RNA-seq analysis of lncRNA’s in human pancreatic islets and β-cells.

Ntoulaveris Grigoris

Introduction

This report presents the results and operations of an RNA-seq analysis. The working data were FASTQ files from a transcriptome analysis of lncRNA’s in human

pancreatic islets and β-cells. There were four files in total, two for each sample. They were paired end data with each part of the file being the left or right read (eg ERR173261_1 was the left read and ERR173261_2 was the right read). The original study defined 1128 islet lncRNA genes, proving them to be an integral component of the dynamic β-cell specific differentiation program.

Data acquirement

The relevant samples were downloaded from ebi.ac.uk by terminal, using the wget command. Since the fastq files were in gun zipped form, they also had to be unzipped.

```
# the same command for all files
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR173/ERR173261/ERR173261_2.fastq.gz

# unzip them all
for file in *.fastq.gz; do gunzip "$file"; done
```

Quality control and preprocessing

FASTQC

This step involves quality control of the raw FASTQ sequencing data. FastQC is a tool that produces a report on the quality of the raw sequencing data, including a detailed summary of various quality metrics, such as per base sequence quality, base composition, and overrepresented sequences.

In order to perform the FASTQC analysis the application needed to be installed. As the VM didn’t have root privileges the application was installed manually from the web.

```
cd FastQC

unzip fastqc_v0.11.9.zip

# grant execute permissions to the file
chmod +x fastqc

# run file
./fastqc
```

The QC analysis is presented below by comparing the different samples in every one of the analysis' results.

Basic sequence statistics

The four samples were of the same file type and nearly identical in their total sequence sum and their GC content.

Basic sequence stats

Measure	Value
Filename	ERR173261_1.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	36899101
Sequences flagged as poor quality	0
Sequence length	51
%GC	48

Basic sequence stats

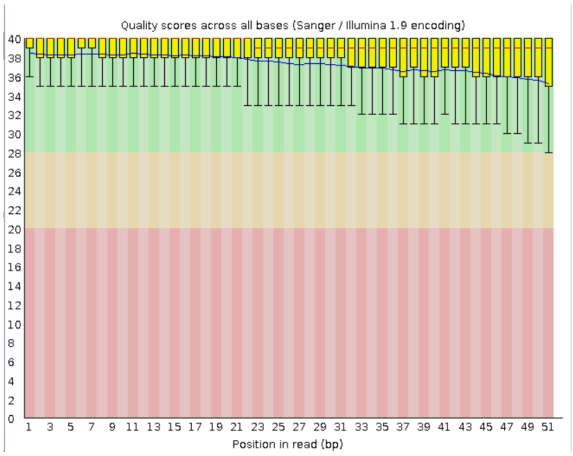
Measure	Value
Filename	ERR173261_2.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	36899101
Sequences flagged as poor quality	0
Sequence length	51
%GC	49

Basic sequence stats

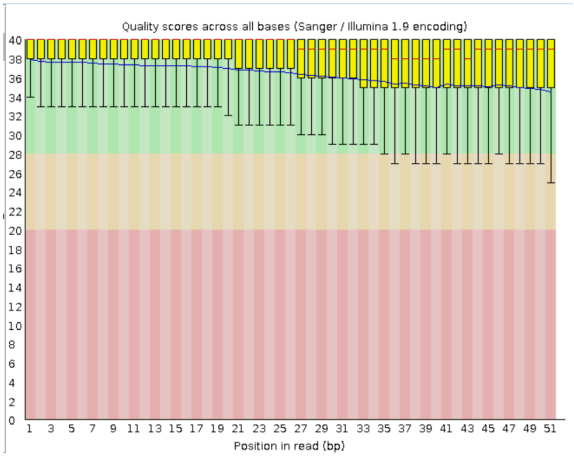
Measure	Value
Filename	ERR173280_1.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	30386851
Sequences flagged as poor quality	0
Sequence length	51
%GC	48

Basic sequence stats	
Measure	Value
Filename	ERR173280_2.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	30386851
Sequences flagged as poor quality	0
Sequence length	51
%GC	48

