Text_1_SuppInfo

Steps for Detection and Removal of Ambiguously Mapped Probes

1. Get probe sequences:
   awk -F"\t" < Sample_methylation_profile.txt '{if ($341 !~ "SOURCESEQ") print ">p"NR"\n"$341}' > probe_sequences.fasta

2. Align them using Bowtie:
   bowtie_submit.sh:

```
#$ -N bowtie
#$ -cwd
#$ -o $JOB_ID.$JOB_NAME.out
#$ -e $JOB_ID.$JOB_NAME.err
#$ -pe omp 12
/Apps/bowtie/bowtie-0.12.8/bowtie-build Reference_genome/hg19
hg19_bowtieIndexes/hg19
/Apps/bowtie/bowtie-0.12.8/bowtie -f -S -p 12 -a -v 2
/hg19_bowtieIndexes/hg19 probe_sequences.fasta aligned.probes.hits .bowtie
```

   qsub bowtie_submit.sh

3. sort aligned.probeIDs.bowtie | uniq --count > aligned.probes.hits.bowtie1.counts

Text_2_SuppInfo

DMR using Minfi (functional normalization):-

```
library(minfi)
targets <- read.450k.sheet("/homeBrahma/neeraja/methylation/tongue_cancer_6","csv$")
targets
RGset <- read.450k.exp(base = "/homeBrahma/neeraja/methylation/tongue_cancer_6",
targets = targets)
RGset
pd <- pData(RGset)
pd
designMatrix <- model.matrix(~pd$Sample_Group)
gRatioSet.Funnorm <- preprocessFunnorm(RGset, bgCorr = TRUE, dyeCorr = TRUE, sex
= NULL)
dmrs <- bumphunter(gRatioSet.Funnorm, designMatrix,cutoff = 0.05, B=1)
write.table(file="func.normalized.dmr.nt",sep="\t",dmrs$table)
```

Text_3_SuppInfo

DMP using Minfi (dasen normalization):-

```
library(minfi)
library(wateRmelon)
targets <- read.450k.sheet("/home/neeraja/methylation/tongue_cancer_6","csv$")
targets
```

```
RGset <- read.450k.exp(base = "/home/neeraja/methylation/tongue_cancer_6", targets =
targets)
RGset
pd <- pData(RGset)
pd
setMethod(
  f= "dasen",
  signature(mns="RGChannelSet"),
  definition=function(mns, fudge=100){
  if(!library(minfi, logical.return=TRUE, quietly=TRUE)){
       stop('can\'t load minfi package')
     }
     mns  <- preprocessRaw(mns)
     dasen ( mns )
  }
)
d <- dasen(RGset)
write.table(file="dasen.normalized.betas",sep="\t",d)
pd <- pData(RGset)
pd
dmp1 <- dmpFinder(d, pheno=pd$Sample_Group, type="categorical")
write.table(file="dasen.normalized.dmp",sep="\t",dmp1)
```

Text_4_SuppInfo

Processing post minfi:

```
awk -F"\t" '{if ($12 <= 0.05 && $13 <= 0.05) print}' func.normalized.dmr.nt >
func.normalized.sig.dmr.nt

awk '{print $2"\t"$3"\t"$4"\t"$5}' <(grep -v start func.normalized.sig.dmr.nt) | sed 's/\"//g' |
intersectBed -a <(sortBed -i -) -b <(sortBed -i <(sed 's/^/chr/g'
/common/Data/projects/internal/OSCC/methylation/1stExon.bed)) -loj | grep -vE '\-1' | awk
'{if ($4 < 0) print }' | cut -f1-4 | sort -u | wc -l

awk '{print $2"\t"$3"\t"$4"\t"$5}' <(grep -v start func.normalized.sig.dmr.nt) | sed 's/\"//g' |
intersectBed -a <(sortBed -i -) -b <(sortBed -i <(sed 's/^/chr/g'
/common/Data/projects/internal/OSCC/methylation/1stExon.bed)) -loj | grep -vE '\-1' | awk
'{if ($4 > 0) print }' | cut -f1-4 | sort -u | wc -l

awk '{print $2"\t"$3"\t"$4"\t"$5}' <(grep -v start func.normalized.sig.dmr.nt) | sed 's/\"//g' |
intersectBed -a <(sortBed -i -) -b <(sortBed -i <(sed 's/^/chr/g'
/common/Data/projects/internal/OSCC/methylation/Island.bed)) -loj | grep -vE '\-1' | awk '{if
($4 < 0) print }' | cut -f1-4 | sort -u | wc -l

awk '{print $2"\t"$3"\t"$4"\t"$5}' <(grep -v start func.normalized.sig.dmr.nt) | sed 's/\"//g' |
intersectBed -a <(sortBed -i -) -b <(sortBed -i <(sed 's/^/chr/g'
/common/Data/projects/internal/OSCC/methylation/Island.bed)) -loj | grep -vE '\-1' | awk '{if
($4 > 0) print }' | cut -f1-4 | sort -u | wc -l

awk '{print $2"\t"$3"\t"$4"\t"$5}' <(grep -v start func.normalized.sig.dmr.nt) | sed 's/\"//g' |
intersectBed -a <(sortBed -i -) -b <(sortBed -i <(sed 's/^/chr/g'
```

```
/common/Data/projects/internal/OSCC/methylation/TSS1500.bed)) -loj | grep -vE '\-1' | awk '{if ($4 < 0) print }' | cut -f1-4 | sort -u | wc -l

awk '{print $2"\t"$3"\t"$4"\t"$5}' <(grep -v start func.normalized.sig.dmr.nt) | sed 's/\"//g' | intersectBed -a <(sortBed -i -) -b <(sortBed -i <(sed 's/^/chr/g' /common/Data/projects/internal/OSCC/methylation/TSS1500.bed)) -loj | grep -vE '\-1' | awk '{if ($4 > 0) print }' | cut -f1-4 | sort -u | wc -l
```

