differential gene expression in Rstudio on data we will use for Chipseq.

In [ ]:
```
# a different dataset was used for Chipseq because sra files weren't avaiable for the da
# source of the data: http://bioconductor.org/packages/release/data/experiment/html/airw
# GEO entry GSE52778, PubMed entry 24926665
# workdirectory: /home/luna.kuleuven.be/r0378624/data/work
```

In [ ]:
```
# we have 8 samples (4 treated with Dexamethasone (Dex) and 4 untreated), the article di
# we want to see which genes are differentially expressed and between which samples are
# we will take a treated and untreated sample with high variability between groups and l
```

In [ ]:
```
# We use less than 12 biological replicates which is why we opt to use EdgeR instead of
# How many biological replicates are needed in an RNA-seq experiment and which different
```

In [23]:
```
# Rscript used to do differential gene expression and MDS plotting on the airway datapac
cat /home/luna.kuleuven.be/r0372261/data/work/Rscript.R
```

```
#install packages (airway data and EdgeR diff gene expression package)
source("https://bioconductor.org/biocLite.R")
biocLite("airway")

source("https://bioconductor.org/biocLite.R")
biocLite("edgeR")

#download these packages into Rstudio
library(airway)
library(edgeR)

#to learn how to use this datapackage:
?airway

#it tells us to use data("airway") to get the values of the datapackage in R
data ("airway")

#to get the data as a data frame into R we make a name for our dataframe
sampleData = as.data.frame(colData(airway))

#we can visualize this dataframe simply by typing sampledata
#this will give us information on the samples, treatments and so on.
sampleData

#the airway package contains a command called assay to put the data of the reads in a
 matrix
#with genes in every row for a sample in every collumn
assay <- assay(airway)
head (assay)

#we are now assigning our matrix assay(airway) to a simpler letter x
x <- assay(airway)

# and we add a grouping factor for our simple list based data object which we will mak
e in the next step.
#for our group we select the treatment column of our sampleData (= dex, see sampledat
a)
#this group will have 2 levels: treated vs untreated.
group = as.factor(sampleData$dex)
#now we have x (our matrix) and our grouping factor which will be combined and stored
 in the DGEList which is a data object
y <- DGEList(counts=x,group=group)

#we now have to normalize the gene counts because we want to know relative- instead of
 absolute differences
#we normalize based on RNA composition to obtain the effective library size (norm fact
ors will now differ from 1)
y <- calcNormFactors(y)
y
#factors > 1 scales up the library size analogues to downscaling the gene counts.
#the library size will be adjusted in our downstream analysis.

#we can visualize the librarysize per sample in a barplot
barplot(y$samples$lib.size*1e-6, xlab="samples", names = (1:8), ylab = "library size i
n millions")

#we first have to look at dispersion before we can look at differential gene axpressio
n
#we want to estimate dispersion of the groups of our matrix so we first provide the gr
oup design
design <- model.matrix(~0+group, data=y$samples)
design
y <- estimateDisp(y, design)
```

```
#estimate the dispersion in a plot
plotBCV (y)

#we can now look for differentially expressed genes between our two groups (treated vs
 untreated)
et <- exactTest(y, pair=c("untrt","trt"))
topTags(et)
#we see that the 10 most significant changes are upregulations of expression in treate
d compared to untreated (exept for 1 gene)

#multidimensional scaling plots of log2 fold changes in distances between samples and
 groups.
Samples <- as.factor(sampleData$Sample)
plotMDS(y, labels = Samples)

plotMDS(y, labels = group)
```

