Imperial College London - Cancer & Surgery

SPIB - A potential biomarker in breast cancer?

Jan-Philipp Cieslik (CID 02139878)

The R code is included in the appendix.

The R-Markdown notebook and the LaTeXsource files are published on GitHub.

`https://github.com/jan-cieslik/mres_assignment`

Word Count: 1277

1 Introduction

Breast cancer is the most common cancer in women. One in every seven women in the UK will be diagnosed with breast cancer in their lifetime[1]. Due to the high incidence, breast cancer is also accountable for the most cancer-associated deaths. To predict patient outcomes and to find new actionable targets for cancer therapy new biomarkers are required. There is already a multitude of established clinical markers like the tumour size, lymphatic node invasion status, distant metastasis status (as in the TNM classification) and molecular characterization (mainly oestrogen, progesterone and HER2 receptor status). High throughput data is becoming more readily available, allowing for in silico analysis of multiomics (e.g., genomics, proteomics, …) of large cohorts. This essay tries to demonstrate the utility of SPIB as a possible breast cancer biomarker. The transcription factor Spi-B is encoded by the gene SPIB and is a member of the Erythroblast Transformation Specific (ETS) group, which is defined by a common highly conserved DNA-binding domain. In the literature SPIB is described as both a tumour suppressor and an oncogenic protein. Studies in lung cancer cells found SPIB to be involved in recruitment of tumour associate macrophages (TAM) [4], further SPIB was found to promotes anoikis resistance [8]. In colorectal cancer cells, on the other hand, SPIB displayed tumour suppressing characteristics by activating NF-κB and JNK signalling pathways [9].
(Introduction word count: 229)

2 Methods

2.1 Data Sources

The Cancer Genome Atlas (TCGA) dataset for breast cancer (BRCA)[2] was acquired from the Xena platform [3].

2.2 Data Pre-Processing and Normalization

Some data is downloaded after pre-processing steps have already been performed. I will outline all important data pre-processing steps to give a complete representation of the resulting data.

2.2.1 RNA-Seq

Gene expression profiles from the Illumina HiSeq 2000 RNA sequencing platform were aligned and annotated using reference transcripts based on the hg19 reference genome. Transcript abundance is calculated with RSEM (RNA-Seq by Expectation Maximization) and transformed to log2(x+1) normalized values.

2.2.2 Methylation

Methylation data from the Illumina Infinium HumanMethylation450 BeadChip (Methylation450k) was transformed into β values ranging from 0 to 1. A higher β value indicates a higher level of DNA methylation. The probes were aligned to the hg19 reference genome. Methlyation probes are then annotated with their corresponding gene symbol.

2.2.3 Copy Number

Copy number data was obtained via a whole genome microarray. The raw data was processing using the GISTIC2 (Genomic Identification of Significant Targets in Cancer) pipeline. Finally, the resulting values were grouped by thresholds and transformed into one of five levels from deep deletion (-2) to high-level amplification (+2).

2.3 Identification of a Possible Novel Prognostic Marker

A multivariate Cox proportional hazards regression model was created with data from 1218 breast cancer patients. Overall survival was defined as the dependent (outcome) variable and mRNA log values (as determined by RNA sequencing from the primary tumour) together with age and tumour stage (I/II as low; III/IV as high) as the independent variables. The calculated p-values were corrected with the false discovery rate (FDR) method and are shown as q-values.

2.4 Survival Analysis

The patient population was divided into SPIB-high and SPIB-low based on the median of the SPIB mRNA expression. A logrank test and an associated Kaplan-Meier plot were generated for the two subpopulations.

2.5 Methylation and mRNA Expression Correlation

Methylation values were tested for normal distribution using Shapiro-Wilk normality test. The β values fail to show a normal distribution (e.g., cg07979271: $p < 2.2 \cdot 10^{-16}$), consequently a nonparametric test was used for correlation. Methylation β values were correlated against the mRNA log values using spearman. P-values were corrected for multiple testing by using FDR and transformed into q-values. The most anti-correlated methylation site is then shown as a scatter plot against the mRNA level.

2.6 Copy Number and mRNA Expression

A chart with multiple box plots (one per copy number threshold) was generated displaying the corresponding mRNA log values. To test for dependence of mRNA expression on copy number I performed a one-way

independent ANOVA (analysis of variance).

2.7 mRNA Co-Expression

mRNA expression values were correlated with the corresponding expression values of SPIB using spearman. The p-values were adjusted for multiple testing by using FDR and transformed into q-values.
(Methods word count: 470)

3 Results

3.1 Identification of a possible novel prognostic marker

In a multivariate Cox analysis in 1218 breast cancer patients, I could identify 2078 significantly ($q < 0.05$) associated genes (mRNA expression) with the overall survival. SPIB was associated with a beneficial hazard ratio ($HR = 0.91$; $q = 3.27\%$; Fig. 1). Further, patients overexpressing SPIB have an overall median survival of 130 months, while low expressing patients survive for a median of 112 months (logrank: $p = 0.01$; Fig. 1). Five years overall survival was 77% (SPIB-low) and 81.4% (SPIB-high) respectively.

3.2 Methylation and mRNA Expression

Next, I focussed on the regulation of SPIB through methylation. From the selected patient cohort, 873 samples had mRNA sequencing and methylation data available. A total of 17 CpG islands associated with SPIB were analysed (Table 1). In the spearman correlation site cg07979271 displayed the most significant negative association between methylation and SPIB mRNA expression levels ($cor = 0.55; q = 9.43 \cdot 10^{-59}$; Fig. 2).

3.3 Copy Number Alteration and mRNA Expression

Another way mRNA expression could be modified is through copy number alteration. A subset of 1078 samples included mRNA sequencing and copy number data. As seen in Fig. 3, no significant difference between the copy number levels could be shown (one-way independent ANOVA; $p > 0.05$).

3.4 Co-Expression Analysis

As SPIB is a known transcription factor, the co-expressed genes are of major interest. The mRNA data of 1218 samples was available and the most significant correlations are shown in Table 2. Many co-expressed genes are involved in immunological pathways, MS4A1 displayed the strongest positive correlation and encodes the CD20 antigen found on B-lymphocytes ($cor = 0.82; q = 2.98 \cdot 10^{-295}$). Further genes include TCL1A,CXCR5,CCL4 (found on T-cells), LCK (found on lymphocytes in general) or MAP4K1 (a JNK-activator). Negative correlations

were also investigated (Table 2) with DCTN4 displaying the strongest one ($cor = -0.42; q = 4.77 \cdot 10^{-51}$).
(Results word count: 294)

4 Conclusion

SPIB is an interesting gene that is currently under explored in breast cancer. It could already be demonstrated to be a viable candidate gene in lung [4, 8] and colorectal [9] cancer. This essay shows a possible positive association of SPIB with survival in breast cancer. Overexpression of SPIB increases the median overall survival in breast cancer by 18 months (112 vs 130 months). The initial regulation analysis found cg07979271 to be the highest anti-correlated CpG site and no association between copy number alterations. Indicating a strong transcriptional regulation of SPIB as an increase in copy number does not increase its expression. As already found in different cancer entities [4, 8] SPIB mRNA expression has a positive correlation with immunologic pathways on B- and T-cells (e.g., CCL4). In lung cancer this seems to induce an increased recruitment of tumour-associated macrophages leading to a shortened overall survival [4]. The data in colorectal cancer, on the other hand, suggest SPIB to act as a tumour suppressor by activating NFκB and JNK signalling [8]. Both effects could be demonstrated in this analysis during correlation of mRNA expression of the selected genes and SPIB mRNA expression data. As this is purely a correlation and SPIB is a known transcription factor, further studies are required to understand the involvement of SPIB in breast cancer. One limitation of this work is the non-directionality of the correlations, as it is not clear if SPIB is the protein affecting the others or if the opposite is true. Additionally SPIB expression shows a negative correlation with known favourable genes in cancer like DCTN4 [5, 6] and FOXA1 [7]. The mechanism behind this counter-intuitive regulation requires further in vitro studies.
(Conclusion word count: 284)

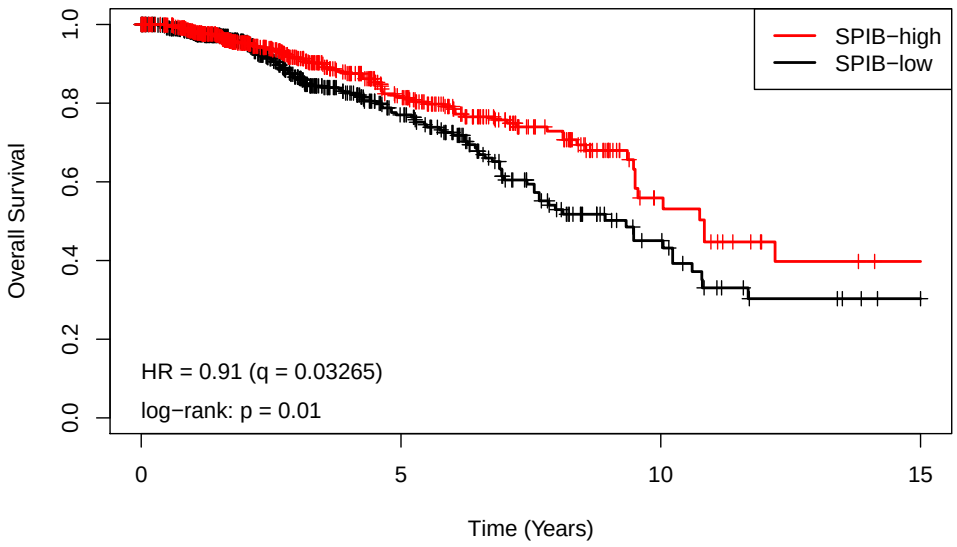


Figure 1: Survival of SPIB-high and SPIB-low breast cancer patients. Data from the TCGA BRCA dataset [2, 3].

Table 1: Methylation Sites

CpG	cor	p-value	q-value	Mean (SPIB-High)	Mean (SPIB-Low)
cg07979271	-0.514	$5.55 \cdot 10^{-60}$	$9.43 \cdot 10^{-59}$	0.816	0.864
cg13403724	0.288	$3.80 \cdot 10^{-18}$	$3.23 \cdot 10^{-17}$	0.248	0.206
cg17774764	-0.281	$3.07 \cdot 10^{-17}$	$1.74 \cdot 10^{-16}$	0.737	0.775
cg03763616	-0.277	$7.68 \cdot 10^{-17}$	$3.26 \cdot 10^{-16}$	0.832	0.859
cg19387862	0.272	$2.70 \cdot 10^{-16}$	$9.20 \cdot 10^{-16}$	0.191	0.168
cg15690347	0.266	$1.25 \cdot 10^{-15}$	$3.54 \cdot 10^{-15}$	0.383	0.301
cg08201854	0.265	$1.87 \cdot 10^{-15}$	$4.55 \cdot 10^{-15}$	0.26	0.226
cg15007959	0.263	$3.08 \cdot 10^{-15}$	$6.55 \cdot 10^{-15}$	0.253	0.203
cg18254819	0.246	$1.53 \cdot 10^{-13}$	$2.90 \cdot 10^{-13}$	0.235	0.212
cg24092179	0.228	$1.00 \cdot 10^{-11}$	$1.71 \cdot 10^{-11}$	0.271	0.244
cg22268231	-0.147	$1.22 \cdot 10^{-05}$	$1.88 \cdot 10^{-05}$	0.45	0.487
cg06512885	-0.133	$7.89 \cdot 10^{-05}$	0.000112	0.875	0.881
cg26522743	-0.0968	0.0042	0.00549	0.394	0.424
cg21152077	0.0921	0.00653	0.00792	0.464	0.448
cg13918544	0.0845	0.0125	0.0141	0.132	0.143
cg22745102	0.0705	0.0372	0.0396	0.469	0.462
cg04508467	$2.02 \cdot 10^{-05}$	1	1	0.696	0.686

Spearman correlation of methylation of CpG islands and the mRNA expression of SPIB. (”cor” is the correlation coefficient ρ)

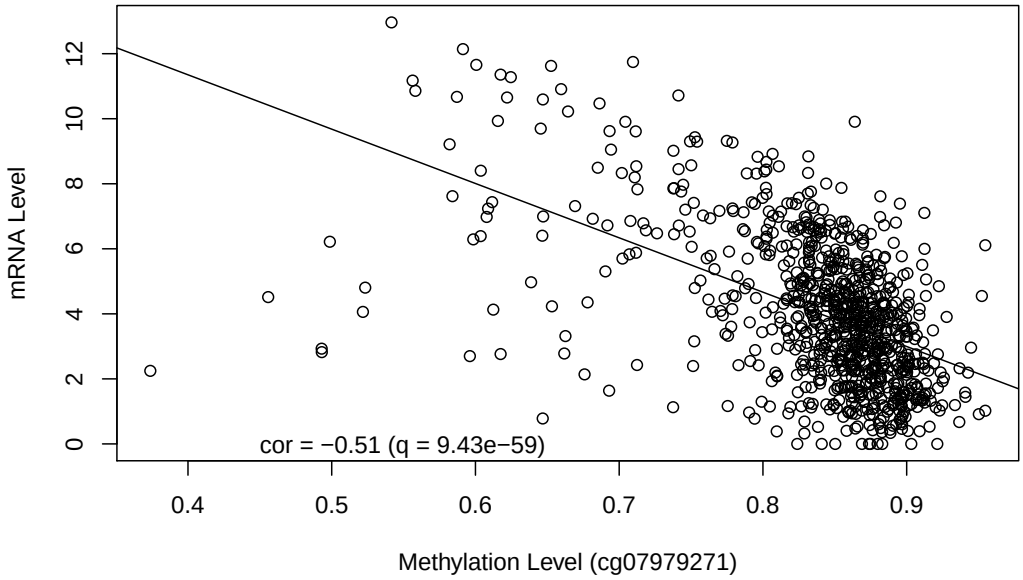


Figure 2: Scatter plot of SPIB methlyation vs mRNA expression values. Spearman correlation shown. Data from the TCGA BRCA dataset [2, 3].

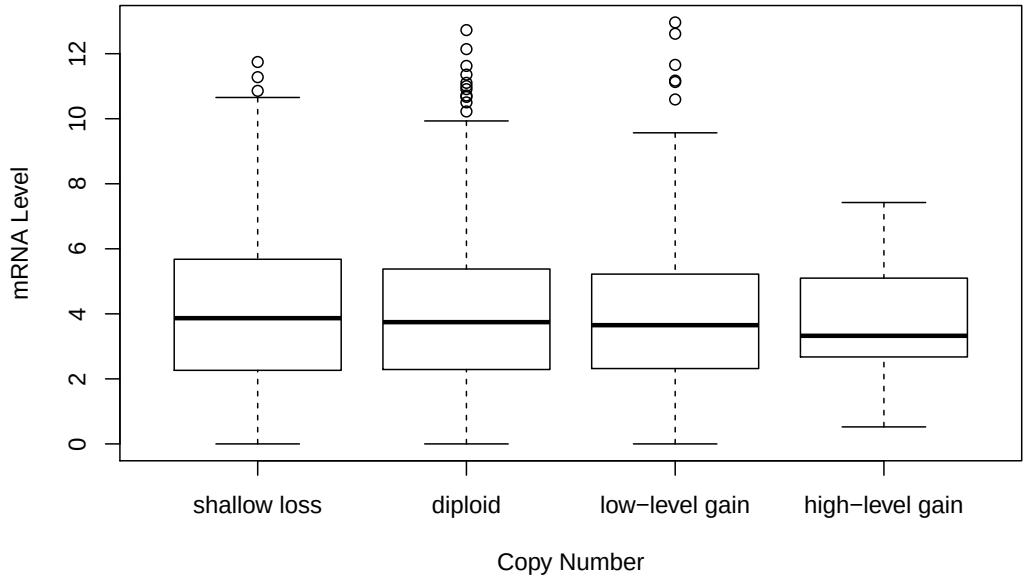


Figure 3: mRNA expression of SPIB compared under different copy number alteration levels. Data from the TCGA BRCA dataset [2, 3].

Table 2: mRNA Co-Expression

Gene	cor	p-value	q-value	Mean (SPIB-High)	Mean (SPIB-Low)
MS4A1	0.822	$2.95 \cdot 10^{-299}$	$2.98 \cdot 10^{-295}$	6.88	3.16
TCL1A	0.820	$1.90 \cdot 10^{-296}$	$1.28 \cdot 10^{-292}$	4.23	1.08
CXCR5	0.819	$3.59 \cdot 10^{-296}$	$1.82 \cdot 10^{-292}$	5.15	2.59
LCK	0.816	$7.29 \cdot 10^{-292}$	$2.95 \cdot 10^{-288}$	7.86	5.62
ACAP1	0.811	$1.52 \cdot 10^{-285}$	$5.12 \cdot 10^{-282}$	7.44	5.15
CCR7	0.809	$5.04 \cdot 10^{-283}$	$1.46 \cdot 10^{-279}$	7.26	4.90
LTB	0.808	$3.61 \cdot 10^{-282}$	$9.13 \cdot 10^{-279}$	7.48	4.68
CD5	0.801	$4.46 \cdot 10^{-273}$	$9.99 \cdot 10^{-270}$	7.41	5.13
IL2RG	0.800	$1.54 \cdot 10^{-272}$	$3.13 \cdot 10^{-269}$	9.45	7.41
CD6	0.799	$9.15 \cdot 10^{-271}$	$1.68 \cdot 10^{-267}$	7.86	5.84
SIT1	0.795	$4.51 \cdot 10^{-266}$	$7.61 \cdot 10^{-263}$	5.92	3.67
CD3D	0.795	$5.73 \cdot 10^{-266}$	$8.93 \cdot 10^{-263}$	7.12	4.79
UBASH3A	0.792	$5.48 \cdot 10^{-263}$	$7.93 \cdot 10^{-260}$	5.31	3.02
MAP4K1	0.792	$1.64 \cdot 10^{-262}$	$2.22 \cdot 10^{-259}$	7.39	5.55
CD27	0.791	$9.02 \cdot 10^{-262}$	$1.14 \cdot 10^{-258}$	7.54	5.23
SPOCK2	0.791	$3.1 \cdot 10^{-261}$	$3.69 \cdot 10^{-258}$	9.56	7.68
CCL4	0.524	$5.7 \cdot 10^{-87}$	$2.66 \cdot 10^{-85}$	7.07	5.92
DCTN4	-0.417	$1.82 \cdot 10^{-52}$	$4.77 \cdot 10^{-51}$	10.7	11.1
FOXA1	-0.407	$6.87 \cdot 10^{-50}$	$1.70 \cdot 10^{-48}$	10.9	12.0
PLA2G12A	-0.401	$2.3 \cdot 10^{-48}$	$5.47 \cdot 10^{-47}$	9.83	10.3
FNIP1	-0.397	$3.88 \cdot 10^{-47}$	$9.03 \cdot 10^{-46}$	9.48	9.96
USP30	-0.396	$5.19 \cdot 10^{-47}$	$1.21 \cdot 10^{-45}$	8.61	8.96
GLRB	-0.393	$2.47 \cdot 10^{-46}$	$5.62 \cdot 10^{-45}$	6.61	7.61
STRN3	-0.392	$6.22 \cdot 10^{-46}$	$1.40 \cdot 10^{-44}$	9.28	9.67
TMEM192	-0.390	$1.31 \cdot 10^{-45}$	$2.92 \cdot 10^{-44}$	8.88	9.32
RNF14	-0.384	$3.89 \cdot 10^{-44}$	$8.39 \cdot 10^{-43}$	10.3	10.6

Spearman correlation of mRNA of multiple genes and the mRNA expression of SPIB. The upper half is sorted based on q-values (ascending). The lower half is sorted based on correlation coefficient (ascending). (”cor” is the correlation coefficient ρ)

References

[1] *Breast cancer statistics*. Sept. 2021. URL: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/breast-cancer.

[2] Giovanni Ciriello et al. "Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer". eng. In: *Cell* 163.2 (Oct. 2015), pp. 506–519. ISSN: 1097-4172. DOI: 10.1016/j.cell.2015.09.033.

[3] Mary Goldman et al. *The UCSC Xena platform for public and private cancer genomics data visualization and interpretatio*. Tech. rep. Type: article. 2018. DOI: 10.1101/326470. URL: https://doi.org/10.1101/326470 (visited on 11/27/2021).

[4] Qiumin Huang et al. "Spi-B Promotes the Recruitment of Tumor-Associated Macrophages via Enhancing CCL4 Expression in Lung Cancer". In: *Frontiers in Oncology* 11 (2021), p. 1877. ISSN: 2234-943X. DOI: 10.3389/fonc.2021.659131. URL: https://www.frontiersin.org/article/10.3389/fonc.2021.659131 (visited on 11/26/2021).

[5] Xiaotao Su et al. "Study on the Prognostic Values of Dynactin Genes in Low-Grade Glioma". en. In: *Technology in Cancer Research & Treatment* 20 (Jan. 2021), p. 15330338211010143. ISSN: 1533-0346. DOI: 10.1177/15330338211010143. URL: https://doi.org/10.1177/15330338211010143 (visited on 12/02/2021).

[6] Shijun Wang et al. "Distinct prognostic value of dynactin subunit 4 (DCTN4) and diagnostic value of DCTN1, DCTN2, and DCTN4 in colon adenocarcinoma". In: *Cancer Management and Research* 10 (Nov. 2018), pp. 5807–5824. ISSN: 1179-1322. DOI: 10.2147/CMAR.S183062. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6248376/ (visited on 12/02/2021).

[7] Ido Wolf et al. "FOXA1: Growth inhibitor and a favorable prognostic factor in human breast cancer". en. In: *International Journal of Cancer* 120.5 (2007), pp. 1013–1022. ISSN: 1097-0215. DOI: 10.1002/ijc.22389. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/ijc.22389 (visited on 12/03/2021).

[8] Hua Zhang et al. "SPIB promotes anoikis resistance via elevated autolysosomal process in lung cancer cells". en. In: *The FEBS Journal* 287.21 (2020), pp. 4696–4709. ISSN: 1742-4658. DOI: 10.1111/febs.15272. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/febs.15272 (visited on 11/26/2021).

[9] Xunping Zhao et al. "SPIB acts as a tumor suppressor by activating the NFkB and JNK signaling pathways through MAP4K1 in colorectal cancer cells". en. In: *Cellular Signalling* 88 (Dec. 2021), p. 110148. ISSN:

0898-6568. DOI: 10.1016/j.cellsig.2021.110148. URL: https://www.sciencedirect.com/science/article/pii/S0898656821002370 (visited on 11/26/2021).

Appendix MRes 2021 Assignment (Jan-Philipp Cieslik)

Setup - Bash (Terminal)

Download data sets from Xena Browser (TCGA BRCA)
https://xenabrowser.net/datapages/?cohort=TCGA%20Breast%20Cancer%20(BRCA)

```
wget -O survival.tsv \
https://tcga-xena-hub.s3.us-east-1.amazonaws.com/download/survival%2FBRCA_survival.txt
wget -O clinical_matrix.tsv \
https://tcga-xena-hub.s3.us-east-1.amazonaws.com/download/TCGA.BRCA.sampleMap%2FBRCA_clinicalMatrix
wget -O HiSeqV2.tsv.gz \
https://tcga-xena-hub.s3.us-east-1.amazonaws.com/download/TCGA.BRCA.sampleMap%2FHiSeqV2.gz
wget -O Methylation450k.tsv.gz \
https://tcga-xena-hub.s3.us-east-1.amazonaws.com/download/TCGA.BRCA.sampleMap%2FHumanMethylation450.gz
wget -O Methylation450k_probemap.tsv \
https://tcga-xena-hub.s3.us-east-1.amazonaws.com/download/probeMap%2FilluminaMethyl450_hg19_GPL16304_TCGAlegacy
wget -O CNV_thresholded.tsv.gz \
https://tcga-xena-hub.s3.us-east-1.amazonaws.com/download/TCGA.BRCA.sampleMap%2FGistic2_CopyNumber_Gistic2_all_thre
```

Unzip the downloaded files

```
gunzip HiSeqV2.tsv.gz
gunzip CNV_thresholded.tsv.gz
gunzip Methylation450k.tsv.gz
```

Setup - R

Load libraries

```
library(survival)
#with the given option the fread function from data.table behaves like read.table
#but is much quicker and memory efficient
library(data.table)
options(datatable.fread.datatable=FALSE)
```

Data Loading - RNA/Survival/Clinical

Load data and adjust row names

```
#set the gene of interest
gene <- "SPIB"
#using check.names=FALSE to prevent the change from hyphens to dots
rna.data <- fread("HiSeqV2.tsv", sep="\t", head=TRUE,
                  stringsAsFactors =FALSE, check.names=FALSE)

rownames(rna.data) <- make.unique(rna.data[, 1])
rna.data <- rna.data[,-1]
rna.data <- as.matrix(rna.data)
```

```
surv.data <- read.table("survival.tsv", sep="\t", header=T, row.names=1)
clin.data <- read.table("clinical_matrix.tsv", sep="\t", header=T, row.names=1, quote = "")
```

Generate survival data

```
os.time <- surv.data[colnames(rna.data),"OS.time"]
os.event <- as.numeric(surv.data[colnames(rna.data),"OS"])
brca.os <- Surv(os.time,os.event)
#Delete local variables (optional, just to keep the environment clean)
rm("os.time", "os.event")
```

Univariate Regression Analysis (RNA/Survival)

Create empty data frame for results

```
rna.survival.univariate<-array(NA, c(nrow(rna.data),4))
colnames(rna.survival.univariate)<-c("HR","LCI","UCI","PVAL")
rownames(rna.survival.univariate)<-rownames(rna.data)
rna.survival.univariate<-as.data.frame(rna.survival.univariate)
```

Iterate through all genes and generate Cox model

```
for(i in 1:nrow(rna.data))
{
  #Check if less than 2 samples are available for correlation
  if(sum(!is.na(rna.data[i,])) < 2){
    next
  }
 coxphmodel <- coxph(brca.os ~ as.numeric(rna.data[i,]))
 summary <-summary(coxphmodel)
 rna.survival.univariate$HR[i] <- summary$coef[1,2]
 rna.survival.univariate$LCI[i] <- summary$conf.int[1,3]
 rna.survival.univariate$UCI[i] <- summary$conf.int[1,4]
 rna.survival.univariate$PVAL[i] <- summary$coef[1,5]
}
rna.survival.univariate <- as.data.frame(rna.survival.univariate)
rna.survival.univariate$FDR <- p.adjust(rna.survival.univariate$PVAL,method="fdr")
rna.survival.univariate <-
  rna.survival.univariate[order(rna.survival.univariate$FDR, decreasing=F),]

#Remove local variables
rm("summary", "i")
```

Print results of univariate analysis

```
kable(rna.survival.univariate[1:5,])
```

	HR	LCI	UCI	PVAL	FDR
LOC729467	1.3728688	1.2272374	1.5357818	0	0.0001751
EPHA5	1.3973600	1.2402194	1.5744108	0	0.0001751
PSME2	0.5940589	0.4930553	0.7157534	0	0.0001751
LOC148145	1.6993532	1.4157505	2.0397671	0	0.0001751
ANO6	1.5970418	1.3539352	1.8837995	0	0.0001751

Multivariate Regression Analysis (RNA/Survival/Clinical)

Clinical data preparation

```
#subset clinical data to patients that also have RNA data
clin.data<-clin.data[colnames(rna.data),]
#create age variable
age<-as.numeric(clin.data$age_at_initial_pathologic_diagnosis)
#create stage high/low variable
x3<-grep("III",clin.data$Converted_Stage_nature2012)
x4<-grep("IV",clin.data$Converted_Stage_nature2012)
stage.high<-rep(0,nrow(clin.data))
stage.high[c(x3,x4)]<-1

#Remove local variable
rm("x3", "x4")
```

Create empty data frame for results

```
rna.survival.multivariate<-array(NA, c(nrow(rna.data),4))
colnames(rna.survival.multivariate)<-c("HR","LCI","UCI","PVAL")
rownames(rna.survival.multivariate)<-rownames(rna.data)
rna.survival.multivariate<-as.data.frame(rna.survival.multivariate)
```

Iterate through all genes to generate multivariate regression model

```
for(i in 1:nrow(rna.data))
{
  #Check if less than 2 samples are available for correlation
  if(sum(!is.na(rna.data[i,])) < 2){
    next
  }
 coxphmodel <- coxph(brca.os ~ rna.data[i,]+age+stage.high)
 summary <- summary(coxphmodel)
 rna.survival.multivariate$HR[i] <- summary$coef[1,2]
 rna.survival.multivariate$LCI[i] <- summary$conf.int[1,3]
 rna.survival.multivariate$UCI[i] <- summary$conf.int[1,4]
 rna.survival.multivariate$PVAL[i] <- summary$coef[1,5]
}
rna.survival.multivariate <- as.data.frame(rna.survival.multivariate)
rna.survival.multivariate$FDR <- p.adjust(rna.survival.multivariate$PVAL,method="fdr")
rna.survival.multivariate <-
  rna.survival.multivariate[order(rna.survival.multivariate$FDR, decreasing=F),]
#Remove local variables
rm("summary", "i")
```

Print results of multivariate analysis

```
kable(rna.survival.multivariate[1:5,])
```

	HR	LCI	UCI	PVAL	FDR
SIAH2	0.6628204	0.5720274	0.7680241	0e+00	0.0006595
LOC148145	1.6765143	1.3900125	2.0220682	1e-07	0.0006595
MRO	1.2952128	1.1721721	1.4311689	4e-07	0.0010961
PSME2	0.6155775	0.5105837	0.7421618	4e-07	0.0010961
PCDHGA3	1.2974274	1.1741901	1.4335992	3e-07	0.0010961

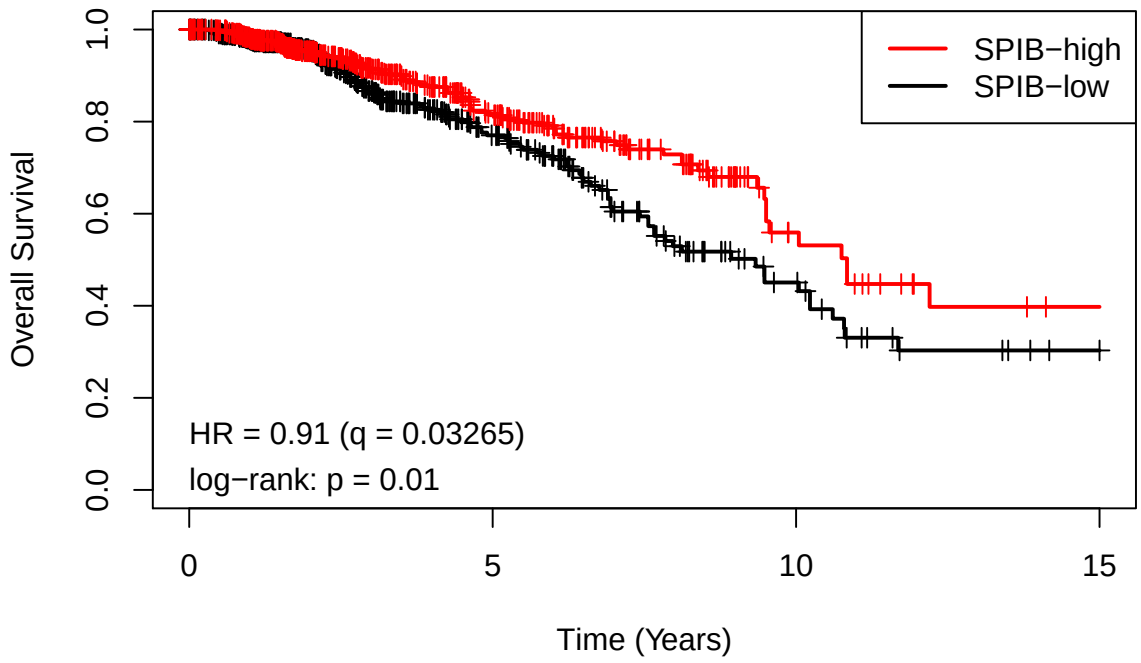
```r
gene.info <- rna.survival.multivariate[gene,]
gene.high <- as.numeric(rna.data[gene,]>median(rna.data[gene,]))

#calculate log rank test
gene.high.logrank <- survdiff(brca.os ~ gene.high)
gene.high.logrank.p <- 1 - pchisq(gene.high.logrank$chisq, length(gene.high.logrank$n) - 1)
print(survfit(brca.os ~ gene.high))
```

```
## Call: survfit(formula = brca.os ~ gene.high)
##
##    4 observations deleted due to missingness
##                n events median 0.95LCL 0.95UCL
## gene.high=0 609    111   3409    2763    3873
## gene.high=1 605     87   3959    3472      NA
```

```r
#create survival plot
plot.text <- paste0("HR = ",round(gene.info["HR"], digits = 2),
                    " (q = ",round(gene.info["FDR"], digits = 5),")")
plot.text2 <- paste0("log-rank: p = ",round(gene.high.logrank.p, digits = 2))
plot.legend <- c(paste0(gene,"-high"),paste0(gene,"-low"))


plot(survfit(brca.os ~ gene.high), col=c("black","red"), lwd=2,
     mark.time=TRUE, xlab="Time (Years)", ylab="Overall Survival",
     xscale = 365.25, xmax = 15*365.25)
legend("topright",legend=plot.legend,col=c("red","black"),lwd=2)
text(0,0.1,plot.text, adj = c(0,0))
text(0,0,plot.text2, adj = c(0,0))
```




```
#delete local variables
rm("plot.text", "plot.text2", "plot.legend")
```

Methylation data

Load methylation data

```
meth.annotation <- read.table("Methylation450k_probemap.tsv", sep="\t",
                              header=T, comment.char="")
meth.data <- fread("Methylation450k.tsv", sep="\t", header=TRUE)

row.names(meth.data) <- meth.data[,1]
row.names(meth.annotation) <- meth.annotation[,1]
meth.data <- meth.data[, -1]
meth.annotation <- meth.annotation[, -1]
meth.probes<- rownames(meth.annotation[grep(gene, meth.annotation$gene),])
```

Display methylation data

```
kable(meth.data[1:3,1:3])
```

	TCGA-OL-A66H-01	TCGA-3C-AALK-01	TCGA-AC-A5EH-01
cg13332474	0.0192	0.2032	0.3003
cg00651829	0.0179	0.2890	0.0892
cg17027195	0.0367	0.0750	0.0333

```
kable(meth.annotation[1:3,])
```

	gene	chrom	chromStart	chromEnd	strand
cg13332474	.	chr7	25935146	25935148	.
cg00651829	RSPH14,GNAZ	chr22	23413065	23413067	.
cg17027195	AUTS2	chr7	69064092	69064094	.

subset data set to samples that have rna, methylation and survival data.

```
samples <- intersect(colnames(meth.data),colnames(rna.data))
meth.intersect.rna <- meth.data[,samples]
rna.intersect.meth <- rna.data[,samples]
surv.intersect.rna.meth <-
  surv.data[intersect(rownames(surv.data),colnames(rna.intersect.meth)),]

meth.intersect.rna <- as.matrix(meth.intersect.rna)
rna.intersect.meth <- as.matrix(rna.intersect.meth)
surv.intersect.rna.meth <-
  as.data.frame(surv.intersect.rna.meth[colnames(meth.intersect.rna),])

meth.intersect.rna <-  meth.intersect.rna[meth.probes,,drop=FALSE]

rm("samples")
```

```
#exclude methylation sites that are not determined in more than 0.5 of samples
na.count <- apply(meth.intersect.rna,1,function(x) sum(as.numeric(is.na(x))))
exclude <- as.numeric(na.count>0.5*ncol(meth.intersect.rna))
meth.intersect.rna <- meth.intersect.rna[which(exclude==0),, drop=FALSE]

#generate empty array for results
results.meth<-array(NA,c(nrow(meth.intersect.rna),5))
rownames(results.meth)<-rownames(meth.intersect.rna)
colnames(results.meth)<-c("Cor","pval","qval","Mean.high","Mean.low")
gene.high.meth.rna <- as.numeric(
  as.numeric(rna.intersect.meth[gene,]) > median( as.numeric(rna.intersect.meth[gene,]) )
  )

#remove local variables
rm("na.count", "exclude")
```

Iterate through every methylation site of the selected gene and perform correlation.

```
for (i in 1:nrow(meth.intersect.rna))
{
  results.meth[i,1] <-
    cor.test(as.numeric(rna.intersect.meth[gene,]),as.numeric(meth.intersect.rna[i,]),
             use="c", method = "spearman",exact=FALSE)$est
  results.meth[i,2] <-
    cor.test(as.numeric(rna.intersect.meth[gene,]),as.numeric(meth.intersect.rna[i,]),
             use="c", method = "spearman",exact=FALSE)$p.value
}
results.meth[,4] <- apply(meth.intersect.rna[,which(gene.high.meth.rna==1), drop=FALSE],
                          1,mean,na.rm=T)
results.meth[,5] <- apply(meth.intersect.rna[,which(gene.high.meth.rna==0), drop=FALSE],
```

```
                          1,mean,na.rm=T)
results.meth[,3] <- p.adjust(results.meth[,2],method="fdr")
results.meth<-results.meth[order(results.meth[,3], decreasing=F),,drop=FALSE]

#remove local variables
rm("i")
```

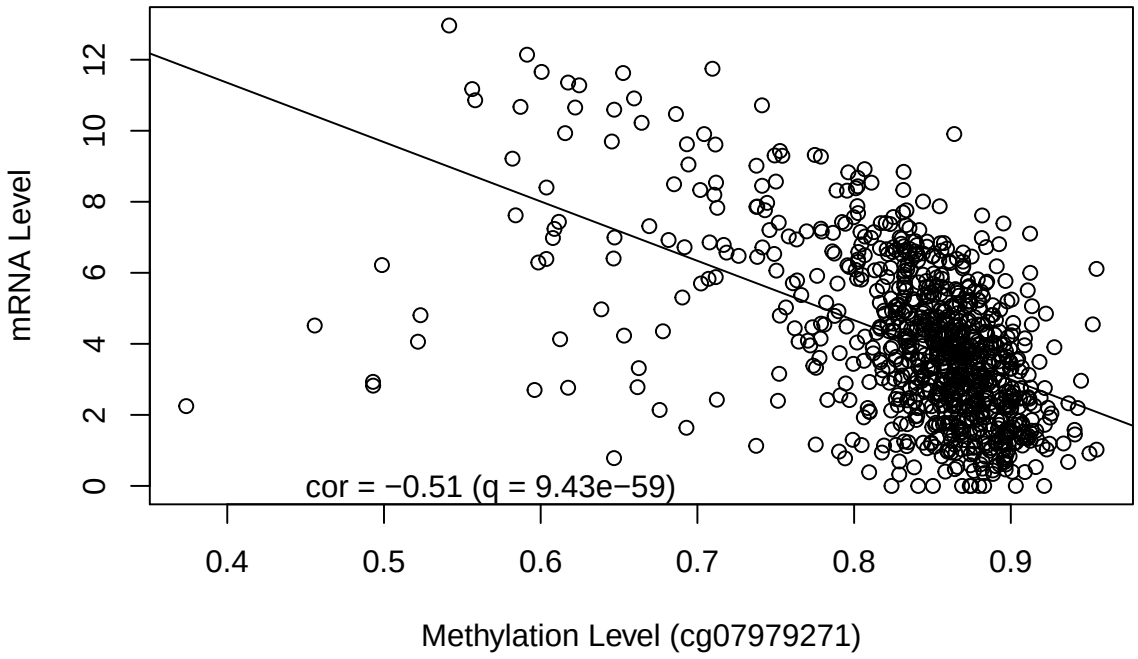
Display results

```
kable(results.meth)
```

	Cor	pval	qval	Mean.high	Mean.low
cg07979271	-5.14e-01	0.00e+00	0.00e+00	0.816	0.864
cg13403724	2.88e-01	0.00e+00	0.00e+00	0.248	0.206
cg17774764	-2.81e-01	0.00e+00	0.00e+00	0.737	0.775
cg13918544	-2.77e-01	0.00e+00	0.00e+00	0.832	0.859
cg06512885	2.72e-01	0.00e+00	0.00e+00	0.191	0.168
cg15690347	2.66e-01	0.00e+00	0.00e+00	0.383	0.301
cg24092179	2.65e-01	0.00e+00	0.00e+00	0.260	0.226
cg15007959	2.63e-01	0.00e+00	0.00e+00	0.253	0.203
cg18254819	2.46e-01	0.00e+00	0.00e+00	0.235	0.212
cg03763616	2.28e-01	0.00e+00	0.00e+00	0.271	0.244
cg22268231	-1.47e-01	1.22e-05	1.88e-05	0.450	0.487
cg21152077	-1.33e-01	7.89e-05	1.12e-04	0.875	0.881
cg26522743	-9.68e-02	4.20e-03	5.49e-03	0.394	0.424
cg04508467	9.21e-02	6.53e-03	7.92e-03	0.464	0.448
cg08201854	8.45e-02	1.25e-02	1.41e-02	0.132	0.143
cg22745102	7.05e-02	3.72e-02	3.96e-02	0.469	0.462
cg19387862	2.02e-05	1.00e+00	1.00e+00	0.696	0.686

```
#set the methylation site of interest
meth.site = "cg07979271"
plot.title <- paste0(gene, " in BRCA")
plot.text <- paste0("cor = ",round(results.meth[meth.site, "Cor"], digits = 2),
                    " (q = ",format(results.meth[meth.site, "qval"], scientific = TRUE),")")
plot.legend <- c(paste0(gene,"-high"),paste0(gene,"-low"))


plot(as.numeric(meth.intersect.rna[meth.site,]), as.numeric(rna.intersect.meth[gene,]),
     xlab=paste0("Methylation Level (", meth.site,")"), ylab="mRNA Level")
text(0.45,0,plot.text, adj = c(0,0.5))
abline(lm(rna.intersect.meth[gene,] ~ meth.intersect.rna[meth.site,]))
```



CNV data

Load CNV data and display table

```
cnv.data <- fread("CNV_thresholded.tsv", sep="\t", header=T)
rownames(cnv.data) <- make.unique(cnv.data[,1])
cnv.data <- cnv.data[,-1]
cnv.data <- as.matrix(cnv.data)
#subset data set to entries that are also available in the RNA data set
cols.intersect <- intersect(colnames(cnv.data), colnames(rna.data))
row.intersect <- intersect(rownames(cnv.data), rownames(rna.data))
cnv.intersect.rna <- cnv.data[row.intersect, cols.intersect]
rna.intersect.cnv <- rna.data[row.intersect, cols.intersect]
kable(cnv.data[1:5,1:3])
```

	TCGA-3C-AAAU-01	TCGA-3C-AALI-01	TCGA-3C-AALJ-01
ACAP3	0	-1	-1
ACTRT2	0	-1	-1
AGRN	0	-1	-1
ANKRD65	0	-1	-1
ATAD3A	0	-1	-1

```
rm("cols.intersect","row.intersect")
```

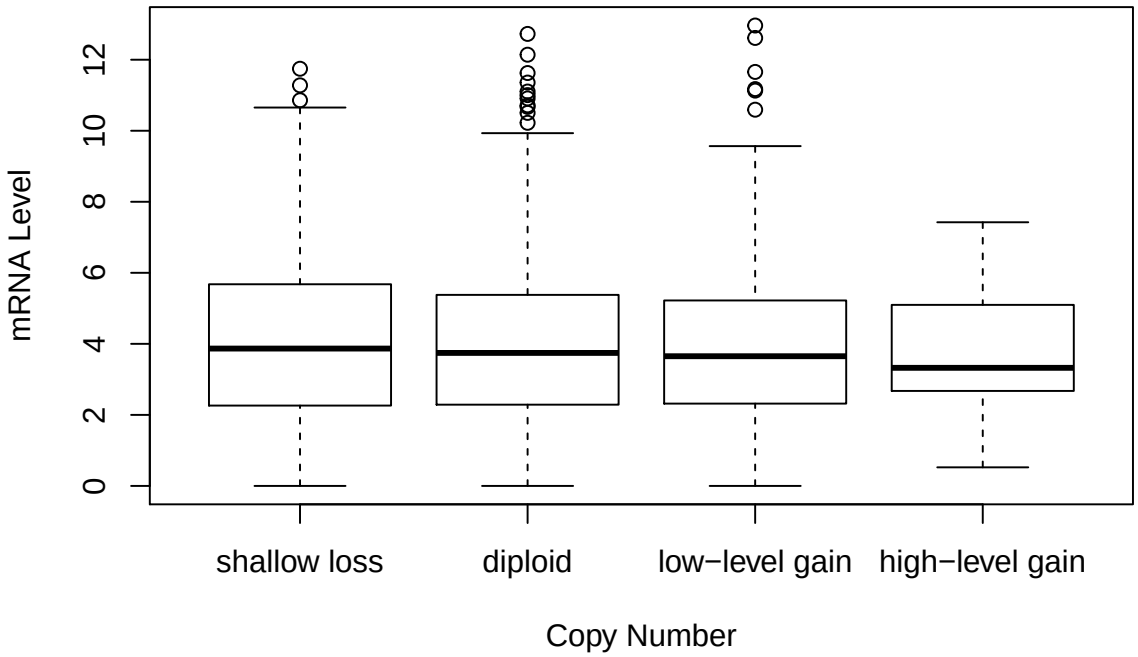
Calculate CNV changes in selected gene

```
cnv.rna.df <- data.frame(CNV = cnv.intersect.rna[gene,],
                         RNA = rna.intersect.cnv[gene,],
                         stringsAsFactors=FALSE)
cnv.rna.df[which(cnv.rna.df$CNV == -2), "CNV"] <- "deep loss"
cnv.rna.df[which(cnv.rna.df$CNV == -1), "CNV"] <- "shallow loss"
cnv.rna.df[which(cnv.rna.df$CNV == 0), "CNV"] <- "diploid"
cnv.rna.df[which(cnv.rna.df$CNV == 1), "CNV"] <- "low-level gain"
cnv.rna.df[which(cnv.rna.df$CNV == 2), "CNV"] <- "high-level gain"
#"deep loss" is missing on purpose,
#as no entry falls into this category for our candidate gene
cnv.rna.df$CNV <- factor(cnv.rna.df$CNV,
                         levels=c("shallow loss", "diploid",
                                  "low-level gain", "high-level gain"),
                         ordered=TRUE)

aov(RNA ~ CNV, data = cnv.rna.df)
```

```
## Call:
##    aov(formula = RNA ~ CNV, data = cnv.rna.df)
##
## Terms:
##                       CNV Residuals
## Sum of Squares     2.716  5920.177
## Deg. of Freedom        3      1072
##
## Residual standard error: 2.350011
## Estimated effects may be unbalanced
## 2 observations deleted due to missingness
```

```
boxplot(RNA~CNV, data=cnv.rna.df,
   xlab="Copy Number", ylab="mRNA Level")
```



mRNA correlation

Create empty data frame for results

```
rna.cor <-array(NA,c(nrow(rna.data),5))
rownames(rna.cor)<-rownames(rna.data)
colnames(rna.cor)<-c("Cor","pval","qval","Mean.high","Mean.low")

gene.high.meth.rna <- as.numeric(
  as.numeric(rna.data[gene,]) > median( as.numeric(rna.data[gene,]) )
  )
```

Loop through every gene and correlate it with the candidate gene

```
for (i in 1:nrow(rna.data)){
  #Check if less than 2 samples are available for correlation
  if(sum(!is.na(rna.data[i,]) & !is.na(rna.data[gene,])) < 2){
    next
  }
  #Check if all values are zero
  if(sum(rna.data[i,]) == 0){
    next
  }
  if(sum(!is.na(rna.data[i,])))
  result.temp <-
    cor.test(as.numeric(rna.data[gene,]), as.numeric(rna.data[i,]),
```

```
            use="c", method = "spearman", exact=FALSE)
  rna.cor[i, 1:2] <- c(result.temp$est, result.temp$p.value)
}
rna.cor[,4] <- apply(rna.data[,which(gene.high.meth.rna==1), drop=FALSE],1,mean,na.rm=T)
rna.cor[,5] <- apply(rna.data[,which(gene.high.meth.rna==0), drop=FALSE],1,mean,na.rm=T)
rna.cor[,3] <- p.adjust(rna.cor[,2],method="fdr")
rna.cor <- rna.cor[order(rna.cor[,3], decreasing=F),,drop=FALSE]
```

Display results

```
kable(rna.cor[1:5,])
```

	Cor	pval	qval	Mean.high	Mean.low
SPIB	1.0000000	0	0	5.687704	2.108523
MS4A1	0.8217339	0	0	6.883461	3.159181
TCL1A	0.8196190	0	0	4.234020	1.084256
CXCR5	0.8194083	0	0	5.152683	2.590970
LCK	0.8161056	0	0	7.855356	5.622874

Save workspace

```
save.image("assignment.RData")
```