High FKBP4 expression is associated with increased mortality risk in lung adenocarcinoma patients

Kenneth Russell Ancheta

A report submitted in partial fulfilment of the requirements for the degree of
MRes in Cancer Biology (Informatics Stream)
Imperial College of Science, Technology and Medicine
December 2020

1. Introduction

Lung cancer is the most common cause of cancer-related death in the UK (Cancer Research UK, 2020). Lung cancer is categorised into two types: small cell (15%) and non-small-cell lung carcinoma (NSCLC; 85%). Squamous cell carcinoma and adenocarcinoma (LUAD) are the main subtypes of NSCLC, with LUAD being the most common (63%) (Gridelli et al., 2015). Global advancements in screening and diagnostic strategies have increased LUAD incidence in a given population and developments in precision of classification system alongside current predictive and prognostic markers have massively improved therapeutic options for individual patients (Rami-Porta et al., 2018; Gridelli et al., 2015). However, staging post-tumour resection remains the strongest prognostic indicator for all types of lung cancer (Rami-Porta et al., 2018). Therefore, there remains an urgency to discover new and less invasive prognostic biomarkers to improve treatment strategies for LUAD patients.

FK506-binding proteins (FKBPs) belongs to the immunophilin family that is known to bind to FK506, an immunosuppressive drug (Hong et al., 2017). FKBPs have been reported to be involved in different pathways such in stress response, neuronal function, foetal development, cardiac function and oncogenesis (Xiong, et al., 2020). An increasing number of reports observed that FKBP4 (also known as FKBP52) was elevated in different cancer types including bladder, breast, colorectal, gastric, leukaemia, lymphoma and ovarian with various oncogenic effects (Xiong, et al., 2020). For example, upregulated FKBP4 gene expression was found to be positively correlated with poor prognosis in hormone-dependent cancers such as breast cancer (Hong et al., 2017). However, the prognostic value of FKBP4 for LUAD patients remains unknown.

The present study aims to highlight the prognostic potential of FKBP4 in LUAD. The TGCA LUAD dataset was investigated to identify the relationship of FKBPA gene expression and survival rate against different clinicopathological features. Various mechanisms that alter gene expressions such as DNA methylation, gene copy number and mature miRNA were investigated to understanding the underlying cause of elevated FKBPA levels.

2. Methods

2.1 Data sources

TGCA LUAD RNAseq, DNA Methylation, copy number profile (CNV), protein expression, mature miRNA strand expression, curated survival and phenotype datasets were obtained from UCSC XenaBrowser (https://xenabrowser.net/datapages/), providing 576, 492, 516, 237, 495, 641 and 704 samples, respectively. RNAseq data were measured using polyA+ IlluminaHiSeq and the gene-level transcription estimates were expressed as $\log_2$(normal count+1) transformed by expectation maximisation normalised count. DNA methylation profile was calculated using Illumina Infinium HumanMethylation450 platform and was reported as beta-values. CNV profiles were quantified using whole-genome microarray. Segmented CNV data was produced using GISTIC2 method (estimate scores). DNA methylation microarray probes, RNAseq genes and CNV microarray were mapped onto UCSC Xena HUGO probeMap. Total protein expression was obtained using reverse-phase protein array (RPPA) and normalised using replicate-base normalisation. miRNA mature strand was quantified using IlluminaHiSeq and the sum of all isoforms for the same miRNA was expressed as $\log_2$(total reads per million + 1) transformed.

2.2 Survival analysis

Patients survival was evaluated using Kaplan–Meier survival curve. Cox proportional-hazard model was performed with RNAseq and RPPA datasets to estimate independent risk factors for the mortality of patients before adjusting for clinicopathological factors. The significance of protein levels and RNA expression on survival was determined using false discovery rate (FDR). Further analysis was done on RNAseq data. Statistically significant genes were stratified into high and low expression based on their median value. Univariate Cox regression was performed with RNAseq data and clinical data including age, gender, histological type, anatomical neoplasm subdivision, location in lung parenchyma, stage (stratified to advanced (III-IV) and low (I-II)), smoking history, new tumour event after initial treatment (NTE) and targeted therapy. Variables with an FDR < 0.05 in univariate analysis were entered into a multivariate Cox regression.

Genes with FDR < 0.05 in multivariate cox regression were considered statistically significant. Hazard ratios (HRs) and 95% confidence intervals (CIs) were estimated.

2.3 DNA methylation and gene expression correlation analysis

DNA methylation profile of CpG sites of genes identified in survival analysis was annotated to using HumanMethylation450K annotation file available from Imperial College London BRC server. CpG sites missing from more than 50% of the samples were excluded from the analysis. Spearman correlation was used to detect the association of gene expression (stratified to high and low base on median) and DNA methylation (M-values). CpG sites with $p < 0.05$ were deemed significant irrespective of the Spearman correlation. The difference in DNA methylation was determined using the absolute change in the mean beta-values between stratified groups.

2.4 CNV analysis

Copy number of the genes was determined based on the median GISTIC2 estimated values across all the samples. GISTIC2 scores estimates ranges from -2 to 2 where: 2 = high-level copy number amplification, 1 = low-level copy number amplification, 0 = diploid normal copy, -1 = single copy deletion and -2 = homozygous deletion.

2.5 RNA levels and miRNA expression heatmap

This analysis was implemented at https://xenabrowser.net/heatmap/. The expression of mature miRNA strands was mapped against the RNA levels of gene of interest to produce a heatmap to detect an association in the miRNA levels and gene expression in TGCA LUAD dataset (Supplementary Data 4b).

2.6 Statistical analyses platform and dependencies

All statistical analyses were performed in R version 1.3.1093. R packages used: “survival”, “dpylr” and “survminer”.

3. Results

3.1 FKBP4 gene is a potential novel prognostic biomarker for LUAD

To identify a novel protein as a prognostic biomarker for LUAD patients, a survival analysis was employed using RPPA data. All the 131 proteins available in TGCA LUAD RPPA dataset showed no statistical significance in patient overall survival (OS) after adjusting for multiple comparisons (FDR > 0.05) (Supplementary Data 1). Therefore, survival analysis in RNAseq data was explored. Around 6% of the genes were statistically associated with LUAD patient survival (FDR < 0.05). Univariate Cox regression was executed using confounding clinicopathological factors that affect OS. No genes were associated with survival when adjusted for age, gender, LUAD histological type (although signet ring adenocarcinoma result in FDR < 0.001, sample size = 1), anatomical neoplasm subdivision, lung tumour location and targeted therapy (FDR > 0.05). Interestingly, univariate analysis showed that smoking history (i.e. non-smoker, ever-smoker, former smokers) and gene expressions associated with survival did not correlate (FDR > 0.05). Contrary to the study by Lee et. al. (2014), ever-smokers, male and advanced age (more than 63 yrs. o.) increased mortality risk for NSCLC patients. However, this might be due to the bias in the demographics and clinical characteristics of their samples, of which, 97% of ever-smokers were male with an average age of 65 yrs. o. (Lee et. al., 2014). This result indicated that smoking history, as a confounding factor, was not associated with prognosis. Nonetheless, smoking remains the biggest risk factor for developing lung cancer (Gridelli et al., 2015). Moreover, NTE and advanced stage were highly correlated with the genes that were associated with survival. Patients that had advanced stage of LUAD (Fig. 1c, HR = 2.64 (95% CI, 1.98-3.52), FDR < 0.001) and had occurrence of new tumour after initial treatment (Fig. 1b, HR = 2.67 (95% CI, 1.97-3.61), FDR < 0.001) have shortened OS. Therefore, NTE and advanced stage were accounted in the multivariate Cox regression. FKPB4 gene was found to be statistically significant in multivariate Cox regression (FDR < 0.001). Multivariate analysis demonstrated that high FKPB4 expression was strongly associated with increased overall death risk of 63% (Fig. 1a, HR = 1.63, 95% CI = 1.43-1.85; FDR < 0.001). The data also indicated that there were 25% more patients that was likely to have the event (i.e. death) after ~2.5 years of diagnosis (Supplementary Data 2). FKBP4 was also determined to be a novel prognostic marker for LUAD and was confirmed through PubMed search using the keywords: “FKBP4 LUNG CANCER PROGNOS*” and “FKBP4 LUNG ADENOCARCINOMA PROGNOS*” (including the alternative names of FKBP4). This observation aligns with previous reports in ER-positive and luminal breast cancer and epithelial ovarian cancer that indicated high FKBP4 expression was associated with worse prognosis (Xiong, et al., 2020; Mangé, et al., 2019; Lawrenson et al., 2015). Overall, these findings suggested a strong positive correlation between high FKPB4 gene expression and survival rate of LUAD patients.

