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 factor 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, 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 repsresses target genes expression depending on its target partners.

Cancer is characterized with changes in the nucleus, 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 compared to 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.

Hypothesis of the study: to understand the relationship between the RUNX1-dependent alterations in higher-order chromatin structure and RUNX1-altered gene expression in breast cancer cells.

- RNAseq: they suppressed RUNX1 in a human MCF-7 breast cancer cell line and evaluated the gene expression
- ChIP-seq: performed to gain insight into RUNX1-mediated regulation of chromatin organization. They probed RUNX1 binding by performing ChIP-seq

ChIP-sequencing

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 antibody specific to it, 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 proceses. 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.

STEP 1: downloading the ChIP-seq data

Download the ChIP-seq data from the GEO site. 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.

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

In [2]:
```
ls
```

```
~               S1_fastqc.zip  S2_fastqc.html  S3_fastqc.html  S4_fastqc.html
S1.fastq        S1.sam         S2_fastqc.zip   S3_fastqc.zip   S4_fastqc.zip
S1_fastqc.html  S2.fastq       S3.fastq        S4.fastq
```

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.

Overview of the data quality after fastqc

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

Figure 1: S1 fastqc data (replicate 1 input)

Figure 2: S2 fastqc data (replicate 1 pulldown)

Figure 3: S3 fastqc data (replicate 2 input)

Figure 4: S4 fastqc data (repliate 2 pulldown)

We checked the reads of our data to check the quality (figure 1-4). We concluded that the overall quality is good for the input and pulldown data of replicate 1 and 2. therefore we worked further with the files without trimming the data first.

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 is. 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 is performed whereafter the samfiles are converted to a bam file and they are sorted.

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

```
~                S2.bam             S3.bam             S4.bam
S1.bam           S2.fastq           S3.fastq           S4.fastq
S1.fastq         S2_fastqc.html     S3_fastqc.html     S4_fastqc.html
S1_fastqc.html   S2_fastqc.zip      S3_fastqc.zip      S4_fastqc.zip
S1_fastqc.zip    S2.sam             S3.sam             S4.sam
S1.sam           S2.sorted.bam      S3.sorted.bam      S4.sorted.bam
S1.sorted.bam    S2.sorted.bam.bai  S3.sorted.bam.bai
```

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 [6]:
```
bowtie2-align -x /mnt/nfs/data/resources/bowtie2/hg19 S2.fastq -S S2.sam
```

In [14]:
```
bowtie2-align -x /mnt/nfs/data/resources/bowtie2/hg19 S1.fastq -S S1.sam
bowtie2-align -x /mnt/nfs/data/resources/bowtie2/hg19 S4.fastq -S S4.sam
```

```
30075142 reads; of these:
  30075142 (100.00%) were unpaired; of these:
    2226261 (7.40%) aligned 0 times
    18605200 (61.86%) aligned exactly 1 time
    9243681 (30.74%) aligned >1 times
92.60% overall alignment rate
29988988 reads; of these:
  29988988 (100.00%) were unpaired; of these:
    1801619 (6.01%) aligned 0 times
    21932161 (73.13%) aligned exactly 1 time
    6255208 (20.86%) aligned >1 times
93.99% overall alignment rate
```

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

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

```
~                S1.sam           S2.sam           S3.sam           S4.sam
S1.fastq         S2.fastq         S3.fastq         S4.fastq
S1_fastqc.html   S2_fastqc.html   S3_fastqc.html   S4_fastqc.html
S1_fastqc.zip    S2_fastqc.zip    S3_fastqc.zip    S4_fastqc.zip
```

In [3]:
```
#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
```

```
[W::sam_read1] parse error at line 22324899
[main_samview] truncated file.
```

In [15]:
```
samtools view -S -b S1.sam > S1.bam
samtools view -S -b S4.sam > S4.bam
```

