RNA-seq and ChIP-Seq analysis of RUNX1 knockdown in MCF-7 cells.

Group members

Florian Delbart

Jens Rummens

Alexander Sneyers

Barutcu AR, Hong D, Lajoie BR, McCord RP, van Wijnen AJ, Lian JB, et al. RUNX1 contributes to higher-order chromatin organization and gene regulation in breast cancer cells. Biochim Biophys Acta - Gene Regul Mech [Internet]. 2016 Nov [cited 2017 Jan 24];1859(11):1389–97. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27514584

The RUNX family of transcription factors consists of three members, RUNX1, RUNX2 and RUNX3. These transcription factors regulate diverse physiological functions. RUNX1, for instance, is involved in haematopoiesis. Next to its physiological role clinical data shows that RUNX1 is involved in cancer. Multiple mutations and translocations in the RUNX1 gene have been found in leukemia. However a large amount of evidence suggests that RUNX1 contributes to breast cancer. RUNX1 belongs to the top 1% of overexpressed genes in breast cancer, however it has been found to be downregulated as well. Nevertheless, the exact role of RUNX1 in breast cancer remains to be elucidated. In a recent study RUNX1 was downregulated by shRNA against RUNX1 mRNA in MCF-7 breast cancer cells. Next, RNA-seq and ChIP-Seq were performed to identify differentially expressed genes and binding sites of RUNX1, respectively.

The RNA-seq and ChIP-Seq data from this study was analyzed in order to find differentially expressed genes. To locate the RUNX1 binding sites ChIP-Seq data was examined.

RNA-seq

Create directory for Assignment using the following command:

mkdir Assignment

RNA seq analyses were performed in this directory.

In [146]:
```
pwd
```

```
/mnt/storage/r0381150/Assignment
```

Obtain data from geo: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE75070 Download .sra files from RNA-seq analysis.

In [24]:
```
wget ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX143/SRX1434845/SRR2919468/*.sra
```

```
--2016-12-31 17:09:00--  ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX143/SRX1434845/SRR2919468/*.sra
           => '.listing'
Resolving ftp-trace.ncbi.nlm.nih.gov (ftp-trace.ncbi.nlm.nih.gov)... 130.14.250.11, 2607:f220:41e:250::11
Connecting to ftp-trace.ncbi.nlm.nih.gov (ftp-trace.ncbi.nlm.nih.gov)|130.14.250.11|:21... connected.
Logging in as anonymous ... Logged in!
==> SYST ... done.    ==> PWD ... done.
==> TYPE I ... done.  ==> CWD (1) /sra/sra-instant/reads/ByExp/sra/SRX/SRX143/SRX1434845/SRR2919468 ... done.
==> PASV ... done.    ==> LIST ... done.

    [ <=>                                   ] 74          --.-K/s   in 0,002s

2016-12-31 17:09:01 (32,4 KB/s) - '.listing' saved [74]

Removed '.listing'.
--2016-12-31 17:09:01--  ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX143/SRX1434845/SRR2919468/SRR2919468.sra
           => 'SRR2919468.sra'
==> CWD not required.
==> PASV ... done.    ==> RETR SRR2919468.sra ... done.
Length: 1054063570 (1005M)

100%[=====================================>] 1.054.063.570 34,7MB/s   in 32s

2016-12-31 17:09:34 (31,0 MB/s) - 'SRR2919468.sra' saved [1054063570]
```

In [25]:
```
wget ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX143/SRX1434847/SRR2919470/*.sra

wget ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX143/SRX1434848/SRR2919471/*.sra
```

```
--2016-12-31 17:13:16--  ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX143/SRX1434847/SRR2919470/*.sra
           => '.listing'
Resolving ftp-trace.ncbi.nlm.nih.gov (ftp-trace.ncbi.nlm.nih.gov)... 130.14.250.11, 2607:f220:41e:250::12
Connecting to ftp-trace.ncbi.nlm.nih.gov (ftp-trace.ncbi.nlm.nih.gov)|130.14.250.11|:21... connected.
Logging in as anonymous ... Logged in!
==> SYST ... done.    ==> PWD ... done.
==> TYPE I ... done.  ==> CWD (1) /sra/sra-instant/reads/ByExp/sra/SRX/SRX143/SRX1434847/SRR2919470 ... done.
==> PASV ... done.    ==> LIST ... done.

    [ <=>                                   ] 74          --.-K/s   in 0,003s

2016-12-31 17:13:17 (22,0 KB/s) - '.listing' saved [74]

Removed '.listing'.
--2016-12-31 17:13:17--  ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instan
```

```
t/reads/ByExp/sra/SRX/SRX143/SRX1434847/SRR2919470/SRR2919470.sra
           => ‘SRR2919470.sra’
==> CWD not required.
==> PASV ... done.    ==> RETR SRR2919470.sra ... done.
Length: 1027219216 (980M)

100%[====================================>] 1.027.219.216 35,7MB/s   in
31s

2016-12-31 17:13:48 (32,0 MB/s) - ‘SRR2919470.sra’ saved [1027219216]

--2016-12-31 17:13:48--  ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instan
t/reads/ByExp/sra/SRX/SRX143/SRX1434848/SRR2919471/*.sra
           => ‘.listing’
Resolving ftp-trace.ncbi.nlm.nih.gov (ftp-trace.ncbi.nlm.nih.gov)... 130
.14.250.7, 2607:f220:41e:250::12
Connecting to ftp-trace.ncbi.nlm.nih.gov (ftp-trace.ncbi.nlm.nih.gov)|13
0.14.250.7|:21... connected.
Logging in as anonymous ... Logged in!
==> SYST ... done.    ==> PWD ... done.
==> TYPE I ... done.  ==> CWD (1) /sra/sra-instant/reads/ByExp/sra/SRX/S
RX143/SRX1434848/SRR2919471 ... done.
==> PASV ... done.    ==> LIST ... done.

    [ <=>                                   ] 74          --.-K/s   in 0
,004s

2016-12-31 17:13:49 (17,3 KB/s) - ‘.listing’ saved [74]

Removed ‘.listing’.
--2016-12-31 17:13:49--  ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instan
t/reads/ByExp/sra/SRX/SRX143/SRX1434848/SRR2919471/SRR2919471.sra
           => ‘SRR2919471.sra’
==> CWD not required.
==> PASV ... done.    ==> RETR SRR2919471.sra ... done.
Length: 1250580245 (1,2G)

100%[====================================>] 1.250.580.245 22,2MB/s   in
38s

2016-12-31 17:14:27 (31,6 MB/s) - ‘SRR2919471.sra’ saved [1250580245]
```

In [26]:

```
wget ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SR
X/SRX143/SRX1434849/SRR2919472/*.sra

wget ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SR
X/SRX143/SRX1434850/SRR2919473/*.sra

#.sra files were deleted from the server after creating the fastq files.
```

```
--2016-12-31 17:14:57--  ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instan
t/reads/ByExp/sra/SRX/SRX143/SRX1434849/SRR2919472/*.sra
           => ‘.listing’
Resolving ftp-trace.ncbi.nlm.nih.gov (ftp-trace.ncbi.nlm.nih.gov)... 130
.14.250.10, 2607:f220:41e:250::13
Connecting to ftp-trace.ncbi.nlm.nih.gov (ftp-trace.ncbi.nlm.nih.gov)|13
0.14.250.10|:21... connected.
Logging in as anonymous ... Logged in!
==> SYST ... done.    ==> PWD ... done.
==> TYPE I ... done.  ==> CWD (1) /sra/sra-instant/reads/ByExp/sra/SRX/S
```

```
RX143/SRX1434849/SRR2919472 ... done.
==> PASV ... done.    ==> LIST ... done.

    [ <=>                                   ] 74          --.-K/s   in 0
,003s

2016-12-31 17:14:58 (21,5 KB/s) - '.listing' saved [74]

Removed '.listing'.
--2016-12-31 17:14:58--  ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instan
t/reads/ByExp/sra/SRX/SRX143/SRX1434849/SRR2919472/SRR2919472.sra
           => 'SRR2919472.sra'
==> CWD not required.
==> PASV ... done.    ==> RETR SRR2919472.sra ... done.
Length: 1805949170 (1,7G)

100%[====================================>] 1.805.949.170 35,6MB/s   in
65s

2016-12-31 17:16:03 (26,7 MB/s) - 'SRR2919472.sra' saved [1805949170]

--2016-12-31 17:16:04--  ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instan
t/reads/ByExp/sra/SRX/SRX143/SRX1434850/SRR2919473/*.sra
           => '.listing'
Resolving ftp-trace.ncbi.nlm.nih.gov (ftp-trace.ncbi.nlm.nih.gov)... 130
.14.250.12, 2607:f220:41e:250::11
Connecting to ftp-trace.ncbi.nlm.nih.gov (ftp-trace.ncbi.nlm.nih.gov)|13
0.14.250.12|:21... connected.
Logging in as anonymous ... Logged in!
==> SYST ... done.    ==> PWD ... done.
==> TYPE I ... done.  ==> CWD (1) /sra/sra-instant/reads/ByExp/sra/SRX/S
RX143/SRX1434850/SRR2919473 ... done.
==> PASV ... done.    ==> LIST ... done.

    [ <=>                                   ] 74          --.-K/s   in 0
,003s

2016-12-31 17:16:05 (26,1 KB/s) - '.listing' saved [74]

Removed '.listing'.
--2016-12-31 17:16:05--  ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instan
t/reads/ByExp/sra/SRX/SRX143/SRX1434850/SRR2919473/SRR2919473.sra
           => 'SRR2919473.sra'
==> CWD not required.
==> PASV ... done.    ==> RETR SRR2919473.sra ... done.
Length: 1618175861 (1,5G)

100%[====================================>] 1.618.175.861 19,6MB/s   in
1m 42s

2016-12-31 17:17:47 (15,1 MB/s) - 'SRR2919473.sra' saved [1618175861]
```

In [27]:
```
ls -l
```

```
total 10915752
-rw-r--r-- 1 r0381150 domain users 3208920040 dec 31 17:08 SRR2919466.sr
a
-rw-r--r-- 1 r0381150 domain users 1054063570 nov 17  2015 SRR2919468.sr
a
-rw-r--r-- 1 r0381150 domain users 1212784571 nov 17  2015 SRR2919469.sr
```

```
a
-rw-r--r-- 1 r0381150 domain users 1027219216 nov 17  2015 SRR2919470.sr
a
-rw-r--r-- 1 r0381150 domain users 1250580245 nov 17  2015 SRR2919471.sr
a
-rw-r--r-- 1 r0381150 domain users 1805949170 nov 17  2015 SRR2919472.sr
a
-rw-r--r-- 1 r0381150 domain users 1618175861 nov 17  2015 SRR2919473.sr
a
```

In [28]:
```
pwd
```

```
/mnt/storage/r0381150/Assignment
```

make directory for Chip-seq data with:

In [29]:
```
mkdir -p ./Chip-seq
```

In [30]:
```
ls
```

```
Chip-seq        SRR2919468.sra  SRR2919470.sra  SRR2919472.sra
SRR2919466.sra  SRR2919469.sra  SRR2919471.sra  SRR2919473.sra
```

In [ ]:
```
fastq-dump --split-3 *.sra
#convert .sra files to .fastq files for analysis. Performed in terminal.
```

In [37]:
```
#Rename RNA-seq data to corresponding experimental conditions.
cd /mnt/storage/r0381150/Assignment
mv SRR2919468.fastq shNS1.fastq
mv SRR2919469.fastq shNS2.fastq
mv SRR2919470.fastq shNS3.fastq
mv SRR2919471.fastq shRUNX1.1.fastq
mv SRR2919472.fastq shRUNX1.2.fastq
mv SRR2919473.fastq shRUNX1.3.fastq

#.fastq files were deleted once data analysis was complete to make room
on the server.
```

In [40]:
```
ls
```

```
Chip-seq                shRUNX1.2.fastq         SRR2919471_fastqc.html
shNS1.fastq             shRUNX1.2_fastqc.html   SRR2919471_fastqc.zip
shNS1_fastqc.html       shRUNX1.2_fastqc.zip    SRR2919471.sra
shNS1_fastqc.zip        shRUNX1.3.fastq         SRR2919472_fastqc.html
shNS2.fastq             SRR2919468_fastqc.html  SRR2919472_fastqc.zip
shNS3.fastq             SRR2919468_fastqc.zip   SRR2919472.sra
shNS3_fastqc.html       SRR2919468.sra          SRR2919473_fastqc.html
shNS3_fastqc.zip        SRR2919469.sra          SRR2919473_fastqc.zip
shRUNX1.1.fastq         SRR2919470_fastqc.html  SRR2919473.sra
shRUNX1.1_fastqc.html   SRR2919470_fastqc.zip
shRUNX1.1_fastqc.zip    SRR2919470.sra
```

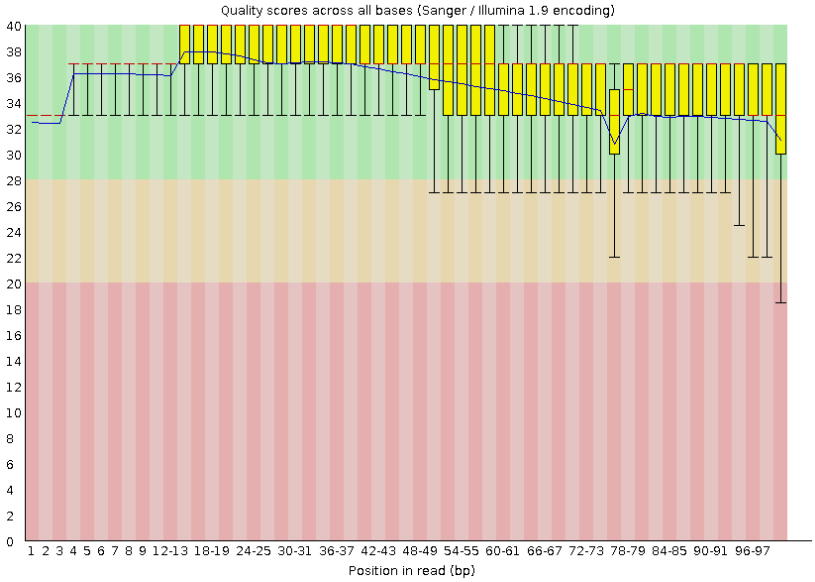
In [ ]:
```
#perform quality control on .fastq files with fastqc.
#Command ran in background on terminal using detach screen command.
/usr/local/bin/FastQC/fastqc *.fastq -o .
```

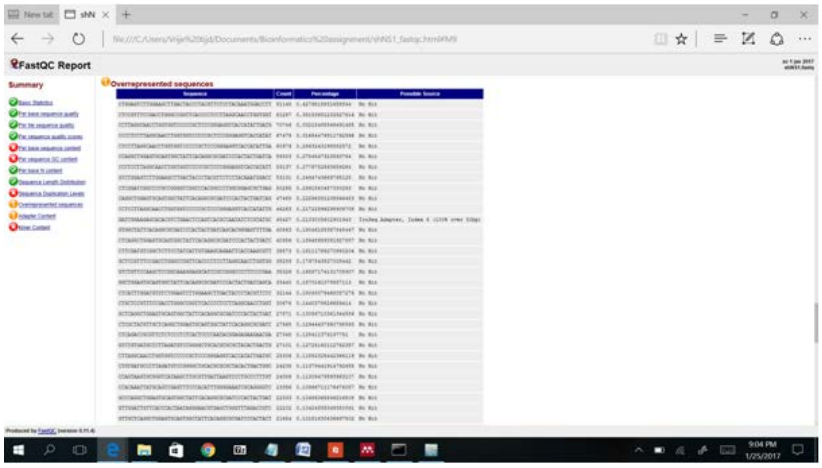
In [147]:
```
ls *.html
```

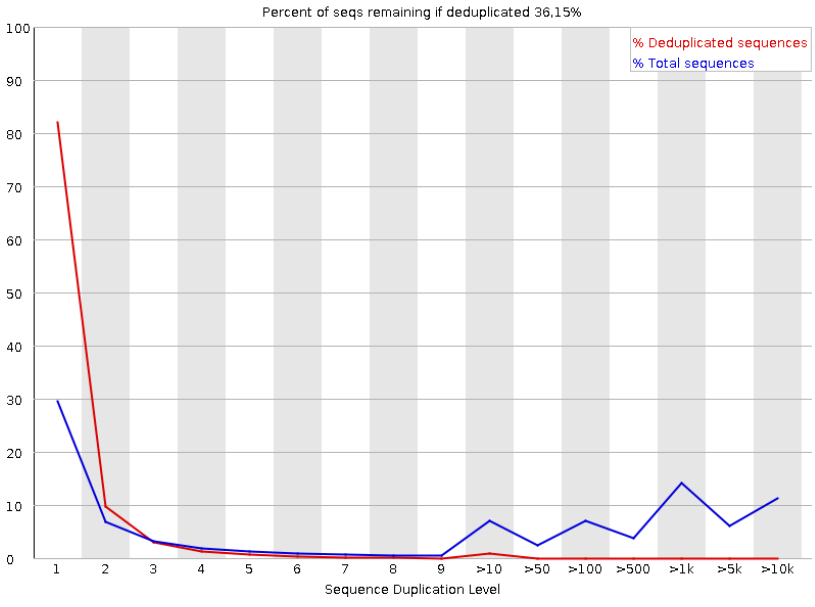
```
shNS1_fastqc.html       shRUNX1.2_fastqc.html   SRR2919471_fastqc.html
shNS2_fastqc.html       shRUNX1.3_fastqc.html   SRR2919472_fastqc.html
shNS3_fastqc.html       SRR2919468_fastqc.html  SRR2919473_fastqc.html
shRUNX1.1_fastqc.html   SRR2919470_fastqc.html
```

Quality control of .html files

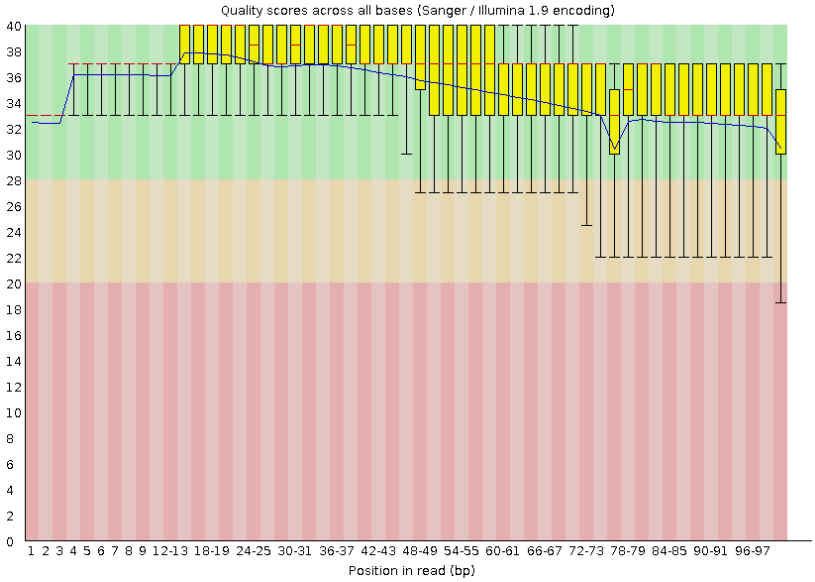
shNS1

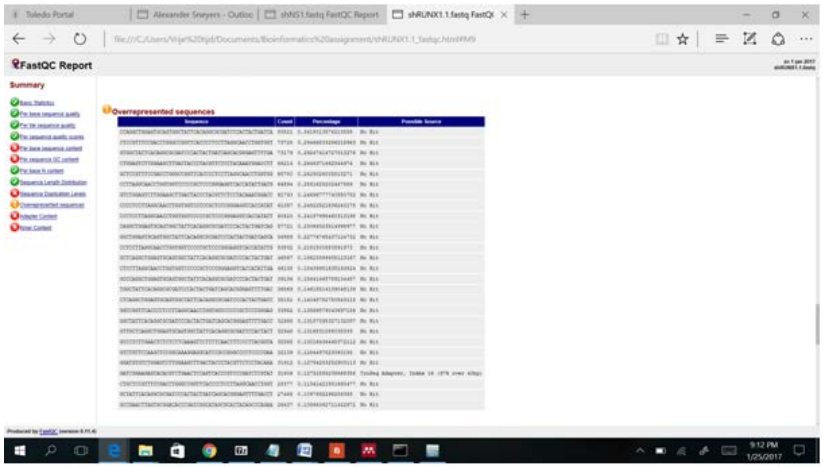


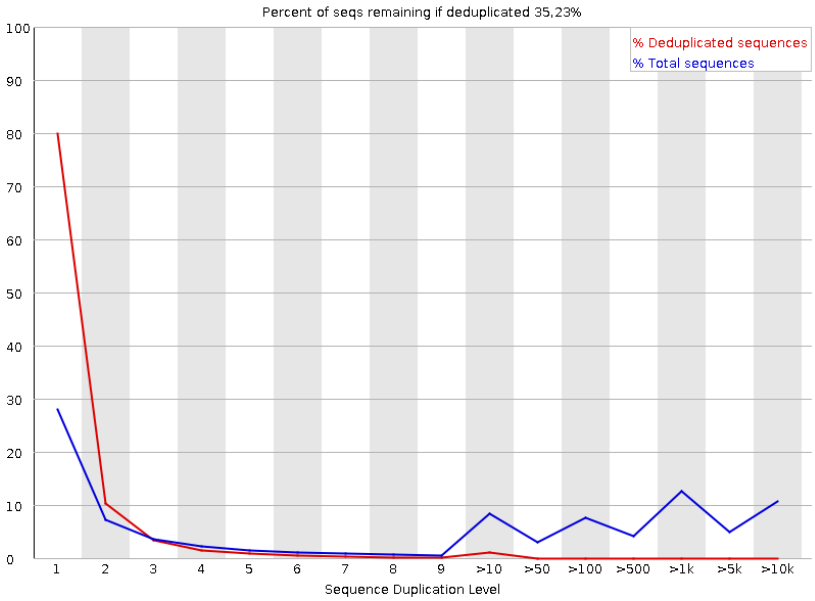




shRUNX1







A lot of sequences are duplicated at high levels. About 35% of the reads is unique. We suspect that the high duplication level might be caused by oversequencing or contamination. Oversequencing is sometimes performed in RNA-seq libraries to detect low-copy number transcripts. We consider adapter contamination to be unlikely as only one of the overrepresented

sequences corresponds to a known adapter sequence.