3.2 High FKBP4 expression is not influenced by DNA methylation and copy number

To understand various factors that can affect FKPB4 expression, DNA methylation was investigated by exploring DNA methylation-gene expression correlation analysis. 16 CpG sites were identified in the FKPB4 gene (10 at TSS1500/200, 1 at 3’-UTR, 2 at exon 1 and 3 in the gene body). CpG site cg04611395 (a TSS200) was determined to be significantly correlated with FKBP4 expression ($p < 0.05$) with the largest difference in mean beta-values between groups of high and low FKBP4 expression (Fig. 2, difference in Beta-value = 0.038). However, this difference was relatively low to detect any direct effect on the total expression of FKPB4. Furthermore, CNV was investigated to confirm if this molecular event was responsible for upregulated FKBP4 expression. The median GISTIC2 estimated gene-level score of FKBP4 gene suggested that the majority of the samples possessed a normal diploid copy of the gene. These suggested that DNA methylation and copy number did not have a significant influence on the expression of FKPB4.

3.3 miRNA-328-3p is upregulated in patients with low FBKP4 levels

To further investigate the molecular mechanism that drives the high expression of FKBP4, the expression of the miRNA was mapped against FKBP4 RNA across all the samples to produce a heatmap. A cohort of anti-FKBP4 miRNA was analysed (Supplementary Data 4a-b). An evidence in the heatmap data demonstrated that miR-328-3p (miRBase accession number: MIMAT0000752) was generally upregulated in LUAD patients with relatively low FKBP4 expression (i.e. higher survival rates) (Fig. 4.). A study by Ma et al. (2016) reported that downregulation of miR-328-3p in NSCLC had a significant correlation with the advanced stage, lymph node metastasis and higher mortality rate. Furthermore, they also reported that increased miR-328-3p level restored radiotherapy sensitivity and suppressed survivability in NSCLC cells, including H23 cell line (LUAD cells). Consistent with this report, this result suggested that high expression of FKBP4 gene might be caused by downregulation miR-328-3p, ultimately leading to worse prognosis in LUAD patients.

4. Conclusion

Lung cancer is currently the leading cause of cancer-related deaths in the UK. There remains an urgency to discover new and less invasive prognostic biomarker for LUAD. FKBP4, a member of the FKBP immunophilin family, have been previously reported to be involved in different biological functions and oncogenic activities in various cancers. High expression of FKPB4 had been observed to be correlated with poor prognosis in breast and ovarian cancers. However, the prognostic merit of FKBP4 was elusive for LUAD patients. In the present study, high FKBP4 expression was shown to be positively associated with OS in LUAD patients, especially after accounting NTE and advanced stage, which were also both independently associated with poor prognosis. Moreover, univariate Cox analysis revealed that age, gender, smoking history, histological subtype, anatomical neoplasm subdivision, parenchymal tumour location and targeted therapy did not have a significant effect on OS. Furthermore, this study also showed that molecular events that control gene expression, particularly DNA methylation and CNV, did not induce high FKBP4 expression. Additionally, a miRNA against FKBP4 translation, miR-328-3p, was demonstrated to be upregulated in many of the patients with low FKBP4 expression. This suggested that increased miR-328-3p levels and consequent FKBP4 suppression results in better LUAD prognosis. Nonetheless, the overall molecular mechanism of FKBP4 remains an open field of research. This study underscored the prognostic potential of FKBP4 in LUAD. This also opened new research opportunities to study the prognostic and therapeutic value of FKBP4 in difference cancers.

Bibliography

Cancer research UK Lung cancer statistics. https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/lung-cancer [Accessed 22 Nov 2020].

Gridelli, C., Rossi, A., Carbone, D., Guarize, J., Karachaliou, N., Mok, T., Petrella, F., Spaggiari, F. & Rosell, R. (2015) Non-small-cell lung cancer. *Nat Rev Dis Primers*. 1 (2015), 1–13.

Hong, C., Li, T., Zhang, F., Wu, X., Chen, X., Cui, X., Zhang, G. & Cui, Y. (2017) Elevated FKBP52 expression indicates a poor outcome in patients with breast cancer. *Oncol Lett*. 14 (5), 5379–5385.

Lawrenson, K., Mhawech-Fauceglia, P., Worthington, J., Spindler, T.J., O'Brien, D., Lee, J.M., Spain, G., Sharifian, M., Wang, G., Darcy, K.M., Pejovic, T., Sowter, H., Timms, J.F., Gayther, S.A. (2015) Identification of novel candidate biomarkers of epithelial ovarian cancer by profiling the secretomes of three-dimensional genetic models of ovarian carcinogenesis. *Int J Cancer.* 137 (8) 1806–1817.

Lee, S. J., Lee, J., Park, Y. S., Lee, C. H., Lee, S. M., Yim, J. J., Yoo, C. G., Han, S. K., & Kim, Y. W. (2014) Impact of smoking on mortality of patients with non-small cell lung cancer. *Thoracic cancer*. 5 (1), 43–49.

Ma, W., Ma, C., Zhou, N., Li, X., & Zhang, Y. (2016) Upregulation of miR-328-3p sensitizes non-small cell lung cancer to radiotherapy. *Sci Rep*. 6 (31651) 1–8.

Mangé, A., Coyaud, E., Desmetz, C., Laurent, E., Béganton, B., Coopman, P., Raught, B. & Solassol, J. (2019) FKBP4 connects mTORC2 and PI3K to activate the PDK1/Akt-dependent cell proliferation signaling in breast cancer. *Theranostics*. 9 (23), 7003–7015.

Rami-Porta, R., Call, S., Dooms, C., Obiols, C., Sánchez, M., Travis, W. D., & Vollmer, I. (2018). Lung cancer staging: a concise update. European Respiratory Journal, 51(5), 1–12.

Xiong, H., Chen, Z., Zheng, W., Sun, J., Fu, Q., Teng, R., Chen, J., Xie, S., Wang, L., Yu, X. F., & Zhou, J. (2020) FKBP4 is a malignant indicator in luminal A subtype of breast cancer. *Journal of Cancer*. 11 (7), 1727–1736.