In [7]: samtools view S4.bam | head -20

```
SRR2919477.1    16      chr2    201828614       42      101M    *       0
0       TTGTTTGGATCTGAGGAAGTTTGCACAGTAGCTGGCTGAGTGGTAGCTGTTTGTCTTGTAGCAATCTTTAAA
AAGTGTTCTTTGGGGTCGGGCGCAGTGTN   BBBBBBBBBBBBB<BBBBBBBFBBBBBBBB<<B7BBBFBFFFFBB0FF
FFFIFFFFFIFFFBFBFFFFFBBFFFBFBFFFBBFFFBIFF<B7F7FFF<00#   AS:i:-1 XN:i:0  XM:i:1
XO:i:0  XG:i:0  NM:i:1  MD:Z:100C0      YT:Z:UU
SRR2919477.2    0       chr3    61375604        42      101M    *       0
0       NTATGTCATAGATATTTTACTACAATAGAAAAGGCAAATATAGAAAGAGAAAGAGATGACTAGTTTTGTTCA
CCACTGTATGCGCAAGGGCTACAGCAATA   #0<FFFFFFFFFFFFFFFFBIIIIIIFIFIIIIIIFFIFFFIIFFIII
FFIIIIIFBFFFFIIIIIIIIFIIIFIIIFFFFFFFBFBBBBBBBBBBBBBBB   AS:i:-1 XN:i:0  XM:i:1
XO:i:0  XG:i:0  NM:i:1  MD:Z:0T100      YT:Z:UU
SRR2919477.3    16      chr14   73491894        33      101M    *       0
0       AGCCTCCCAAGGAGCTGGGATTACACGCGCTAGCCACAATGCCCAGCTAATTTTTGTATTTTTAGTAGAGAC
GGGGTTTCACCATGTTGGCCAGGCTGGTN   7FB<BBB<BBFFBBBBFFFFFB<B7BBBFBFFFFFFFFFIIIIIIIII
IIIIIFIIIFFIIIIIIIIFIIFIIIIIIIIIFIIIIIIIFFFFFFFFFF<0#   AS:i:-1 XS:i:-60
XN:i:0  XM:i:1  XO:i:0  XG:i:0  NM:i:1  MD:Z:100C0      YT:Z:UU
SRR2919477.4    16      chr20   49356059        42      101M    *       0
0       TAAGCATTAGACCACTCCCAGGGCTCTGGAATTCCACTGATTATGTATCTTTGCTTTCCTGGGTTTGGAGGA
ACTAAAGTCCACCAACAAAATAAACCTCC   FFBBBBBBFBBB7<BFFFFFBFFFFFFFFFFFFFIIFFIIIIIFIF<III
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIFIIIIIIFFFFFFFFFFBBB   AS:i:0  XN:i:0  XM:i:0
XO:i:0  XG:i:0  NM:i:0  MD:Z:101        YT:Z:UU
SRR2919477.5    0       chr1    149832555       1       101M    *       0
0       NAGATCCGCTTAACGCCGCCACGCCGAGCTAGACGCCGAATGGCAGGCTTGGTGATGCCCTGAATGTTGTCT
CTCAAGACCTTGCGGTGGCGCTTAGCGCC   #0<FFFFFFFFFFIIIIIIIIIIIIIIIIIIIIIIFFFFFFFFFFFFFF
FFFFBBBFFFFFFFBFBFFFFFFFFFBFFFFFBBFFFFFFBFFFFFBFFBFFBFF   AS:i:-1 XS:i:-1 XN:i:0
XM:i:1  XO:i:0  XG:i:0  NM:i:1  MD:Z:0G100      YT:Z:UU
SRR2919477.6    0       chr11   66273531        37      101M    *       0
0       TGCCCCAGTCTCCCGAGTAGCTGGTATAACAGGCGCCCGCCACCACACCCAGCTAATTTTTGTATTTTTTTT
AGTAGAGACGGGGTTTTACTATGTTGGCT   BB<BBFB<B0BBBFF<B<F<FBFB<BFFBFF<BBF7BFBBBFFFFF<F
BF<B<<B'<BBBFB<7BBFBBBBBBBBBFBB<<<<BBB77<BBBFFB0<00B0   AS:i:0  XS:i:-48
XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:101        YT:Z:UU
SRR2919477.7    4       *       0       0       *       *       0       0
GATCGGAAGAGCACACGTCTGAACTCCAGTCACTCGAAGATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAAAAGAAAA
CACGGAGGGATGCCGCAGGCG   BBBFFFBFFFFFFFIIIFFFFFFIIFFFFFFFBFFFFFIFFFI<FBFFFFFFFFFF
