DNA Hydroxymethylation in Hepatocellular Carcinoma

Francois Collin

2020-09-06

Contents

```
  # file rmarkdown file management options: cache, figures
figures_DIR <- file.path('Static', 'figures/')
suppressMessages(dir.create(figures_DIR, recursive=T))
```

```
## Warning in dir.create(figures_DIR, recursive = T): 'Static/figures' already
## exists
```

```
knitr::opts_chunk$set(fig.path=paste0(figures_DIR))
```

Preamble

This vignette offers some exploratory data analyses of data available from the NCBI GEO web site.

License

Chapter 1

Introduction

The goal of detecting cancer at the earliest stage of development with a non-invasive procedure has busied many groups with the task of perfecting techniques to support what has become commonly known as a liquid biopsy - the analysis of biomarkers circulating in fluids such as blood, saliva or urine. Epigenetic biomarkers present themselves as good candidates for this application (Gai and Sun (2019) [1]). In particular, given their prevalence in the human genome, close correlation with gene expression and high chemical stability, DNA modifications such as 5-methylcytosine (5mC) and 5-hydroxymethylcytosine (5hmC) are DNA epigenetic marks that provide much promise as cancer diagnosis biomarkers that could be profitably analyzed in liquid biopsies [2–5].

Li et al. (2017) [3] used a sensitive and selective chemical labeling technology to extract genome-wide 5hmC profiles from circulating cell-free DNA (cfDNA) as well as from genomic DNA (gDNA) collected from a cohort of 260 patients recently diagnosed with colorectal, gastric, pancreatic, liver or thyroid cancer and normal tissues from 90 healthy individuals They found 5hmC-based biomarkers of circulating cfDNA to be highly predictive of some cancer types. Similar small sample size findings were reported in Song et al. (2017) [4].

Focusing on hepatocellular carcinoma, Cai et al. (2019) [2] assembled a sizable dataset to demonstrate the feasibility of using features derived from 5-hydroxymethylcytosines marks in circulating cell-free DNA as a non-invasive approach for the early detection of hepatocellular carcinoma. The data that are the basis of that report are available on the NCBI GEO web site (Series GSE112679). The data have also been bundled in a R data package which can be installed from github:

```
if (!requireNamespace("devtools", quietly = TRUE))
    install.packages("devtools")
devtools::install_github("12379Monty/GSE112679")
```

An important question in the early development of classifiers of the sorts that are the basis of any liquid biopsy diagnostic tool is how many samples should be collected to make properly informed decisions. In this report we will explore the GSE112679 data to shed some light on the relationship between sample size and model performance in the context classifying samples based on 5hmC data.

In Section 2 we preprocess the data that we will use for the classification analysis and perform some light QC analyses. In Section 3 we provide some background to our modeling approach. In Section **??** we explore some glmnet fits that discriminate between early stage HCC and control samples. In Section 6 we examine the results of fitting a suite of models to investigate the effect of sample size on model performance.

Chapter 2

Preprocessing

2.1 Load the data

The data that are available from NCBI GEO Series GSE112679 can be conveniently accessed through an R data package. Attaching the GSE112679 package makes the count data tables available as well as a gene annotation table and a sample description table. See GSE112679 R Data Package page. For the Cai et al. [2] model fitting and analysis, samples were separated into `Train` and `Val-1` subsets. `Val-2` was an external validation set.

```
if (!("GSE112679" %in% rownames(installed.packages()))) {
  if (!requireNamespace("devtools", quietly = TRUE)) {
    install.packages("devtools")
  }
  devtools::install_github("12379Monty/GSE112679")
}
library(GSE112679)
sampDesc$DxStage <- with(sampDesc, ifelse(outcome=='HCC',
   paste0(outcome,':',stage), outcome))

with(
  sampDesc %>% dplyr::filter(sampType == "blood"),
  knitr::kable(table(DxStage, trainValGroup, exclude = NULL),
   caption="GSE112679 Samples by Dx Group and Subset") %>%
  kableExtra::kable_styling(full_width = F)
)
```

For this analysis, we will consider early stage cancer samples and healthy or benign samples from the `Train` or `Val-1` subsets. The appropriate outcome variable will be renamed or aliased `group`

Table 2.1: GSE112679 Samples by Dx Group and Subset

	Train	Val-1	Val-2
Benign	253	132	3
CHB	190	96	0
Cirrhosis	73	33	0
HCC:Early	335	220	24
HCC:Late	0	442	13
HCC:NA	0	147	23
Healthy	269	124	177

Table 2.2: Samples used in this analysis

group	Freq
Control	778
HCC	555

The features are counts of reads captured by chemical labeling, and indicate the level of 5-hydroxymethylcytosines within each gene body. Cai et al. (2019), Li et al. (2017) and Song et al. (2017) [2–4] all analyze 5hmC gene body counts using standard RNA-Seq methodologies, and we will do the same here.

Note that before conducting any substantive analyses, the data would normally be very carefully examined for any sign of quality variation between groups of samples. This analysis would integrate sample meta data - where and when were the blood samples collected - as well as library preparation and sequencing metrics in order to detect any sign of processing artifacts that may be present in the dataset. This is particularly important when dealing with blood samples as variable DNA quality degradation is a well known challenge that is encountered when dealing with such samples [6]. Although blood specimen handling protocols can be put in place to minimize quality variation [7], variability can never be completely eradicated, especially in the context of blood samples collected by different groups, working in different environments. The problem of variable DNA quality becomes paricularly pernicuous when it is compounded with a confounding factor that sneaks in when the control sample collection events are separated in time and space from the cancer sample collection events; an all too common occurence.

As proper data QC requires an intimate familiarity with the details of data collection and processing, such a task cannot be untertaken here. We will simply run a *minimal set of QC sanity checks* to make sure that there are no apparent systematic effects in the data.

We first look at coverage - make sure there isn't too much disparity of coverage across samples. To detect shared variability, samples can be annotated

Table 2.3: Coverage Summary - Columns are sample coverage quantiles and total coverage Rows are quartiles across samples

	15%	25%	50%	75%	totCovM
25%	4	24	111	321	5.5
50%	5	30	135	391	6.7
75%	6	35	162	468	8.0

and ordered according to sample features that may be linked to sample batch processing. Here we the samples have been ordered by group and sample id (an alias of geoAcc).

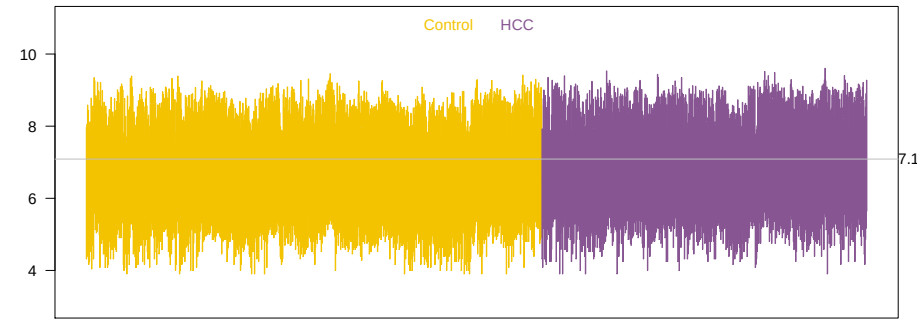


Figure 2.1: Sample log2 count boxplots

From this table, we see that 25% of the samples have total coverage exceeding 8M reads, 25% of samples have a 15 percentile of coverage lower than 4, etc.

2.2 Differential representation analysis

In the remainder of this section, we will process the data and perform differential expression analysis as outlined in Law et al. (2018) [8]. The main analysis steps are:

- remove lowly expressed genes
- normalize gene expression distributions
- remove heteroscedascity
- fit linear models and examine DE results

It is good practice to perform this differential expression analysis prior to fitting models to get an idea of how difficult it will be to discriminate between samples

belonging to the different subgroups. The pipeline outlined in Law et al. (2018) [8] also provides some basic quality assessment opportunities.

Remove lowly expressed genes

Genes that are not expressed at a biologically meaningful level in any condition should be discarded to reduce the subset of genes to those that are of interest, and to reduce the number of tests carried out downstream when looking at differential expression. Carrying un-informative genes may also be a hindrance to classification and other downtream analyses.

To determine a sensible threshold we can begin by examining the shapes of the distributions.

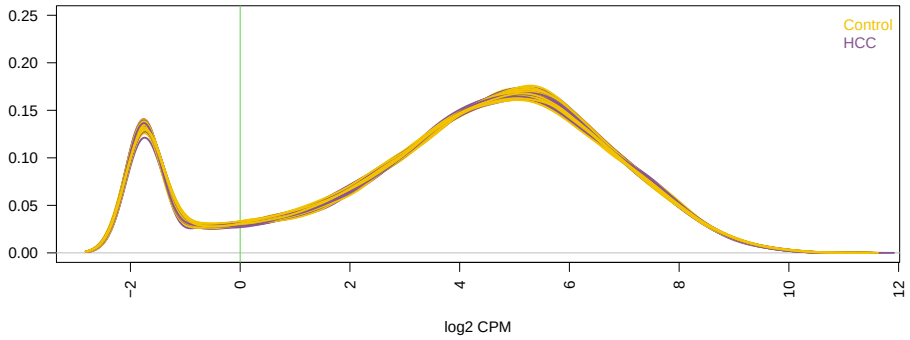


Figure 2.2: Sample log_2 CPM densities

As is typically the case with RNA-Seq data, we notice many weakly represented genes in this dataset. A cpm value of 1 appears to adequatly separate the expressed from the un-expressed genes, but we will be slightly more strict here and require a CPM threshold of 3 . Using a nominal CPM value of 3, genes are deeemed to be `represented` if their expression is above this threshold, and not represented otherwise. For this analysis we will require that genes be `represented` in at least 25 samples across the entire dataset to be retained for downstream analysis. Here, a CPM value of 3 means that a gene is represented if it has at least 9 reads in the sample with the lowest sequencing depth (library size 2.9 million). Note that the thresholds used here are arbitrary as there are no hard and fast rules to set these by. The voom-plot, which is part of analyses done to remove heteroscedasticity, can be examined to verify that the filtering performed is adequate.

Remove weakly represented genes and replot densities.

Removing 17.5% of genes. . .