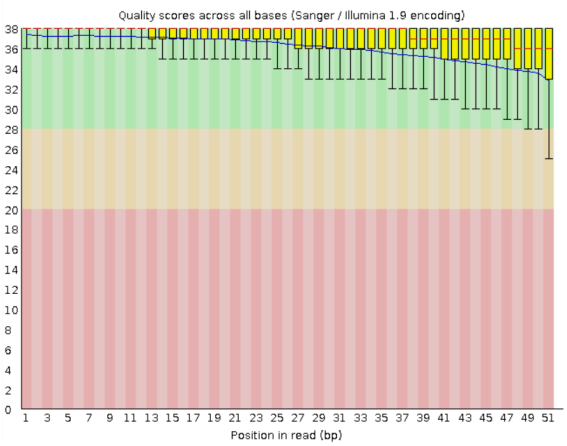
Sequence quality per base



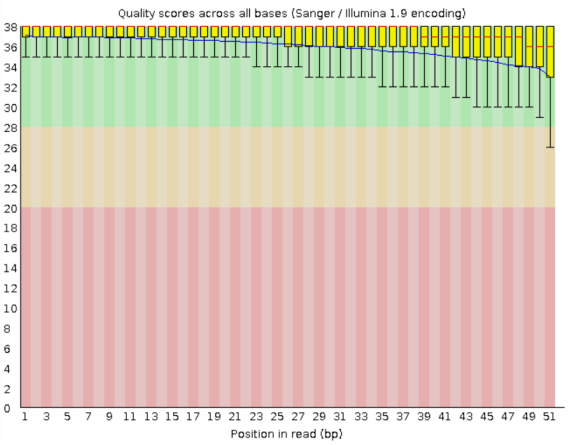
ERR173261_1



ERR173261_2

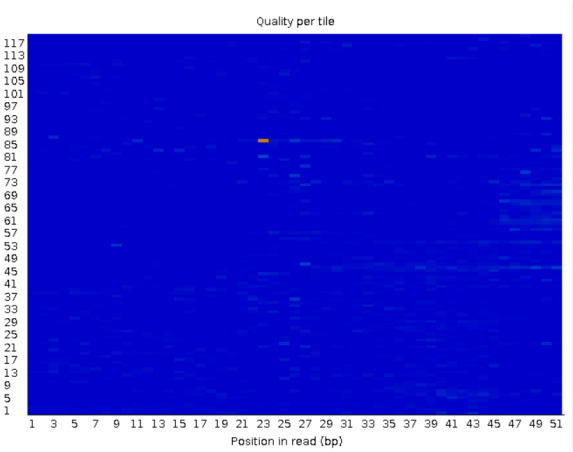


ERR173280_1

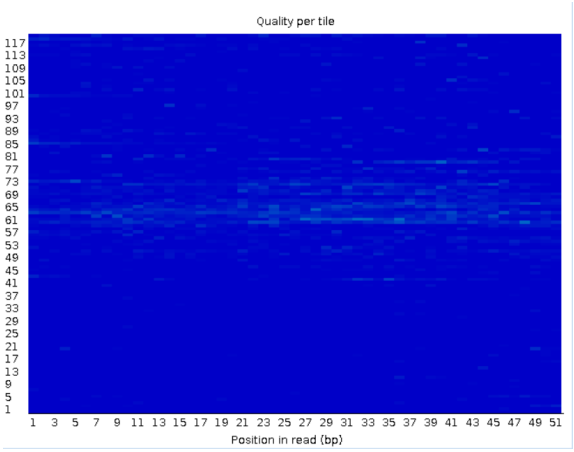


ERR173280_2

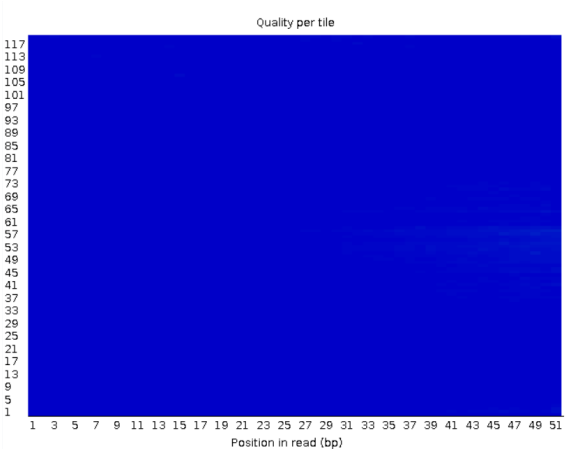
Sequence quality per tile



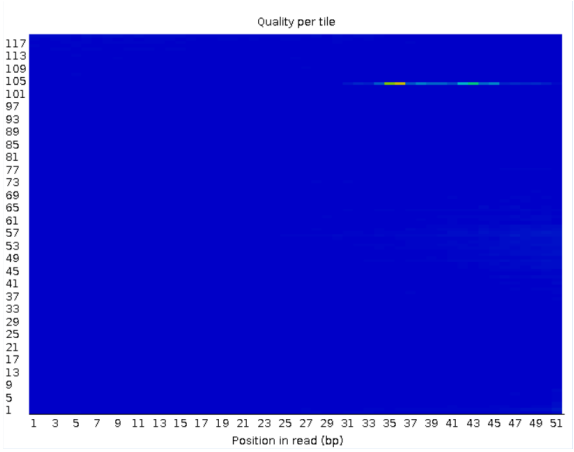
ERR173261_1



ERR173261_2

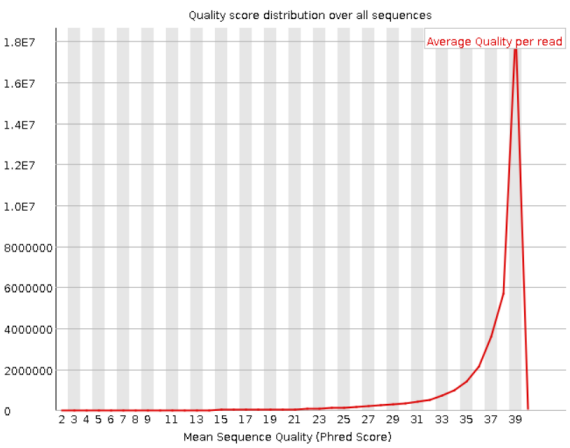


ERR173280_1

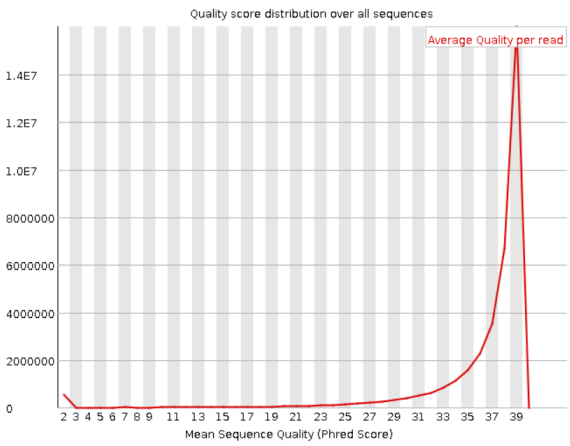


ERR173280_2

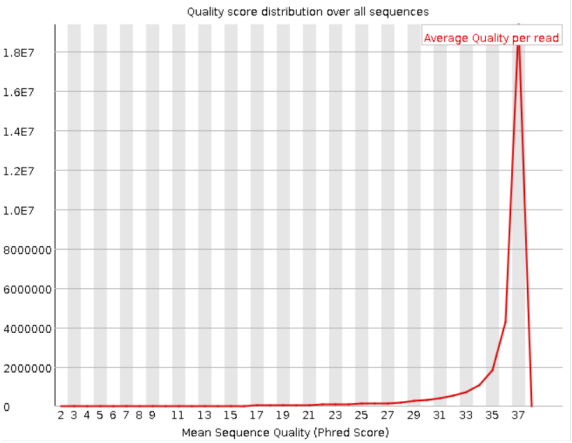
Quality scores per sequence



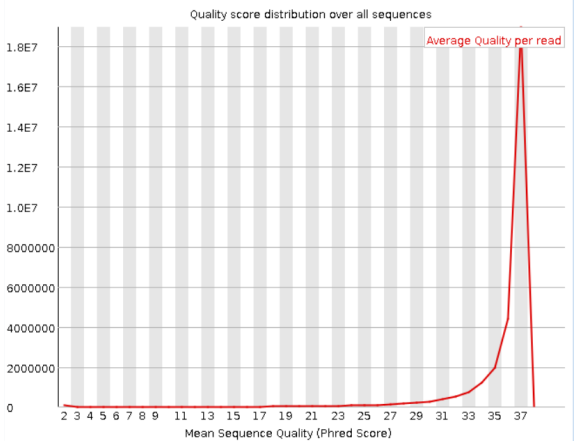
ERR173261_1



ERR173261_2

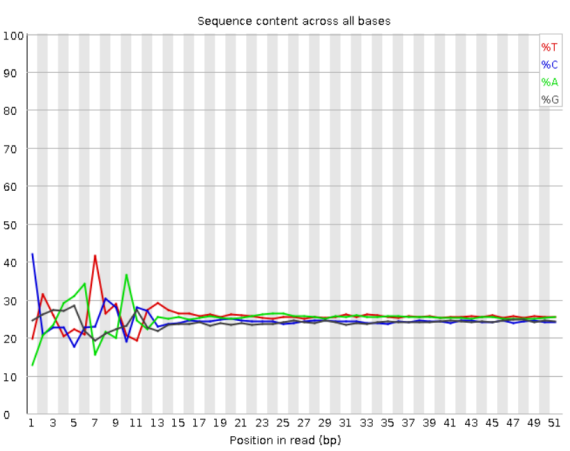


ERR173280_1

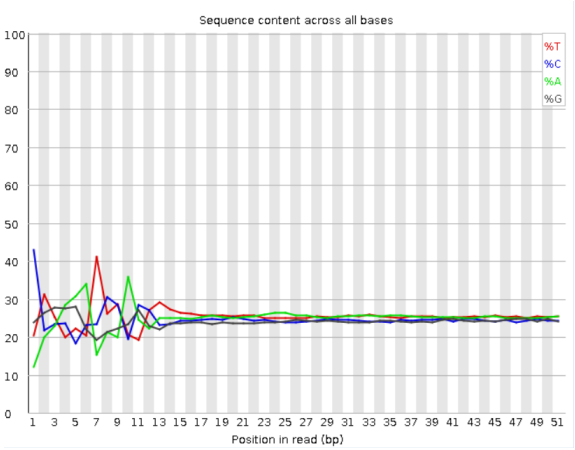


ERR173280_2

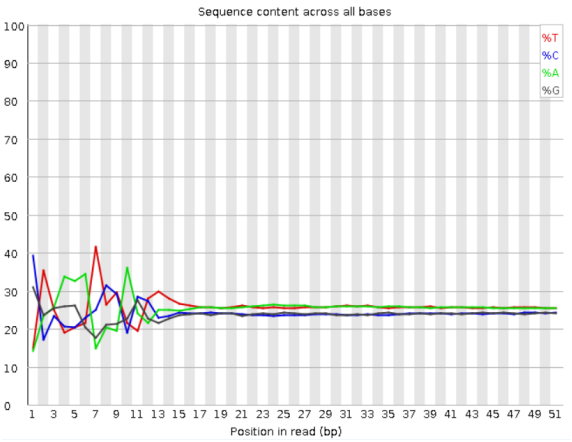
Sequence content per base



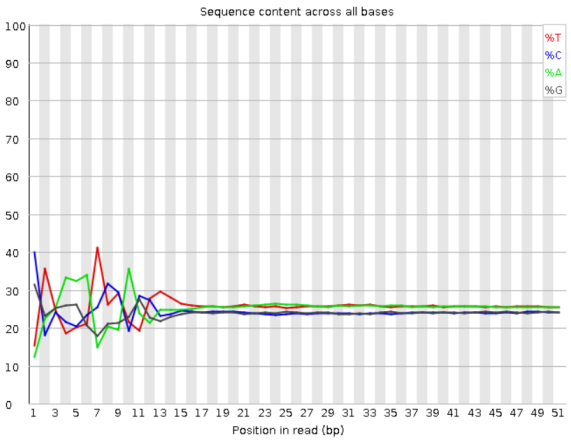
ERR173261_1



ERR173261_2

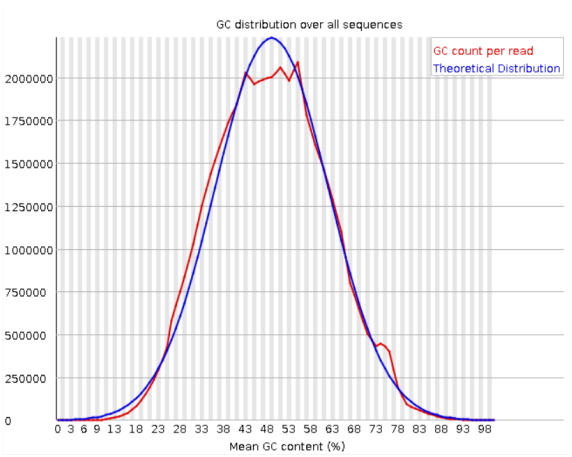


ERR173280_1

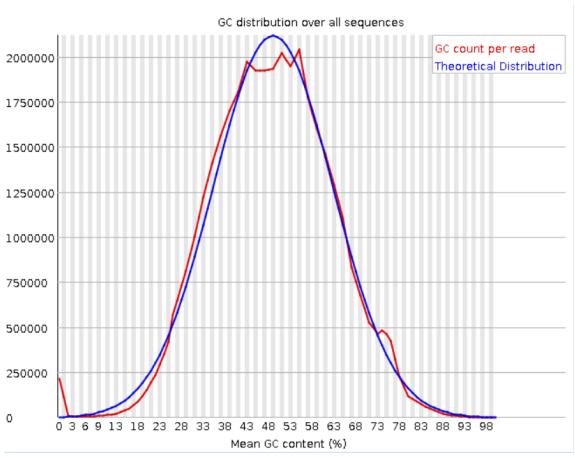


ERR173280_2

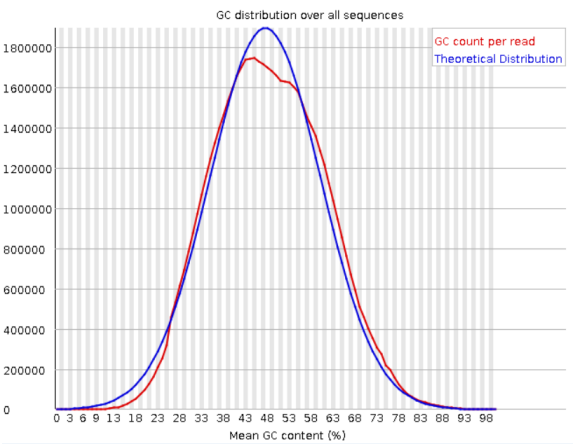
GC content per sequence



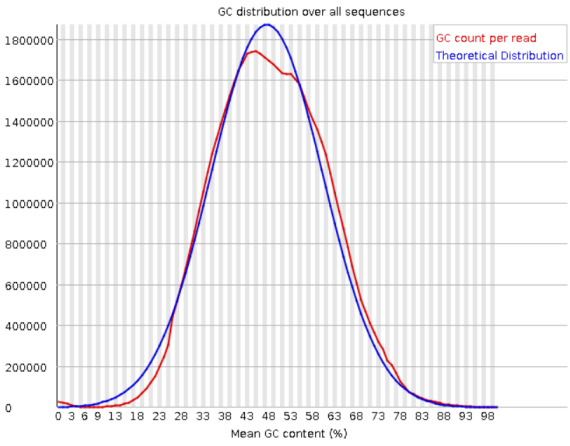
ERR173261_1



ERR173261_2

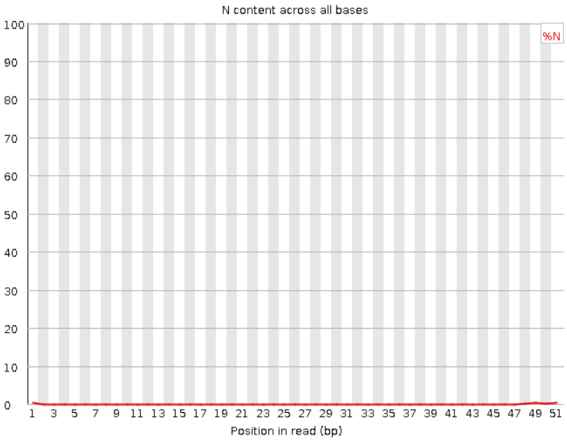


ERR173280_1

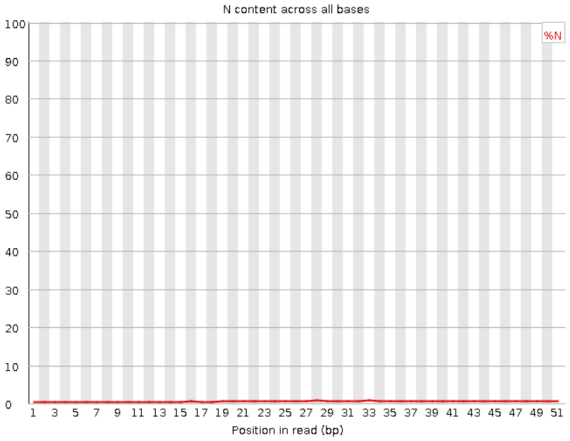