Figures

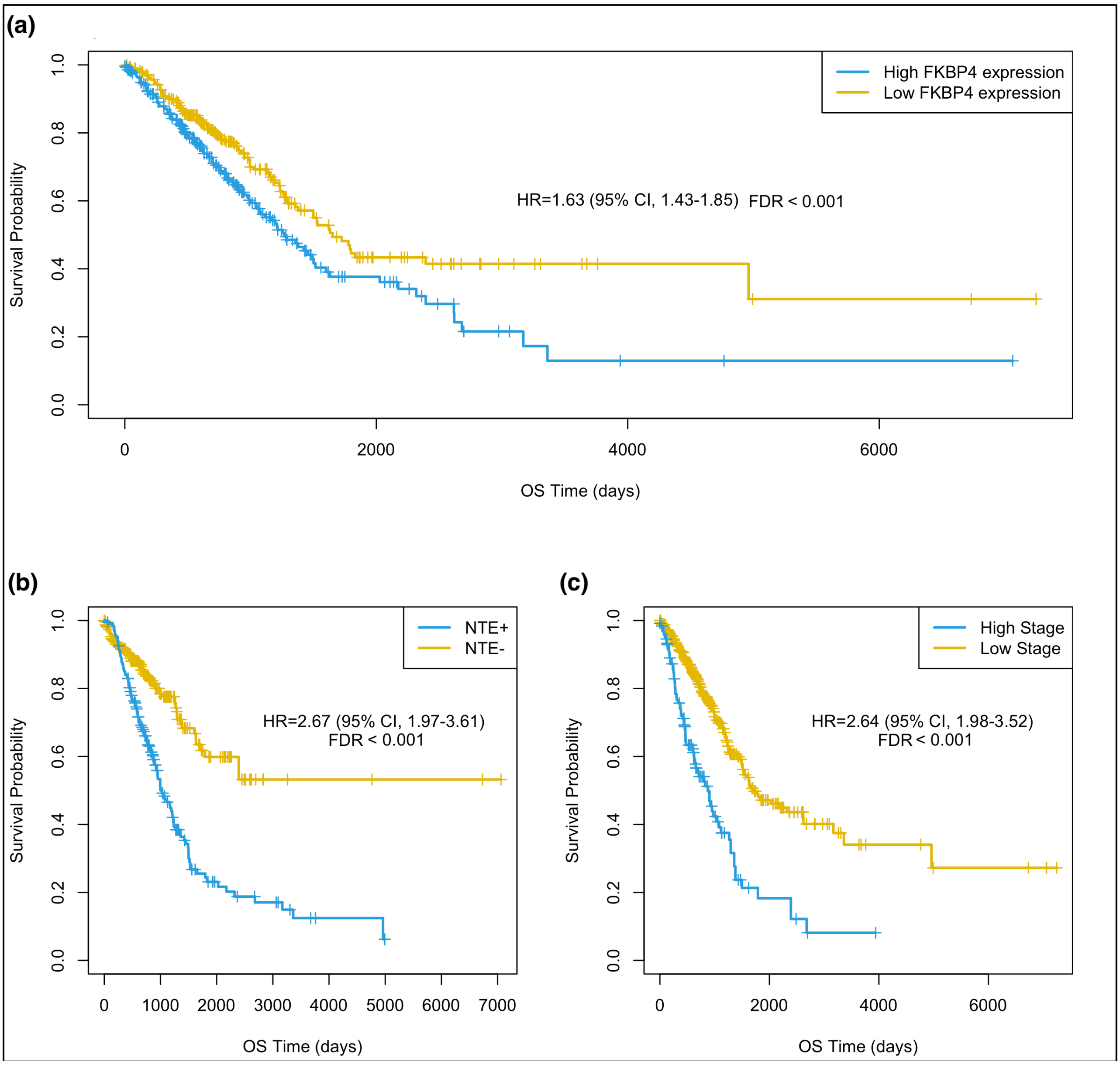


Figure 1. The Kaplan-Meier survival curves. (a) Patients with high expression of FKBP4 have an increased death rate by 63% (HR = 1.63 (95% CI, 1.43-1.85), FDR < 0.001); (b) NTE or new tumour event after initial treatment. NTE+ (NTE-positive) patients showed lower survival rate (HR = 2.67 (95% CI, 1.97-3.61), FDR < 0.001) compared to NTE- (NTE-negative); (c) Patients with advanced stages of LUAD (stage III-IV) has at least 2.6-fold higher risk of death compared to lower stages (I-II) (HR = 2.64 (95% CI, 1.98-3.52), FDR < 0.001).

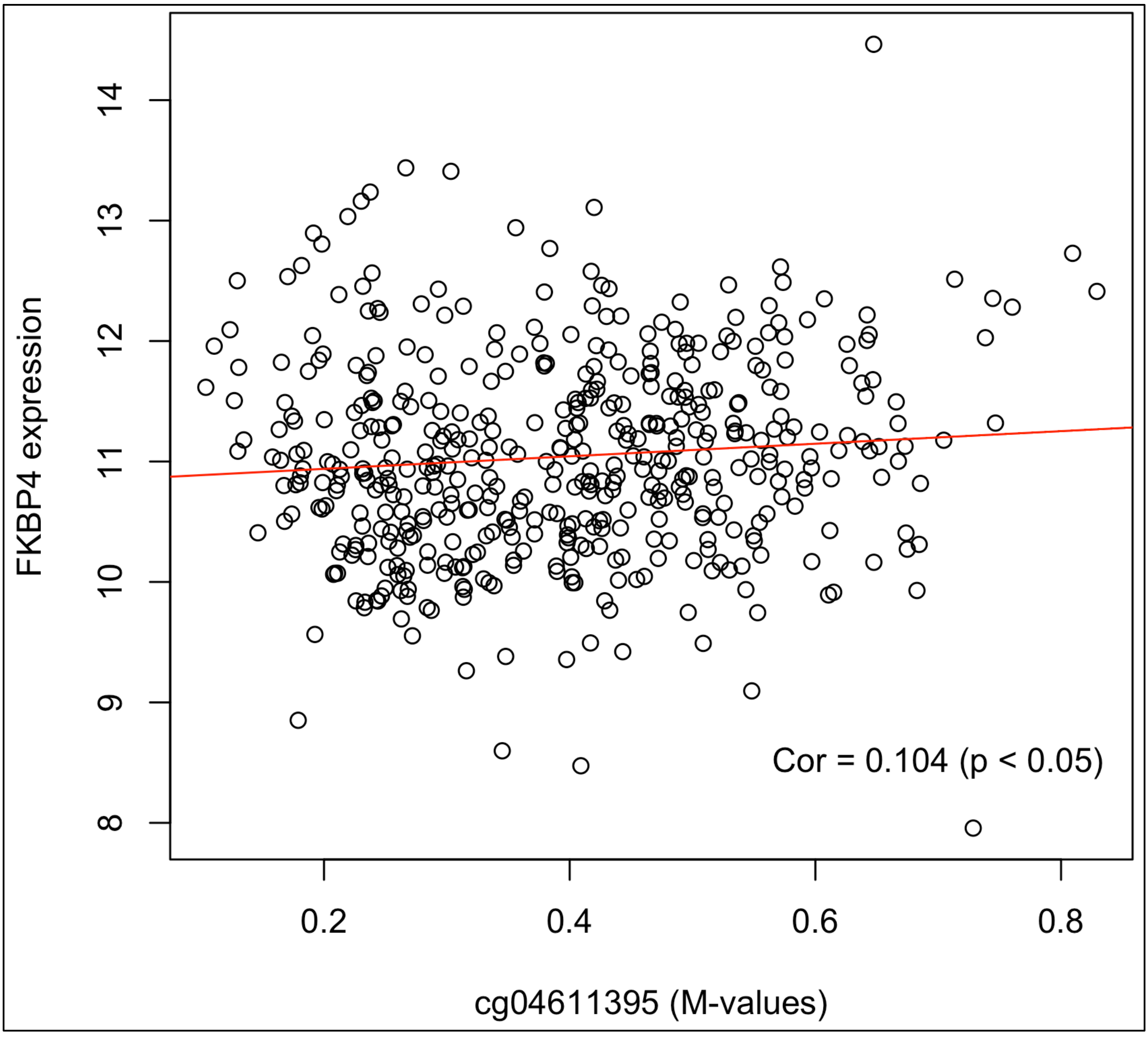


Figure 2. Correlation of FKBP4 expression and TSS200 CpG site cg04611395. The mean methylation beta-value of the group with high FKBP4 expression was 0.0383 higher relative to the group with low FKBP4 expression group.

