Bioinformatics - Main report

Students:

- Marieke Bamps r0618772
- Lara Hemeryck r0638328
- Hanne Hillen r0637408

The following data analysis (RNA-seq and ChIP-seq) is based on the following article. **RUNX1 contributes to higher-order chromatin organization and gene regulation in breast cancer cells**

Here (https://www.ncbi.nlm.nih.gov/pubmed/27514584), you can find the link to the pubmed page of the article (doi: 10.1016/j.bbagrm.2016.08.003)

Introduction and aim of the study

RUNX is a family of mammalian transcription factors consisting of three members, RUNX1, RUNX2 and RUNX3. They each have tissue-specific expression patterns and cell-context dependent functions through interactions with their binding factors and each have another physiological function. The RUNX factors play a key role in different types of cancer.

The article focuses on the contributions of RUNX1, which is involved in hematopoiesis and in the development in breast cancer. It functions both as a tumor suppressor and an oncogene, and is thus shown to be both down-regulated and upregulated in breast cancer. Whole genome sequencing has shown deletions and point mutations in the RUNX1 gene in human breast tumors.

RUNX1 activates and represses target genes expression of different genes depending on its target partners.

Cancer is characterized with changes gene expressions, mostly alterations in the folding of the genome. In breast cancer, alterations in the nuclear architecture are seen, such as alterations in the chromatin and at telomeres when comparing with epithelial cells. Through the importance of the higher-order genome folding and the involvement of RUNX1 in breast cancer, it is necessary to understand the relation between the function of RUNX1 in chromatin organization and its regulatory role in breast cancer.

The aim of this study is to investigate the relationship between RUNX1 mediated transcription and genome organisation through the comparison of control and RUNX1-depleted MCF-7 breast cancer cells. In order to do this RNA-seq and ChIP-seq were performed.

- RNAseq: suppression of RUNX1 in a human MCF-7 breast cancer cell line and evaluation of the gene expression through mRNA levels at a given moment in the cell. Additionally, a functional analysis was preformed using DESeq.
- ChIP-seq: performed to gain insight into RUNX1-mediated regulation of chromatin organization. RUNX1 binding was probed by performing ChIP-seq.

RNA-seq analysis

The RNA-seq analysis reveals the presence and quantity of RNA in a biological sample at a given moment in time. Four samples are compared: shNSR1, shNSR2, shRUNR1 and shRUNR2. The expression of genes in the RNA-seq analysis is quantified by counting the number of reads that mapped to each locus in the transcriptome assembly step.

Subsequently, these results can be used in order to perform a functional analysis of the gene counts searching for differenctial gene expression between the different samples

Step 1: Make a new directory (mkdir command)

cd command (change directory) is used to go to another directory

pwd command displays where you are (in which directory)

ls command can be used to display the files present in the current directory

In [1]:
```
cd /home/luna.kuleuven.be/r0637408/data
pwd
```
```
/home/luna.kuleuven.be/r0637408/data
```

In [2]:
```
mkdir -p ~/data/work
cd ~/data/work
pwd
ls
```
```
/home/luna.kuleuven.be/r0637408/data/work
~              shNSR1._STARtmp  shNSR2._STARtmp  shRUNR2.fastq
shNSR1.fastq  shNSR2.fastq     shRUNR1.fastq
```

Step 2: Extract raw data

The raw data of this experiment are available on the Gene Expression Omnibus (GEO) website under the accession number GSE75070. (link) (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE75070)

The subtitle 'Samples' indicates that 3 replicates are analysed: Replicate 1,2 and 3. For the analysis in this report, we will compare the raw data of replicate 1 and 2. Each sample has a specific GSM accession code and clicking on it, it redirects you to the page specific for that sample.

- MCF-7 shNS RNA-seq

```
SSR2919468, named shNSR1
SSR2919469, named shNSR2
```

- MCF-7 shRUNX1 RNA-seq

```
SSR2919471, named shRUNR1
SSR2919472, named shRUNR2
```

On this webpage you click on the SRA = Sequence Read Archive relation link which will lead you to the next page. On this page the SRR number is indicated which can be used to convert this file to a fastq file. First we download the datafiles with the following commando's (in SSH, bitvise):

mkdir ~/data/ncbi

vdb-config -s /repository/user/main/public/root=~/data/ncbi

In order to do the conversion we use the sra toolkit with the following commando:

fastq-dump --split-files SRRnumber

Now we have fastq files in our work directory.

Step 3: Quality control

Check the quality of the raw data by using the FastQC toolkit.

General commando: fastqc fastqfile -o outputdir

- **shNSR1**

In [16]:
```
/usr/local/bin/FastQC/fastqc -o ~/data shNSR1.fastq
```

```
Started analysis of shNSR1.fastq
Approx 5% complete for shNSR1.fastq
Approx 10% complete for shNSR1.fastq
Approx 15% complete for shNSR1.fastq
Approx 20% complete for shNSR1.fastq
Approx 25% complete for shNSR1.fastq
Approx 30% complete for shNSR1.fastq
Approx 35% complete for shNSR1.fastq
Approx 40% complete for shNSR1.fastq
Approx 45% complete for shNSR1.fastq
Approx 50% complete for shNSR1.fastq
Approx 55% complete for shNSR1.fastq
Approx 60% complete for shNSR1.fastq
Approx 65% complete for shNSR1.fastq
Approx 70% complete for shNSR1.fastq
Approx 75% complete for shNSR1.fastq
Approx 80% complete for shNSR1.fastq
Approx 85% complete for shNSR1.fastq
Approx 90% complete for shNSR1.fastq
Approx 95% complete for shNSR1.fastq
Analysis complete for shNSR1.fastq
```

This analysis generates a html file in which the results are represented.

In [5]:
```
cd ~/data/work
```

```
/mnt/storage/r0637408/work
```

In [6]:
```
ls *html
```

```
shNSR1_fastqc.html  shRUNX1_R1_fastqc.html
```

Analysis of the Quality control results shNSR1:

Sequences flagged as poor quality: 0

- Per base sequence quality report: This report shows a boxplot for eac h position in the sequence. A boxplox consists of the following elements: T he central red line is the median value, the yellow box represents the inte r-quartile range (25-75%), the upper and lower whiskers represent the 10% a nd 90% points and the blue line represents the mean quality.

The y-axis shows the quality scores. The higher the score, the better t he base call. The background of the graph divides the y axis into very good quality calls (green), calls of reasonable quality (orange), and calls of p oor quality (red). In this analysis the quality scores of the bases stay ve ry high, which indicates a good base calling.

Per base sequence content plots out the proportion of each base positi on in a file for which each of the four normal DNA bases has been called. I n a random library you would expect that the proportions of the bases are e qual. Also this analysis shows a good base calling. However, in many cases the results will point this out as 'not good' because of the presence of th e adaptor sequences. These should be deleted, but this is not always workin g very well.

- Sequences duplication levels report: The first 100 000 sequences are analysed and are representative for the duplication levels of the entire fi le. In a diverse library most sequences will occur only once in the final s et. A low level of duplication may indicate a very high level of coverage o f the target sequence, but a high level of duplication is more likely to in dicate some kind of enrichment bias (technical duplicates due to PCR artefa cts or biological duplicates.

This module counts the degree of duplication for every sequence in a li brary and creates a plot showing the relative number of sequences with diff erent degrees of duplication.

Even so, longer reads are more likely to contain sequencing errors, whi ch will artificially increase the observed diversity and will tend to under represent highly duplicated sequences.

If it is a diverse library, most of the sequences will fall into the fa r left of the plot. The presence of spikes towards the right of the plot (w hich is the case here) indicate more specific enirchments or the presence o f contaminants. This parameter will warn you if non-unique sequences make u p of more than 20% of the total. (A failure indication is showed when the s equencs make up more than 50% of the total.

- Overrepresented sequences report: A single sequence very overrepresen ted in the set means that it is highly biologically significant or that it means that the library is contaminated. The graph shows these sequences tha t make up more than 0,1% of the total. The analysis program searches for ma tches in a database of common contaminants and will report the best hit it finds. Important to mention is that adapter sequences are very similar to e ach other so you may get a hit reported that is not technically correct, bu t which has very similar sequence to the actual match.

In our data set, there are several overrepresented sequences but for no n of them a possible source is found. Except for one that is a TruSeq adapt er like pointed out above.

Longer reads are more likely to contain sequencing errors, that will ar tificially increase the observed diversity and will tend to underrepresent highly duplicated sequences.

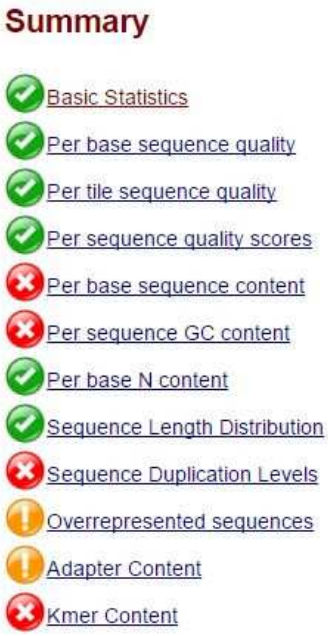

Figure 1: Summary of the quality control results of shNSR1

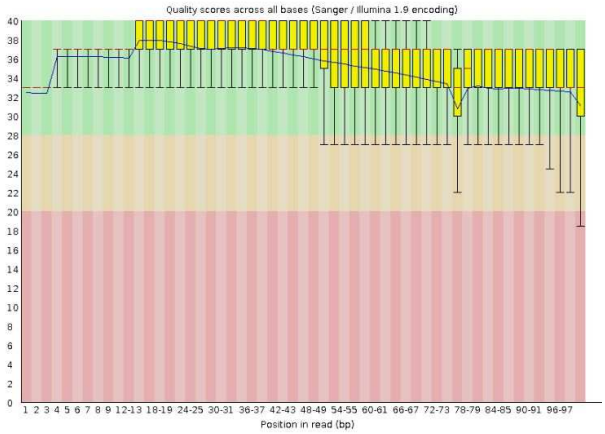

Figure 2: Per base sequence quality of shNSR1

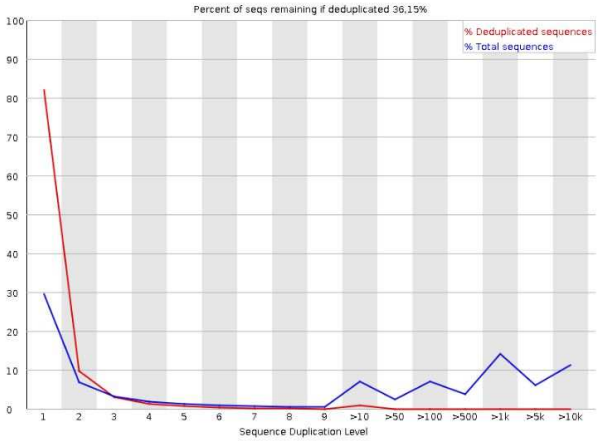

Figure 3: Sequence duplication levels of shNSR1

- **shNSR2**

In [4]:
```
/usr/local/bin/FastQC/fastqc -o ./data shNSR2.fastq
```

```
Started analysis of shNSR2.fastq
Approx 5% complete for shNSR2.fastq
Approx 10% complete for shNSR2.fastq
Approx 15% complete for shNSR2.fastq
Approx 20% complete for shNSR2.fastq
Approx 25% complete for shNSR2.fastq
Approx 30% complete for shNSR2.fastq
Approx 35% complete for shNSR2.fastq
Approx 40% complete for shNSR2.fastq
Approx 45% complete for shNSR2.fastq
Approx 50% complete for shNSR2.fastq
Approx 55% complete for shNSR2.fastq
Approx 60% complete for shNSR2.fastq
Approx 65% complete for shNSR2.fastq
Approx 70% complete for shNSR2.fastq
Approx 75% complete for shNSR2.fastq
Approx 80% complete for shNSR2.fastq
Approx 85% complete for shNSR2.fastq
Approx 90% complete for shNSR2.fastq
Approx 95% complete for shNSR2.fastq
Analysis complete for shNSR2.fastq
```

In [38]:
```
ls *html
```

```
shNSR1_fastqc.html  shRUNX1_R1_fastqc.html
shNSR2_fastqc.html  shRUNX1_R2_fastqc.html
```

Analysis of the Quality control results shNSR2:

Sequences flagged as poor quality: 0

- Per base sequence quality report: The higher the score the better the base call. In this analysis, the quality scores of the bases stay very high , which indicates a good base calling.

Per base sequence content plots out the proportion of each base positi on in a file of which each of the four normal DNA bases has been called. Th is library shows equal proportions of the bases as expected. Also this anal ysis shows a good base calling.

- Sequences duplication levels report: The trend of the curve shows a p eak on the right side of the curve, which implements more specific enrichme nts or the presence of contaminants.

- Overrepresented sequences report: In our data set, there are several overrepresented sequences but for non of them a possible source is found.
The warning means that one sequence is found that represents more than 0,1% of the total.

In the figures below you can see some selected results of the quality control.

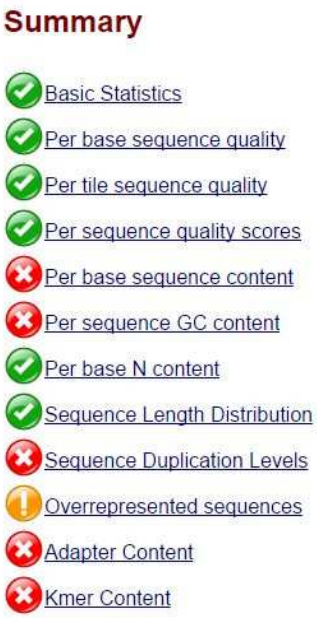

Figure 4: Summary of the quality control results of shNSR2

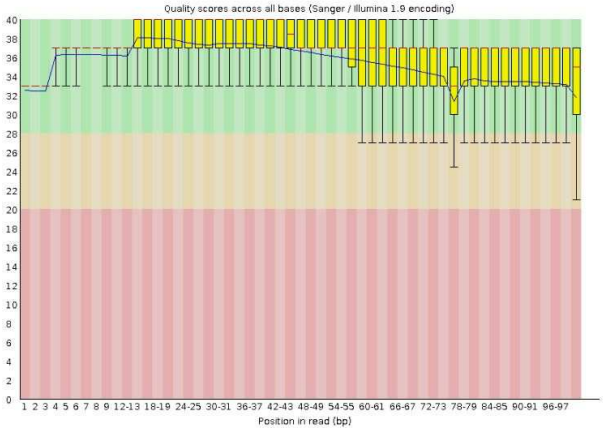

Figure 5: Per base sequence quality of shNSR2

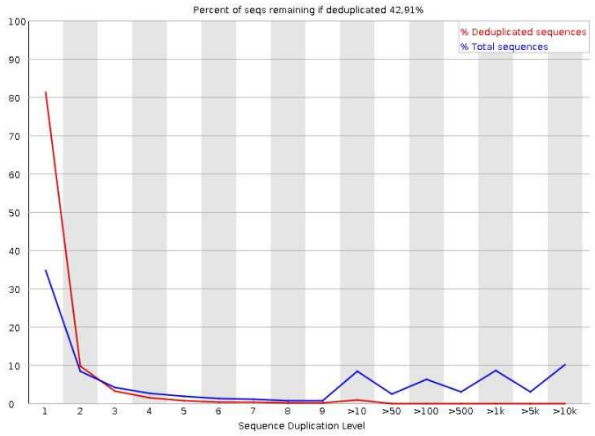

Figure 6: Sequence duplication levels of shNSR2

- **shRUNX1_R1**

In [20]:
```
/usr/local/bin/FastQC/fastqc -o ./data shRUNX1_R1.fastq
```

```
Started analysis of shRUNX1_R1.fastq
Approx 5% complete for shRUNX1_R1.fastq
Approx 10% complete for shRUNX1_R1.fastq
Approx 15% complete for shRUNX1_R1.fastq
Approx 20% complete for shRUNX1_R1.fastq
Approx 25% complete for shRUNX1_R1.fastq
Approx 30% complete for shRUNX1_R1.fastq
Approx 35% complete for shRUNX1_R1.fastq
Approx 40% complete for shRUNX1_R1.fastq
Approx 45% complete for shRUNX1_R1.fastq
Approx 50% complete for shRUNX1_R1.fastq
Approx 55% complete for shRUNX1_R1.fastq
Approx 60% complete for shRUNX1_R1.fastq
Approx 65% complete for shRUNX1_R1.fastq
Approx 70% complete for shRUNX1_R1.fastq
Approx 75% complete for shRUNX1_R1.fastq
Approx 80% complete for shRUNX1_R1.fastq
Approx 85% complete for shRUNX1_R1.fastq
Approx 90% complete for shRUNX1_R1.fastq
Approx 95% complete for shRUNX1_R1.fastq
Analysis complete for shRUNX1_R1.fastq
```

In [39]:
```
ls *html
```

```
shNSR1_fastqc.html  shRUNX1_R1_fastqc.html
shNSR2_fastqc.html  shRUNX1_R2_fastqc.html
```

Analysis of the Quality control results shRUNX1:

Sequences flagged as poor quality: 0

- Per base sequence quality report: The score of the base calling decreases a little bit along the sequence, however the scores still remain in the green section of the y-axis, which means that the base calling is of good quality.