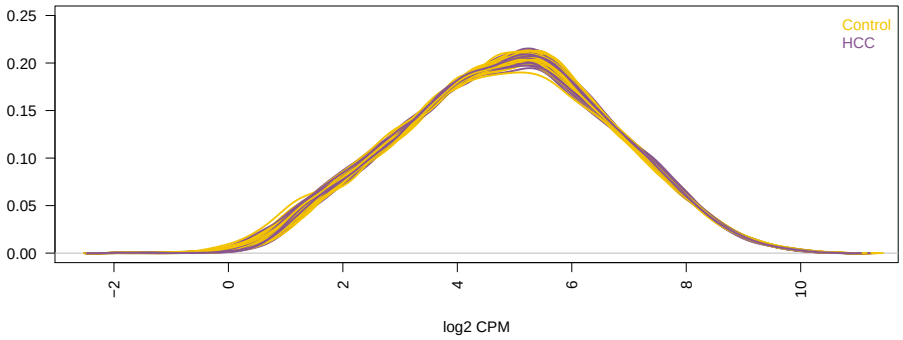


Figure 2.3: Sample log_2 CPM densities after removing weak genes

As another sanity check, we will look at a multidimensional scaling plot of distances between gene expression profiles. We use `plotMDS` in limma package [9]), which plots samples on a two-dimensional scatterplot so that distances on the plot approximate the typical log2 fold changes between the samples.

Before producing the MDS plot we will normalize the distributions. We will store the data into s `DGEList` object as this is convenient when running many of the analyses implemented in the edgeR and limma packages. Call the set 'AF', for set 'A', 'Filtered'.

```
AF_dgel <- edgeR::DGEList(
  counts = featureCountsAF,
  genes = genes_annotAF,
  samples = sampDescA,
  group = sampDescA$group
)
AF_dgel <- edgeR::calcNormFactors(AF_dgel)
AF_lcmp_mtx <- edgeR::cpm(AF_dgel, log = T)

# Save AF_dgel to facilitate restarting
# remove from final version
save(list = "AF_dgel", file = "RData/AF_dgel")
```

Verify that the counts are properly normalized.

Proceed with MDS plots.

The MDS plot, which is analogous to a PCA plot adapted to gene exression data, does not indicate strong clustering of samples. The fanning pattern observed in the first two dimensions indicates that a few samples are drifting way from the core set, but in no particular direction. There is some structure in the 3rd and

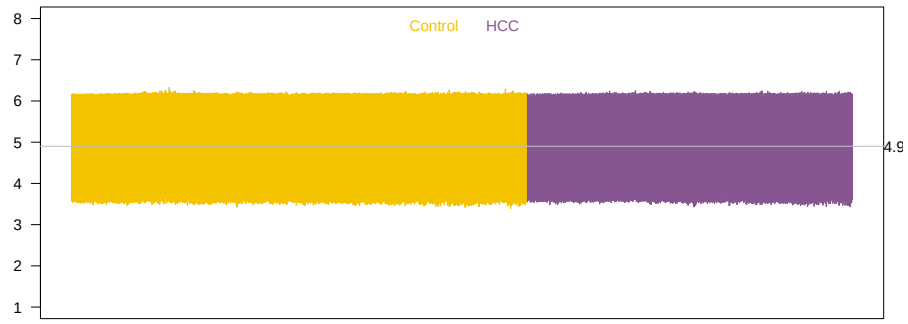


Figure 2.4: Sample log2 count boxplots

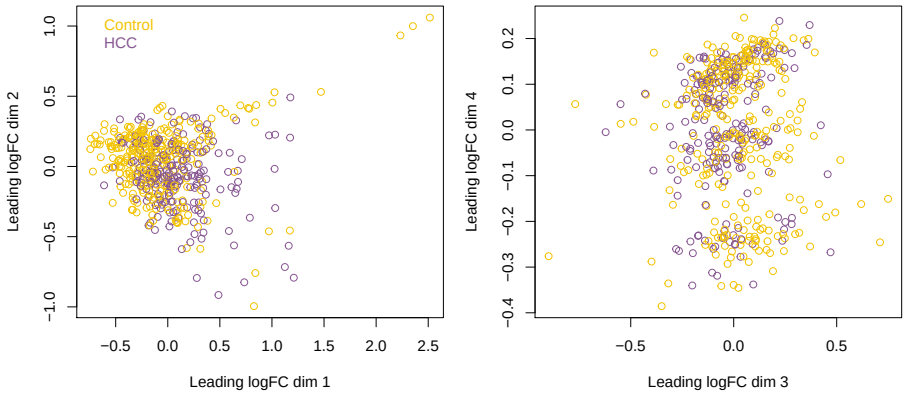


Figure 2.5: MDS plots of log-CPM values

4th dimension plot which should be investigated.
`glMDSPlot` from package `Glimma` provides an interactive MDS plot that can extremely usful for exploration

Link to glMDSPlot: Here

No obvious factor links the samples in the 3 clusters observed on the 4th MDS dimensions. The percent of variance exaplained by this dimension or $\approx 4\%$. The glMDSPlot indicates further segregation along the 6th dimension. The percent of variance exaplained by this dimension or $\approx 2\%$. Tracking down this source of variability may be quite challenging, especially without having the complete information about the sample attributes and provenance.

Unwanted variability is a well-documented problem in the analysis of RNA-Seq data (see Peixoto et al. (2015) [10]), and many procedures have been proposed to reduce the effect of unwanted variation on RNA-Seq analsys results ([10–12]). There are undoubtedly some similar sources of systematic variation in the 5hmC data, but it is beyond the scope of this work to investigate these in this particular dataset. Given that the clustering of samples occurs in MDS dimensions that explain a small fraction of variability, and that these is no assocation with the factor of interest, HCC vs Control, these sources of variability should not interfere too much with our classification analysis. It would nonetheless be interesting to assess whether downstream results can be improved by removing this variability.

Creating a design matrix and contrasts

Before proceeding with the statistical modeling used for the differential expression analysis, we need to set up a model design matrix.

```
## colSums(Design_mtx):
```

```
## Control     HCC
##     778     555
```

```
## Contrasts:
```

```
##          Contrasts
## Levels    HCCvsControl
##   Control           -1
##   HCC                1
```

Removing heteroscedasticity from the count data

As for RNA-Seq data, for 5hmC count data the variance is not independent of the mean. In `limma`, the R package we are using for our analyses, linear modeling is carried out on the log-CPM values which are assumed to be normally distributed and the mean-variance relationship is accommodated using precision weights calculated by the voom function. We apply this transformation next.

```
par(mfrow=c(1,1))
filteredCountsAF_voom <- limma::voom(AF_dgel, Design_mtx, plot=T)
```

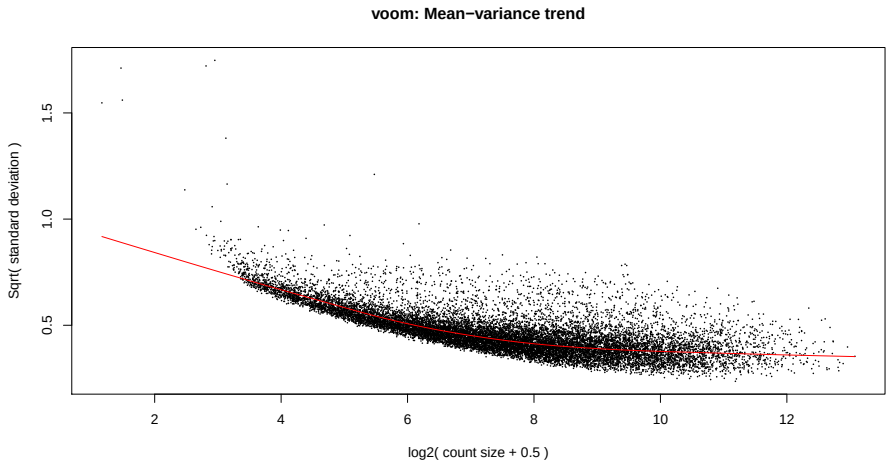


Figure 2.6: Removing heteroscedascity

Note that the voom-plot provides a visual check on the level of filtering performed upstream. If filtering of lowly-expressed genes is insufficient, a drop in variance levels can be observed at the low end of the expression scale due to very small counts.

Fit linear models and examine the results

Having properly filtered and normalized the data, the linear models can be fitted to each gene and the results examined to assess differential expression between the two groups of interest, in our case HCC vs Control.

Table 2.4 displays the counts of genes in each DE category:

Table 2.4: DE Results at FDR = 0.05

	Down	NotSig	Up
HCCvsControl	5214	5280	5258

Table 2.5: log FC quartiles by gene identification

	down	notDE	up
25%	-0.07	-0.01	0.04
50%	-0.05	0.00	0.06
75%	-0.03	0.01	0.09

Graphical representations of DE results: MD Plots

To summarise results for all genes visually, mean-difference plots (aka MA plot), which display log-FCs from the linear model fit against the average log-CPM values can be generated using the plotMD function, with the differentially expressed genes highlighted.

We may also be interested in whether certain gene features are related to gene identification. Gene GC content, for example, might be of interest.

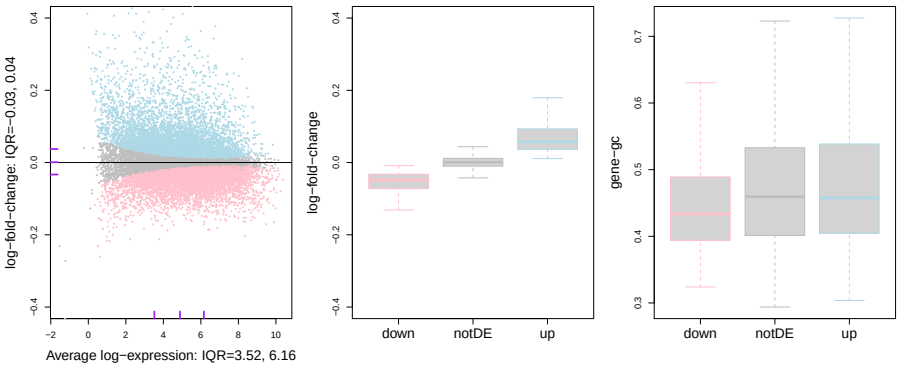


Figure 2.7: HCC vs Control - Genes Identified at FDR = 0,05

While many genes are identified, the effect sizes are quite small, which results in a low signal-to-noise ratio context. See Section 2.3 below.

The log-fold-change distribution for up-represented genes is long-tailed, with many high log fold-change values. By contrast, log-fold-change distribution for down-represented genes closer to symmetric and has few genes with low log

fold-change values. We will see how this affects the results of identifying genes with an effect size requirement.

The GC content of down regulated genes tends to be slightly lower than the rest of the genes. A statistical test would find that the difference between the mean of the down regulated gene population is singificantly different than the mean of the other gene population even though the difference is quite small (-0.028).

