Real_dataset_tutorial

Code

1.1 Data and Goals

Human embryonic stem cells (hESCs) typically exhibit “primed” pluripotency, analogous to stem cells derived from the mouse post-implantation epiblast. Since primed hESC have limited differentiation capacity, scientists have tried various method trying to revert “primed” hESCs to a more “naive” state which have higher pluripotency capacity (could have more clinical application potentials, and can also help us understand the early embryo development). By investigating the differentially expressed genes and their involved pathways, we can understand the mechanism undelying the differentiation capacity difference between naive hESCs and primed hESCs.

In this tutorial, we will use the 8 RNAseq dataset from *William Pastor et al., 2016, Cell Stem Cell*, with 4 replicates of naive hESCs and 4 replicates of primed hESCs.

Accession	ID	Replicate	CellType
GSM2041708	1	rep1	Primed_hESC
GSM2041709	2	rep2	Primed_hESC
GSM2041710	3	rep3	Primed_hESC
GSM2041711	4	rep4	Primed_hESC
GSM2041712	5	rep1	Naive_hESC
GSM2041713	6	rep2	Naive_hESC
GSM2041714	7	rep3	Naive_hESC
GSM2041715	8	rep4	Naive_hESC

Hide

```
#read in the primed hESC RNAseq raw count
primed_hESC_rep1 <- read.delim('/Users/jesi/Documents/real_Data/GSM2041708_RNAseq_UCLA1_
Primed_rep1_readsCount.txt',row.names = 1)
primed_hESC_rep2 <- read.delim('/Users/jesi/Documents/real_Data/GSM2041709_RNAseq_UCLA1_
Primed_rep2_readsCount.txt',,row.names = 1)
primed_hESC_rep3  <-read.delim('/Users/jesi/Documents/real_Data/GSM2041710_RNAseq_UCLA1_
Primed_rep3_readsCount.txt',row.names = 1)
primed_hESC_rep4  <-read.delim('/Users/jesi/Documents/real_Data/GSM2041711_RNAseq_UCLA1_
Primed_rep4_readsCount.txt',row.names = 1)

#read in the naive hESC RNAseq raw count
naive_hESC_rep1 <- read.delim('/Users/jesi/Documents/real_Data/GSM2041712_RNAseq_SSEA4_n
eg_rep1_readCounts.txt',row.names = 1)
naive_hESC_rep2 <- read.delim('/Users/jesi/Documents/real_Data/GSM2041713_RNAseq_SSEA4_n
eg_rep2_readCounts.txt',row.names = 1)
naive_hESC_rep3  <-read.delim('/Users/jesi/Documents/real_Data/GSM2041714_RNAseq_SSEA4_n
eg_rep3_readCounts.txt',row.names = 1)
naive_hESC_rep4  <-read.delim('/Users/jesi/Documents/real_Data/GSM2041715_RNAseq_SSEA4_n
eg_rep4_readsCount.txt',row.names = 1)

head(primed_hESC_rep4)
```

	X116 <int>
A1BG	18
A1BG-AS1	15
A1CF	13
A2LD1	30
A2M	36
A2ML1	393

6 rows

Hide

```
readcount = data.frame(naive_hESC_rep1,
                       naive_hESC_rep2,
                       naive_hESC_rep3,
                       naive_hESC_rep4,
                       primed_hESC_rep1,
                       primed_hESC_rep2,
                       primed_hESC_rep3,
                       primed_hESC_rep4)

colnames(readcount) = c(paste(rep('naive_hESC_rep',4),c(1:4),sep=''),paste(rep('primed_h
ESC_rep',4),c(1:4),sep=''))
# convert variable name as str: deparse(substitute(data))
```

Hide

```
remotes::install_github("Zjx01/Generalized-RNAseq-analysis-pipeline")
```

```
Error: Failed to install 'unknown package' from GitHub:
  HTTP error 404.
  Not Found

  Did you spell the repo owner (`Zjx01`) and repo name (`Generalized-RNAseq-analysis-pip
eline`) correctly?
  - If spelling is correct, check that you have the required permissions to access the r
epo.
```

