Report Bioinformatics 2016-2017 - Katrien Neyrinck & Eleonora Persoons

We start our report with making a new directory in our jupyternotebook (account of Eleonora Persoons: r0384602) called 'groepswerk' where we will store all our data. This notebook (Bash) will serve as our main report, while the R-notebook will serve as an addendum.

In [11]:

```
mkdir -p /mnt/storage/$USER/jupyternotebooks/groepswerk
cd /mnt/storage/$USER/jupyternotebooks/groepswerk
```

Introduction

Transactive response DNA-binding protein (TDP-43) is a nuclear protein that can bind RNA and DNA in most tissues. TDP-43 is involved in RNA metabolism as it regulates transcription and splicing within the cell. In addition, it plays a role in miRNA maturation, RNA transport and it is a subunit of the stress granule to protect mRNA during stress conditions. In the neurodegenerative diseases amyotrophic lateral sclerosis (ALS) and frontotemporal dementia (FTD), often TDP-43 aggregates are found within the cytosol of affected neurons. As a consequence, loss of the nuclear function of TDP-43 is observed as well as gain of new toxic functions within the cytoplasm which contributes to the degeneration of the neuron.

Within this report, we searched for articles that performed knockdown of TDP-43. The observed effects upon TDP-43 silencing, could give us more insights in the pathologic mechanism of neurodegeneration in ALS and FTD.

Microarray data analysis

- PubMed: https://www.ncbi.nlm.nih.gov/pubmed/?term=Knockdown+of+transactive+response+DNA-binding+protein+TDP-43+downregulates+histone+deacetylase+6 (https://www.ncbi.nlm.nih.gov/pubmed/?term=Knockdown+of+transactive+response+DNA-binding+protein+TDP-43+downregulates+histone+deacetylase+6)
- GeoData: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE18632 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE18632)

In the article mentioned above, they performed knockdown of TDP-43 in HEK293E cells via RNA silencing using siRNA against TDP-43 compared to scrambled siRNA (negative control). Afterwards, microarray analysis was performed to asses the genes that are differentially expressed upon TDP-43 knockdown. The article states to have found histone deacetylase 6 (HDAC6) to be significantly downregulated when TDP-43 is silenced. HDAC6 is an cytoplasmic enzyme that connects ubiquitinated/misfolded proteins to the microtubules inducing their retrograde transport and subsequent degradation via autophagy. So in

situations of HDAC6 deficiency, this mechanism is impaired inducing an increased proteotoxicity to the cell. This can be one possible mechanism contributing to the neurodegeneration seen in ALS and FTD characterized by TDP-43 loss of function.

We now try to reproduce this microarray data analysis performed in this article.

First, we download the package 'Affy'. See our R-jupyternotebook, COMMANDO 1.

In order to upload the celfiles into our directory, groepswerk, we use the following commando and we check whether the files are indeed there.

In [12]:

```
cd /mnt/storage/$USER/jupyternotebooks/groepswerk/ARRAY_GSE18632_RAW/
ls
```

```
Data_norm.eset    GSM463017.CEL.gz  GSM463020.CEL.gz
GSM463015.CEL.gz  GSM463018.CEL.gz  GSM463021.CEL.gz
GSM463016.CEL.gz  GSM463019.CEL.gz  GSM463022.CEL.gz
```

We group these 8 gz-files under the name 'celfiles'. See COMMANDO 2. Because the 'celfiles' only contain a list of names, we now load the raw data as well. See COMMANDO 3.

Annotation of the samples was performed, using COMMANDO 4.

The next step is to merge the sampleNames with the actual tranfection used on the HEK293E cells. See COMMANDO 5.

We convert this dataframe into an annotated dataframe. See COMMANDO 6.

Since we have 8 different samples, we normalise our data. See COMMANDO 7.

The current sample names are not very informative to us. Therefore we change them manually. See COMMANDO 8.

Now we quickly check the normalized expression levels for a few probes. See COMMANDO 9.

Before we can go further with the workflow, two libraries need to be loaded. See COMMANDO 10.

Now gene names are assigned to the probes. See COMMANDO 11.

To make our data table more informative, we add an extra (last) column with the gene symbols. See COMMANDO 12.

Some of the probes do not have a gene symbol assigned to them. Therefore we will remove them out of the dataset. See COMMANDO 13

As one gene can have multiple probes, we will just take the median expression in further analysis. See COMMANDO 14.

A gene count table was made, using COMMANDO 15.

Because the column gene.symbols is 'just a column', we now want to link each gene symbol to a row. See COMMANDO 16.

An important step in this work flow, is to check the quality of the data. Therefore we make a MDS plot. See COMMANDO 17. The control samples (scrambled) are grouped together on the left side of the graph. The TDP-43 samples on the other hand, are located on the right handside. -> So there is a lot of difference between the control and treated samples. -> Within the contol and treated samples, there is some difference. For example, scrambled_replicate 2 and 3 are quite different from the other two. -> This indicates some variance within the groups.

We now check the expression levels of HDAC6 and we indeed see that the values are lower upon TDP-43 knockdown. See COMMANDO 18. In order to know if this difference in expression is significant, we will perform a differential expression analysis using Limma (cf. infra).

First, we need to define the different experimental conditions. See COMMANDO 19.

We build a model matrix, because Limma uses a linear model in the background. See COMMANDO 20.

For the differential expression analysis, a new library Limma is needed. Next, we make a MA plot to get a general overview of our differential expression analysis. The dots in the upper half (above the orange line) symbolise upregulated genes, those in the lower half downregulated genes. See COMMANDO 21.

As 8 different samples are used, we will correct for multiple testing. See COMMANDO 22.

We use the Benjamini Hochberg method with a cut off value of 0.05. -> the expression of 3515 genes is significantly different after TDP-43 knockdown.

Now, we make a tsv-file from this differential expression analysis. See COMMANDO 23.

When we assess the tsv-file in Excel, we see that HDAC6 has a log-fold-change of -0.25 and a adjusted p-value of 0.013. So according to the cut-offs we chose, HDAC6 is indeed significantly downregulated when TDP-43 is silenced.

Extra: When we run commando 23, the first few genes of the TopTable are shown. Here, we find TARDBP, which is the gene for TDP-43. This is a good control for the transfection that was performed: TDP-43 is indeed silenced (LogFC=-1.832799 ; p adj.=2.675694e-08).

RNASeq Analysis

- PubMed: https://www.ncbi.nlm.nih.gov/pubmed/?term=Distinct+and+shared+functions+of+ALS-associated+TDP-43%2C+FUS%2C+and+TAF15+revealed+by+comprehensive+multi-system+integrative+analyses (https://www.ncbi.nlm.nih.gov/pubmed/?term=Distinct+and+shared+functions+of+ALS-associated+TDP-43%2C+FUS%2C+and+TAF15+revealed+by+comprehensive+multi-system+integrative+analyses)
- GeoData: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE77702 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE77702)

In the article mentioned above, TDP-43 knockdown is performed in an identical way (RNA silencing). However, a more clinical relevant celltype is used, namely motor neurons generated from induced pluripotent stem cells from healthy humans. To check the effect on gene expression levels after TDP-43 silencing, RNASeq is perfomed.

We redo the RNASeq analysis and see if HDAC6 is also significantly downregulated upon TDP-43 knockdown in this setting.

Before we can start this analysis, we need to get RNAseq fastq data from the NCBI GEO/SRA db. (Commando given by prof.)