- Sequences duplication levels report: The trend of the curve is similar to the previous one with peak on the right side of the curve. This might implicate more specific enirchments or the presence of contaminants. The warning denotes that non-unique sequences make up of more than 20% of the total.

- Overrepresented sequences report: A single sequence very overrepresented in the set means that it is highly biologically significant or that it means that the library is contaminated.
  In our data set, there are several overrepresented sequences but for non of them a possible source is found. Except for one that is a TruSeq adapter, which can happen as pointed out above. The warning means that one sequence is found that represents more than 0,1% of the total.

In the figures below you can see some selected results of the quality control.

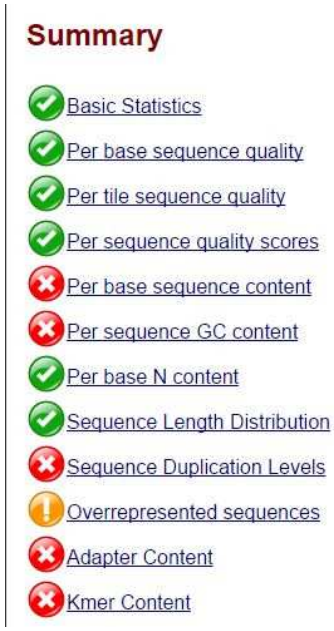

Figure 7: Summary of the quality control results of shRUNX1

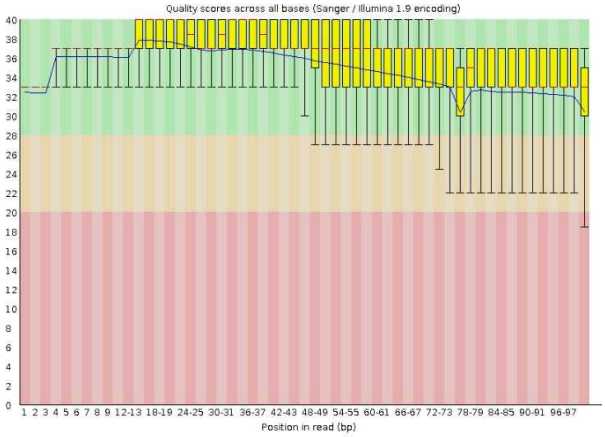

Figure 8: Per base sequence qualityof shRUNX1

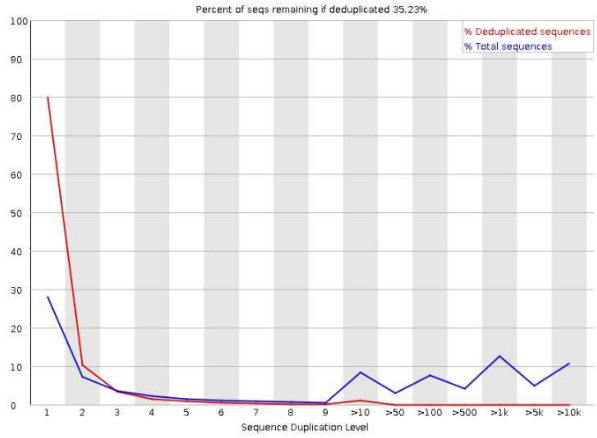

Figure 9: Sequence duplication levels of shRUNX1

- **shRUNX1_R2**

In [35]:
```
pwd
cd ../data
```

```
/home/luna.kuleuven.be/r0637408/data/work
```

In [29]:
```
ls
```

```
~                       shNSR1.Log.progress.out     shRUNX1_R1_fastqc.zip
shNSR1.Aligned.out.sam  shNSR1.SJ.out.tab           shRUNX1_R1.Log.final.out
shNSR1.bam              shNSR2.fastq                shRUNX1_R1.Log.out
shNSR1.bam.bai          shNSR3.fastq                shRUNX1_R1.Log.progress.out
shNSR1.fastq            shRUNX1_R1.Aligned.out.sam  shRUNX1_R1.SJ.out.tab
shNSR1_fastqc.html      shRUNX1_R1.bam              shRUNX1_R2.fastq
shNSR1_fastqc.zip       shRUNX1_R1.bam.bai          shRUNX1_R3.fastq
shNSR1.Log.final.out    shRUNX1_R1.fastq
shNSR1.Log.out          shRUNX1_R1_fastqc.html
```

In [30]:
```
/usr/local/bin/FastQC/fastqc ../data/shRUNX1_R2.fastq
```

```
Started analysis of shRUNX1_R2.fastq
Approx 5% complete for shRUNX1_R2.fastq
Approx 10% complete for shRUNX1_R2.fastq
Approx 15% complete for shRUNX1_R2.fastq
Approx 20% complete for shRUNX1_R2.fastq
Approx 25% complete for shRUNX1_R2.fastq
Approx 30% complete for shRUNX1_R2.fastq
Approx 35% complete for shRUNX1_R2.fastq
Approx 40% complete for shRUNX1_R2.fastq
Approx 45% complete for shRUNX1_R2.fastq
Approx 50% complete for shRUNX1_R2.fastq
Approx 55% complete for shRUNX1_R2.fastq
Approx 60% complete for shRUNX1_R2.fastq
Approx 65% complete for shRUNX1_R2.fastq
Approx 70% complete for shRUNX1_R2.fastq
Approx 75% complete for shRUNX1_R2.fastq
Approx 80% complete for shRUNX1_R2.fastq
Approx 85% complete for shRUNX1_R2.fastq
Approx 90% complete for shRUNX1_R2.fastq
Approx 95% complete for shRUNX1_R2.fastq
Analysis complete for shRUNX1_R2.fastq
```

In [40]:
```
ls *html
```

```
shNSR1_fastqc.html  shRUNX1_R1_fastqc.html
shNSR2_fastqc.html  shRUNX1_R2_fastqc.html
```

Analysis of the Quality control results shRUNX2:

Sequences flagged as poor quality: 0

```
    - Per base sequence quality report: The score of the base calling decre
ases a little bit along the sequence, however the scores still remain in th
e green section of the y-axis which means that the base calling is of good
quality.

    - Sequences duplication levels report: The trend of the curve is simila
r to the previous one with peak on the right side of the curve. This might
implicate more specific enirchments or the presence of contaminants. The wa
rning denotes that non-unique sequences make up of more than 20% of the tot
al.

    - Overrepresented sequences report: A single sequence very overrepresen
ted in the set means that it is highly biologically significant or that it
means that the library is contaminated.
    In our data set, there are several overrepresented sequences but for no
n of them a possible source is found. Except for one that is a TruSeq adapt
er which can happen as pointed out above. The warning means that one sequen
ce is found that represents more than 0,1% of the total.
```

In the figures below you can see some selected results of the quality control.



Figure 10: Summary of the quality control results of shRUNXR2

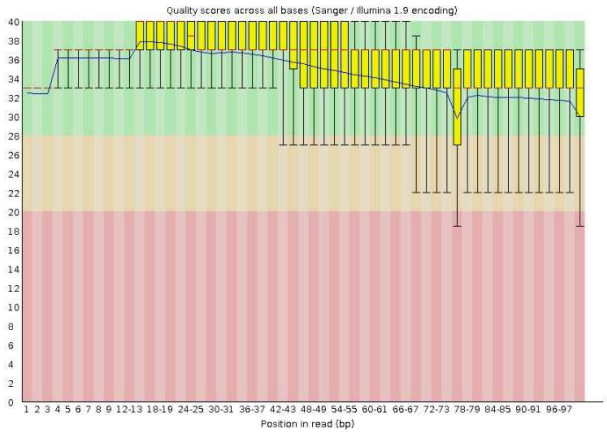

Figure 11: Per base sequence quality of shRUNXR2

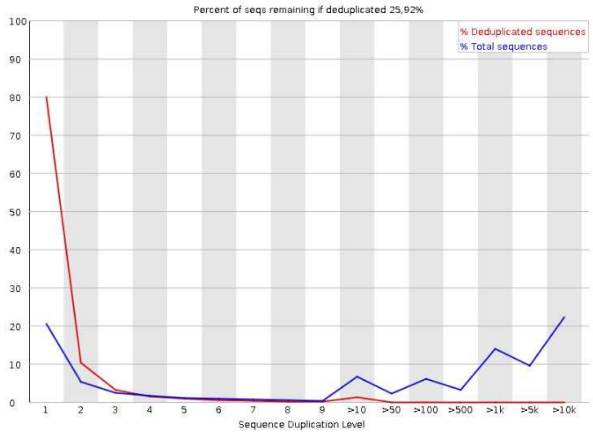

Figure 12: Sequence duplication levels of shRUNXR2

In [21]:
```
total 55439644
drwxr-xr-x 3 r0637408 domain users        4096 nov 16 12:04 ~
-rw-r--r-- 1 r0637408 domain users  7537729432 nov 16 13:46 shNSR1.fastq
-rw-r--r-- 1 r0637408 domain users      364400 nov 26 14:41 shNSR1_fastqc.html
-rw-r--r-- 1 r0637408 domain users      494039 nov 26 14:41 shNSR1_fastqc.zip
-rw-r--r-- 1 r0637408 domain users  8862217020 nov 16 13:27 shNSR2.fastq
-rw-r--r-- 1 r0637408 domain users  7288568384 nov 16 13:24 shNSR3.fastq
-rw-r--r-- 1 r0637408 domain users  8859434212 nov 16 12:42 shRUNX1_R1.fastq
-rw-r--r-- 1 r0637408 domain users      375672 nov 26 14:35 shRUNX1_R1_fastqc.ht
ml
-rw-r--r-- 1 r0637408 domain users      511874 nov 26 14:35 shRUNX1_R1_fastqc.zi
p
-rw-r--r-- 1 r0637408 domain users 12681835778 nov 16 13:04 shRUNX1_R2.fastq
-rw-r--r-- 1 r0637408 domain users 11538621938 nov 16 13:05 shRUNX1_R3.fastq
```

Conclusion of step 3: The quality control of the four samlpes had indicated a good quality, this implements that these samples can be used for the further analysis.

Step 4: Mapping to the genome

STAR toolkit is used to map the genome. This program also needs a database with the human genome (hg19) and gene annotation. Path of the datase that will be used:

/mnt/nfs/mfiers/STAR/hg19_star_db

In [3]:
```
ls /mnt/nfs/mfiers/STAR/hg19_star_db/
```

```
chrLength.txt      Genome                       SAindex
chrNameLength.txt  genomeParameters.txt         sjdbInfo.txt
chrName.txt        JAZZYCHIP.Aligned.out.sam    sjdbList.fromGTF.out.tab
chrStart.txt       JAZZYCHIP.Log.out            sjdbList.out.tab
exonGeTrInfo.tab   JAZZYCHIP.Log.progress.out  transcriptInfo.tab
exonInfo.tab       JAZZYCHIP._STARtmp
geneInfo.tab       SA
```

In [4]:
```
STAR --genomeLoad LoadAndExit --genomeDir /mnt/nfs/mfiers/STAR/hg19_star_db
```

```
Jan 24 14:16:11 ..... started STAR run
Jan 24 14:16:11 ..... loading genome
```

In [5]:
```
cd /mnt/nfs/r0637408/work
```

In [6]:
```
STAR --genomeDir /mnt/nfs/mfiers/STAR/hg19_star_db \
     --genomeLoad LoadAndKeep \
     --runThreadN 2 \
     --readFilesIn shNSR1.fastq \
     --outFileNamePrefix shNSR1
```

```
Jan 24 14:16:15 ..... started STAR run
Jan 24 14:16:15 ..... loading genome
Jan 24 14:16:15 ..... started mapping
Jan 24 14:27:36 ..... finished successfully
```

The output file of the mapping is of the **.Aligned.out.sam** type. The mapping also generates some additional log files which we don't need now.

In [7]:
```
ls -l shNSR1.*
```

```
-rw-r--r-- 1 r0637408 domain users 7237134968 jan 24 14:27 shNSR1.Aligned.out.sa
m
-rw-r--r-- 1 r0637408 domain users 7537729432 jan 24 11:47 shNSR1.fastq
-rw-r--r-- 1 r0637408 domain users       1862 jan 24 14:27 shNSR1.Log.final.out
-rw-r--r-- 1 r0637408 domain users      15540 jan 24 14:27 shNSR1.Log.out
-rw-r--r-- 1 r0637408 domain users       1544 jan 24 14:27 shNSR1.Log.progress.o
ut
-rw-r--r-- 1 r0637408 domain users    5897545 jan 24 14:27 shNSR1.SJ.out.tab
```

Step 5: Inspecting the structure of the SAM file

The sam output file of the STAR analysis is a text file containing the alignment of all reads.

The head commando reads the first few lines of text and displays them on the screen. This enables us to check if the file is build up well - with the right structure.

In [10]:
```
head -34 shNSR1.Aligned.out.sam | grep -v '^@'
```

```
SRR2919468.1    0       chr6    74227928        255     1S60M89N40M     *
0       0       NGCAATCCAATACAGGGGCATAGCCGGCGCTTATTTGGCCTGGATGGTTCAGGATAATCACCTG
AGCAGTGAAGCCAGCTGCTTCCATTGGTGGGTCATTT   #0<FFFFFFFFFFIIIIIIIIIIFFFIIIIIIIIIIIII
IIIIFFFBFFFFFFBFFFF77<<BBBFFFBBFFFBBFFFFFFBFFFFFFFFBBB<7BBFFF   NH:i:1  HI:i:1
AS:i:100        nM:i:0
SRR2919468.2    16      chr6    52860628        255     99M2S   *       0
0       CCTGCTAGAACCTCCAAACAAGCTCTCAAGGTCCATTTGTAGGAGAACGTAGGGTAGTCAAGCTTCCAAGAC
TCCAGACACATCCAAATGAGGCGCTGCGN    FFBFFFBBB<BBFFFFFFFFBFFBFFFBFFFFFFFBFFIIIIIFIFI
IIFFIIFIIIIIIIFFFIIIIFFIIIIIIIIIIIIIIIIIFFFFFFFFF<0#   NH:i:1  HI:i:1  AS:i:97
nM:i:0
SRR2919468.3    0       chr15   41282673        255     101M    *       0
0       CTCAGCTTCCCAAAGTGTTGGAATTACAGGTATGAGCCACTGCACCTGGCCAAAATGGTTTCCCTCCATTTT
TAGCACAGGTTAGAATCCTTGGGAAGAAC    BBBFFFFFFFFFFIIFFFIFFIIIIIIFFFFFFFIIIIIIIIIIII
IIIIIIIIIFFFFIIIIIIIIIIFFFBFFFFFFFFBBFFBFFFFFBBBFFFB   NH:i:1  HI:i:1  AS:i:99
nM:i:0
SRR2919468.4    16      chr17   31149639        255     12S89M  *       0
0       TCCGATCTGTGGCGGTCCCCGTCCCGGGGGTTGGCCGCGCGGGCCCCGGTGGGGCGGCCACCCGGGGTCCCG
GCCCTCGCGCGTCCTTCCTCCTCGCTCCT    '''7''00''''''''00'''7<700700''''''0B7'770<<0'7B
7<77'0<707''777B<7B<7''<07'7<FB'BB7FBBFFFFFFFFFFBBBBB   NH:i:1  HI:i:1  AS:i:75
nM:i:6
SRR2919468.5    16      chr22   21243207        255     101M    *       0
0       TTTGCTAAGGGGGCAAAAAGCCCAGATACCATTTTAACTGGCCCAGCTCATTTTCTCCTTTCATTTTACAAA
AGAAATCATGCAGAAGCATCAATTAAAAC    BBBBBBBBBBBBBBFFFFFFFFIFB<BFIIFBB<FIFFIIIIIIIFF
FFFFFFIFIIFF<IIIFFFIIIIIIFIIFIIFFFFFFBIIFFFBBBBFFFBBB   NH:i:1  HI:i:1  AS:i:97
nM:i:1
SRR2919468.6    0       chr17   40500418        255     1S100M  *       0
0       NCTCAATCCAAGGGGCCAGAAACTGCCGCAGCTCCATTGGGAAGCTGTCACTGTAGAGCTGATGGAGCTGCT
CCAGGTACCGTGTGTCAAGCTGCTGTAGC    #0<FFBFFFFBFBFFIIIIIFFFIBFFFFFFIIIIIIIIIIFFBFFFB
BBFFFBFFFFFFFFFFFB<BFBBBB<B<B7<<BB7B7B<BBBBBBBBBBBBB   NH:i:1  HI:i:1  AS:i:98
nM:i:0
```

In [11]:
```
ls /mnt/nfs/r0637408/work
```

```
~                       shNSR1.Log.final.out    shNSR2.Log.out
Aligned.out.sam         shNSR1.Log.out          shNSR2.Log.progress.out
Log.out                 shNSR1.Log.progress.out shNSR2._STARtmp
Log.progress.out        shNSR1.SJ.out.tab       shRUNR1.fastq
shNSR1.Aligned.out.sam  shNSR2.Aligned.out.sam  shRUNR2.fastq
shNSR1.fastq            shNSR2.fastq            _STARtmp
```

The .sam files were analysed and showed a good structure. They contain additional information about the sequences, such as chromosone location and SRR number.