Repeat the above for other sub-regions

Text_5_SuppInfo

Top 100 DMRs for boxPlot

```
awk '{print $2"\t"$3"\t"$4"\t"$5}' <(grep -v start func.normalized.sig.dmr.nt) | sed 's/\"//g' | intersectBed -a <(sortBed -i -) -b <(sortBed -i <(sed 's/^/chr/g' /common/Data/projects/internal/OSCC/methylation/Island.bed)) -loj | grep -vE '\-1' | awk -F"\t" '{print $0, $4 <0?$4*-1:$4}' | sort -grk8,8 | head -100 > Island.top100dmr

awk '{print $2"\t"$3"\t"$4"\t"$5}' <(grep -v start func.normalized.sig.dmr.nt) | sed 's/\"//g' | intersectBed -a <(sortBed -i -) -b <(sortBed -i <(sed 's/^/chr/g' /common/Data/projects/internal/OSCC/methylation/TSS1500.bed)) -loj | grep -vE '\-1' | awk -F"\t" '{print $0, $4 <0?$4*-1:$4}' | sort -grk8,8 | head -100 > TSS1500.top100dmr

cat <(cut -f4 Island.top100dmr | awk '{print "OTSCC\tIsland\t"$0}') <(cut -f4 ../tcga/Island.top100dmr | awk '{print "TCGA_HNSCC\tIsland\t"$0}') <(cut -f4 TSS1500.top100dmr | awk '{print "OTSCC\tTSS1500\t"$0}') <(cut -f4 ../tcga/TSS1500.top100dmr | awk '{print "TCGA_HNSCC\tTSS1500\t"$0}') > top100dmr.forBoxPlot

df<-read.table("top100dmr.forBoxPlot",sep="\t",header=TRUE)
library(plyr)
library(ggplot2)
plot_Data <-ddply(df, .(Data Source, DMR Region), mutate,med=median(DMR DeltaBeta),min=min(DMR DeltaBeta),max=max(DMR DeltaBeta),Q1=quantile(DMR DeltaBeta, 1/4),Q3=quantile(DMR DeltaBeta, 3/4), IQR=Q3-Q1, upper.limit=Q3+1.5*IQR, lower.limit=Q1-1.5*IQR)
p <-ggplot(data = plot_Data, aes(x = Data Source))
p <-p + geom_boxplot(aes(lower = Q1, upper = Q3, middle = med, ymin = min, ymax = max), stat= "identity", colour="blue", outlier.colour = "black", outlier.shape = 16, outlier.size = 16)
p <-p +geom_point(data=plot_Data[plot_Data$DMR DeltaBeta > plot_Data$upper.limit | plot_Data$DMR DeltaBeta < plot_Data$lower.limit,], aes(y=DMR DeltaBeta))
p <-p + facet_grid( ~ DMR Region, scales="free", space="free")
p <-p + theme(plot.title = element_text("DMR DeltaBeta"),axis.text.x = element_text(angle = 90, hjust = 1, size = 8, colour = "grey50"),plot.title = element_text(face="bold", size=11),axis.title.x = element_text(face="bold", size=9),axis.title.y = element_text(face="bold", size=9, angle=90),panel.grid.major = element_blank(),panel.grid.minor = element_blank())
p <-p + scale_fill_hue(c=45, l=80)
ggsave("DMR DeltaBeta.png",plot=p,dpi=600)
```

Text_6_SuppInfo

Top 100 DMRS for boxPlot along with Tumor Staging:-

```
cat <(cut -f4 ES/Island.top100dmr | awk '{print "OTSCC\tIsland\tEarly\t"$0}') <(cut -f4
../tcga/ES/Island.top100dmr | awk '{print "TCGA_HNSCC\tIsland\tEarly\t"$0}') <(cut -f4
ES/TSS1500.top100dmr | awk '{print "OTSCC\tTSS1500\tEarly\t"$0}') <(cut -f4
../tcga/ES/TSS1500.top100dmr | awk '{print "TCGA_HNSCC\tTSS1500\tEarly\t"$0}') <(cut
-f4 LS/Island.top100dmr | awk '{print "OTSCC\tIsland\tLate\t"$0}') <(cut -f4
../tcga/LS/Island.top100dmr | awk '{print "TCGA_HNSCC\tIsland\tLate\t"$0}') <(cut -f4
LS/TSS1500.top100dmr | awk '{print "OTSCC\tTSS1500\tLate\t"$0}') <(cut -f4
../tcga/LS/TSS1500.top100dmr | awk '{print "TCGA_HNSCC\tTSS1500\tLate\t"$0}') >
top100dmr.Staged.forBoxPlot

df<-read.table("top100dmr.Staged.forBoxPlot",sep="\t",header=TRUE)
library(plyr)
library(ggplot2)
plot_Data <-ddply(df, .(dataSource, Region, Stage),
mutate,med=median(dmrDeltaBeta),min=min(dmrDeltaBeta),max=max(dmrDeltaBeta),Q1
=quantile(dmrDeltaBeta, 1/4),Q3=quantile(dmrDeltaBeta, 3/4), IQR=Q3-Q1,
upper.limit=Q3+1.5*IQR, lower.limit=Q1-1.5*IQR)
p <-ggplot(data = plot_Data, aes(x = dataSource))
p <-p + geom_boxplot(aes(lower = Q1, upper = Q3, middle = med, ymin = min, ymax =
max), stat= "identity", colour="blue", outlier.colour = "black", outlier.shape = 16, outlier.size
= 16)
p <-p +geom_point(data=plot_Data[plot_Data$dmrDeltaBeta > plot_Data$upper.limit |
plot_Data$dmrDeltaBeta < plot_Data$lower.limit,], aes(y=dmrDeltaBeta))
p <-p + facet_grid( Region ~ Stage, scales="free", space="free")
p <-p + theme(plot.title = element_text("dmrDeltaBeta"),axis.text.x = element_text(angle =
90, hjust = 1, size=12, colour = "grey50"),plot.title = element_text(face="bold",
size=11),axis.title.x = element_text(face="bold", size=12),axis.title.y =
element_text(face="bold", size=12, angle=90),panel.grid.major =
element_blank(),panel.grid.minor = element_blank())
p <-p + scale_fill_hue(c=45, l=80)
ggsave("dmrDeltaBeta.png",plot=p,dpi=600)
```