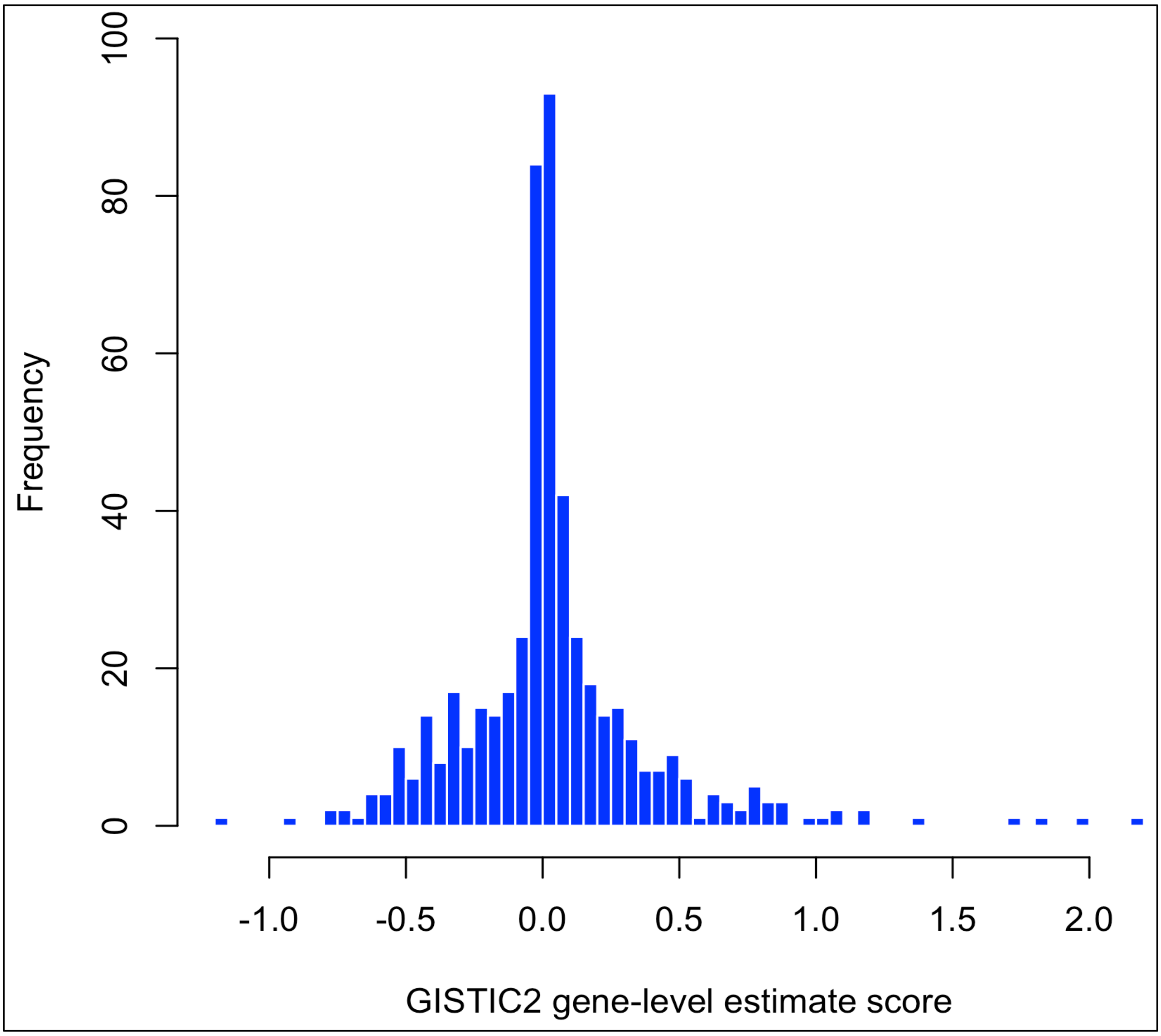


Figure 3. FKBP4 CNV histogram. Copy gene number of FKBP4 was estimated using GISTIC2 method. The copy number of FKBP4 was based on the median score. FKBP4 copy number was determined to be diploid normal copy.

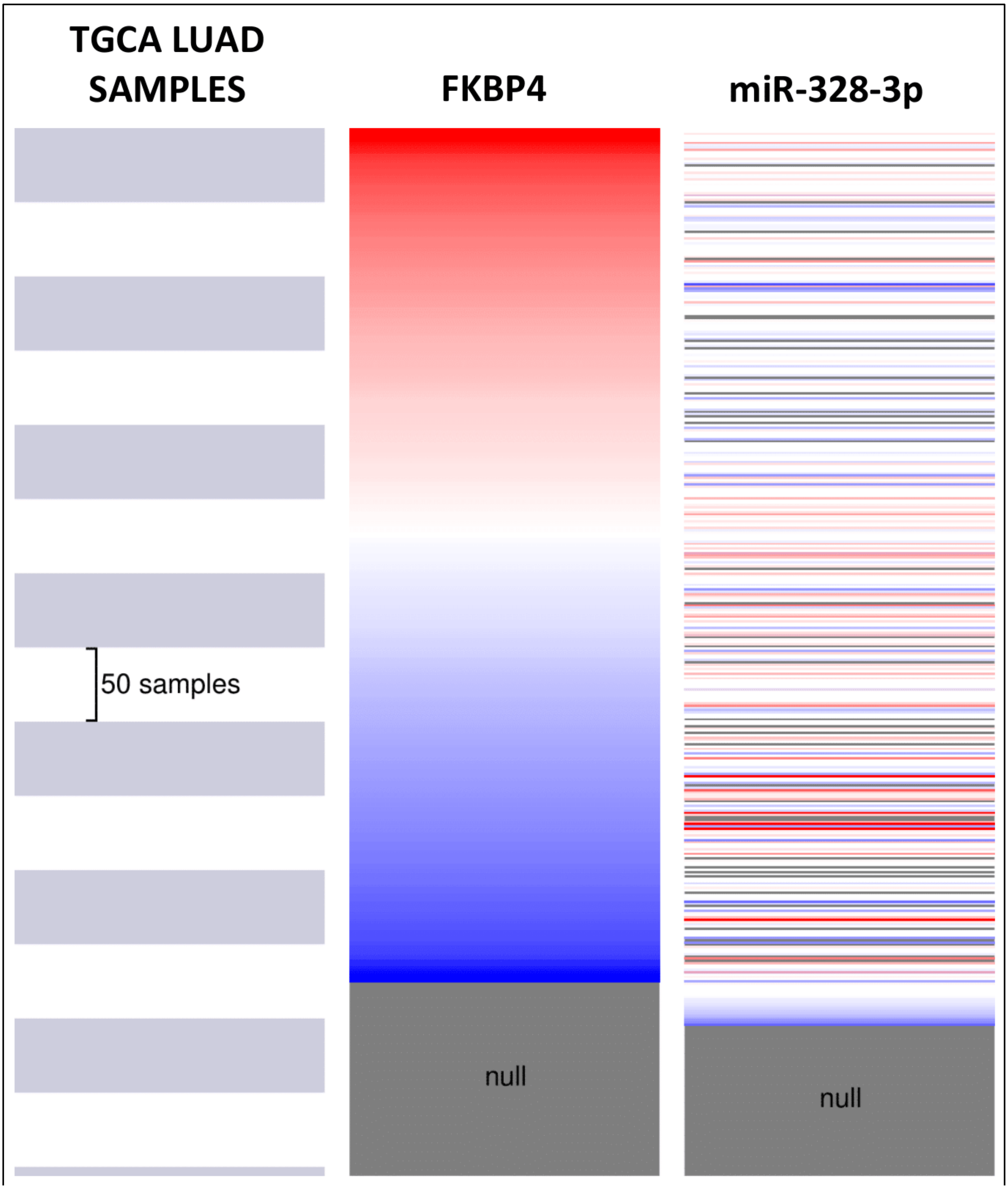


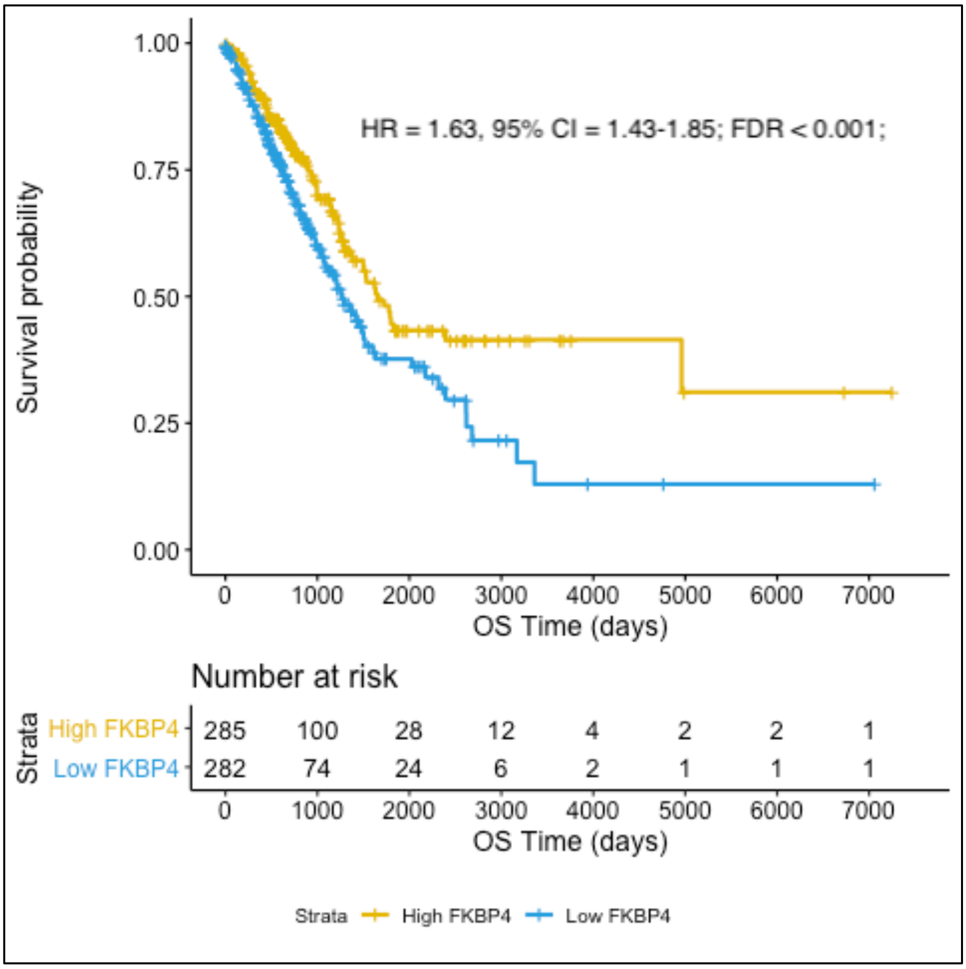
Figure 4. miR-328-3p level and FKBP4 expression in TGCA LUAD dataset heatmap. Patients with low FKBP4 expression have a general increase in miR-328-3p level.