FFIFFFF<<7<777B70<0'00'0'07707'070'0''''''700   YT:Z:UU
SRR2919477.8    16      chr2    187296576       27      101M    *       0
0       GGTAGAAACCTGGGTGTTGGTAGCTGGGTCCCTCATGAGCATGGAGAAGGCACTGAAGCAGCTGGAAGTGCA
GAGCACCAAGAAGCAGTGAGCTTTTGCTN   BBFBBBBB77B7BBBBBBBFFBFFFFFFB<07<<BFB<FFFFIFFFFB
B<FBFFFFFFFBFFFFFBFFIIIFBBFBFBFIFFF<FFFFFFBFFBBFFF<0#   AS:i:-1 XS:i:-55
XN:i:0  XM:i:1  XO:i:0  XG:i:0  NM:i:1  MD:Z:100G0      YT:Z:UU
SRR2919477.9    0       chr5    108632692       42      101M    *       0
0       NTGGAAGCCATTATTCTAAGTGAAGTAACTAAGGAATGGAACATCAAATACTGCACATTCTCACTTACAAGG
GGGAGCTAAGCTATGGGTACACAAAGCCT   #0<FFFFFFFFFFIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIFI
IIIIIIIIIIIIIIIIIIIIIIIIIIFFFFFFFFFFFFFFFFFFFFFFFFFB   AS:i:-1 XN:i:0  XM:i:1
XO:i:0  XG:i:0  NM:i:1  MD:Z:0C100      YT:Z:UU
SRR2919477.10   16      chr4    116809655       42      101M    *       0
0       ACAAATTATTGAAGTGACTATGAGACATTAGTGGTCTTTTAAAATAATACAACATTCAGGAACTAATCAACT
TACCCCAAAATTATTTTTAGCATTGAAAN   <FFFFFFFFFFFFFFFFFFFFFFFFIIIFIIIIIIFFIIIIIIIIII
IIIFIIIIIFIIFFIIIIIIFFFFFBIIIIIIIIIIFFFFFFFFFFFFFF<0#   AS:i:-1 XN:i:0  XM:i:1
XO:i:0  XG:i:0  NM:i:1  MD:Z:100A0      YT:Z:UU
SRR2919477.11   0       chr8    135362563       38      101M    *       0
0       TCTTGTGCCTGTTTTCAAATGGAATGCTTCCAGCTTTTGCCCATTCAGTATGATATTGGCTGTGGATTTGTC
ATAAATGGCTATTATTTTGAGTATGTTCC   BBBFFFFFFFFFFFIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
IFIIFIIIIIIIIIFFFBFIIIFFIIIIIIIIIFFFFFFFFFFBFBFBFFFFB   AS:i:0  XS:i:-49
XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:101        YT:Z:UU
SRR2919477.12   0       chr4    151332999       42      101M    *       0
0       NCAATCATTGCATAAAATCTGTGGAATTGAAGTGGTAAACCAAAAAGAGGAGGGGGATGAAACTCTTATGTG
ACAATATGCTAACCAATACATAAGATACA   #0<FFFFFFFFFFFFIIIIIIIFFFFIIIIIIIIIIFFIIIIIIIIIF
FIBFIIIIFFFFFFFFFFFFFFFFFFFFBFFFFFFFFFFFFFFFFFFFFFB   AS:i:-1 XN:i:0  XM:i:1
XO:i:0  XG:i:0  NM:i:1  MD:Z:0G100      YT:Z:UU
SRR2919477.13   0       chr3    132101967       42      101M    *       0
0       NCCTCCTATGCTCATCAGAGTCCAAAGTCCAGAGGGGGCCAAAGCAGCAGGGGCTGGCATGTCAGCACTGCT
CCAAGTGTTTATACTCCCAGCCGGGCTGC   #0<FFFFFFFFFFIIIIIIIIIIIIIIIFIIIIIIIIIIIIIIIIIII
IIIIFFFFFFFFFFBFFFFFFBBBFFBFBBBFFFFFFFBFFFFBFFFFFFFFB   AS:i:-1 XN:i:0  XM:i:1
XO:i:0  XG:i:0  NM:i:1  MD:Z:0T100      YT:Z:UU
SRR2919477.14   0       chrX    125452454       42      101M    *       0
0       NCTGTTCAGTTATTCGACCCAACACTGCATATGATTCTCTCTTATCTTTGCTAATTTAACAGGTAAAAATGT
ATTCCTATTTGTGTTTGTTTGACTATTGG   #0<FFFFFFFFFFIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
```

In [8]:
```
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
```

```
[bam_sort_core] merging from 13 files...
[W::bam_hdr_read] EOF marker is absent. The input is probably truncated.
[W::bam_hdr_read] EOF marker is absent. The input is probably truncated.
[bam_merge_core] 'S1.bam.tmp.0005.bam' is truncated. Continue anyway.
[bam_merge_core] 'S1.bam.tmp.0004.bam' is truncated. Continue anyway.
[bam_sort_core] merging from 13 files...
[W::bam_hdr_read] EOF marker is absent. The input is probably truncated.
[bam_merge_core] 'S2.bam.tmp.0004.bam' is truncated. Continue anyway.
[bam_sort_core] merging from 5 files...
[bam_sort_core] merging from 10 files...
```

In [16]:
```
samtools sort -O bam -o S1.sorted.bam S1.bam
samtools sort -O bam -o S4.sorted.bam S4.bam
```

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

In [9]:
```
ls
```

```
~               S1.sam          S2_fastqc.zip  S3_fastqc.html  S4.fastq
S1.bam          S1.sorted.bam   S2.sam         S3_fastqc.zip   S4_fastqc.html
S1.fastq        S2.bam          S2.sorted.bam  S3.sam          S4_fastqc.zip
S1_fastqc.html  S2.fastq        S3.bam         S3.sorted.bam   S4.sam
S1_fastqc.zip   S2_fastqc.html  S3.fastq       S4.bam          S4.sorted.bam
```

In [10]:
```
samtools index S1.sorted.bam
samtools index S2.sorted.bam
samtools index S3.sorted.bam
samtools index S4.sorted.bam
```

```
[W::bam_hdr_read] EOF marker is absent. The input is probably truncated.
samtools index: "S1.sorted.bam" is corrupted or unsorted
samtools index: "S4.sorted.bam" is corrupted or unsorted
```

In [17]:
```
samtools index S1.sorted.bam
samtools index S4.sorted.bam
```

In [18]:
```
ls
```

```
~                   S2.bam              S3.fastq            S4_fastqc.html
S1.bam              S2.fastq            S3_fastqc.html      S4_fastqc.zip
S1.fastq            S2_fastqc.html      S3_fastqc.zip       S4.sam
S1_fastqc.html      S2_fastqc.zip       S3.sam              S4.sorted.bam
S1_fastqc.zip       S2.sam              S3.sorted.bam       S4.sorted.bam.bai
S1.sam              S2.sorted.bam       S3.sorted.bam.bai
S1.sorted.bam       S2.sorted.bam.bai   S4.bam
S1.sorted.bam.bai   S3.bam              S4.fastq
```

In [ ]:
```
# look at bam files in IGV
```

Then the regions of the genome where the chiped protein is bound has to be predicted by finding regions with significant numbers of mapped reads (peaks). The input DNA can be used as a control sample for the use of the mapped reads.

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 strans 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 eliminates peaks that are not significant with respect to local background
- Uses the control sample to eliminate peaks that are also present there

In [19]:
```
macs2
```

```
usage: macs2 [-h] [--version]
             {callpeak,bdgpeakcall,bdgbroadcall,bdgcmp,bdgopt,cmbreps,bdgdiff,fi
lterdup,predictd,pileup,randsample,refinepeak}
             ...
macs2: error: too few arguments
```

In [20]:
```
#macs2 gebruikt en de sorted files ingevoegd met telkens -c S1 en S3 als de control
mkdir macs2_RUNX
macs2 callpeak -t  S2.sorted.bam -c S1.sorted.bam -f BAM -g hs -n RUNX
```

```
INFO  @ Tue, 24 Jan 2017 12:16:35:
# Command line: callpeak -t S2.sorted.bam -c S1.sorted.bam -f BAM -g hs -n RUNX
# ARGUMENTS LIST:
# name = RUNX
# format = BAM
# ChIP-seq file = ['S2.sorted.bam']
# control file = ['S1.sorted.bam']
# effective genome size = 2.70e+09
# band width = 300
# model fold = [5, 50]
# qvalue cutoff = 5.00e-02
# 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  @ Tue, 24 Jan 2017 12:16:35: #1 read tag files...
INFO  @ Tue, 24 Jan 2017 12:16:35: #1 read treatment tags...
INFO  @ Tue, 24 Jan 2017 12:16:41:  1000000
INFO  @ Tue, 24 Jan 2017 12:16:45:  2000000
INFO  @ Tue, 24 Jan 2017 12:16:51:  3000000
INFO  @ Tue, 24 Jan 2017 12:16:56:  4000000
INFO  @ Tue, 24 Jan 2017 12:17:01:  5000000
INFO  @ Tue, 24 Jan 2017 12:17:06:  6000000
INFO  @ Tue, 24 Jan 2017 12:17:12:  7000000
INFO  @ Tue, 24 Jan 2017 12:17:17:  8000000
INFO  @ Tue, 24 Jan 2017 12:17:23:  9000000
INFO  @ Tue, 24 Jan 2017 12:17:28:  10000000
INFO  @ Tue, 24 Jan 2017 12:17:32:  11000000
INFO  @ Tue, 24 Jan 2017 12:17:38:  12000000
INFO  @ Tue, 24 Jan 2017 12:17:43:  13000000
INFO  @ Tue, 24 Jan 2017 12:17:47:  14000000
INFO  @ Tue, 24 Jan 2017 12:17:52:  15000000
INFO  @ Tue, 24 Jan 2017 12:17:57:  16000000
INFO  @ Tue, 24 Jan 2017 12:18:02:  17000000
INFO  @ Tue, 24 Jan 2017 12:18:07:  18000000
INFO  @ Tue, 24 Jan 2017 12:18:11:  19000000
INFO  @ Tue, 24 Jan 2017 12:18:17:  20000000
INFO  @ Tue, 24 Jan 2017 12:18:21:  21000000
INFO  @ Tue, 24 Jan 2017 12:18:26:  22000000
INFO  @ Tue, 24 Jan 2017 12:18:29:  23000000
INFO  @ Tue, 24 Jan 2017 12:18:32: #1.2 read input tags...
INFO  @ Tue, 24 Jan 2017 12:18:39:  1000000
INFO  @ Tue, 24 Jan 2017 12:18:45:  2000000
INFO  @ Tue, 24 Jan 2017 12:18:51:  3000000
INFO  @ Tue, 24 Jan 2017 12:18:58:  4000000
INFO  @ Tue, 24 Jan 2017 12:19:06:  5000000
INFO  @ Tue, 24 Jan 2017 12:19:12:  6000000
INFO  @ Tue, 24 Jan 2017 12:19:17:  7000000
INFO  @ Tue, 24 Jan 2017 12:19:23:  8000000
INFO  @ Tue, 24 Jan 2017 12:19:29:  9000000
INFO  @ Tue, 24 Jan 2017 12:19:35:  10000000
INFO  @ Tue, 24 Jan 2017 12:19:41:  11000000
INFO  @ Tue, 24 Jan 2017 12:19:46:  12000000
INFO  @ Tue, 24 Jan 2017 12:19:53:  13000000
INFO  @ Tue, 24 Jan 2017 12:20:00:  14000000
INFO  @ Tue, 24 Jan 2017 12:20:06:  15000000
INFO  @ Tue, 24 Jan 2017 12:20:12:  16000000
INFO  @ Tue, 24 Jan 2017 12:20:18:  17000000
INFO  @ Tue, 24 Jan 2017 12:20:25:  18000000
INFO  @ Tue, 24 Jan 2017 12:20:31:  19000000
INFO  @ Tue, 24 Jan 2017 12:20:37:  20000000
INFO  @ Tue, 24 Jan 2017 12:20:42:  21000000
INFO  @ Tue, 24 Jan 2017 12:20:47:  22000000
INFO  @ Tue, 24 Jan 2017 12:20:53:  23000000
INFO  @ Tue, 24 Jan 2017 12:20:59:  24000000
```

In [21]: `macs2 callpeak -t S4.sorted.bam -c S3.sorted.bam -f BAM -g hs -n RUNX1`

```
INFO  @ Tue, 24 Jan 2017 12:25:21:
# Command line: callpeak -t S4.sorted.bam -c S3.sorted.bam -f BAM -g hs -n RUNX1
# ARGUMENTS LIST:
# name = RUNX1
# format = BAM
# ChIP-seq file = ['S4.sorted.bam']
# control file = ['S3.sorted.bam']
# effective genome size = 2.70e+09
# band width = 300
# model fold = [5, 50]
# qvalue cutoff = 5.00e-02
# 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  @ Tue, 24 Jan 2017 12:25:21: #1 read tag files...
INFO  @ Tue, 24 Jan 2017 12:25:21: #1 read treatment tags...
INFO  @ Tue, 24 Jan 2017 12:25:25:  1000000
INFO  @ Tue, 24 Jan 2017 12:25:29:  2000000
INFO  @ Tue, 24 Jan 2017 12:25:33:  3000000
INFO  @ Tue, 24 Jan 2017 12:25:37:  4000000
INFO  @ Tue, 24 Jan 2017 12:25:42:  5000000
INFO  @ Tue, 24 Jan 2017 12:25:46:  6000000
INFO  @ Tue, 24 Jan 2017 12:25:50:  7000000
INFO  @ Tue, 24 Jan 2017 12:25:54:  8000000
INFO  @ Tue, 24 Jan 2017 12:25:58:  9000000
INFO  @ Tue, 24 Jan 2017 12:26:02:  10000000
INFO  @ Tue, 24 Jan 2017 12:26:06:  11000000
INFO  @ Tue, 24 Jan 2017 12:26:10:  12000000
INFO  @ Tue, 24 Jan 2017 12:26:15:  13000000
INFO  @ Tue, 24 Jan 2017 12:26:19:  14000000
INFO  @ Tue, 24 Jan 2017 12:26:23:  15000000
INFO  @ Tue, 24 Jan 2017 12:26:27:  16000000
INFO  @ Tue, 24 Jan 2017 12:26:31:  17000000
INFO  @ Tue, 24 Jan 2017 12:26:35:  18000000
INFO  @ Tue, 24 Jan 2017 12:26:39:  19000000
INFO  @ Tue, 24 Jan 2017 12:26:43:  20000000
INFO  @ Tue, 24 Jan 2017 12:26:47:  21000000
INFO  @ Tue, 24 Jan 2017 12:26:51:  22000000
INFO  @ Tue, 24 Jan 2017 12:26:56:  23000000
INFO  @ Tue, 24 Jan 2017 12:27:00:  24000000
INFO  @ Tue, 24 Jan 2017 12:27:04:  25000000
INFO  @ Tue, 24 Jan 2017 12:27:08:  26000000
INFO  @ Tue, 24 Jan 2017 12:27:12:  27000000
INFO  @ Tue, 24 Jan 2017 12:27:16:  28000000
INFO  @ Tue, 24 Jan 2017 12:27:19:  29000000
INFO  @ Tue, 24 Jan 2017 12:27:22: #1.2 read input tags...
INFO  @ Tue, 24 Jan 2017 12:27:27:  1000000
INFO  @ Tue, 24 Jan 2017 12:27:32:  2000000
INFO  @ Tue, 24 Jan 2017 12:27:37:  3000000
INFO  @ Tue, 24 Jan 2017 12:27:42:  4000000
INFO  @ Tue, 24 Jan 2017 12:27:47:  5000000
INFO  @ Tue, 24 Jan 2017 12:27:52:  6000000
INFO  @ Tue, 24 Jan 2017 12:27:58:  7000000
INFO  @ Tue, 24 Jan 2017 12:28:03:  8000000
INFO  @ Tue, 24 Jan 2017 12:28:08:  9000000
INFO  @ Tue, 24 Jan 2017 12:28:13:  10000000
INFO  @ Tue, 24 Jan 2017 12:28:16: #1 tag size is determined as 101 bps
INFO  @ Tue, 24 Jan 2017 12:28:16: #1 tag size = 101
INFO  @ Tue, 24 Jan 2017 12:28:16: #1  total tags in treatment: 28187369
INFO  @ Tue, 24 Jan 2017 12:28:16: #1 user defined the maximum tags...
INFO  @ Tue, 24 Jan 2017 12:28:16: #1 filter out redundant tags at the same loca
tion and the same strand by allowing at most 1 tag(s)
INFO  @ Tue, 24 Jan 2017 12:28:17: #1  tags after filtering in treatment: 720388
9
INFO  @ Tue, 24 Jan 2017 12:28:17: #1  Redundant rate of treatment: 0.74
INFO  @ Tue, 24 Jan 2017 12:28:17: #1  total tags in control: 9872550
```