In [ ]:
```
#Map to genome using STAR, ran in terminal with detached screen since it
 takes long.
STAR --genomeDir /mnt/nfs/mfiers/STAR/hg19_star_db \
     --genomeLoad LoadAndKeep \
     --runThreadN 2 \
     --readFilesIn shNS1.fastq \
     --outFileNamePrefix shNS1.
```

In [ ]:
```
STAR --genomeDir /mnt/nfs/mfiers/STAR/hg19_star_db \
     --genomeLoad LoadAndKeep \
     --runThreadN 2 \
     --readFilesIn shNS2.fastq \
     --outFileNamePrefix shNS2.
```

In [ ]:
```
STAR --genomeDir /mnt/nfs/mfiers/STAR/hg19_star_db \
     --genomeLoad LoadAndKeep \
     --runThreadN 2 \
     --readFilesIn shNS3.fastq \
     --outFileNamePrefix shNS3.
```

In [ ]:
```
STAR --genomeDir /mnt/nfs/mfiers/STAR/hg19_star_db \
     --genomeLoad LoadAndKeep \
     --runThreadN 2 \
     --readFilesIn shRUNX1.1.fastq \
     --outFileNamePrefix shRUNX1.1.
```

In [ ]:
```
STAR --genomeDir /mnt/nfs/mfiers/STAR/hg19_star_db \
     --genomeLoad LoadAndKeep \
     --runThreadN 2 \
     --readFilesIn shRUNX1.2.fastq \
     --outFileNamePrefix shRUNX1.2.
```

In [ ]:
```
STAR --genomeDir /mnt/nfs/mfiers/STAR/hg19_star_db \
     --genomeLoad LoadAndKeep \
     --runThreadN 2 \
     --readFilesIn shRUNX1.3.fastq \
     --outFileNamePrefix shRUNX1.3.
```

In [44]:
```
# Control if .sam files look correct.
head -40 shNS1.Aligned.out.sam | grep '^@'
```

```
@HD     VN:1.4
@SQ     SN:chrM LN:16571
@SQ     SN:chr1 LN:249250621
@SQ     SN:chr2 LN:243199373
@SQ     SN:chr3 LN:198022430
@SQ     SN:chr4 LN:191154276
@SQ     SN:chr5 LN:180915260
@SQ     SN:chr6 LN:171115067
@SQ     SN:chr7 LN:159138663
@SQ     SN:chr8 LN:146364022
@SQ     SN:chr9 LN:141213431
@SQ     SN:chr10        LN:135534747
@SQ     SN:chr11        LN:135006516
@SQ     SN:chr12        LN:133851895
@SQ     SN:chr13        LN:115169878
```

```
@SQ     SN:chr14        LN:107349540
@SQ     SN:chr15        LN:102531392
@SQ     SN:chr16        LN:90354753
@SQ     SN:chr17        LN:81195210
@SQ     SN:chr18        LN:78077248
@SQ     SN:chr19        LN:59128983
@SQ     SN:chr20        LN:63025520
@SQ     SN:chr21        LN:48129895
@SQ     SN:chr22        LN:51304566
@SQ     SN:chrX LN:155270560
@SQ     SN:chrY LN:59373566
@PG     ID:STAR PN:STAR VN:STAR_2.5.2b  CL:STAR   --runThreadN 2   --gen
omeDir /mnt/nfs/mfiers/STAR/hg19_star_db   --genomeLoad LoadAndKeep   --
readFilesIn shNS1.fastq      --outFileNamePrefix shNS1.
@CO     user command line: STAR --genomeDir /mnt/nfs/mfiers/STAR/hg19_st
ar_db --genomeLoad LoadAndKeep --runThreadN 2 --readFilesIn shNS1.fastq
--outFileNamePrefix shNS1.
```

In [45]:
```
head -30 shNS1.Aligned.out.sam | grep -v '^@'
```

```
SRR2919468.sra.1.96656  16      chr14   92628560        255     101M
*       0       0       CATACGTGCTATGACTCTCTCCAGTTTGAATATGCAATTGTTTTCACA
GGCAGGATGTCTGTTTTCTGCCTGTATTTCCCAGTGATTTACTCTAGGGTAAG   BBBBFFFFFFFBBBBB
BBFFFFFFFBFFFFIFFFFFIFFIFIIFFFIFFBFFBFIIIFFIIIIIIIIIIIIIIIFIIIIIIIIIFFF
BFFFFFFFFFBBB   NH:i:1  HI:i:1  AS:i:99 nM:i:0
SRR2919468.sra.1.96657  16      chr1    91852983        255     101M
*       0       0       CGCTTTCACGGTCTGTATTCGTACTGAAAATCAAGATCAAGCGAGCTT
TTGCCCTTCTGCTCCACGGGAGGTTTCTGTCCTCCCTGAGCTCGCCTTAGGAC   BBFFFFFBFFBFFFBF
FFFFBFBFFBFFBFFFFFFFFFBFFFFBFBFFFFFFFFFIIFIIIIFIIFIIIIIIIIIIIFFIIIIIIIII
FFFFFFFFFBBB   NH:i:1  HI:i:1  AS:i:99 nM:i:0
```

In [46]:
```
head -40 shNS2.Aligned.out.sam | grep '^@'
```

```
@HD     VN:1.4
@SQ     SN:chrM LN:16571
@SQ     SN:chr1 LN:249250621
@SQ     SN:chr2 LN:243199373
@SQ     SN:chr3 LN:198022430
@SQ     SN:chr4 LN:191154276
@SQ     SN:chr5 LN:180915260
@SQ     SN:chr6 LN:171115067
@SQ     SN:chr7 LN:159138663
@SQ     SN:chr8 LN:146364022
@SQ     SN:chr9 LN:141213431
@SQ     SN:chr10        LN:135534747
@SQ     SN:chr11        LN:135006516
@SQ     SN:chr12        LN:133851895
@SQ     SN:chr13        LN:115169878
@SQ     SN:chr14        LN:107349540
@SQ     SN:chr15        LN:102531392
@SQ     SN:chr16        LN:90354753
@SQ     SN:chr17        LN:81195210
@SQ     SN:chr18        LN:78077248
@SQ     SN:chr19        LN:59128983
@SQ     SN:chr20        LN:63025520
@SQ     SN:chr21        LN:48129895
@SQ     SN:chr22        LN:51304566
@SQ     SN:chrX LN:155270560
@SQ     SN:chrY LN:59373566
@PG     ID:STAR PN:STAR VN:STAR_2.5.2b  CL:STAR   --runThreadN 2   --gen
omeDir /mnt/nfs/mfiers/STAR/hg19_star_db   --genomeLoad LoadAndKeep   --
```

```
readFilesIn shNS2.fastq      --outFileNamePrefix shNS2.
@CO     user command line: STAR --genomeDir /mnt/nfs/mfiers/STAR/hg19_st
ar_db --genomeLoad LoadAndKeep --runThreadN 2 --readFilesIn shNS2.fastq
--outFileNamePrefix shNS2.
```

In [47]:
```
head -30 shNS2.Aligned.out.sam | grep -v '^@'
```

```
SRR2919469.sra.1.1      0       chr4    125590651       255     1S99M1S
*       0       0       NTGGGCTTTACTTGACTCCACTCAAATTCACTACTTGGTGAATTATGA
CTCTGTCCTAAGGACTGTGTTGTGGAAGATGGGGAAACAATTGACTGTCGACA   #0<FFFFFFFFFIII
IIIIIIIFIFIIIIIIIIIIIFFFFIIIIIIIIIIIFIIFFFIFIIIFIIIIFFIBF<BFFFFFFFBBFF
FFFFFFFFFF<BB   NH:i:1  HI:i:1  AS:i:97 nM:i:0
SRR2919469.sra.1.2      0       chr14   50329414        3       1S100M
*       0       0       NCCCGCTCCCGGGAGGTCACCATATTGATGCCGAACTTAGTGCGGACA
CCCGATCGGCATAGCGCACTACAGCCCAGAACTCCTGGGCTCAAGCGATCCTC   #0<FBFFFFFFFFIII
BFFFIFIIIIIIFIFFIIIIIIIFFIIIIFFFFFFFFFFBBFFFBFBFBFFFBFBFFBFBBBBFFFFBBFFF
FFFFFFFFFFFB   NH:i:2  HI:i:1  AS:i:98 nM:i:0
```

In [48]:
```
head -30 shNS3.Aligned.out.sam | grep '^@'
```

```
@HD     VN:1.4
@SQ     SN:chrM LN:16571
@SQ     SN:chr1 LN:249250621
@SQ     SN:chr2 LN:243199373
@SQ     SN:chr3 LN:198022430
@SQ     SN:chr4 LN:191154276
@SQ     SN:chr5 LN:180915260
@SQ     SN:chr6 LN:171115067
@SQ     SN:chr7 LN:159138663
@SQ     SN:chr8 LN:146364022
@SQ     SN:chr9 LN:141213431
@SQ     SN:chr10        LN:135534747
@SQ     SN:chr11        LN:135006516
@SQ     SN:chr12        LN:133851895
@SQ     SN:chr13        LN:115169878
@SQ     SN:chr14        LN:107349540
@SQ     SN:chr15        LN:102531392
@SQ     SN:chr16        LN:90354753
@SQ     SN:chr17        LN:81195210
@SQ     SN:chr18        LN:78077248
@SQ     SN:chr19        LN:59128983
@SQ     SN:chr20        LN:63025520
@SQ     SN:chr21        LN:48129895
@SQ     SN:chr22        LN:51304566
@SQ     SN:chrX LN:155270560
@SQ     SN:chrY LN:59373566
@PG     ID:STAR PN:STAR VN:STAR_2.5.2b  CL:STAR   --runThreadN 2   --gen
omeDir /mnt/nfs/mfiers/STAR/hg19_star_db   --genomeLoad LoadAndKeep   --
readFilesIn shNS3.fastq      --outFileNamePrefix shNS3.
@CO     user command line: STAR --genomeDir /mnt/nfs/mfiers/STAR/hg19_st
ar_db --genomeLoad LoadAndKeep --runThreadN 2 --readFilesIn shNS3.fastq
--outFileNamePrefix shNS3.
```

In [49]:
```
head -30 shNS3.Aligned.out.sam | grep -v '^@'
```

```
SRR2919470.sra.1.96663  16      chr14   50053471        3       101M
*       0       0       GGAGGGGTGAACCGGCCCAGGTCGGAAACGGAGCAGGTCAAAACTCCC
GTGCTGATCAGTAGTGGGATCGCGCCTGTGAATAGCCACTGCACTCCAGCCTG   FFFFFFFBB<BFFFBB
BFFBFFFBBBFFFFFFFBFBFFFFFFBF<FFIIIIFIIIIIIIIIIIIIIIIIIIIIIIIIIIIIFFFF
FFFFFFFFFFBBB   NH:i:2  HI:i:1  AS:i:99 nM:i:0
SRR2919470.sra.1.96663  256     chr14   50329294        3       101M
```

```
*       0       0       CAGGCTGGAGTGCAGTGGCTATTCACAGGCGCGATCCCACTACTGATC
AGCACGGGAGTTTTGACCTGCTCCGTTTCCGACCTGGGCCGGTTCACCCCTCC   BBBFFFFFFFFFFFFF
FIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIFF<FBFFFFFFBFBFFFFFFFBBBFFFBFFBBBF
FFB<BBFFFFFFF   NH:i:2  HI:i:2  AS:i:99 nM:i:0
```

In [50]:
```
head -30 shRUNX1.1.Aligned.out.sam | grep '^@'
```

```
@HD     VN:1.4
@SQ     SN:chrM LN:16571
@SQ     SN:chr1 LN:249250621
@SQ     SN:chr2 LN:243199373
@SQ     SN:chr3 LN:198022430
@SQ     SN:chr4 LN:191154276
@SQ     SN:chr5 LN:180915260
@SQ     SN:chr6 LN:171115067
@SQ     SN:chr7 LN:159138663
@SQ     SN:chr8 LN:146364022
@SQ     SN:chr9 LN:141213431
@SQ     SN:chr10        LN:135534747
@SQ     SN:chr11        LN:135006516
@SQ     SN:chr12        LN:133851895
@SQ     SN:chr13        LN:115169878
@SQ     SN:chr14        LN:107349540
@SQ     SN:chr15        LN:102531392
@SQ     SN:chr16        LN:90354753
@SQ     SN:chr17        LN:81195210
@SQ     SN:chr18        LN:78077248
@SQ     SN:chr19        LN:59128983
@SQ     SN:chr20        LN:63025520
@SQ     SN:chr21        LN:48129895
@SQ     SN:chr22        LN:51304566
@SQ     SN:chrX LN:155270560
@SQ     SN:chrY LN:59373566
@PG     ID:STAR PN:STAR VN:STAR_2.5.2b  CL:STAR   --runThreadN 2   --gen
omeDir /mnt/nfs/mfiers/STAR/hg19_star_db   --genomeLoad LoadAndKeep   --
readFilesIn shRUNX1.1.fastq      --outFileNamePrefix shRUNX1.1.
@CO     user command line: STAR --genomeDir /mnt/nfs/mfiers/STAR/hg19_st
ar_db --genomeLoad LoadAndKeep --runThreadN 2 --readFilesIn shRUNX1.1.fa
stq --outFileNamePrefix shRUNX1.1.
```

In [51]:
```
head -30 shRUNX1.1.Aligned.out.sam | grep -v '^@'
```

```
SRR2919471.100440       16      chr17   9145706 255     101M    *
0       0       CTGCTGGGGCGGTTGTCCTATTTCTGTCTATGCCAGTGTGGTTTCTTCACCCTGCC
CAGAAGGGCTGTGGCAGCCCCACGATATCCCACCCTGGGTCTGGG   <FFFFFFFFFFFFFFFBBFFFBFB
B<FBBFFFBFFFFFFFFFFFFFFBBFFFFIIIIIIIIIIIIIIIIIIIIFFFIIIIIIIFFBFFFFFFFF
FFBBB   NH:i:1  HI:i:1  AS:i:99 nM:i:0
SRR2919471.100441       16      chr1    91852783        255     4S97M
*       0       0       AAACCCAGCTCACGTTCCCTATTAGTGGGTGAACAATCCAACGCTTGG
TGAATTCTGCTTCACAATGATAGGAAGAGCCGACATCGAAGGATCAAAAAGCG   BBBFBBBBBFFBFFFB
<BBBBFFBBBFFFBFFFFFFBFFFBFFIIIIFIIFIFFIIIFIFIIFFFIIIIIIIFFFIFFFIIIIIIIII
FFFFFFFFFFBBB   NH:i:1  HI:i:1  AS:i:95 nM:i:0
```

In [52]:
```
head -30 shRUNX1.2.Aligned.out.sam | grep '^@'
```

```
@HD     VN:1.4
@SQ     SN:chrM LN:16571
@SQ     SN:chr1 LN:249250621
@SQ     SN:chr2 LN:243199373
@SQ     SN:chr3 LN:198022430
@SQ     SN:chr4 LN:191154276
```

```
@SQ     SN:chr5 LN:180915260
@SQ     SN:chr6 LN:171115067
@SQ     SN:chr7 LN:159138663
@SQ     SN:chr8 LN:146364022
@SQ     SN:chr9 LN:141213431
@SQ     SN:chr10        LN:135534747
@SQ     SN:chr11        LN:135006516
@SQ     SN:chr12        LN:133851895
@SQ     SN:chr13        LN:115169878
@SQ     SN:chr14        LN:107349540
@SQ     SN:chr15        LN:102531392
@SQ     SN:chr16        LN:90354753
@SQ     SN:chr17        LN:81195210
@SQ     SN:chr18        LN:78077248
@SQ     SN:chr19        LN:59128983
@SQ     SN:chr20        LN:63025520
@SQ     SN:chr21        LN:48129895
@SQ     SN:chr22        LN:51304566
@SQ     SN:chrX LN:155270560
@SQ     SN:chrY LN:59373566
@PG     ID:STAR PN:STAR VN:STAR_2.5.2b  CL:STAR   --runThreadN 2   --gen
omeDir /mnt/nfs/mfiers/STAR/hg19_star_db   --genomeLoad LoadAndKeep   --
readFilesIn shRUNX1.2.fastq      --outFileNamePrefix shRUNX1.2.
@CO     user command line: STAR --genomeDir /mnt/nfs/mfiers/STAR/hg19_st
ar_db --genomeLoad LoadAndKeep --runThreadN 2 --readFilesIn shRUNX1.2.fa
stq --outFileNamePrefix shRUNX1.2.
```

In [53]:
```
head -30 shRUNX1.2.Aligned.out.sam | grep -v '^@'
```

```
SRR2919472.100452       256     chrX    88735819        3       88M13S
*       0       0       GTTGTTGCCATGGTAATCCTGCTCAGTACGAGAGGAACCGCAGGTTCA
GACATTTGGTGTATGTGCTTGGCTGAGGAGCCAATGGGGCGAAGCTACCATCT   BBBFFFFFFFFFFFII
IIIIIIIIIIIIIIIIIIIIIIIIIIIFFIIIIIIIIIIIFFFFFFFFFFFFFFFFFFBBFFFFFFFBBBB
BBFFFFFFFFFB   NH:i:2  HI:i:1  AS:i:76 nM:i:5
SRR2919472.100452       0       chrY    3005051 3       91M10S  *
0       0       GTTGTTGCCATGGTAATCCTGCTCAGTACGAGAGGAACCGCAGGTTCAGACATTTG
GTGTATGTGCTTGGCTGAGGAGCCAATGGGGCGAAGCTACCATCT   BBBFFFFFFFFFFFIIIIIIIIII
IIIIIIIIIIIIIIIIIIIFFIIIIIIIIIIIFFFFFFFFFFFFFFFFFFBBFFFFFFFBBBBBBFFFFFF
FFFFB   NH:i:2  HI:i:2  AS:i:77 nM:i:6
```

In [54]:
```
head -30 shRUNX1.3.Aligned.out.sam | grep '^@'
```

```
@HD     VN:1.4
@SQ     SN:chrM LN:16571
@SQ     SN:chr1 LN:249250621
@SQ     SN:chr2 LN:243199373
@SQ     SN:chr3 LN:198022430
@SQ     SN:chr4 LN:191154276
@SQ     SN:chr5 LN:180915260
@SQ     SN:chr6 LN:171115067
@SQ     SN:chr7 LN:159138663
@SQ     SN:chr8 LN:146364022
@SQ     SN:chr9 LN:141213431
@SQ     SN:chr10        LN:135534747
@SQ     SN:chr11        LN:135006516
@SQ     SN:chr12        LN:133851895
@SQ     SN:chr13        LN:115169878
@SQ     SN:chr14        LN:107349540
@SQ     SN:chr15        LN:102531392
@SQ     SN:chr16        LN:90354753
@SQ     SN:chr17        LN:81195210
```

```
@SQ     SN:chr18        LN:78077248
@SQ     SN:chr19        LN:59128983
@SQ     SN:chr20        LN:63025520
@SQ     SN:chr21        LN:48129895
@SQ     SN:chr22        LN:51304566
@SQ     SN:chrX LN:155270560
@SQ     SN:chrY LN:59373566
@PG     ID:STAR PN:STAR VN:STAR_2.5.2b  CL:STAR   --runThreadN 2   --gen
omeDir /mnt/nfs/mfiers/STAR/hg19_star_db   --genomeLoad LoadAndKeep   --
readFilesIn shRUNX1.3.fastq      --outFileNamePrefix shRUNX1.3.
@CO     user command line: STAR --genomeDir /mnt/nfs/mfiers/STAR/hg19_st
ar_db --genomeLoad LoadAndKeep --runThreadN 2 --readFilesIn shRUNX1.3.fa
stq --outFileNamePrefix shRUNX1.3.
```

In [55]:
```
head -30 shRUNX1.3.Aligned.out.sam | grep -v '^@'
```

```
SRR2919473.3    16      chr3    58156333        255     100M1S  *
0       0       GCCCACCACCCCCTGCGAGGAGGTCTCCATGAAGCATGTAGGCAACCAGCAATACA
ACGTCACATACGTCGTCAAGGAGAGGGGCGATTATGTGCTGGCTN   770B7BBBBBB<77<BBB7BBBBB
BB<<<B<BFBBBFBBFFFBBBFFF<7IIIIIIFFFBFFFFFFFFFIIIFFIFBIFFFF<IIIIFBFFFFFFF
BB00#   NH:i:1  HI:i:1  AS:i:98 nM:i:0
SRR2919473.5    16      chr9    9442161 1       101M    *       0
0       ATCGGGTGTCCGCACTAAGTTCGGCATCAATATGGTGACCTCCCGGGAGCGGGGGACCACCAGG
TTGCCTAAGGAGGGGTGAACCGGCCCAGGTCGGAAAC   <BFFBFBFBBFBBFFFFFFBFFFFBBBFFFFB
FFBBB<FBBFFFFBFFFFFFFIFFIIFIIFIIIIIIIIIIIIIIIIIIIFFIIIIIFFFFFFFFFFBBB
NH:i:4  HI:i:1  AS:i:99 nM:i:0
```

In [56]:
```
#Sort and convert .sam file to .bam
samtools sort -o shNS1.bam shNS1.Aligned.out.sam
```

```
[bam_sort_core] merging from 11 files...
```

In [57]:
```
ls -l shNS1.Aligned.out.sam shNS1.bam
```

```
-rw-r--r-- 1 r0381150 domain users 7389124516 jan  1 22:48 shNS1.Aligned
.out.sam
-rw-r--r-- 1 r0381150 domain users 1067056227 jan 13 10:04 shNS1.bam
```

In [ ]:
```
samtools sort -o shNS2.bam shNS2.Aligned.out.sam
```

In [58]:
```
samtools sort -o shNS3.bam shNS3.Aligned.out.sam
```

```
[bam_sort_core] merging from 10 files...
```

In [59]:
```
samtools sort -o shRUNX1.1.bam shRUNX1.1.Aligned.out.sam
```

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

In [ ]:
```
samtools sort -o shRUNX1.2.bam shRUNX1.2.Aligned.out.sam
```

.sam files were deleted after analyses were performed to make space on the server.

