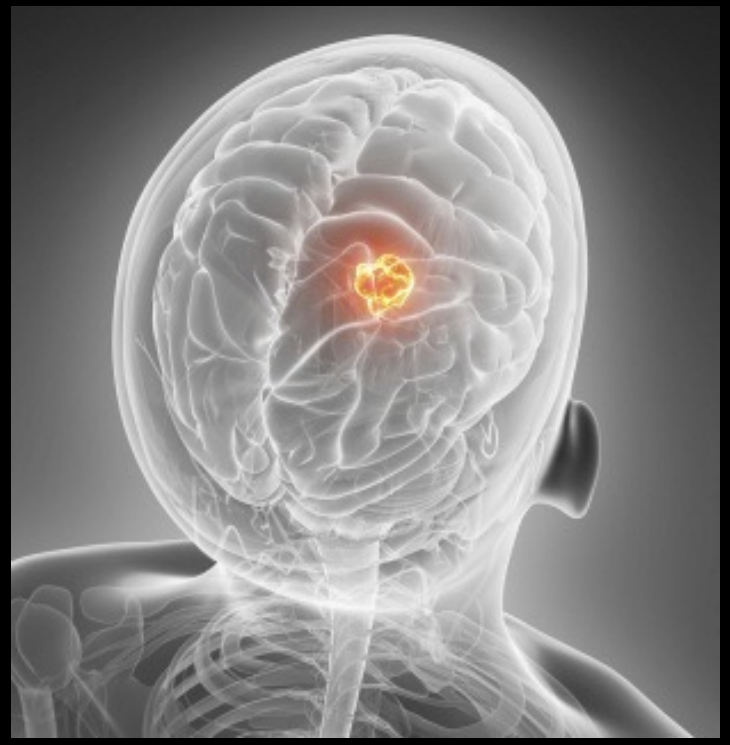


DATA201 GROUP WORK PRESENTATION

BRAIN TUMOUR

Zhuoheng Chen
zch70@uclive.ac.nz

Source of Data

UNIVERSITY OF CALIFORNIA SANTA CRUZ

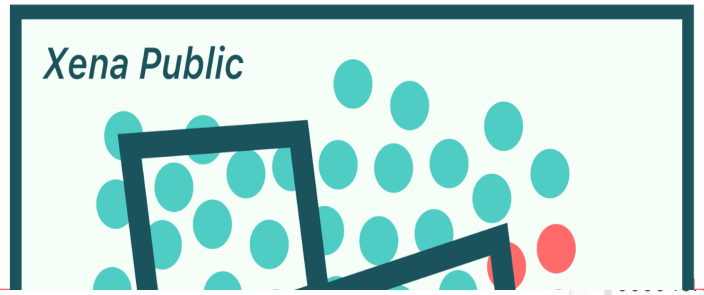
Overview Analysis Tutorials What's New Cite Us Subscribe

UCSC Xena

See the bigger picture

An online exploration tool for public and private, multi-omic and clinical/phenotype data

Launch Xena



UNIVERSITY OF CALIFORNIA SANTA CRUZ

DATA SETS VISUALIZATION TRANSCRIPTS DATA HUBS VIEW MY DATA HELP MORE TOOLS

129 Cohorts, 1571 Datasets

- Acute lymphoblastic leukemia (Mullighan 2008) (3 datasets)
- Breast Cancer (Caldas 2007) (3 datasets)
- Breast Cancer (Chin 2006) (3 datasets)
- Breast Cancer (Haverty 2008) (2 datasets)
- Breast Cancer (Hess 2006) (2 datasets)
- Breast Cancer (Miller 2005) (2 datasets)
- Breast Cancer (vantVeer 2002) (2 datasets)
- Breast Cancer (Vijver 2002) (2 datasets)
- Breast Cancer (Yau 2010) (2 datasets)
- Breast Cancer Cell Lines (Heiser 2012) (4 datasets)
- Breast Cancer Cell Lines (Neve 2006) (2 datasets)
- Cancer Cell Line Encyclopedia (Breast) (4 datasets)
- Cancer Cell Line Encyclopedia (CCLE) (9 datasets)
- Connectivity Map (2 datasets)
- GBM (Parsons 2008) (2 datasets)
- GDC MMRF-COMMPASS (6 datasets)
- GDC Pan-Cancer (PANCAN) (17 datasets)
- GDC TARGET-ALL-P3 (6 datasets)
- GDC TARGET-AML (6 datasets)
- GDC TARGET-CCSK (6 datasets)
- GDC TARGET-NBL (5 datasets)
- GDC TARGET-OS (6 datasets)
- GDC TARGET-RT (6 datasets)
- GDC TARGET-WT (6 datasets)

Active Data Hubs

- [x] My computer hub
- [x] UCSC Public Hub
- [x] TCGA Hub
- [x] Pan-Cancer Atlas Hub
- [x] ICGC Hub
- [x] PCAWG Hub
- [x] UCSC Toil RNA-seq Recompute
- [ ] Treehouse Hub
- [x] GDC Hub
- [x] ATAC-seq Hub
- [x] Kids First Hub
- [x] jupyter notebook

samples	173
version	07-19-2019
type of data	gene expression RNAseq
unit	log2(fpkm-uq+1)
platform	Illumina
ID/Gene Mapping	https://gdc-hub.s3.us-east-1.amazonaws.com/download/gencode.v22.annotation.gene.probeMap; Full metadata
author	Genomic Data Commons
raw data	https://docs.gdc.cancer.gov/Data/Release_Notes/Data_Release_Notes/#data-release-180
raw data	https://api.gdc.cancer.gov/data/

ENSG00000000460.15	14.76									
ENSG00000000938.11	16.5									
ENSG00000000971.14	17.7	15.05	17.11	16.85	16.67	17.3	17.04	18.77	16.8	15.86
ENSG00000001036.12	19.18	18.29	17.76	19.49	19.19	18.01	18.93	18.32	19.56	18.59
ENSG00000001084.9	16.86	15.82	18.48	15.88	17.38	17.57	15.97	17.18	16.96	16.87
ENSG00000001167.13	17.51	17.2	17.44	16.98	17.35	17.15	18.92	17.05	16.89	17.76