Appendix I – Supplementary Data

Supplementary Data 1: Top 10 hits in RPPA data and survival

Genes	HR	LCI	UCI	PVAL	FDR
CD49B	2.525966	1.455195	4.384638	0.000991	0.129758
PAI1	1.469763	1.149141	1.879842	0.002161	0.141551
CYCLINB1	1.387614	1.106266	1.740516	0.004604	0.20105
CKIT	0.686119	0.485953	0.968733	0.032318	0.839184
ERALPHA	0.494518	0.230582	1.060568	0.070466	0.839184
ERALPHAPS118	0.382791	0.147548	0.993093	0.048357	0.839184
FIBRONECTIN	1.724289	0.960563	3.095237	0.067975	0.839184
GATA3	1.897208	0.99306	3.624554	0.052516	0.839184
KU80	1.925999	0.970653	3.821626	0.060825	0.839184
P70S6KPT389	0.409727	0.15671	1.07125	0.068819	0.839184

Supplementary Data 2: Kaplan-Meier curve - FKBP4 expression correlation with survival and risk ratio table.



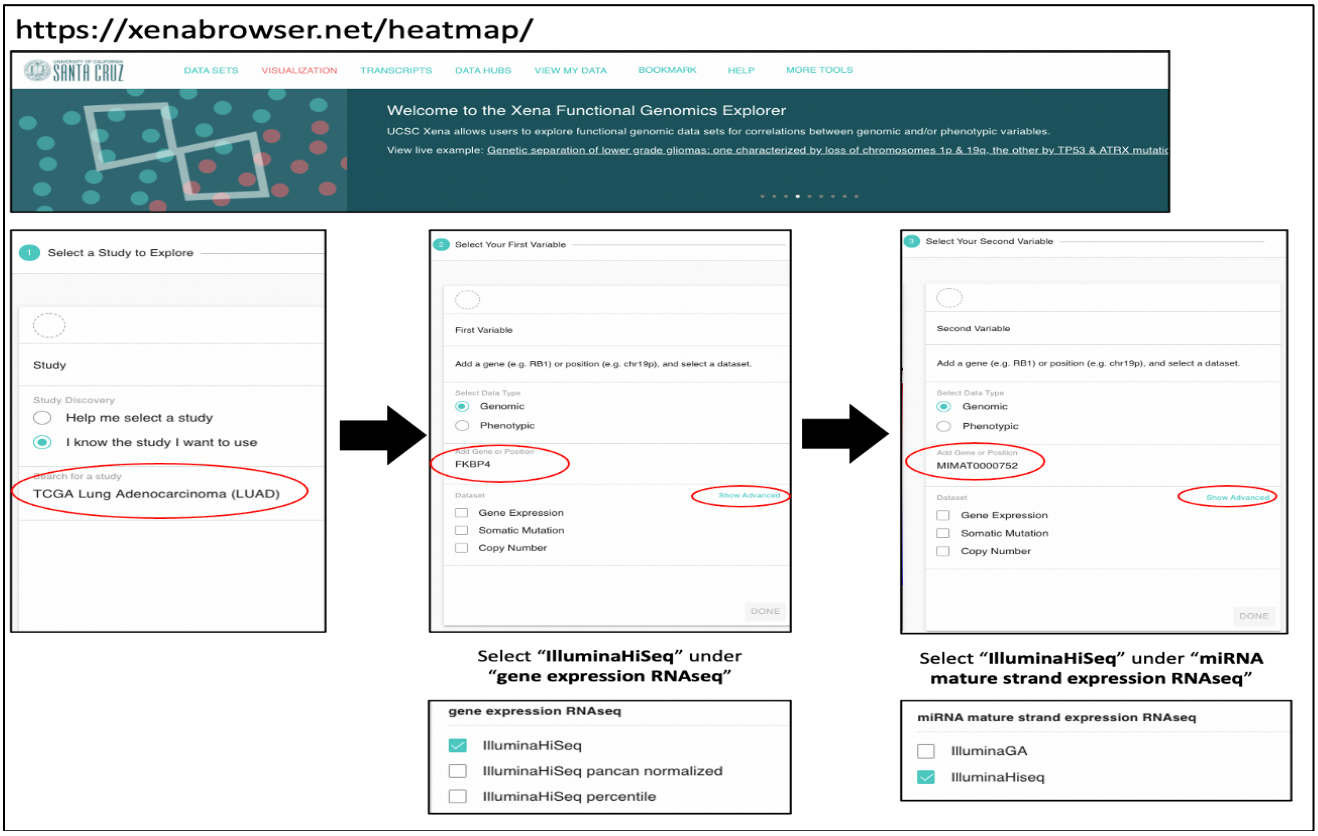
Supplementary Data 3: DNA methylation (Beta- and M-Values) and RNA expression Spearman correlation

			Means (Beta-value)		**Mean (M-values)**		**Absolute Difference in High Vs Low FKBP4**	
CpG Sites	**Spearman R**	**P-value**	**High FKBP4**	**Low FKBP4**	**High FKBP4**	**Low FKBP4**	**Beta-Values**	**M-Values**
cg00779206	0.1450	0.0015	0.7123	0.6982	1.3558	1.2437	0.0141	0.1120
cg00862618	-0.0620	0.1761	0.0249	0.0284	-5.4970	-5.4178	0.0035	0.0793
cg00970015	-0.0942	0.0398	0.0732	0.0776	-3.6946	-3.6410	0.0044	0.0537
cg01044331	-0.0627	0.1713	0.1090	0.1128	-3.0442	-3.0117	0.0038	0.0325
cg01601306	-0.0533	0.2451	0.0259	0.0283	-Inf	-5.2878	0.0023	Inf
cg02238069	0.0098	0.8315	0.1396	0.1385	-2.6424	-2.6650	0.0011	0.0225
cg03310242	0.0293	0.5234	0.3697	0.3256	-0.9985	-1.2562	0.0441	0.2578
cg04611395	0.1044	0.0226	0.4122	0.3738	-0.5767	-0.7979	0.0383	0.2212
cg04915277	-0.0011	0.9806	0.0303	0.0319	-Inf	-5.2007	0.0016	Inf
cg06401966	-0.0430	0.3488	0.0186	0.0216	-Inf	-5.8897	0.0030	Inf
cg08501815	-0.0529	0.2485	0.0141	0.0187	-6.3702	-6.2298	0.0046	0.1404
cg09446995	-0.0164	0.7206	0.0286	0.0289	-5.2338	-5.2746	0.0003	0.0407
cg11518240	0.0075	0.8708	0.8760	0.8812	3.0064	3.0471	0.0052	0.0407
cg13846563	-0.0974	0.0335	0.0172	0.0228	-Inf	-5.8520	0.0056	Inf
cg15260466	-0.0665	0.1469	0.0177	0.0222	-Inf	-Inf	0.0045	NaN
cg18776056	-0.0336	0.4635	0.0578	0.0610	-4.0728	-4.0372	0.0032	0.0356

Supplementary Data 4a: FKBP4 miRNA from GeneCards and accession from miRBase

miRBase ID (GeneCards)	Accession	miRBase ID	Accession
hsa-miR-760 (MIRT036772)	MIMAT0004957	hsa-miR-4476 (MIRT745039)	MIMAT0019003
hsa-miR-877-3p (MIRT037008)	MIMAT0004950	hsa-miR-4519 (MIRT745785)	MIMAT0019056
hsa-miR-744-5p (MIRT037530)	MIMAT0004945	hsa-miR-4710 (MIRT747398)	MIMAT0019815
hsa-miR-423-5p (MIRT038093)	MIMAT0004748	hsa-miR-4787-5p (MIRT748889)	MIMAT0019956
hsa-miR-99b-3p (MIRT038526)	MIMAT0004678	hsa-miR-5197-5p (MIRT750101)	MIMAT0021131
hsa-miR-615-3p (MIRT039622)	MIMAT0003283	hsa-miR-5680 (MIRT751130)	MIMAT0022468
hsa-miR-92b-3p (MIRT040589)	MIMAT0003218	hsa-miR-6750-5p (MIRT754571)	MIMAT0027400
hsa-miR-18a-3p (MIRT040812)	MIMAT0002891	hsa-miR-6752-5p (MIRT754632)	MIMAT0027404
hsa-miR-484 (MIRT041945)	MIMAT0002174	hsa-miR-6822-5p (MIRT756903)	MIMAT0027544
hsa-miR-328-3p (MIRT043773)	MIMAT0000752	hsa-miR-6842-5p (MIRT757366)	MIMAT0027586
hsa-miR-149-5p (MIRT045461)	MIMAT0000450	hsa-miR-6876-5p (MIRT758149)	MIMAT0027652
hsa-miR-197-3p (MIRT048140)	MIMAT0000227	hsa-miR-7110-5p (MIRT758992)	MIMAT0028117
hsa-miR-92a-3p (MIRT049584)	MIMAT0000092	hsa-miR-7515 (MIRT759379)	MIMAT0029310
hsa-miR-23a-3p (MIRT050423)	MIMAT0004496	hsa-miR-3915 (MIRT742882)	MIMAT0018189
hsa-miR-1253 (MIRT736538)	MIMAT0005904	hsa-miR-3928-3p (MIRT743038)	MIMAT0018205
hsa-miR-3202 (MIRT741143)	MIMAT0015089	hsa-miR-4274 (MIRT743661)	MIMAT0016906

Supplementary Data 4b: miRNA and RNA expression heatmap in XenaBrowser – Workflow



https://xenabrowser.net/heatmap/

Welcome to the Xena Functional Genomics Explorer

UCSC Xena allows users to explore functional genomic data sets for correlations between genomic and/or phenotypic variables.

View live example: Genetic separation of lower grade gliomas: one characterized by loss of chromosomes 1p & 19q, the other by TP53 & ATRX mutatic

1 Select a Study to Explore

Study

Study Discovery

○ Help me select a study

● I know the study I want to use

Search for a study: TCGA Lung Adenocarcinoma (LUAD)

2 Select Your First Variable

First Variable

Add a gene (e.g. RB1) or position (e.g. chr19p), and select a dataset.

Select Data Type

● Genomic

○ Phenotypic

Add Gene or Position: FKBP4

Dataset — Show Advanced

□ Gene Expression

□ Somatic Mutation

□ Copy Number

DONE

3 Select Your Second Variable

Second Variable

Add a gene (e.g. RB1) or position (e.g. chr19p), and select a dataset.

Select Data Type

● Genomic

○ Phenotypic

Add Gene or Position: MIMAT0000752

Dataset — Show Advanced

□ Gene Expression

□ Somatic Mutation

□ Copy Number

DONE

Select “**IlluminaHiSeq**” under “**gene expression RNAseq**”

gene expression RNAseq

☑ IlluminaHiSeq

□ IlluminaHiSeq pancan normalized

□ IlluminaHiSeq percentile

Select “**IlluminaHiSeq**” under “**miRNA mature strand expression RNAseq**”

miRNA mature strand expression RNAseq

□ IlluminaGA

☑ IlluminaHiseq

Appendix II – Pipeline and Data source

Data sources:

