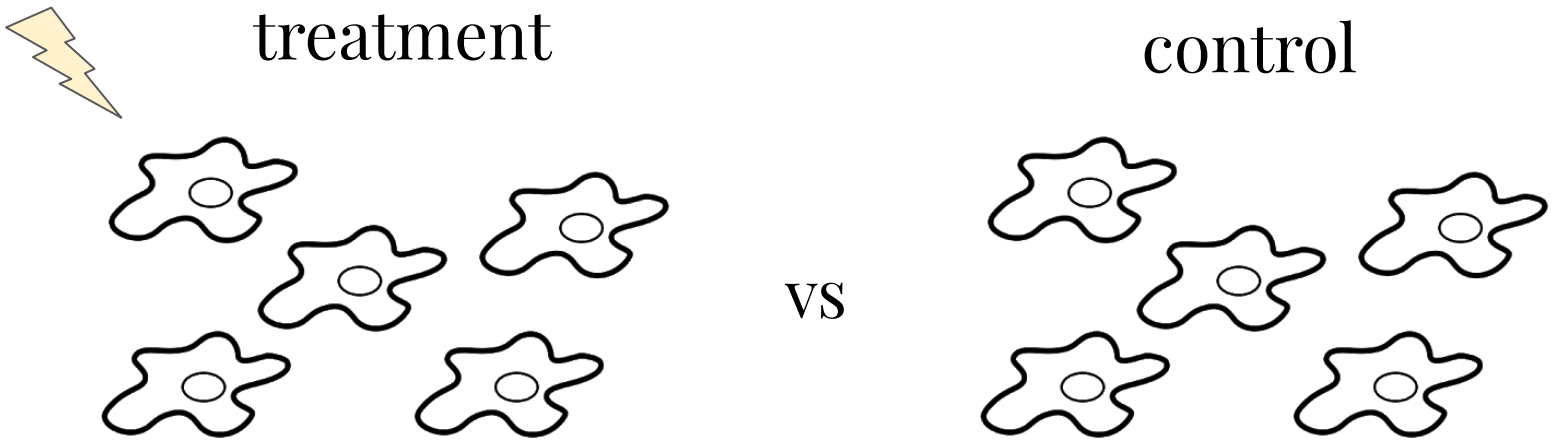
# Анализ дифференциальной экспрессии



# Как можно получить данные для анализа?

fastq → hisat2 → htseq-count

Есть много других программ и подходов.

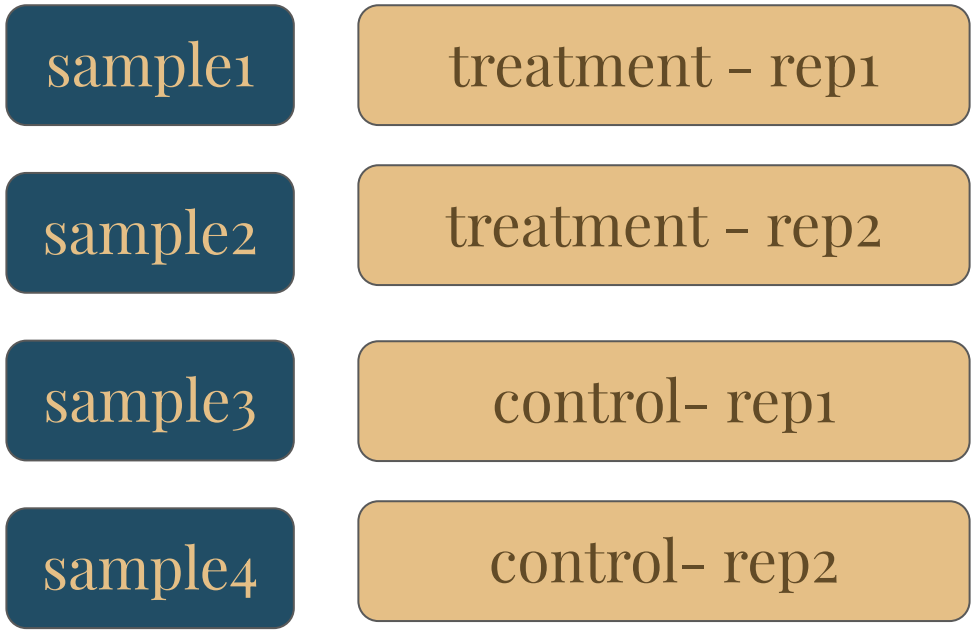
Для работы “с нуля” необходимы:

- чтения
- геном
- разметка генов
- информация о дизайне эксперимента

# В итоге нужно получить таблицу с числом чтений, попавших на каждый ген, для каждого образца

| | sample1 | sample2 | sample3 | sample4 |
|---|---|---|---|---|
| gene1 | 36 | 25 | 2 | 8 |
| gene2 | 412 | 520 | 784 | 840 |
| gene3 | 0 | 1 | 4 | 3 |

# Дизайн эксперимента




**Задача:** узнать, экспрессия каких генов изменилась в ответ на воздействие (treatment) по сравнению с контролем (control). При этом в каждой группе как минимум 2 повторности!

# Пакеты для анализа дифференциальной экспрессии

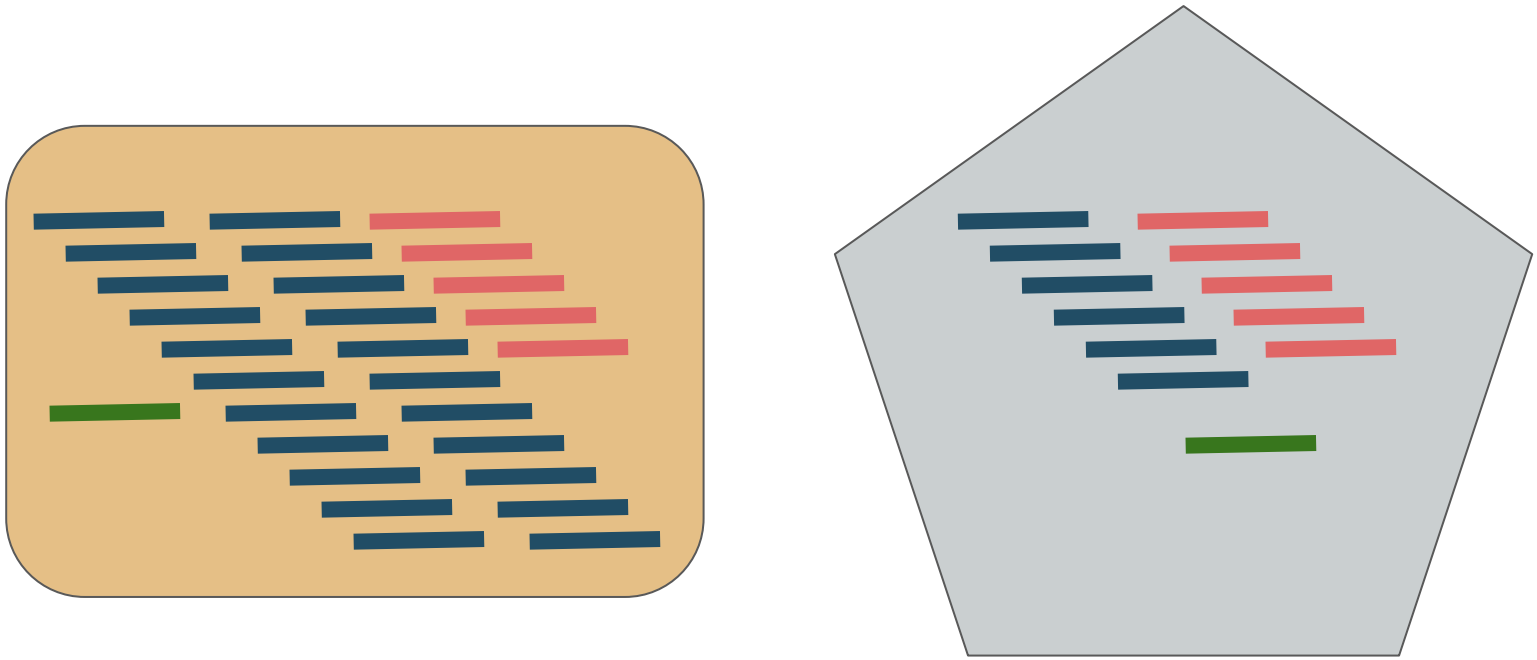
**DESeq2**
edgeR
limma
DEXSeq
Cuffdiff
Ballgown
SAMseq
NOIseq

http://bioconductor.org/packages/release/bioc/html/DESeq2.html

# Размер библиотеки

| | sample1 | sample2 |
|---|---|---|
| gene1 | 100 | 200 |
| gene2 | 50 | 100 |
| gene3 | 150 | 300 |
| lib size | 300 | 600 |

# Композиция библиотеки



Глубина секвенирования
от 30 млн чтений

# Матрица каунтов

```r
counts <- fread("GSE147507_RawReadCounts_Human_A549.tsv")
counts
```

Description: df[,20] [21,797 x 20]

| Gene_Name <chr> | Series2_A549_Mock_1 <int> | Series2_A549_Mock_2 <int> | Series2_A549_Mock_3 <int> |
|---|---|---|---|
| DDX11L1 | 0 | 0 | 0 |
| WASH7P | 68 | 43 | 33 |
| FAM138A | 0 | 0 | 0 |
| FAM138F | 0 | 0 | 0 |
| OR4F5 | 0 | 0 | 0 |
| LOC729737 | 11 | 3 | 6 |
| LOC100132... | 0 | 0 | 0 |
| LOC100132... | 0 | 0 | 0 |
| LOC100133... | 54 | 23 | 20 |
| OR4F29 | 0 | 0 | 0 |

1-10 of 21,797 rows | 1-4 of 20 columns Previous 1 2 3 4 5 6 ... 100 Next

```r
counts_mtx <- as.matrix(counts[,-1])
rownames(counts_mtx) <- counts$Gene_Name
```

# Метаданные

```r
{r}
meta <- fread("GSE147507_MetaData_A549_Human.tsv")
```

```r
{r}
meta <- data.frame(
  Sample = colnames(counts_mtx),
  Cell_line = "A549",
  Condition = unlist(tstrsplit(colnames(counts_mtx), "_", keep = 3)),
  batch = unlist(tstrsplit(colnames(counts_mtx), "_", keep = 1)),
  row.names = colnames(counts_mtx))
meta
```