INITIAL CHALLENGE

- Understand the meaning of the data
- Figure out what I'm going to do and how

id	gene	chrom	chromStart	chromEnd	strand
ENSG00000223972.5	DDX11L1	chr1	11869	14409	+
ENSG00000227232.5	WASH7P	chr1	14404	29570	-
ENSG00000278267.1	MIR6859-3	chr1	17369	17436	-
ENSG00000243485.3	RP11-34P13.3	chr1	29554	31109	+
ENSG00000274890.1	MIR1302-9	chr1	30366	30503	+
ENSG00000237613.2	FAM138A	chr1	34554	36081	-
ENSG00000268020.3	OR4G4P	chr1	52473	53312	+
ENSG00000240361.1	OR4G11P	chr1	62948	63887	+
ENSG00000186092.4	OR4F5	chr1	69091	70008	+
ENSG00000238009.5	RP11-34P13.7	chr1	89295	133723	-
ENSG00000239945.1	RP11-34P13.8	chr1	89551	91105	-
ENSG00000233750.3	CICP27	chr1	131025	134836	+
ENSG00000268903.1	RP11-34P13.15	chr1	135141	135895	-
ENSG00000269981.1	RP11-34P13.16	chr1	137682	137965	-
ENSG00000239906.1	RP11-34P13.14	chr1	139790	140339	-
ENSG00000241860.5	RP11-34P13.13	chr1	141474	173862	-

Distinguish between normal cells and cancer cells

- Import data

```
exp <- read.table("TCGA-GBM.htseq_fpkm-uq.tsv", header=T,sep = "\t",row.names = 1)
ID <- read.table("gencode.v22.annotation.gene.probeMap", header=T, sep = "\t")
pdata <- data.frame(sample=colnames(exp),
                    group=ifelse(substr(colnames(exp),14,14)=="0","tumor","normal"))
head(exp)
```

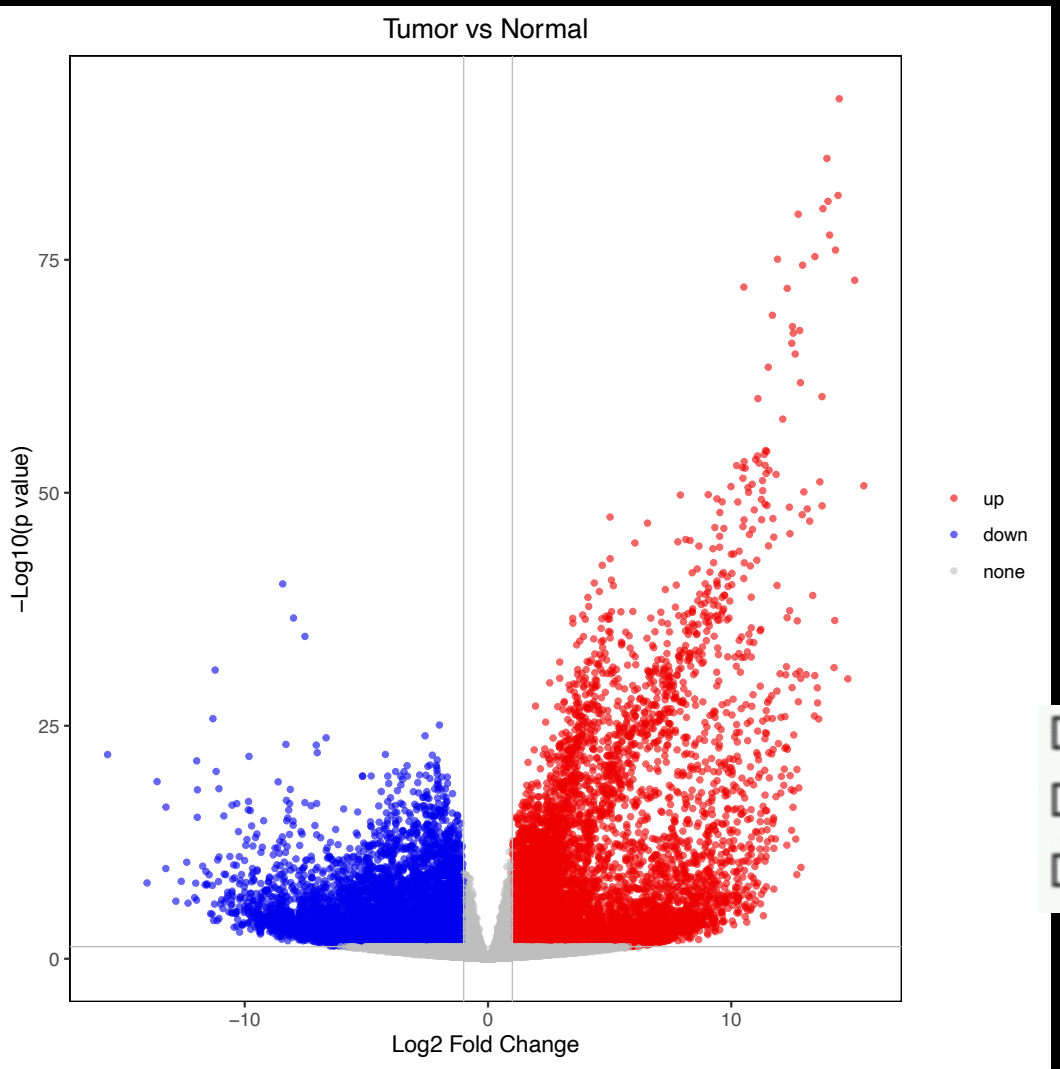
A data.frame: 6 × 173

	TCGA.06.0878.01A	TCGA.26.5135.01A	TCGA.06.5859.01A	TCGA.06.2563.01A	TCGA.27.1834.01A	TCGA.28.5
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	
ENSG00000242268.2	14.48288	12.02771	12.09509	13.75609	14.254343	
ENSG00000270112.3	10.86865	12.96473	13.23419	10.61799	8.653561	
ENSG00000167578.15	16.57162	16.75191	17.06651	17.04936	16.334255	

- In biology, the fourteenth digit is used to distinguish between normal cells and cancer cells