These asymmetries are minor, but it would still be good to establish that they relfect biology rather than processing artifacts.

DE genes at 10% fold change

For a stricter definition on significance, one may require log-fold-changes (log-FCs) to be above a minimum value. The treat method (McCarthy and Smyth 2009 [13]) can be used to calculate p-values from empirical Bayes moderated t-statistics with a minimum log-FC requirement. The number of differentially expressed genes are greatly reduced if we impose a minimal fold-change requirement of 10%.

```
## 10% FC Gene Identification Summary - voom, adjust.method = BH, p.value = 0.05:
```

```
##        HCCvsControl
## Down              3
## NotSig        15550
## Up              199
```

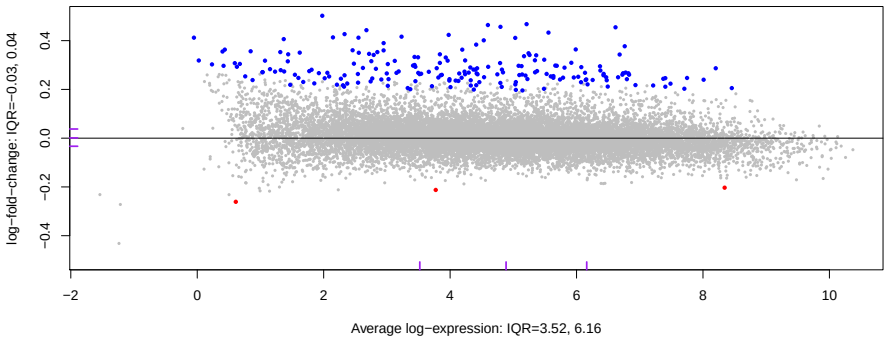


Figure 2.8: HCC vs Control - Identified Genes at FDR = 0,05 and logFC > 10%

Table 2.6: SNR Quantiles

25%	50%	75%	90%
0.018	0.036	0.06	0.082

As noted above, the log-fold-change distribution for the up-represented genes is long-tailes in comparison to log-fold-change distribution for the down-represented genes. As a result fewer down-represented than up-regulated genes are identified when a minimum log-FC requirement is imposed.

2.3 Signal-to-noise ratio regime

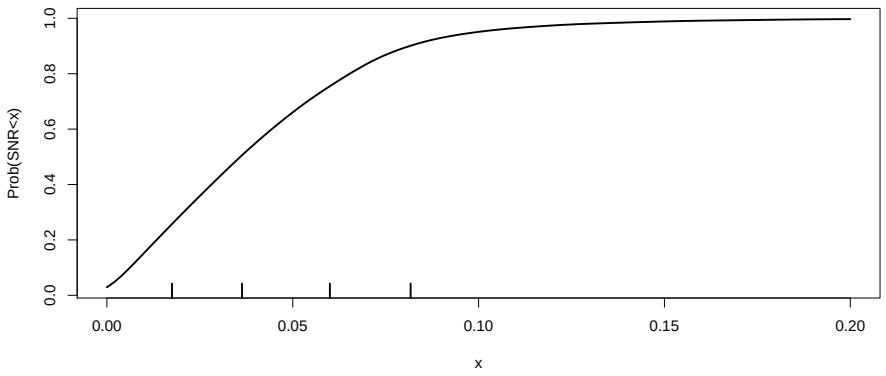
In Hastie et al. (2017) [14]) results from `lasso` fits are compared with `best subset` and `forward selection` fits and it is argued that while `best subset` is optimal for high signal-to-noise regimes, the lasso gains some competitive advantage when the prevailing signal-to-noise ratio of the dataset is lowered.

We can extract sigma and signal from the fit objects to get SNR values for each gene to see in what SNR regime the 5hmC gene body data are.

```
lib.size <- colSums(AF_dgel$counts)

fit <- filteredCountsAF_voom_efit
sx <- fit$Amean + mean(log2(lib.size + 1)) - log2(1e+06)
sy <- sqrt(fit$sigma)

CV <- sy/sx
```



These SNR values are in the range where the lasso and relaxed lasso gain some advantage over best subset and forward selection fits (see Hastie et al. (2017) [14]).

Chapter 3

Modeling - Background

Refer to first pass study for relevant exploratory data analyis results.

In the section we look at some models fitted to discriminate between early stage HCC and healthy and benign samples (grouped as Controls here) from the GSE112679 data set.

- Some questions to addredd with the baseline model
  - how separable are the data: what accuracy do we expect
  - individual sample quality scores: which samples are hard to classify? Compute a score in [0, 1], where 1 is perfectly good classification and 0 is poerfectly bad.

3.1 Predictive modeling for genomic data

The main challenge in calibrating predictive models to genomic data is that there are many more features than there are example cases to fit to; the now classic $n << p$ problem.
In this scenario, fitting methods tend to overfit. The problem can be addressed by selecting variables, regularizng the fit or both. See the Trevor Hastie talk: Statistical Learning with Big Data - Trevor Hastie for a good discussion of this problem and potential solutions.

3.1.1 caret for model evaluation

The `caret` Package provide a set of functions that streamline the process for fitting and evalluating a large numbet of predictive models in parallel. The package contains tools for:

- data splitting
- pre-processing
- feature selection
- model tuning using resampling
- variable importance estimation

The tools facilitate the process of automating randomly spliting data sets into training, testing and evaluating so that predictive models can be evaluated on a comparable and exhaustive basis. Especially useful is the functionality that is provided to repeatedly randomly stratify samples into train and test set so that any sample selection bias is removed.

What makes the `caret` package extremely useful is that a common interface is provided to an exhaustive collection of fitting procedures. Without this common interface, one has to learn the programming interfaces that are used in all fitting procedures to be included in a comparative analysis, which can be quite burdenful.

Some of the models which can be evaluated with caret include: (only some of these can be used with multinomial responses)

- FDA - Flexible Discriminant Analysis
- stepLDA - Linear Discriminant Analysis with Stepwise Feature Selection
- stepQDA - Quadratic Discriminant Analysis with Stepwise Feature Selection
- knn - k nearest neighbors
- pam - Nearest shrunken centroids
- rf - Random forests
- svmRadial - Support vector machines (RBF kernel)
- gbm - Boosted trees
- xgbLinear - eXtreme Gradient Boosting
- xgbTree - eXtreme Gradient Boosting

- neuralnet - neural network

Many more models can be implemented and evaluated with `caret`, including some `deep learning` methods.

`Simulated Annealing Feature Selection` and `Genetic Algorithms`. Many methods found here are also worth investigating.

We only mention `caret` here because it is an exteremely useful tool for anyone interested in comparing many predictive models. We have done that in the past and have found that regularized regression models perform as welll as any in the context of classifiction based on genomic scale data.

3.2 glmnet

In this investigation we will focus on models that can be analyzed with the the `glmnet` R package [15]. Several factors favor this choice:

- the glmnet package is a well supported package providing extensive functionality for regularized regression and classification models.
- the hyper-parameters of the elastic net enable us to explore the relationship between model size, or sparsity, and predictive accuracy. ie. we can investigate the "bet on sparsity" principle: *Use a procedure that does well in sparse problems, since no procedure does well in dense problems.*
- in our experience building classifiers from genomic scale data, regularized classification models using the elastic net penalty do as well as any other, and are more economical in terms of computing time, espacially in comparison to the more exotic boosting algorithms.
- the `lasso` has been shown to be near optimal for the $n << p$ problem over a wide range of signal-to-noise regiments.

Much of the following comes from the Glmnet Vignette.

Glmnet is a package that fits a generalized linear model via penalized maximum likelihood. The regularization path is computed for the lasso or elasticnet penalty at a grid of values for the regularization parameter lambda ([15–18]).

`glmnet` solves the following problem:

$$\min_{\beta_0,\beta} \frac{1}{N} \sum_{i=1}^{N} w_i l(y_i, \beta_0 + \beta^T x_i) + \lambda \left[(1-\alpha)||\beta||_2^2/2 + \alpha||\beta||_1 \right],$$

over a grid of values of λ. Here $l(y, \eta)$ is the negative log-likelihood contribution for observation i; e.g. for the Gaussian case it is $\frac{1}{2}(y - \eta)^2$.

alpha hyper-parameter

The elastic-net penalty is controlled by α, and bridges the gap between lasso (α=1, the default) and ridge (α=0). The tuning parameter λ controls the overall strength of the penalty.

It is known that the ridge penalty shrinks the coefficients of correlated predictors towards each other while the lasso tends to pick one of them and discard the others. The elastic-net penalty mixes these two; if predictors are correlated in groups, an α=0.5 tends to select the groups in or out together. This is a higher level parameter, and users might pick a value upfront, else experiment with a few different values. One use of α is for numerical stability; for example, the *elastic net with* $\alpha = 1 - \epsilon$ *for some small* ϵ*>0 performs much like the lasso, but removes any degeneracies and wild behavior caused by extreme correlations.*

Lasso vs Best Subset

- Best subset selection

$$\min_{\beta \in \mathcal{R}^p} ||Y - X\beta||_2^2 \; subject\, to \; ||\beta||_0 \leq k$$

- lasso

$$\min_{\beta \in \mathcal{R}^p} ||Y - X\beta||_2^2 \; subject\, to \; ||\beta||_1 \leq t$$

- Bertsimas et al. (2016) [19]
  - presented a mixed integer optimization (MIO) formulation for the best subset selection problem
  - Using these MIO solvers, can solve problems with p in the hundreds and even thousands
  - demonstrated that best subset selection generally gives superior prediction accuracy compared to forward stepwise selection and the lasso, over a variety of problem setups.
- Hastie et al. (2017) [14]

  - neither best subset selection nor the lasso uniformly dominate the other, with best subset selection generally performing better in high signal-to-noise (SNR) ratio regimes, and the lasso better in low SNR regimes;
  - best subset selection and forward stepwise perform quite similarly throughout;
  - the relaxed lasso is the overall winner, performing just about as well as the lasso in low SNR scenarios, and as well as best subset selection in high SNR scenarios. We conclude that it is able to use its auxiliary shrinkage parameter (γ) to get the "best of both worlds": it accepts the heavy shrinkage from the lasso when such shrinkage is helpful, and reverses it when it is not.