ERR173280_2

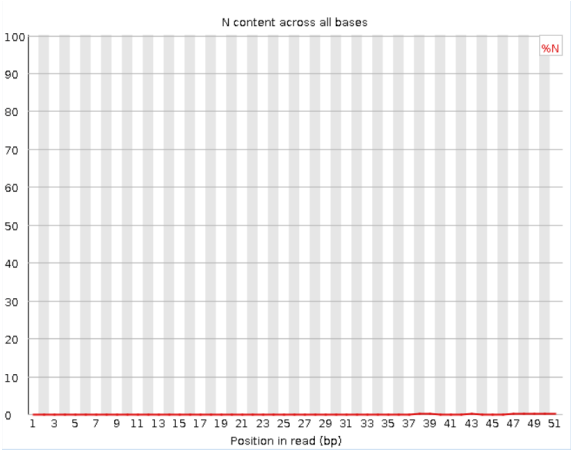
N content per base



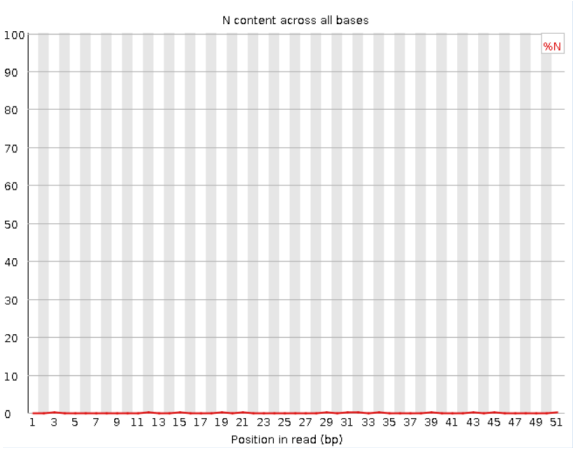
ERR173261_1



ERR173261_2

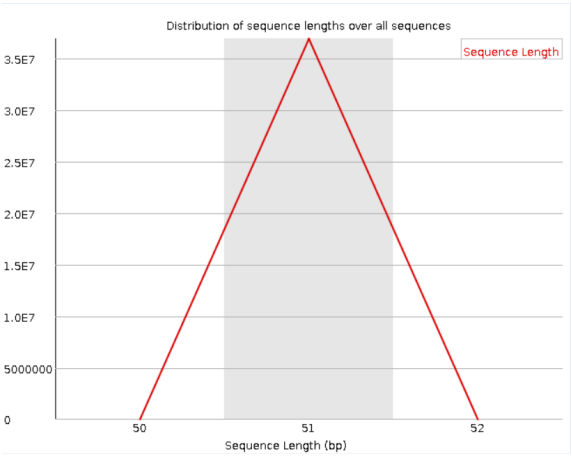


ERR173280_1

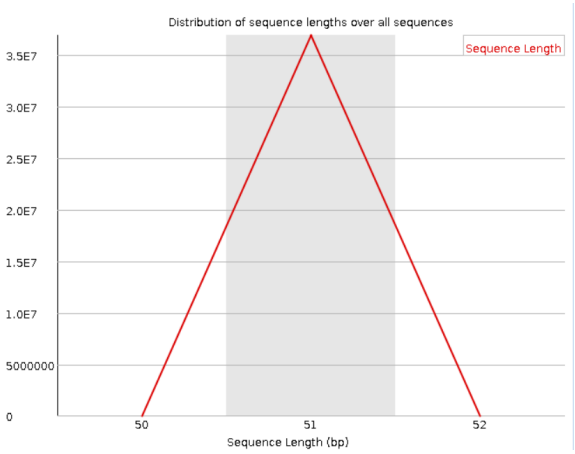


ERR173280_2

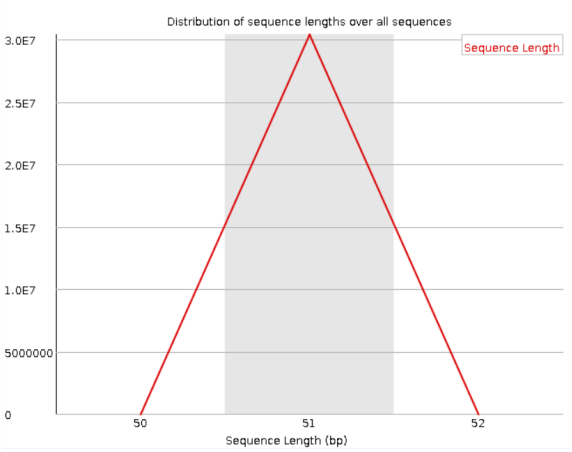
Sequence length distribution



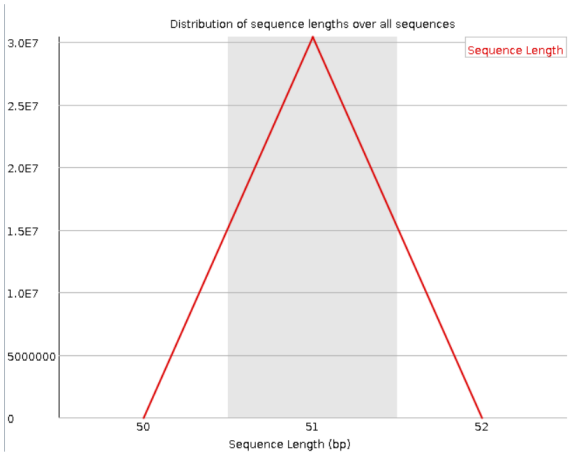
ERR173261_1



ERR173261_2

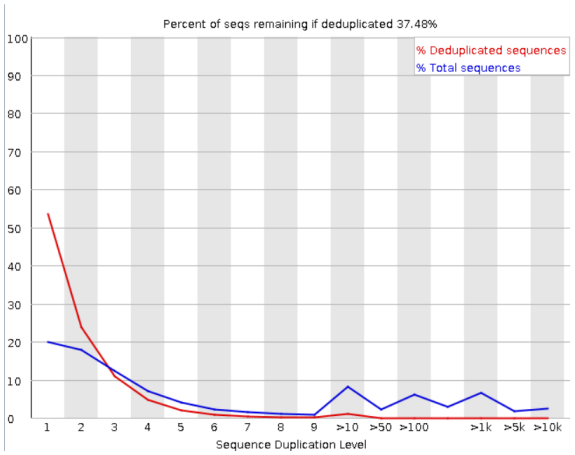


ERR173280_1

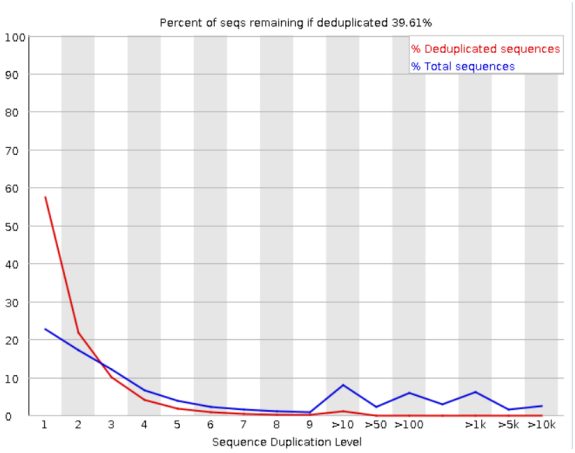


ERR173280_2

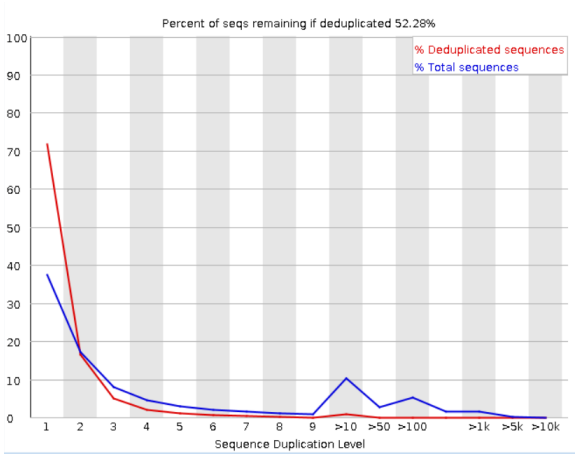
Sequence duplication levels



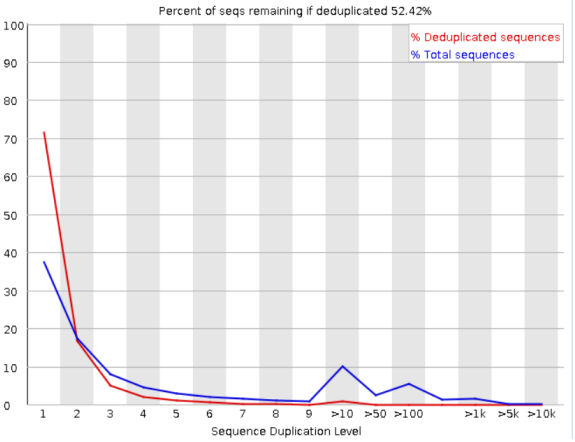
ERR173261_1



ERR173261_2



ERR173280_1



ERR173280_2

Overrepresented sequences

Overrepresented sequences			
Sequence	Count	Percentage	Possible Source
GGACAGGCTGCATCAGAA...	63239	0.171	No Hit
CTGCATCAGAAGAGGCCA...	42484	0.115	No Hit
CAAATATTCAAACGAGAA...	38802	0.105	No Hit

ERR173261_1

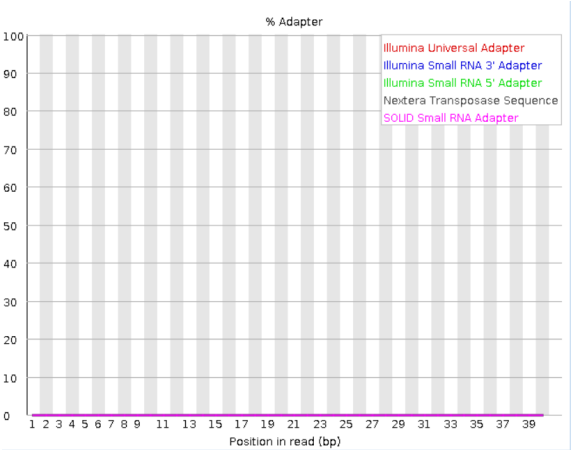
Overrepresented sequences			
Sequence	Count	Percentage	Possible Source
NNNNNNNNNNNNNNNN...	209950	0.569	No Hit
GGACAGGCTGCATCAGAA...	51200	0.139	No Hit

ERR173261_2

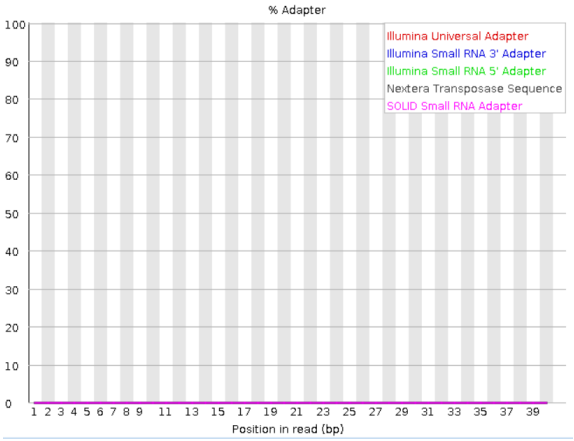
For the other two files there are no overrepresented sequences.