TCGA-06-0211-02A	TCGA-06-0675-11A	TCGA-06-0210-02A	TCGA-76-4927-01A
11.780789552475403	10.300762739967114	13.862896083699075	12.23319136333867
12.28175918031517	13.483760165373726	11.51638485744961	8.066946331297661
17.045247987824286	17.043956417124253	16.53286433807512	16.594590490052294

Differential genes expression analysis

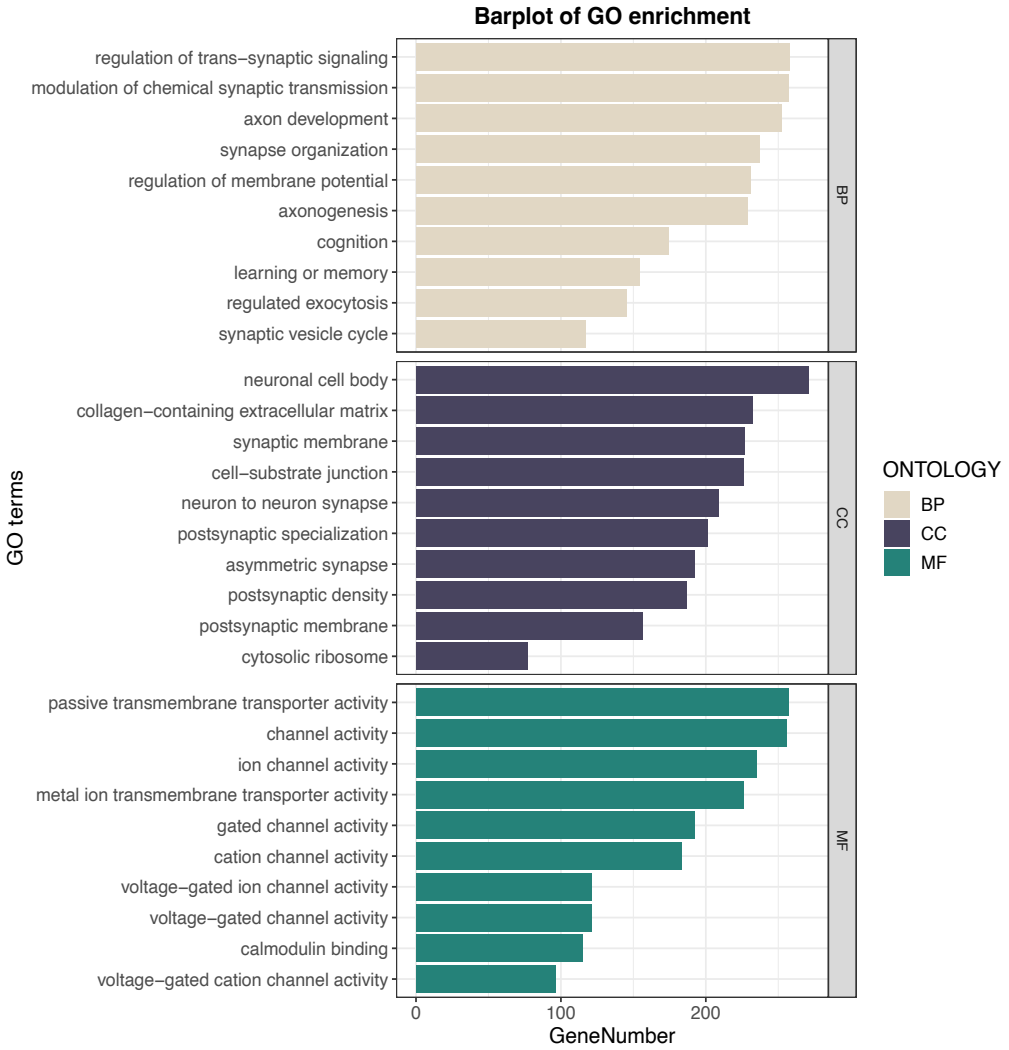


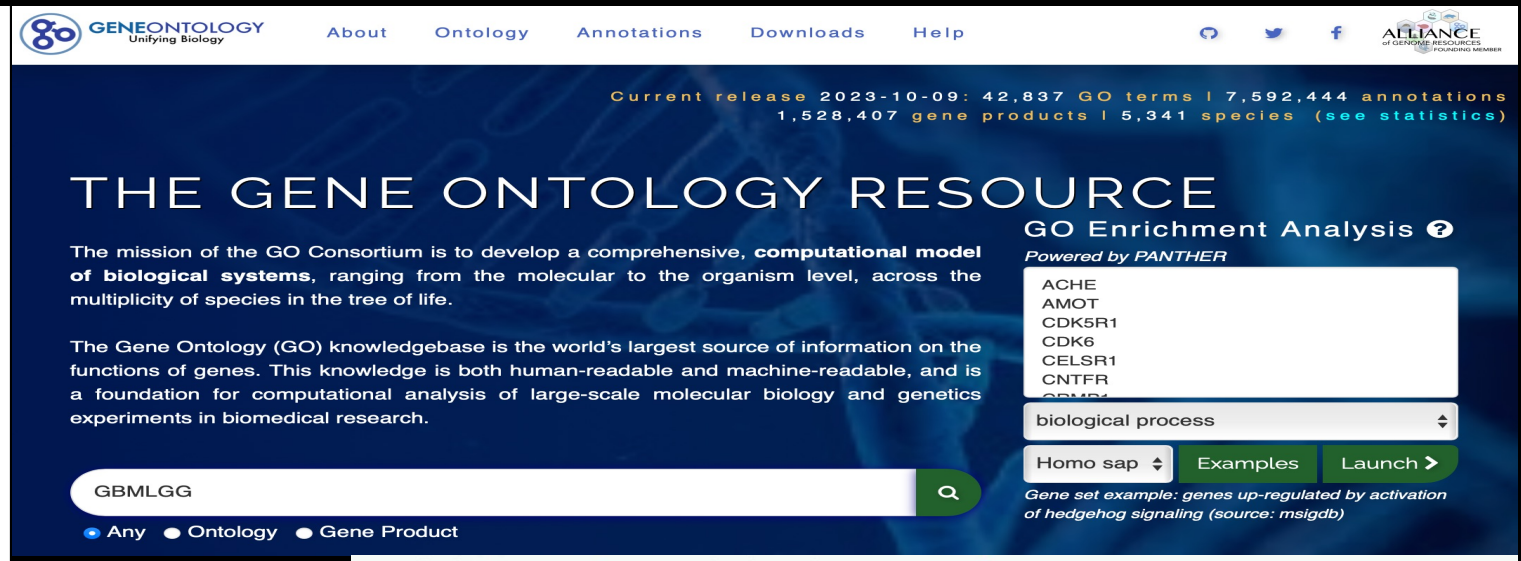
The ggplot2 package was used to create the volcano plot.
Volcano plots were used to demonstrate the significance of each gene in the gene expression data.
In this figure, up-regulated genes, down-regulated genes, and genes with no significant differences are shown in red, blue, and grey.

```
DEG[which(DEG$P.Value<0.05 & DEG$logFC <=-1),'sig']<-'down'
DEG[which(DEG$P.Value<0.05 & DEG$logFC >=1),'sig']<-'up'
DEG[which(DEG$P.Value>=0.05 | abs(DEG$logFC) < 1),'sig']<-'none'
```