- relaxed lasso

$$\hat{\beta}^{relax}(\lambda, \gamma) = \gamma\beta^{lasso}(\lambda) + (1 - \gamma)(\beta^{LS}(\lambda)$$

- shrunken relaxed lasso (aka the blended fit)

Suppose the **glmnet** fitted linear predictor at λ is $\hat{\eta}_\lambda(x)$ and the relaxed version is $\tilde{\eta}_\lambda(x)$, then the shrunken relaxed lasso fit is

$$\tilde{\eta}_{\lambda,\gamma}(x) = (1 - \gamma)\tilde{\eta}_\lambda(x) + \gamma\hat{\eta}_\lambda(x)$$

$\gamma \in [0, 1]$ is an additional tuning parameter which can be selected by cross validation.

The debiasing will potentially improve prediction performance, and CV will typically select a model with a smaller number of variables. This procedure is very competitive with forward-stepwise and best-subset regression, and has a considerable speed advantage when the number of variables is large. This is especially true for best-subset, but even so for forward stepwise. The latter has to plod through the variables one-at-a-time, while glmnet will just plunge in and find a good active set.

Further details may be found in Friedman, Hastie, and Tibshirani (2010), Tibshirani et al. (2012), Simon et al. (2011), Simon, Friedman, and Hastie (2013) and Hastie, Tibshirani, and Tibshirani (2017) ([14–18]).

- SNR
  - $y_0 = f(x_0) + \epsilon_0$
  - $SNR = \frac{var(f(x_0))}{var(\epsilon_0)}$
  - $PVE(g) = 1 - \frac{\mathbb{E}(y_0 - g(x_0))^2}{Var(y_0)}$

- $PVE(f) = 1 - \frac{Var(\epsilon_0)}{Var(y_0)} = \frac{SNR}{1+SNR}$
- $SNR = \frac{PVE}{1-PVE}$
- $c_v = \frac{\sigma}{\mu} = \frac{Var(y)}{\mathbb{E}(y)}$
    - a PVE of 0.5 is rare for noisy observational data, and 0.2 may be more typical
    - A PVE of 0.86, corresponding to an SNR of 6, is unheard of!
    - For small SNR, SNR $\approx$ PVE
    - See Xiang et al. (2020) [20], Lozoya et al. (2018) [21], Simonson et al. (2018) [22] and Rapaport et al. (2013) [23] for SNR in RNA-Seq

Chapter 4

The bet on sparsity

In this section we explore various fits that can be computed ans analyzed with tools provided in the `glmnet` package. Refer to the Glmnet Vignette for a quick reference guide.

4.1 CV analysis setup

```
K_FOLD <- 10
trainP <- 0.8
EPS <- 0.02    # Have no idea what "small" epsilon means
```

First we divide the analysis dataset into `train` and `test` in a 4:1 ratio.

```
set.seed(1)
train_sampID_vec <- with(AF_dgel$samples,
AF_dgel$samples$sampID[caret::createDataPartition(y=group, p=trainP, list=F)]
)

test_sampID_vec <- with(AF_dgel$samples,
setdiff(sampID, train_sampID_vec)
)

train_group_vec <- AF_dgel$samples[train_sampID_vec, 'group']
test_group_vec <- AF_dgel$samples[test_sampID_vec, 'group']

knitr::kable(table(train_group_vec),
  caption="Train set") %>%
   kableExtra::kable_styling(full_width = F)
```

Table 4.1: Train set

train_group_vec	Freq
Control	623
HCC	444

Table 4.2: Test set

test_group_vec	Freq
Control	155
HCC	111

```
knitr::kable(table(test_group_vec),
  caption="Test set") %>%
   kableExtra::kable_styling(full_width = F)
```

```
train_lcpm_mtx <- t(lcpm_mtx[,train_sampID_vec])
test_lcpm_mtx <- t(lcpm_mtx[,test_sampID_vec])
```

We explore some glmnet fits and the "bet on sparsity"

- Consider models:
  - lasso: $\alpha = 1.0$ - sparse model
  - ridge $\alpha = 0$ - shrunken coefficients model
  - elastic net: $\alpha = 0.5$ - semi sparse model
- Does the relaxed lasso improve performance?
- Does the shrunken relaxed lasso (aka the blended mix) improve performance
- How sparse is the model undelying best 5hmC classifier for Early HCC vs Control?
- Is the degree of sparsity, or the size of the model, a stable feature of the problem and data set?

In this analysis, we will only evaluate models in terms of model size, stability and performance. We leave the question of significance testing of hypotheses about model parameters completely out. See Lockhart et al. (2014) [24] and Wassermam (2014) [25] for a discussion of this topic.

Table 4.3: training samples fold composition

	1	2	3	4	5	6	7	8	9	10
Control	62	62	62	63	62	62	63	62	63	62
HCC	45	44	44	45	44	45	44	45	44	44

Next we create folds for 10-fold cross-validation of models fitted to training data. We'll use caret::createFolds to assign samples to folds while keeping the outcome ratios constant across folds.

```
# This is too variable, both in terms of fold size And composition
#foldid_vec <- sample(1:10, size=length(train_group_vec), replace=T)

set.seed(1)
train_foldid_vec <- caret::createFolds(
 factor(train_group_vec),
 k=K_FOLD,
 list=F)

knitr::kable(sapply(split(train_group_vec, train_foldid_vec),
  table), caption="training samples fold composition") %>%
   kableExtra::kable_styling(full_width = F)
```

Note that the folds identify samples that are left-out of the training data for each fold fit.

4.2 Fit and compare models

- cross-validated accuracy
- test set accuracy

- sparsity
  - for lasso, enet , examine number of selected variables

Although "the one standard error rule" can produce a model with fewer predictors, it usually results in increased MSE and more biased parameter estimates (see Engebretsen et al. (2019) [26] for example). We will look at both the minimum cv error and the one standard error rule model preformance.

4.2.1 Logistic regression in `glmnet`

`glmnet` provides functionality to extract various predicted of fitted values from calibrated models. Note in passsing that some folks make a distinction between

fitted or **estimated** values for sample points in the training data vurses **predicted** values for sample points that are not in the training dataset. `glmnet` makes no such distinction and the `predict` function is used to produce both fitted as well as predicted values. For logistic regressions, which is the model fitted in a regularized fashion when models are fitted by glmnet with the parameter `family='binomial'`, three fitted or predicted values can be extracted at a given design point.
Suppose our response variable Y is either 0 or 1 (Control or HCC in our case). These are specified by the `type` parameter. `type='resp'` returns the fitted or predicted probability of $Y = 1$. `type='class'` returns the fitted or predicted class for the design point, which is simply dichotomizing the response: class = 1 if the fitted or predicted probability is greater than 0.5 (check to make sure class is no the Bayes estimate). `type='link'` returns the fitted or predicted vealue of the linear predictor $\beta' x$. The relationship between the linear predictor and the response can be derided from the logis toc regression model:

$$P(Y = 1|x, \beta) = g^{-1}(\beta' x) = h(\beta' x) = \frac{e^{\beta' x}}{1 + e^{\beta' x}}$$

where g is the link function, g^{-1} the mean function. The link function is given by:

$$g(y) = h^{-1}(y) = ln(\frac{y}{1-y})$$

This link function is called the logit function, and its inverse the logistic function.

```
logistic_f <- function(x) exp(x)/(1+exp(x))
```

It is important to note that all *predicted* values extracted from `glmnet` fitted models by the **predict()** extraction method yield **fitted** values for design points that are part of the training data set. This includes the predicted class for training data which ae used to estimate misclassification error rates. As a result, the cv error rates quoted in various `glmnet` summaries are generally optimistic. `glmnet` fitting functions have a parameter, *keep*, which instructs the fitting function to keep the `out-of-fold`, or `prevalidated`, predictions as part of the returned object. The `out-of-fold` predictions are predicted values for the samples in the left-out folds, pooled across all cv folds. For each hyper-parameter specification, we get one full set of `out-of-fold` predictions for the training set samples. Performance assessments based on these values are usually more generalizable - ie. predictive of performance in unseen data - than assessments based on values produced from the full fit, which by default is what `glmnet` extraction methods provide. See Höfling and Tibshirani (2008) [27] for a description of the use of pre-validation in model assessment.

Because the `keep=T` option will store predicted values for all models evaluated in the cross-validation process, we will limit the number of models tested by setting **nlambda=30** when calling the fitting functions. This has no effect on performance in this data set.

```
start_time <-  proc.time()

cv_lasso <- glmnet::cv.glmnet(
 x=train_lcpm_mtx,
 y=train_group_vec,
 foldid=train_foldid_vec,
 alpha=1,
 family='binomial',
 type.measure = "class",
 keep=T,
 nlambda=30
)

message("lasso time: ", round((proc.time() - start_time)[3],2),"s")
```

```
## lasso time: 12.59s
```

```
start_time <-  proc.time()

cv_ridge <- glmnet::cv.glmnet(
 x=train_lcpm_mtx,
 y=train_group_vec,
 foldid=train_foldid_vec,
 alpha=0,
 family='binomial',
 type.measure = "class",
 keep=T,
 nlambda=30
)

message("ridge time: ", round((proc.time() - start_time)[3],2),"s")
```

```
## ridge time: 104.08s
```

```
start_time <-  proc.time()

cv_enet <- glmnet::cv.glmnet(
 x=train_lcpm_mtx,
 y=train_group_vec,
```

```
 foldid=train_foldid_vec,
 alpha=0.5,
 family='binomial',
 type.measure = "class",
 keep=T,
 nlambda=30
)

message("enet time: ", round((proc.time() - start_time)[3],2),"s")
```

```
## enet time: 12.1s
```

The ridge regression model takes over 10 times longer to compute.