Text_7_SuppInfo

Differential expression analyses:-

```
library(lumi)
fileName <- 'DASL.raw.expr.txt'
dasl.lumi <- lumiR.batch(fileName)

# VST and Quantile/SSN/RSN/RI
dasl.lumi.VST <- lumiT(dasl.lumi)
dasl.lumi.VST.Q <- lumiN(dasl.lumi.VST, method='quantile')
dasl.lumi.VST.SSN <- lumiN(dasl.lumi.VST, method='ssn')
dasl.lumi.VST.RSN <- lumiN(dasl.lumi.VST, method='rsn')
dasl.lumi.VST.RI <- lumiN(dasl.lumi.VST, method='rankinvariant')
dasl.lumi.VST.LOESS <- lumiN(dasl.lumi.VST, method='loess')
dasl.lumi.VST.VSN <- lumiN(dasl.lumi.VST, method='vsn')
```

```
write.exprs(dasl.lumi,"dasl.raw.expr.txt")
write.exprs(dasl.lumi.VST,"dasl.VST.expr.txt")
write.exprs(dasl.lumi.VST.Q,"dasl.VST.Q.expr.txt")
write.exprs(dasl.lumi.VST.SSN,"dasl.VST.SSN.expr.txt")
write.exprs(dasl.lumi.VST.RSN,"dasl.VST.RSN.expr.txt")
write.exprs(dasl.lumi.VST.RI,"dasl.VST.RI.expr.txt")
write.exprs(dasl.lumi.VST.LOESS,"dasl.VST.LOESS.expr.txt")
write.exprs(dasl.lumi.VST.VSN,"dasl.VST.VSN.expr.txt")
```