GSEA- Gene Set Enrichment Analysis

Gene Ontology Enrichment Analysis



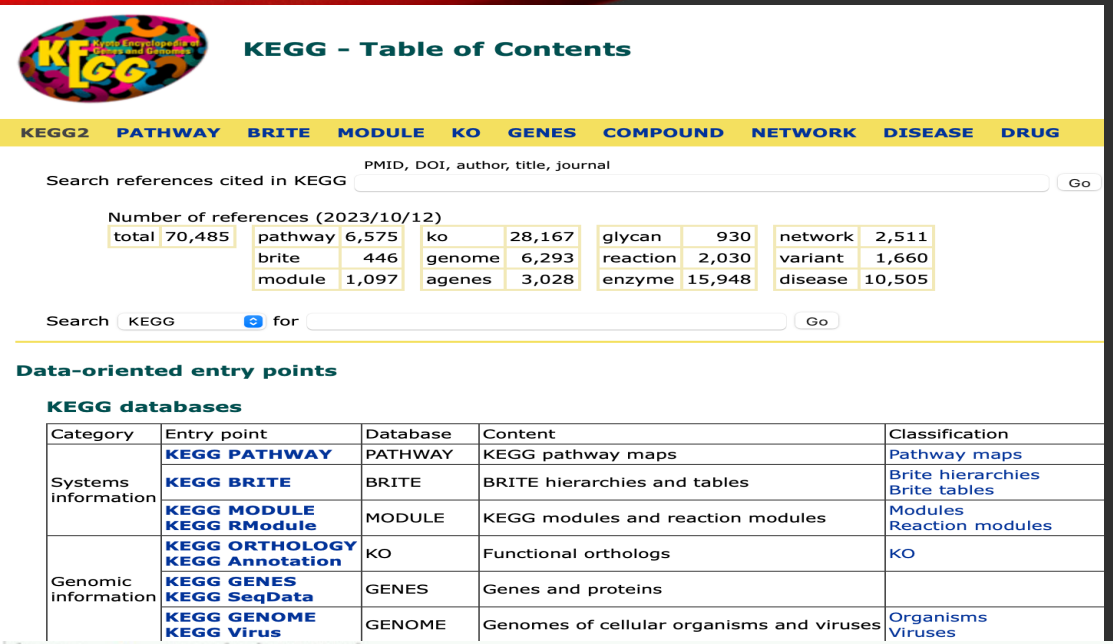


BP-Biological Process
MF- Molecular Function
CC-Cellular Component

```
enrichGO_res<- enrichGO(ID_trans$ENTREZID,
                        OrgDb = org.Hs.eg.db,
                        keyType = "ENTREZID",
                        ont = "ALL",
                        pvalueCutoff = 0.2,
                        pAdjustMethod = "BH",
                        qvalueCutoff = 0.2)
GO_res <- as.data.frame(enrichGO_res)
GO_res <- GO_res[GO_res$qvalue<0.05,]
GO <- GO_top(GO_res,10)
GO$foldEnrich <- enrichment_factor(GO$GeneRatio,GO$BgRatio)
```

KEGG (Kyoto Encyclopedia of Genes and Genomes) Enrichment Analysis

The main role of the KEGG model is to help us understand the high-level function and utility of biological systems such as cells, organisms, and ecosystems, how genes and proteins interact through complex networks and thus influence various biological processes within organisms



KEGG - Table of Contents

KEGG2 PATHWAY BRITE MODULE KO GENES COMPOUND NETWORK DISEASE DRUG

Search references cited in KEGG (PMID, DOI, author, title, journal) Go

Number of references (2023/10/12)

total	70,485	pathway	6,575	ko	28,167	glycan	930	network	2,511
		brite	446	genome	6,293	reaction	2,030	variant	1,660
		module	1,097	agenes	3,028	enzyme	15,948	disease	10,505