Adapter content

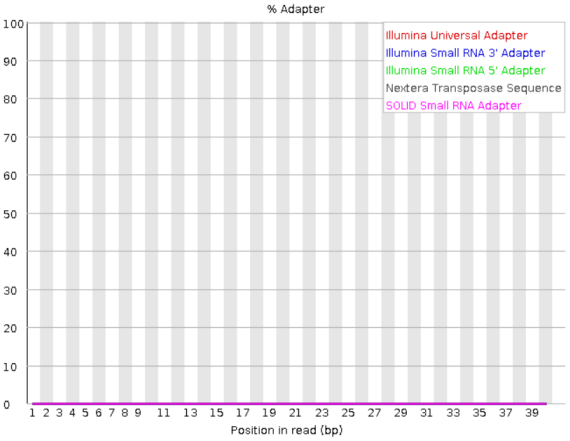
It seems that no sample had any of the known adapter sequences that are listed on the legend of the graphs.



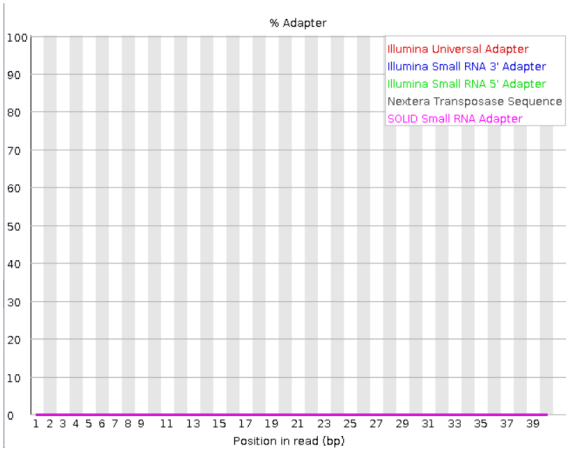
ERR173261_1



ERR173261_2



ERR173280_1



ERR173280_2

Minion

This step was performed in order to identify unknown adapters in the samples. FastQC may have indicated the abscence of any of the known adapters, but further examination is needed to ensure that the sequences are free from adapter sequences.

```
cd ~/tools
wget https://genomics-lab.fleming.gr/fleming/uoa/vm/minion
chmod u+x minion
~/tools/minion --help

~/tools/minion search-adapter -i ERR173261_1.fastq
~/tools/minion search-adapter -i ERR173261_2.fastq

~/tools/minion search-adapter -i ERR173280_1.fastq
~/tools/minion search-adapter -i ERR173280_2.fastq
```

The results provided by minion were:

- ERR173261_1

```
criterion=sequence-density
sequence-density=0.20
sequence-density-rank=1
fanout-score=2.96
fanout-score-rank=21
prefix-density=0.20
prefix-fanout=2.8
sequence=GGGAGGCTGAGGCAGGAGAAT

criterion=fanout-score
sequence-density=0.01
sequence-density-rank=36
fanout-score=24.16
fanout-score-rank=1
prefix-density=0.05
prefix-fanout=3.2
sequence=CCCCGCCCCCGGAGCCCCGCGGACGCTACGCCGCGACGAGTAGGAGGGCCGCTGCGGTGAGCCTTGAAGCCTAGGGC
GCGGGCCCGGGTGGAGCCGCCGCAGGTGCAGATCTTGGTGGTAGTAGCAAATATTCAAACGAGAACTTTGAAGGCCGAAGTGGAGA
AGGGTTCCATGTGAACAGCAGTTGAACATGGGTCAGTCGGTCCTGAGAGATGGGCGAGCGCCGTTCCGAAGGGACGGGCGATGGCC
TCCGTTGCCCTCGGCCGATCGAAAGGGAGTCGGGTTCAGATCCCCGAATCCGGAGTGGCGGAGATGGGCGCCGCGAGGCGTCCAGT
GCGGTAACGCGACCGATCCCGGAGAAGCCGGCGGGAGCCCCGGGGAGAGTTCTCTTTTCTTTGTGAAGGGCAGGGCGCCCTGGAAT
GGGTTCGCCCCGAGAGAGGGGCCCGTGCCTTGGAAAGCGTCGCGGTTCCGGCGGCGTCCGGTGAGCTCTCGCTGGCCCTTGAAAAT
CCGG
```

- ERR173261_2

```
criterion=sequence-density
sequence-density=0.19
sequence-density-rank=1
fanout-score=2.97
fanout-score-rank=25
prefix-density=0.20
prefix-fanout=2.9
sequence=GGGAGGCTGAGGCAGGAGAAT


criterion=fanout-score
sequence-density=0.01
sequence-density-rank=41
fanout-score=21.60
fanout-score-rank=1
prefix-density=0.10
prefix-fanout=1.1
sequence=CCCGCCCCCGGAGCCCCGCGGACGCTACGCCGCGACGAGTAGGAGGGCCGCTGCGGTGAGCCTTGAAGCCTAGGGCG
CGGGCCCGGGTGGAGCCGCCGCAGGTGCAGATCTTGGTGGTAGTAGCAAATATTCAAACGAGAACTTTGAAGGCCGAAGTGGAGAA
GGGTTCCATGTGAACAGCAGTTGAACATGGGTCAGTCGGTCCTGAGAGATGGGCGAGCGCCGTTCCGAAGGGACGGGCGATGGCCT
CCGTTGCCCTCGGCCGATCGAAAGGGAGTCGGGTTCAGATCCCCGAATCCGGAGTGGCGGAGATGGGCGCCGCGAGGCGTCCAGTG
CGGTAACGCGACCGATCCCGGAGAAGCCGGCGGGAGCCCCGGGGAGAGTTCTCTTTTCTTTGTGAAGGGCAGGGCGCCCTGGAATG
GGTTCGCCCCGAGAGAGGGGCCCGTGCCTTGGAAAGCGTCGCGGTTCCGGCGGCGTCCGGTGAGCTCTCGCTGGCCCTTGAAAATC
CGGG
```

- ERR173280_1

```
criterion=sequence-density
sequence-density=0.12
sequence-density-rank=1
fanout-score=3.01
fanout-score-rank=20
prefix-density=0.11
prefix-fanout=3.0
sequence=GCCTGTAATCCCAGCACTTTGGGAGGCC


criterion=fanout-score
sequence-density=0.03
sequence-density-rank=24
fanout-score=9.69
fanout-score-rank=1
prefix-density=0.11
prefix-fanout=2.6
sequence=CCCAGCTACTCGGGAGGCTGAG
```

- ERR173280_2

```
criterion=sequence-density
sequence-density=0.12
sequence-density-rank=1
fanout-score=3.35
fanout-score-rank=15
prefix-density=0.12
prefix-fanout=3.4
sequence=GCCTGTAATCCCAGCACTTTGGGAGGCC


criterion=fanout-score
sequence-density=0.03
sequence-density-rank=27
fanout-score=9.95
fanout-score-rank=1
prefix-density=0.12
prefix-fanout=2.6
sequence=CCCAGCTACTCGGGAGGCTGAG
```

Since in all cases the sequence density was extremely low the sequencies were not deemed intrusive to the overall quality of the sequencies and therefore were not removed.

Cutadapt

Cutadapt was used to remove the sequences of the adaptors and of unwanted overrepresented sequences that were found with FastQC and minion. It was installed in the VM by downloading it from the official website (**https://pypi.org/project/cutadapt/)**, as it was not possible to be installed from terminal without being a root. Also pip was needed to install setuptools_scm (needed to install cutadapt), which was once again downloaded manually due to a lack of root priviledges.

```
# install pip
tar xvfz pip-23.0.tar.gz
cd pip-23.0
python setup.py install --user
# make sure the installation was successful
```

```
/home/grad/.local/bin/pip3.8

# pip was added to bash PATH so that it could be called from any directory
nano ~/.bashrc # open bash configuration file
export PATH=$PATH:/home/grad/.local/bin # written at the end of the file

# reload terminal to set the change in effect
source ~/.bashrc

# setuptools were requested to install cutadapt
pip install setuptools_scm

pip install cutadapt
```

In ERR173261_2 there was an overexpressed sequence that was "`NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN`". This was probably a contaminant such as adapter dimers, primer-dimer artifacts, or sequencing artifacts. Therefore it was removed with cutadapt. For the other overexpressed sequences one cannot be sure whether they are contaminants or have a biological significance. They were cross tested however with the adapter sequences that correspond to the sequencing platforms that were used in the experiment.

For Illumina GAIIx, the adapter sequences are typically:

- Read 1: AGATCGGAAGAGCACACGTCTGAACTCCAGTCA
- Read 2: AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT

For Illumina HiSeq2000 and HiSeq2500, the adapter sequences are typically:

- Read 1: AGATCGGAAGAGCACACGTCTGAACTCCAGTCA
- Read 2: AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT

For the SOLiD 4 platform, the adapter sequences are typically:

- Read 1: CTGTCTCTTATACACATCTCCGAGCCCACGAGAC
- Read 2: CTGTCTCTTATACACATCTGACGCTGCCGACGA

To see if any of these sequences was the same as an adapter sequence the following script was build. No similarities were found, therefore only the sequence with the many N's was removed with cutadapt. The adapter sequence themselves

were not removed because the FastQC analysis dictated a zero adapter content for all four samples.

```
# Create arrays
fastq1=("GGACAGGCTGCATCAGAAGAGGCCATCAAGCAGATCACTGTCCTTCTGCCA" "CTGCATCAGAAGAGGCCATCAAG
CAGATCACTGTCCTTCTGCCATGGCCCT" "CAAATATTCAAACGAGAACTTTGAAGGCCGAAGTGGAGAAGGGTTCCATGT")
fastq2=("GGACAGGCTGCATCAGAAGAGGCCATCAAGCAGATCACTGTCCTTCTGCCA")
adapters=("AGATCGGAAGAGCACACGTCTGAACTCCAGTCA" "AGATCGGAAGAGCACACGTCTGAACTCCAGTCA" "CTG
TCTCTTATACACATCTCCGAGCCCACGAGAC")

# Check duplicates between fastq1 and adapters
for seq1 in "${fastq1[@]}"; do
  for seq2 in "${adapters[@]}"; do
    if [ "$seq1" == "$seq2" ]; then
      echo "Duplicate found: $seq1"
    fi
  done
done

# Check duplicates between fastq2 and adapters
for seq1 in "${fastq2[@]}"; do
  for seq2 in "${adapters[@]}"; do
    if [ "$seq1" == "$seq2" ]; then
      echo "Duplicate found: $seq1"
    fi
  done
done
```

```
# remove the overrepresented sequence
cutadapt -u 30 -o ERR173261_2_trimmed.fastq ERR173261_2.fastq

=== Summary ===

Total reads processed:              36,899,101
Reads written (passing filters):    36,899,101 (100.0%)

Total basepairs processed: 1,881,854,151 bp
Total written (filtered):    774,881,121 bp (41.2%)
```

Spliced alignment

Spliced alignment is the process of aligning the reads from RNA-seq experiments to a reference genome. The goal of spliced alignment is to map the reads to the correct

exonic regions, while properly accounting for introns, in order to accurately quantify the expression levels of individual genes.

The whole hg38 (human genome) was used as a reference genome, because it was the most updated version of the human genome in NCBI and also because the human genome had been used in the original analysis. Two files were needed for the ensuing analysis: a FASTA file and a GTF format file. The FASTA file contained the entire genome sequence, including exons, introns, and intergenic regions and it was used to create the bowtie index files that were needed for the spliced alignment. The GTF file contain information about the location and structure of transcripts, including exon locations and gene structure information and it was used to later perform the differential expression analysis.

```
tar xvfz genome_assemblies_genome_gtf.tar
cd ncbi-genomes-2023-02-06
gunzip GCF_000001405.40_GRCh38.p14_genomic.gtf.gz
mv GCF_000001405.40_GRCh38.p14_genomic.gtf hg38.gtf
mv GCF_000001405.40_GRCh38.p14_genomic.gtf ~/  #move to home

tar -xvf genome_assemblies_genome_fasta.tar
cd ncbi-genomes-2023-02-08
gunzip GCF_000001405.40_GRCh38.p14_genomic.fna.gz
mv GCF_000001405.40_GRCh38.p14_genomic.fna hg38.fasta
mv hg38.fasta ~/  #move to home

cd ~/
```

For the alignment itself, TopHat2 was used. It is a spliced alignment tool that is specifically designed for RNA-seq data. It can be used to align paired-end reads to a reference genome.

```
# a bowtie index file of hg38 must be created first
bowtie2-build hg38.fa hg38

# alignment with tophat
tophat2 -p 8 --no-coverage-search -o ERR173261_aligned hg38 ERR173261_1.fastq.gz ERR17
3261_2.fastq.gz

tophat2 -p 8 --no-coverage-search -o ERR173280_aligned hg38 ERR173280_1.fastq.gz ERR17
3280_1.fastq.gz
```

The output files of the alignment for each of the two samples were:

1. `accepted_hits.bam`: a BAM file containing the aligned reads.
2. `align_summary.txt`: a summary of the alignment statistics.
3. `insertions.bed`: a BED file containing information about insertions.
4. `deletions.bed`: a BED file containing information about deletions.
5. `junctions.bed`: a BED file containing information about junctions.
6. `prep_reads.info`: the number of reads that were initially read from the input files, the number of reads that were discarded due to various filters, and the number of reads that were used for the alignment.
7. `log files`

Differential expression analysis

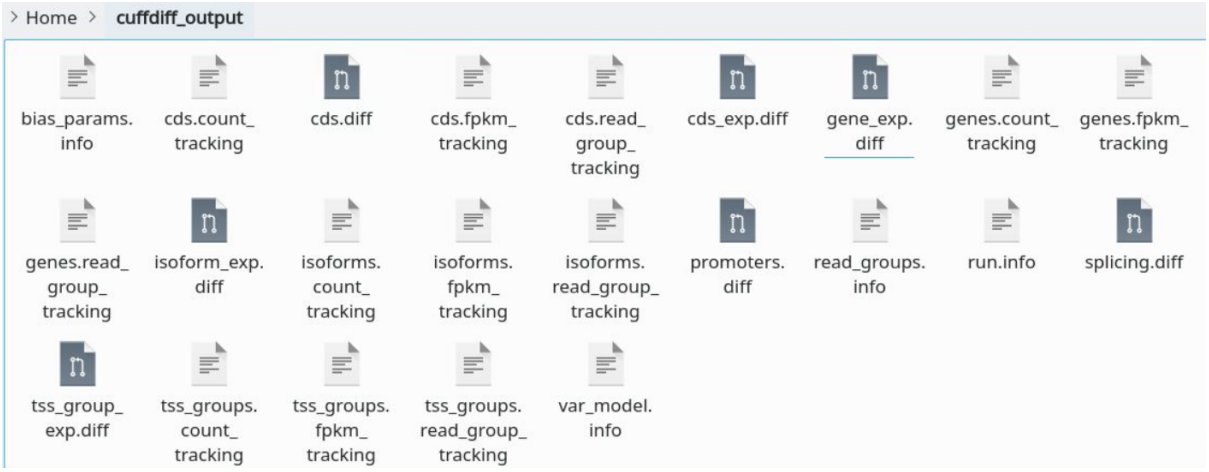
Differential expression analysis is a statistical analysis that is performed to identify genes that are differentially expressed between two or more experimental conditions. In RNA-seq experiments, it is used to compare the transcriptomes of different samples to identify genes that are up- or down-regulated in response to a specific experimental condition or treatment. It provides insights for the biological processes and pathways and it can be used to identify genes that are involved in a specific disease or condition, which can lead to the development of new diagnostic or therapeutic approaches.

To perform the differential expression analysis the accepted_hits.bam file from the spliced alignment was used. The CLI tool that was chosen to perform the analysis was cuffdiff, which is a famous tool for RNA-seq differential expression analysis. It is part of the Cufflinks package and allows for the comparison of gene expression between two or more experimental conditions. Cuffdiff calculates significant differences in gene expression using a statistical framework and generates output files that can be used for downstream analysis and visualization.

The goal of this analysis step was to identify genes that were differentially expressed between the two samples ERR173261 and ERR173280.

```
cuffdiff -p 8 -o cuffdiff_output -g hg38.gtf ERR173261_accepted_hits.bam ERR173280_acc
epted_hits.bam
```

The analysis output consisted of a number of files, shown in the image below.



The gene_exp.diff file contained information about the differences in expression levels of the genes between the two samples. It presented the gene ID the status of the files and some statistics, while also listing if according to them the level of differential expression was significant between the two files. The genes with significant differential expression were extracted using the following bash snippet.

```
grep "yes" "gene_exp.diff" | grep -v "^test_id" > significant_diff_expressed_genes.txt
```

```
test_id gene_id gene  locus sample_1  sample_2  status  value_1 value_2 log2(fold_chan
ge) test_stat p_value q_value significant

ADH1C ADH1C ADH1C NC_000004.12:99336496-99352746  q1  q2  OK  0 21.827  inf -nan  5e-0
5 0.0173774 yes
```

```
AKR1B1P2  AKR1B1P2  AKR1B1P2  NC_000003.12:74135926-74137588  q1  q2  OK  0 1.32817 in
f -nan  0.0002  0.0478442 yes
APOL3 APOL3 APOL3 NC_000022.11:36140322-36166177  q1  q2  OK  0 0.895043  inf -nan  5e
-05 0.0173774 yes
C3  C3  C3  NC_000019.10:6677703-6720650  q1  q2  OK  3.03324 108.954 5.16672 5.13128
5e-05 0.0173774 yes
CALM2P1 CALM2P1 CALM2P1 NC_000017.11:70241254-70242378  q1  q2  OK  0 2.33754 inf -nan
0.0001  0.0287813 yes
CBWD6 CBWD6 CBWD6 NC_000009.12:41131308-41205735  q1  q2  OK  0.00264303  1.13156 8.74
19  0.101578  5e-05 0.0173774 yes
CCDC74B CCDC74B CCDC74B NC_000002.12:130129622-130145112  q1  q2  OK  0 0.616557  inf
-nan  5e-05 0.0173774 yes
CCL20 CCL20 CCL20 NC_000002.12:227813841-227817556  q1  q2  OK  0 6.97253 inf -nan  5e
-05 0.0173774 yes
CEACAM7 CEACAM7 CEACAM7 NC_000019.10:41673302-41688270  q1  q2  OK  0 1.14253 inf -nan
5e-05 0.0173774 yes
CFTR  CFTR  CFTR  NC_000007.14:117480024-117668665  q1  q2  OK  1.60421 59.2457 5.2067
8 4.75351 5e-05 0.0173774 yes
COX20P2 COX20P2 COX20P2 NC_000002.12:230907327-230961201  q1  q2  OK  0 2.78923 inf -n
an  0.0002  0.0478442 yes
CXCL6 CXCL6 CXCL6 NC_000004.12:73836677-73838760  q1  q2  OK  0 12.7132 inf -nan  5e-0
5 0.0173774 yes
EGR1  EGR1  EGR1  NC_000005.10:138465478-138469303  q1  q2  OK  256.112 13.4506 -4.251
03  -4.52141  5e-05 0.0173774 yes
EI24P2  EI24P2  EI24P2  NC_000001.11:158454115-158455468  q1  q2  OK  0 2.0399  inf -n
an  0.0002  0.0478442 yes
EIF3LP3 EIF3LP3 EIF3LP3 NC_000010.11:38079900-38080863  q1  q2  OK  0 3.07348 inf -nan
5e-05 0.0173774 yes
FAM153A FAM153A FAM153A NC_000005.10:177672378-177783425  q1  q2  OK  0.0542842 3.8707
1 6.15592 2.22569 5e-05 0.0173774 yes
FCGR2B  FCGR2B  FCGR2B  NC_000001.11:161647242-161678654  q1  q2  OK  0 0.452199  inf
-nan  5e-05 0.0173774 yes
FN1 FN1 FN1 NC_000002.12:215312058-215436068  q1  q2  OK  0.764097  78.0633 6.67474 3.
1862  5e-05 0.0173774 yes
FOSB  FOSB  FOSB  NC_000019.10:45467995-45475179  q1  q2  OK  64.3312 0.733522  -6.454
53  -5.20448  0.00015 0.0406324 yes
GNAI2P1 GNAI2P1 GNAI2P1 NC_000012.12:14254913-14255486  q1  q2  OK  0 8.40423 inf -nan
0.0001  0.0287813 yes
HLA-A_4 HLA-A_4 HLA-A NT_167246.2:1197073-1200428 q1  q2  OK  0 37.58 inf -nan  5e-05
0.0173774 yes
HLA-C_3 HLA-C_3 HLA-C NT_167246.2:2577800-2581132 q1  q2  OK  0 5.35862 inf -nan  5e-0
5 0.0173774 yes
HLA-DQB1  HLA-DQB1  HLA-DQB1  NC_000006.12:32659466-32666657  q1  q2  OK  0 3.89065 in
f -nan  0.00015 0.0406324 yes
HLA-DRB1_5  HLA-DRB1_5  HLA-DRB1  NT_167249.2:3979127-3993841 q1  q2  OK  0 3.82772 in
f -nan  5e-05 0.0173774 yes
HNRNPH1_1 HNRNPH1_1 HNRNPH1 NW_016107298.1:380456-401011  q1  q2  OK  0.0328291 0.4389
69  3.74107 0.633085  5e-05 0.0173774 yes
HTR1A HTR1A HTR1A NC_000005.10:63957873-63962445  q1  q2  OK  0 0.506026  inf -nan  5e
-05 0.0173774 yes
IL32  IL32  IL32  NC_000016.10:3065402-3069651  q1  q2  OK  1.58152 94.6545 5.90329 4.
29691 5e-05 0.0173774 yes
IL7R  IL7R  IL7R  NC_000005.10:35856890-35879603  q1  q2  OK  0 0.57033 inf -nan  5e-0
5 0.0173774 yes
KCNE3 KCNE3 KCNE3 NC_000011.10:74454840-74467549  q1  q2  OK  0 1.0515  inf -nan  5e-0
5 0.0173774 yes
KRT6B KRT6B KRT6B NC_000012.12:52446650-52452146  q1  q2  OK  0 1.88659 inf -nan  5e-0
5 0.0173774 yes
```

```
LAPTM4BP2 LAPTM4BP2 LAPTM4BP2 NC_000003.12:72884231-72884879  q1  q2  OK  0 2.90269 in
f -nan  5e-05 0.0173774 yes
LDHAP1  LDHAP1  LDHAP1  NC_000004.12:4893623-4895268  q1  q2  OK  0 1.94407 inf -nan
5e-05 0.0173774 yes
LDHAP5  LDHAP5  LDHAP5  NC_000010.11:118932097-118933248  q1  q2  OK  0 6.10287 inf -n
an  5e-05 0.0173774 yes
LGI1  LGI1  LGI1  NC_000010.11:93757935-93798159  q1  q2  OK  0 0.918462  inf -nan  0.
0001  0.0287813 yes
LINC00229 LINC00229 LINC00229 NC_000022.11:44606327-44625419  q1  q2  OK  0 0.401583
inf -nan  5e-05 0.0173774 yes
LOC124901275  LOC124901275  LOC124901275  NC_000006.12:22200988-22205065  q1  q2  OK
0 0.394102  inf -nan  0.0001  0.0287813 yes
LOC124902122  LOC124902122  LOC124902122  NC_000009.12:14927927-14931668  q1  q2  OK
0 0.449977  inf -nan  5e-05 0.0173774 yes
LOC124903889  LOC124903889  LOC124903889  NC_000017.11:198796-200330  q1  q2  OK  0 1.
96515 inf -nan  0.0001  0.0287813 yes
LOC124907888  LOC124907888  LOC124907888  NC_000002.12:128742411-128746430  q1  q2  OK
0 0.466191  inf -nan  0.0002  0.0478442 yes
LOC389249 LOC389249 LOC389249 NC_000004.12:187942163-187992879  q1  q2  OK  0 2.05115
inf -nan  0.0001  0.0287813 yes
MIR210HG  MIR210HG  MIR210HG  NC_000011.10:565656-568457  q1  q2  OK  0 0.688089  inf
-nan  0.0001  0.0287813 yes
MIR210HG_1  MIR210HG_1  MIR210HG  NT_187586.1:95311-98112 q1  q2  OK  0 0.819153  inf
-nan  5e-05 0.0173774 yes
MIR492  MIR492  MIR492  NC_000012.12:94834397-94835028  q1  q2  OK  0 30.9709 inf -nan
0.00015 0.0406324 yes
MT2A  MT2A  MT2A  NC_000016.10:56608583-56609497  q1  q2  OK  1491.58 55.9001 -4.73784
-5.18118  5e-05 0.0173774 yes
MTND4P24  MTND4P24  MTND4P24  NC_000023.11:126472884-126473283  q1  q2  OK  0 5.82348
inf -nan  0.0001  0.0287813 yes
NPIPB4  NPIPB4  NPIPB4  NC_000016.10:21833630-21857756  q1  q2  OK  0.775211  0 -inf
-nan  5e-05 0.0173774 yes
NTM NTM NTM NC_000011.10:131370614-132336822  q1  q2  OK  0.000622555 0.713914  10.163
3 0.0543556 5e-05 0.0173774 yes
PARPBP  PARPBP  PARPBP  NC_000012.12:102120182-102197833  q1  q2  OK  0.135661  0.6477
9 2.25552 0.694265  0.0002  0.0478442 yes
PKHD1 PKHD1 PKHD1 NC_000006.12:51615298-52087615  q1  q2  OK  0.0158083 1.96439 6.9572
5 2.26637 5e-05 0.0173774 yes
PKMP5 PKMP5 PKMP5 NC_000006.12:5972323-5974405  q1  q2  OK  0 4.05086 inf -nan  5e-05
0.0173774 yes
PLA2G2A PLA2G2A PLA2G2A NC_000001.11:19975430-19980434  q1  q2  OK  0 2.55135 inf -nan
5e-05 0.0173774 yes
PPP1R15A  PPP1R15A  PPP1R15A  NC_000019.10:48872420-48876058  q1  q2  OK  310.813 19.0
742 -4.02635  -4.27642  0.0002  0.0478442 yes
PROM1 PROM1 PROM1 NC_000004.12:15968227-16084023  q1  q2  OK  0.225932  11.8493 5.7127
7 4.13511 5e-05 0.0173774 yes
PSMA6P2 PSMA6P2 PSMA6P2 NC_000023.11:12825810-12826996  q1  q2  OK  0 1.98242 inf -nan
0.0002  0.0478442 yes
PTGES3P4  PTGES3P4  PTGES3P4  NC_000010.11:102845600-102846026  q1  q2  OK  0 4.61864
inf -nan  5e-05 0.0173774 yes
RPH3AL  RPH3AL  RPH3AL  NC_000017.11:212388-352807  q1  q2  OK  0.0254559 7.08755 8.12
115 3.6739  5e-05 0.0173774 yes
RPL37AP5  RPL37AP5  RPL37AP5  NC_000007.14:139227470-139227859  q1  q2  OK  0 9.26577
inf -nan  5e-05 0.0173774 yes
RPL3P12 RPL3P12 RPL3P12 NC_000023.11:122538231-122539516  q1  q2  OK  0 3.01915 inf -n
an  5e-05 0.0173774 yes
RPS20P10  RPS20P10  RPS20P10  NC_000002.12:71984140-71984471  q1  q2  OK  0 19.2058 in
f -nan  5e-05 0.0173774 yes
```

```
RPS7  RPS7  RPS7  NC_000002.12:3575259-3580920  q1  q2  OK  1.22482 37.2529 4.92671 4.
5076  5e-05 0.0173774 yes
SCN7A SCN7A SCN7A NC_000002.12:166403572-166494249  q1  q2  OK  0 0.621388  inf -nan
5e-05 0.0173774 yes
SERPINA3  SERPINA3  SERPINA3  NC_000014.9:94612390-94624053 q1  q2  OK  18.1911 638.51
5.13341 4.8124  0.00015 0.0406324 yes
SLC17A4 SLC17A4 SLC17A4 NC_000006.12:25723742-25832052  q1  q2  OK  0.0292022 2.24983
6.2676  1.12989 5e-05 0.0173774 yes
SLC37A4_1 SLC37A4_1 SLC37A4 NW_009646203.1:40365-52476  q1  q2  OK  0 1.03056 inf -nan
0.0001  0.0287813 yes
SNCAIP  SNCAIP  SNCAIP  NC_000005.10:122311352-122479087  q1  q2  OK  0.000255296 0.38
683 10.5653 0.0208904 5e-05 0.0173774 yes
SORD2P_1  SORD2P_1  SORD2P  NT_187605.1:150086-216644 q1  q2  OK  0 0.623281  inf -nan
0.0002  0.0478442 yes
SULT1B1 SULT1B1 SULT1B1 NC_000004.12:69721166-69760620  q1  q2  OK  0 0.38664 inf -nan
5e-05 0.0173774 yes
SZRD1_1 SZRD1_1 SZRD1 NW_025791756.1:171376-241687  q1  q2  OK  0 0.742174  inf -nan
5e-05 0.0173774 yes
TCP11 TCP11 TCP11 NC_000006.12:35118074-35141339  q1  q2  OK  0.512821  0 -inf  -nan
5e-05 0.0173774 yes
TGFBI TGFBI TGFBI NC_000005.10:136028987-136063818  q1  q2  OK  2.91068 49.4607 4.0868
6 3.96961 0.0002  0.0478442 yes
THBD  THBD  THBD  NC_000020.11:23045632-23049672  q1  q2  OK  1.9661  54.6308 4.79631
4.64269 5e-05 0.0173774 yes
TNFSF14 TNFSF14 TNFSF14 NC_000019.10:6661252-6670588  q1  q2  OK  0 0.936752  inf -nan
5e-05 0.0173774 yes
TRNH  TRNH  TRNH  NC_012920.1:12137-12206 q1  q2  OK  0 127.816 inf -nan  0.0001  0.02
87813 yes
TRNT  TRNT  TRNT  NC_012920.1:15887-15953 q1  q2  OK  0 114.448 inf -nan  5e-05 0.0173
774 yes
TUBBP2  TUBBP2  TUBBP2  NC_000013.11:41383681-41385076  q1  q2  OK  0 9.1124  inf -nan
5e-05 0.0173774 yes
TXNIP TXNIP TXNIP NC_000001.11:145992434-145996579  q1  q2  OK  6.50734 112.9 4.11683
4.42998 5e-05 0.0173774 yes
UFM1P2  UFM1P2  UFM1P2  NC_000017.11:35375235-35377746  q1  q2  OK  0 0.584927  inf -n
an  0.0001  0.0287813 yes
```

In a more concise way only the gene ID’s and gene names that were significantly differentially expressed are presented below:

```
grep "yes" "gene_exp.diff" | grep -v "^test_id" | cut -f2,3
```

```
gene_id gene
ADH1C ADH1C
AKR1B1P2  AKR1B1P2
APOL3 APOL3
C3  C3
CALM2P1 CALM2P1
CBWD6 CBWD6
CCDC74B CCDC74B
```

```
CCL20 CCL20
CEACAM7 CEACAM7
CFTR  CFTR
COX20P2 COX20P2
CXCL6 CXCL6
EGR1  EGR1
EI24P2  EI24P2
EIF3LP3 EIF3LP3
FAM153A FAM153A
FCGR2B  FCGR2B
FN1 FN1
FOSB  FOSB
GNAI2P1 GNAI2P1
HLA-A_4 HLA-A
HLA-C_3 HLA-C
HLA-DQB1  HLA-DQB1
HLA-DRB1_5  HLA-DRB1
HNRNPH1_1 HNRNPH1
HTR1A HTR1A
IL32  IL32
IL7R  IL7R
KCNE3 KCNE3
KRT6B KRT6B
LAPTM4BP2 LAPTM4BP2
LDHAP1  LDHAP1
LDHAP5  LDHAP5
LGI1  LGI1
LINC00229 LINC00229
LOC124901275  LOC124901275
LOC124902122  LOC124902122
LOC124903889  LOC124903889
LOC124907888  LOC124907888
LOC389249 LOC389249
MIR210HG  MIR210HG
MIR210HG_1  MIR210HG
MIR492  MIR492
MT2A  MT2A
MTND4P24  MTND4P24
NPIPB4  NPIPB4
NTM NTM
PARPBP  PARPBP
PKHD1 PKHD1
PKMP5 PKMP5
PLA2G2A PLA2G2A
PPP1R15A  PPP1R15A
PROM1 PROM1
PSMA6P2 PSMA6P2
PTGES3P4  PTGES3P4
RPH3AL  RPH3AL
RPL37AP5  RPL37AP5
RPL3P12 RPL3P12
RPS20P10  RPS20P10
RPS7  RPS7
SCN7A SCN7A
SERPINA3  SERPINA3
SLC17A4 SLC17A4
SLC37A4_1 SLC37A4
SNCAIP  SNCAIP
```

```
SORD2P_1  SORD2P
SULT1B1 SULT1B1
SZRD1_1 SZRD1
TCP11 TCP11
TGFBI TGFBI
THBD  THBD
TNFSF14 TNFSF14
TRNH  TRNH
TRNT  TRNT
TUBBP2  TUBBP2
TXNIP TXNIP
UFM1P2  UFM1P2
```

IGV visualization of the top 5 differentially expressed genes

Integrative Genomics Viewer (IGV) is a high-performance visualization tool for interactive exploration of large, integrated datasets. It allows users to view and analyze genomic data, such as gene expression, copy number variation, and sequence alignments, in an intuitive and customizable interface. IGV can be used to visualize data from a variety of sources, including RNA-seq, ChIP-seq, and whole genome sequencing. Here IGV was used to visualize the top five differentially expressed genes.

The top 5 differentially expressed genes were inferred from the gene_exp.diff output file of the cuffdiff analysis.

```
grep "yes" gene_exp.diff | grep -v "^test_id" | sort -k 12n,12 -k 11nr,11 | head -n 15
| cut -f 2,3

CALM2P1 CALM2P1
GNAI2P1 GNAI2P1
LGI1  LGI1
LOC124901275  LOC124901275
LOC124903889  LOC124903889
LOC389249 LOC389249
MIR210HG  MIR210HG
MTND4P24  MTND4P24
SLC37A4_1 SLC37A4
TRNH  TRNH
UFM1P2  UFM1P2
SERPINA3  SERPINA3
```

```
HLA-DQB1  HLA-DQB1
MIR492  MIR492
FOSB  FOSB
```

CALM2P1 is a pseudogene near CALM2 gene that has lost its protein-coding ability and does not produce a functional protein. Therefore it wasn't considered as a gene and could also not be visualized in IGV. Similarly the LOC regions and any other pseudogenes from the list above were excluded. The top 5 differentially expressed genes were:

1. LGI1
2. MIR210HG
3. SLC37A4
4. SERPINA3
5. MIR492

For IGV to access the BAM files of the samples that were generated with spliced alignment an index file for each one was needed, which was created with samtools. Also the BAM files needed to be aligned to be able to work with IGV, which they indeed were.

```
# create the index bai files
samtools index ERR173261_accepted_hits.bam

samtools index ERR173280_accepted_hits.bam
```

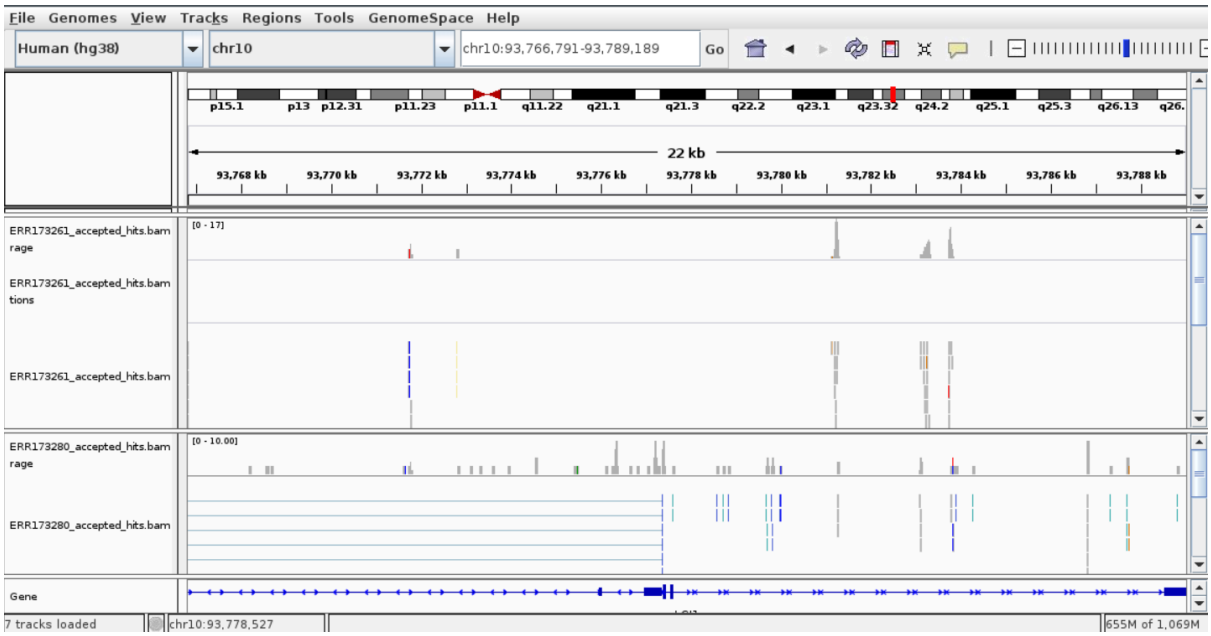
```
# check if BAM files are properly aligned
samtools view -H ERR173261_accepted_hits.bam | grep -w "SO:coordinate"
@HD VN:1.0  SO:coordinate

samtools view -H ERR173280_accepted_hits.bam | grep -w "SO:coordinate"
@HD VN:1.0  SO:coordinate
```

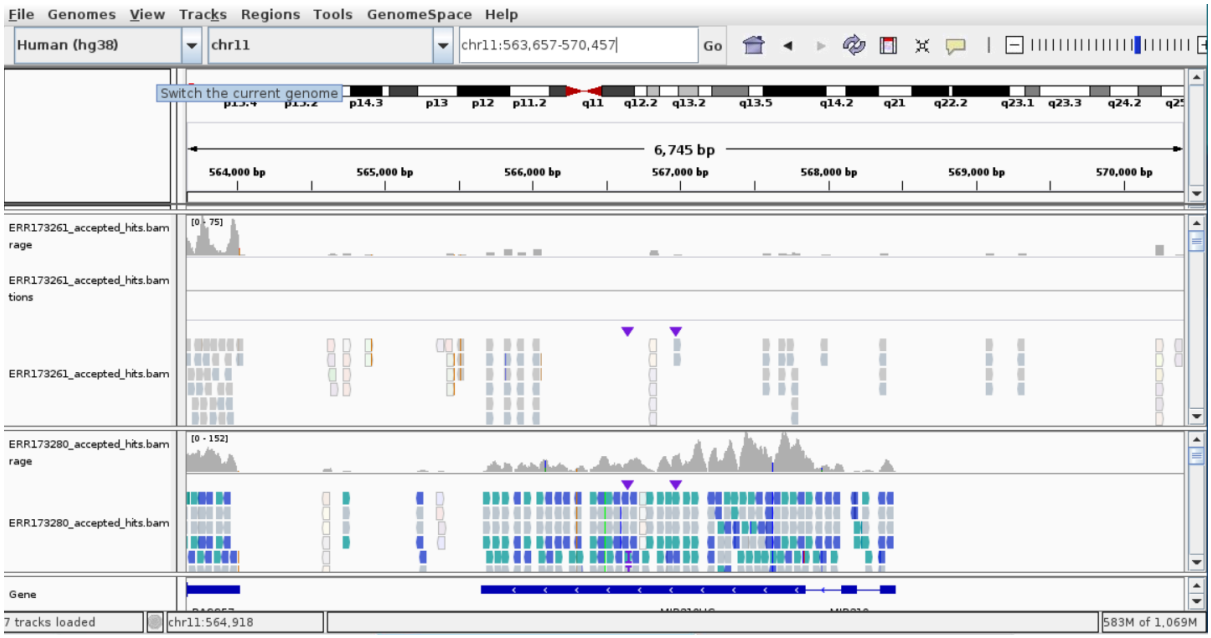
```
samtools view -h ERR173261_accepted_hits.bam | head -30
```

The IGV visualizations for each gene are presented below:

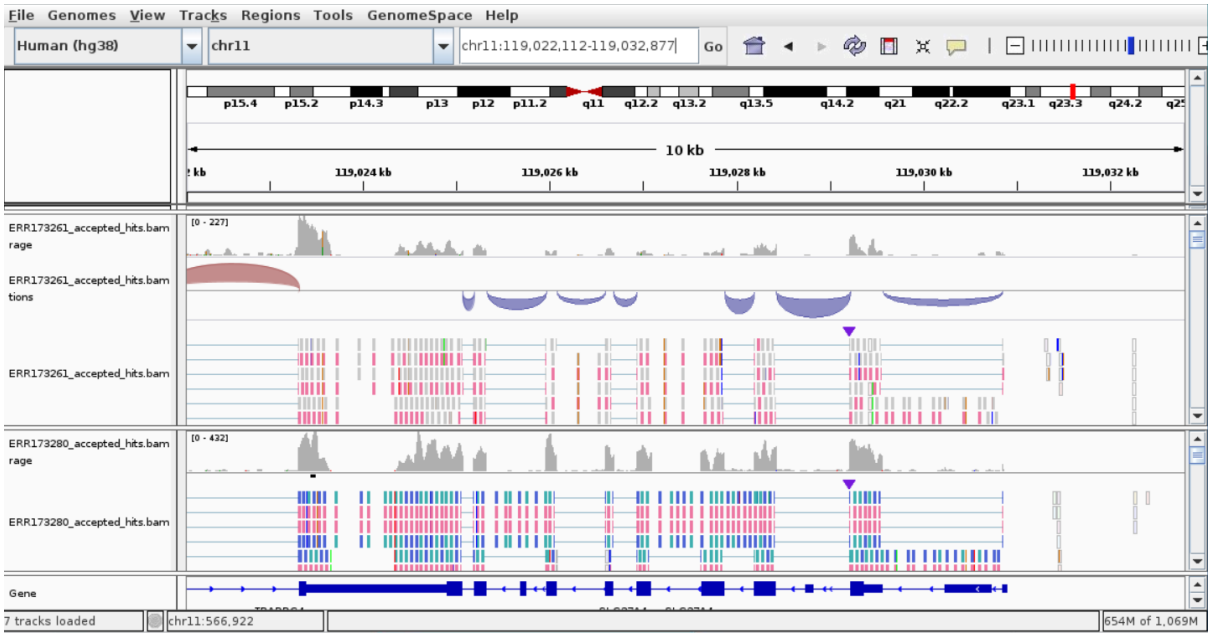
- LGI1



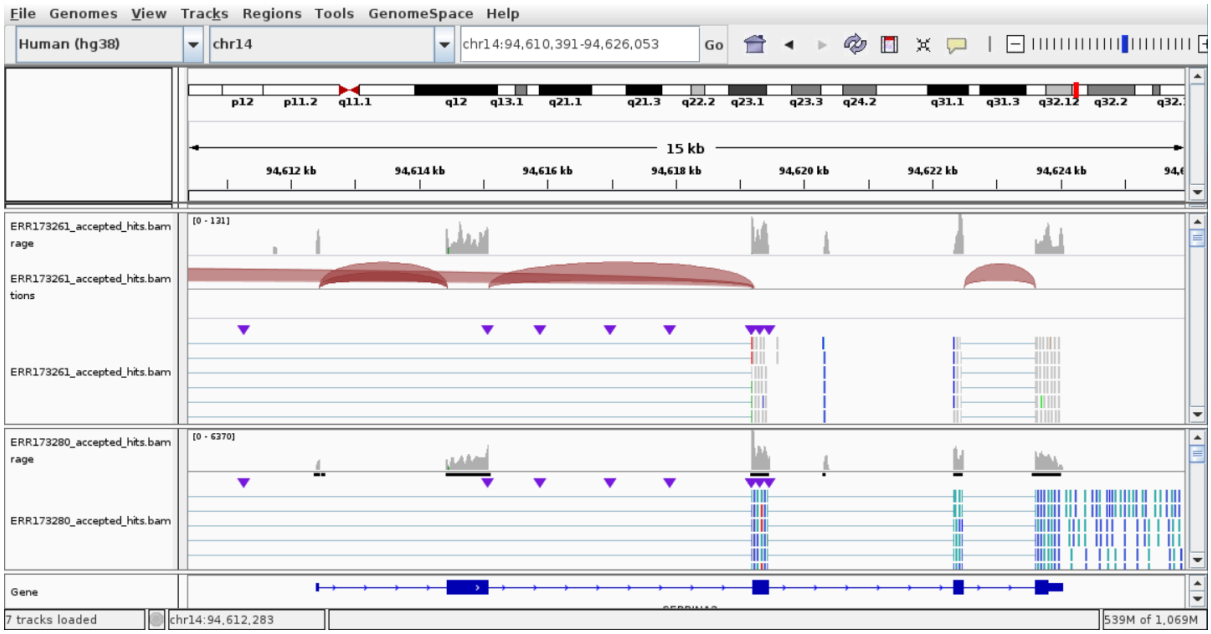
- MIR210HG



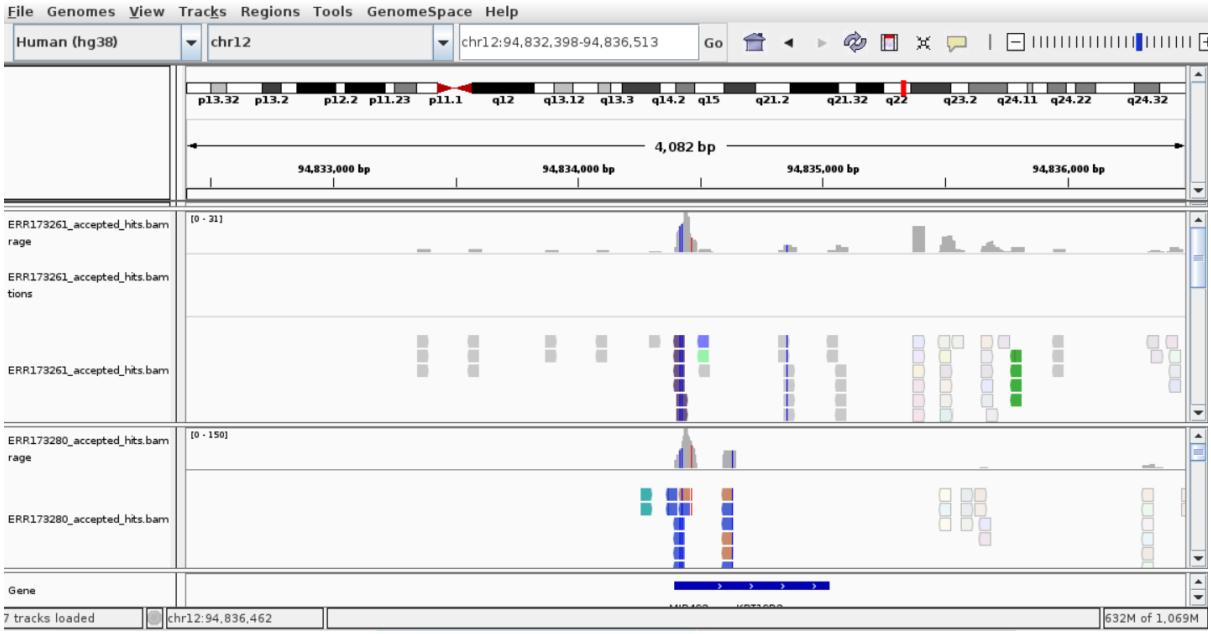
- SLC37A4



- SERPINA3



- MIR492



GO term enrichment for significant genes

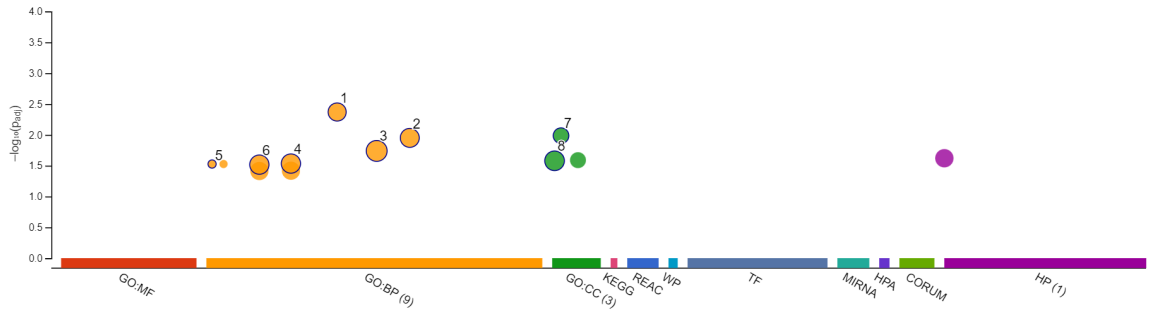
GO (Gene Ontology) term enrichment analysis is a statistical method used to determine whether a set of genes or proteins are enriched for particular biological functions, processes or cellular components, based on their GO annotations. GO terms are assigned to genes based on their predicted or known biological roles, and represent functional categories such as "molecular function", "biological process", and "cellular component".

In RNA-seq analysis, GO term enrichment analysis is often used to identify the biological pathways or functions that are overrepresented in a list of differentially expressed genes compared to the background set of all genes in the genome. This analysis can provide insights into the underlying biological processes that are affected by a particular condition or treatment, and help generate hypotheses for further investigation.

The tools that was used was g:Profiler (https://biit.cs.ut.ee/gprofiler/gost) a web server for functional enrichment analysis and conversions of gene lists. The significantly differentially expressed genes entries were filtered to keep only the gene

ID and gene name and then they were passed on to the Query Search of g:Profiler for Homo Sapiens. The results can be seen in the image below.

```
cut -f1,2 significant_diff_expressed_genes.txt > filtered_genes.txt
```



ID	Source	Term ID	Term Name	p_{adj} (query_1)
1	GO:BP	GO:0042592	homeostatic process	4.270×10^{-3}
2	GO:BP	GO:0065008	regulation of biological quality	1.121×10^{-2}
3	GO:BP	GO:0050896	response to stimulus	1.824×10^{-2}
4	GO:BP	GO:0023051	regulation of signaling	2.926×10^{-2}
5	GO:BP	GO:0001796	regulation of type IIa hypersensitivity	2.977×10^{-2}
6	GO:BP	GO:0010646	regulation of cell communication	3.045×10^{-2}
7	GO:CC	GO:0016324	apical plasma membrane	1.034×10^{-2}
8	GO:CC	GO:0005576	extracellular region	2.640×10^{-2}

version	e108_eg55_p17_d098162
date	25/2/2023, 5:55:56 μ.μ.
organism	hsapiens

g:Profiler

Conclusion

The five highest differentially expressed genes were LGI1, MIR210HG, SLC37A4, SERPINA3 and MIR492. The gene LGI1 is involved in the development and morphogenesis of neuronal projections, MIR210HG plays a role in the cellular response to hypoxia and angiogenesis, SLC37A4 is involved in glucose and carbohydrate transport, SERPINA3 is involved in the acute-phase and inflammatory response, and MIR492 is involved in cell proliferation and transcriptional regulation.

The seven significantly GO terms that were identified in the GO terms enrichment analysis were homeostatic process, regulation of biological quality, response to stimulus, regulation of signaling, regulation of type IIa hypersensitivity, regulation of cell communication, apical plasma membrane and extracellular region. The significant GO terms suggest that the biological processes that the significantly expressed genes are involved with are regulated and maintained by homeostatic processes, response to stimuli, regulation of signaling and communication, and the extracellular environment.

Based on these results of the transcriptome analysis of lncRNAs in human pancreatic islets and β-cells, it can be infered that these cells play important roles in regulating glucose and carbohydrate transport, acute-phase and inflammatory response, and cellular response to hypoxia and angiogenesis. The significant GO terms identified provide additional insights into the potential regulatory pathways involved in the maintenance and function of these cells, which are consistent with their known functions.