In [ ]:
```
# Generate index
samtools index shNS1.bam
```

In [60]:
```
samtools index shNS2.bam
```

In [61]:
```
samtools index shNS3.bam
```

In [62]:
```
samtools index shRUNX1.1.bam
samtools index shRUNX1.2.bam
samtools index shRUNX1.3.bam
```

In [63]:
```
ls -l *.bai
```

```
-rw-r--r-- 1 r0381150 domain users 2797256 jan 13 10:17 shNS1.bam.bai
-rw-r--r-- 1 r0381150 domain users 2959688 jan 13 15:46 shNS2.bam.bai
-rw-r--r-- 1 r0381150 domain users 2698016 jan 13 15:46 shNS3.bam.bai
-rw-r--r-- 1 r0381150 domain users 2792472 jan 13 15:47 shRUNX1.1.bam.ba
i
-rw-r--r-- 1 r0381150 domain users 2793944 jan 13 15:47 shRUNX1.2.bam.ba
i
-rw-r--r-- 1 r0381150 domain users 2790968 jan 13 15:49 shRUNX1.3.bam.ba
i
```

In [65]:
```
samtools view shNS1.bam | head -4
```

```
SRR2919468.sra.1.1257417        0       chrM    1       255     1S95M5S
*       0       0       GGATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATG
CATTTGGTATTTTCGTCTGGGGGGTATGCACGCGATAGCATTGCGAGAGATCG   BBBF<BFBB<BFFBF0
FF0BFFBFBBF<BFIIFI<FF<FFFFFFFFFBBBFFIII'7B70BBF7BFFBFFBB77<<BBBB<BBBBBBB
<BBBB<<BBBBBB   NH:i:1  HI:i:1  AS:i:91 nM:i:1
SRR2919468.sra.1.1307827        0       chrM    1       255     3S98M
*       0       0       ATGGATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCA
TGCATTTGGTATTTTCGTCTGGGGGGTATGCACGCGATAGCATTGCGAGACGC   BBBFFFFFFFFFFFII
IFIIIIIIIFIIIIIIIIIIIIIIIIIIIIIIIIIIIF7BFIIIIIFFFFFFFFF77BBFFFFFFFFBF
FFFFFFBFFFFFF   NH:i:1  HI:i:1  AS:i:94 nM:i:1
SRR2919468.sra.1.2735860        0       chrM    1       255     4S97M
*       0       0       GATGGATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCC
ATGCATTTGGTATTTTCGTCTGGGGGGTATGCACGCGATAGCATTGCGAGACG   BBBFFFFFFFFFFFFF
FIIIIFIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIFFIIIIBFFIIIIIFFFFFFFFFFBFFFFBBBFBBF
FFFFBFFFFF<<B   NH:i:1  HI:i:1  AS:i:93 nM:i:1
SRR2919468.sra.1.4686903        0       chrM    1       255     6S95M
*       0       0       ACGATGGATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCT
CCATGCATTTGGTATTTTCGTCTGGGGGGTATGCACGCGATAGCATTGCGAGA   BBBFFFFFFFFFFIIF
FIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIFIIIIIFFFFFFFFF<B7<BFBFFFFFF
FFFFFFFFBBFF7   NH:i:1  HI:i:1  AS:i:91 nM:i:1
samtools view: writing to standard output failed: Broken pipe
samtools view: error closing standard output: -1
```

In [66]:
```
samtools view shNS2.bam | head -4
```

```
SRR2919469.sra.1.356717 0       chrM    1       255     3S98M   *
0       0       ATGGATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTG
GTATTTTCGTCTGGGGGGTATGCACGCGATAGCATTGCGAGACGC   BBBBF<0B<BFF<BB<BBBFFBFB
F<FFFFIIIIBFBFFIFFBF<F<B7B7FBBB<FFFB<FBB00<77'B<BB0'0<BFB<BBB'07<B0<<<<B
BB<B'   NH:i:1  HI:i:1  AS:i:94 nM:i:1
SRR2919469.sra.1.1461893        0       chrM    1       255     9S92M
*       0       0       ATCACGATGGATCACAGGTCTATCACCCTATTAACCACTCACGGGAGC
TCTCCATGCATTTGGTATTTTCGTCTGGGGGGTATGCACGCGATAGCATTGCG   BBBFFFFFFFFFFIII
FIBFFFFFFFIIFFFIFFIB<FFFFFFIIBFIIIIIBFIFBFFIFIFFFFFFFFFFFFBBFFF<<BBBFBF
FFBBFFFBBFFFF   NH:i:1  HI:i:1  AS:i:88 nM:i:1
SRR2919469.sra.1.4462712        0       chrM    1       255     5S96M
*       0       0       CGATGGATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTC
CATGCATTTGGTATTTTCGTCTGGGGGGTATGCACGCGATAGCATTGCGAGAC   BBBFFFFFFFFFFIFF
IIIIIIIIIIIIIIIIIIIIIIIFIIIIIIIIIIIIIFIIIII<BFFIIIFFFFFFFFFF<BFFFFFFFBBB
FFFFFFBFFFFFB   NH:i:1  HI:i:1  AS:i:92 nM:i:1
```

```
SRR2919469.sra.1.4707862        0       chrM    1       255     9S92M
*       0       0       ATCACGATGGATCACAGGTCTATCACCCTATTAACCACTCACGGGAGC
TCTCCATGCATTTGGTATTTTCGTCTGGGGGGTATGCACGCGATAGCATTGCG   <BBFFFFFFFBB<BFB
FFFFFFBFFFFFBBFFFBFBFFBFIFBFFF<B<BBFBBFFFFBBBFI<<BFFBFF<BBBBBBBB07<BBFBB
BBBBBBBBFBBBB   NH:i:1  HI:i:1  AS:i:88 nM:i:1
samtools view: writing to standard output failed: Broken pipe
samtools view: error closing standard output: -1
```

In [67]:
```
samtools view shNS3.bam | head -4
```

```
SRR2919470.sra.1.67970  0       chrM    1       255     9S92M   *
0       0       ATCACGATGGATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATG
CATTTGGTATTTTCGTCTGGGGGGTATGCACGCGATAGCATTGCG   BBBFFFFFFFFFFIIIIIFFIIII
IIFIIIIIIIIIIIIIIIFFFIIIIIIIIIIFFIIIIIFFFFFFFFFFFBBFFFF<BFFFFFFFFFFFBF
FFFFF   NH:i:1  HI:i:1  AS:i:88 nM:i:1
SRR2919470.sra.1.1406057        0       chrM    1       255     9S92M
*       0       0       GTCACGATGGATCACAGGTCTATCACCCTATTAACCACTCACGGGAGC
TCTCCATGCATTTGGTATTTTCGTCTGGGGGGTATGCACGCGATAGCATTGCG   BBBFFFFFFFFFFIII
IIFIIIIIIIIIIFIIIIIIIIIIIIIFIFIIIIIIIIFIFFIIIIIIFBFFFFFFFFFBBBFFB<BFFFFFF
FFFFFBBFFFFFB   NH:i:1  HI:i:1  AS:i:88 nM:i:1
SRR2919470.sra.1.3950986        0       chrM    1       255     5S95M1S
*       0       0       CGATGGATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTC
CATGCATTTGGTATTTTCGTCTGGGGGGTATGCACGCGATAGCATTGCGAGAT   BBBFFFFFFFFFFIFF
FIIIIIIIIFIIIIIIIIIIIIIIIIIIIBIIIIIIIIIIIFBFFIIIIFFFFFFFBFBBFFFFFFFFFFF
BFFBFFFBF<BBB   NH:i:1  HI:i:1  AS:i:91 nM:i:1
SRR2919470.sra.1.3872019        0       chrM    1       255     10S91M
*       0       0       CATCACGATGGATCACAGGTCTATCACCCTATTAACCACTCACGGGAG
CTCTCCATGCATTTGGTATTTTCGTCTGGGGGGTATGCACGCGATAGCATTGC   BBBFFFFFFFFFFIII
IIIFFFIIIIFFIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIFFFFFFFFFFFB<BFFFF0<BFFFF
FFFBFFFFFFFFB   NH:i:1  HI:i:1  AS:i:87 nM:i:1
samtools view: writing to standard output failed: Broken pipe
samtools view: error closing standard output: -1
```

In [68]:
```
samtools view shRUNX1.1.bam | head -4
```

```
SRR2919471.40625        0       chrM    1       255     9S92M   *
0       0       ATCACGATGGATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATG
CATTTGGTATTTTCGTCTGGGGGGTATGCACGCGATAGCATTGCG   BBBFFFFFFFFFFIIIIIFFIIII
FIIIIFIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIFFFFFFFBFFFBBFFF7BBBBFFFFFBBFFFB
FBBF<   NH:i:1  HI:i:1  AS:i:88 nM:i:1
SRR2919471.1316495      0       chrM    1       255     7S94M   *
0       0       CACGATGGATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCA
TTTGGTATTTTCGTCTGGGGGGTATGCACGCGATAGCATTGCGAG   BBBFFFFFFFFFFIIIIIIIIIII
IIIIIIIIIIIIIIIIFFIIIIIIIIIIIIIIIIIBFFIFFFFFFFFFFBFF7<FBFFFFFFFBFFFFF
FFFFF   NH:i:1  HI:i:1  AS:i:90 nM:i:1
SRR2919471.7269090      0       chrM    1       255     9S92M   *
0       0       ATCACGATGGATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATG
CATTTGGTATTTTCGTCTGGGGGGTATGCACGCGATAGCATTGCG   BBBFFFFFFFFFFFFIIIBFFFIF
FIIIIIIFIIFFFFFIIIIFIIIIFFIIFFIIFFFFFIIBFBFFFFFFFBFBFBFF7<BFBFFFFBFBFBFF
<<BBB   NH:i:1  HI:i:1  AS:i:88 nM:i:1
SRR2919471.8655236      0       chrM    1       255     5S96M   *
0       0       CGATGGATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATT
TGGTATTTTCGTCTGGGGGGTATGCACGCGATAGCATTGCGAGAC   BBBFFFFFFFFFFIFFFIIIIIII
IIIIIIIIIIIIIIIIIIIFIFFFIIIIIIIIIIFFFIIIIFFFFFFFFFF<BBFFFFFBFFFBFBBFFFF
BFBB<   NH:i:1  HI:i:1  AS:i:92 nM:i:1
samtools view: writing to standard output failed: Broken pipe
samtools view: error closing standard output: -1
```

In [69]:
```
samtools view shRUNX1.2.bam | head -4
```

```
SRR2919472.2337006      0       chrM    1       255     1S100M  *
```

```
0       0       GGATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGT
ATTTTCGTCTGGGGGGTATGCACGCGATAGCATTGCGAGACGCTG   BBBFFFFFFFFFFIIIIIIIIIII
IIIIIIIIIIIIIFIIIIIIIIIIIFIIIII7BFIFFIIFFFFFFFFF<BFBFFFFFFFBBFBBFBFFFFFF
FFFFB   NH:i:1  HI:i:1  AS:i:96 nM:i:1
SRR2919472.4117301      0       chrM    1       255     2S99M   *
0       0       TGGATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGG
TATTTTCGTCTGGGGGGTATGCACGCGATAGCATTGCGAGACGCT   BBBFFFFFFFFFFIIIIIIIIII
IIIIIIIIFIIIIIIFIIIIIIIIIIIIIIFI<BFFIIIIIFFFFFFFF<BBBFFFFFBFFFFFFFFBFFFF
FFFF<   NH:i:1  HI:i:1  AS:i:95 nM:i:1
SRR2919472.9773707      0       chrM    1       255     7S94M   *
0       0       CACGATGGATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCA
TTTGGTATTTTCGTCTGGGGGGTATGCACGCGATAGCATTGCGAG   BBBFFFFFFFFFFIIIFFIIIIII
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIFIIIFFFFFFFFF<BF<BBFBFFFFFFFFFFFF
BFFFF   NH:i:1  HI:i:1  AS:i:90 nM:i:1
SRR2919472.11225302     0       chrM    1       255     5S95M1S *
0       0       CGATGGATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATT
TGGTATTTTCGTCTGGGGGGTATGCACGCGATAGCATTGCGAGAT   BBBFFFFFFFFFFIFIIIIIIIII
IIIIIIIIFIIIIIFIIFIIIIIIIFIIIIIIIIIBFFIIIFFFFFFFFFFB<<BFFFFFFFFFBFFFFFBB
BBFFF   NH:i:1  HI:i:1  AS:i:91 nM:i:1
samtools view: writing to standard output failed: Broken pipe
samtools view: error closing standard output: -1
```

In [70]:
```
samtools view shRUNX1.3.bam | head -4
```

```
SRR2919473.5804939      0       chrM    1       255     9S85M7S *
0       0       ATCACGATGGATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATG
CATTTGGTATTTTCGTCTGGGGGGTATGCACGCGATAGAGATCGG   BBBFFFFFFFFFFIIIIIFFIIII
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIFIIIIIIFFFFFFFFFFFFFBFFFF<BFFFFFFFFFFFFF
FFFFF   NH:i:1  HI:i:1  AS:i:81 nM:i:1
SRR2919473.6571287      0       chrM    1       255     5S96M   *
0       0       CGATGGATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATT
TGGTATTTTCGTCTGGGGGGTATGCACGCGATAGCATTGCGAGAC   BBBFFFBFFFFFFIBFFIIFFIFF
IFFFFFIIIIIIIIIIIIIIIIIIIFIFIFFIIII<FFIIIIFFFFFFFFF<07BFBFFFFFFBFFBFFFB
FBFFF   NH:i:1  HI:i:1  AS:i:92 nM:i:1
SRR2919473.7544612      0       chrM    1       255     95M6S   *
0       0       GATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTA
TTTTCGTCTGGGGGGTATGCACGCGATAGCATTGCGAGATCGGAA   BB<BFFFFFBBFFFFFFFFFFFFF
FFIFIIIIFBFIIFFFFBBFBFFFBFFFII<<B<FFFFBFFFFFFBB'0<BBFBBBB<77BBBFBBBBBBBB
BBBBB   NH:i:1  HI:i:1  AS:i:91 nM:i:1
SRR2919473.7271102      0       chrM    1       255     9S82M10S
*       0       0       CTCACGATGGATCACAGGTCTATCACCCTATTAACCACTCACGGGAGC
TCTCCATGCATTTGGTATTTTCGTCTGGGGGGTATGTACGCGAGATCGGAAGA   BBBFFFFFFFFFFIII
IIFIIIIIIIIIIIIIIIIIIIIIIIIIFIIIIIIIIIIIIIIIIBFFFFFFFFFFFFFFFFF7<BFFFFF
FF<BBFFFFFFFB   NH:i:1  HI:i:1  AS:i:76 nM:i:2
samtools view: writing to standard output failed: Broken pipe
samtools view: error closing standard output: -1
```

In [71]:
```
samtools idxstats shNS1.bam
```

```
chrM    16571   672011  0
chr1    249250621       2807106 0
chr2    243199373       1708269 0
chr3    198022430       1236724 0
chr4    191154276       366282  0
chr5    180915260       760904  0
chr6    171115067       1424282 0
chr7    159138663       556079  0
chr8    146364022       644763  0
chr9    141213431       1610494 0
chr10   135534747       458957  0
chr11   135006516       1071912 0
```

```
chr12   133851895       1345691 0
chr13   115169878       208785  0
chr14   107349540       4513376 0
chr15   102531392       404278  0
chr16   90354753        1295121 0
chr17   81195210        1300534 0
chr18   78077248        155624  0
chr19   59128983        577278  0
chr20   63025520        435843  0
chr21   48129895        422616  0
chr22   51304566        308276  0
chrX    155270560       624839  0
chrY    59373566        421548  0
*       0       0       0
```

In [72]:
```
samtools idxstats shNS2.bam
```

```
chrM    16571   296560  0
chr1    249250621       3208755 0
chr2    243199373       1718128 0
chr3    198022430       1700737 0
chr4    191154276       512527  0
chr5    180915260       1010650 0
chr6    171115067       1980074 0
chr7    159138663       840415  0
chr8    146364022       817029  0
chr9    141213431       2090551 0
chr10   135534747       683608  0
chr11   135006516       1426051 0
chr12   133851895       1559560 0
chr13   115169878       268596  0
chr14   107349540       6115118 0
chr15   102531392       606697  0
chr16   90354753        1035220 0
chr17   81195210        1775622 0
chr18   78077248        226841  0
chr19   59128983        792064  0
chr20   63025520        669359  0
chr21   48129895        340778  0
chr22   51304566        454438  0
chrX    155270560       670820  0
chrY    59373566        166595  0
*       0       0       0
```

In [73]:
```
samtools idxstats shNS3.bam
```

```
chrM    16571   871270  0
chr1    249250621       2814510 0
chr2    243199373       1989448 0
chr3    198022430       788329  0
chr4    191154276       318887  0
chr5    180915260       716809  0
chr6    171115067       959478  0
chr7    159138663       411468  0
chr8    146364022       662526  0
chr9    141213431       1103290 0
chr10   135534747       340497  0
chr11   135006516       925341  0
chr12   133851895       1350569 0
chr13   115169878       166037  0
chr14   107349540       2621110 0
```

```
chr15   102531392       292842  0
chr16   90354753        1680375 0
chr17   81195210        954999  0
chr18   78077248        118079  0
chr19   59128983        565019  0
chr20   63025520        320845  0
chr21   48129895        629767  0
chr22   51304566        194506  0
chrX    155270560       777932  0
chrY    59373566        606904  0
*       0       0       0
```

In [74]:
```
samtools idxstats shRUNX1.1.bam
```

```
chrM    16571   271715  0
chr1    249250621       2910305 0
chr2    243199373       1539109 0
chr3    198022430       1494137 0
chr4    191154276       435757  0
chr5    180915260       870253  0
chr6    171115067       1764180 0
chr7    159138663       712821  0
chr8    146364022       751771  0
chr9    141213431       1833833 0
chr10   135534747       533973  0
chr11   135006516       1387154 0
chr12   133851895       1482013 0
chr13   115169878       228545  0
chr14   107349540       5657791 0
chr15   102531392       489108  0
chr16   90354753        910049  0
chr17   81195210        2028703 0
chr18   78077248        181599  0
chr19   59128983        723543  0
chr20   63025520        558481  0
chr21   48129895        469982  0
chr22   51304566        421055  0
chrX    155270560       725105  0
chrY    59373566        206506  0
*       0       0       0
```

In [75]:
```
samtools idxstats shRUNX1.2.bam
```

```
chrM    16571   810807  0
chr1    249250621       4879473 0
chr2    243199373       3140441 0
chr3    198022430       1379193 0
chr4    191154276       566438  0
chr5    180915260       1301836 0
chr6    171115067       1437693 0
chr7    159138663       731233  0
chr8    146364022       1351121 0
chr9    141213431       1900089 0
chr10   135534747       539006  0
chr11   135006516       1498713 0
chr12   133851895       2189318 0
chr13   115169878       221253  0
chr14   107349540       4512690 0
chr15   102531392       504438  0
chr16   90354753        2215925 0
chr17   81195210        2013871 0
```

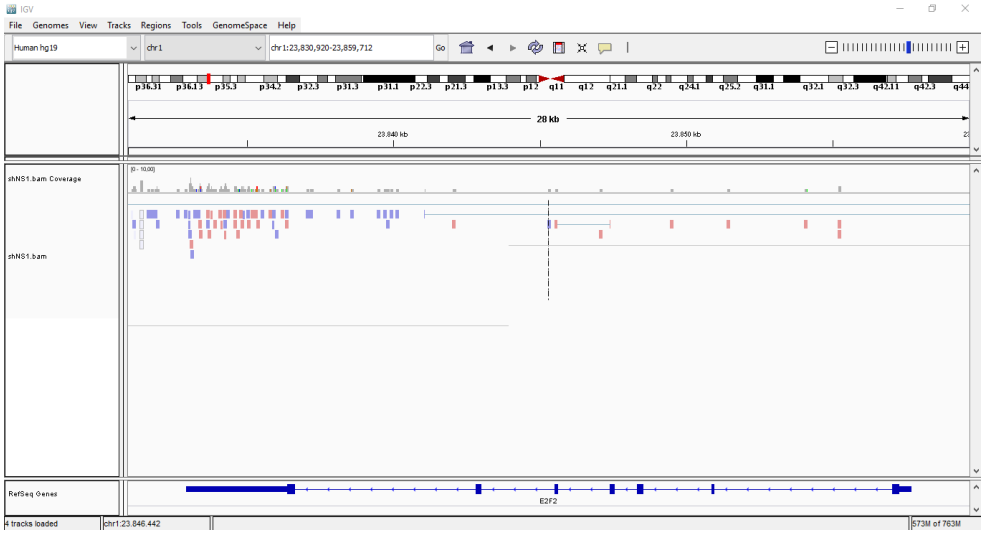
```
chr18   78077248        178502  0
chr19   59128983        1202400 0
chr20   63025520        560888  0
chr21   48129895        963271  0
chr22   51304566        320812  0
chrX    155270560       1925742 0
chrY    59373566        700532  0
*       0       0       0
```