In [16]:

```
mkdir -p ~/data/ncbi
vdb-config -s /repository/user/main/public/root=~/data/ncbi
mkdir -p ~/data/work && cd ~/data/work
```

In [1]:

```
fastq-dump --split-files SRR3153260
fastq-dump --split-files SRR3153259
fastq-dump --split-files SRR3153256
fastq-dump --split-files SRR3153255
```

```
Read 29905668 spots for SRR3153260
Written 29905668 spots for SRR3153260
Read 44386427 spots for SRR3153259
Written 44386427 spots for SRR3153259
Read 33675077 spots for SRR3153256
Written 33675077 spots for SRR3153256
Read 22513222 spots for SRR3153255
Written 22513222 spots for SRR3153255
```

In [2]:

```
cd /mnt/storage/$USER/jupyternotebooks
mv SRR3153260_1.fastq TDP-43_2.fastq
mv SRR3153259_1.fastq TDP-43_1.fastq
mv SRR3153256_1.fastq scr_2.fastq
mv SRR3153255_1.fastq scr_1.fastq
```

We now move the fastq files to the appropriate directory.

In [4]:

```
mkdir -p /mnt/storage/$USER/jupyternotebooks/groepswerk/RNAseq
cd /mnt/storage/$USER/jupyternotebooks

mv TDP-43_2.fastq /mnt/storage/$USER/jupyternotebooks/groepswerk/RNAseq
mv TDP-43_1.fastq /mnt/storage/$USER/jupyternotebooks/groepswerk/RNAseq
mv scr_2.fastq /mnt/storage/$USER/jupyternotebooks/groepswerk/RNAseq
mv scr_1.fastq /mnt/storage/$USER/jupyternotebooks/groepswerk/RNAseq

cd /mnt/storage/$USER/jupyternotebooks/groepswerk/RNAseq
ls -l
```

```
mv: cannot stat ‘TDP-43_2.fastq’: No such file or directory
mv: cannot stat ‘TDP-43_1.fastq’: No such file or directory
mv: cannot stat ‘scr_2.fastq’: No such file or directory
mv: cannot stat ‘scr_1.fastq’: No such file or directory
total 44127588
-rw-r--r-- 1 r0384602 domain users  7789707940 nov 16 12:12 scr_1.fa
stq
-rw-r--r-- 1 r0384602 domain users 11662802940 nov 16 11:59 scr_2.fa
stq
-rw-r--r-- 1 r0384602 domain users 15379403148 nov 16 11:56 TDP-43_1
.fastq
-rw-r--r-- 1 r0384602 domain users 10354709946 nov 16 11:53 TDP-43_2
.fastq
```

It's important to check the quality of the data. We do this via FastQC and then look at per base sequence quality and overrepresented sequences.

In [7]:

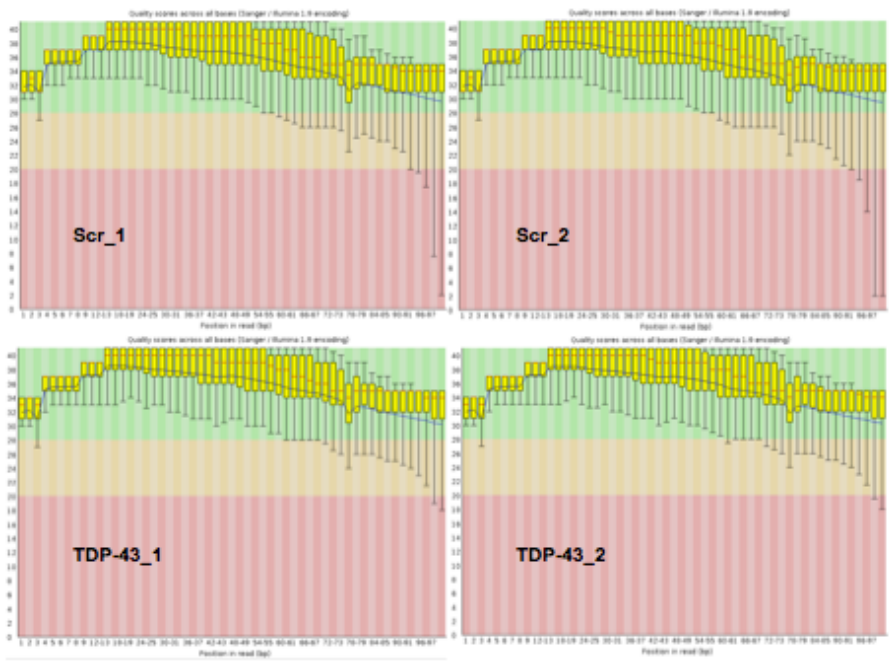
```
cd /mnt/storage/$USER/jupyternotebooks/groepswerk/RNAseq
/usr/local/bin/FastQC/fastqc -o . TDP-43_2.fastq
/usr/local/bin/FastQC/fastqc -o . TDP-43_1.fastq
/usr/local/bin/FastQC/fastqc -o . scr_2.fastq
/usr/local/bin/FastQC/fastqc -o . scr_1.fastq
```

```
Started analysis of TDP-43_2.fastq
Approx 5% complete for TDP-43_2.fastq
Approx 10% complete for TDP-43_2.fastq
Approx 15% complete for TDP-43_2.fastq
Approx 20% complete for TDP-43_2.fastq
Approx 25% complete for TDP-43_2.fastq
Approx 30% complete for TDP-43_2.fastq
Approx 35% complete for TDP-43_2.fastq
Approx 40% complete for TDP-43_2.fastq
Approx 45% complete for TDP-43_2.fastq
Approx 50% complete for TDP-43_2.fastq
Approx 55% complete for TDP-43_2.fastq
Approx 60% complete for TDP-43_2.fastq
Approx 65% complete for TDP-43_2.fastq
Approx 70% complete for TDP-43_2.fastq
Approx 75% complete for TDP-43_2.fastq
Approx 80% complete for TDP-43_2.fastq
Approx 85% complete for TDP-43_2.fastq
```

```
Approx 90% complete for TDP-43_2.fastq
Approx 95% complete for TDP-43_2.fastq
Analysis complete for TDP-43_2.fastq
Started analysis of TDP-43_1.fastq
Approx 5% complete for TDP-43_1.fastq
Approx 10% complete for TDP-43_1.fastq
Approx 15% complete for TDP-43_1.fastq
Approx 20% complete for TDP-43_1.fastq
Approx 25% complete for TDP-43_1.fastq
Approx 30% complete for TDP-43_1.fastq
Approx 35% complete for TDP-43_1.fastq
Approx 40% complete for TDP-43_1.fastq
Approx 45% complete for TDP-43_1.fastq
Approx 50% complete for TDP-43_1.fastq
Approx 55% complete for TDP-43_1.fastq
Approx 60% complete for TDP-43_1.fastq
Approx 65% complete for TDP-43_1.fastq
Approx 70% complete for TDP-43_1.fastq
Approx 75% complete for TDP-43_1.fastq
Approx 80% complete for TDP-43_1.fastq
Approx 85% complete for TDP-43_1.fastq
Approx 90% complete for TDP-43_1.fastq
Approx 95% complete for TDP-43_1.fastq
Analysis complete for TDP-43_1.fastq
Started analysis of scr_2.fastq
Approx 5% complete for scr_2.fastq
Approx 10% complete for scr_2.fastq
Approx 15% complete for scr_2.fastq
Approx 20% complete for scr_2.fastq
Approx 25% complete for scr_2.fastq
Approx 30% complete for scr_2.fastq
Approx 35% complete for scr_2.fastq
Approx 40% complete for scr_2.fastq
Approx 45% complete for scr_2.fastq
Approx 50% complete for scr_2.fastq
Approx 55% complete for scr_2.fastq
Approx 60% complete for scr_2.fastq
Approx 65% complete for scr_2.fastq
Approx 70% complete for scr_2.fastq
Approx 75% complete for scr_2.fastq
Approx 80% complete for scr_2.fastq
Approx 85% complete for scr_2.fastq
Approx 90% complete for scr_2.fastq
Approx 95% complete for scr_2.fastq
Analysis complete for scr_2.fastq
Started analysis of scr_1.fastq
Approx 5% complete for scr_1.fastq
Approx 10% complete for scr_1.fastq
Approx 15% complete for scr_1.fastq
Approx 20% complete for scr_1.fastq
Approx 25% complete for scr_1.fastq
Approx 30% complete for scr_1.fastq
Approx 35% complete for scr_1.fastq
Approx 40% complete for scr_1.fastq
Approx 45% complete for scr_1.fastq
Approx 50% complete for scr_1.fastq
Approx 55% complete for scr_1.fastq
```

```
Approx 60% complete for scr_1.fastq
Approx 65% complete for scr_1.fastq
Approx 70% complete for scr_1.fastq
Approx 75% complete for scr_1.fastq
Approx 80% complete for scr_1.fastq
Approx 85% complete for scr_1.fastq
Approx 90% complete for scr_1.fastq
Approx 95% complete for scr_1.fastq
Analysis complete for scr_1.fastq
```

