```
library(DESeq2);load("CancerAndParacancer.RData")
```

```
## Loading required package: S4Vectors
```

```
## Loading required package: stats4
```

```
## 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, saveRDS, setdiff,
##     table, tapply, union, unique, unsplit, which.max, which.min
```

```
##
## 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
```

```
## Loading required package: GenomicRanges
```

```
## Loading required package: GenomeInfoDb
```

```
## Loading required package: SummarizedExperiment
```

```
## Loading required package: MatrixGenerics
```

```
## Loading required package: matrixStats
```

```
##
## 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
```

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

```
##
## Attaching package: 'Biobase'
```

```
## The following object is masked from 'package:MatrixGenerics':
##
##     rowMedians
```

```
## The following objects are masked from 'package:matrixStats':
##
##     anyMissing, rowMedians
```

```
expr <- ExprMat;pdat <- pDataMat;feat <- FeatureMat
same_samples <- intersect(colnames(expr), rownames(pdat))
expr <- expr[, same_samples, drop = F]
pdat <- pdat[same_samples, , drop = F]
expr <- expr[rowSums(expr) >= 954, , drop = F]
expr <- expr[!duplicated(rownames(expr)), , drop = F]
feat <- feat[!duplicated(rownames(feat)), , drop = F]
new_col <- intersect(c("gene_symbol", "symbol", "chromosome", "description", "gene_biotype"), colnames(feat))
feat <- feat[, new_col, drop = F]
feat <- feat[rownames(expr), , drop = F]
colnames(pdat)[1] <- "condition"
pdat$condition <- factor(pdat$condition)
pdat$condition <- relevel(pdat$condition, ref = "paracancerous")
dds <- DESeqDataSetFromMatrix(countData = expr, colData = pdat, design = ~ condition)
dds <- DESeq(dds);res <- results(dds)
```

```
## estimating size factors
```

```
## estimating dispersions
```

```
## gene-wise dispersion estimates
```

```
## mean-dispersion relationship
```

```
## final dispersion estimates
```

```
## fitting model and testing
```

```
## -- replacing outliers and refitting for 1371 genes
## -- DESeq argument 'minReplicatesForReplace' = 7
## -- original counts are preserved in counts(dds)
```

```
## estimating dispersions
```

```
## fitting model and testing
```

```
res_ordered <- res[order(res$padj, na.last = NA), ]
des_order <- as.data.frame(res_ordered)
normal_des <- as.data.frame(counts(dds, normalized = T))
des_order$gene <- rownames(des_order)
top_genes <- rownames(des_order)[1:15]
top_genes <- top_genes[top_genes %in% rownames(FeatureMat)]
cutoff <- min(rowSums(expr)[top_genes]);cat("cutoff is:", cutoff, "\n")
```

```
## cutoff is: 9296
```

```
top <- cbind(GeneID = top_genes,FeatureMat[top_genes, , drop = F],
  des_order[top_genes, c("log2FoldChange","pvalue", "padj","stat","lfcSE","baseMean"), drop = F])
up_order <- order(des_order$log2FoldChange, decreasing = T)
up_genes <- rownames(des_order)[up_order][1:10]
up_genes <- up_genes[up_genes %in% rownames(FeatureMat)]
cutoff <- min(rowSums(expr)[up_genes]);cat("cutoff is:", cutoff, "\n")
```

```
## cutoff is: 1043
```

```
top_up <- cbind(GeneID = up_genes,FeatureMat[up_genes, , drop = F],
  des_order[up_genes, c("log2FoldChange","pvalue", "padj","stat","lfcSE","baseMean"), drop = F])
down_order <- order(des_order$log2FoldChange, decreasing = F)
down_genes <- rownames(des_order)[down_order][1:10]
down_genes <- down_genes[down_genes %in% rownames(FeatureMat)]
cutoff <- min(rowSums(expr)[down_genes]);cat("cutoff is:", cutoff, "\n")
```

```
## cutoff is: 954
```

```
top_down <- cbind(GeneID = down_genes,FeatureMat[down_genes, , drop = F],
  des_order[down_genes, c("log2FoldChange","pvalue", "padj","stat","lfcSE","baseMean"), drop = F])
summary(top);knitr::kable(top)
```

```
##     GeneID              GeneID          Symbol          Description       
##  Length:15          Min.   :  1589   Length:15          Length:15         
##  Class :character   1st Qu.:  4898   Class :character   Class :character  
##  Mode  :character   Median :  9340   Mode  :character   Mode  :character  
##                     Mean   : 79939                                        
##                     3rd Qu.:182952                                        
##                     Max.   :286133                                        
##    Synonyms           GeneType         EnsemblGeneID         Status         
##  Length:15          Length:15          Length:15          Length:15         
##  Class :character   Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character  
##                                                                             
##                                                                             
##                                                                             
##     ChrAcc            ChrStart           ChrStop          Orientation       
##  Length:15          Length:15          Length:15          Length:15         
##  Class :character   Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character  
##                                                                             
##                                                                             
##                                                                             
##      Length      GOFunctionID       GOProcessID        GOComponentID     
##  Min.   :  947   Length:15          Length:15          Length:15         
##  1st Qu.: 2186   Class :character   Class :character   Class :character  
##  Median : 2429   Mode  :character   Mode  :character   Mode  :character  
##  Mean   : 4340                                                           
##  3rd Qu.: 5370                                                           
##  Max.   :13899                                                           
##   GOFunction         GOProcess         GOComponent        log2FoldChange  
##  Length:15          Length:15          Length:15          Min.   :-7.681  
##  Class :character   Class :character   Class :character   1st Qu.:-4.934  
##  Mode  :character   Mode  :character   Mode  :character   Median :-4.777  
##                                                           Mean   :-4.722  
##                                                           3rd Qu.:-4.088  
##                                                           Max.   :-3.357  
##      pvalue              padj               stat            lfcSE       
##  Min.   :0.000e+00   Min.   :0.000e+00   Min.   :-19.64   Min.   :0.2385  
##  1st Qu.:0.000e+00   1st Qu.:0.000e+00   1st Qu.:-15.38   1st Qu.:0.2853  
##  Median :2.520e-45   Median :6.770e-42   Median :-14.13   Median :0.3125  
##  Mean   :4.419e-42   Mean   :6.613e-39   Mean   :-14.87   Mean   :0.3156  
##  3rd Qu.:6.347e-43   3rd Qu.:1.164e-39   3rd Qu.:-13.75   3rd Qu.:0.3529  
##  Max.   :3.319e-41   Max.   :4.766e-38   Max.   :-13.44   Max.   :0.3911  
##     baseMean     
##  Min.   : 136.3  
##  1st Qu.: 217.3  
##  Median : 435.4  
##  Mean   :1081.2  
##  3rd Qu.:1428.7  
##  Max.   :4452.2  
```

