code\_feb28

# Initialization

## Packages import

First, we import packages and datasets

# Install Bioconductor packages  
#if (!require("BiocManager", quietly = TRUE))  
 #install.packages("BiocManager")  
#BiocManager::install(version = "3.17")  
  
## Package to download data from acession numbers  
# BiocManager::install("Biobase")  
# BiocManager::install("GEOquery")  
library(Biobase)

Loading required package: BiocGenerics

Attaching package: 'BiocGenerics'

The following objects are masked from 'package:stats':  
  
 IQR, mad, sd, var, xtabs

The following objects are masked from 'package:base':  
  
 anyDuplicated, aperm, append, as.data.frame, basename, cbind,  
 colnames, dirname, do.call, duplicated, eval, evalq, Filter, Find,  
 get, grep, grepl, intersect, is.unsorted, lapply, Map, mapply,  
 match, mget, order, paste, pmax, pmax.int, pmin, pmin.int,  
 Position, rank, rbind, Reduce, rownames, sapply, setdiff, sort,  
 table, tapply, union, unique, unsplit, which.max, which.min

Welcome to Bioconductor  
  
 Vignettes contain introductory material; view with  
 'browseVignettes()'. To cite Bioconductor, see  
 'citation("Biobase")', and for packages 'citation("pkgname")'.

library(GEOquery)

Setting options('download.file.method.GEOquery'='auto')

Setting options('GEOquery.inmemory.gpl'=FALSE)

## Packages for DESeq2  
# Install DESeq2 package  
# BiocManager::install("DESeq2")  
library(DESeq2)

Warning: package 'DESeq2' was built under R version 4.3.1

Loading required package: S4Vectors

Warning: package 'S4Vectors' was built under R version 4.3.1

Loading required package: stats4

Attaching package: 'S4Vectors'

The following object is masked from 'package:utils':  
  
 findMatches

The following objects are masked from 'package:base':  
  
 expand.grid, I, unname

Loading required package: IRanges

Warning: package 'IRanges' was built under R version 4.3.1

Attaching package: 'IRanges'

The following object is masked from 'package:grDevices':  
  
 windows

Loading required package: GenomicRanges

Warning: package 'GenomicRanges' was built under R version 4.3.1

Loading required package: GenomeInfoDb

Warning: package 'GenomeInfoDb' was built under R version 4.3.1

Loading required package: SummarizedExperiment

Loading required package: MatrixGenerics

Warning: package 'MatrixGenerics' was built under R version 4.3.1

Loading required package: matrixStats

Warning: package 'matrixStats' was built under R version 4.3.2

Attaching package: 'matrixStats'

The following objects are masked from 'package:Biobase':  
  
 anyMissing, rowMedians

Attaching package: 'MatrixGenerics'

The following objects are masked from 'package:matrixStats':  
  
 colAlls, colAnyNAs, colAnys, colAvgsPerRowSet, colCollapse,  
 colCounts, colCummaxs, colCummins, colCumprods, colCumsums,  
 colDiffs, colIQRDiffs, colIQRs, colLogSumExps, colMadDiffs,  
 colMads, colMaxs, colMeans2, colMedians, colMins, colOrderStats,  
 colProds, colQuantiles, colRanges, colRanks, colSdDiffs, colSds,  
 colSums2, colTabulates, colVarDiffs, colVars, colWeightedMads,  
 colWeightedMeans, colWeightedMedians, colWeightedSds,  
 colWeightedVars, rowAlls, rowAnyNAs, rowAnys, rowAvgsPerColSet,  
 rowCollapse, rowCounts, rowCummaxs, rowCummins, rowCumprods,  
 rowCumsums, rowDiffs, rowIQRDiffs, rowIQRs, rowLogSumExps,  
 rowMadDiffs, rowMads, rowMaxs, rowMeans2, rowMedians, rowMins,  
 rowOrderStats, rowProds, rowQuantiles, rowRanges, rowRanks,  
 rowSdDiffs, rowSds, rowSums2, rowTabulates, rowVarDiffs, rowVars,  
 rowWeightedMads, rowWeightedMeans, rowWeightedMedians,  
 rowWeightedSds, rowWeightedVars