This commando creates a html-file for every fastq file. When you open this, you get an elaborate rapport of the quality of the sample.



- per base sequence quality is good (see image)
- There are 16 overrepresentative sequences in the scrambled samples and in the TDP-43 samples. So, there is some contamination of adaptors. However, we will continue the analysis without correcting for this.

We now make SAM-files out of the fastq-files with the following commando. We sort and compress the data in these files, and thereby converting them into BAM-files.

In [4]:

```
cd /mnt/storage/$USER/jupyternotebooks/groepswerk/RNAseq
STAR --genomeLoad LoadAndExit --genomeDir /mnt/nfs/mfiers/STAR/hg19_star_db
STAR --genomeDir /mnt/nfs/mfiers/STAR/hg19_star_db \
     --genomeLoad LoadAndKeep \
     --runThreadN 2 \
     --readFilesIn TDP-43_2.fastq \
     --outFileNamePrefix TDP-43_2.

STAR --genomeLoad LoadAndExit --genomeDir /mnt/nfs/mfiers/STAR/hg19_star_db
STAR --genomeDir /mnt/nfs/mfiers/STAR/hg19_star_db \
     --genomeLoad LoadAndKeep \
     --runThreadN 2 \
     --readFilesIn TDP-43_1.fastq \
     --outFileNamePrefix TDP-43_1.

STAR --genomeLoad LoadAndExit --genomeDir /mnt/nfs/mfiers/STAR/hg19_star_db
STAR --genomeDir /mnt/nfs/mfiers/STAR/hg19_star_db \
     --genomeLoad LoadAndKeep \
     --runThreadN 2 \
     --readFilesIn scr_2.fastq \
     --outFileNamePrefix scr_2.

STAR --genomeLoad LoadAndExit --genomeDir /mnt/nfs/mfiers/STAR/hg19_star_db
STAR --genomeDir /mnt/nfs/mfiers/STAR/hg19_star_db \
     --genomeLoad LoadAndKeep \
     --runThreadN 2 \
     --readFilesIn scr_1.fastq \
     --outFileNamePrefix scr_1.
```

```
Nov 16 15:39:10 ..... started STAR run
Nov 16 15:39:12 ..... loading genome
Nov 16 15:39:12 ..... started mapping
```

In [12]:

```
cd /mnt/storage/$USER/jupyternotebooks/groepswerk/RNAseq
samtools sort -o TDP-43_1.bam TDP-43_1.Aligned.out.sam
samtools sort -o TDP-43_2.bam TDP-43_2.Aligned.out.sam
samtools sort -o scr_1.bam scr_1.Aligned.out.sam
samtools sort -o scr_2.bam scr_2.Aligned.out.sam
```

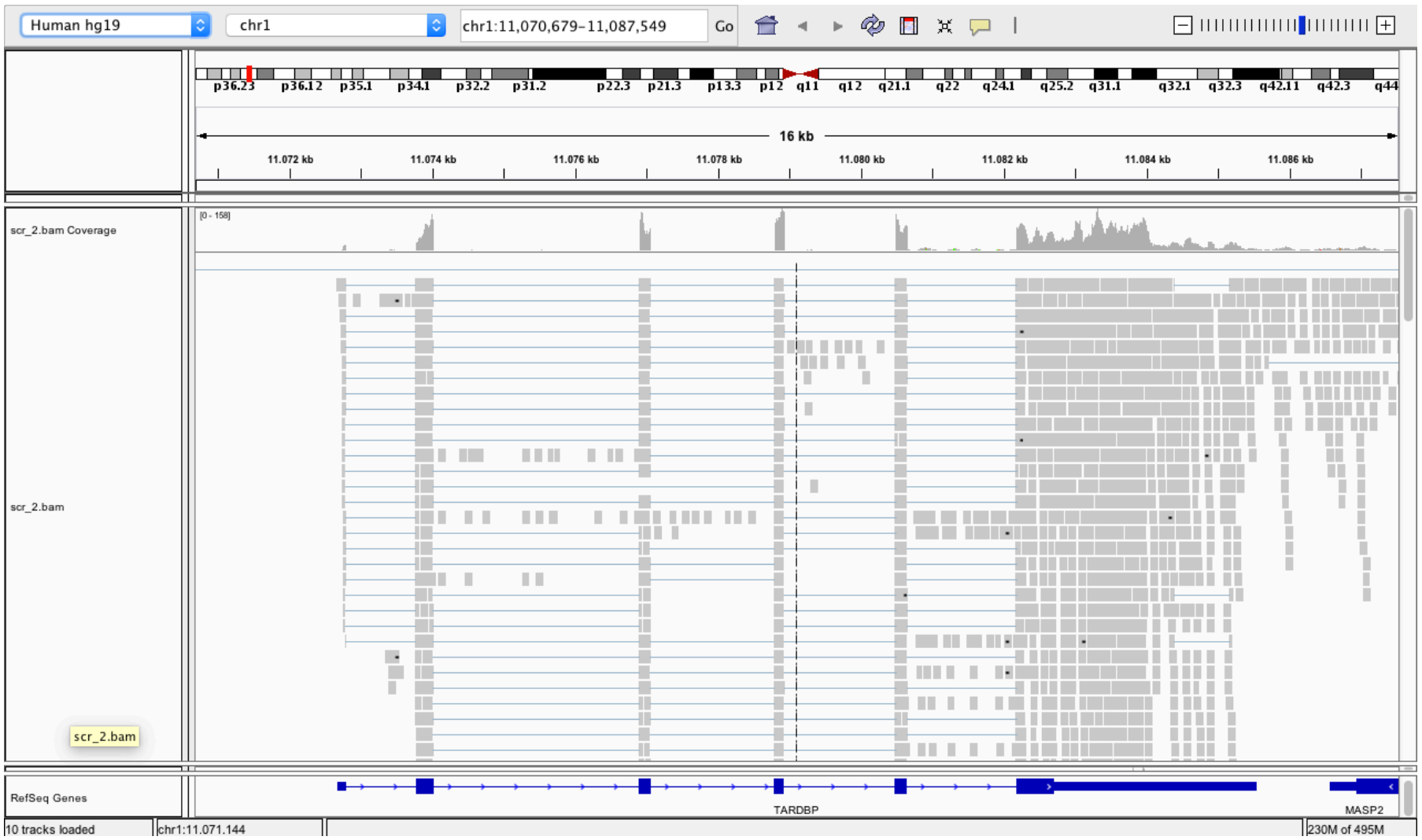
```
[bam_sort_core] merging from 3 files...
[bam_sort_core] merging from 3 files...
[bam_sort_core] merging from 11 files...
[bam_sort_core] merging from 16 files...
```

