MITRA BIO

A predictive model of melanoma using 450k methylation data: A re-analysis of GEO120878

Contents

1 Introduction and Summary

The work herein details a reanalysis of the datasets published in GSE120878 along with external validation in GSE12209, GSE86355 and a TCGA cohort (Conway et al. 2019, Wouters et al. 2017, Fujiwara et al. 2019). A brief description of the datasets can be found below. The overaching aim is to develop a model based on DNA methylation that discriminates between Melanoma and Nevus samples. A multi-stage model that further classifies primary from metastatic melanoma samples may be possible by leveraging the primary vs metastatic samples in the TCGA data. Given the distinct epigenomic profiles between Nevus and Melanoma samples there is likely scope to be selective for probes on the 450k arrays that are also covered in the TWIST library and also probes that are detected less variably using Mitra's patch-based technology. The arrays from citet{Conway2019} were reprocessed from raw IDAT files. Briefly, samples were subjected to various quality control measures; probes with a detection p value of > 0.01 were filtered from the arrays to remove unreliable methylation signals, CpG probes that mapped to X & Y chromosomes, cross hyrbridised to multiple genomic locations or were located within 2 base pairs of a SNP were also removed. Samples were normalised in a multi-step process, first dye-based correction was preformed using the Noob method then probe type distribution was corrected using the regression of correlated probes method (Niu et al. 2016, Triche et al. 2013). This work focuses on multivariate modelling to better capture the relationship between CpGs with the sole focus on biomarker discovery/development. However, univariate statistics, particularly on datasets GSE120878 and GSE86355 could still be of value to understanding the biology.

Table 1: Description of Publicly Available Methylation Datasets

GEO ID	Melanoma Samples	Nevus	Notes
GSE120878	89	73	FFPE Biospies
GSE122909	11	0	11 Melanomas + 9 Cultured Melanocytes
GSE86355	33 + 28 (Primary and Metastatic)	14	Could potentially be used to increase model dataset
TCGA-SKCM	473 (105 Primary, 367 Metastatic)		

Table 2: Additional Gene Expression Data Generated on Melanoma vs Nevus Samples

GEO ID	Samples	Array Type
GSE12391	114	Agilent Microarray
GSE8401	83	Affymetrix Microarray
GSE7553	87	Affymetrix Microarray

```
library(minfi)
library(tidyverse)
library(lumi)
library(DMRcate)
library(GEOquery)


theme_set(theme_light() + theme(text = element_text(face = "bold"),
    panel.grid = element_blank()))
```

```
utils <- list.files("~/Documents/MitraStuff/", pattern = ".R",
    full.names = T)
lapply(utils[-grep(".Rmd", utils)], source)

# for i in $(ls *Grn*); do mv $i $(basename $i
# .idat.gz)_Grn.idat.gz ;done for i in $(ls *Red*); do mv
# $i $(basename $i .idat.gz)_Red.idat.gz ;done for file in
# $(ls *Grn*); do mv '${file}' '${file/Grn_/}'; done for
# file in $(ls *Red*); do mv '${file}' '${file/Red_/}';
# done for i in $(ls *Red*); do mv $i $(basename $i
# .idat.gz)_Red.idat.gz ;done
# GSM3417694_melanocytic_Red_1391.idat.gz >>>
# GSM3417694_melanocytic_Red_1391_Red.idat.gz

# for file in $(ls *Red*); do mv '${file}' '${file/Red_/}';
# done GSM3417694_melanocytic_Red_1391_Red.idat.gz >>>
# GSM3417694_melanocytic_1391_Red.idat.gz
```

```
pdata <- getGEO("GSE120878")[[1]] %>%
    pData

pdata$file_link <- pdata$supplementary_file
pdata$file_link[grep("Grn", pdata$file_link)] <- gsub(".*\\/",
    "", pdata$file_link[grep("Grn", pdata$file_link)]) %>%
    gsub("Grn_", "", .) %>%
    gsub(".idat.gz", "", .)

pdata$Basename <- paste0("./GSE120878_RAW/", pdata$file_link)
pdata$Type <- if_else(pdata$`tissue:ch1` == "primary invasive melanoma",
    "PrimaryMelanoma", "Nevus")

idats <- read.metharray.exp(targets = pdata, verbose = T)
# idats_backup <- idats idats <- idats_backup
anno <- getAnnotation(idats) %>%
    as.data.frame()
```

```r
pvals <- detectionP(idats)


filter.p <- function(x, detected.p.vals) {
    detected.p.vals <- detected.p.vals[match(featureNames(x),
        rownames(detected.p.vals)), ]
    filtered.01 <- rowSums(detected.p.vals < 0.01) == ncol(x)
    cat("A total of", sum(filtered.01 == FALSE), "low quality probes were removed")
    x <- x[filtered.01, ]
}

idats <- preprocessNoob(idats)
idats <- filter.p(idats, pvals)


remove.snps <- function(x, dist = 2, maf = 0.05, xhyb = T, xy = F) {
    a <- nrow(x)
    x <- DMRcate::rmSNPandCH(x, dist = dist, mafcut = maf, rmcrosshyb = xhyb,
        rmXY = xy)
    b <- nrow(x)
    cat("A total of", sum(a - b), "probes removed with distance from CpG of",
        dist, "and a minor allele freq of", maf)
    return(x)
}

library(ENmix)
betas <- rcp(idats)
M <- lumi::beta2m(betas + 0.001)


betas <- remove.snps(betas, dist = 2, maf = 0.05, xhyb = T, xy = T)
M <- remove.snps(M, dist = 2, maf = 0.05, xhyb = T, xy = T)
```

2 Exploratory Data Analysis

2.1 Unnormalised Beta Value Distribution

```
probe_type <- data.frame(CpG = anno$Name, ProbeType = as.factor(paste0("Type",
    anno$Type, anno$Color)))

raw <- getBeta(idats) %>%
    rowMeans() %>%
    as.data.frame() %>%
    `colnames<-`("Beta") %>%
    rownames_to_column(var = "CpG") %>%
    inner_join(probe_type)

raw$ProbeType <- factor(raw$ProbeType, levels = c("TypeIRed",
    "TypeIGrn", "TypeII"))

ggplot(raw, aes(x = Beta), col = "black") + geom_density(aes(fill = ProbeType),
    alpha = 0.5)
```

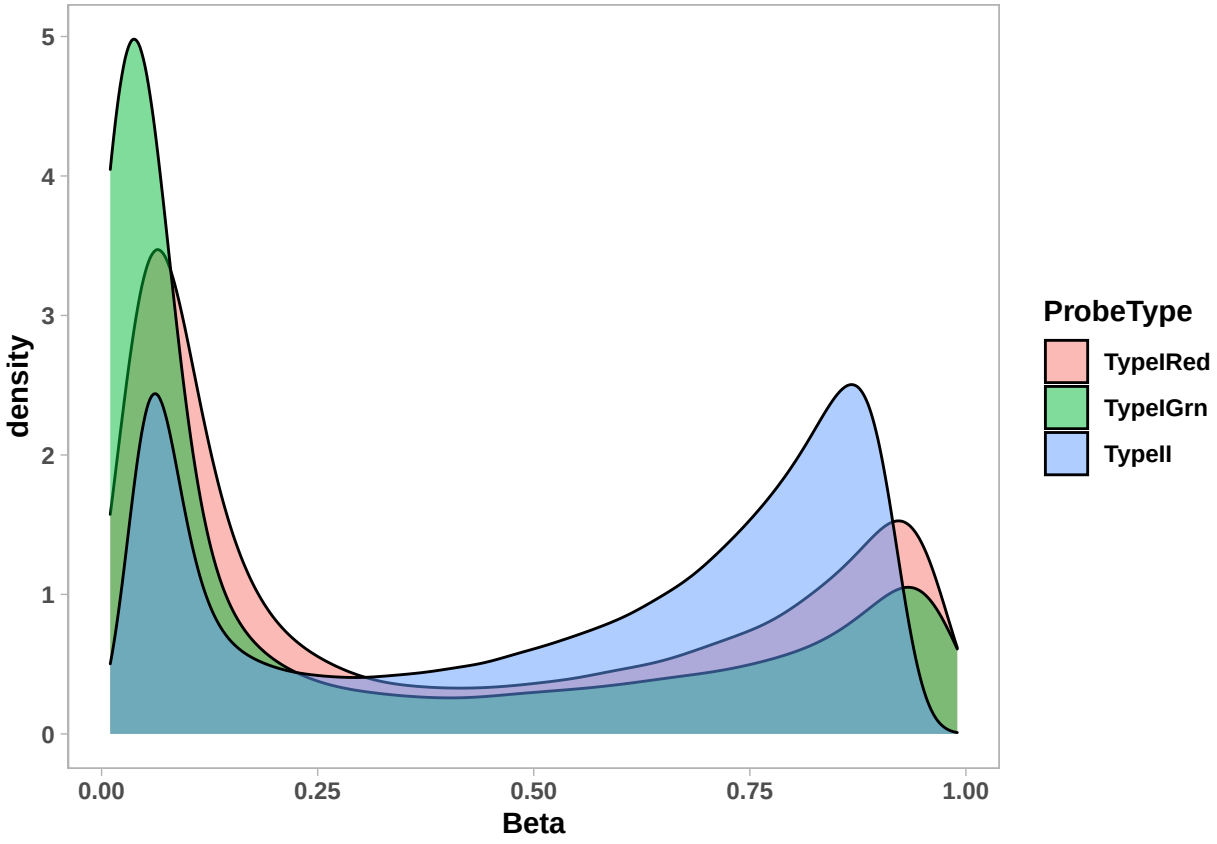


Figure 1: Unnormalised distribution of beta values

2.2 Normalised Beta Value Distribution

```
rcp <- betas %>%
    rowMeans() %>%
    as.data.frame() %>%
    `colnames<-`("Beta") %>%
    rownames_to_column(var = "CpG") %>%
    inner_join(probe_type)


ggplot(rcp, aes(x = Beta), col = "black") + geom_density(aes(fill = ProbeType),
    alpha = 0.5)
```