To calculate Inter-array Correlation Coefficients:

```
x=read.table("dasl.raw.expr.txt",header=TRUE)
y=as.matrix(cor(x))
mean(y)

x=read.table("dasl.VST.expr.txt",header=TRUE)
y=as.matrix(cor(x))
mean(y)

x=read.table("dasl.VST.Q.expr.txt",header=TRUE)
y=as.matrix(cor(x))
mean(y)

x=read.table("dasl.VST.SSN.expr.txt",header=TRUE)
y=as.matrix(cor(x))
mean(y)

x=read.table("dasl.VST.RSN.expr.txt",header=TRUE)
y=as.matrix(cor(x))
mean(y)

x=read.table("dasl.VST.RI.expr.txt",header=TRUE)
y=as.matrix(cor(x))
mean(y)

x=read.table("dasl.VST.LOESS.expr.txt",header=TRUE)
y=as.matrix(cor(x))
mean(y)

x=read.table("dasl.VST.VSN.expr.txt",header=TRUE)
y=as.matrix(cor(x))
mean(y)

# LRT and Quantile/SSN/RSN/RI/LOESS/VSN
dasl.lumi.LRT <- lumiT(dasl.lumi, method='log2')
dasl.lumi.LRT.Q <- lumiN(dasl.lumi.LRT, method='quantile')
dasl.lumi.LRT.SSN <- lumiN(dasl.lumi.LRT, method='ssn')
dasl.lumi.LRT.RSN <- lumiN(dasl.lumi.LRT, method='rsn')
dasl.lumi.LRT.RI <- lumiN(dasl.lumi.LRT, method='rankinvariant')
dasl.lumi.LRT.LOESS <- lumiN(dasl.lumi.LRT, method='loess')
dasl.lumi.VSN <- lumiN(dasl.lumi, method='vsn')  # VSN does not work with LRT
```

```
write.exprs(dasl.lumi.LRT,"dasl.LRT.expr.txt")
write.exprs(dasl.lumi.LRT.Q,"dasl.LRT.Q.expr.txt")
write.exprs(dasl.lumi.LRT.SSN,"dasl.LRT.SSN.expr.txt")
write.exprs(dasl.lumi.LRT.RSN,"dasl.LRT.RSN.expr.txt")
write.exprs(dasl.lumi.LRT.RI,"dasl.LRT.RI.expr.txt")
write.exprs(dasl.lumi.LRT.LOESS,"dasl.LRT.LOESS.expr.txt")
write.exprs(dasl.lumi.VSN,"dasl.VSN.expr.txt")

x=read.table("dasl.LRT.expr.txt",header=TRUE)
y=as.matrix(cor(x))
mean(y)

x=read.table("dasl.LRT.Q.expr.txt",header=TRUE)
y=as.matrix(cor(x))
mean(y)

x=read.table("dasl.LRT.SSN.expr.txt",header=TRUE)
y=as.matrix(cor(x))
mean(y)

x=read.table("dasl.LRT.RSN.expr.txt",header=TRUE)
y=as.matrix(cor(x))
mean(y)

x=read.table("dasl.LRT.RI.expr.txt",header=TRUE)
y=as.matrix(cor(x))
mean(y)

x=read.table("dasl.LRT.LOESS.expr.txt",header=TRUE)
y=as.matrix(cor(x))
mean(y)

x=read.table("dasl.VSN.expr.txt",header=TRUE)
y=as.matrix(cor(x))
mean(y)

# CRT and Quantile/SSN/RSN/RI/LOESS/VSN
dasl.lumi.CRT <- lumiT(dasl.lumi, method='cubicRoot')
dasl.lumi.CRT.Q <- lumiN(dasl.lumi.CRT, method='quantile')
dasl.lumi.CRT.SSN <- lumiN(dasl.lumi.CRT, method='ssn')
dasl.lumi.CRT.RSN <- lumiN(dasl.lumi.CRT, method='rsn')
dasl.lumi.CRT.RI <- lumiN(dasl.lumi.CRT, method='rankinvariant')
dasl.lumi.CRT.LOESS <- lumiN(dasl.lumi.CRT, method='loess')
dasl.lumi.CRT.VSN <- lumiN(dasl.lumi.CRT, method='vsn')

write.exprs(dasl.lumi.CRT,"dasl.CRT.expr.txt")
write.exprs(dasl.lumi.CRT.Q,"dasl.CRT.Q.expr.txt")
write.exprs(dasl.lumi.CRT.SSN,"dasl.CRT.SSN.expr.txt")
write.exprs(dasl.lumi.CRT.RSN,"dasl.CRT.RSN.expr.txt")
write.exprs(dasl.lumi.CRT.RI,"dasl.CRT.RI.expr.txt")
write.exprs(dasl.lumi.CRT.LOESS,"dasl.CRT.LOESS.expr.txt")
write.exprs(dasl.lumi.CRT.VSN,"dasl.CRT.VSN.expr.txt")
```

```
x=read.table("dasl.CRT.expr.txt",header=TRUE)
y=as.matrix(cor(x))
mean(y)

x=read.table("dasl.CRT.Q.expr.txt",header=TRUE)
y=as.matrix(cor(x))
mean(y)

x=read.table("dasl.CRT.SSN.expr.txt",header=TRUE)
y=as.matrix(cor(x))
mean(y)

x=read.table("dasl.CRT.RSN.expr.txt",header=TRUE)
y=as.matrix(cor(x))
mean(y)

x=read.table("dasl.CRT.RI.expr.txt",header=TRUE)
y=as.matrix(cor(x))
mean(y)

x=read.table("dasl.CRT.LOESS.expr.txt",header=TRUE)
y=as.matrix(cor(x))
mean(y)
```

 # dasl.sample.info file format for ComBat

```
Sample      Array  Batch Covariate 1
OT1_N       8981245084 I       N
OT1_T       8981245084 I       T
.
.
.


R
source('ComBat.R')
ComBat('dasl.log2.Q.expr.txt','dasl.sample.info.txt')

cut -f 2-29 Adjusted_log2.Q.expr.txt_.xls > T2
cut -f 1 dasl.log2.Q.expr.txt > T1
paste -d"\t" T1 T2 > Adjusted_dasl.log2.Q.expr.txt

x=read.table("Adjusted_dasl.log2.Q.expr.txt ",header=TRUE)
y=as.matrix(cor(x))
mean(y)

awk '{print NR"\t"$0}' Adjusted_dasl.log2.Q.expr.txt | cut -f 1,3-30 >
Adjusted_dasl.log2.Q.expr.geneIDRem.txt

R
```

```
dataMatrix <- read.table("Adjusted_dasl.log2.Q.expr.geneIDRem.txt",header=TRUE)
sampleType <- c('N', 'N', 'N', 'N', 'N', 'N', 'N', 'N', 'N', 'N', 'N', 'N', 'N', 'N', 'T', 'T', 'T', 'T', 'T', 'T',
'T', 'T', 'T', 'T', 'T', 'T', 'T', 'T', 'T')
if(require(limma)) {
design <- model.matrix(~ factor(sampleType))
fit <- lmFit(dataMatrix, design)
fit <- eBayes(fit)
diffExpDASL <- topTable(fit, coef='factor(sampleType)T', adjust='fdr', number=50000)
write.table(file="Differential.Expression.DASL.txt",diffExpDASL,sep="\t")
}
```