Description: df[,4] [19 x 4]

| | Sample <fctr> | Cell_line <fctr> | Condition <fctr> |
|---|---|---|---|
| Series2_A549_Mock_1 | Series2_A549_Mock_1 | A549 | Mock |
| Series2_A549_Mock_2 | Series2_A549_Mock_2 | A549 | Mock |
| Series2_A549_Mock_3 | Series2_A549_Mock_3 | A549 | Mock |
| Series3_A549_Mock_1 | Series3_A549_Mock_1 | A549 | Mock |
| Series3_A549_Mock_2 | Series3_A549_Mock_2 | A549 | Mock |
| Series4_A549_Mock_1 | Series4_A549_Mock_1 | A549 | Mock |
| Series4_A549_Mock_2 | Series4_A549_Mock_2 | A549 | Mock |
| Series5_A549_Mock_1 | Series5_A549_Mock_1 | A549 | Mock |
| Series5_A549_Mock_2 | Series5_A549_Mock_2 | A549 | Mock |
| Series5_A549_Mock_3 | Series5_A549_Mock_3 | A549 | Mock |

1-10 of 19 rows | 1-4 of 4 columns

Previous 1 2 Next

Взять из файла или создать самостоятельно

# DESeqDataSet

- матрица каунтов
- метаданные в виде dataframe (названия строк = названия образцов)
- формула (design) для модели, включающая переменные из метаданных
  *~ some_variable + batch +* ***variable_of_interest***

```{r}
dds <- DESeqDataSetFromMatrix(countData = counts_mtx,
                              colData = meta,
                              design = ~ Condition)
```

Для ускорения работы DESeq2 можно заранее исключить гены с низкой и нулевой экспрессией.
В некоторых случаях стоит исключить топ высоко-экспрессируемых генов.

# Нормировочные коэффициенты

```r
dds <- estimateSizeFactors(dds)
sizeFactors(dds)
```

```
cc01-20covid1 cc01-20covid2  cc01-20mock2  cc01-20mock1  cc03-19covid   cc03-19mock cc03-20covid1
    1.0717796     1.1296282     0.9654354     1.0505590     1.1163488     0.9980212     0.8693471

cc03-20covid2  cc03-20mock1  cc03-20mock2
    0.8887331     1.0591175     1.0122730
```

```r
counts(dds, normalized = TRUE)[,1:4]
```

```
                cc01-20covid1 cc01-20covid2 cc01-20mock2 cc01-20mock1
TSPAN6           3.613616e+03  3.693251e+03 4.829945e+03 3.605699e+03
TNMD             0.000000e+00  0.000000e+00 0.000000e+00 0.000000e+00
DPM1             2.037732e+03  2.068822e+03 1.422156e+03 1.425907e+03
SCYL3            6.904405e+02  7.391813e+02 9.550095e+02 8.462161e+02
C1orf112         6.997708e+02  7.807879e+02 1.864444e+02 1.913267e+02
FGR              9.330277e-01  1.770494e+00 0.000000e+00 8.566867e+00
CFH              4.214486e+03  4.430661e+03 6.457190e+03 7.742544e+03
FUCA2            6.503203e+03  6.510107e+03 2.909568e+03 2.892746e+03
GCLC             6.913735e+03  6.968665e+03 2.907496e+03 2.633836e+03
NFYA             1.275449e+03  1.300428e+03 9.632959e+02 1.061340e+03
STPG1            8.509213e+02  7.551157e+02 5.417245e+02 4.407177e+02
NIPAL3           3.996158e+03  3.803907e+03 3.281421e+03 3.678042e+03
LAS1L            2.092781e+03  2.077675e+03 1.163206e+03 1.294549e+03
```

# Как происходит нормализация?

- размер библиотеки
- структура библиотеки
- ~~длина гена~~

$$\hat{s}_j = \underset{i}{\text{median}} \frac{k_{ij}}{\left(\prod_{v=1}^{m} k_{iv}\right)^{1/m}}$$

# Как происходит нормализация?

$$\hat{s}_j = \underset{i}{\text{median}} \frac{k_{ij}}{\left(\prod_{v=1}^{m} k_{iv}\right)^{1/m}}$$

1. среднее геометрическое для каждого гена (устойчиво к выбросам) - псевдо-референс
2. отношение экспрессии к псевдо-референсу
3. медиана отношений по каждому образцу

Лишь относительно небольшое число генов действительно дифференциально экспрессируются.

# Как происходит нормализация?

1.

| gene | sampleA | sampleB | pseudo-reference sample |
|---|---|---|---|
| EF2A | 1489 | 906 | sqrt(1489 * 906) = **1161.5** |
| ABCD1 | 22 | 13 | sqrt(22 * 13) = **17.7** |
| ... | ... | ... | ... |

2.

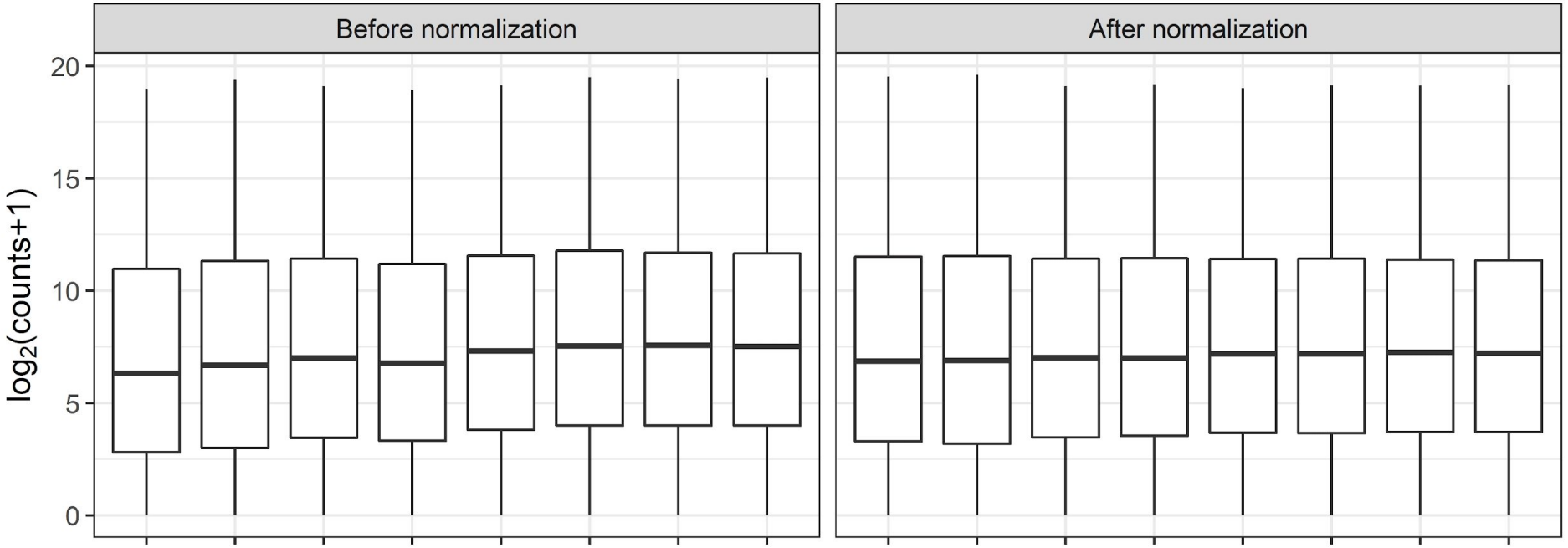
| gene | sampleA | sampleB | pseudo-reference sample | ratio of sampleA/ref | ratio of sampleB/ref |
|---|---|---|---|---|---|
| EF2A | 1489 | 906 | 1161.5 | 1489/1161.5 = **1.28** | 906/1161.5 = **0.78** |
| ABCD1 | 22 | 13 | 16.9 | 22/16.9 = **1.30** | 13/16.9 = **0.77** |
| MEFV | 793 | 410 | 570.2 | 793/570.2 = **1.39** | 410/570.2 = **0.72** |
| BAG1 | 76 | 42 | 56.5 | 76/56.5 = **1.35** | 42/56.5 = **0.74** |
| MOV10 | 521 | 1196 | 883.7 | 521/883.7 = **0.590** | 1196/883.7 = **1.35** |
| ... | ... | ... | ... | | |

3.

```
normalization_factor_sampleA <- median(c(1.28, 1.3, 1.39, 1.35, 0.59))
```

```
normalization_factor_sampleB <- median(c(0.78, 0.77, 0.72, 0.74, 1.35))
```

# Успешная нормализация



Нет ли выбросов среди образцов?

Кластеризация

Корреляция

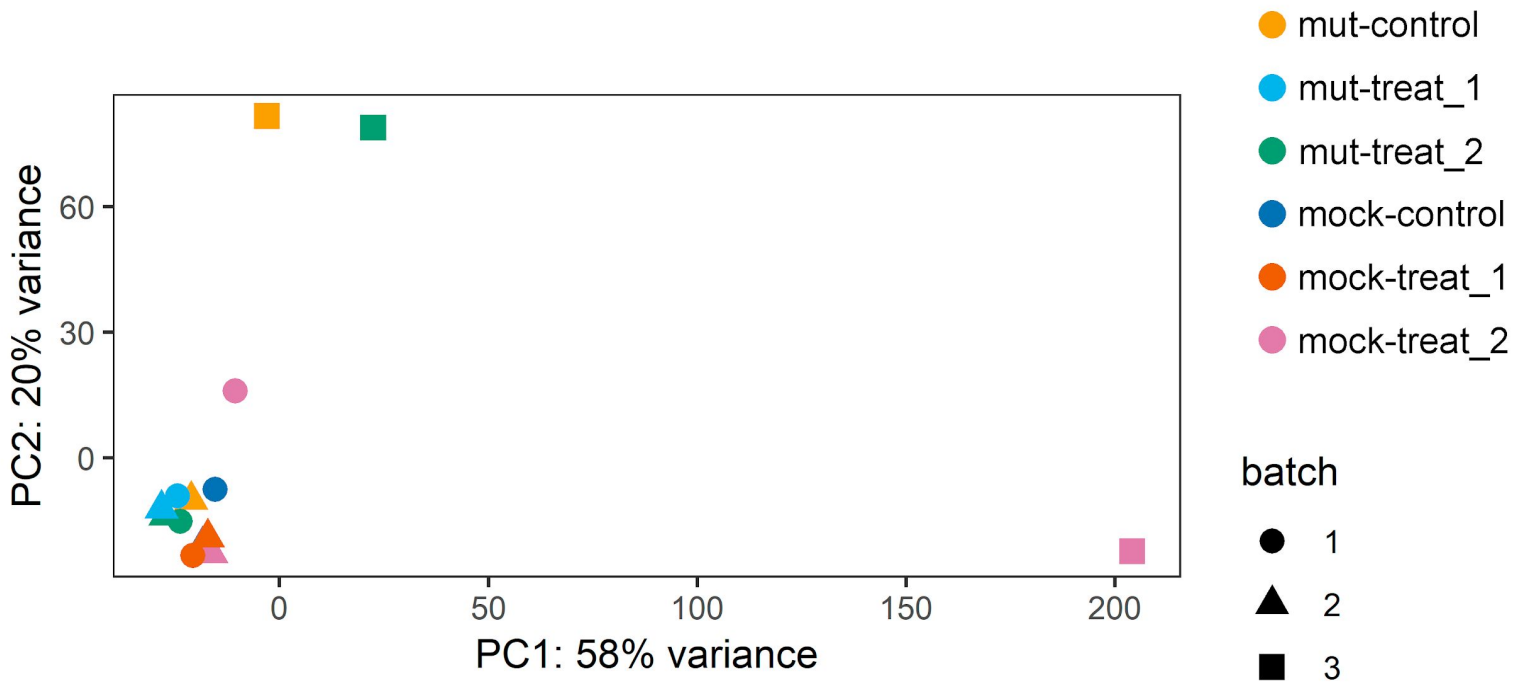
Реплики! Нужны обязательно!

| | ensgene | SRR1039508 | SRR1039509 | SRR1039512 | SRR1039513 | SRR1039516 | SRR1039517 | SRR1039520 | SRR1039521 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | ENSG00000000003 | 723 | 486 | 904 | 445 | 1170 | 1097 | 806 | 604 |
| 2 | ENSG00000000005 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | ENSG00000000419 | 467 | 523 | 616 | 371 | 582 | 781 | 417 | 509 |
| 4 | ENSG00000000457 | 347 | 258 | 364 | 237 | 318 | 447 | 330 | 324 |
| 5 | ENSG00000000460 | 96 | 81 | 73 | 66 | 118 | 94 | 102 | 74 |
| 6 | ENSG00000000938 | 0 | 0 | 1 | 0 | 2 | 0 | 0 | 0 |

У каждого гена есть среднее и дисперсия по образцам

# PCA

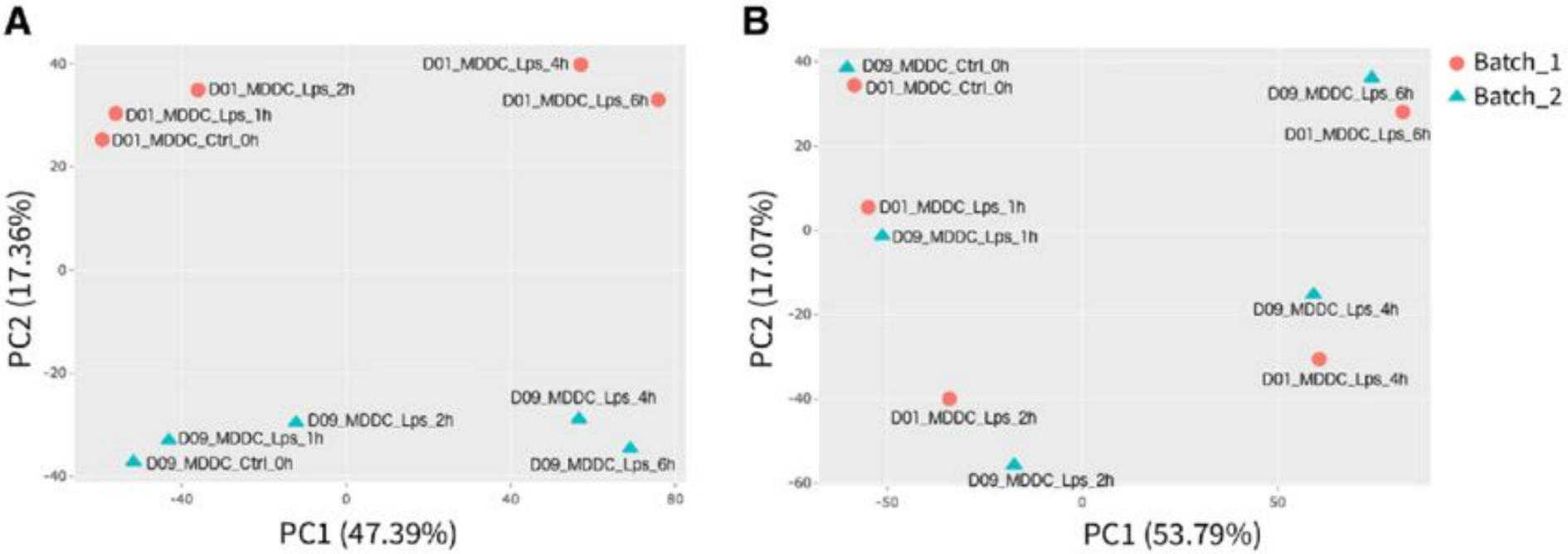
Сильно отличающиеся образцы – выбросы




Поэтому лучше секвенировать больше повторностей на всякий случай.

# PCA

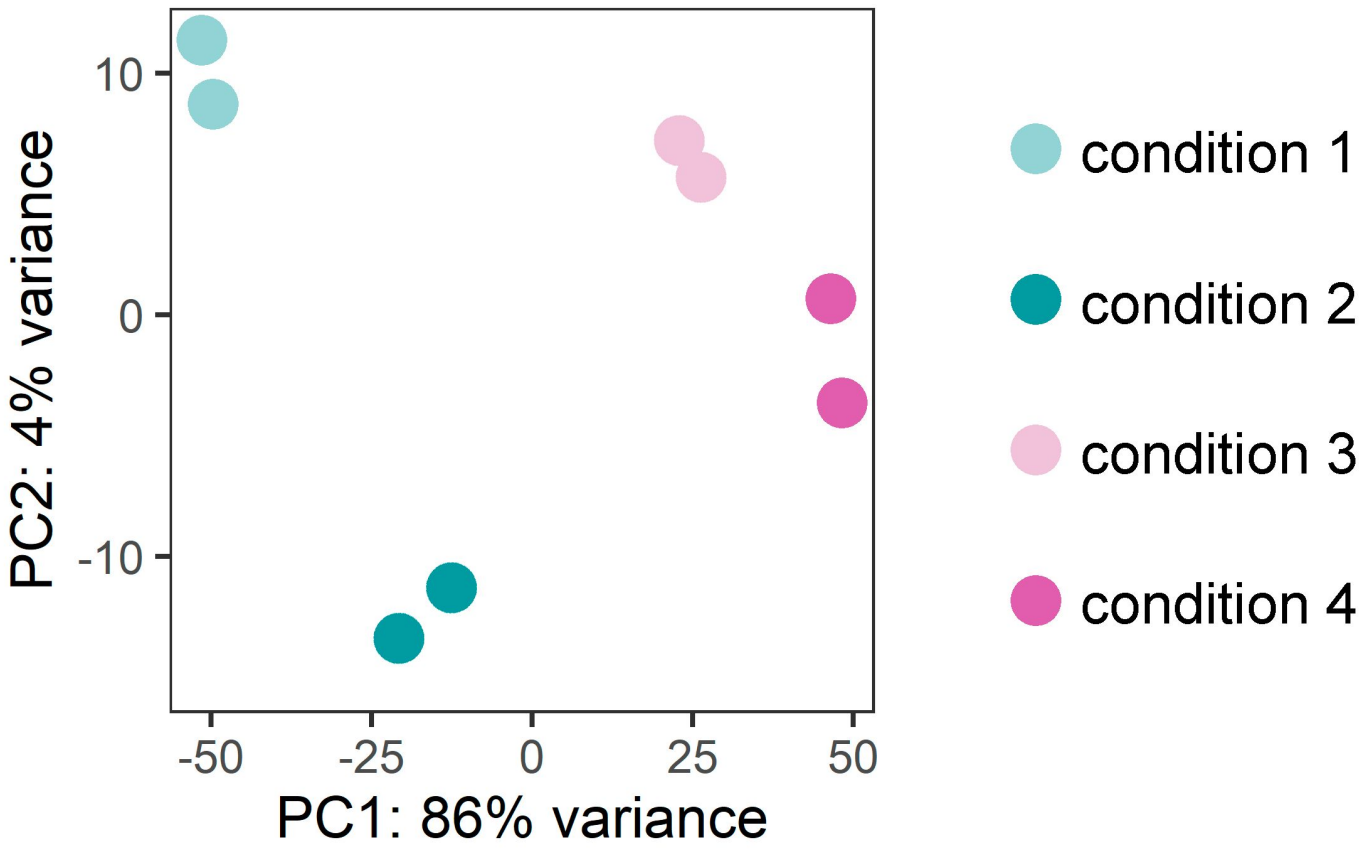
## Batch effect




https://doi.org/10.1186/s12864-018-5362-x

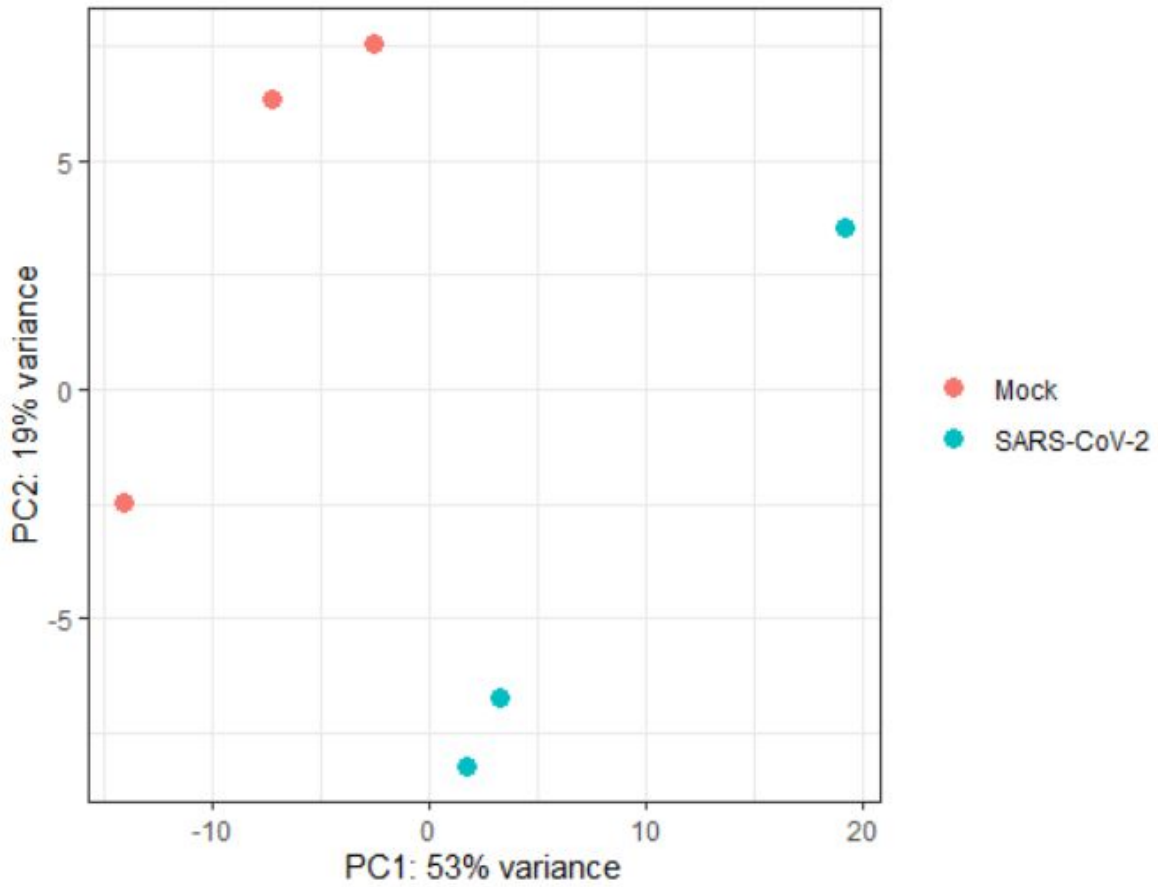
Можно добавить переменную batch в формулу для модели DESeq2

# PCA


10
0
-10
PC2: 4% variance
-50
-25
0
25
50
PC1: 86% variance
condition 1
condition 2
condition 3
condition 4

# PCA

```{r}
plotPCA(vst(dds), intgroup = "Condition", ntop = Inf) +
  theme_bw() + theme(aspect.ratio = 1, legend.title = element_blank())
```



# DESeq - анализ дифференциальной экспрессии

```r
{r}
dds <- DESeq(dds)

res <- results(dds, alpha = 0.05)
summary(res)
```

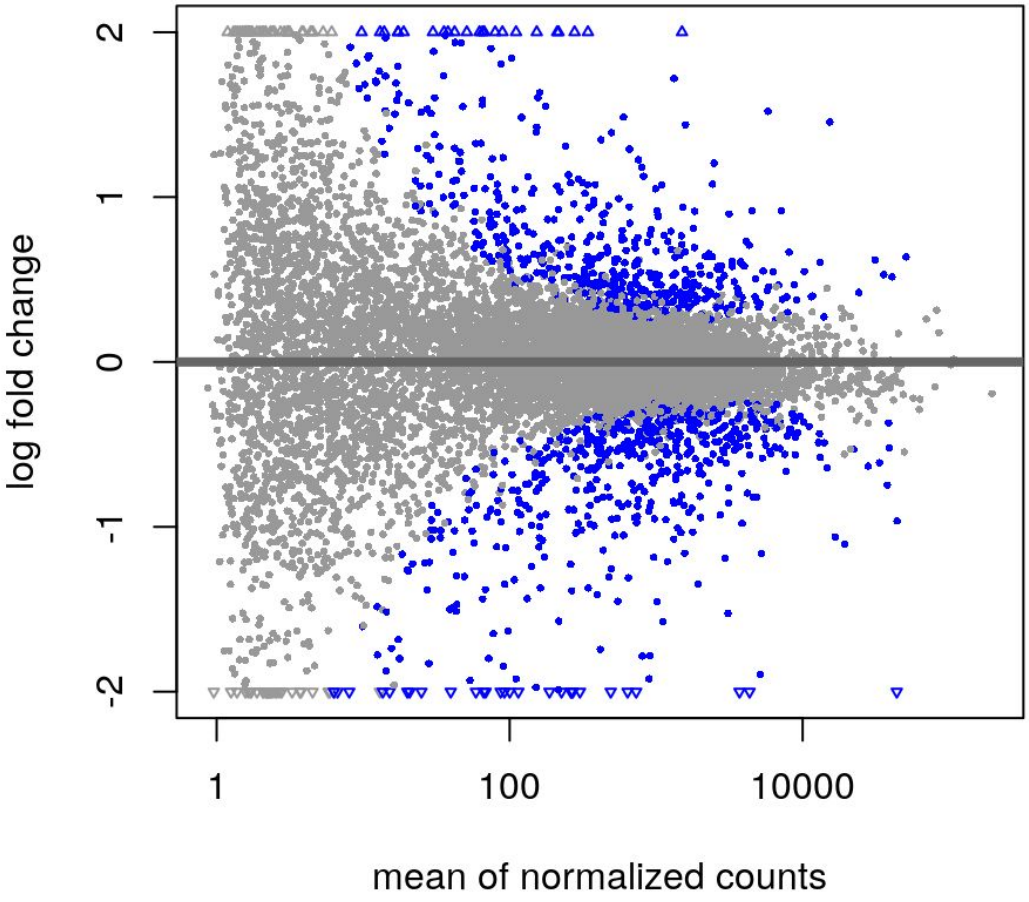
```
out of 16992 with nonzero total read count
adjusted p-value < 0.05
LFC > 0 (up)       : 166, 0.98%
LFC < 0 (down)     : 63, 0.37%
outliers [1]       : 0, 0%
low counts [2]     : 3895, 23%
(mean count < 5)
[1] see 'cooksCutoff' argument of ?results
[2] see 'independentFiltering' argument of ?results
```

- Отрицательное биномиальное распределение
- Поправка на множественное тестирование

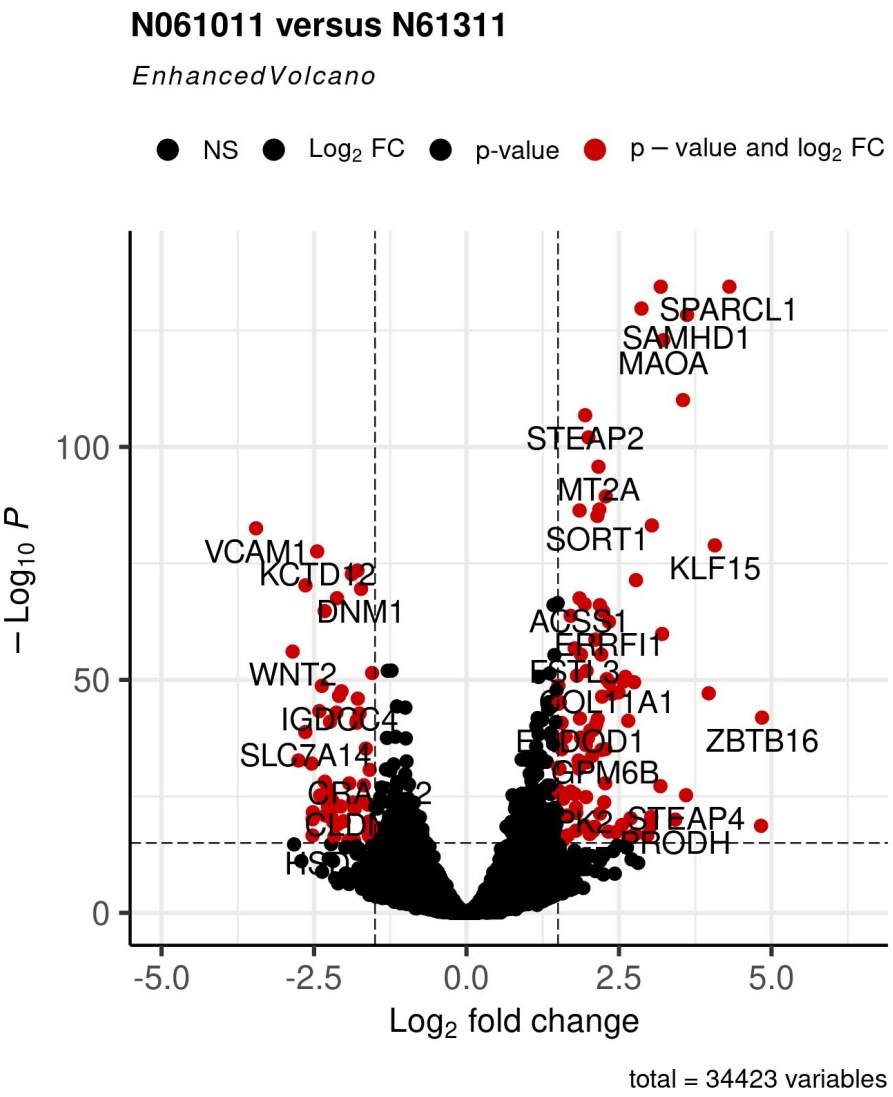
```r
str(res)
```

```
Formal class 'DESeqResults' [package "DESeq2"] with 7 slots
  ..@ priorInfo      : list()
  ..@ rownames       : chr [1:21797] "DDX11L1" "WASH7P" "FAM138A" "FAM138F" ...
  ..@ nrows          : int 21797
  ..@ listData       :List of 6
  .. ..$ baseMean      : num [1:21797] 0.234 42.984 0 0 0 ...
  .. ..$ log2FoldChange: num [1:21797] 1.296 -0.237 NA NA NA ...
  .. ..$ lfcSE         : num [1:21797] 3.576 0.286 NA NA NA ...
  .. ..$ stat          : num [1:21797] 0.362 -0.829 NA NA NA ...
  .. ..$ pvalue        : num [1:21797] 0.717 0.407 NA NA NA ...
  .. ..$ padj          : num [1:21797] NA 0.796 NA NA NA ...
  ..@ elementType    : chr "ANY"
  ..@ elementMetadata:Formal class 'DFrame' [package "S4Vectors"] with 6 slots
  .. .. ..@ rownames       : NULL
  .. .. ..@ nrows          : int 6
  .. .. ..@ listData       :List of 2
  .. .. .. ..$ type       : chr [1:6] "intermediate" "results" "results" "results"
...
  .. .. .. ..$ description: chr [1:6] "mean of normalized counts for all samples"
"log2 fold change (MLE): Condition SARS.CoV.2 vs Mock" "standard error: Condition
SARS.CoV.2 vs Mock" "Wald statistic: Condition SARS.CoV.2 vs Mock" ...
  .. .. ..@ elementType    : chr "ANY"
  .. .. ..@ elementMetadata: NULL
  .. .. ..@ metadata       : list()
  ..@ metadata       :List of 6
  .. ..$ filterThreshold: Named num 4.68
  .. .. ..- attr(*, "names")= chr "39.91102%"
  .. ..$ filterTheta    : num 0.399
  .. ..$ filterNumRej   :'data.frame':  50 obs. of  2 variables:
  .. .. ..$ theta : num [1:50] 0.22 0.235 0.25 0.265 0.28 ...
  .. .. ..$ numRej: num [1:50] 205 206 206 210 212 213 214 215 217 219 ...
  .. ..$ lo.fit         :List of 2
  .. .. ..$ x: num [1:50] 0.22 0.235 0.25 0.265 0.28 ...
  .. .. ..$ y: num [1:50] 204 206 208 209 211 ...
  .. ..$ alpha          : num 0.05
  .. ..$ lfcThreshold   : num 0
```

# MA-plot



# Volcano plot






https://bioconductor.org/packages/devel/bioc/vignettes/EnhancedVolcano/inst/doc/EnhancedVolcano.html

# Функциональная аннотация генов

# bitr - конвертация gene IDs

```
x <- c("GPX3",  "GLRX",   "LBP",   "CRYAB", "DEFB1", "HCLS1",   "SOD2",   "HSPA2",
       "ORM1",  "IGFBP1", "PTHLH", "GPC3",  "IGFBP3","TOB1",    "MITF",   "NDRG1",
       "NR1H4", "FGFR3",  "PVR",   "IL6",   "PTPRM", "ERBB2",   "NID2",   "LAMB1",
       "COMP",  "PLS3",   "MCAM",  "SPP1",  "LAMC1", "COL4A2",  "COL4A1", "MYOC",
       "ANXA4", "TFPI2",  "CST6",  "SLPI",  "TIMP2", "CPM",     "GGT1",   "NNMT",
       "MAL",   "EEF1A2", "HGD",   "TCN2",  "CDA",   "PCCA",    "CRYM",   "PDXK",
       "STC1",  "WARS",  "HMOX1", "FXYD2", "RBP4",   "SLC6A12", "KDELR3", "ITM2B")
eg = bitr(x, fromType="SYMBOL", toType="ENTREZID", annoDb="org.Hs.eg.db")
head(eg)
```

```
##   SYMBOL ENTREZID
## 1   GPX3     2878
## 2   GLRX     2745
## 3    LBP     3929
## 4  CRYAB     1410
## 5  DEFB1     1672
## 6  HCLS1     3059
```

Еще есть Uniprot и пакет R biomaRt.

```
idType("org.Hs.eg.db")
```

```
##  [1] "ENTREZID"     "PFAM"         "IPI"          "PROSITE"      "ACCNUM"
##  [6] "ALIAS"        "ENZYME"       "MAP"          "PATH"         "PMID"
## [11] "REFSEQ"       "SYMBOL"       "UNIGENE"      "ENSEMBL"      "ENSEMBLPROT"
## [16] "ENSEMBLTRANS" "GENENAME"     "UNIPROT"      "GO"           "EVIDENCE"
## [21] "ONTOLOGY"     "GOALL"        "EVIDENCEALL"  "ONTOLOGYALL"  "OMIM"
## [26] "UCSCKG"
```

```
ids <- bitr(x, fromType="SYMBOL", toType=c("UNIPROT", "ENSEMBL"), annoDb="org.Hs.eg.db")
head(ids)
```

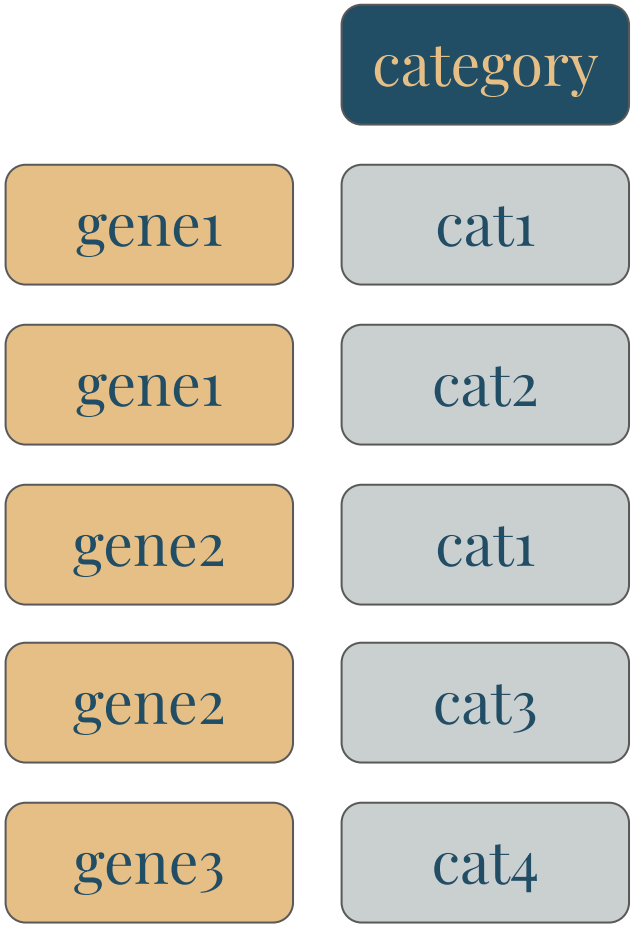
```
##   SYMBOL    UNIPROT         ENSEMBL
## 1   GPX3     P22352 ENSG00000211445
## 2   GLRX A0A024RAM2 ENSG00000173221
## 3   GLRX     P35754 ENSG00000173221
## 4    LBP     P18428 ENSG00000129988
## 5    LBP     Q8TCF0 ENSG00000129988
## 6  CRYAB     P02511 ENSG00000109846
```

Есть не только для человека

Для ~20 модельных организмов

# Откуда брать аннотации для генов?

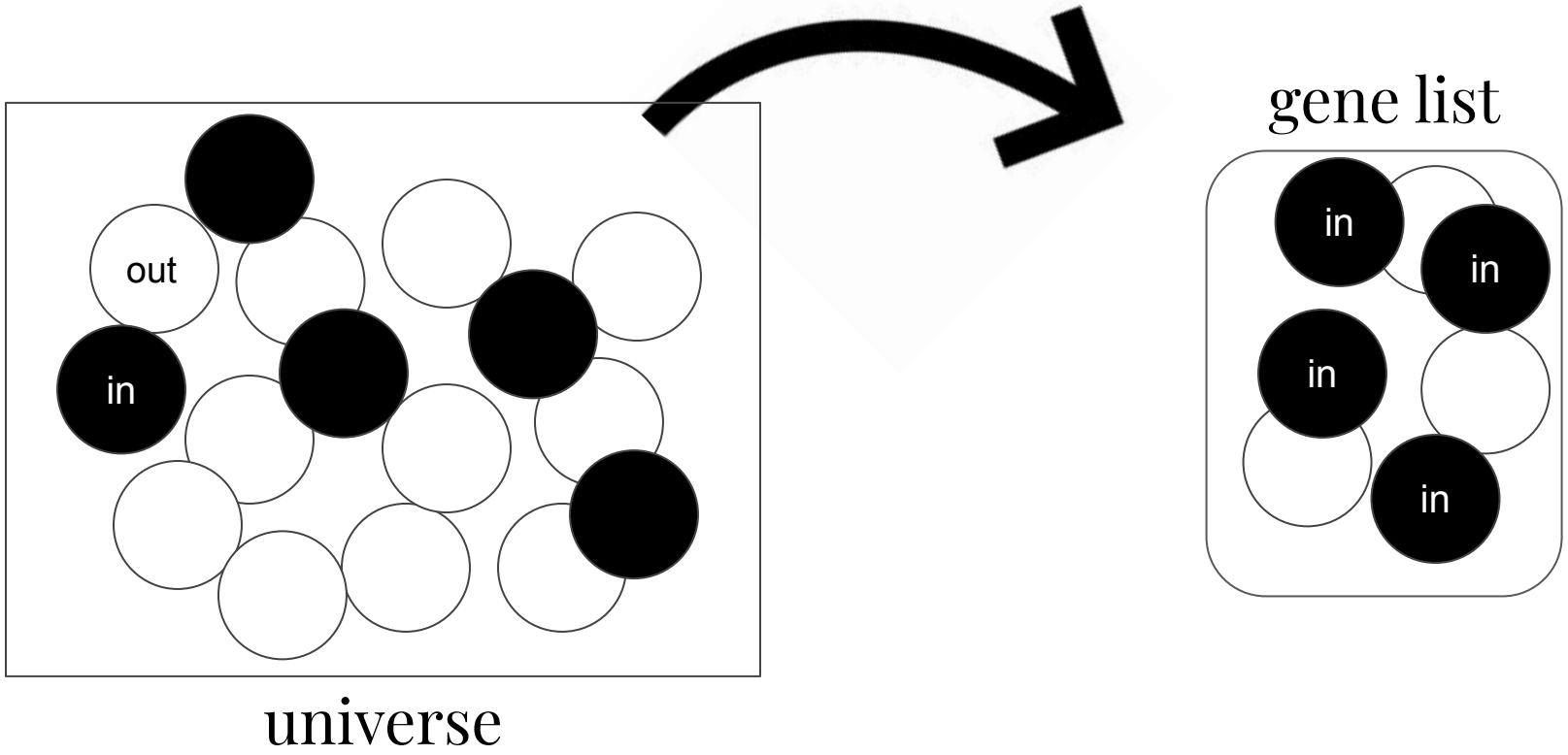
- Gene Ontology
- KEGG
- Reactome
- MSigDB
- WikiPathways
- ...
- можно создать самому:



# Over Representation Analysis

Список генов:

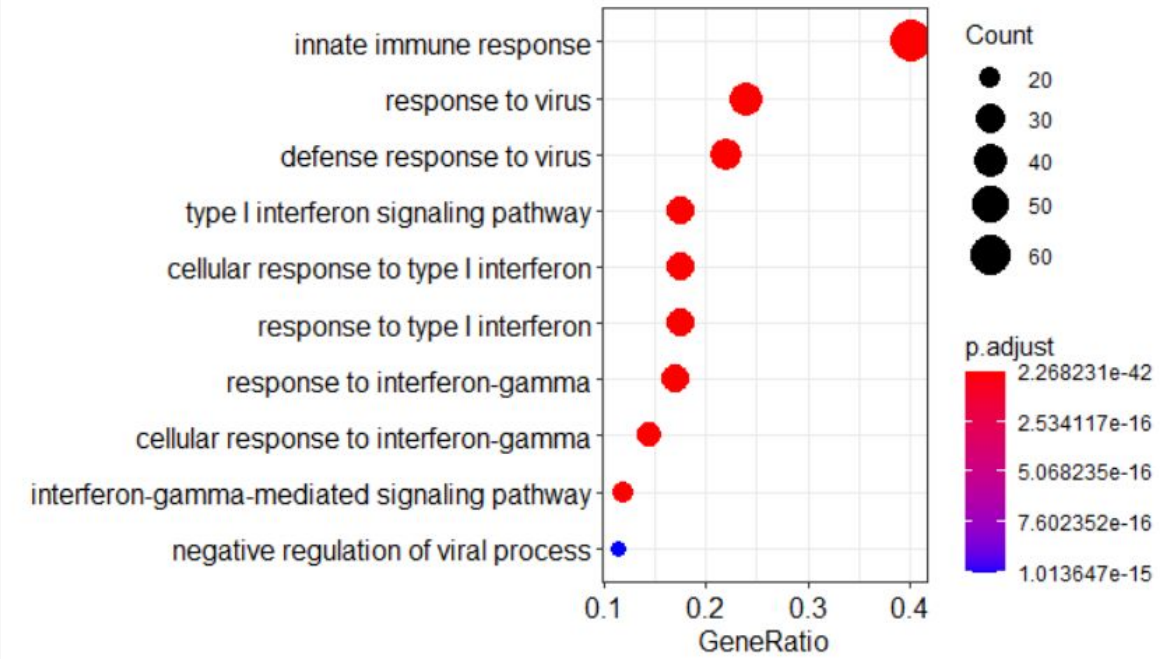
- ДЭ генов
- гены, значимо повысившие/понизившие экспрессию
- любой другой список интересных/значимых генов



# ORA используя clusterProfiler

```r
{r}
library(clusterProfiler)
library(org.Hs.eg.db)
```

```r
{r}
up_DE_genes <- res_df %>%
  filter(padj < 0.05, log2FoldChange >= 0) %>%
  pull(Gene_Name)

ego <- enrichGO(gene = up_DE_genes,
                OrgDb = "org.Hs.eg.db",
                keyType = "SYMBOL",
                ont = "BP",
                universe = res_df$Gene_Name)

dotplot(ego)
```



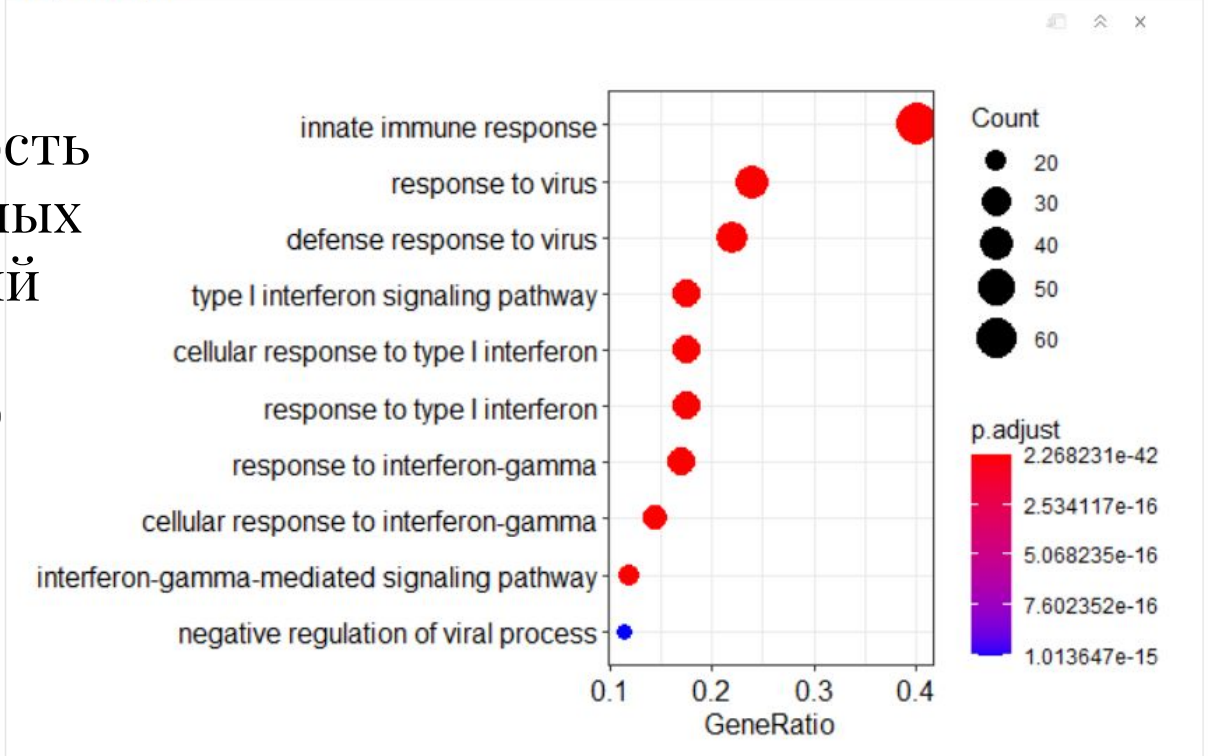
# ORA используя clusterProfiler

```r
{r}
library(clusterProfiler)
library(org.Hs.eg.db)
```

```r
{r}
up_DE_genes <- res_df %>%
  filter(padj < 0.05, log2FoldChange >= 0) %>%
  pull(Gene_Name)

ego <- enrichGO(gene = up_DE_genes,
                OrgDb = "org.Hs.eg.db",
                keyType = "SYMBOL",
                ont = "BP",
                universe = res_df$Gene_Name)

dotplot(ego)
```

избыточность обогащенных категорий

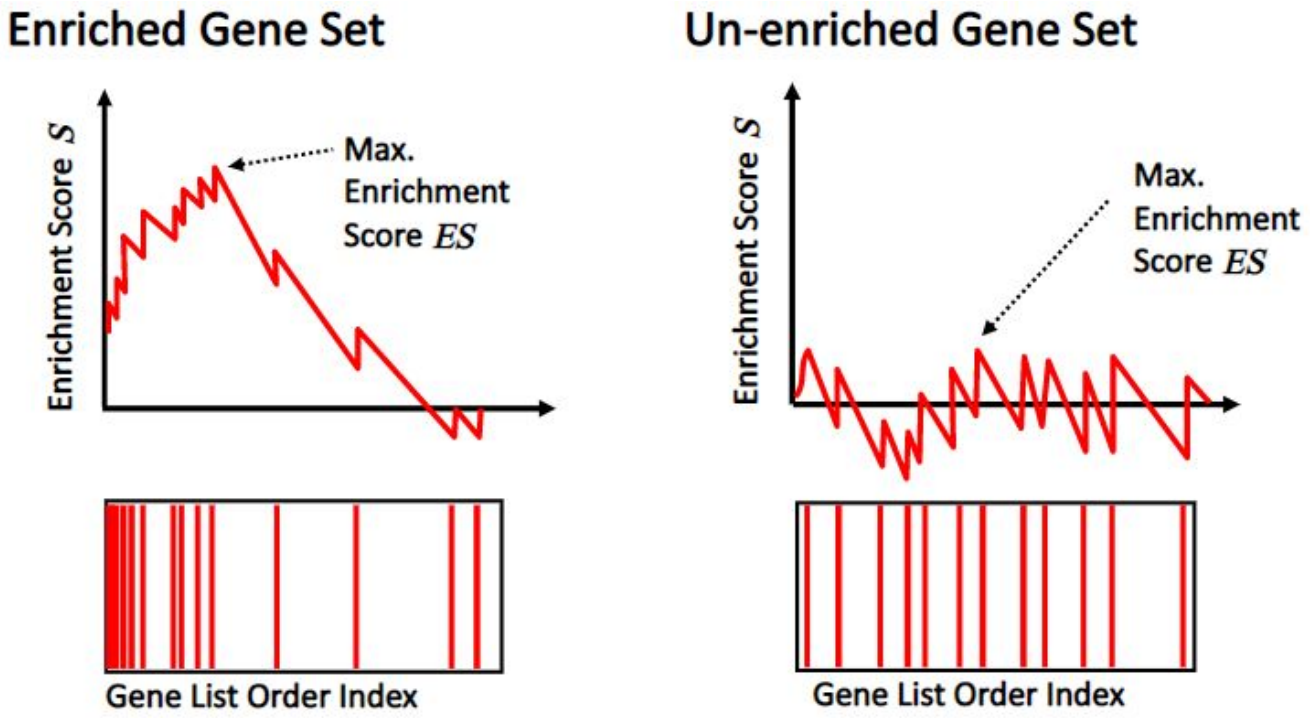
REVIGO



# GSEA
# Gene Set Enrichment Analysis

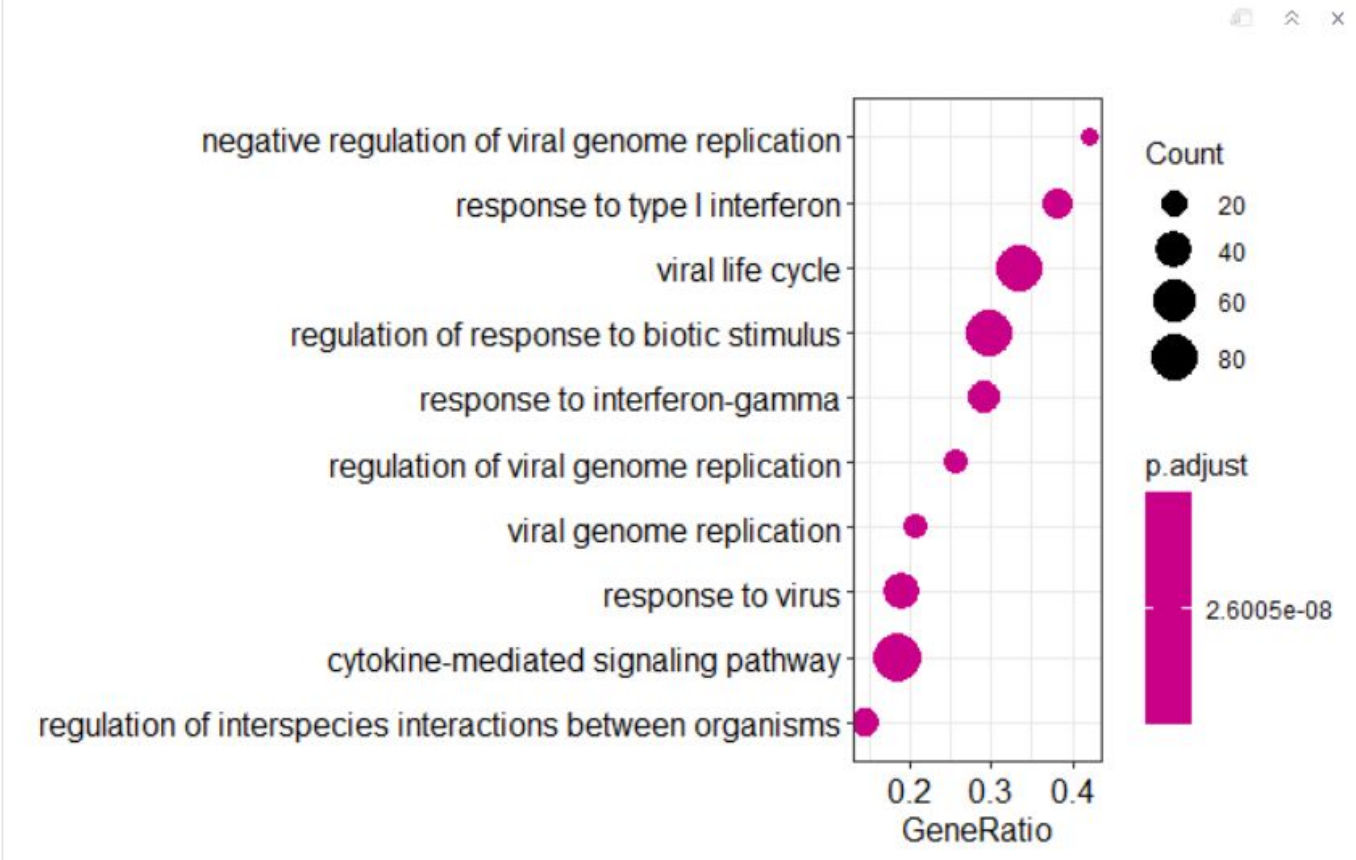
Ранжированный список [всех экспрессируемых] генов:

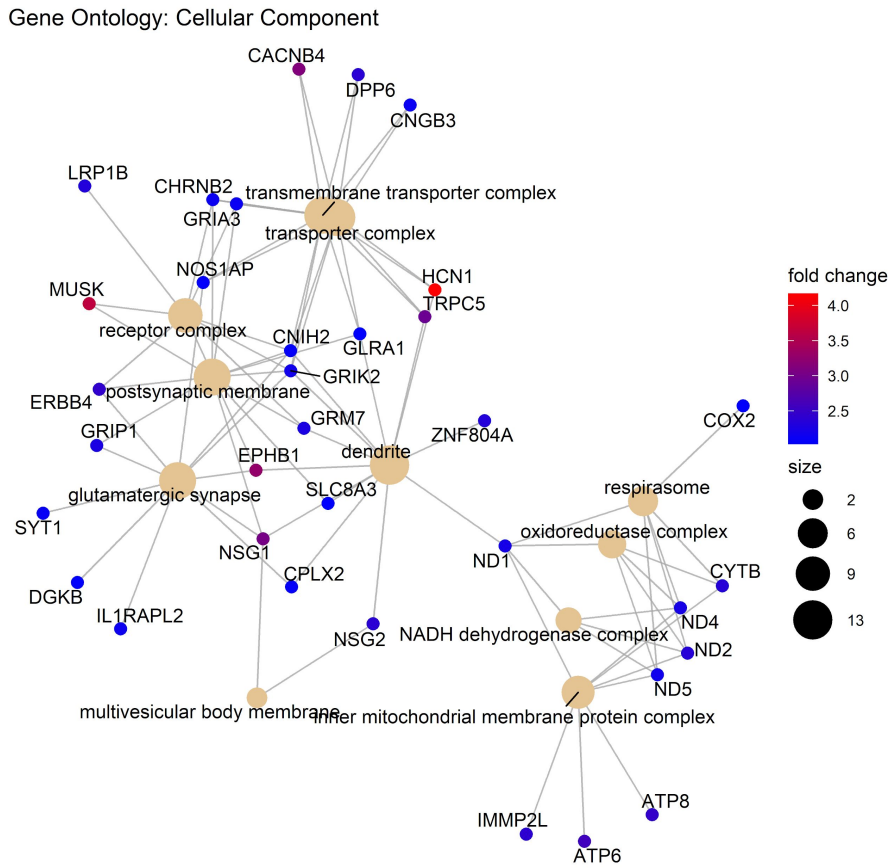
- Fold Change
- комбинация padj и Fold Change
- ...

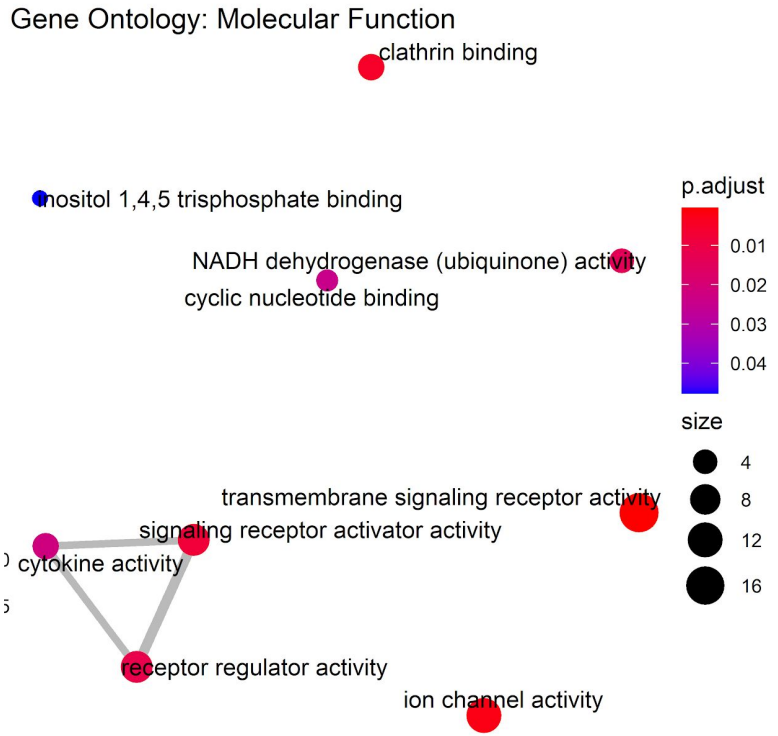


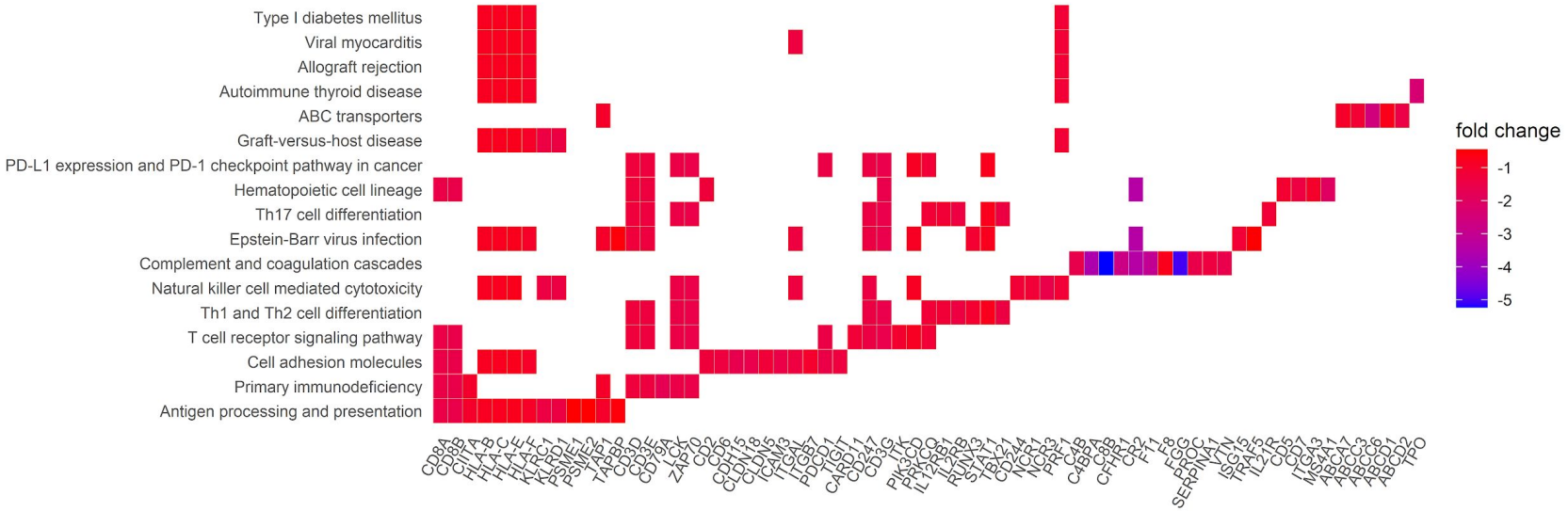
# GSEA используя clusterProfiler

```r
{r}
gene_list <- res_df$log2FoldChange
names(gene_list) <- res_df$Gene_Name
gene_list <- sort(gene_list, decreasing = TRUE)

gsea <- gseGO(geneList = gene_list,
              OrgDb = org.Hs.eg.db,
              keyType = "SYMBOL",
              ont = "BP")

dotplot(gsea)
```

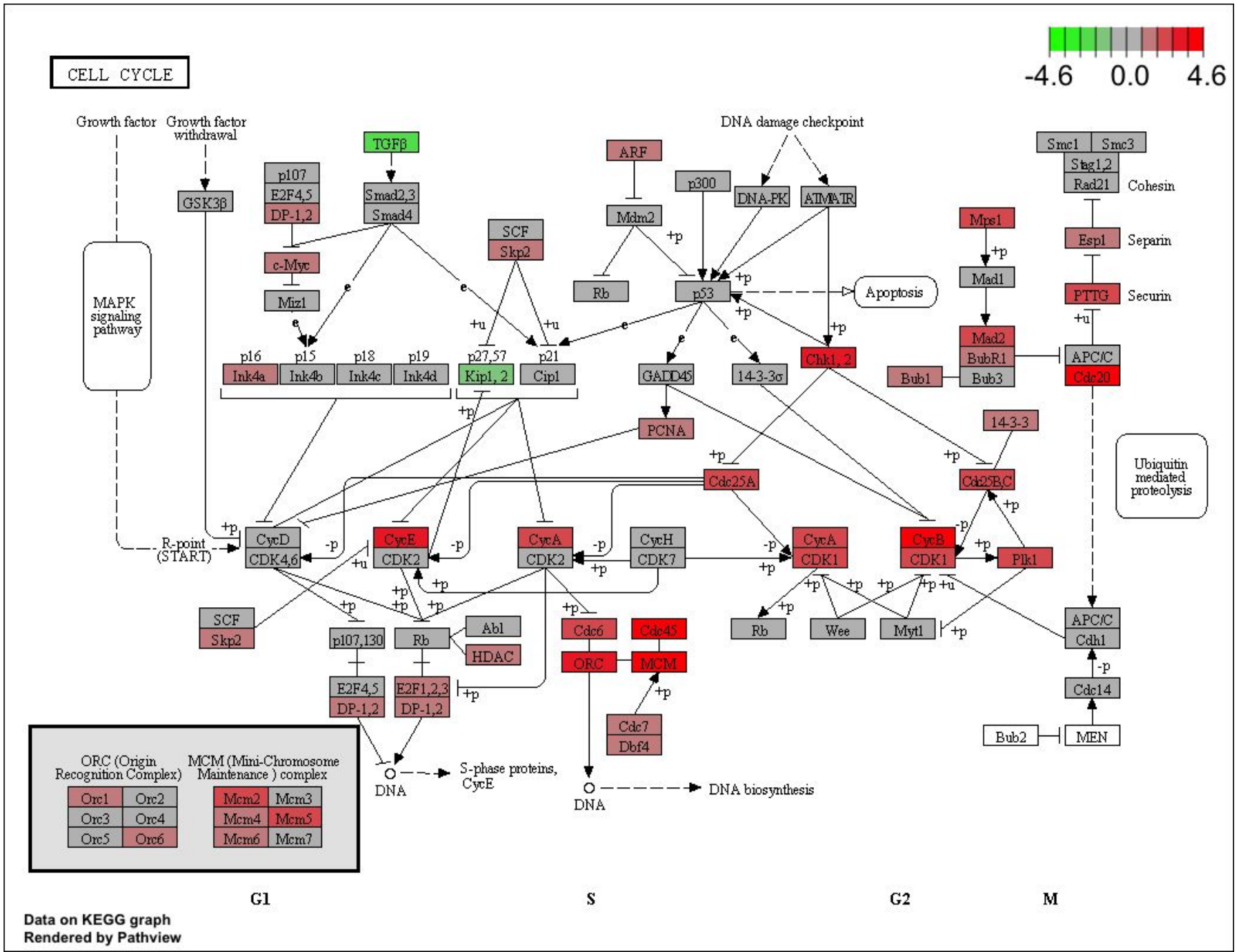



Gene Ontology: Cellular Component
CACNB4
DPP6
CNGB3
LRP1B
CHRNB2
transmembrane transporter complex
GRIA3
transporter complex
NOS1AP
HCN1
MUSK
TRPC5
receptor complex
CNIH2
GLRA1
GRIK2
ERBB4
postsynaptic membrane
GRM7
GRIP1
ZNF804A
EPHB1
dendrite
glutamatergic synapse
SLC8A3
SYT1
NSG1
CPLX2
DGKB
IL1RAPL2
NSG2
multivesicular body membrane
COX2
respirasome
oxidoreductase complex
ND1
CYTB
ND4
NADH dehydrogenase complex
ND2
ND5
inner mitochondrial membrane protein complex
ATP8
IMMP2L
ATP6
fold change
4.0
3.5
3.0
2.5
size
2
6
9
13


Gene Ontology: Molecular Function
clathrin binding
inositol 1,4,5 trisphosphate binding
NADH dehydrogenase (ubiquinone) activity
cyclic nucleotide binding
transmembrane signaling receptor activity
signaling receptor activator activity
cytokine activity
receptor regulator activity
ion channel activity
p.adjust
0.01
0.02
0.03
0.04
size
4
8
12
16


Type I diabetes mellitus
Viral myocarditis
Allograft rejection
Autoimmune thyroid disease
ABC transporters
Graft-versus-host disease
PD-L1 expression and PD-1 checkpoint pathway in cancer
Hematopoietic cell lineage
Th17 cell differentiation
Epstein-Barr virus infection
Complement and coagulation cascades
Natural killer cell mediated cytotoxicity
Th1 and Th2 cell differentiation
T cell receptor signaling pathway
Cell adhesion molecules
Primary immunodeficiency
Antigen processing and presentation
CD8A CD8B CIITA HLA-B HLA-C HLA-E HLA-F KLRC1 KLRD1 PSME1 PSME2 TAP1 TAPBP CD3D CD3E CD3G LCK ZAP70 CD2 CD6 CDH15 CLDN5 ICAM3 ITGAL ITGB7 PDCD1 TIGIT CARD11 CD247 CD3G ITK PIK3CD PRKCQ IL12RB1 IL2RB RUNX3 STAT1 TBX21 CD244 NCR1 NCR3 PRF1 C4B C4BPA C8B CFHR1 CR2 F11 F8 FGG PROC SERPINA1 VTN ISG15 TRAF5 IL2IR CD5 CD7 ITGA3 MS4A1 ABCA7 ABCC3 ABCC6 ABCD1 ABCD2 TPO
fold change
-1
-2
-3
-4
-5

# KEGG


-4.6 0.0 4.6
CELL CYCLE
Growth factor
Growth factor withdrawal
GSK3β
TGFβ
p107
E2F4,5
DP-1,2
Smad2,3
Smad4
c-Myc
Miz1
MAPK signaling pathway
SCF
Skp2
ARF
Mdm2
p300
DNA-PK
ATM/ATR
DNA damage checkpoint
Rb
p53
Apoptosis
p16 Ink4a
p15 Ink4b
p18 Ink4c
p19 Ink4d
p27,57 Kip1, 2
p21 Cip1
GADD45
14-3-3σ
Chk1, 2
PCNA
Mps1
Mad1
Mad2
BubR1
Bub3
Bub1
Smc1
Smc3
Stag1,2
Rad21
Cohesin
Esp1
Separin
PTTG
Securin
APC/C
Cdc20
14-3-3
Cdc25A
Cdc25B,C
Ubiquitin mediated proteolysis
R-point (START)
CycD
CDK4,6
CycE
CDK2
CycA
CDK2
CycH
CDK7
CycA
CDK1
CycB
CDK1
Plk1
SCF
Skp2
p107,130
Rb
Abl
HDAC
E2F4,5
DP-1,2
E2F1,2,3
DP-1,2
Cdc6
ORC
Cdc45
MCM
Cdc7
Dbf4
Rb
Wee
Myt1
APC/C
Cdh1
Cdc14
Bub2
MEN
DNA
S-phase proteins, CycE
DNA
DNA biosynthesis
ORC (Origin Recognition Complex)
Orc1 Orc2
Orc3 Orc4
Orc5 Orc6
MCM (Mini-Chromosome Maintenance ) complex
Mcm2 Mcm3
Mcm4 Mcm5
Mcm6 Mcm7
G1
S
G2
M
Data on KEGG graph
Rendered by Pathview

# Что почитать?

Анализ дифференциальной экспрессии:

- DESeq2
  https://doi.org/10.1186/s13059-014-0550-8
  http://www.bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.html
- https://bioconductor.org/packages/release/workflows/vignettes/rnaseqGene/inst/doc/rnaseqGene.html
- https://doi.org/10.1186/s12859-019-2599-6
- https://chagall.med.cornell.edu/RNASEQcourse/Intro2RNAseq.pdf
- https://www.biorxiv.org/content/10.1101/2020.06.10.144063v1.full.pdf

Функциональная аннотация генов:

- clusterProfiler
  http://yulab-smu.top/clusterProfiler-book/index.html
- https://bioinformatics-core-shared-training.github.io/cruk-summer-school-2018/RNASeq2018/html/06_Gene_set_testing.nb.html

R пакеты swirl и BiocSwirl - интерактивное обучение в R.