	GeneID	GeneID	Symbol	Description	Synonyms	GeneType	EnsemblGeneID	Status	ChrAcc
221476	221476	221476	PI16	peptidase inhibitor 16	CD364|CRISP9|MSMBBP|PSPBP	protein-coding	ENSG00000164530	active	NC_000
1675	1675	1675	CFD	complement factor D	ADIPSIN|ADN|DF|PFD	protein-coding	ENSG00000197766	active	NC_000
286133	286133	286133	SCARA5	scavenger receptor class A member 5	NET33|Tesr	protein-coding	ENSG00000168079	active	NC_000

	GeneID	GeneID	Symbol	Description	Synonyms	GeneType	EnsemblGeneID	Status	ChrAcc
7123	7123	7123	CLEC3B	C-type lectin domain family 3 member B	MCDR4|TN|TNA	protein-coding	ENSG00000163815	active	NC_000
9340	9340	9340	GLP2R	glucagon like peptide 2 receptor		protein-coding	ENSG00000065325	active	NC_000
11170	11170	11170	FAM107A	family with sequence similarity 107 member A	DRR1|TU3A	protein-coding	ENSG00000168309	active	NC_000
219348	219348	219348	PLAC9	placenta associated 9		protein-coding	ENSG00000189129	active	NC_000
146556	146556	146556	C16orf89	chromosome 16 open reading frame 89		protein-coding	ENSG00000153446	active	NC_000
221091	221091	221091	LRRN4CL	LRRN4 C-terminal like		protein-coding	ENSG00000177363	active	NC_000
1589	1589	1589	CYP21A2	cytochrome P450 family 21 subfamily A member 2	CA21H|CAH1|CPS1|CYP21|CYP21B|P450c21B	protein-coding	ENSG00000231852	active	NC_000
2674	2674	2674	GFRA1	GDNF family receptor alpha 1	GDNFR|GDNFR-alpha-1|GDNFRA|GFR-ALPHA-1|GFRalpha-1|RET1L|RETL1|RHDA4|TRNR1	protein-coding	ENSG00000151892	active	NC_000
1590	1590	1590	CYP21A1P	cytochrome P450 family 21 subfamily A member 1, pseudogene	CYP21A|CYP21P|P450c21A	pseudo	ENSG00000290788	active	NC_000
7146	7146	7146	TNXA	tenascin XA (pseudogene)	D6S103E|HXBL|TNX|XA	pseudo		active	NC_000

	GeneID	GeneID	Symbol	Description	Synonyms	GeneType	EnsemblGeneID	Status	ChrAcc
55022	55022	55022	PID1	phosphotyrosine interaction domain containing 1	HMFN2073|NYGGF4|P-CLI1|PCLI1	protein-coding	ENSG00000153823	active	NC_000
7148	7148	7148	TNXB	tenascin XB	EDS3|EDSCLL|EDSCLL1|HXBL|TENX|TN-X|TNX|TNXB1|TNXB2|TNXBS|VUR8|XB|XBS	protein-coding	ENSG00000168477	active	NC_000

```
summary(top_up);knitr::kable(top_up)
```

```
##     GeneID              GeneID             Symbol           Description       
##  Length:10          Min.   :     1472   Length:10          Length:10         
##  Class :character   1st Qu.:    19677   Class :character   Class :character  
##  Mode  :character   Median :   109906   Mode  :character   Mode  :character  
##                     Mean   : 41924921                                        
##                     3rd Qu.:101929231                                        
##                     Max.   :109729169                                        
##    Synonyms           GeneType         EnsemblGeneID         Status         
##  Length:10          Length:10          Length:10          Length:10         
##  Class :character   Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character  
##                                                                             
##                                                                             
##                                                                             
##     ChrAcc            ChrStart           ChrStop          Orientation       
##  Length:10          Length:10          Length:10          Length:10         
##  Class :character   Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character  
##                                                                             
##                                                                             
##                                                                             
##      Length     GOFunctionID       GOProcessID        GOComponentID     
##  Min.   : 679   Length:10          Length:10          Length:10         
##  1st Qu.:1236   Class :character   Class :character   Class :character  
##  Median :1519   Mode  :character   Mode  :character   Mode  :character  
##  Mean   :1804                                                           
##  3rd Qu.:2160                                                           
##  Max.   :4158                                                           
##   GOFunction         GOProcess          GOComponent        log2FoldChange 
##  Length:10          Length:10          Length:10          Min.   :5.448  
##  Class :character   Class :character   Class :character   1st Qu.:5.788  
##  Mode  :character   Mode  :character   Mode  :character   Median :5.931  
##                                                           Mean   :6.141  
##                                                           3rd Qu.:6.124  
##                                                           Max.   :7.555  
##      pvalue               padj                stat           lfcSE       
##  Min.   :0.000e+00   Min.   :0.000e+00   Min.   :3.806   Min.   :0.6162  
##  1st Qu.:0.000e+00   1st Qu.:0.000e+00   1st Qu.:5.354   1st Qu.:0.7731  
##  Median :9.300e-10   Median :1.230e-08   Median :6.512   Median :0.9580  
##  Mean   :1.763e-05   Mean   :6.626e-05   Mean   :6.558   Mean   :1.0348  
##  3rd Qu.:1.525e-07   3rd Qu.:1.219e-06   3rd Qu.:7.230   3rd Qu.:1.1248  
##  Max.   :1.410e-04   Max.   :5.105e-04   Max.   :9.996   Max.   :1.9849  
##     baseMean     
##  Min.   : 10.84  
##  1st Qu.: 21.86  
##  Median : 33.03  
##  Mean   : 41.44  
##  3rd Qu.: 45.57  
##  Max.   :132.55  
```