In [76]:
```
samtools idxstats shRUNX1.3.bam
```

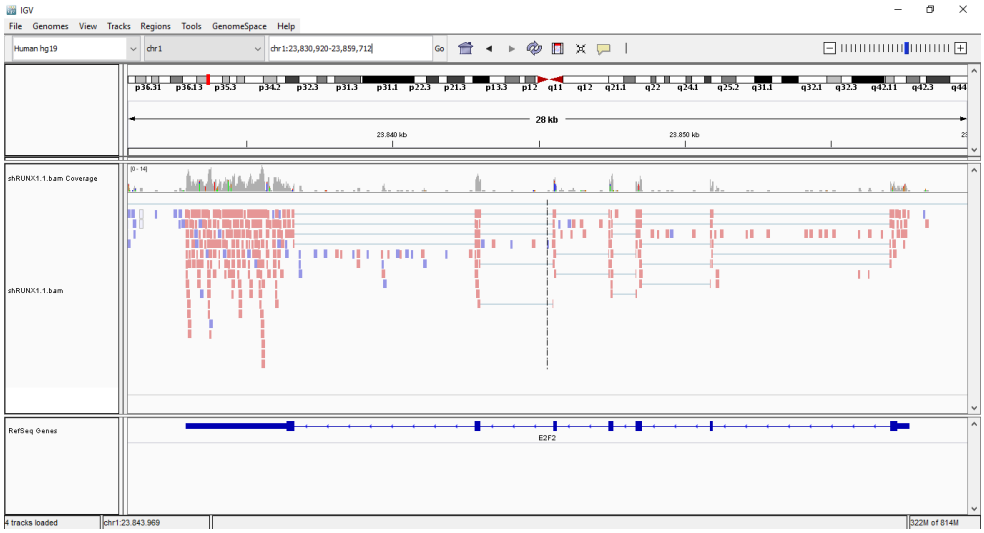
```
chrM    16571   670342  0
chr1    249250621       4045046 0
chr2    243199373       2540821 0
chr3    198022430       1385210 0
chr4    191154276       517693  0
chr5    180915260       1114479 0
chr6    171115067       1387313 0
chr7    159138663       730864  0
chr8    146364022       1125394 0
chr9    141213431       1825852 0
chr10   135534747       537007  0
chr11   135006516       1394466 0
chr12   133851895       1958834 0
chr13   115169878       219599  0
chr14   107349540       4502778 0
chr15   102531392       503738  0
chr16   90354753        1847375 0
chr17   81195210        2043330 0
chr18   78077248        174256  0
chr19   59128983        1038121 0
chr20   63025520        564558  0
chr21   48129895        807148  0
chr22   51304566        322870  0
chrX    155270560       1495841 0
chrY    59373566        540660  0
*       0       0       0
```

.bam files were checked in IGV.

shNS1



shRUNX1



In [77]:
```
#samtools flagstat
samtools flagstat shNS1.bam
```

```
25331592 + 0 in total (QC-passed reads + QC-failed reads)
6182322 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
25331592 + 0 mapped (100.00% : N/A)
0 + 0 paired in sequencing
0 + 0 read1
0 + 0 read2
0 + 0 properly paired (N/A : N/A)
0 + 0 with itself and mate mapped
0 + 0 singletons (N/A : N/A)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)
```

In [78]:
```
samtools flagstat shNS2.bam
```

```
30966793 + 0 in total (QC-passed reads + QC-failed reads)
7492951 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
30966793 + 0 mapped (100.00% : N/A)
0 + 0 paired in sequencing
0 + 0 read1
0 + 0 read2
0 + 0 properly paired (N/A : N/A)
0 + 0 with itself and mate mapped
0 + 0 singletons (N/A : N/A)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)
```

In [79]:
```
samtools flagstat shNS3.bam
```

```
22180837 + 0 in total (QC-passed reads + QC-failed reads)
4381714 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
22180837 + 0 mapped (100.00% : N/A)
0 + 0 paired in sequencing
0 + 0 read1
0 + 0 read2
```

```
0 + 0 properly paired (N/A : N/A)
0 + 0 with itself and mate mapped
0 + 0 singletons (N/A : N/A)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)
```

In [80]:
```
samtools flagstat shRUNX1.1.bam
```

```
28587488 + 0 in total (QC-passed reads + QC-failed reads)
7273525 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
28587488 + 0 mapped (100.00% : N/A)
0 + 0 paired in sequencing
0 + 0 read1
0 + 0 read2
0 + 0 properly paired (N/A : N/A)
0 + 0 with itself and mate mapped
0 + 0 singletons (N/A : N/A)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)
```

In [81]:
```
samtools flagstat shRUNX1.2.bam
```

```
37045685 + 0 in total (QC-passed reads + QC-failed reads)
7203259 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
37045685 + 0 mapped (100.00% : N/A)
0 + 0 paired in sequencing
0 + 0 read1
0 + 0 read2
0 + 0 properly paired (N/A : N/A)
0 + 0 with itself and mate mapped
0 + 0 singletons (N/A : N/A)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)
```

In [82]:
```
samtools flagstat shRUNX1.3.bam
```

```
33293595 + 0 in total (QC-passed reads + QC-failed reads)
6855992 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
33293595 + 0 mapped (100.00% : N/A)
0 + 0 paired in sequencing
0 + 0 read1
0 + 0 read2
0 + 0 properly paired (N/A : N/A)
0 + 0 with itself and mate mapped
0 + 0 singletons (N/A : N/A)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)
```

All reads passed quality control.

In [83]:
```
#Create symbolic link to human chromosomes
ln -sf /mnt/nfs/data/RNA-seq/gencode.v19.nopseudo.plus.sort.gtf
```

In [84]:
```
ls
```

```
Aligned.out.sam                        shRUNX1.2.bam
Chip-seq                               shRUNX1.2.bam.bai
gencode.v19.nopseudo.plus.sort.gtf     shRUNX1.2.fastq
Log.out                                shRUNX1.2_fastqc.html
Log.progress.out                       shRUNX1.2_fastqc.zip
shNS1.Aligned.out.sam                  shRUNX1.2.Log.final.out
shNS1.bam                              shRUNX1.2.Log.out
shNS1.bam.bai                          shRUNX1.2.Log.progress.out
shNS1.fastq                            shRUNX1.2.SJ.out.tab
shNS1_fastqc.html                      shRUNX1.3.Aligned.out.sam
shNS1_fastqc.zip                       shRUNX1.3.bam
shNS1.Log.final.out                    shRUNX1.3.bam.bai
shNS1.Log.out                          shRUNX1.3.fastq
shNS1.Log.progress.out                 shRUNX1.3_fastqc.html
shNS1.SJ.out.tab                       shRUNX1.3_fastqc.zip
shNS21.Aligned.out.sam                 shRUNX1.3.Log.final.out
shNS2.Aligned.out.sam                  shRUNX1.3.Log.out
shNS2.bam                              shRUNX1.3.Log.progress.out
shNS2.bam.bai                          shRUNX1.3.SJ.out.tab
shNS2.fastq                            SRR2919468_fastqc.html
shNS2_fastqc.html                      SRR2919468_fastqc.zip
shNS2_fastqc.zip                       SRR2919468.sra
shNS2.Log.final.out                    SRR2919468.sra.1
shNS2.Log.out                          SRR2919469.sra
shNS2.Log.progress.out                 SRR2919470_fastqc.html
shNS2.SJ.out.tab                       SRR2919470_fastqc.zip
shNS3.Aligned.out.sam                  SRR2919470.sra
shNS3.bam                              SRR2919470.sra.1
shNS3.bam.bai                          SRR2919471_fastqc.html
shNS3.fastq                            SRR2919471_fastqc.zip
shNS3_fastqc.html                      SRR2919471.sra
shNS3_fastqc.zip                       SRR2919472_fastqc.html
shNS3.Log.final.out                    SRR2919472_fastqc.zip
shNS3.Log.out                          SRR2919472.sra
shNS3.Log.progress.out                 SRR2919473_fastqc.html
shNS3.SJ.out.tab                       SRR2919473_fastqc.zip
shRUNX1.1.Aligned.out.sam              SRR2919473.sra
shRUNX1.1.bam                          *STAR.Aligned.out.sam
shRUNX1.1.bam.bai                      *STAR.Log.out
shRUNX1.1.fastq                        *STAR.Log.progress.out
shRUNX1.1_fastqc.html                  *STAR._STARtmp
shRUNX1.1_fastqc.zip                   _STARtmp
shRUNX1.1.Log.final.out                test.Aligned.out.sam
shRUNX1.1.Log.out                      test.Log.final.out
shRUNX1.1.Log.progress.out             test.Log.out
shRUNX1.1.SJ.out.tab                   test.Log.progress.out
shRUNX1.2.Aligned.out.sam              test.SJ.out.tab
```

In [87]:
```
head -8 gencode.v19.nopseudo.plus.sort.gtf
```

```
##description: evidence-based annotation of the human genome (GRCh37), v
ersion 19 (Ensembl 74)
##provider: GENCODE
##contact: gencode@sanger.ac.uk
##format: gtf
##date: 2013-12-05
chr1    HAVANA  exon    29554   30039   .       +       .       gene_id
"ENSG00000243485.2"; transcript_id "ENST00000473358.1"; gene_type "lincR
NA"; gene_status "NOVEL"; gene_name "MIR1302-11"; transcript_type "lincR
```

```
NA"; transcript_status "KNOWN"; transcript_name "MIR1302-11-001"; exon_n
umber 1;  exon_id "ENSE00001947070.1";  level 2; tag "not_best_in_genome
_evidence"; havana_gene "OTTHUMG00000000959.2"; havana_transcript "OTTHU
MT00000002840.1";
chr1    HAVANA  gene    29554   31109   .       +       .       gene_id
"ENSG00000243485.2"; transcript_id "ENSG00000243485.2"; gene_type "lincR
NA"; gene_status "NOVEL"; gene_name "MIR1302-11"; transcript_type "lincR
NA"; transcript_status "NOVEL"; transcript_name "MIR1302-11"; level 2; h
avana_gene "OTTHUMG00000000959.2";
chr1    HAVANA  transcript      29554   31097   .       +       .
gene_id "ENSG00000243485.2"; transcript_id "ENST00000473358.1"; gene_typ
e "lincRNA"; gene_status "NOVEL"; gene_name "MIR1302-11"; transcript_typ
e "lincRNA"; transcript_status "KNOWN"; transcript_name "MIR1302-11-001"
; level 2; tag "not_best_in_genome_evidence"; havana_gene "OTTHUMG000000
00959.2"; havana_transcript "OTTHUMT00000002840.1";
```

The number of reads mapped to each gene were quantified using the featureCounts command.

In [94]:
```
# Genecounts of all files
featureCounts -Q 10 -g gene_name -a gencode.v19.nopseudo.plus.sort.gtf -
o allgene.counts shNS1.bam shNS2.bam shNS3.bam \
shRUNX1.1.bam shRUNX1.2.bam shRUNX1.3.bam

#-Q count reads with minimum mapping quality score of 10
#-g use gene name field, more informative than gene id
#-a and specify which gtf file we will use
```

```
 ==========     _____ _    _ ____  _____  ______          _____
 =====         / ____| |  | |  _ \|  __ \|  ____|   /\   |  __ \
   =====      | (___ | |  | | |_) | |__) | |__     /  \  | |  | |
     ====      \___ \| |  | |  _ <|  _  /|  __|   / /\ \ | |  | |
        ====   ____) | |__| | |_) | | \ \| |____ / ____ \| |__| |
 ==========   |_____/ \____/|____/|_|  \_\______/_/    \_\_____/
 v1.5.0-p1

//========================== featureCounts setting ====================
======\\
||
      ||
||             Input files : 6 BAM files
      ||
||                           S shNS1.bam
      ||
||                           S shNS2.bam
      ||
||                           S shNS3.bam
      ||
||                           S shRUNX1.1.bam
      ||
||                           S shRUNX1.2.bam
      ||
||                           S shRUNX1.3.bam
      ||
||
      ||
||             Output file : allgene.counts
      ||
||             Annotations : gencode.v19.nopseudo.plus.sort.gtf (GTF)
      ||
||
```

```
      ||
||                Threads : 1
      ||
||                  Level : meta-feature level
      ||
||              Paired-end : no
      ||
||         Strand specific : no
      ||
||      Multimapping reads : not counted
      ||
|| Multi-overlapping reads : not counted
      ||
||       Read orientations : fr
      ||
||
      ||
\\====================== http://subread.sourceforge.net/ ================
======//

//================================== Running ============================
======\\
||
      ||
|| Load annotation file gencode.v19.nopseudo.plus.sort.gtf ...
      ||
||    Features : 1156385
      ||
||    Meta-features : 41863
      ||
||    Chromosomes/contigs : 25
      ||
||
      ||
|| Process BAM file shNS1.bam...
      ||
||    Single-end reads are included.
      ||
||    Assign reads to features...
      ||
||    Total reads : 25331592
      ||
||    Successfully assigned reads : 9202975 (36.3%)
      ||
||    Running time : 0.40 minutes
      ||
||
      ||
|| Process BAM file shNS2.bam...
      ||
||    Single-end reads are included.
      ||
||    Assign reads to features...
      ||
||    Total reads : 30966793
      ||
||    Successfully assigned reads : 13072801 (42.2%)
      ||
||    Running time : 0.50 minutes
      ||
||
```

```
                    ||
                || Process BAM file shNS3.bam...
                    ||
                ||    Single-end reads are included.
                    ||
                ||    Assign reads to features...
                    ||
                ||    Total reads : 22180837
                    ||
                ||    Successfully assigned reads : 6566875 (29.6%)
                    ||
                ||    Running time : 0.44 minutes
                    ||
                ||
                    ||
                || Process BAM file shRUNX1.1.bam...
                    ||
                ||    Single-end reads are included.
                    ||
                ||    Assign reads to features...
                    ||
                ||    Total reads : 28587488
                    ||
                ||    Successfully assigned reads : 11202455 (39.2%)
                    ||
                ||    Running time : 0.48 minutes
                    ||
                ||
                    ||
                || Process BAM file shRUNX1.2.bam...
                    ||
                ||    Single-end reads are included.
                    ||
                ||    Assign reads to features...
                    ||
                ||    Total reads : 37045685
                    ||
                ||    Successfully assigned reads : 10788451 (29.1%)
                    ||
                ||    Running time : 0.70 minutes
                    ||
                ||
                    ||
                || Process BAM file shRUNX1.3.bam...
                    ||
                ||    Single-end reads are included.
                    ||
                ||    Assign reads to features...
                    ||
                ||    Total reads : 33293595
                    ||
                ||    Successfully assigned reads : 10563869 (31.7%)
                    ||
                ||    Running time : 0.52 minutes
                    ||
                ||
                    ||
                ||                          Read assignment finished.
                    ||
                ||
                    ||
```

```
\\====================  http://subread.sourceforge.net/  ================
======//
```

In [95]:
```
ls -l *.counts
head -25 allgene.counts.summary | column -t
```

```
-rw-r--r-- 1 r0381150 domain users 8888773 jan 16 13:55 allgene.counts
-rw-r--r-- 1 r0381150 domain users 8344230 jan 16 13:43 shNS1.counts
Status                      shNS1.bam  shNS2.bam  shNS3.bam  shRUNX1.1.ba
m  shRUNX1.2.bam  shRUNX1.3.bam
Assigned                    9202975    13072801   6566875    11202455
   10788451       10563869
Unassigned_Ambiguity        612021     861139     410369     767761
   696060         703926
Unassigned_MultiMapping     0          0          0          0
   0              0
Unassigned_NoFeatures       6080016    5722461    8131286    5458528
   13720904       11001321
Unassigned_Unmapped         0          0          0          0
   0              0
Unassigned_MappingQuality   9436580    11310392   7072307    11158744
   11840270       11024479
Unassigned_FragmentLength   0          0          0          0
   0              0
Unassigned_Chimera          0          0          0          0
   0              0
Unassigned_Secondary        0          0          0          0
   0              0
Unassigned_Nonjunction      0          0          0          0
   0              0
Unassigned_Duplicate        0          0          0          0
   0              0
```

In [97]:
```
head -15 allgene.counts
```

```
# Program:featureCounts v1.5.0-p1; Command:"featureCounts" "-Q" "10" "-g
" "gene_name" "-a" "gencode.v19.nopseudo.plus.sort.gtf" "-o" "allgene.co
unts" "shNS1.bam" "shNS2.bam" "shNS3.bam" "shRUNX1.1.bam" "shRUNX1.2.bam
" "shRUNX1.3.bam"
Geneid  Chr     Start   End     Strand  Length  shNS1.bam       shNS2.ba
m       shNS3.bam       shRUNX1.1.bam   shRUNX1.2.bam   shRUNX1.3.bam
MIR1302-11      chr1;chr1;chr1  29554;30267;30976       30039;30667;3110
9       +;+;+   1021    0       0       0       0       0
FAM138A chr1;chr1;chr1  34554;35245;35721       35174;35481;36081
-;-;-   1219    0       0       0       0       0       0
OR4F5   chr1    69091   70008   +       918     0       0       0
0       0       0
RP11-34P13.7    chr1;chr1;chr1;chr1;chr1;chr1;chr1      89295;92091;1109
53;112700;120721;129055;133374  91629;92240;111357;112804;120932;129223;
133566  -;-;-;-;-;-;-   3569    2       0       0       0       0
1
RP11-34P13.8    chr1;chr1       89551;90287     90050;91105     -;-
1319    0       0       0       0       0       0
AL627309.1      chr1;chr1       134901;137621   135802;139379   -;-
2661    13      14      21      6       15      12
RP11-34P13.14   chr1;chr1       139790;140075   139847;140339   -;-
323     0       0       0       0       0       0
RP11-34P13.13   chr1;chr1;chr1;chr1;chr1;chr1;chr1;chr1;chr1;chr1
141474;146386;155767;164263;165884;168100;168610;169049;172557;173753
143011;149707;155831;164791;165942;168165;168767;169264;172688;173862
```

```
-;-;-;-;-;-;-;-;-;-     6195    9       5       8       5       11
11
RNU6-1100P      chr1    157784  157887  -       104     0       0
0       0       0       0
RP11-34P13.9    chr1;chr1       160446;161314   160690;161525   +;+
457     0       0       1       0       0       1
AP006222.2      chr1;chr1;chr1;chr1;chr1;chr1;chr1      227615;234355;23
6112;238418;257268;259017;267096        228776;234484;237955;238567;2576
72;259121;267253        -;-;-;-;-;-;-   3954    1       2       2
2       1       2
RP4-669L17.10   chr1;chr1;chr1;chr1;chr1;chr1;chr1;chr1;chr1;chr1;chr1;c
hr1;chr1;chr1;chr1;chr1;chr1;chr1       317720;319968;320881;321032;3220
38;322672;323861;324288;324439;327552;329784;334129;342392;439365;446014
;450887;453217;453645   317781;320653;320938;321290;322228;323073;324060
;324345;326514;328453;329976;334505;342806;439568;446193;451086;453318;4
53948   +;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+     6869    0       1
0       1       0       1
OR4F29  chr1    367640  368634  +       995     0       0       0
0       0       0
```

In [98]:
```
cut -f-6 allgene.counts > allgene.genedata.tsv
#take out first 6 columns of all.counts and compile it into allgene gene
data.tsv file
#first 6 columns contain metadata
```

In [99]:
```
ls -l *.tsv
```

```
-rw-r--r-- 1 r0381150 domain users 8236281 jan 16 14:06 allgene.genedata
.tsv
```

In [100]:
```
head allgene.genedata.tsv
```

```
# Program:featureCounts v1.5.0-p1; Command:"featureCounts" "-Q" "10" "-g
" "gene_name" "-a" "gencode.v19.nopseudo.plus.sort.gtf" "-o" "allgene.co
unts" "shNS1.bam" "shNS2.bam" "shNS3.bam" "shRUNX1.1.bam" "shRUNX1.2.bam
" "shRUNX1.3.bam"
Geneid  Chr     Start   End     Strand  Length
MIR1302-11      chr1;chr1;chr1  29554;30267;30976       30039;30667;3110
9       +;+;+   1021
FAM138A chr1;chr1;chr1  34554;35245;35721       35174;35481;36081
-;-;-   1219
OR4F5   chr1    69091   70008   +       918
RP11-34P13.7    chr1;chr1;chr1;chr1;chr1;chr1;chr1      89295;92091;1109
53;112700;120721;129055;133374  91629;92240;111357;112804;120932;129223;
133566  -;-;-;-;-;-;-   3569
RP11-34P13.8    chr1;chr1       89551;90287     90050;91105     -;-
1319
AL627309.1      chr1;chr1       134901;137621   135802;139379   -;-
2661
RP11-34P13.14   chr1;chr1       139790;140075   139847;140339   -;-
323
RP11-34P13.13   chr1;chr1;chr1;chr1;chr1;chr1;chr1;chr1;chr1;chr1;chr1
141474;146386;155767;164263;165884;168100;168610;169049;172557;173753
143011;149707;155831;164791;165942;168165;168767;169264;172688;173862
-;-;-;-;-;-;-;-;-;-     6195
```

In [101]:
```
cut -f1,7- allgene.counts | grep -v '^#' > allgene.gene.counts.tsv
```

In [103]:
```
head allgene.gene.counts.tsv | column -t
```

```
Geneid          shNS1.bam  shNS2.bam  shNS3.bam  shRUNX1.1.bam  shRUNX1.2
.bam  shRUNX1.3.bam
MIR1302-11      0          0          0          0              0
      0
FAM138A         0          0          0          0              0
      0
OR4F5           0          0          0          0              0
      0
RP11-34P13.7    2          0          0          0              0
      1
RP11-34P13.8    0          0          0          0              0
      0
AL627309.1      13         14         21         6              15
      12
RP11-34P13.14   0          0          0          0              0
      0
RP11-34P13.13   9          5          8          5              11
      11
RNU6-1100P      0          0          0          0              0
      0
```

In [106]:
```
grep RUNX1 allgene.gene.counts.tsv | column -t
```

```
Geneid    shNS1.bam  shNS2.bam  shNS3.bam  shRUNX1.1.bam  shRUNX1.2.bam
shRUNX1.3.bam
RUNX1     333        466        245        498            501
558
RUNX1T1   10         11         7          11             21
17
```

The obtained allgene.gene.counts.tsv data were further analyzed with DeSEQ. This can be found in the attached R notebook.

DESeq analysis of RNAseq data

Group members

Florian Delbart

Jens Rummens

Alexander Sneyers

The DESeq tool was used to analyze the count data of RNAseq and identify upregulated and downregulated genes.

In [1]:
```
library("DESeq2")
library(scales)
options(repr.plot.width=4, repr.plot.height=3)
```

```
Loading required package: S4Vectors
Loading required package: stats4
Loading required package: BiocGenerics
Loading required package: parallel

Attaching package: ‘BiocGenerics’

The following objects are masked from ‘package:parallel’:

    clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,
    clusterExport, clusterMap, parApply, parCapply, parLapply,
    parLapplyLB, parRapply, parSapply, parSapplyLB

The following object is masked from ‘package:stats’:

    xtabs

The following objects are masked from ‘package:base’:

    anyDuplicated, append, as.data.frame, as.vector, cbind, colnames,
    do.call, duplicated, eval, evalq, Filter, Find, get, intersect,
    is.unsorted, lapply, Map, mapply, match, mget, order, paste, pmax,
    pmax.int, pmin, pmin.int, Position, rank, rbind, Reduce, rep.int,
    rownames, sapply, setdiff, sort, table, tapply, union, unique,
    unlist, unsplit

Creating a generic function for ‘nchar’ from package ‘base’ in package ‘
S4Vectors’
Loading required package: IRanges
Loading required package: GenomicRanges
Loading required package: GenomeInfoDb
Loading required package: Rcpp
Loading required package: RcppArmadillo
Warning message:
: replacing previous import ‘S4Vectors::Position’ by ‘ggplot2::Position’
```

```
when loading ‘DESeq2’
```

In [2]:
```
setwd(path.expand("~/data/Assignment/DeSEQ"))
#set working directory
```

To be able to perform the DESeq analysis a matrix of the gene count data was constructed

In [3]:
```
matrixFile = path.expand("~/data/Assignment/DeSEQ/allgene.gene.counts.tsv")
#define matrix file
```

In [4]:
```
countData <- as.matrix(read.csv(matrixFile, sep="\t", row.names="Geneid"))
head(countData)
#making matrix of gene count data
```

	shNS1.bam	shNS2.bam	shNS3.bam	shRUNX1.1.bam	shRUNX1.2.bam	shRU
MIR1302-11	0	0	0	0	0	0
FAM138A	0	0	0	0	0	0
OR4F5	0	0	0	0	0	0
RP11-34P13.7	2	0	0	0	0	1
RP11-34P13.8	0	0	0	0	0	0
AL627309.1	13	14	21	6	15	12

Changing the layout of the matrix

In [5]:
```
colnames(countData) <- c("shNS1", "shNS2", "shNS3", "shRUNX1.1", "shRUNX1.2", "shRUNX1.3")
head(countData)
#Change column names to experimental conditions rather than file name
```

	shNS1	shNS2	shNS3	shRUNX1.1	shRUNX1.2	shRUNX1.3
MIR1302-11	0	0	0	0	0	0
FAM138A	0	0	0	0	0	0
OR4F5	0	0	0	0	0	0
RP11-34P13.7	2	0	0	0	0	1
RP11-34P13.8	0	0	0	0	0	0
AL627309.1	13	14	21	6	15	12

Defining which samples correspond to control or knockdown conditions

In [6]:
```
#Name of experimental conditions must be the same as data, very importan
t.
```

```
condition <- factor(c("shNS", "shNS", "shNS", "shRUNX1", "shRUNX1", "shR
UNX1"))
print(condition)
```

```
[1] shNS    shNS    shNS    shRUNX1 shRUNX1 shRUNX1
Levels: shNS shRUNX1
```

In [7]:
```
#Make table with sample names and corresponding conditions
colData <- data.frame(sampleName = colnames(countData),
                      condition = condition)
colData
```

sampleName	condition
shNS1	shNS
shNS2	shNS
shNS3	shNS
shRUNX1.1	shRUNX1
shRUNX1.2	shRUNX1
shRUNX1.3	shRUNX1

In [1]:
```
#Reformatting the generated dataset to a DESeq dataset to be able to use
 the DESeq tool
dds <- DESeqDataSetFromMatrix(countData = countData,
                              colData = colData,
                              design = ~ condition)
#Make sure the column names are correct.
colnames(dds) <- colnames(countData)
```

```
Error in eval(expr, envir, enclos): could not find function "DESeqDataSe
tFromMatrix"
Traceback:
```

In [9]:
```
#Summary of the DESeq dataset
dds
```

```
class: DESeqDataSet
dim: 41863 6
exptData(0):
assays(1): counts
rownames(41863): MIR1302-11 FAM138A ... AC007965.1 RNU6-1314P
rowRanges metadata column names(0):
colnames(6): shNS1 shNS2 ... shRUNX1.2 shRUNX1.3
colData names(2): sampleName condition
```

In [10]:
```
#Check if conditions contain all samples
dds$condition
```

shNS shNS shNS shRUNX1 shRUNX1 shRUNX1

Analysis of differentially expressed genes

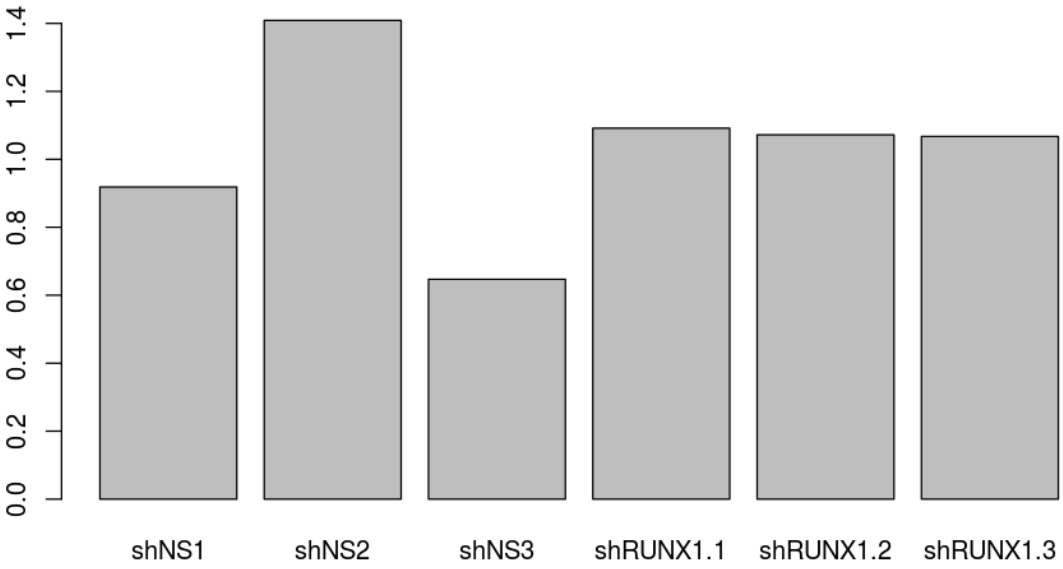
In [11]:
```
dds <- DESeq(dds)
```

```
estimating size factors
estimating dispersions
```

```
gene-wise dispersion estimates
mean-dispersion relationship
final dispersion estimates
fitting model and testing
```

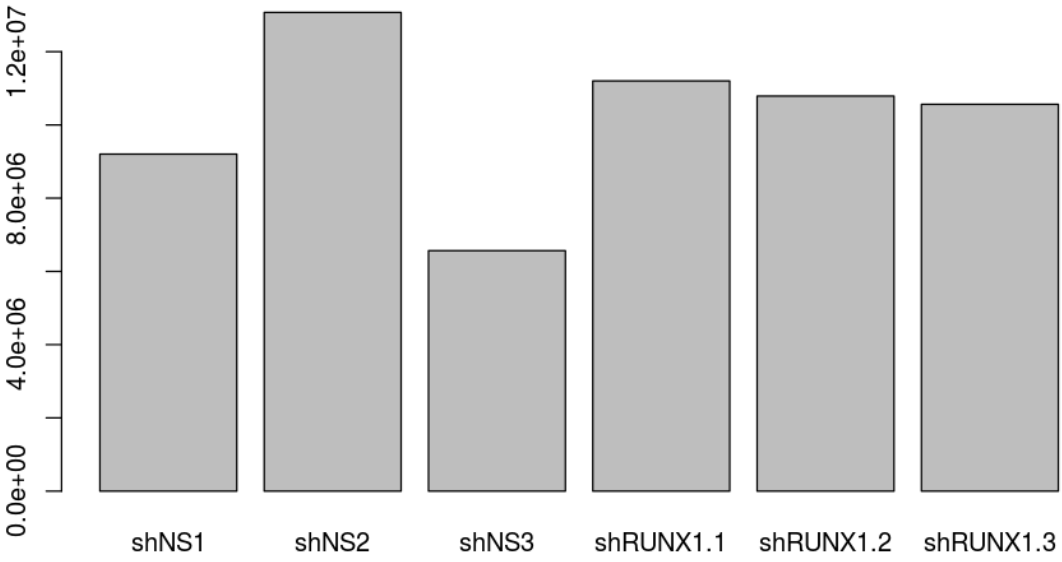
Quality control for library size.

In [19]:
```
#Plot of relative library sizes of the different samples
options(repr.plot.width=8, repr.plot.height=5)
barplot(sizeFactors(dds))
```



Large variation in library size was observed between the shNS conditions. Much less variation in library size in shRUNX1 samples. DeSEQ corrects for variations in library size.

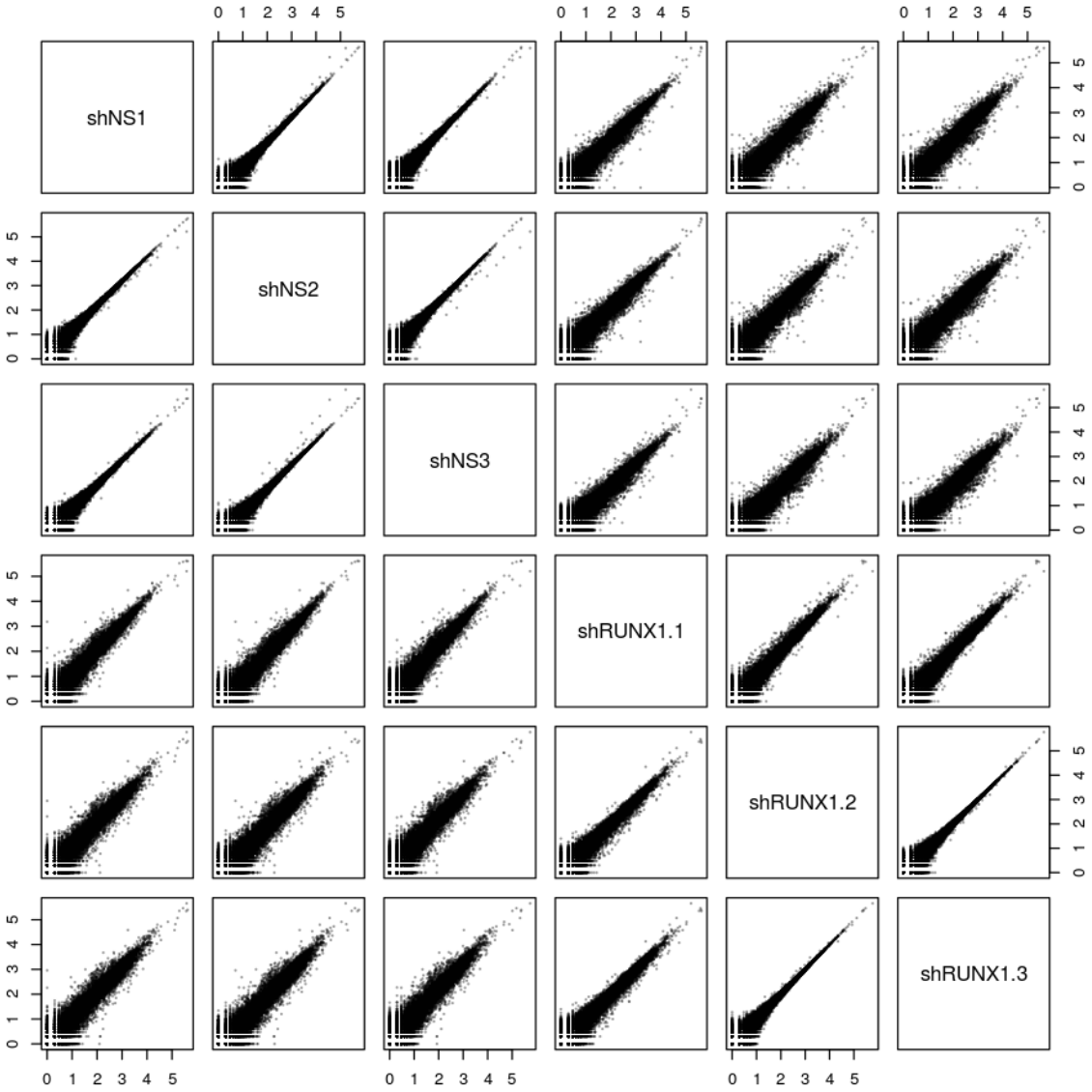
In [20]:
```
#Plot of absolute library sizes of the different samples
barplot(colSums(countData))
```



Checking for correlation in gene expression between different samples

Gene expression varies less in samples from the same experimental condition

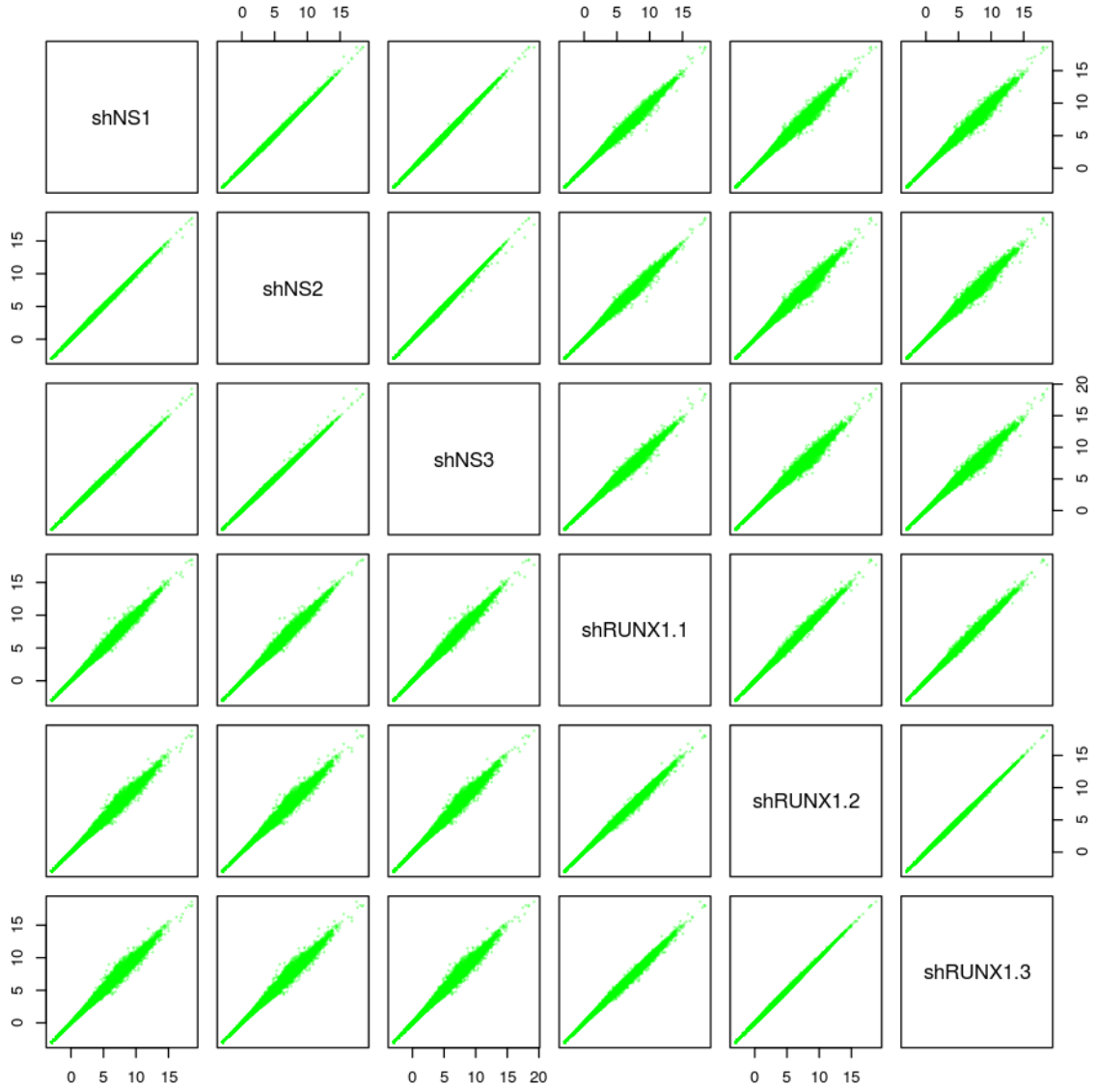
In [21]:
```
options(repr.plot.width=8, repr.plot.height=8)
plot(data.frame(log10(counts(dds))), cex=0.1, col=alpha("black", 0.4))
```



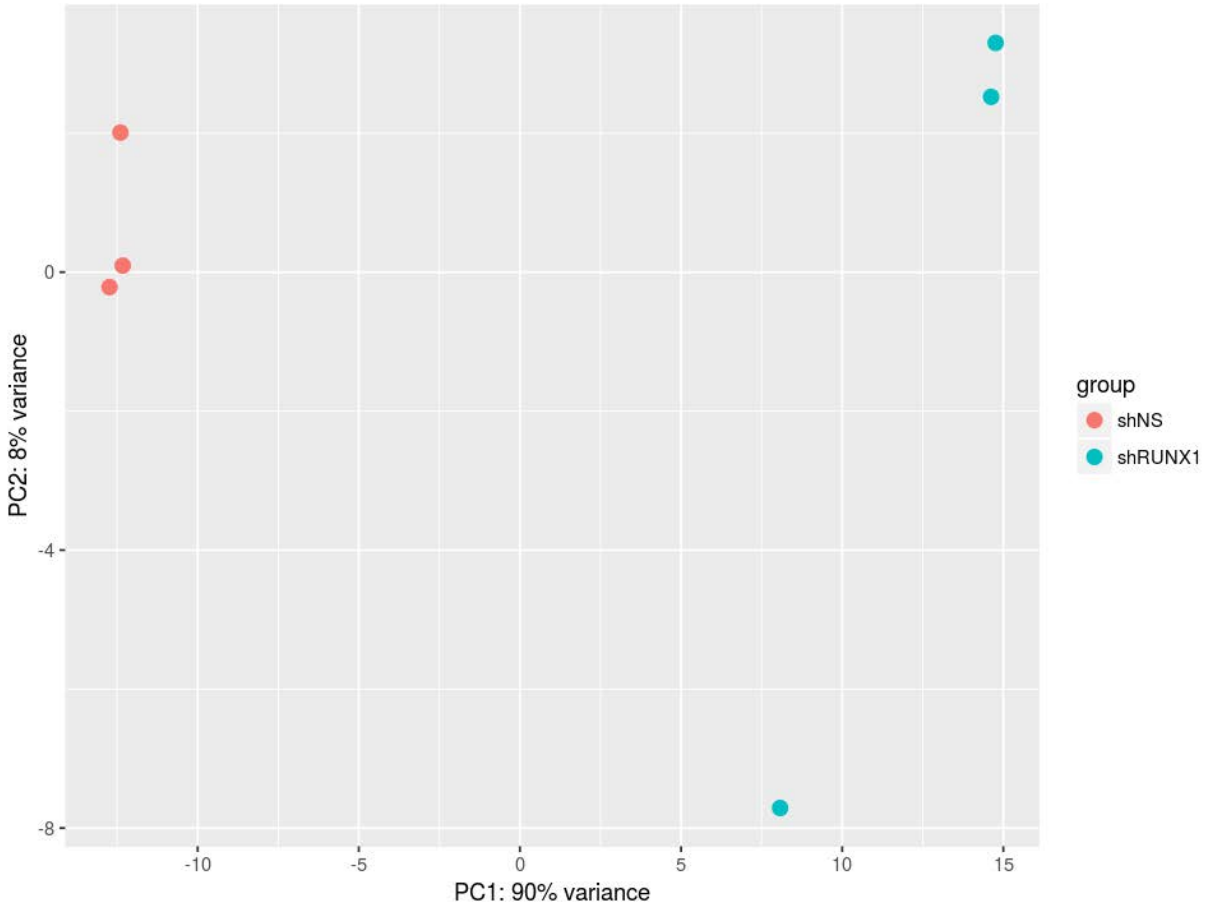
To reduce the large variability at low expression levels a log transformation was performed.

In [22]:
```
rld <- rlogTransformation(dds, blind=TRUE)
```

In [23]:
```
options(repr.plot.width=8, repr.plot.height=8)
plot(data.frame(assay(rld)), cex=0.1, col=alpha("green", 0.4))
```

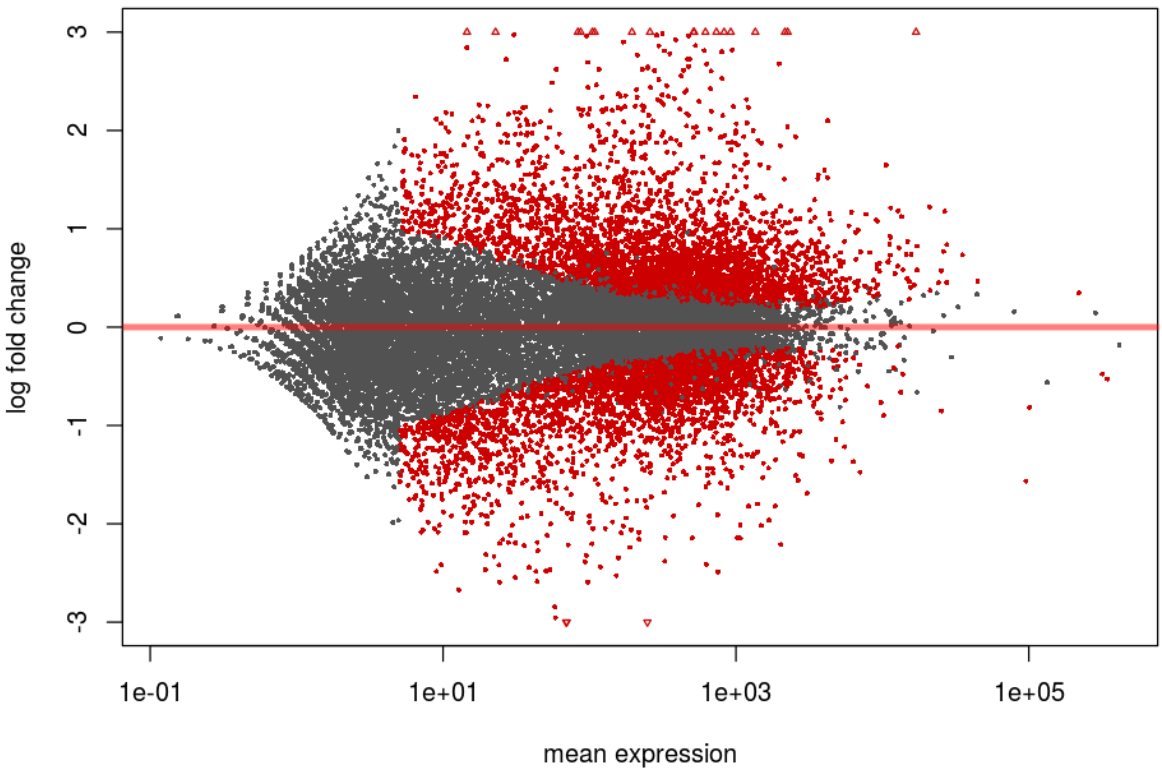


In [33]:
```
plotPCA(rld, intgroup=c("condition"))
```



Making a MA plot to visualize significantly differentialy expressed genes

In [25]:
```
options(repr.plot.width=8, repr.plot.height=6)
plotMA(dds,ylim=c(-3,3))
```



Plots log2fold change vs mean expression. Significant differentially expressed genes are shown in red.

Exporting the results of DESeq analysis

In [44]:
```
res <- results(dds)
res <- res[order(res$log2FoldChange),]
head(res)
#Ordering the results according to log2foldchange
```

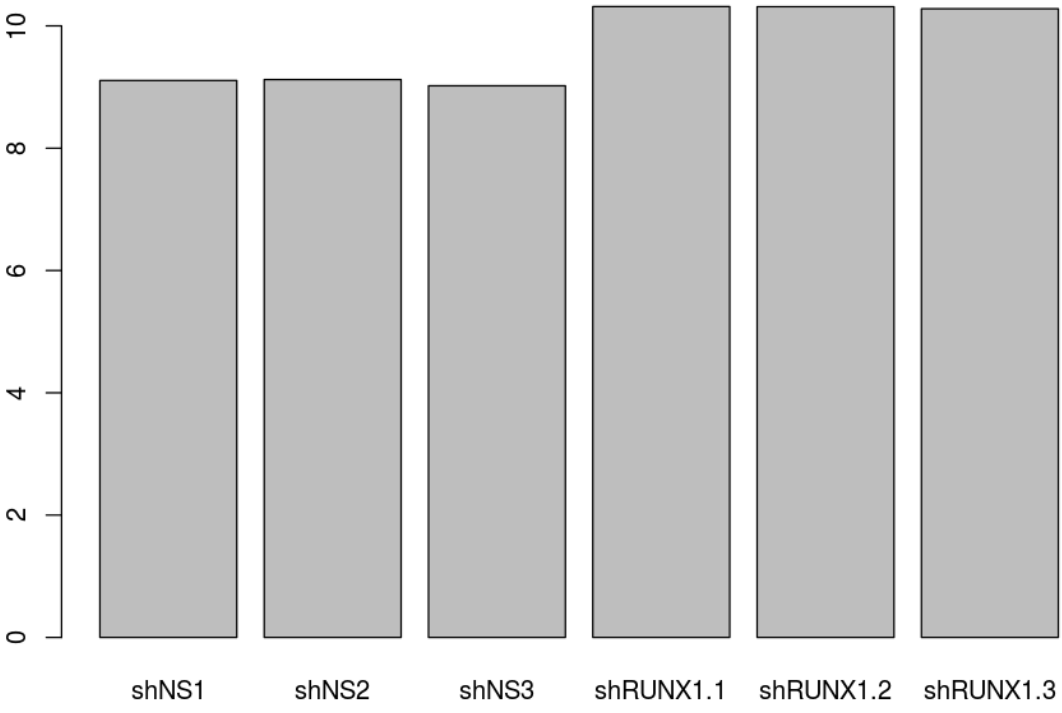
```
log2 fold change (MAP): condition shRUNX1 vs shNS
Wald test p-value: condition shRUNX1 vs shNS
DataFrame with 6 rows and 6 columns
        baseMean log2FoldChange     lfcSE       stat       pvalue
  padj
       <numeric>      <numeric> <numeric>  <numeric>    <numeric>    <nu
meric>
CYP1A1  69.32013      -4.700095 0.3563651 -13.188988 1.015396e-39 2.5416
01e-37
PCSK2   69.91344      -4.376865 0.3415849 -12.813402 1.379511e-37 3.0152
99e-35
MME    248.48050      -3.003275 0.2448537 -12.265589 1.386007e-34 2.5606
49e-32
MMP13   58.59528      -2.953780 0.2861810 -10.321368 5.641329e-25 5.2739
63e-23
TPM2    57.97398      -2.846650 0.2900655  -9.813820 9.818026e-23 7.8946
08e-21
CHKB    12.77972      -2.670959 0.4457153  -5.992521 2.066125e-09 3.6855
39e-08
```

In [45]:
```
res.sign <- res[(!is.na(res$padj)) & (res$padj < 0.05),]
dim(res.sign)
#Counting the number of significantly expressed genes, not including NA
```

```
5395  6
```

In [46]:
```
write.table(res, "DeSEQRUNX1logfc.results.tsv", sep="\t", col.names=NA,
quote=FALSE)
#Exporting the results as .tsv file
```

In [40]:
```
barplot(assay(rld)["GLA",])
#Visualizing the expression levels of a specific gene within the differe
nt samples
```



In [ ]:

ChIP-seq data analysis

In [149]:
```
cd Chip-seq
```

ChIP-seq data was obtained from ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX143/SRX1434851/

In [33]:
```
wget ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX143/SRX1434851/SRR2919474/*.sra

wget ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX143/SRX1434852/SRR2919475/*.sra

wget ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX143/SRX1434853/SRR2919476/*.sra

wget ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX143/SRX1434854/SRR2919477/*

#.sra files were deleted from the server after creating the fastq files.
```

```
--2016-12-31 17:22:36--  ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX143/SRX1434851/SRR2919474/*.sra
           => '.listing'
Resolving ftp-trace.ncbi.nlm.nih.gov (ftp-trace.ncbi.nlm.nih.gov)... 130.14.250.7, 2607:f220:41e:250::13
Connecting to ftp-trace.ncbi.nlm.nih.gov (ftp-trace.ncbi.nlm.nih.gov)|130.14.250.7|:21... connected.
Logging in as anonymous ... Logged in!
==> SYST ... done.    ==> PWD ... done.
==> TYPE I ... done.  ==> CWD (1) /sra/sra-instant/reads/ByExp/sra/SRX/SRX143/SRX1434851/SRR2919474 ... done.
==> PASV ... done.    ==> LIST ... done.

    [ <=>                                   ] 74          --.-K/s   in 0,003s

2016-12-31 17:22:38 (24,3 KB/s) - '.listing' saved [74]

Removed '.listing'.
--2016-12-31 17:22:38--  ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX143/SRX1434851/SRR2919474/SRR2919474.sra
           => 'SRR2919474.sra'
==> CWD not required.
==> PASV ... done.    ==> RETR SRR2919474.sra ... done.
Length: 1400174326 (1,3G)

100%[=====================================>] 1.400.174.326 35,2MB/s   in 47s

2016-12-31 17:23:25 (28,1 MB/s) - 'SRR2919474.sra' saved [1400174326]

--2016-12-31 17:23:26--  ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX143/SRX1434852/SRR2919475/*.sra
           => '.listing'
Resolving ftp-trace.ncbi.nlm.nih.gov (ftp-trace.ncbi.nlm.nih.gov)... 130.14.250.7, 2607:f220:41e:250::11
Connecting to ftp-trace.ncbi.nlm.nih.gov (ftp-trace.ncbi.nlm.nih.gov)|13
```

```
0.14.250.7|:21... connected.
Logging in as anonymous ... Logged in!
==> SYST ... done.    ==> PWD ... done.
==> TYPE I ... done.  ==> CWD (1) /sra/sra-instant/reads/ByExp/sra/SRX/S
RX143/SRX1434852/SRR2919475 ... done.
==> PASV ... done.    ==> LIST ... done.

    [ <=>                                   ] 74          --.-K/s   in 0
,003s

2016-12-31 17:23:27 (26,6 KB/s) - '.listing' saved [74]

Removed '.listing'.
--2016-12-31 17:23:27--  ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instan
t/reads/ByExp/sra/SRX/SRX143/SRX1434852/SRR2919475/SRR2919475.sra
           => 'SRR2919475.sra'
==> CWD not required.
==> PASV ... done.    ==> RETR SRR2919475.sra ... done.
Length: 1403912398 (1,3G)

100%[====================================>] 1.403.912.398 35,6MB/s   in
43s

2016-12-31 17:24:10 (30,9 MB/s) - 'SRR2919475.sra' saved [1403912398]

--2016-12-31 17:24:11--  ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instan
t/reads/ByExp/sra/SRX/SRX143/SRX1434853/SRR2919476/*.sra
           => '.listing'
Resolving ftp-trace.ncbi.nlm.nih.gov (ftp-trace.ncbi.nlm.nih.gov)... 130
.14.250.10, 2607:f220:41e:250::12
Connecting to ftp-trace.ncbi.nlm.nih.gov (ftp-trace.ncbi.nlm.nih.gov)|13
0.14.250.10|:21... connected.
Logging in as anonymous ... Logged in!
==> SYST ... done.    ==> PWD ... done.
==> TYPE I ... done.  ==> CWD (1) /sra/sra-instant/reads/ByExp/sra/SRX/S
RX143/SRX1434853/SRR2919476 ... done.
==> PASV ... done.    ==> LIST ... done.

    [ <=>                                   ] 73          --.-K/s   in 0
,003s

2016-12-31 17:24:12 (25,5 KB/s) - '.listing' saved [73]

Removed '.listing'.
--2016-12-31 17:24:12--  ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instan
t/reads/ByExp/sra/SRX/SRX143/SRX1434853/SRR2919476/SRR2919476.sra
           => 'SRR2919476.sra'
==> CWD not required.
==> PASV ... done.    ==> RETR SRR2919476.sra ... done.
Length: 513525795 (490M)

100%[=====================================>] 513.525.795 22,5MB/s   in
23s

2016-12-31 17:24:35 (21,5 MB/s) - 'SRR2919476.sra' saved [513525795]

--2016-12-31 17:24:35--  ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instan
t/reads/ByExp/sra/SRX/SRX143/SRX1434854/SRR2919477/*
           => '.listing'
Resolving ftp-trace.ncbi.nlm.nih.gov (ftp-trace.ncbi.nlm.nih.gov)... 130
.14.250.10, 2607:f220:41e:250::12
```

```
Connecting to ftp-trace.ncbi.nlm.nih.gov (ftp-trace.ncbi.nlm.nih.gov)|13
0.14.250.10|:21... connected.
Logging in as anonymous ... Logged in!
==> SYST ... done.    ==> PWD ... done.
==> TYPE I ... done.  ==> CWD (1) /sra/sra-instant/reads/ByExp/sra/SRX/S
RX143/SRX1434854/SRR2919477 ... done.
==> PASV ... done.    ==> LIST ... done.

    [ <=>                                   ] 74          --.-K/s   in 0
,002s

2016-12-31 17:24:36 (31,5 KB/s) - '.listing' saved [74]

Removed '.listing'.
--2016-12-31 17:24:36--  ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instan
t/reads/ByExp/sra/SRX/SRX143/SRX1434854/SRR2919477/SRR2919477.sra
           => 'SRR2919477.sra'
==> CWD not required.
==> PASV ... done.    ==> RETR SRR2919477.sra ... done.
Length: 1407172002 (1,3G)

100%[====================================>] 1.407.172.002 35,2MB/s   in
44s

2016-12-31 17:25:20 (30,8 MB/s) - 'SRR2919477.sra' saved [1407172002]
```

In [34]:
```
ls
```

```
SRR2919474.sra  SRR2919475.sra  SRR2919476.sra  SRR2919477.sra
```

In [ ]:
```
fastq-dump --split-3 *.sra
#convert Chip-seq .sra files to .fastq
```

In [109]:
```
ls -l *.fastq
#.fastq files were deleted after data analysis was completed.
```

```
-rw-r--r-- 1 r0381150 domain users 10592714752 dec 31 18:49 SRR2919474.f
astq
-rw-r--r-- 1 r0381150 domain users 10472259584 dec 31 18:52 SRR2919475.f
astq
-rw-r--r-- 1 r0381150 domain users  3825071378 dec 31 18:54 SRR2919476.f
astq
-rw-r--r-- 1 r0381150 domain users 10562343272 dec 31 18:58 SRR2919477.f
astq
```

In [110]:
```
#Rename .fastq files
# Input: control
# Pulldown
# 1 and 2 specify experimental replicate 1 and 2, respectively.

mv -n SRR2919474.fastq Input1.fastq
mv -n SRR2919475.fastq Pulldown1.fastq
mv -n SRR2919476.fastq Input2.fastq
mv -n SRR2919477.fastq Pulldown2.fastq
```

In [111]:
```
ls -l *.fastq
```

```
-rw-r--r-- 1 r0381150 domain users 10592714752 dec 31 18:49 Input1.fastq
```

```
-rw-r--r-- 1 r0381150 domain users  3825071378 dec 31 18:54 Input2.fastq
-rw-r--r-- 1 r0381150 domain users 10472259584 dec 31 18:52 Pulldown1.fa
stq
-rw-r--r-- 1 r0381150 domain users 10562343272 dec 31 18:58 Pulldown2.fa
stq
```

In [112]:
```
#The bowtie read aligner was used to map the fastq reads to the referenc
e genome
bowtie2 --help
```

```
Bowtie 2 version 2.1.0 by Ben Langmead (langmea@cs.jhu.edu, www.cs.jhu.e
du/~langmea)
Usage:
  bowtie2 [options]* -x <bt2-idx> {-1 <m1> -2 <m2> | -U <r>} [-S <sam>]

  <bt2-idx>  Index filename prefix (minus trailing .X.bt2).
             NOTE: Bowtie 1 and Bowtie 2 indexes are not compatible.
  <m1>       Files with #1 mates, paired with files in <m2>.
             Could be gzip'ed (extension: .gz) or bzip2'ed (extension: .
bz2).
  <m2>       Files with #2 mates, paired with files in <m1>.
             Could be gzip'ed (extension: .gz) or bzip2'ed (extension: .
bz2).
  <r>        Files with unpaired reads.
             Could be gzip'ed (extension: .gz) or bzip2'ed (extension: .
bz2).
  <sam>      File for SAM output (default: stdout)

  <m1>, <m2>, <r> can be comma-separated lists (no whitespace) and can b
e
  specified many times.  E.g. '-U file1.fq,file2.fq -U file3.fq'.

Options (defaults in parentheses):

 Input:
  -q                 query input files are FASTQ .fq/.fastq (default)
  --qseq             query input files are in Illumina's qseq format
  -f                 query input files are (multi-)FASTA .fa/.mfa
  -r                 query input files are raw one-sequence-per-line
  -c                 <m1>, <m2>, <r> are sequences themselves, not files
  -s/--skip <int>    skip the first <int> reads/pairs in the input (none
)
  -u/--upto <int>    stop after first <int> reads/pairs (no limit)
  -5/--trim5 <int>   trim <int> bases from 5'/left end of reads (0)
  -3/--trim3 <int>   trim <int> bases from 3'/right end of reads (0)
  --phred33          qualities are Phred+33 (default)
  --phred64          qualities are Phred+64
  --int-quals        qualities encoded as space-delimited integers

 Presets:                 Same as:
  For --end-to-end:
   --very-fast           -D 5 -R 1 -N 0 -L 22 -i S,0,2.50
   --fast                -D 10 -R 2 -N 0 -L 22 -i S,0,2.50
   --sensitive           -D 15 -R 2 -N 0 -L 22 -i S,1,1.15 (default)
   --very-sensitive      -D 20 -R 3 -N 0 -L 20 -i S,1,0.50

  For --local:
   --very-fast-local     -D 5 -R 1 -N 0 -L 25 -i S,1,2.00
   --fast-local          -D 10 -R 2 -N 0 -L 22 -i S,1,1.75
   --sensitive-local     -D 15 -R 2 -N 0 -L 20 -i S,1,0.75 (default)
   --very-sensitive-local -D 20 -R 3 -N 0 -L 20 -i S,1,0.50
```

```
 Alignment:
  -N <int>           max # mismatches in seed alignment; can be 0 or 1 (
0)
  -L <int>           length of seed substrings; must be >3, <32 (22)
  -i <func>          interval between seed substrings w/r/t read len (S,
1,1.15)
  --n-ceil <func>    func for max # non-A/C/G/Ts permitted in aln (L,0,0
.15)
  --dpad <int>       include <int> extra ref chars on sides of DP table
(15)
  --gbar <int>       disallow gaps within <int> nucs of read extremes (4
)
  --ignore-quals     treat all quality values as 30 on Phred scale (off)
  --nofw             do not align forward (original) version of read (of
f)
  --norc             do not align reverse-complement version of read (of
f)

  --end-to-end       entire read must align; no clipping (on)
   OR
  --local            local alignment; ends might be soft clipped (off)

 Scoring:
  --ma <int>         match bonus (0 for --end-to-end, 2 for --local)
  --mp <int>         max penalty for mismatch; lower qual = lower penalt
y (6)
  --np <int>         penalty for non-A/C/G/Ts in read/ref (1)
  --rdg <int>,<int>  read gap open, extend penalties (5,3)
  --rfg <int>,<int>  reference gap open, extend penalties (5,3)
  --score-min <func> min acceptable alignment score w/r/t read length
                     (G,20,8 for local, L,-0.6,-0.6 for end-to-end)

 Reporting:
  (default)          look for multiple alignments, report best, with MAP
Q
   OR
  -k <int>           report up to <int> alns per read; MAPQ not meaningf
ul
   OR
  -a/--all           report all alignments; very slow, MAPQ not meaningf
ul

 Effort:
  -D <int>           give up extending after <int> failed extends in a r
ow (15)
  -R <int>           for reads w/ repetitive seeds, try <int> sets of se
eds (2)

 Paired-end:
  -I/--minins <int>  minimum fragment length (0)
  -X/--maxins <int>  maximum fragment length (500)
  --fr/--rf/--ff     -1, -2 mates align fw/rev, rev/fw, fw/fw (--fr)
  --no-mixed         suppress unpaired alignments for paired reads
  --no-discordant    suppress discordant alignments for paired reads
  --no-dovetail      not concordant when mates extend past each other
  --no-contain       not concordant when one mate alignment contains oth
er
  --no-overlap       not concordant when mates overlap at all

 Output:
```

```
  -t/--time           print wall-clock time taken by search phases
  --un <path>           write unpaired reads that didn't align to <path>
  --al <path>           write unpaired reads that aligned at least once
to <path>
  --un-conc <path>      write pairs that didn't align concordantly to <p
ath>
  --al-conc <path>      write pairs that aligned concordantly at least o
nce to <path>
  (Note: for --un, --al, --un-conc, or --al-conc, add '-gz' to the optio
n name, e.g.
  --un-gz <path>, to gzip compress output, or add '-bz2' to bzip2 compre
ss output.)
  --quiet             print nothing to stderr except serious errors
  --met-file <path>   send metrics to file at <path> (off)
  --met-stderr        send metrics to stderr (off)
  --met <int>         report internal counters & metrics every <int> secs
 (1)
  --no-head           supppress header lines, i.e. lines starting with @
  --no-sq             supppress @SQ header lines
  --rg-id <text>      set read group id, reflected in @RG line and RG:Z:
opt field
  --rg <text>         add <text> ("lab:value") to @RG line of SAM header.
                      Note: @RG line only printed when --rg-id is set.
  --omit-sec-seq      put '*' in SEQ and QUAL fields for secondary alignm
ents.

 Performance:
  -o/--offrate <int> override offrate of index; must be >= index's offra
te
  -p/--threads <int> number of alignment threads to launch (1)
  --reorder           force SAM output order to match order of input read
s
  --mm                use memory-mapped I/O for index; many 'bowtie's can
 share

 Other:
  --qc-filter         filter out reads that are bad according to QSEQ fil
ter
  --seed <int>        seed for random number generator (0)
  --non-deterministic seed rand. gen. arbitrarily instead of using read
attributes
  --version           print version information and quit
  -h/--help           print this usage message
```

In [113]:
```
bowtie2-align -x /mnt/nfs/data/resources/bowtie2/hg19 Input1.fastq -S In
put1.sam
bowtie2-align -x /mnt/nfs/data/resources/bowtie2/hg19 Input2.fastq -S In
put2.sam
bowtie2-align -x /mnt/nfs/data/resources/bowtie2/hg19 Pulldown1.fastq -S
 Pulldown1.sam
bowtie2-align -x /mnt/nfs/data/resources/bowtie2/hg19 Pulldown2.fastq -S
 Pulldown2.sam
#Commands ran in a detached screen
```

samtools was used to convert the .sam files obtained from bowtie2 to .bam files .sam files were deleted after all analyses were completed.