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

In [22]:
```
# to count the total of peaks
cat RUNX_summits.bed | wc -l
```

```
10914
```

In [23]:
```
cat RUNX_peaks.narrowPeak | sort -k9 -g -r | head -5
```

```
chr1    145208751       145209328       RUNX_peak_528   2841    .       59.26764
        292.91473       284.12222       282
chr1    121483988       121485495       RUNX_peak_515   2094    .       9.78081
216.87074       209.49983       1269
chr15   70755033        70755782        RUNX_peak_3148  1937    .       73.76989
        200.79581       193.73245       394
chr16   12009687        12010180        RUNX_peak_3456  1641    .       63.54353
        170.92520       164.10767       246
chr2    11638662        11639087        RUNX_peak_5169  1491    .       60.41454
        155.78967       149.14822       194
sort: write failed: standard output: Broken pipe
sort: write error
```

In [24]:
```
cat RUNX1_summits.bed | wc -l
```

```
12812
```

In [25]:
```
cat RUNX1_peaks.narrowPeak | sort -k9 -g -r | head -5
```

```
chr16   12009727        12010253        RUNX1_peak_4271 3975    .       109.1363
3       407.07169       397.58020       242
chr2    11638620        11639103        RUNX1_peak_6318 3414    .       100.2632
9       348.93518       341.46072       250
chr1    145208737       145209479       RUNX1_peak_542  3389    .       38.02754
        346.41064       338.97223       305
chr1    205180321       205181031       RUNX1_peak_974  3062    .       72.78316
        313.35925       306.24615       521
chr15   50646618        50647924        RUNX1_peak_3640 2878    .       69.33554
        294.79651       287.85281       980
sort: write failed: standard output: Broken pipe
sort: write error
```

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

Figure 5: loading of bed files to IGV

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. (invoegen foto's van IGV)

We perfomed icis-target, you can find the links for respectively 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).

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 haven't found 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 expressionlevel of this set of genes.

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

overall conclusion of the experiment (chip seq data en rna-seq data beiden in IGV --> neat en malat)