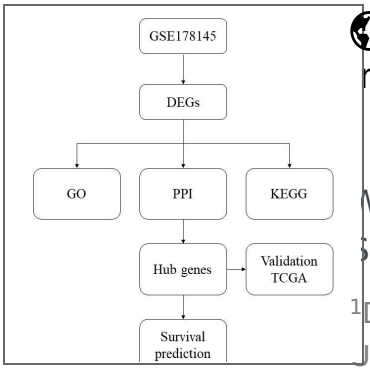


AUG 22, 2023

OPEN ACCESS

# CXCL2 and IL-1β: prognostic markers and immune cell infiltration targets of colorectal cancer

Wen-Juan Zhang[1,2], Ke-Yun Li[1], Sheng Peng[1], Gao-Lu Cao[1], Sheng-Cheng Shi[3], Zi-Qing Xiao[3], Shui-Qin Chen[1], Shao-Gui Wan[4]

[1]Department of Immunology, School of Basic Medicine, Gannan Medical University, Ganzhou 341000, Jiangxi, P.R. China;
[2]2Key Laboratory of Prevention and Treatment of Cardiovascular and Cerebrovascular Diseases of Ministry of Education, Gannan Medical University, Ganzhou 341000, Jiangxi, P.R. China;
[3]The First Clinical Medical College, Gannan Medical University, Ganna 341000, China;
[4]Center for Molecular Pathology, School of Basic Medicine, Gannan Medical University, Ganzhou 341000, Jiangxi, P.R. China

Sheng Peng
Department of Immunology, School of Basic Medicine, Gannan M...

DISCLAIMER

The authors have no conflicts of interest concerning the work reported in this manuscript.

**Protocol Citation:** Wen-Jua Zhang, Ke-Yun Li, Sheng Peng, Gao-Lu Cao, Sheng-Cheng Shi, Zi-Qing Xiao, Shui-Qin Chen, Shao-Gui Wa 2023. CXCL2 and IL-1β: prognostic markers and immune cell infiltration targets of colorectal cancer. **protocols.io** https://protocols.io/view/cxcl2-and-il-1-prognostic-markers-and-immune-cell-cyyxxxxn



**Protocol status:** Working
We use this protocol and it's working

**Created:** Aug 22, 2023

**Last Modified:** Aug 22, 2023

**PROTOCOL integer ID:** 86775

**Keywords:** Colorectal carcinoma, Differentially expressed genes, Bioinformatics analysis, Protein-Protein interaction network, prognosis, Tumor-infiltrating immune cells

ABSTRACT

Aims In this paper, the key genes in the occurrence and development of Colorectal carcinoma (CRC) are identified through survival analysis and immunoinfiltration analysis using bioinformatics methods, which help predict prognostic markers of CRC and develop therapeutic drugs as targets. Method The gene expression profiles (GSE178145) were downloaded from the Gene Expression Omnibus (GEO) database and processed by R software to screen the differentially expressed genes (DEGs) between normal tissues and tumor tissue samples of the CRC. These DEGs were analyzed by Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) to determine the biological role of the DEGs. Subsequently, protein-protein interaction network (PPI) was used to identify key genes that predicted CRC progression, and then GEPIA2 was used to verify the expression levels of key genes and analyze survival prognosis. Finally, the correlation between genes related to CRC prognosis and immune infiltrating cells was analyzed using TIMER database. Result 897 DEGs were selected. GO analysis showed that DEGs was mainly concentrated in the regulation of metal ion transport, ion channel activity, shrinkage fiber and other functions. KEGG pathway analysis showed that DEGs was mainly involved in cytokine - receptor interaction, calcium signaling and other pathways. Get 10 key genes through the analysis of the PPI, which were CXCL2, IL-1β, CCL2, TNF, CXCL10, IL-6, IFIT1, IFIT2, USP18 and OASL2.Survival analysis of key genes showed that CXCL2 and IL-1β affect the overall survival of CRC patients. The immunoinfiltration analysis of prognostic genes showed that the expression of CXCL2 and IL-1β was correlated with the infiltration of macrophages. Conclusion CXCL2 and IL-1ß may be key genes for the occurrence and development of CRC, and can be considered the molecular biological basis for early diagnosis, prognosis and targeted therapy of CRC.

GUIDELINES

no

MATERIALS

no

SAFETY WARNINGS

no

ETHICS STATEMENT

no

## Data preprocessing of colorectal cancer tissue transcriptome.

1 Download the Gene expression matrix of GSE178145 and convert them to blast databases. The GSE178145 is located on the Affymetrix GPL24247 platform (Affymetrix Illumina NovaSeq 6000) and is submitted by Wu M.

All raw data were processed using the R software(4.2.3).

Group information was extracted from Metadata using the GEOquery package and divided into normal control and colorectal cancer groups. From the GENCODE website (https://www.gencodegenes.org/), download the gene annotation file (GRCh38) in gif format. Use the tidyverse package to identify the gene ensemble ID and convert it into the standard gene name (HGNC symbol) to obtain the standardized gene expression matrix.The base package was used to filter out genes with low expression levels.

Dataset

**GSE178145_count_table.txt** NAME

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE178145 LINK

Dataset

**gencode.vM25.chr_patch_hapl_scaff.annotation.gtf** NAME

https://www.gencodegenes.org/mouse/ LINK

Software

**R Studio Desktop** NAME

The R Studio, Inc. DEVELOPER

Command

```
###
proj = "GSE178145"
library(GEOquery)
eSet = getGEO("GSE178145",destdir = ".",getGPL = F)
eSet = eSet[[1]]
```

```
exp = exprs(eSet)
pd = pData(eSet)
dat <- read.table("GSE178145_count_table.txt",header = T)

###
library('rtracklayer')
gtf_data = as.data.frame(import('gencode.vM25.chr_patch_hapl_scaff.annotation.gtf'))
s2e      = gtf_data[,c(10,12)]
library('dplyr')
s2e <- distinct(s2e)
library("tidyverse")
s2e <- separate(s2e,1,into = c('gene_id','drop'),sep = '[.]') %>% select(-drop)
ids=data.frame(gene_id=dat$gene_id, median=apply(dat,1,median))
table(ids$gene_id %in% s2e$gene_id)
ids=ids[ids$gene_id %in% s2e$gene_id,]
ids$symbol=s2e[match(ids$gene_id,s2e$gene_id),2]
length(unique(ids$symbol))
ids=ids[!duplicated(ids$symbol),]
dim(ids)
dat1= dat[rownames(ids),]
dat1$symbol=ids[match(dat1$gene_id,ids$gene_id),3]

###
dat1=dat1[!duplicated(dat1$symbol),]
sum(is.na(dat1$symbol))
which(is.na(dat1$symbol))
#dat1<-dat1[-1350,]
sum(is.na(dat1$symbol))

rownames(dat1)=as.vector(as.matrix(dat1[,17]))
dat1$ID_REF<-NULL
dat1$symbol<-NULL
exp = as.matrix(dat1)

###
nrow(exp)
exp = exp[apply(exp, 1, function(x) sum(x > 0) > 0.5*ncol(exp)), ]
#exp = exp[rowSums(exp)>0,]
nrow(exp)
exp = as.matrix(exp)

###
Group = str_remove(colnames(exp),"\\d")
Group = factor(Group,levels = c("WT","CRC"),nmax = 12)
table(Group)

###
```

```
save(exp,Group,proj,file = paste0(proj,".Rdata"))
```

## Analysis of differential gene expression

2 The DESeq2, edgeR and limma packages were used to analyze the differential expression of standardized gene expression matrix.The overlapping DEGs were visualized by using the tinyarray and ggplot2 packages. The identification standard of DEGs is $|\log_2 \text{fold change}| \geq 2$ and $P$ value < 0.05.

Command

```
rm(list = ls())
load("GSE178145.Rdata")
fix(Group)
table(Group)

class(exp)

#deseq2----
library(DESeq2)
colData <- data.frame(row.names =colnames(exp),
                      condition=Group)
if(!file.exists(paste0(proj,"_dd.Rdata"))){
  dds <- DESeqDataSetFromMatrix(
    countData = exp,
    colData = colData,
    design = ~ condition)
  dds <- DESeq(dds)
  save(dds,file = paste0(proj,"_dd.Rdata"))
}
load(file = paste0(proj,"_dd.Rdata"))
class(dds)
res <- results(dds, contrast = c("condition",rev(levels(Group))))#constrast :
c("condition",rev(levels(Group)))
class(res)
DEG1 <- as.data.frame(res)
DEG1 <- DEG1[order(DEG1$pvalue),]
DEG1 = na.omit(DEG1)
head(DEG1)
```

```
logFC_t = 2
pvalue_t = 0.05

k1 = (DEG1$pvalue < pvalue_t)&(DEG1$log2FoldChange < -logFC_t);table(k1)
k2 = (DEG1$pvalue < pvalue_t)&(DEG1$log2FoldChange > logFC_t);table(k2)
DEG1$change = ifelse(k1,"DOWN",ifelse(k2,"UP","NOT"))
table(DEG1$change)
head(DEG1)
write.csv(DEG1,file ="DEG1.csv" )

#edgeR----
library(edgeR)
exp = na.omit(exp)
dge <- DGEList(counts=exp,group=Group)
dge$samples$lib.size <- colSums(dge$counts)
dge <- calcNormFactors(dge)

design <- model.matrix(~Group)

dge <- estimateGLMCommonDisp(dge, design)
dge <- estimateGLMTrendedDisp(dge, design)
dge <- estimateGLMTagwiseDisp(dge, design)

fit <- glmFit(dge, design)
fit <- glmLRT(fit)

DEG2=topTags(fit, n=Inf)
class(DEG2)
DEG2=as.data.frame(DEG2)
head(DEG2)

k1 = (DEG2$PValue < pvalue_t)&(DEG2$logFC < -logFC_t);table(k1)
k2 = (DEG2$PValue < pvalue_t)&(DEG2$logFC > logFC_t);table(k2)
DEG2$change = ifelse(k1,"DOWN",ifelse(k2,"UP","NOT"))

head(DEG2)
table(DEG2$change)

###limma----
library(limma)
dge <- edgeR::DGEList(counts=exp)
dge <- edgeR::calcNormFactors(dge)
design <- model.matrix(~Group)
v <- voom(dge,design, normalize="quantile")

design <- model.matrix(~Group)
```

```
fit <- lmFit(v, design)
fit= eBayes(fit)

DEG3 = topTable(fit, coef=2, n=Inf)
DEG3 = na.omit(DEG3)

k1 = (DEG3$P.Value < pvalue_t)&(DEG3$logFC < -logFC_t);table(k1)
k2 = (DEG3$P.Value < pvalue_t)&(DEG3$logFC > logFC_t);table(k2)
DEG3$change = ifelse(k1,"DOWN",ifelse(k2,"UP","NOT"))
table(DEG3$change)
head(DEG3)

tj = data.frame(deseq2 = as.integer(table(DEG1$change)),
                edgeR = as.integer(table(DEG2$change)),
                limma_voom = as.integer(table(DEG3$change)),
                row.names = c("down","not","up")
);tj
save(DEG1,DEG2,DEG3,Group,tj,file = paste0(proj,"_DEG.Rdata"))

###
library(ggplot2)
#BiocManager::install("org.Mm.eg.db")
library(tinyarray)


dat = log2(cpm(exp)+1)
pca.plot = draw_pca(dat,Group);pca.plot
save(pca.plot,file = paste0(proj,"_pcaplot.Rdata"))

cg1 = rownames(DEG1)[DEG1$change !="NOT"]
cg2 = rownames(DEG2)[DEG2$change !="NOT"]
cg3 = rownames(DEG3)[DEG3$change !="NOT"]

h1 = draw_heatmap(dat[cg1,],Group,n_cutoff = 2)
h2 = draw_heatmap(dat[cg2,],Group,n_cutoff = 2)
h3 = draw_heatmap(dat[cg3,],Group,n_cutoff = 2)

v1 = draw_volcano(DEG1,pkg = 1,logFC_cutoff = logFC_t)
v2 = draw_volcano(DEG2,pkg = 2,logFC_cutoff = logFC_t)
v3 = draw_volcano(DEG3,pkg = 3,logFC_cutoff = logFC_t)

library(patchwork)
(h1 + h2 + h3) / (v1 + v2 + v3) +plot_layout(guides = 'collect') &theme(legend.position =
"none")

ggsave(paste0(proj,"_heat_vo.png"),width = 15,height = 10)
```

```
###
UP=function(df){
  rownames(df)[df$change=="UP"]
}
DOWN=function(df){
  rownames(df)[df$change=="DOWN"]
}

up = intersect(intersect(UP(DEG1),UP(DEG2)),UP(DEG3))
down = intersect(intersect(DOWN(DEG1),DOWN(DEG2)),DOWN(DEG3))
dat = log2(cpm(exp)+1)
hp = draw_heatmap(dat[c(up,down),],Group,n_cutoff = 2)

up_total <- as.data.frame(up)
down_total <- as.data.frame(down)
colnames(up_total) <- "SYMBOL"
colnames(down_total) <- "SYMBOL"
library(dplyr)
DEG <- bind_rows(up_total,down_total)
write.csv(DEG,file = "raw_gene.csv")
##############################

up_genes = list(Deseq2 = UP(DEG1),
                edgeR = UP(DEG2),
                limma = UP(DEG3))

down_genes = list(Deseq2 = DOWN(DEG1),
                  edgeR = DOWN(DEG2),
                  limma = DOWN(DEG3))

install.packages('VennDiagram')

up.plot <- draw_venn(up_genes,"UPgene")
down.plot <- draw_venn(down_genes,"DOWNgene")

library(patchwork)
up.plot + down.plot

pca.plot + hp+up.plot +down.plot+ plot_layout(guides = "collect")
ggsave(paste0(proj,"_heat_ve_pca.png"),width = 15,height = 10)
up.plot +down.plot+ plot_layout(guides = "collect")
```

## GO and KEGG pathway enrichment analysis

3 The clusterProfiler package were used for enrichment analysis, and the enrichment results were visualized by enrichplot and DOSE packages. The adjusted *P* value (*P*.adj)< 0.05 was considered a statistically significant difference.

Command

```
###KEGG、GO
DEG <- as.vector(as.matrix(DEG[,1]))
class(DEG)
###table(DEG123$SYMBOL %in% keys(org.Mm.eg.db, "SYMBOL"))
###DEG123 <- subset(DEG,DEG$SYMBOL %in% keys(org.Mm.eg.db, "SYMBOL"),select =
SYMBOL)
###duplicated(DEG123)
library(clusterProfiler)
DEG_entrez <- as.character(na.omit(bitr(DEG,
                                fromType="SYMBOL",
                                toType="ENTREZID",
                                OrgDb="org.Mm.eg.db")[,2]))
library(R.utils)
R.utils::setOption("clusterProfiler.download.method",'auto')
library(clusterProfiler)

kegg_enrich_results <- enrichKEGG(gene  = DEG_entrez,
                            organism  = "mmu",
                            pvalueCutoff = 0.05,
                            qvalueCutoff = 0.2)
kegg_enrich_results <- DOSE::setReadable(kegg_enrich_results,
                                 OrgDb="org.Mm.eg.db",
                                 keyType='ENTREZID')#ENTREZID to gene Symbol
write.csv(kegg_enrich_results@result,'KEGG_enrichresults.csv')
save(kegg_enrich_results, file ='KEGG_enrichresults.Rdata')

go_enrich_results <- enrichGO(gene = DEG_entrez,
                        OrgDb = "org.Mm.eg.db",
                        ont   = "ALL"  ,     #One of "BP", "MF"  "CC"  "ALL"
                        pvalueCutoff  = 0.05,
                        qvalueCutoff  = 0.2,
                        readable      = TRUE)
write.csv(go_enrich_results@result, 'GO_enrichresults.csv')
save(go_enrich_results, file ='GO_enrichresults.Rdata')
```

```
###
options(stringsAsFactors = F)
library(enrichplot)
library(tidyverse)
library(DOSE)

library(pathview)
#GO enrichment
### dotplot
dotp <- enrichplot::dotplot(go_enrich_results,font.size =14)+
  theme(legend.key.size = unit(10, "pt"),
        plot.margin=unit(c(1,1,1,1),'lines'))
if (T) {
  dotp <- enrichplot::dotplot(go_enrich_results,font.size =20,split = 'ONTOLOGY')+
    scale_y_discrete(labels=function(go_enrich_results) str_wrap(go_enrich_results,width =
180))+
    scale_size(range=c(1, 20))+
    facet_grid(ONTOLOGY~., scale="free")
    theme(legend.key.size = unit(20, "pt"),
          plot.margin=unit(c(1,1,1,1),'lines'))
};dotp
ggsave(dotp,filename = paste0("go_dotplot",'.jpg'),width =18,height =18)

#KEGG enrichment
library(stringr)
### dotplot
kegg_dotplot <- dotplot(kegg_enrich_results,showCategory = 10,font.size =25)+
  scale_size(range=c(2, 25))+
  scale_y_discrete(labels=function(kegg_enrich_results) str_wrap(kegg_enrich_results,width
= 100));kegg_dotplot
ggsave(kegg_dotplot,filename = paste0("kegg_dotplot",'.jpg'),width =20,height =10)
```

## Protein-protein interaction network

**4** The screened DEGs was submitted to the String database to analyze the potential interaction of proteins. "Homo sapiens" was taken as the research species, and the confidence score > 0.07 was considered of great significance. Cytoscape (3.9.1) was used to analyze the interaction network of different genes. The plugin cytohubba was used to select the top 10 hub genes from the PPI network. The cytohubba uses the Maximal Clique Centrality (MCC) algorithm.

Dataset

STRING

NAME

https://cn.string-db.org/cgi/input?sessionId=b26fDU8a5QLP&input_page_active_form=multiple_sequences

LINK

Software

Cytoscape

NAME

## Verification of the expression of key genes and analysis of su...

5 The data of 362 patients with CRC from the TCGA and GTEx were selected to analyze the survival and prognosis. The CRC and normal tissues were compared between groups to verify the differences in the expression of key genes. The data of CRC patients were divided into high expression group and low expression group with the median as the cut-off limit,so as to explore the correlation between the expression level of key genes and the overall survival (OS) rate of the patients.

Dataset

GEPIA2

NAME

GEPIA2 (http://gepia2.cancer-pku.cn/

LINK

## Analysis of the relationship between prognosis-related genes...

6 The key genes were introduced into SCNA module toselect six immune cells for analysis, and then use the Correlation module to verify the relationship between the key genes and the immune cell markers. The *P* value < 0.05 was considered a statistically significant difference.

| Dataset | |
| --- | --- |
| **TIMER** | NAME |
| https://cistrome.shinyapps.io/timer/ | LINK |