	GeneID	GeneID	Symbol	Description	Synonyms	GeneType	EnsemblGeneID	Status	ChrAcc
51350	51350	51350	KRT76	keratin 76	HUMCYT2A|KRT2B|KRT2P	protein-coding	ENSG00000185069	active	NC_00001
9119	9119	9119	KRT75	keratin 75	CK-75|K6HF|K75|KB18|PFB|hK6hf	protein-coding	ENSG00000170454	active	NC_00001
1472	1472	1472	CST4	cystatin S		protein-coding	ENSG00000101441	active	NC_00002
101929412	101929412	101929412	LINC02212	long intergenic non-protein coding RNA 2212		ncRNA	ENSG00000249396	active	NC_00000
101928687	101928687	101928687	LNCAROD	lncRNA activating regulator of DKK1	A-ROD|LINC01468|lnc-MBL2-4	ncRNA	ENSG00000231131	active	NC_00001

	GeneID	GeneID	Symbol	Description	Synonyms	GeneType	EnsemblGeneID	Status	ChrAcc
6704	6704	6704	SPRR2E	small proline rich protein 2E		protein-coding	ENSG00000203785	active	NC_00000
56033	56033	56033	BARX1	BARX homeobox 1		protein-coding	ENSG00000131668	active	NC_00000
163778	163778	163778	SPRR4	small proline rich protein 4		protein-coding	ENSG00000184148	active	NC_00000
105373485	105373485	105373485	LOC105373485	uncharacterized LOC105373485		ncRNA		active	NC_00002
109729169	109729169	109729169	LINC02154	long intergenic non-protein coding RNA 2154		ncRNA	ENSG00000235385	active	NC_00002

```
summary(top_down);knitr::kable(top_down)
```

```
##     GeneID             GeneID             Symbol          Description       
##  Length:10          Min.   :       70   Length:10          Length:10         
##  Class :character   1st Qu.:     4770   Class :character   Class :character  
##  Mode  :character   Median :     7276   Mode  :character   Mode  :character  
##                     Mean   : 10283197                                        
##                     3rd Qu.:   187198                                        
##                     Max.   :101927499                                        
##    Synonyms           GeneType         EnsemblGeneID         Status         
##  Length:10          Length:10          Length:10          Length:10         
##  Class :character   Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character  
##                                                                             
##                                                                             
##                                                                             
##     ChrAcc            ChrStart           ChrStop          Orientation       
##  Length:10          Length:10          Length:10          Length:10         
##  Class :character   Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character  
##                                                                             
##                                                                             
##                                                                             
##      Length       GOFunctionID       GOProcessID        GOComponentID     
##  Min.   : 382.0   Length:10          Length:10          Length:10         
##  1st Qu.: 767.5   Class :character   Class :character   Class :character  
##  Median :1741.0   Mode  :character   Mode  :character   Mode  :character  
##  Mean   :2201.6                                                           
##  3rd Qu.:2241.2                                                           
##  Max.   :6769.0                                                           
##   GOFunction         GOProcess         GOComponent        log2FoldChange  
##  Length:10          Length:10          Length:10          Min.   :-7.681  
##  Class :character   Class :character   Class :character   1st Qu.:-6.584  
##  Mode  :character   Mode  :character   Mode  :character   Median :-6.087  
##                                                           Mean   :-6.317  
##                                                           3rd Qu.:-5.924  
##                                                           Max.   :-5.751  
##      pvalue               padj                stat             lfcSE       
##  Min.   :0.000e+00   Min.   :0.000e+00   Min.   :-19.638   Min.   :0.3872  
##  1st Qu.:0.000e+00   1st Qu.:0.000e+00   1st Qu.:-12.127   1st Qu.:0.5233  
##  Median :0.000e+00   Median :0.000e+00   Median :-11.327   Median :0.5599  
##  Mean   :5.015e-17   Mean   :2.762e-15   Mean   :-11.969   Mean   :0.5514  
##  3rd Qu.:1.000e-22   3rd Qu.:1.900e-20   3rd Qu.: -9.968   3rd Qu.:0.6075  
##  Max.   :5.015e-16   Max.   :2.762e-14   Max.   : -8.111   Max.   :0.7293  
##     baseMean       
##  Min.   :   19.44  
##  1st Qu.:  145.78  
##  Median :  242.39  
##  Mean   : 4194.07  
##  3rd Qu.: 1797.90  
##  Max.   :34495.23  
```