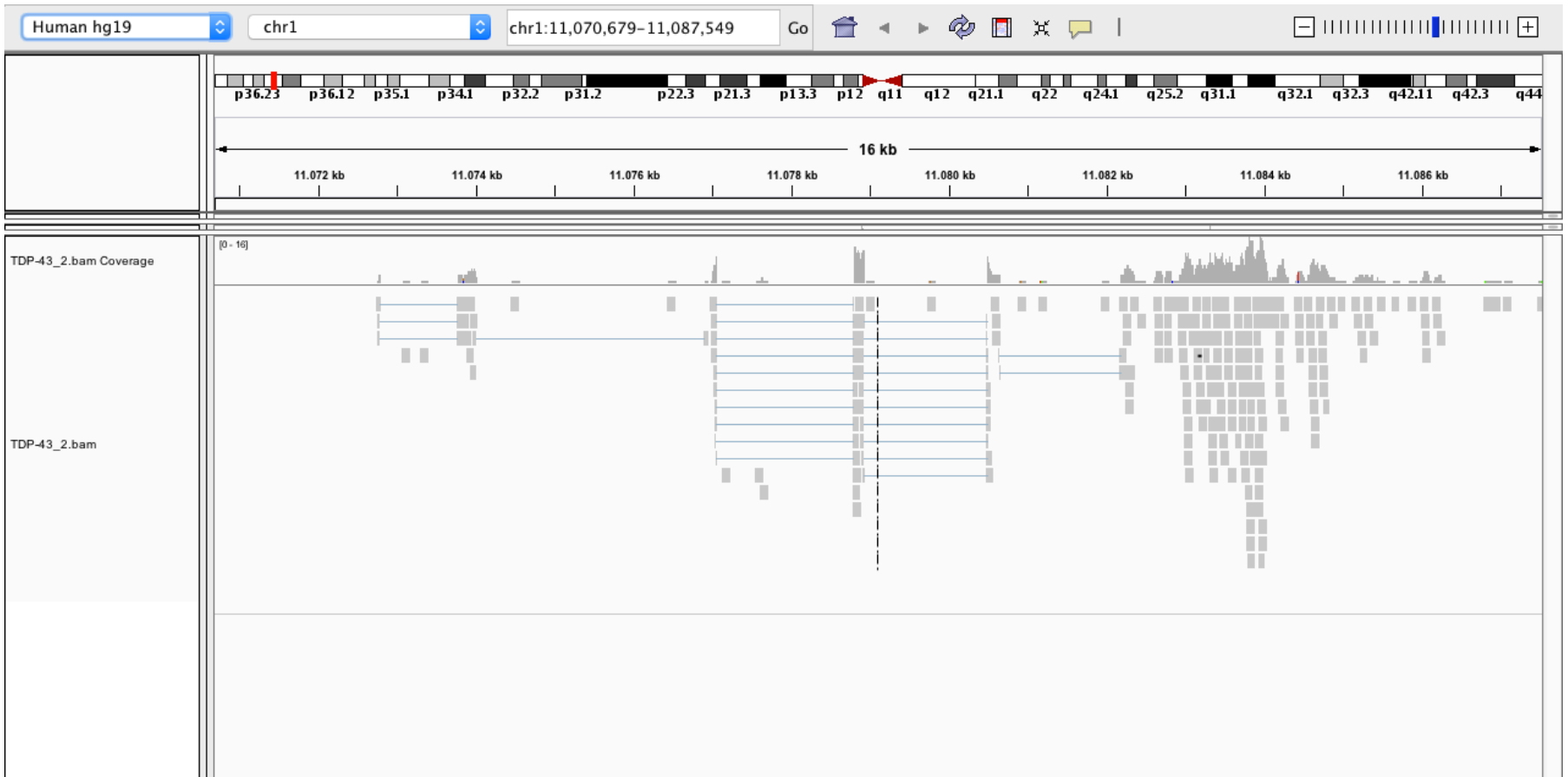
BAM-files are indexed, using samtools.

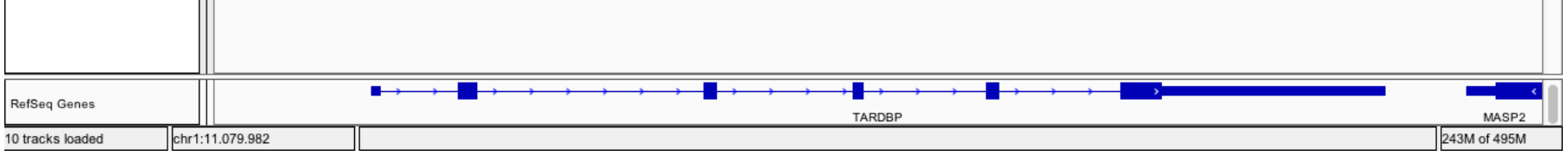
In [13]:

```
samtools index TDP-43_1.bam
samtools index TDP-43_2.bam
samtools index scr_1.bam
samtools index scr_2.bam
```

We check the bam- and bai-files in IGV. When we look at TARDBP, we can confirm that the knockdown of TARDBP was done succesfully, as the amount of reads mapped in the scrambled condition is much more than upon TDP-knockdown.







We now want to know how many reads were mapped to the individual chromosomes.

