

Supporting Information for Hilvo et al. (2015)

May 13, 2015

Contents

1	Libraries and functions	2
2	External Data Sources	2
2.1	TCGA data	2
2.2	Copy Number Alterations (CNA)	3
3	Survival analysis for the TCGA cohort	4
3.1	Supplementary Table S7: survival analyses using TCGA data	6
3.2	Figure 2E and Figure S5: Low gene expression and loss of ALDH5A1 copy number were both associated to poor overall survival	7
4	Session info: R-packages and their versions used for this analysis	43

1 Libraries and functions

The R scripts provided in this vignette fully reproduces the TCGA survival analyses of Hilvo et al (2015). The following R packages may be required:

- `survival`

We use a local library for some low level functions. This will be loaded from `http://www.markowetzlab.org/supplements/` if not present on the local machine.

2 External Data Sources

The package `Hilvo2015-supplement-data.Rdata` contains the publicly available data described in the following subsections. This will be loaded from `http://www.markowetzlab.org/supplements/` if not present on the local machine.

```
## load data file from local copy or from URL
if (file.exists("Hilvo2015-supplementary-data.Rdata")){
  load("Hilvo2015-supplementary-data.Rdata")
  cat("Hilvo2015-supplementary-data.Rdata loaded from local copy")
} else {
  load(url("http://www.markowetzlab.org/supplements/Hilvo2015-supplementary-data.Rdata"))
  cat("Hilvo2015-supplementary-data.Rdata loaded from URL") }

## Hilvo2015-supplementary-data.Rdata loaded from local copy
```

2.1 TCGA data

Data from 489 high-grade serous ovarian cancers reported by The Cancer Genome Atlas Research Network (Nature, 2011, 609-615) were obtained from `https://tcga-data.nci.nih.gov/docs/publications/ov_2011/`. The `exp.TCGAOC` object contains the gene expression matrix obtained from `https://tcga-data.nci.nih.gov/docs/publications/ov_2011/`:

```
dim(exp.TCGAOC)

## [1] 11864 489
```

The `clinical.TCGAOC` object contains the clinical annotations for all samples obtained from: `https://tcga-data.nci.nih.gov/docs/dictionary/TCGA_BCR_DataDictionary.xml`

```
dim(clinical.TCGAOC)

## [1] 488 12

colnames(clinical.TCGAOC)

## [1] "AgeAtDiagnosis..yrs." "VITALSTATUS"
## [3] "TUMORSTAGE" "TUMORGRADE"
## [5] "TUMORRESIDUALDISEASE" "PRIMARYTHERAPYOUTCOMESUCCESS"
## [7] "PERSONNEOPLASMCANCERSTATUS" "OverallSurvival.mos."
## [9] "ProgressionFreeStatus" "ProgressionFreeSurvival..mos.."
## [11] "PlatinumFreeInterval..mos.." "PlatinumStatus"
```

2.2 Copy Number Alterations (CNA)

The CNA data was downloaded from cBioPortal (<http://cbioportal.org>) using the TCGA set of 489 high-grade serous ovarian cancers. Only 316 tumours have CNA data available:

```
dim(CNAtable)

## [1] 316 69
```

3 Survival analysis for the TCGA cohort

To validate the metabolomics findings with gene expression data, survival analyses were performed for The Cancer Genome Atlas (TCGA) data. Survival analysis based on gene expression and copy number was performed for genes of those selected KEGG pathways that showed most interesting metabolomics results: fatty acid import and beta oxidation, omega oxidation, ketone body production, pentose phosphate pathway reactions related to 3-erythritol accumulation, leucine degradation reactions related to 3-hydroxyisovaleric acid accumulation, peroxisomal biogenesis factors related to Zellweger syndrome as well as ALDH5A1 gene encoding for SSADH enzyme.

```
colnames(CNAtable)
```

```
## [1] "ACAA1" "ACAA2" "ACADL" "ACADM" "ACADS" "ACADSB" "ACADVL"
## [8] "ACAT1" "ACOX1" "ACSBG1" "ACSBG2" "ACSL1" "ACSL3" "ACSL5"
## [15] "ACSL6" "ADH1A" "ADH1B" "ADH1C" "ADH5" "ADH6" "ADH7"
## [22] "ALDH1B1" "ALDH2" "ALDH3A2" "ALDH5A1" "ALDH7A1" "ALDH9A1" "AUH"
## [29] "BDH1" "BDH2" "BTD" "CPT1A" "CPT1B" "CPT2" "CYP4A11"
## [36] "CYP4B1" "CYP4F11" "DLD" "ECHS1" "ECI1" "ECI2" "EHHADH"
## [43] "FABP4" "HADH" "HADHA" "HADHB" "HLCS" "HMGCL" "HMGCS1"
## [50] "HMGCS2" "MCCC1" "MCCC2" "OXCT1" "OXCT2" "PEX1" "PEX10"
## [57] "PEX12" "PEX13" "PEX14" "PEX16" "PEX19" "PEX3" "PEX5"
## [64] "PEX6" "PEX7" "RPIA" "SHPK" "TALDO1" "TKT"
```

For each gene of interest, we subdivided the TCGA Ovarian Cancer cohort according to:

- Four quartiles of gene expression
- Different copy number alterations: Loss, wt and gain

First, we simplified the Gistic scores obtained from the downloaded table as "loss", "wt", "gain":

```
CNAallgenes <- CNAtable
CNAallgenes[CNAallgenes== -2] <- "loss" #hom del
CNAallgenes[CNAallgenes== -1] <- "loss" #het loss
CNAallgenes[CNAallgenes== 0] <- "wt"
CNAallgenes[CNAallgenes== 1] <- "gain" #gain
CNAallgenes[CNAallgenes== 2] <- "gain" #amp
CNAallgenes <- t(CNAallgenes)
```

```
# Survival wrapper function
survwrapp <- function(gene, plot = TRUE) {
```

```

# par(mfrow=c(1,2),mar=c(20,4,5,4))

# by quartiles
group <- cut(as.numeric(exp.TCGAOC[gene, ]), quantile(as.numeric(exp.TCGAOC[gene,
  ]), probs = c(0, 0.25, 0.5, 0.75, 1)))
names(group) <- names(exp.TCGAOC[gene, ])
survmat <- cbind(groups = group[rownames(clinical.TCGAOC)], clinical.TCGAOC)
pvalue1 <- plotsurv(survmat, colors = c("blue", "lightblue", "gray", "red"),
  maint = "", plot = plot)
if (plot) {
  title(main = paste(gene, "\nBy expression quartile"))
}

# by mutation type
group <- factor(CNAallgenes[gene, ], levels = c("loss", "wt", "gain"))
survmat <- cbind(groups = group[rownames(clinical.TCGAOC)], clinical.TCGAOC)
pvalue2 <- plotsurv(survmat, colors = c("turquoise4", "darkgray", "orangered"),
  maint = "", plot = plot)

if (plot) {
  title(main = paste(gene, "\nBy CNA type"))
}

# this function returns the chisq p-values obtained in both cases
# 'by_ExpQuartiles' and 'by_CNA'
return(c(by_ExpQuartiles = pvalue1, by_CNA = pvalue2))
}

plotsurv <- function(survmat, colors, maint, timecol = "OverallSurvival.mos.",
  eventcol = "VITALSTATUS", plot = TRUE) {
  library(survival)

  # fix survmat
  if (eventcol == "VITALSTATUS") {
    survmat$VITALSTATUS <- as.character(survmat$VITALSTATUS)
    survmat$VITALSTATUS[survmat$VITALSTATUS == "LIVING"] <- 0
    survmat$VITALSTATUS[survmat$VITALSTATUS == "DECEASED"] <- 1
    survmat$VITALSTATUS <- as.numeric(survmat$VITALSTATUS)
  }
  if (eventcol == "ProgressionFreeStatus") {
    survmat$ProgressionFreeStatus <- as.character(survmat$ProgressionFreeStatus)
    survmat$ProgressionFreeStatus[survmat$ProgressionFreeStatus == "DiseaseFree"] <-
    survmat$ProgressionFreeStatus[survmat$ProgressionFreeStatus == "Recurred/Progression"]
  }
}

```

```

    survmat$ProgressionFreeStatus <- as.numeric(survmat$ProgressionFreeStatus)
  }
  survmat <- subset(survmat, select = c("groups", timecol, eventcol))
  survmat <- na.omit(survmat)

  # plot
  ss <- Surv(as.numeric(survmat[, timecol]), survmat[, eventcol])
  if (plot) {
    plot(survfit(ss ~ survmat$groups), col = colors, xlab = paste(timecol,
      "[months]"), ylab = "Survival Probability", lwd = 3, las = 1, main = maint)
    legendtext = paste(levels(survmat$group), ": n=", table(survmat$group),
      sep = "")
    legend("topright", col = colors, legend = legendtext, lty = 1)
  }

  # Pvalue and HR
  res <- survdiff(ss ~ survmat$group)

  if (plot) {
    p <- format(1 - pchisq(res$chisq, 1), scientific = TRUE, nsmall = 3)
    text(100, 0.6, paste("p-value=", p))
    options(scipen = 10)
  }

  p <- 1 - pchisq(res$chisq, 1)
  return(p)
}

```

3.1 Supplementary Table S7: survival analyses using TCGA data

Survival analysis was performed for the following genes:

```

genes <- c('ALDH5A1', 'DLD', 'PEX6', 'HMGCL', 'ACSL3', 'ADH1B', 'HLCS', 'AUH', 'ALDH7A1',
'MCCC2', 'PEX12', 'OXCT1', 'ALDH1B1', 'EHHADH', 'HMGCS1', 'FABP4', 'ECI1', 'HADHA', 'CPT1B')

```

For each one of the selected genes, we investigated if low expression and loss in copy numbers of the gene were associated with worse overall survival. The following code reproduces the p-values in supplementary table 7.

```

library(knitr)
pvalues <- t(sapply(genes, survwrapp, plot=FALSE))
kable(pvalues)

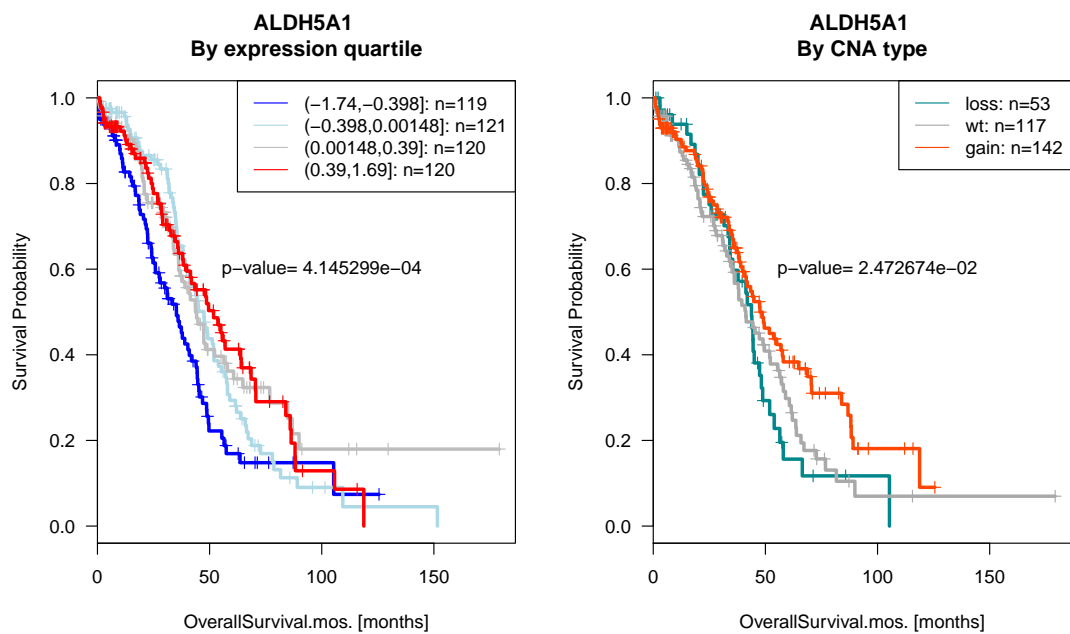
```

	by_ExpQuartiles	by_CNA
ALDH5A1	0.0004145	0.0247267
DLD	0.0004627	0.1200124
PEX6	0.0008578	0.2345744
HMGCL	0.0043549	0.5601570
ACSL3	0.0092199	0.1371937
ADH1B	0.0270184	0.6412994
HLCS	0.0285858	0.3880522
AUH	0.0332415	0.0927564
ALDH7A1	0.0285707	0.0224340
MCCC2	0.8866716	0.0001703
PEX12	0.4274987	0.0031190
OXCT1	0.1775536	0.0035922
ALDH1B1	0.2173628	0.0048617
EHHADH	0.0625108	0.0096193
HMGCS1	0.0815992	0.0119629
FABP4	0.0587003	0.0178226
ECI1	0.3547530	0.0251559
HADHA	0.0429957	0.0310051
CPT1B	0.0880067	0.0403244

3.2 Figure 2E and Figure S5: Low gene expression and loss of ALDH5A1 copy number were both associated to poor overall survival

Intriguingly, the most significant findings were obtained for the ALDH5A1 gene, as in the TCGA data set both low expression and loss in copy numbers of the gene were associated with worse overall survival of the patients.

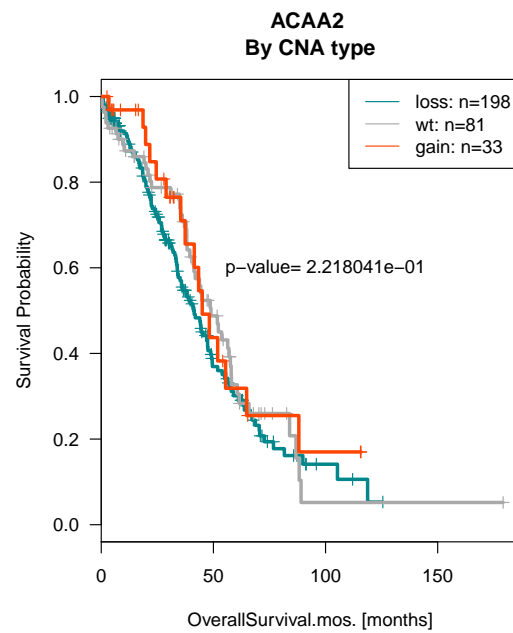
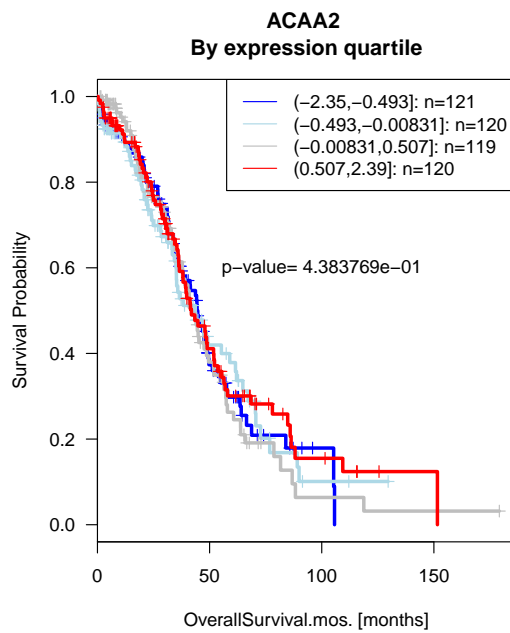
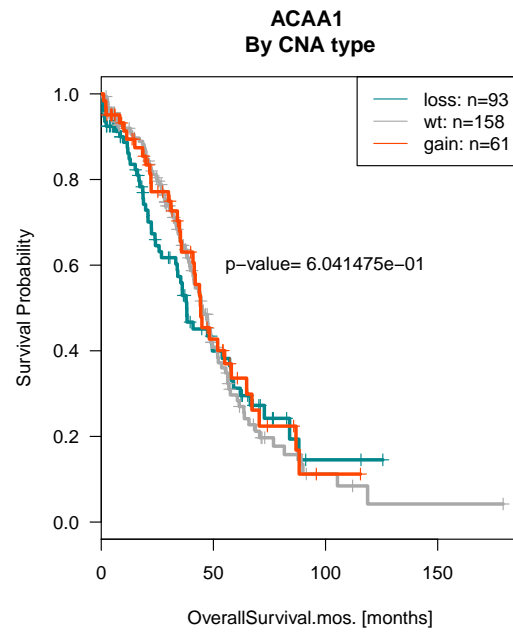
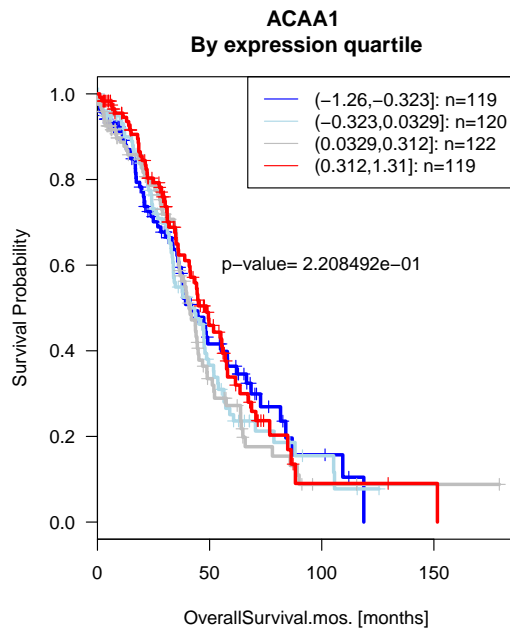
```
par(mfrow=c(1,2))
survwrapp("ALDH5A1")
```

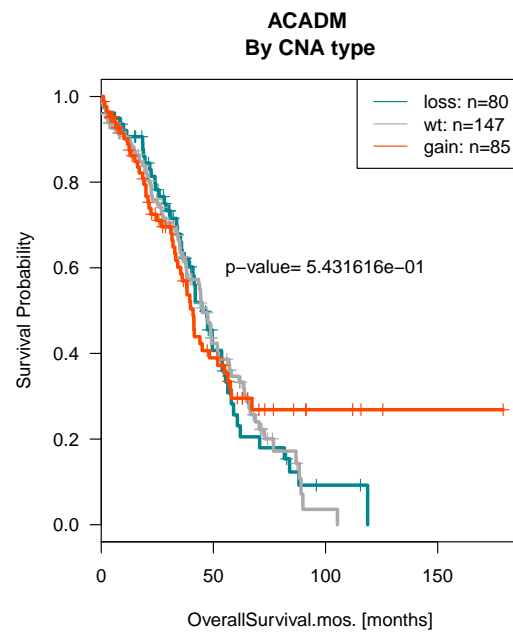
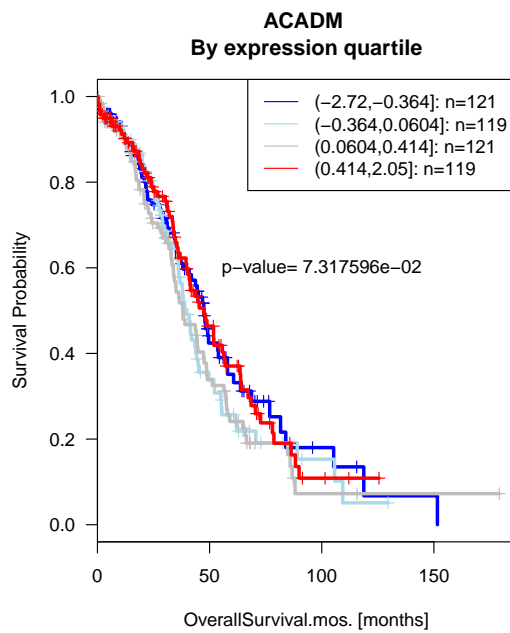
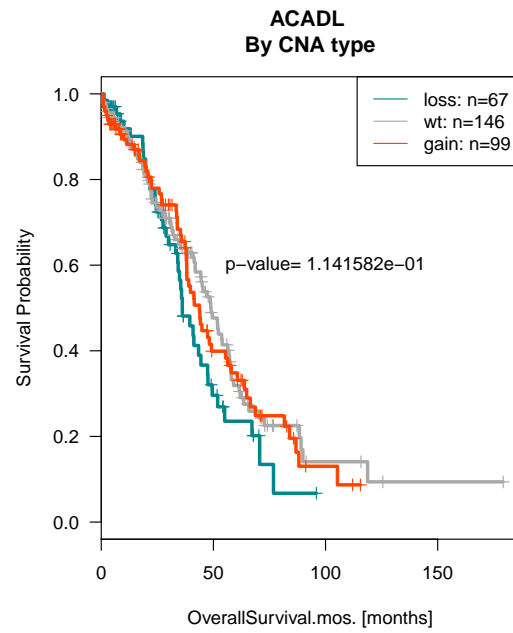
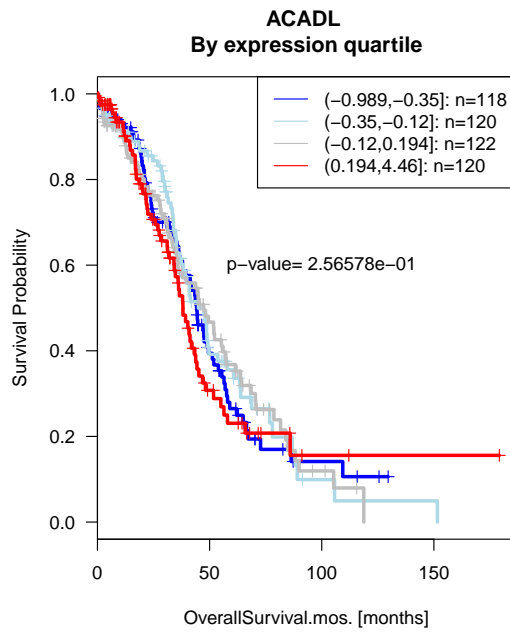