Step 6: Sorting of the SAM file

Sorting and conversion of the SAM file into a BAM file, sorts the data and compresses the file. In order to do this we use **samtools sort**.

General command: samtools sort -o (output file name) (input file name)

In [12]:
```
samtools sort -o shNSR1.bam shNSR1.Aligned.out.sam
```

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

ls -l command to display the files in a specific directory + extra information about them, including the size of the file.

In [13]:
```
ls -l shNSR1.bam
```

```
-rw-r--r-- 1 r0637408 domain users 1063491935 jan 24 14:35 shNSR1.bam
```

Step 7: Indexing the bam file

The tool **samtools index** is used to generate an index to visualise and to enable a fast access to the bam file. The resulting **.bai** file allows programs to jump directly to specific parts of the bam file without reading through all of the sequences and must always be kept together with the bam file.

In [14]:
```
samtools index shNSR1.bam
```

In [15]:
```
ls -l shNSR1.bam.bai
```

```
-rw-r--r-- 1 r0637408 domain users 2795216 jan 24 14:37 shNSR1.bam.bai
```

Step 8: Inspecting the BAM file

The square brackets are used to include the 2 possiblities at that position in the command. This is used in order to display all the files with the suffix .bam or .bai

In [16]:
```
ls *.ba[mi]
```

```
shNSR1.bam  shNSR1.bam.bai
```

In [17]:
```
samtools view shNSR1.bam | head -3
```

```
SRR2919468.1257417      0       chrM    1       255     1S95M5S *       0
0       GGATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTTCGTCTGGGGGG
TATGCACGCGATAGCATTGCGAGAGATCG    BBBF<BFBB<BFFBF0FF0BFFBFBBF<BFIIFI<FF<FFFFFFFFFB
BBFFIII'7B70BBF7BFFBFFBB77<<BBBB<BBBBBBB<BBBB<<BBBBBB   NH:i:1  HI:i:1  AS:i:91
nM:i:1
SRR2919468.1307827      0       chrM    1       255     3S98M   *       0
0       ATGGATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTTCGTCTGGGG
GGTATGCACGCGATAGCATTGCGAGACGC    BBBFFFFFFFFFFFFIIIFIIIIIIIFIIIIIIIIIIIIIIIIIIII
IIIIIIIIF7BFIIIIIFFFFFFFFF77BBFFFFFFFFFBFFFFFFFFBFFFFFF   NH:i:1  HI:i:1  AS:i:94
nM:i:1
SRR2919468.2735860      0       chrM    1       255     4S97M   *       0
0       GATGGATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTTCGTCTGGG
GGGTATGCACGCGATAGCATTGCGAGACG    BBBFFFFFFFFFFFFFFIIIIFIIIIIIIIIIIIIIIIIIIIIIIII
IIIIFFIIIIBFFIIIIIFFFFFFFFFFBFFFFBBBFBBFFFFFBFFFFF<<B   NH:i:1  HI:i:1  AS:i:93
nM:i:1
samtools view: writing to standard output failed: Broken pipe
samtools view: error closing standard output: -1
```

The structure of the BAM and SAM file is the same.

Visualisation by using IGV

The Integrative Genomics Viewer (IGV) is a visualization tool which supports a large variety of data types, such as next-generation sequence data, genomic annotations and array-based data. It provides an interactive exploration of large, integrated genomic datasets.

To visualize the .bam and .bai files of both shNSR and shRUNX we used IGV.



Figure 13: Visualisation in IGV of the .bam and .bai files of both shNSR1 and shRUNX1. Zoomed in on MALAT1.

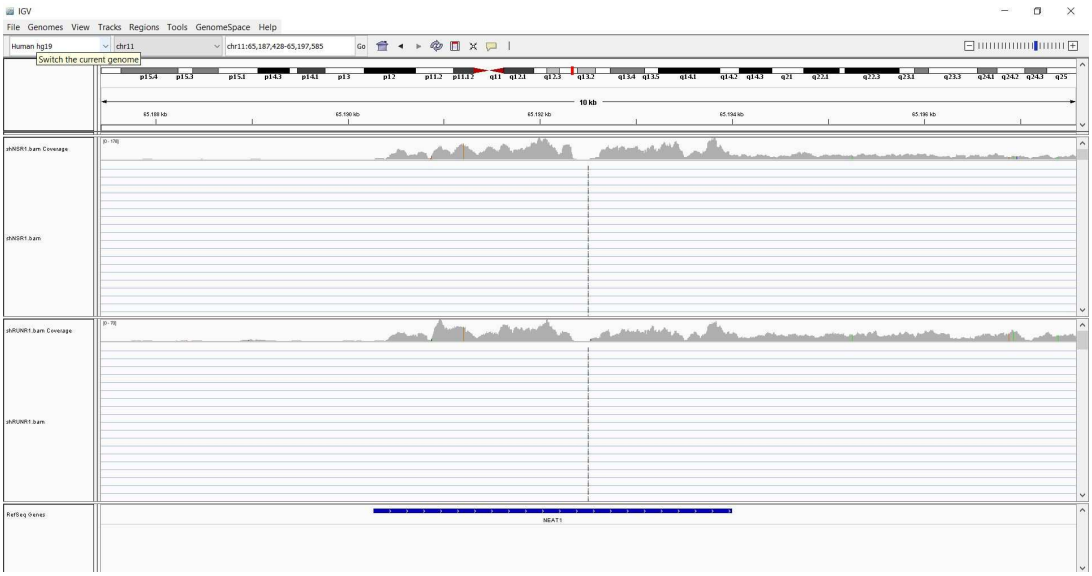

Figure 14: Visualisation in IGV of the .bam and .bai files of both shNSR1 and shRUNX1. Zoomed in on NEAT1.

We use **samtools idxstats** to calculate how many reads map to each chromosome:

In [18]:
```
samtools idxstats chNSR1.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
```

The first colomn in the results shows the reference sequence name and the length of this sequence in the second column. In the third and fourth column the numer of mapped and unmapped reads respectively. In the chr 14 reference sequence name the most reads are mapped.

With the tool **samtools flagstat** it is possible to display the distribution of mapping flags. This can also be found in the second column of the sam/bam file.

The first column of output gives the total number of reads that are QC passed and failed (in second column) --> seperated with a +

In [19]:
```
samtools flagstat chNSR1.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)
```

There are in total of 25.331.592 reads in the input file. All the reads are quality control (QC) passed and all of the reads are mapped.

The procedure step 4-8 is exactly the same for the other RNA-seq files: shNSR2.fastq, shRUNR1.fastq, shRUNR2.fastq

The steps for the shRUNXR1.fastq are showed. The steps of the others are not included in this report.

Step 4: Mapping to the genome (tool: STAR)

Step 5: Inspecting the structure of the sam file (commando: head)

Step 6: Sorting of the sam file --> output: Bam file (commando: samtools sort)

Step 7: Indexing --> output: bam.bai file (commando: samtools index)

Step 8: Inspecting the BAM file (commando's: samtools idxstat, flagstat)

In [20]:
```
STAR --genomeDir /mnt/nfs/mfiers/STAR/hg19_star_db \
     --genomeLoad LoadAndKeep \
     --runThreadN 2 \
     --readFilesIn shRUNR1.fastq \
```

```
Jan 24 14:38:57 ..... started STAR run
Jan 24 14:38:57 ..... loading genome
Jan 24 14:38:57 ..... started mapping
Jan 24 14:52:56 ..... finished successfully
```

In [21]:
```
-rw-r--r-- 1 r0637408 domain users 8182917631 jan 24 14:52 shRUNR1.Aligned.out.s
am
-rw-r--r-- 1 r0637408 domain users 8859434212 jan 24 11:53 shRUNR1.fastq
-rw-r--r-- 1 r0637408 domain users       1864 jan 24 14:52 shRUNR1.Log.final.out
-rw-r--r-- 1 r0637408 domain users      15551 jan 24 14:52 shRUNR1.Log.out
-rw-r--r-- 1 r0637408 domain users       1780 jan 24 14:52 shRUNR1.Log.progress.
out
-rw-r--r-- 1 r0637408 domain users    6254202 jan 24 14:52 shRUNR1.SJ.out.tab
```

In [24]:
```
head -34 shRUNR1.Aligned.out.sam | grep -v '^@'
```

```
SRR2919471.1    0       chr8    144662711       255     2S99M   *       0
0       NGCCAGTGCAGGCTTCTTGGCCTTCTTCTCCGCGTACTGCCGTAGCCGCTCCTCCCGCAGCTGTGCCGCCTC
CTTGTCCTCCTCCTCATTGTCACTGCCAA   #0<FFFFFFFFFFIIIIIIIIIIIIIIIIIIIIIFFIIIIIIFIIIII
IIFFFFFFFFFBFFFFFBBFFFFFFBF<BBFFFFFFFFB<BBBFBBBFFFFFF   NH:i:1  HI:i:1  AS:i:97
nM:i:0
SRR2919471.2    16      chr3    10321639        255     18S82M1S        *
0       0       CGTGTGCTCTTCCGATCTGGCTGCTGTGTGGGTGTGTTAAGCTGATAAGGTTCAGTTTGCGGCG
GAAACTACCTGCTAGTCTTGTGTCAGGGGCTGCCAGN   7FFFFFFF<FFFFFFFFFFFFFBBBBFFFBFFFFFFFFFF
FFFFFFFFFIIIFIIFFIIIIIIIIIIFIIIIFIIIIIIIIIIIIIIIFFFFFFFFFF<0#   NH:i:1  HI:i:1
AS:i:80 nM:i:0
SRR2919471.3    16      chr3    96777395        255     101M    *       0
0       CTTGGCCTCCCAAAGTGCTGGGATTACAGGTGTGAGCCACCATGCCCGGCCTGGAGTAAGGAGTTTTTAATC
AAGTTCTTTTATTGTTACTTTCAATATAG   FFFFFFFBFBFFFFFFBFFFFFFFFBFFFFFFFFFFFIIIIIIIIIII
FIFFBIIIIIIFFIIIIIIIIIIIIIIIIIIIIIIIIIIIFFFFFFFFFBBB   NH:i:1  HI:i:1  AS:i:99
nM:i:0
SRR2919471.4    16      chr14   45433630        255     2S41M31278N57M1S
*       0       0       TCTGGGAGAAAACCAGACCTCCACTTCCAAGGATATAGAGCAGTTTTCAACATATG
ATTTCATCCCATCTGCAAAATTAAAGCTTTCTCAAGGAATGCCAN   BBBB<BFBFFFFFFBBBFFFFFBFIBIIFFIF
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIFFIIFIIIFIIIFIIIIIIIIIIFFIFFFFFFFFFFF<0#   NH:i:1
HI:i:1  AS:i:96 nM:i:0
SRR2919471.5    16      chrX    70517274        255     38M374N63M      *
0       0       AACCAAGAGGTGCAAAAACGAAAGCAACTGGAGCTCAGGCAGGAGGAAGAGCGCAGGCGCCGTG
AAGAAGAGATGCGGCGGCAGCAAGAAGAAATGATGCG   BBBBBBBBBB<B7<<7B<BBBBBB<0B<<B<BBBBBBBBB
BBBBBB<<7B<FBBB<<FFBBFBBFFFFFFBB<BIFIFFB<IIFIIFIFFFBFFB<FFBBB   NH:i:1  HI:i:1
AS:i:101        nM:i:0
SRR2919471.6    16      chr19   33076803        255     11M950N72M323N16M2S
*       0       0       TAAGTGAGAAGGTATCAGAACAAGGTTTAATAGAAATCCTTAAAAAAGTAAGCCAA
CAAACAGAAAAGACAACAACAGTGAAATTCAACAGAAGAAAAGCN   <FBBFFFFFBBBBFFFFFFFBFFBFBFIIIII
IFFFFFIIIIIIIIIFFIIIIFIIIIFFIIIIIIIFFFFIIIFIIIIIIIIIIIIIFFFFFFFFFF<0#   NH:i:1
HI:i:1  AS:i:100        nM:i:0
```

In [25]:
```
samtools sort -o shRUNR1.bam shRUNR1.Aligned.out.sam
```

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

In [26]:
```
ls -l shRUNR1.bam
```

```
-rw-r--r-- 1 r0637408 domain users 1226869972 jan 24 15:07 shRUNR1.bam
```

In [27]:
```
samtools index shRUNR1.bam
```

In [29]:
```
ls *.ba[mi]
```

```
shNSR1.bam                              shRUNR2.Aligned.out.sam.tmp.0006.bam
shNSR1.bam.bai                          shRUNR2.Aligned.out.sam.tmp.0007.bam
shNSR2.bam                              shRUNR2.Aligned.out.sam.tmp.0008.bam
shNSR2.bam.bai                          shRUNR2.Aligned.out.sam.tmp.0009.bam
shRUNR1.bam                             shRUNR2.Aligned.out.sam.tmp.0010.bam
shRUNR1.bam.bai                         shRUNR2.Aligned.out.sam.tmp.0011.bam
shRUNR2.Aligned.out.sam.tmp.0000.bam    shRUNR2.Aligned.out.sam.tmp.0012.bam
shRUNR2.Aligned.out.sam.tmp.0001.bam    shRUNR2.Aligned.out.sam.tmp.0013.bam
shRUNR2.Aligned.out.sam.tmp.0002.bam    shRUNR2.Aligned.out.sam.tmp.0014.bam
shRUNR2.Aligned.out.sam.tmp.0003.bam    shRUNR2.Aligned.out.sam.tmp.0015.bam
shRUNR2.Aligned.out.sam.tmp.0004.bam    shRUNR2.bam
shRUNR2.Aligned.out.sam.tmp.0005.bam
```

In [30]:
```
samtools view chRUNR1.bam | head -3
```

```
SRR2919471.40625        0       chrM    1       255     9S92M   *       0
0       ATCACGATGGATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTTCGT
CTGGGGGGTATGCACGCGATAGCATTGCG   BBBFFFFFFFFFFFIIIIIFFIIIIFIIIIFIIIIIIIIIIIIIIIII
IIIIIIIIIIIIIIIFFFFFFFFBFFFBBFFF7BBBBFFFFFBBFFFBFBBF<   NH:i:1  HI:i:1  AS:i:88
nM:i:1
SRR2919471.1316495      0       chrM    1       255     7S94M   *       0
0       CACGATGGATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTTCGTCT
GGGGGGTATGCACGCGATAGCATTGCGAG   BBBFFFFFFFFFFFIIIIIIIIIIIIIIIIIIIIIIIIIIIFFIIIII
IIIIIIIIIIIIBFFIFFFFFFFFFFFBFF7<FBFFFFFFFFBFFFFFFFFF   NH:i:1  HI:i:1  AS:i:90
nM:i:1
SRR2919471.7269090      0       chrM    1       255     9S92M   *       0
0       ATCACGATGGATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTTCGT
CTGGGGGGTATGCACGCGATAGCATTGCG   BBBFFFFFFFFFFFFFIIIBFFFIFFIIIIIIFIIFFFFFIIIIFIIII
FFIIFFIIFFFFFIIBFBFFFFFFFFBFBFBFF7<BFBFFFFBFBFBFBFF<<BBB   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 [31]:
```
samtools idxstats chRUNR1.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 [32]:
```
samtools flagstat chRUNR1.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)
```

There are in total of 28.587.488 reads in the input file. All the reads are quality control (QC) passed and all of the reads are mapped.

Step 9: Reads to gene counts

This is the conversion of the results of the analysis (the reads) to information about genes. In this way it is possible to relate certain gene expressions in the 2 conditions (presence / absence of RUNX1).