Search KEGG for Go

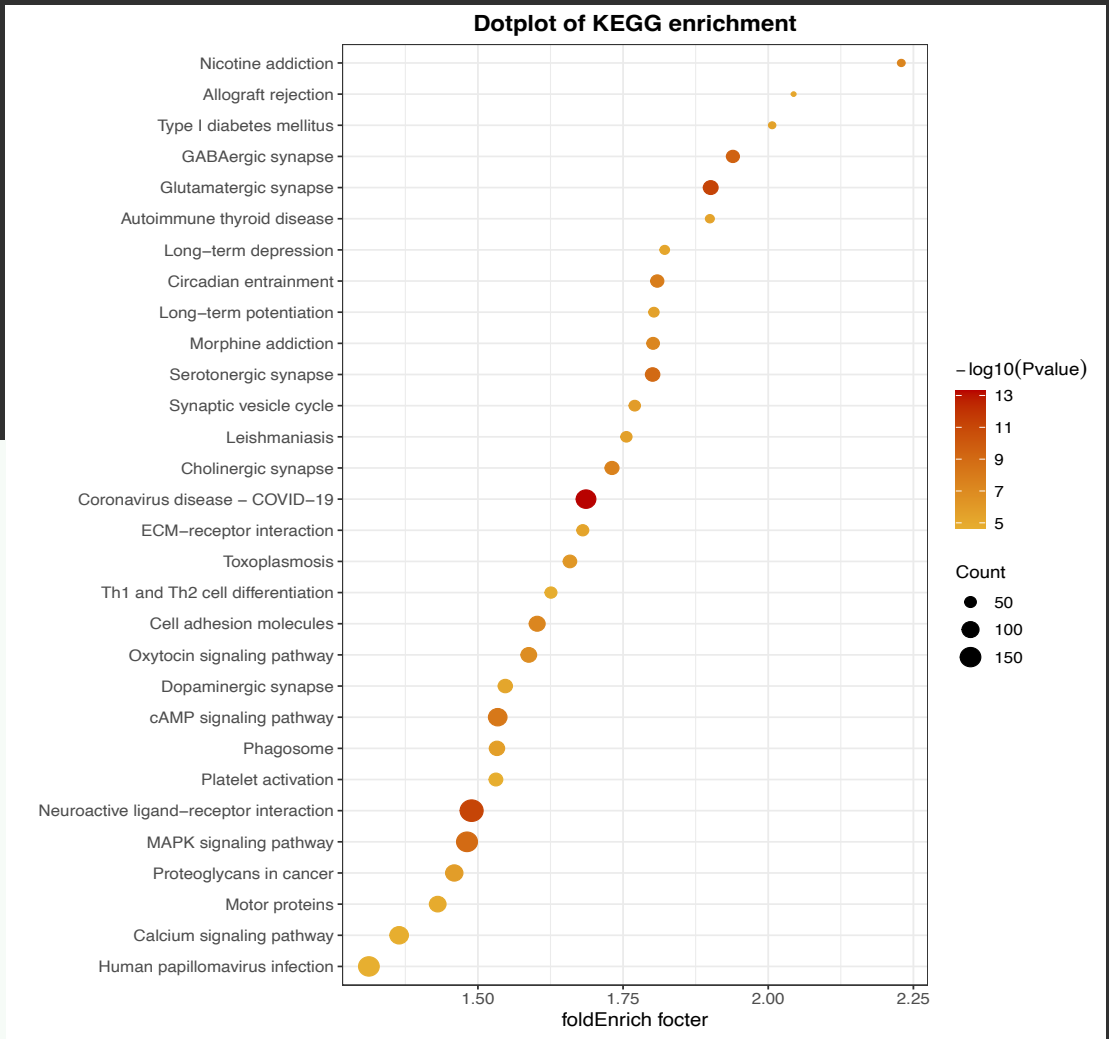
Data-oriented entry points

KEGG databases

Category	Entry point	Database	Content	Classification
Systems information	KEGG PATHWAY	PATHWAY	KEGG pathway maps	Pathway maps
	KEGG BRITE	BRITE	BRITE hierarchies and tables	Brite hierarchies Brite tables
	KEGG MODULE KEGG RModule	MODULE	KEGG modules and reaction modules	Modules Reaction modules
Genomic information	KEGG ORTHOLOGY KEGG Annotation	KO	Functional orthologs	KO
	KEGG GENES KEGG SeqData	GENES	Genes and proteins	
	KEGG GENOME KEGG Virus	GENOME	Genomes of cellular organisms and viruses	Organisms Viruses

```
kegg <- enrichKEGG(
  gene = ID_trans$ENTREZID,
  keyType = 'ncbi-geneid',  #Gene type
  organism = 'hsa',  #hsa means Human being
  pvalueCutoff = 0.2,  #P-value range and cut-off value
  qvalueCutoff = 0.2)  #q-value range and cut-off value

KEGG_res <- as.data.frame(kegg)
KEGG_res <- KEGG_res[1:30,]
KEGG_res$foldEnrich <- enrichment_factor(KEGG_res$GeneRatio,KEGG_res$BgRatio)
```



Import new data

dataset: phenotype - survival data

VISUALIZE

hub: https://gdc.xenahubs.net

cohort	GDC TCGA Glioblastoma (GBM)
dataset ID	TCGA-GBM.survival.tsv
download	https://gdc-hub.s3.us-east-1.amazonaws.com/download/TCGA-GBM.survival.tsv; Full metadata
samples	649
version	07-19-2019
type of data	phenotype
author	Genomic Data Commons
raw data	https://docs.gdc.cancer.gov/Data/Release_Notes/Data_Release_Notes/#data-release-180
raw data	https://api.gdc.cancer.gov/data/
input data format	ROWs (samples) x COLUMNs (identifiers) (i.e. clinicalMatrix)