Examine model performance.

```
## Warning: package 'glmnet' was built under R version 4.0.2
```

```
## Loading required package: Matrix
```

```
## Loaded glmnet 4.0-2
```

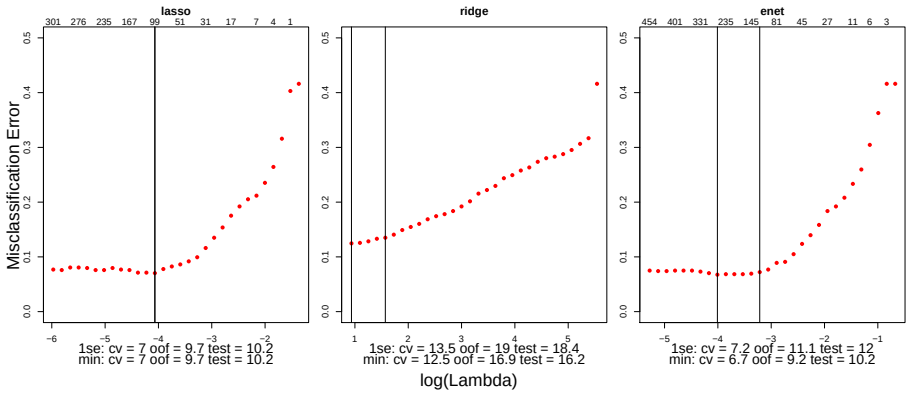


Figure 4.1: compare fits

```
errors_frm <- data.frame(
  lasso = lasso_errors_mtx, ridge = rifge_errors_mtx, enet = enet_errors_mtx
)
```

Table 4.4: Misclassifiaction error rates

	train_cv	train_oof	test
lasso.error_1se	7.0	9.7	10.2
lasso.error_min	7.0	9.7	10.2
ridge.error_1se	13.5	19.0	18.4
ridge.error_min	12.5	16.9	16.2
enet.error_1se	7.2	11.1	12.0
enet.error_min	6.7	9.2	10.2

```
knitr::kable(t(errors_frm)*100,
 caption = 'Misclassifiaction error rates',
 digits=1) %>%
  kableExtra::kable_styling(full_width = F)
```

We see that the lasso and enet models do better than the ridge model. These is very little difference between the min lambda and the the one standard error rule lambda models (the two are the same for the lasso in this data set). We also see that the training data out-of-fold estimates of misclassification error rates are much closer to the test set estimates than are the cv estimated rates. This has been our experience with regularized regression models fitted to genomic scale data. It should also be noted thay the cv estimates of misclassification rates become more biased as the sample size decreases, as we will show in Section 6.

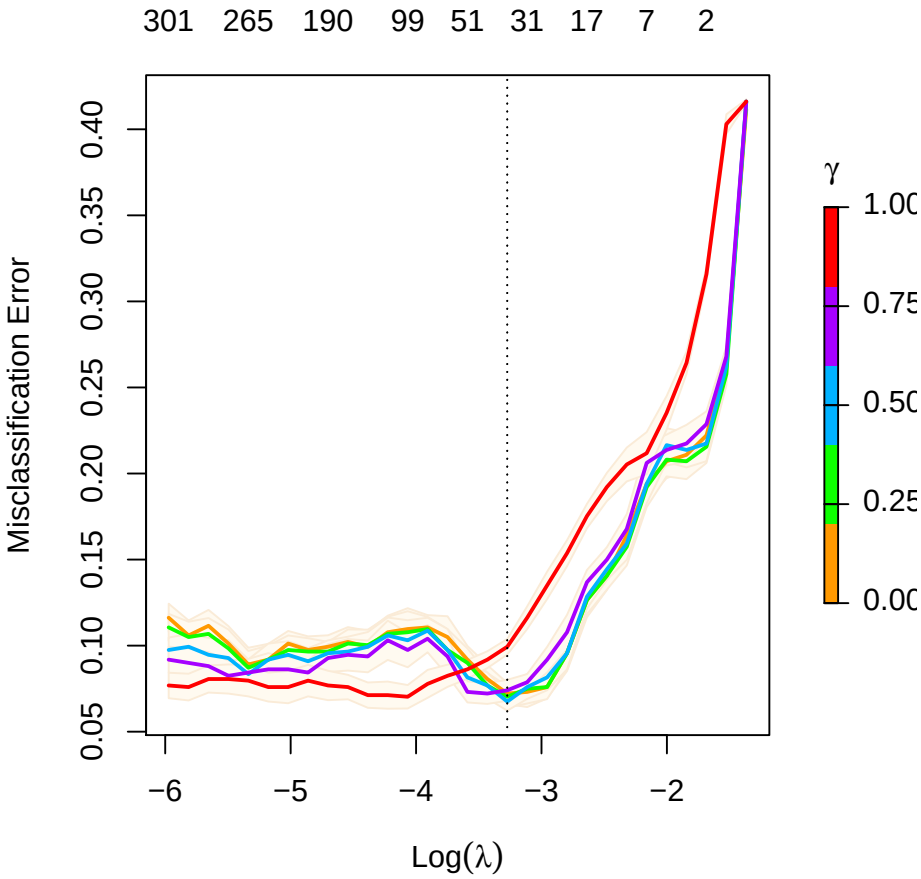
4.3 Relaxed lasso and blended mix

Next we look at the so-called `relaxed lasso` and the `blended mix` which is an optimized shrinkage between the relaxed lasso and the regular lasso.

```
##
## Call:  glmnet::cv.glmnet(x = train_lcpm_mtx, y = train_group_vec, type.measure = "class",
##
## Measure: Misclassification Error
##
##     Gamma Lambda Measure       SE Nonzero
## min   0.5 0.0379 0.06748 0.005274      35
## 1se   0.5 0.0379 0.06748 0.005274      35
```

Table 4.5: Relaxed lasso and blended mix error rates

train_blended_cv	6.7
train_blended_oof	10.2
train_relaxed_oof	11.1
test_blended_oof	10.2
test_relaxed_oof	10.9



The relaxed lasso and blended mix error rates are comparable to the regular lasso fit error rate. We see here too that the reported cv error rates are quite optimistic, while out-of-fold error rates continue to be good indicaters of unseen data error rates.

4.4 Examination of sensitivity vs specificity

In the results above we reported error rates without inspecting the sensitivity vurses specificity trade off. Here we look at this question with the help of ROC curves.

4.4.1 Training data out-of-fold ROC curves

```
# train
# lasso
ndx_1se <- match(cv_lasso$lambda.1se,cv_lasso$lambda)
train_lasso_oofProb_vec <- logistic_f(cv_lasso$fit.preval[,ndx_1se])
train_lasso_roc <- pROC::roc(
 response = as.numeric(train_group_vec=='HCC'),
 predictor = train_lasso_oofProb_vec)
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
# lasso - relaxed
ndx_1se <- match(cv_lassoR$lambda.1se,cv_lassoR$lambda)
train_lassoR_oofProb_vec <- logistic_f(cv_lassoR$fit.preval[['g:0']][,ndx_1se])
train_lassoR_roc <- pROC::roc(
 response = as.numeric(train_group_vec=='HCC'),
 predictor = train_lassoR_oofProb_vec)
```

```
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```

```
# blended mix (gamma=0.5)
ndx_1se <- match(cv_lassoR$lambda.1se,cv_lassoR$lambda)
train_blended_oofProb_vec <- logistic_f(cv_lassoR$fit.preval[['g:0.5']][,ndx_1se])
train_blended_roc <- pROC::roc(
 response = as.numeric(train_group_vec=='HCC'),
 predictor = train_blended_oofProb_vec)
```

```
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```

```
plot(train_lasso_roc)
lines(train_lassoR_roc, col='blue')
lines(train_blended_roc, col='green')

legend('bottomright', title='AUC',
 legend=c(
  paste('lasso =', round(train_lasso_roc[['auc']],3)),
  paste('lassoR =', round(train_lassoR_roc[['auc']],3)),
  paste('blended =', round(train_blended_roc[['auc']],3))
 ),
 text.col = c('black', 'blue', 'green'))
```

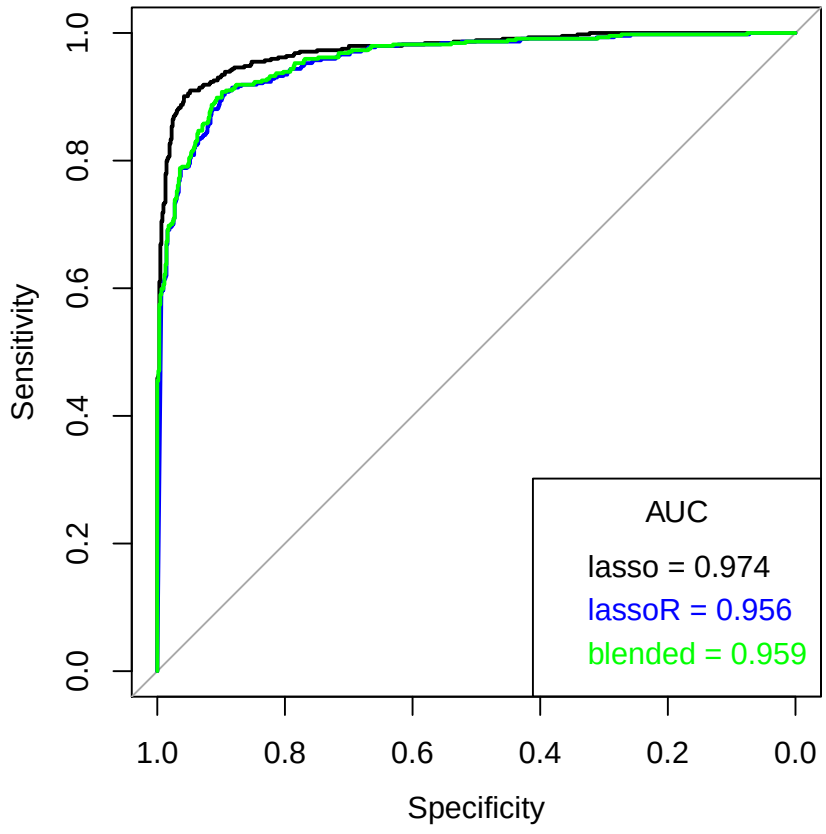


Figure 4.2: Train data out-of-sample ROCs

Table 4.6: Specificity = .90 Coordinates

	threshold	specificity	sensitivity
lasso	0.337	0.9	0.932
lassoR	0.003	0.9	0.894
blended	0.031	0.9	0.899

Compare thresholds for 90% Specificity:

```
lasso_ndx <- with(as.data.frame(pROC::coords(train_lasso_roc, transpose=F)),
  min(which(specificity >= 0.9)))

lassoR_ndx <- with(as.data.frame(pROC::coords(train_lassoR_roc, transpose=F)),
  min(which(specificity >= 0.9)))

blended_ndx <- with(as.data.frame(pROC::coords(train_blended_roc, transpose=F)),
  min(which(specificity >= 0.9)))

 spec90_frm <- data.frame(rbind(
 lasso=as.data.frame(pROC::coords(train_lasso_roc, transpose=F))[lasso_ndx,],
 lassoR=as.data.frame(pROC::coords(train_lassoR_roc, transpose=F))[lassoR_ndx,],
 blended=as.data.frame(pROC::coords(train_blended_roc, transpose=F))[blended_ndx,]
))


knitr::kable(spec90_frm,
  digits=3,
  caption="Specificity = .90 Coordinates"
) %>%
  kableExtra::kable_styling(full_width = F)
```

This is strange.

```
par(mfrow=c(1,3))

# lasso
plot(density(train_lasso_oofProb_vec[train_group_vec=='Control']),
 xlim=c(0,1),main='', xlab='', col='green')
lines(density(train_lasso_oofProb_vec[train_group_vec=='HCC']),
   co='red')
title("lasso")

# lassoR
plot(density(train_lassoR_oofProb_vec[train_group_vec=='Control']),
```

```
 xlim=c(0,1),main='', xlab='', col='green')
lines(density(train_lassoR_oofProb_vec[train_group_vec=='HCC']),
   co='red')
title("lassoR")

sapply(split(train_lassoR_oofProb_vec,train_group_vec), summary)
```

```
##              Control          HCC
## Min.    1.328376e-69 3.337696e-41
## 1st Qu. 6.241943e-28 1.000000e+00
## Median  8.812711e-19 1.000000e+00
## Mean    7.668147e-02 8.481142e-01
## 3rd Qu. 4.672541e-10 1.000000e+00
## Max.    1.000000e+00 1.000000e+00
```

```
# blended
plot(density(train_blended_oofProb_vec[train_group_vec=='Control']),
 xlim=c(0,1),main='', xlab='', col='green')
lines(density(train_blended_oofProb_vec[train_group_vec=='HCC']),
   co='red')
title("blended")
```

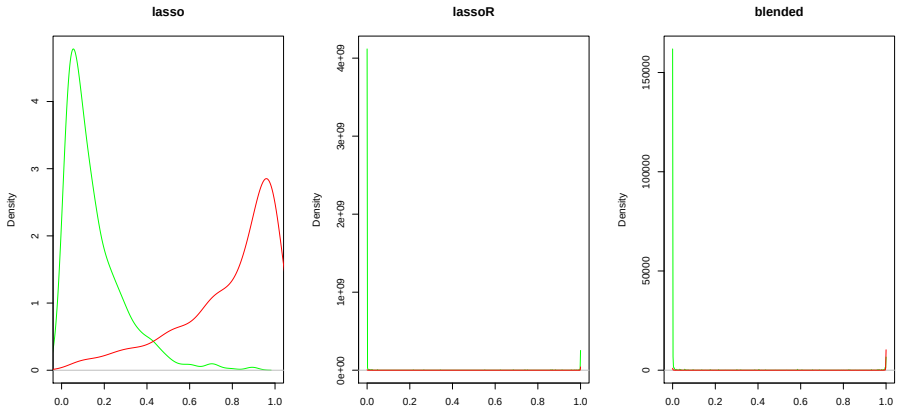


Figure 4.3: Train data out-of-fold predicted probabilities

This makes no sense.

Look at test data ROC.

```
# train
# lasso
test_lasso_oofProb_vec <- predict(
 cv_lasso,
 type='resp',
 lambda='1se',
 newx=test_lcpm_mtx
)

test_lasso_roc <- pROC::roc(
 response = as.numeric(test_group_vec=='HCC'),
 predictor = test_lasso_oofProb_vec)
```

```
## Setting levels: control = 0, case = 1
```

```
## Warning in roc.default(response = as.numeric(test_group_vec == "HCC"), predictor
## = test_lasso_oofProb_vec): Deprecated use a matrix as predictor. Unexpected
## results may be produced, please pass a numeric vector.
```

```
## Setting direction: controls < cases
```

```
# lassoR
test_lassoR_oofProb_vec <- predict(
 cv_lassoR,
 type='resp',
 lambda='1se',
 newx=test_lcpm_mtx,
 gamma=0,
)

test_lassoR_roc <- pROC::roc(
 response = as.numeric(test_group_vec=='HCC'),
 predictor = test_lassoR_oofProb_vec)
```

```
## Setting levels: control = 0, case = 1
```

```
## Warning in roc.default(response = as.numeric(test_group_vec == "HCC"), predictor
## = test_lassoR_oofProb_vec): Deprecated use a matrix as predictor. Unexpected
## results may be produced, please pass a numeric vector.
```

```
## Setting direction: controls < cases
```

```
# blended mix (gamma=0.5)
test_blended_oofProb_vec <- predict(
 cv_lassoR,
 type='resp',
 lambda='1se',
 newx=test_lcpm_mtx,
 gamma=0.5,
)

test_blended_roc <- pROC::roc(
 response = as.numeric(test_group_vec=='HCC'),
 predictor = test_blended_oofProb_vec)
```

```
## Setting levels: control = 0, case = 1
```

```
## Warning in roc.default(response = as.numeric(test_group_vec == "HCC"), predictor
## = test_blended_oofProb_vec): Deprecated use a matrix as predictor. Unexpected
## results may be produced, please pass a numeric vector.
```

```
## Setting direction: controls < cases
```

```
plot(test_lasso_roc)
lines(test_lassoR_roc, col='blue')
lines(test_blended_roc, col='green')

legend('bottomright', title='AUC',
 legend=c(
  paste('lasso =', round(test_lasso_roc[['auc']],3)),
  paste('lassoR =', round(test_lassoR_roc[['auc']],3)),
  paste('blended =', round(test_blended_roc[['auc']],3))
 ),
 text.col = c('black', 'blue', 'green'))
```

Look at desnities of predicted probabilities.

```
par(mfrow=c(1,3))

# lasso
plot(density(test_lasso_oofProb_vec[test_group_vec=='Control']),
 xlim=c(0,1),main='', xlab='', col='green')
lines(density(test_lasso_oofProb_vec[test_group_vec=='HCC']),
   co='red')
title("lasso")
```

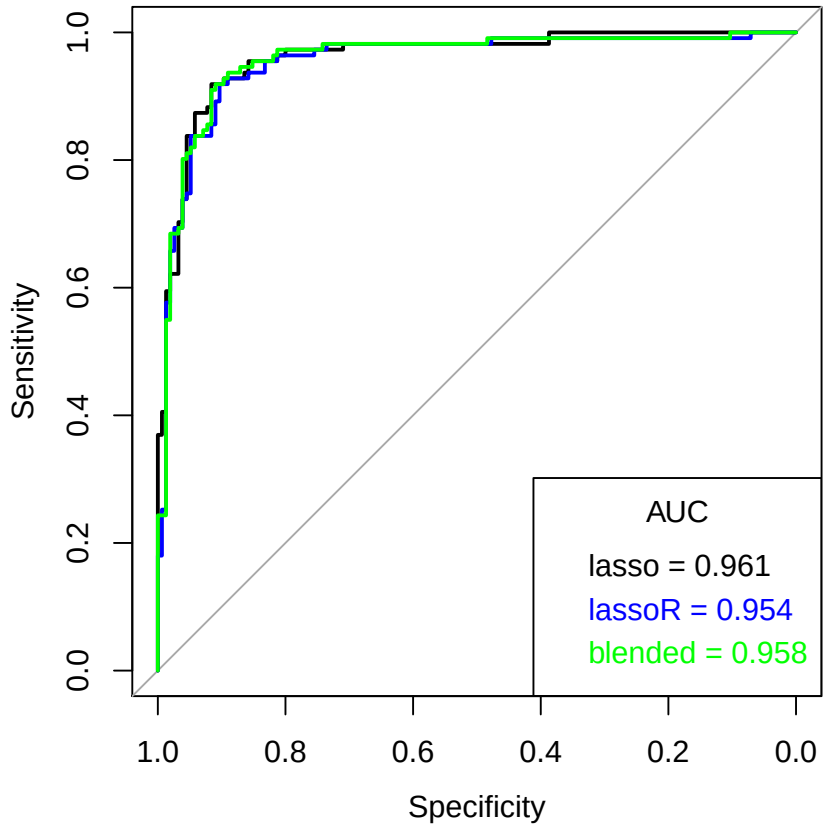


Figure 4.4: Test data out-of-sample ROCs

```
# lassoR
plot(density(test_lassoR_oofProb_vec[test_group_vec=='Control']),
 xlim=c(0,1),main='', xlab='', col='green')
lines(density(test_lassoR_oofProb_vec[test_group_vec=='HCC']),
   co='red')
title("lassoR")

sapply(split(test_lassoR_oofProb_vec,test_group_vec), summary)
```

```
##              Control          HCC
## Min.    2.352635e-09 4.183949e-05
## 1st Qu. 4.850862e-04 8.373417e-01
## Median  3.050687e-03 9.925541e-01
## Mean    9.277576e-02 8.424789e-01
## 3rd Qu. 3.704167e-02 9.999439e-01
## Max.    9.999883e-01 1.000000e+00
```

```
# blended
plot(density(test_blended_oofProb_vec[test_group_vec=='Control']),
 xlim=c(0,1),main='', xlab='', col='green')
lines(density(test_blended_oofProb_vec[test_group_vec=='HCC']),
   co='red')
title("blended")
```

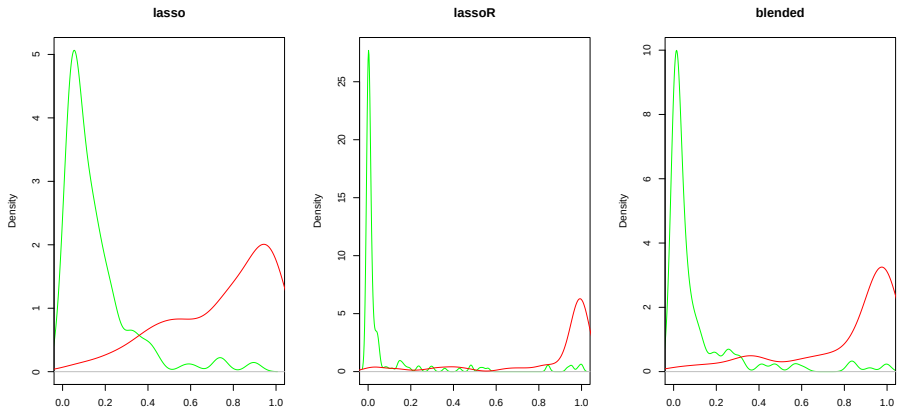


Figure 4.5: Test data out-of-fold predicted probabilities

```
# Train - preval is out-of-fold linear predictor for training design points
onese_ndx <- match(cv_lasso$lambda.1se,cv_lasso$lambda)
train_1se_preval_vec <- cv_lasso$fit.preval[,onese_ndx]
train_1se_predProb_vec <- logistic_f(train_1se_preval_vec)

#Test
test_1se_predProb_vec <- predict(
 cv_lasso,
 newx=test_lcpm_mtx,
 s="lambda.1se",
 type='resp'
)

tmp <- c(
 train=split(train_1se_predProb_vec, train_group_vec),
 test=split(test_1se_predProb_vec, test_group_vec))
names(tmp) <- sub('\\.','\t',names(tmp))


boxplot(tmp)
```

```
## Warning in axis(side = 1, at = 1:4, labels = c("train\tControl", "train\tHCC", :
## font width unknown for character 0x9

## Warning in axis(side = 1, at = 1:4, labels = c("train\tControl", "train\tHCC", :
## font width unknown for character 0x9

## Warning in axis(side = 1, at = 1:4, labels = c("train\tControl", "train\tHCC", :
## font width unknown for character 0x9

## Warning in axis(side = 1, at = 1:4, labels = c("train\tControl", "train\tHCC", :
## font width unknown for character 0x9

## Warning in axis(side = 1, at = 1:4, labels = c("train\tControl", "train\tHCC", :
## font width unknown for character 0x9

## Warning in axis(side = 1, at = 1:4, labels = c("train\tControl", "train\tHCC", :
## font width unknown for character 0x9

## Warning in axis(side = 1, at = 1:4, labels = c("train\tControl", "train\tHCC", :
## font width unknown for character 0x9

## Warning in axis(side = 1, at = 1:4, labels = c("train\tControl", "train\tHCC", :
## font width unknown for character 0x9
```

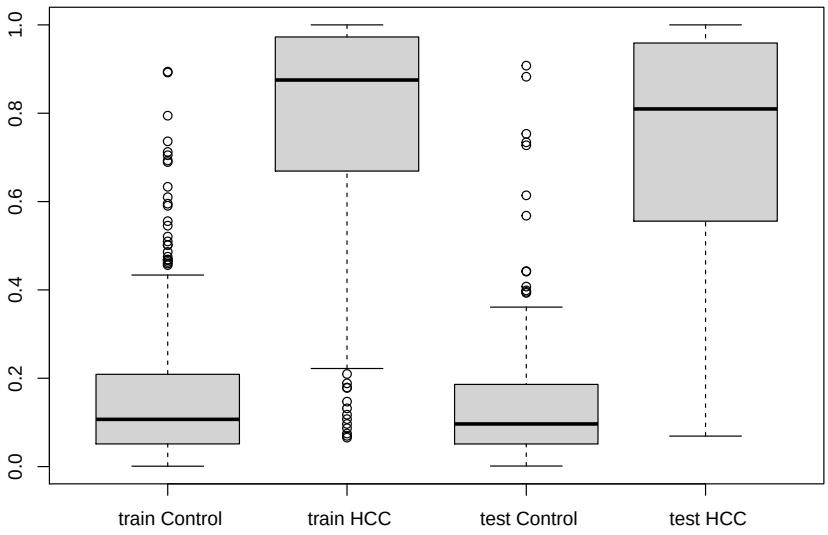


Figure 4.6: Predicted Probabilities - Train and Test

Table 4.7: cv lasso confusion matrix: train set

	Control	HCC
Control	615	31
HCC	8	413

Table 4.8: cv lassoR confusion matrix: train set

	Control	HCC
Control	607	34
HCC	16	410

Table 4.9: cv lasso confusion matrix: test set

	Control	HCC
Control	148	22
HCC	7	89

Table 4.10: cv lassoR confusion matrix: test set

	Control	HCC
Control	146	17
HCC	9	94

4.5 Refit with "auc" as optimization

```
start_time <-  proc.time()

cv_lasso2 <- glmnet::cv.glmnet(
 x=train_lcpm_mtx,
 y=train_group_vec,
 foldid=train_foldid_vec,
 alpha=1,
 family='binomial',
 type.measure = "auc")

message("lasso time: ", round((proc.time() - start_time)[3],2),"s")
```

```
## lasso time: 23.72s
```

```
start_time <-  proc.time()

cv_ridge2 <- glmnet::cv.glmnet(
 x=train_lcpm_mtx,
 y=train_group_vec,
 foldid=train_foldid_vec,
 alpha=0,
 family='binomial',
 type.measure = "auc")

message("ridge time: ", round((proc.time() - start_time)[3],2),"s")
```

```
## ridge time: 268.22s
```

```
start_time <-  proc.time()

cv_enet2 <- glmnet::cv.glmnet(
 x=train_lcpm_mtx,
 y=train_group_vec,
```

```
 foldid=train_foldid_vec,
 alpha=0.5,
 family='binomial',
 type.measure = "auc")

message("enet time: ", round((proc.time() - start_time)[3],2),"s")
```

```
## enet time: 23.57s
```

```
start_time <-  proc.time()

cv_lassoC2 <-  glmnet::cv.glmnet(
 x=train_lcpm_mtx,
 y=train_group_vec,
 foldid=train_foldid_vec,
 alpha=1-EPS,
 family='binomial',
 type.measure = "class")

message("lassoC time: ", round((proc.time() - start_time)[3],2),"s")
```

```
## lassoC time: 21.52s
```

The ridge regression model takes over 10 times longer to compute.

Examine model performance.

```
## Setting levels: control = Control, case = HCC

## Setting direction: controls < cases

## Setting levels: control = Control, case = HCC

## Setting direction: controls < cases

## Setting levels: control = Control, case = HCC

## Setting direction: controls < cases

## Setting levels: control = Control, case = HCC

## Setting direction: controls < cases
```

```
## Setting levels: control = Control, case = HCC

## Setting direction: controls < cases

## Setting levels: control = Control, case = HCC

## Setting direction: controls < cases
```

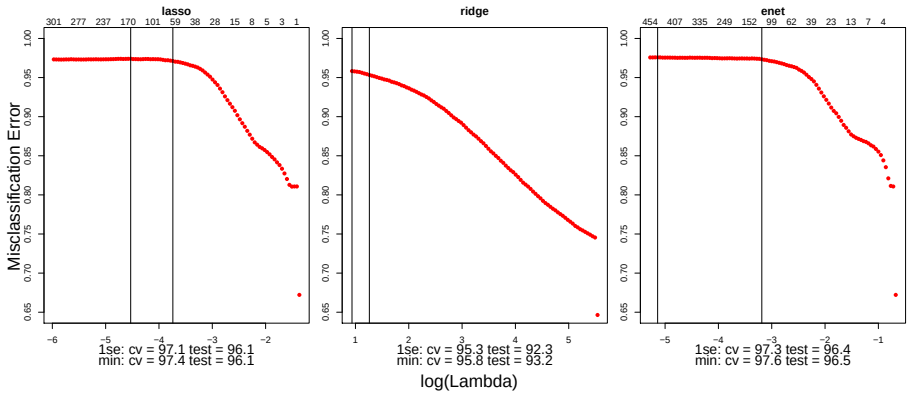


Figure 4.7: compare fits

All models produce cv assessments of Misclassification Error that are slightly better than the test set assessments. `lasso` performs comparably to `enet` and better than the `ridge` model.

4.6 Relaxed lasso and blended mix

```
library(glmnet)

cv_lassoR2_sum <- print(cv_lassoR2)
```

```
##
## Call:  glmnet::cv.glmnet(x = train_lcpm_mtx, y = train_group_vec, type.measure = "auc",      f
##
## Measure: AUC
##
##      Gamma  Lambda Measure       SE Nonzero
## min   0.25 0.04153  0.9765 0.003813      32
## 1se   0.50 0.04558  0.9729 0.003849      31
```

```
plot(cv_lassoR2)
```

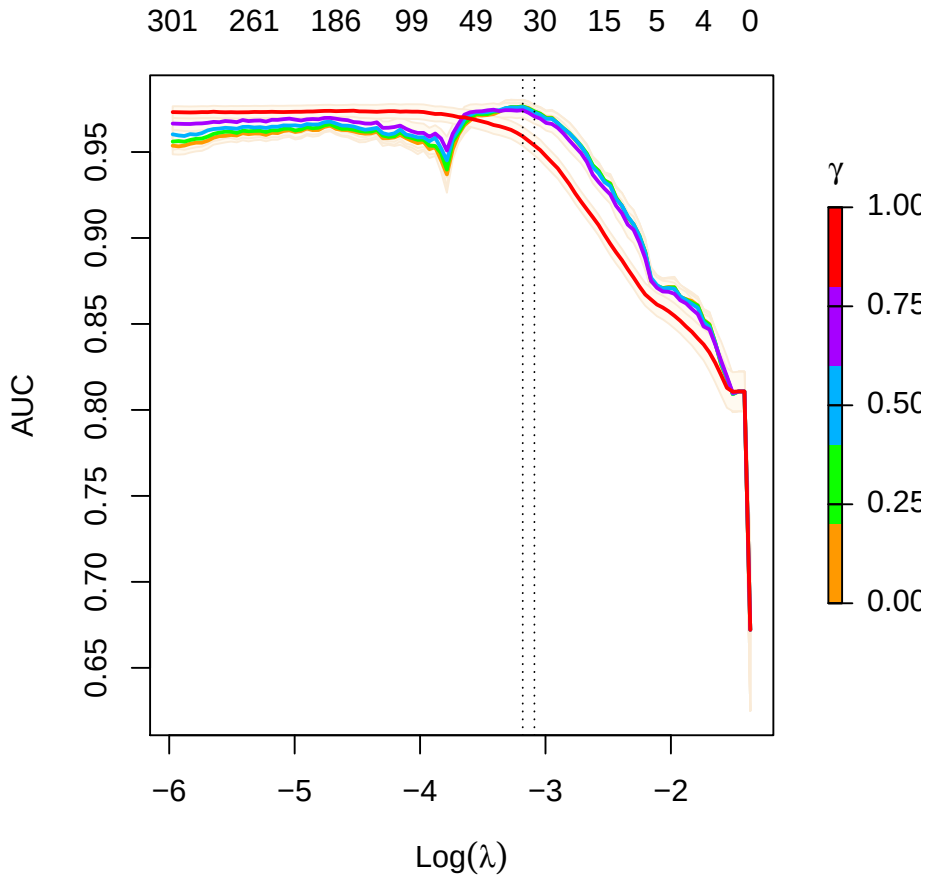


Figure 4.8: lassoR fit

```
test_pred_1se_vec <- predict(
  cv_lassoR2,
  newx=test_lcpm_mtx,
  s="lambda.1se",
  type="response"
)
test_pred_1se_auc <- suppressWarnings(pROC::auc(test_group_vec,test_pred_1se_vec)[1])
```

```
## Setting levels: control = Control, case = HCC
```

Table 4.11: CV vs test Errors

	cv_eucg	test_auc
onese	97.3	95.8
min	97.7	95.6

Table 4.12: cv lasso confusion matrix: train set

	Control	HCC
Control	614	48
HCC	9	396

```
## Setting direction: controls < cases
```

```
test_pred_min_vec <- predict(
  cv_lassoR2,
  newx=test_lcpm_mtx,
  s="lambda.min",
  type="response"
)
test_pred_min_auc <- suppressWarnings(pROC::auc(test_group_vec,test_pred_min_vec)[1])
```

```
## Setting levels: control = Control, case = HCC
## Setting direction: controls < cases
```

```
cv_lassoR2_1se_auc <- cv_lassoR2_sum['1se','Measure']
cv_lassoR2_min_auc <- cv_lassoR2_sum['min','Measure']

knitr::kable(rbind(
 onese=c(cv_eucg=cv_lassoR2_1se_auc, test_auc=test_pred_1se_auc)*100,
 min=c(cv_eucg=cv_lassoR2_min_auc, test_auc=test_pred_min_auc)*100
),
 caption="CV vs test Errors", digits=1
) %>%
  kableExtra::kable_styling(full_width = F)
```

Look at confusion matrices

In all models the sensitivity weak compared to the specificity. Let's examine the ROC curves to see where the trade-off is.