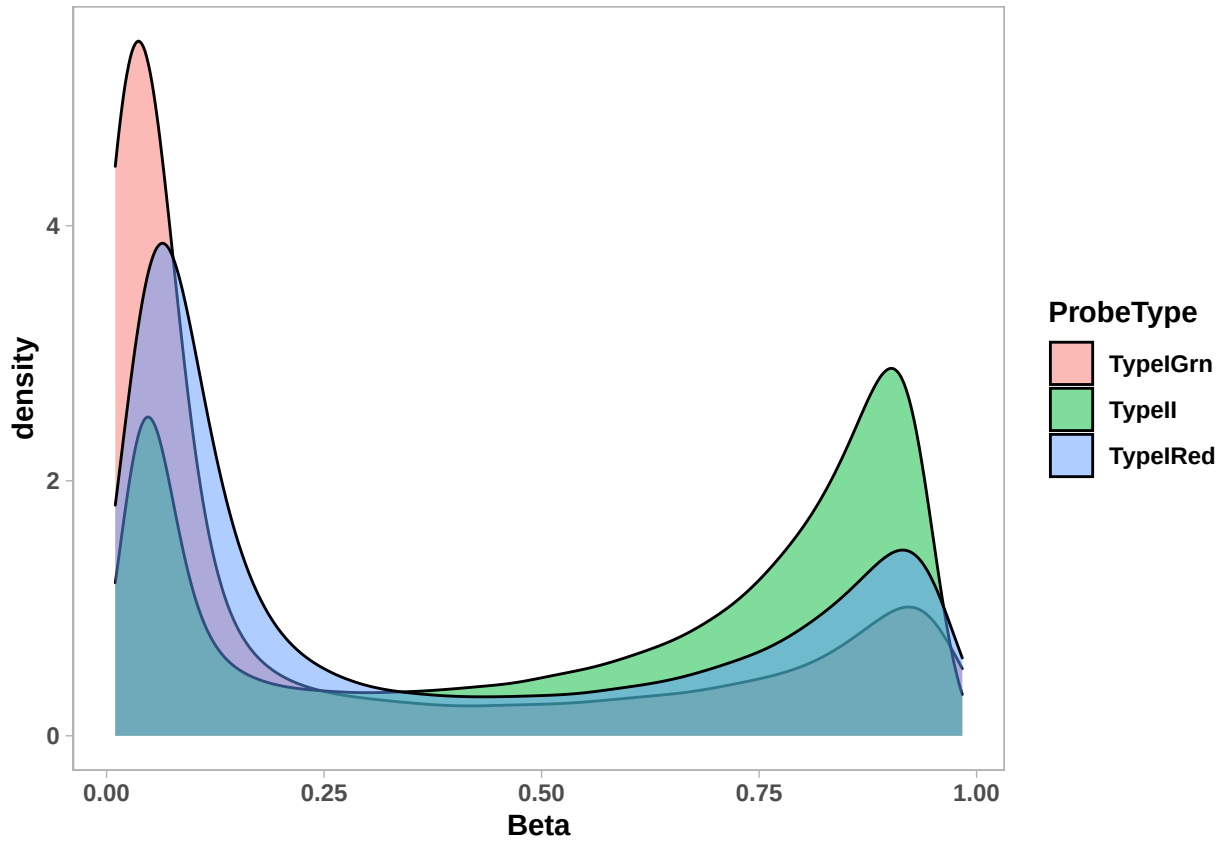


Figure 2: Normalised distribution of beta values following Noob dye bias correction and regression of correlated probes (RCP) normalisation

2.3 Preprocessing and QC of 450k Probes

Number of CpGs filtered at each processing step. The detection PValue metric describes the confidence in the signal from a probe being real dependent on whether its background intensity was above that of the control (non-human/non-targeting oligos) on the array. The number of probes that failed based on their detection PValue being > 0.01 is quite substanial. However, these arrays were hybridised from DNA extracted from FFPE biopsies which may affect the quality/purity of the DNA. The remaining probes all hybridised with sufficient confidence.

Processing Steps	CpGs
Detection PValue	109,460
X&Y, SNPs	40,042

2.4 Conversion of Beta Values to M Values for Downstream Statistical Modelling

Conversion of Beta to M values

Beta values were converted to M values as follows;

$$M_i = log2\left(\frac{Beta_i}{1 - Beta_i}\right)$$

Distributions of Beta and M Values can be seen below.

```
plot(density(betas[, 1]), ylim = c(0, 5))
for (i in 2:ncol(betas)) {
    lines(density(betas[, i]))
}
```

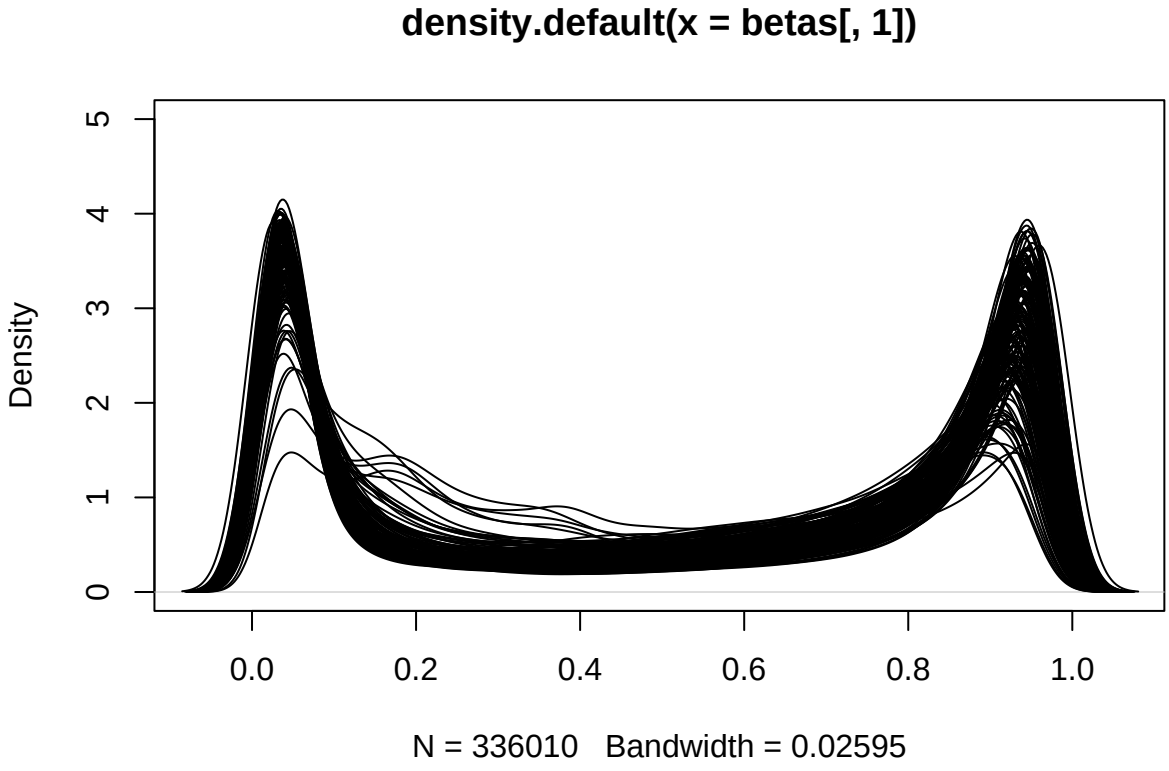


Figure 3: Density plot of distribution of Beta Values across all samples

```r
plot(density(M[, 1]), ylim = c(0, 0.2))
for (i in 2:ncol(M)) {
    lines(density(M[, i]))
}
```