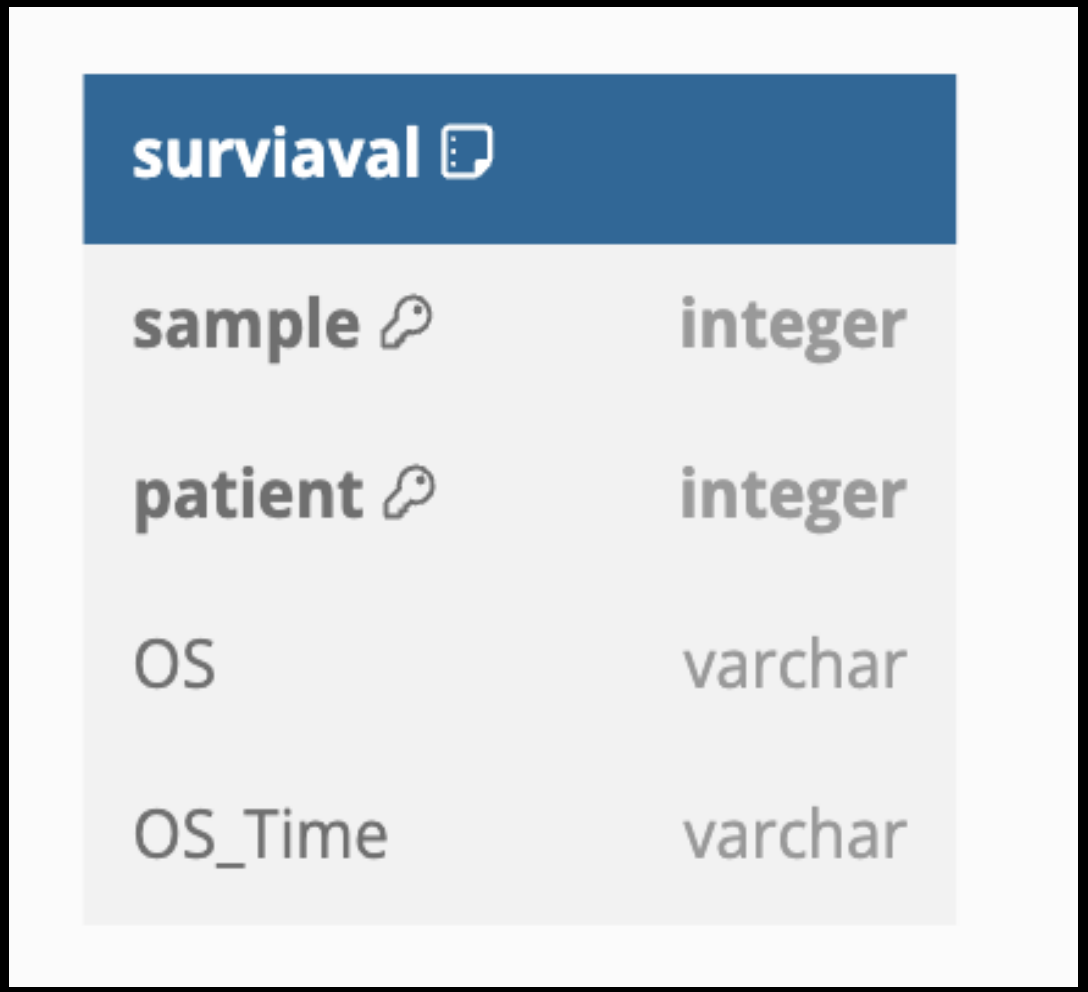
Data were analysed with the use of Univariate Cox regression analysis, which is a survival analysis method used to assess the effect of a single factor on the risk of survival or recurrence. It is based on the Cox proportional hazards model, which is a commonly used model for survival analysis

	sample	OS	_PATIENT	OS.time
1	TCGA-12-0657-01A	1	TCGA-12-0657	3
2	TCGA-32-1977-01A	0	TCGA-32-1977	3
3	TCGA-19-1791-01A	0	TCGA-19-1791	4
4	TCGA-28-1757-01A	0	TCGA-28-1757	4
5	TCGA-19-2624-01A	1	TCGA-19-2624	5
6	TCGA-41-4097-01A	1	TCGA-41-4097	6
7	TCGA-06-0140-01A	1	TCGA-06-0140	6
8	TCGA-28-1746-01A	0	TCGA-28-1746	6
9	TCGA-06-0402-01A	1	TCGA-06-0402	8
10	TCGA-06-0201-01A	1	TCGA-06-0201	12

A data.frame: 6 × 5

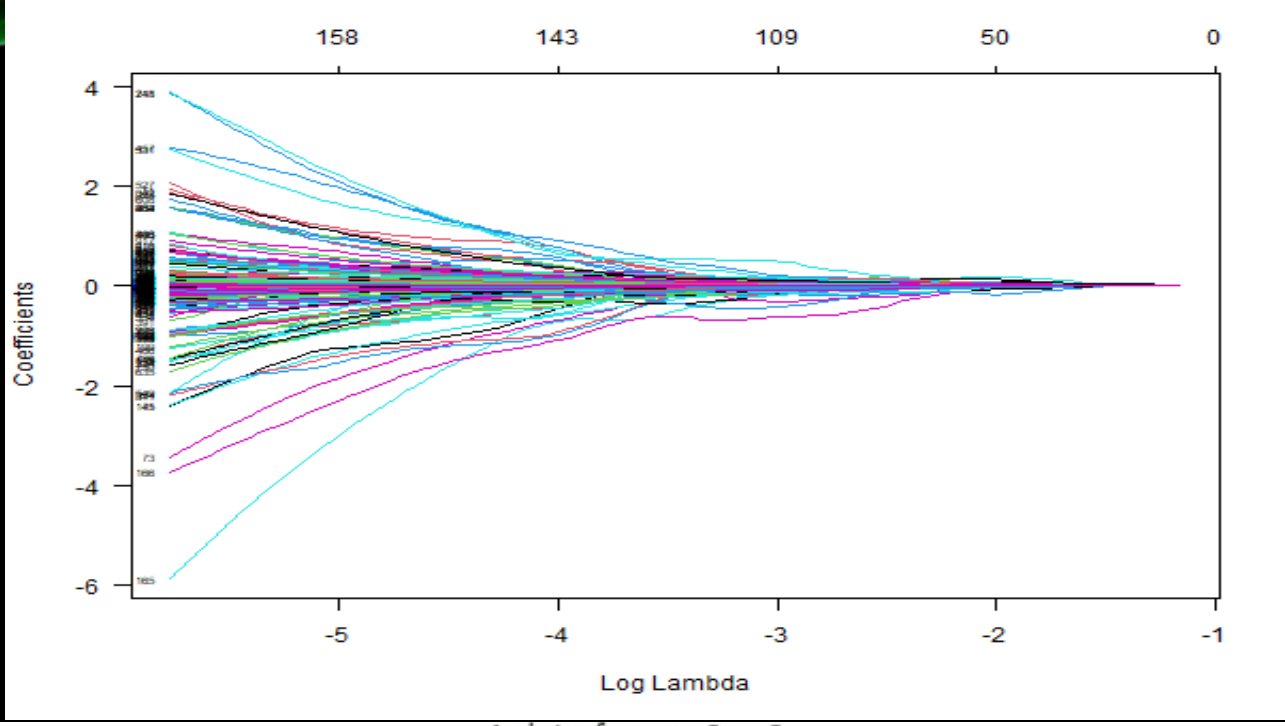
	gene	HR	L95CI	H95CI	pvalue
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
1	ENSG00000239198.1	0.7900129	0.6487125	0.9620909	0.01906069
2	ENSG00000237425.1	0.8756353	0.7753368	0.9889086	0.03238238
3	ENSG00000231313.2	1.2127788	1.0199388	1.4420790	0.02900546
4	ENSG00000254012.1	0.8032369	0.6514770	0.9903487	0.04029282
5	ENSG00000235962.5	0.8230164	0.6801860	0.9958394	0.04519496
6	ENSG00000230568.1	0.8032919	0.6533847	0.9875926	0.03766941