	GeneID	GeneID	Symbol	Description	Synonyms	GeneType	EnsemblGeneID
221476	221476	221476	PI16	peptidase inhibitor 16	CD364|CRISP9|MSMBBP|PSPBP	protein-coding	ENSG00000164530

	GeneID	GeneID	Symbol	Description	Synonyms	GeneType	EnsemblGeneID
4653	4653	4653	MYOC	myocilin	GLC1A|GPOA|JOAG|JOAG1|TIGR	protein-coding	ENSG00000034971
572558	572558	572558	PGM5-AS1	PGM5 antisense RNA 1	FAM233A	ncRNA	ENSG00000224958
5212	5212	5212	VIT	vitrin	VIT1	protein-coding	ENSG00000205221
84366	84366	84366	PRAC1	PRAC1 small nuclear protein	C17orf92|PRAC	protein-coding	ENSG00000159182
70	70	70	ACTC1	actin alpha cardiac muscle 1	ACTC|ASD5|CMD1R|CMH11|LVNC4	protein-coding	ENSG00000159251

	GeneID	GeneID	Symbol	Description	Synonyms	GeneType	EnsemblGeneID
9340	9340	9340	GLP2R	glucagon like peptide 2 receptor		protein-coding	ENSG00000065325
101927499	101927499	101927499	LOC101927499	uncharacterized LOC101927499		ncRNA	
1674	1674	1674	DES	desmin	CDCD3|CSM1|CSM2|LGMD1D|LGMD1E|LGMD2R	protein-coding	ENSG00000175084
5121	5121	5121	PCP4	Purkinje cell protein 4	PEP-19	protein-coding	ENSG00000183036

STEP 4- showing the top 15 genes

```
# find significant DEGs
sum( res$padj < 0.1, na.rm=TRUE )
```

```
## [1] 13250
```

```
# clean and delete na
res_df <- as.data.frame(res)
res_df <- res_df[!is.na(res_df$padj), ]

#order by significance
res_ordered <- res_df[order(res_df$padj), ]

# extract top15
top15 <- head(res_ordered, 15)
top15
```

```
##         baseMean log2FoldChange     lfcSE      stat       pvalue         padj
## 221476 1945.5333      -7.681227 0.3911341 -19.63835 7.272694e-86 1.566175e-81
## 1675   4452.1590      -4.964090 0.2787639 -17.80751 6.180072e-71 6.654392e-67
## 286133 1803.0834      -5.509867 0.3535451 -15.58462 9.260587e-55 6.647558e-51
## 7123   1054.3341      -4.777304 0.3102290 -15.39928 1.654936e-53 8.909761e-50
## 9340    161.6683      -5.950893 0.3872252 -15.36804 2.681586e-53 1.154959e-49
## 11170   844.1215      -4.797705 0.3233931 -14.83552 8.632901e-50 3.098492e-46
## 219348  435.3765      -3.394889 0.2385134 -14.23354 5.673588e-46 1.745439e-42
## 146556  199.0275      -4.732812 0.3349705 -14.12904 2.515499e-45 6.771409e-42
## 221091  235.4840      -3.405595 0.2424912 -14.04420 8.360530e-45 2.000489e-41
## 1589    136.3388      -4.366090 0.3124977 -13.97159 2.323788e-44 5.004278e-41
## 2674    262.9957      -4.177935 0.3030151 -13.78788 3.014997e-43 5.902542e-40
## 1590    172.1416      -3.998853 0.2918134 -13.70346 9.680062e-43 1.737168e-39
## 7146    581.8593      -4.808086 0.3523215 -13.64687 2.107349e-42 3.490905e-39
## 55022   407.2087      -3.356706 0.2495187 -13.45272 2.967443e-41 4.564563e-38
## 7148   3527.1622      -4.903672 0.3647362 -13.44443 3.319385e-41 4.765530e-38
```

```
top15_table <- top15[, c("log2FoldChange", "pvalue", "padj")]
top15_table
```

```
##        log2FoldChange       pvalue         padj
## 221476      -7.681227 7.272694e-86 1.566175e-81
## 1675        -4.964090 6.180072e-71 6.654392e-67
## 286133      -5.509867 9.260587e-55 6.647558e-51
## 7123        -4.777304 1.654936e-53 8.909761e-50
## 9340        -5.950893 2.681586e-53 1.154959e-49
## 11170       -4.797705 8.632901e-50 3.098492e-46
## 219348      -3.394889 5.673588e-46 1.745439e-42
## 146556      -4.732812 2.515499e-45 6.771409e-42
## 221091      -3.405595 8.360530e-45 2.000489e-41
## 1589        -4.366090 2.323788e-44 5.004278e-41
## 2674        -4.177935 3.014997e-43 5.902542e-40
## 1590        -3.998853 9.680062e-43 1.737168e-39
## 7146        -4.808086 2.107349e-42 3.490905e-39
## 55022       -3.356706 2.967443e-41 4.564563e-38
## 7148        -4.903672 3.319385e-41 4.765530e-38
```

STEP 4- showing the top 15 genes

```
# find significant DEGs
sum( res$padj < 0.1, na.rm=TRUE )
```

```
## [1] 13250
```

```
# clean and delete na
res_df <- as.data.frame(res)
res_df <- res_df[!is.na(res_df$padj), ]

#order by significance
res_ordered <- res_df[order(res_df$padj), ]

# extract top15
top15 <- head(res_ordered, 15)
top15
```

```
##         baseMean log2FoldChange     lfcSE      stat       pvalue         padj
## 221476 1945.5333      -7.681227 0.3911341 -19.63835 7.272694e-86 1.566175e-81
## 1675   4452.1590      -4.964090 0.2787639 -17.80751 6.180072e-71 6.654392e-67
## 286133 1803.0834      -5.509867 0.3535451 -15.58462 9.260587e-55 6.647558e-51
## 7123   1054.3341      -4.777304 0.3102290 -15.39928 1.654936e-53 8.909761e-50
## 9340    161.6683      -5.950893 0.3872252 -15.36804 2.681586e-53 1.154959e-49
## 11170   844.1215      -4.797705 0.3233931 -14.83552 8.632901e-50 3.098492e-46
## 219348  435.3765      -3.394889 0.2385134 -14.23354 5.673588e-46 1.745439e-42
## 146556  199.0275      -4.732812 0.3349705 -14.12904 2.515499e-45 6.771409e-42
## 221091  235.4840      -3.405595 0.2424912 -14.04420 8.360530e-45 2.000489e-41
## 1589    136.3388      -4.366090 0.3124977 -13.97159 2.323788e-44 5.004278e-41
## 2674    262.9957      -4.177935 0.3030151 -13.78788 3.014997e-43 5.902542e-40
## 1590    172.1416      -3.998853 0.2918134 -13.70346 9.680062e-43 1.737168e-39
## 7146    581.8593      -4.808086 0.3523215 -13.64687 2.107349e-42 3.490905e-39
## 55022   407.2087      -3.356706 0.2495187 -13.45272 2.967443e-41 4.564563e-38
## 7148   3527.1622      -4.903672 0.3647362 -13.44443 3.319385e-41 4.765530e-38
```

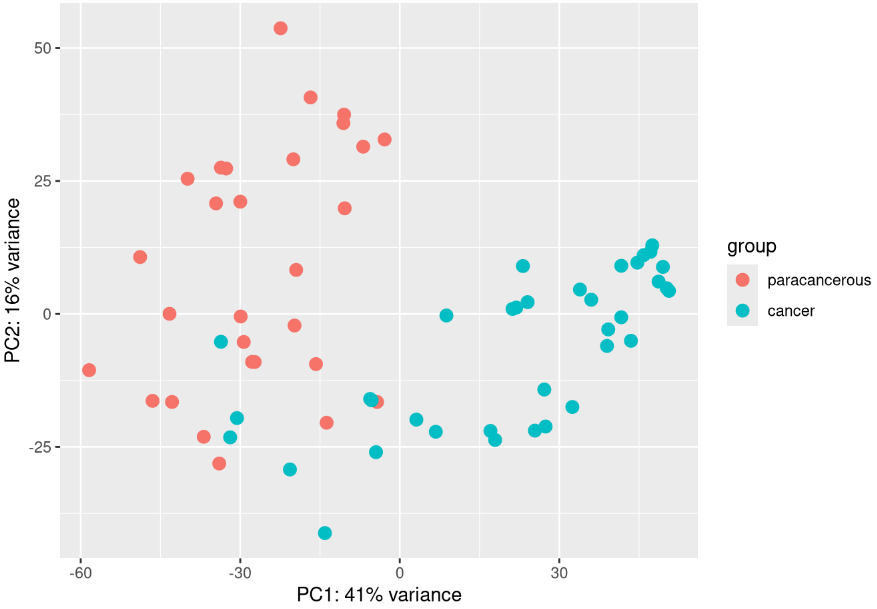
```
top15_table <- top15[, c("log2FoldChange", "pvalue", "padj")]
top15_table
```

```
##        log2FoldChange       pvalue         padj
## 221476      -7.681227 7.272694e-86 1.566175e-81
## 1675        -4.964090 6.180072e-71 6.654392e-67
## 286133      -5.509867 9.260587e-55 6.647558e-51
## 7123        -4.777304 1.654936e-53 8.909761e-50
## 9340        -5.950893 2.681586e-53 1.154959e-49
## 11170       -4.797705 8.632901e-50 3.098492e-46
## 219348      -3.394889 5.673588e-46 1.745439e-42
## 146556      -4.732812 2.515499e-45 6.771409e-42
## 221091      -3.405595 8.360530e-45 2.000489e-41
## 1589        -4.366090 2.323788e-44 5.004278e-41
## 2674        -4.177935 3.014997e-43 5.902542e-40
## 1590        -3.998853 9.680062e-43 1.737168e-39
## 7146        -4.808086 2.107349e-42 3.490905e-39
## 55022       -3.356706 2.967443e-41 4.564563e-38
## 7148        -4.903672 3.319385e-41 4.765530e-38
```

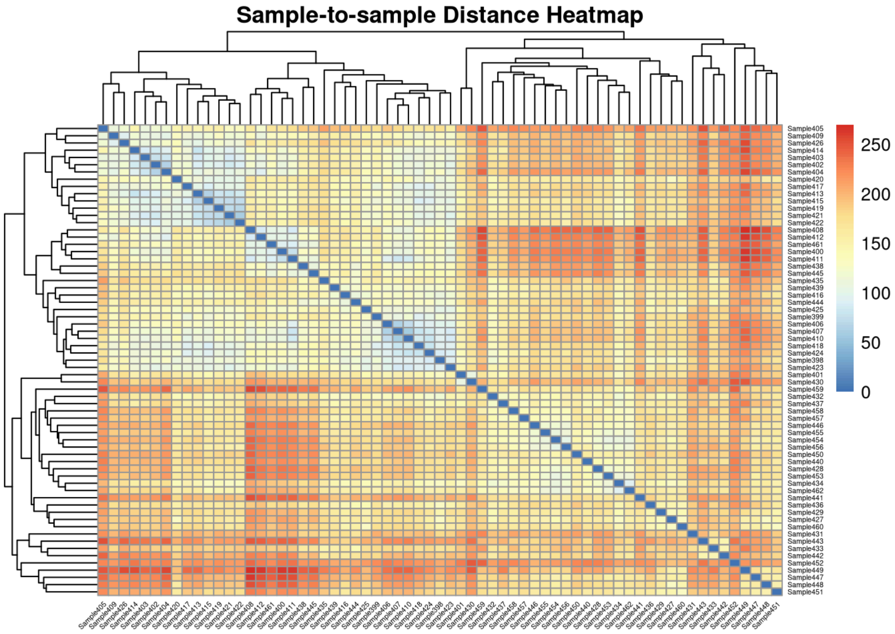
STEP 5- plotting for visualization

```
library(pheatmap)
library(ggplot2)

# 1-Data quality control plots
# PCA plot

vsd <- vst(dds, blind=FALSE)
plotPCA(vsd, intgroup = "condition", ntop = 500)
```

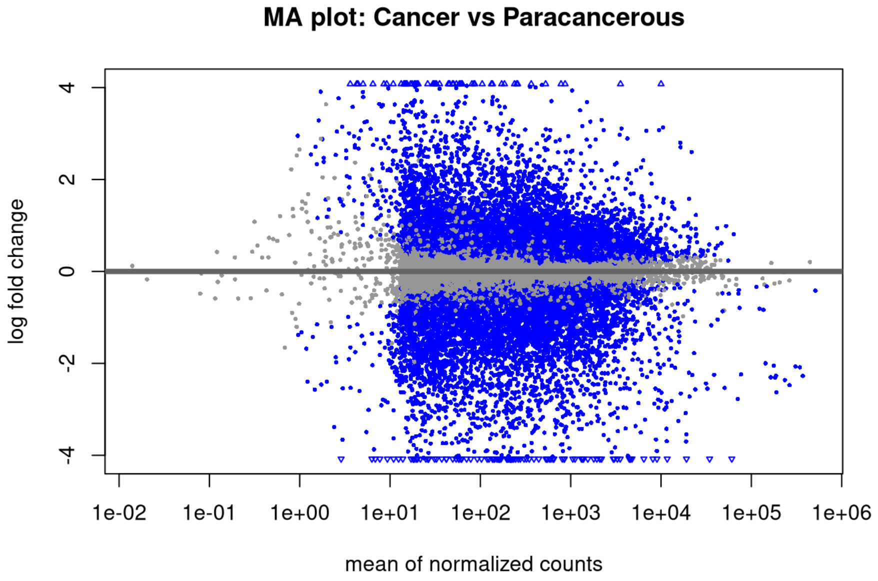
```
## using ntop=500 top features by variance
```



```
# simple distance heatmap- overall gene-expression profiles.
# compute sample-to-sample distances
dists <- dist(t(assay(vsd)))

# convert to matrix
mat <- as.matrix(dists)

pheatmap(mat,
         clustering_distance_rows = dists,
         clustering_distance_cols = dists,
         show_rownames = TRUE,
         show_colnames = TRUE,
         fontsize_row = 4,
         fontsize_col = 4,
         angle_col = 45,
         main = "Sample-to-sample Distance Heatmap")
```



```
## 2 – Differential expression plots

# MA plot: log2 fold change versus mean expression
# shows the pattern of genes that are up- or down-regulated
plotMA(
  res,
  main = "MA plot: Cancer vs Paracancerous"
)
```



```
install.packages("BiocManager")
```

```
## Installing package into '/opt/app-root/src/Rpackages/4.3'
## (as 'lib' is unspecified)
```

```
BiocManager::install("EnhancedVolcano")
```

```
## 'getOption("repos")' replaces Bioconductor standard repositories, see
## 'help("repositories", package = "BiocManager")' for details.
## Replacement repositories:
##     CRAN: https://cran.rstudio.com/
```

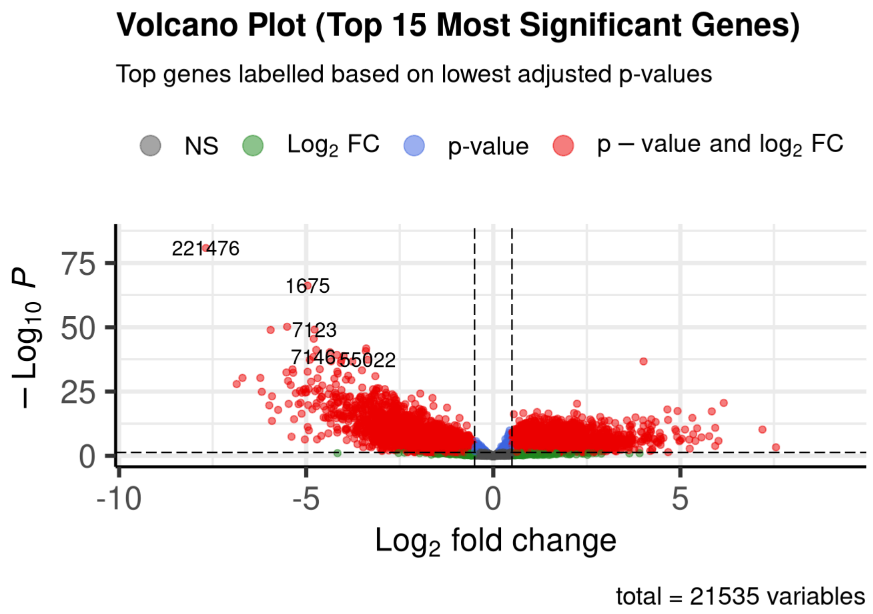
```
## Bioconductor version 3.20 (BiocManager 1.30.27), R 4.4.3 (2025-02-28)
```

```
## Warning: package(s) not installed when version(s) same as or greater than current; use
##   `force = TRUE` to re-install: 'EnhancedVolcano'
## Installation paths not writeable, unable to update packages
##   path: /usr/lib64/R/library
##   packages:
##     AER, aplot, arrow, bayesplot, bigD, BiocManager, BiocParallel, bookdown,
##     boot, BradleyTerry2, brglm, broom, Cairo, checkmate, cli, clock, cluster,
##     colorspace, commonmark, covr, cowplot, credentials, crosstalk, Cubist,
##     curl, data.table, dbplyr, dbscan, dendextend, DEoptimR, Deriv, devtools,
##     dials, diffobj, digest, doBy, doFuture, doRNG, DOSE, downlit, DT, dtplyr,
##     effects, emdbook, emmeans, entropy, Epi, epiR, etm, evaluate, extrafont,
##     extrafontdb, FactoMineR, fastcluster, fgsea, fitdistrplus, flextable,
##     foghorn, forcats, forecast, foreign, fs, future, future.apply,
##     future.callr, gam, gapminder, gargle, gclus, gcookbook, gdtools, generics,
##     gert, GGally, gganimate, ggforce, ggfortify, ggfun, ggnewscale, ggplot2,
##     ggplotify, ggpp, ggpubr, ggraph, ggridges, ggsci, ggstats, ggtangle,
##     ggvenn, gh, glmnet, globals, googledrive, googlesheets4, GPfit, gprofiler2,
##     gss, gt, hardhat, harmony, haven, here, Hmisc, HMM, hms, httpuv, httr2,
##     hunspell, igraph, insight, jsonlite, keras, KFAS, kinship2, knitr,
##     labelled, Lahman, later, lattice, lava, leidenbase, lgr, listenv, littler,
##     lme4, lobstr, logger, magrittr, mapproj, maps, markdown, mathjaxr, Matrix,
##     MatrixModels, mclust, memisc, mgcv, mime, miniUI, mlr, mlr3, mlr3measures,
##     mlr3misc, mockery, mockr, modeltools, multcomp, MuMIn, nanotime, ncdf4,
##     nlme, nloptr, officer, openssl, openxlsx, optimx, ottr, parallelly,
##     parsnip, partykit, patchwork, pbapply, pbdZMQ, pbkrtest, permute, pheatmap,
##     pillar, pixmap, pkgbuild, pkgdown, pkgload, plotly, plotrix, poppr, pracma,
##     pROC, prodlim, profmem, progressr, promises, PRROC, ps, purrr, qtl2,
##     quantmod, questionr, QuickJSR, R.cache, R.oo, ragg, raster, rbibutils,
##     Rcpp, RcppArmadillo, RcppDate, RcppParallel, RCurl, Rdpack, readr, recipes,
##     reformulas, rentrez, reshape, reshape2, restfulr, reticulate, rgl, rlang,
##     rmarkdown, rms, RMySQL, robustbase, roxygen2, RPostgreSQL, rprojroot,
##     rsample, RSQLite, rstantools, rstatix, Rttf2pt1, RUnit, rversions, rvest,
##     s2, S7, safetensors, sass, scales, scatterpie, sctransform, seriation,
##     Seurat, SeuratObject, sf, sfsmisc, shapr, shiny, slider, sparsevctrs,
##     spatstat, spatstat.data, spatstat.explore, spatstat.geom, spatstat.linnet,
##     spatstat.model, spatstat.random, spatstat.univar, spatstat.utils, spelling,
##     spls, statmod, statnet.common, stringi, stringr, styler, survey, svglite,
##     systemfonts, tensor, tensorflow, terra, testthat, textshaping, tfruns,
##     TH.data, threejs, tibble, timeDate, tinytex, TSP, tune, tzdb, units,
##     usethis, utf8, uwot, V8, vcdExtra, vegan, visNetwork, vroom, waldo, warp,
##     workflows, WriteXLS, xfun, xgboost, XML, xml2, yulab.utils, zeallot, zip,
##     zoo
```

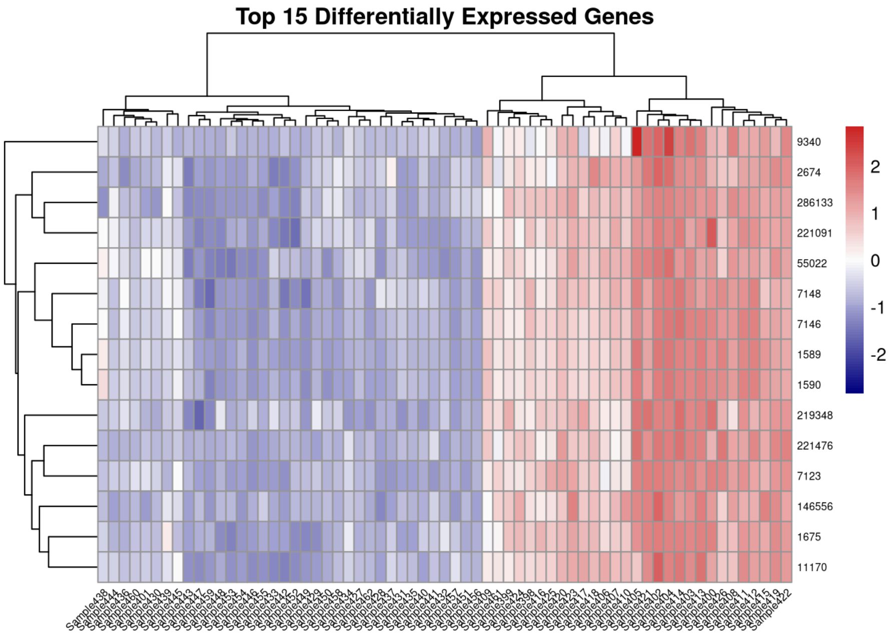
```r
# Volcano plot
library(EnhancedVolcano)
```

```
## Loading required package: ggrepel
```

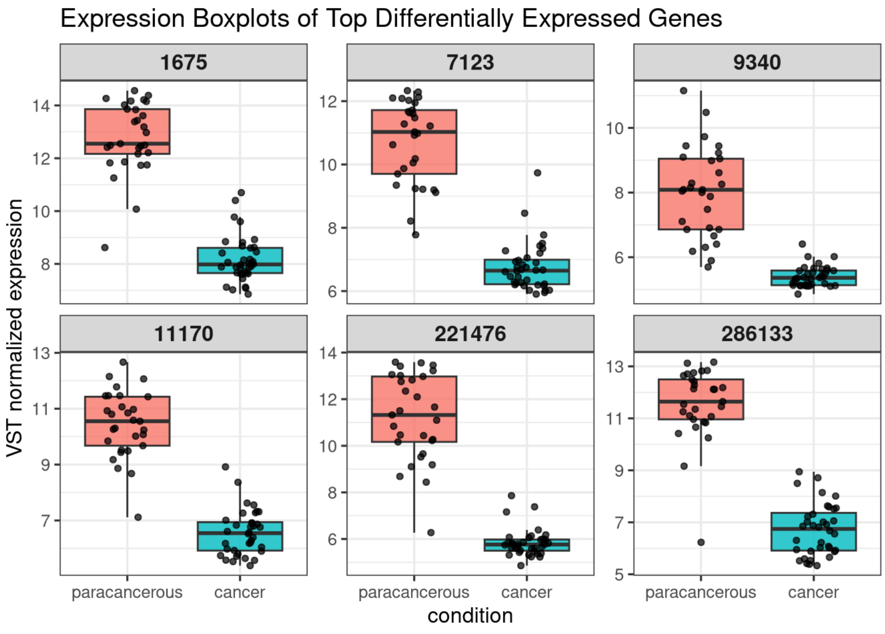
```r
# get the row names of the top 15 genes.
# these row names correspond to the gene IDs in the DESeq2 result object.
top15_genes <- rownames(top15_table)

# making a volcano plot, highlighting only the top 15 most significant genes
EnhancedVolcano(
  res,
  lab       = rownames(res),      # labels for all genes (by row name)
  x         = "log2FoldChange",   # x-axis: effect size (Cancer / Paracancerous)
  y         = "padj",             # y-axis: adjusted p-value
  pCutoff   = 0.05,               # significance threshold for padj
  FCcutoff  = 0.5,                # threshold for |log2 fold change|
  selectLab = top15_genes,        # only these genes are labelled on the plot
  pointSize = 1.5,
  labSize   = 4,
  col       = c("grey30", "forestgreen", "royalblue", "red2"),
  title     = "Volcano Plot (Top 15 Most Significant Genes)",
  subtitle  = "Top genes labelled based on lowest adjusted p-values"
)
```



```
# 3- Biological interpretation plots

library(RColorBrewer)

# Heatmap for Top 15 DEGs

# extract expression matrix from VST
vsd_mat <- assay(vsd)

# taking the rownames of top15 genes
top15_genes <- rownames(top15_table)

# intersect to avoid missing genes (just in case)
top15_genes <- intersect(top15_genes, rownames(vsd_mat))

# subset the VST matrix
heatmap_top15 <- vsd_mat[top15_genes, ]

# plot heatmap
pheatmap(
  heatmap_top15,
  scale = "row",
  clustering_distance_rows = "euclidean",
  clustering_distance_cols = "euclidean",
  clustering_method = "complete",
  show_rownames = TRUE,
  show_colnames = TRUE,
  fontsize_row = 6,
  fontsize_col = 6,
  angle_col = 45,
  color = colorRampPalette(c("navy", "white", "firebrick3"))(100),
  main = "Top 15 Differentially Expressed Genes"
)
```



```
library(reshape2)
library(ggplot2)

# Choosing how many genes to show in the boxplots (we went with top 6)
top_genes <- rownames(top15_table)[1:6]

# subset the VST expression matrix to only these genes
expr_top <- vsd_mat[top_genes, ]

# convert the matrix to long format for ggplot2
df <- melt(expr_top)
colnames(df) <- c("Gene", "Sample", "Expression")

# add group (condition) information for each sample
# 'condition' is the column we created earlier in pdat / colData(dds)
df$condition <- colData(vsd)$condition[match(df$Sample, rownames(colData(vsd)))]

#  making boxplots of expression per condition for each gene
ggplot(df, aes(x = condition, y = Expression, fill = condition)) +
  geom_boxplot(outlier.shape = NA, alpha = 0.8) +
  geom_jitter(width = 0.2, size = 1.3, alpha = 0.7) +
  facet_wrap(~ Gene, scales = "free_y", ncol = 3) +
  theme_bw(base_size = 12) +
  theme(
    strip.text       = element_text(size = 12, face = "bold"),
    legend.position  = "none"
  ) +
  ylab("VST normalized expression") +
  ggtitle("Expression Boxplots of Top Differentially Expressed Genes")
```


Expression Boxplots of Top Differentially Expressed Genes
VST normalized expression
condition
paracancerous
cancer