БИ ДЗ-6. Анализ дифференциальной экспрессии генов

SRA. FastQC

1. Find information about all .fastq files in accession_list.txt file

```
cp -i common/Differential_expression/accession_list.txt common/K
cat common/K/accession_list.txt
```

Output:

> SRR18066729
> SRR18066739

2. Write two sentences about the differences between RNA-seq experiment and WGS.

RNA-seq - это метод секвенирования РНК (как кодирующих, так и некодирующих).
WGS - это метод, при котором происходит секвенирование всего генома.

3. Answers the following questions for ALL entries in accessions_list.txt.

Ответы нашла здесь:

1️⃣ SRR18066729: https://www.ncbi.nlm.nih.gov/sra/?term=SRR18066729

2️⃣ SRR18066739: https://www.ncbi.nlm.nih.gov/sra/?term=SRR18066739

From what place of organism, biological material was taken to obtain these.fastq files?

🍒gut of human

What university submitted them?

🍒Gan Nan Medical University

On which sequenator was the experiment conducted?

🍒Illumina NovaSeq 6000

It is pair end data. What is the length of the reads?

🍒250 pb

How many nucleotides were sequenced? How many clusters of DNA fragments were formed in each run?

1️⃣ 31.4M bases, 62 810 spots

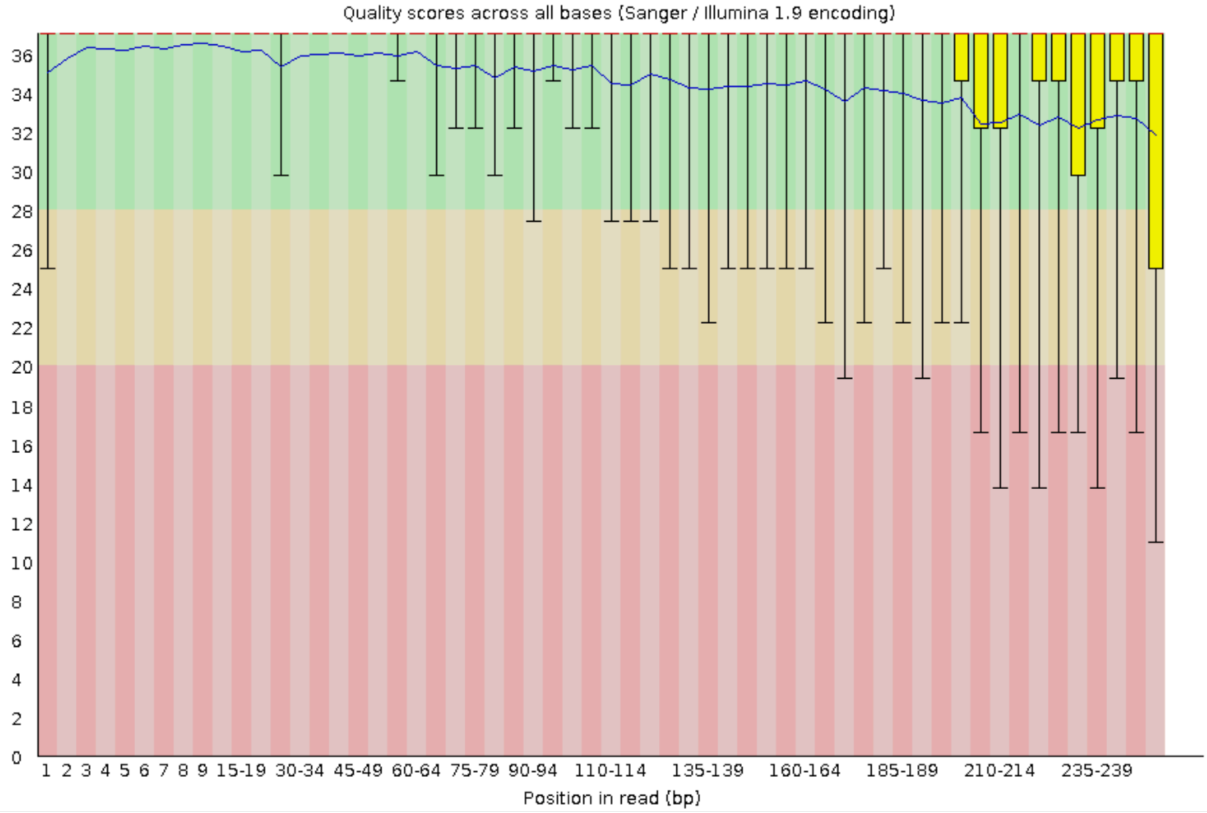
2️⃣ 32.7M bases, 65,394 spots

4. Для fastq из первого задания нужно запустить fastQC

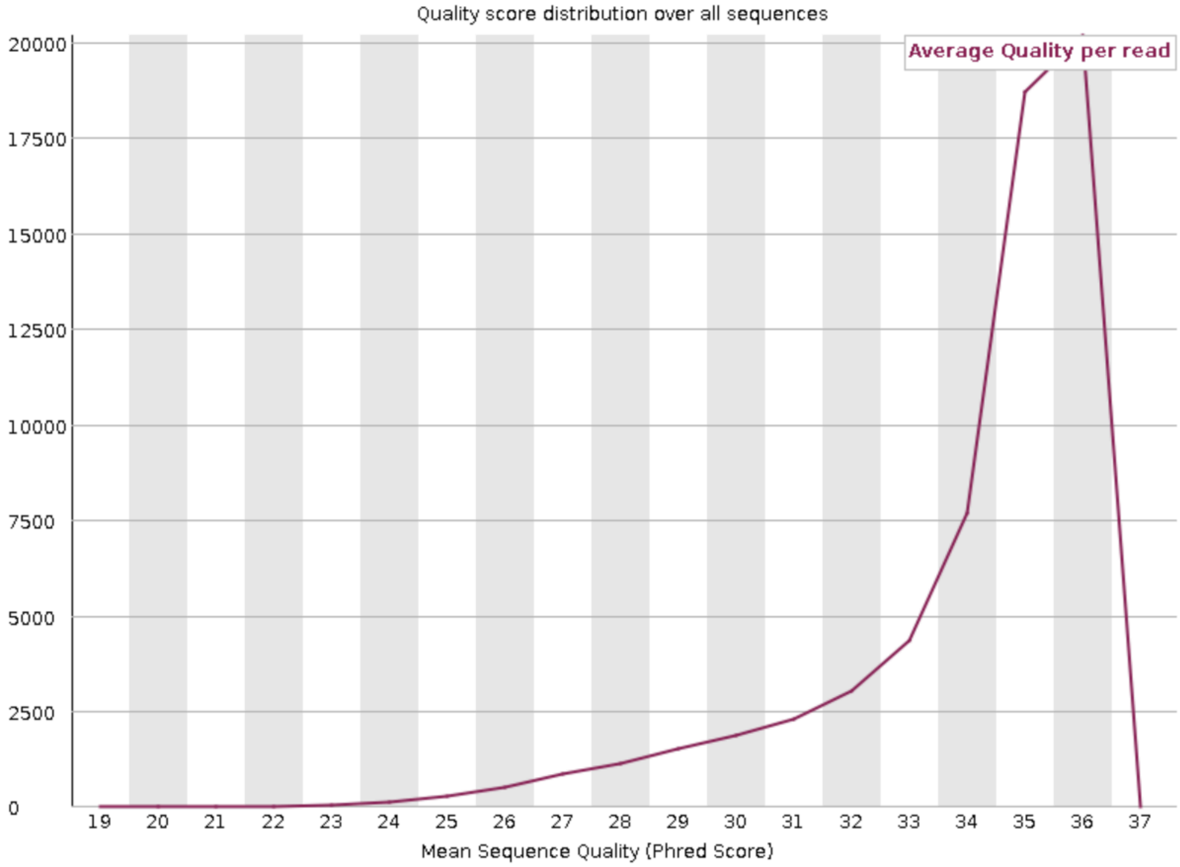
Скопировала себе все файлы из папки common/Differential_expression/transcriptomics на сервере. Далее запустила fastqc для одного из файлов и получила report.html.

```
fastqc common/K/SRR18066729_1.fastq
```

✓ Per base sequence quality



✓ Per sequence quality scores

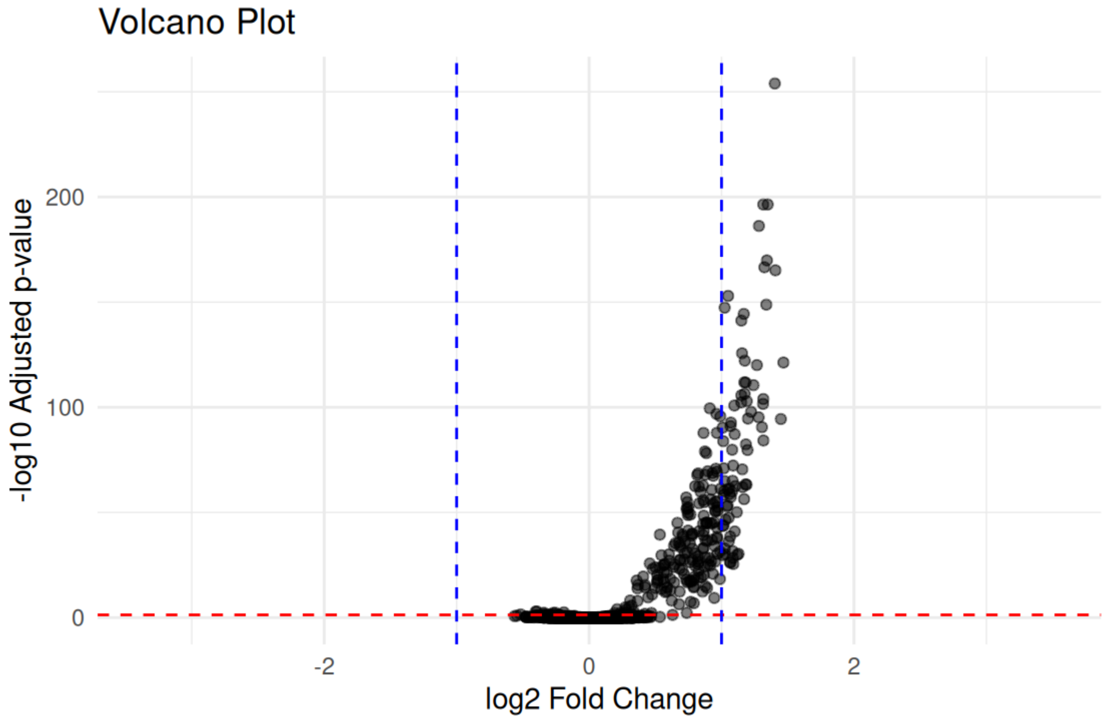


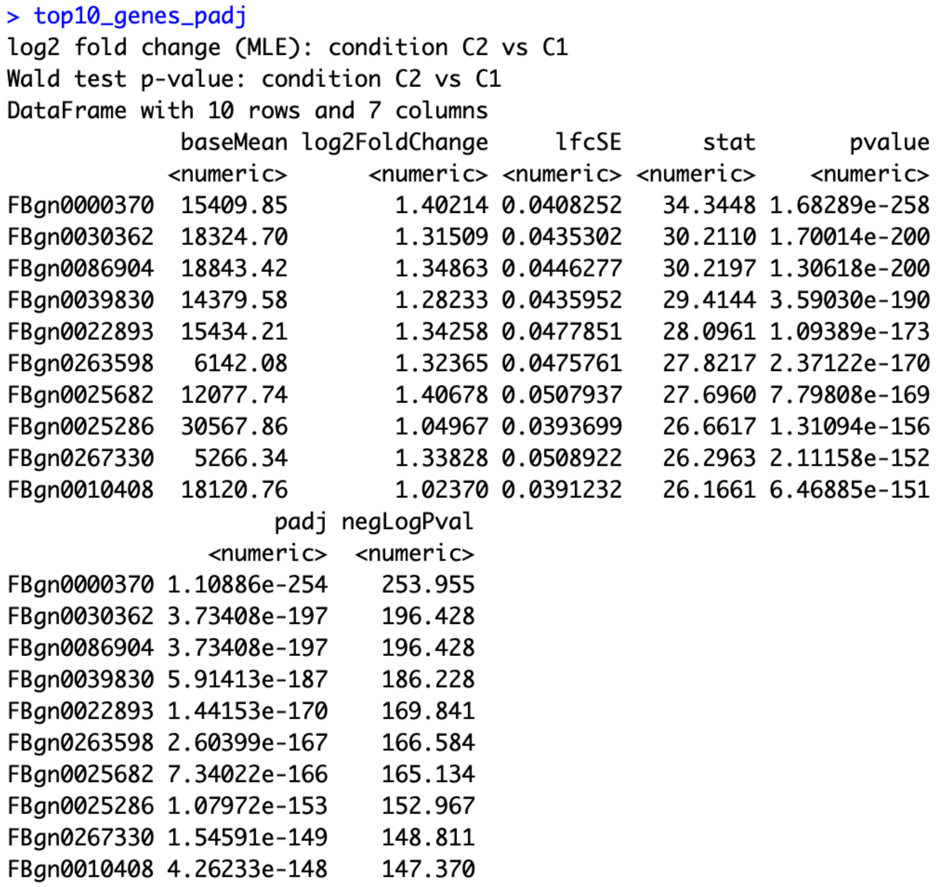
Differential Gene Expression

💅 Tutorial 1: Using DESeq2 and edgeR

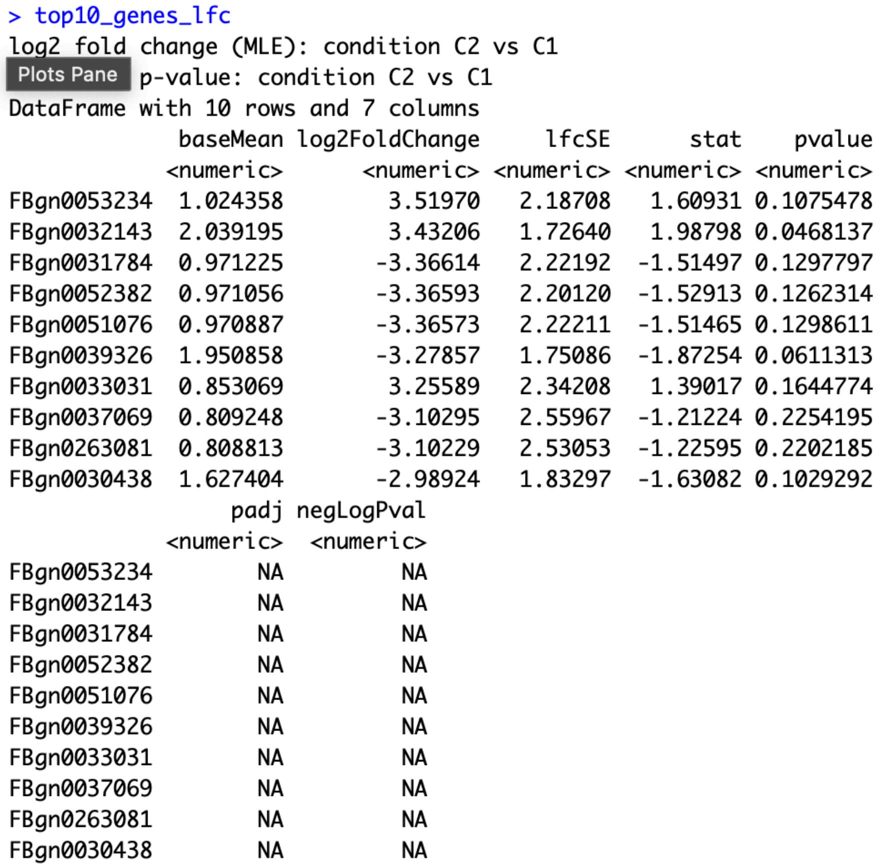
https://gtpb.github.io/ADER18F/pages/tutorial1.html

1. **provide a volcano plot split based on p-adj and lo2FC and provide top10 genes according to each metric. Which one is more preferable to use in our example? why?**





```
> top10_genes_padj
log2 fold change (MLE): condition C2 vs C1
Wald test p-value: condition C2 vs C1
DataFrame with 10 rows and 7 columns
             baseMean log2FoldChange     lfcSE      stat       pvalue
            <numeric>      <numeric> <numeric> <numeric>    <numeric>
FBgn0000370  15409.85        1.40214 0.0408252   34.3448 1.68289e-258
FBgn0030362  18324.70        1.31509 0.0435302   30.2110 1.70014e-200
FBgn0086904  18843.42        1.34863 0.0446277   30.2197 1.30618e-200
FBgn0039830  14379.58        1.28233 0.0435952   29.4144 3.59030e-190
FBgn0022893  15434.21        1.34258 0.0477851   28.0961 1.09389e-173
FBgn0263598   6142.08        1.32365 0.0475761   27.8217 2.37122e-170
FBgn0025682  12077.74        1.40678 0.0507937   27.6960 7.79808e-169
FBgn0025286  30567.86        1.04967 0.0393699   26.6617 1.31094e-156
FBgn0267330   5266.34        1.33828 0.0508922   26.2963 2.11158e-152
FBgn0010408  18120.76        1.02370 0.0391232   26.1661 6.46885e-151
                    padj negLogPval
               <numeric>  <numeric>
FBgn0000370 1.10886e-254    253.955
FBgn0030362 3.73408e-197    196.428
FBgn0086904 3.73408e-197    196.428
FBgn0039830 5.91413e-187    186.228
FBgn0022893 1.44153e-170    169.841
FBgn0263598 2.60399e-167    166.584
FBgn0025682 7.34022e-166    165.134
FBgn0025286 1.07972e-153    152.967
FBgn0267330 1.54591e-149    148.811
FBgn0010408 4.26233e-148    147.370
```



```
> top10_genes_lfc
log2 fold change (MLE): condition C2 vs C1
           p-value: condition C2 vs C1
DataFrame with 10 rows and 7 columns
             baseMean log2FoldChange     lfcSE      stat    pvalue
            <numeric>      <numeric> <numeric> <numeric> <numeric>
FBgn0053234  1.024358        3.51970   2.18708   1.60931 0.1075478
FBgn0032143  2.039195        3.43206   1.72640   1.98798 0.0468137
FBgn0031784  0.971225       -3.36614   2.22192  -1.51497 0.1297797
FBgn0052382  0.971056       -3.36593   2.20120  -1.52913 0.1262314
FBgn0051076  0.970887       -3.36573   2.22211  -1.51465 0.1298611
FBgn0039326  1.950858       -3.27857   1.75086  -1.87254 0.0611313
FBgn0033031  0.853069        3.25589   2.34208   1.39017 0.1644774
FBgn0037069  0.809248       -3.10295   2.55967  -1.21224 0.2254195
FBgn0263081  0.808813       -3.10229   2.53053  -1.22595 0.2202185
FBgn0030438  1.627404       -2.98924   1.83297  -1.63082 0.1029292
                 padj negLogPval
            <numeric>  <numeric>
FBgn0053234        NA         NA
FBgn0032143        NA         NA
FBgn0031784        NA         NA
FBgn0052382        NA         NA
FBgn0051076        NA         NA
FBgn0039326        NA         NA
FBgn0033031        NA         NA
FBgn0037069        NA         NA
FBgn0263081        NA         NA
FBgn0030438        NA         NA
```

Лучше использовать p-adj (скорректированные p-значения), т.к. они учитывают множественные сравнения и вероятность ложноположительных результатов. Однако, log2FC помогает понять, насколько сильно изменилась экспрессия генов. Так что в первую очередь смотрим на значение p-adj, однако также учитываем и значение log2FC.

2. **repeat the workflow with and without normalization, and give all plots and top10 genes (from each side) for the comparison**

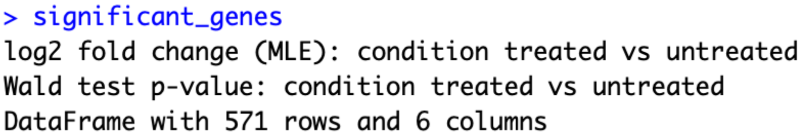
🍄

💅Tutorial 2: Pasilla

https://introtogenomics.readthedocs.io/en/latest/2021.11.11.DeseqTutorial.html

1. how many significant genes considering alpha= 0.01 for FDR

```
res05['padj'] <- as.numeric(res05['padj'])
significant_genes = res05[!is.na(res05$padj) & res05$padj < 0.01, ]
significant_genes
```

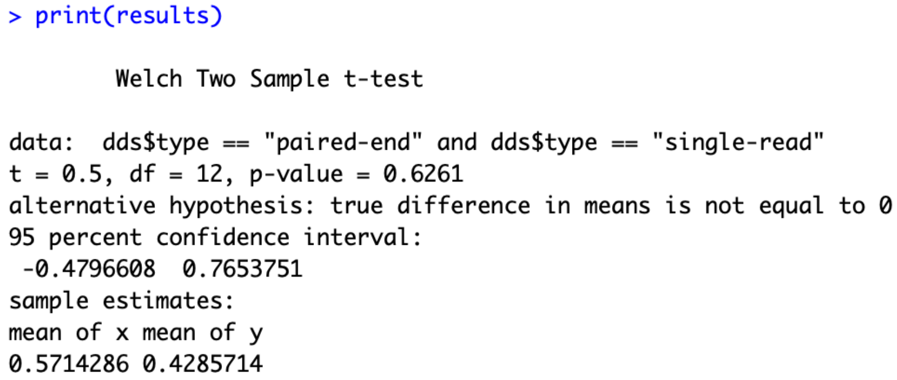


```
> significant_genes
log2 fold change (MLE): condition treated vs untreated
Wald test p-value: condition treated vs untreated
DataFrame with 571 rows and 6 columns
```

Ответ: 571

2. Establish a simple model between read type (paired/single end) and condition, is there any significant bias introduced by read type?

```
results <- t.test(dds$type == "paired-end", dds$type == "single-read")
```



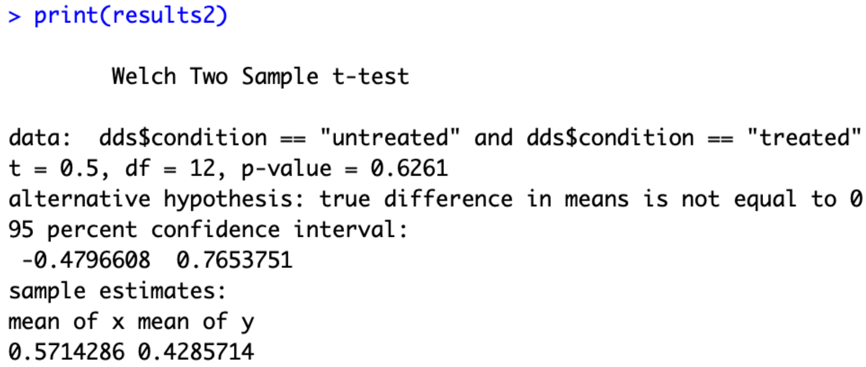
```
> print(results)

        Welch Two Sample t-test

data:  dds$type == "paired-end" and dds$type == "single-read"
t = 0.5, df = 12, p-value = 0.6261
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.4796608  0.7653751
sample estimates:
mean of x mean of y
0.5714286 0.4285714
```

Так как p > 0.05 (p-value = 0.6261), значимых различий экспрессии при разных типах чтения нет.

```
results2 <- t.test(dds$condition == "untreated", dds$condition == "treated")
```



```
> print(results2)

        Welch Two Sample t-test

data:  dds$condition == "untreated" and dds$condition == "treated"
t = 0.5, df = 12, p-value = 0.6261
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.4796608  0.7653751
sample estimates:
mean of x mean of y
0.5714286 0.4285714
```

Так как p > 0.05 (p-value = 0.6261), значимых различий экспрессии при разных типах чтения нет.

Забавно, что результаты t-test одинаковые. Но это чистейшее совпадение.