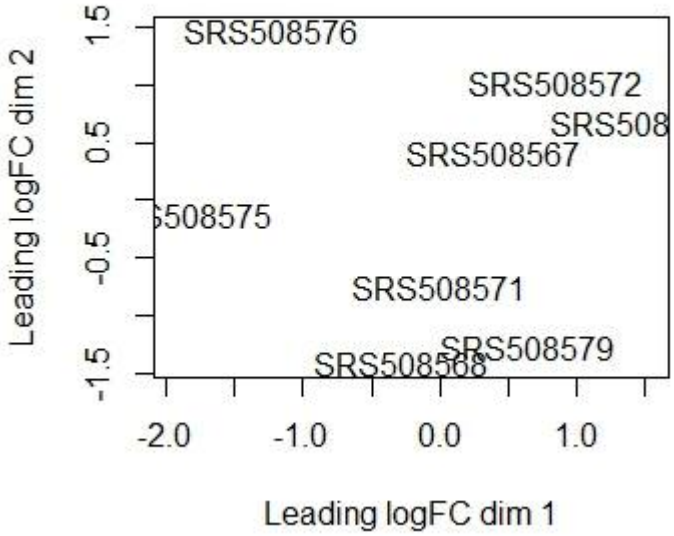
#image of the sampledata

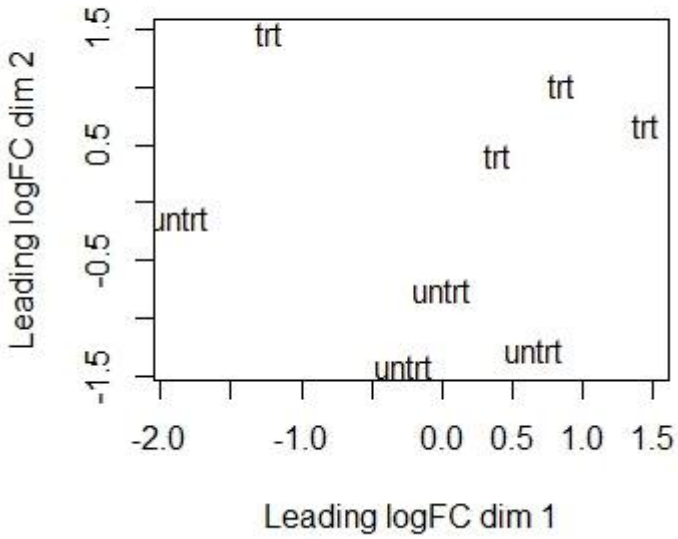
```
> sampleData
           SampleName    cell   dex albut        Run avgLength Experiment    Sample    BioSample
SRR1039508 GSM1275862  N61311 untrt untrt SRR1039508       126  SRX384345 SRS508568 SAMN02422669
SRR1039509 GSM1275863  N61311   trt untrt SRR1039509       126  SRX384346 SRS508567 SAMN02422675
SRR1039512 GSM1275866 N052611 untrt untrt SRR1039512       126  SRX384349 SRS508571 SAMN02422678
SRR1039513 GSM1275867 N052611   trt untrt SRR1039513        87  SRX384350 SRS508572 SAMN02422670
SRR1039516 GSM1275870 N080611 untrt untrt SRR1039516       120  SRX384353 SRS508575 SAMN02422682
SRR1039517 GSM1275871 N080611   trt untrt SRR1039517       126  SRX384354 SRS508576 SAMN02422673
SRR1039520 GSM1275874 N061011 untrt untrt SRR1039520       101  SRX384357 SRS508579 SAMN02422683
SRR1039521 GSM1275875 N061011   trt untrt SRR1039521        98  SRX384358 SRS508580 SAMN02422677
```

#toptag of differential gene expression between groups

```
Comparison of groups:  trt-untrt
                     logFC   logCPM       PValue          FDR
ENSG00000152583  4.584682 5.536240 2.377738e-81 1.524178e-76
ENSG00000179094  3.167542 5.177198 1.615845e-46 5.178945e-42
ENSG00000120129  2.932826 7.310715 5.199321e-40 1.110956e-35
ENSG00000189221  3.288819 6.769209 2.721966e-39 4.362087e-35
ENSG00000125148  2.188282 7.412816 2.334310e-38 2.992678e-34
ENSG00000178695 -2.514624 6.959467 7.557134e-38 8.073790e-34
ENSG00000109906  7.127071 4.162779 2.492621e-36 2.282599e-32
ENSG00000101347  3.842551 9.207488 4.338321e-34 3.476188e-30
ENSG00000179593 10.100807 1.658949 3.829289e-32 2.727390e-28
ENSG00000096060  3.921644 6.898829 6.424938e-32 4.118514e-28
```