The following object is masked from 'package:Biobase':  
  
 rowMedians

# package for logistic regression + basic calculation  
# install.packages("glmnet", repos = "http://cran.us.r-project.org")  
  
library(glmnet)

Warning: package 'glmnet' was built under R version 4.3.2

Loading required package: Matrix

Warning: package 'Matrix' was built under R version 4.3.2

Attaching package: 'Matrix'

The following object is masked from 'package:S4Vectors':  
  
 expand

Loaded glmnet 4.1-8

library(dplyr)

Warning: package 'dplyr' was built under R version 4.3.2

Attaching package: 'dplyr'

The following object is masked from 'package:matrixStats':  
  
 count

The following objects are masked from 'package:GenomicRanges':  
  
 intersect, setdiff, union

The following object is masked from 'package:GenomeInfoDb':  
  
 intersect

The following objects are masked from 'package:IRanges':  
  
 collapse, desc, intersect, setdiff, slice, union

The following objects are masked from 'package:S4Vectors':  
  
 first, intersect, rename, setdiff, setequal, union

The following object is masked from 'package:Biobase':  
  
 combine

The following objects are masked from 'package:BiocGenerics':  
  
 combine, intersect, setdiff, union

The following objects are masked from 'package:stats':  
  
 filter, lag

The following objects are masked from 'package:base':  
  
 intersect, setdiff, setequal, union

library(ggplot2)

Warning: package 'ggplot2' was built under R version 4.3.2

## Packages for GOSeq (pathway examination)  
#BiocManager::install("goseq")  
#BiocManager::install("clusterProfiler")  
#BiocManager::install("AnnotationDbi")  
#BiocManager::install("org.Hs.eg.db")   
#BiocManager::install("DOSE")  
library(clusterProfiler)

Warning: package 'clusterProfiler' was built under R version 4.3.1

Registered S3 methods overwritten by 'treeio':  
 method from   
 MRCA.phylo tidytree  
 MRCA.treedata tidytree  
 Nnode.treedata tidytree  
 Ntip.treedata tidytree  
 ancestor.phylo tidytree  
 ancestor.treedata tidytree  
 child.phylo tidytree  
 child.treedata tidytree  
 full\_join.phylo tidytree  
 full\_join.treedata tidytree  
 groupClade.phylo tidytree  
 groupClade.treedata tidytree  
 groupOTU.phylo tidytree  
 groupOTU.treedata tidytree  
 inner\_join.phylo tidytree  
 inner\_join.treedata tidytree  
 is.rooted.treedata tidytree  
 nodeid.phylo tidytree  
 nodeid.treedata tidytree  
 nodelab.phylo tidytree  
 nodelab.treedata tidytree  
 offspring.phylo tidytree  
 offspring.treedata tidytree  
 parent.phylo tidytree  
 parent.treedata tidytree  
 root.treedata tidytree  
 rootnode.phylo tidytree  
 sibling.phylo tidytree

clusterProfiler v4.8.3 For help: https://yulab-smu.top/biomedical-knowledge-mining-book/  
  
If you use clusterProfiler in published research, please cite:  
T Wu, E Hu, S Xu, M Chen, P Guo, Z Dai, T Feng, L Zhou, W Tang, L Zhan, X Fu, S Liu, X Bo, and G Yu. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. The Innovation. 2021, 2(3):100141

Attaching package: 'clusterProfiler'

The following object is masked from 'package:IRanges':  
  
 slice

The following object is masked from 'package:S4Vectors':  
  
 rename

The following object is masked from 'package:stats':  
  
 filter

library(org.Hs.eg.db)

Loading required package: AnnotationDbi

Warning: package 'AnnotationDbi' was built under R version 4.3.1

Attaching package: 'AnnotationDbi'

The following object is masked from 'package:clusterProfiler':  
  
 select

The following object is masked from 'package:dplyr':  
  
 select

library(AnnotationDbi)  
library(DOSE)

