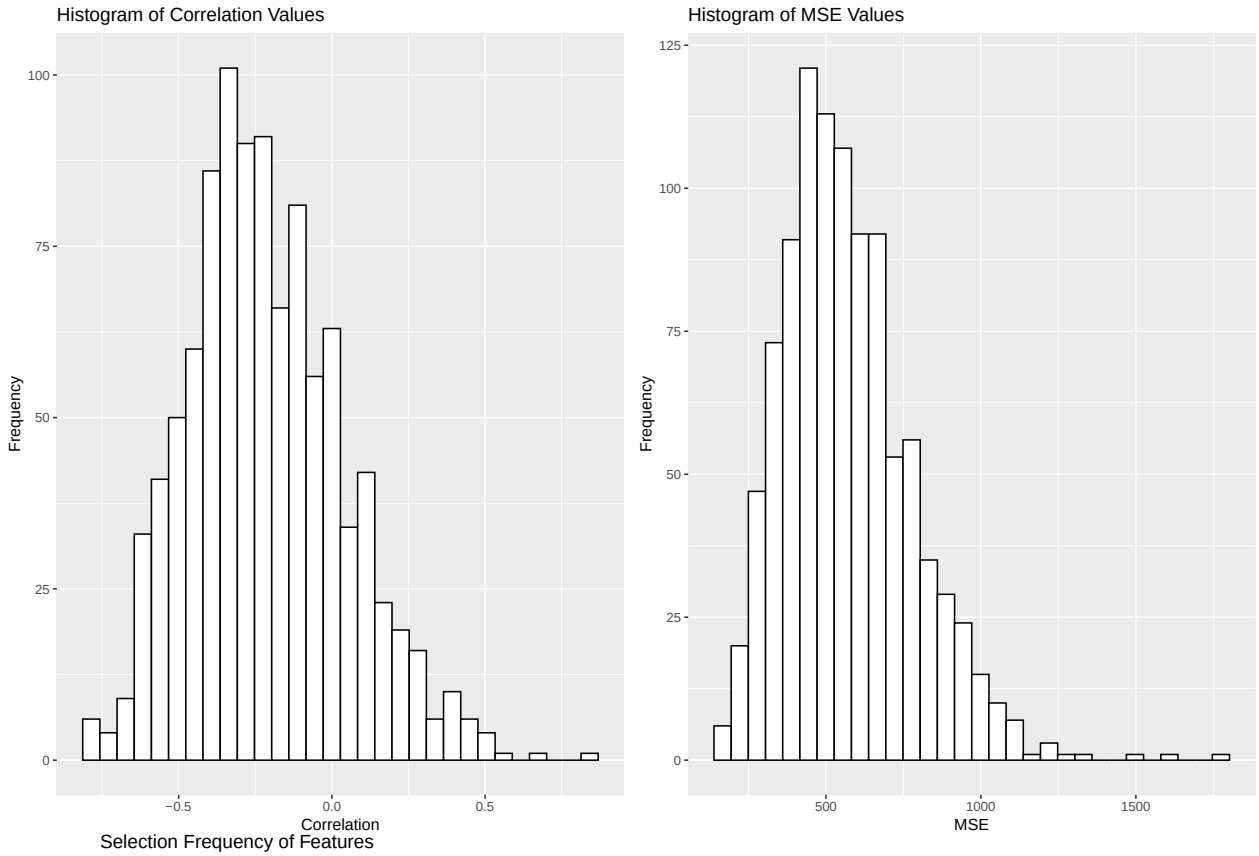


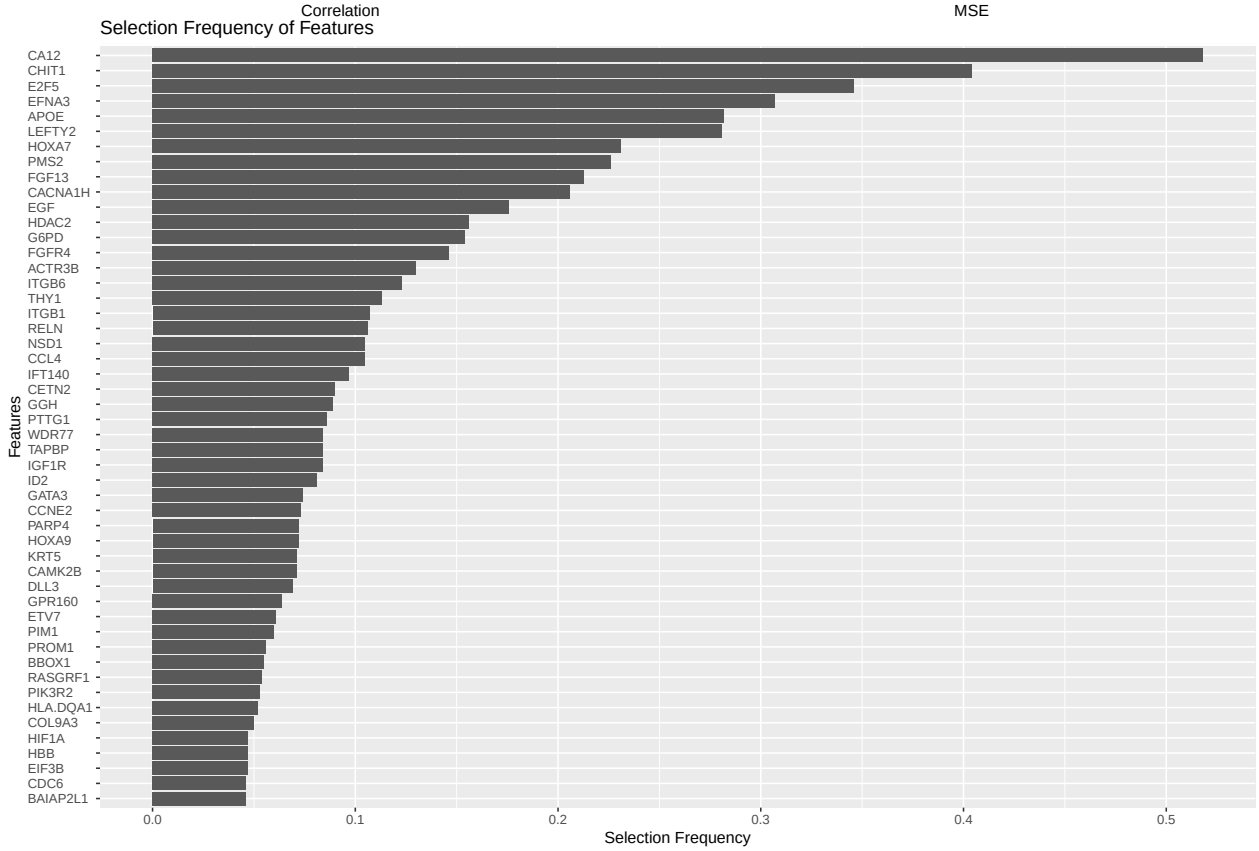




Mechnaistic + Residuals (multiplicative) -> ROR-proliferation score (lasso - repeated cross-validation)

```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0
##
## CORRELATIONS RESULTS
## Mean: -0.2145253
## Median: -0.2415987
## st.dev.: 0.2512231
##
## MSE RESULTS
## Mean: 568.8063
## Median: 541.2023
## st.dev.: 208.5911
##
## 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"
```


Histogram of Correlation Values
Frequency
Correlation
Histogram of MSE Values
Frequency
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
Selection Frequency

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

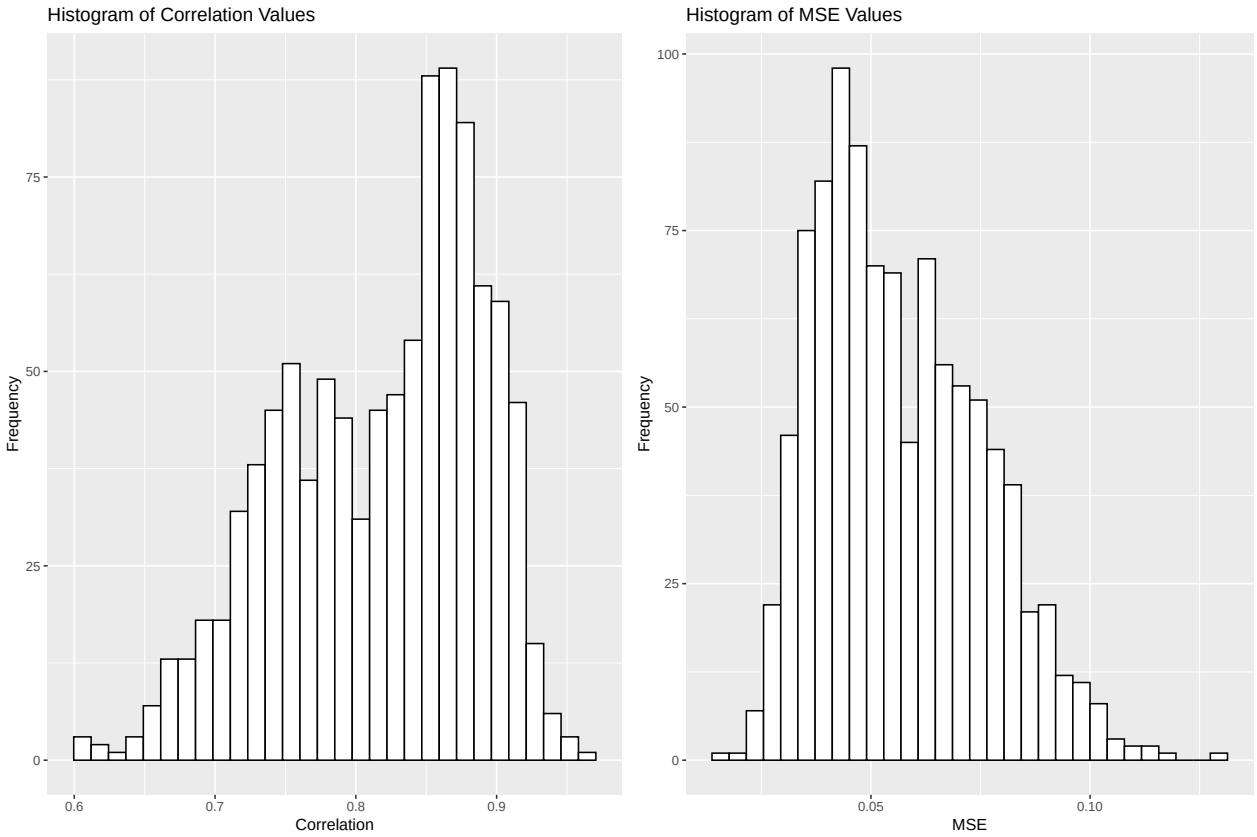
Summery results: lasso ROR+proliferation score (repeated cross-validation)

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

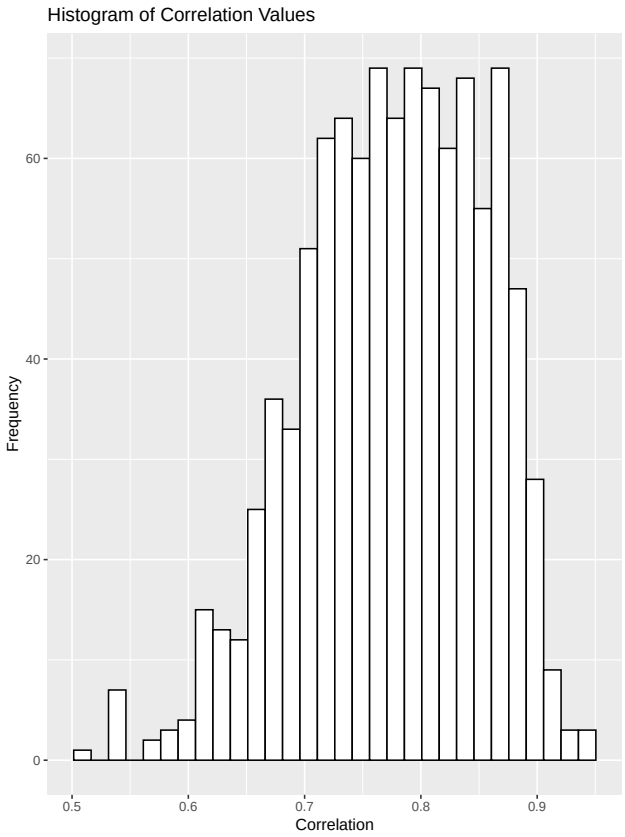
771 genes -> proliferation score (ridge bootstrap)

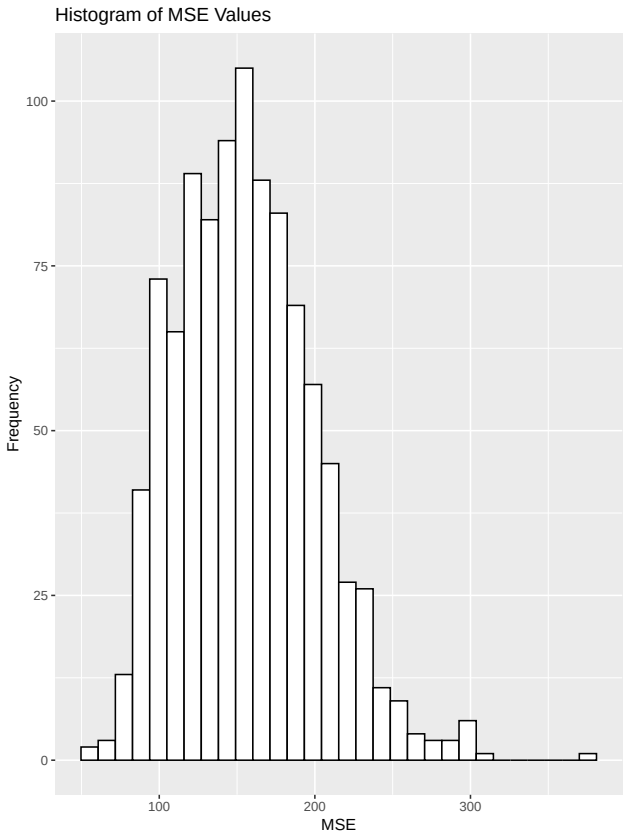
```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0
##
## CORRELATIONS RESULTS
## Mean: 0.8189568
## Median: 0.8355063
## st.dev.: 0.07102764
##
## MSE RESULTS
## Mean: 0.05665241
## Median: 0.0532761
## st.dev.: 0.0186773
```



771 genes -> ROR-proliferation score (ridge bootstrap)

```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0
##
## CORRELATIONS RESULTS
## Mean: 0.7761924
## Median: 0.7811637
## st.dev.: 0.07746885
##
## MSE RESULTS
## Mean: 156.065
## Median: 154.0679
## st.dev.: 44.18173
```

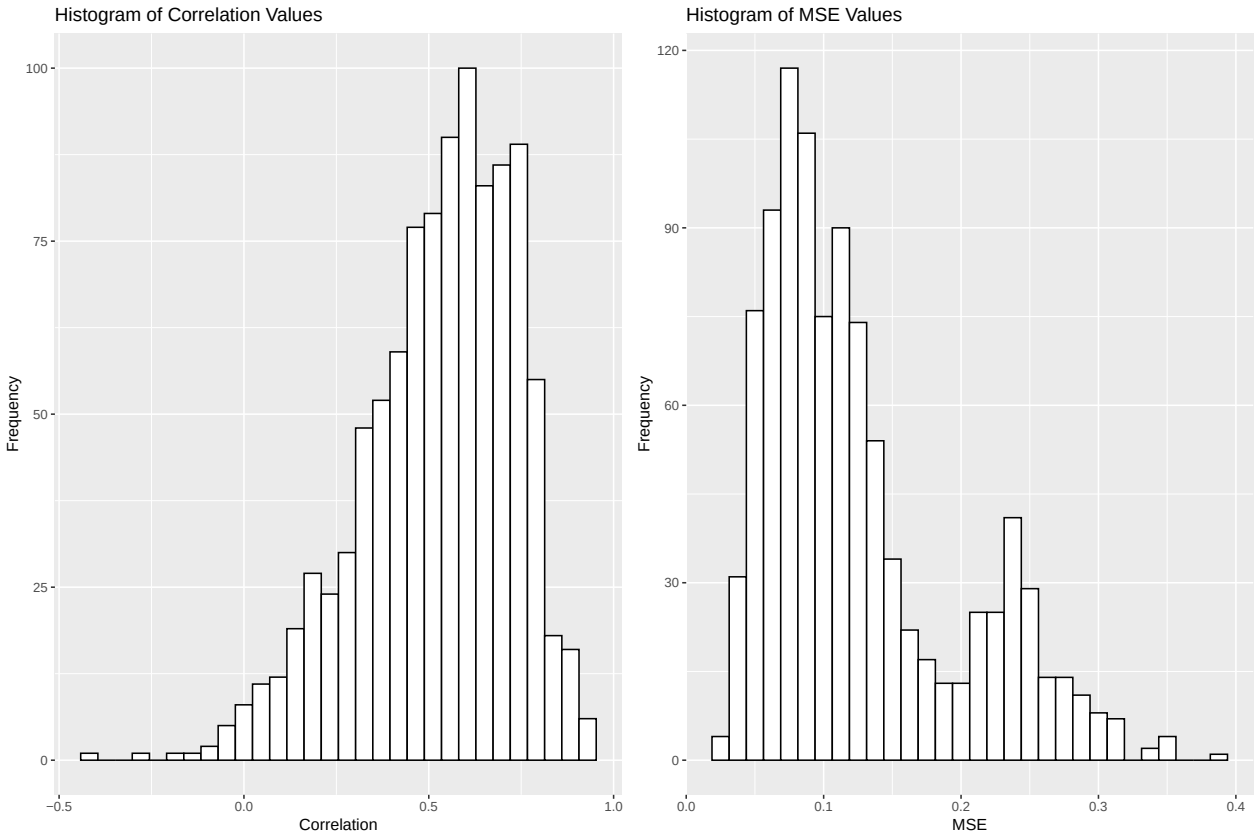




Ridge repeated cross-validation

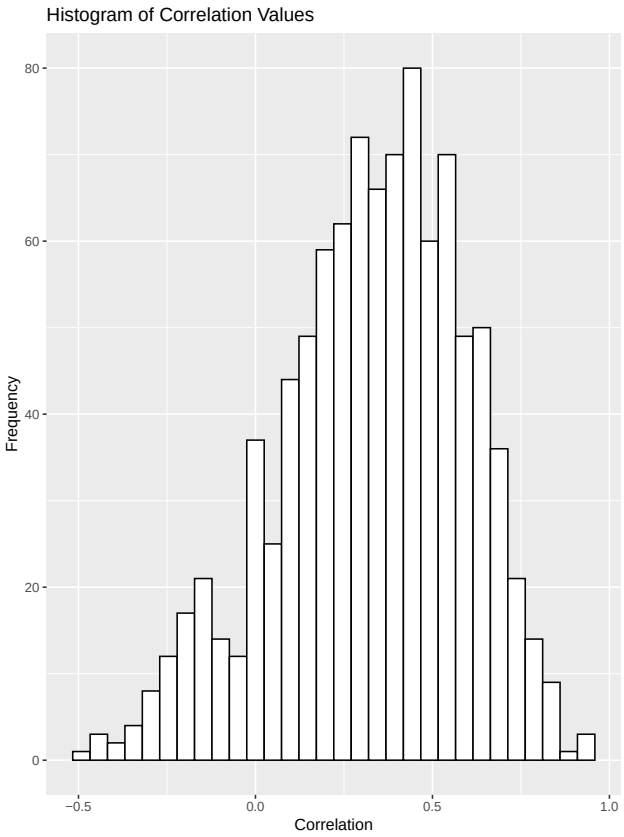
771 genes -> proliferation score (ridge repeated cross-validation)

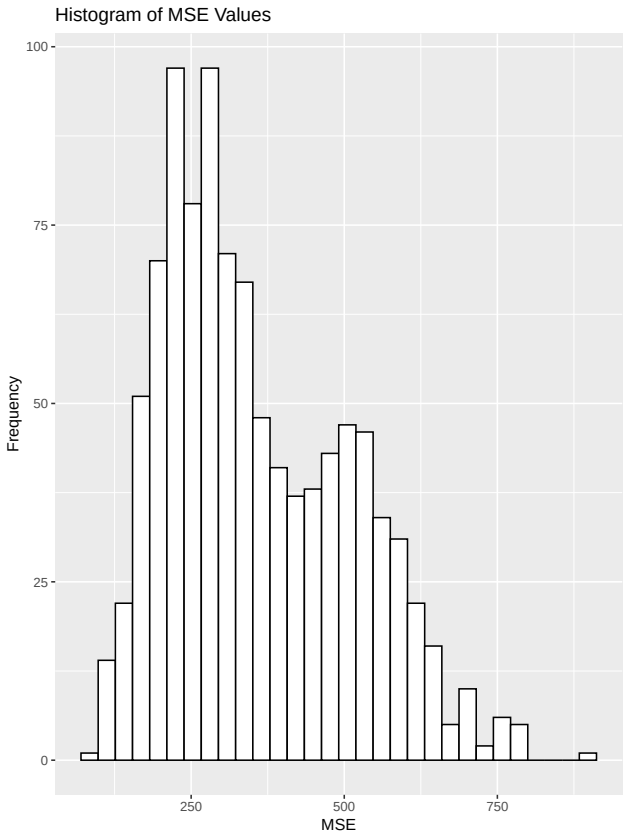
```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0
##
## CORRELATIONS RESULTS
## Mean: 0.5268342
## Median: 0.5562175
## st.dev.: 0.2071537
##
## MSE RESULTS
## Mean: 0.1256548
## Median: 0.1059589
## st.dev.: 0.06993178
```



771 genes -> ROR-proliferation score (ridge repeated cross-validation)

```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0.029
##
## CORRELATIONS RESULTS
## Mean: 0.3286101
## Median: 0.3486304
## st.dev.: 0.2621708
##
## MSE RESULTS
## Mean: 357.5098
## Median: 322.5114
## st.dev.: 150.0655
```





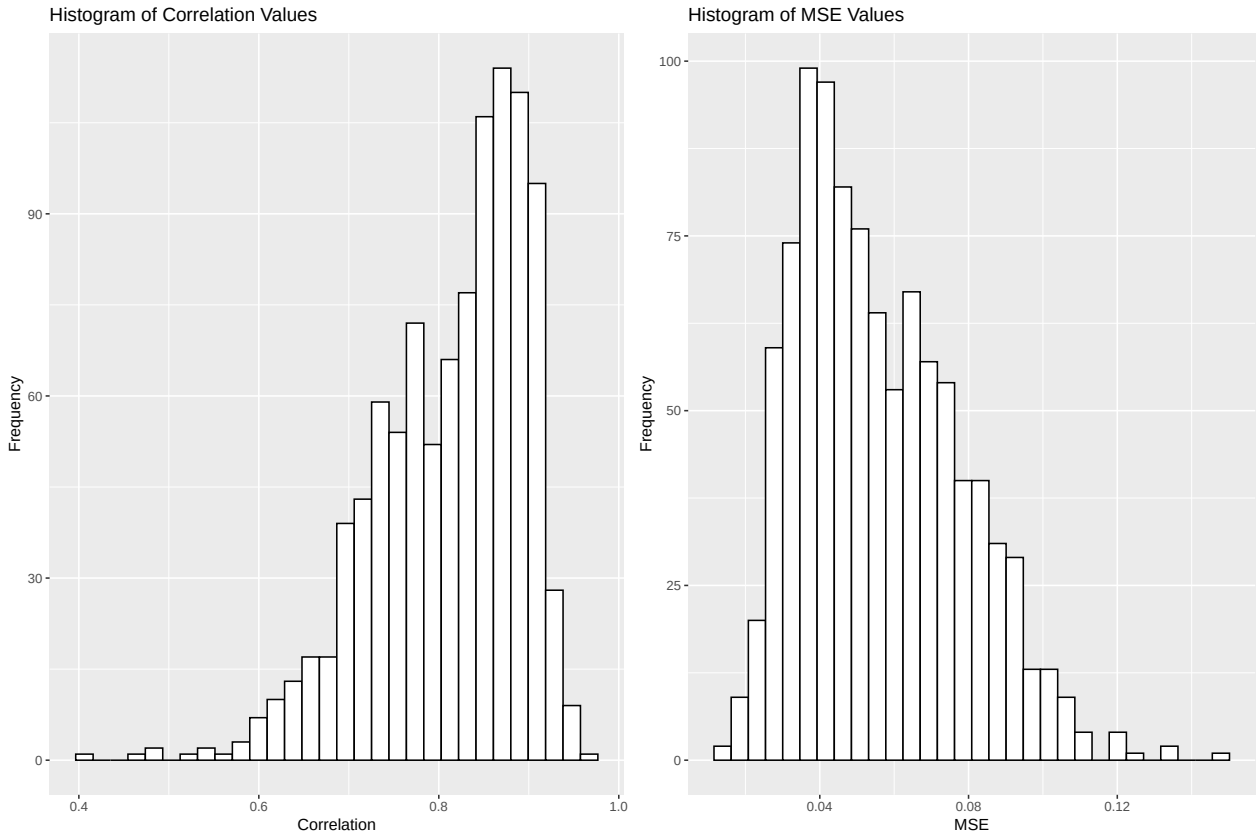
Summery results: ridge 771 genes bootstrap and repeated cross-validation

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	0.3286101	0.2621708	357.5098096	150.0654797

Elastic Net - bootstrap

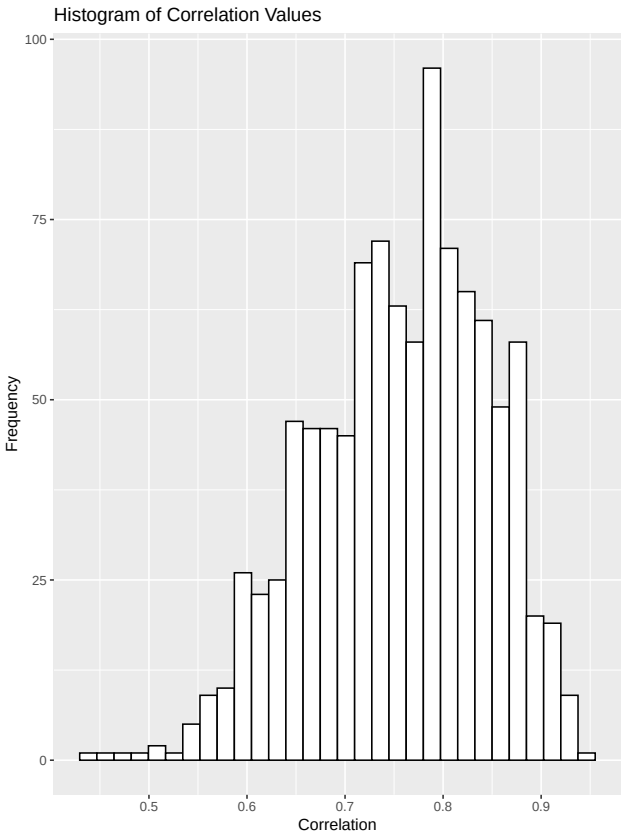
771 genes -> proliferation score (elastic Net - bootstrap)

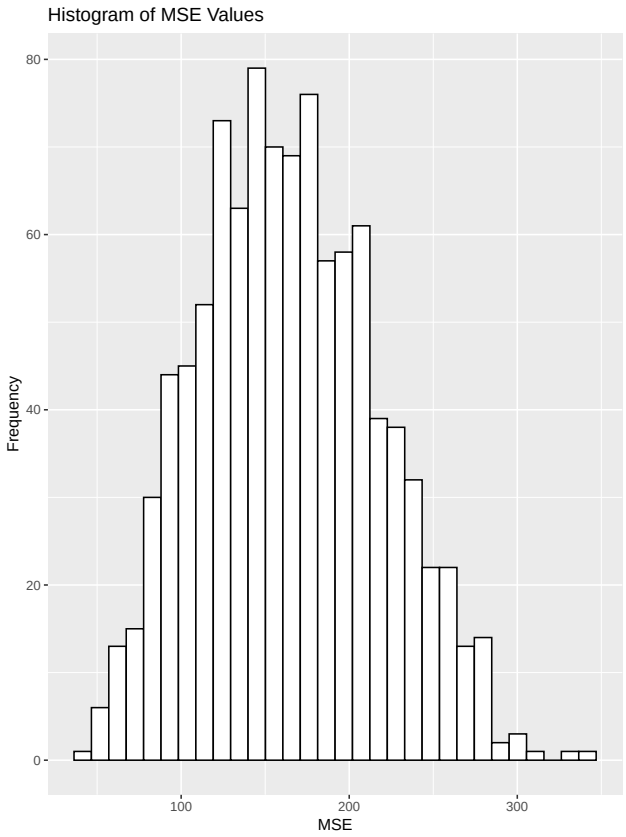
```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0
##
## CORRELATIONS RESULTS
## Mean: 0.8125591
## Median: 0.8324809
## st.dev.: 0.0838975
##
## MSE RESULTS
## Mean: 0.0558955
## Median: 0.05152763
## st.dev.: 0.02170316
```



771 genes -> ROR-proliferation score (elastic Net - bootstrap)

```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0
##
## CORRELATIONS RESULTS
## Mean: 0.7565123
## Median: 0.7644687
## st.dev.: 0.08881431
##
## MSE RESULTS
## Mean: 164.4116
## Median: 162.2643
## st.dev.: 52.77518
```

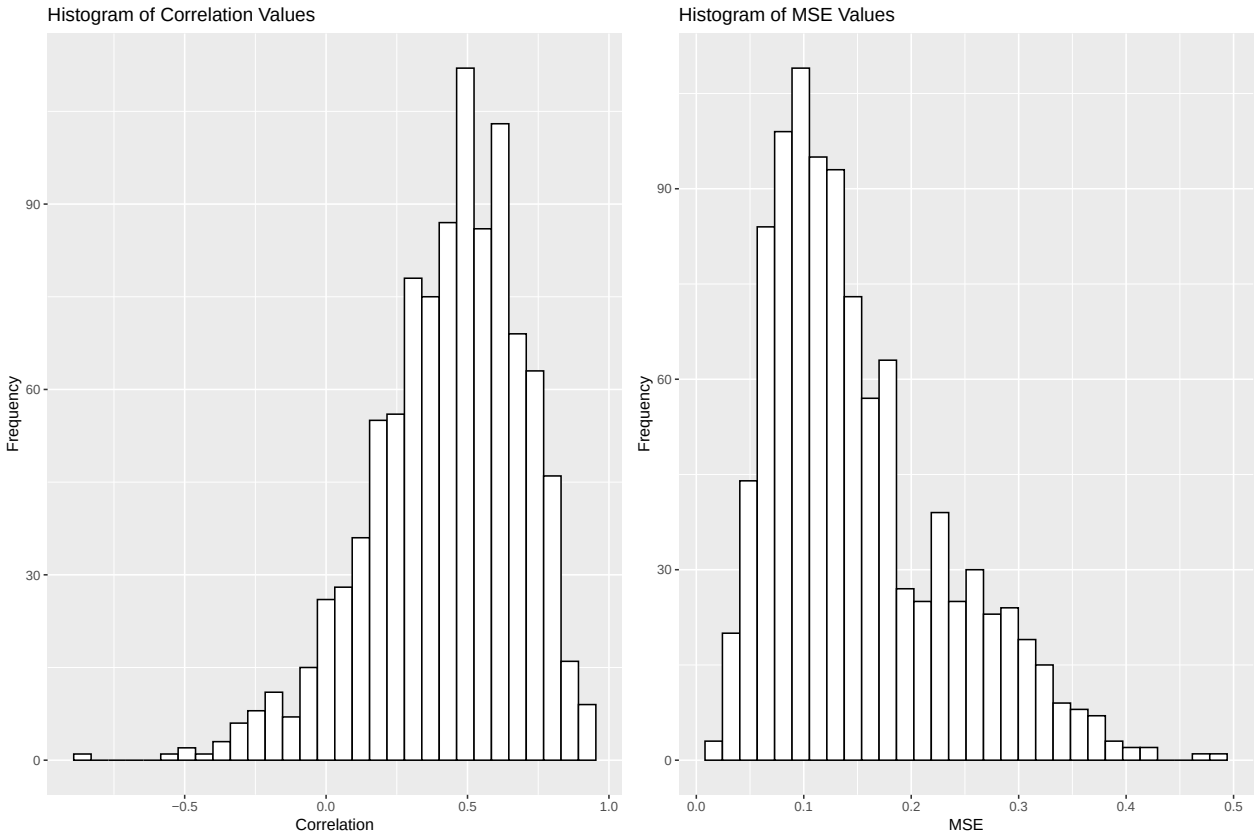




Elastic Net: cross-validation

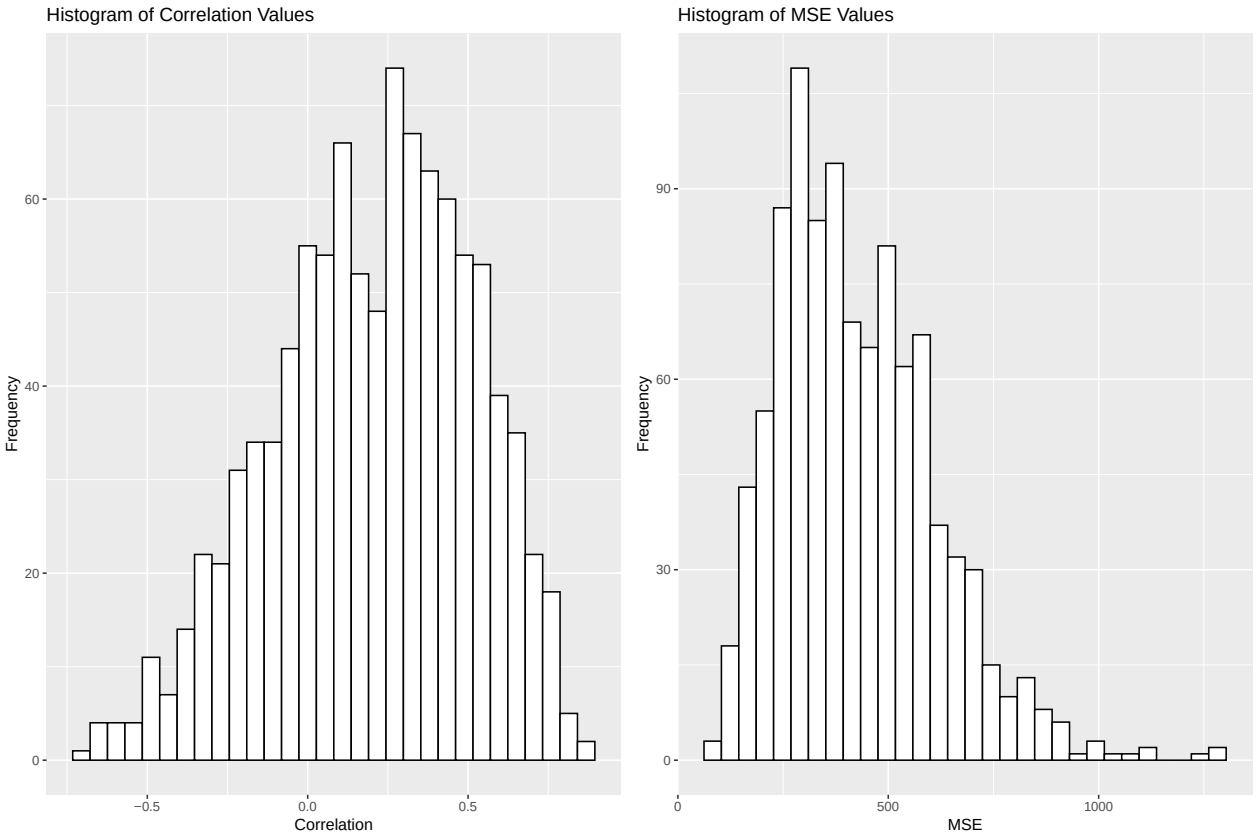
771 genes -> proliferation score (elastic Net - repeated cross-validation)

```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0
##
## CORRELATIONS RESULTS
## Mean: 0.427189
## Median: 0.4644365
## st.dev.: 0.2637029
##
## MSE RESULTS
## Mean: 0.1501025
## Median: 0.1284895
## st.dev.: 0.08224004
```



771 genes -> ROR-proliferation score (elastic Net - repeated cross-validation)

```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0.002
##
## CORRELATIONS RESULTS
## Mean: 0.2049341
## Median: 0.226514
## st.dev.: 0.3111969
##
## MSE RESULTS
## Mean: 427.3517
## Median: 396.2622
## st.dev.: 186.2465
```



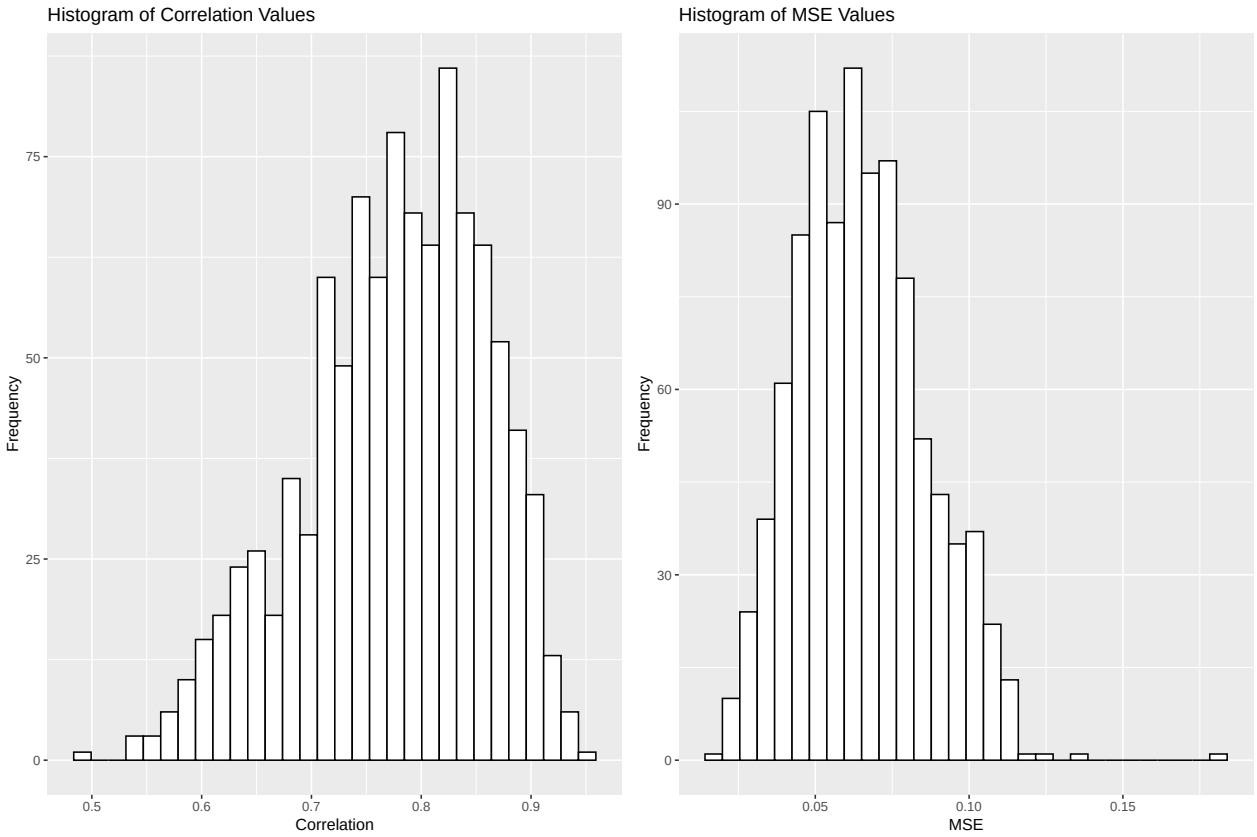
Summery results: elastic net 771 genes bootstrap and repeated cross-validation

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

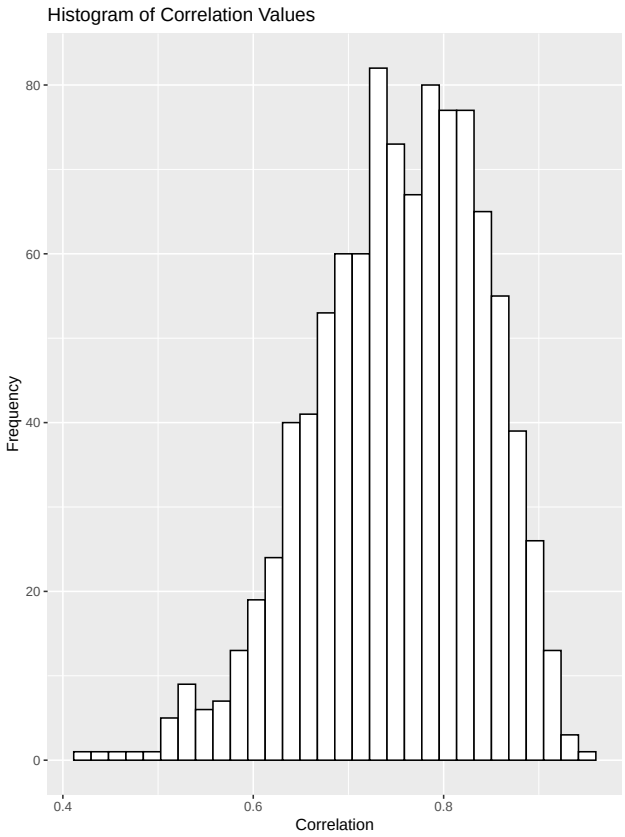
771 genes -> proliferation score (boosting - bootstrap)

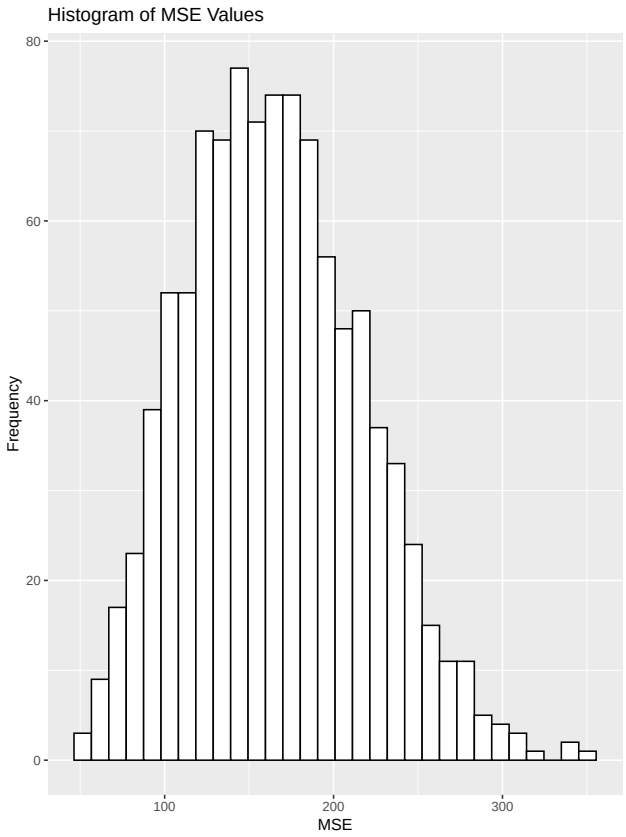
```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0
##
## CORRELATIONS RESULTS
## Mean: 0.7760479
## Median: 0.7841718
## st.dev.: 0.08278286
##
## MSE RESULTS
## Mean: 0.06537103
## Median: 0.06397191
## st.dev.: 0.02100152
```



771 genes -> ROR-proliferation score (boosting - bootstrap)

```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0
##
## CORRELATIONS RESULTS
## Mean: 0.7530515
## Median: 0.7593775
## st.dev.: 0.08824321
##
## MSE RESULTS
## Mean: 165.145
## Median: 162.2361
## st.dev.: 51.86608
```

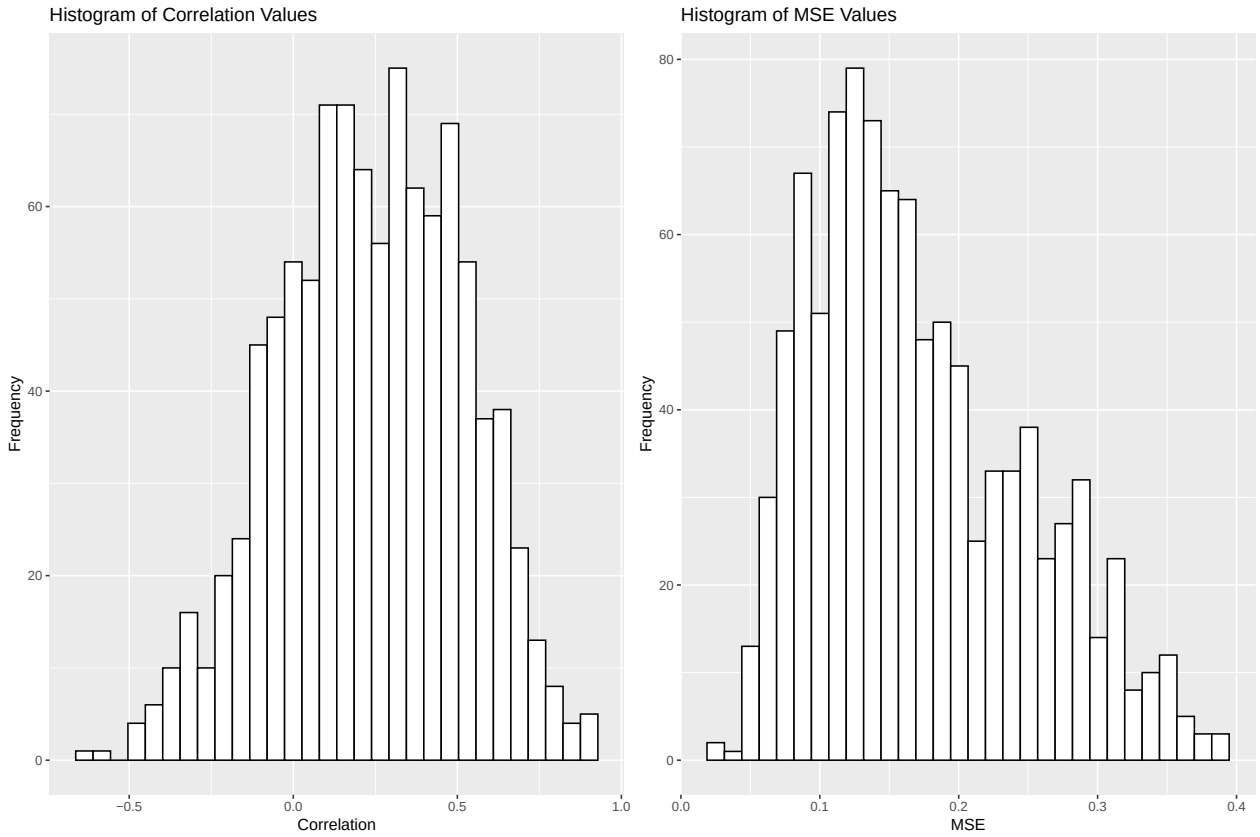




Boosting with stumps as base learner cross-validation

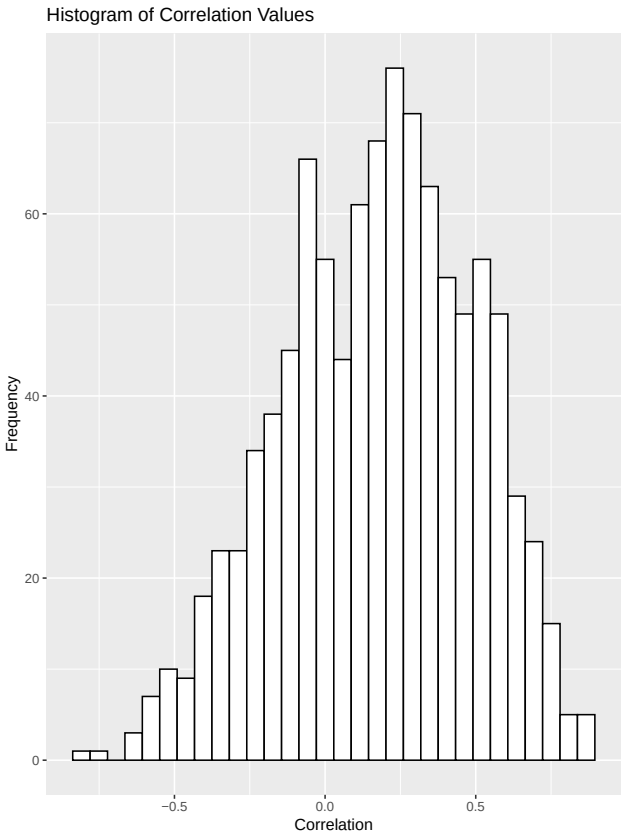
771 genes -> proliferation score (boosting - rep cross-val)

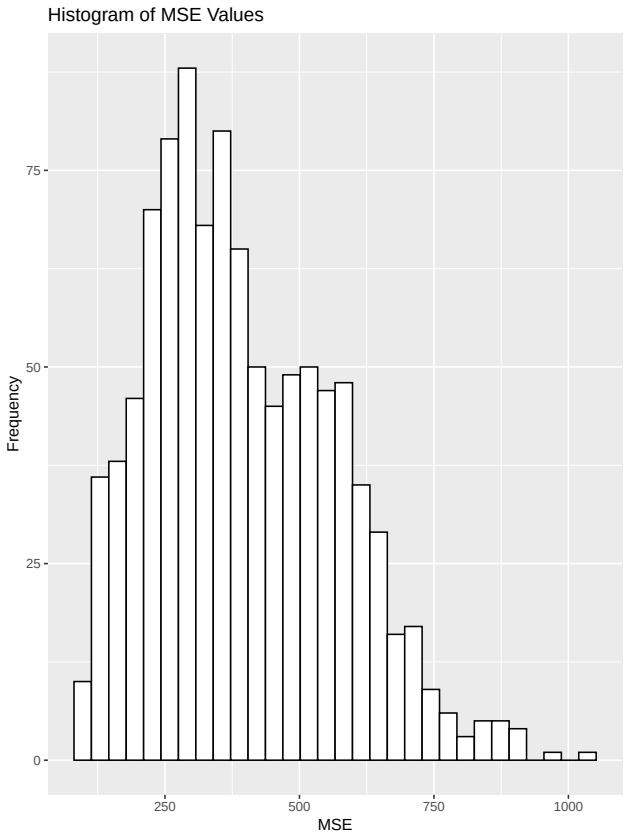
```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0
##
## CORRELATIONS RESULTS
## Mean: 0.2364594
## Median: 0.2426694
## st.dev.: 0.2795041
##
## MSE RESULTS
## Mean: 0.1712618
## Median: 0.1554505
## st.dev.: 0.07640141
```



771 genes -> ROR-proliferation score (boosting - rep cross-val)

```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0
##
## CORRELATIONS RESULTS
## Mean: 0.1744792
## Median: 0.1948469
## st.dev.: 0.3151549
##
## MSE RESULTS
## Mean: 394.2634
## Median: 366.3266
## 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

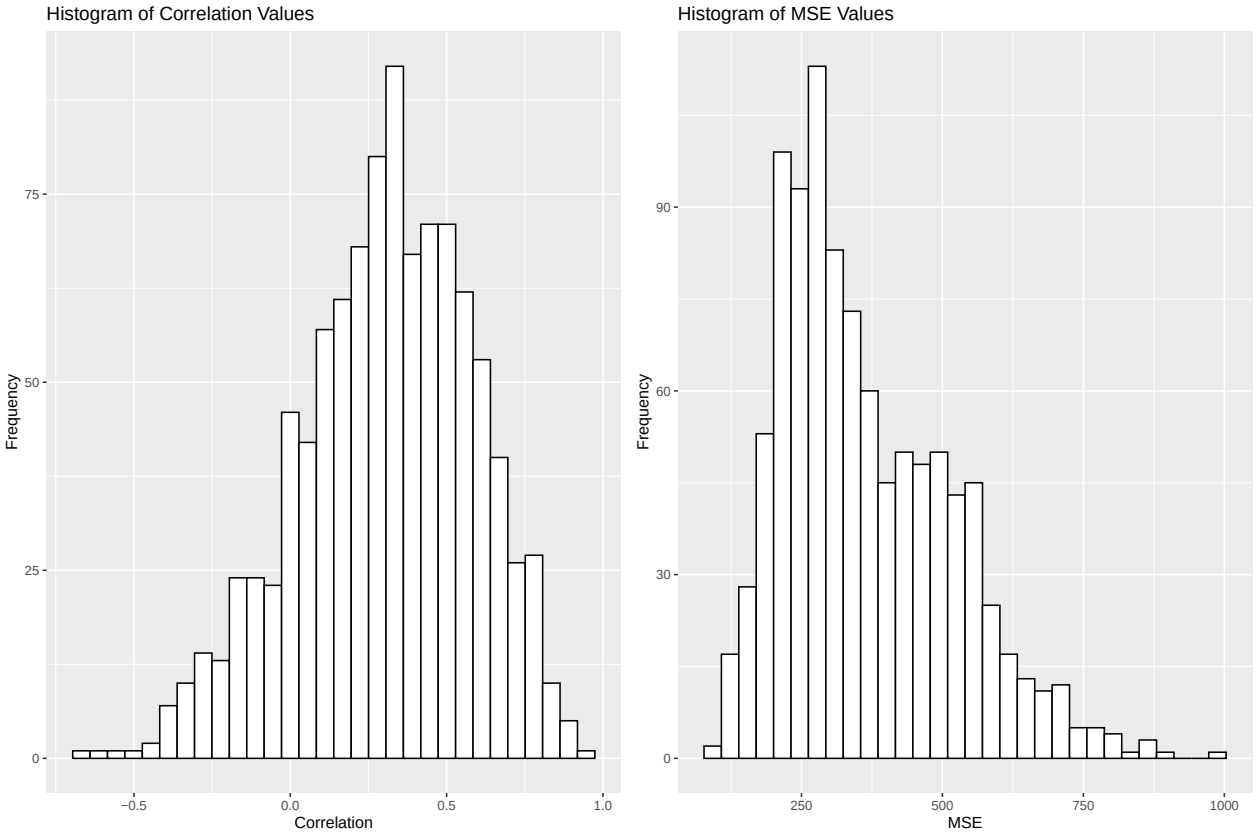
START USING DOMAIN KNOWLEDGE

PCA on signature gene sets using repeated cross-validation

Ridge: 771 genes -> ROR-proliferation score

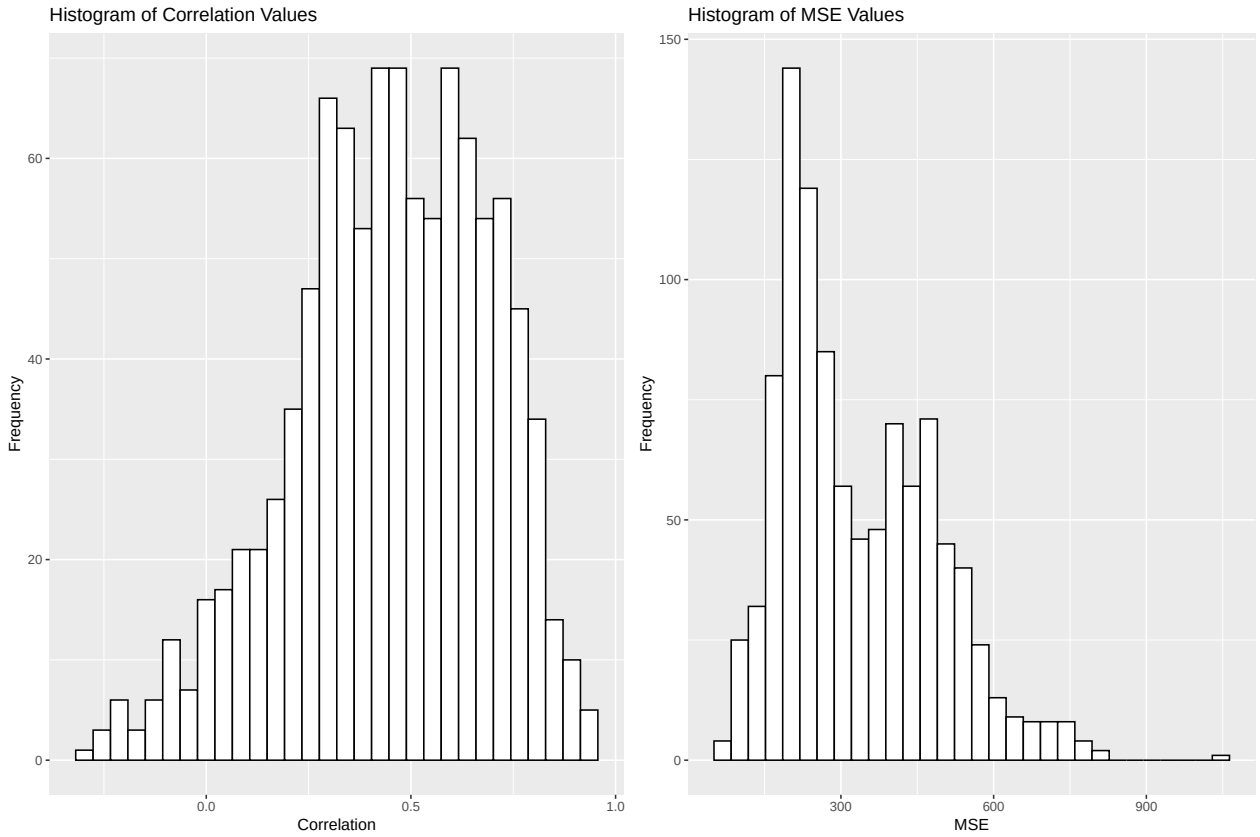
```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0
##
## CORRELATIONS RESULTS
## Mean: 0.3028091
## Median: 0.319312
## st.dev.: 0.2779432
##
## MSE RESULTS
## Mean: 364.5733
## Median: 327.0747
```

```
## st.dev.: 151.0208
```



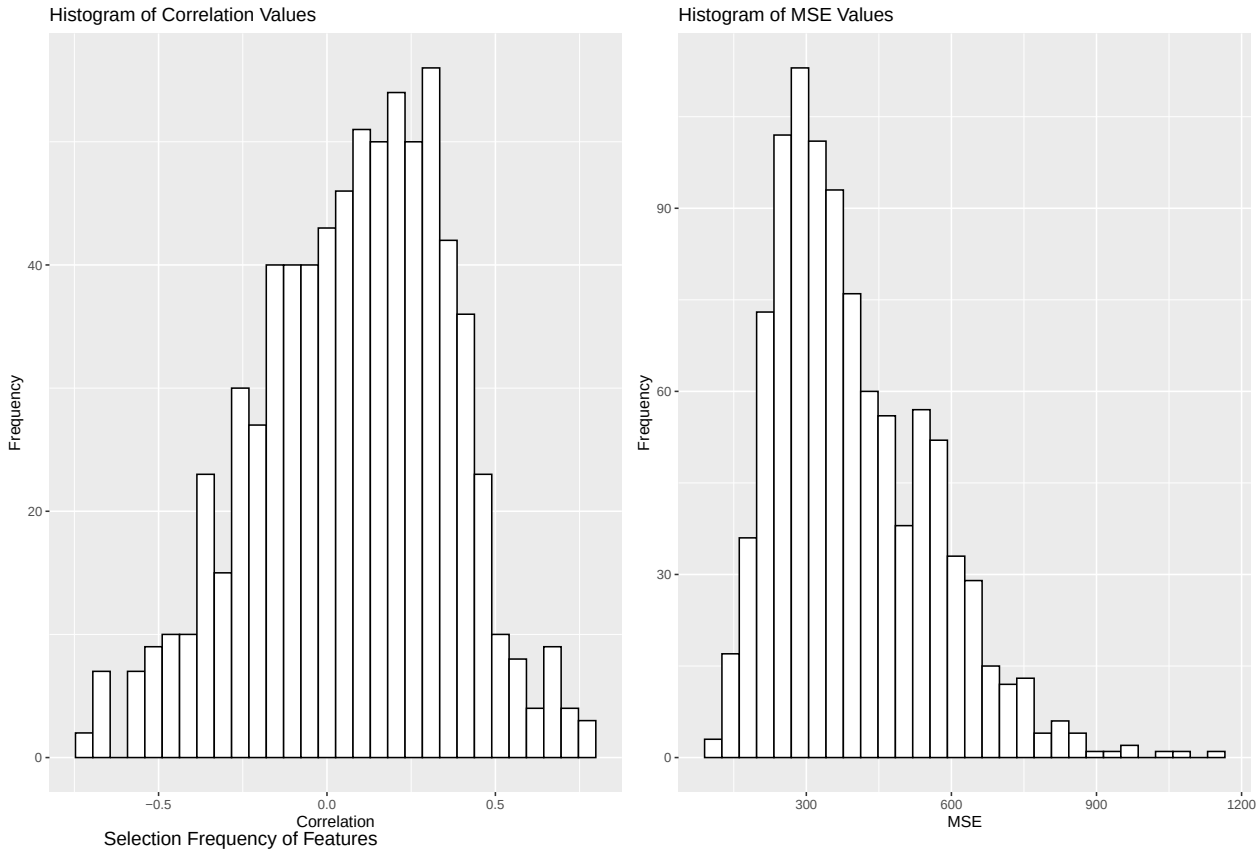
Ridge: 771 genes -> ROR-proliferation score + interactions between PCs

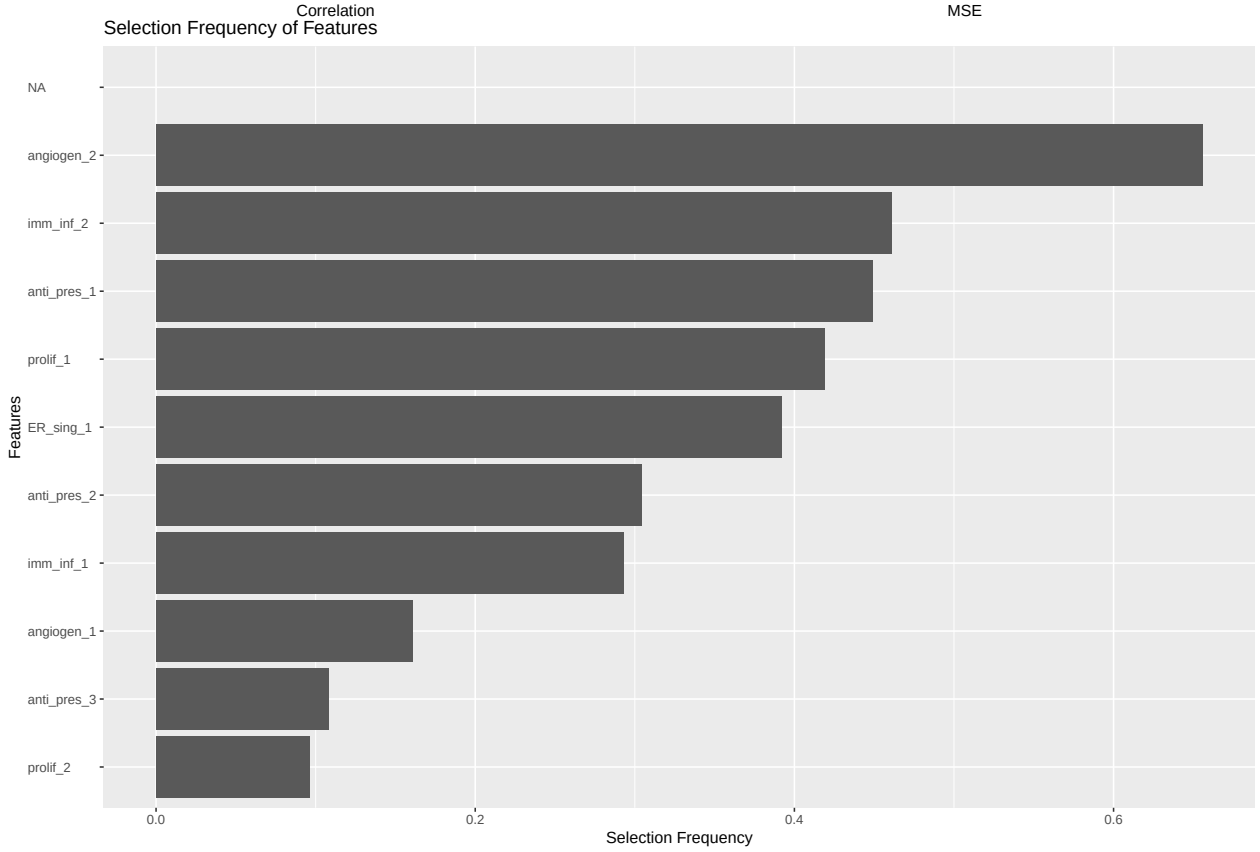
```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0
##
## CORRELATIONS RESULTS
## Mean: 0.4505837
## Median: 0.463892
## st.dev.: 0.2404457
##
## MSE RESULTS
## Mean: 333.3975
## Median: 291.1335
## st.dev.: 151.6577
```



Lasso: 771 genes -> ROR-proliferation score

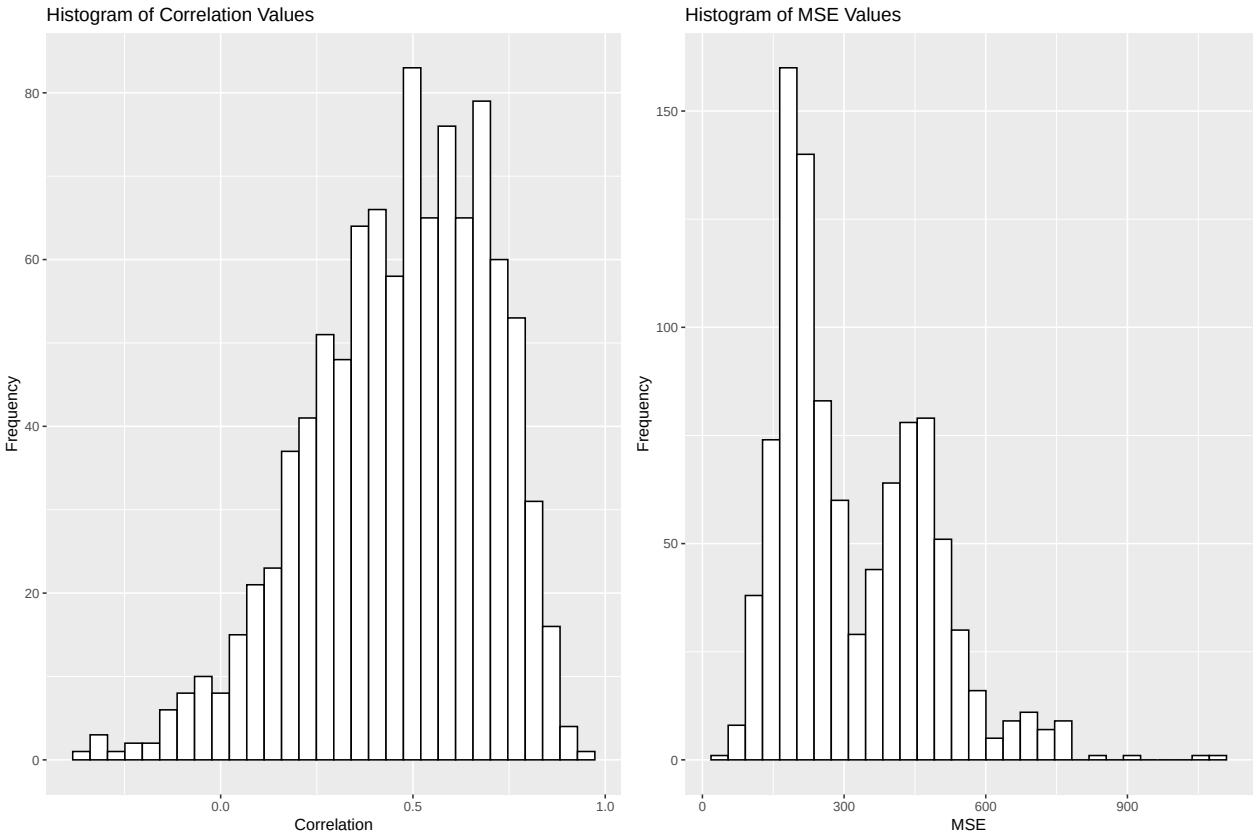
```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0.251
##
## CORRELATIONS RESULTS
## Mean: 0.07849613
## Median: 0.1028317
## st.dev.: 0.2856382
##
## MSE RESULTS
## Mean: 396.4902
## Median: 361.1201
## st.dev.: 160.1303
##
## Features selected 50% or more times:
## 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
```

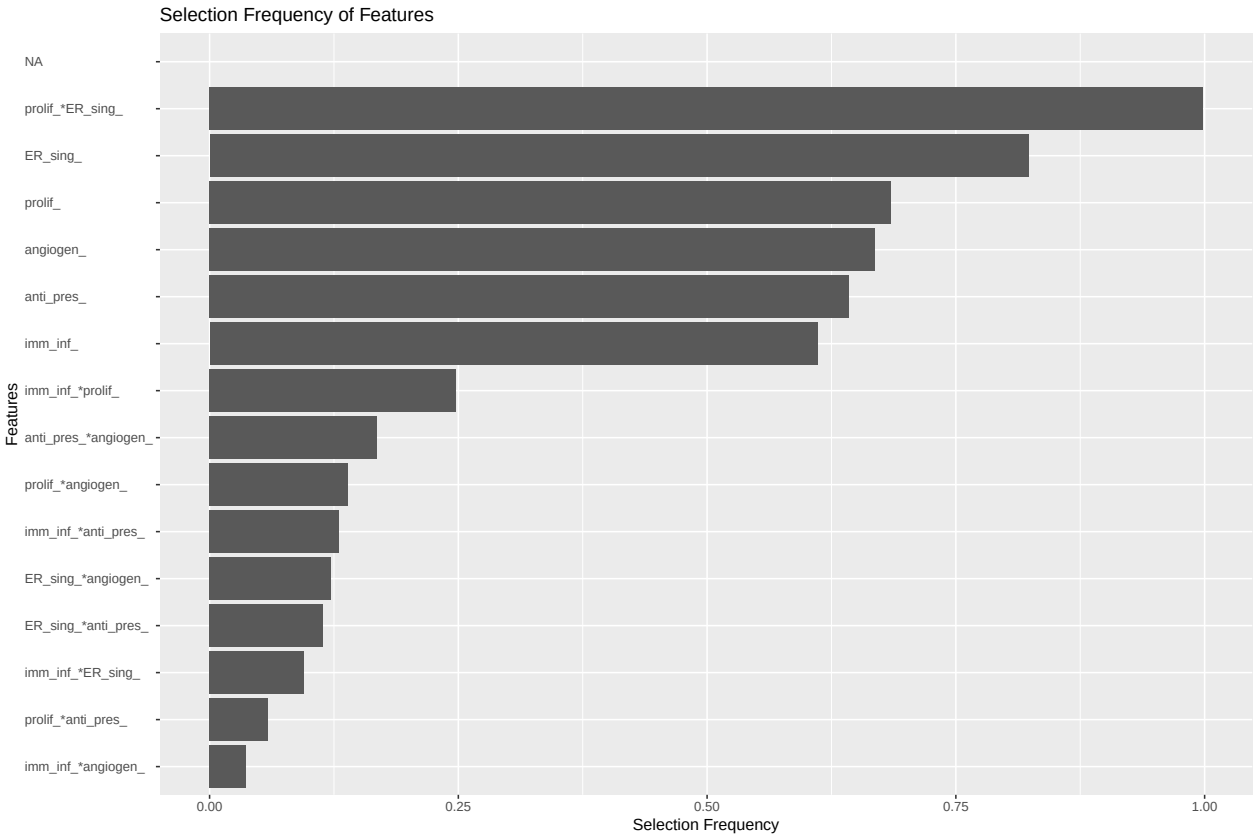

Histogram of Correlation Values
Frequency
Correlation
Histogram of MSE Values
Frequency
MSE


Selection Frequency of 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
Features
Selection Frequency

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

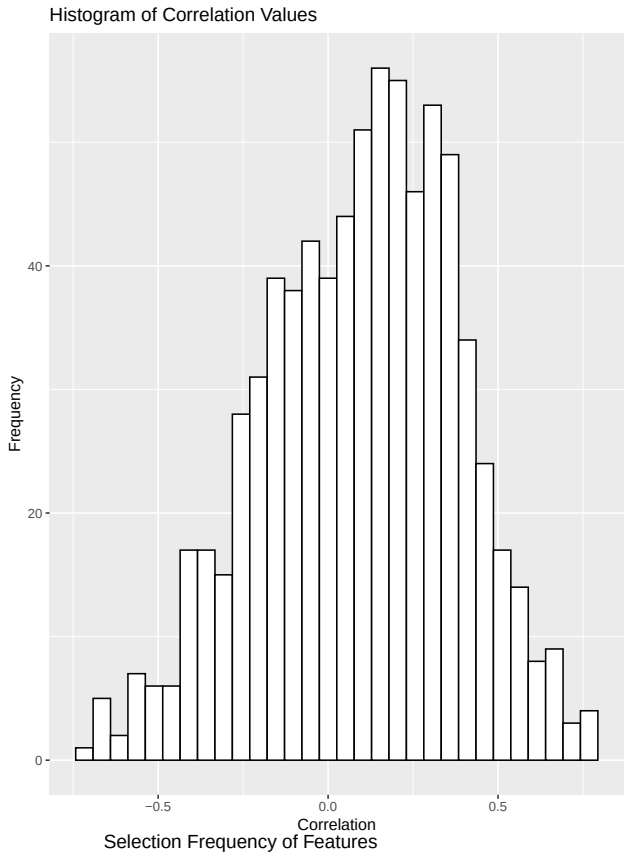
```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0.002
##
## CORRELATIONS RESULTS
## Mean: 0.4687056
## Median: 0.4972697
## st.dev.: 0.2329426
##
## MSE RESULTS
## Mean: 321.0086
## Median: 265.9525
## st.dev.: 157.5895
##
## Features selected 50% or more times:
## 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
```

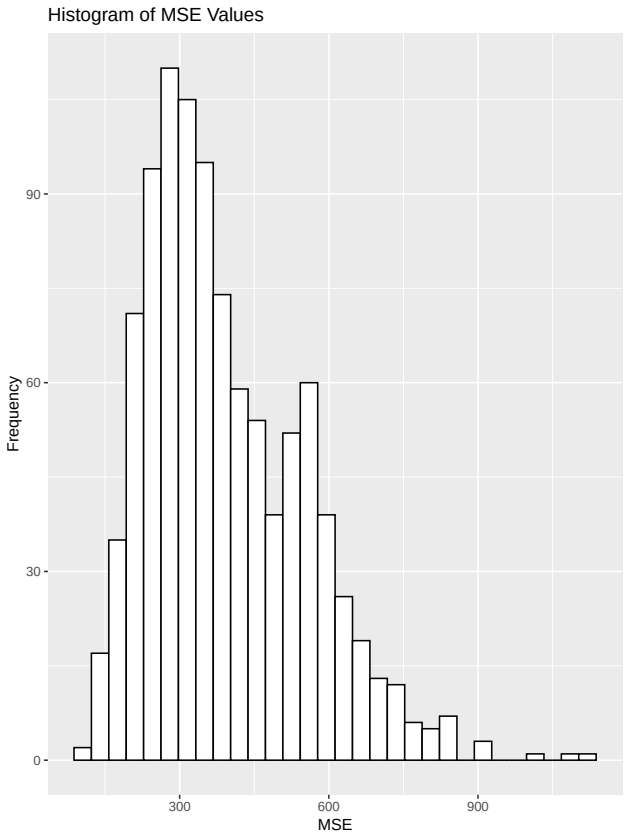


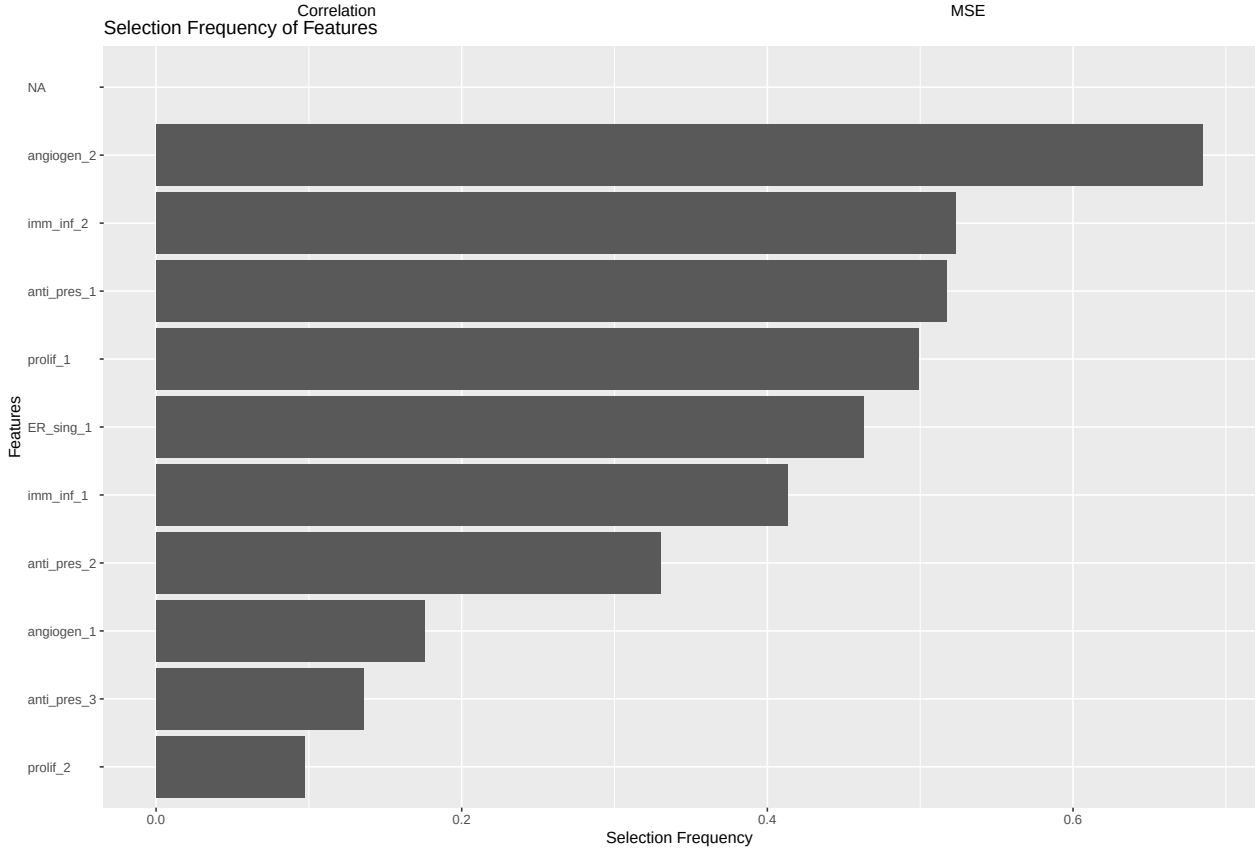


ElasticNet: 771 genes -> ROR-proliferation score

```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0.24
##
## CORRELATIONS RESULTS
## Mean: 0.09606929
## Median: 0.1167623
## st.dev.: 0.2844609
##
## MSE RESULTS
## Mean: 392.3862
## Median: 358.5637
## st.dev.: 158.0853
##
## 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
```

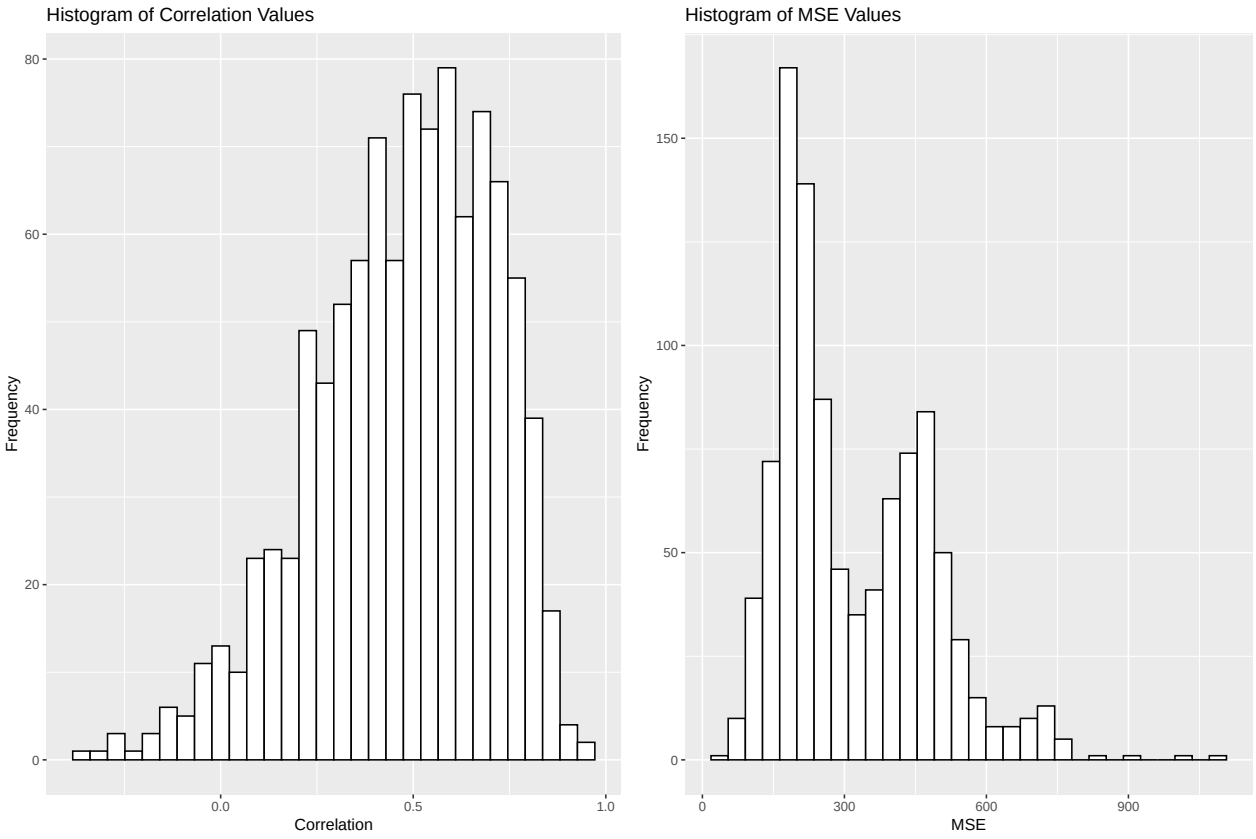

Histogram of Correlation Values
Frequency
Correlation
0
20
40
−0.5
0.0
0.5

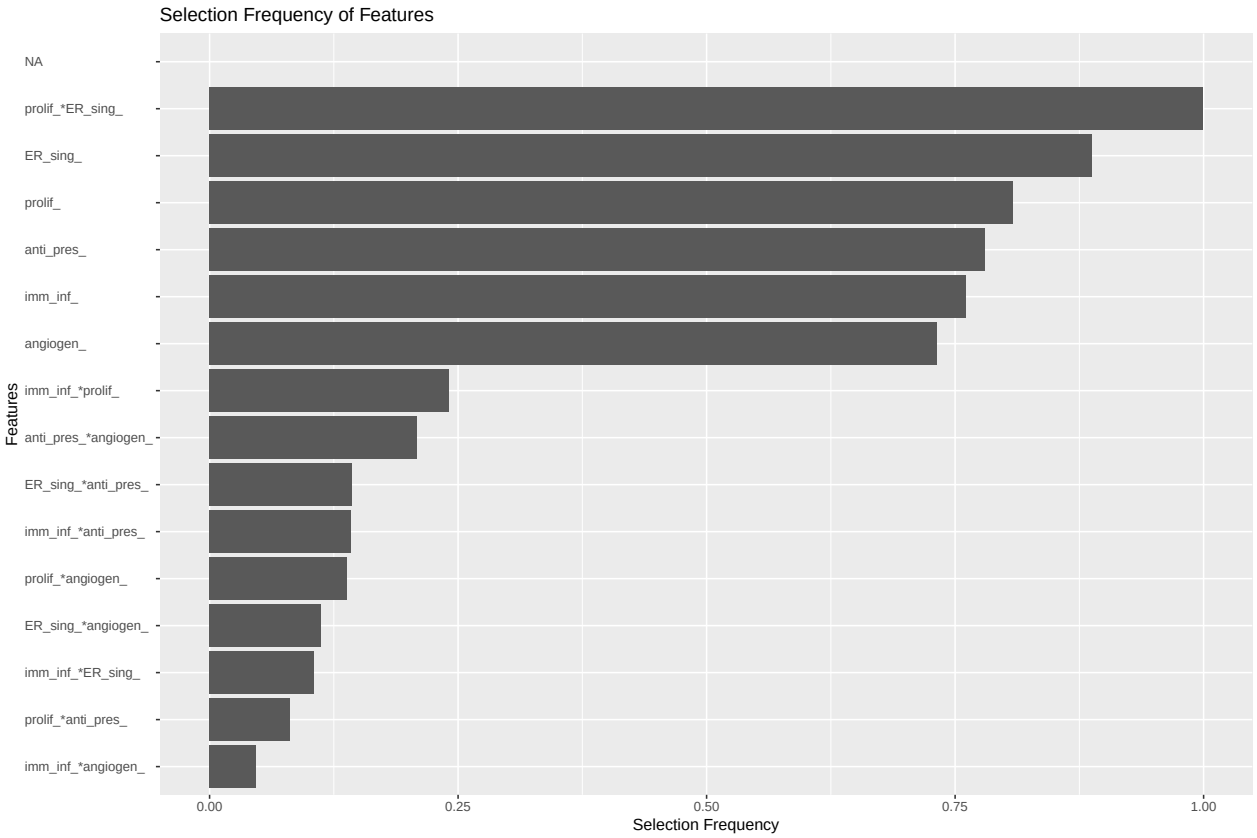

Histogram of MSE Values
Frequency
MSE
0
30
60
90
300
600
900


Selection Frequency of Features
Features
Selection Frequency
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

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
## st.dev.: 0.2335133
##
## MSE RESULTS
## Mean: 319.7269
## Median: 263.5692
## st.dev.: 157.8497
##
## 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: PCA ROR+proliferation score (repeated cross-validation)

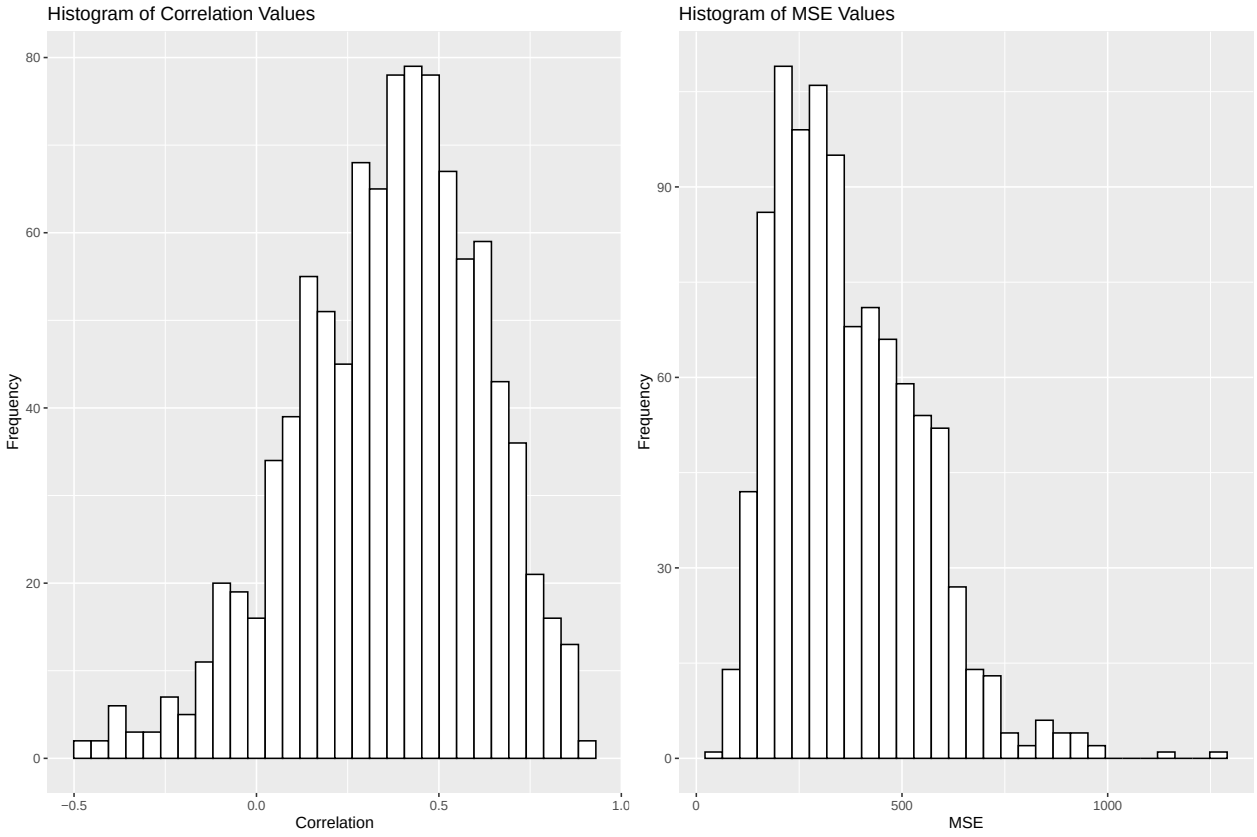
Model	cor_mean	sd_cor	MSE_mean	MSE_sd
ridge	0.3028091	0.2779432	364.5733	151.0208
ridge_interac	0.4505837	0.2404457	333.3975	151.6577
lasso	0.0784961	0.2856382	396.4902	160.1303
lasso_interact	0.4687056	0.2329426	321.0086	157.5895
elastic	0.0960693	0.2844609	392.3862	158.0853
elastic_interact	0.4749342	0.2335133	319.7269	157.8497

Stacking on signature gene sets using repeated cross-validation

Ridge: 771 genes -> ROR-proliferation score

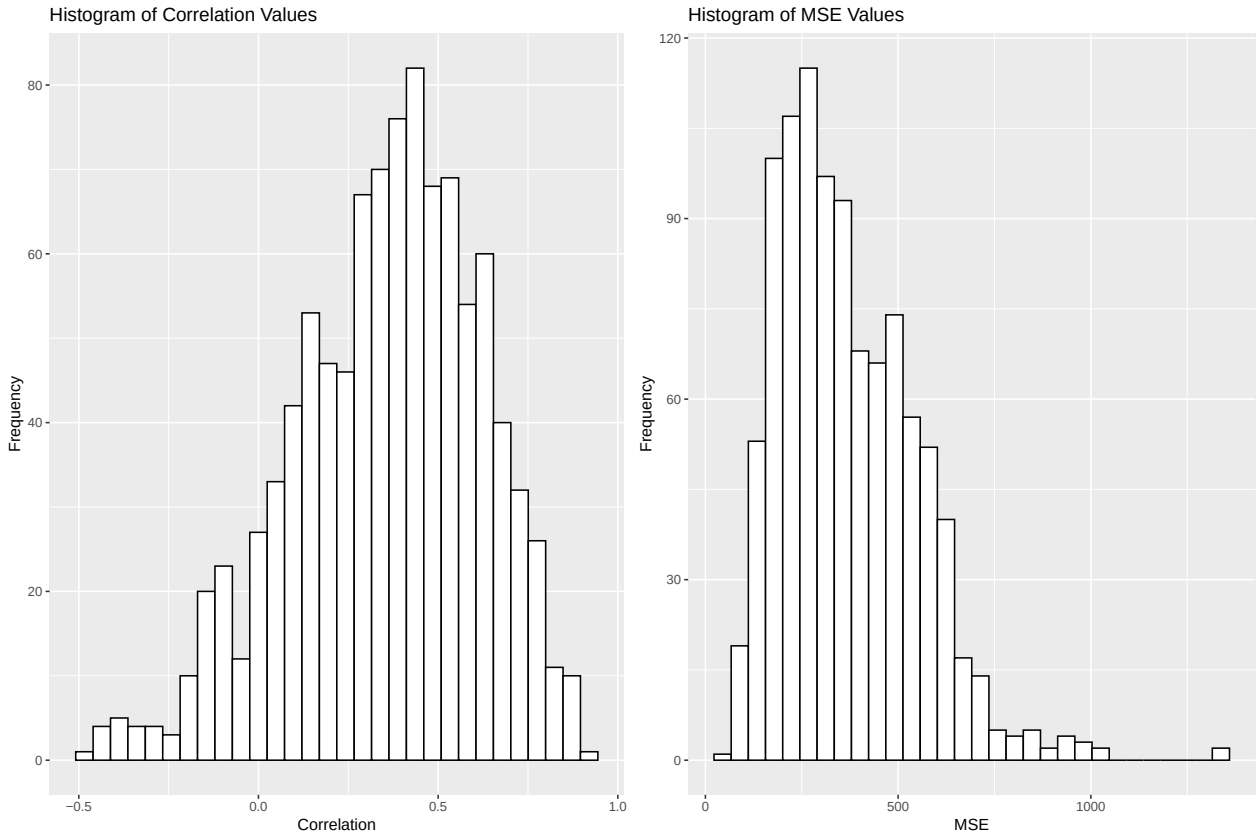
```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0
##
## CORRELATIONS RESULTS
## Mean: 0.363015
## Median: 0.3907333
## st.dev.: 0.2544613
##
## MSE RESULTS
## Mean: 367.3704
## Median: 335.367
```

```
## st.dev.: 171.8509
```



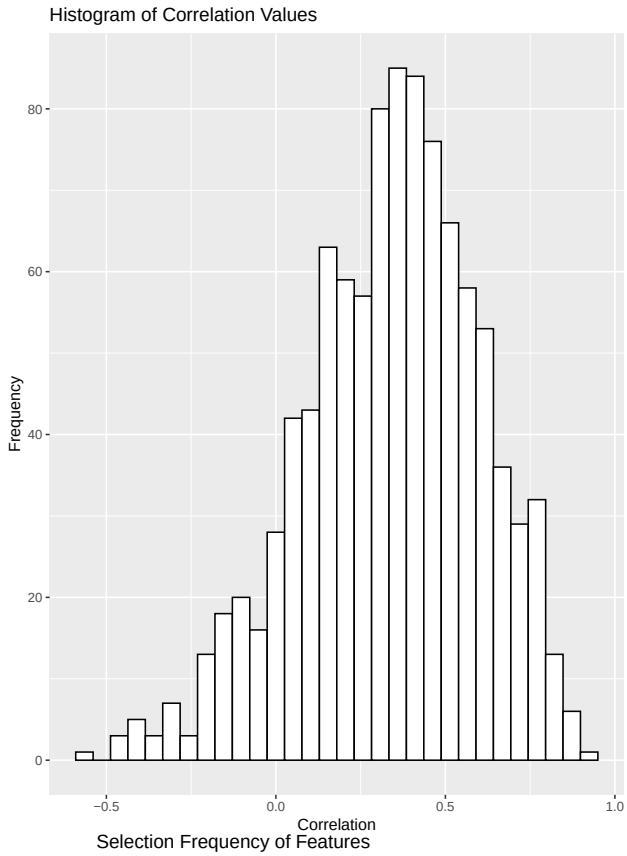
Ridge: 771 genes -> ROR-proliferation score + interactions between PCs

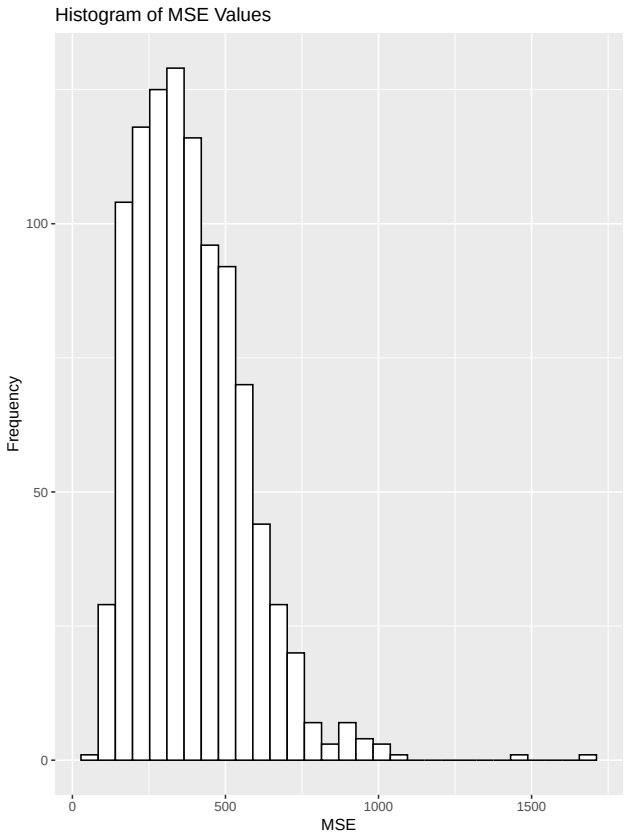
```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0
##
## CORRELATIONS RESULTS
## Mean: 0.3555283
## Median: 0.3864856
## st.dev.: 0.2622557
##
## MSE RESULTS
## Mean: 369.2144
## Median: 337.1152
## st.dev.: 176.9371
```

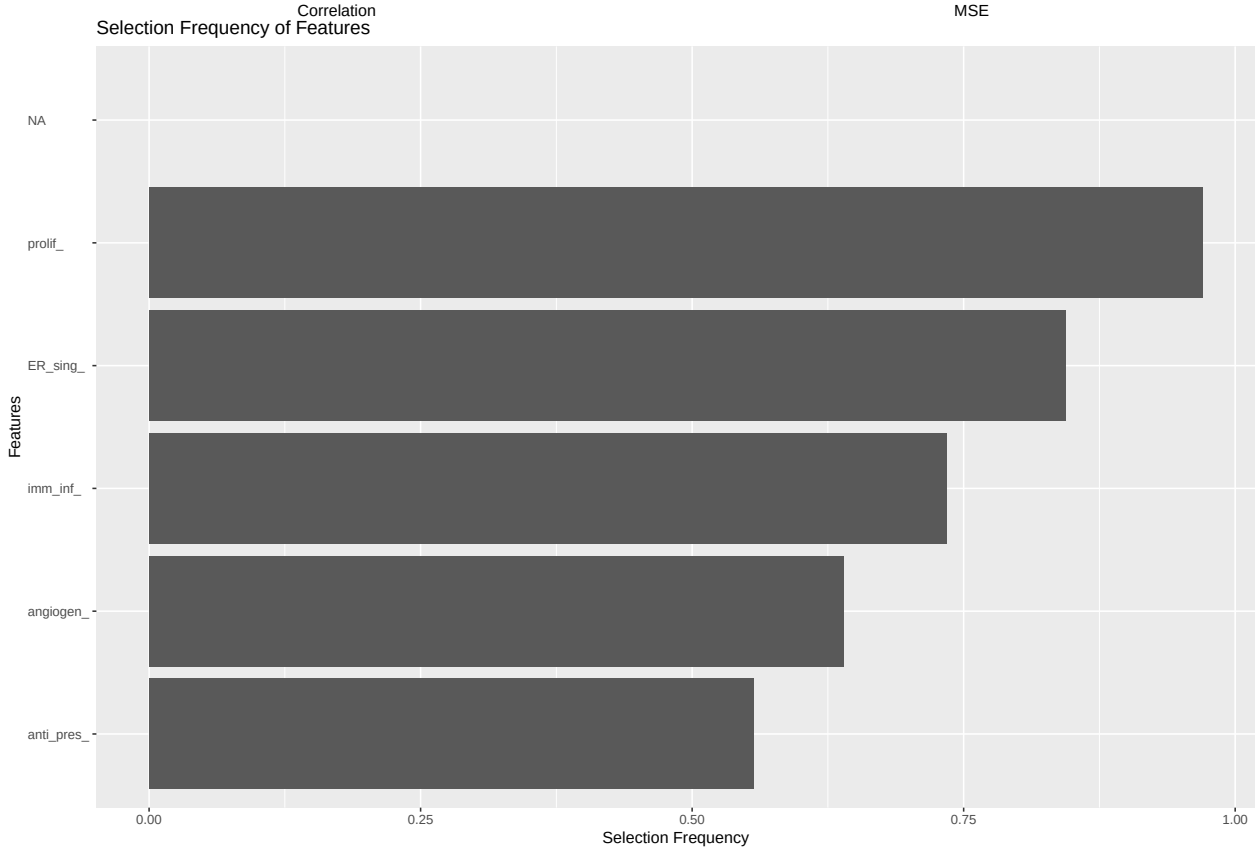


Lasso: 771 genes -> ROR-proliferation score

```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0
##
## CORRELATIONS RESULTS
## Mean: 0.3373612
## Median: 0.3601618
## st.dev.: 0.2605003
##
## MSE RESULTS
## Mean: 386.6455
## Median: 360.5423
## st.dev.: 179.9151
##
## Features selected 50% or more times:
## imm_inf_ prolif_ ER_sing_ anti_pres_ angiogen_
##
## Top 20 featrues:
##  [1] "prolif_"    "ER_sing_"   "imm_inf_"   "angiogen_"  "anti_pres_"
##  [6] NA           NA           NA           NA           NA
## [11] NA           NA           NA           NA           NA
## [16] NA           NA           NA           NA           NA
```

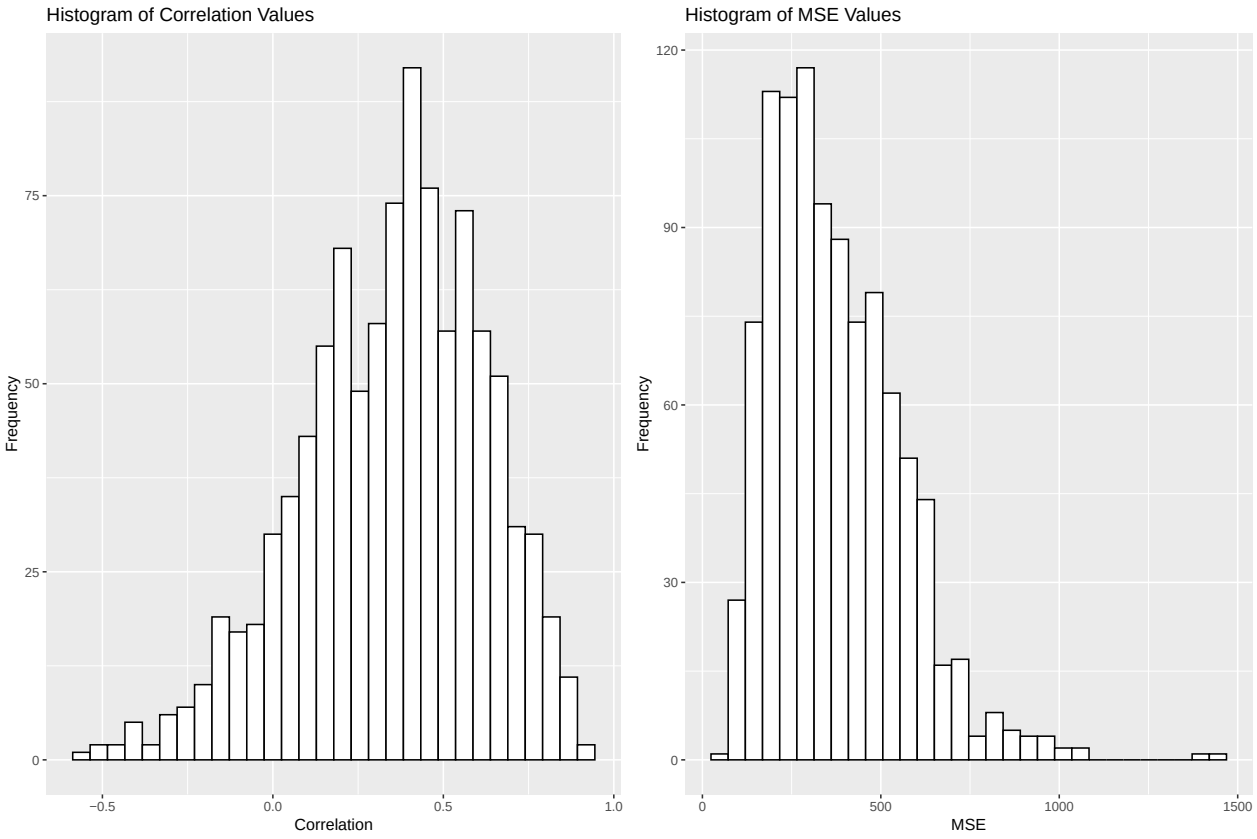

Histogram of Correlation Values
Frequency
Correlation
0
20
40
60
80
−0.5
0.0
0.5
1.0

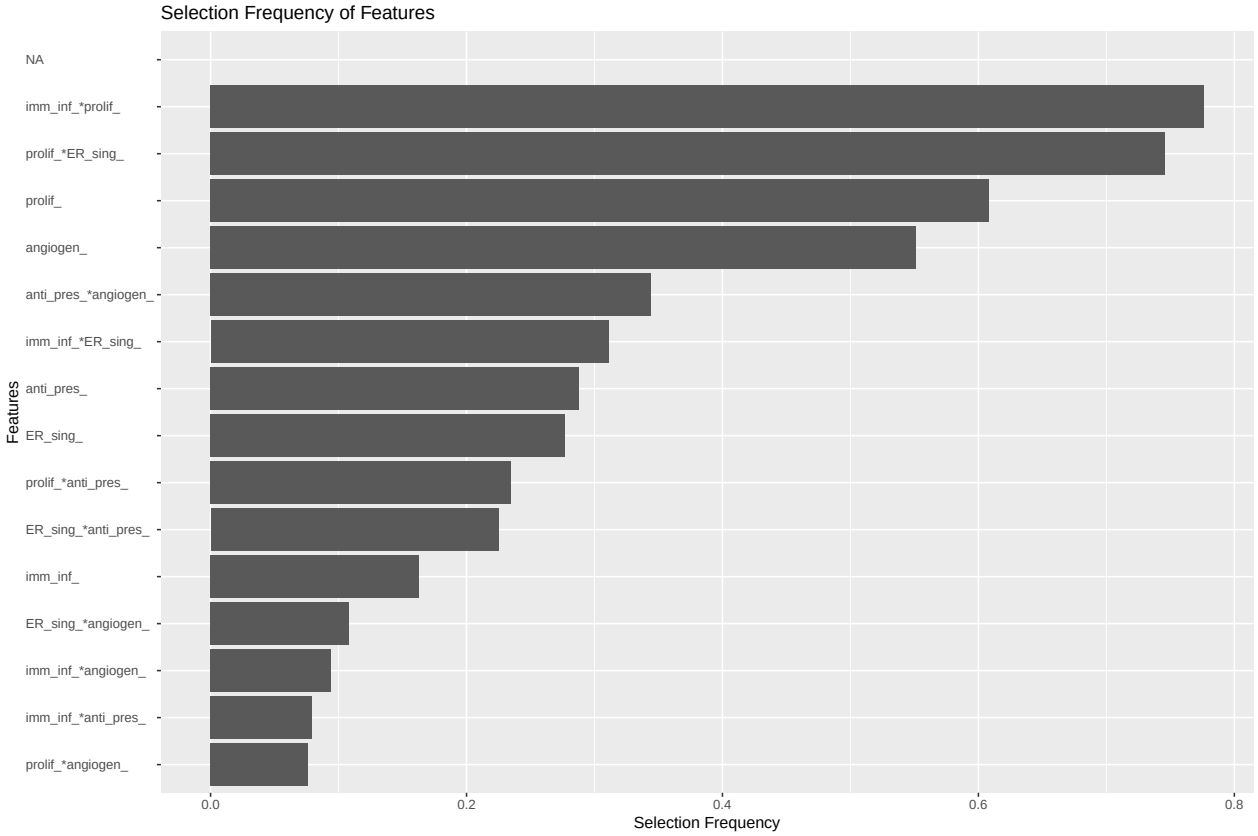

Histogram of MSE Values
Frequency
MSE
0
50
100
0
500
1000
1500


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

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

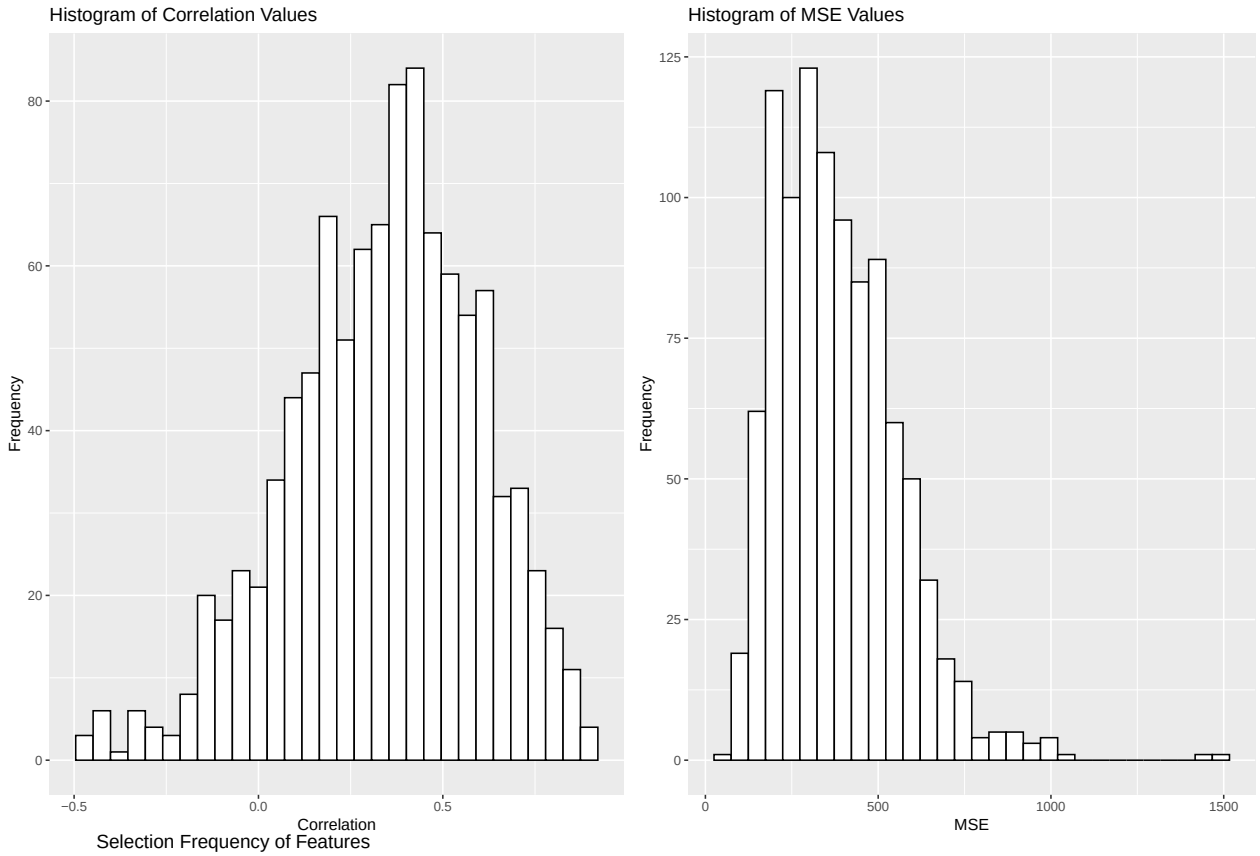
```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0
##
## CORRELATIONS RESULTS
## Mean: 0.3521323
## Median: 0.381805
## st.dev.: 0.2687827
##
## MSE RESULTS
## Mean: 373.2299
## Median: 341.6008
## st.dev.: 182.5783
##
## Features selected 50% or more times:
## prolif_ angiogen_ imm_inf_*prolif_ prolif_*ER_sing_
##
## Top 20 featrues:
##  [1] "imm_inf_*prolif_"     "prolif_*ER_sing_"     "prolif_"
##  [4] "angiogen_"            "anti_pres_*angiogen_" "imm_inf_*ER_sing_"
##  [7] "anti_pres_"           "ER_sing_"             "prolif_*anti_pres_"
## [10] "ER_sing_*anti_pres_"  "imm_inf_"             "ER_sing_*angiogen_"
## [13] "imm_inf_*angiogen_"   "imm_inf_*anti_pres_"  "prolif_*angiogen_"
## [16] NA                     NA                     NA
## [19] NA                     NA
```

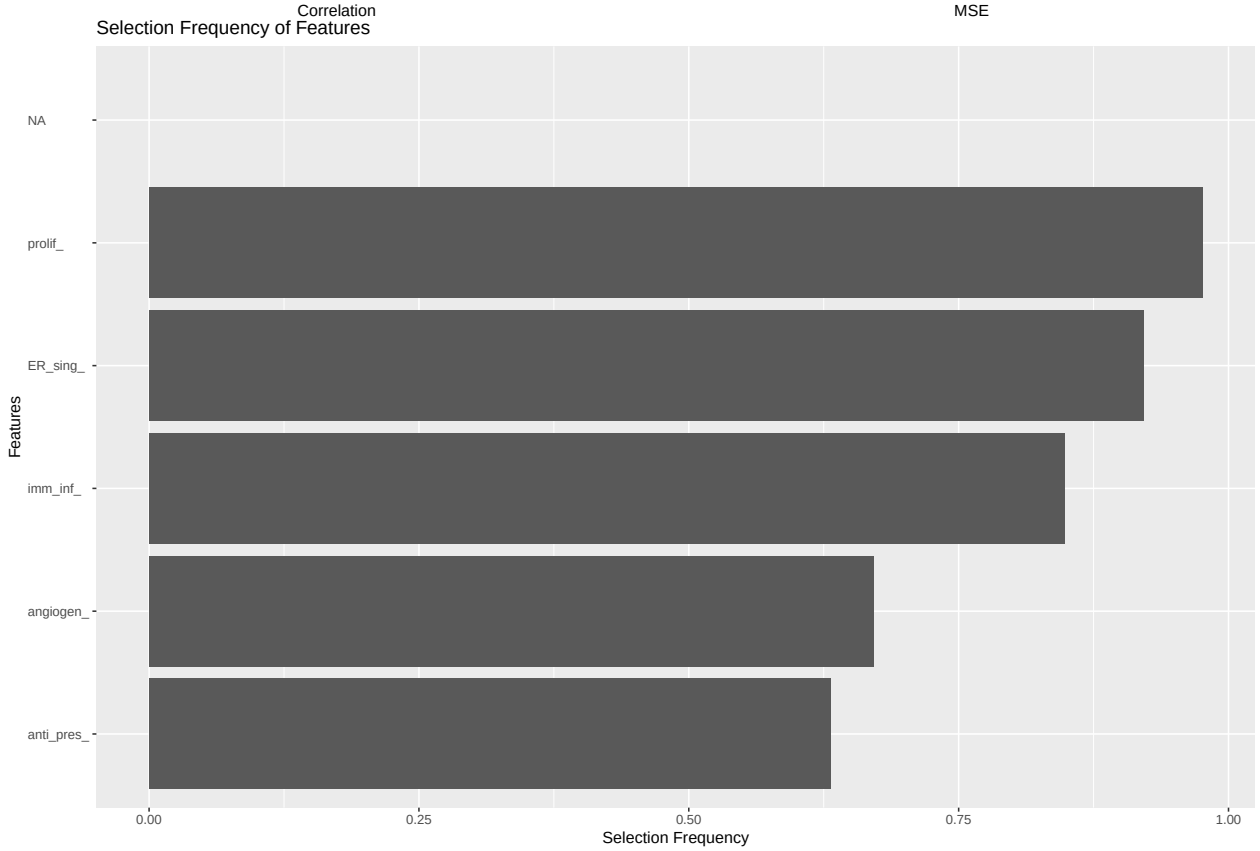




ElasticNet: 771 genes -> ROR-proliferation score

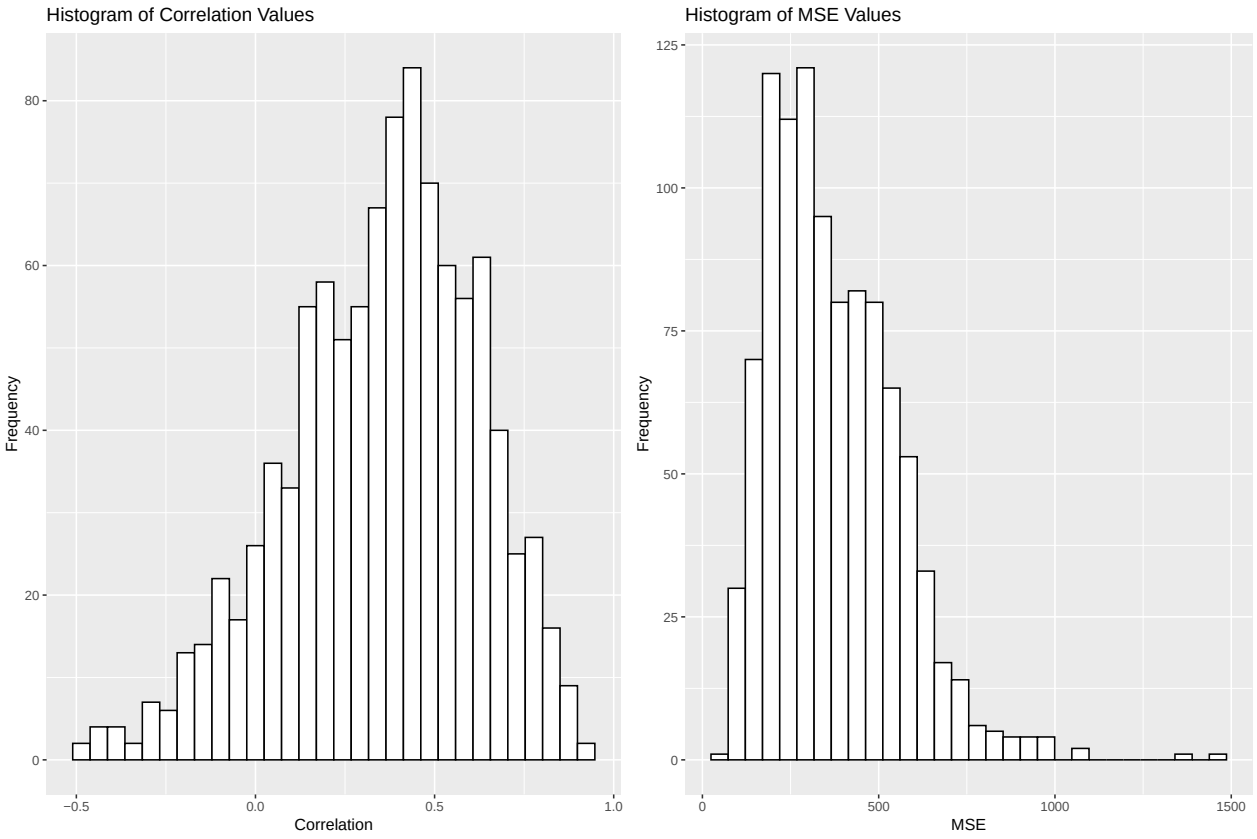
```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0
##
## CORRELATIONS RESULTS
## Mean: 0.3417178
## Median: 0.3669518
## st.dev.: 0.2583745
##
## MSE RESULTS
## Mean: 383.7542
## Median: 356.8975
## st.dev.: 177.3923
##
## Features selected 50% or more times:
## imm_inf_ prolif_ ER_sing_ anti_pres_ angiogen_
##
## Top 20 featrues:
##  [1] "prolif_"    "ER_sing_"   "imm_inf_"   "angiogen_"  "anti_pres_"
##  [6] NA           NA           NA           NA           NA
## [11] NA           NA           NA           NA           NA
## [16] NA           NA           NA           NA           NA
```

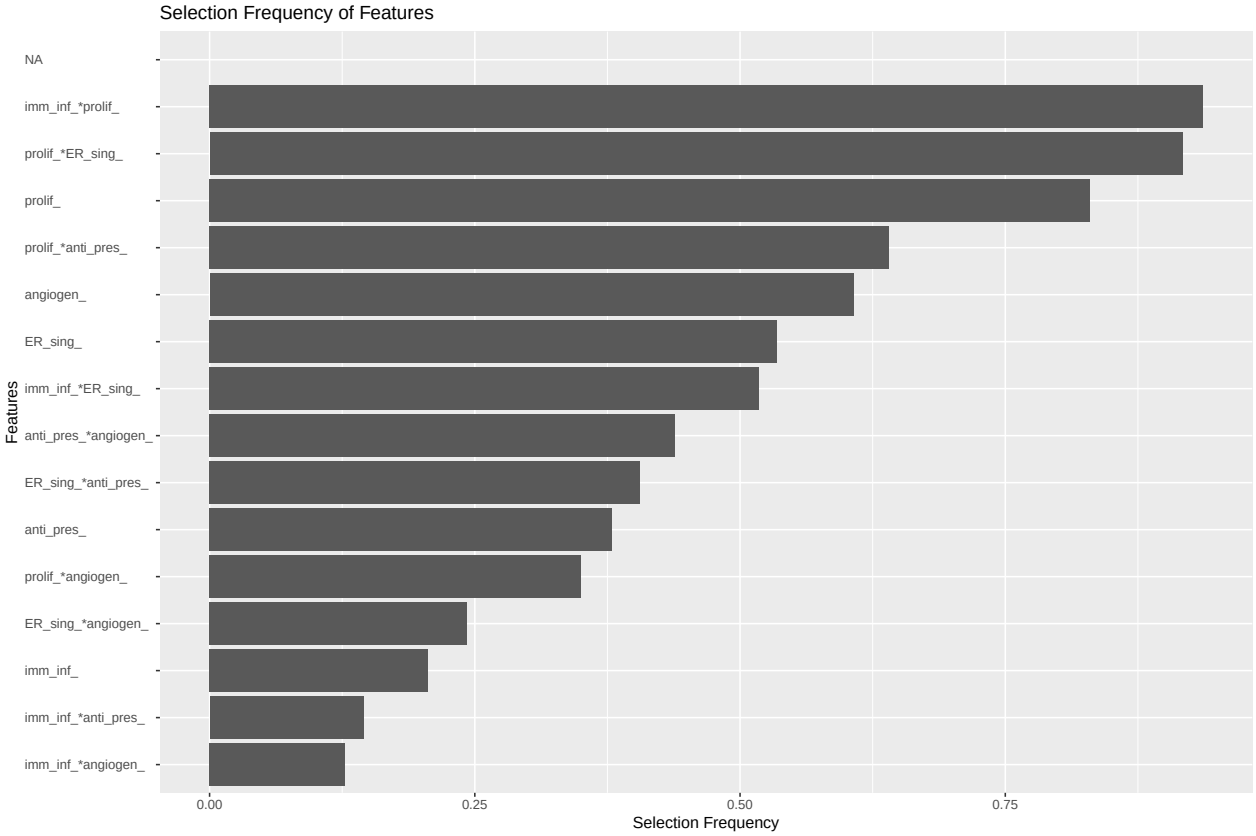

Histogram of Correlation Values
Frequency
Correlation
Histogram of MSE Values
Frequency
MSE


Selection Frequency of Features
Features
NA
prolif_
ER_sing_
imm_inf_
angiogen_
anti_pres_
Selection Frequency

ElasticNet: 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: 0.353272
## Median: 0.3797595
## st.dev.: 0.2654461
##
## MSE RESULTS
## Mean: 371.1879
## Median: 339.6459
## st.dev.: 179.5494
##
## Features selected 50% or more times:
## prolif_ ER_sing_ angiogen_ imm_inf_*prolif_ imm_inf_*ER_sing_ prolif_*ER_sing_ prolif_*anti_pres_
##
## Top 20 featrues:
##  [1] "imm_inf_*prolif_"     "prolif_*ER_sing_"     "prolif_"
##  [4] "prolif_*anti_pres_"   "angiogen_"            "ER_sing_"
##  [7] "imm_inf_*ER_sing_"    "anti_pres_*angiogen_" "ER_sing_*anti_pres_"
## [10] "anti_pres_"           "prolif_*angiogen_"    "ER_sing_*angiogen_"
## [13] "imm_inf_"             "imm_inf_*anti_pres_"  "imm_inf_*angiogen_"
## [16] NA                     NA                     NA
## [19] NA                     NA
```





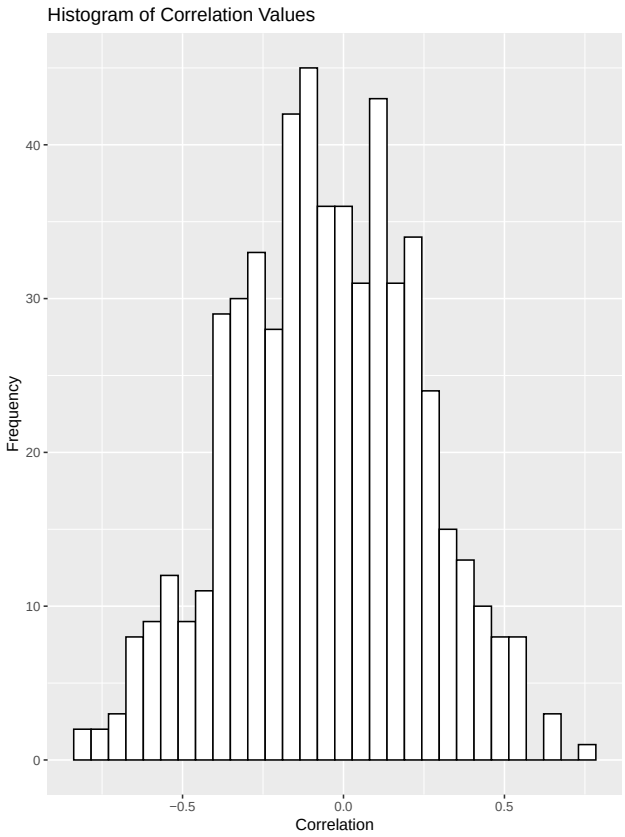
Summery results: Stacking ROR+proliferation score (repeated cross-validation)

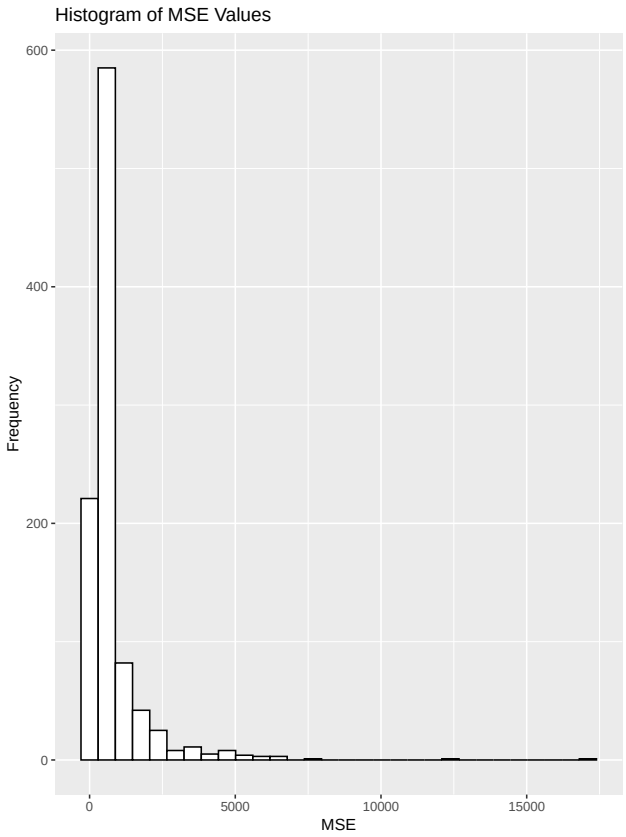
Model	cor_mean	sd_cor	MSE_mean	MSE_sd
ridge	0.3630150	0.2544613	367.3704	171.8509
ridge_interac	0.3555283	0.2622557	369.2144	176.9371
lasso	0.3373612	0.2605003	386.6455	179.9151
lasso_interact	0.3521323	0.2687827	373.2299	182.5783
elastic	0.3417178	0.2583745	383.7542	177.3923
elastic_interact	0.3532720	0.2654461	371.1879	179.5494

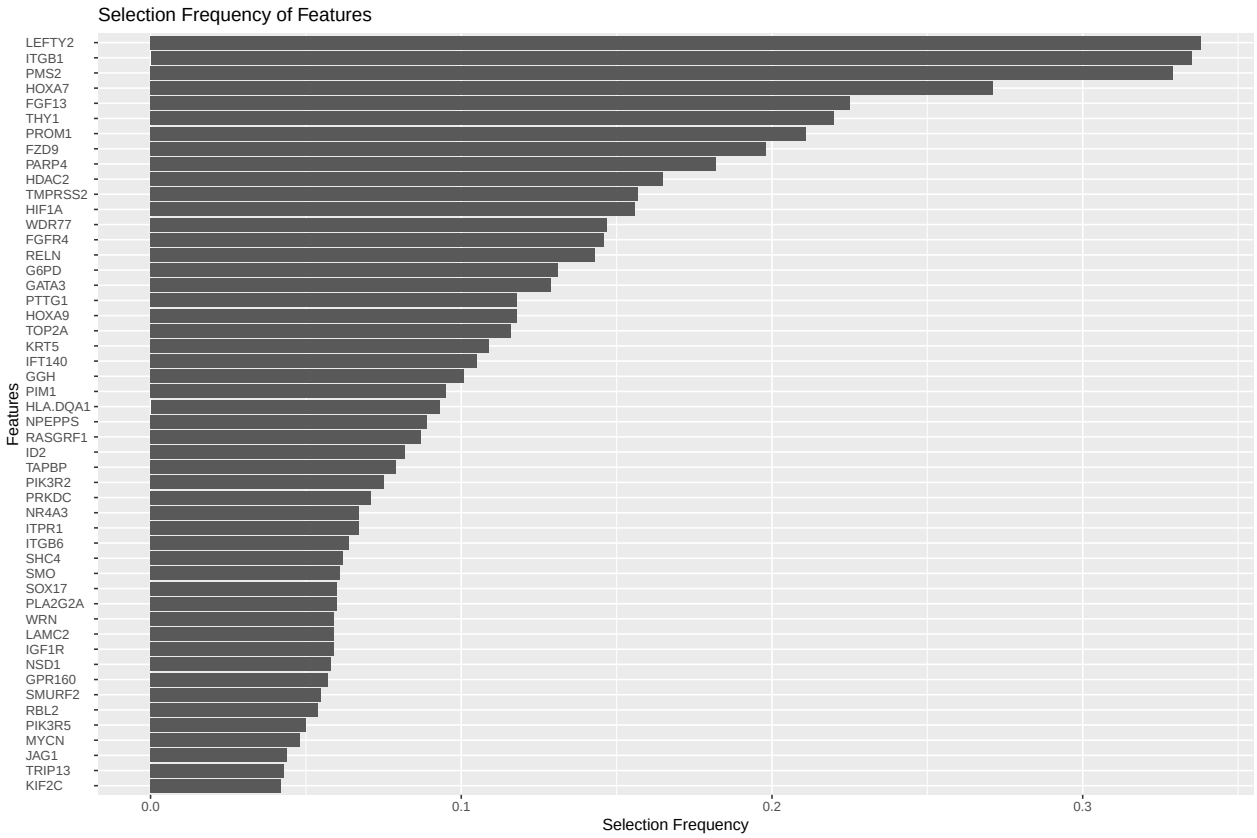
Sparse Group Lasso: 771 genes -> ROR-proliferation score + interactions between PCs

```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0.444
##
## CORRELATIONS RESULTS
## Mean: -0.05441624
## Median: -0.05224598
## st.dev.: 0.2856222
##
## MSE RESULTS
## Mean: 772.1238
## Median: 428.2544
## st.dev.: 1123.466
##
## Features selected 50% or more times:
```

```
## Non selected that many times
##
## Top 20 featrues:
##  [1] "LEFTY2"  "ITGB1"   "PMS2"    "HOXA7"   "FGF13"   "THY1"    "PROM1"
##  [8] "FZD9"    "PARP4"   "HDAC2"   "TMPRSS2" "HIF1A"   "WDR77"   "FGFR4"
## [15] "RELN"    "G6PD"    "GATA3"   "HOXA9"   "PTTG1"   "TOP2A"
```







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	0.350	0.063	0.149	0.012
lasso 771 genes	0.794	0.090	0.062	0.024
Nodes	0.414	0.064	0.148	0.192
Residual additive	0.784	0.085	0.064	0.021
Residual multiplicative	0.727	0.090	0.081	0.023

Summery results: lasso ROR+proliferation score (bootstrap)

Model	cor_mean	sd_cor	MSE_mean	MSE_sd
lasso 6 genes	0.128	0.053	378.094	23.440
lasso 771 genes	0.697	0.100	203.408	61.349
Nodes	0.296	0.083	417.667	1027.062
Residual additive	0.693	0.109	202.323	61.042
Residual multiplicative	0.543	0.191	291.019	82.792

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

Model	cor_mean	sd_cor	MSE_mean	MSE_sd
lasso 6 genes	0.092	0.307	0.166	0.072
lasso 771 genes	0.474	0.231	0.062	0.075
Nodes	0.284	0.277	0.156	0.076
Residual additive	0.463	0.223	0.133	0.065
Residual multiplicative	0.403	0.230	0.146	0.068

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

Model	cor_mean	sd_cor	MSE_mean	MSE_sd
lasso 6 genes	-0.482	0.172	374.152	144.156
lasso 771 genes	0.081	0.277	393.807	159.645
Nodes	0.181	0.285	380.116	164.587
Residual additive	0.164	0.279	392.544	158.873
Residual multiplicative	-0.215	0.251	568.806	208.591

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

Model	cor_mean	sd_cor	MSE_mean	MSE_sd
prolif boot	0.819	0.071	0.057	0.019
ROR-prolif boot	0.776	0.077	156.065	44.182
prolif rep cross-val	0.527	0.207	0.126	0.070
ROR-prolif rep cross-val	0.329	0.262	357.510	150.065

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

Model	cor_mean	sd_cor	MSE_mean	MSE_sd
prolif boot	0.813	0.084	0.056	0.022
ROR-prolif boot	0.757	0.089	164.412	52.775
prolif rep cross-val	0.427	0.264	0.150	0.082
ROR-prolif rep cross-val	0.205	0.311	427.352	186.246

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

Model	cor_mean	sd_cor	MSE_mean	MSE_sd
prolif boot	0.776	0.083	0.065	0.021
ROR-prolif boot	0.753	0.088	165.145	51.866
prolif rep cross-val	0.236	0.280	0.171	0.076
ROR-prolif rep cross-val	0.174	0.315	394.263	171.644

Summary using domain knowledge

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

Model	cor_mean	sd_cor	MSE_mean	MSE_sd
ridge	0.303	0.278	364.573	151.021
ridge_interac	0.451	0.240	333.397	151.658
lasso	0.078	0.286	396.490	160.130
lasso_interact	0.469	0.233	321.009	157.590
elastic	0.096	0.284	392.386	158.085
elastic_interact	0.475	0.234	319.727	157.850

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

Model	cor_mean	sd_cor	MSE_mean	MSE_sd
ridge	0.363	0.254	367.370	171.851
ridge_interac	0.356	0.262	369.214	176.937
lasso	0.337	0.261	386.646	179.915
lasso_interact	0.352	0.269	373.230	182.578
elastic	0.342	0.258	383.754	177.392
elastic_interact	0.353	0.265	371.188	179.549

Summery most interesting maybe

Response is just ROR+proliferation score, and only used repeated cross-validation.

Model	cor_mean	sd_cor	MSE_mean	MSE_sd
Lasso	0.081	0.277	393.807	159.645
Ridge	0.329	0.262	357.51	150.065
ElasticNet	0.205	0.311	427.352	186.246
Boosting	0.174	0.315	394.263	171.644
Residual (lasso/additive)	0.164	0.279	392.544	158.873
Sparse group lasso	-0.054	0.286	772.124	1123.466
PCA ON GENE SETS				
ridge	0.303	0.278	364.573	151.021
ridge_interac	0.451	0.24	333.397	151.658
lasso	0.078	0.286	396.49	160.13
lasso_interact	0.469	0.233	321.009	157.59
elastic	0.096	0.284	392.386	158.085
elastic_interact	0.475	0.234	319.727	157.85
STACKING ON GENE SETS				
ridge	0.363	0.254	367.37	171.851
ridge_interac	0.356	0.262	369.214	176.937
lasso	0.337	0.261	386.646	179.915
lasso_interact	0.352	0.269	373.23	182.578
elastic	0.342	0.258	383.754	177.392
elastic_interact	0.353	0.265	371.188	179.549