 # nondasl.sample.info file format for ComBat

```
Sample      Array  Batch Covariate 1
OT2_N       6116733012 I       N
OT2_T       6116733012 I       T
.
.
.
```

```
R
source('ComBat.R')
ComBat('nondasl.VST.RSN.expr.txt','nondasl.sample.info.txt')

cut -f 2-16 Adjusted_nondasl.VST.RSN.expr.txt_.xls > T2
cut -f 1 nondasl.VST.RSN.expr.txt > T1
paste -d"\t" T1 T2 > Adjusted_nondasl.VST.RSN.expr.txt

x=read.table("Adjusted_nondasl.VST.RSN.expr.txt ",header=TRUE)
y=as.matrix(cor(x))
mean(y)

awk '{print NR"\t"$0}' Adjusted_nondasl.VST.RSN.expr.txt | cut -f 1,3-30 >
Adjusted_nondasl.VST.RSN.expr.geneIDRem.txt

R

dataMatrix <-
read.table("Adjusted_nondasl.VST.RSN.expr.geneIDRem.txt",header=TRUE)
sampleType <- c('N', 'N', 'N', 'N', 'N', 'N', 'N', 'N', 'T', 'T', 'T', 'T', 'T', 'T', 'T')
if(require(limma)) {
design <- model.matrix(~ factor(sampleType))
fit <- lmFit(dataMatrix, design)
fit <- eBayes(fit)
diffExpNonDASL <- topTable(fit, coef='factor(sampleType)T', adjust='fdr', number=50000)
write.table(file="Differential.Expression.NonDASL.txt",diffExpNonDASL,sep="\t")
}
```

Text_8_SuppInfo

Correlation between differential expression and differential methylation:-

Bubble-plots with varying Y axes for samples, and X axes for genes. Colors and sizes of bubbles indicate differential methylation and differential expression magnitudes and directions.

Text_9_SuppInfo

R commands implemented to perform variable elimination using random forest analyses to identify differentially methylated probes predicting tissue type

```
library(randomForest)
library(varSelRF)
DS=read.table("/homeBrahma/neeraja/methylation/tongue_cancer_6/dasen.normalized.funNormdmr.categs.betas",header=TRUE,na.strings="NA")
DS<-na.omit(DS)
for (i in 1:500) {
        set.seed(i)
        DS.rf.vsf=varSelRF(DS[,-c(1,2)],DS[,1],ntree=3000,ntreeIterat=2000,vars.drop.frac=0.2,whole.range = FALSE,keep.forest = TRUE)
        print(DS.rf.vsf$selected.vars)
        print(predict(DS.rf.vsf$rf.model,subset(DS[,-c(1,2)],select=DS.rf.vsf$selected.vars)))
}
```

Text_10_SuppInfo
R commands used to implement and visualize 0.632+ bootstrapping of OTSCC data to assess error in prediction of the tumor-specific minimal methylation signature

```
library(randomForest)
library(varSelRF)
DS=read.table("func.normalized.sig.dmr.nt.betas_cat_diff.betas.25Aug2015",header=TRUE,na.strings="NA")
DS<-na.omit(DS)
for (i in 1:500) {
        set.seed(i)
        DS.rf.vsf=varSelRF(DS[,-1],DS[,1],ntree=3000,ntreeIterat=2000,vars.drop.frac=0.2,whole.range = FALSE,keep.forest = TRUE)

DS.bootrf.vsf=varSelRFBoot(DS[,c(DS.rf.vsf$selected.vars)],DS[,1],ntree=3000,ntreeIterat=2000,vars.drop.frac=0.2,whole.range = FALSE,keep.forest = TRUE, bootnumber=50, usingCluster = FALSE)
      file1=paste("cat.27Aug2015.classPredictions",i,sep="")
      file2=paste("cat.27Aug2015.allDataRandomForest",i,sep="")
      write.table(file=file1,DS.bootrf.vsf$class.predictions,sep="\t",quote=FALSE)

write.table(file=file2,DS.bootrf.vsf$all.data.randomForest$confusion,sep="\t",quote=FALSE)
}

df<-read.table("/storage/rfs/meth/tongue_cancer_6/otscc.meth.error.3Sep2015",sep="\t",header=TRUE)
```

```
library(plyr)
library(ggplot2)
plot_Data <-ddply(df, .(tissue),
mutate,med=median(error),min=min(error),max=max(error),Q1=quantile(error,
1/4),Q3=quantile(error, 3/4), IQR=Q3-Q1, upper.limit=Q3+1.5*IQR, lower.limit=Q1-
1.5*IQR)
p <-ggplot(data = plot_Data, aes(x = dfsType))
p <-p + geom_boxplot(aes(lower = Q1, upper = Q3, middle = med, ymin = min, ymax =
max), stat= "identity", colour="blue", outlier.colour = "black", outlier.shape = 16, outlier.size
= 16)
p <-p +geom_point(data=plot_Data[plot_Data$error > plot_Data$upper.limit |
plot_Data$error < plot_Data$lower.limit,], aes(y=error))
p <-p + facet_grid( RF ~ survival, scales="free", space="free")
p <-p + theme(plot.title = element_text("0.632 error"),axis.text.x = element_text(angle = 90,
hjust = 1, size = 8, colour = "grey50"),plot.title = element_text(face="bold",
size=11),axis.title.x = element_text(face="bold", size=9),axis.title.y =
element_text(face="bold", size=9, angle=90),panel.grid.major =
element_blank(),panel.grid.minor = element_blank())
p <-p + scale_fill_hue(c=45, l=80)
ggsave("otscc.meth.error.632.3Sep2015.png",plot=p,dpi=600)
```

Text11_SuppInfo

example shown below is for training set 3

```
R
library(varSelRF)
DS=read.table("/storage/rfs/meth/tongue_cancer_6/only.homogenous.cat.dmrs.31Aug201
5",header=TRUE,na.strings="NA",sep="\t")
DS<-na.omit(DS)
for (i in 1:500) {
        set.seed(i)
        DS.rf.vsf=varSelRF(DS[,-
1],DS[,1],ntree=3000,ntreeIterat=2000,vars.drop.frac=0.2,whole.range=FALSE,keep.forest
=TRUE)
        DS.rf=randomForest(DS[,-
1],DS[,1],ntree=3000,keep.forest=FALSE,importance=TRUE)
        fileb=paste("meth.TminN.tr.set3.beforeFDR.1Dec2015.",i,sep="")
        filea=paste("meth.TminN.tr.set3.afterFDR.1Dec2015.",i,sep="")
        before<-DS.rf.vsf$initialImportances
        after<-p.adjust(DS.rf$importance[,3],method="BH",n=5337)
        write.table(x=before,file=fileb,sep="\t",quote=FALSE)
        write.table(x=after,file=filea,sep="\t",quote=FALSE)
}
q()
```

```
cat meth.TminN.tr.set3.*.1Dec2015.iid | grep -v x | grep -v Mean | grep -v V1 | grep -v V2 |
sort -k1,1 | paste - - | cut -f1-2,4 > meth.TminN.tr.set3.best.1Dec2015.iid
```

where iid to be substituted by iterations resulting in the best prediction set.

```
R
```

```
library(varSelRF)
for (i in c(iid)) {
	set.seed(i)
	file=paste("meth.TminN.tr.set3.best.1Dec2015.",i,sep="")
	t=read.table(file,sep="\t")
	mymat=matrix(1:10674,5337,2)
	mymat[,1]<-rank(t$V2)
	mymat[,2]<-rank(t$V3)
	fil=paste("meth.TminN.tr.set3.best.Rank.2Dec2015.",i,sep="")
	write.table(mymat,file=fil,sep="\t",quote=FALSE)
}
q()
```

again, where iid to be substituted by iterations (comma-separated) resulting in the best prediction set.

cat meth.TminN.tr.set3.best.Rank.2Dec2015.482 | grep -E $'\t5337|\t5336|\t5335|\t5334' > meth.TminN.tr.set3.best.Rank.2Dec2015 (grep the top-ranking variables)

awk 'function abs(x){return ((x < 0.0) ? -x : x)} {print abs($2-$3)}' meth.TminN.tr.set3.best.Rank.2Dec2015 | awk '{if ($1 <= 889) print}' | wc -l

difference threshhold decided as the number of variables (5336 in this case) divided by a factor of 6.

Text12_SuppInfo

Discovery & Validation

A. Habits

```
library(survival)
time<-
c(1.10,3.70,3.80,3.97,7.73,8.47,11.77,12.8,12.93,14.27,18.20,19.70,19.53,20.63,20.60,23.37,23.90,25.93,27.9,28.53,28.87,31.27,47.67,54.63,64.37,74.93,76.90,91.30,89.10,97.07)
status<-c(1,1,0,1,1,1,1,1,0,0,1,0,1,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0)
treatment<-c(2,2,2,1,2,2,2,1,1,2,1,2,2,1,2,2,2,2,2,1,2,2,2,2,1,1,1,1,1,2)
fit <-survdiff(Surv(time, status) ~ treatment)
fit

library(survival)
time1<-c(3.97,12.8,12.93,18.20,20.63,28.53,64.37,74.93,76.90,91.30,89.10)
status1<-c(1,1,0,1,0,0,0,0,0,0,0)
treatment1<-c(1,1,1,1,1,1,1,1,1,1,1)
time2<-
c(1.10,3.70,3.80,7.73,8.47,11.77,14.27,19.70,19.53,20.60,23.37,23.90,25.93,27.9,28.87,31.27,47.67,54.63,97.07,NA)
status2<-c(1,1,0,1,1,1,0,0,1,0,0,0,0,1,0,0,0,0,0,1)
treatment2<-c(2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2)
fit1 <-survfit(Surv(time1, status1) ~ 1)
fit2 <-survfit(Surv(time2, status2) ~ 1)
png("EVS.Alive_0.Dead.REC_1.Habits.Discover_Validation.4Feb2016.png")
```

```
plot(fit1,conf.int="none",col='blue',xlab='follow-up (months)',ylab= 'Cumulative Survival Probability',xlim=c(0,60))
lines(fit2,conf.int="none",col='red')
legend(20,0.9,c('Habits- [11]','Habits+ [19]'),col= c('blue','red'),bty='n',lty=1)
text(30,0.5,"pval=0.494")
title(main='Habits- vs Habits+')
dev.off()
```

B. *NR4A3*

```
library(survival)
time<-
c(1.10,3.70,3.80,3.97,7.73,8.47,11.77,12.8,12.93,14.27,18.20,19.70,19.53,20.63,20.60,23.37,23.90,25.93,27.9,28.53,28.87,31.27,47.67,54.63,64.37,74.93,76.90,91.30,89.10,97.07)
status<-c(1,1,0,1,1,1,1,1,0,0,1,0,1,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0)
treatment<-c(2,2,1,2,2,1,2,1,1,1,1,2,1,1,1,1,2,1,1,1,1,2,1,1,1,1,1,1,2,1)
fit <-survdiff(Surv(time, status) ~ treatment)
fit

library(survival)
time1<-
c(3.80,8.47,12.8,12.93,14.27,18.20,19.53,20.63,20.60,23.37,25.93,27.9,28.53,28.87,47.67,54.63,64.37,74.93,76.90,91.30,97.07,NA,NA)
status1<-c(0,1,1,0,0,1,1,0,0,0,0,1,0,0,0,0,0,0,0,0,0,1,1)
treatment1<-c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1)
time2<-c(1.10,3.70,3.97,7.73,11.77,19.70,23.90,31.27,89.10)
status2<-c(1,1,1,1,1,0,0,0,0)
treatment2<-c(2,2,2,2,2,2,2,2,2)
fit1 <-survfit(Surv(time1, status1) ~ 1)
fit2 <-survfit(Surv(time2, status2) ~ 1)
png("EVS.Alive_0.Dead.REC_1.NR4A3.Discover_Validation.4Feb2016.png")
plot(fit1,conf.int="none",col='blue',xlab='follow-up (months)',ylab= 'Cumulative Survival Probability',xlim=c(0,60))
lines(fit2,conf.int="none",col='red')
legend(20,0.9,c('NR4A3-Down [21]','NR4A3-Up [9]'),col= c('blue','red'),bty='n',lty=1)
text(30,0.6,"pval=0.034")
title(main='NR4A3-Down- vs NR4A3-Up')
dev.off()
```

C. Habits or *NR4A3*

```
library(survival)
time<-
c(1.10,3.70,3.80,3.97,7.73,8.47,11.77,12.8,12.93,14.27,18.20,19.70,19.53,20.63,20.60,23.37,23.90,25.93,27.9,28.53,28.87,31.27,47.67,54.63,64.37,74.93,76.90,91.30,89.10,97.07)
status<-c(1,1,0,1,1,1,1,1,0,0,1,0,1,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0)
treatment<-c(2,2,1,1,2,1,2,1,1,1,1,2,1,1,1,1,2,1,1,1,1,2,1,1,1,1,1,1,1,1)
fit <-survdiff(Surv(time, status) ~ treatment)
fit

library(survival)
time1<-
```

```
c(3.80,3.97,8.47,12.8,12.93,14.27,18.20,19.53,20.63,20.60,23.37,25.93,27.9,28.53,28.87,
47.67,54.63,64.37,74.93,76.90,91.30,89.10,97.07,NA,NA)
status1<-c(0,1,1,1,0,0,1,1,0,0,0,0,1,0,0,0,0,0,0,0,0,0,1,1)
treatment1<-c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1)
time2<-c(1.10,3.70,7.73,11.77,19.70,23.90,31.27)
status2<-c(1,1,1,1,0,0,0)
treatment2<-c(2,2,2,2,2,2,2)
fit1 <-survfit(Surv(time1, status1) ~ 1)
fit2 <-survfit(Surv(time2, status2) ~ 1)
png("EVS.Alive_0.Dead.REC_1.HabitsORNR4A3.Discover_Validation.4Feb2016.png")
plot(fit1,conf.int="none",col='blue',xlab='follow-up (months)',ylab= 'Cumulative Survival
Probability',xlim=c(0,60))
lines(fit2,conf.int="none",col='red')
legend(20,0.9,c('Habits- OR Nr4A3-Down [23]','Habits+ & NR4A3-Up [7]'),col=
c('blue','red'),bty='n',lty=1)
text(30,0.6,"pval=0.0464")
title(main='Habits- OR NR4A3-Down vs Habits+ & NR4A3-Up')
dev.off()
```

Discovery & Validation & TCGA_OralTongue

D. Habits

```
library(survival)
time<-
c(1.10,3.70,3.80,3.97,7.73,8.47,11.77,12.8,12.93,14.27,18.20,19.70,19.53,20.63,20.60,23.
37,23.90,25.93,27.9,28.53,28.87,31.27,47.67,54.63,64.37,74.93,76.90,91.30,89.10,97.07,
91.3666666667,2.1666666667,22.2,7.2333333333,8.5333333333,2.5333333333,53.0333
333333,48.1333333333,2.1333333333,6.4666666667,23.1666666667,3.7)
status<-c(1,1,0,1,1,1,1,1,0,0,1,0,1,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,1,1,1,1,1,1,1,0,1,1,0,0)
treatment<-
c(2,2,2,1,2,2,2,1,1,2,1,2,2,1,2,2,2,2,2,1,2,2,2,2,2,1,1,1,1,1,2,2,2,1,2,2,2,2,1,2,1,2,1)
fit <-survdiff(Surv(time, status) ~ treatment)
fit

library(survival)
time1<-
c(3.97,12.8,12.93,18.20,20.63,28.53,64.37,74.93,76.90,91.30,89.10,22.2,48.1333333333,
6.4666666667,3.7)
status1<-c(1,1,0,1,0,0,0,0,0,0,0,1,0,1,0)
treatment1<-c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1)
time2<-
c(1.10,3.70,3.80,7.73,8.47,11.77,14.27,19.70,19.53,20.60,23.37,23.90,25.93,27.9,28.87,3
1.27,47.67,54.63,97.07,NA,91.3666666667,2.1666666667,7.2333333333,8.5333333333,2
.5333333333,53.0333333333,2.1333333333,23.1666666667)
status2<-c(1,1,0,1,1,1,0,0,1,0,0,0,0,1,0,0,0,0,0,1,1,1,1,1,1,1,1,0)
treatment2<-c(2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2)
fit1 <-survfit(Surv(time1, status1) ~ 1)
fit2 <-survfit(Surv(time2, status2) ~ 1)
png("EVS.Alive_0.Dead.REC_1.Habits.Discover_Validation_TCGA_OT.4Feb2016.png")
plot(fit1,conf.int="none",col='blue',xlab='follow-up (months)',ylab= 'Cumulative Survival
Probability',xlim=c(0,60))
```

```
lines(fit2,conf.int="none",col='red')
legend(20,0.9,c('Habits- [15]','Habits+ [27]'),col= c('blue','red'),bty='n',lty=1)
text(30,0.6,"pval=0.283")
title(main='Habits- vs Habits+')
dev.off()
```

E. *NR4A3*

```
library(survival)
time<-
c(1.10,3.70,3.80,3.97,7.73,8.47,11.77,12.8,12.93,14.27,18.20,19.70,19.53,20.63,20.60,23.
37,23.90,25.93,27.9,28.53,28.87,31.27,47.67,54.63,64.37,74.93,76.90,91.30,89.10,97.07,
91.3666666667,2.1666666667,22.2,7.2333333333,8.5333333333,2.5333333333,53.0333
333333,48.1333333333,2.1333333333,6.4666666667,23.1666666667,3.7)
status<-c(1,1,0,1,1,1,1,1,0,0,1,0,1,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,1,1,1,1,1,1,1,0,1,1,0,0)
treatment<-
c(2,2,1,2,2,1,2,1,1,1,1,2,1,1,1,1,2,1,1,1,1,2,1,1,1,1,1,1,2,1,1,2,2,1,2,2,2,1,1,1,2,2)
fit <-survdiff(Surv(time, status) ~ treatment)
fit
```

```
library(survival)
time1<-
c(3.80,8.47,12.8,12.93,14.27,18.20,19.53,20.63,20.60,23.37,25.93,27.9,28.53,28.87,47.67
,54.63,64.37,74.93,76.90,91.30,97.07,NA,NA,91.3666666667,7.2333333333,48.13333333
33,2.1333333333,6.4666666667)
status1<-c(0,1,1,0,0,1,1,0,0,0,0,1,0,0,0,0,0,0,0,0,1,1,1,1,0,1,1)
treatment1<-c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1)
time2<-
c(1.10,3.70,3.97,7.73,11.77,19.70,23.90,31.27,89.10,2.1666666667,22.2,8.5333333333,2.
5333333333,53.0333333333,23.1666666667,3.7)
status2<-c(1,1,1,1,1,0,0,0,0,1,1,1,1,1,0,0)
treatment2<-c(2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2)
fit1 <-survfit(Surv(time1, status1) ~ 1)
fit2 <-survfit(Surv(time2, status2) ~ 1)
png("EVS.Alive_0.Dead.REC_1.NR4A3.Discover_Validation_TCGA_OT.4Feb2016.png")
plot(fit1,conf.int="none",col='blue',xlab='follow-up (months)',ylab= 'Cumulative Survival
Probability',xlim=c(0,60))
lines(fit2,conf.int="none",col='red')
legend(20,0.9,c('NR4A3-Down [26]','NR4A3-Up [16]'),col= c('blue','red'),bty='n',lty=1)
text(30,0.6,"pval=0.0191")
title(main='NR4A3-Down- vs NR4A3-Up')
dev.off()
```

F. Habits or *NR4A3*

```
library(survival)
time<-
c(1.10,3.70,3.80,3.97,7.73,8.47,11.77,12.8,12.93,14.27,18.20,19.70,19.53,20.63,20.60,23.
37,23.90,25.93,27.9,28.53,28.87,31.27,47.67,54.63,64.37,74.93,76.90,91.30,89.10,97.07,
91.3666666667,2.1666666667,22.2,7.2333333333,8.5333333333,2.5333333333,53.0333
333333,48.1333333333,2.1333333333,6.4666666667,23.1666666667,3.7)
status<-c(1,1,0,1,1,1,1,1,0,0,1,0,1,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,1,1,1,1,1,1,1,0,1,1,0,0)
```

```
treatment<-
c(2,2,1,1,2,1,2,1,1,1,1,2,1,1,1,1,2,1,1,1,1,2,1,1,1,1,1,1,1,1,2,1,1,2,1,1,1,2,1,1,1,1,2,1,1,1,1,2,1,1,1,1,1,1,1,1,2,1,1,2,2,2,1,1,1,2,1)
fit <-survdiff(Surv(time, status) ~ treatment)
fit

library(survival)
time1<-
c(3.80,3.97,8.47,12.8,12.93,14.27,18.20,19.53,20.63,20.60,23.37,25.93,27.9,28.53,28.87,
47.67,54.63,64.37,74.93,76.90,91.30,89.10,97.07,NA,NA,91.3666666667,22.2,7.2333333
333,48.1333333333,2.1333333333,6.4666666667,3.7)
status1<-c(0,1,1,1,0,0,1,1,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,1,1,1,1,1,0,1,1,0)
treatment1<-c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1)
time2<-
c(1.10,3.70,7.73,11.77,19.70,23.90,31.27,2.1666666667,8.5333333333,2.5333333333,53.
0333333333,23.1666666667)
status2<-c(1,1,1,1,0,0,0,1,1,1,1,0)
treatment2<-c(2,2,2,2,2,2,2,2,2,2,2,2)
fit1 <-survfit(Surv(time1, status1) ~ 1)
fit2 <-survfit(Surv(time2, status2) ~ 1)
png("EVS.Alive_0.Dead.REC_1.HabitsORNR4A3.Discover_Validation_TCGA_OT.4Feb20
16.png")
plot(fit1,conf.int="none",col='blue',xlab='follow-up (months)',ylab= 'Cumulative Survival
Probability',xlim=c(0,60))
lines(fit2,conf.int="none",col='red')
legend(20,0.9,c('Habits- OR Nr4A3-Down [30]','Habits+ & NR4A3-Up [12]'),col=
c('blue','red'),bty='n',lty=1)
text(30,0.5,"pval=0.0215")
title(main='Habits- OR NR4A3-Down vs Habits+ & NR4A3-Up')
```