Analysis of Pixel-Wise Correlations

So far we've only looked at the mean stain levels between different patients. However, this ignores any spatial processes that might play a role. In order to start gaining a first insight into what spatial processes might play a role we'll here analyse the pixel-wise correlation matrices for the cores from different patients. This will for example highlight the presence/absence of specific cell types/meta-phenotypes.

1 A First Look at the Data

I compute the correlation matrices using python and save the lower triangular parts of these matrices to file. In order to adjust for the different scales of the stains I calculate the standardised correlations using the `np.corrcoef()` function.

Let's load in the results and label the columns with the correlation they measure:

```
corrArr = read.csv("pixelcorrelations.csv",header=F)
dim(corrArr)

## [1] 121 668

# Label the columns
labelArr = c("CoreId","PtSnty")
markerLabelsVec = c('SrBCK', 'RR101', 'RR102', 'AvantiLipid', 'XeBCK', 'CD196', 'CD19', 'Vimentin',
                    'CD163', 'CD20', 'CD16', 'CD25', 'p53', 'CD134', 'CD45', 'CD44s', 'CD14', 'FoxP3
                    'CD4', 'E-cadherin', 'p21', 'CD152', 'CD8a', 'CD11b', 'Beta-catenin', 'B7-H4', '
                    'CollagenI', 'CD3', 'CD68', 'PD-L2', 'B7-H3', 'HLA-DR', 'pS6', 'HistoneH3', 'DNA
                    'DNA193')
for (i in seq(2,37)) {
  for (j in seq(i-1)) {
    labelArr = c(labelArr,paste0(markerLabelsVec[i],".",markerLabelsVec[j]))
  }
}
names(corrArr) = labelArr
```

Let's plot the correlations for each patient to see if there's an obvious difference between responders and non-responders.

```
library(ggplot2)
library(reshape2)
corrArr = corrArr[with(corrArr, order(PtSnty)), ]
corrArr_idxd = data.frame(corrArr,LinId=seq(nrow(corrArr)))
corrArr_reshaped = melt(corrArr_idxd[,-1],id.vars=c("LinId"))
ggplot(corrArr_reshaped, aes(variable, LinId)) +
  geom_tile(aes(fill = value),colour="white") +
  scale_fill_gradient(low="white",high="steelblue") +
  theme_bw() +
  labs(x="",y="Core") +
  theme(axis.text.x = element_text(angle=90, hjust=1))
```

There doesn't seem anything obvious. Let's test for statistically significant differences.

2 A Logistic Regression Model

Let's use a logistic regression model to find if there are any significant differences in the correlations between responders and non-responders.

```
initModel = glm(PtSnty ~.,family=binomial(link='logit'),
                data=corrArr)

## Warning:  glm.fit:  algorithm did not converge
```

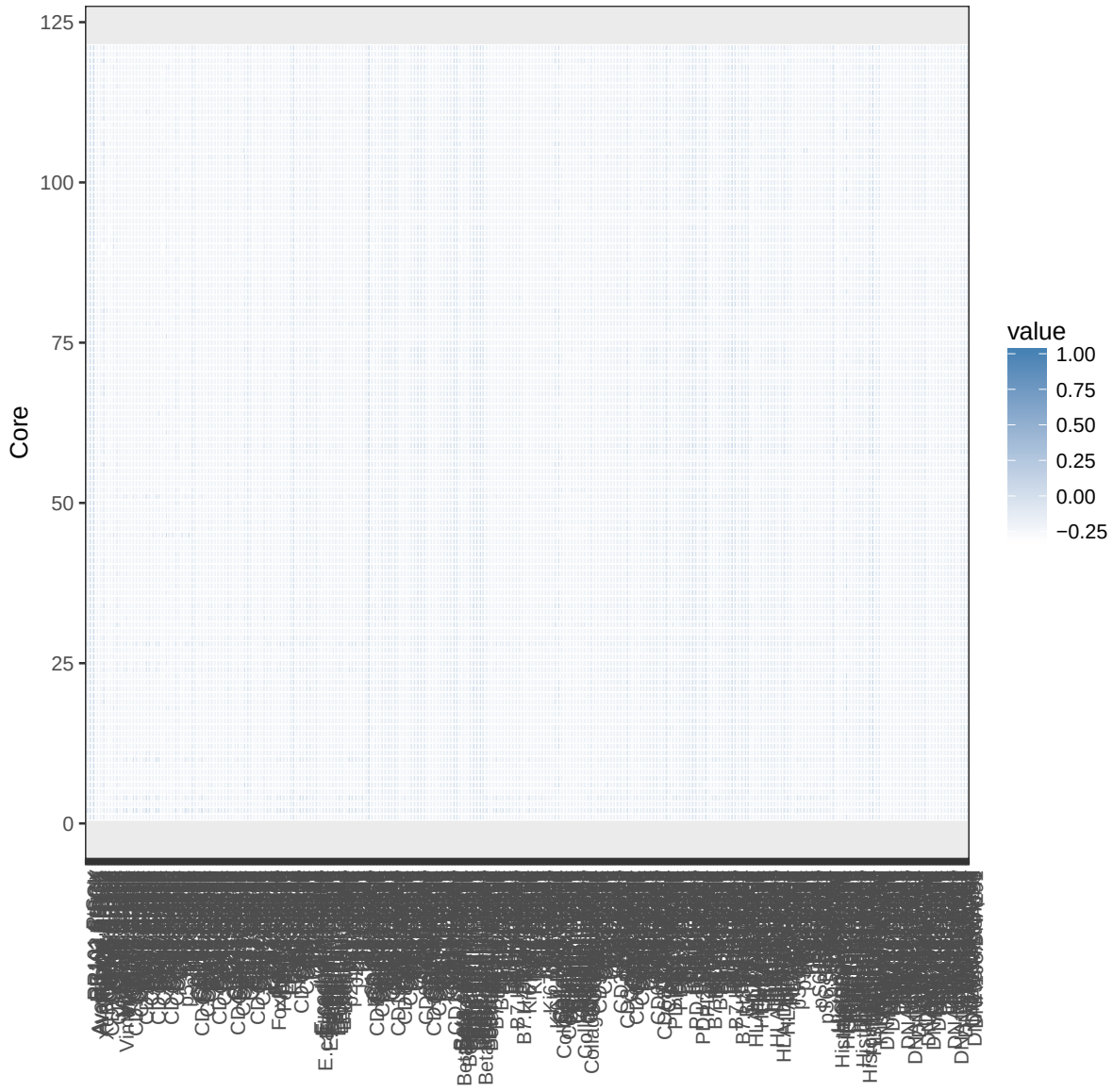


Figure 1: Pixel-wise correlation of the different stains for responders (top-half) and non-responders (bottom-half).

There seems to be a lot of co-linearity which prevents the model from being fitted. Maybe there are too many optima...