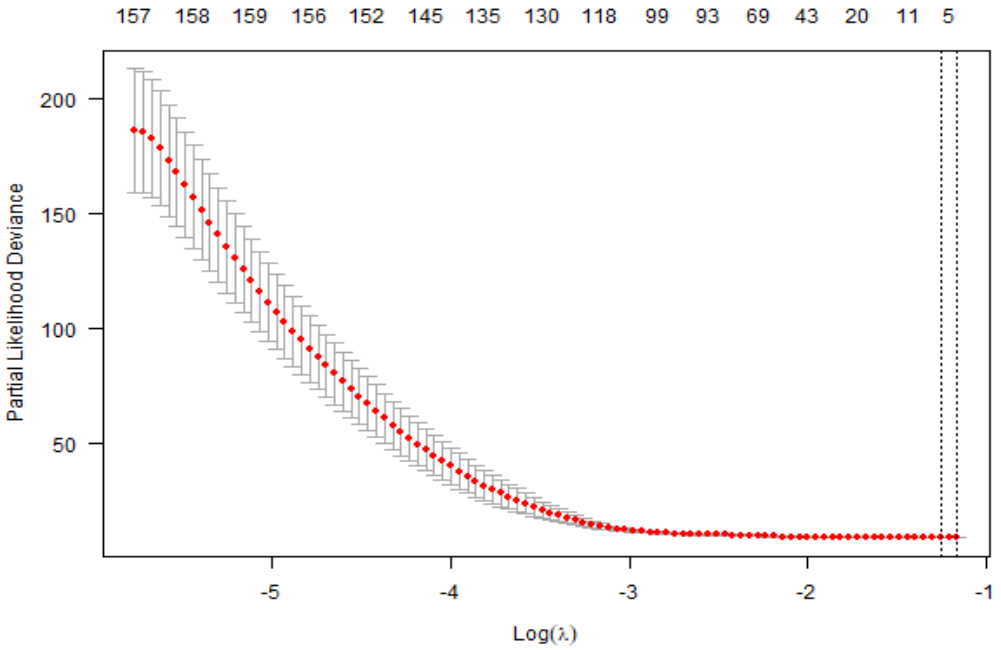
RELATIONAL DATA MODEL



LASSO

Lasso regression can make the coefficients of some features smaller, and even make some coefficients with small absolute values directly zero. In this way, Lasso regression can filter out the main features from the many features. By introducing penalty terms, Lasso regression can control the complexity of the model to avoid overcomplexity, thereby improving its generalization ability. Whether the dependent variable is continuous or discrete, Lasso can handle it



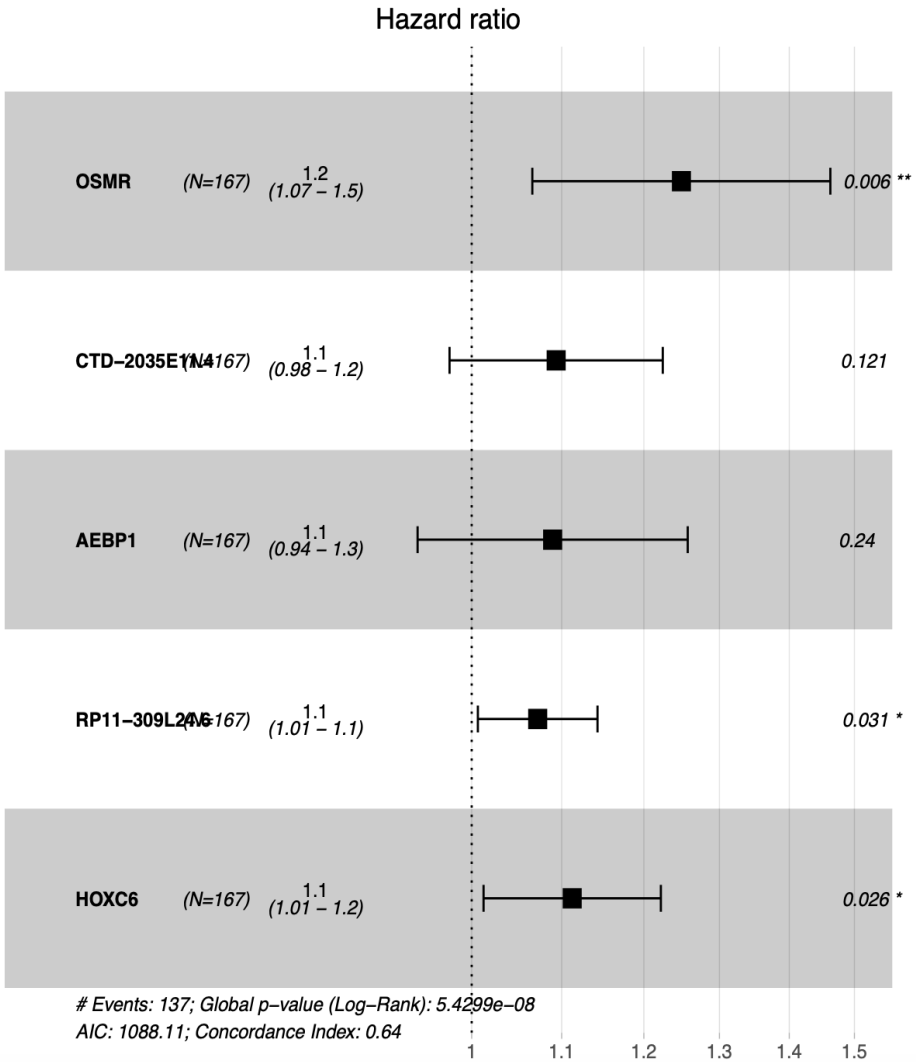


A data.frame: 6 × 8

	sample	OS.time	OS	OSMR	CTD-2035E11.4	AEBP1	RP11-309L2
	<chr>	<int>	<int>	<dbl>	<dbl>	<dbl>	<dl
1	TCGA-19-2624-01A	5	1	14.49971	15.99230	21.17678	11.70(
2	TCGA-41-4097-01A	6	1	18.35223	12.67729	20.98071	11.85:
3	TCGA-06-2569-01A	13	0	15.59140	11.90376	19.48246	13.12:
4	TCGA-06-0219-01A	22	1	17.34499	13.38933	18.69600	11.15(
5	TCGA-41-2571-01A	26	1	16.57018	14.77731	20.92882	13.24:
6	TCGA-06-0750-01A	28	1	18.19763	15.80249	22.63977	12.87(