#images of the MDS plots with sample labels and group labels





Chipseq and visualization in IGV

In [1]:
```
mkdir ~/data/ncbi
vdb-config -s /repository/user/main/public/root=~/data/ncbi
```

```
mkdir: cannot create directory ‘/home/luna.kuleuven.be/r0372261/data/ncbi’: File exist
s
```

In [2]:
```
mkdir ~/data/work
cd ~/data/work
```

```
mkdir: cannot create directory ‘/home/luna.kuleuven.be/r0372261/data/work’: File exist
s
```

In [3]:
```
# de sra files omzetten in fastq (we kozen een Dex treated sample vs een untreated sampl
fastq-dump SRR1039513
fastq-dump SRR1039512
```

```
Read 43356464 spots for SRR1039513
Written 43356464 spots for SRR1039513
Read 28136282 spots for SRR1039512
Written 28136282 spots for SRR1039512
```

In [4]:
```
#naam van de fastq files aanpassen (optioneel)
mv SRR1039513.fastq treated_sammple.fastq
mv SRR1039512.fastq untreated_sammple.fastq
```

In [5]:
```
# de fastq files alignen met ons referentiegenoom h19
bowtie2-align -x /mnt/nfs/data/resources/bowtie2/hg19 treated_sammple.fastq -S treated_s
bowtie2-align -x /mnt/nfs/data/resources/bowtie2/hg19 untreated_sammple.fastq -S untreat
```

```
43356464 reads; of these:
  43356464 (100.00%) were unpaired; of these:
    19953227 (46.02%) aligned 0 times
    17694247 (40.81%) aligned exactly 1 time
    5708990 (13.17%) aligned >1 times
53.98% overall alignment rate
28136282 reads; of these:
  28136282 (100.00%) were unpaired; of these:
    28127740 (99.97%) aligned 0 times
    7533 (0.03%) aligned exactly 1 time
    1009 (0.00%) aligned >1 times
0.03% overall alignment rate
```

In [6]:
```
#.sam files omzetten in .bam files
samtools view -S -b treated_sample_aligned.sam > treated_sample_aligned.bam
samtools view -S -b untreated_sample_aligned.sam > untreated_sample_aligned.bam
```

In [7]:
```
# sort and index the bam file
samtools sort -O bam -o treated_sample_aligned_sorted.bam treated_sample_aligned.bam
samtools index treated_sample_aligned_sorted.bam

samtools sort -O bam -o untreated_sample_aligned_sorted.bam untreated_sample_aligned.bam
samtools index untreated_sample_aligned_sorted.bam
```

```
[bam_sort_core] merging from 18 files...
[bam_sort_core] merging from 12 files...
```

In [8]:
```
#rename the files
mv treated_sample_aligned_sorted.bam treatedS1.bam
mv untreated_sample_aligned_sorted.bam untreatedS2.bam
```