In [114]:
```
samtools view -S -b Pulldown1.sam > Pulldown1.bam
#-S : auto detect input file
```

```
#-b : output .bam file
```

In [115]:
```
samtools view -S -b Pulldown2.sam > Pulldown2.bam
```

In [133]:
```
samtools view -S -b Input1.sam > Input1.bam
```

In [134]:
```
samtools view -S -b Input2.sam > Input2.bam
```

In [ ]:
```
samtools sort -O bam -o Pulldown1.sorted.bam Pulldown1.bam
#sort -O output file type is .bam
#-o output file name
#Run in terminal using detached screen
```

```
[bam_sort_core] merging from 13 files...
```

In [ ]:
```
samtools sort -O bam -o Pulldown2.sorted.bam Pulldown2.bam
```

```
[bam_sort_core] merging from 13 files...
```

In [130]:
```
samtools index Pulldown1.sorted.bam
```

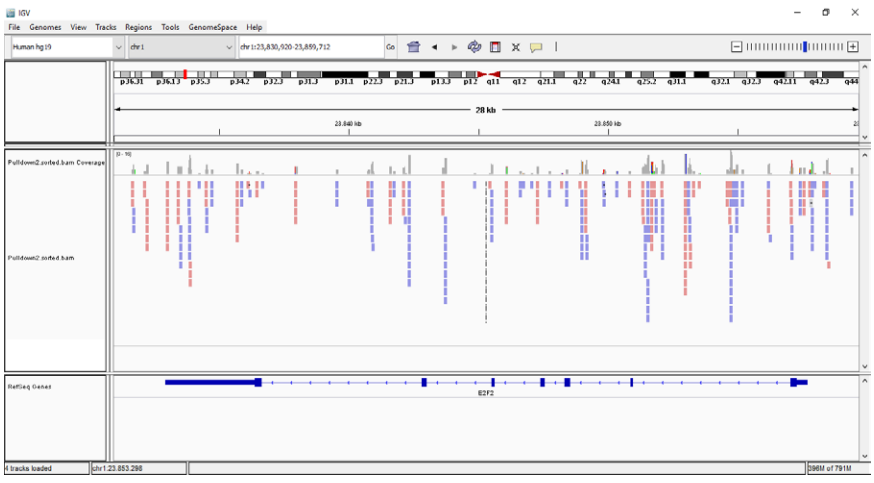
In [131]:
```
samtools index Pulldown2.sorted.bam
```

In [132]:
```
ls -l *.sorted.bam
```

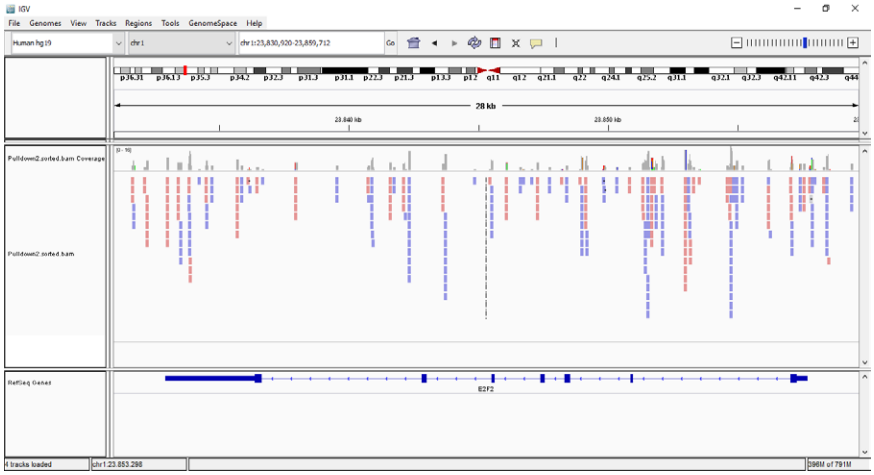
```
-rw-r--r-- 1 r0381150 domain users 1205411305 jan 17 13:51 Pulldown1.sor
ted.bam
-rw-r--r-- 1 r0381150 domain users 1391927222 jan 17 13:52 Pulldown2.sor
ted.bam
```

sorted.bam files were inspected in IGV.

Pulldown1.sorted.bam



Pulldown2.sorted.bam



MACS was performed to identify areas in the genome that were significantly enriched in RUNX1 binding sites.

In [121]:
```
mkdir MACS
```

In [135]:
```
macs2 callpeak -t Pulldown1.sorted.bam -c Input1.bam --outdir MACS -n Pu
lldown1 -g hs -q 0.005

#-t : specify treatment file
#-c : Control file
#-n : name of experiment
#-g : genome size, hs for default homo sapiens
#-q : q-value minimal false discovery rate cutoff for peak calling, here
 = 0.005
```

```
INFO  @ Thu, 19 Jan 2017 14:03:42:
# Command line: callpeak -t Pulldown1.sorted.bam -c Input1.bam --outdir
MACS -n Pulldown1 -g hs -q 0.005
# ARGUMENTS LIST:
# name = Pulldown1
# format = AUTO
# ChIP-seq file = ['Pulldown1.sorted.bam']
# control file = ['Input1.bam']
# effective genome size = 2.70e+09
# 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: 1000 bps and 10000 bps
# Broad region calling is off

INFO  @ Thu, 19 Jan 2017 14:03:42: #1 read tag files...
INFO  @ Thu, 19 Jan 2017 14:03:42: #1 read treatment tags...
INFO  @ Thu, 19 Jan 2017 14:03:42: Detected format is: BAM
INFO  @ Thu, 19 Jan 2017 14:03:42: * Input file is gzipped.
INFO  @ Thu, 19 Jan 2017 14:03:47:  1000000
INFO  @ Thu, 19 Jan 2017 14:03:52:  2000000
INFO  @ Thu, 19 Jan 2017 14:03:57:  3000000
INFO  @ Thu, 19 Jan 2017 14:04:01:  4000000
INFO  @ Thu, 19 Jan 2017 14:04:06:  5000000
INFO  @ Thu, 19 Jan 2017 14:04:11:  6000000
INFO  @ Thu, 19 Jan 2017 14:04:15:  7000000
INFO  @ Thu, 19 Jan 2017 14:04:20:  8000000
INFO  @ Thu, 19 Jan 2017 14:04:25:  9000000
```

```
INFO  @ Thu, 19 Jan 2017 14:04:29:  10000000
INFO  @ Thu, 19 Jan 2017 14:04:34:  11000000
INFO  @ Thu, 19 Jan 2017 14:04:39:  12000000
INFO  @ Thu, 19 Jan 2017 14:04:44:  13000000
INFO  @ Thu, 19 Jan 2017 14:04:48:  14000000
INFO  @ Thu, 19 Jan 2017 14:04:53:  15000000
INFO  @ Thu, 19 Jan 2017 14:04:58:  16000000
INFO  @ Thu, 19 Jan 2017 14:05:03:  17000000
INFO  @ Thu, 19 Jan 2017 14:05:07:  18000000
INFO  @ Thu, 19 Jan 2017 14:05:12:  19000000
INFO  @ Thu, 19 Jan 2017 14:05:17:  20000000
INFO  @ Thu, 19 Jan 2017 14:05:22:  21000000
INFO  @ Thu, 19 Jan 2017 14:05:26:  22000000
INFO  @ Thu, 19 Jan 2017 14:05:31:  23000000
INFO  @ Thu, 19 Jan 2017 14:05:36:  24000000
INFO  @ Thu, 19 Jan 2017 14:05:41:  25000000
INFO  @ Thu, 19 Jan 2017 14:05:45:  26000000
INFO  @ Thu, 19 Jan 2017 14:05:50:  27000000
INFO  @ Thu, 19 Jan 2017 14:05:54:  28000000
INFO  @ Thu, 19 Jan 2017 14:05:57:  29000000
INFO  @ Thu, 19 Jan 2017 14:06:00: #1.2 read input tags...
INFO  @ Thu, 19 Jan 2017 14:06:00: Detected format is: BAM
INFO  @ Thu, 19 Jan 2017 14:06:00: * Input file is gzipped.
INFO  @ Thu, 19 Jan 2017 14:06:05:  1000000
INFO  @ Thu, 19 Jan 2017 14:06:10:  2000000
INFO  @ Thu, 19 Jan 2017 14:06:16:  3000000
INFO  @ Thu, 19 Jan 2017 14:06:22:  4000000
INFO  @ Thu, 19 Jan 2017 14:06:27:  5000000
INFO  @ Thu, 19 Jan 2017 14:06:32:  6000000
INFO  @ Thu, 19 Jan 2017 14:06:38:  7000000
INFO  @ Thu, 19 Jan 2017 14:06:43:  8000000
INFO  @ Thu, 19 Jan 2017 14:06:48:  9000000
INFO  @ Thu, 19 Jan 2017 14:06:54:  10000000
INFO  @ Thu, 19 Jan 2017 14:07:00:  11000000
INFO  @ Thu, 19 Jan 2017 14:07:06:  12000000
INFO  @ Thu, 19 Jan 2017 14:07:11:  13000000
INFO  @ Thu, 19 Jan 2017 14:07:16:  14000000
INFO  @ Thu, 19 Jan 2017 14:07:22:  15000000
INFO  @ Thu, 19 Jan 2017 14:07:27:  16000000
INFO  @ Thu, 19 Jan 2017 14:07:33:  17000000
INFO  @ Thu, 19 Jan 2017 14:07:39:  18000000
INFO  @ Thu, 19 Jan 2017 14:07:44:  19000000
INFO  @ Thu, 19 Jan 2017 14:07:48:  20000000
INFO  @ Thu, 19 Jan 2017 14:07:53:  21000000
INFO  @ Thu, 19 Jan 2017 14:07:58:  22000000
INFO  @ Thu, 19 Jan 2017 14:08:04:  23000000
INFO  @ Thu, 19 Jan 2017 14:08:09:  24000000
INFO  @ Thu, 19 Jan 2017 14:08:15:  25000000
INFO  @ Thu, 19 Jan 2017 14:08:20:  26000000
INFO  @ Thu, 19 Jan 2017 14:08:26:  27000000
INFO  @ Thu, 19 Jan 2017 14:08:31:  28000000
INFO  @ Thu, 19 Jan 2017 14:08:37:  29000000
INFO  @ Thu, 19 Jan 2017 14:08:42:  30000000
INFO  @ Thu, 19 Jan 2017 14:08:44: #1 tag size is determined as 101 bps
INFO  @ Thu, 19 Jan 2017 14:08:44: #1 tag size = 101
INFO  @ Thu, 19 Jan 2017 14:08:44: #1  total tags in treatment: 27809590

INFO  @ Thu, 19 Jan 2017 14:08:44: #1 user defined the maximum tags...
INFO  @ Thu, 19 Jan 2017 14:08:44: #1 filter out redundant tags at the s
ame location and the same strand by allowing at most 1 tag(s)
INFO  @ Thu, 19 Jan 2017 14:08:45: #1  tags after filtering in treatment
```

```
: 2674426
INFO  @ Thu, 19 Jan 2017 14:08:45: #1  Redundant rate of treatment: 0.90

INFO  @ Thu, 19 Jan 2017 14:08:45: #1  total tags in control: 27848802
INFO  @ Thu, 19 Jan 2017 14:08:45: #1 user defined the maximum tags...
INFO  @ Thu, 19 Jan 2017 14:08:45: #1 filter out redundant tags at the s
ame location and the same strand by allowing at most 1 tag(s)
INFO  @ Thu, 19 Jan 2017 14:08:46: #1  tags after filtering in control:
25809618
INFO  @ Thu, 19 Jan 2017 14:08:46: #1  Redundant rate of control: 0.07
INFO  @ Thu, 19 Jan 2017 14:08:46: #1 finished!
INFO  @ Thu, 19 Jan 2017 14:08:46: #2 Build Peak Model...
INFO  @ Thu, 19 Jan 2017 14:08:46: #2 looking for paired plus/minus stra
nd peaks...
INFO  @ Thu, 19 Jan 2017 14:08:55: #2 number of paired peaks: 306939
INFO  @ Thu, 19 Jan 2017 14:08:55: start model_add_line...
INFO  @ Thu, 19 Jan 2017 14:08:55: start X-correlation...
INFO  @ Thu, 19 Jan 2017 14:08:55: end of X-cor
INFO  @ Thu, 19 Jan 2017 14:08:55: #2 finished!
INFO  @ Thu, 19 Jan 2017 14:08:55: #2 predicted fragment length is 177 b
ps
INFO  @ Thu, 19 Jan 2017 14:08:55: #2 alternative fragment length(s) may
 be 177 bps
INFO  @ Thu, 19 Jan 2017 14:08:55: #2.2 Generate R script for model : MA
CS/Pulldown1_model.r
WARNING @ Thu, 19 Jan 2017 14:08:55: #2 Since the d (177) calculated fro
m paired-peaks are smaller than 2*tag length, it may be influenced by un
known sequencing problem!
WARNING @ Thu, 19 Jan 2017 14:08:55: #2 You may need to consider one of
the other alternative d(s): 177
WARNING @ Thu, 19 Jan 2017 14:08:55: #2 You can restart the process with
 --nomodel --extsize XXX with your choice or an arbitrary number. Nonthe
less, MACS will continue computing.
INFO  @ Thu, 19 Jan 2017 14:08:55: #3 Call peaks...
INFO  @ Thu, 19 Jan 2017 14:08:55: #3 Pre-compute pvalue-qvalue table...

INFO  @ Thu, 19 Jan 2017 14:10:15: #3 Call peaks for each chromosome...
INFO  @ Thu, 19 Jan 2017 14:10:59: #4 Write output xls file... MACS/Pull
down1_peaks.xls
INFO  @ Thu, 19 Jan 2017 14:10:59: #4 Write peak in narrowPeak format fi
le... MACS/Pulldown1_peaks.narrowPeak
INFO  @ Thu, 19 Jan 2017 14:10:59: #4 Write summits bed file... MACS/Pul
ldown1_summits.bed
INFO  @ Thu, 19 Jan 2017 14:10:59: Done!
```

In [136]:
```
macs2 callpeak -t Pulldown2.sorted.bam -c Input2.bam --outdir MACS -n Pu
lldown2 -g hs -q 0.005
```

```
INFO  @ Thu, 19 Jan 2017 14:13:41:
# Command line: callpeak -t Pulldown2.sorted.bam -c Input2.bam --outdir
MACS -n Pulldown2 -g hs -q 0.005
# ARGUMENTS LIST:
# name = Pulldown2
# format = AUTO
# ChIP-seq file = ['Pulldown2.sorted.bam']
# control file = ['Input2.bam']
# effective genome size = 2.70e+09
# 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: 1000 bps and 10000 bps
# Broad region calling is off

INFO  @ Thu, 19 Jan 2017 14:13:41: #1 read tag files...
INFO  @ Thu, 19 Jan 2017 14:13:41: #1 read treatment tags...
INFO  @ Thu, 19 Jan 2017 14:13:41: Detected format is: BAM
INFO  @ Thu, 19 Jan 2017 14:13:41: * Input file is gzipped.
INFO  @ Thu, 19 Jan 2017 14:13:47:  1000000
INFO  @ Thu, 19 Jan 2017 14:13:52:  2000000
INFO  @ Thu, 19 Jan 2017 14:13:58:  3000000
INFO  @ Thu, 19 Jan 2017 14:14:03:  4000000
INFO  @ Thu, 19 Jan 2017 14:14:09:  5000000
INFO  @ Thu, 19 Jan 2017 14:14:14:  6000000
INFO  @ Thu, 19 Jan 2017 14:14:20:  7000000
INFO  @ Thu, 19 Jan 2017 14:14:25:  8000000
INFO  @ Thu, 19 Jan 2017 14:14:31:  9000000
INFO  @ Thu, 19 Jan 2017 14:14:36:  10000000
INFO  @ Thu, 19 Jan 2017 14:14:42:  11000000
INFO  @ Thu, 19 Jan 2017 14:14:47:  12000000
INFO  @ Thu, 19 Jan 2017 14:14:52:  13000000
INFO  @ Thu, 19 Jan 2017 14:14:58:  14000000
INFO  @ Thu, 19 Jan 2017 14:15:05:  15000000
INFO  @ Thu, 19 Jan 2017 14:15:10:  16000000
INFO  @ Thu, 19 Jan 2017 14:15:15:  17000000
INFO  @ Thu, 19 Jan 2017 14:15:20:  18000000
INFO  @ Thu, 19 Jan 2017 14:15:26:  19000000
INFO  @ Thu, 19 Jan 2017 14:15:31:  20000000
INFO  @ Thu, 19 Jan 2017 14:15:36:  21000000
INFO  @ Thu, 19 Jan 2017 14:15:41:  22000000
INFO  @ Thu, 19 Jan 2017 14:15:46:  23000000
INFO  @ Thu, 19 Jan 2017 14:15:51:  24000000
INFO  @ Thu, 19 Jan 2017 14:15:56:  25000000
INFO  @ Thu, 19 Jan 2017 14:16:01:  26000000
INFO  @ Thu, 19 Jan 2017 14:16:07:  27000000
INFO  @ Thu, 19 Jan 2017 14:16:15:  28000000
INFO  @ Thu, 19 Jan 2017 14:16:20:  29000000
INFO  @ Thu, 19 Jan 2017 14:16:25: #1.2 read input tags...
INFO  @ Thu, 19 Jan 2017 14:16:25: Detected format is: BAM
INFO  @ Thu, 19 Jan 2017 14:16:25: * Input file is gzipped.
INFO  @ Thu, 19 Jan 2017 14:16:30:  1000000
INFO  @ Thu, 19 Jan 2017 14:16:35:  2000000
INFO  @ Thu, 19 Jan 2017 14:16:41:  3000000
INFO  @ Thu, 19 Jan 2017 14:16:46:  4000000
INFO  @ Thu, 19 Jan 2017 14:16:52:  5000000
INFO  @ Thu, 19 Jan 2017 14:16:57:  6000000
INFO  @ Thu, 19 Jan 2017 14:17:02:  7000000
INFO  @ Thu, 19 Jan 2017 14:17:08:  8000000
INFO  @ Thu, 19 Jan 2017 14:17:13:  9000000
INFO  @ Thu, 19 Jan 2017 14:17:18:  10000000
INFO  @ Thu, 19 Jan 2017 14:17:24: #1 tag size is determined as 101 bps
INFO  @ Thu, 19 Jan 2017 14:17:24: #1 tag size = 101
INFO  @ Thu, 19 Jan 2017 14:17:24: #1  total tags in treatment: 28187369

INFO  @ Thu, 19 Jan 2017 14:17:24: #1 user defined the maximum tags...
INFO  @ Thu, 19 Jan 2017 14:17:24: #1 filter out redundant tags at the s
ame location and the same strand by allowing at most 1 tag(s)
INFO  @ Thu, 19 Jan 2017 14:17:24: #1  tags after filtering in treatment
: 7203889
INFO  @ Thu, 19 Jan 2017 14:17:24: #1  Redundant rate of treatment: 0.74

INFO  @ Thu, 19 Jan 2017 14:17:24: #1  total tags in control: 9872550
```

```
INFO  @ Thu, 19 Jan 2017 14:17:24: #1 user defined the maximum tags...
INFO  @ Thu, 19 Jan 2017 14:17:24: #1 filter out redundant tags at the s
ame location and the same strand by allowing at most 1 tag(s)
INFO  @ Thu, 19 Jan 2017 14:17:25: #1  tags after filtering in control:
8946186
INFO  @ Thu, 19 Jan 2017 14:17:25: #1  Redundant rate of control: 0.09
INFO  @ Thu, 19 Jan 2017 14:17:25: #1 finished!
INFO  @ Thu, 19 Jan 2017 14:17:25: #2 Build Peak Model...
INFO  @ Thu, 19 Jan 2017 14:17:25: #2 looking for paired plus/minus stra
nd peaks...
INFO  @ Thu, 19 Jan 2017 14:17:28: #2 number of paired peaks: 67064
INFO  @ Thu, 19 Jan 2017 14:17:28: start model_add_line...
INFO  @ Thu, 19 Jan 2017 14:17:28: start X-correlation...
INFO  @ Thu, 19 Jan 2017 14:17:28: end of X-cor
INFO  @ Thu, 19 Jan 2017 14:17:28: #2 finished!
INFO  @ Thu, 19 Jan 2017 14:17:28: #2 predicted fragment length is 173 b
ps
INFO  @ Thu, 19 Jan 2017 14:17:28: #2 alternative fragment length(s) may
 be 173 bps
INFO  @ Thu, 19 Jan 2017 14:17:28: #2.2 Generate R script for model : MA
CS/Pulldown2_model.r
WARNING @ Thu, 19 Jan 2017 14:17:28: #2 Since the d (173) calculated fro
m paired-peaks are smaller than 2*tag length, it may be influenced by un
known sequencing problem!
WARNING @ Thu, 19 Jan 2017 14:17:28: #2 You may need to consider one of
the other alternative d(s): 173
WARNING @ Thu, 19 Jan 2017 14:17:28: #2 You can restart the process with
 --nomodel --extsize XXX with your choice or an arbitrary number. Nonthe
less, MACS will continue computing.
INFO  @ Thu, 19 Jan 2017 14:17:28: #3 Call peaks...
INFO  @ Thu, 19 Jan 2017 14:17:28: #3 Pre-compute pvalue-qvalue table...

INFO  @ Thu, 19 Jan 2017 14:18:14: #3 Call peaks for each chromosome...
INFO  @ Thu, 19 Jan 2017 14:18:39: #4 Write output xls file... MACS/Pull
down2_peaks.xls
INFO  @ Thu, 19 Jan 2017 14:18:39: #4 Write peak in narrowPeak format fi
le... MACS/Pulldown2_peaks.narrowPeak
INFO  @ Thu, 19 Jan 2017 14:18:39: #4 Write summits bed file... MACS/Pul
ldown2_summits.bed
INFO  @ Thu, 19 Jan 2017 14:18:39: Done!
```

In [137]:
```
ls -l MACS/
```

```
total 2728
-rw-r--r-- 1 r0381150 domain users  78213 jan 19 14:08 Pulldown1_model.r
-rw-r--r-- 1 r0381150 domain users 330074 jan 19 14:10 Pulldown1_peaks.n
arrowPeak
-rw-r--r-- 1 r0381150 domain users 372040 jan 19 14:10 Pulldown1_peaks.x
ls
-rw-r--r-- 1 r0381150 domain users 220810 jan 19 14:10 Pulldown1_summits
.bed
-rw-r--r-- 1 r0381150 domain users  78073 jan 19 14:17 Pulldown2_model.r
-rw-r--r-- 1 r0381150 domain users 607502 jan 19 14:18 Pulldown2_peaks.n
arrowPeak
-rw-r--r-- 1 r0381150 domain users 685023 jan 19 14:18 Pulldown2_peaks.x
ls
-rw-r--r-- 1 r0381150 domain users 404459 jan 19 14:18 Pulldown2_summits
.bed
```