Warning: package 'DOSE' was built under R version 4.3.1

DOSE v3.26.2 For help: https://yulab-smu.top/biomedical-knowledge-mining-book/  
  
If you use DOSE in published research, please cite:  
Guangchuang Yu, Li-Gen Wang, Guang-Rong Yan, Qing-Yu He. DOSE: an R/Bioconductor package for Disease Ontology Semantic and Enrichment analysis. Bioinformatics 2015, 31(4):608-609

#BiocManager::install("illuminaHumanv4.db")  
library("illuminaHumanv4.db")

## Data import

#US\_1 - Arendt et al., 2015  
data\_u <- getGEO("GSE89632")

Found 1 file(s)

GSE89632\_series\_matrix.txt.gz

data\_u <- data\_u$GSE89632\_series\_matrix.txt.gz  
  
#US\_2  
data\_u2 <- getGEO("GSE163211")

Found 1 file(s)

GSE163211\_series\_matrix.txt.gz

data\_u2 <- data\_u2$GSE163211\_series\_matrix.txt.gz

# Preprocessing

## Get to know the data

#US\_1  
# dimension of data  
dim(data\_u)

Features Samples   
 29377 63

# check for zeros expression values and/or NA, just in case  
zeros <- apply(exprs(data\_u), 1, function(x) sum(x==0))  
data\_u <- data\_u[zeros!=63,]  
dim(data\_u)

Features Samples   
 29377 63

#US\_2  
# dimension of data  
dim(data\_u2)

Features Samples   
 800 318

# check for zeros expression values and/or NA, just in case  
zeros <- apply(exprs(data\_u2), 1, function(x) sum(x==0))  
data\_u2 <- data\_u2[zeros!=318,]  
dim(data\_u2)

Features Samples   
 800 318

All datasets are cleaned and do not have zero expression values.

In the analysis, we will focus on gene expression values and its relationship with each other and phenotype characteristics.

# Differentially expressed analysis

## Andrent et al (USA data 1)

### Get values for progression of NAFLD

# Values for diabetes   
data\_uu <- data\_u[, !is.na(data\_u@phenoData@data[["characteristics\_ch1.22"]]) & data\_u@phenoData@data[["characteristics\_ch1.22"]] == "diabetes: yes" | data\_u@phenoData@data[["characteristics\_ch1.22"]] == "diabetes: no" ]   
  
data\_uu$diabetes <- ifelse(data\_uu@phenoData@data[["characteristics\_ch1.22"]] == "diabetes: no", '0', '1')

# Values for progression of NAFLD   
data\_uu$stages <- gsub("[^:]+: (.\*)", "\\1", data\_uu$characteristics\_ch1.1)  
  
data\_uu$stages <- ifelse(data\_uu$stages == 'HC', 1,  
 ifelse(data\_uu$stages == 'NASH', 2,  
 ifelse(data\_uu$stages == 'SS', 3, data\_uu$stages)))  
  
  
   
  
# get the back-transformed data into a new variable  
data\_u\_expr = data\_uu@assayData$exprs   
data\_u\_expr = round((2^data\_u\_expr-1),0)

# Fit DESeq2   
dds\_u <- DESeqDataSetFromMatrix(countData = data\_u\_expr, colData = pData(data\_uu), design=~stages+diabetes)

converting counts to integer mode

Warning in DESeqDataSet(se, design = design, ignoreRank): some variables in  
design formula are characters, converting to factors

dds\_u <- DESeq(dds\_u)

estimating size factors

estimating dispersions

gene-wise dispersion estimates

mean-dispersion relationship

-- note: fitType='parametric', but the dispersion trend was not well captured by the  
 function: y = a/x + b, and a local regression fit was automatically substituted.  
 specify fitType='local' or 'mean' to avoid this message next time.

final dispersion estimates

fitting model and testing

-- replacing outliers and refitting for 451 genes  
-- DESeq argument 'minReplicatesForReplace' = 7   
-- original counts are preserved in counts(dds)

estimating dispersions

fitting model and testing

## Compare healthy vs simple steatosis

# get the result and summary  
head(dds\_u)

class: DESeqDataSet   
dim: 6 29   
metadata(1): version  
assays(6): counts mu ... replaceCounts replaceCooks  
rownames(6): ILMN\_1343291 ILMN\_1651209 ... ILMN\_1651235 ILMN\_1651236  
rowData names(31): baseMean baseVar ... maxCooks replace  
colnames(29): GSM2385720 GSM2385723 ... GSM2385771 GSM2385773  
colData names(86): title geo\_accession ... sizeFactor replaceable

dds.results\_u <- results(dds\_u, contrast = c('stages', "1", "2"))   
summary(dds.results\_u, alpha = 0.05) # p-value = 0.05

out of 29377 with nonzero total read count  
adjusted p-value < 0.05  
LFC > 0 (up) : 2587, 8.8%  
LFC < 0 (down) : 1031, 3.5%  
outliers [1] : 292, 0.99%  
low counts [2] : 2278, 7.8%  
(mean count < 280)  
[1] see 'cooksCutoff' argument of ?results  
[2] see 'independentFiltering' argument of ?results

# Only get differential expressed genes (p-val <= 0.05 and log2FC outside of [-2,2])  
dds.results\_u.h\_ss <- dds.results\_u[!is.na(dds.results\_u$padj) & dds.results\_u$padj <= 0.05 & dds.results\_u$log2FoldChange > 2 | dds.results\_u$log2FoldChange < -2 ,]  
  
head(dds.results\_u.h\_ss)

log2 fold change (MLE): stages 1 vs 2   
Wald test p-value: stages 1 vs 2   
DataFrame with 6 rows and 6 columns  
 baseMean log2FoldChange lfcSE stat pvalue  
 <numeric> <numeric> <numeric> <numeric> <numeric>  
ILMN\_1651498 3177.259 2.35824 0.568095 4.15113 3.30844e-05  
ILMN\_1651838 2099.424 2.25276 0.458222 4.91631 8.81914e-07  
ILMN\_1652464 994.163 2.05476 0.509094 4.03611 5.43439e-05  
ILMN\_1652866 961.951 2.14865 0.555177 3.87020 1.08744e-04  
ILMN\_1653447 1787.750 2.34395 0.684971 3.42197 6.21691e-04  
ILMN\_1656812 1147.062 2.56844 0.660839 3.88664 1.01642e-04  
 padj  
 <numeric>  
ILMN\_1651498 0.001636334  
ILMN\_1651838 0.000115324  
ILMN\_1652464 0.002255105  
ILMN\_1652866 0.003676049  
ILMN\_1653447 0.011703411  
ILMN\_1656812 0.003490106

Number of upregulated and downregulated transcripts

#upregulated  
sum(dds.results\_u.h\_ss$log2FoldChange > 0)

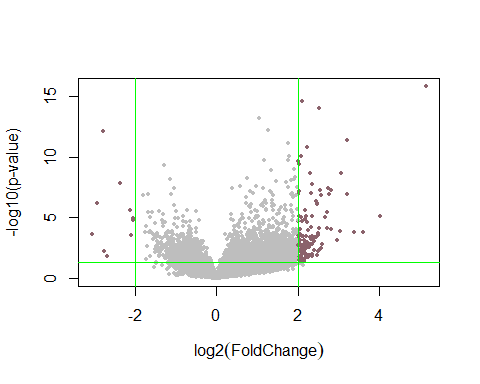
[1] 133

#downregulated  
sum(dds.results\_u.h\_ss$log2FoldChange < 0)

[1] 10

Plot - Healthy versus SS

plot(dds.results\_u$log2FoldChange, -log10(dds.results\_u$padj), col = c("gray","pink4", "blue")[(dds.results\_u$padj < 0.05 & abs(dds.results\_u$log2FoldChange) > 2) + 1 ], xlab = expression(log2(FoldChange)), ylab = "-log10(p-value)", cex = 0.8, pch = 20)  
abline(v = c(-2, 2), col = "green")  
abline(h = -log10(0.05), col = "green")



### Compare healthy versus NASH

# get the result and summary  
head(dds\_u)

class: DESeqDataSet   
dim: 6 29   
metadata(1): version  
assays(6): counts mu ... replaceCounts replaceCooks  
rownames(6): ILMN\_1343291 ILMN\_1651209 ... ILMN\_1651235 ILMN\_1651236  
rowData names(31): baseMean baseVar ... maxCooks replace  
colnames(29): GSM2385720 GSM2385723 ... GSM2385771 GSM2385773  
colData names(86): title geo\_accession ... sizeFactor replaceable

dds.results\_u\_h\_nash <- results(dds\_u, contrast = c('stages', "1", "3"))   
summary(dds.results\_u\_h\_nash, alpha = 0.05) # p-value = 0.05

out of 29377 with nonzero total read count  
adjusted p-value < 0.05  
LFC > 0 (up) : 484, 1.6%  
LFC < 0 (down) : 227, 0.77%  
outliers [1] : 292, 0.99%  
low counts [2] : 6259, 21%  
(mean count < 414)  
[1] see 'cooksCutoff' argument of ?results  
[2] see 'independentFiltering' argument of ?results

# Only get differential expressed genes (p-val <= 0.05 and log2FC outside of [-2,2])  
dds.results\_u.h\_nash <- dds.results\_u[!is.na(dds.results\_u\_h\_nash$padj) & dds.results\_u\_h\_nash$padj <= 0.05 & dds.results\_u\_h\_nash$log2FoldChange > 2 | dds.results\_u\_h\_nash$log2FoldChange < -2 ,]  
  
head(dds.results\_u.h\_nash)

log2 fold change (MLE): stages 1 vs 2   
Wald test p-value: stages 1 vs 2   
DataFrame with 6 rows and 6 columns  
 baseMean log2FoldChange lfcSE stat pvalue  
 <numeric> <numeric> <numeric> <numeric> <numeric>  
ILMN\_1661178 464.023 2.2262018 0.284945 7.8127503 5.59534e-15  
ILMN\_1670669 1661.999 -0.0113593 0.711730 -0.0159602 9.87266e-01  
ILMN\_1672295 1456.443 2.4743108 0.413143 5.9889921 2.11145e-09  
ILMN\_1680618 2919.909 2.8115637 0.431878 6.5100852 7.51082e-11  
ILMN\_1694472 4282.311 -3.0537817 0.647165 -4.7187091 2.37346e-06  
ILMN\_1717639 2299.082 2.5146822 0.500098 5.0283780 4.94646e-07  
 padj  
 <numeric>  
ILMN\_1661178 1.66660e-11  
ILMN\_1670669 9.93982e-01  
ILMN\_1672295 8.20315e-07  
ILMN\_1680618 5.29849e-08  
ILMN\_1694472 2.41908e-04  
ILMN\_1717639 7.09089e-05

#upregulated  
sum(dds.results\_u.h\_nash$log2FoldChange > 0)

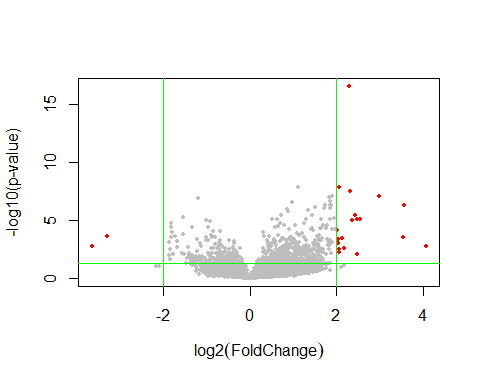
[1] 19

#downregulated  
sum(dds.results\_u.h\_nash$log2FoldChange < 0)

[1] 4

Plot healthy versus NASH

plot(dds.results\_u\_h\_nash$log2FoldChange, -log10(dds.results\_u\_h\_nash$padj), col = c("gray","red", "blue")[(dds.results\_u\_h\_nash$padj < 0.05 & abs(dds.results\_u\_h\_nash$log2FoldChange) > 2) + 1 ], xlab = expression(log2(FoldChange)), ylab = "-log10(p-value)", cex = 0.8, pch = 20)  
abline(v = c(-2, 2), col = "green")  
abline(h = -log10(0.05), col = "green")



## Gene Identification Analysis (still working on)

We want to find which molecular function pathway associated with those differentially expressed genes.

#### 1. Healthy versus SS

# FDR <= 0.05 and foldchange outside abs 2  
sigs\_FC\_u <- dds.results\_u[!is.na(dds.results\_u$padj) & dds.results\_u$padj < 0.05 &( dds.results\_u$log2FoldChange > 2 | dds.results\_u$log2FoldChange < -2) ,]  
genes\_sigs\_FC\_u <- rownames(sigs\_FC\_u)  
  
  
# Gene identification  
data.frame(Gene=unlist(mget(x = genes\_sigs\_FC\_u,envir = illuminaHumanv4SYMBOL)))

Gene  
ILMN\_1651498 GADD45G  
ILMN\_1651838 RND1  
ILMN\_1652464 TUBA3E  
ILMN\_1652866 LMX1B  
ILMN\_1653447 PSORS1C2  
ILMN\_1656812 NMRK2  
ILMN\_1657148 CIRBP-AS1  
ILMN\_1658632 NLRP7  
ILMN\_1660086 MYH11  
ILMN\_1661178 NR4A1  
ILMN\_1661197 CLCF1  
ILMN\_1665372 FXYD3  
ILMN\_1665471 MAB21L4  
ILMN\_1669523 FOS  
ILMN\_1670019 PEG10  
ILMN\_1672148 AKR1B10  
ILMN\_1672295 ZC3H12A  
ILMN\_1673566 ADAMTS1  
ILMN\_1674353 MIOX  
ILMN\_1680139 MAFF  
ILMN\_1680399 KAZN  
ILMN\_1680618 MYC  
ILMN\_1682599 GPRC5A  
ILMN\_1684401 FMO1  
ILMN\_1687481 HIF3A  
ILMN\_1690253 SYNPO2L  
ILMN\_1691881 IL4R  
ILMN\_1692280 NKX2-2  
ILMN\_1692332 ALOX12B  
ILMN\_1694034 LGALS4  
ILMN\_1694472 GCK  
ILMN\_1696666 DBF4B  
ILMN\_1697499 HLA-DRB5  
ILMN\_1698846 SLC8A2  
ILMN\_1699651 IL6  
ILMN\_1706186 PIANP  
ILMN\_1711331 MAP1LC3C  
ILMN\_1712506 DPP6  
ILMN\_1715748 FLNC  
ILMN\_1717639 SIK1  
ILMN\_1721449 VSX2  
ILMN\_1722825 LINC00304  
ILMN\_1723522 APOLD1  
ILMN\_1727364 TRIM7  
ILMN\_1727589 SULT2B1  
ILMN\_1728677 CREB5  
ILMN\_1731742 TNFRSF13C  
ILMN\_1732143 LINC00173  
ILMN\_1736969 SLITRK3  
ILMN\_1738401 FOXC1  
ILMN\_1740426 RASD1  
ILMN\_1740915 <NA>  
ILMN\_1740917 SCNN1B  
ILMN\_1743445 FAM107A  
ILMN\_1745395 CTCFL  
ILMN\_1748751 C2CD4B  
ILMN\_1748915 S100A12  
ILMN\_1749575 CYP7A1  
ILMN\_1749984 <NA>  
ILMN\_1751495 FGF14  
ILMN\_1751607 FOSB  
ILMN\_1755796 BEST2  
ILMN\_1758066 DSCR8  
ILMN\_1758529 P2RX1  
ILMN\_1759787 THBD  
ILMN\_1767129 ABCC8  
ILMN\_1767643 CTRB2  
ILMN\_1768099 SCGB1C1  
ILMN\_1768469 TCN1  
ILMN\_1769471 <NA>  
ILMN\_1769492 CACNG2  
ILMN\_1769891 CDH4  
ILMN\_1770014 KCNQ4  
ILMN\_1770711 TNFRSF8  
ILMN\_1771841 FOSL1  
ILMN\_1773459 SOX11  
ILMN\_1774350 MYOZ3  
ILMN\_1774705 ELOA2  
ILMN\_1774733 SOCS1  
ILMN\_1775257 PROK2  
ILMN\_1777658 SCARF1  
ILMN\_1778668 TAGLN  
ILMN\_1778681 EBF1  
ILMN\_1779043 CACNG6  
ILMN\_1780170 APOD  
ILMN\_1781001 SOCS3  
ILMN\_1781812 NR4A3  
ILMN\_1782305 NR4A2  
ILMN\_1782704 CD19  
ILMN\_1789112 TMEM145  
ILMN\_1793570 <NA>  
ILMN\_1794959 SLC35F3  
ILMN\_1795325 ACTG2  
ILMN\_1799329 TTLL10  
ILMN\_1804993 FAM178B  
ILMN\_1807034 CALCA  
ILMN\_1810054 CNN1  
ILMN\_1815023 PIM1  
ILMN\_2054297 PTGS2  
ILMN\_2086077 JUNB  
ILMN\_2108735 EEF1A2  
ILMN\_2110908 MYC  
ILMN\_2118129 ITLN2  
ILMN\_2137789 KLF4  
ILMN\_2149815 CYP2F1  
ILMN\_2159721 <NA>  
ILMN\_2163790 <NA>  
ILMN\_2184373 CXCL8  
ILMN\_2230178 DAND5  
ILMN\_2232463 ARL14  
ILMN\_2242900 IL1RL1  
ILMN\_2264177 <NA>  
ILMN\_2275760 TFAP2A  
ILMN\_2277877 MEN1  
ILMN\_2291010 PSMF1  
ILMN\_2296450 GH2  
ILMN\_2297626 PEG10  
ILMN\_2313672 IL1RL1  
ILMN\_2336094 TENM3  
ILMN\_2339955 NR4A2  
ILMN\_2365091 FCAR  
ILMN\_2368834 MYH11  
ILMN\_2373670 UGT2B17  
ILMN\_2394841 NKX2-1  
ILMN\_2408566 NR4A1  
ILMN\_2410145 NR4A1  
ILMN\_3232161 FAM86B3P  
ILMN\_3238130 <NA>  
ILMN\_3238990 USP17L9P  
ILMN\_3240236 SMCR5  
ILMN\_3241034 <NA>  
ILMN\_3243129 <NA>  
ILMN\_3243788 CATSPERD  
ILMN\_3246489 SBK2  
ILMN\_3248631 PTX4  
ILMN\_3249444 POM121L4P  
ILMN\_3250412 CNGB1  
ILMN\_3251332 POM121L10P  
ILMN\_3307841 AGR2  
ILMN\_3308856 <NA>  
ILMN\_3308986 MIR1272  
ILMN\_3310010 MIR146B  
ILMN\_3310840 MIR21

#### 2. Healthy versus NASH

# FDR <= 0.05 and foldchange outside abs 2  
sigs\_FC\_u\_h\_nash <- dds.results\_u[!is.na(dds.results\_u\_h\_nash$padj) & dds.results\_u\_h\_nash$padj <= 0.05 &( dds.results\_u\_h\_nash$log2FoldChange > 2 | dds.results\_u\_h\_nash$log2FoldChange < -2) ,]  
genes\_sigs\_FC\_u\_h\_nash <- rownames(sigs\_FC\_u\_h\_nash)  
  
  
# Gene identification  
data.frame(Gene=unlist(mget(x = genes\_sigs\_FC\_u\_h\_nash,envir = illuminaHumanv4SYMBOL)))

Gene  
ILMN\_1661178 NR4A1  
ILMN\_1672295 ZC3H12A  
ILMN\_1680618 MYC  
ILMN\_1694472 GCK  
ILMN\_1717639 SIK1  
ILMN\_1723522 APOLD1  
ILMN\_1728677 CREB5  
ILMN\_1751607 FOSB  
ILMN\_1768469 TCN1  
ILMN\_1781001 SOCS3  
ILMN\_1785732 TNFAIP6  
ILMN\_1807034 CALCA  
ILMN\_2086077 JUNB  
ILMN\_2110908 MYC  
ILMN\_2184373 CXCL8  
ILMN\_2196479 XRN2  
ILMN\_2230178 DAND5  
ILMN\_2232463 ARL14  
ILMN\_2339955 NR4A2  
ILMN\_2373670 UGT2B17  
ILMN\_2410145 NR4A1

Similar genes between these two analyses

h\_ss\_h\_nash <- intersect(unlist(mget(x = genes\_sigs\_FC\_u\_h\_nash,envir = illuminaHumanv4SYMBOL)), unlist(mget(x = genes\_sigs\_FC\_u,envir = illuminaHumanv4SYMBOL)))

Molecular function pathway

# run goseq, find molecular function pathways  
GO\_results\_sigs <- enrichGO(gene = h\_ss\_h\_nash, OrgDb = "org.Hs.eg.db", keyType = "SYMBOL", ont = "MF")  
  
# check results  
dfGO\_sigs <- data.frame(GO\_results\_sigs@result) # transform the result file to dataframe  
head(dfGO\_sigs)

ID  
GO:0001228 GO:0001228  
GO:0001216 GO:0001216  
GO:0035259 GO:0035259  
GO:0035198 GO:0035198  
GO:0004879 GO:0004879  
GO:0098531 GO:0098531  
 Description  
GO:0001228 DNA-binding transcription activator activity, RNA polymerase II-specific  
GO:0001216 DNA-binding transcription activator activity  
GO:0035259 nuclear glucocorticoid receptor binding  
GO:0035198 miRNA binding  
GO:0004879 nuclear receptor activity  
GO:0098531 ligand-activated transcription factor activity  
 GeneRatio BgRatio pvalue p.adjust qvalue  
GO:0001228 5/17 468/18369 5.046270e-05 0.001625130 0.0009649895  
GO:0001216 5/17 472/18369 5.254974e-05 0.001625130 0.0009649895  
GO:0035259 2/17 13/18369 6.250500e-05 0.001625130 0.0009649895  
GO:0035198 2/17 37/18369 5.267705e-04 0.009171301 0.0054458468  
GO:0004879 2/17 46/18369 8.146316e-04 0.009171301 0.0054458468  
GO:0098531 2/17 46/18369 8.146316e-04 0.009171301 0.0054458468  
 geneID Count  
GO:0001228 NR4A1/MYC/FOSB/JUNB/NR4A2 5  
GO:0001216 NR4A1/MYC/FOSB/JUNB/NR4A2 5  
GO:0035259 NR4A1/NR4A2 2  
GO:0035198 ZC3H12A/SOCS3 2  
GO:0004879 NR4A1/NR4A2 2  
GO:0098531 NR4A1/NR4A2 2

## For US 2

# data\_u2\_design <- data\_u2@phenoData$`nafld stage:ch1`

For diabetes

# Values for diabetes   
data\_u2u <- data\_u2[, !is.na(data\_u2@phenoData@data[["characteristics\_ch1.5"]]) & data\_u2@phenoData@data[["characteristics\_ch1.5"]] == "diabetes: Yes" | data\_u2@phenoData@data[["characteristics\_ch1.5"]] == "diabetes: No" ]   
  
data\_u2u$diabetes <- ifelse(data\_u2u@phenoData@data[["characteristics\_ch1.5"]] == "diabetes: No", '0', '1')

We would not need back transform for this dataset since it has raw counts already

# Fit DESeq2   
#dds\_u2 <- DESeqDataSetFromMatrix(round(exprs(data\_u2)), colData = pData(data\_u2), design=~data\_u2\_design)  
#dds\_u2 <- DESeq(dds\_u2)