Dataset	Source (TGCA LUAD)
Copy number	https://tcga.xenahubs.net/download/TCGA.LUAD.sampleMap/Gistic2_CopyNumber_Gistic2_all_data_by_genes.gz
DNA methylation	https://tcga.xenahubs.net/download/TCGA.LUAD.sampleMap/HumanMethylation450.gz
Gene expression	https://tcga.xenahubs.net/download/TCGA.LUAD.sampleMap/HiSeqV2.gz
Curated survival data	https://tcga.xenahubs.net/download/survival/LUAD_survival.txt.gz
Clinical data	https://tcga.xenahubs.net/download/TCGA.LUAD.sampleMap/LUAD_clinicalMatrix
Protein expression	https://tcga.xenahubs.net/download/TCGA.LUAD.sampleMap/RPPA_RBN.gz
DNA methylation annotation	Imperial College BRC Server: /data/seqtools/CancerInformaticsData/day6/annot450k.rds

```
# Download data from the sources and unzip

# R version 1.3.1093.

# Load packages
library("survival")
library("survminer")
library("dplyr")

# Load data
rna.seq <- as.matrix(read.table("HiSeqV2", sep="\t",header=TRUE,row.names = 1))
survival <- read.table("LUAD_survival.txt ", sep="\t",header=TRUE, row.names = 1 )
clin.data <- read.table("LUAD_clinicalMatrix", sep= "\t", header=T, row.names=1)
prot.exp <- as.matrix (read.table("RPPA_RBN ", sep="\t", header=T, row.names=1))
dna.meth <- read.table("HumanMethylation450", sep="\t", header=T, row.names=1)
CNV <-
as.matrix(read.table("Gistic2_CopyNumber_Gistic2_all_data_by_genes",sep="\t",head=T,row.names=1))

# Loading DNA Methylation annotation
annot.dna.meth <- readRDS("annot450k.rds") # annotation was taken from Imperial College BRC Server

# Fixing PX ID format
rownames(clin.data)<-gsub(rownames(clin.data), pattern="-", replace=".")
rownames(survival)<-gsub(rownames(survival), pattern="-", replace=".")
```

```
# RPPA and survival
OS.Time.prot <- survival[colnames(prot.exp),"OS.time"]
OS.Event.prot <- as.numeric(survival[colnames(prot.exp),"OS"])
OS.prot <- Surv(OS.Time.prot,OS.Event.prot)

# RRPA cox regression
Results.OS_prot.exp<- array(NA, c(nrow(prot.exp),4))
colnames(Results.OS_prot.exp)<-c("HR","LCI","UCI","PVAL")
rownames(Results.OS_prot.exp)<-rownames(prot.exp)
Results.OS_prot.exp <- as.data.frame(Results.OS_prot.exp)

for(i in 1:nrow(prot.exp)){
  coxphmodel2 <- coxph(OS.prot~ as.numeric(prot.exp[i,]))
  Results.OS_prot.exp$HR[i] <- summary(coxphmodel2)$coef[1,2]
  Results.OS_prot.exp$LCI[i] <- summary(coxphmodel2)$conf.int[1,3]
  Results.OS_prot.exp$UCI[i] <- summary(coxphmodel2)$conf.int[1,4]
  Results.OS_prot.exp$PVAL[i] <- summary(coxphmodel2)$coef[1,5]
}

# Adjusting for multiple testing using FDR method
Results.OS_prot.exp$FDR <- p.adjust(Results.OS_prot.exp$PVAL,method="fdr")
Results.OS_prot.exp<-Results.OS_prot.exp[order(Results.OS_prot.exp$FDR, decreasing=F),]

# Check RPPA data with FDR < 0.05
Results.OS_prot.exp   # no protein were statistically significant after adjusting the p-values


###


# RNAseq and survival
OS.Time.rna <- survival[colnames(rna.seq),"OS.time"]
OS.Event.rna <- as.numeric(survival[colnames(rna.seq),"OS"])
OS.rna <- Surv(OS.Time.rna,OS.Event.rna)

# RNAseq cox regression
Results.OS_rna.seq<- array(NA, c(nrow(rna.seq),4))
colnames(Results.OS_rna.seq)<-c("HR","LCI","UCI","PVAL")
rownames(Results.OS_rna.seq)<-rownames(rna.seq)
Results.OS_rna.seq <- as.data.frame(Results.OS_rna.seq)

for(i in 1:nrow(rna.seq)){
  coxphmodel <- coxph(OS.rna~ as.numeric(rna.seq[i,]))
  Results.OS_rna.seq$HR[i] <- summary(coxphmodel)$coef[1,2]
  Results.OS_rna.seq$LCI[i] <- summary(coxphmodel)$conf.int[1,3]
  Results.OS_rna.seq$UCI[i] <- summary(coxphmodel)$conf.int[1,4]
  Results.OS_rna.seq$PVAL[i] <- summary(coxphmodel)$coef[1,5]
}

# Adjusting for multiple testing using FDR method
Results.OS_rna.seq$FDR <- p.adjust(Results.OS_rna.seq$PVAL,method="fdr")
Results.OS_rna.seq<-Results.OS_rna.seq[order(Results.OS_rna.seq$FDR, decreasing=F),]

# Check gene data with FDR < 0.05
Results.OS_rna.seq # statiscally significant genes were detected - proceed to univariate analysis

# Percentage of genes that have FDR < 0.05
AA <- length(which(Results.OS_rna.seq[,5] < 0.05))
Aa <- length(Results.OS_rna.seq[,5])
((AA/Aa)*100) # 6.02%

# RNASeq multivariate cox regression
clin.data <- clin.data[colnames(rna.seq),]

# confounding factors from clinical data
# Gender
gender <- rep(NA, nrow(clin.data))
gender[clin.data$gender=="FEMALE"] <- 0
gender[clin.data$gender=="MALE"] <- 1

# age
age<-as.numeric(clin.data$age_at_initial_pathologic_diagnosis)

# histological type
histology <- as.factor(clin.data$histological_type)

# anatomic neoplasm subdivisioon
neoplasm <- as.factor(clin.data$anatomic_neoplasm_subdivision)

# location
location <- as.factor(clin.data$location_in_lung_parenchyma)

# Pathologic stage
stage.III <- grep("III", clin.data$pathologic_stage)
stage.IV <- grep("IV", clin.data$pathologic_stage)
stage.high <- rep(0, nrow((clin.data)))
stage.high [c(stage.III,stage.IV)] <- 1

# Smoking
# 0 = non-smokers, 1 = active smokers and Current reformed smoker for <, > or = 15 years
smoke.pack <-clin.data$number_pack_years_smoked
smoke.pack[which(clin.data$tobacco_smoking_history==1)]<-0
smoke <-rep(NA,nrow(clin.data))
smoke[which(clin.data$tobacco_smoking_history==1)]<-0
smoke[which(clin.data$tobacco_smoking_history>1)]<-1

# new tumour event prior initial treatment (NTE)
NTE <- rep(NA, nrow(clin.data))
NTE[clin.data$new_tumor_event_after_initial_treatment=="YES"] <- 1
NTE[clin.data$new_tumor_event_after_initial_treatment=="NO"] <- 0

# PX received target molecular therapy
```

```
targeted.therapy <-rep(NA,nrow(clin.data))
targeted.therapy[which(clin.data$targeted_molecular_therapy=="NO")]<-0
targeted.therapy[which(clin.data$targeted_molecular_therapy=="YES")]<-1

# Summary of confounding factors
summary(coxph(OS.rna ~ gender))$coef              # no statistical difference
summary(coxph(OS.rna ~ age))$coef                 # no statistical difference
summary(coxph(OS.rna ~ histology))$coef           # no statistical difference
summary(coxph(OS.rna ~ neoplasm))$coef             # no statistical difference
summary(coxph(OS.rna ~ stage.high))$coef           # SIGNIFICANT (p < 3.378704e-11)
summary(coxph(OS.rna ~ smoke))$coef               # no statistical difference
summary(coxph(OS.rna ~ smoke.pack))$coef           # no statistical difference
summary(coxph(OS.rna ~ NTE))$coef                 # SIGNIFICANT (p < 2.614033e-10)
summary(coxph(OS.rna ~ targeted.therapy))$coef     # no statistical difference
summary(coxph(OS.rna ~ location))$coef            # no statistical difference

# NTE cox regression
coxph(OS.rna ~ NTE)
Results.OS_NTE<-array(NA, c(nrow(rna.seq),4))
colnames(Results.OS_NTE)<-c("HR","LCI","UCI","PVAL")
rownames(Results.OS_NTE)<-rownames(rna.seq)
Results.OS_NTE<-as.data.frame(Results.OS_NTE)

for(i in 1:nrow(rna.seq)){
  coxphmodel_NTE <- coxph(OS.rna ~ NTE)
  Results.OS_NTE$HR[i] <- summary(coxphmodel_NTE)$coef[1,2]
  Results.OS_NTE$LCI[i] <- summary(coxphmodel_NTE)$conf.int[1,3]
  Results.OS_NTE$UCI[i] <- summary(coxphmodel_NTE)$conf.int[1,4]
  Results.OS_NTE$PVAL[i] <- summary(coxphmodel_NTE)$coef[1,5]
}
Results.OS_NTE$FDR <- p.adjust(Results.OS_NTE$PVAL, method =  "fdr")

summary(coxphmodel_NTE)$coef[1,2]       # HR = 2.666519
summary(coxphmodel_NTE)$conf.int[1,3]   # LCI = 1.967217
summary(coxphmodel_NTE)$conf.int[1,4]   # HCI= 3.614407
summary(coxphmodel_NTE)$coef[1,5]       # p = 2.614033e-10; FDR = 2.614033e-10


# Stage cox regression
coxph(OS.rna ~ stage.high)
Results.OS_stage<-array(NA, c(nrow(rna.seq),4))
colnames(Results.OS_stage)<-c("HR","LCI","UCI","PVAL")
rownames(Results.OS_stage)<-rownames(rna.seq)
Results.OS_stage<-as.data.frame(Results.OS_stage)

for(i in 1:nrow(rna.seq)){
  coxphmodel_stage <- coxph(OS.rna ~ stage.high)
  Results.OS_stage$HR[i] <- summary(coxphmodel_stage)$coef[1,2]
  Results.OS_stage$LCI[i] <- summary(coxphmodel_stage)$conf.int[1,3]
  Results.OS_stage$UCI[i] <- summary(coxphmodel_stage)$conf.int[1,4]
  Results.OS_stage$PVAL[i] <- summary(coxphmodel_stage)$coef[1,5]
}