In [141]:
```
head MACS/Pulldown1_peaks.narrowPeak | column -t
```

```
chr1  10134   10398   Pulldown1_peak_1   40   .  4.24468   7.65429   4.0
9239   32
chr1  29106   29463   Pulldown1_peak_2   109  .  8.09518   15.18149  10.
95348  190
chr1  441493  441749  Pulldown1_peak_3   218  .  14.32239  26.54005  21.
81782  143
chr1  459858  460035  Pulldown1_peak_4   66   .  7.00779   10.58695  6.6
9987   68
chr1  522829  523006  Pulldown1_peak_5   78   .  7.45501   11.79238  7.8
0072   81
chr1  523464  523659  Pulldown1_peak_6   88   .  8.07472   12.88016  8.8
0794   66
chr1  541154  541422  Pulldown1_peak_7   221  .  14.48395  26.83014  22.
10177  114
chr1  706679  706856  Pulldown1_peak_8   46   .  5.92844   8.33201   4.6
9246   119
chr1  762645  762851  Pulldown1_peak_9   63   .  6.86570   10.24796  6.3
8752   138
chr1  912096  912415  Pulldown1_peak_10  189  .  12.76244  23.53026  18.
91455  152
```

In [140]:
```
head MACS/Pulldown2_peaks.narrowPeak | column -t
```

```
chr1  10221   10430   Pulldown2_peak_1   96   .  6.50372   13.32020  9.6
0609   109
chr1  11834   12040   Pulldown2_peak_2   60   .  6.43098   9.51633   6.0
6998   81
chr1  28894   29868   Pulldown2_peak_3   355  .  19.49479  40.19792  35.
57613  364
chr1  384566  384777  Pulldown2_peak_4   53   .  6.15771   8.76155   5.3
6702   72
chr1  441448  441734  Pulldown2_peak_5   259  .  15.73637  30.32679  25.
94702  138
chr1  459716  460122  Pulldown2_peak_6   123  .  9.50905   16.25258  12.
38234  252
chr1  460468  460692  Pulldown2_peak_7   151  .  10.52488  19.20435  15.
19797  118
chr1  522744  522991  Pulldown2_peak_8   138  .  9.97093   17.83057  13.
88195  102
chr1  523493  523754  Pulldown2_peak_9   179  .  12.31542  22.03965  17.
92287  131
chr1  533625  533907  Pulldown2_peak_10  136  .  9.80474   17.55158  13.
62280  157
```

In [142]:
```
wc -l MACS/Pulldown1_peaks.narrowPeak
```

```
4228 MACS/Pulldown1_peaks.narrowPeak
```

In [150]:
```
wc -l MACS/Pulldown2_peaks.narrowPeak
```

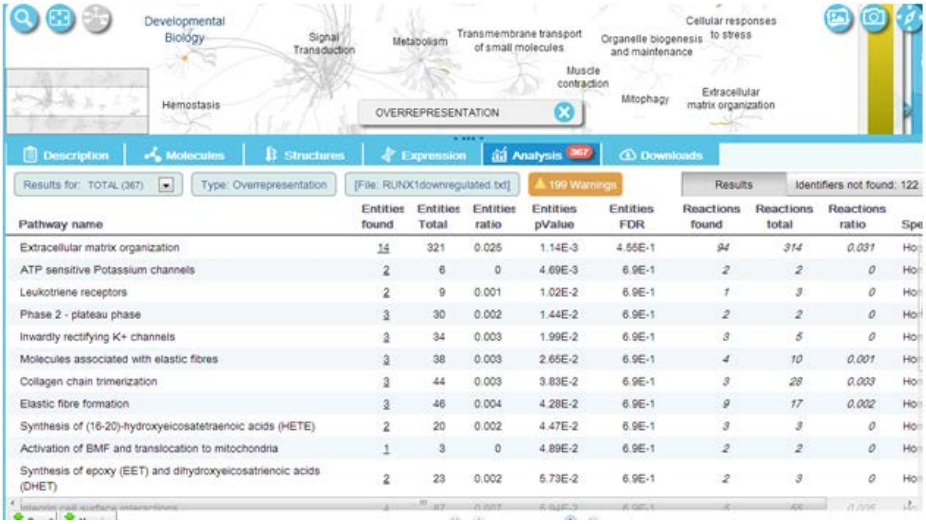
```
7726 MACS/Pulldown2_peaks.narrowPeak
```

RUNX1 is a transcription factor that has been frequently implicated in cancer development. To investigate which genes are transcriptionally regulated by this transcription factor, we analyzed RNA- and ChIP-Seq data from the previously mentioned study. RNAseq analysis was performed to identify genes that were differentially expressed in a MCF-7 breast cancer cell in which RUNX1 was knocked down with a short hairpin RNA against RUNX1. MCF-7 cells expressing a non-specific shRNA was used as control cell line. We obtained a list of 200 most significantly up- and downregulated genes. REACTOME pathway analysis was performed to identify pathways in which

these genes are involved.

200 most down- and upregulated genes can be found in a file attachment.

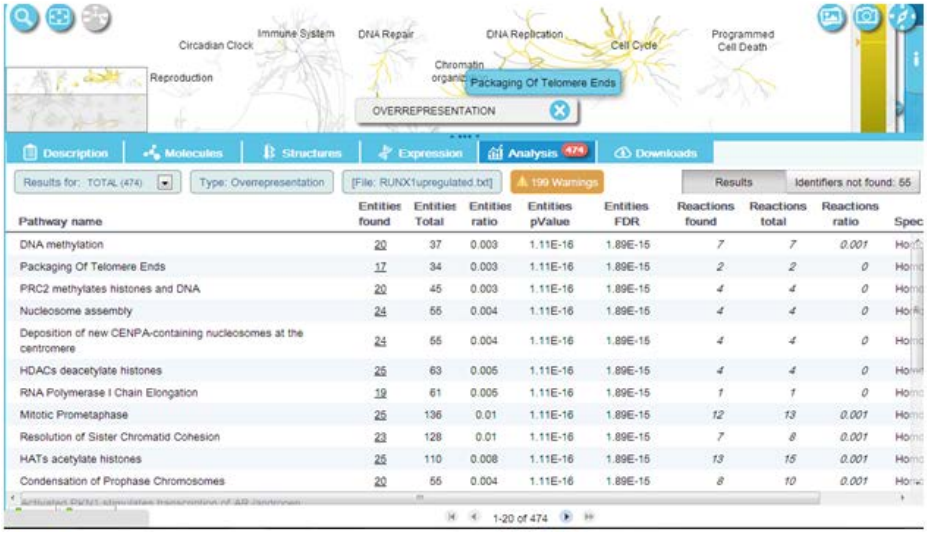
REACTOME downregulated genes



Pathway name	Entities found	Entities Total	Entities ratio	Entities pValue	Entities FDR	Reactions found	Reactions total	Reactions ratio	Spe
Extracellular matrix organization	14	321	0.025	1.14E-3	4.55E-1	94	314	0.031	Ho
ATP sensitive Potassium channels	2	6	0	4.69E-3	6.9E-1	2	2	0	Ho
Leukotriene receptors	2	9	0.001	1.02E-2	6.9E-1	1	3	0	Ho
Phase 2 - plateau phase	3	30	0.002	1.44E-2	6.9E-1	2	2	0	Ho
Inwardly rectifying K+ channels	3	34	0.003	1.99E-2	6.9E-1	3	5	0	Ho
Molecules associated with elastic fibres	3	38	0.003	2.65E-2	6.9E-1	4	10	0.001	Ho
Collagen chain trimerization	3	44	0.003	3.83E-2	6.9E-1	3	28	0.003	Ho
Elastic fibre formation	3	46	0.004	4.28E-2	6.9E-1	9	17	0.002	Ho
Synthesis of (16-20)-hydroxyeicosatetraenoic acids (HETE)	2	20	0.002	4.47E-2	6.9E-1	3	3	0	Ho
Activation of BMF and translocation to mitochondria	1	3	0	4.89E-2	6.9E-1	2	2	0	Ho
Synthesis of epoxy (EET) and dihydroxyeicosatrienoic acids (DHET)	2	23	0.002	5.73E-2	6.9E-1	2	3	0	Ho

Genes that were downregulated upon RUNX1 knockdown turned out to be most frequently implicated in extracellular matrix organization. (For example, genes encoding for ECM cell-adhesion molecules LRFN5 and FBN2). Other pathways we find enriched among downregulated genes are for instance ATP-sensitive Potassium channels and Leukotriene receptors. In accordance to previous observations we also find that RUNX1 knockdown results in downregulation of the long non-coding RNAs NEAT1 and MALAT1.

REACTOME upregulated genes



Pathway name	Entities found	Entities Total	Entities ratio	Entities pValue	Entities FDR	Reactions found	Reactions total	Reactions ratio	Spec
DNA methylation	20	37	0.003	1.11E-16	1.89E-15	7	7	0.001	Ho
Packaging Of Telomere Ends	17	34	0.003	1.11E-16	1.89E-15	2	2	0	Ho
PRC2 methylates histones and DNA	20	45	0.003	1.11E-16	1.89E-15	4	4	0	Ho
Nucleosome assembly	24	55	0.004	1.11E-16	1.89E-15	4	4	0	Ho
Deposition of new CENPA-containing nucleosomes at the centromere	24	55	0.004	1.11E-16	1.89E-15	4	4	0	Ho
HDACs deacetylate histones	25	63	0.005	1.11E-16	1.89E-15	4	4	0	Ho
RNA Polymerase I Chain Elongation	19	61	0.005	1.11E-16	1.89E-15	1	1	0	Ho
Mitotic Prometaphase	25	136	0.01	1.11E-16	1.89E-15	12	13	0.001	Ho
Resolution of Sister Chromatid Cohesion	23	128	0.01	1.11E-16	1.89E-15	7	8	0.001	Ho
HATs acetylate histones	25	110	0.008	1.11E-16	1.89E-15	13	15	0.001	Ho
Condensation of Prophase Chromosomes	20	55	0.004	1.11E-16	1.89E-15	8	10	0.001	Ho

1-20 of 474

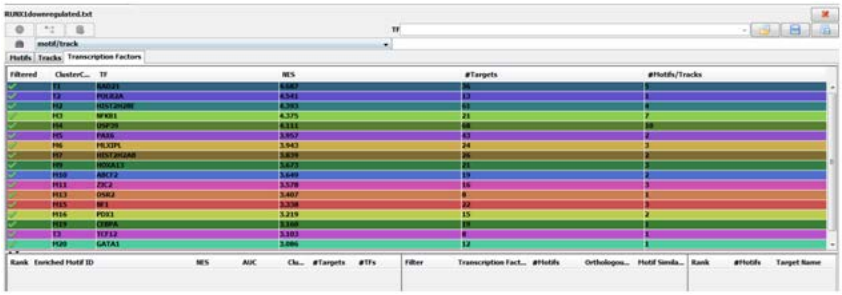
shRNA mediated knockdown of RUNX1 resulted mainly in upregulation of genes involved in chromatin remodeling. Indeed, among the most significantly enriched pathways were DNA methylation and packaging of telomere ends.

Interestingly many of the identified pathways, such as ECM organization or chromatin remodeling, are known to be involved in cancer development. ECM disruption is one of the necessary steps for tissue invasion, which favours metastasis. Disruption of genomic maintenance leads to genomic

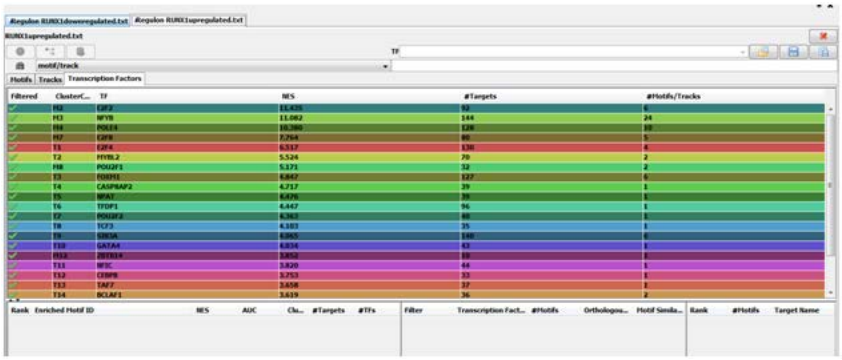
instability, leading to additional mutations. The accumulation of mutations can cause cancer progression.

The top 200 most down and upregulated genes were analyzed with iRegulon. This tool searches enriched transcription factor motifs 20kb around the transcription start site (TSS) and directly links these motifs to corresponding transcription factors. The final output is a list of enriched motifs alongside with candidate transcription factors and possible target genes. In contrast to our expectations, the RUNX1 motif was not identified as significantly enriched. We hypothesize RUNX1 might not directly down or upregulate the identified genes, but instead their expression might be regulated by transcription factors downstream of RUNX1. This will be further investigated during the ChIP-Seq analysis. Another possible explanation could be that iRegulon searches for targets within 20kb of the TSS, while RUNX1 has been shown to frequently bind intergenic regions and thus the corresponding motif might not be enriched within this interval.

iRegulon Transcription factors identified from downregulated genes



iRegulon Transcription factors identified from upregulated genes

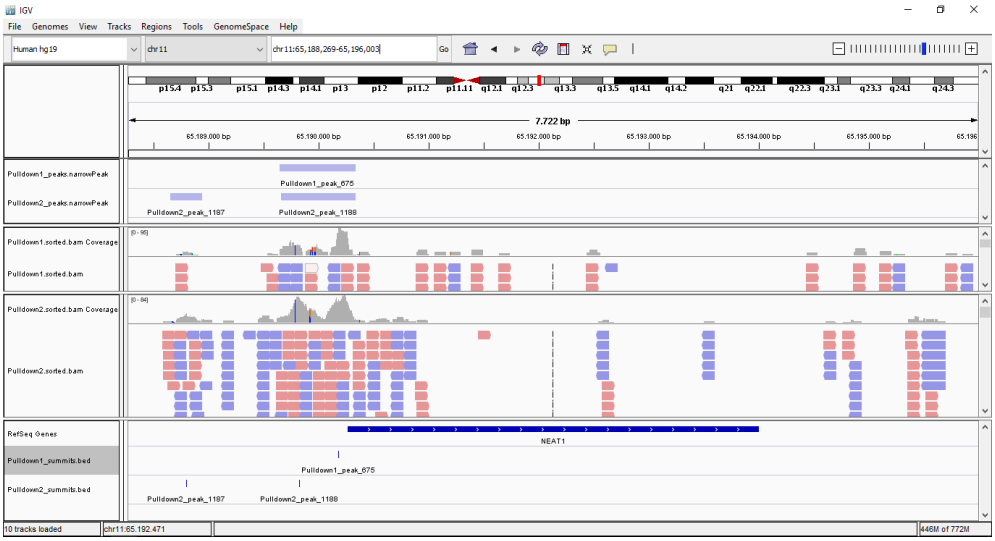


To identify which of these genes were directly regulated by the RUNX1 transcription factor we performed ChIP-Seq analysis. This allows us to localize RUNX1 binding sites. Firstly, we visually checked the aligned ChIP-Seq reads and the called peaks in IGV. We previously observed that RUNX1 knockdown results in downregulation of MALAT1 and NEAT. To check whether RUNX1 physically interacts with the promotor region of these genes we visualized the corresponding ChIP-Seq reads in IGV.

IGV MALAT1 gene from Pulldown 1 and 2



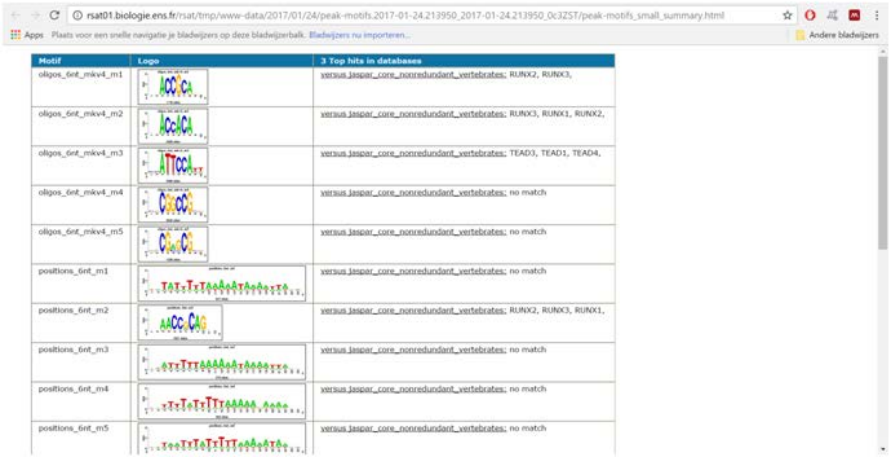
IGV NEAT1 gene from Pulldown 1 and 2



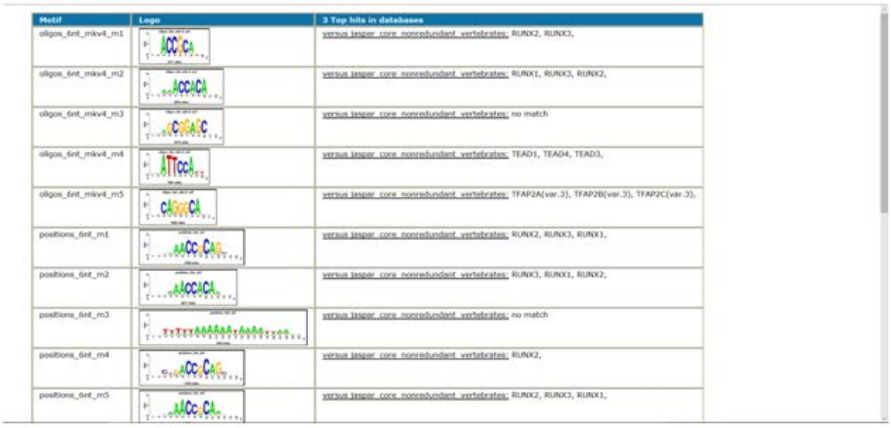
Binding sites for RUNX1 were predicted in promotor regions of both non-coding RNA genes. Since expression of these genes was reduced upon knockdown of RUNX1, these data suggest that transcription of these genes is directly activated by RUNX1 binding. However, further inspection of ChIP-Seq peaks showed that RUNX1 binding peaks were absent in most of the genes that were identified as differentially expressed in the RNA-seq analysis. This further supports the hypothesis that RUNX1 mainly acts through downstream transcription factors.

Finally, de novo motif analysis was performed on the RUNX1 peaks to find significantly enriched motifs within the DNA regions that were pulled down during the ChIP-Seq assay. Herefore we used the analysis tool RSAT.

RSAT Pulldown 1 motifs

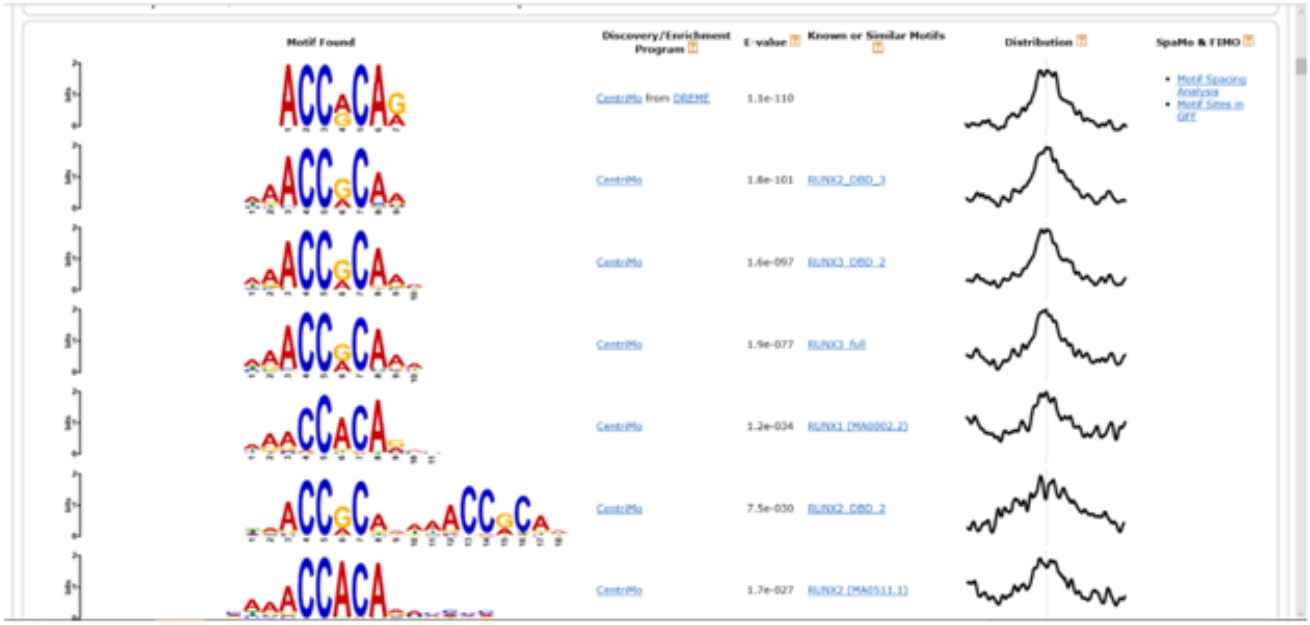


RSAT Pulldown 2 motifs



According to our expectations, the sequence of the RUNX1 binding site was identified as one of the most significantly enriched motifs, along with the RUNX family members RUNX2 and 3. This confirms that during ChIP-Seq pulldown the correct transcription factor was isolated. The analysis was repeated using MEME, an alternative motif analysis tool, to validate previous results. As expected, the RUNX binding motif was again identified as most significantly enriched.

MEME Pulldown 1 motifs



MEME Pulldown 2 motif