Table 4.13: cv lassoR2 confusion matrix: train set

	Control	HCC
Control	608	39
HCC	15	405

Table 4.14: cv lasso confusion matrix: test set

	Control	HCC
Control	148	25
HCC	7	86

Table 4.15: cv lassoR2 confusion matrix: test set

	Control	HCC
Control	147	19
HCC	8	92

Chapter 5

Examine feature selection

Recall

```
It is known that the ridge penalty shrinks the coefficients of correlated predictors
towards each other while the lasso tends to pick one of them and discard the others.
The elastic-net penalty mixes these two; if predictors are correlated in groups,
an $\alpha$=0.5 tends to select the groups in or out together.
This is a higher level parameter, and users might pick a value upfront,
else experiment with a few different values. One use of $\alpha$ is for numerical stability;
for example, the *elastic net with $\alpha = 1 - \epsilon$ for some small $\epsilon$>0
performs much like the lasso, but removes any degeneracies and wild behavior caused
by extreme correlations*.
```

5.1 Sparsity stability

Chapter 6

Fitted Model Suite

We examine the results of fitting a suite of models to investigate the effect of sample size on model performance.

Predicted probabilities can be transformed into a sample quality score: $Q_i = p_i^{y_i}(1-p_i)^{1-y_i}$, where p_i is the out-of-fold estimated probability of HCC for sample i and y_i is 1 for HCC samples and 0 for Controls. ie. we use the fitted likelihood as a sample quality score. The quality scores derived from a particular cv run will depend to some extent on the random assignment of samples to folds. To remove this dependency, we can derive quality scores by averaging over many cv runs, 30 say. Hard to classify samples will have low quality scores. In the results that we discuss below, when we look at variability across repeated random sampling of different sizes, we can use sample quality scores to investigate how much of the variability is due to sample selection. Note that quality here is not used to say anything about the sample data quality. Low quality here only means that a sample is different from the core of the data set in a way that makes it hard to properly classify. That could happen if the sample were mislabeled, in which case we could think of this sample as being poor quality of course.

Chapter 7

Conclusions

We have found that ...

Other questions ...

1. Gai, W., and Sun, K. Epigenetic biomarkers in cell-free dna and applications in liquid biopsy. Genes *10*, 32. Available at: https://pubmed.ncbi.nlm.nih.gov/30634483.

2. Cai, J., Chen, L., Zhang, Z., Zhang, X., Lu, X., Liu, W., Shi, G., Ge, Y., Gao, P., and Yang, Y. *et al.* Genome-wide mapping of 5-hydroxymethylcytosines in circulating cell-free dna as a non-invasive approach for early detection of hepatocellular carcinoma. Gut, gutjnl–2019–318882. Available at: http://gut.bmj.com/content/early/2019/07/28/gutjnl-2019-318882.abstract.

3. Li, W., Zhang, X., Lu, X., You, L., Song, Y., Luo, Z., Zhang, J., Nie, J., Zheng, W., and Xu, D. *et al.* DNA 5-hydroxymethylcytosines from cell-free circulating dna as diagnostic biomarkers for human cancers. bioRxiv, 163204. Available at: http://biorxiv.org/content/early/2017/07/13/163204.abstract.

4. Song, C.-X., Yin, S., Ma, L., Wheeler, A., Chen, Y., Zhang, Y., Liu, B., Xiong, J., Zhang, W., and Hu, J. *et al.* (2017). 5-hydroxymethylcytosine signatures in cell-free dna provide information about tumor types and stages. Cell Research *27*, 1231–1242. Available at: https://doi.org/10.1038/cr.2017.106.

5. Collin, F., Ning, Y., Phillips, T., McCarthy, E., Scott, A., Ellison, C., Ku, C.-J., Guler, G.D., Chau, K., and Ashworth, A. *et al.* Detection of early stage pancreatic cancer using 5-hydroxymethylcytosine signatures in circulating cell free dna. bioRxiv, 422675. Available at: http://biorxiv.org/content/early/2018/09/26/422675.abstract.

6. Huang, L.-H., Lin, P.-H., Tsai, K.-W., Wang, L.-J., Huang, Y.-H., Kuo, H.-C., and Li, S.-C. The effects of storage temperature and duration of blood

samples on dna and rna qualities. PloS one *12*, e0184692–e0184692. Available at: https://pubmed.ncbi.nlm.nih.gov/28926588.

7. Permenter, J., Ishwar, A., Rounsavall, A., Smith, M., Faske, J., Sailey, C.J., and Alfaro, M.P. (2015). Quantitative analysis of genomic dna degradation in whole blood under various storage conditions for molecular diagnostic testing. Molecular and Cellular Probes *29*, 449–453. Available at: http://www.sciencedirect.com/science/article/pii/S0890850815300207.

8. Law, C., Alhamdoosh, M., Su, S., Dong, X., Tian, L., Smyth, G., and Ritchie, M. (2018). RNA-seq analysis is easy as 1-2-3 with limma, glimma and edgeR [version 3; peer review: 3 approved]. F1000Research *5*. Available at: https://dx.doi.org/10.12688%2Ff1000research.9005.3.

9. Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. Limma powers differential expression analyses for rna-sequencing and microarray studies. Nucleic acids research *43*, e47–e47. Available at: https://pubmed.ncbi.nlm.nih.gov/25605792.

10. Peixoto, L., Risso, D., Poplawski, S.G., Wimmer, M.E., Speed, T.P., Wood, M.A., and Abel, T. How data analysis affects power, reproducibility and biological insight of rna-seq studies in complex datasets. Nucleic acids research *43*, 7664–7674. Available at: https://pubmed.ncbi.nlm.nih.gov/26202970.

11. Gandolfo, L.C., and Speed, T.P. RLE plots: Visualizing unwanted variation in high dimensional data. PloS one *13*, e0191629–e0191629. Available at: https://pubmed.ncbi.nlm.nih.gov/29401521.

12. Risso, D., Ngai, J., Speed, T.P., and Dudoit, S. (2014). Normalization of rna-seq data using factor analysis of control genes or samples. Nature Biotechnology *32*, 896–902. Available at: https://doi.org/10.1038/nbt.2931.

13. McCarthy, D.J., and Smyth, G.K. Testing significance relative to a fold-change threshold is a treat. Bioinformatics *25*, 765–771. Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2654802/.

14. Hastie, T., and Tibshirani, R. (2017). Extended comparisons of best subset selection, forward stepwise selection, and the lasso. arXiv: Methodology.

15. Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. J Stat Softw *33*, 1–22.

16. Tibshirani, R., Bien, J., Friedman, J., Hastie, T., Simon, N., Taylor, J., and Tibshirani, R.J. Strong rules for discarding predictors in lasso-type problems. Journal of the Royal Statistical Society. Series B, Statistical methodology *74*, 245–266. Available at: https://pubmed.ncbi.nlm.nih.gov/25506256.

17. Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. Regularization paths for cox's proportional hazards model via coordinate descent. J Stat Softw *39*, 1–13.

18. Simon, N., Friedman, J., and Hastie, T. (2013). A blockwise descent algorithm for group-penalized multiresponse and multinomial regression. arXiv preprint arXiv:1311.6529.

19. Bertsimas, D., King, A., and Mazumder, R. (2016). Best subset selection via a modern optimization lens. Ann. Statist. *44*, 813–852. Available at: https://projecteuclid.org:443/euclid.aos/1458245736.

20. Xiang, G., Keller, C.A., Giardine, B., An, L., Li, Q., Zhang, Y., and Hardison, R.C. (2020). S3norm: Simultaneous normalization of sequencing depth and signal-to-noise ratio in epigenomic data. Nucleic Acids Research *48*, e43–e43. Available at: https://doi.org/10.1093/nar/gkaa105.

21. Lozoya, O.A., Santos, J.H., and Woychik, R.P. A leveraged signal-to-noise ratio (lstnr) method to extract differentially expressed genes and multivariate patterns of expression from noisy and low-replication rnaseq data. Frontiers in genetics *9*, 176–176. Available at: https://pubmed.ncbi.nlm.nih.gov/29868123.

22. Simonsen, A.T., Hansen, M.C., Kjeldsen, E., Møller, P.L., Hindkjær, J.J., Hokland, P., and Aggerholm, A. (2018). Systematic evaluation of signal-to-noise ratio in variant detection from single cell genome multiple displacement amplification and exome sequencing. BMC Genomics *19*, 681. Available at: https://doi.org/10.1186/s12864-018-5063-5.

23. Rapaport, F., Khanin, R., Liang, Y., Pirun, M., Krek, A., Zumbo, P., Mason, C.E., Socci, N.D., and Betel, D. (2013). Comprehensive evaluation of differential gene expression analysis methods for rna-seq data. Genome biology *14*, R95–R95. Available at: https://pubmed.ncbi.nlm.nih.gov/24020486.

24. Lockhart, R., Taylor, J., Tibshirani, R.J., and Tibshirani, R. A significance test for the lasso. Ann Stat *42*, 413–468.

25. Wasserman, L. (2014). Discussion: "A significance test for the lasso". Ann. Statist. *42*, 501–508. Available at: https://projecteuclid.org:443/euclid.aos/1400592166.

26. Engebretsen, S., and Bohlin, J. Statistical predictions with glmnet. Clinical epigenetics *11*, 123–123. Available at: https://pubmed.ncbi.nlm.nih.gov/31443682.

27. Höfling, H., and Tibshirani, R. (2008). A study of pre-validation. *2*, 643–664. Available at: http://www.jstor.org/stable/30244221.