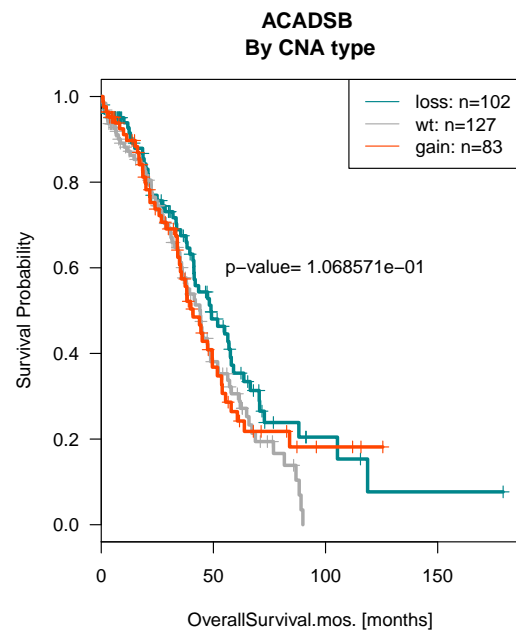
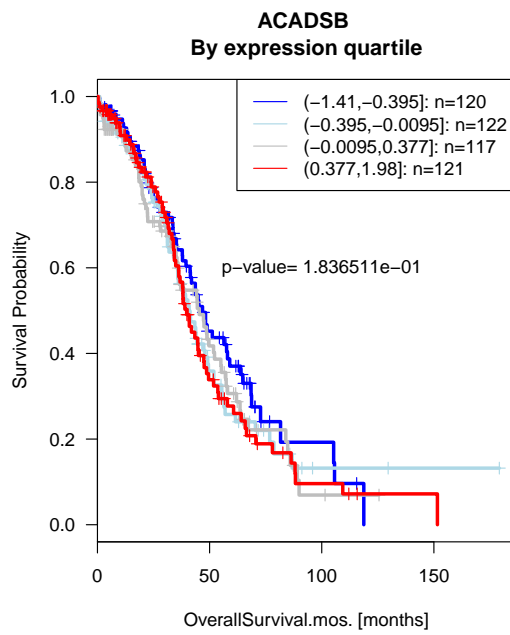
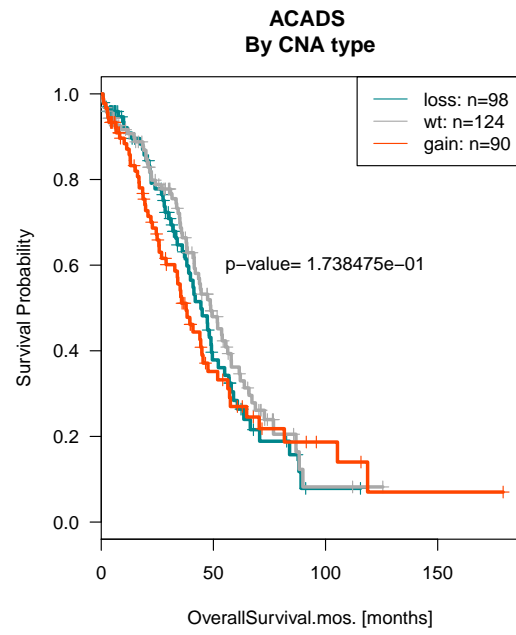
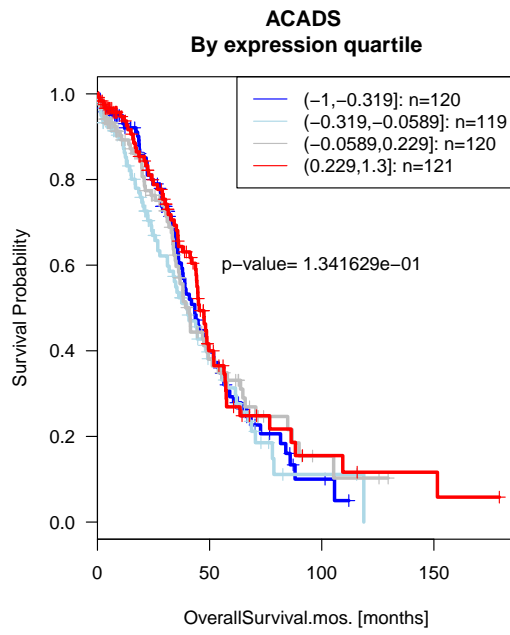
```
## by_ExpQuartiles      by_CNA
##      0.0004145299    0.0247267415
```

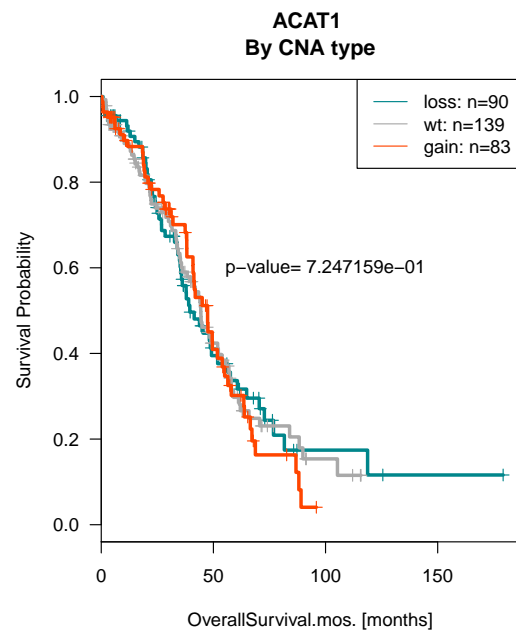
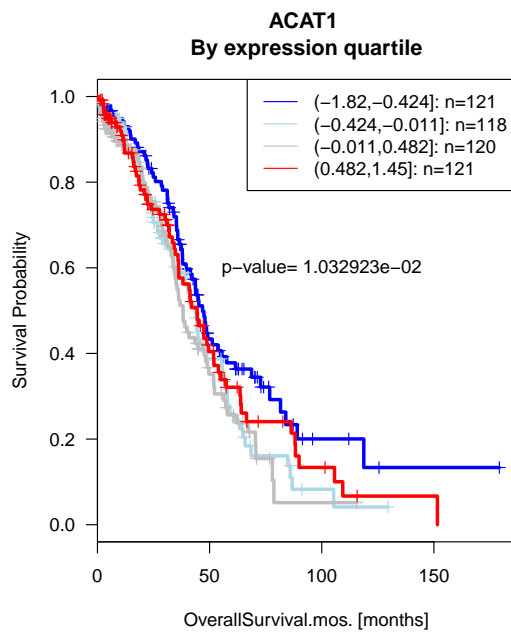
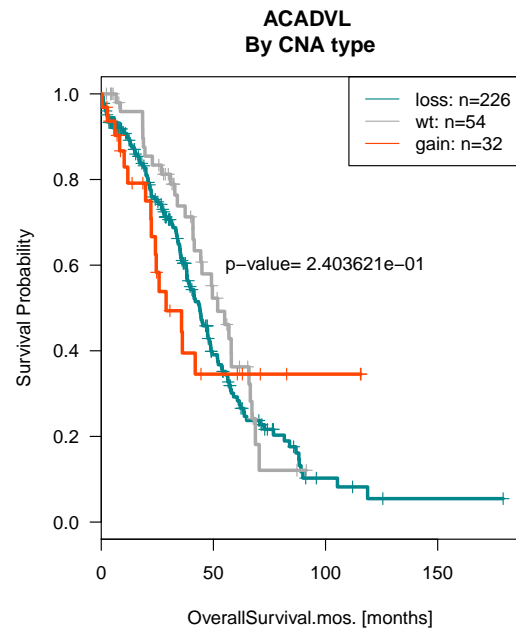
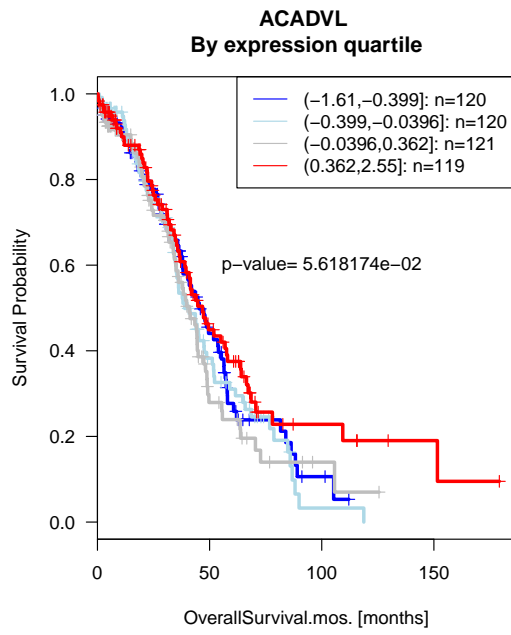
The survival plots for all analysed genes can be inspected with the following code:

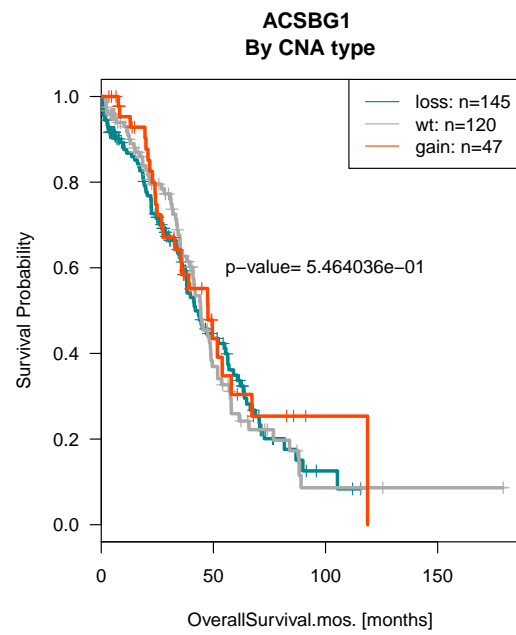
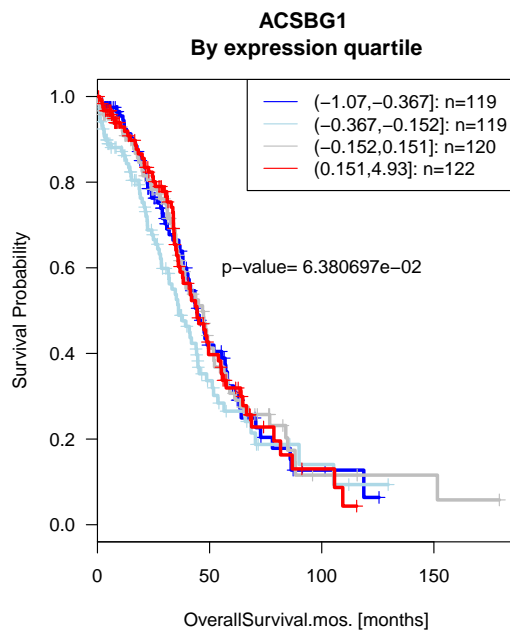
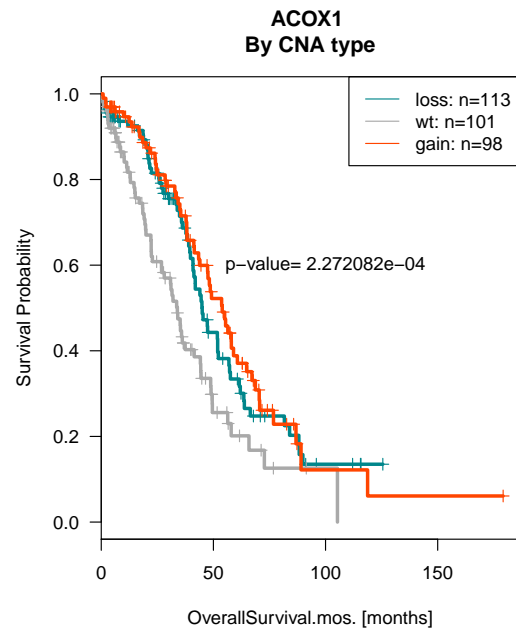
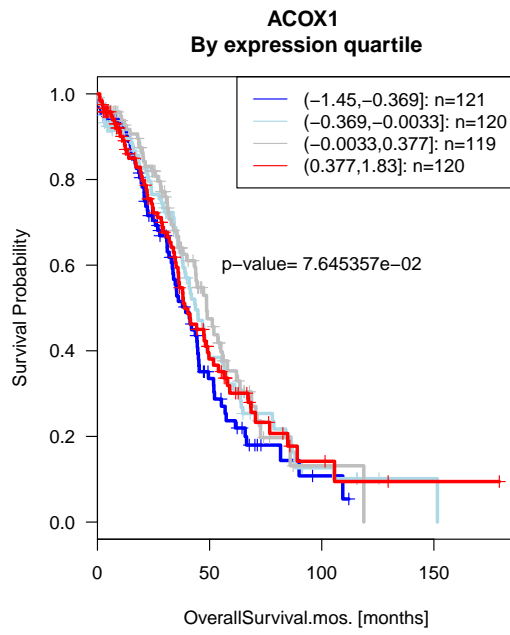
```
par(mfrow=c(1,2))
pvalues <- sapply(colnames(CNatable), survwrapp, plot=TRUE)
```

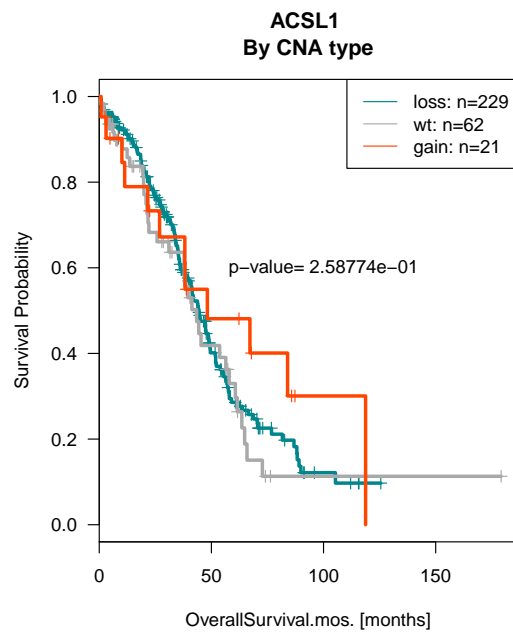
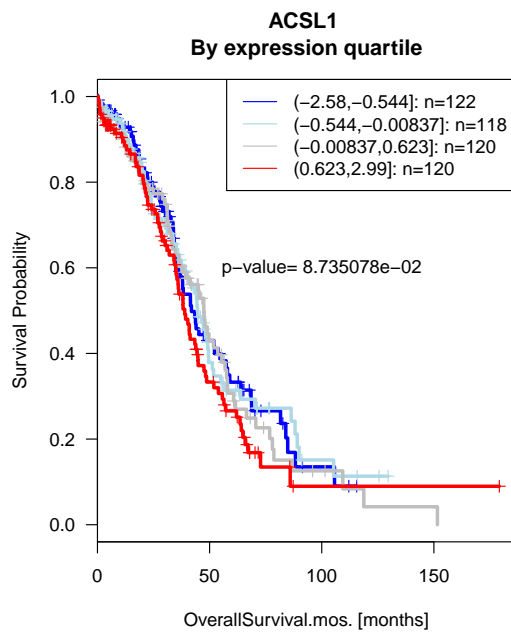
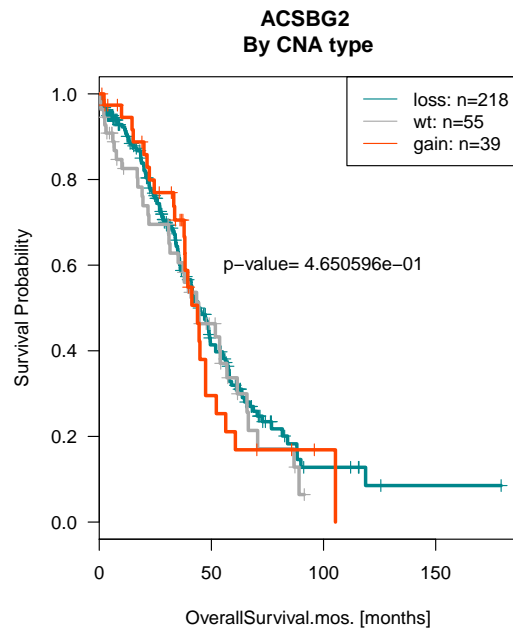
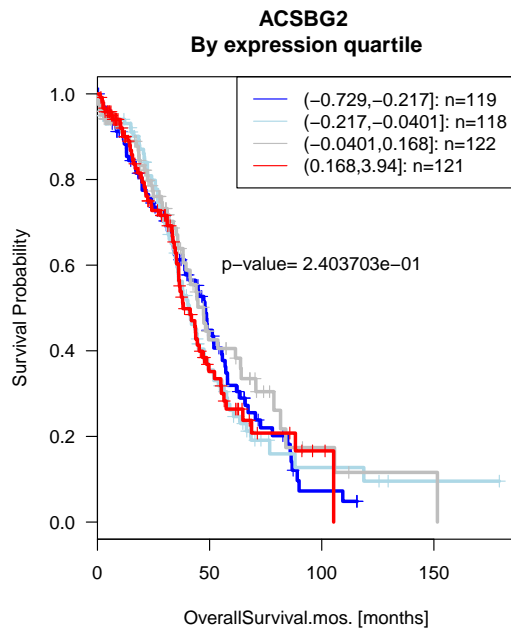



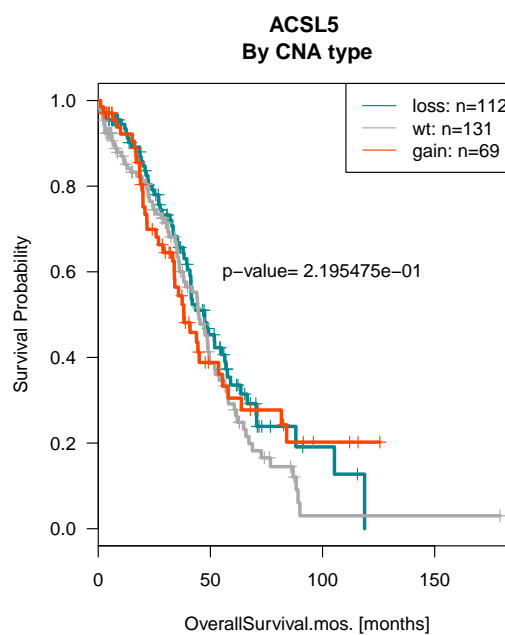
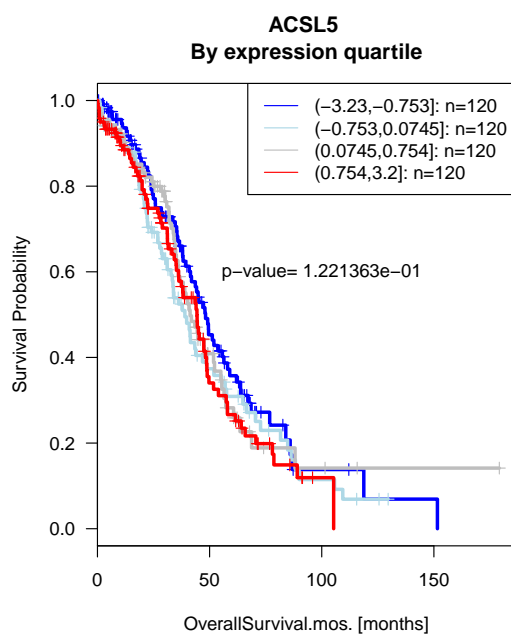
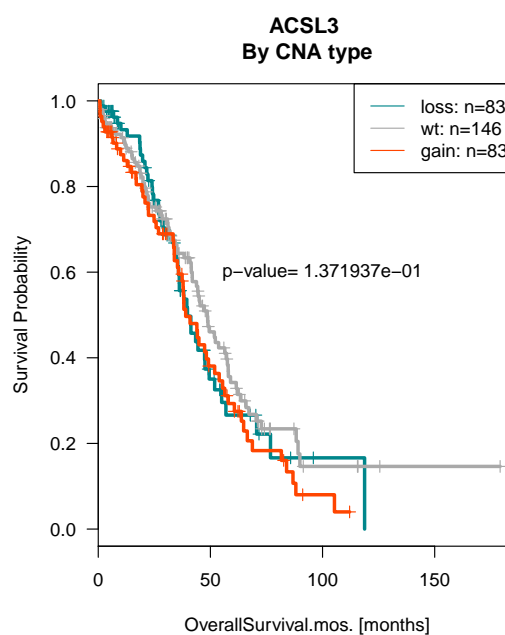
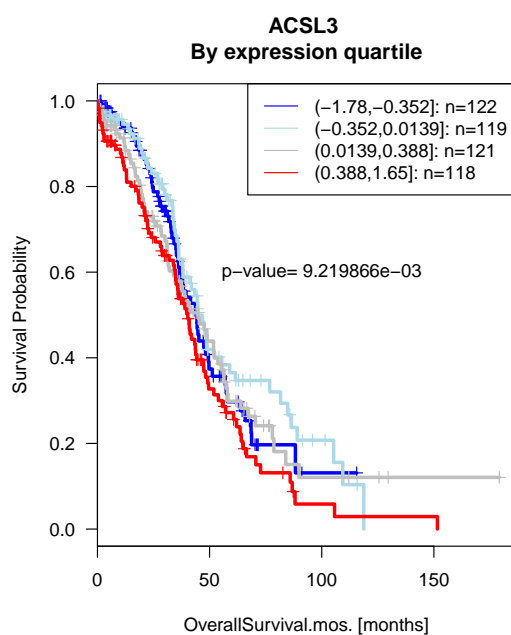


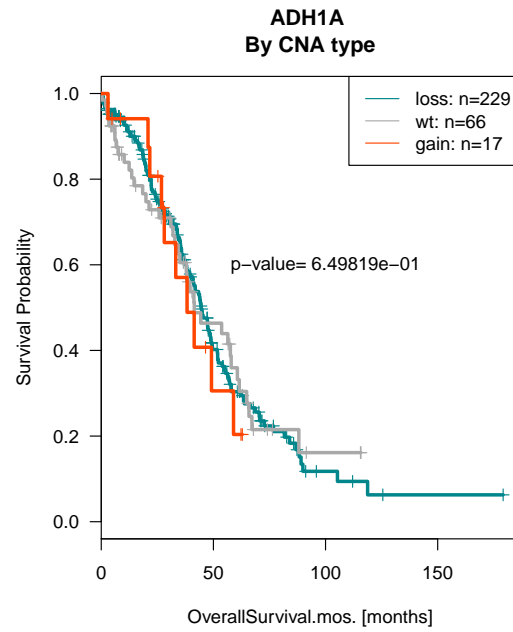
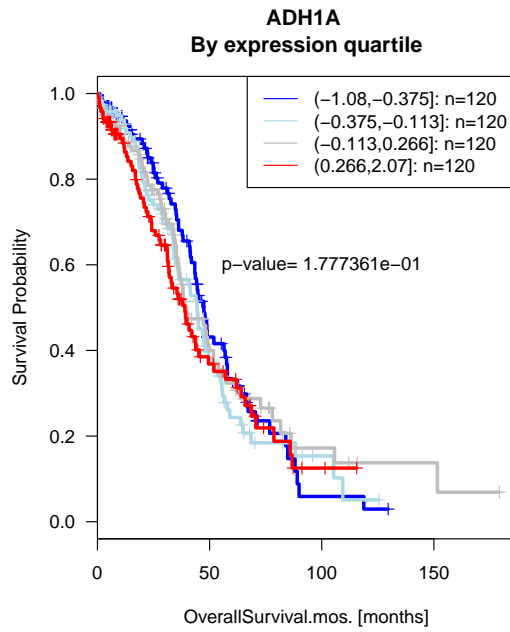
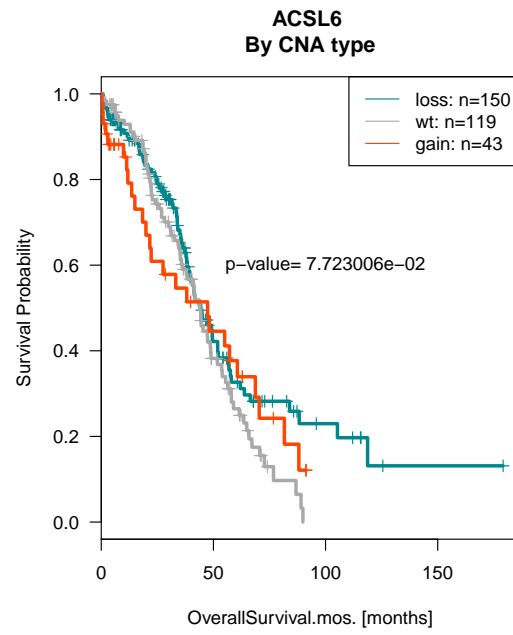
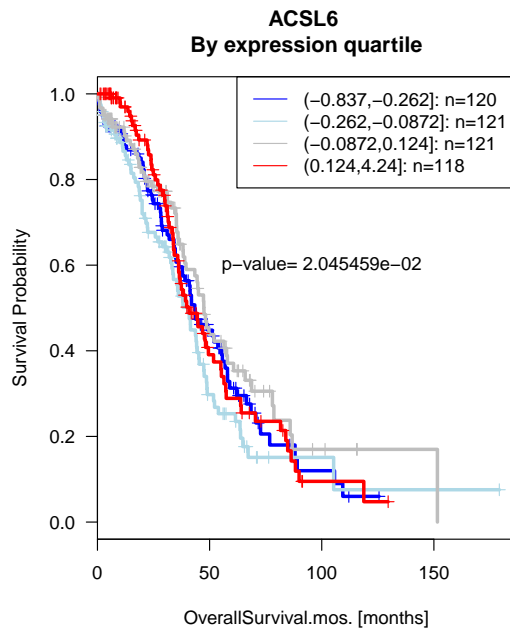


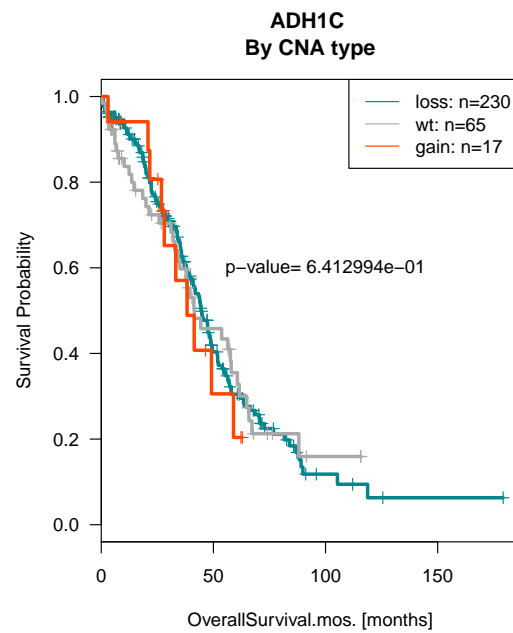
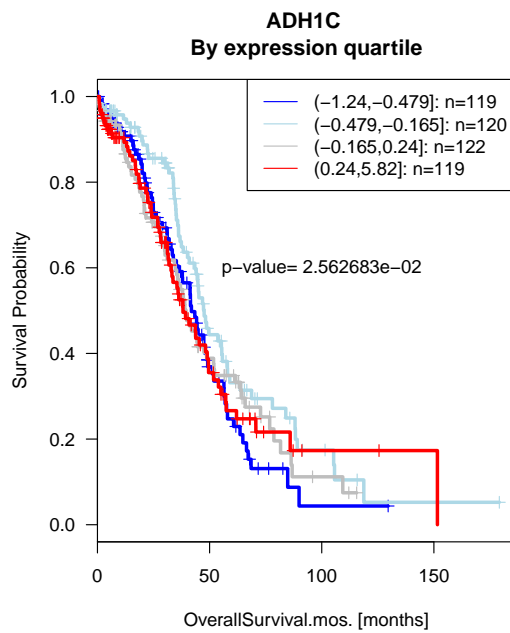
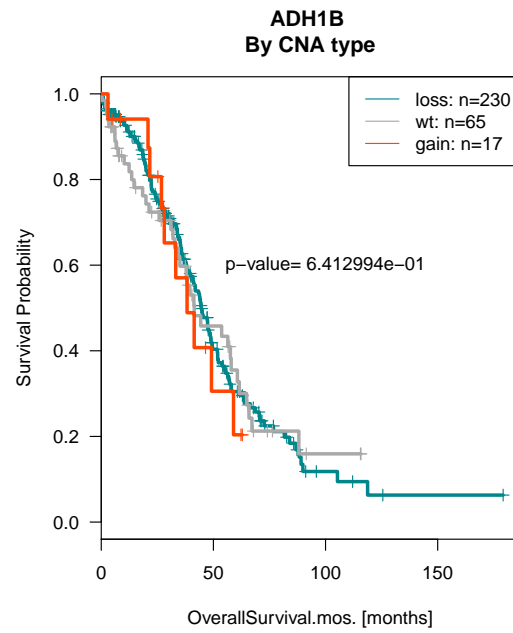
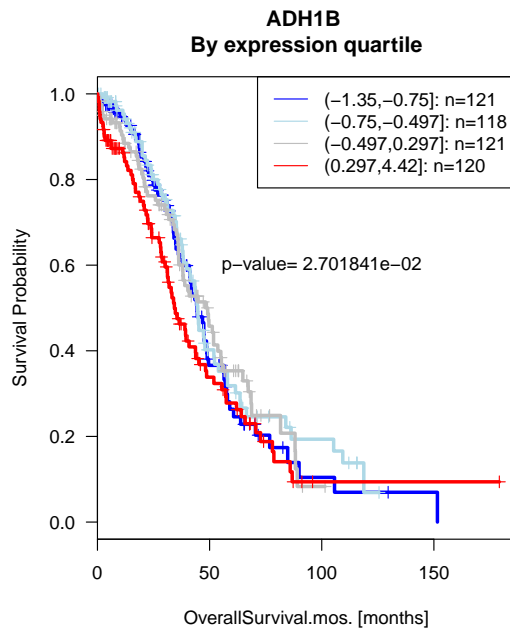


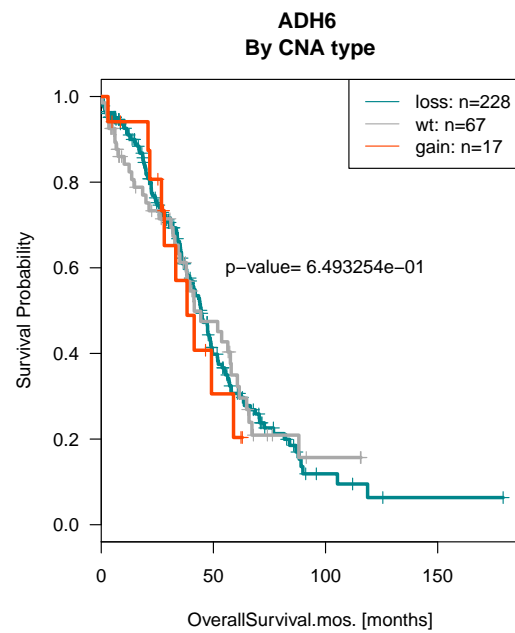
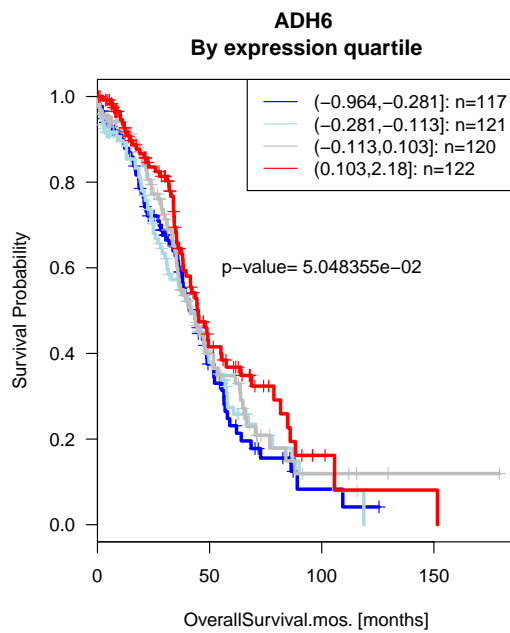
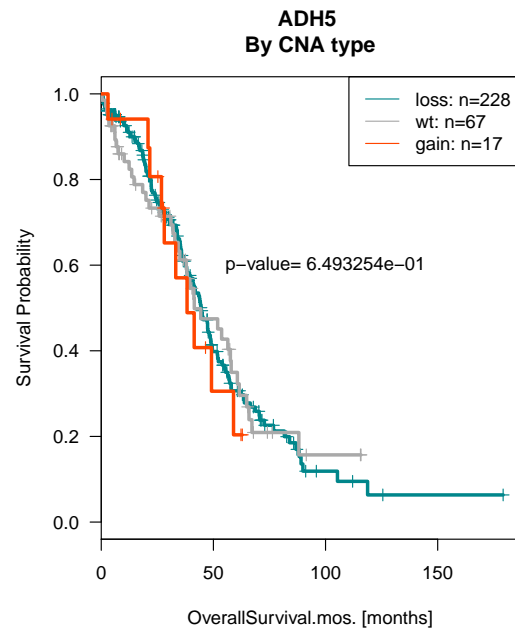
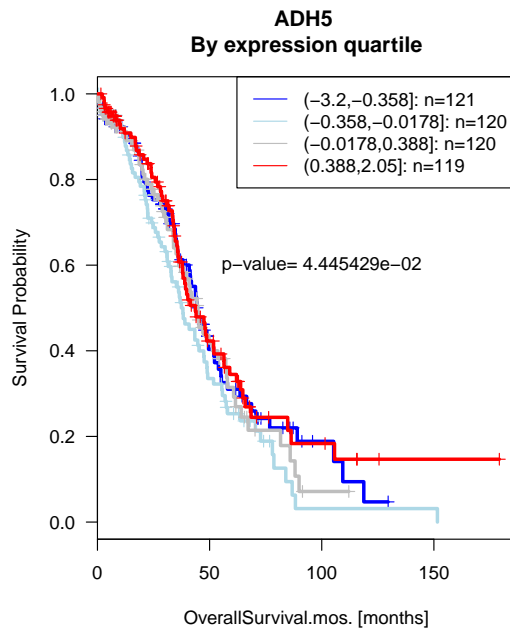


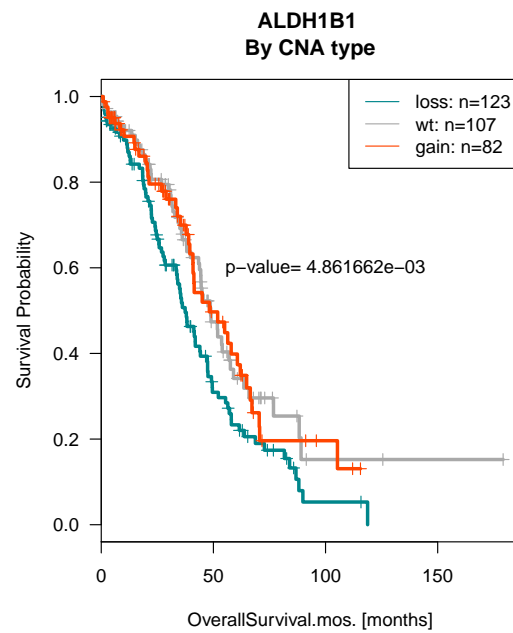
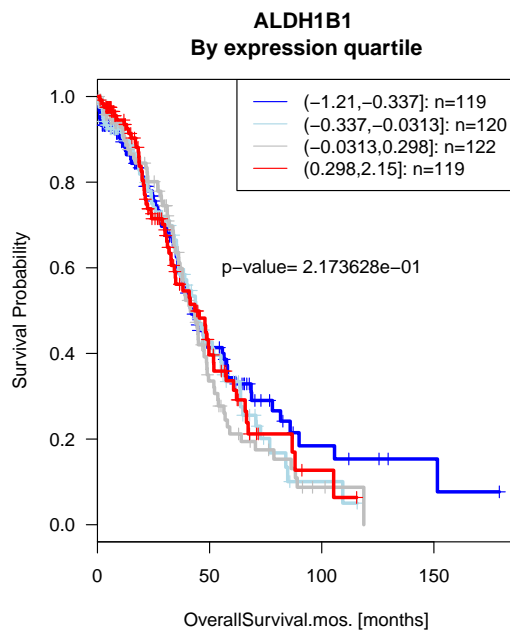
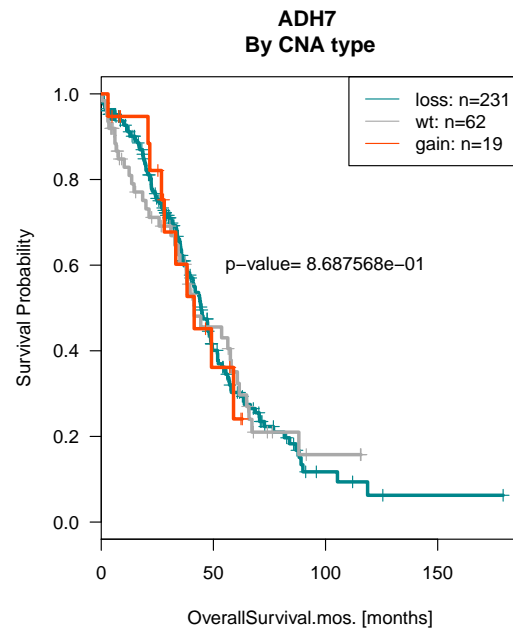
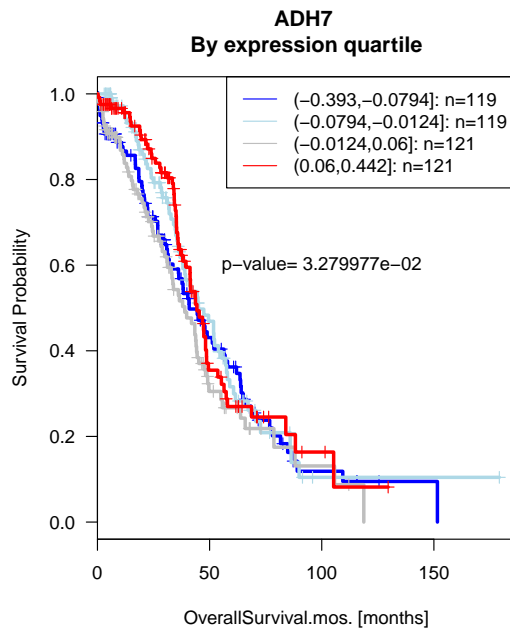


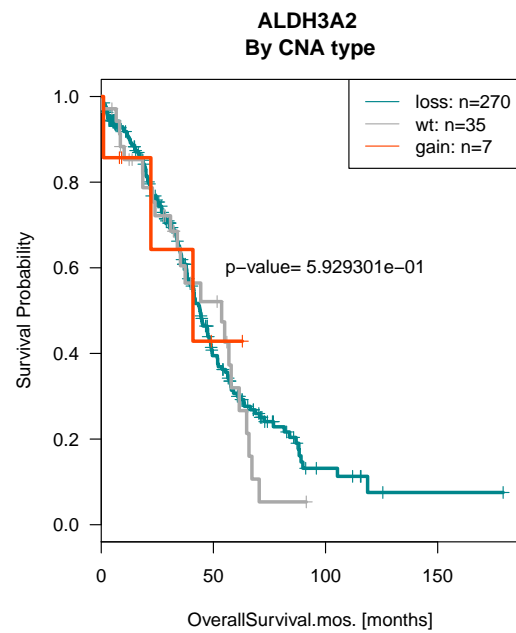
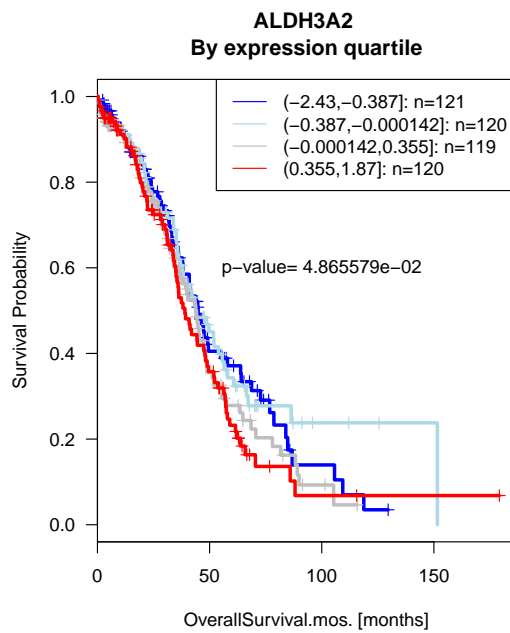
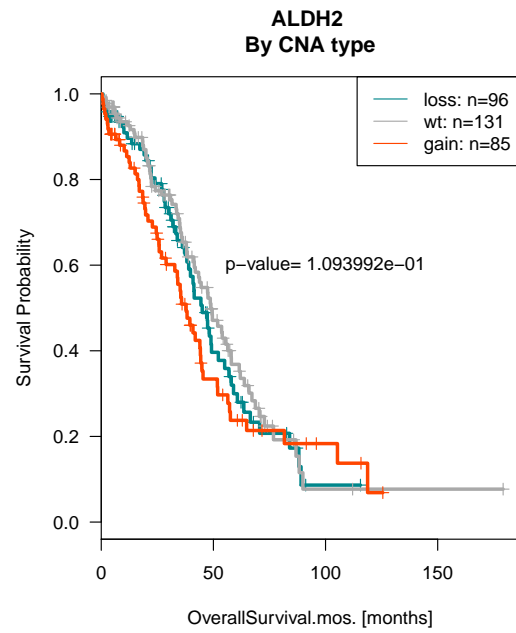
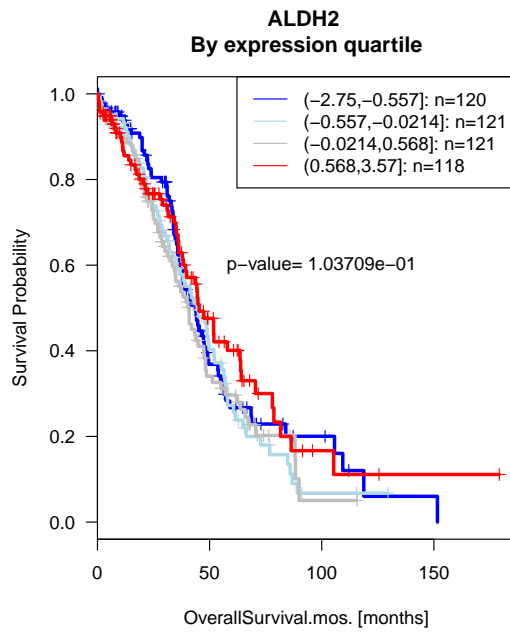


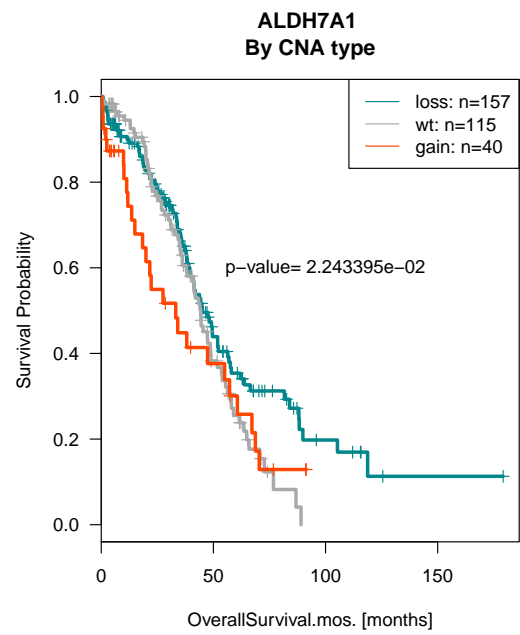
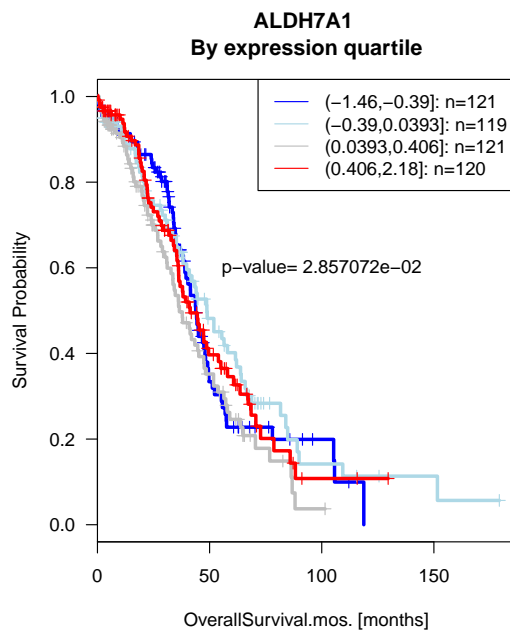
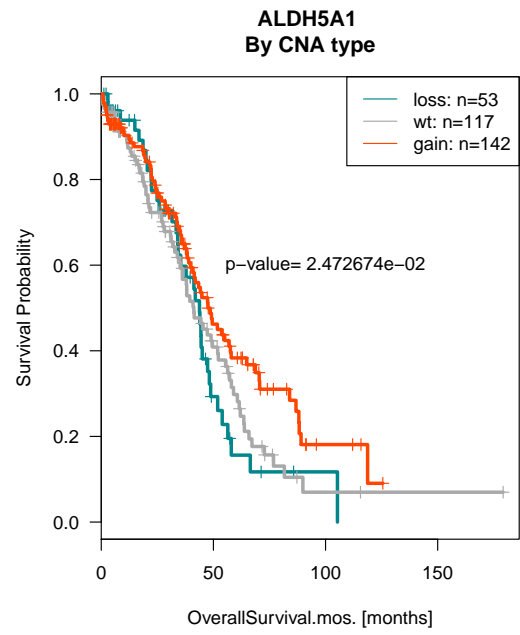
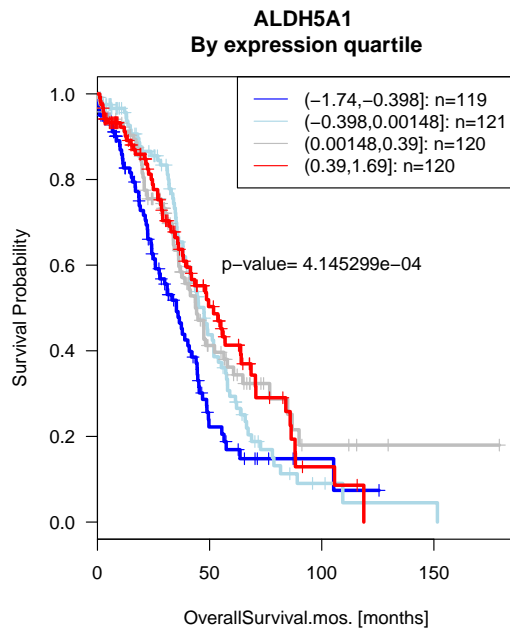


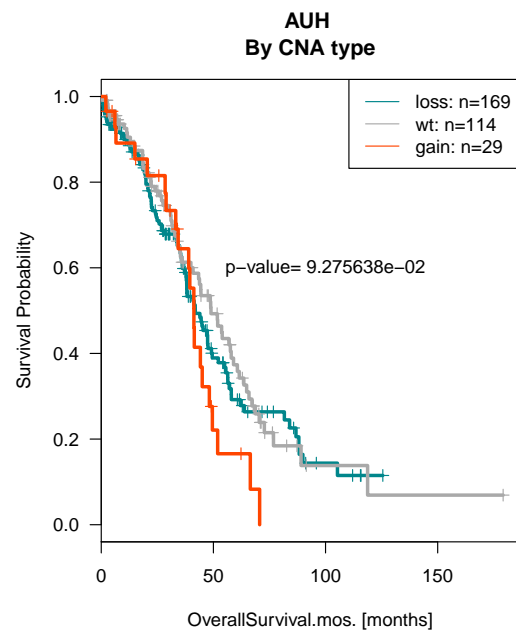
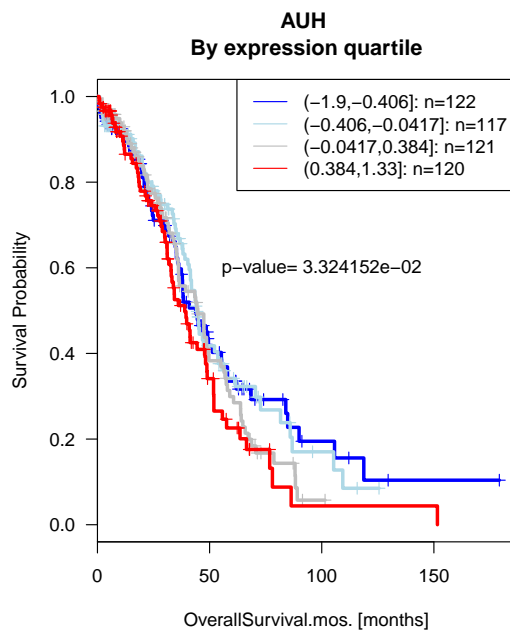
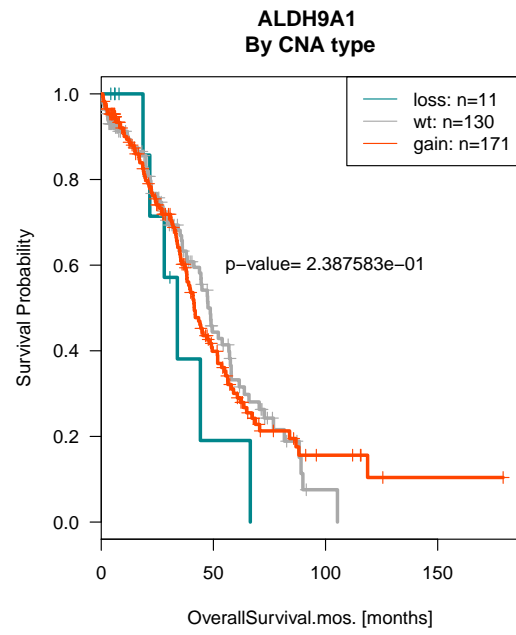
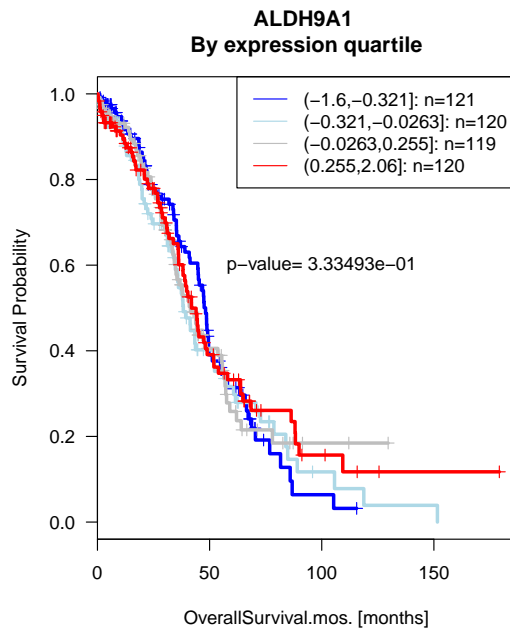


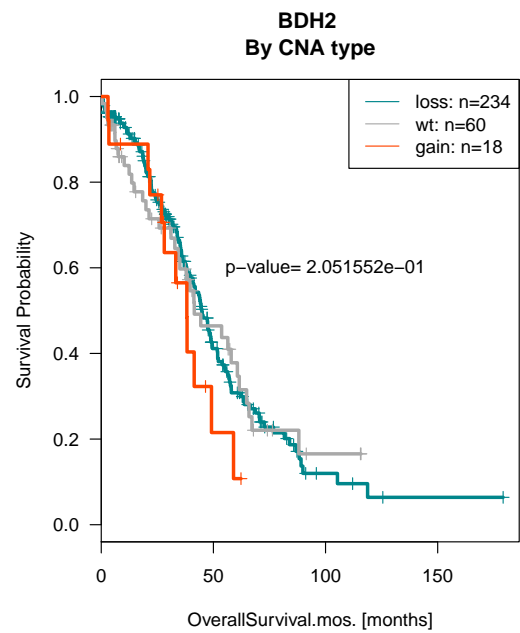
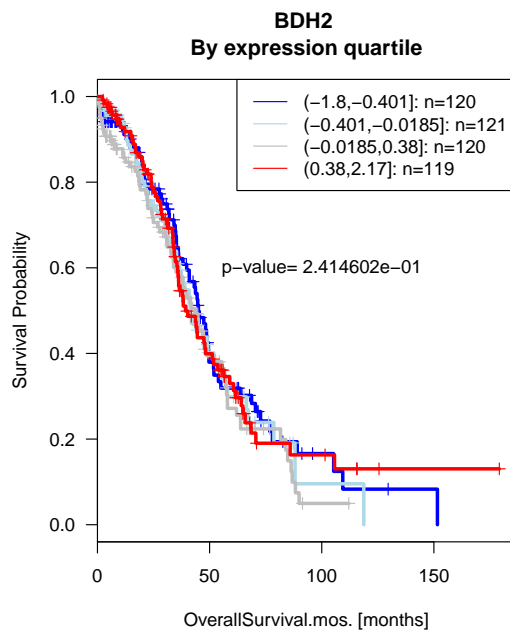
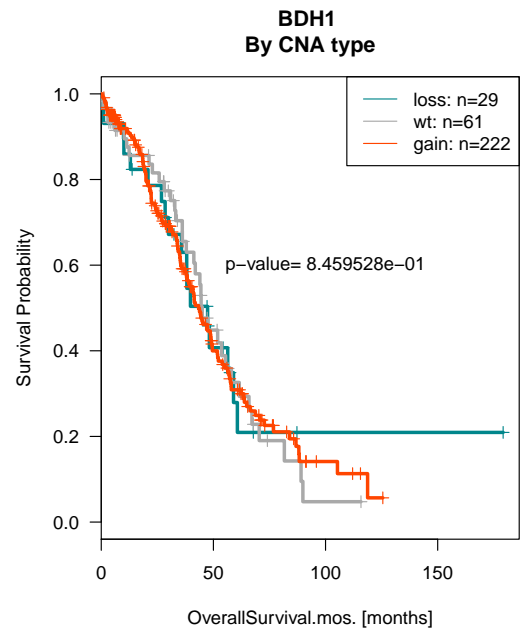
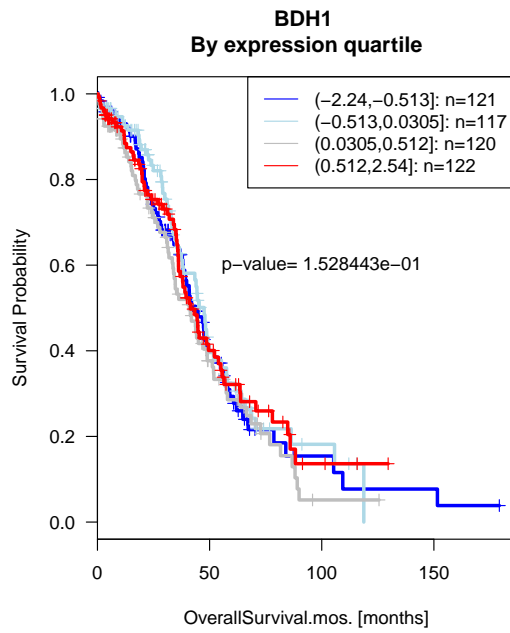


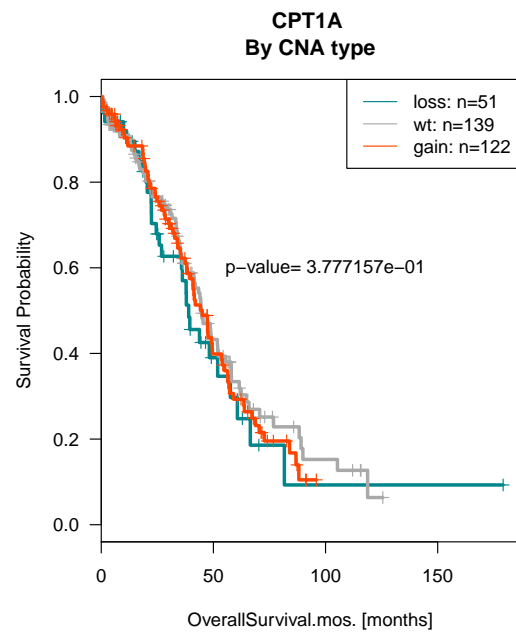
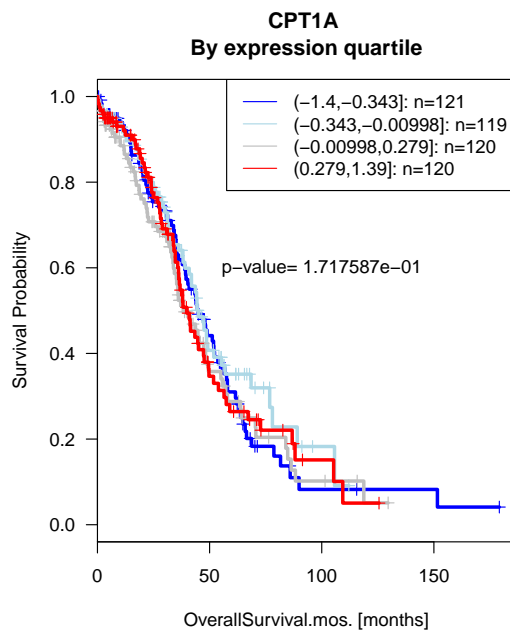
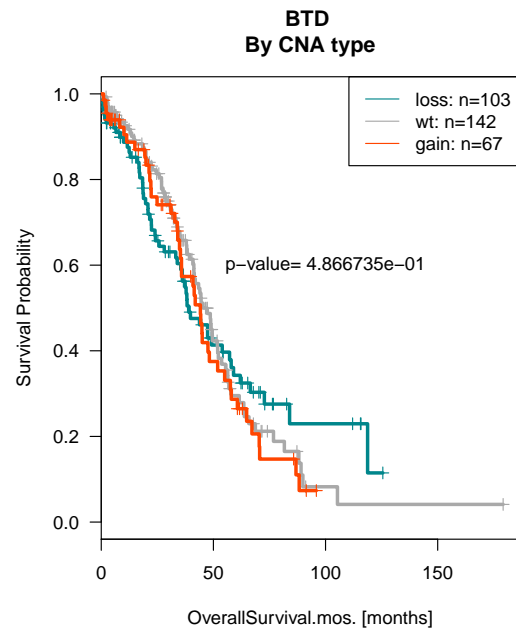
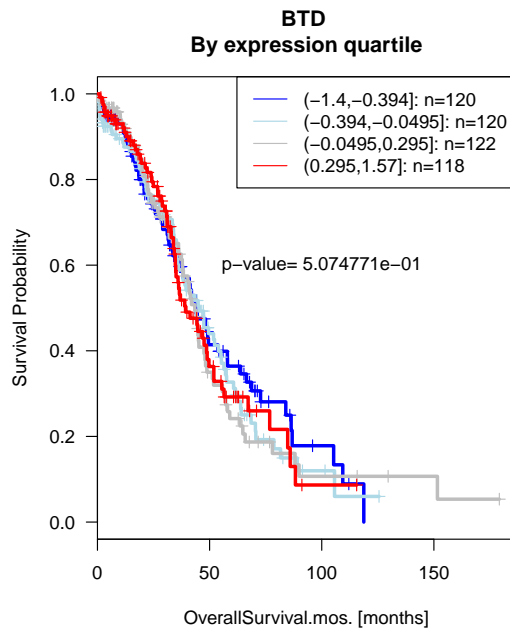


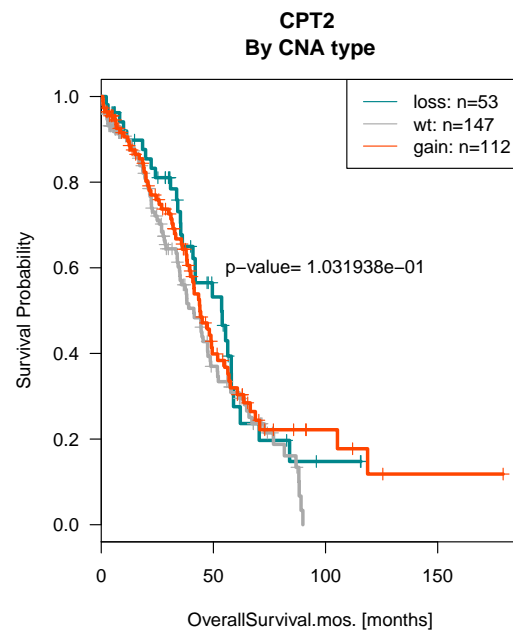
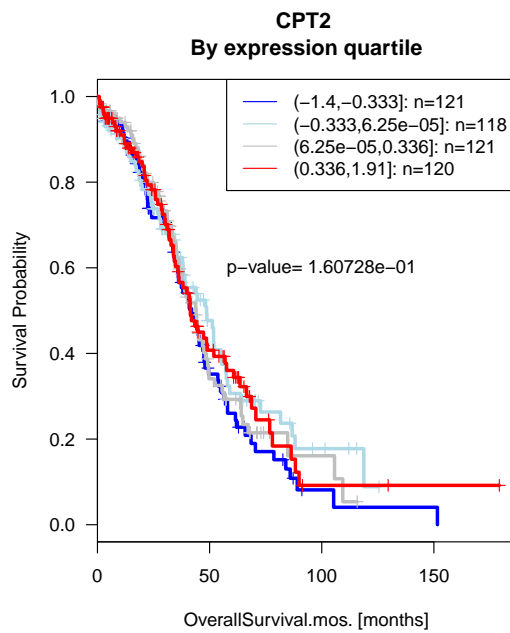
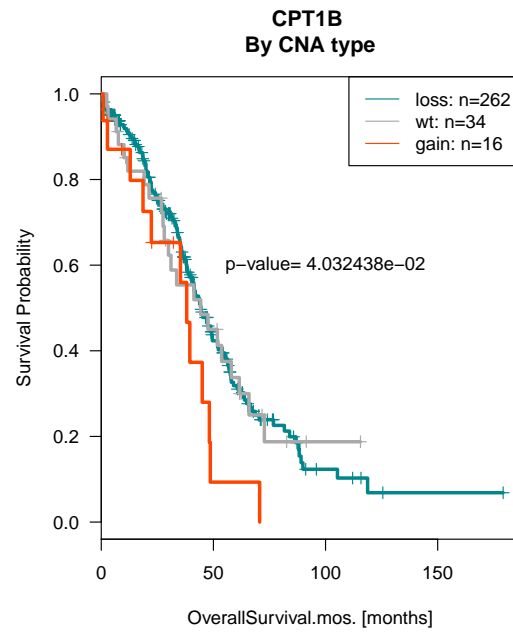
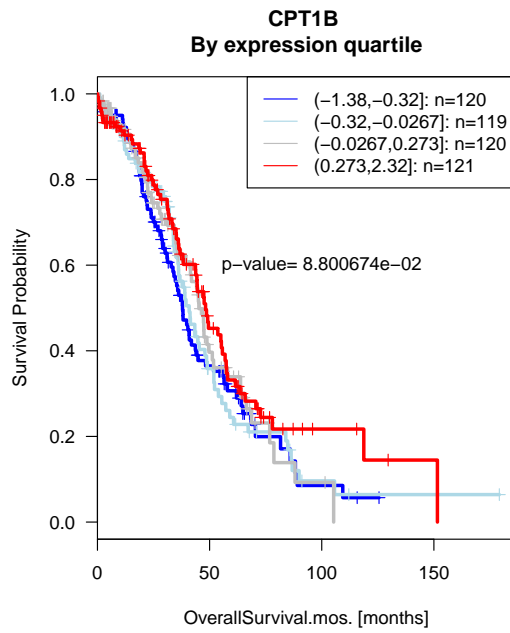


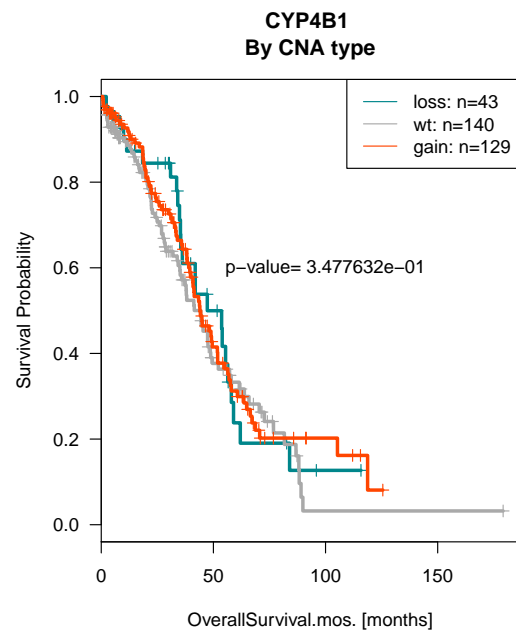
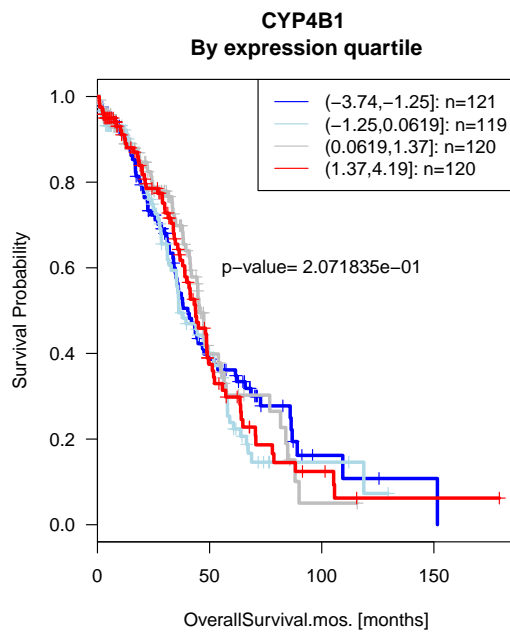
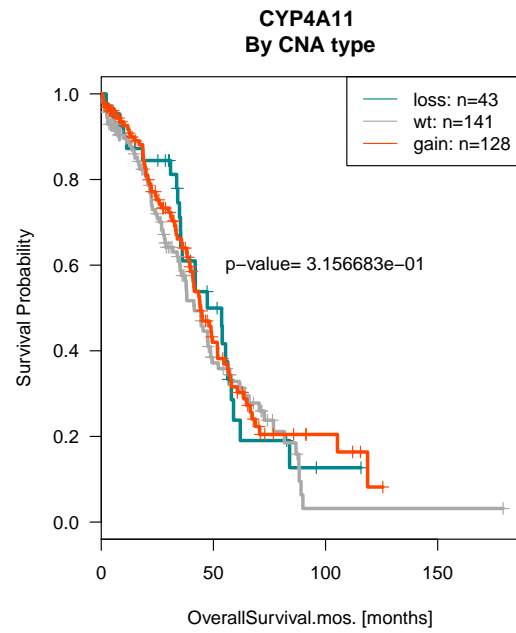
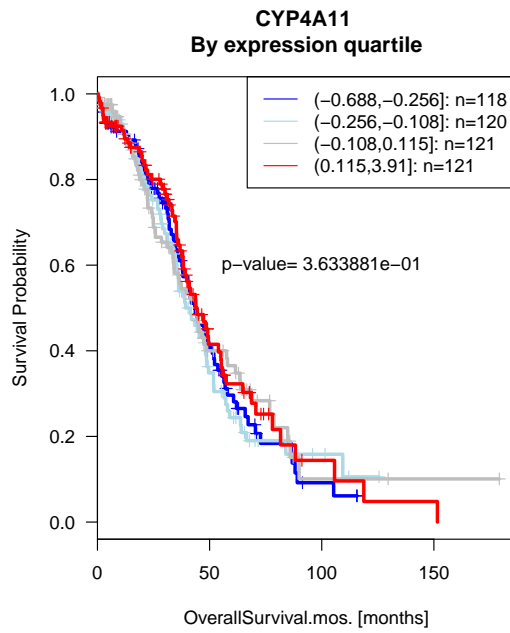


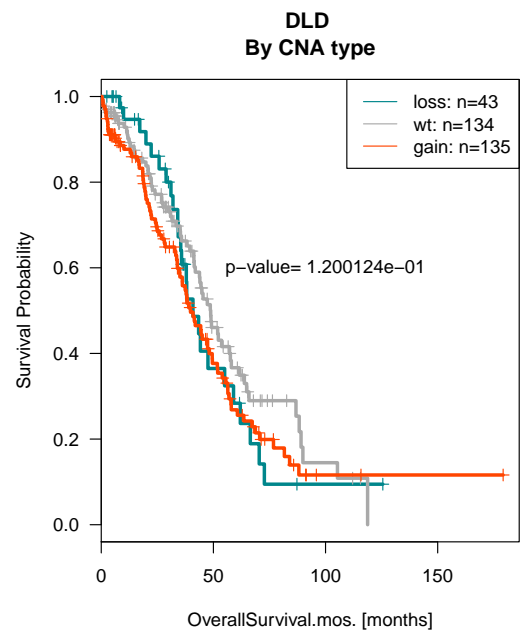
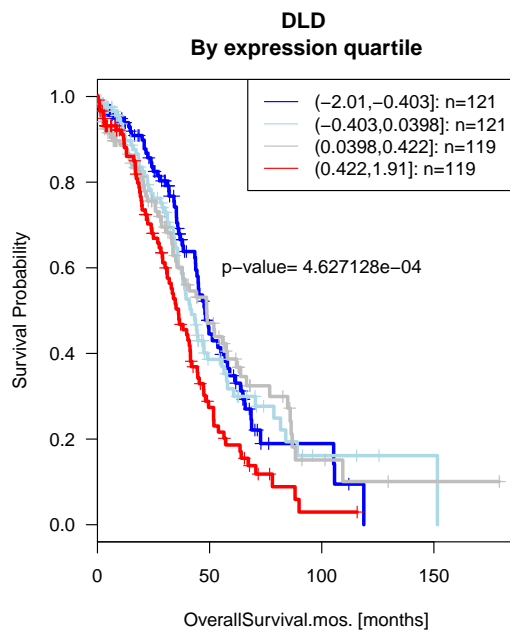
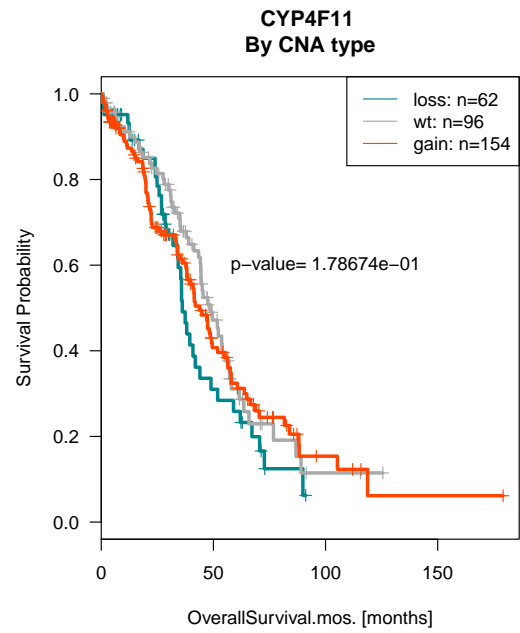
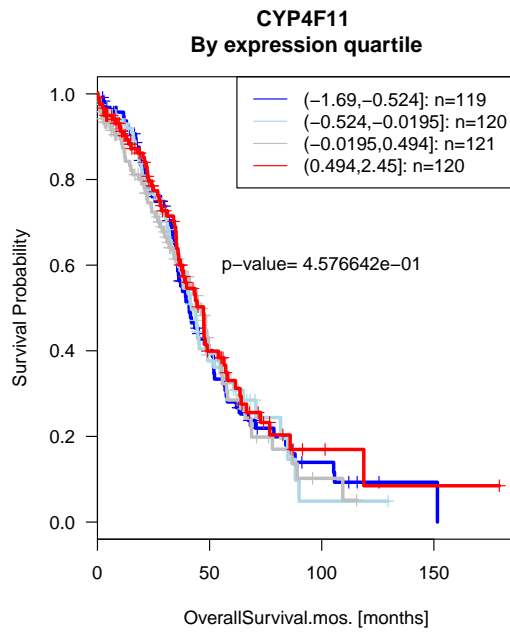


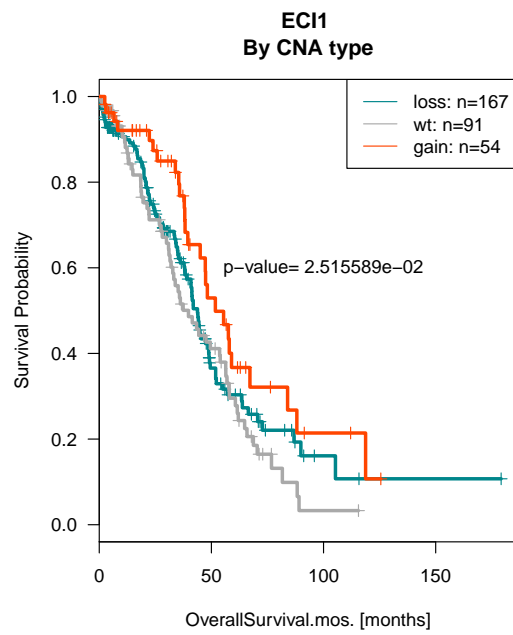
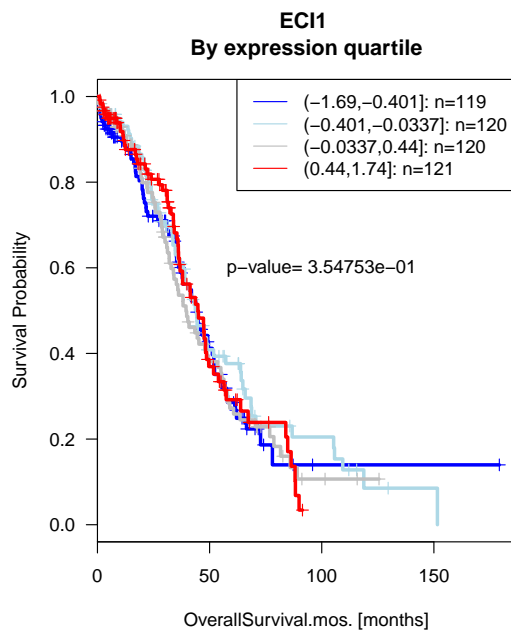
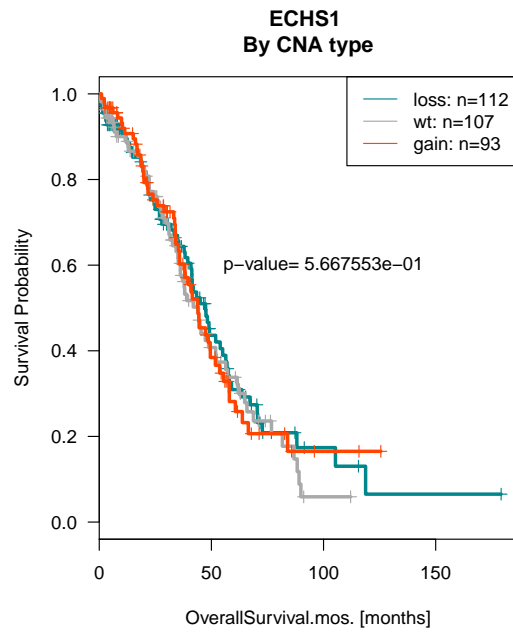
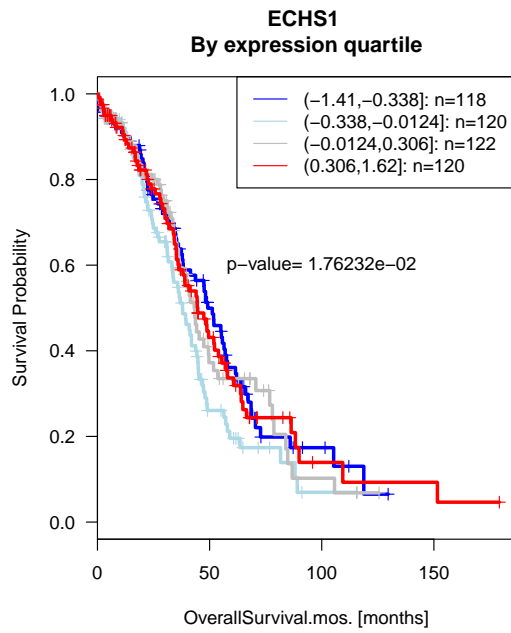


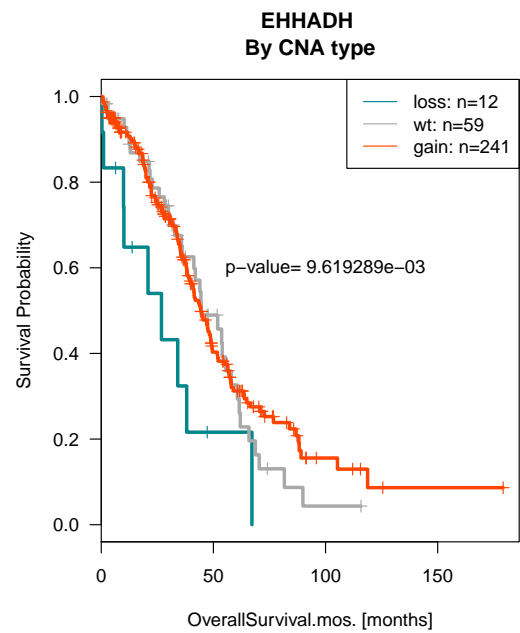
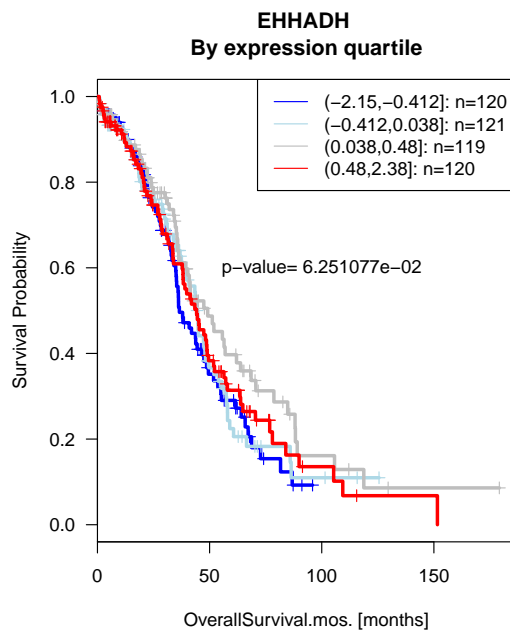
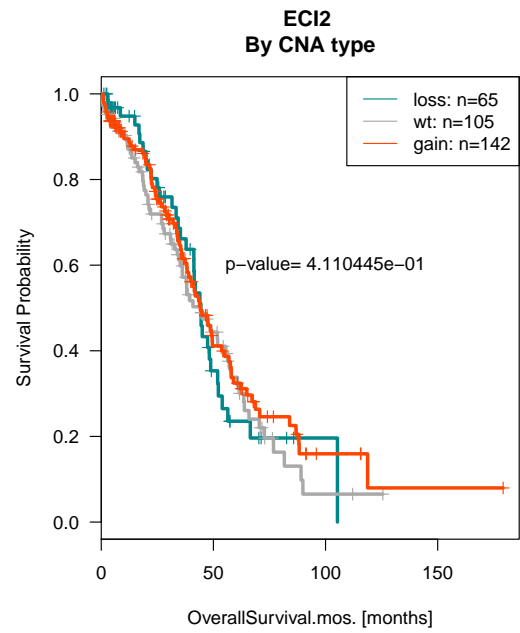
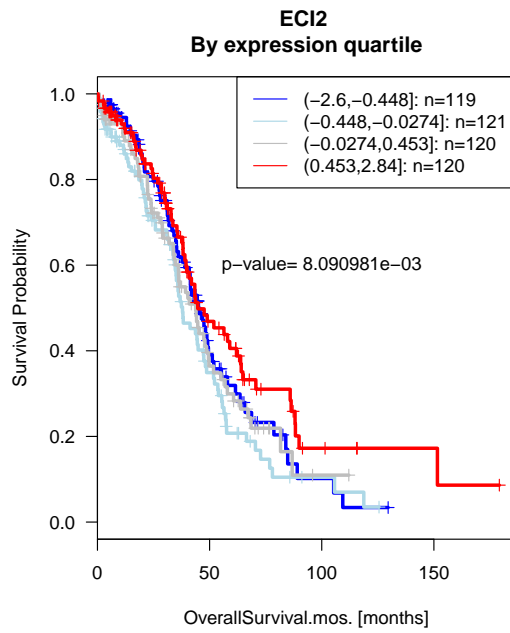


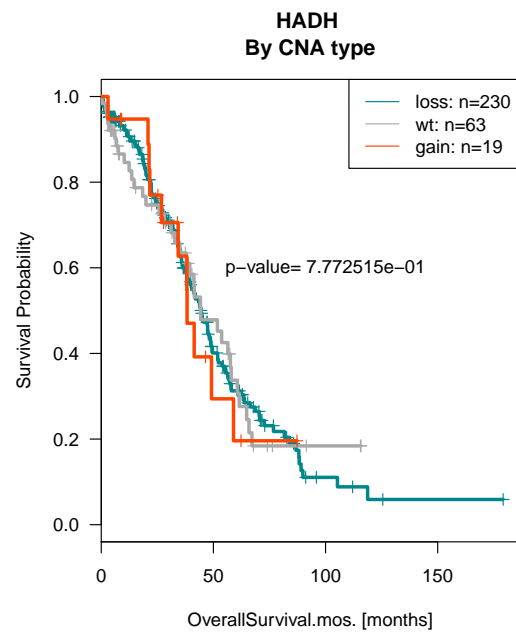
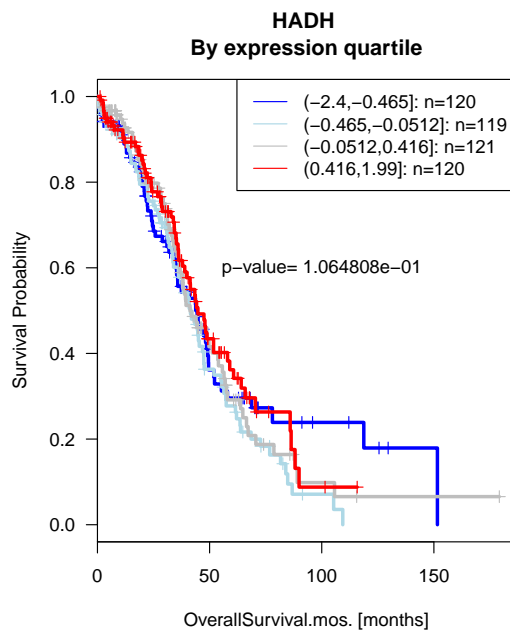
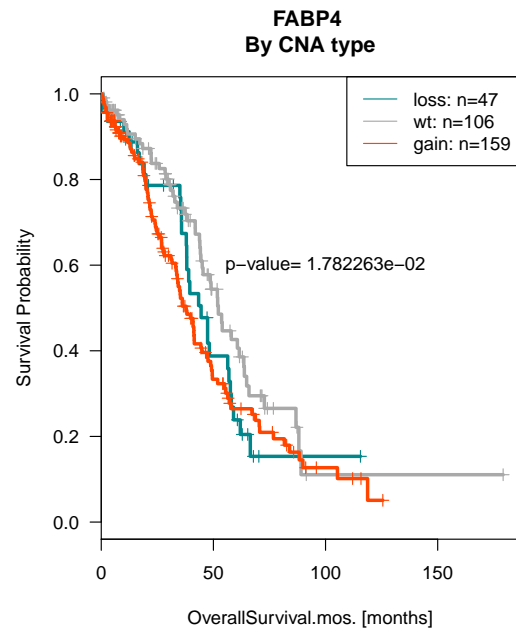
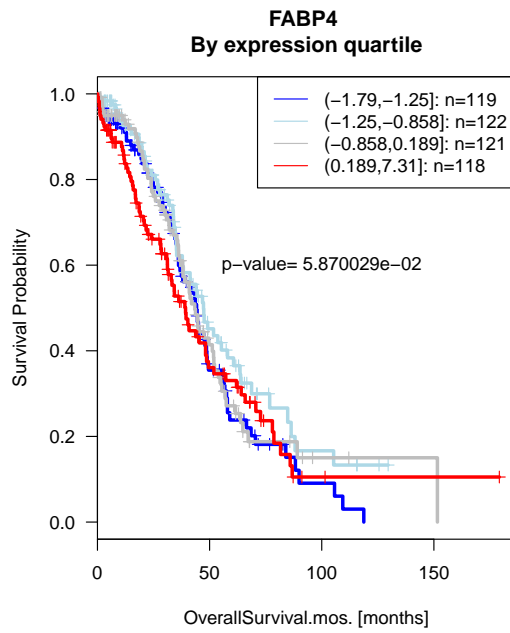


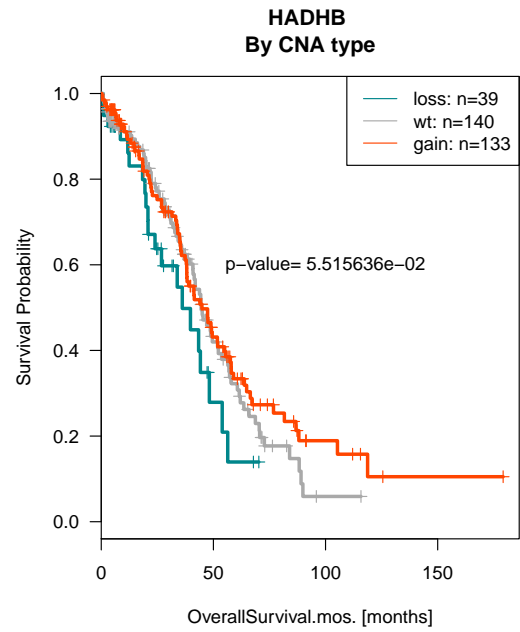
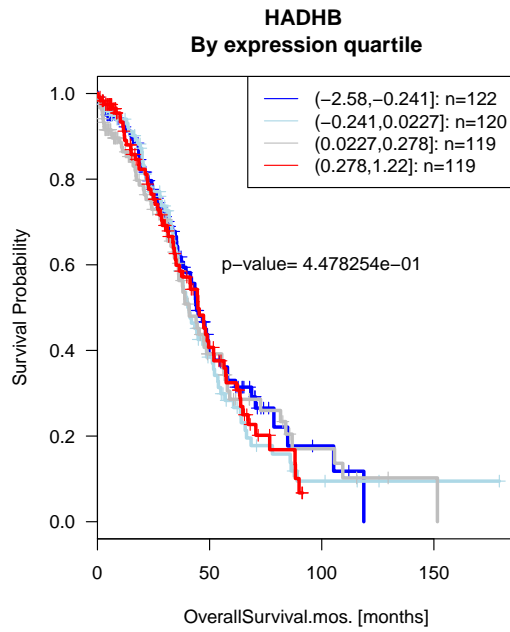
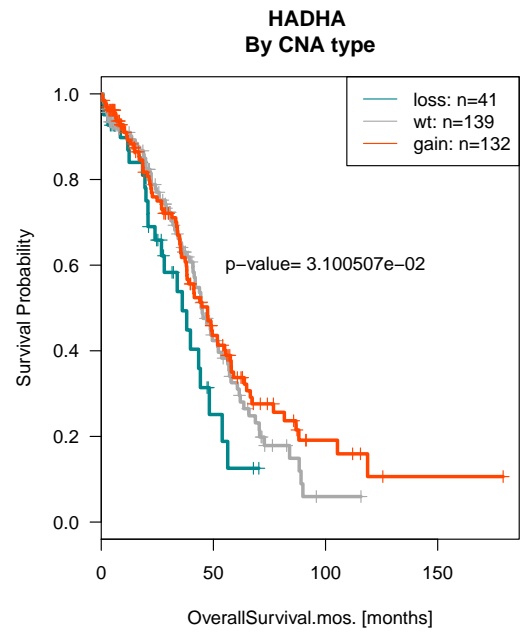
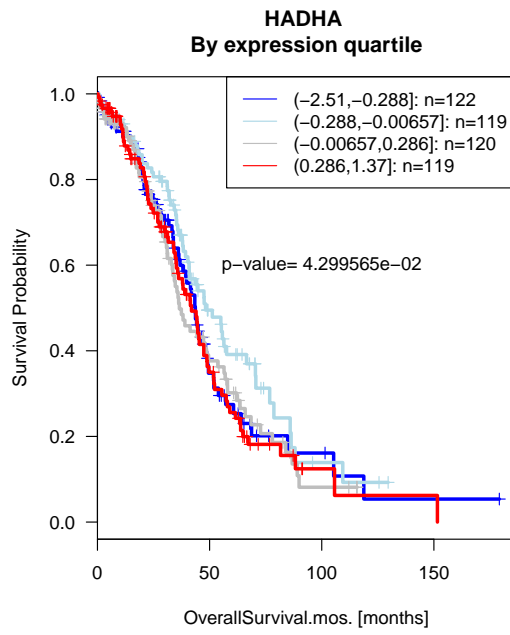


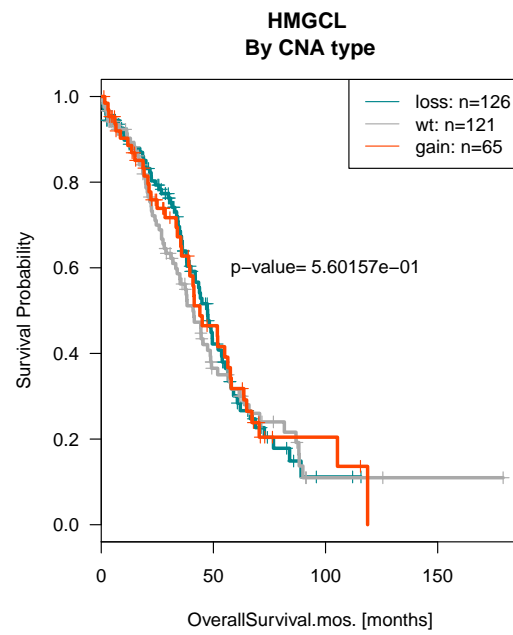
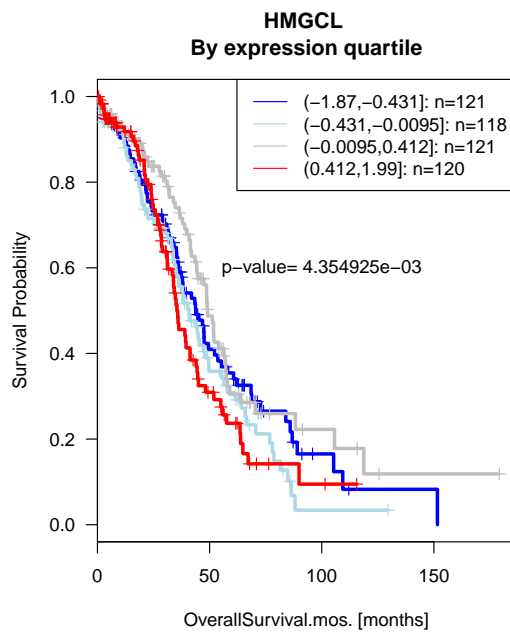
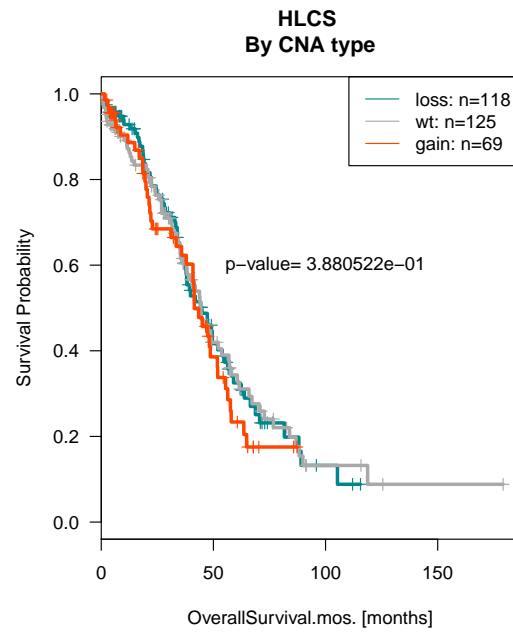
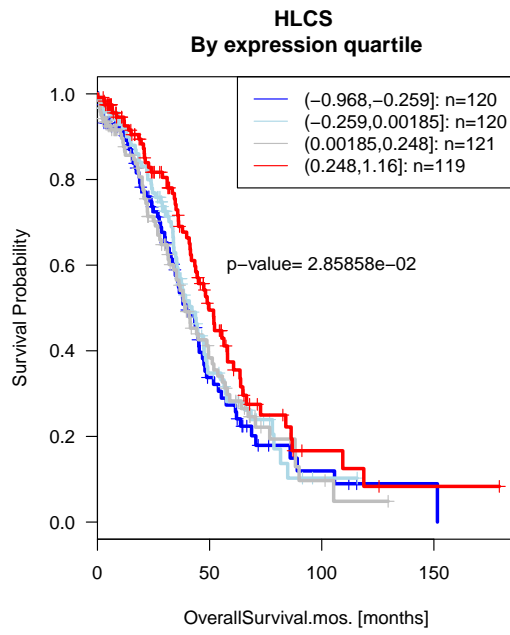


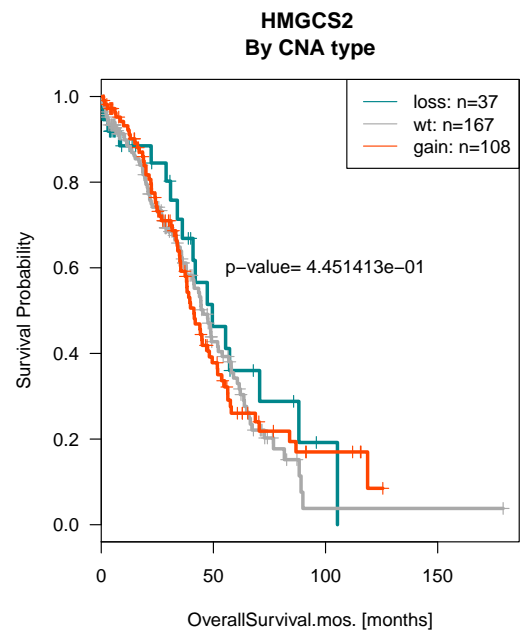
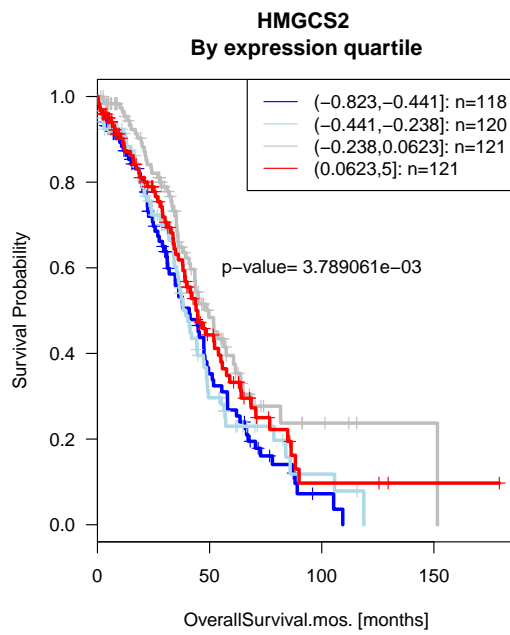
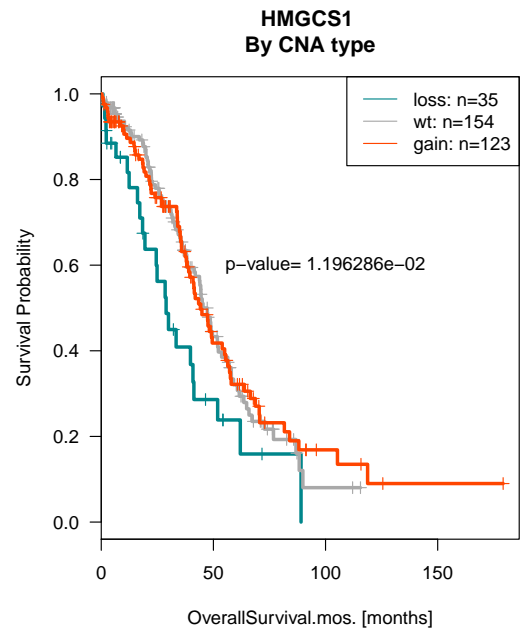
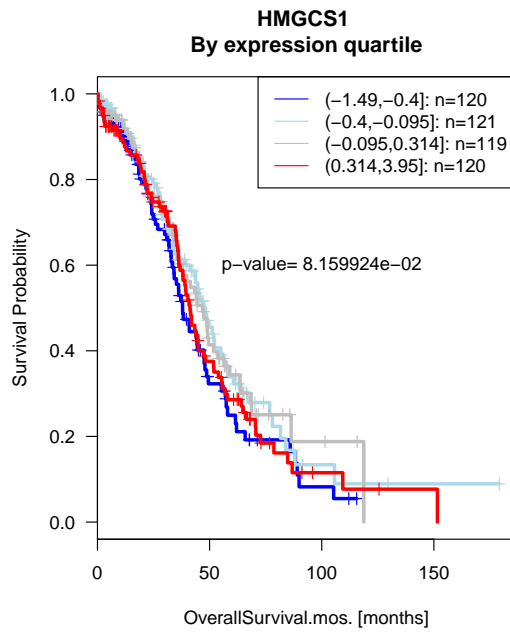


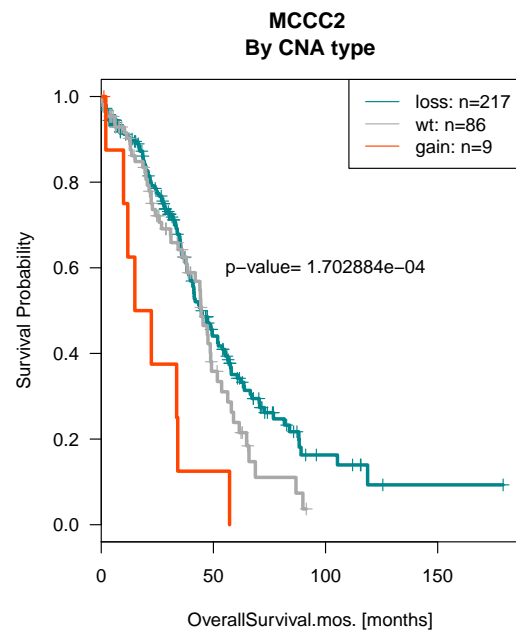
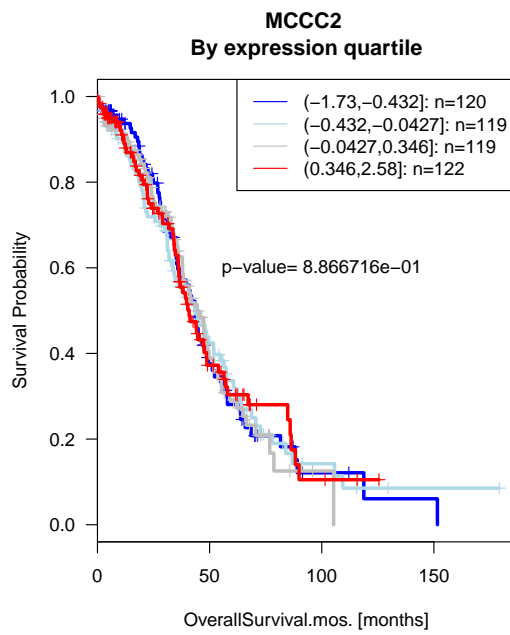
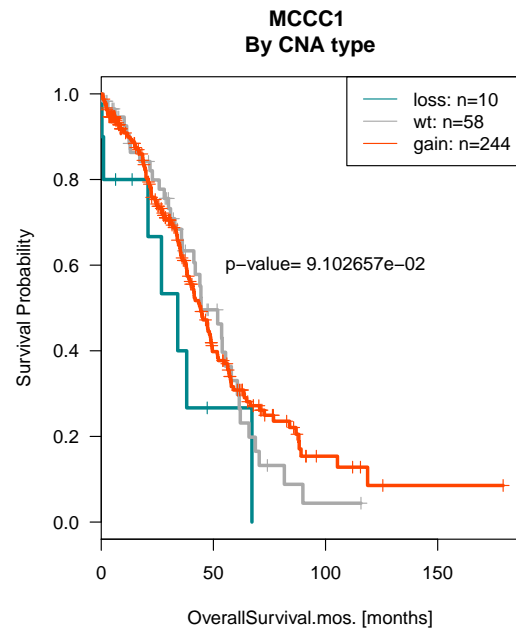
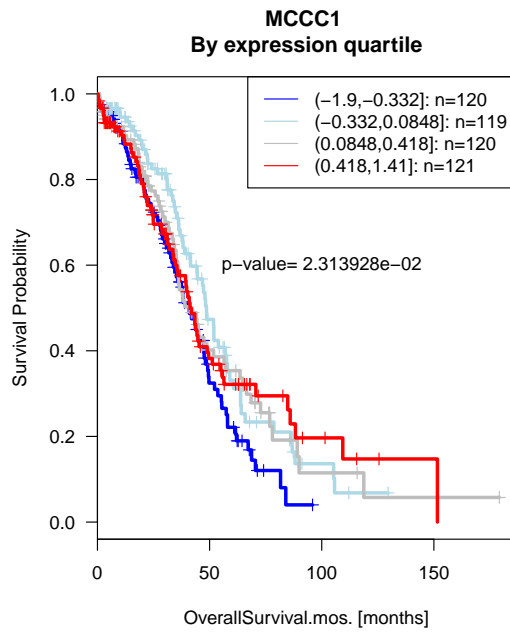


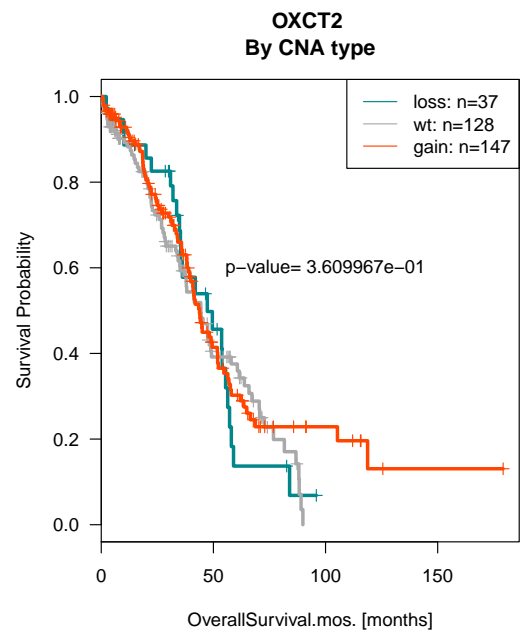
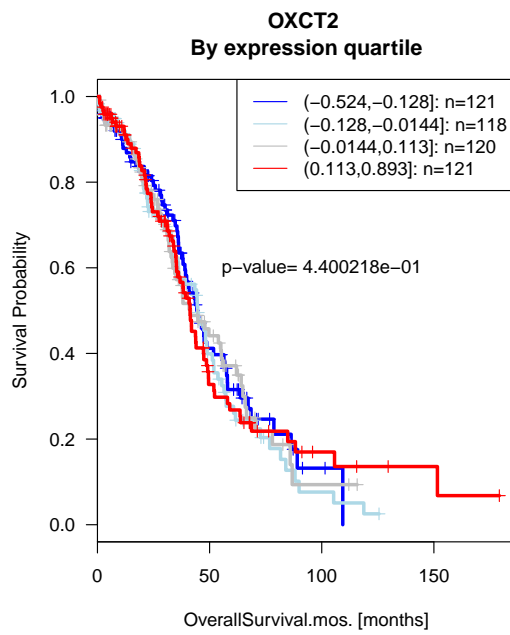
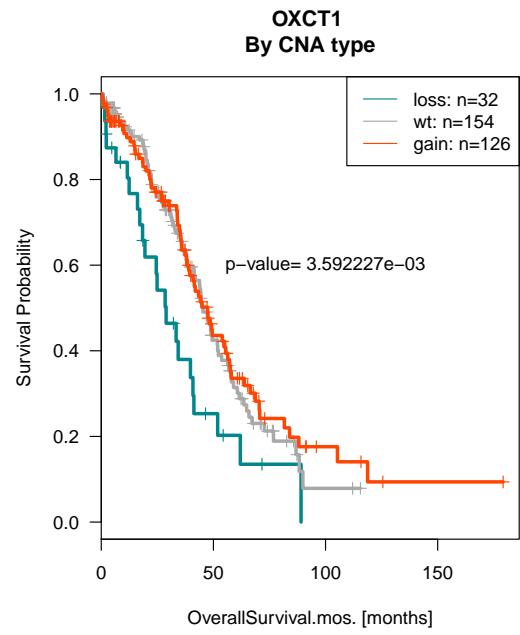
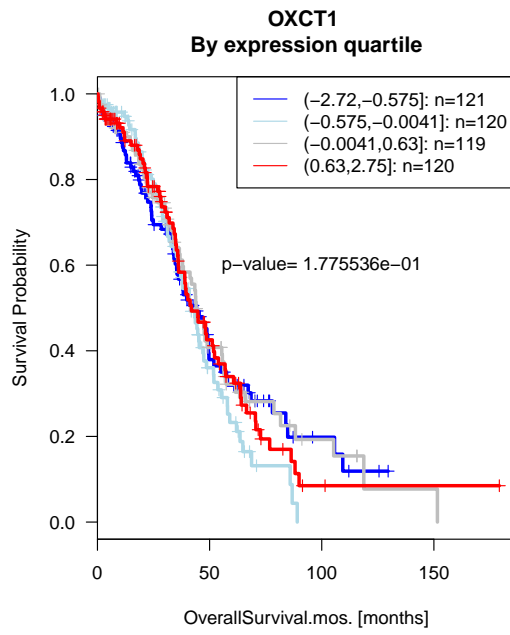


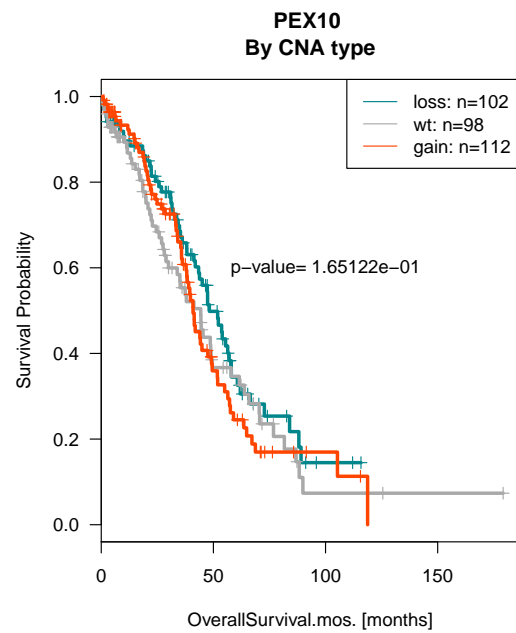
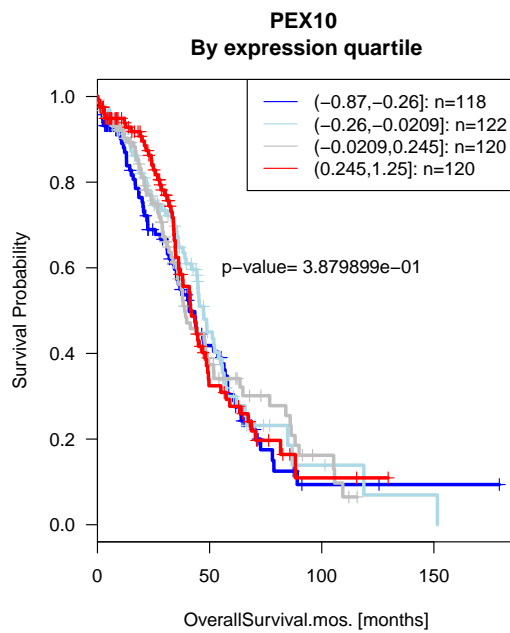
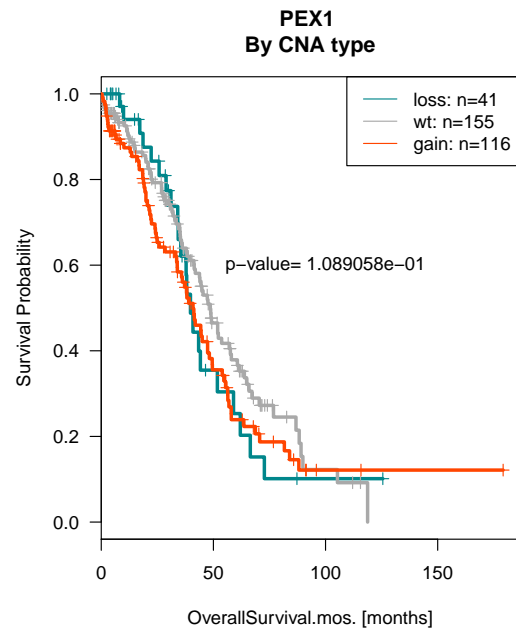
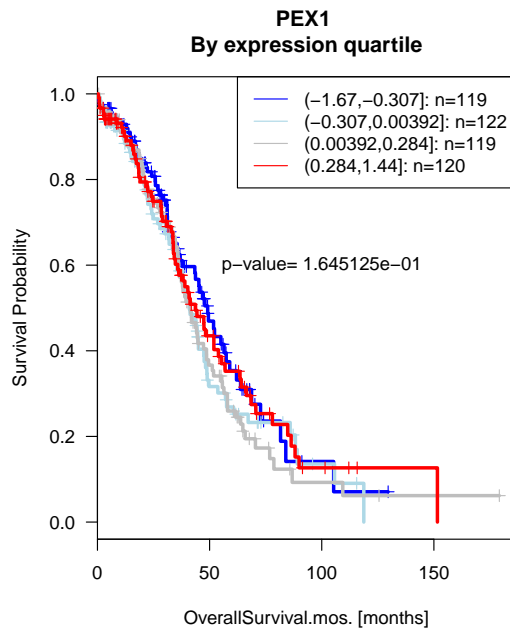


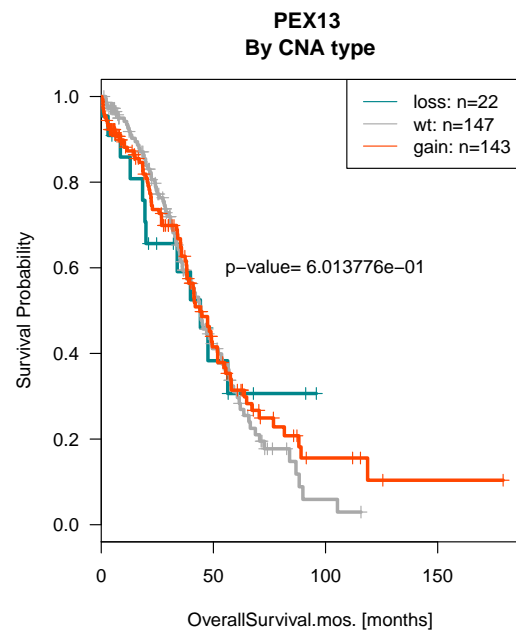
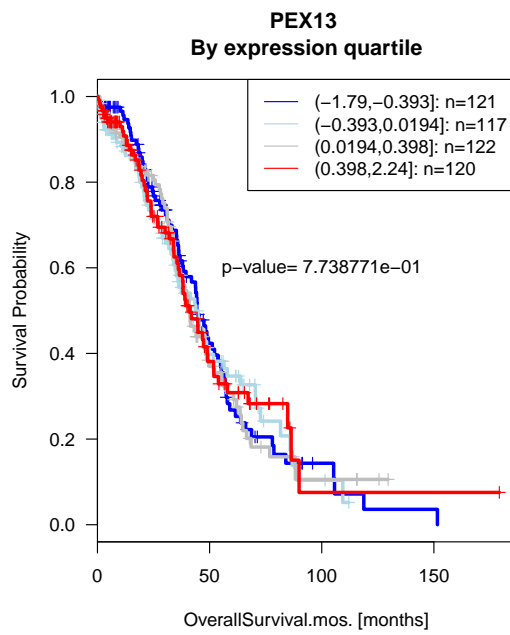
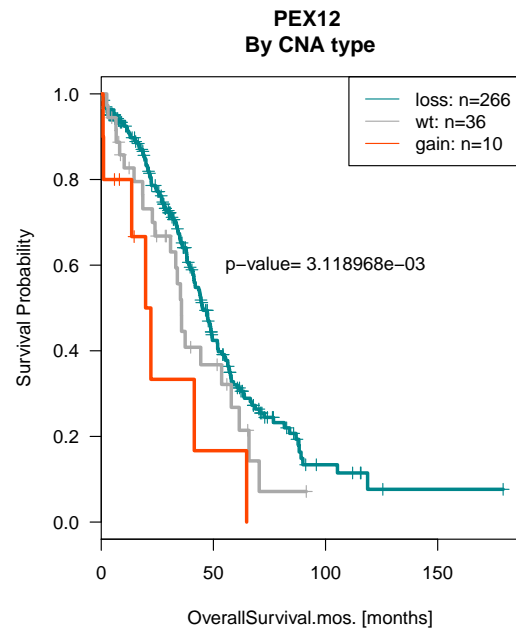
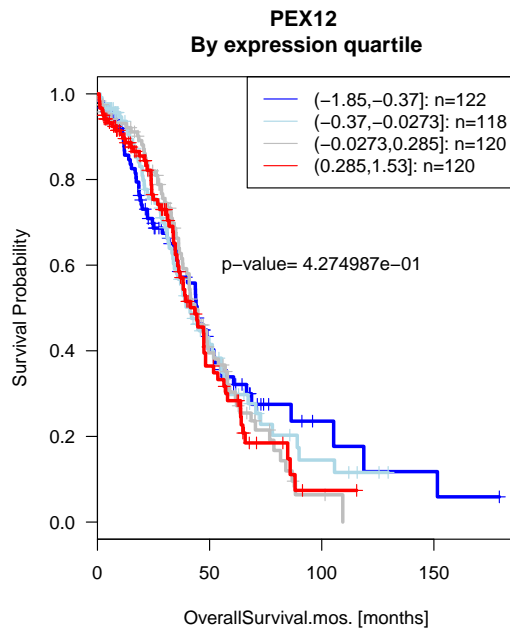


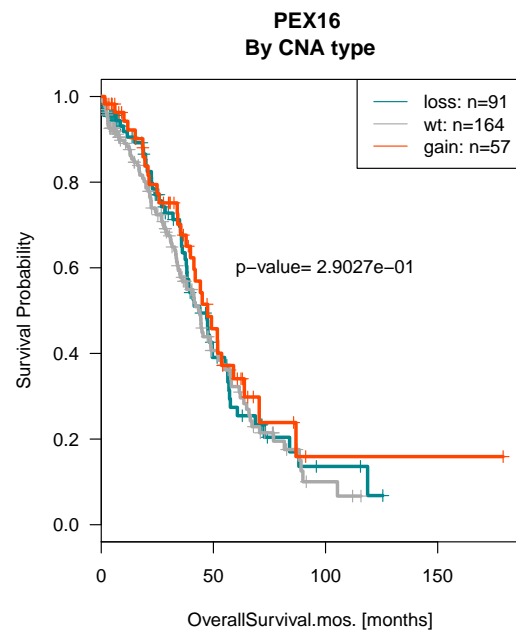
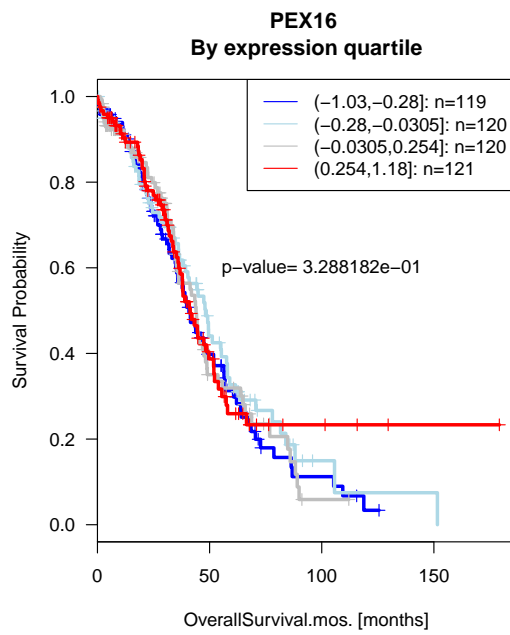
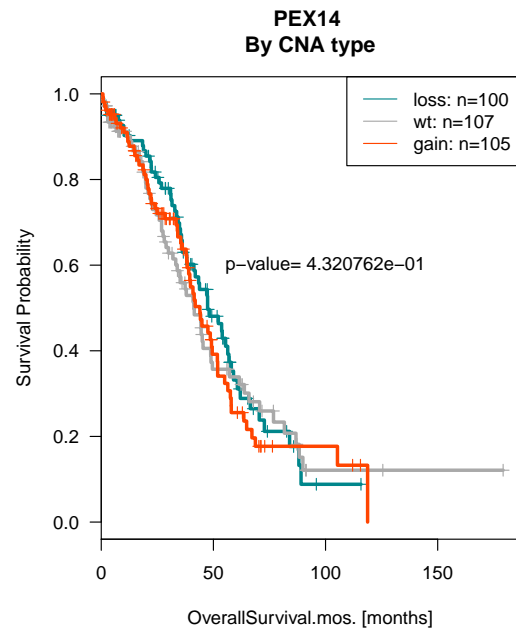
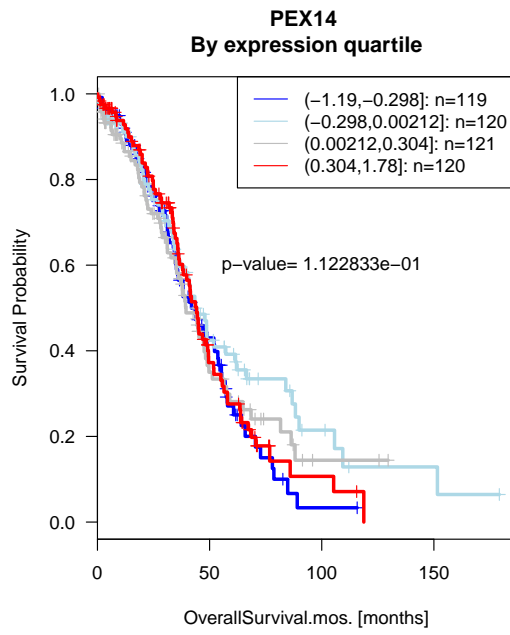


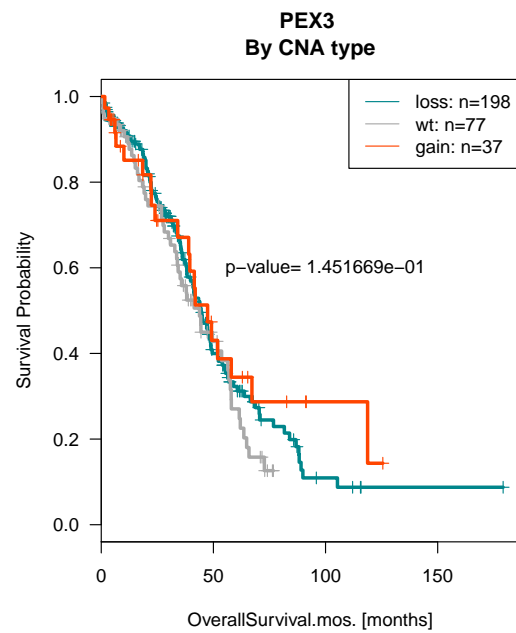
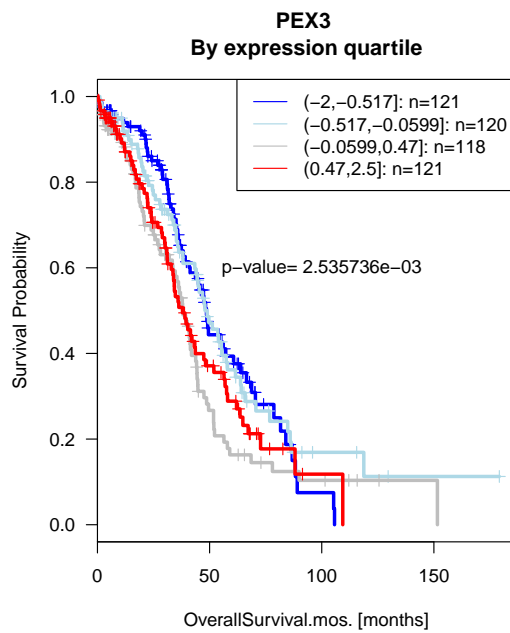
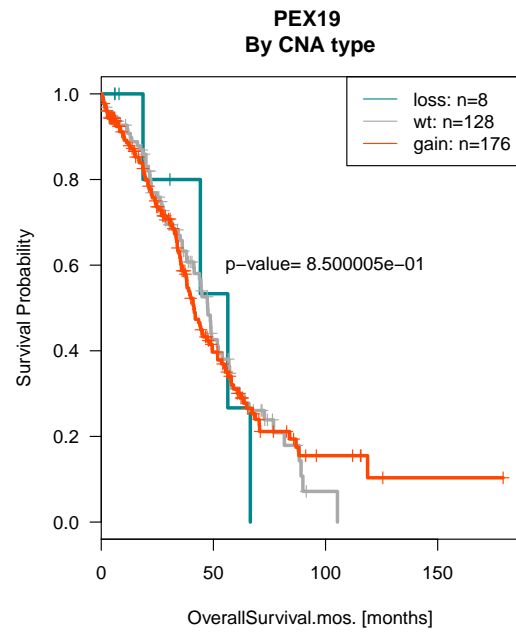
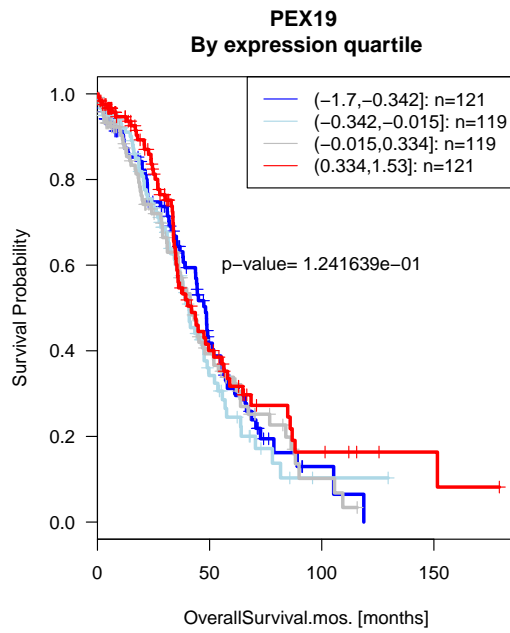


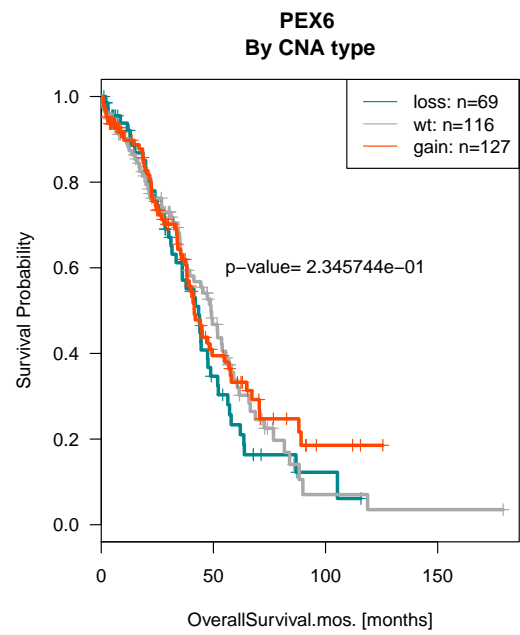
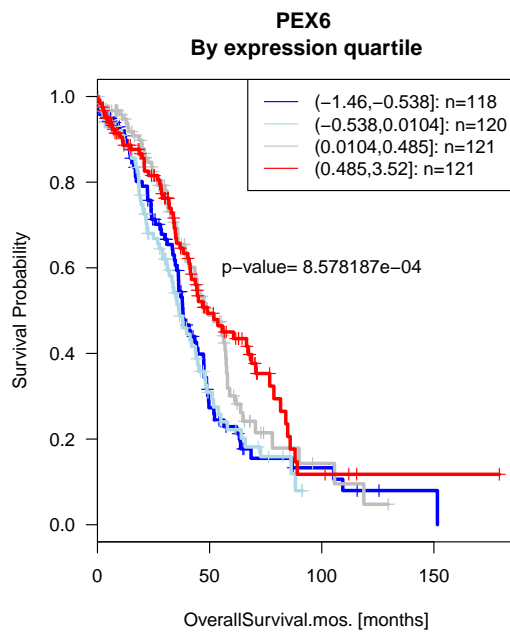
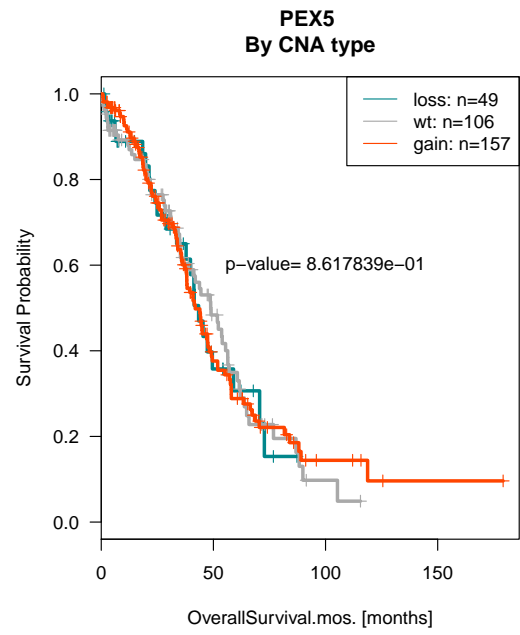
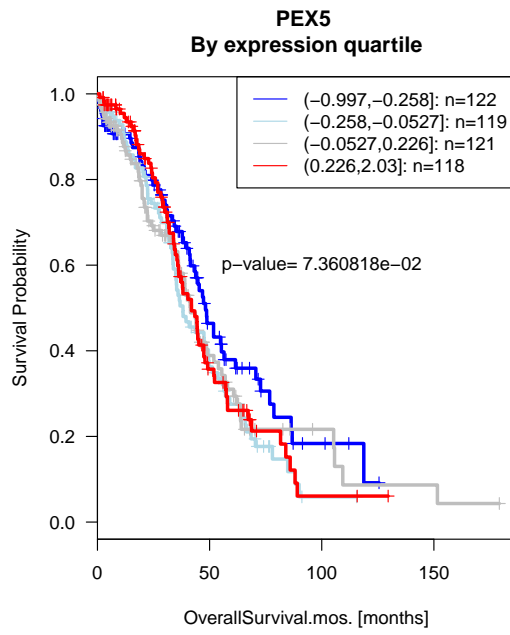


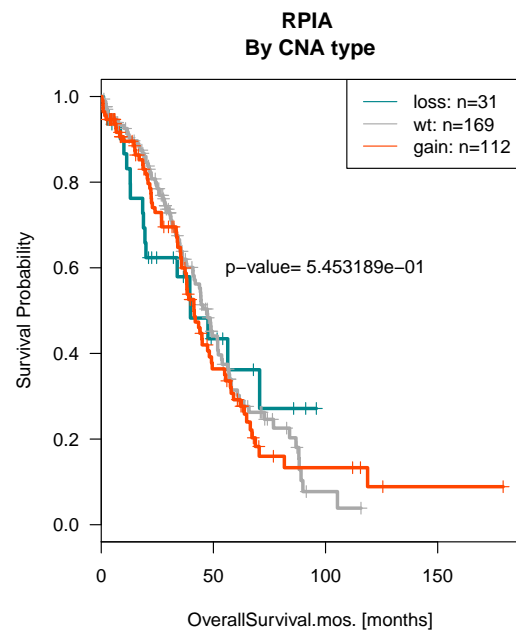
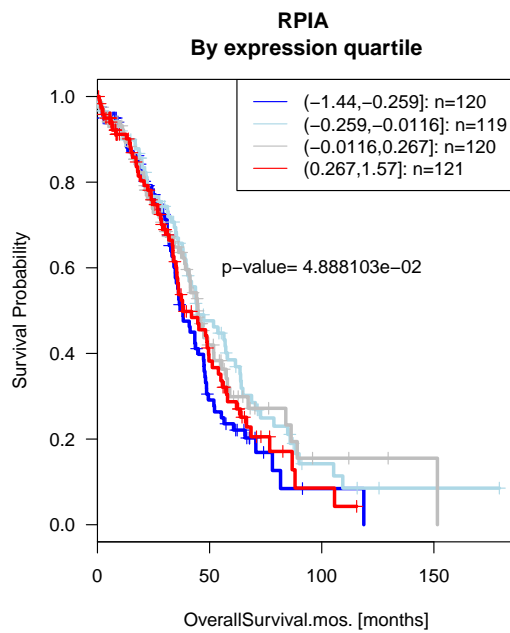
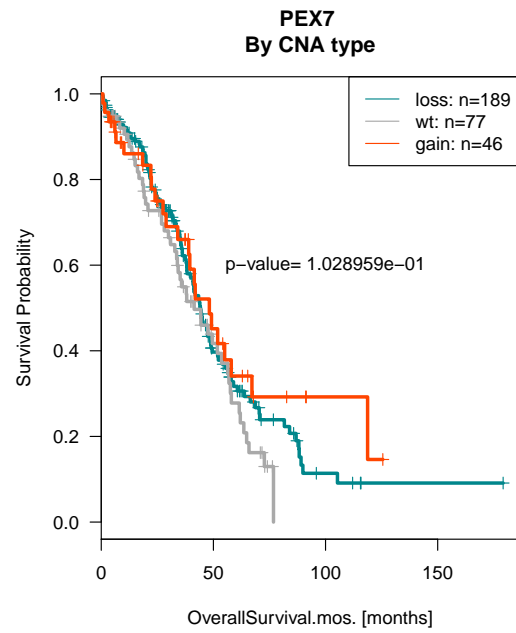
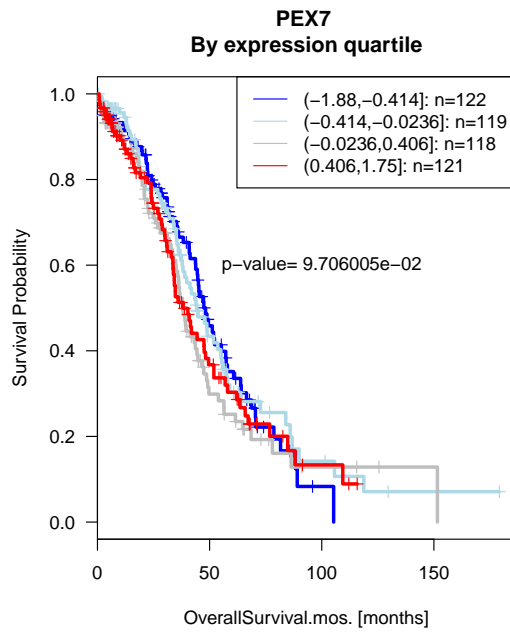


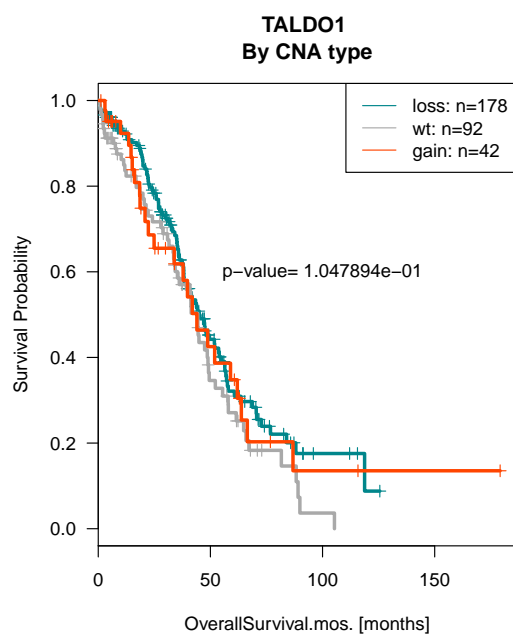
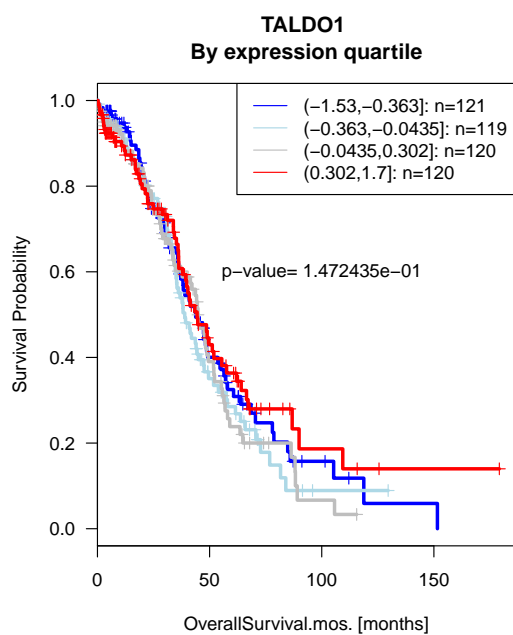
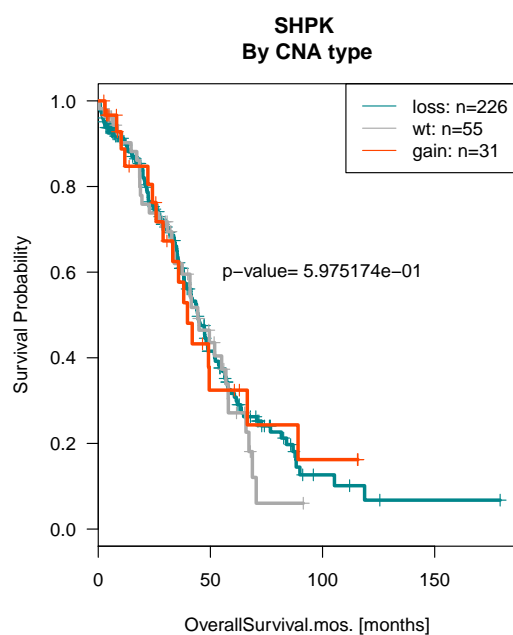
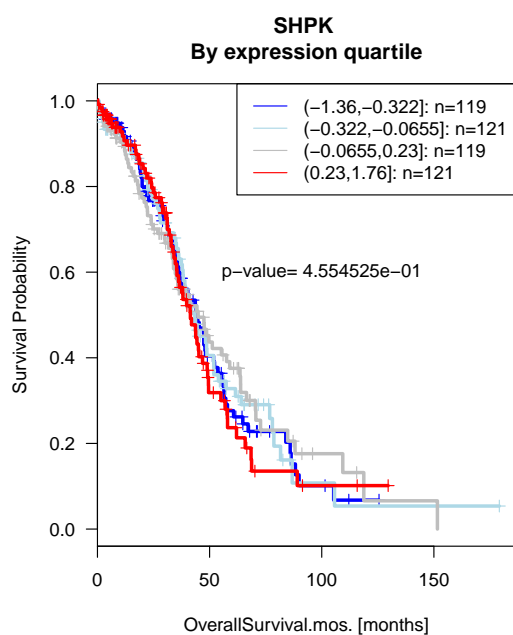


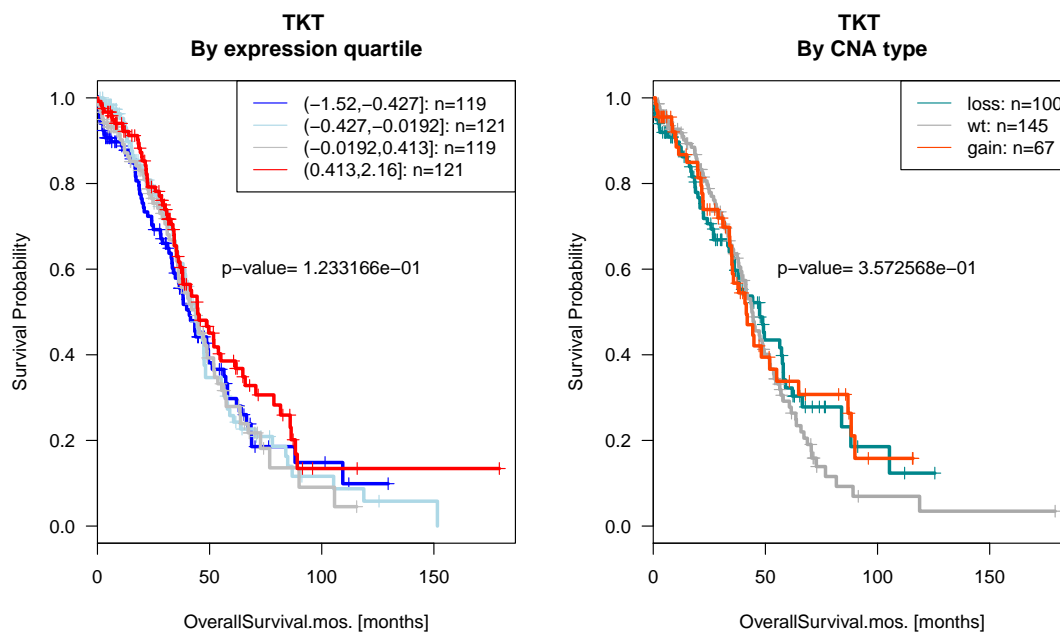












4 Session info: R-packages and their versions used for this analysis

```
sessionInfo()

## R version 3.1.2 (2014-10-31)
## Platform: x86_64-apple-darwin10.8.0 (64-bit)
##
## locale:
## [1] en_GB.UTF-8/en_GB.UTF-8/en_GB.UTF-8/C/en_GB.UTF-8/en_GB.UTF-8
##
## attached base packages:
## [1] stats      graphics  grDevices  utils      datasets  methods   base
##
## other attached packages:
## [1] survival_2.38-1 knitr_1.9
##
## loaded via a namespace (and not attached):
## [1] evaluate_0.6  formatR_1.1  highr_0.4.1  splines_3.1.2 stringr_0.6.2
## [6] tools_3.1.2
```