Results.OS_stage$FDR <- p.adjust(Results.OS_stage$PVAL, method =  "fdr")

summary(coxphmodel_stage)$coef[1,2]       # HR = 2.643894
summary(coxphmodel_stage)$conf.int[1,3]   # LCI = 1.983362
summary(coxphmodel_stage)$conf.int[1,4]   # HCI =3.524406
summary(coxphmodel_stage)$coef[1,5]       # p = 3.378704e-11; FDR = 3.378704e-11


# RNASeq multivariate cox regression
# Only accounted for stage.high and new.tumour.pre
# Did not adjust for other factors
Results.OS_rna.seq_fctrs<-array(NA, c(nrow(rna.seq),4))
colnames(Results.OS_rna.seq_fctrs)<-c("HR","LCI","UCI","PVAL")
rownames(Results.OS_rna.seq_fctrs)<-rownames(rna.seq)
Results.OS_rna.seq_fctrs<-as.data.frame(Results.OS_rna.seq_fctrs)

for(i in 1:nrow(rna.seq)){
  coxphmodel3 <- coxph(OS.rna ~ as.numeric(rna.seq[i,]+stage.high+NTE))
  Results.OS_rna.seq_fctrs$HR[i] <- summary(coxphmodel3)$coef[1,2]
  Results.OS_rna.seq_fctrs$LCI[i] <- summary(coxphmodel3)$conf.int[1,3]
  Results.OS_rna.seq_fctrs$UCI[i] <- summary(coxphmodel3)$conf.int[1,4]
  Results.OS_rna.seq_fctrs$PVAL[i] <- summary(coxphmodel3)$coef[1,5]
}

Results.OS_rna.seq_fctrs <- as.data.frame(Results.OS_rna.seq_fctrs) # Multivariate analysis

# Adjust for multiple testing using FDR
Results.OS_rna.seq_fctrs$FDR <- p.adjust(Results.OS_rna.seq_fctrs$PVAL, method =  "fdr")
Results.OS_rna.seq_fctrs<-Results.OS_rna.seq_fctrs[order(Results.OS_rna.seq_fctrs$FDR, decreasing=F),]

# Identify potential candidates
Results.OS_rna.seq_fctrs

# Identified FKBP4 to be a good novel candidate based on adj.p.vals and Pubmed Search
# Stratifying PX based on the expression of FKBP4 (median)
summary(rna.seq["FKBP4",])
Results.OS_rna.seq["FKBP4",]       # Univariate cox regression
Results.OS_rna.seq_fctrs["FKBP4",]  # Multivariate cox regression
FKBP4.high <- as.numeric(rna.seq["FKBP4",]>median(rna.seq["FKBP4",]))

# KM plot
png("KM_plots.png", width=9,height=9,units='in',res=300)
grid <- matrix(c(1,1,2,3), nrow = 2, ncol = 2, byrow = T)
layout(grid)
plot(survfit(OS.rna ~ FKBP4.high), col=c("#E7B800", "#2E9FDF"),lwd=2,mark.time=TRUE, xlab="OS Time
(days)", ylab="Survival Probability")
  legend("topright",legend=c("High FKBP4 expression","Low FKBP4
expression"),col=c("#2E9FDF","#E7B800"),lwd=2)
  text(4000,0.6,"HR=1.63 (95% CI, 1.43-1.85)")
  text(5100,0.6, "FDR")
  text(5500,0.6, "< 0.001")
plot(survfit(OS.rna ~ NTE), col=c("#E7B800", "#2E9FDF"),lwd=2,mark.time=TRUE, xlab="OS Time (days)",
ylab="Survival Probability")
  legend("topright",legend=c("NTE+", "NTE-"),col=c("#2E9FDF","#E7B800"),lwd=2)
  text(4800,0.7,"HR=2.67 (95% CI, 1.97-3.61)")
  text(4300,0.65, "FDR")
  text(5200,0.65, "< 0.001")
plot(survfit(OS.rna ~ stage.high), col=c("#E7B800", "#2E9FDF"),lwd=2,mark.time=TRUE, xlab="OS Time
(days)", ylab="Survival Probability")
  legend("topright",legend=c("High Stage", "Low Stage"),col=c("#2E9FDF","#E7B800"),lwd=2)
  text(4800,0.7,"HR=2.64 (95% CI, 1.98-3.52)")
  text(4300,0.65, "FDR")
  text(5200,0.65, "< 0.001")
dev.off()

# FKBP4 KM-plot: dependencies" survminer" and "dpylr"
png("KMplot_RiskTable.png")
ggsurvplot(survfit(Surv(OS.Time.rna, OS.Event.rna) ~ FKBP4.high, data = survival), conf.int = F, risk.table =
"absolute", risk.table.y.text.col = TRUE, palette = c("#E7B800", "#2E9FDF"),
        xlab = "OS Time (days)", legend = "bottom", xlim = c(0,7500), break.time.by = 1000, legend.labs =
c("High FKBP4", "Low FKBP4"))
dev.off()


##


# Higher FKBP4 expression results to poorer prognosis
# DNA Methylation of FKBP4
rna.seq.meth<-rna.seq[,which(is.element(colnames(rna.seq),colnames(dna.meth)))]
dna.meth2<- dna.meth[,which(is.element(colnames(dna.meth),colnames(rna.seq.meth)))]

# Align PX IDs
rna.seq.meth <- as.matrix(rna.seq.meth[,order(colnames(rna.seq.meth))])
dna.meth2 <- as.matrix(dna.meth2[,order(colnames(dna.meth2))])

# Check for methylated CpG site on FKBP4 gene
FKBP4.meth <- rownames(annot.dna.meth[which(annot.dna.meth$UCSC_RefGene_Name=="FKBP4"),])
FKBP4.meth

# Check annotation
annot.dna.meth[FKBP4.meth,]

# Filtering methylated CpG islands in FKBP4 gene
meth.data.FKBP4 <- dna.meth2[FKBP4.meth,]

# Exclusion CpG islands that were missing in 50% of the PX
NA.Count_FKBP4.meth<-apply(meth.data.FKBP4,1,function(x) sum(as.numeric(is.na(x))))
Exclude<-as.numeric(NA.Count_FKBP4.meth>0.5*ncol(meth.data.FKBP4))
meth.data.FKBP4<-meth.data.FKBP4[which(Exclude==0),]
```

```
# Correlation test between DNA methylation (beta-values) and RNAseq (log2(x+1), RBN): Spearman
correlation
Result.FKBP4.meth <- array(NA,c(nrow(meth.data.FKBP4),4))
rownames(Result.FKBP4.meth)<-rownames(meth.data.FKBP4)
colnames(Result.FKBP4.meth)<-c("Cor.FKBP4","Cor.test.FKBP4","Mean.high.FKBP4","Mean.low.FKBP4")
FKBP4.high.meth <- as.numeric(rna.seq.meth["FKBP4",]>median(rna.seq.meth["FKBP4",]))

for (i in 1:nrow(meth.data.FKBP4)){
  Result.FKBP4.meth [i,1]<-
cor.test(as.numeric(rna.seq.meth["FKBP4",]),as.numeric(meth.data.FKBP4[i,]),method="spearman",
use="c")$est
  Result.FKBP4.meth [i,2]<-
cor.test(as.numeric(rna.seq.meth["FKBP4",]),as.numeric(meth.data.FKBP4[i,]),method="spearman",
use="c")$p.value
}
Result.FKBP4.meth[,3]<-apply(meth.data.FKBP4[,which(FKBP4.high.meth==1)],1,mean,na.rm=T)
Result.FKBP4.meth[,4]<-apply(meth.data.FKBP4[,which(FKBP4.high.meth==0)],1,mean,na.rm=T)
Result.FKBP4.meth

# Beta-value - more intuitive in biological interpretion
# M-value - more statistically valid for the differential analysis of methylation levels.

# Produce an object for DNA Methylation M-values
# Convert beta-values to m-values. Formula: mvalues = log2(x/(1-x))
meth.data.FKBP4.Mvals <- apply(meth.data.FKBP4, MARGIN = 2, function(x) log2(x/(1-x)))

# Correlation test between DNA methylation (M-values) and RNAseq (log2(x+1), RBN): Spearman
correlation
Result.FKBP4.meth.MVals <- array(NA,c(nrow(meth.data.FKBP4.Mvals),4))
rownames(Result.FKBP4.meth.MVals)<-rownames(meth.data.FKBP4.Mvals)
colnames(Result.FKBP4.meth.MVals)<-
c("Cor.FKBP4","Cor.test.FKBP4","Mean.high.FKBP4","Mean.low.FKBP4")
FKBP4.high.meth.Mvals <- as.numeric(rna.seq.meth["FKBP4",]>median(rna.seq.meth["FKBP4",]))

for (i in 1:nrow(meth.data.FKBP4.Mvals)){
  Result.FKBP4.meth.MVals[i,1]<-
cor.test(as.numeric(rna.seq.meth["FKBP4",]),as.numeric(meth.data.FKBP4.Mvals[i,]),method="spearman",
use="c")$est
  Result.FKBP4.meth.MVals[i,2]<-
cor.test(as.numeric(rna.seq.meth["FKBP4",]),as.numeric(meth.data.FKBP4.Mvals[i,]),method="spearman",
use="c")$p.value
}
Result.FKBP4.meth.MVals[,3]<-
apply(meth.data.FKBP4.Mvals[,which(FKBP4.high.meth.Mvals==1)],1,mean,na.rm=T)
Result.FKBP4.meth.MVals[,4]<-
apply(meth.data.FKBP4.Mvals[,which(FKBP4.high.meth.Mvals==0)],1,mean,na.rm=T)


# Check for CpG island
Result.FKBP4.meth.MVals
Result.FKBP4.meth
```

```
# Identified "cg04611395" to be correlated with largest change in Beta-value abd M-values between groups of high and low FKBP4 expression
#              Cor.FKBP4 Cor.test.FKBP4 Mean.high.FKBP4 Mean.low.FKBP4
# cg04611395  0.104384099    0.022605706      0.41217710     0.37383766
# Difference in Beta-value  =  0.038
# Difference in M-value     =  0.22

# Most significant CpG Island =cg04611395
# Plots RRA Expression and DNA Methylation for FKBP4 gene
png("GeneExpVsMeth.png",width=6,height=6,units='in',res=300)
plot(as.numeric(meth.data.FKBP4["cg04611395",]),as.numeric(rna.seq.meth["FKBP4",]), xlab="cg04611395 (M-values)", ylab="FKBP4 expression")
abline(lm(rna.seq.meth["FKBP4",]~dna.meth2["cg04611395",]), col="red")
text(0.7, 8.5, "Cor = 0.104 (p < 0.05)")
dev.off()


##

# CNV and FKBP4
# Visually check the raw data
rna.seq.CNV<-rna.seq[,which(is.element(colnames(rna.seq),colnames(CNV)))]
CNV<- CNV[,which(is.element(colnames(CNV),colnames(rna.seq.CNV)))]

FKBP4.CNV <- CNV["FKBP4",]
summary(FKBP4.CNV)

# CNV plot
png("hist_CNV.png",width=6,height=6,units='in',res=300)
hist(FKBP4.CNV,xlab = "GISTIC2 gene-level estimate score", ylab = "Frequency",ylim = c(0,100), breaks=50 , col="blue", border=F,main="")
dev.off()
```