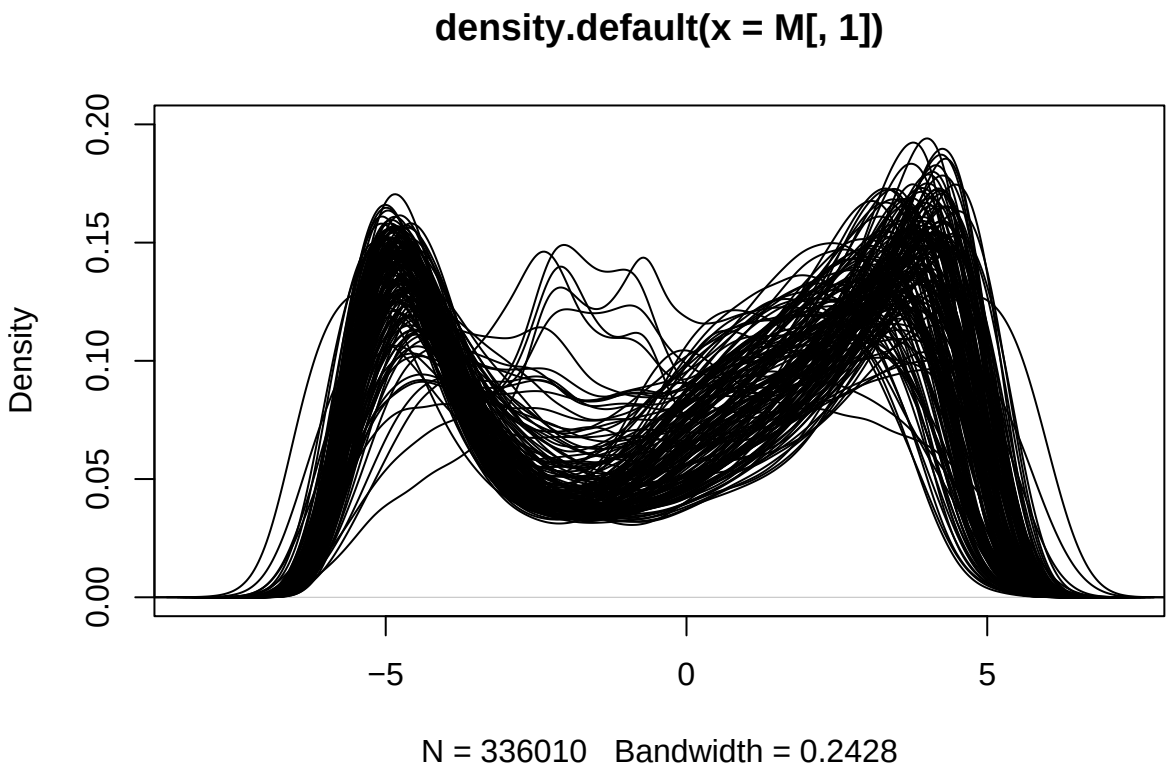


Figure 4: Density plot of distribution of M Values across all samples

3 Principal Component Analysis Shows Good Separation Between Melanoma and Nevus Samples

```
full_pca <- plot_pca_alpha(M, phenotype = pdata, scale = T, variable = "Type")
```

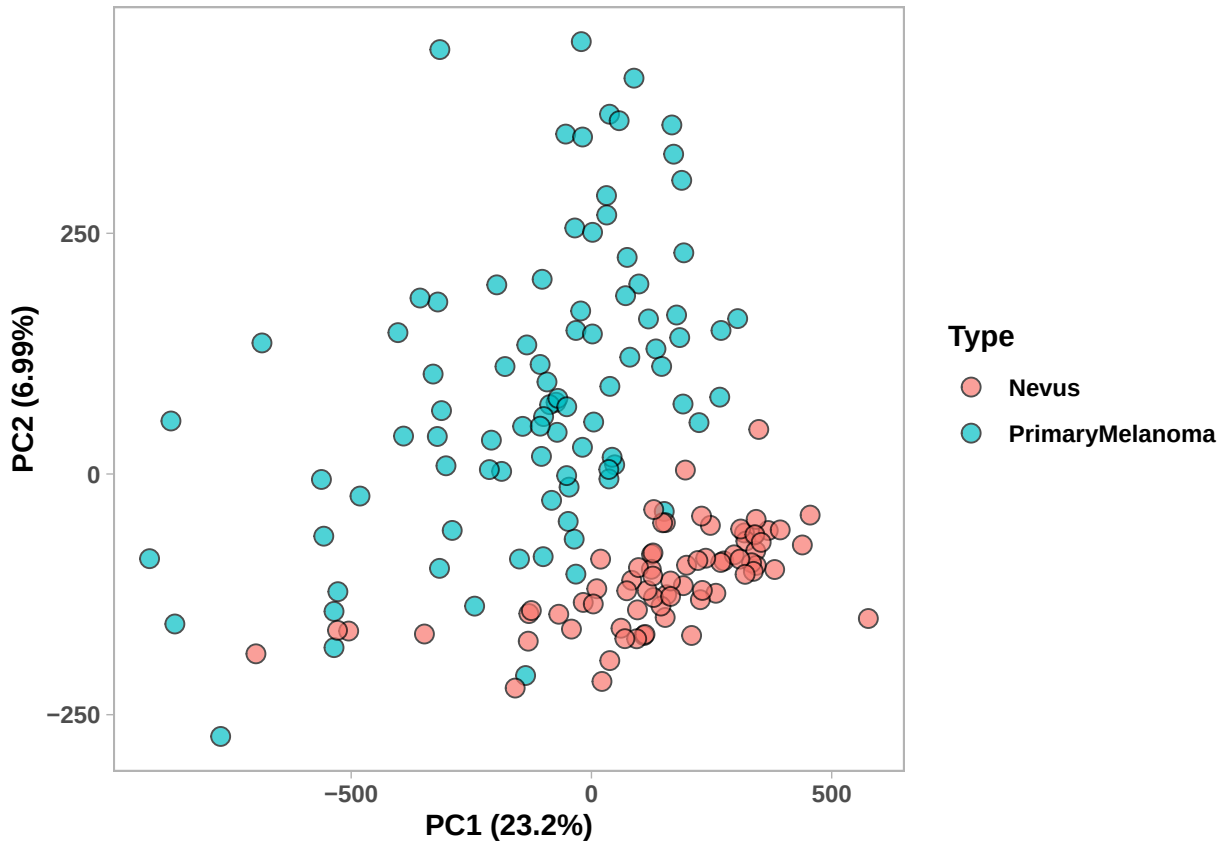


Figure 5: Scores plot from principal component analysis of M values from the 336,010 CpG probes that remained after filtering and preprocessing. Points (samples) are coloured by the variable Type which describes whether the DNA was extracted from a Primary Melanoma or a Nevus sample. Clear separation can be observed between the sample types across the first two principal components indicating that Nevus and Primary Melanoma FFPE samples have distinct epigenome-wide methylation differences.

4 A Preliminary Model To Discriminate Between Melanoma and Nevus Samples

4.1 Building a preliminary model of Melanoma vs Nevus

```
set.seed(123)
inTrain <- createDataPartition(pdata$Type, p = 0.6, list = F)

train_mel <- t(M[, inTrain])
test_mel <- t(M[, -inTrain])

# train_mel <- train_mel[rowVars(train_mel) > 0.01,]

train_pheno <- pdata[inTrain, ]
test_pheno <- pdata[-inTrain, ]

opt_alpha <- data.frame(Alpha = seq(0, 1, length.out = 3), Error = 0)
seq_alphas <- seq(0, 1, length.out = 3)



set.seed(123)
foldid <- sample(1:10, size = length(train_pheno$Type), replace = TRUE)
for (i in 1:3) {
    opt_lambda <- cv.glmnet(train_mel, as.factor(train_pheno$Type),
        type.measure = "class", nfolds = 10, alpha = seq_alphas[i],
        family = "binomial", parallel = F, foldid = foldid)
    opt_alpha$Error[i] <- opt_lambda$cvm[opt_lambda$lambda ==
        opt_lambda$lambda.1se]
    print(i)
}

opt_lambda_final <- cv.glmnet(train_mel, as.factor(train_pheno$Type),
    type.measure = "class", nfolds = 10, alpha = 0.5, family = "binomial",
    parallel = F, foldid = foldid)
```

4.2 Cross Validated Error of Various Regularised Modelling Approaches

Table 4: Out of bag cross validated error for different values of alpha in the regularised linear models

Alpha	Cross-Validated Error
0 (Ridge)	0.031
0.5 (Elastic Net)	0.010
1 (Lasso)	0.010

4.3 Identification of a 104 CpG Signature To Predict Melanoma In Skin Biopsies

Either Elastic Net or Lasso models appear to offer the best cross-validated model performance. Given Elastic net has some benefit over Lasso in terms of its additional quadratic penality I'll use Elastic Net for the final model. The final elastic net model has a total of 104 non-zero coefficients (CpGs). PCA's of the selected CpG coefficients can be seen in the two figures below, Figure X is a PCA of the variance from the 104 CpGs from all of the data, and Figure X is a PCA of the variance from the 104 CpGs on the independent test data alone.

```
glm_mod <- glmnet(train_mel, as.factor(train_pheno$Type), type.measure = "class",
    alpha = 0.5, family = "binomial")

elas_glm_coefficients <- extract_glm_coefficients(glm_mod, opt_lambda = opt_lambda_final$lambda
elas_glm_coefficients <- elas_glm_coefficients %>%
    left_join(anno, by = c(name = "Name"))
```

4.4 Principal Component Analysis Of The 104 CpG Signature

Using the 104 CpGs selected there is clear separation across the entire dataset and also the independent test dataset.

```
glm_pca <- plot_pca_alpha(M[rownames(M) %in% elas_glm_coefficients$name,
    ], phenotype = pdata, scale = T, variable = "Type")
```