For this conversion we use the tool: **featureCounts**

General command: FeatureCounts -Q 10 -g gene_name -a [human annotation of specific chromosome.gtf] -o output file input file

In [7]:
```
cd ~/data/work
pwd
```

```
/home/luna.kuleuven.be/r0637408/data/work
```

In [33]:
```
featureCounts -Q 10 -g gene_name \
-a /mnt/nfs/data/RNA-seq/gencode.v19.nopseudo.plus.sort.gtf \
-o all.counts.RUN \
shNSR1.bam shNSR2.bam shRUNR1.bam shRUNR2.bam
```

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

//========================== featureCounts setting ===========================\\
||                                                                            ||
||             Input files : 4 BAM files                                      ||
||                           S shNSR1.bam                                     ||
||                           S shNSR2.bam                                     ||
||                           S shRUNR1.bam                                    ||
||                           S shRUNR2.bam                                    ||
||                                                                            ||
||             Output file : all.counts.RUN                                   ||
||             Annotations : /mnt/nfs/data/RNA-seq/gencode.v19.nopseudo.p ... ||
||                                                                            ||
||                 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/ (http://subread.sourcefo
rge.net/) ======================//

//================================= Running ==================================\\
||                                                                            ||
|| Load annotation file /mnt/nfs/data/RNA-seq/gencode.v19.nopseudo.plus.s ... ||
||    Features : 1156385                                                      ||
||    Meta-features : 41863                                                   ||
||    Chromosomes/contigs : 25                                                ||
||                                                                            ||
|| Process BAM file shNSR1.bam...                                             ||
||    Single-end reads are included.                                          ||
||    Assign reads to features...                                             ||
||    Total reads : 25331592                                                  ||
||    Successfully assigned reads : 9202975 (36.3%)                           ||
||    Running time : 0.47 minutes                                             ||
||                                                                            ||
|| Process BAM file shNSR2.bam...                                             ||
||    Single-end reads are included.                                          ||
||    Assign reads to features...                                             ||
||    Total reads : 30966793                                                  ||
||    Successfully assigned reads : 13072801 (42.2%)                          ||
||    Running time : 0.53 minutes                                             ||
||                                                                            ||
|| Process BAM file shRUNR1.bam...                                            ||
||    Single-end reads are included.                                          ||
||    Assign reads to features...                                             ||
||    Total reads : 28587488                                                  ||
||    Successfully assigned reads : 11202455 (39.2%)                          ||
||    Running time : 0.60 minutes                                             ||
||                                                                            ||
|| Process BAM file shRUNR2.bam...                                            ||
||    Single-end reads are included.                                          ||
||    Assign reads to features...                                             ||
||    Total reads : 37045685                                                  ||
||    Successfully assigned reads : 10788451 (29.1%)                          ||
||    Running time : 0.68 minutes                                             ||
```

In [36]:
```
ls -l *.counts*
```

```
-rw-r--r-- 1 r0637408 domain users 8673915 jan 24 15:18 all.counts.RUN
-rw-r--r-- 1 r0637408 domain users     472 jan 24 15:18 all.counts.RUN.summary
```

In [38]:
```
head -20 all.counts.RUN.summary | column -t
```

```
Status                      shNSR1.bam  shNSR2.bam  shRUNR1.bam  shRUNR2.bam
Assigned                    9202975     13072801    11202455     10788451
Unassigned_Ambiguity        612021      861139      767761       696060
Unassigned_MultiMapping     0           0           0            0
Unassigned_NoFeatures       6080016     5722461     5458528      13720904
Unassigned_Unmapped         0           0           0            0
Unassigned_MappingQuality   9436580     11310392    11158744     11840270
Unassigned_FragmentLength   0           0           0            0
Unassigned_Chimera          0           0           0            0
Unassigned_Secondary        0           0           0            0
Unassigned_Nonjunction      0           0           0            0
Unassigned_Duplicate        0           0           0            0
```

In [41]:
```
head -10 all.counts.RUN
```

```
# Program:featureCounts v1.5.0-p1; Command:"featureCounts" "-Q" "10" "-g" "gene_
name" "-a" "/mnt/nfs/data/RNA-seq/gencode.v19.nopseudo.plus.sort.gtf" "-o" "all.
counts.RUN" "shNSR1.bam" "shNSR2.bam" "shRUNR1.bam" "shRUNR2.bam"
Geneid  Chr     Start   End     Strand  Length  shNSR1.bam      shNSR2.bam
shRUNR1.bam     shRUNR2.bam
MIR1302-11      chr1;chr1;chr1  29554;30267;30976       30039;30667;31109
+;+;+   1021    0       0       0       0
FAM138A chr1;chr1;chr1  34554;35245;35721       35174;35481;36081       -;-;-
1219    0       0       0       0
OR4F5   chr1    69091   70008   +       918     0       0       0       0
RP11-34P13.7    chr1;chr1;chr1;chr1;chr1;chr1;chr1      89295;92091;110953;11270
0;120721;129055;133374  91629;92240;111357;112804;120932;129223;133566  -;-;-;-;
-;-;-;-   3569    2       0       0       0
RP11-34P13.8    chr1;chr1       89551;90287     90050;91105     -;-     1319
0       0       0       0
AL627309.1      chr1;chr1       134901;137621   135802;139379   -;-     2661
13      14      6       15
RP11-34P13.14   chr1;chr1       139790;140075   139847;140339   -;-     323
0       0       0       0
RP11-34P13.13   chr1;chr1;chr1;chr1;chr1;chr1;chr1;chr1;chr1;chr1       141474;1
46386;155767;164263;165884;168100;168610;169049;172557;173753   143011;149707;15
5831;164791;165942;168165;168767;169264;172688;173862   -;-;-;-;-;-;-;-;-;-
6195    9       5       5       11
```

With the tool featurecounts we were able to merge the .bam files of the different samples in order to compare the gene counts. The output of the featurecounts is a document that lists the gene counts of specific genes. And gives additional information such as the chromosome location and the length. Hereafter, we will only continue with the gene counts of the different samples.

In [42]:
```
cut -f1,7- all.counts.RUN > all.genedata.RUN.tsv
```

In [43]:
```
ls *.tsv
```

```
all.genedata.RUN.tsv
```

In [44]:
```
cd ~/data/work
pwd
```

```
/home/luna.kuleuven.be/r0637408/data/work
```

In [46]:
```
head all.genedata.RUN.txt
```

```
# Program:featureCounts v1.5.0-p1; Command:"featureCounts" "-Q" "10" "-g" "gene_
name" "-a" "/mnt/nfs/data/RNA-seq/gencode.v19.nopseudo.plus.sort.gtf" "-o" "all.
counts.RUN" "shNSR1.bam" "shNSR2.bam" "shRUNR1.bam" "shRUNR2.bam"
Geneid  Chr     Start   End     Strand  Length
MIR1302-11      chr1;chr1;chr1  29554;30267;30976       30039;30667;31109
+;+;+   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;110953;11270
0;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       141474;1
46386;155767;164263;165884;168100;168610;169049;172557;173753   143011;149707;15
5831;164791;165942;168165;168767;169264;172688;173862   -;-;-;-;-;-;-;-;-;-
6195
```

In [47]:
```
cut -f1,7- all.counts.RUN | grep -v '^#' > all.gene.counts.RUN.txt
```

In [3]:
```
head -50 all_gene_counts_RUN.tsv
```

```
Geneid  shNSR1.bam      shNSR2.bam      shRUNR1.bam     shRUNR2.bam
MIR1302-11      0       0       0       0
FAM138A 0       0       0       0
OR4F5   0       0       0       0
RP11-34P13.7    2       0       0       0
RP11-34P13.8    0       0       0       0
AL627309.1      13      14      6       15
RP11-34P13.14   0       0       0       0
RP11-34P13.13   9       5       5       11
RNU6-1100P      0       0       0       0
RP11-34P13.9    0       0       0       0
AP006222.2      1       2       2       1
RP4-669L17.10   0       1       1       0
OR4F29  0       0       0       0
RP4-669L17.2    0       0       2       0
RP5-857K21.15   0       0       0       0
RP5-857K21.1    3       4       0       4
RP5-857K21.2    0       0       0       0
RP5-857K21.3    0       0       0       0
RP5-857K21.4    0       0       1       0
RP5-857K21.5    0       0       0       0
OR4F16  0       0       0       0
RP11-206L10.3   2       0       1       1
RNU6-1199P      0       1       1       0
RP11-206L10.5   0       0       1       0
RP11-206L10.4   0       0       0       0
RP11-206L10.2   12      11      3       13
RP11-206L10.9   18      32      12      19
AL669831.1      1       1       0       2
RP11-206L10.10  6       9       5       4
FAM87B  2       5       3       0
LINC00115       8       19      2       2
RP11-206L10.11  76      115     95      46
FAM41C  0       0       0       0
AL645608.2      0       0       0       0
RP11-54O7.16    3       1       0       1
RP11-54O7.1     4       6       6       6
RP11-54O7.2     0       0       0       0
RP11-54O7.3     42      79      72      71
SAMD11  52      107     58      86
AL645608.1      4       3       1       2
NOC2L   729     1095    803     1217
KLHL17  186     288     53      41
PLEKHN1 26      45      14      30
C1orf170        9       6       5       7
RP11-54O7.17    7       7       6       8
HES4    129     176     176     132
RP11-54O7.11    1       2       1       2
ISG15   318     463     257     534
AGRN    1601    2642    1319    1865
```

In [1]:
```
cd ~/data/work
pwd
```

```
/home/luna.kuleuven.be/r0637408/data/work
```

In [50]:
```
grep MALAT1 all_gene_counts_RUN.tsv
```

```
MALAT1  126596  195321  74491   36787
```

In [4]:
```
grep NEAT1 all_gene_counts_RUN.tsv
```

```
NEAT1   7768    14094   4295    3736
```

The expression level of a lot of genes is lower in the mutated samples, this suggests that RUNX1 cannot activate these genes due to its inactivation. Through knockdown of RUNX1 in breast cancer cells, we find a downregulation of 2 important genes after inspection of the excel files (which are not shown), NEAT1 and MALAT1. These genes are involved in SC-35 speckle organisation. The 2 genes and their counts per file/condition are shown. Clearly the expression level is much lower in the cell lines with knockdown of RUNX1.

- Function of NEAT1: This gene produces a long non-coding RNA (lncRNA). This lncRNA is retained in the nucleus where it forms the core structural component of the paraspeckle sub-organelles. It may act as a transcriptional regulator for numerous genes, including some genes involved in cancer progression.
- Function of MALAT1: This gene also produces a long non-coding RNA (lncRNA). This transcript is retained in the nucleus where it is thought to form molecular scaffolds for ribonucleoprotein complexes. It may act as a transcriptional regulator for numerous genes, including some genes involved in cancer metastasis and cell migration and it is involved in cell cycle regulation. Its upregulation in multiple cancerous tissues has been associated with the proliferation and metastasis of tumor cells.

Therefore, the upregulation of these 2 genes by RUNX1 is beneficial for breast cancer cells.

Functional analysis of the RNAseq data

In the functional analysis, we analyse which genes are most (and significantly) affected with knockdown of RUNX1 (shRUNX1). This is done in order to investigate the role of RUNX1 in breast cancer in MCF-7 breast cancer cell lines.

RNA-Seq is generally used to compare gene expression between conditions, such as a drug treatment vs non-treated, and find out which genes are up- or down regulated in each condition. In principle, RNA-Seq will allow to account for all the transcripts in the cell for each condition. Differently expressed genes can be identified by using tools that count the sequencing reads per gene and compare them between samples. The most commonly used tools for this type of analysis are DESeq and edgeR. Both these tools use a model based on the negative binomial distribution.

The functional analysis uses the results of the RNA-seq in order to look for significantly differentially expressed genes. These results can be represented in a heat map using the program 'Multi Experiment Viewer (MeV). Before usage all data was normalized. The ouputs represent the top 24 of the most significant changes in gene expression. To do this the FeatureCounts results is uploaded and a specific genecluster is selected (= result DESeq).

DE-seq analysis

For the DESeq - differential gene expression analysis a R-notebook was used. You can find the workflow in the other notebook file named **"Bioinformatics_DESeq2.ipynb"**.

Interpretatie

In [13]:
```
cd ~/data/work
pwd
```

```
/home/luna.kuleuven.be/r0637408/data/work
```

In [26]:
```
head -10 deseq.results.RUN.tsv | column -t
```

```
baseMean   log2FoldChange    lfcSE             stat               pvalue
padj
MIR663A    17657.6449167025  3.95057855902894  0.206539361749826  19.12748507383
37  1.49095580537939e-81  1.51795745915827e-77
MIR30A     613.735581696374  6.66475614114898  0.348909670370175  19.10166644013
63  2.4455573693544e-81   1.51795745915827e-77
ANKRD30BL  2306.08489269738  4.19799897436267  0.222470669667772  18.86989858317
85  2.01685179250905e-79  8.34573271740243e-76
RRM2       1012.72008544986  3.21630483725746  0.203776175804426  15.78351750179
23  4.04030692060258e-56  1.25390925280901e-52
MIR3687    589.413352493287  4.50184986826774  0.29287659418576   15.37114934289
49  2.55594406268351e-53  6.34589791883062e-50
HIST1H3B   1444.33335960224  2.90787104671683  0.201422769712973  14.43665505573
44  3.04253521448053e-47  6.29500535876021e-44
MIR3648    943.248473149307  3.40832276440333  0.241165508308419  14.13271237794
31  2.38767626351928e-45  4.2343733050469e-42
HIST1H1B   2348.42359361963  2.81889786479314  0.210060901962102  13.41943140519
18  4.65269250746914e-41  7.21981559846524e-38
HIST1H3F   1018.65342528922  2.63130496091174  0.199168988562413  13.21141900606
26  7.53896796410326e-40  1.03987498118198e-36
```

In [16]:
```
head -25 deseq.results.RUN.tsv | column -t > deseq.results.RUN.Top20.tsv
```

In [17]:
```
head -25 deseq.results.RUN.Top20.tsv
```

```
baseMean   log2FoldChange    lfcSE              stat               pvalue
padj
MIR663A    17657.6449167025  3.95057855902894   0.206539361749826  19.1274850738
337   1.49095580537939e-81  1.51795745915827e-77
MIR30A     613.735581696374  6.66475614114898   0.348909670370175  19.1016664401
363   2.4455573693544e-81   1.51795745915827e-77
ANKRD30BL  2306.08489269738  4.19799897436267   0.222470669667772  18.8698985831
785   2.01685179250905e-79  8.34573271740243e-76
RRM2       1012.72008544986  3.21630483725746   0.203776175804426  15.7835175017
923   4.04030692060258e-56  1.25390925280901e-52
MIR3687    589.413352493287  4.50184986826774   0.29287659418576   15.3711493428
949   2.55594406268351e-53  6.34589791883062e-50
HIST1H3B   1444.33335960224  2.90787104671683   0.201422769712973  14.4366550557
344   3.04253521448053e-47  6.29500535876021e-44
MIR3648    943.248473149307  3.40832276440333   0.241165508308419  14.1327123779
431   2.38767626351928e-45  4.2343733050469e-42
HIST1H1B   2348.42359361963  2.81889786479314   0.210060901962102  13.4194314051
918   4.65269250746914e-41  7.21981559846524e-38
HIST1H3F   1018.65342528922  2.63130496091174   0.199168988562413  13.2114190060
626   7.53896796410326e-40  1.03987498118198e-36
```

This files contains the 20 genes that have the most significantly changed gene expression.

In [28]:
```
cat deseq.results.RUN.tsv | awk '$3 > 0' > deseq.results.RUN.posFOLD.tsv
ls -l deseq.results.RUN.posFOLD.tsv
```

```
-rw-r--r-- 1 r0637408 domain users 1836089 jan 27 18:53 deseq.results.RUN.posFOL
D.tsv
```

In [30]:
```
head -20 deseq_results_RUN_resFOLD.tsv | column -t
```