In [14]:

```
samtools idxstats TDP-43_1.bam
samtools idxstats TDP-43_2.bam
samtools idxstats scr_1.bam
samtools idxstats scr_2.bam
```

```
chrM    16571   14816   0
chr1    249250621       576343  0
chr2    243199373       418804  0
chr3    198022430       338965  0
chr4    191154276       221175  0
chr5    180915260       301436  0
chr6    171115067       326260  0
chr7    159138663       236543  0
chr8    146364022       200271  0
chr9    141213431       329797  0
chr10   135534747       198443  0
chr11   135006516       271560  0
chr12   133851895       295461  0
chr13   115169878       99903   0
chr14   107349540       645305  0
chr15   102531392       173606  0
chr16   90354753        155582  0
chr17   81195210        353707  0
chr18   78077248        85061   0
chr19   59128983        175095  0
chr20   63025520        105059  0
chr21   48129895        52614   0
chr22   51304566        94787   0
chrX    155270560       157245  0
chrY    59373566        21301   0
*       0       0       0
chrM    16571   18445   0
chr1    249250621       671438  0
chr2    243199373       478474  0
chr3    198022430       387927  0
chr4    191154276       248162  0
chr5    180915260       345362  0
chr6    171115067       382758  0
chr7    159138663       272313  0
chr8    146364022       228102  0
chr9    141213431       379236  0
chr10   135534747       219743  0
chr11   135006516       315091  0
chr12   133851895       342648  0
chr13   115169878       114592  0
chr14   107349540       756224  0
chr15   102531392       198196  0
```

```
chr16   90354753        173200  0
chr17   81195210        402814  0
chr18   78077248        94440   0
chr19   59128983        197413  0
chr20   63025520        118243  0
chr21   48129895        58743   0
chr22   51304566        107555  0
chrX    155270560       179957  0
chrY    59373566        21672   0
*       0       0       0
chrM    16571   53713   0
chr1    249250621       2537773 0
chr2    243199373       1811120 0
chr3    198022430       1581077 0
chr4    191154276       932087  0
chr5    180915260       1311461 0
chr6    171115067       1504294 0
chr7    159138663       994839  0
chr8    146364022       876665  0
chr9    141213431       1562970 0
chr10   135534747       825921  0
chr11   135006516       1190390 0
chr12   133851895       1322312 0
chr13   115169878       427545  0
chr14   107349540       3591786 0
chr15   102531392       723144  0
chr16   90354753        604307  0
chr17   81195210        1606659 0
chr18   78077248        357581  0
chr19   59128983        694321  0
chr20   63025520        428160  0
chr21   48129895        223475  0
chr22   51304566        393490  0
chrX    155270560       665557  0
chrY    59373566        64934   0
*       0       0       0
chrM    16571   87870   0
chr1    249250621       3913935 0
chr2    243199373       2655724 0
chr3    198022430       2341986 0
chr4    191154276       1376593 0
chr5    180915260       1928126 0
chr6    171115067       2166948 0
chr7    159138663       1477604 0
chr8    146364022       1290118 0
chr9    141213431       2278092 0
chr10   135534747       1229805 0
chr11   135006516       1763127 0
chr12   133851895       1953873 0
chr13   115169878       642010  0
chr14   107349540       5292658 0
chr15   102531392       1068140 0
chr16   90354753        914281  0
chr17   81195210        2549963 0
chr18   78077248        535990  0
chr19   59128983        1034169 0
chr20   63025520        644745  0
```

```
chr21    48129895        331705  0
chr22    51304566        580199  0
chrX     155270560       978206  0
chrY     59373566        103204  0
*        0       0       0
```

In the 3rd column we see the number of reads mapped to each chromosome (high number). The 4th column, represents the amount that did not map (0). In our case, this means that all the reads are mapped.

Next, we make a symbolic link to the human annotation.

In [16]:

```
ln -sf /mnt/nfs/data/RNA-seq/gencode.v19.nopseudo.plus.sort.gtf /mnt/storage/$USE
ls -l *gtf
```

```
lrwxrwxrwx 1 r0384602 domain users 56 nov 16 16:21 gencode.v19.nopse
udo.plus.sort.gtf -> /mnt/nfs/data/RNA-seq/gencode.v19.nopseudo.plus
.sort.gtf
```

Using featureCounts, we create gene counts for each sample.

In [18]:

```
featureCounts -Q 10 -g gene_name -a gencode.v19.nopseudo.plus.sort.gtf -o all.cou
```

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

//========================= featureCounts setting ================
==========\\
||
||
||             Input files : 4 BAM files
||
||                           S scr_1.bam
||
||                           S scr_2.bam
||
||                           S TDP-43_1.bam
||
||                           S TDP-43_2.bam
||
||
```

```
||
||              Output file : all.counts
||
||              Annotations : gencode.v19.nopseudo.plus.sort.gtf (GTF
)         ||
||
||
||                  Threads : 1
||
||                    Level : meta-feature level
||
||               Paired-end : no
||
||         Strand specific : no
||
||      Multimapping reads : not counted
||
|| Multi-overlapping reads : not counted
||
||       Read orientations : fr
||
||
||
\\==================== http://subread.sourceforge.net/
(http://subread.sourceforge.net/) ====================//

//================================ Running ===========================
==========\\
||
|| Load annotation file gencode.v19.nopseudo.plus.sort.gtf ...
||
||    Features : 1156385
||
||    Meta-features : 41863
||
||    Chromosomes/contigs : 25
||
||
||
|| Process BAM file scr_1.bam...
||
||    Single-end reads are included.
||
||    Assign reads to features...
||
||    Total reads : 26285581
||
||    Successfully assigned reads : 13367385 (50.9%)
||
||    Running time : 0.55 minutes
||
||
||
|| Process BAM file scr_2.bam...
||
||    Single-end reads are included.
```

```
||
||    Assign reads to features...
||
||    Total reads : 39139071
||
||    Successfully assigned reads : 18839854 (48.1%)
||
||    Running time : 0.78 minutes
||
||
||
|| Process BAM file TDP-43_1.bam...
||
||    Single-end reads are included.
||
||    Assign reads to features...
||
||    Total reads : 5849139
||
||    Successfully assigned reads : 2868258 (49.0%)
||
||    Running time : 0.14 minutes
||
||
||
|| Process BAM file TDP-43_2.bam...
||
||    Single-end reads are included.
||
||    Assign reads to features...
||
||    Total reads : 6712748
||
||    Successfully assigned reads : 3327819 (49.6%)
||
||    Running time : 0.14 minutes
||
||
||
||                          Read assignment finished.
||
||
||
\\===================== http://subread.sourceforge.net/
(http://subread.sourceforge.net/) ======================//
```

When you open the all.counts file, you see that the file is made of several columns.

In [19]:

```
head all.counts
```

```
# Program:featureCounts v1.5.0-p1; Command:"featureCounts" "-Q" "10"
"-g" "gene_name" "-a" "gencode.v19.nopseudo.plus.sort.gtf" "-o" "all
.counts" "scr_1.bam" "scr_2.bam" "TDP-43_1.bam" "TDP-43_2.bam"
Geneid  Chr     Start   End     Strand  Length  scr_1.bam
scr_2.bam       TDP-43_1.bam    TDP-43_2.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;112700;120721;129055;133374
91629;92240;111357;112804;120932;129223;133566  -;-;-;-;-;-;-   3569
2       1       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    6       6       0       2
RP11-34P13.14   chr1;chr1       139790;140075   139847;140339   -;-
323     1       0       0       0
RP11-34P13.13   chr1;chr1;chr1;chr1;chr1;chr1;chr1;chr1;chr1;chr1
141474;146386;155767;164263;165884;168100;168610;169049;172557;17375
3
143011;149707;155831;164791;165942;168165;168767;169264;172688;17386
2       -;-;-;-;-;-;-;-;-;-     6195    10      13      1       3
```

Column 1, 2, 3, 4, 5 and 6 only give information about the position in the genome (GeneID, chromosome, start, end, strand and length). Using the following commando, we separate these first 6 columns and paste them into a new file, called all.genedata.tsv

In [20]:

```
cut -f-6 all.counts > all.genedata.tsv
```

Column 7, 8, 9 and 10 only give information about the different samples. Using the following commando, we separate these 4 columns together with the first GeneID column into a new file, called all.gene.counts.tsv

In [21]:

```
cut -f1,7- all.counts | grep -v '^#' > all.gene.counts.tsv
```

For the differential expression analysis of the data, we load the packages DESeq2 and scales. See COMMANDO 1.

Now we make sure that we are working in the right directory. See COMMANDO 2.

We define the file we just made containing the raw counts. See COMMANDO 3.

We define the file we just made containing the raw counts. See COMMANDO 3.

We rename the column to have a better overview of the table. See COMMANDO 4.

Now we say which different experimental conditions were used. See COMMANDO 5. We first assigned scrambled, then TDP-43 knockdown. This means that if we see a positive Log2FoldChange, the gene is upregulated upon TDP-43 knockdown and vice versa.

Next we make a dataframe with the sample names and conditions. See COMMANDO 6.

A DESeq Object was created for using in the further analysis with the DESeq package. See COMMANDO 7.

We want to confirm that the order of the conditions has remained. See COMMANDO 8.

Next, the actual differential analysis was done using DESeq2. See COMMANDO 9.

As with all analysis, quality control needs to be done.

We check whether the scrambled replicates are similar to each other, and likewise for the TDP-43 replicates. Further, we check if the scambled replicates are more "distant" from the TDP-43 replicates. See COMMANDO 10. --> The diagonals are indeed what we expect. Great variability between the two conditions and small variability within the conditions.

There is some noise visible in the diagonals for the low expressing genes. Using DESeq2, we can minimise this. See COMMANDO 11.

Another way to look at the data and to understand the between-sample variation, is by making a PCA-plot. See COMMANDO 12. Again you can see a larger variability between the conditions than within.

To visualise the results of the differentional expression analysis, we create a MA plot. The significantly differentially expressed genes are coloured red. See COMMANDO 13.

We export these results. See COMMANDO 14.

If we set the p-value at 0.05, we see that 1579 genes are differentially expressed. See COMMANDO 15.

Now, we are interested in the expression levels of HDAC6. If we make a bar plot (see COMMANDO 16), we do not see a downregulation of HDAC6 after TDP-43 knockdown.

If we look at the actual data (therefore we make a tsv-file, using COMMANDO 17), we see that HDAC6 is downregulated (Log2FoldChange=-0,036849), but not significantly (p-adjusted=0,9018339).

270	C10orf12	210,166686	0,06035274	0,17703347	0,34091144	0,73317025	0,90155077
271	TMUB2	220,719841	-0,0603392	0,17701212	-0,3408759	0,73319698	0,90155077
272	GDPD1	370,365265	-0,0523687	0,15356476	-0,3410205	0,73308817	0,90155077
273	ANAPC1	600,38515	0,04076998	0,11956811	0,34097705	0,73312086	0,90155077
274	ISL2	59,5699246	-0,0846913	0,24882872	-0,34036	0,73358544	0,9018339
275	HDAC6	808,209177	-0,036849	0,10824555	-0,3404201	0,7335402	0,9018339

276	GPATCH3	78,4670788	-0,0807665	0,23784281	-0,3395792	0,73417345	0,902251
277	PCGF2	61,0139133	0,08582658	0,25258416	0,33979399	0,73401167	0,902251
278	NGLY1	342,965054	-0,0503222	0,14815796	-0,3396524	0,73411829	0,902251
279	HMBOX1	406,364729	0,04707789	0,13867275	0,33948914	0,73424128	0,902251
280	CYTH2	455,957863	0,04530493	0,13354921	0,33923776	0,73443063	0,90238642
281	GTF2H2C	71,9662532	0,08355584	0,24652471	0,33893494	0,73465874	0,90256943

So in this clincally more relevant setting, we observe in our analysis that the downregulation of HDAC6 after TDP-43 silencing is not significant.

We think the difference in results between the 2 articles can be present because the research is perfomed by two different labs using different techniques, different cells, different tools,...

In addition, we did not control for adaptor contamination which may also influence the outcome of the analysis.

Ranking based analysis

Now, we are interested in the function of the genes that are differentially expressed after TDP-43 knockdown.

For this analysis, we used the Deseq.results.tsv file and opened it in Excel. Here we ranked the data on decreasing Log2FoldChange. Then, we uploaded the ranked_up list in GOrilla with the following result (http://cbl-gorilla.cs.technion.ac.il/GOrilla/10n6crwt/GOResults.html (http://cbl-gorilla.cs.technion.ac.il/GOrilla/10n6crwt/GOResults.html)).

GO term	Description	P-value	FDR q-value	Enrichment (N, B, n, b)	Genes
GO:0050911	detection of chemical stimulus involved in sensory perception of smell	1.64E-156	2.37E-152	14.49 (18359,393,548,170)	[+] Show genes
GO:0050907	detection of chemical stimulus involved in sensory perception	2.84E-148	2.06E-144	13.11 (18359,437,548,171)	[+] Show genes
GO:0009593	detection of chemical stimulus	7.58E-142	3.67E-138	15.86 (18359,470,367,149)	[+] Show genes
GO:0050906	detection of stimulus involved in sensory perception	2.53E-139	9.16E-136	15.34 (18359,486,367,149)	[+] Show genes
GO:0051606	detection of stimulus	3.7E-122	1.07E-118	11.97 (18359,627,367,150)	[+] Show genes
GO:0007186	G-protein coupled receptor signaling pathway	2.42E-97	5.86E-94	5.81 (18359,1135,548,197)	[+] Show genes
GO:0006958	complement activation, classical pathway	4.91E-52	1.02E-48	14.27 (18359,95,772,57)	[+] Show genes
GO:0006956	complement activation	2.21E-48	4E-45	12.54 (18359,110,772,58)	[+] Show genes
GO:0006955	immune response	5.27E-46	8.49E-43	3.37 (18359,937,1022,176)	[+] Show genes
GO:0018149	peptide cross-linking	6.73E-46	9.76E-43	101.47 (18359,56,84,26)	[+] Show genes

In the top of the table (which is ranked by increasing p-value), you see that the most significant cellular function of the upregulated genes after knockdown of TDP-43 is 'detection of chemical stimulus involved in sensory perception of smell'. Because loss of function of TDP-43 is associated with neurodegeneration, we hypothesized that mainly functions needed for normal neuronal functioning would come out of this anaysis. If we look further into the list we find for example neurological system process.

GO:0019585	glucuronate metabolic process	1.77E-5	2.45E-3	7.21 (18313,27,847,9)	[+] Show genes
GO:0050871	positive regulation of B cell activation	1.82E-5	2.49E-3	3.88 (18313,85,944,17)	[+] Show genes
GO:0050877	neurological system process	1.96E-5	2.67E-3	2.31 (18313,846,346,37)	[+] Show genes
GO:0000183	chromatin silencing at rDNA	3.64E-5	4.91E-3	37.35 (18313,37,53,4)	[+] Show genes
GO:0030593	neutrophil chemotaxis	4.95E-5	6.61E-3	6.17 (18313,62,479,10)	[+] Show genes

Then we ranked the data on increasing Log2FoldChange. Then, we uploaded the ranked_down list in GOrilla with the following result (http://cbl-gorilla.cs.technion.ac.il/GOrilla/bz704sar/GOResults.html (http://cbl-gorilla.cs.technion.ac.il/GOrilla/bz704sar/GOResults.html)).

GO term	Description	P-value	FDR q-value	Enrichment (N, B, n, b)	Genes
GO:0030198	extracellular matrix organization	2.04E-28	2.95E-24	13.84 (18359,293,163,36)	[+] Show genes
GO:0032502	developmental process	2.29E-28	1.66E-24	1.55 (18359,4551,1401,538)	[+] Show genes

GO:0043062	extracellular structure organization	2.31E-28	1.12E-24	13.79 (18359,294,163,36)	[+] Show genes
GO:0044767	single-organism developmental process	4.32E-27	1.56E-23	1.58 (18359,4125,1401,496)	[+] Show genes
GO:0044699	single-organism process	5.5E-23	1.59E-19	1.22 (18359,10203,1729,1169)	[+] Show genes
GO:0030199	collagen fibril organization	9E-22	2.18E-18	71.30 (18359,35,103,14)	[+] Show genes
GO:0007155	cell adhesion	9.88E-22	2.05E-18	2.03 (18359,1247,1555,214)	[+] Show genes
GO:0048856	anatomical structure development	1.33E-21	2.42E-18	1.84 (18359,3086,842,260)	[+] Show genes
GO:0022610	biological adhesion	1.52E-21	2.45E-18	2.02 (18359,1251,1555,214)	[+] Show genes
GO:0050793	regulation of developmental process	2.08E-21	3.02E-18	1.81 (18359,2191,1303,281)	[+] Show genes

In the top of the table (which is ranked by increasing p-value), you see that the most significant cellular function of the downregulated genes after knockdown of TDP-43 is 'extracellular matrix organisation'. Again we hypothesized that mainly functions needed for normal neuronal functioning would come out of this anaysis. If we look further into the list we find for example neuron projection guidance.

GO:0009991	response to extracellular stimulus	2.58E-7	1.98E-5	2.21 (18359,370,1212,54)	[+] Show genes
GO:1902533	positive regulation of intracellular signal transduction	3.08E-7	2.35E-5	1.79 (18359,880,1096,94)	[+] Show genes
GO:0097485	neuron projection guidance	3.08E-7	2.34E-5	2.12 (18359,195,2352,53)	[+] Show genes
GO:0051338	regulation of transferase activity	3.13E-7	2.36E-5	1.70 (18359,908,1311,110)	[+] Show genes
GO:0051716	cellular response to stimulus	3.17E-7	2.38E-5	1.40 (18359,2271,1373,238)	[+] Show genes
GO:1901342	regulation of vasculature development	3.32E-7	2.48E-5	3.00 (18359,233,815,31)	[+] Show genes

CLiP-Seq

The article (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4935974/ (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4935974/)) also investigates the ALS-associated protein TAF15. TAF15 is an RNA-binding protein just like TDP-43. They performed CLiP-seq on whole-brain tissue from adult mice to check which genes TAF15 binds. CLiP stands for "cross-linking immunoprecipitation" and is a technique to identify protein interactions with RNA.

After applying a cluster-finding algorithm, the article identified TAF-binding clusters. We could find these clusters in a bed-file generated by the researches of the article (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM2056777 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM2056777)).

We download and rename this bed file to "TAF15.bed", copy it manually to our directory and take a look at it.

In [9]:

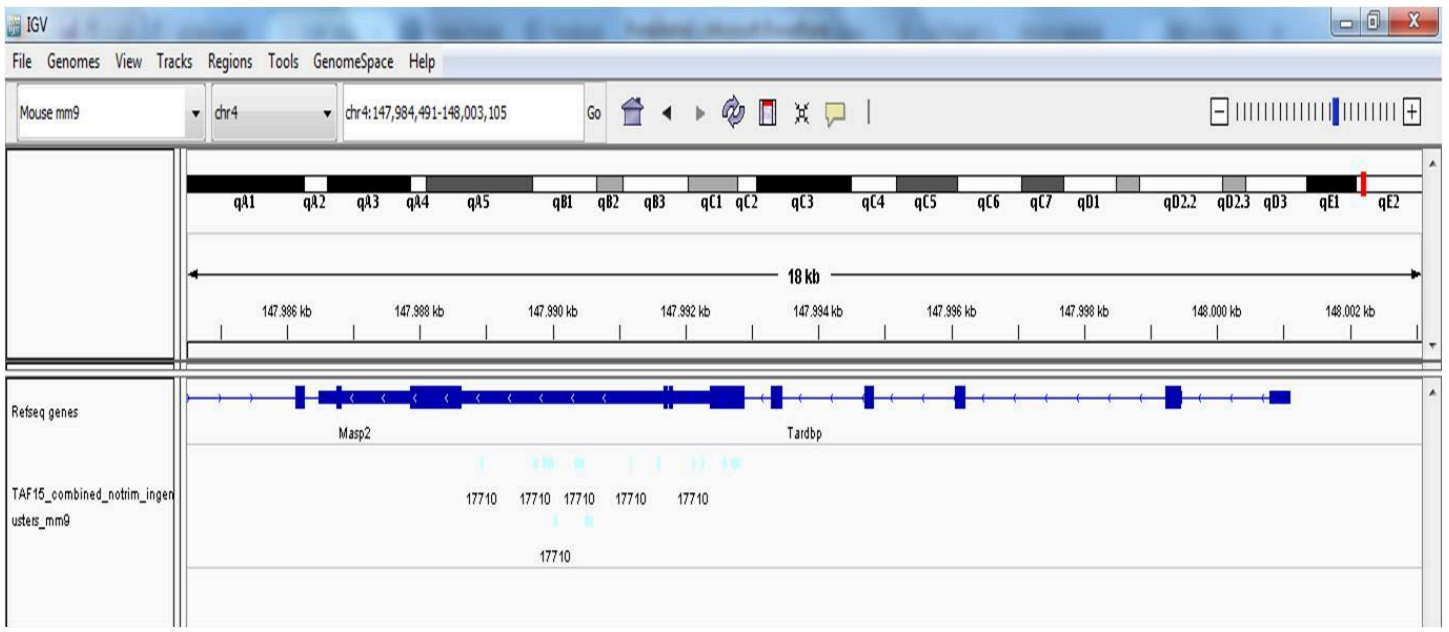
```
mkdir -p /mnt/storage/$USER/jupyternotebooks/groepswerk/CLIPseq
cd ~/data/jupyternotebooks/groepswerk/CLIPseq
head TAF15.bed
```

```
#TAF15_combined_data/TAF15_combined_notrim_s0_50-150_g20_corrected_p
remRNA_withrepeats_clusters
track name=TAF15_combined_notrim_ingenes_clusters_mm9 description=TA
F15_combined_notrim_ingenes_clusters_mm9 visibility=2 itemRgb="On" u
seScore=1
chr1    3280460 3280518 1       25      -       3280490 3280499 0,25
5,255
chr1    3290124 3290173 1       41      -       3290130 3290139 0,25
5,255
chr1    3310940 3311005 1       65      -       3310965 3310974 0,25
5,255
chr1    3344692 3344741 1       28      -       3344719 3344728 0,25
5,255
chr1    3346065 3346118 1       30      -       3346082 3346091 0,25
5,255
chr1    3350701 3350757 1       50      -       3350731 3350740 0,25
5,255
chr1    3362888 3362936 1       21      -       3362905 3362914 0,25
5,255
chr1    3367009 3367061 1       24      -       3367032 3367041 0,25
5,255
```

We now upload this file in IGV and see to which genes TAF15 can bind in the brain tissue of mice. Therefore, we first have to make sure that in IGV the genome is changed to "Mouse mm9".

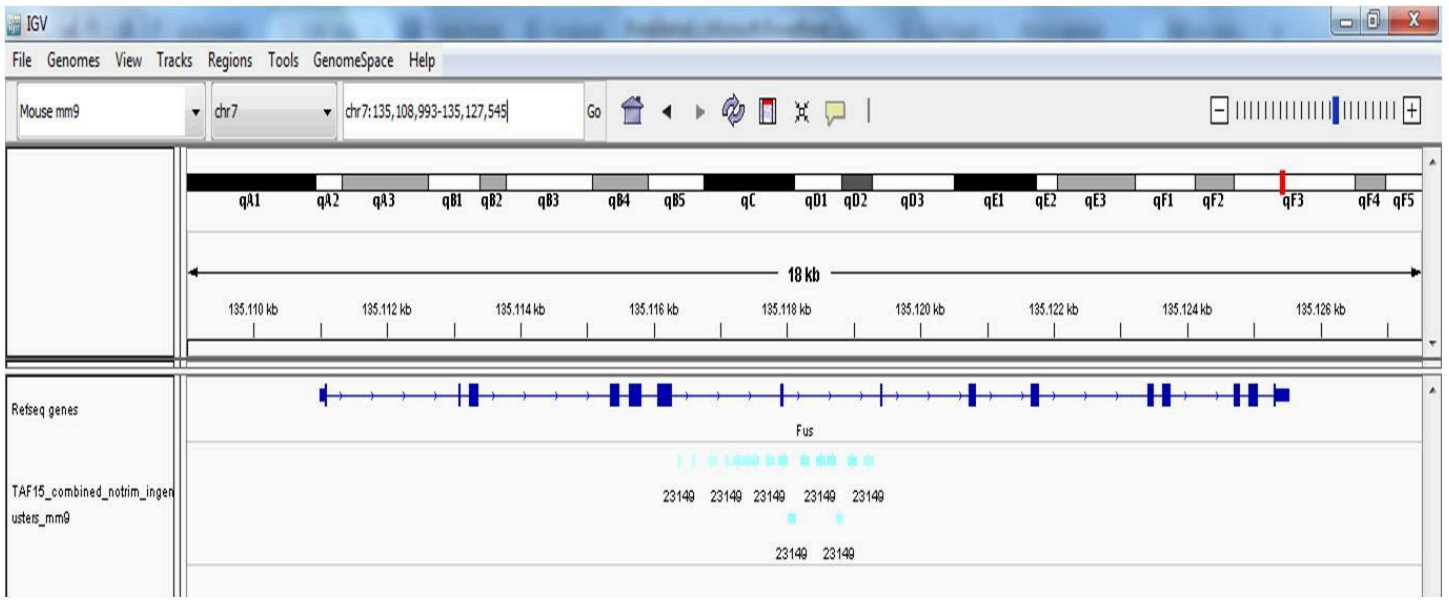
Now, we are curious if TAF15 binds the RNA of TDP-43, FUS and/or itself as these are all proteins involved in RNA metabolism and maybe regulate in that way each others expression. And indeed, if we look for these proteins in IGV we see overlap of the bed file, meaning TAF15 binds here.

TDP-43 (tardbp)

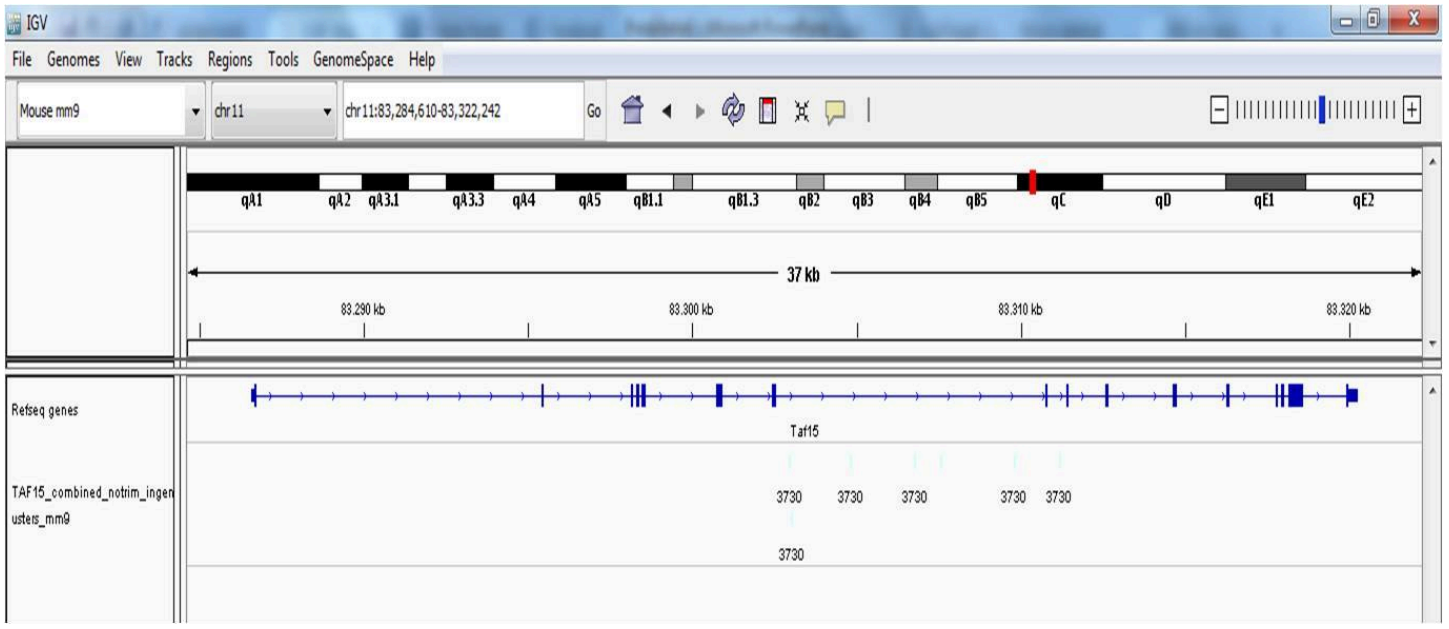


Fus

Fus



Taf15



Discussion

In the neurodegenerative disorders ALS and FTD, the nuclear protein TDP-43 is known to form cytoplasmic aggregates in the affected neurons which possibly contributes to neuronal death. In this report, we tried to get more insight in the role of TDP-43 in neurodegeneration by looking at the effect of TDP-43 silencing. We found two articles that performed TDP-43 knockdown using siRNA against TDP-43.

The first article silenced TDP-43 in HEK293E cells and performed a micro-array analysis afterwards. By analysing the data, we could confirm TDP-43 knockdown and the finding that the cytoplasmic enzyme HDAC6 is significantly downregulated after TDP-43 knockdown. HDAC6 has a protective role as it primes misfolded proteins to degradation via autophagy. Hence, a possible mechanism involved in neurodegeneration could be proteotoxicity caused by HDAC6 deficiency.

The second article performed TDP-43 silencing in a more clinical relevant setting as they used motor neurons generated from induced pluripotent stem cells from healthy humans instead of HEK293E cells. This group used RNA-sequencing and not micro-array to assess the differentially expressed genes upon TDP-43 knockdown. First, we could confirm the silencing of TDP-43, when we visualized the mapped reads from the test (TDP-43 knockdown) and control situation in IGV. Second, we were particularly interested in the

expression levels of HDAC6 RNA after TDP-43 knockdown. However, we did not see a significant downregulation of HDAC6 after silencing TDP-43. There can be numerous of reasons for this difference in result between the two articles as a different cell type is used and the research is perfomed by two different labs using different techniques and different analyzing tools. Another contributing factor that can possibly influence the results, is the fact that we did not control for adaptor contamination.

Next, we continued with the results of the second article for functional analysis. As loss of function of TDP-43 is associated with neurodegeneration, we hypothesized that after knockdown of TDP-43 in vitro, mostly the expression of genes important for normal neuronal functioning would be affected. Indeed, via a ranking-based analysis, we found 2352 (of 18359 genes in total) downregulated genes involved in neuron projection guidance and 346 (of 18313 genes in total) upregulated genes involved in neurological system processes after TDP-43 knockdown.

Lastly, the second article also used the interesting technique CLiP-Seq to identify protein-RNA interactions. The article focused on the ALS-associated protein TAF15 which is an RNA-binding protein just like TDP-43. Here, we could confirm the interaction of TAF15 in brain tissue of mice with RNA of three important proteins involved in RNA-metabolism, namely TDP-43, FUS and TAF15 itself.

In [ ]: