Projekt zaliczeniowy

Ksenia Kvitko

26 06 2020

1. Ładowanie bibliotek

```
library(DESeq2)
library(ggplot2)
library(ComplexHeatmap)
```

2. Informacje o danych bioprojektu (PRJNA313294)

```
SRAruns <- c("SRR3191542", "SRR3191543", "SRR3191544", "SRR3191545", "SRR3194428", "SRR3194429", "SRR31
SRAitems <- c("Mock1-1", "Mock2-1", "ZIKV1-1", "ZIKV2-1", "Mock1-2", "Mock2-2", "ZIKV1-2", "ZIKV2-2")
SRAdevices <- c(rep("Illumina MiSeq", times = 4), rep("Illumina NextSeq 500", times = 4))
SRAshortcuts <- c("M1_MiSeq", "M2_MiSeq", "Z1_MiSeq", "Z2_MiSeq", "M1_NextSeq", "M2_NextSeq", "Z1_NextS

SRAdata <- rbind(SRAruns, SRAitems, SRAdevices, SRAshortcuts)
print(SRAdata)
```

```
##              [,1]             [,2]             [,3]
## SRAruns      "SRR3191542"     "SRR3191543"     "SRR3191544"
## SRAitems     "Mock1-1"        "Mock2-1"        "ZIKV1-1"
## SRAdevices   "Illumina MiSeq" "Illumina MiSeq" "Illumina MiSeq"
## SRAshortcuts "M1_MiSeq"       "M2_MiSeq"       "Z1_MiSeq"
##              [,4]             [,5]                   [,6]
## SRAruns      "SRR3191545"     "SRR3194428"           "SRR3194429"
## SRAitems     "ZIKV2-1"        "Mock1-2"              "Mock2-2"
## SRAdevices   "Illumina MiSeq" "Illumina NextSeq 500" "Illumina NextSeq 500"
## SRAshortcuts "Z2_MiSeq"       "M1_NextSeq"           "M2_NextSeq"
##              [,7]                   [,8]
## SRAruns      "SRR3194430"           "SRR3194431"
## SRAitems     "ZIKV1-2"              "ZIKV2-2"
## SRAdevices   "Illumina NextSeq 500" "Illumina NextSeq 500"
## SRAshortcuts "Z1_NextSeq"           "Z2_NextSeq"
```

3. Wczytanie danych zmapowanych do genomu hg19

```
counts_all <- read.delim("~/projekt/analizaTranskryptomu/projekt/hg19/COUNTS/counts_ALL.txt", comment.c
colnames(counts_all)[7:14] <- SRAshortcuts

counts_paired <- read.delim("~/projekt/analizaTranskryptomu/projekt/hg19/COUNTS/counts_PE.txt", comment
colnames(counts_paired)[7:10] <- SRAshortcuts[1:4]
```

```
counts_single <- read.delim("~/projekt/analizaTranskryptomu/projekt/hg19/COUNTS/counts_SE.txt", comment
colnames(counts_single)[7:10] <- SRAshortcuts[5:8]
```

4. Analiza DE

```
dds <- function(data, runs) {
  countData <- data[,7:(7+runs-1)]
  rownames(countData) = data$Geneid
  samples <- names(countData)
  if(runs > 4) {
    condition <- factor(c("mock", "mock", "zika", "zika", "mock", "mock", "zika", "zika"))
  }
  else {
    condition <- factor(c("mock", "mock", "zika", "zika"))
  }
  colData <- data.frame(samples = samples, condition = condition)
  dds <- DESeqDataSetFromMatrix(countData = countData, colData = colData, design = ~condition)

  return(dds)
}

analyseDE = function(data) {
  dds <- DESeq(data)
  res <- results(dds)

  r = res[res$baseMean!=0,]
  r = r[r$log2FoldChange > 1 | r$log2FoldChange < -1,]
  x = !is.na(r$padj)
  r = r[x,]
  r = r[r$padj<0.05,]

  print(head(r))

  return(dds)
}
```

4.1. Zbiorcza

```
DE_all <- analyseDE(dds(counts_all, 8))
```

```
## log2 fold change (MLE): condition zika vs mock
## Wald test p-value: condition zika vs mock
## DataFrame with 6 rows and 6 columns
##                       baseMean    log2FoldChange              lfcSE
##                      <numeric>         <numeric>          <numeric>
## VWA1         334.426316217386 -1.32578817670352 0.0969376879419521
## MMP23B       64.2662818719584 -1.20821536435696  0.256532925580776
## CFAP74       29.9194006545816 -1.81721702770035  0.335123059831466
## LOC100129534 48.4952917103514 -1.24258583913354  0.248226437135996
## TNFRSF14-AS1 50.0132552352106  5.55556197239441  0.508976370901608
## TNFRSF14     39.5661970198706  3.71195816205217  0.369528973318885
##                           stat               pvalue                 padj
##                      <numeric>            <numeric>            <numeric>
## VWA1         -13.6767051582396 1.39889139887406e-42 1.04063200715454e-40
## MMP23B       -4.70978671303745 2.47976153514423e-06 1.20388217385792e-05
## CFAP74        -5.4225365112572 5.87592075449711e-08 3.60238555016946e-07
## LOC100129534  -5.0058561588779 5.56142771889026e-07 2.96677517640969e-06
## TNFRSF14-AS1  10.9151667739569 9.75500944865484e-28 3.30917241530221e-26
## TNFRSF14      10.0451072312777 9.65426209400259e-24 2.47513273088438e-22
```

4.2. Dla każdego z urządzeń

Illumina MiSeq

```
DE_paired <- analyseDE(dds(counts_paired, 4))
```

```
## log2 fold change (MLE): condition zika vs mock
## Wald test p-value: condition zika vs mock
## DataFrame with 6 rows and 6 columns
##                      baseMean    log2FoldChange             lfcSE
##                     <numeric>         <numeric>         <numeric>
## VWA1         75.5398925477733 -1.27710020252877 0.310291019365312
## TNFRSF14-AS1 15.7621492406834  5.98721330457336  1.54916859724598
## TNFRSF14     10.0893229839815  4.30846211355709  1.26716465965906
## MEGF6        417.300756458076   -1.197162158804  0.14170808792281
## LNCTAM34A    18.4667693307206 -1.60355478272455 0.635892821555479
## TMEM51       76.7978830409192 -1.08678049119696  0.30771355360715
##                           stat               pvalue                 padj
##                      <numeric>            <numeric>            <numeric>
## VWA1         -4.11581426088652  3.8581491608349e-05 0.000309091607984836
## TNFRSF14-AS1  3.86479129206278 0.000111184229442529 0.000783191516013005
## TNFRSF14       3.4000807083085 0.000673659657991891  0.00373635297489782
## MEGF6        -8.44808631851775 2.96116419348808e-17 1.55932466518511e-15
## LNCTAM34A    -2.52173751356721   0.0116776810442466   0.0415245815584491
## TMEM51       -3.53179272884556 0.000412752741205068  0.00243836283249716
```

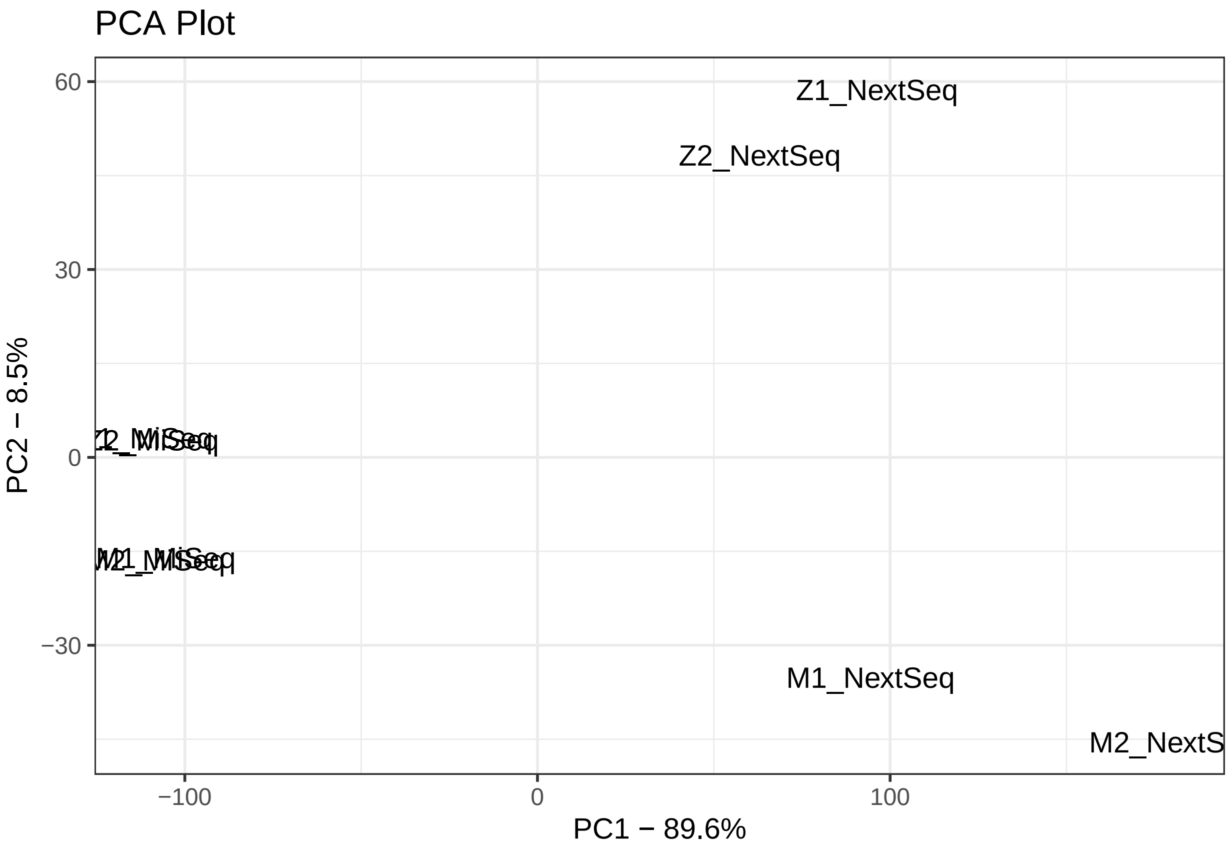
Illumina NextSeq 500

```
DE_single <- analyseDE(dds(counts_single, 4))
```

```
## log2 fold change (MLE): condition zika vs mock
## Wald test p-value: condition zika vs mock
## DataFrame with 6 rows and 6 columns
##                      baseMean    log2FoldChange             lfcSE
##                     <numeric>         <numeric>         <numeric>
## LINC02593      155.3179313799  -1.0667898894932 0.224527222396563
## VWA1         750.709188115296 -1.35230758965782 0.110248830255318
## MMP23B       114.073377704044 -1.16619139667573 0.261206019938957
## CFAP74       67.2196258284694 -2.00458276889009 0.366954567515667
## GABRD        26.0893918827756 -1.82675486692474  0.57611650293165
## LOC100129534 111.017280389548 -1.29314098690466 0.273434690240048
##                           stat               pvalue                 padj
##                      <numeric>            <numeric>            <numeric>
## LINC02593    -4.75127193088872 2.02141068919781e-06 7.35889188230861e-06
## VWA1         -12.2659586185731 1.37969709828036e-34 3.22381776525957e-33
## MMP23B       -4.46464211256793 8.02027001183273e-06 2.73707395838519e-05
## CFAP74       -5.46275464688011 4.68802339767858e-08 2.02722708021764e-07
## GABRD        -3.17080808765075  0.00152015537041154  0.00383060370870692
## LOC100129534  -4.7292499198599 2.25350844351798e-06 8.15950963445154e-06
```

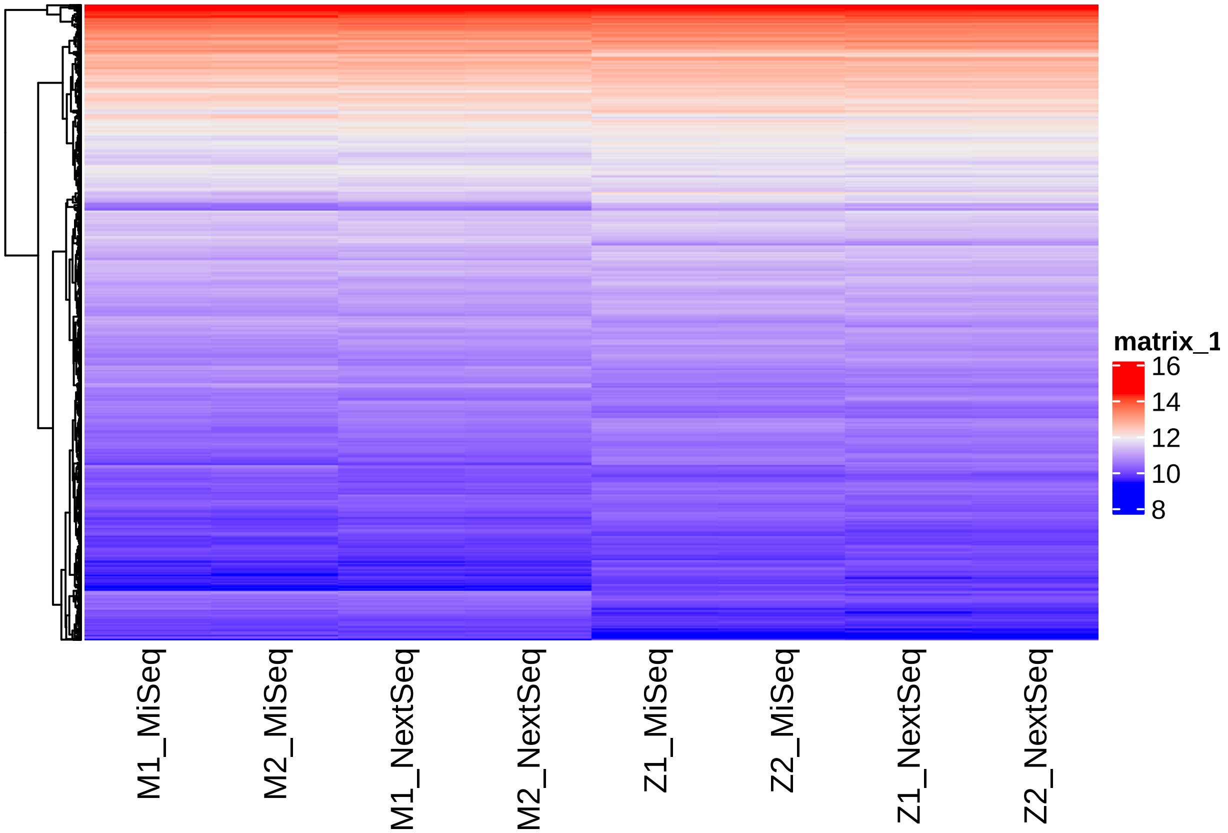
5. Analiza PCA

```
PCAnalysis <- function(data, runs) {
  gene_data <- data[7:14]
  rownames(gene_data) <- data$Geneid

  row_sub <- apply(gene_data, 1, function(row) all(row != 0))
  gene_data <- gene_data[row_sub,]
  gene_data_matrix <- as.matrix(gene_data)

  pca <- prcomp(t(gene_data_matrix), scale = T)
  pca.data <- data.frame(Sample = rownames(pca$x), X = pca$x[,1], Y = pca$x[,2])
  pca.var <- pca$sdev^2
  pca.var.per <- round(pca.var/sum(pca.var)*100, 1)

  ggplot(data = pca.data, aes(x = X, y = Y, label = Sample)) +
    geom_text() +
    xlab(paste("PC1 - ", pca.var.per[1], "%", sep="")) +
    ylab(paste("PC2 - ", pca.var.per[2], "%", sep="")) +
    theme_bw() +
    ggtitle("PCA Plot")

}

PCAnalysis(counts_all)
```



6. Heatmapa

```
normalize <- function(data) {
  log_data <- rlog(data)
  normalized_data<- assay(log_data)
  normalized_data <- as.data.frame(normalized_data)

  return(normalized_data)
}

drawHeatmap <- function(data){
  threshold <- 10
  data <- data[
      data$M1_MiSeq > threshold |
      data$M2_MiSeq > threshold |
      data$Z1_MiSeq > threshold |
      data$Z2_MiSeq > threshold |
      data$M1_NextSeq > threshold |
      data$M2_NextSeq > threshold |
      data$Z1_NextSeq > threshold |
      data$Z2_NextSeq > threshold
    , ]
  data <- data.frame(data$M1_MiSeq, data$M2_MiSeq, data$M1_NextSeq, data$M2_NextSeq, data$Z1_NextSeq, da
  names(data) = c(SRAshortcuts[1], SRAshortcuts[2], SRAshortcuts[5], SRAshortcuts[6], SRAshortcuts[3], S
  Heatmap(data , cluster_columns = FALSE,
          row_names_side = "left",
          row_dend_sid = "left",
          row_names_gp=gpar(cex = 0.8))
}

drawHeatmap(normalize(dds(counts_all, 8)))
```


matrix_1
M1_MiSeq
M2_MiSeq
M1_NextSeq
M2_NextSeq
Z1_MiSeq
Z2_MiSeq
Z1_NextSeq
Z2_NextSeq