FINAL RESULT

```
model <- coxph( Surv(OS.time, OS) ~ OSMR+`CTD-2035E11.4`+AEBP1+`RP11-309L24.6`+HOXC
               data = targetGene )
summary(model)
#pdf("04.CoxModel.pdf")
ggforest(model,
         data = targetGene,
         main = "Hazard ratio", # Titled
         cpositions = c(0.08, 0.2, 0.35), #distance for the first three columns
         fontsize = 0.8, # set fontsize
         refLabel = "reference",
         noDigits = 2) #set the digits to keep after the decimal point
```




```
                     coef exp(coef) se(coef)     z Pr(>|z|)
OSMR              0.22209   1.24869  0.08056 2.757  0.00583 **
`CTD-2035E11.4`   0.08944   1.09356  0.05761 1.553  0.12054
AEBP1             0.08574   1.08952  0.07299 1.175  0.24011
`RP11-309L24.6`   0.06989   1.07239  0.03233 2.162  0.03062 *
HOXC6             0.10644   1.11231  0.04790 2.222  0.02627 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                  exp(coef) exp(-coef) lower .95 upper .95
OSMR                  1.249     0.8008    1.0663     1.462
`CTD-2035E11.4`       1.094     0.9144    0.9768     1.224
AEBP1                 1.090     0.9178    0.9443     1.257
`RP11-309L24.6`       1.072     0.9325    1.0066     1.143
HOXC6                 1.112     0.8990    1.0126     1.222

Concordance= 0.643  (se = 0.028 )
Likelihood ratio test= 42.17  on 5 df,   p=5e-08
Wald test            = 30.27  on 5 df,   p=1e-05
Score (logrank) test = 30.97  on 5 df,   p=1e-05
```

ARTICLE PROOF

Huang *et al. Cancer Cell International* (2022) 22:170
https://doi.org/10.1186/s12935-022-02589-9

Cancer Cell International

RESEARCH Open Access

HOXC6 impacts epithelial-mesenchymal transition and the immune microenvironment through gene transcription in gliomas

Hui Huang†, Zhengyuan Huo†, Jiantong Jiao†, Wei Ji, Jin Huang, Zheng Bian, Bin Xu, Junfei Shao* and Jun Sun*

Abstract

Background: Gliomas are the most common primary malignant tumours of the central nervous system (CNS). To improve the prognosis of glioma, it is necessary to identify molecular markers that may be useful for glioma therapy. HOXC6, an important transcription factor, is involved in multiple cancers. However, the role of HOXC6 in gliomas is not clear.

Methods: Bioinformatic and IHC analyses of collected samples (n = 299) were performed to detect HOXC6 expression and the correlation between HOXC6 expression and clinicopathological features of gliomas. We collected clinical information from 177 to 299 patient samples and estimated the prognostic value of HOXC6. Moreover, cell proliferation assays were performed. We performed Gene Ontology (GO) analysis and gene set enrichment analysis (GSEA) based on ChIP-seq and public datasets to explore the biological characteristics of HOXC6 in gliomas. RNA-seq was conducted to verify the relationship between HOXC6 expression levels and epithelial-mesenchymal transition (EMT) biomarkers. Furthermore, the tumour purity, stromal and immune scores were evaluated. The relationship between HOXC6 expression and infiltrating immune cell populations and immune checkpoint proteins was also researched.

Results: HOXC6 was overexpressed and related to the clinicopathological features of gliomas. In addition, knockdown of HOXC6 inhibited the proliferation of glioma cells. Furthermore, increased HOXC6 expression was associated

Hindawi
Journal of Oncology
Volume 2022, Article ID 8016102, 18 pages
https://doi.org/10.1155/2022/8016102

Research Article

HOXC6 Regulates the Epithelial-Mesenchymal Transition through the TGF-β/Smad Signaling Pathway and Predicts a Poor Prognosis in Glioblastoma

Sun Eryi, Li Zheng, Cai Honghua, Zhao Su, Xie Han, Pan Donggang, Zhou Zhou, Zhan Liping, and Chen Bo

Department of Neurosurgery, Affiliated People's Hospital of Jiangsu University, Zhenjiang, Jiangsu Province 212002, China
Affiliated Hospital of Jiangsu University, Zhenjiang, Jiangsu Province 212002, China
Wujin Traditional Chinese Medicine Hospital, Changzhou, Jiangsu Province, China

Correspondence should be addressed to Chen Bo; chenbo_19771202@163.com

Received 2 March 2022; Revised 13 April 2022; Accepted 15 April 2022; Published 5 May 2022

Academic Editor: Jimei Wang

Background. The HOX gene family of transcription factors, characterized by conserved homeodomains, is positively correlated with the resistance to chemotherapy drugs and poor prognosis, as well as the initiating potential of gliomas. However, there are few studies regarding the HOXC6 gene in glioma cells. Therefore, in the present study, we explored the regulatory roles

Research Article

Expression Profiles of HOXC6 Predict the Survival of Glioblastoma Patients and Correlate with Cell Cycle

Zhen-Hang Li, Yue Ma, Yan Zhou, and Zhen-Hua Huang

Department of Neurosurgery, Tianjin Huanhu Hospital, Tianjin, China
Clinical College of Neurology, Neurosurgery and Neurorehabilitation, Tianjin Medical University, Tianjin, China

Correspondence should be addressed to Yue Ma; drmayue@outlook.com

Zhen-Hang Li and Yue Ma contributed equally to this work.

Received 1 July 2021; Revised 14 December 2021; Accepted 13 March 2022; Published 6 April 2022

Academic Editor: Ferdinand Frauscher


Q&A