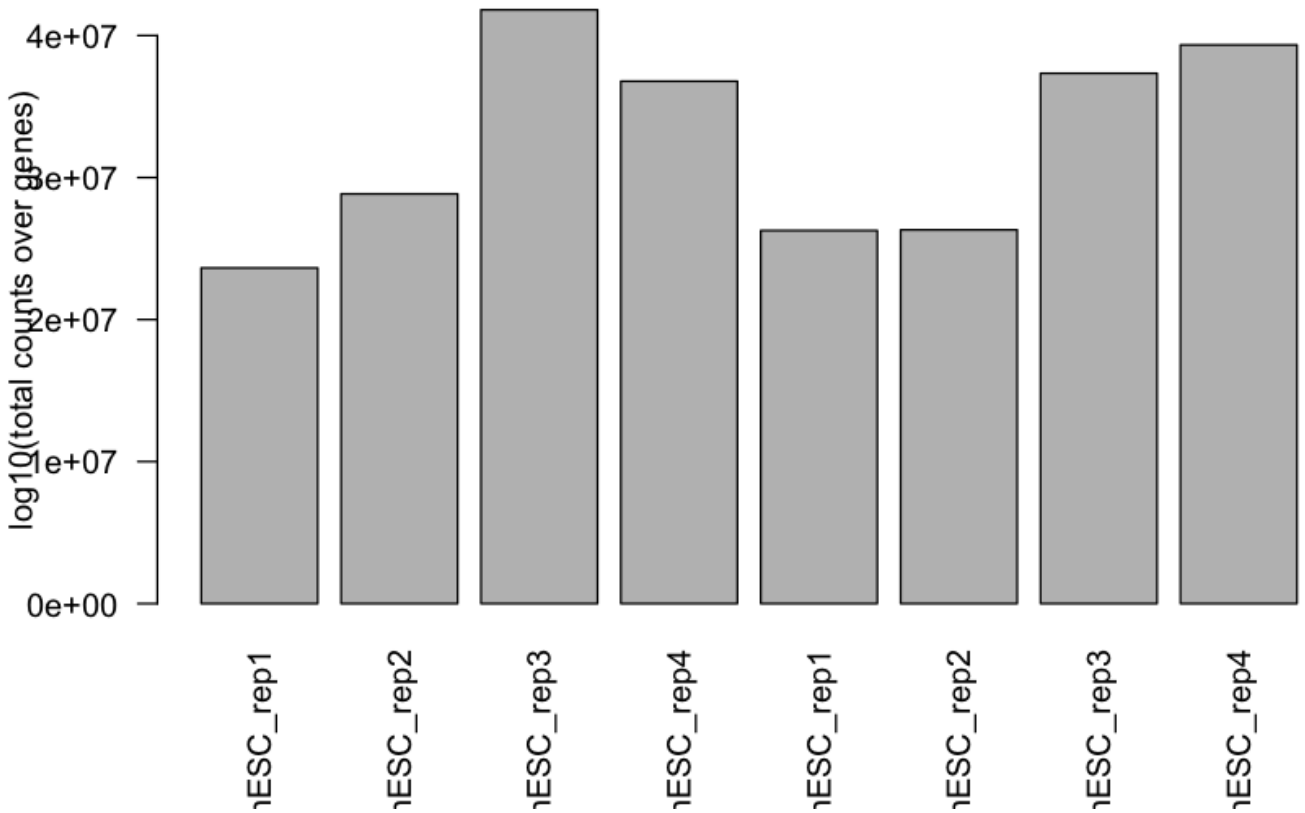
Hide

```
library(grnaeR)
library(DESeq2)
library('ggplot2')
library("pheatmap")
library("RColorBrewer")
library('AnnotationDbi')
library('org.Hs.eg.db')
```

1.2 Check the data quality

Hide

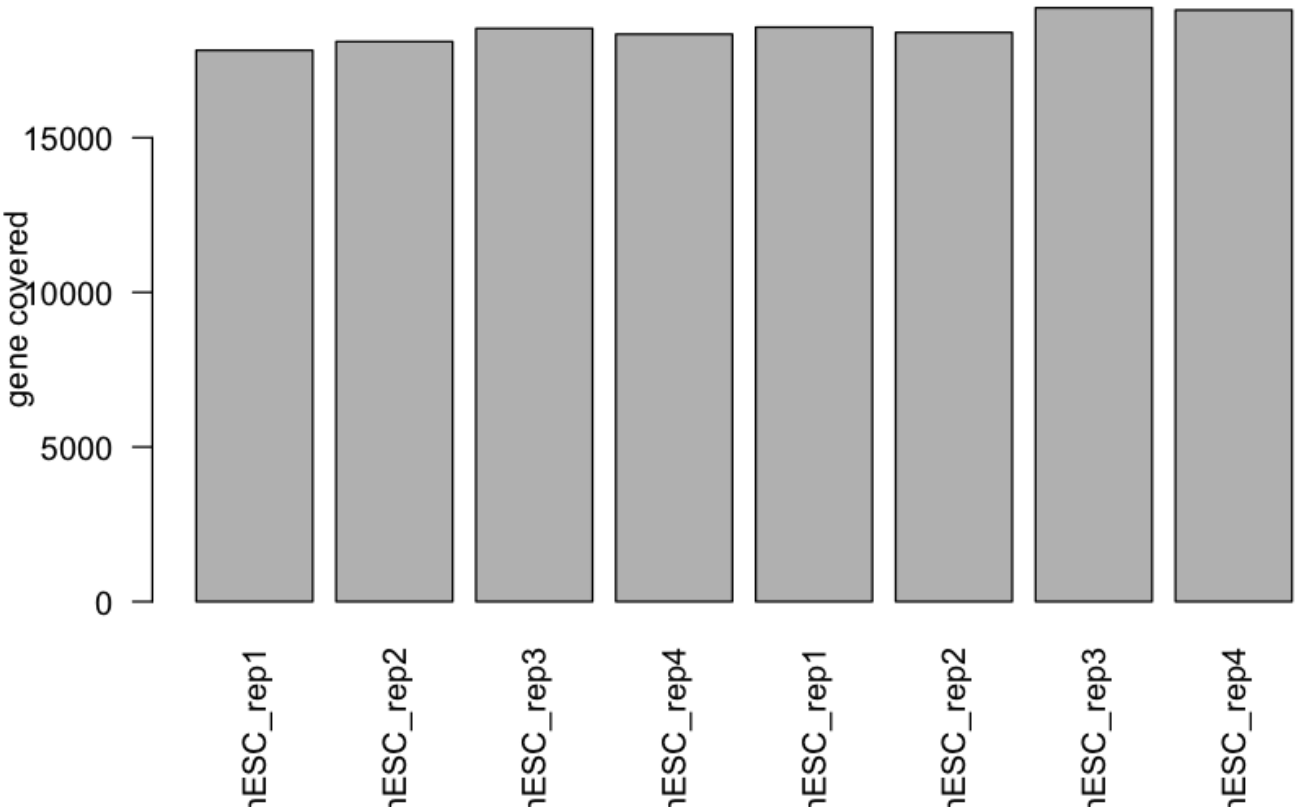
```
total.cov = check_totalcov_quality(readcount)
```



From this plot, we can see that the coverage of all those 8 samples are in the same magnitude indicating they got sequenced evenly.

Hide

```
gene.cov = check_genecovered_quality(readcount)
```



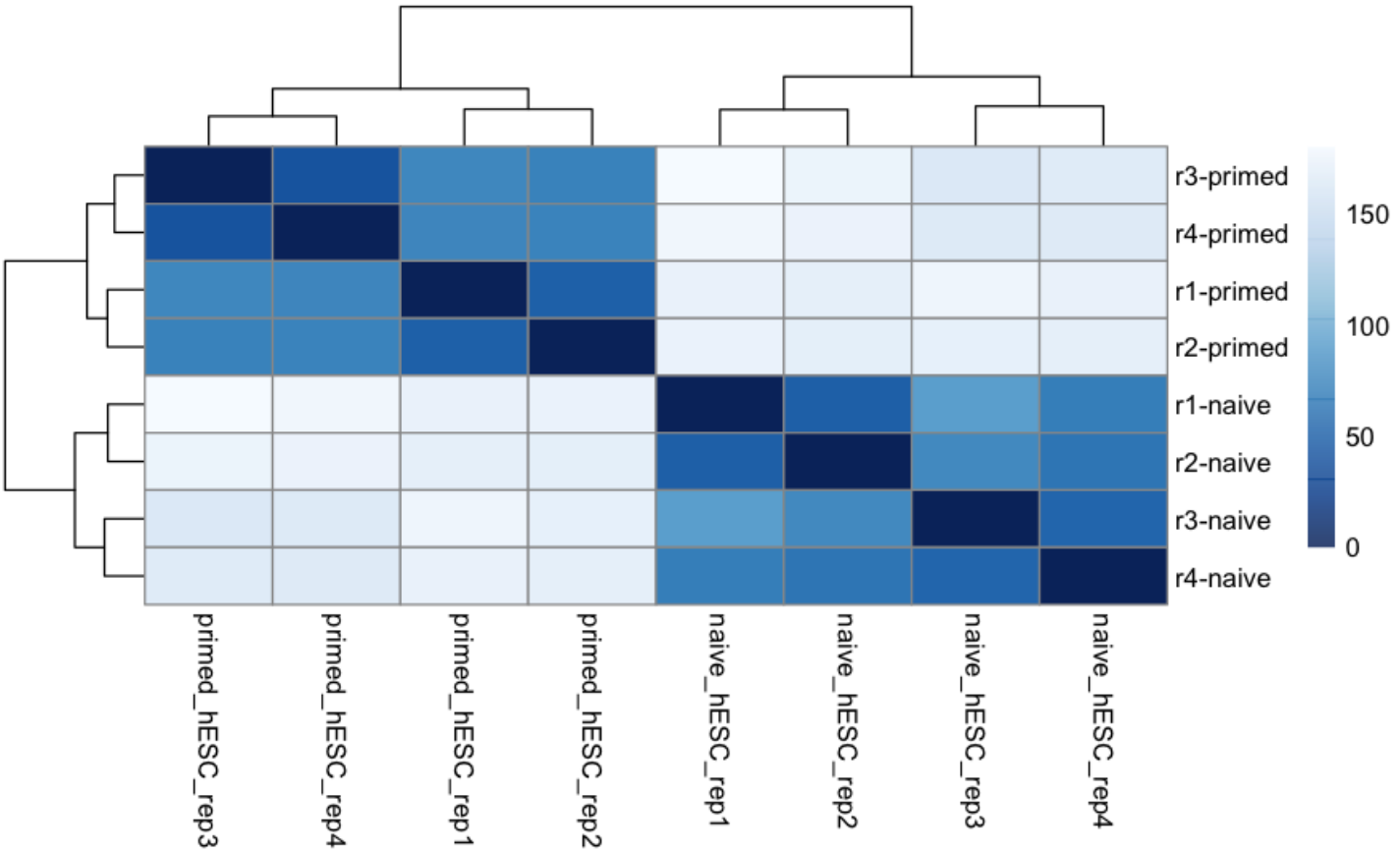
From this plot, we can see that most genes (18k+) are covered in all 8 samples, and the number of genes detected in each library is similar.

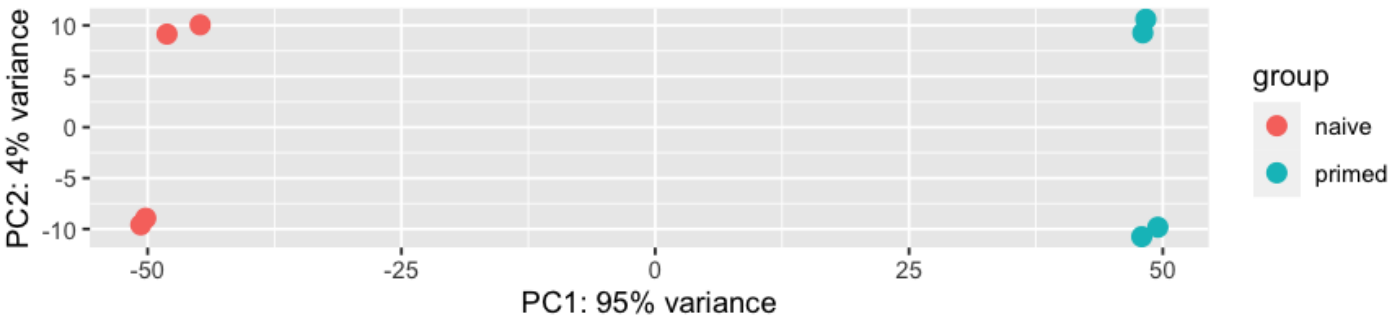
1.2 Load data into Deseq2

Hide

```
condition_vector = c(c(rep('naive',4)),c(rep('primed',4)))
type_vector = c(rep(paste(rep('r',4),c(1:4),sep=''),2))

dds = load_data_for_DESeq2(readcount,condition_vector,type_vector)
normalized_dds = normalize_dataset(dds)
check_sample_distance(normalized_dds)
```





Hide

```
select_DEGs = select_DEG(dds = dds,filter_thresh = 0,log2_fc = 1, padjust = 0.05)
```

```
[1] "filtering 2489 genes with low counts"
```

```
using pre-existing size factors
estimating dispersions
gene-wise dispersion estimates
mean-dispersion relationship
final dispersion estimates
fitting model and testing
```

```
out of 20883 with nonzero total read count
adjusted p-value < 0.05
LFC > 1.00 (up)    : 1988, 9.5%
LFC < -1.00 (down) : 1182, 5.7%
outliers [1]       : 1, 0.0048%
low counts [2]     : 2025, 9.7%
(mean count < 1)
[1] see 'cooksCutoff' argument of ?results
[2] see 'independentFiltering' argument of ?results
```

Here, we displayed the differentially expressed genes in the primed and naive human embryonic stem cells

Hide

```
select_DEGs
```

	baseMean <dbl>	log2FoldChange <dbl>	lfcSE <dbl>	stat <dbl>	pvalue <dbl>	padj <dbl>
A4GALT	398.511748	-4.370871	0.2506463	-13.448718	3.132572e-41	3.211801e-39
AADAC	3.615533	-5.495015	1.3855628	-3.244180	1.177893e-03	8.406033e-03
AADACL3	16.371728	3.365400	0.7923113	2.985443	2.831683e-03	1.884823e-02
AARS2	2285.624212	-2.587097	0.2117793	-7.494107	6.675119e-14	1.636836e-12
ABCA1	7578.004807	-3.240110	0.3196038	-7.009022	2.399900e-12	5.166086e-11
ABCA13	187.323098	-2.190139	0.2360508	-5.041876	4.609896e-07	5.622820e-06
ABCB1	3.524982	5.064268	1.2967236	3.134259	1.722885e-03	1.193112e-02
ABCB10	574.587749	-1.790584	0.1493826	-5.292343	1.207592e-07	1.600250e-06
ABCB4	10.869679	4.185474	0.8504698	3.745546	1.800021e-04	1.507907e-03
ABCB8	292.623634	-2.977612	0.2038041	-9.703498	2.913341e-22	1.222326e-20

1-10 of 3,170 rows Previous **1** 2 3 4 5 6 … 100 Next

Later, we try to convert the DEGS from SYMBOL IS to ENSEMBLE ID to enable the further visualization

Hide

```
gene_name = mapIds(org.Hs.eg.db,
                     keys = row.names(select_DEGs),
                     column = "ENTREZID",
                     keytype = "SYMBOL",
                     multiVals = "first")
```

```
'select()' returned 1:many mapping between keys and columns
```

Hide

```
target_genes = as.data.frame(cbind('symbol_name' = row.names(select_DEGs),gene_name))
target_genes <- target_genes[complete.cases(target_genes), ]

DEGS <- subset(normalized_dds@assays@data@listData[[1]],rownames(normalized_dds@assays@d
ata@listData[[1]]) %in% rownames(target_genes)==TRUE)
```

Hide

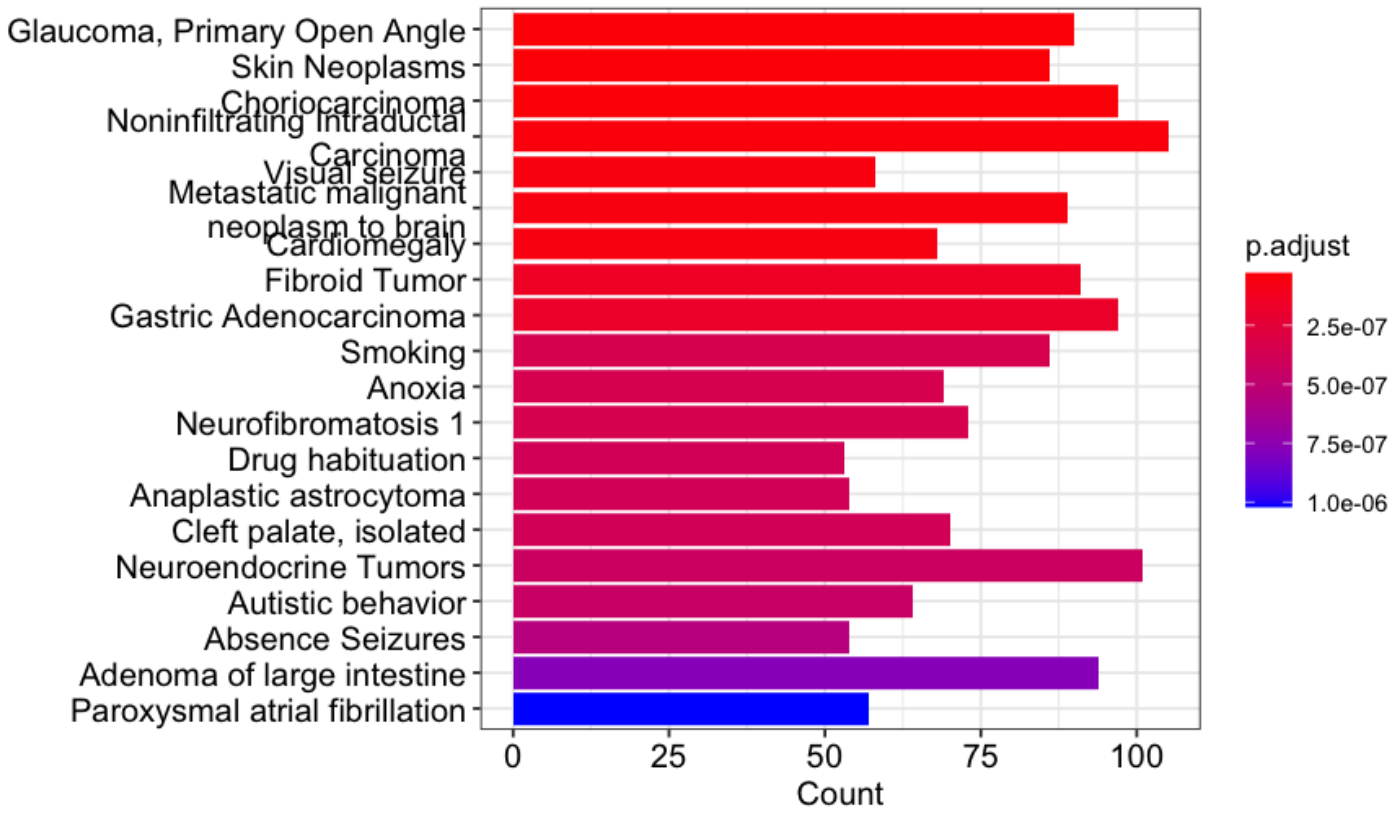
```
DEG_genename <- c()
for(i in rownames(DEGS)){
  genename = subset(target_genes,symbol_name == i)$gene_name
  DEG_genename <- c(DEG_genename,genename)
}
rownames(DEGS) <- DEG_genename
naive_mean <- rowMeans(DEGS[,1:4])
edo <- filter_genelist(naive_mean,standard_fc = 2)
```

```
[1] "enrichResult object generated"
```

We can see that the differentially expressed genes in naive cells are involved in following enriched terms. It depicts the enrichment scores (e.g. p values) and gene count or ratio as bar height and color.

Hide

```
barplot <- show_barplot(edo,showCategory_num = 20)
barplot
```



And we can view the over representation analysis and gene set enrichment analysis in the naive human embryonic stem cells, to see the enriched pathways.

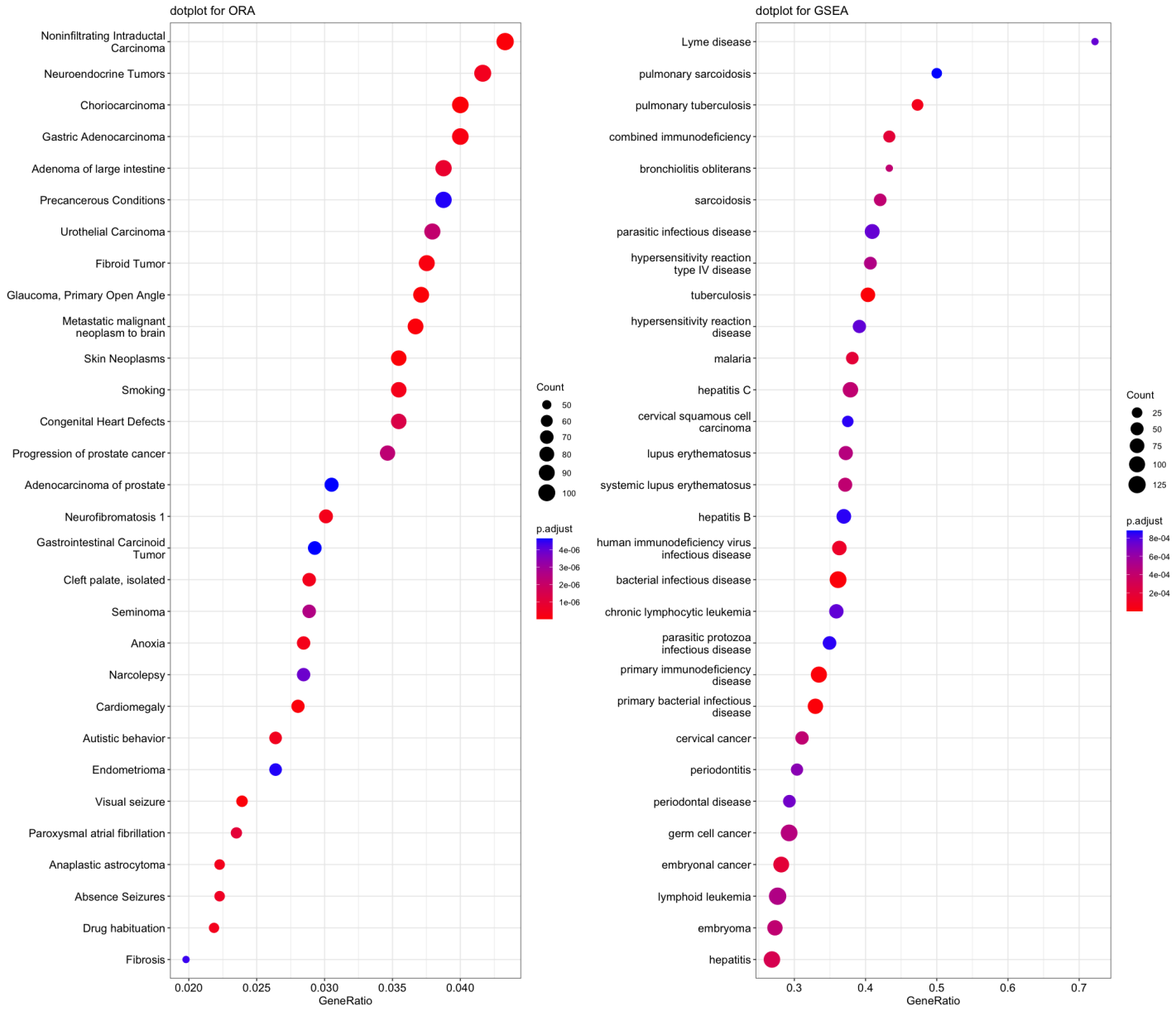
Hide

```
dotplot <- show_dotplot(edo,showCategory_num=30)
```

```
preparing geneSet collections...
GSEA analysis...
Warning: For some pathways, in reality P-values are less than 1e-10. You can set the `ep
s` argument to zero for better estimation.leading edge analysis...
done...
```

Hide

```
dotplot
```



Hide

```
gene_network <-develop_Gene_Network(edo,naive_mean)
```

```
Scale for size is already present.
Adding another scale for size, which will replace the existing scale.
```

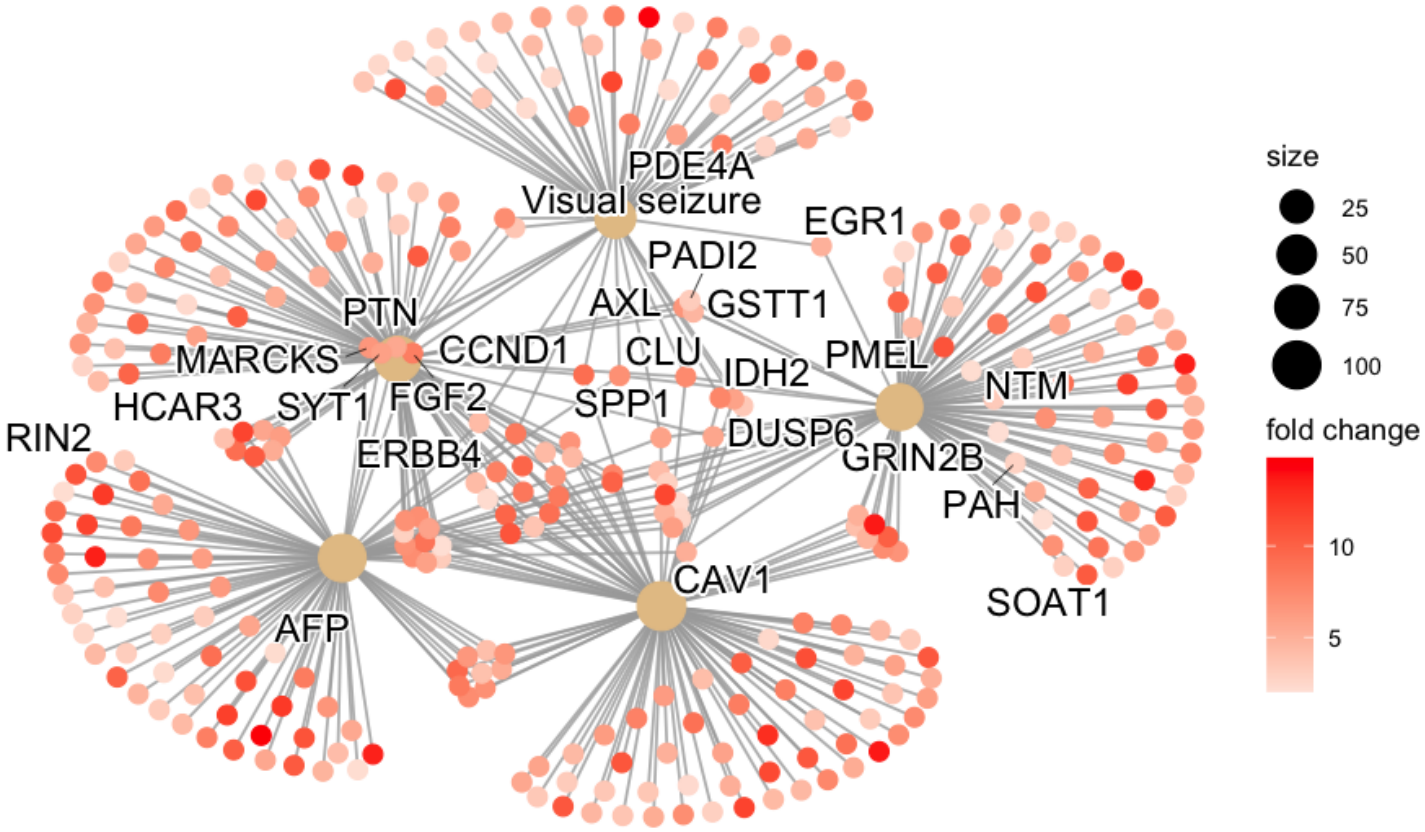
Hide

```
gene_network <-develop_Gene_Network(edo,naive_mean)
```

```
Scale for size is already present.
Adding another scale for size, which will replace the existing scale.
```

Hide

```
gene_network
```


PDE4A
Visual seizure
EGR1
PADI2
AXL
GSTT1
PTN
MARCKS
CCND1
CLU
IDH2
PMEL
NTM
HCAR3
SYT1
FGF2
SPP1
DUSP6
RIN2
ERBB4
GRIN2B
PAH
CAV1
SOAT1
AFP
size
25
50
75
100
fold change
10
5