In [20]:
```
#peakcalling on the aligned, sorted and indexed file
mkdir ~/data/macs2_peakcalling
macs2 callpeak -t  treatedS1.bam --outdir macs2_peakcalling -n SRR1039513 -g 59128983 -q

macs2 callpeak -t  untreatedS2.bam --outdir macs2_peakcalling -n SRR1039512 -g 59128983
```

```
mkdir: cannot create directory ‘/home/luna.kuleuven.be/r0372261/data/macs2_peakcallin
g’: File exists
INFO  @ Sun, 29 Jan 2017 11:30:51:
# Command line: callpeak -t treatedS1.bam --outdir macs2_peakcalling -n SRR1039513 -g
 59128983 -q 0.005
# ARGUMENTS LIST:
# name = SRR1039513
# format = AUTO
# ChIP-seq file = ['treatedS1.bam']
# control file = None
# effective genome size = 5.91e+07
# band width = 300
# model fold = [5, 50]
# qvalue cutoff = 5.00e-03
# Larger dataset will be scaled towards smaller dataset.
# Range for calculating regional lambda is: 10000 bps
# Broad region calling is off

INFO  @ Sun, 29 Jan 2017 11:30:51: #1 read tag files...
INFO  @ Sun, 29 Jan 2017 11:30:51: #1 read treatment tags...
Traceback (most recent call last):
  File "/usr/local/bin/macs2", line 614, in <module>
    main()
  File "/usr/local/bin/macs2", line 56, in main
    run( args )
  File "/usr/local/lib/python2.7/dist-packages/MACS2/callpeak_cmd.py", line 70, in run
    else:       (treat, control) = load_tag_files_options  (options)
  File "/usr/local/lib/python2.7/dist-packages/MACS2/callpeak_cmd.py", line 392, in lo
ad_tag_files_options
    tp = options.parser(options.tfile[0], buffer_size=options.buffer_size)
  File "MACS2/IO/Parser.pyx", line 68, in MACS2.IO.Parser.guess_parser (MACS2/IO/Parse
r.c:2754)
  File "MACS2/IO/Parser.pyx", line 90, in MACS2.IO.Parser.guess_parser (MACS2/IO/Parse
r.c:2410)
  File "MACS2/IO/Parser.pyx", line 670, in MACS2.IO.Parser.BAMParser.__init__ (MACS2/I
O/Parser.c:8989)
  File "/usr/lib/python2.7/gzip.py", line 34, in open
    return GzipFile(filename, mode, compresslevel)
  File "/usr/lib/python2.7/gzip.py", line 94, in __init__
    fileobj = self.myfileobj = __builtin__.open(filename, mode or 'rb')
IOError: [Errno 2] No such file or directory: 'treatedS1.bam'
INFO  @ Sun, 29 Jan 2017 11:30:52:
# Command line: callpeak -t untreatedS2.bam --outdir macs2_peakcalling -n SRR1039512 -
g 59128983 -q 0.005
# ARGUMENTS LIST:
# name = SRR1039512
# format = AUTO
# ChIP-seq file = ['untreatedS2.bam']
# control file = None
# effective genome size = 5.91e+07
# band width = 300
# model fold = [5, 50]
# qvalue cutoff = 5.00e-03
# Larger dataset will be scaled towards smaller dataset.
# Range for calculating regional lambda is: 10000 bps
# Broad region calling is off

INFO  @ Sun, 29 Jan 2017 11:30:52: #1 read tag files...
INFO  @ Sun, 29 Jan 2017 11:30:52: #1 read treatment tags...
Traceback (most recent call last):
  File "/usr/local/bin/macs2", line 614, in <module>
    main()
  File "/usr/local/bin/macs2", line 56, in main
```

```
    run( args )
  File "/usr/local/lib/python2.7/dist-packages/MACS2/callpeak_cmd.py", line 70, in run
    else:       (treat, control) = load_tag_files_options  (options)
  File "/usr/local/lib/python2.7/dist-packages/MACS2/callpeak_cmd.py", line 392, in lo
ad_tag_files_options
    tp = options.parser(options.tfile[0], buffer_size=options.buffer_size)
  File "MACS2/IO/Parser.pyx", line 68, in MACS2.IO.Parser.guess_parser (MACS2/IO/Parse
r.c:2754)
  File "MACS2/IO/Parser.pyx", line 90, in MACS2.IO.Parser.guess_parser (MACS2/IO/Parse
r.c:2410)
  File "MACS2/IO/Parser.pyx", line 670, in MACS2.IO.Parser.BAMParser.__init__ (MACS2/I
O/Parser.c:8989)
  File "/usr/lib/python2.7/gzip.py", line 34, in open
    return GzipFile(filename, mode, compresslevel)
  File "/usr/lib/python2.7/gzip.py", line 94, in __init__
    fileobj = self.myfileobj = __builtin__.open(filename, mode or 'rb')
IOError: [Errno 2] No such file or directory: 'untreatedS2.bam'
```

In [21]:
```
#see which files are in your directory now
ls
```

```
macs2_peakcalling            SRR1039513_model.r
SRR1039512_model.r           SRR1039513_peaks.narrowPeak
SRR1039512_peaks.narrowPeak  SRR1039513_peaks.xls
SRR1039512_peaks.xls         SRR1039513_summits.bed
SRR1039512_summits.bed
```

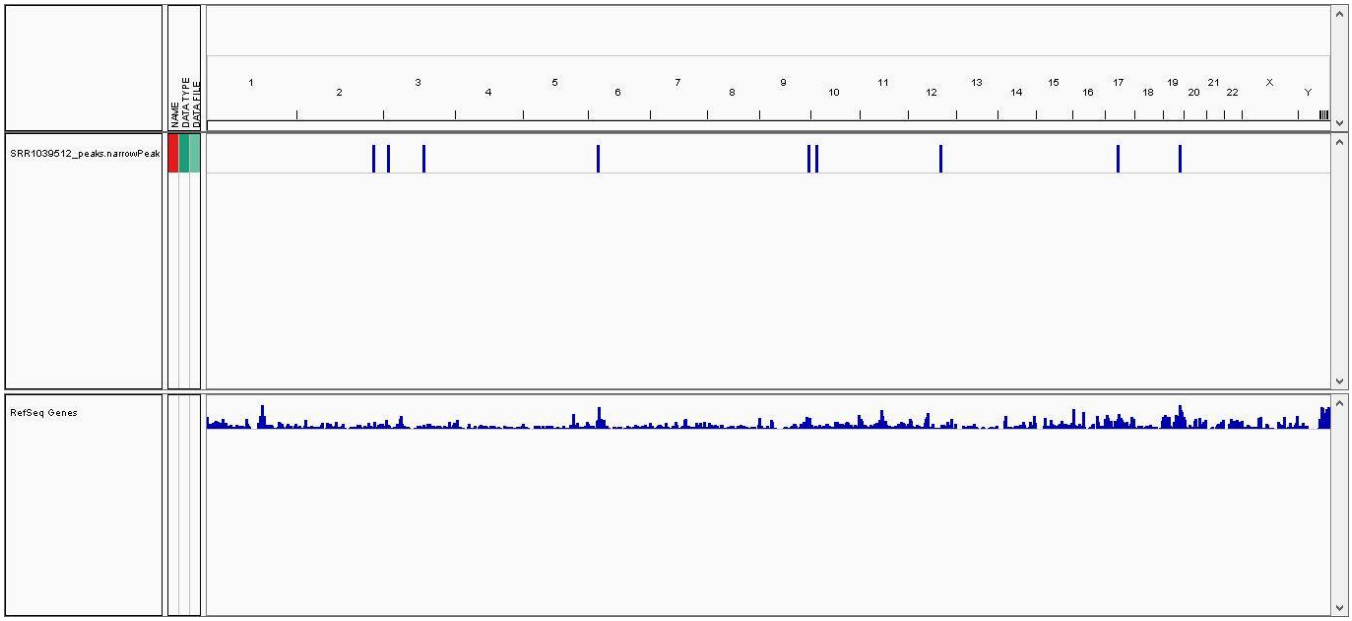
In [17]:
```
#look at the files with the narrow peaks and count them = how many TF binding sites (loo
head SRR1039512_peaks.narrowPeak
wc -lc SRR1039512_peaks.narrowPeak
```

```
chr10   17271402        17271623        SRR1039512_peak_1       172     .       13.192
61      24.34644        17.27386        59
chr12   91539495        91539964        SRR1039512_peak_2       100     .       9.7775
6       16.72079        10.09417        341
chr12   91552105        91552298        SRR1039512_peak_3       165     .       13.058
62      23.55883        16.51532        55
chr17   38600142        38600314        SRR1039512_peak_4       95      .       9.6711
8       16.19693        9.59222 51
chr19   49469877        49470095        SRR1039512_peak_5       221     .       14.810
43      29.53810        22.19667        81
chr2    216236825       216236997       SRR1039512_peak_6       75      .       8.6261
0       13.87448        7.52775 63
chr3    13612480        13612656        SRR1039512_peak_7       125     .       11.128
78      19.28225        12.51070        115
chr3    112357675       112357898       SRR1039512_peak_8       125     .       10.607
41      19.34377        12.56887        111
chr6    27620392        27620584        SRR1039512_peak_9       227     .       16.686
01      32.27706        22.78951        89
chr9    135895081       135895253       SRR1039512_peak_10      35      .       6.4499
5       9.36982 3.59710 16
 10 774 SRR1039512_peaks.narrowPeak
```

In [22]:
```
#do the same for the untreated sample to see the differences.
head SRR1039513_peaks.narrowPeak
wc -l SRR1039513_peaks.narrowPeak
```

```
0 SRR1039513_peaks.narrowPeak
```