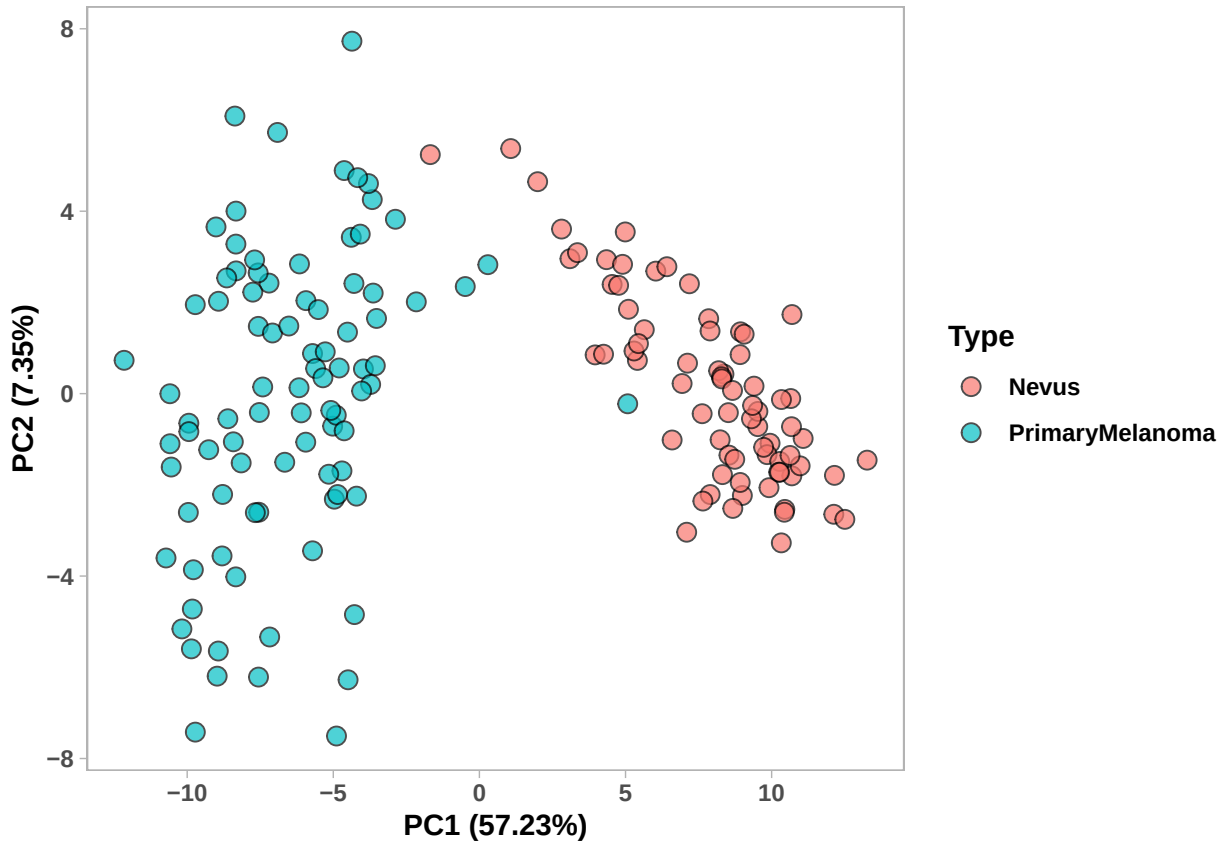


Figure 6: Scores plot from a principal component analysis based on the 104 coefficients selected by the elastic net model on the full GSE120878 dataset, samples are coloured by Origin.

```
glm_pca_test <- plot_pca_alpha(t(test_mel[, colnames(test_mel) %in%
    elas_glm_coefficients$name]), phenotype = test_pheno, scale = T,
    variable = "Type")
```

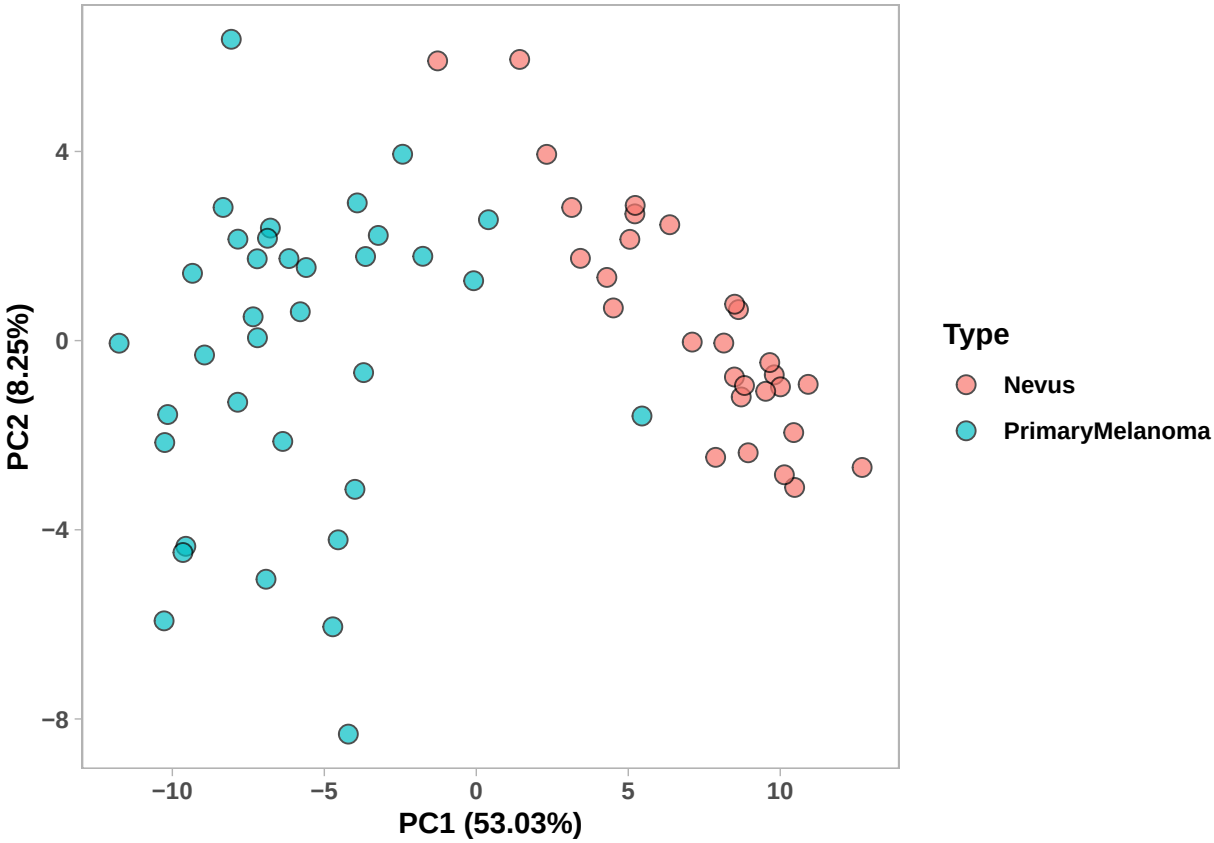


Figure 7: Scores plot from a principal component analysis based on the 104 coefficients selected by the elastic net model on the independent test dataset partitioned from GSE120878, samples are coloured by Origin.

4.5 Independent Test Set Predictions

```
glm_preds <- predict(glm_mod, test_mel, s = opt_lambda_final$lambda.min,
    type = "link")

glm_preds_class <- predict(glm_mod, test_mel, s = opt_lambda_final$lambda.min,
    type = "class")

glm_preds_resp <- predict(glm_mod, test_mel, s = opt_lambda_final$lambda.min,
    type = "response")

glm_preds_class <- data.frame(SampleID = rownames(test_mel),
    Predictions = if_else(glm_preds_class == "PrimaryMelanoma",
        "Melanoma", "Nevus"), Link = as.numeric(glm_preds[, 1]),
    Response = as.numeric(glm_preds_resp[, 1]), Labels = test_pheno$Type)
```

```
ggplot(glm_preds_class, aes(x = Link, y = Response, fill = Predictions)) +
    geom_point(pch = 21, size = 3)
```