```
baseMean   log2FoldChange    lfcSE             stat               pvalue
padj
MIR663A    17657.6449167025  3.95057855902894  0.206539361749826  19.12748507383
37  1.49095580537939e-81  1.51795745915827e-77
MIR30A     613.735581696374  6.66475614114898  0.348909670370175  19.10166644013
63  2.4455573693544e-81   1.51795745915827e-77
ANKRD30BL  2306.08489269738  4.19799897436267  0.222470669667772  18.86989858317
85  2.01685179250905e-79  8.34573271740243e-76
RRM2       1012.72008544986  3.21630483725746  0.203776175804426  15.78351750179
23  4.04030692060258e-56  1.25390925280901e-52
MIR3687    589.413352493287  4.50184986826774  0.29287659418576   15.37114934289
49  2.55594406268351e-53  6.34589791883062e-50
HIST1H3B   1444.33335960224  2.90787104671683  0.201422769712973  14.43665505573
44  3.04253521448053e-47  6.29500535876021e-44
MIR3648    943.248473149307  3.40832276440333  0.241165508308419  14.13271237794
31  2.38767626351928e-45  4.2343733050469e-42
HIST1H1B   2348.42359361963  2.81889786479314  0.210060901962102  13.41943140519
18  4.65269250746914e-41  7.21981559846524e-38
HIST1H3F   1018.65342528922  2.63130496091174  0.199168988562413  13.21141900606
26  7.53896796410326e-40  1.03987498118198e-36
HIST1H3G   848.885327202892  2.68921708605218  0.214346829606054  12.54610152617
92  4.17567108885407e-36  5.18367808970345e-33
HIST1H3C   495.886013819933  2.90337618699046  0.237520604184494  12.22368138106
99  2.32320853593567e-34  2.6218464331914e-31
XIST       193.158333912231  4.34282433038071  0.356601703026426  12.17836116183
35  4.05344986833827e-34  4.19329388879594e-31
ANLN       735.288618025918  2.45585390337792  0.213112357818148  11.52375173603
74  1.00153176216165e-30  9.56385791959595e-28
HIST1H2AJ  409.061790195616  2.62987414098246  0.231166973426073  11.37651327093
01  5.47449873595696e-30  4.85431623629784e-27
HIST1H2BH  849.591709187777  2.41746503393854  0.213156632012939  11.34126117076
1   8.19547758692224e-30  6.35866617275329e-27
HIST1H3J   341.77401963615   2.79308864088033  0.246180840432552  11.34567838818
3   7.79195927494915e-30  6.35866617275329e-27
HIST1H2BJ  495.31892500409   2.60216771950724  0.235137780538307  11.06656579623
24  1.82251699615717e-28  1.25692922168306e-25
HIST1H2BL  546.768269813001  2.74276602621427  0.248505594258671  11.03703936483
34  2.53240232024346e-28  1.65459170544749e-25
MYBL2      611.252806851485  3.01238707902343  0.27400972885721   10.99372307540
67  4.09670864233363e-28  2.54282705429649e-25
```

In [31]:
```
head -20 deseq_results_RUN_resFOLD.tsv | column -t > deseq_results_RUN_Top20.tsv
```

In [35]:
```
head -20 deseq_results_RUN_Top20.tsv | column -t
```

```
baseMean   log2FoldChange    lfcSE             stat               pvalue
padj
MIR663A    17657.6449167025  3.95057855902894  0.206539361749826  19.12748507383
37  1.49095580537939e-81  1.51795745915827e-77
MIR30A     613.735581696374  6.66475614114898  0.348909670370175  19.10166644013
63  2.4455573693544e-81   1.51795745915827e-77
ANKRD30BL  2306.08489269738  4.19799897436267  0.222470669667772  18.86989858317
85  2.01685179250905e-79  8.34573271740243e-76
RRM2       1012.72008544986  3.21630483725746  0.203776175804426  15.78351750179
23  4.04030692060258e-56  1.25390925280901e-52
MIR3687    589.413352493287  4.50184986826774  0.29287659418576   15.37114934289
49  2.55594406268351e-53  6.34589791883062e-50
HIST1H3B   1444.33335960224  2.90787104671683  0.201422769712973  14.43665505573
44  3.04253521448053e-47  6.29500535876021e-44
MIR3648    943.248473149307  3.40832276440333  0.241165508308419  14.13271237794
31  2.38767626351928e-45  4.2343733050469e-42
HIST1H1B   2348.42359361963  2.81889786479314  0.210060901962102  13.41943140519
18  4.65269250746914e-41  7.21981559846524e-38
HIST1H3F   1018.65342528922  2.63130496091174  0.199168988562413  13.21141900606
26  7.53896796410326e-40  1.03987498118198e-36
```

This files contains the 20 genes that have the most significantly changed gene expression, but only those with a positive log2FoldChange --> expression went up after RUNX1 knockdown.

In [100]:
```
cat deseq_results_RUN.tsv | awk '$3 < 0' > deseq_results_RUN_negFOLD.tsv
```

In [101]:
```
sed -i '1i  baseMean log2FoldChange lfcSE stat pvalue padj' deseq_results_RUN_negF
```

In [102]:
```
head -20 deseq_results_RUN_negFOLD.tsv | column -t
```

```
baseMean  log2FoldChange    lfcSE              stat               pvalue
padj
NEURL4    684.487083595738  -2.33507477501541  0.209068905033253  -11.1689243058
12   5.787753634023e-29    4.22642197722127e-26
CCNL2     2252.41318052693  -2.08661061177139  0.195806851324268  -10.6564739571
643  1.62647471221833e-26  8.77872046846883e-24
CYP1A1    78.1809129285149  -3.96343897315976  0.396823442668725  -9.98791539760
04   1.72164164783306e-23  6.67889356756239e-21
PRPF39    1205.25731423419  -1.91781125853713  0.199211386604976  -9.62701626257
953  6.14889914046102e-22  1.9572418956329e-19
PCSK2     79.749148416811   -3.69997393787923  0.39243766898804   -9.42818243574
878  4.17256343981929e-21  1.2332905367123e-18
GDF15     1362.49359278684  -1.82805381183228  0.194442635627494  -9.40150706110
772  5.37870602224722e-21  1.51752855818584e-18
MME       289.6044331914    -2.79628158108627  0.299263148164472  -9.34388880902
056  9.28593766423483e-21  2.50599196008285e-18
CDK5RAP3  1881.66963690715  -1.87409617835637  0.201420208924523  -9.30440986216
356  1.34738477241828e-20  3.55881586485118e-18
PLCG1     3385.59236642117  -1.61614669802047  0.177554521545975  -9.10225593777
395  8.84746515687227e-20  1.96129343673951e-17
ST8SIA4   980.319731098971  -1.58987885154324  0.193532095264935  -8.21506556505
155  2.12048367822143e-16  3.37483133095396e-14
SCIN      372.530682231728  -2.43686212653378  0.298492391789327  -8.16390029885
146  3.24376670524345e-16  5.09723036441673e-14
TIMP3     1681.02117987098  -1.88766545615523  0.232438779051305  -8.12112963189
581  4.6186437879571e-16   7.07849925724684e-14
EPHA4     2067.48052100966  -1.67309125640371  0.20677414126846   -8.09139501748
186  5.89852202864146e-16  8.9297868857994e-14
TSPYL2    373.872528040766  -1.94377959079083  0.240546787239219  -8.08067159449
435  6.44110537054065e-16  9.519033579749e-14
TFPI      858.265076097687  -2.26150145968266  0.280007036724137  -8.07658795343
31   6.66041274522324e-16  9.72733691990603e-14
GOLGA8A   167.119497250549  -2.26745001393169  0.281699842604525  -8.04917032600
169  8.33571739131325e-16  1.18942064018118e-13
MAPK8IP3  602.537402157244  -1.74205157437545  0.216402144825404  -8.05006612009
766  8.2749319198718e-16   1.18942064018118e-13
PLK2      7082.71811917194  -1.38001092924027  0.17292517731476   -7.98039331617
024  1.45867728939209e-15  1.97109093165529e-13
KIFC2     419.817202593929  -1.75421440125648  0.222072521916109  -7.89928617066
435  2.80505199614455e-15  3.48219154801384e-13
```

In [18]:
```
head -25 deseq_results_RUN_negFOLD.tsv | column -t > deseq_results_RUN_Top20min.tsv
```

This files contains the 20 genes that have the most significantly changed gene expression, but only those with a positive log2FoldChange --> expression went up after RUNX1 knockdown.

Deletion of the lines that contain O counts for all the samples using awk command.

In [5]:
```
awk '(NR>1){s=0; for (i=3;i<=NF;i++) s+=$i; if (s!=0)print}' all.gene.counts.RUN.ts
```

```
-rw-r--r-- 1 r0637408 domain users 527381 jan 27 20:30 all.gene.counts.RUN.notZe
ro.tsv
```

In [7]:
```
sed -i '1i  Geneid shNSB1.bam shNSB2.bam shRUNB1.bam shRUNB2.bam' all.gene.counts.R
```

In [8]:

```
Geneid shNSR1.bam shNSR2.bam shRUNR1.bam shRUNR2.bam
AL627309.1      13      14      6       15
RP11-34P13.13   9       5       5       11
AP006222.2      1       2       2       1
RP4-669L17.10   0       1       1       0
RP4-669L17.2    0       0       2       0
RP5-857K21.1    3       4       0       4
RP5-857K21.4    0       0       1       0
RP11-206L10.3   2       0       1       1
RNU6-1199P      0       1       1       0
RP11-206L10.5   0       0       1       0
RP11-206L10.2   12      11      3       13
RP11-206L10.9   18      32      12      19
AL669831.1      1       1       0       2
RP11-206L10.10  6       9       5       4
FAM87B  2       5       3       0
LINC00115       8       19      2       2
RP11-206L10.11  76      115     95      46
RP11-54O7.16    3       1       0       1
RP11-54O7.1     4       6       6       6
```

The following figures are the heatmaps that represent the differential expression of the four samples. The scale of the heatmap varies from 2 to 12, which is coloured from blue to red, respectively.

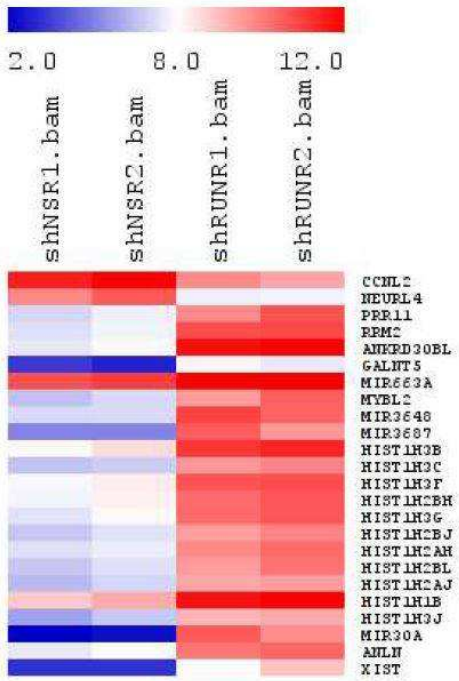

Figure 15: Heatmap representing the individual values of the four samples. We selected a top 24 of the most significantly upregulated genes after RUNX knockdown.

The function of each of these genes is described below. We selected the most important and relevant genes for breast cancer.

- MIR663A (MicroRNA 663a) and other MicroRNAs: microRNAs have an impact on both the stability and translation of mRNAs.
- MYBL2 (MYB Proto-Oncogene Like 2): involved in cell cycle progression.
- HIST1H3B (Histone Cluster 1 H3 Family Member B): the protein encoded by this gene is a histone. A histone is a basic nuclear protein that has an influence on the nucleosome structure of the chromosal fiber (in eukaryotes). Note: Multiple other histones were found to be upregulated after knock down of RUNX, such as HIST1H3C and HIST1H2BJ.

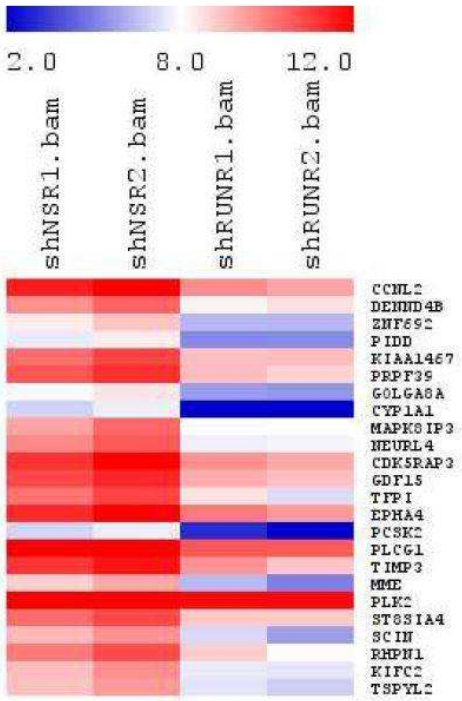

Figure 16: Heatmap representing the individual values of the four samples. We selected a top 24 of the most significantly downregulated genes after RUNX knockdown.

The function of the downregulated genes are discribed below. We selected the most relevant genes again.

- PRPF39: protein involved in pre-mRNA splicing
- GDF15: protein involved in the inflammatory and apoptotic pathways during injury and disease
- CDK5RAP3: tumor suppressor
- TIMP3: The proteins encoded by this gene family are inhibitors of the matrix metalloproteinases, a group of peptidases involved in degradation of the extracellular matrix (ECM). Expression of this gene is induced in response to mitogenic stimulation and this netrin domain-containing protein is localized to the ECM.
- TSPYL2: The encoded protein is localized to the nucleolus where it functions in chromatin remodeling and as an inhibitor of cell-cycle progression. This protein may play a role in the suppression of tumor growth.
- PLK2: This is a tumor suppressor serine/threonine-protein kinase involved in synaptic plasticity, centriole duplication and G1/S phase transition.

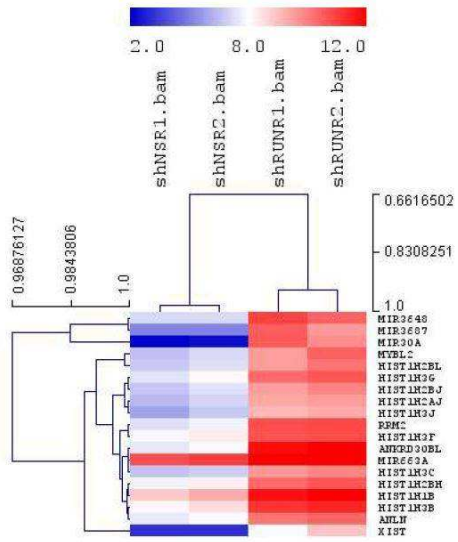

Figure 17: Heat map with added hierarchial clustering.

We selected the genes which are most important in breast cancer, or general in cancer. We didn't display all genes in the heatmap, because it wouldn't be clear. We observed a downregulation of a lot of genes, further in the list are NEAT1 and MALAT1. But on the other hand, we observed an up-regulation of genes, the most important are associated with chromatin regulation.

Further, we used the transcription factor RUNX1 to create a network of genes related to this TF in Genemania. This program gives the genes that are linked with the TF RUNX1. In figure 18, we can see that RUNX has an influence on chromatin remodelling(left column), which is the most important feature for our article, but RUNX1 is involved in many other functions. Here we retrieve the network of RUNX1 and use this as an input for iRegulon (Cytoscape). This sequence-based tool searches for Regulons. These may include TF, the targets and the motifs/tracks from a set of genes. We found RUNX1 as enriched transcription factor and made a network of related transcription factors. Figure 19 presents the output after using iRegulon. We found a list of related transcription factors to RUNX1.

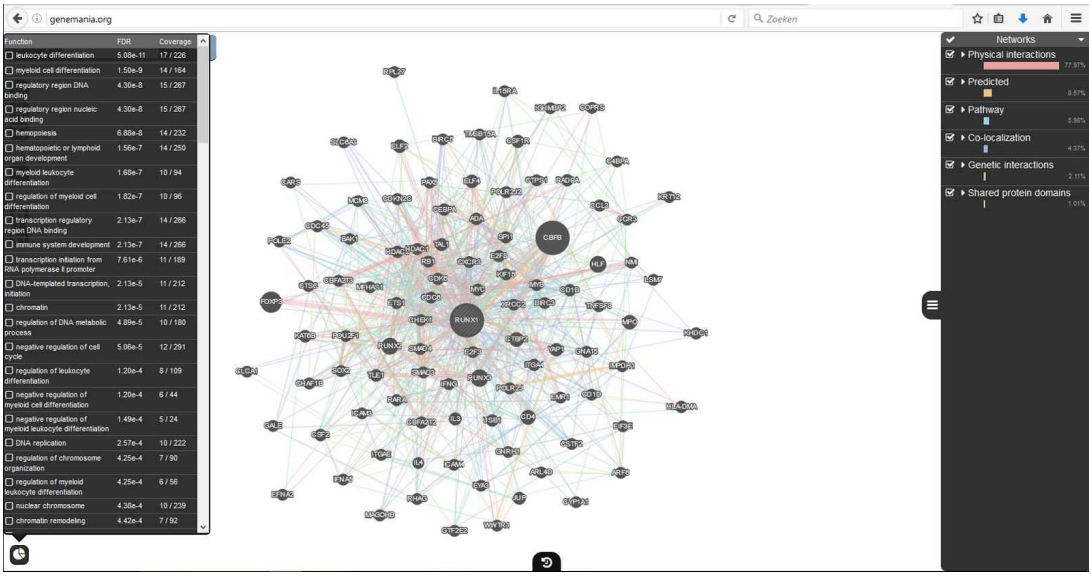

Figure 18: Transcription factor RUNX1 with its related genes in Genemania.

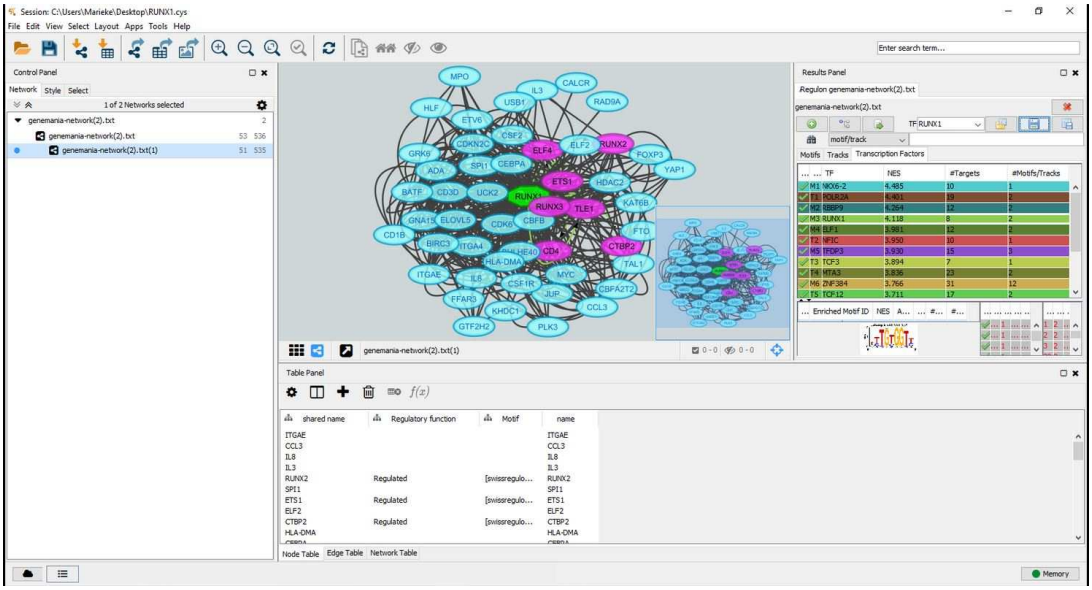

Figure 19: Sequence-based discovery of regulons for the TF RUNX1.

ChIP-seq analysis

For the ChIP-seq analysis we will explain every step we did in this jupyternotebook and will refer to the **"TaskChiP-seqFINAL.ipynb"** notebook for the output of every step (agreed by Dr. Mark Fiers).

Chromatin immunoprecipitation followed by high-throughput sequencing is the standard method to predict DNA binding sites from ChIP-seq read count data and thereby perform peak calling methods.

Aim of ChIP-seq

ChIP-seq is performed to determine where a protein binds the genome, which can be transcription factors, DNA-binding enzymes, histones, chaperones, or nucleosomes.

ChIP-seq first cross-links bound proteins to chromatin, fragments the chromatin, captures the DNA fragments bound to one protein using an RUNX1 antibody, and sequences the ends of the captured fragments using next-generation sequencing. Mapping of the sequenced DNA identifies the genomic locations of bound DNA-enzymes, modified histones, chaperones, nucleosomes and transcription factors, thereby illuminating the role of these protein-DNA interactions in gene expression and other cellular processes. The use of NGS provides relatively high solution, low noise, and high genomic coverage compared with ChIP-seq assays (ChIP followed by microarray hybridisation). ChIP-seq is the most widely used procedure for DNA-protein interaction.

We performed ChIP-seq to determine RUNX1-dependent differences in gene expression. The RNA-seq results are directly related to the binding of RUNX1.

STEP 1: downloading the ChIP-seq data

Download the ChIP-seq data from the GEO site (analogue to RNA-seq). They performed ChIP-seq on 2 replicates, namely MCF-7 wildtype RUNX1 replicate 1 and 2. We downloaded the input and pulldown data of both replicates.

SRA numbers:

- SRR2919474, replicate 1 input named S1
- SRR2919475, replicate 1 pulldown named S2
- SRR2919476, replicate 2 input named S3
- SRR2919477, replicate 2 pulldown named S4

The input files are used as control.

Note: for the output of the following commands go to the title **STEP 1: dowloading the ChIP-seq data** in **TaskChIP-seqFINAL.ipynb** jupyternotebook.

In [ ]:
```
cd ~/data/wonk
```

In [ ]:
```
ls
```

Before we map the reads to the reference genome, we should apply quality control to asses possible sequencing errors or biases. We performed quality control in bitvise with fastqc (same method as RNA-seq). See above for further details.

Overview of the data quality after fastqc

The phred quality scores are analysed for both input and pulldown data.

Check the quality of the raw data by using the FastQC toolkit. General commando: fastqc fastqfile -o outputdir

- **S1**

Analysis of the Quality control results S1:

Sequences flagged as poor quality: 0

- Per base sequence quality report: The graph represents a very good quality score of the bases, which is indicative for a good base calling. At the end of the sequence the score decreases a little bit, however the scores still remain in the green area of the y-axis. Therefore, it is still a good base calling of good quality.
- Sequences duplication levels report: The trend of the curve shows a peak on the right side of the curve, which implements more specific enrichments or the presence of contaminants.
- Overrepresented sequences report: no overrepresented sequences had been found.

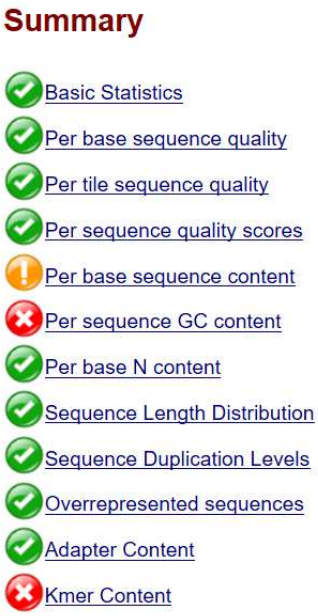

Figure 20: Summary of the quality control results of S1

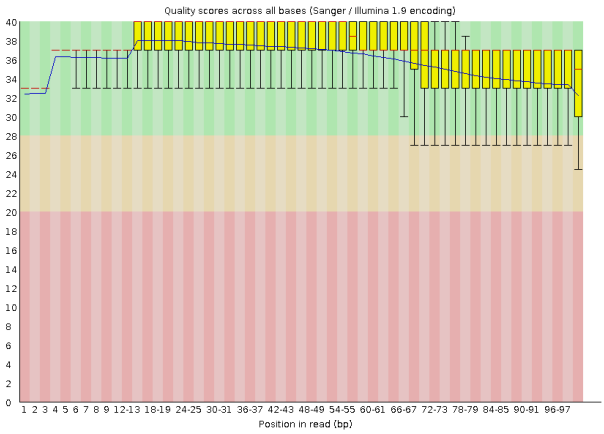

Figure 21: Per base sequence quality of S1

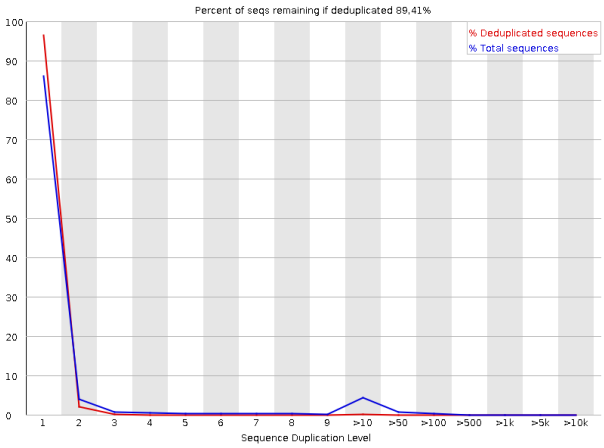

Figure 22: Sequence duplication levels of S1

- **S2**

Analysis of the Quality control results S2:

Sequences flagged as poor quality: 0

- Per base sequence quality report: The graph represents a very good quality score of the bases similar to the previous sample. At the end of the sequence the scores show a reduction of quality score, however the scores still remain in the green area of the y-axis. Therefore, it is still a good base calling of good quality.
- Sequences duplication levels report: There is a large peak on the right side, which represents more specific enrichments or the presence of contaminants
- Overrepresented sequences report: There is one overrepresented sequence in S2 of which the suggested source could be TruSeq Adapter, Index 6.

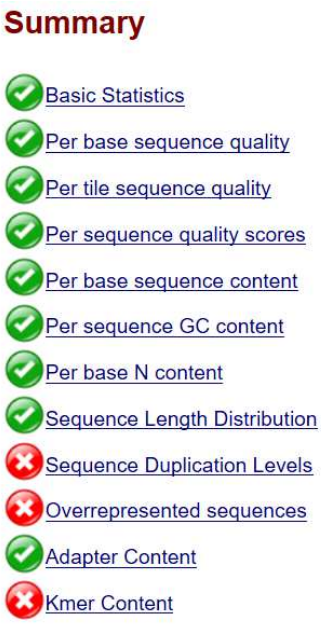

Figure 23: Summary of the quality control results of S2

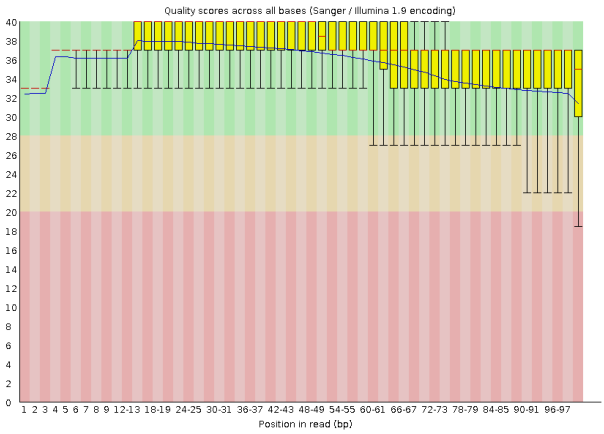

Figure 24: Per base sequence quality of S2

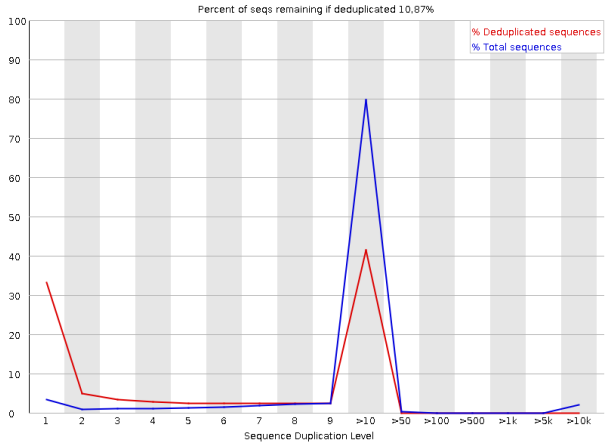

Figure 25: Sequence duplication levels of S2

- **S3**

Analysis of the Quality control results S3:

Sequences flagged as poor quality: 0

- Per base sequence quality report: A very good quality score of all bases has been found (green area). However, a lower quality score has been observed in the end of the sequence. Nevertheless, since the scores still reside in the green area, it is related with a very high quality score.
- Sequences duplication levels report: A little peak at the right side is observed. This can represent more specific enrichments or contaminations.
- Overrepresented sequences report: no overrepresented sequences had been found.

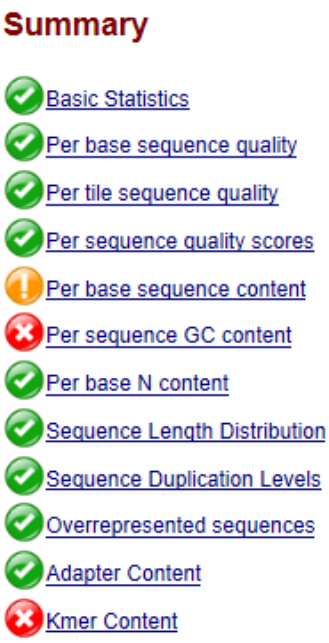

Figure 26: Summary of the quality control results of S3

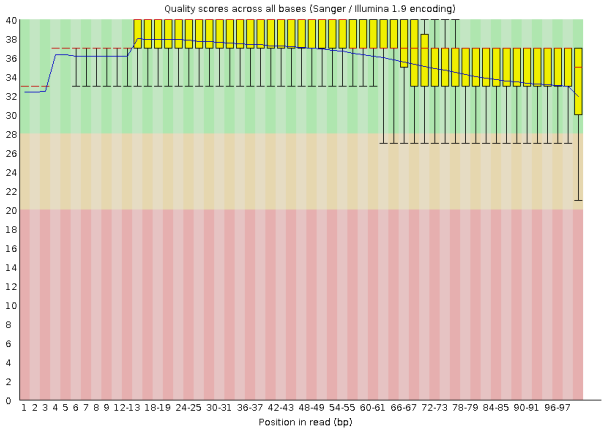

Figure 27: Per base sequence quality of S3

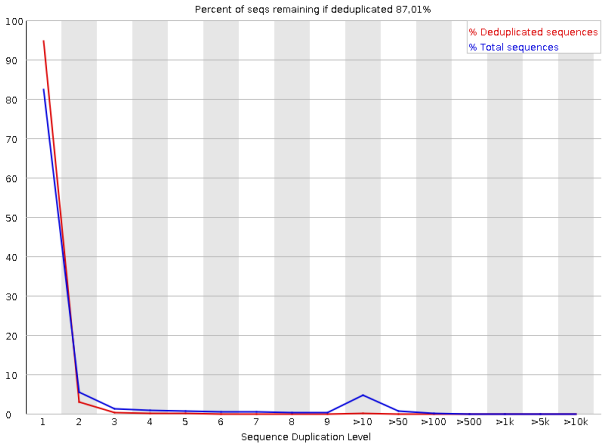

Figure 28: Sequence duplication levels of S3

- **S4**

Analysis of the Quality control results S4:

Sequences flagged as poor quality: 0

- Per base sequence quality report: The quality scores across all bases show a very high quality. Same as all previous samples: at the end of the sequence the quality score decreases. However, all scores remain in the green area, which is related with a very good base calling.
- Sequences duplication levels report: The curve shows a small peak on the right side, which is associated with more specific enrichments or contaminations.
- Overrepresented sequences report: One overrepresented sequence has been found of which the possible source could be: TruSeq Adapter, Index 10.

Summary

- Basic Statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content
- Kmer Content

Figure 29: Summary of the quality control results of S4

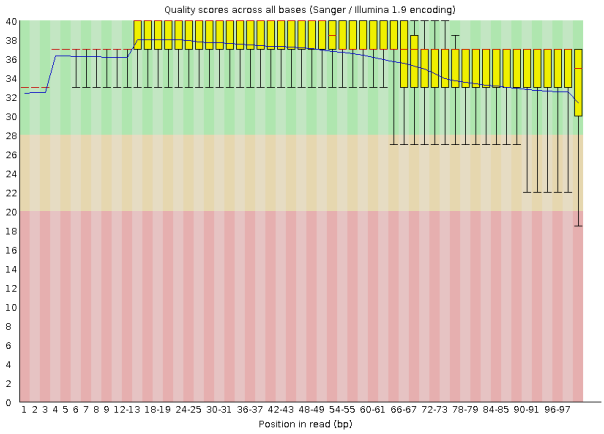

Figure 30: Per base sequence quality of S4

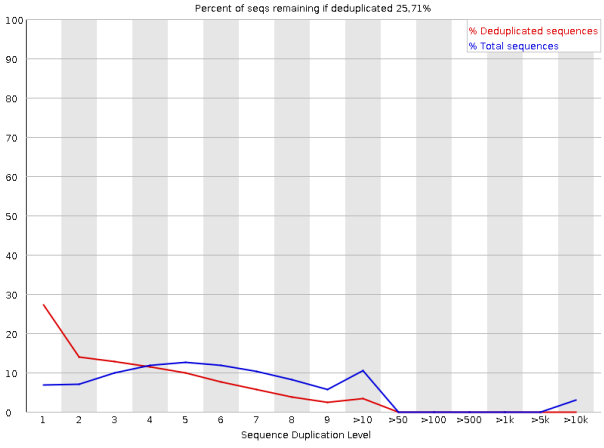

Figure 31: Sequence duplication levels of S4

As a conclusion, we decided that the overall quality was good for all reads, since no sequences were 'flagged as poor quality'. Nevertheless, multiple notifications were present in the quality report, especially those from the overrepresented sequences. However, if the overrepresented sequences are adapters, this could be solved by trimming of the reads. In practice, this step is skipped most of the time, as the reads will not be mapped anyway. In our cases, the overrepresented sequences were adapters, therefore we decided not to trim the reads.

Note: You can find a more detailed information about every report in the paragraph of RNA-seq analysis (step 3).

STEP 2: alignment with BOWTIE

The reads should be mapped using Bowtie, which allows gapped alignment, but detection of indels is not necessary for most ChIP-seq experiments. We use Bowtie for mapping, because it is ultra fast. Bowtie needs a reference genome to align each read and the genome has to be in an index for Bowtie to be able to use it. All the samples of our genome are human, so we choose the human reference genome hg19, which is not the latest but a stable version of human genome assembly.

Alignment with Bowtie2 was performed whereafter the samfiles were converted to a bam file and were sorted.

Note: for the output of the following commands go to the title **STEP 2: alignment with BOWTIE** in **TaskChIP-seqFINAL.ipynb** jupyternotebook.

In [ ]:
```
cd ~/data/work
ls
```

In [ ]:
```
bowtie2-align -x /mnt/nfs/data/resources/bowtie2/hg19 S1.fastq -S S1.sam
bowtie2-align -x /mnt/nfs/data/resources/bowtie2/hg19 S2.fastq -S S2.sam
bowtie2-align -x /mnt/nfs/data/resources/bowtie2/hg19 S3.fastq -S S3.sam
bowtie2-align -x /mnt/nfs/data/resources/bowtie2/hg19 S4.fastq -S S4.sam
```

In [ ]:
```
cd ~/data/work
ls
```

In [ ]:
```
#evaluation of the mapped reads: view, sort and index files
samtools view -S -b S1.sam > S1.bam
samtools view -S -b S2.sam > S2.bam
samtools view -S -b S3.sam > S3.bam
samtools view -S -b S4.sam > S4.bam
```

In [ ]:
```
#view the bamfiles to check the structure of the files (do this for every file)
samtools view S4.bam | head -20
```

In [ ]:
```
samtools sort -O bam -o S1.sorted.bam S1.bam
samtools sort -O bam -o S2.sorted.bam S2.bam
samtools sort -O bam -o S3.sorted.bam S3.bam
samtools sort -O bam -o S4.sorted.bam S4.bam
```

In [ ]:
```
ls
```

In [ ]:
```
samtools index S1.sorted.bam
samtools index S2.sorted.bam
samtools index S3.sorted.bam
samtools index S4.sorted.bam
```

In [ ]:
```
ls
```

Now we look at the bam files in IGV. We see peaks in every sorted bam file in RUNX and other genes. Thereafter, the regions of the genome where the chipped protein, RUNX in our case, is bound have to be predicted. Thus we have to find regions in the genome with significant numbers of mapped reads or peaks.

STEP 3: Peak calling with Macs2

Macs (model-based analysis of ChIP-seq) is a peak caller which supports both single and paired-end sequencing. It uses cross-correlation to find the lag between the reads mapped to the minus and the plus strains as the size of actual protein-DNA interacting regions.

- Remove duplicate tags
- Find candidate peaks with a significant tag enrichment
- Merge overlapping peaks
- Look at background levels and eliminate peaks that are not significant with respect to local background
- Use of the control sample to eliminate peaks that are also present there

We used the input files, S1 and S3, of replicate 1 and 2 as control samples for peak calling of the pulldown files, S2 and S4, of replicate 1 and 2. Furthermore we used the whole human genome, hs, for performing the peak calling.

Note: for the output of the following commands go to the title **STEP 3: Peak calling with Macs2** in **TaskChIP-seqFINAL.ipynb** jupyternotebook.

In [ ]:
```
macs2
```

In [ ]:
```
mkdir macs2_RUNX
```

In [ ]:
```
macs2 callpeak -t S2_sorted.bam -c S1_sorted.bam -f BAM -g hs -n RUNX
```

In [ ]:
```
macs2 callpeak -t S4_sorted.bam -c S3_sorted.bam -f BAM -g hs -n RUNX1
```

The file RUNX and RUNX1 are respectively replicate 1 and replicate 2.

In [ ]:
```
cd macs2_RUNX
ls
```

In [ ]:
```
# to count the total of peaks
cat RUNX_summits.bed | wc -l
```

In [ ]:
```
cat RUNX_peaks.narrowPeak | sort -k9 -g -r | head -5
```

In [ ]:
```
cat RUNX1_summits.bed | wc -l
```

In [ ]:
```
cat RUNX1_peaks.narrowPeak | sort -k9 -g -r | head -5
```

In [ ]:
```
# load bed files to igv
```



Figure 32: loading the bed files of RUNX in IGV

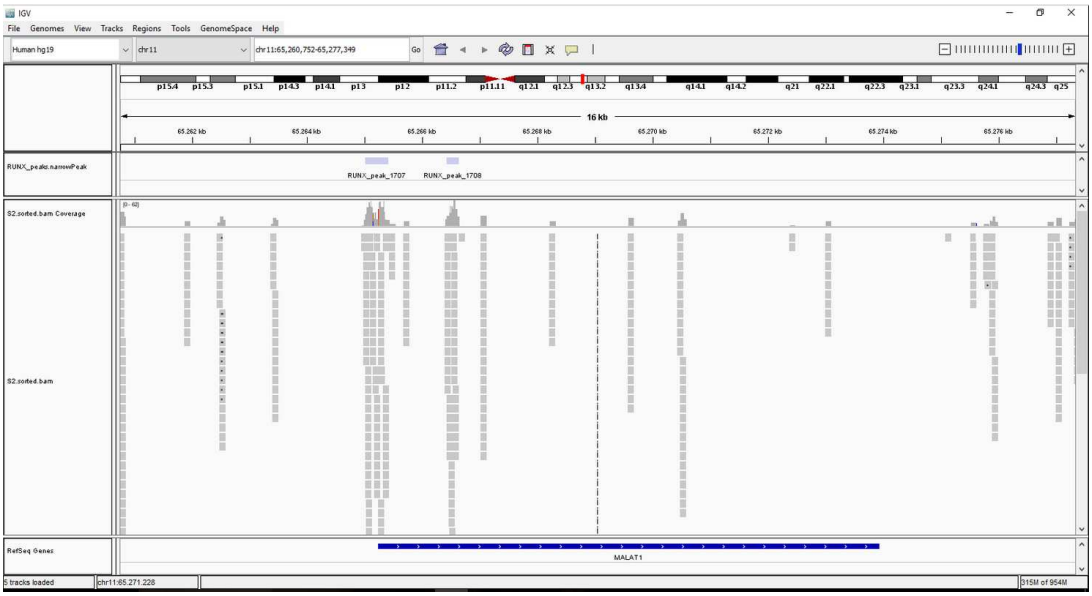

Figure 33: Loading of .bam and .bai files in IGV of S2. Zoomed in on MALAT1.



Figure 34: Loading of .bam and .bai files in IGV of S2. Zoomed in on NEAT1.

In the results, we found peaks in the promotor region of different genes. Here we show as an example MALAT1 and NEAT1 (figure 26 and 27 respectively), because these are found to be downregulated in absence of RUNX1 in RNA-seq. Now we visualise these genes in the presence of RUNX1 and it is clear that RUNX1 binds to the genes themselves or their promotor region. In this way RUNX can have an influence on the expression of these genes. The results are analogous for replicate 2.

STEP 4 : Motif analysis

Motif analysis is useful for much more than just identifying the causal DNA-binding motif in TF ChIP-seq peaks. When the motif of the ChIPed protein is already known, motif analysis provides validation of the success of the experiment. Even when the motif is not known beforehand, identifying a centrally located motif in a large fraction of the peaks by motif analysis is indicative of a successful experiment. Motif analysis can also identify the DNA-binding motifs of other proteins that bind in complex or in conjunction with the ChIPed protein, illuminating the mechanisms of transcriptional regulation. Motif analysis is also useful with histone modification ChIP-seq because it can discover unanticipated sequence signals associated with such marks.

We loaded the bed files from replicate 1 and 2 (RUNX and RUNX1 are the names of the files respectively) in IGV. We found that RUNX1 binds to the promotor regions of NEAT1 and MALAT1, which corresponds with the results of the article.



Figure 35: Screenshot of IGV output with .bam and .bamCoverage files of S2 and S4. Zoomed in on MALAT1.



Figure 36: Screenshot of IGV output with .bam and .bamCoverage files of S2 and S4. Zoomed in on NEAT1.

After loading our .bam and .bamCoverage files of both S2 and S4 in IGV, we could visually see that our data was of good quality (as there were high peaks). We also found that the peaks were higher in the promotor region of both genes NEAT1 and MALAT1. This corresponds with the findings of the article.

Further, we perfomed icis-target, you can find the links for replicate 1 and 2 Here (https://gbiomed.kuleuven.be/apps/lcb/i-cisTarget/reports/4823c681009bb3eae6bb6189fab55b96426ab95d/report.html) and Here (https://gbiomed.kuleuven.be/apps/lcb/i-cisTarget/reports/e62ba80c6d20e39357d621d3aec666bd06e7be79/report.html) respectively. Here we didn't find the list of genes we expected, therefore we searched the genes of the article in IGV and later on in UCSC.

Additionally, we used RSAT (Regulatory Sequence Analysis Tools) which is a software for the analysis of cis-regulatory elements in genome sequences. It can be used for motif discovery (used with genome-wide data sets like ChIP-seq), transcription factor binding motif analysis, comparative genomics and analysis of regulatory variations.

We used the consensus sequence of RUNX to discover motifs. The consensus sequence of RUNX is AAACCACAAA. You can find the result of RSAT Here (http://rsat.sb-roscoff.fr/tmp/apache/2017/01/26/gs-dna-pattern_2017-01-26.210204_LqJTDH.dnapat).

STEP 5: enhancer-to-gene association

We have linked the peaks to genes with UCSC, you can find the link to the report Here (https://genome.ucsc.edu/cgi-bin/hgTracks?db=hg19&lastVirtModeType=default&lastVirtModeExtraState=&virtModeType=default&virtMode=0&nonVirtPosition=&position=chr11%3A64073044%2D64084210&hgsid=576694635_ZuucNUI0Otw93sy1QbNEoDMiFaSK) .

Although, we did not find the genes from the article via icis-target, we searched the set of genes from the article. Here we found that our gene RUNX1 has an influence on the expression level of the following set of genes: MALAT1, NEAT1, EBF1, EGR1, ESRR, FOXP3.

Thus these results confirm again an influence on the expression of both MALAT1 and NEAT1 (and the other genes listed above) after the binding of RUNX.

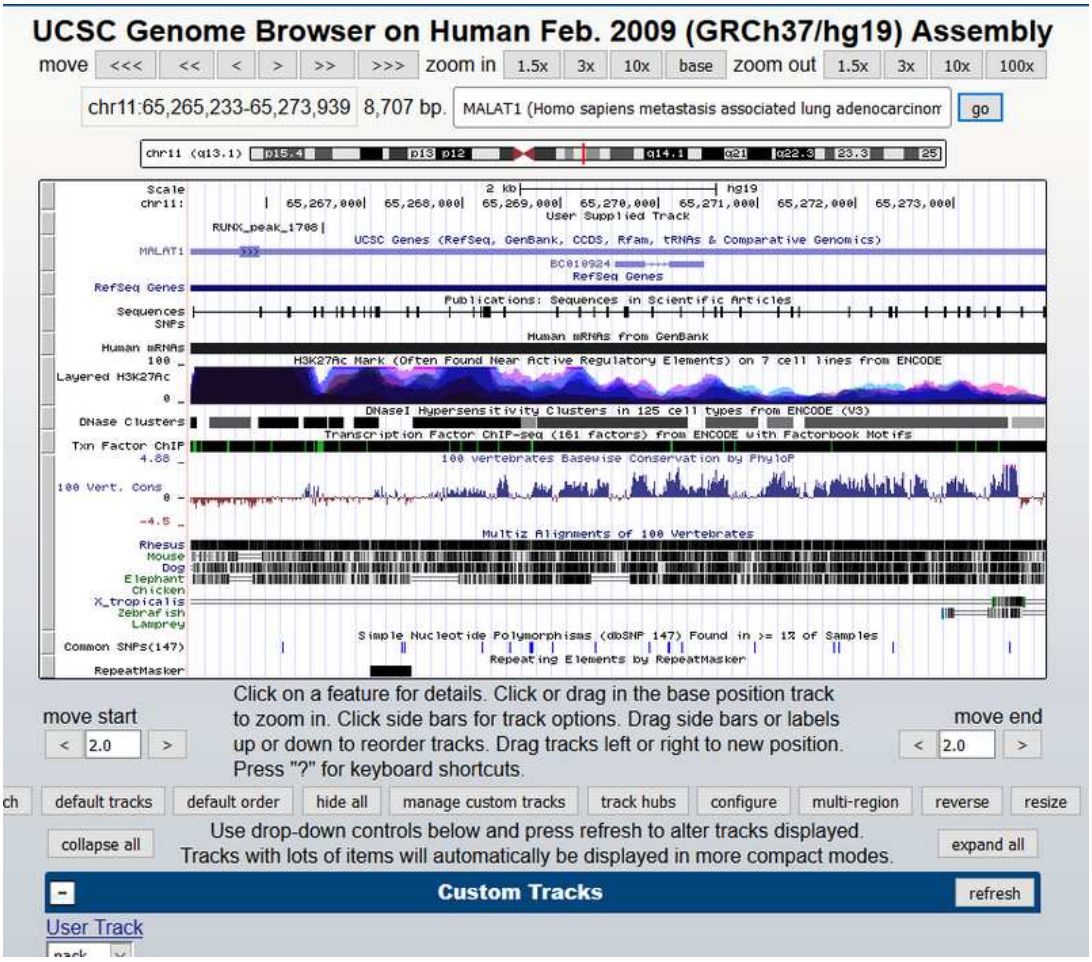

Figure 37: UCSC report with a close up of the MALAT1 gene

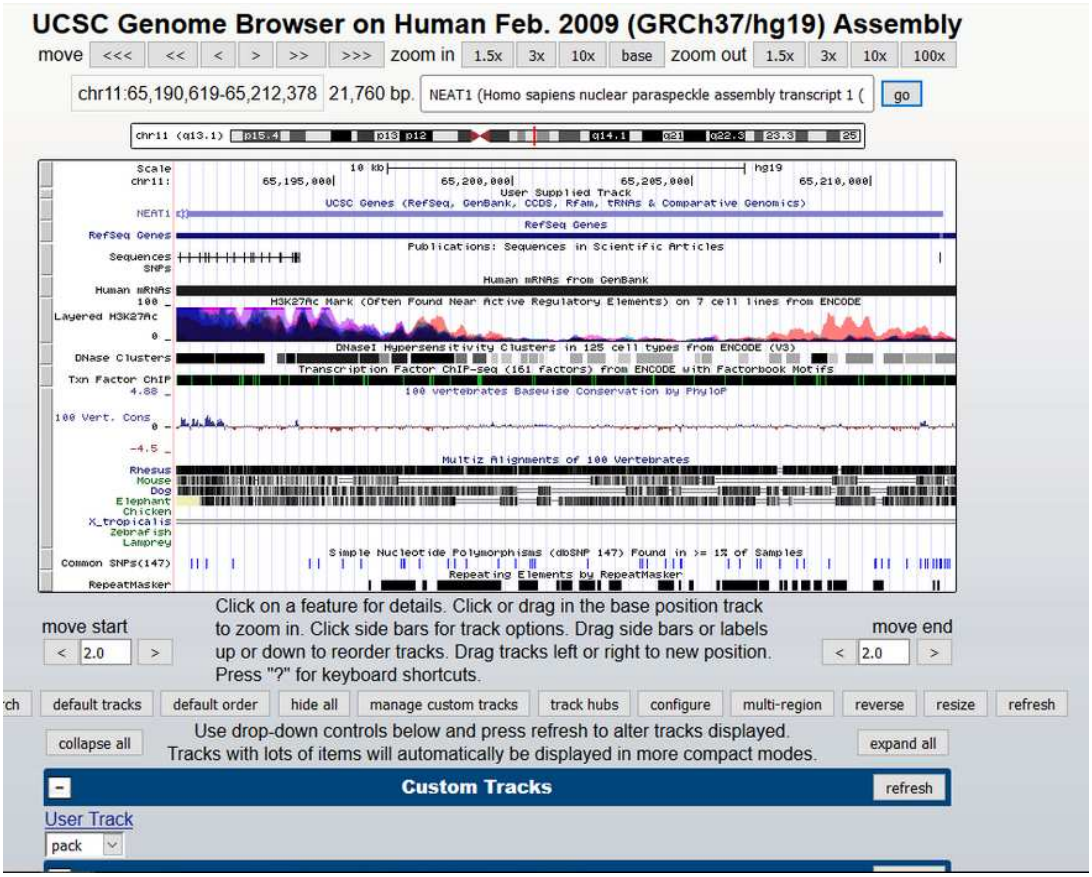

Figure 38: UCSC report with a close up of the NEAT1 gene

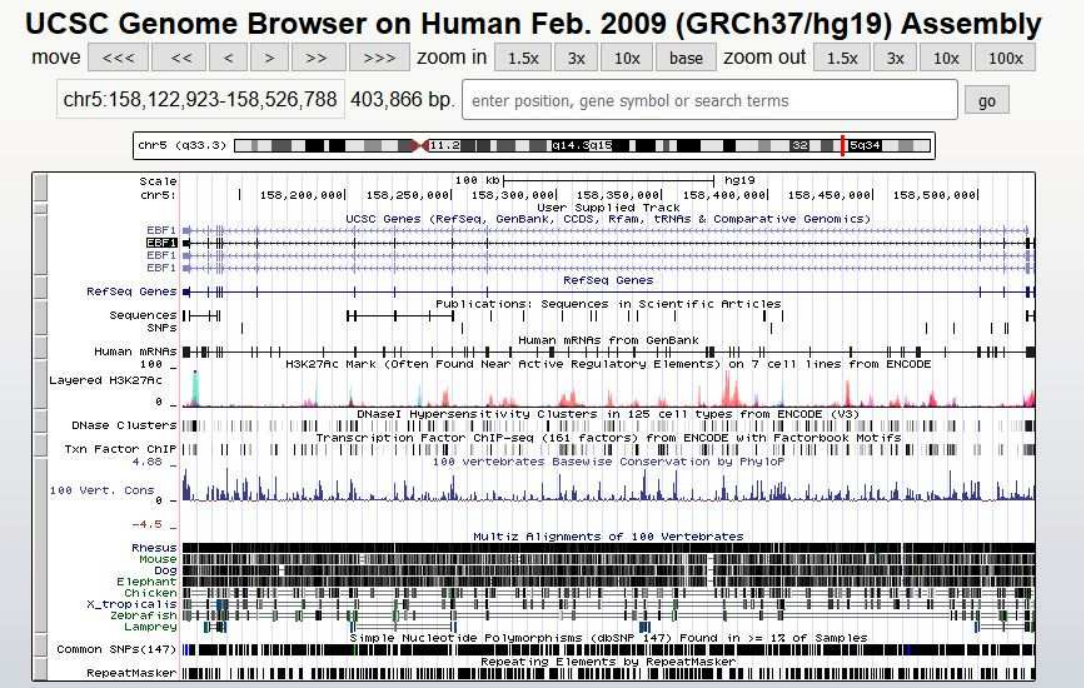

Figure 39: UCSC report with a close up of the EBF1 gene

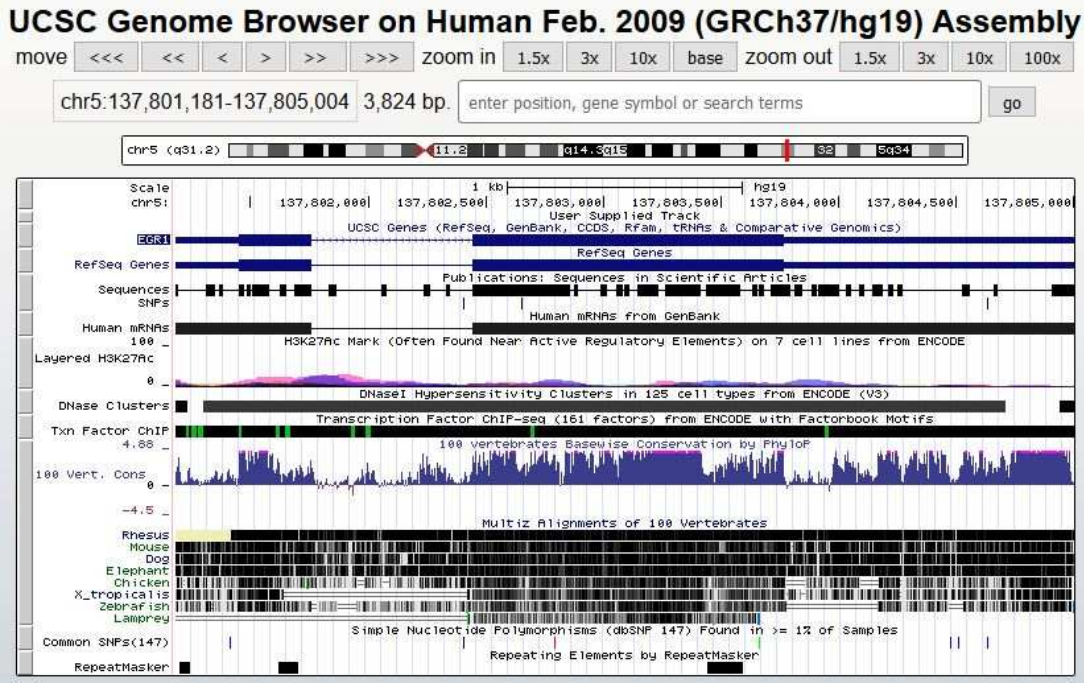

Figure 40: UCSC report with a close up of the EGR1 gene

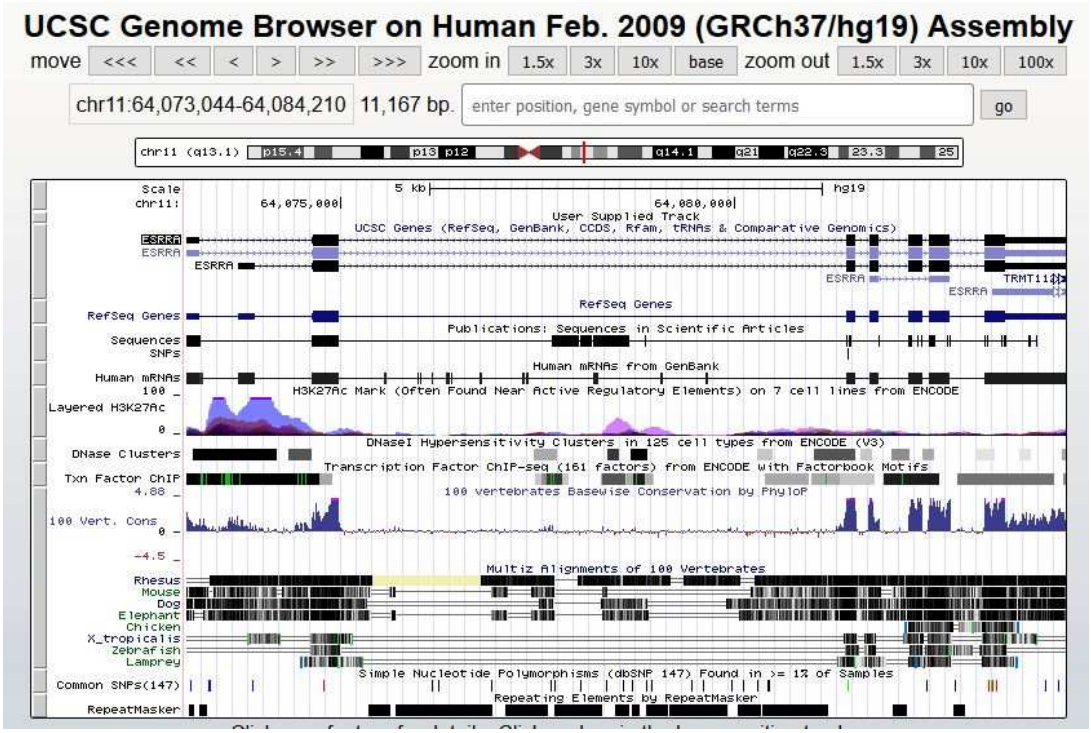

Figure 41: UCSC report with a close up of the ESRR gene

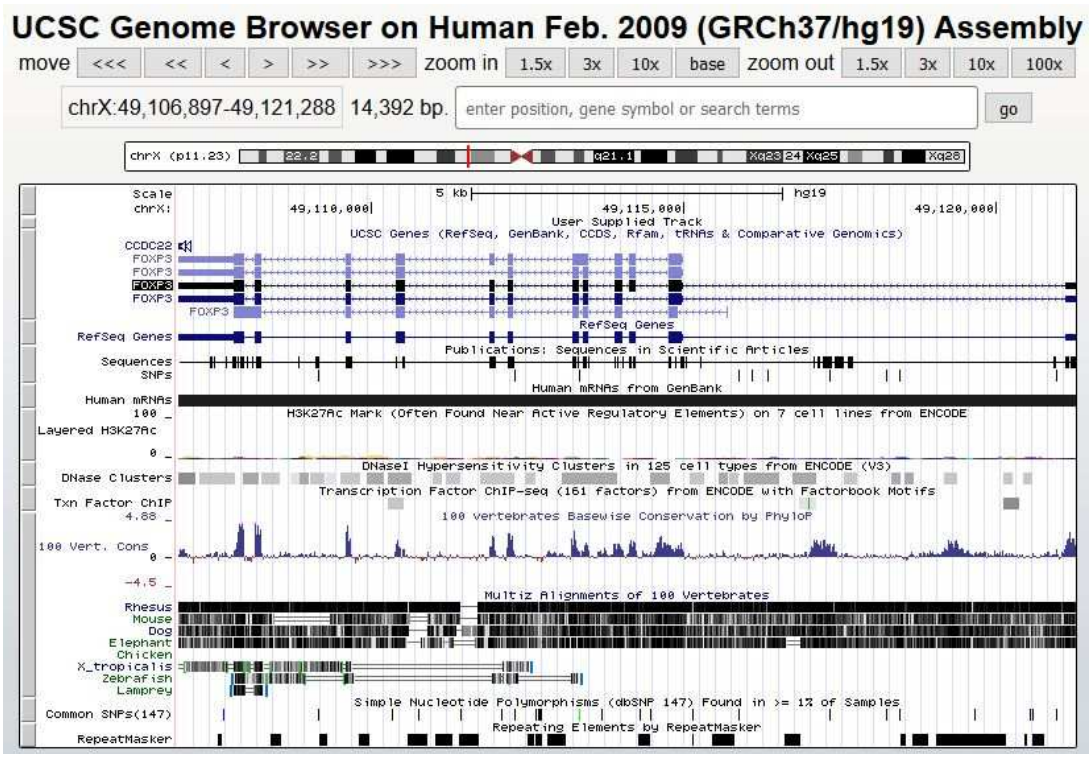

Figure 42: UCSC report with a close up of the FOXP3 gene

STEP 6: link ChIP-seq data to RNA-seq data

We loaded both RNA-seq data and ChIP-seq data in IGV. Based on the ChIP-seq data, we found that RUNX binds to the promotor region of NEAT1 and MALAT1. Additionally, from the RNA-seq data we could conclude that this binding of RUNX to both genes (NEAT1 and MALAT1) stimulates the expression of these genes (see Figure 32 and 33).

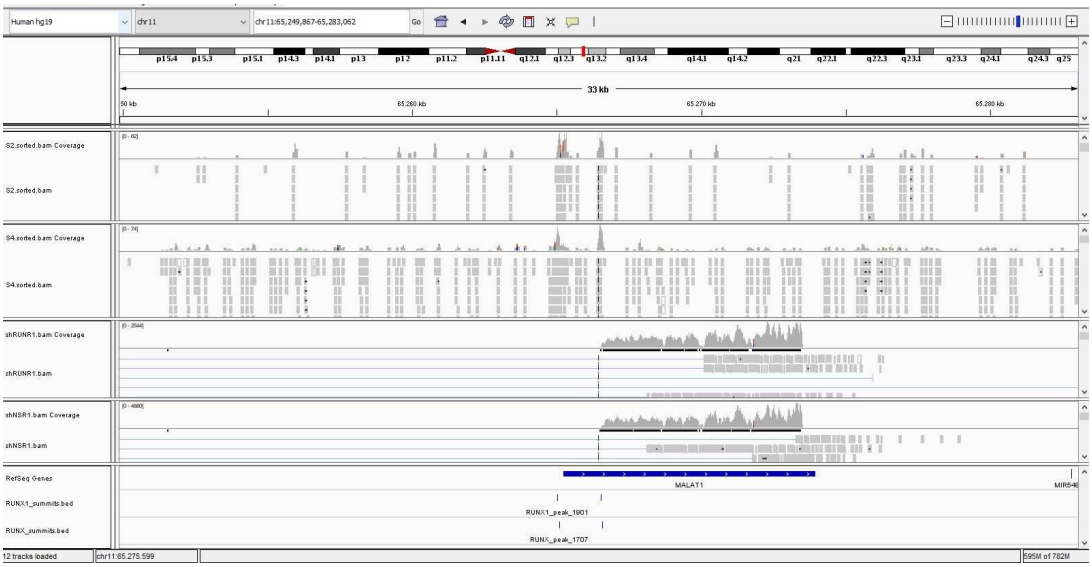

Figure 43: Both RNA-seq and ChIP-seq data loaded in IGV. Zoomed in at MALAT1 gene.

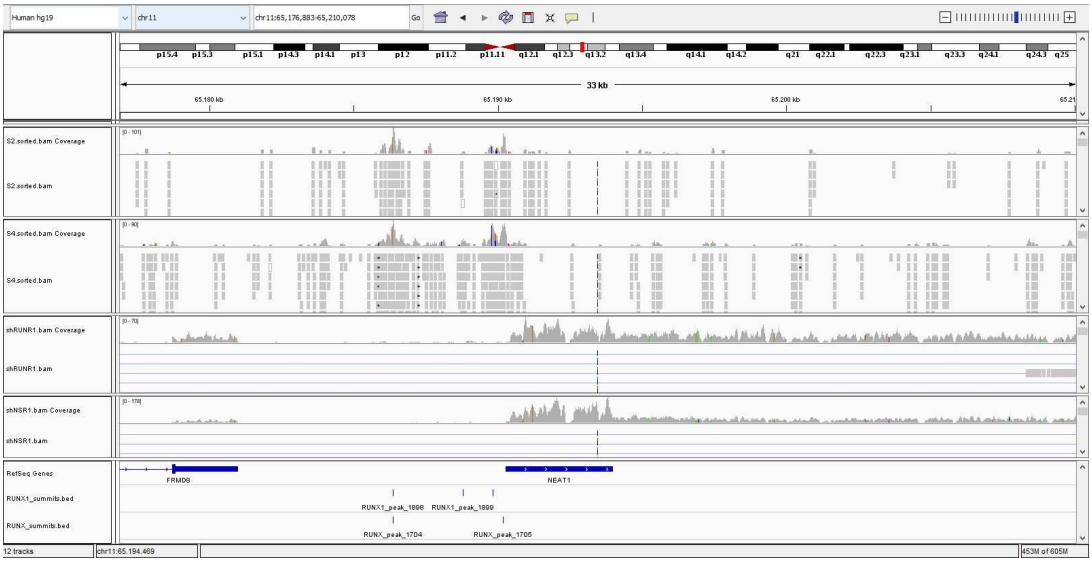

Figure 44: Both RNA-seq and ChIP-seq data loaded in IGV. Zoomed in at NEAT1 gene.

Conclusion

When we merge the RNA-seq data and the ChIP-seq data, we see the same result. Thus RUNX transcription factor has an influence on the expression level of different genes, which is proved by comparing the wild type sample with the sample where RUNX1 is knocked down in breast cancer MCF-7 cells. RUNX1 has a positive regulation on the expression of NEAT1 and MALAT1, which are involved in the nuclear structure and in cancer. Further ChIP-seq data showed that RUNX1 influences the expression of different genes by binding to their promotor regions, especially MALAT1 and NEAT1. By performing ChIP-seq we see that RUNX1 is important in chromatin regulation and folding. Thus different genes are upregulated and downregulated upon the protein binding of RUNX1. In breast cancer, through the dysfunction of RUNX1 there are effects on the nuclear structure and chromatine folding.

References:

http://www.bioinformatics.babraham.ac.uk/ (http://www.bioinformatics.babraham.ac.uk/)

http://www.sthda.com/english/wiki/rna-sequencing-data-analysis-counting-normalization-and-differential-expression (http://www.sthda.com/english/wiki/rna-sequencing-data-analysis-counting-normalization-and-differential-expression)