Let's try de-correlate the data. Since we can't compute VIFs, let's start by working with the correlation matrix. Remove any variables that are highly correlated with other variables. This stackexchange post (https://stackoverflow.com/questions/18275639/remove-highly-correlated-variables) suggests a caret function. Let's try it:

```
library(caret)

## Loading required package:  lattice

covariatesArr = corrArr[,-c(1,2)]
coMat = cor(covariatesArr)
hc = findCorrelation(coMat,cutoff=0.7,exact=TRUE) # put any value as a "cutoff"
hc = sort(hc)
corrArr_Reduced = data.frame(corrArr[,c(1,2)],covariatesArr[,-c(hc)])
dim(corrArr_Reduced)
```

```
## [1] 121  76
```

Let's try fitting a model again.

```
initModel = glm(PtSnty ~.,family=binomial(link='logit'),
                control = list(maxit = 100),
                data=corrArr_Reduced)

## Warning:  glm.fit:  fitted probabilities numerically 0 or 1 occurred

summary(initModel)

##
## Call:
## glm(formula = PtSnty ~ ., family = binomial(link = "logit"),
##     data = corrArr_Reduced, control = list(maxit = 100))
##
## Deviance Residuals:
##        Min          1Q      Median          3Q         Max
## -3.923e-06  -2.110e-08   2.110e-08   1.514e-06   4.288e-06
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)            1.344e+02  8.903e+06       0        1
## CoreId                -3.416e-01  1.159e+04       0        1
## AvantiLipid.RR102      1.636e+02  5.743e+06       0        1
## CD196.SrBCK           -3.671e+02  1.088e+08       0        1
## CD196.XeBCK           -2.319e+03  6.183e+07       0        1
## CD19.CD196             7.670e+01  9.137e+06       0        1
## Vimentin.XeBCK        -7.413e+03  6.014e+08       0        1
## Vimentin.CD196        -3.712e+02  8.639e+06       0        1
## Vimentin.CD19         -2.440e+02  3.443e+07       0        1
## CD163.XeBCK            3.459e+03  4.876e+08       0        1
## CD16.CD196             2.096e+01  6.055e+07       0        1
## CD25.XeBCK             1.950e+03  2.244e+08       0        1
## CD25.CD196             3.676e+02  8.925e+06       0        1
## CD25.CD19             -1.749e+02  5.415e+06       0        1
## p53.CD196             -5.312e+02  3.677e+07       0        1
## p53.CD20              -6.319e+02  1.307e+07       0        1
## CD45.CD25             -9.419e+02  2.634e+08       0        1
## CD44s.AvantiLipid      2.065e+02  1.319e+07       0        1
## CD44s.CD196            5.248e+02  2.413e+07       0        1
## CD44s.Vimentin        -3.300e+02  1.771e+07       0        1
## CD44s.CD20             3.495e+02  2.386e+07       0        1
## CD44s.p53              4.463e+02  3.932e+07       0        1
## CD14.Vimentin          7.197e+02  1.166e+07       0        1
## CD14.CD16              1.275e+01  1.335e+07       0        1
## CD14.CD44s             1.604e+02  9.006e+06       0        1
## FoxP3.XeBCK            9.015e+02  1.399e+08       0        1
## FoxP3.CD19             1.018e+02  7.634e+06       0        1
## E.cadherin.CD196      -7.354e+02  9.047e+06       0        1
## E.cadherin.FoxP3       7.302e+02  7.729e+07       0        1
## CD152.AvantiLipid     -1.295e+02  2.067e+07       0        1
## CD152.CD163           -1.426e+03  6.374e+07       0        1
## CD152.E.cadherin      -2.238e+02  1.201e+07       0        1
## CD8a.CD196             4.430e+02  1.295e+07       0        1
## CD8a.CD25              7.373e+02  1.016e+07       0        1
## CD8a.CD14              6.357e+02  8.246e+06       0        1
## CD11b.CD44s           -6.163e+02  1.345e+07       0        1
```

```
## Beta.catenin.p53     -3.388e+02  1.155e+07       0        1
## B7.H4.E.cadherin     -4.870e+00  2.817e+07       0        1
## Ki67.AvantiLipid      4.265e+03  8.133e+07       0        1
## Ki67.p53              5.929e+02  4.184e+07       0        1
## CollagenI.SrBCK      -8.634e+02  5.406e+07       0        1
## CollagenI.XeBCK       2.907e+03  2.904e+08       0        1
## CollagenI.CD196       9.256e+01  3.603e+06       0        1
## CollagenI.CD19       -2.138e+02  4.469e+07       0        1
## CollagenI.CD163      -1.110e+03  1.238e+08       0        1
## CollagenI.p53         2.193e+02  4.543e+07       0        1
## CollagenI.CD45        1.791e+02  2.486e+07       0        1
## CollagenI.CD44s       1.585e+02  1.379e+07       0        1
## CollagenI.CD4        -3.040e+02  3.494e+07       0        1
## CD3.CD45             -3.213e+02  3.242e+07       0        1
## B7.H3.RR101          -5.465e+02  3.102e+07       0        1
## HLA.DR.SrBCK         -3.618e+02  4.750e+07       0        1
## HLA.DR.AvantiLipid   -1.413e+03  1.855e+07       0        1
## HLA.DR.Ki67          -4.861e+02  3.508e+07       0        1
## pS6.AvantiLipid      -7.810e+02  1.704e+07       0        1
## pS6.CD134             6.408e+01  5.361e+06       0        1
## pS6.HLA.DR            3.378e+02  2.262e+07       0        1
## HistoneH3.RR101       9.638e+01  6.528e+06       0        1
## HistoneH3.Vimentin   -2.272e+03  2.399e+07       0        1
## HistoneH3.CD20       -1.416e+03  3.515e+07       0        1
## HistoneH3.CD134       3.473e+02  1.294e+07       0        1
## HistoneH3.CD45       -7.034e+02  2.117e+07       0        1
## HistoneH3.CD44s       7.448e+02  1.840e+07       0        1
## HistoneH3.FoxP3       6.345e+02  9.344e+07       0        1
## HistoneH3.E.cadherin -1.408e+03  2.931e+07       0        1
## HistoneH3.p21         5.473e+02  2.818e+07       0        1
## HistoneH3.CollagenI   1.230e+03  2.195e+07       0        1
## DNA191.CD163          1.798e+03  9.048e+07       0        1
## DNA191.CD20           1.010e+03  6.875e+07       0        1
## DNA191.p53           -5.000e+02  4.245e+07       0        1
## DNA193.XeBCK          3.233e+03  3.018e+08       0        1
## DNA193.CD19          -7.426e+02  3.263e+07       0        1
## DNA193.FoxP3          1.075e+03  2.149e+07       0        1
## DNA193.Ki67          -2.017e+03  3.746e+07       0        1
## DNA193.HLA.DR         1.229e+03  1.961e+07       0        1
## DNA193.DNA191         8.968e+00  6.846e+06       0        1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1.6167e+02  on 120  degrees of freedom
## Residual deviance: 4.5840e-10  on  45  degrees of freedom
## AIC: 152
##
## Number of Fisher Scoring iterations: 27

# Look at the VIFs
library(car)
vif(initModel)

##               CoreId    AvantiLipid.RR102          CD196.SrBCK
##            274.35914             68.25919           1954.29438
##          CD196.XeBCK           CD19.CD196       Vimentin.XeBCK
##            102.05953            274.98336           1557.08134
##       Vimentin.CD196        Vimentin.CD19          CD163.XeBCK
```

```
##            224.58496            357.88146           1218.78489
##          CD16.CD196           CD25.XeBCK           CD25.CD196
##          2312.59149            973.28338            761.41927
##           CD25.CD19            p53.CD196             p53.CD20
##           108.57328           3380.94038            255.96801
##           CD45.CD25    CD44s.AvantiLipid          CD44s.CD196
##          4649.79762            343.83970           1867.11544
##       CD44s.Vimentin           CD44s.CD20            CD44s.p53
##           765.94335            771.53682           1081.48400
##        CD14.Vimentin            CD14.CD16           CD14.CD44s
##           504.19056            147.77164            372.83141
##          FoxP3.XeBCK           FoxP3.CD19     E.cadherin.CD196
##            73.89555             80.58028            312.58433
##     E.cadherin.FoxP3    CD152.AvantiLipid          CD152.CD163
##          3448.22071           1210.87395           1377.49540
##     CD152.E.cadherin           CD8a.CD196            CD8a.CD25
##           772.99695           1433.62996             76.45263
##           CD8a.CD14          CD11b.CD44s     Beta.catenin.p53
##           520.64894            206.49233            343.10751
##     B7.H4.E.cadherin     Ki67.AvantiLipid             Ki67.p53
##           641.89802           3154.26256           5788.75184
##      CollagenI.SrBCK      CollagenI.XeBCK      CollagenI.CD196
##           236.73695           1529.22184            238.72848
##       CollagenI.CD19      CollagenI.CD163        CollagenI.p53
##          2790.62066           1658.11388           2179.86355
##       CollagenI.CD45      CollagenI.CD44s        CollagenI.CD4
##           182.18580            304.64228           2983.36298
##            CD3.CD45          B7.H3.RR101         HLA.DR.SrBCK
##          1983.57259           4779.22317           6153.76017
##   HLA.DR.AvantiLipid          HLA.DR.Ki67      pS6.AvantiLipid
##          1039.58441           2740.51106            349.97523
##           pS6.CD134            pS6.HLA.DR      HistoneH3.RR101
##           135.44281           1209.71779            142.41971
##   HistoneH3.Vimentin       HistoneH3.CD20      HistoneH3.CD134
##           699.20282            622.34000            333.71593
##       HistoneH3.CD45      HistoneH3.CD44s      HistoneH3.FoxP3
##           260.51664            513.92370           4363.62660
## HistoneH3.E.cadherin        HistoneH3.p21  HistoneH3.CollagenI
##          2812.36908            905.03840           1040.90722
##         DNA191.CD163          DNA191.CD20           DNA191.p53
##          1119.80585            690.10707           5483.24577
##         DNA193.XeBCK          DNA193.CD19         DNA193.FoxP3
##           591.36331            621.73769            186.65420
##          DNA193.Ki67        DNA193.HLA.DR        DNA193.DNA191
##          1370.72245           2329.65186            727.66523
```

Pretty high VIFs, but let's do a stepping search.

```
source("../Utils_Maxi.R")
reducedCoLinModelArr200 = AICVIFCoElimination(DecorrelateVariables(initModel,200,verbose=F)
                                              ,verbose=F)
reducedCoLinModelArr100 = AICVIFCoElimination(DecorrelateVariables(initModel,100,verbose=F)
                                              ,verbose=F)
reducedCoLinModelArr20 = AICVIFCoElimination(DecorrelateVariables(initModel,20,verbose=F)
                                             ,verbose=F)
reducedCoLinModelArr10 = AICVIFCoElimination(DecorrelateVariables(initModel,10,verbose=F)
                                             ,verbose=F)
```

Say we tolerate a maximum VIF of 25. What are the best AICs we get?

```
targetVIF = 25
best200 = reducedCoLinModelArr200[unlist(reducedCoLinModelArr200$maxVIF)<targetVIF,]
best200 = best200[which.min(unlist(best200$V1)),]
best100 = reducedCoLinModelArr100[unlist(reducedCoLinModelArr100$maxVIF)<targetVIF,]
best100 = best100[which.min(unlist(best100$V1)),]
best20 = reducedCoLinModelArr20[unlist(reducedCoLinModelArr20$maxVIF)<targetVIF,]
best20 = best20[which.min(unlist(best20$V1)),]
best10 = reducedCoLinModelArr10[unlist(reducedCoLinModelArr10$maxVIF)<targetVIF,]
best10 = best10[which.min(unlist(best10$V1)),]
print(best200[1:4])

##         V1  accuracy   maxVIF nVariables
## 5 139.1947 0.7933884 4.558939         14

print(best100[1:4])

##         V1  accuracy   maxVIF nVariables
## 5 139.1947 0.7933884 4.558939         14

print(best20[1:4])

##         V1  accuracy   maxVIF nVariables
## 1 147.5249 0.7024793 16.69107          8

print(best10[1:4])

##         V1  accuracy   maxVIF nVariables
## 1 144.6403 0.7107438 8.840991          8
```

Nice, so we get a model with fairly de-correlated variables (maxVIF around xxx) and pretty decent predictive power (around xxx) accuracy)!

What does the model consist of?

```
best100Model = glm(paste0(best100[,5]),family=binomial(link='logit'),
                           data=corrArr_Reduced)
PlotCoefficients(best10Model,yLim=c(-30,30),yPos=22,errBarWidth=.4)

## Error in PlotCoefficients(best10Model, yLim = c(-30, 30), yPos = 22, errBarWidth = 0.4):
object 'best10Model' not found
```

Strange... It's picking up XeBCK which should be background control.

3 Cleaning up the Data

I just spoke to Olya and I now know all the different stains. They are all meaningful to a certain extend, but there is a certain amount of redundancy in them. Let's clean the data up to remove some of that redundancy.

```
stainsToOmitVec = c('SrBCK','RR101','XeBCK','DNA193')
colToOmitVec = c()

# Calculate the index of the columns with correlations with the above stains and
# collect them in a vector.
k = 3
for (i in seq(2,37)) {
  for (j in seq(i-1)) {
    if (any(markerLabelsVec[c(i,j)] %in% stainsToOmitVec)) {
      colToOmitVec = c(colToOmitVec,k)
    }
```

```
    k = k + 1
  }
}

corrArr_Curated = corrArr[,-colToOmitVec]
dim(corrArr_Curated) # Should be removing 36*4-4*3/2 = 138, so expect 530

## [1] 121 530
```

Let's do de-correlation:

```
covariatesArr = corrArr_Curated[,-c(1,2)]
coMat = cor(covariatesArr)
hc = findCorrelation(coMat,cutoff=0.7,exact=TRUE) # put any value as a "cutoff"
hc = sort(hc)
corrArrCurated_Reduced = data.frame(corrArr_Curated[,c(1,2)],covariatesArr[,-c(hc)])
dim(corrArrCurated_Reduced)

## [1] 121  67
```

Let's try fitting a model again.

```
initModel = glm(PtSnty ~.,family=binomial(link='logit'),
                control = list(maxit = 100),
                data=corrArrCurated_Reduced[,-1])

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

summary(initModel)

##
## Call:
## glm(formula = PtSnty ~ ., family = binomial(link = "logit"),
##     data = corrArrCurated_Reduced[, -1], control = list(maxit = 100))
##
## Deviance Residuals:
##        Min          1Q      Median          3Q         Max
## -5.552e-06  -2.110e-08   2.110e-08   1.121e-06   5.050e-06
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)             3.405e+02  1.280e+07       0        1
## AvantiLipid.RR102       1.317e+00  7.363e+06       0        1
## CD19.CD196             -5.601e+02  7.556e+06       0        1
## Vimentin.CD196         -6.878e+02  1.351e+07       0        1
## Vimentin.CD19           2.909e+03  4.602e+07       0        1
## CD16.AvantiLipid       -1.798e+03  3.170e+07       0        1
## CD16.CD196              4.008e+01  5.607e+07       0        1
## CD25.CD196              8.684e+02  4.995e+06       0        1
## CD25.CD19               3.439e+02  5.642e+06       0        1
## CD25.CD163             -9.170e+03  7.482e+07       0        1
## p53.AvantiLipid         4.200e+00  1.479e+07       0        1
## p53.CD196              -1.038e+03  1.362e+07       0        1
## CD44s.AvantiLipid       4.999e+02  2.563e+07       0        1
## CD44s.CD196             7.765e+02  2.066e+07       0        1
## CD44s.Vimentin          2.546e-01  7.909e+06       0        1
## CD44s.CD16              4.344e+02  1.885e+07       0        1
## CD44s.p53               6.577e+02  1.713e+07       0        1
## CD14.Vimentin           9.333e+02  7.928e+06       0        1
```

```
## CD14.CD16                5.656e+01  5.480e+07       0        1
## CD14.CD134              -9.134e+01  8.754e+06       0        1
## CD14.CD44s               3.673e+02  1.224e+07       0        1
## FoxP3.CD19              -6.080e+01  5.718e+06       0        1
## CD4.p53                 -3.973e+02  7.407e+06       0        1
## CD4.CD44s               -1.501e+03  1.833e+07       0        1
## E.cadherin.CD196        -4.787e+02  8.307e+06       0        1
## E.cadherin.FoxP3         1.985e+02  4.652e+07       0        1
## CD152.CD163             -9.288e+02  1.887e+07       0        1
## CD8a.AvantiLipid        -6.787e+01  1.852e+07       0        1
## CD8a.CD196               4.868e+02  5.200e+06       0        1
## CD8a.CD25               -1.092e+02  2.531e+07       0        1
## CD8a.E.cadherin          2.650e+02  9.350e+06       0        1
## CD11b.CD45              -1.829e+03  3.490e+07       0        1
## B7.H4.AvantiLipid        6.632e+02  7.476e+07       0        1
## Ki67.AvantiLipid         1.397e+03  5.332e+07       0        1
## Ki67.B7.H4              -1.278e+03  2.240e+07       0        1
## CollagenI.CD196          7.406e+01  4.564e+06       0        1
## CollagenI.CD19          -5.153e+02  1.112e+07       0        1
## CollagenI.CD163          1.095e+03  6.770e+07       0        1
## CollagenI.p53            2.492e+01  1.167e+07       0        1
## CollagenI.CD45           5.466e+02  2.579e+07       0        1
## CollagenI.CD44s         -2.569e+02  1.011e+07       0        1
## CollagenI.CD14          -2.147e+02  7.525e+06       0        1
## CollagenI.Beta.catenin -2.092e+02  8.784e+06       0        1
## B7.H3.RR102             -3.406e+02  1.097e+07       0        1
## HLA.DR.RR102             6.536e+02  3.298e+07       0        1
## HLA.DR.p53               1.996e+01  8.928e+06       0        1
## HLA.DR.Ki67              6.116e+02  1.979e+07       0        1
## pS6.AvantiLipid         -8.196e+02  2.210e+07       0        1
## pS6.CD134               -1.851e+02  5.026e+06       0        1
## pS6.HLA.DR               3.293e+02  1.690e+07       0        1
## HistoneH3.RR102         -3.525e+01  1.611e+07       0        1
## HistoneH3.Vimentin      -6.300e+02  3.165e+07       0        1
## HistoneH3.CD20          -1.863e+03  7.297e+07       0        1
## HistoneH3.CD134          2.839e+02  1.339e+07       0        1
## HistoneH3.CD45          -1.138e+03  3.145e+07       0        1
## HistoneH3.CD44s          1.112e+03  1.498e+07       0        1
## HistoneH3.p21            1.418e+03  1.795e+07       0        1
## HistoneH3.CD152         -1.130e+03  1.639e+07       0        1
## HistoneH3.B7.H4         -2.397e+02  3.157e+07       0        1
## HistoneH3.CollagenI      6.994e+02  1.542e+07       0        1
## HistoneH3.HLA.DR        -5.854e+02  1.693e+07       0        1
## DNA191.CD19              1.109e+03  3.681e+07       0        1
## DNA191.CD163            -5.344e+00  3.153e+07       0        1
## DNA191.CD20              3.272e+03  9.246e+07       0        1
## DNA191.FoxP3            -4.923e+02  1.854e+07       0        1
## DNA191.HLA.DR            5.183e+02  1.285e+07       0        1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1.6167e+02  on 120  degrees of freedom
## Residual deviance: 4.2703e-10  on  55  degrees of freedom
## AIC: 132
##
## Number of Fisher Scoring iterations: 28

# Look at the VIFs
```

```
vif(initModel)
```

```
##      AvantiLipid.RR102             CD19.CD196         Vimentin.CD196
##              100.15973              219.70611              495.54802
##          Vimentin.CD19       CD16.AvantiLipid             CD16.CD196
##              631.56673              279.07232             1609.11311
##             CD25.CD196             CD25.CD19             CD25.CD163
##              231.22305              165.44975              136.73243
##        p53.AvantiLipid             p53.CD196      CD44s.AvantiLipid
##              382.66081              481.22662              959.80087
##            CD44s.CD196         CD44s.Vimentin             CD44s.CD16
##             1146.46674              167.05345              223.04109
##             CD44s.p53          CD14.Vimentin             CD14.CD16
##              134.74573              174.23425             1805.15391
##             CD14.CD134             CD14.CD44s             FoxP3.CD19
##              364.87324              961.56553              21.00971
##                CD4.p53             CD4.CD44s       E.cadherin.CD196
##              167.42148              706.93091              402.95821
##       E.cadherin.FoxP3            CD152.CD163       CD8a.AvantiLipid
##              622.05083              122.53286              702.37535
##             CD8a.CD196             CD8a.CD25        CD8a.E.cadherin
##              350.03564              632.53709              414.57072
##             CD11b.CD45      B7.H4.AvantiLipid       Ki67.AvantiLipid
##              437.91610              623.99836             1355.56080
##             Ki67.B7.H4        CollagenI.CD196         CollagenI.CD19
##              467.23494              327.03865              158.66681
##        CollagenI.CD163          CollagenI.p53         CollagenI.CD45
##              389.17491              134.05468              115.64912
##        CollagenI.CD44s         CollagenI.CD14 CollagenI.Beta.catenin
##              141.13691              123.27490              276.12938
##            B7.H3.RR102           HLA.DR.RR102             HLA.DR.p53
##              358.40130             3930.97560             1025.99257
##            HLA.DR.Ki67        pS6.AvantiLipid              pS6.CD134
##             1210.47989              343.48040               96.12266
##             pS6.HLA.DR       HistoneH3.RR102     HistoneH3.Vimentin
##              419.03356              590.11377              618.96788
##         HistoneH3.CD20       HistoneH3.CD134         HistoneH3.CD45
##             1677.44839              260.58545              668.89301
##        HistoneH3.CD44s          HistoneH3.p21        HistoneH3.CD152
##              202.32437              247.04700              618.23921
##        HistoneH3.B7.H4    HistoneH3.CollagenI       HistoneH3.HLA.DR
##              362.90191              425.16517             2776.28349
##            DNA191.CD19          DNA191.CD163            DNA191.CD20
##              464.45397              114.82713              747.01580
##           DNA191.FoxP3          DNA191.HLA.DR
##              164.06612              448.21844
```

Pretty high VIFs, but let's do a stepping search.

```
reducedCoLinModelArr200 = AICVIFCoElimination(DecorrelateVariables(initModel,200,verbose=F)
                                              ,verbose=F)
reducedCoLinModelArr100 = AICVIFCoElimination(DecorrelateVariables(initModel,100,verbose=F)
                                              ,verbose=F)
reducedCoLinModelArr20 = AICVIFCoElimination(DecorrelateVariables(initModel,20,verbose=F)
                                             ,verbose=F)
reducedCoLinModelArr10 = AICVIFCoElimination(DecorrelateVariables(initModel,10,verbose=F)
                                             ,verbose=F)
# Print out the results
```

```
reducedCoLinModelArr200[,1:4]

##         V1  accuracy   maxVIF nVariables
## 1 141.4547  0.768595 16.05183         15
## 2 151.3508 0.7438017 4.519692         14
## 3 150.3381 0.7190083 4.519692         13
## 4 160.2219 0.6942149 3.224513         12
## 5 151.9751 0.7190083 3.224513          6
## 6 156.1263 0.6694215  1.91365          5
## 7 153.0455  0.661157  1.91365          2
## 8 159.8677 0.6033058 1.003952          1
## 9 159.8677 0.6033058 1.003952          1

reducedCoLinModelArr100[,1:4]

##         V1  accuracy   maxVIF nVariables
## 1 141.4547  0.768595 16.05183         15
## 2 151.3508 0.7438017 4.519692         14
## 3 150.3381 0.7190083 4.519692         13
## 4 160.2219 0.6942149 3.224513         12
## 5 151.9751 0.7190083 3.224513          6
## 6 156.1263 0.6694215  1.91365          5
## 7 153.0455  0.661157  1.91365          2
## 8 159.8677 0.6033058 1.003952          1
## 9 159.8677 0.6033058 1.003952          1

reducedCoLinModelArr20[,1:4]

##         V1  accuracy   maxVIF nVariables
## 1 141.4547  0.768595 16.05183         15
## 2 151.3508 0.7438017 4.519692         14
## 3 150.3381 0.7190083 4.519692         13
## 4 160.2219 0.6942149 3.224513         12
## 5 151.9751 0.7190083 3.224513          6
## 6 156.1263 0.6694215  1.91365          5
## 7 153.0455  0.661157  1.91365          2
## 8 159.8677 0.6033058 1.003952          1
## 9 159.8677 0.6033058 1.003952          1

reducedCoLinModelArr10[,1:4]

##         V1  accuracy   maxVIF nVariables
## 1 144.4886 0.7768595 9.742738         10
## 2 156.4682 0.7272727 2.460953          9
## 3 151.1057 0.7272727 2.460953          4
## 4 159.8545 0.6446281  2.22998          3
## 5 156.0279 0.6528926  2.22998          1
```

Say we tolerate a maximum VIF of 25. What are the best AICs we get?

```
targetVIF = 5
best200 = reducedCoLinModelArr200[unlist(reducedCoLinModelArr200$maxVIF)<targetVIF,]
best200 = best200[which.min(unlist(best200$V1)),]
best100 = reducedCoLinModelArr100[unlist(reducedCoLinModelArr100$maxVIF)<targetVIF,]
best100 = best100[which.min(unlist(best100$V1)),]
best20 = reducedCoLinModelArr20[unlist(reducedCoLinModelArr20$maxVIF)<targetVIF,]
best20 = best20[which.min(unlist(best20$V1)),]
best10 = reducedCoLinModelArr10[unlist(reducedCoLinModelArr10$maxVIF)<targetVIF,]
best10 = best10[which.min(unlist(best10$V1)),]
```

```
print(best200[1:4])

##         V1  accuracy   maxVIF nVariables
## 3 150.3381 0.7190083 4.519692         13

print(best100[1:4])

##         V1  accuracy   maxVIF nVariables
## 3 150.3381 0.7190083 4.519692         13

print(best20[1:4])

##         V1  accuracy   maxVIF nVariables
## 3 150.3381 0.7190083 4.519692         13

print(best10[1:4])

##         V1  accuracy   maxVIF nVariables
## 3 151.1057 0.7272727 2.460953          4
```

Starting from a VIF of 10 seems to be giving the best results. It gives a model with 4 coefficients and 72% accuracy!

```
model4Coef = glm(paste0(best10[,5]),family=binomial(link='logit'),
                          data=corrArrCurated_Reduced)
p = PlotCoefficients(model4Coef,yLim=c(-100,100),yPos=110,errBarWidth=.4)
# Annotate the markers
# yPos = 132.5
# tSize = 2.5
# # Positive
# p = p + annotate("text", x = "CollagenI.CD163", y = yPos,
#                          label = "Immature Dendritic Cells\n Memory T-Cells\n and Collagen", size =
# p = p + annotate("text", x = "CD44s.AvantiLipid", y = yPos,
#                          label = "Cancer Stem Cell Markers\n and Cell Membrane", size = tSize)
# p = p + geom_rect(aes(xmin = "Ki67.B7.H4", xmax = 4.5, ymin = 115, ymax = 150),
#                fill = "transparent", color = "green4", size = 1.5)
#
# # Negative
# yPos = -132.5
# p = p + annotate("text", x = "Ki67.B7.H4", y = yPos,
#                          label = "Cell Proliferation \n and Immune Check Point", size = tSize)
# p = p + annotate("text", x = "HistoneH3.Vimentin", y = yPos,
#                          label = "Cell Nucleus Marker \n and Motile Phenotype", size = tSize)
# p = p + geom_rect(aes(xmin = 0, xmax = "CD44s.AvantiLipid", ymin = -115, ymax = -150),
#                fill = "transparent", color = "red", size = 1.5)
p

## Warning:  Removed 1 rows containing missing values (geom_text).
## Warning:  Removed 1 rows containing missing values (geom_text).
## Warning:  Removed 1 rows containing missing values (geom_text).
```

Alternatively there is a model with 13 coefficients:

```
model13Coef = glm(paste0(best20[,5]),family=binomial(link='logit'),
                          data=corrArrCurated_Reduced)
PlotCoefficients(model13Coef,yLim=c(-100,100),yPos=22,errBarWidth=.4)
```

How do the two compare in cross-validation?

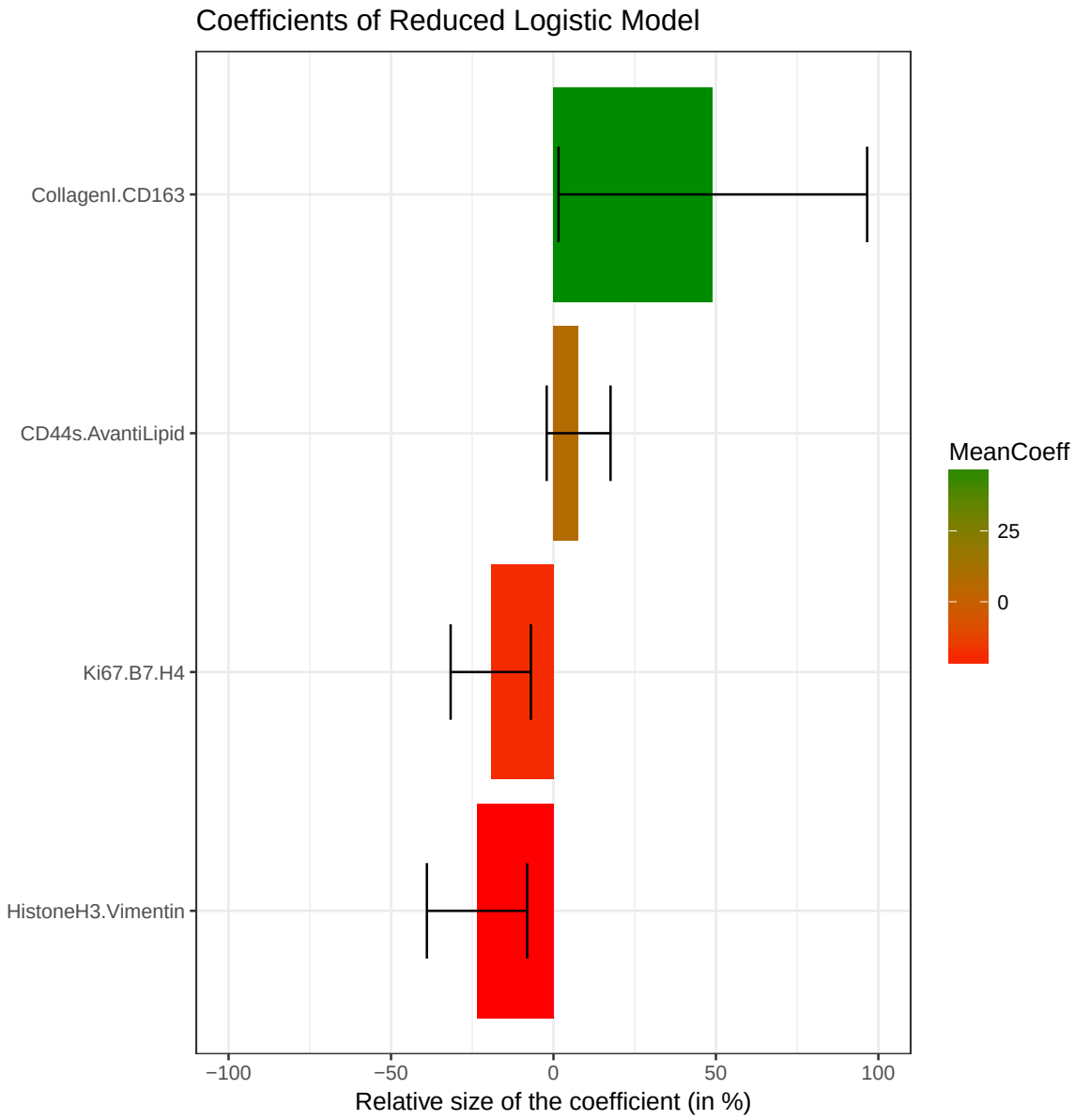


Figure 2: Importance of the different stains according to the logistic model with maxVIF 100. Asterisk indicates level of statistical support for non-zero contribution from this stain (T-test: *p<0.05,**p<0.01).

```
# Cross-validation
modelVec = c(model4Coef$formula,model13Coef$formula)
labelVec = c("4 Coefficient Model",
             "13 Coefficient Model")
PlotCrossValidation(modelVec,corrArrCurated_Reduced,nIter=100,nFolds=5,labelVec=labelVec)
```

Interesting. The model with only 4 variables does much better than the one with 13. That's a bit strange. Maybe I should check the ROC curves? Anyhow, for now the results suggest that maybe macrophage infiltration and CD44s expression are positively related to response and B7H4 and Vimentin are negatively related.

CD163 is a macrophage marker, so a strong collagen-cd163 correlation might indicate macrophage infiltration? CD44s is a membrane protein involved in cell-cell interactions. Thus, its correlation with AvantiLipid, which marks cell membranes is not too surprising. However, the fact that it is correlated with good outcome is something that the mean level model gives as well, and has been found in other studies as well. I see this as a little bit of a confirmation that were not just picking up noise. It shows that were picking up a correlation between two stains that we would expect to correlate and a result that

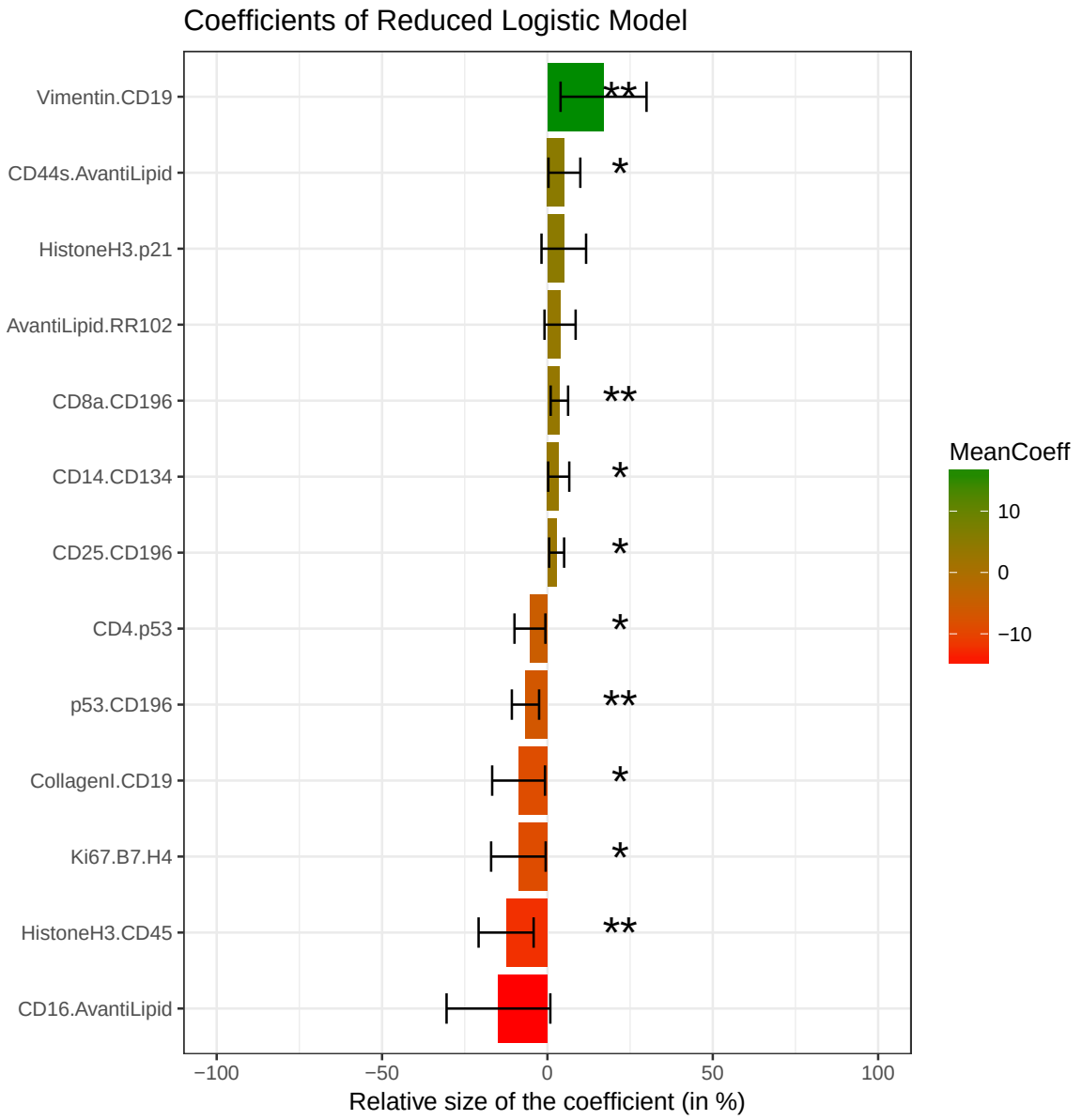


Figure 3: Importance of the different stains according to the logistic model with maxVIF 100. Asterisk indicates level of statistical support for non-zero contribution from this stain (T-test: *p<0.05,**p<0.01).

is biologically valid as well. B7H4 is a checkpoint inhibitor. Its interaction with Ki67 perhaps means that these cells are using this inhibitor for immune evasion? Histones and vimentin perhaps hints at very aggressive tumour cells? Vimentin is part of the cytoskeleton and involved in actively moving cells.

In principle this could be interesting, however, it is not clear from this whether it's just because it likes the CD163 levels in general, or whether it is really about CD163 and Collagen being in the same place.

To if this is the case, let's first plot out the correlation for the patients to see if it really does separate them now and then colour in the images for those patients where the signal is strongest.

```
only4CoefDataArr = corrArrCurated_Reduced[,c("CoreId","PtSnty",names(model4Coef$coefficients)[-1])]
only4CoefDataArr = data.frame(only4CoefDataArr,Prediction=predict(model4Coef,only4CoefDataArr,type='
# only4CoefDataArr[,2:5] = t(apply(only4CoefDataArr,1,function(row){row[-1]*model4Coef$coefficients[
only4CoefDataArr = only4CoefDataArr[with(only4CoefDataArr, order(PtSnty)), ]
only4CoefDataArr_idxd = data.frame(only4CoefDataArr,LinId=seq(nrow(only4CoefDataArr)))
only4CoefDataArr_reshaped = melt(only4CoefDataArr_idxd[,-1],id.vars=c("LinId"))
ggplot(only4CoefDataArr_reshaped, aes(variable, LinId)) +
```

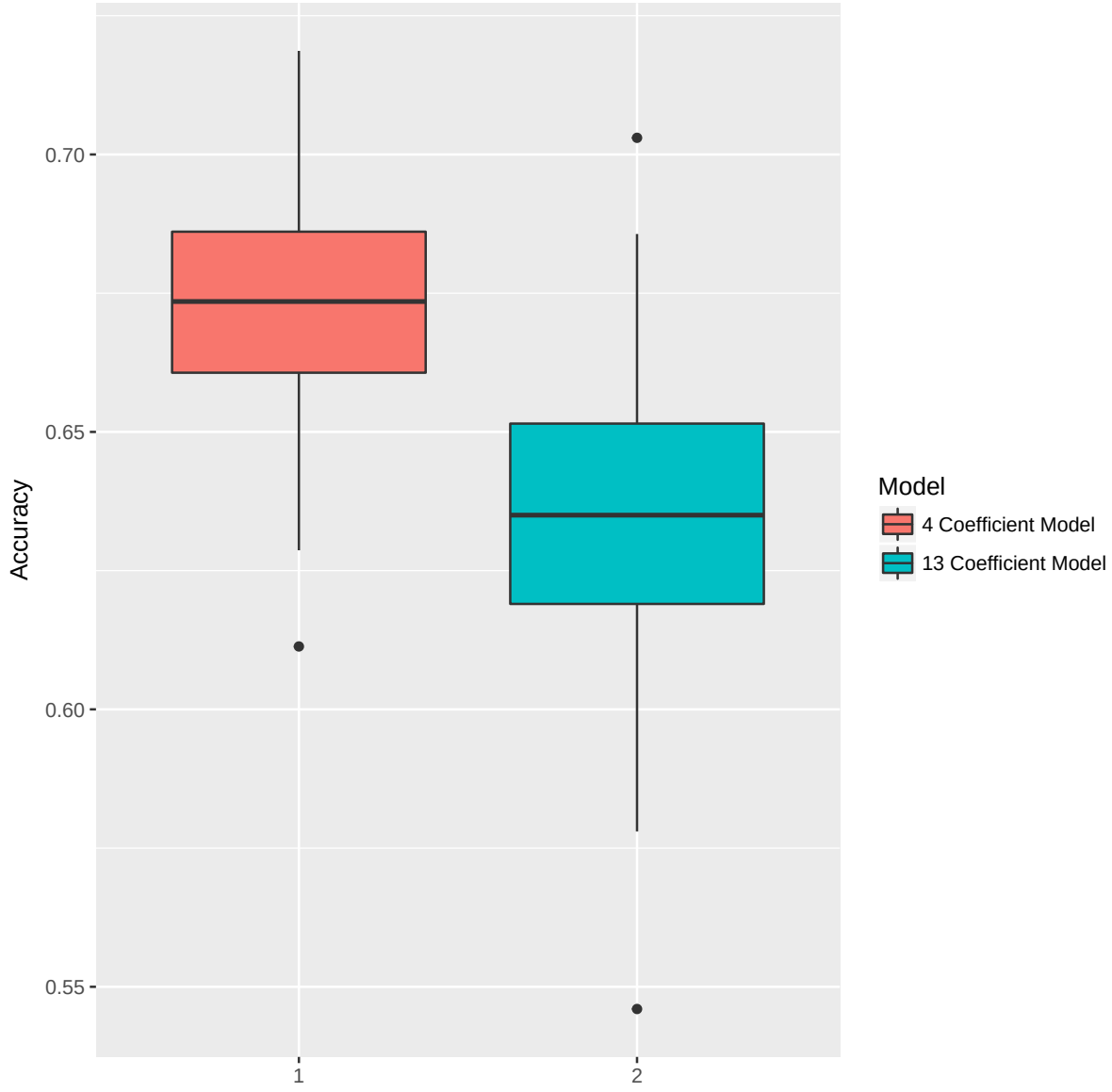


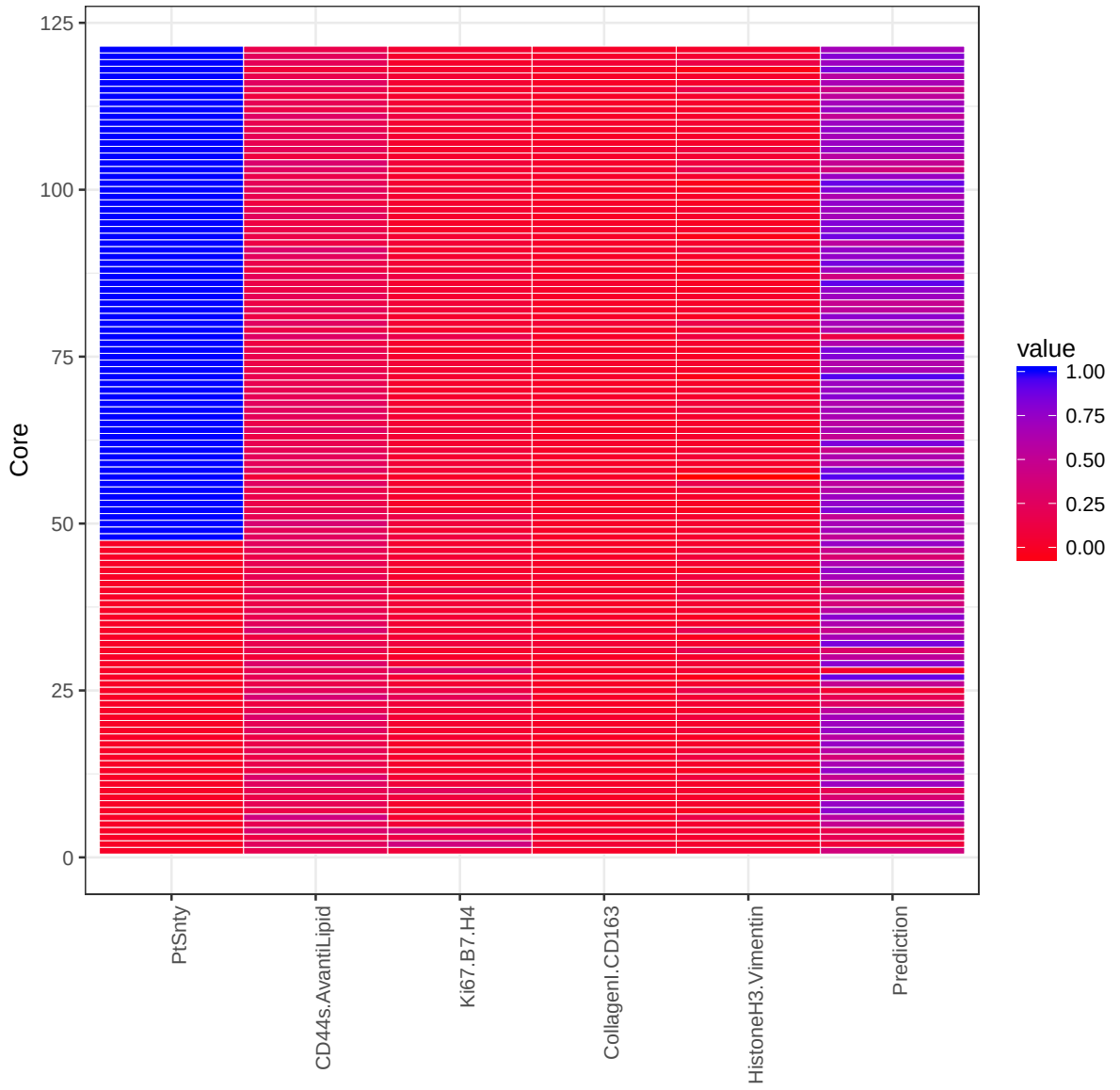
Figure 4: Comparison of the model with 4 and the model with 13 coefficients in cross-validation (5-fold, 100 iterations). Intriguingly the 4 Coefficient model does significantly better!

```
  geom_tile(aes(fill = value),colour="white") +
  scale_fill_gradient(low="red",high="blue") +
  theme_bw() +
  labs(x="",y="Core") +
  theme(axis.text.x = element_text(angle=90, hjust=1))
```

```
# predictions = ifelse(predictions > 0.5,1,0)
# predictions==corrArrCurated_Reduced£PtSnty
#
# mean((predictions==0)[corrArrCurated_Reduced£PtSnty==0])

only4CoefDataArr[which.min(only4CoefDataArr$Prediction),]

##    CoreId PtSnty CD44s.AvantiLipid Ki67.B7.H4 CollagenI.CD163
## 69    200      0         0.1864656  0.2755805    -0.002569898
```




```
##    HistoneH3.Vimentin Prediction
## 69          0.0420468 0.02315692

only4CoefDataArr[which.max(only4CoefDataArr$Prediction),]

##    CoreId PtSnty CD44s.AvantiLipid Ki67.B7.H4 CollagenI.CD163
## 38    161      1         0.1709083 0.03831591      0.02387417
##    HistoneH3.Vimentin Prediction
## 38         -0.0437627  0.9293528

write.csv(file="correlationModelScores.csv",x=only4CoefDataArr,row.names = FALSE)

only4CoefDataArr = only4CoefDataArr[with(only4CoefDataArr, order(Prediction)), ]
```

What else might be helpful is to look at the images for patients with elevated correlations to see what they correspond to. Let's find a patient who has particularly high