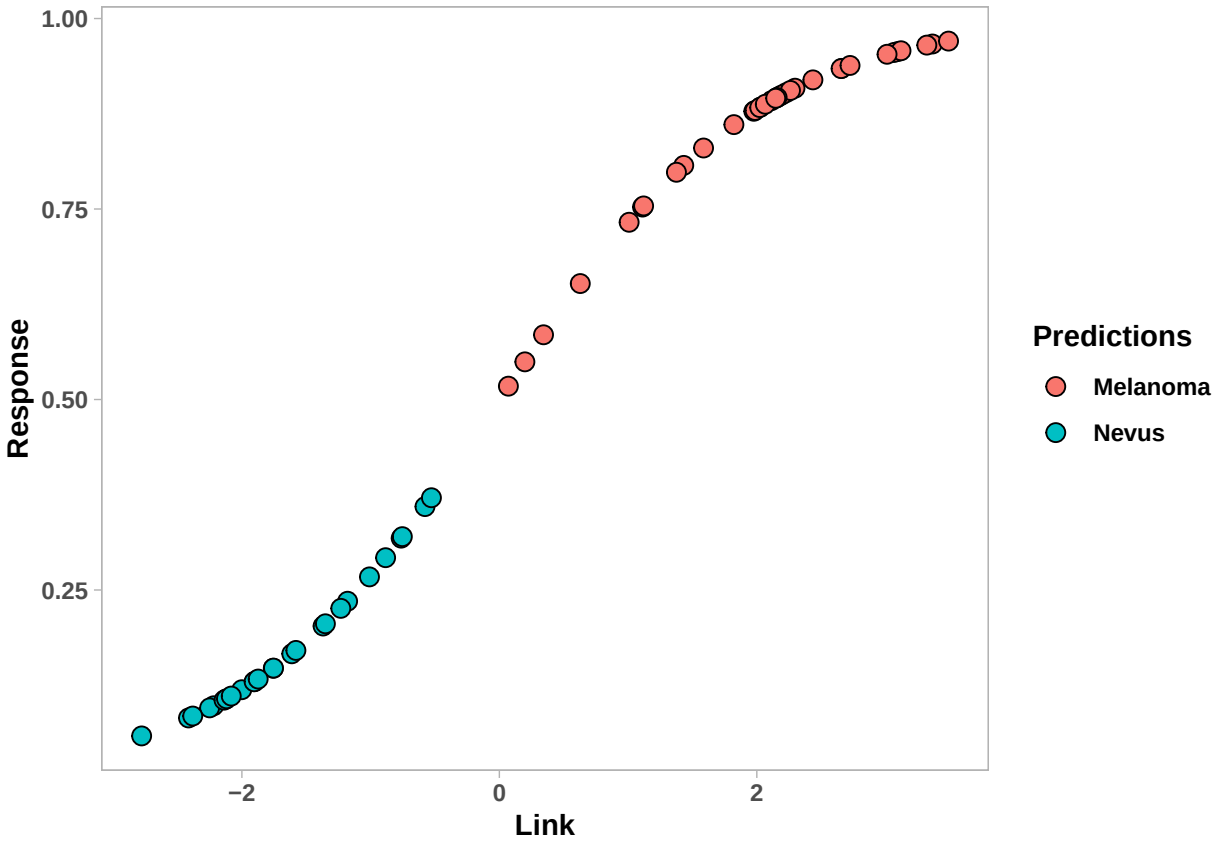


Figure 8: Sanity check of the classification attributed between link and response

4.6 Recievier Operating Curve Of Model Performance On Independent Test Data

The model performs well with very high Sensitivity and Specificity. A ROC curve can be seen below in Figure X.

```
my_roc <- function(labels, scores) {
    labels <- labels[order(scores, decreasing = TRUE)]
    data.frame(TPR = cumsum(labels)/sum(labels), FPR = cumsum(!labels)/sum(!labels),
        labels)
}

simp_roc <- my_roc(if_else(glm_preds_class$Labels == "PrimaryMelanoma",
    1, 0), glm_preds_class$Link)
simp_roc$`1 - FDR` <- 1 - simp_roc$FPR
colnames(simp_roc) <- c("Sensitivity", "FPR", "Labels", "Specificity")
```

```
ggplot(simp_roc, aes_string(y = "Sensitivity", x = "Specificity")) +
    geom_line(size = 1.5, col = "red", lty = 1) + scale_x_reverse() +
    geom_abline(intercept = 1, lty = 3)
```

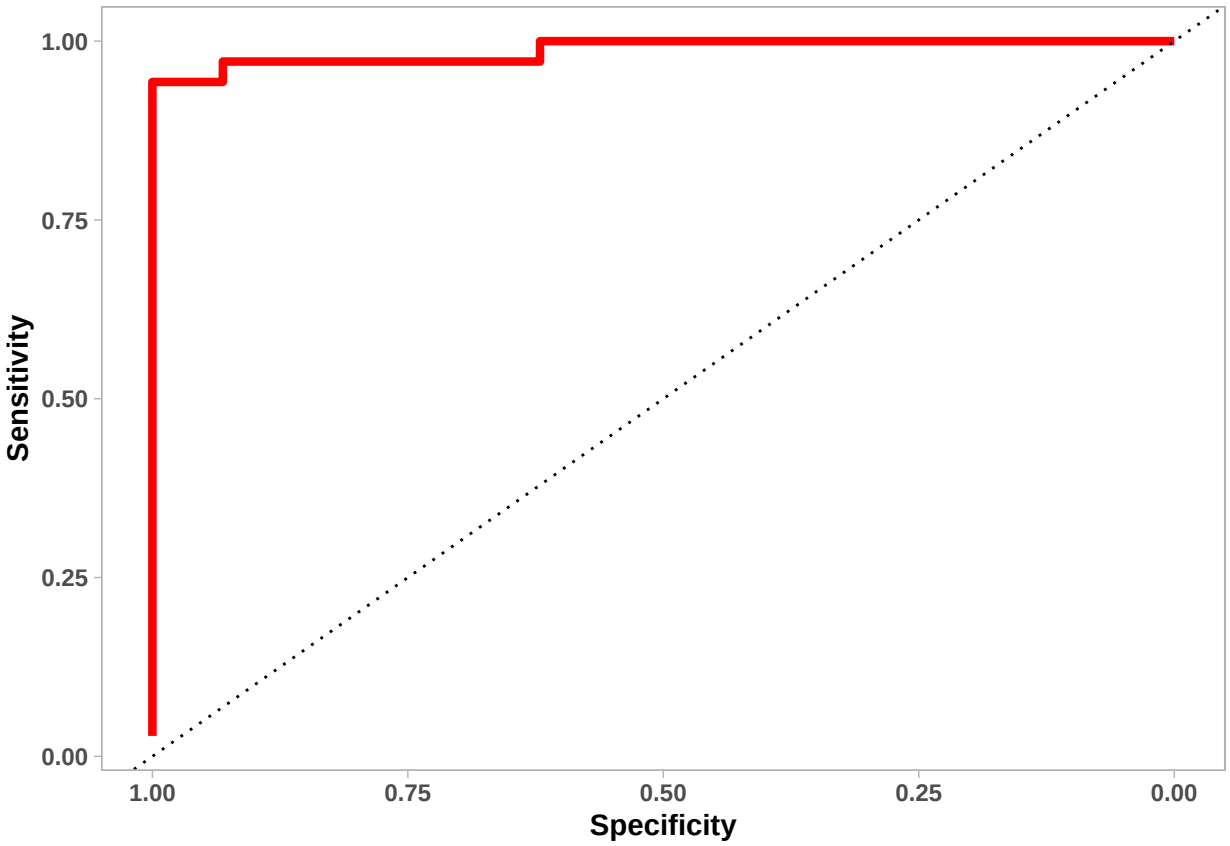


Figure 9: ROC curve of independent test set predictions using the 104 CpG signature

4.7 Additional Metrics Of Model Performance On Independent Test Data

Table X shows the exact scores for numerous metrics for the melanoma classifier.

```
library(caret)
glm_cm <- confusionMatrix(as.factor(glm_preds_class$Predictions),
    as.factor(if_else(glm_preds_class$Labels == "Nevus", "Nevus",
        "Melanoma")))

glm_cm <- glm_cm$byClass %>%
    as.data.frame() %>%
    rownames_to_column(var = "Metric") %>%
    `colnames<-`(c("Metric", "Score"))
```

```
knitr::kable(glm_cm, caption = "Scores for various metrics from the 104 CpG Melanoma classifica
```

Table 5: Scores for various metrics from the 104 CpG Melanoma classification signature on independent test data.

Metric	Score
Sensitivity	0.9714286
Specificity	0.9310345
Pos Pred Value	0.9444444
Neg Pred Value	0.9642857
Precision	0.9444444
Recall	0.9714286
F1	0.9577465
Prevalence	0.5468750
Detection Rate	0.5312500
Detection Prevalence	0.5625000
Balanced Accuracy	0.9512315

4.8 Heatmap and Hierachical Clustering of the 104 CpGs On Indepenent Test Data

We can observe good distance based clustering of the Melanoma vs Nevus samples in the independent test set using the 104 CpG signature. Interestingly, we can see that the large majority of CpGs that are hypermethylated in Melanoma vs Nevus samples are found within CpG islands.

```
library(ComplexHeatmap)
library(circlize)
library(RColorBrewer)
p_anno_test <- data.frame(Nevus = if_else(test_pheno$Type ==
    "Nevus", 1, 0), Melanoma = if_else(test_pheno$Type == "PrimaryMelanoma",
    1, 0), Probability = as.numeric(glm_preds[, 1]), row.names = rownames(test_mel))

ha <- HeatmapAnnotation(Melanoma = if_else(test_pheno$Type ==
    "PrimaryMelanoma", 1, 0), Nevus = if_else(test_pheno$Type ==
    "Nevus", 1, 0), Probability = anno_barplot(if_else(p_anno_test$Probability <
    0.5, p_anno_test$Probability - 1, p_anno_test$Probability),
    baseline = 0, gp = gpar(fill = if_else(p_anno_test$Probability >
        0.5, 1, 2))))

cols1 <- colorRamp2(c(-2, 0, 2), colors = c("navy", "white",
    "red"))

# scale_test <-
# scale(lumi::m2beta(test_mel[,colnames(test_mel) %in%
# elas_glm_coefficients$name]), scale =T, center=T)
scale_test <- scale(test_mel[, colnames(test_mel) %in% elas_glm_coefficients$name],
    scale = T, center = T)

glm_cpg_info <- elas_glm_coefficients[match(colnames(scale_test),
    elas_glm_coefficients$name), ]
glm_cpg_info$CGI <- gsub(";.*", "", glm_cpg_info$Relation_to_Island)

cgi_cols <- brewer.pal(11, "RdBu")
cgi_cols <- cgi_cols[c(2, 4, 5, 7, 8, 6)]
names(cgi_cols) <- c("Island", "N_Shore", "S_Shore", "N_Shelf",
    "S_Shelf", "OpenSea")
ra <- rowAnnotation(CGI = glm_cpg_info$CGI, col = list(CGI = cgi_cols))

Heatmap(t(scale_test), cluster_rows = T, cluster_columns = T,
    clustering_distance_rows = "spearman", clustering_distance_columns = "spearman",
    show_row_names = F, show_column_names = F, top_annotation = ha,
    col = cols1, name = " ", split = c(), right_annotation = ra)
```

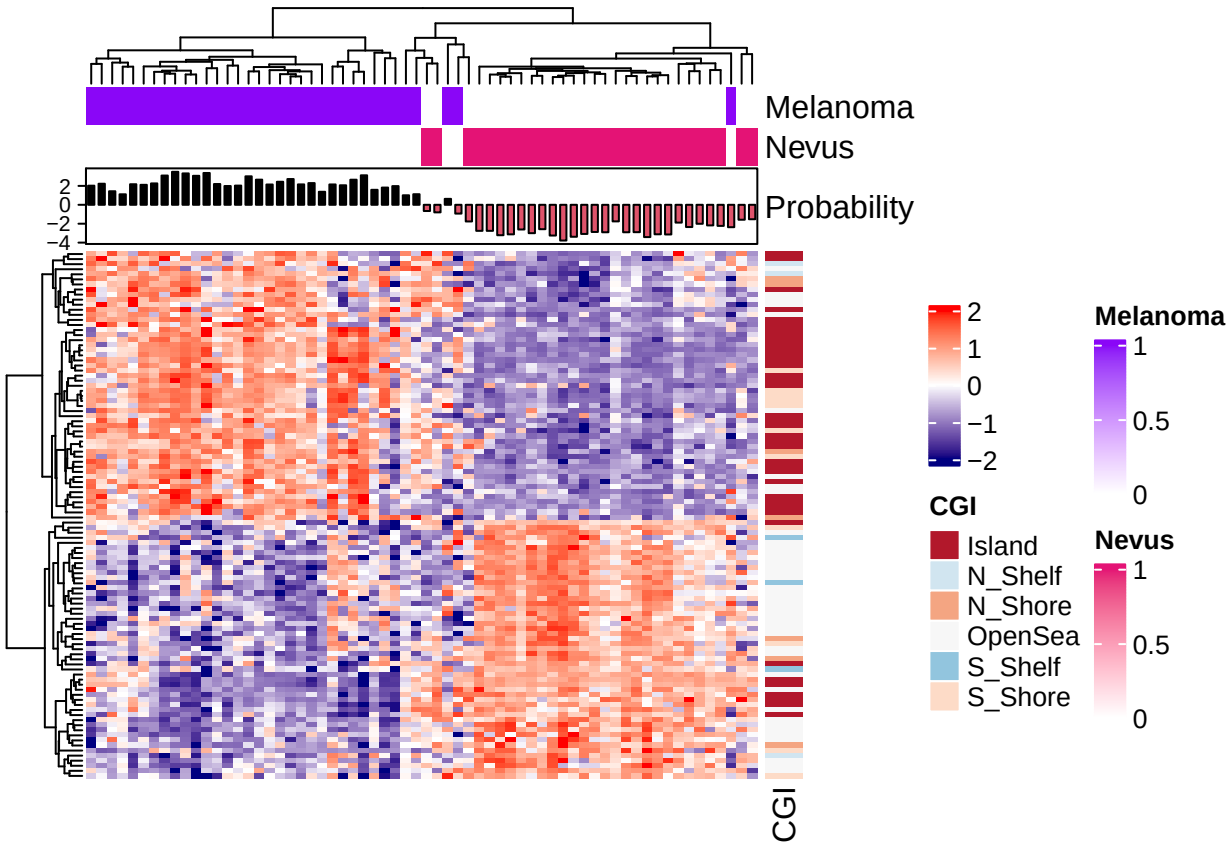


Figure 10: Heatmap of the CpGs selected from the Elastic Net Regression model on the independent test dataset. The heatmap is generated on scaled M values. Probability refers to a transformed probability from the model predictions of a sample being Melanoma or Nevus. The probability was conditionally transformed so that its center was 0 and negative numbers were indiciative of prediction of Nevus and positive values a prediction of Melanoma

5 Attempting to Explore the Biology

5.1 Exploring potential pathways for the CpGs within, or closely related to the model.

The lambda regularisation parameters for elastic net are depicted at the end of the formula here. As mentioned above, its quadratic term helps it improve on the sparsity Lasso introduces (which drops colinear variables). However, it still contains a linear combination of lasso and ridge regression penalties which force variable selection in high dimensional settings. Here we work backwards to obtain all CpGs that show high correlations (absolutely pearson correlation > 0.8) with the selected CpGs in the model to obtain a set in which to explore potential biology, rather than serve solely as biomarkers. This identified set is mainly of interest for enrichment based approaches.

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda_1 \sum_{j=1}^{p}\beta_j^2 + \lambda_2 \sum_{j=1}^{p}|\beta_j|$$

```
library(foreach)
library(doParallel)
library(HiClimR)

cl <- parallel::makeCluster(6)
doParallel::registerDoParallel(cl)


write.csv(M, "./M_Values_Correlation_Julia.csv", row.names = T,
    quote = F)
co_correlated_cpgs <- cor(t(M), method = "pearson")


# Needs to be CpGs in columns
co_correlated_cpgs <- fastCor(t(M), upperTri = T)

transpose_M <- t(M)
M_glm_cpgs <- M[rownames(M) %in% elas_glm_coefficients$name,
    ]

co_correlated_cpgs <- foreach(i = seq_len(nrow(M_glm_cpgs)),
    .combine = rbind, .multicombine = TRUE, .inorder = FALSE,
    .packages = c("data.table", "doParallel")) %dopar% {
    xtmp <- cor(as.numeric(M_glm_cpgs[i, ]), transpose_M, method = "pearson")
    rownames(xtmp) <- rownames(M_glm_cpgs)[i]
    xtmp

}
```

5.2 Identification of Co-Correlated CpGs to the 104 CpG Signature

```
install.packages("corrr")
library(corrr)
library(reshape2)

long_co_correlated_cpgs <- melt(co_correlated_cpgs)

long_co_correlated_cpgs <- long_co_correlated_cpgs[abs(long_co_correlated_cpgs$value) >
    0.8, ]

long_co_correlated_cpgs <- long_co_correlated_cpgs[as.character(long_co_correlated_cpgs$Var1)
    as.character(long_co_correlated_cpgs$Var2), ]

cor_cpgs <- as.character(unique(c(long_co_correlated_cpgs$Var1,
    long_co_correlated_cpgs$Var2)))
```

5.3 Visualisation of 742 CpGs selected by, or highly correlated to the 104 CpG Signature

```
library(RColorBrewer)
library(ComplexHeatmap)

p_anno_test <- data.frame(Nevus = if_else(test_pheno$Type ==
    "Nevus", 1, 0), Melanoma = if_else(test_pheno$Type == "PrimaryMelanoma",
    1, 0), Probability = as.numeric(glm_preds[, 1]), row.names = rownames(test_mel))

ha <- HeatmapAnnotation(Melanoma = if_else(test_pheno$Type ==
    "PrimaryMelanoma", 1, 0), Nevus = if_else(test_pheno$Type ==
    "Nevus", 1, 0), Probability = anno_barplot(if_else(p_anno_test$Probability <
    0.5, p_anno_test$Probability - 1, p_anno_test$Probability),
    baseline = 0, gp = gpar(fill = if_else(p_anno_test$Probability >
        0.5, 1, 2))))

cols1 <- colorRamp2(c(-2, 0, 2), colors = c("navy", "white",
    "red"))

# scale_test <- scale(test_mel[,colnames(test_mel) %in%
# elas_glm_coefficients$name], scale = T, center = T)

scale_test <- scale(test_mel[, colnames(test_mel) %in% cor_cpgs],
    scale = T, center = T)

glm_cpg_info <- anno[match(colnames(scale_test), anno$Name),
    ]

# glm_cpg_info <-
# elas_glm_coefficients[match(colnames(scale_test),
# elas_glm_coefficients$name),]
glm_cpg_info$CGI <- gsub(";.*", "", glm_cpg_info$Relation_to_Island)

cgi_cols <- brewer.pal(11, "RdBu")
cgi_cols <- cgi_cols[c(2, 4, 5, 7, 8, 6)]
names(cgi_cols) <- c("Island", "N_Shore", "S_Shore", "N_Shelf",
    "S_Shelf", "OpenSea")
ra <- rowAnnotation(CGI = glm_cpg_info$CGI, col = list(CGI = cgi_cols))


Heatmap(t(scale_test), cluster_rows = T, cluster_columns = T,
    clustering_distance_rows = "spearman", clustering_distance_columns = "spearman",
    show_row_names = F, show_column_names = F, top_annotation = ha,
    col = cols1, name = " ", split = c(), right_annotation = ra)
```

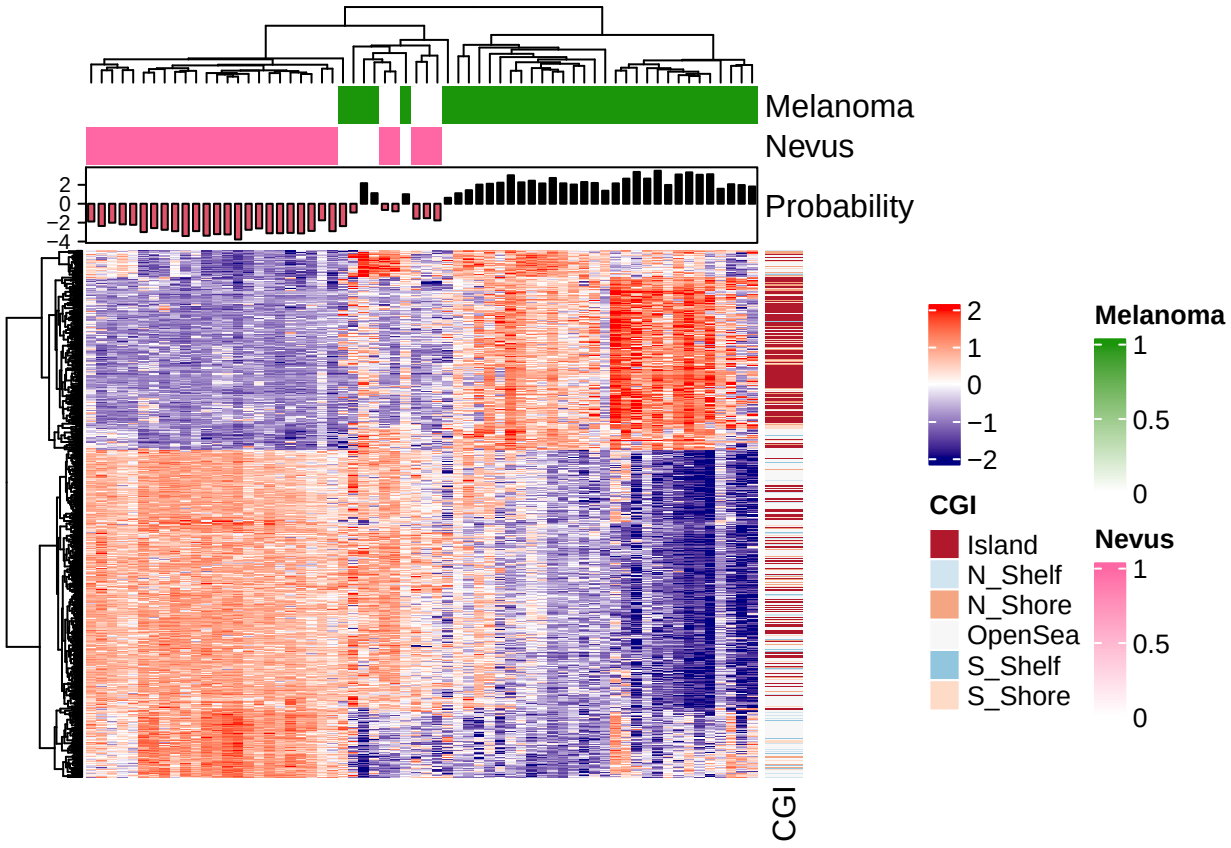


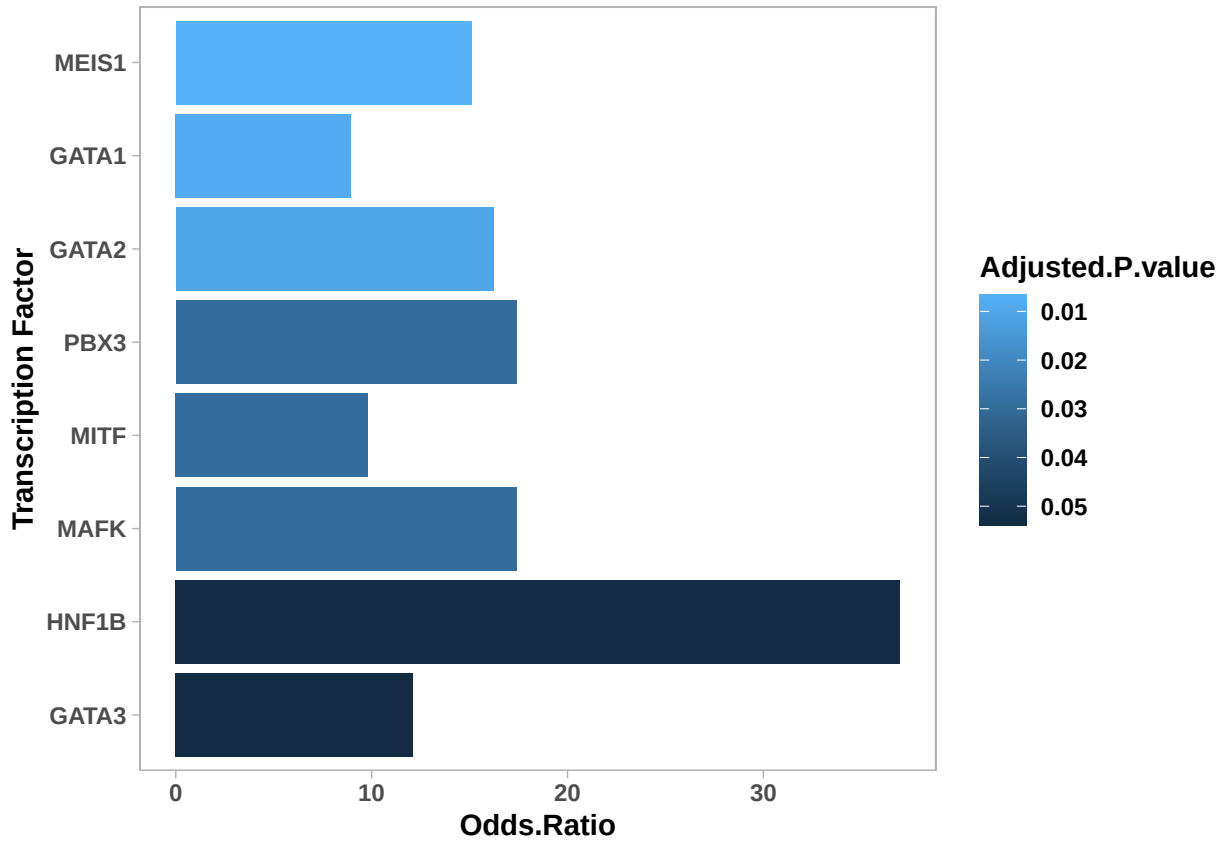
Figure 11: Heatmap of the CpGs selected from the Elastic Net Regression model and those with an absolute correlation value to them of > 0.8 on the independent test dataset. The heatmap is generated on scaled M values. Probability refers to a transformed probability from the model predictions of a sample being Melanoma or Nevus. The probability was conditionally transformed so that its center is 0 and negative numbers were indiciative of prediction of Nevus and positive values a prediction of Melanoma

```
# pheatmap(t(test_mel[,colnames(test_mel) %in%
# elas_glm_coefficients$name]), scale = 'row',
# clustering_method = 'ward.D2', clustering_distance_rows =
# 'euclidean', clustering_distance_cols = 'euclidean',
# show_colnames = F, annotation = p_anno_test, fontsize_row
# = 3, breaks = seq(-2,2, length.out = 101),
# annotation_colors = annot_colors)
```

5.4 Enrichment for Transcription Factor Protein-Protein Interaction Networks Identifies Melanocyte Inducing Transcription Factor (MITF) as an Enriched Regulator of Hypermethylated CpGs in Melanoma.

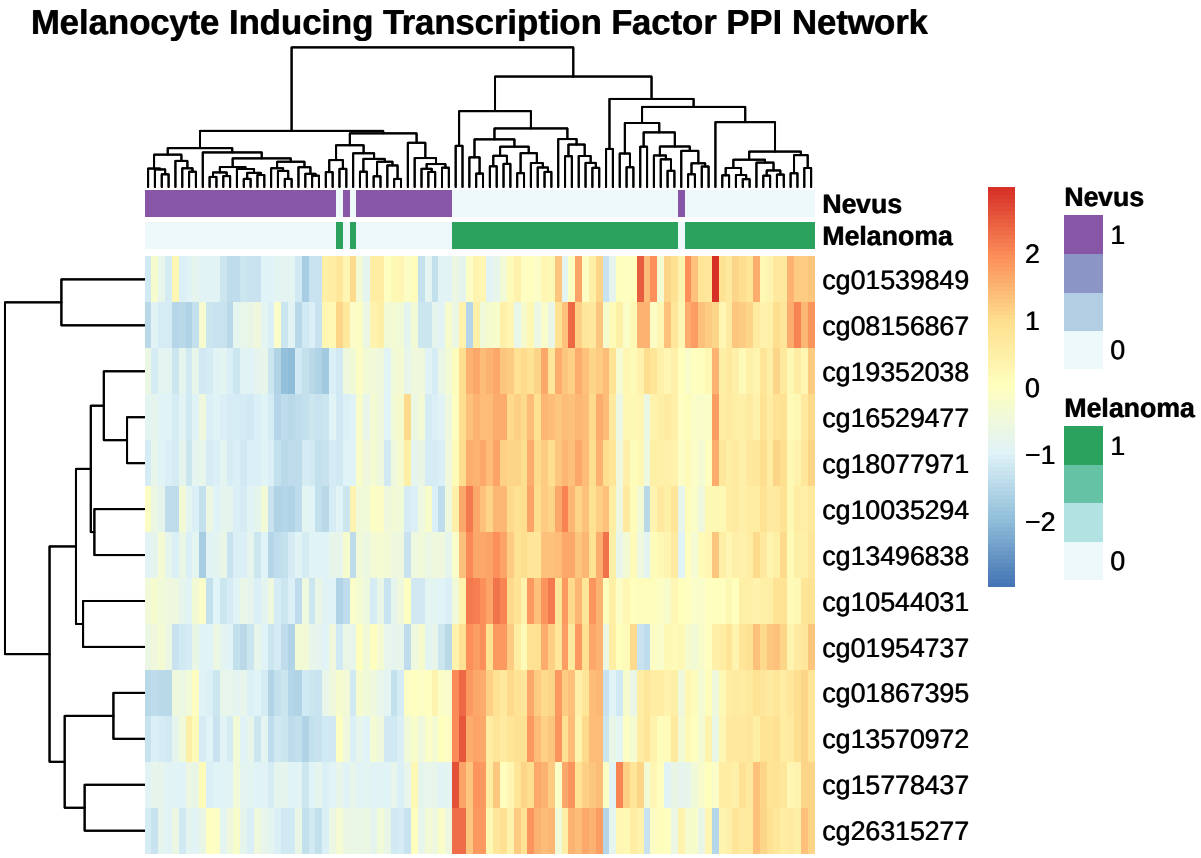
```
library(enrichR)
unique_genes <- gsub(";.*", "", glm_cpg_info$UCSC_RefGene_Name)
unique_genes <- unique_genes[unique_genes != ""]

dbs <- listEnrichrDbs()

dbs <- c("Cancer_Cell_Line_Encyclopedia", "KEGG_2021_Human",
    "Reactome_2016", "GO_Biological_Process_2021", "MSigDB_Hallmark_2020",
    "Transcription_Factor_PPIs", "Reactome_2016", "ENCODE_Histone_Modifications_2015",
    "Epigenomics_Roadmap_HM_ChIP-seq")

enrichR_obj <- enrichr(unique_genes, dbs)
```

```
library(forcats)
ggplot(enrichR_obj$Transcription_Factor_PPIs[1:8, ], aes(y = fct_reorder(Term,
    desc(Adjusted.P.value)), x = Odds.Ratio, fill = Adjusted.P.value)) +
    geom_bar(stat = "identity") + ylab("Transcription Factor") +
    scale_fill_continuous(trans = "reverse")
```



```
tfs_list <- list(strsplit(enrichR_obj$Transcription_Factor_PPIs[1,
    ]$Genes, ";")[[1]], strsplit(enrichR_obj$Transcription_Factor_PPIs[2,
    ]$Genes, ";")[[1]], strsplit(enrichR_obj$Transcription_Factor_PPIs[3,
    ]$Genes, ";")[[1]], strsplit(enrichR_obj$Transcription_Factor_PPIs[4,
    ]$Genes, ";")[[1]], strsplit(enrichR_obj$Transcription_Factor_PPIs[5,
    ]$Genes, ";")[[1]], strsplit(enrichR_obj$Transcription_Factor_PPIs[6,
    ]$Genes, ";")[[1]], strsplit(enrichR_obj$Transcription_Factor_PPIs[7,
    ]$Genes, ";")[[1]], strsplit(enrichR_obj$Transcription_Factor_PPIs[8,
    ]$Genes, ";")[[1]])

names(tfs_list) <- enrichR_obj$Transcription_Factor_PPIs$Term[1:8]


p_anno_tfs <- data.frame(Melanoma = if_else(train_pheno$Type ==
    "PrimaryMelanoma", 1, 0), Nevus = if_else(train_pheno$Type ==
    "Nevus", 1, 0), row.names = rownames(train_mel))


for (i in 1:8) {

    tfs_cpgs <- glm_cpg_info[gsub(";.*", "", glm_cpg_info$UCSC_RefGene_Name) %in%
        tfs_list[[i]], ]$Name

    print(pheatmap::pheatmap(t(train_mel[, tfs_cpgs]), scale = "row",
        annotation = p_anno_tfs, show_colnames = F, clustering_distance_cols = "euclidean",
        main = paste0(names(tfs_list)[i])))

}
```

```
tfs_cpgs <- glm_cpg_info[gsub(";.*", "", glm_cpg_info$UCSC_RefGene_Name) %in%
    tfs_list[[6]], ]$Name

print(pheatmap::pheatmap(t(train_mel[, tfs_cpgs]), scale = "row",
    annotation = p_anno_tfs, show_colnames = F, clustering_distance_cols = "euclidean",
    main = paste0("Melanocyte Inducing Transcription Factor PPI Network")))
```



6 Concluding Statements

This work re-analysed the dataset of Primary Melanoma vs Nevus samples from GSE120878 and built a multivariate classification model to discriminate Melanoma from Nevus samples. The model peformed well on independent test data (Sensitivity = 0.97, Specificity = 0.93, F1 Score = 0.96). The 104 CpGs selected in the model showed enrichment of hypermethylation in CpG islands in Melanoma. A correlation based analysis identified a further 638 CpGs that had high association (absolute pearson correlation > 0.8) with the 104 CpGs in the model. When performing enrichment for the 742 (104 + 638) CpGs a number of transcription factor's protein-to-protein interaction (PPI) networks were enriched. By plotting the CpG's in the data that mapped to the genes in the transcription factor PPI networks and performing hierachical clustering it was observed that **MEIS1**, **MAFK** and **MITF** showed clear independent clustering between Melanoma and Nevus samples. Interestingly, MITF is the Melanocyte Inducing Transcription Factor gene. Mechanistically, there is evidence to suggest that hypermethylation of MITF and genes within its pathway/network leads to a decrease in gene expression (Lauss et al. 2015). How the MITF network is regulated at the level of expression can also be tested with existing and publicly available data. MEIS1 and MAFK are less directly associated with Melanoma but are worth further exploration.

This provides some intial framework to building a Melanoma methylation biomarker panel and uncovers some plausible biology in relation to the CpGs associated with the model. It is unlikely that an Elastic net regression model is the optimal final solution. However, given the sample size in the current study there is high propensity to overfit the data with slightly more non-linear modelling alternatives. It should be plausible to adapt this model to a multi-stage predictor that also leverages TCGA data to further classify Primary Melanoma and Metastatic Melanoma. There is also the possibility to attempt to integrate the current dataset with Primary Melanoma and Nevus samples from GSE86355 to increase the sample size of the training.

```
# save.image('~/Documents/Melanoma_Nevi_GEO/Melanoma_Reanalysis_Completed_1.RData')
```

7 Bibliography

Conway, K., Edmiston, S. N., Parker, J. S., Kuan, P. F., Tsai, Y.-H., Groben, P. A., Zedek, D. C., Scott, G. A., Parrish, E. A., Hao, H., Pearlstein, M. V., Frank, J. S., Carson, C. C., Wilkerson, M. D., Zhao, X., Slater, N. A., Moschos, S. J., Ollila, D. W. & Thomas, N. E. (2019), 'Identification of a robust methylation classifier for cutaneous melanoma diagnosis', *Journal of Investigative Dermatology* **139**, 1349–1361.

Fujiwara, S., Nagai, H., Jimbo, H., Jimbo, N., Tanaka, T., Inoie, M. & Nishigori, C. (2019), 'Gene expression and methylation analysis in melanomas and melanocytes from the same patient: Loss of npm2 expression is a potential immunohistochemical marker for melanoma', *Frontiers in Oncology* **8**.

Lauss, M., Haq, R., Cirenajwis, H., Phung, B., Harbst, K., Staaf, J., Rosengren, F., Holm, K., Aine, M., Jirström, K., Åke Borg, Busch, C., Geisler, J., Lønning, P. E., Ringnér, M., Howlin, J., Fisher, D. E. & Jönsson, G. (2015), 'Genome-wide dna methylation analysis in melanoma reveals the importance of cpg methylation in mitf regulation', *Journal of Investigative Dermatology* **135**, 1820–1828.

Niu, L., Xu, Z. & Taylor, J. A. (2016), 'Rcp: a novel probe design bias correction method for illumina methylation beadchip', *Bioinformatics* **32**, 2659–2663.

Triche, T. J., Weisenberger, D. J., Berg, D. V. D., Laird, P. W. & Siegmund, K. D. (2013), 'Low-level processing of illumina infinium dna methylation beadarrays', *Nucleic Acids Research* **41**, e90–e90.

Wouters, J., Vizoso, M., Martinez-Cardus, A., Carmona, F. J., Govaere, O., Laguna, T., Joseph, J., Dynoodt, P., Aura, C., Foth, M., Cloots, R., van den Hurk, K., Balint, B., Murphy, I. G., McDermott, E. W., Sheahan, K., Jirström, K., Nodin, B., Mallya-Udupi, G., van den Oord, J. J., Gallagher, W. M. & Esteller, M. (2017), 'Comprehensive dna methylation study identifies novel progression-related and prognostic markers for cutaneous melanoma', *BMC Medicine* **15**, 101.