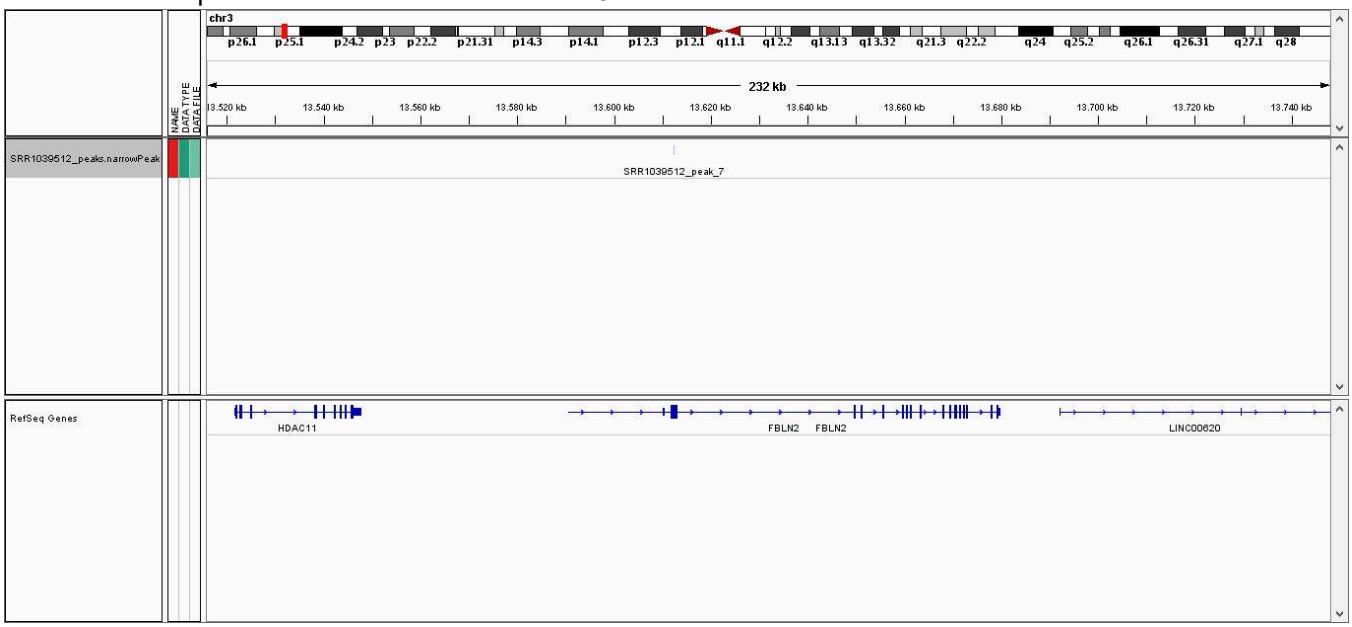
#IGV visualization of the peaks



there are 10 peaks (on chromosomes 2, 2x3, 6, 9, 10, 2x12, 17 and 19)

these possibly new TF binding sites are located in the following genes and are involved in the following biological process (GO) VIM (chr10) = vimentin = protein/glycogen binding DCN (ch12) = decorin = protein binding and placenta/kidney development IGFBP4 (chr17) = insulin like growth factor binding protein 4 = proliferation, cellular growth and inflammation FTL (chr19) = ferritin light chain = iron ion binding FN1 (chr2) = fibronectin = integrin binding and angiogenesis FBLN2 (chr3) = fubulin 2 = extracellular matrix constituent CCDC80 (chr3) = membrane/matrix organisation

visualization of peak 7 located in chromosome 3



We first did differential gene expression analysis with EdgeR in Rstudio to look at the differences in gene expression between treated and untreated samples. The variability within groups is very small whereas the variability between in groups is high suggesting a differences in gene expression after treatment. We have chosen a sample from the treated group and a sample from the untreated group of the same celline to ChipSeq analysis. The untreated sample had 10 peaks suggesting 10 TF binding sites which aren't present in the treated sample (downregulation in treated sample). Most of these peaks are located in genes which are involved in protein binding and matrix constitution (according to gene ontology).

In [ ]: