Norwegian University
of Life Sciences

Master's Thesis 2019 60 ECTS

Department of Chemistry, Biotechnology and Food Science (KBM)

The effect of chromatin structure on duplicate gene expression in Atlantic salmon

Cathrine Horntvedt Kristiansen

Bioinformatics and applied statistics

Acknowledgements

The analysis was performed at the CIGEN (Center for Integrative Genetics)/NMBU Orion cluster environment. The thesis work was started autumn 2018 and continued into spring 2019, with help from my main supervisor Professor Torgeir R. Hvidsten and the co-supervisor Associate professor Simen R. Sandve. In addition, Torfinn Nome, Gareth Gillard and Lars Grønvold have also been helpful with navigating and providing the data. Thank you all so much for providing insight into the complex genome and regulatory environment of the Atlantic salmon and for the help to analyse parts of it.

Thank you to the Norwegian University of Life Science for having interesting courses that have supplied me with lots of knowledge throughout the years that I have been here.

At last I want to thank Andreas for being a great support from home.

Cathrine Horntvedt Kristiansen
May 14. 2019

Sammendrag

Atlanterhavslaksen har gjennomgått flere helgenomduplikasjoner og står igjen med nesten halvparten av sine gener som duplikater. Forskjellen i genuttrykk hos duplikatene er ekstra spennende med tanke på epigenetikk. Dersom det regulatoriske miljøet ble arvet sammen med genet etter helgenomdupliseringen og duplikatene har lignende genuttrykk, så kan kromatinstruktur i nærområdet til genet gi innsikt i genreguleringen. ATAC-seq data har blitt brukt til å bestemme hvilke regioner som har åpen kromatinstruktur som kan legge til rette for transkripsjon. Påvirker kromatinstrukturen som ble detektert av ATAC-seq data genuttrykket i Atlanterhavslaksen? Det var et forhold som ble funnet mellom de åpne kromatinområdene i nærområdet til genene og deres genuttrykk. En sammenheng ble funnet mellom prosentandel av promotoren som var dekket i peaks og en økning i genuttrykk, genuttrykket sank samtidig som en høy andel av regionen var dekket av peaks.

En EVE-analyse av lakseduplikatene ble gjort for å finne ut av opp- og nedregulerte duplikater i forhold til en utgruppe som ikke har vært igjennom den laksespesifikke helgenomdupliseringen. Er kromatinstrukturen rundt genduplikatene lignende for duplikatene som har lignende genuttrykk? Når man ser nærmere på forskjellen i ATAC-seq peaks og forskjellen i genuttrykk for duplikatene, så har noen duplikater en høy likhet i både genuttrykk og antall peaks, men noen duplikater har også en stor forskjell i peaks og genuttrykk. For de dupliserte genene virker det ikke som om et større antall peaks fører til økt genuttrykk.

Abstract

The Atlantic salmon has gone through several WGD events and is left with almost half of its gene as duplicates. How expression differs in the duplicates is extra intriguing in regards to epigenetics. If the regulatory environment was inherited together with the gene after the WGD and the duplicates are similarly expressed can chromatin structure surrounding the gene give some insight into the gene regulation. ATAC-seq data has been used to determine open chromatin regions that might facilitate transcription. Does the open chromatin structure detected by ATAC-seq data affect the expression of genes in the Atlantic Salmon? There was a relationship found between the open chromatin regions surrounding the genes and their expression. A connection was found between the percentage of the promoter covered in peaks and an increase in expression, the expression diminished with a high coverage level in the region.

An EVE analysis of the salmons duplicates has been done to determine up- and downregulated duplicates in regards to an outgroup that has not gone through the salmonid specific WGD .Is the chromatin structure surrounding the gene duplicates similar for the duplicates with a similar expression? Looking into the difference in ATAC-seq peaks and difference in expression for the duplicates there are duplicates that have a high similarity in both expression and number of peaks, but some duplicates also have a high difference in peaks and expression. For the duplicated genes a higher number of peaks does not seem to give a higher level of expression.

Contents

Introduction

Chromatin structure and transcription

More than 90% of regions in the deoxyribonucleic acid(DNA) that are bound by transcription factors (TFs) are found in the accessible genome which is only approximately 2-3% of the DNA in the human genome.[1] Genes in regions that are bound by TFs and have been marked by open chromatin are more likely to be expressed. [2] The Assay for Transposon Accessible Chromatin sequencing (ATAC-seq) data can be used to predict the areas in the genome that have open chromatin and is not occupied by nucleosomes. [3] Promoter and enhancer activity is dependent on chromatin accessibility, but there can still be open chromatin even for genes that are not transcribed and have inactive promoters and enhancers. [4,5] In the human genome, no simple correlation was found between open chromatin structure and gene expression, which was unexpected. [2] Here we used ATAC-seq and expression data from four different individuals of Atlantic Salmon (Salmo salar) to gain additional insight into the chromatin and gene expression relationship after whole genome duplication (WGD).

ATAC-seq

ATAC-seq is a sequencing method that can be used to identify open regions of DNA. By identifying available DNA regions we get information about which parts of the DNA that is available for transcription. Analysing ATAC-seq data can be helpful in finding important information about nucleosome packing and positioning, patterns of nucleosome-TF spacing, and TF co-occupancy at genome wide resolution. [3] Tagmentation uses sequencing adaptors to do fragmentation and tagging of a genome at the same time. [6] The main step of ATAC-seq is that Tn5 transposase in a mutated hyperactive form extracts DNA from open chromatin that is long enough. Tn5 is preloaded with DNA adapters and tags genomic DNA that will be extracted and fragmented downstream.(illustrated in figure 1) The tagged fragments are then processed further and sequenced before analysis of the reads can begin. The reads are mapped to the genome which are analyzed to identify probable peaks that indicates positions where the chromatin is most likely open. [3]

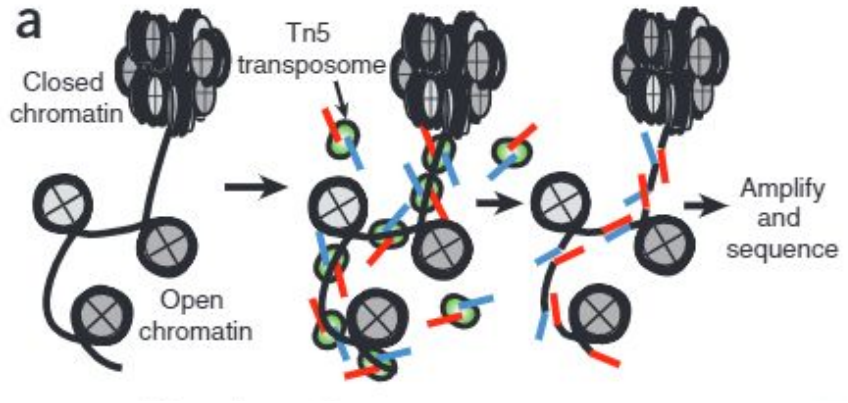


Figure 1. Illustration of ATAC-seq reaction from figure 1 a in Buenrostro et al. 2013 [3]

RNA-seq

The gene expression data used was derived from Ribonucleic acid sequencing (RNA-seq) data. RNA-seq is an approach for transcriptome profiling using deep sequencing technologies. RNA-seq provides measurements that are more precise, both with respect to the levels of transcripts and their isoforms, than other methods [7] The number of reads mapped to a gene is used to determine how expressed that gene is. More reads usually stems from more mRNA copies being present in the cell at the time of sampling. A long gene can end up having equally many reads as a shorter gene, but in the long gene the reads can be more spread out over the length of the gene and so this does not necessarily indicate a higher expression for the long gene. Normalization for gene length and number of reads from a sample is an important step in the processing of the RNA count data when determining the expression of a gene. Transcripts per million (TPM) is one normalizing method that takes into account the length of the gene and the total number of reads produced from the sample. It can suffer some bias but normalizes the data within sample and is comparable between samples. [8]

About the project

This study of the epigenetics in the Atlantic Salmon investigates the relationship between gene expression and chromatin structure using RNA-seq and ATAC-seq data as well gene data from the assembly of the Atlantic salmon. Chromatin accessibility is determined by chromatin binding factors and the organization of nucleosomes and refers the degree that molecules in the cells nucleus can bind to that part of the DNA. [9] Does the chromatin accessibility affect the expression of

genes in the Atlantic Salmon? Is there a causation between the open chromatin in near proximity to a gene and the genes expression in the Atlantic Salmon genome?

The Atlantic Salmon has been through several whole genome duplication events which leves it with several gene duplicates scattered across its genome. [10] Do the gene duplicates have the same expression levels as well as similar chromatin structure surrounding them?

The Atlantic salmon is especially interesting because of the salmonid specific fourth vertebrate WGD event that happened ~80 million years ago (Mya). [10] Today around half of all Atlantic salmon genes remain in duplicates. To gain insight into evolution after WGD it is interesting to know more about how these duplicates have evolved after the WGD. In particular, we want to know if they have the same level of gene expression and, if not, whether differences in chromatin structure plays a part in differentiating the expression of the duplicated pairs. Rediploidization has continuously been taking place after the WGD, but the Atlantic Salmon is still considered a pseudo-tetraploid as it can have quadruple sets of chromosomes. [11]

In this project we want to explore gene expression of gene duplicates and the relationship between their expression and the chromatin structure surrounding each duplicate pair. The way the duplicates evolve during rediploidization and how the chromatin structure and gene expression in each of the duplicates has transformed in relation to each other can give some insight into chromatin assisted gene regulation. Finding out if differences in expression between gene duplicates in part can be explained by difference in chromatin structure will aid in the understanding of how duplicated genes can evolve after a WGD.

Materials and methods

The samples

The sequenced samples were originally derived from Atlantic salmon liver samples from four different fish. The RNA-seq [12] and ATAC-seq [13] data is from the same four fish. Gene duplicates with a shift in gene expression where obtained from another project using the EVE method. [14] The gene information is taken from the reference sequence of the Atlantic salmon, assembly ICSASG_v2. The GFF assembly was downloaded and a subset only containing the rows where the type

was “gene” was used. The chromosome name had to be translated from RefSeq to Name in the gene data. Only the genes that were placed on the chromosomes or mitochondria were used.

Computing resources

The Orion Computer Cluster was used for the data analysis. The bioinformatic tools used in the analysis were all open source and available on the Orion cluster at CIGENE-NMBU (Center of Integrative Genetics, Norwegian University of Life Science). R studio Anaconda3 was used for scripting, using the web extension [15] to run it on the computer cluster. Some of the R packages used were data.table for handling large data files and ggplot2 for visualization and plotting. To run Bedtools through R the RLinuxModules[16] package was used. An R-markdown file was made and added in the appendix, showing the most important scripts used to generate the results.

Peak calling

The former official ATAC-seq pipeline of ENCODE was used. [17] The input to this pipeline was four replicates of paired end reads of ATAC-seq data in the form of raw FASTQ files from the Atlantic Salmon. Irreproducible Discovery Rate (IDR) was turned on, 10 threads were used, 8 fastq files and a folder for the output was chosen, otherwise the standards parameter settings for the pipeline were used. The pipeline first trims adapters, then aligns the reads, then filters reads , then removes duplicates and calls peaks using MACS2 before doing an IDR analysis for all pairs of replicates. [18] Peak calling being the most interesting step of the pipeline. It was used to generate peaks associated with open chromatin structure in the Atlantic Salmon. MACS2 automatically detects the read length, filters duplicated reads and calculates the maximum number of duplicated reads in a single position. The output from MACS2 is in the narrow peak file format. [17]

The pipeline generates a main html report containing all the results, including three sets of narrow peak files containing chromosome, start and stop position of the peaks, different score values, etc. There is one naive data set, one optimal and one conservative that all have had different threshold values for weather or not a peak is significant enough to be considered. These data sets are all compiled from the four replicates and contains peaks from all the different replicate fish. The “optimal” data set was used in the following analysis. The html report with the rest of the results is on the Orion cluster. [13]

Bedtools

Bedtools is a toolkit for the exploration of high-throughput genomics datasets. [19] The functions *closest* and *coverage* are the tools from bedtools that were used. First Bedtools

closest was used to find the distance between gene start and the closest peak. Bedtools coverage helped in determining the upstream peak distribution for each gene in the Atlantic salmon genome. To use Bedtools, a .bed file with three columns containing chromosome information, start position(lower than stop position) and stop position for a feature is the minimum requirement to perform any analysis. Additional columns of information can be added and used in the analysis with different settings. One can choose to do the analysis against one or more .bed files. The input .bed files have to be sorted by chromosome and by start position or Bedtools own sorting function can be used. [19]

Bedtools closest compares positions on file a and b, with or without overlaps between the features. Closest was used to look into the distances between the genes and the closest peak in the Atlantic Salmon. File A contains the Atlantic Salmon genes and their chromosome, start and stop position and file B contains the same info for the peaks. The output then is each gene peak relationship with an additional column containing the distance in base pairs between the start positions of the features. This was done for each of the three data sets, the naive, optimal and conservative. [20] With bedtools closest, the options used were -D and ref. -D gives an extra output column with the distance in base pairs between the start position of the gene and peak, overlapping features gets set to 0 and the distance for upstream features is negative. The ref option reports the distance with respect to the reference genome.

Bedtools coverage was used to identify what the peak distribution near a gene looks like on a more detailed level, including looking into which and how many base pairs (bp) near the genes were covered by peaks. A .bed file was made for each of the different distances (in relation to gene start) that were analysed for their peak distribution. Area 1 was defined from 100 bp downstream of gene start to 1000 bp upstream of gene start, area 2 was 1001 bp upstream to 2000 bp upstream of gene start, area 3 was 2001 bp upstream to 3000 bp upstream of gene start, area 4 was 3001 bp upstream to 4000 bp upstream of gene start, area 5 was 4001 bp upstream to 5000 bp upstream of gene start, area 6 was 5001 bp upstream to 6000 bp upstream of gene start, area 7 was 6001 bp upstream to 7000 bp upstream of gene start, area 8 was 7001 bp upstream to 8000 bp upstream of gene start, area 9 was 8001 bp upstream to 9000 bp upstream of gene start and area 10 was 9001 bp upstream to 10 000 bp upstream of gene start. Some positions for the areas ended up being negative so instead the position was changed to 0, since a negative position on the chromosome is not defined and Bedtools does not accept that as an input. The Bedtools coverage analysis was done for each area as the A file and the peaks from the optimal dataset as the B file. No other options were chosen other than the two input files and a file path for the output. closest then adds four columns to the data.

1. The number of peaks that overlapped the area surrounding the genes.
2. The length in bps of the area that had at least one peak overlapping that position.
3. The length in bp of the area surrounding the gene.
4. column 2 divided by column 3, which gives the fraction of the area that was covered by peaks.

The column 1 with number of peaks and 4 with coverage of peaks were used in the following analysis. [21]

Expression

The count data for the four fish were TPM normalized and log2 adjusted to be normalized. The mean expression for the different genes were calculated based on the expression for the fish that also has ATAC-seq data.

Jbrowse

Jbrowse was used for visualizing ATAC-reads on a genome. The filtered and deduped bam files were used from each replicate. salmonbase.org was used for the visualization. On the website Species->Atlantic Salmon->JBrowse takes you to a genome browser. Then the reference sequence for Atlantic Salmon with its genes were chosen to be displayed and each file for the replicates were added as a new track with the file URL and .bai at the end for indexing and then the URL again. (URL example for one track for one replicate: https://orion.nmbu.no/users/torfn/FAASG/2017-07-Pilot-ATAC/align/rep1/2-ATAC-S3-50-3_S2_R1_001.trim.PE2SE.nodup.bam.bai https://orion.nmbu.no/users/torfn/FAASG/2017-07-Pilot-ATAC/align/rep1/2-ATAC-S3-50-3_S2_R1_001.trim.PE2SE.nodup.bam)

EVE duplicate data

The expression variance and evolution model (EVE) was used to identify duplicates with expression divergence. [14] The EVE analysis was based on liver expression data replicated for several teleost and salmonid species. [22] An EVE results table containing the duplicate analysis (Table 1) and a table with clan information containing the geneIDs of the duplicated genes were used to define upregulated and downregulated duplicates. For both the up and down regulated duplicates, only the data where the likelihood ratio test (LRT) was larger than 4 were analyzed. Furthermore, test.type was Ss4R, data.type was BSNsgl,

gene.type was either dupA or dupB and clan.maxLRT was TRUE. For the duplicates with a shift in expression a subset was created like previously described, but with shift.direction as up for upregulated duplicates and down for the downregulated duplicates. LRT above 4 indicate that a shift in the optimum expression level most likely has taken place. test.type equal to Ss4R is data collected from Atlantic salmon. data.type as BSNsgl is a type of normalization, gene.type indicates which duplicate had a shift in expression. clan.maxLRT is TRUE for single clans. [22]

Table 1. The EVE results table. The different columns meaning, clan is the name of the clan tested, data.type is the normalization method used(BSNsgl was used for between species normalization, using single orthologs for only for factors), gene.type is single for clans with 1:1 orthologs, dupA for clan with first duplicate of 1:2 orthologs, dupB clan with second duplicate og 1:2 orthologs, the test.type is Ss4R for branch specific shift on the salmonid branch, LRT is log 2 likelihood ratio test score, clan.maxLRT is tthetahe LRT the highest for that test for that clan, always true for single clans, shows for duplicate clans if e.g. dupA LRT > dupB LRT, theta is non-shift branch expression level, thetaShif is shifting branch expression level, shift.direction is up or down dependent on if thetaShift is higher or lower than theta and alpha, beta and sigma.sq are likelihood model parameters.

	clan	data.type	gene.type	test.type	LRT	clan.maxLRT
1	OG0000026_6	BSNcmb	single	Var	9.543547e-01	TRUE
2	OG0000026_6	BSNcmb	single	Eluc	1.613166e-04	TRUE
3	OG0000026_6	BSNcmb	single	Ss4R	4.256237e+00	TRUE

theta	thetaShift	shift.direction	alpha	beta	sigma.sq
NA	*NA*	*NA*	*NA*	*NA*	*NA*
1.3779345	1.332720438	down	5.827372e+00	9.523563e-01	2.131141e+01
0.7660724	3.111410213	up	9.425142e+04	4.794749e+00	7.089735e+04

Analysis

A venn diagram was made of all the overlapping peaks between the different dataset. A peak was defined as overlapping between two datasets if the values of the peak start position added to the peak point source were the same.

The Bedtools closest data was used to visualize the distribution of peaks in comparison to the gene start for all the genes. A density plot was created using the density() function in R incorporating all the datasets to look into their distribution of peaks. The density plot was used to help decide which data set to continue with, resulting in us choosing the optimal one. A histogram for the optimal plot was also created after removing all results having -1 as the start position for the peak as there were no peaks found on the chromosome/scaffold/contig the gene was on.

All the areas from the Bedtools coverage data was merged together and then merged together with the expression data into a large data.frame. Columns with total coverage and total no. peaks were also created by adding up the values for areas 1 through 10. A correlation plot was made between some of the different columns to investigate any correlation further. An boxplot was made with the number of peaks from 100 bp downstream to 10 000 pb upstream of gene start and mean expression for the four samples using ggplot2. Another boxplot containing the coverage in the promoter region (100 bp downstream to 10 000 pb upstream) and mean expression of the four fish was made, here varwidth was set to TRUE to show the amount of genes each box was based on in comparison to each other. The cor.test() function was applied on different sets of data to see if there was any significant correlation between expression and peak distribution surrounding a gene.

The duplicate data was also merged with peak and expression data from the four samples and incorporated into the large data.frame, one table for the upregulated genes and one for the downregulated genes. Gene duplicates that are up- or

downregulated in salmon compared to their expression level in outgroup fish species without the salmonid specific WGD. Using ggplot2, a boxplot was made containing both the downregulated duplicate genes and for the upregulated duplicates, with the number of peaks from 100 bp downstream to 10 000 pb upstream of gene start and mean expression for the four samples. The difference in number of peaks and difference in expression was plotted against each other to look for correlation, as well as significance using the cor.test().

Some translation of names had to be done between CIGEN geneIDs and NCBI geneIDs to be able to compile the information, and for this the R-package Ssa.RefSeq.db [23] was used. Functions were made in Rstudio to avoid having to copy and paste code for things that had to be done several times with different datasets.

Writing

Written using google documents (docs.google.com) and the F1000 addon to help managing references.

Results

The datasets

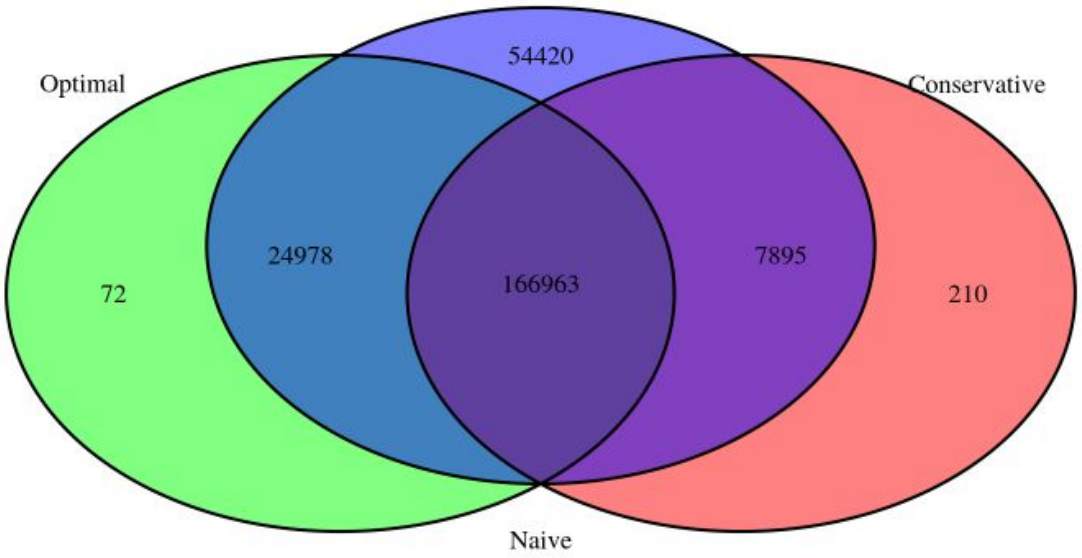


Figure 2. Number of peaks in each data set and overlapping peaks between the different data sets. Based on peak start position and number to identify "the same" peak.

The naive dataset had 254256 peaks, the optimal dataset had 192013 peaks and the conservative dataset had 175068 peaks. As seen in figure 2, all the datasets had some unique peaks, the naive data set had the most with 54420 peaks while the conservative and optimal sets had a lot less. All the data sets had 166 963 peaks in common. The naive and optimal set had more peaks in common than the conservative and naive data sets.

Bedtools

Table 2. Bedtools closest result, gene data and closest peak data (from optimal dataset) with the distance in bp between gene start and the peak. A distance equal to zero means that the features are overlapping.

Chromosome	Start	Stop	Strand	Chromosome	PeakStart	PeakStop	Distance
ssa02	4195587	4223098	-	ssa02	4193434	4194144	-1444
ssa02	4240645	4250882	-	ssa02	4247539	4247797	0
ssa02	4315639	4318404	-	ssa02	4308541	4308804	-6836
ssa02	4322523	4325190	-	ssa02	4331102	4331669	5913
ssa02	4329133	4330212	-	ssa02	4331102	4331669	891

Bedtools closest reveals that the closest peak to a gene's start position can be both downstream (positive distance) and upstream (negative distance) as seen in table 2.

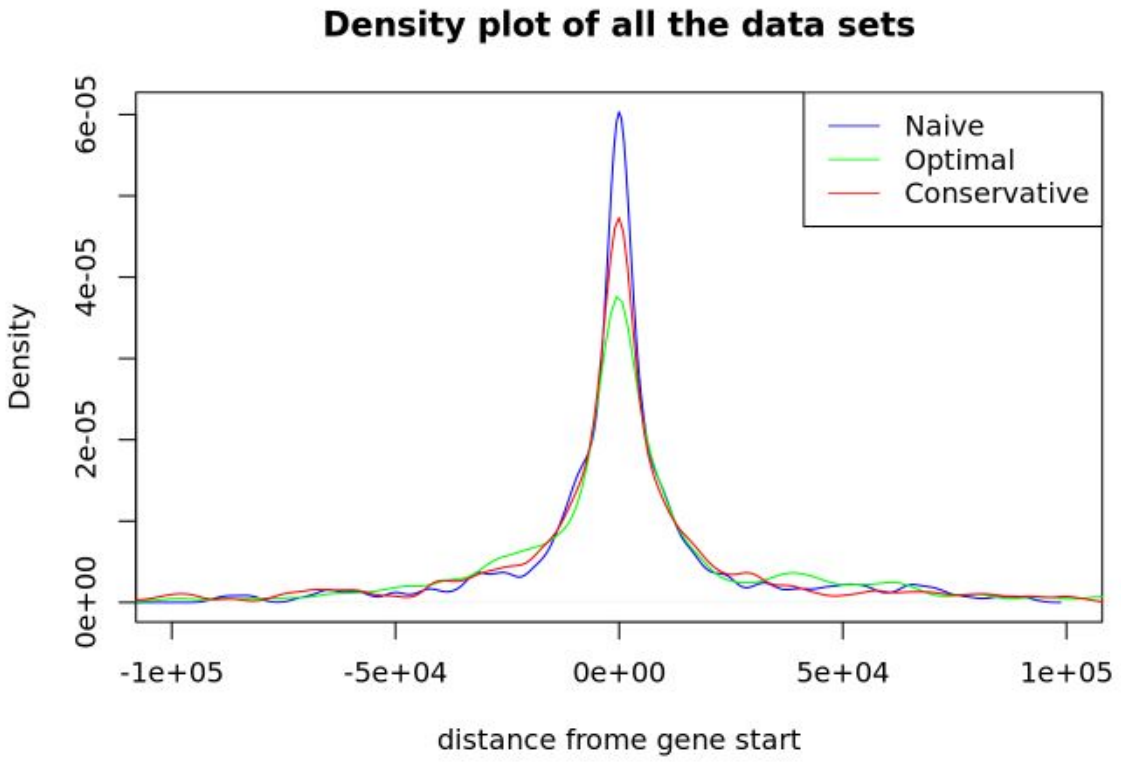


Figure 3. Density plot of all the data sets. Peak density is on the y-axis and the distance in base pairs from gene start on the x-axis, where negative values represent upstream positions and positive values represent downstream positions. The naive data set is blue, optimal is green and conservative is red.

The optimal set has the least peaks at gene start, the conservative has more and the naive set has the most peaks at gene start, as seen in figure 3. It doesn't look like a huge difference in peak distribution compared to gene start for the three data sets. For all the three sets the decline in peaks as you move further away from gene start is more rapid downstream of gene start than it is upstream.

Table 3. The output table after Bedtools coverage looks like this after applying column names and only keeping the number of peaks and coverage results. Chromosome, geneIDs, start and stop position for the area, which strand, the number of peaks and the coverage of peaks in that area. This type of table was made for all the chosen areas surrounding the genes.

	Chromosome	GeneID	ID	Start	Stop	Strand	No.peaks	Coverage
1	NC_001960.1	808311	gene81573	3846	4820	+	14	1.0000000
2	NC_001960.1	808316	gene81574	5036	6085	+	17	0.9427273
3	NC_001960.1	808314	gene81575	6476	8026	+	2	0.9327273

In table 3 representing the bedtools coverage data it is seen that the number of peaks can be lower for one gene that has full coverage than a gene that doesn't because the length of the peaks differ. For instance upstream of gene81573 there are 14 peaks and a coverage of 100%, which means that the entire feature was covered by peaks. gene81574 has 17 peaks that covers ~94.3% of the feature, so even though it has more peaks the coverage for this gene is lower.

Table 4. Merged table with all the areas for each gene. A1 is area one which is a 1000 bp upstream of gene start and 100 bp downstream, A2 is area two which is 1001 bp upstream to 2000 bp upstream etc. all the way up to A10 which is 10 000 bp upstream to 9001 bp upstream. Both coverage information and peak count is present for each of the areas.

	Chromosome	GeneID	ID	Start	Stop	Strand	A1Peaks	A1Cov	A2Peaks
1	NC_001960.1	808311	gene81573	3846	4820	+	14	1.0000000	14
2	NC_001960.1	808316	gene81574	5036	6085	+	17	0.9427273	14
3	NC_001960.1	808314	gene81575	6476	8026	+	2	0.9327273	17

	A2Cov	A3Peaks	A3Cov	A4Peaks	A4Cov	A5Peaks	A5Cov	A6Peaks
1	1.0000000	14	1	14	0.9976332	0	0	0
2	1.0000000	14	1	14	1.0000000	14	1	14
3	0.8968969	14	1	14	1.0000000	14	1	14

	A6Cov	A7Peaks	A7Cov	A8Peaks	A8Cov	A9Peaks	A9Cov	A10Peaks	A10Cov
1	0.0000000	0	0.0000000	0	0	0	0	0	0
2	0.9428571	0	0.0000000	0	0	0	0	0	0
3	1.0000000	14	0.9957895	0	0	0	0	0	0

As seen in table 4 a gene can have high coverage in some of the areas and no coverage in others. Typically ,when the coverage has dropped to zero in an area it stays at zero further upstream as well. The number of peaks for a gene can increase in an area further upstream of gene start without elevating the degree of coverage. Out of 50644 genes with zero expression 42234 genes also has zero ATAC-seq peaks.

Table 5. TPM normalized and log2 adjusted expression for each gene for the four replicates. Containing gene IDs for all the genes as well as a mean expression for the four fish.

GeneID	ID	Fish1	Fish2	Fish3	Fish4	meansof4
100135779	gene50914	0.000000000	0.000000000	0.04762517	0.032788977	0.020103536
100136349	gene44219	4.227226597	4.487942466	4.58723942	4.177734264	4.370035686
100136351	gene45104	0.785208961	1.582490951	1.80968407	0.433106550	1.152622633
100136352	gene8441	9.703669980	10.198026455	10.35192395	10.115163035	10.092195854
100136353	gene40008	4.689775626	4.337988219	4.21174904	4.591575272	4.457772039

42 229 genes out off 79 030 had zero in mean expression in the liver cells for these four fish. In table 5 the results show that the expression can vary quite a bit between the fishes, gene35810 has an expression of ~0.296 for fish 1 while the mean ends up being ~0.872, the mean expression ends up being quite close to the expression of some of the fish, but not all.

Table 6. A table only containing, geneIDs, gene expression and peak information for all genes. Each row represents a gene and each column represents information about that gene.

```
      ID    GeneID Chromosome  Start   Stop Strand A1Peaks     A1Cov
1  gene0 106560212      ssa01   5501  62139      -       1 0.5345455
2  gene1 106607996      ssa01 160437 198815      -       0 0.0000000
3 gene10 106599499      ssa01 516060 519262      +       0 0.0000000
  A2Peaks A2Cov A3Peaks A3Cov A4Peaks A4Cov A5Peaks A5Cov A6Peaks A6Cov
1       0     0       0     0       0     0       0     0       0     0
2       0     0       0     0       0     0       0     0       0     0
3       0     0       0     0       0     0       0     0       0     0
  A7Peaks A7Cov A8Peaks A8Cov A9Peaks A9Cov A10Peaks A10Cov      Fish1
1       0     0       0     0       0     0        0      0 0.00000000
2       0     0       0     0       0     0        0      0 0.47508136
3       0     0       0     0       0     0        0      0 0.04840276
      Fish2     Fish3      Fish4  meansof4 totpeaks     totcov
1 0.0000000 0.0000000 0.00000000 0.0000000        1 0.05345455
2 0.1151858 0.0487502 0.73042042 0.3423595        0 0.00000000
3 0.2621157 0.1142145 0.07919524 0.1259821        0 0.00000000
```

Box plots

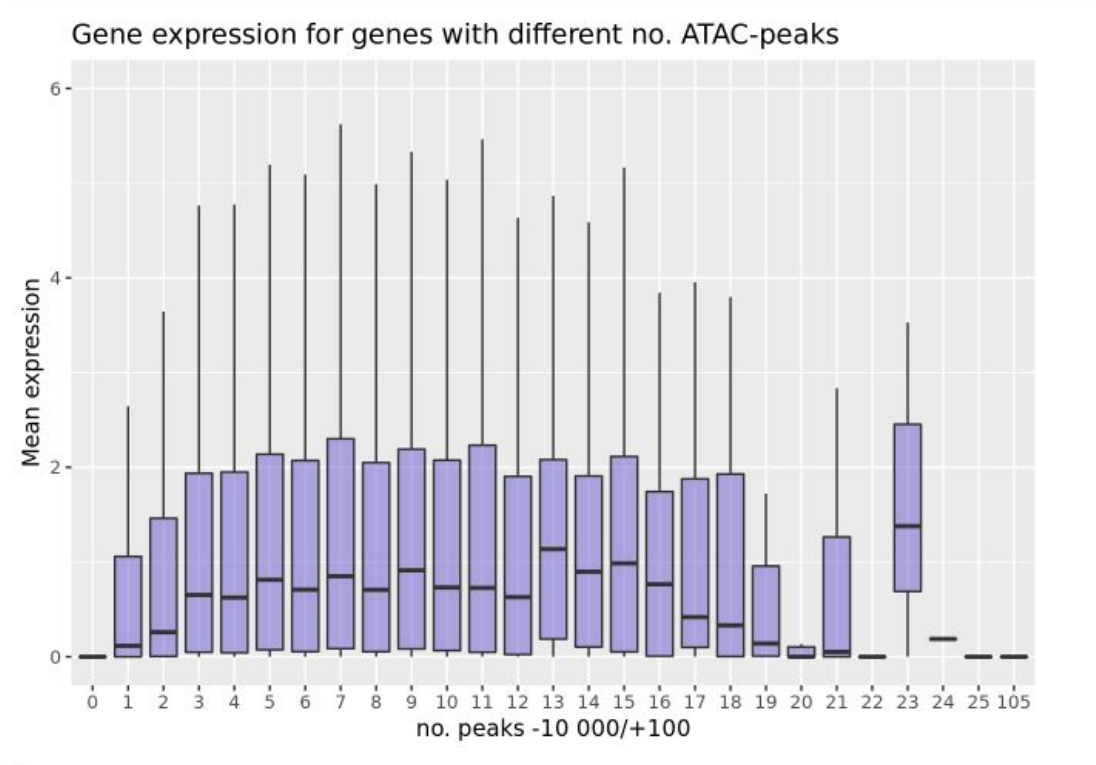


Figure 4. Mean expression for genes with different number of ATAC-peaks in the region from 100 bp downstream to 10 000 bp upstream of gene start.

For 0 to 9 peaks there seems to be a trend that the more peaks a gene has in the region from 100 bp downstream to 10 000 bp upstream of gene start the more it is expressed (Figure 4). For more peaks the trend is much more variable but here the data rely on increasingly fewer genes. The highest median expression level is for 23 peaks, the highest upper quartile is at 14 peaks.

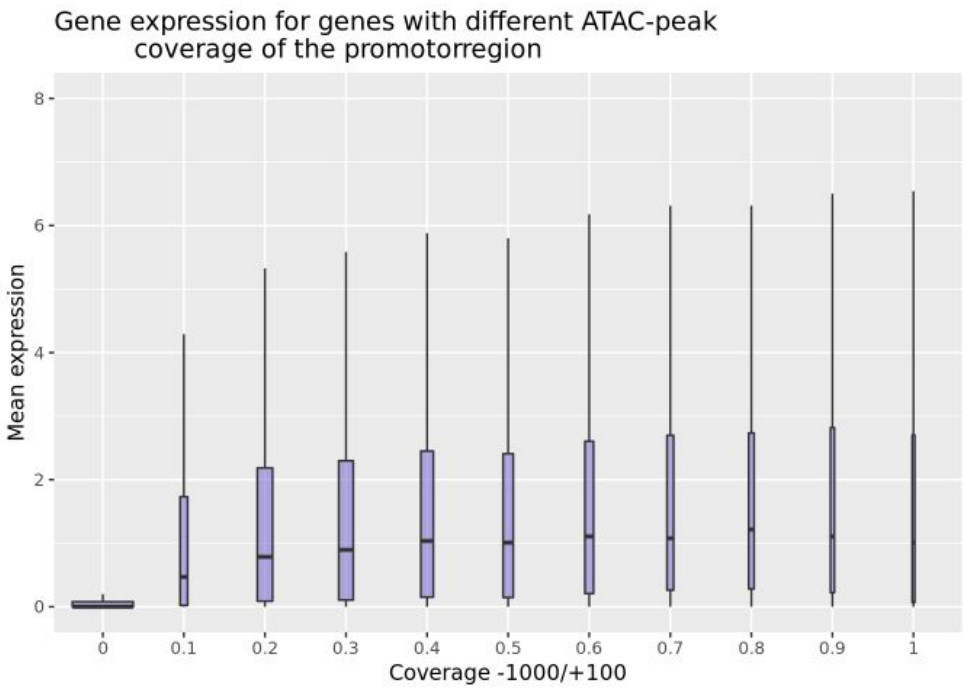


Figure 5. Mean expression for genes with different ATAC-peak coverage of the promoter region(100 bp downstream to 1000 bp upstream). The coverage is the number of bp covered by a peak in the designated area divided by the total number of bps for that area.

There is a general trend that the expression increases as the peak coverage goes up, but only from 0 coverage to 0.4 coverage (Figure 5). After that it varies a bit and especially at 0.9 coverage the mean expression is lower. The amount of genes decreases while the coverage increases. The lowest number of genes is when the coverage is 1, which means the entire region is overlapped by peaks. Some of the genes with higher coverage levels even have lower mean expression than those with lower coverage.

Duplicates

Table 7. After getting necessary information from the clan and adding it to the EVE subset this was the result. A similarly structured table for downregulated duplicates was also created. The columns meaning is the same as in table 1, but with some additional columns. Ssal.dupA is the CIGEN gene name of duplicate A and similar for dupB, geneID is the NCBI geneID for the shifted gene duplicate(dupA if gene.type is dupA) and product is a short description of the function of the genes.

clan	data.type	gene.type	test.type	LRT	clan.maxLRT	theta	thetaShift	shift.direction
OG0000244_1	BSNsgl	dupB	Ss4R	4.713791	TRUE	0.016125015	28.7330362	up
OG0000370_1	BSNsgl	dupA	Ss4R	5.207431	TRUE	0.087395936	4.2845637	up
OG0000426_1	BSNsgl	dupB	Ss4R	13.323130	TRUE	1.069494899	2.9206009	up
OG0000431_1	BSNsgl	dupB	Ss4R	5.561906	TRUE	0.722423165	1.8712880	up
OG0000442_2	BSNsgl	dupA	Ss4R	4.276883	TRUE	0.005213866	1.7097612	up

alpha	beta	sigma.sq	Ssal.dupA	Ssal.dupB	geneID	product
6.978247e+00	8.372628e-01	8.708539e+01	XP_014065962	NP_001133178	gene460:100194621	adenylosuccinate synthase like 2
5.944222e+02	8.721134e-01	1.678621e+03	NP_001167347	XP_014060091	gene32328:100380591	Hippocalcin-like protein 1
1.833790e+02	1.059273e+01	6.531306e+00	XP_013994844	XP_014045294	gene6086:106598815	F-box-like/WD repeat-containing protein TBL1XR1
1.123797e+07	1.318202e+00	4.050060e+06	XP_014032844	XP_013997257	gene31164:106569978	glycogen synthase kinase-3 beta-like
4.898320e+07	6.673389e-01	6.088461e+07	XP_014069939	XP_014018114	gene19834:106612869	arfaptin-2-like

The NCBI geneID for the upregulated gene is in table 7. In table 7 it is seen that one of the upregulated genes has a arfaptin-2-like product, and arfaptin 2 is suggested to be involved in regulating huntingtin protein aggregation in humans. [24]

Correlation

Table 8. The cor.test results for the different relationships between gene duplicates expression and ATAC-seq peak data. The region is where the peak data is from, 100 bp downstream of gene start to either 1000 bp upstream of gene start or 10 000 bp upstream of gene start. The significant correlations (p-value < 0.05) have a light blue background color.

Regulation	Expression data	Peak-data	Region	Correlation coefficient	P-value
up	difference in mean expression	difference in peaks	+100/-1000	0.024	0.801

up	difference in mean expression	difference in coverage	+100/-1000	0.092	0.324
up	difference in mean expression	difference in coverage	+100/-10000	0.093	0.321
up	mean expression	number of peaks	+100/-1000	0.166	0.073
up	mean expression	coverage	+100/-1000	0.253	0.006
up	mean expression	number of peaks	+100/-10000	0.176	0.058
up	mean expression	coverage	+100/-10000	0.182	0.049
down	difference in mean expression	difference in peaks	+100/-1000	-0.006	0.887
down	difference in mean expression	difference in coverage	+100/-1000	-0.032	0.424
down	difference in mean expression	difference in coverage	+100/-10000	0.040	0.323
down	mean expression	number of peaks	+100/-1000	0.062	0.120
down	mean expression	coverage	+100/-1000	0.021	0.595
down	mean expression	number of peaks	+100/-10000	0.099	0.013
down	mean expression	coverage	+100/-10000	0.082	0.040

As seen in table 8, the upregulated genes had a correlation coefficient below 0.1 for difference in coverage and peak number correlated with difference in

expression. The p-value for difference in peaks and expression was 0.8, the correlation was not significant. The p-value for the Pearson's correlation test between difference in coverage and difference in expression was ~0.32 for both the promoter region and for the area 100 bp downstream to 10000 bp upstream of gene start for the upregulated duplicates. (Table8) Out of all the correlation relationships that were tested only 2 of the upregulated duplicates and 2 of the downregulated duplicates had a significant p-value < 0.05. The highest correlation coefficient with a significant p-value was 0.253 which was for mean expression and coverage in the promoter region of upregulated duplicates. (Table 8)
The correlation between the expression and peak data for the duplicates were low. The correlation coefficient was under 0.3 for all the tests, and the p-values were quite varying as seen in table 8.

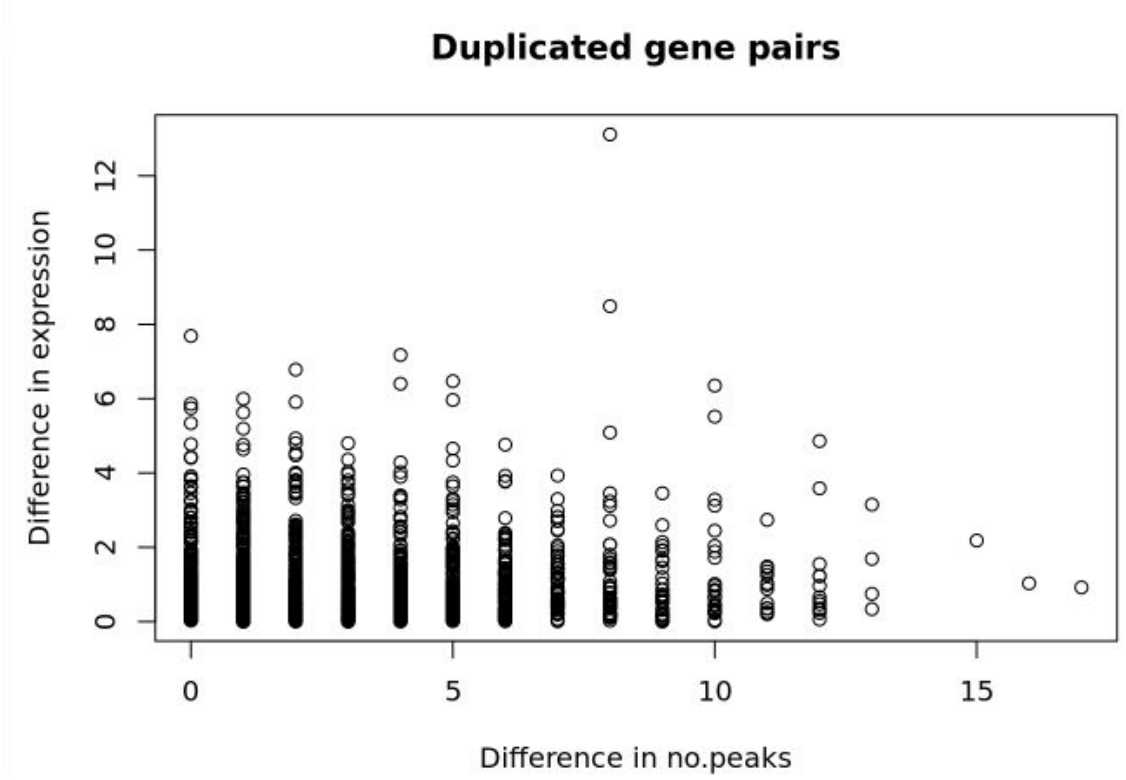


Figure 6. Plot for the absolute difference in expression from the mean of the four fish and absolute difference in peaks 10 000 bp upstream to 100 bp downstream of gene start for all the duplicated gene pairs.

In figure 6 there is no clear correlation between the difference in expression and difference in the number of peaks. When the difference in number of peaks increases the difference in expression seems to go down towards the end, but there is no clear pattern in the data. Some of the genes that have a high similarity in the number of peaks also have a high similarity in expression, but there is also some duplicates having the same number of peaks with a large difference in the expression level. A cor.test of these two variables shows no significant correlation with a p-value of 0.573.

Boxplot

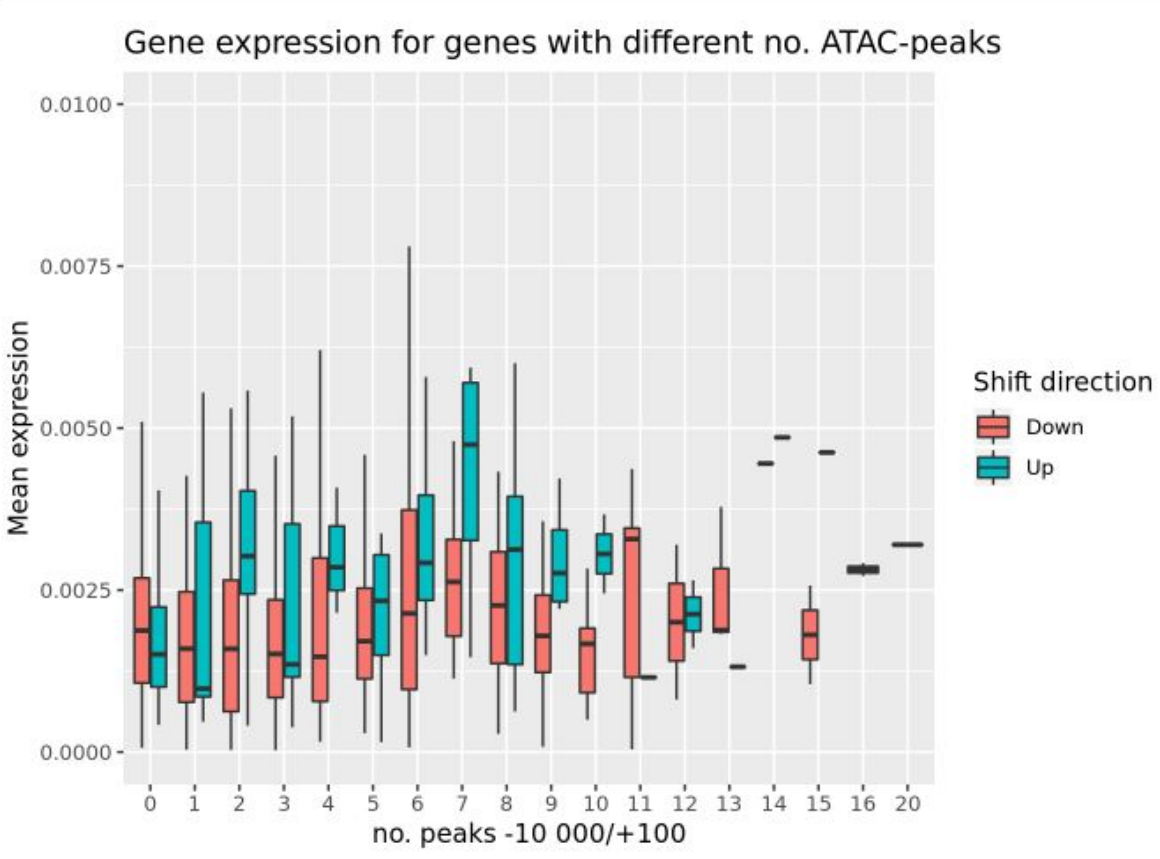


Figure 7. Mean expression for genes with different number of ATAC-seq peaks 100 bp downstream to 10 000 bp upstream, for both the upregulated duplicates and the downregulated duplicates.

Figure 7 shows that the upregulated gene duplicates mostly have a higher expression than the downregulated duplicates. Out of the 17 different levels of number of peaks, the up regulated ones were the most expressed in 12 of them. For zero peaks the mean expression was highest for the down regulated genes. For 16 and 20 peaks which are the two highest categories, the median for both the upregulated duplicates and the downregulated ones is the same. The downregulated duplicates have been shown in figure 7 to have a tendency to be both a little positively and negatively skewed. The up regulated duplicates has more positively skewed boxes than the downregulated duplicates. In figure 7 the higher number of peaks does not seem to give a higher level of expression for the duplicates.

Discussion

Were the results as expected?

A correlation between open chromatin upstream for a gene and its expression was not expected to be found genome wide. [2,3] In figure 6 containing all the gene duplicates there is no apparent pattern for the plot of the difference in expression against the difference in number of peaks and the correlation test showed no significant correlation. Hence for the duplicated genes the results were as expected, as a significant relationship between the chromatin structure surrounding the duplicates and the duplicates expression was not found.

We expected a low number of peaks upstream of most of the genes that had no or very low expression. [2] In this study this was mostly the case for all the genes with zero expression, where out of 50644 genes with zero expression 42234 genes also has zero ATAC-seq peaks.

A high number of peaks upstream of most of the genes that had high expression, was expected but this was not always the case. [2] Several genes with high expression had no ATAC-seq peaks in the areas that were checked.

The relationship between the chromatin structure and expression of a gene was more difficult to determine than expected. A relationship was expected to be found, but it seems as if only ATAC-seq and RNA-seq expression data was not enough to determine that relationship. [25] There could simply be to many other factors in play, that are regulating the gene expression in a way so that chromatin accessibility does not get a big part in determining the level of gene expression.

Possible problems and faults

The genes that had high expression, but no peaks can be due to N's in the reference genome making ATAC-seq reads not map even when the regions are upstream of a gene that is highly expressed.

Other regulatory elements (such as promoters and enhancers) play a role in the expression of a gene. [1] It is very possible that the chromatin structure is important for gene regulation, but only up to a certain point. The open chromatin in itself does not result in transcription. Other factors such as TF binding sites and TFs need to be present in the region. [4] This will hide the chromatin and gene expression relationship to some extent, and have not been accounted for in this study. Maybe the relationship between open chromatin and gene regulation is more important in regions with partly open chromatin structure upstream of a gene and if the region is either very open or very closed, other mechanisms may play an more important role in the gene expression levels. ATAC-seq data seems to be better at determining which regions that have an accessible chromatin structure and which don't, than to predict how expressed a gene is depended on the number of peaks in a upstream region of that gene.

When looking into weather or not open chromatin was correlated with expression, only the peaks from 100 bp downstream to 10000 bp upstream were taken into account. It might be better to also look into the chromatin structure further away from gene start both upstream and downstream. This might catch expression signals from different types of open chromatin. Open chromatin can have different states that have not been accounted for in this study, a region in the genome was classified as open or closed based on the ATAC-seq peak data. Open chromatin can be enhancers, promotor, in a transcribed state or poised state and still give a signal of open chromatin. [25]

The sample size is quite small, only four fish, even with the expression data and open chromatin data about the same individuals the power of the study is not very large. The four replicates might not be enough to say something generally about Atlantic salmon and its relationship between expression and open chromatin. Looking into individual data for the samples could give a more precise image of how open chromatin structure and gene expression is connected.

Conclusion

There was no clear significant correlation between the number of peaks upstream of a gene duplicate and its expression compared to the other duplicate. The only significant correlation between number of peaks and mean expression had a correlation coefficient of ~0.099, even though this relationship was significant it is still very weak. Looking into the difference in ATAC-seq peaks and difference in expression for the duplicates there are duplicates that have a high similarity in both expression and number of peaks, but also duplicates that have a high difference in peaks and expression. For the duplicated genes a higher number of peaks does not seem to give a higher level of expression. (Figure 6) ATAC-seq paired with expression data is not enough to find any meaningful relationship between the gene expression and accessible chromatin regions for the gene duplicates.

A connection was found between the percentage of the promoter covered in peaks and an increased in expression, this diminished with a very high coverage of the region together with the number of genes that had that high amount of coverage. (Figure 5) Some relationship is detectable between all the genes with expression and ATAC-seq data.

Further analysis

The up- and downregulated gene duplicates are up- or downregulated in salmon compared to their expression level in outgroup fish species without the salmonid specific WGD. [22] It would be interesting with a similar analysis between the different cell types in the Atlantic salmon to see how the expression differs in different cell tissue.

In further analysis it would be ideal to continue expanding the data types overlapping for the different genes. A start would be to overlap known TF-motifs and occupancy by different TFs. [4]

A clustering method could be used to find clusters of genes that are co-expressed and then investigate if the co-expressed genes have a similar open chromatin environment surrounding them. [26]

ATAC-seq protocol and peak calling algorithm could be done using newer methods. [27] [28]

A gene ontology analysis for both the up- and downregulated genes. Performing an enrichment analysis on the gene sets using its annotations can find gene ontology terms that are over represented. [29]

There is a lot that can be done to continue analysing how the chromatin structure and epigenetics affect the gene expression in Atlantic salmon, some there wasn't enough time for and other methods go beyond the scope of a master thesis.

Literature

1. Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489,** 75–82 (2012).
2. Gilbert, N. *et al.* Chromatin architecture of the human genome: gene-rich domains are enriched in open chromatin fibers. *Cell* **118,** 555–566 (2004).
3. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10,** 1213–1218 (2013).
4. Dogan, N. *et al.* Occupancy by key transcription factors is a more accurate predictor of enhancer activity than histone modifications or chromatin accessibility. *Epigenetics Chromatin* **8,** 16 (2015).
5. Corces, M. R. *et al.* Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat. Genet.* **48,** 1193–1203 (2016).
6. Adey, A. *et al.* Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome Biol.* **11,** R119 (2010).
7. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10,** 57–63 (2009).
8. Conesa, A. *et al.* A survey of best practices for RNA-seq data analysis. *Genome*

Biol. **17,** 13 (2016).

9. Klemm, S. L., Shipony, Z. & Greenleaf, W. J. Chromatin accessibility and the regulatory epigenome. *Nat. Rev. Genet.* **20,** 207–220 (2019).
10. Lien, S. *et al.* The Atlantic salmon genome provides insights into rediploidization. *Nature* **533,** 200–205 (2016).
11. Davidson, W. S. *et al.* Sequencing the genome of the Atlantic salmon (Salmo salar). *Genome Biol.* **11,** 403 (2010).
12. *RNA-seq count data*.
13. orion.nmbu.no/users/torfn/FAASG/2017-07-Pilot-ATAC/REPLICATES-NSC-2017-07_report.html. at <https://orion.nmbu.no/users/torfn/FAASG/2017-07-Pilot-ATAC/REPLICATES-NSC-2017-07_report.html>
14. Rohlfs, R. V. & Nielsen, R. Phylogenetic ANOVA: the expression variance and evolution model for quantitative trait evolution. *Syst. Biol.* **64,** 695–708 (2015).
15. RStudio - markdown. at <https://orion.nmbu.no/rstudio/anaconda3_latest/>
16. GitHub - larsgr/RLinuxModules: R package that makes linux environment modules available from R. at <https://github.com/larsgr/RLinuxModules>
17. GitHub - kundajelab/atac_dnase_pipelines: ATAC-seq and DNase-seq processing pipeline. at <https://github.com/kundajelab/atac_dnase_pipelines>
18. Frozen, S. ATAC-Seq pipeline v1 specifications.
19. Quinlan, A. R. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr. Protoc. Bioinformatics* **47,** 11.12.1-34 (2014).
20. closest — bedtools 2.28.0 documentation. at

<https://bedtools.readthedocs.io/en/latest/content/tools/closest.html?>

21. coverage — bedtools 2.28.0 documentation. at <https://bedtools.readthedocs.io/en/latest/content/tools/coverage.html>
22. Gillard, G. B. Evolution of gene expression following the whole genome duplication in salmonid fish. (2019).
23. CIGENE / R / Ssa.RefSeq.db · GitLab. at <https://gitlab.com/cigene/R/Ssa.RefSeq.db>
24. Peters, P. J. *et al.* Arfaptin 2 regulates the aggregation of mutant huntingtin protein. *Nat. Cell Biol.* **4,** 240–245 (2002).
25. Jiang, S. & Mortazavi, A. Integrating ChIP-seq with other functional genomics data. *Brief. Funct. Genomics* **17,** 104–115 (2018).
26. D'haeseleer, P., Liang, S. & Somogyi, R. Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics* **16,** 707–726 (2000).
27. Corces, M. R. *et al.* An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat. Methods* **14,** 959–962 (2017).
28. ATAC-seq Data Standards and Prototype Processing Pipeline – ENCODE. at <https://www.encodeproject.org/atac-seq/>
29. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **37,** 1–13 (2009).

Appendix

Appendix

Cathrine H. Kristiansen

5/13/2019

ATAC-seq peak data

Setup

Loading packages and files to use later.

```
library(data.table)
library(VennDiagram) # for plotting a venn diagram
```

```
## Loading required package: grid
```

```
## Loading required package: futile.logger
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:data.table':
##
##     between, first, last
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(purrr)
```

```
##
## Attaching package: 'purrr'
```

```
## The following object is masked from 'package:data.table':
##
##     transpose
```

```
library(ggplot2) # for plotting
library(Ssa.RefSeq.db) # for translation of geneIDs
```

```
## Loading required package: RSQLite
```

```
library(RLinuxModules)
```

```
moduleInit( modulesHome = "/local/genome/Modules/3.2.10")
```

```
module("load bedtools") # loads bedtools into the environment

# Data
atacNaive <- fread('/mnt/users/ckristia/R/Data/naive.narrowPeak')
atacOptimal<- fread('/mnt/users/ckristia/R/Data/optimal.narrowPeak')
atacConservative <-
  fread('/mnt/users/ckristia/R/Data/conservative.narrowPeak')
load("/mnt/users/ckristia/R/Data/AllGenes.Rdata") # subset of the "genes"
# from the GFF file, sorted and added GeneIDs
load("/mnt/users/ckristia/R/Data/coveragetable.Rdata") #finished coverage
table
load("/mnt/users/ckristia/R/markdown/Data/TPM4Fish.Rdata") #TPM normalized
and log2 adusted values for the 4 fishes
```

Overlapping peaks between datasets

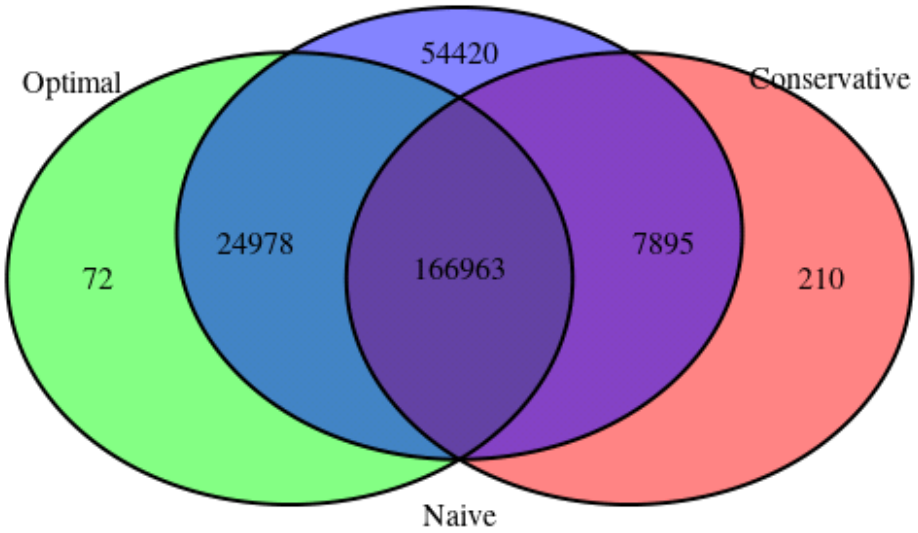
Venn diagram with the removal of duplicated peaks before comparing the datasets. Duplicated peaks defined as peaks with the same peak value which is determined by the peaks' start position (column 2 in narrowpeak format) added to the point source of the peak (column 10 in narrowpeak format)

```
## Number of peaks in each data set

# Naive:
no_peaksN <- length(atacNaive$V1)

# Optimal:
no_peaksO <- length(atacOptimal$V1)

# Conservative:
no_peaksC <- length(atacConservative$V1)

cat("Peaks in naive set:", no_peaksN,"\n", "Peaks in optimal
set:",no_peaksO,"\n",
    "Peaks in conservative set:",no_peaksC, "\n")
```

```
## Peaks in naive set: 254256
## Peaks in optimal set: 192013
## Peaks in conservative set: 175068
```

```
# Making data.frames with relevant info and removing any duplicates of
# peaks to easier check for duplicates between sets
naive <- unique(data.frame(atacNaive$V1, atacNaive$V2+atacNaive$V10))
optimal <- unique(data.frame(atacOptimal$V1,
atacOptimal$V2+atacOptimal$V10))
conservative <- unique(data.frame(atacConservative$V1,
atacConservative$V2+atacConservative$V10))

# Changing the column names to merge data.frames later
colnames(naive) <- c("Chromosome", "PeakValue")
colnames(optimal) <- c("Chromosome", "PeakValue")
colnames(conservative) <- c("Chromosome", "PeakValue")
```

```
# Merging the data.frames before checking for duplicates/overlaps between
the sets
NaiveOptimal <- which(duplicated(rbind(naive, optimal)))

NaiveConservative <- which(duplicated(rbind(naive,conservative)))

ConservativeOptimal <- which(duplicated(rbind(conservative, optimal)))

# Making a venn diagram

draw.triple.venn(no_peaksN, no_peaksO, no_peaksC, length(NaiveOptimal),
                 length(ConservativeOptimal), length(NaiveConservative),
                 length(ConservativeOptimal),
                 category = c("Naive", "Optimal", "Conservative"),
                 fill=c("blue", "green", "red"))
```




```
## (polygon[GRID.polygon.1], polygon[GRID.polygon.2],
polygon[GRID.polygon.3], polygon[GRID.polygon.4], polygon[GRID.polygon.5],
polygon[GRID.polygon.6], text[GRID.text.7], text[GRID.text.8],
text[GRID.text.9], text[GRID.text.10], text[GRID.text.11],
text[GRID.text.12], text[GRID.text.13], text[GRID.text.14],
text[GRID.text.15])
```

Bedtools closest -D ref

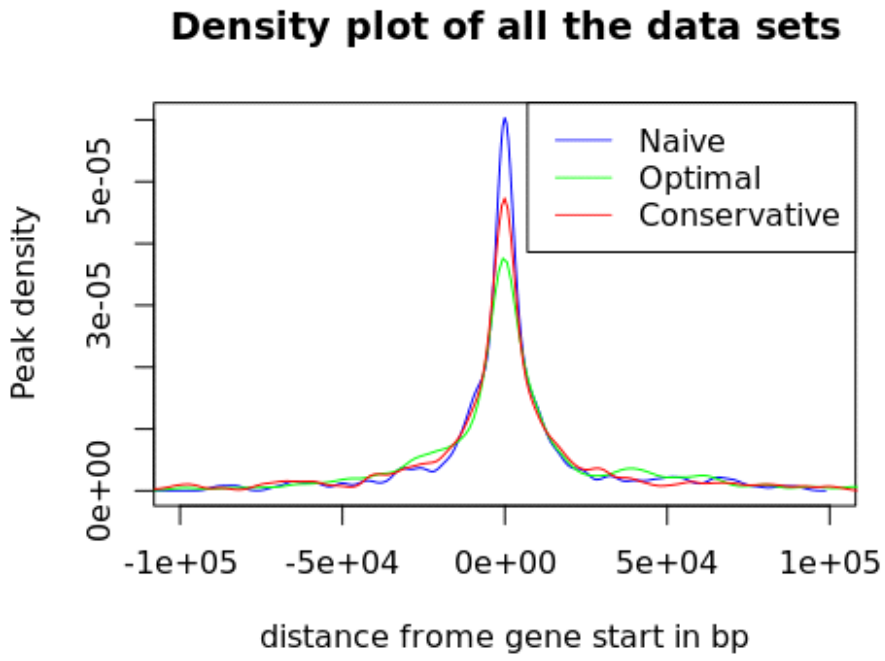
```
# # Bedtools closest
# A.in <-"/mnt/users/ckristia/R/bedtools/genes.bed" # bed file with the
salmon genes
# B1.in <-"/mnt/users/ckristia/R/bedtools/naive.bed" # a sorted bed file
with the naive dataset
# B2.in <- "/mnt/users/ckristia/R/bedtools/optimal.bed" # a sorted bed file
```

```
with the optimal dataset
# B3.in <- "/mnt/users/ckristia/R/bedtools/conservative.bed" # a sorted bed
file with the conservative dataset
#
#
# ## Building the command line,
# cmd <- paste("bedtools closest",
#              "-D",           # Gives distance between gene start and
peaks, gives negative distance for upstream distance
#              "ref",          # Report distance with respect to reference
genome
#              "-a",
#              A.in,
#              "-b",
#              B1.in,
#              "> /mnt/users/ckristia/R/markdown/Data/distN.bed") #
defining outfile
#
#
# system(cmd)
#
# cmd <- paste("bedtools closest",
#              "-D",           # Gives distance between gene start and
peaks, gives negative distance for upstream distance
#              "ref",          # Report distance with respect to reference
genome
#              "-a",
#              A.in,
#              "-b",
#              B2.in,
#              "> /mnt/users/ckristia/R/markdown/Data/distO.bed") #
defining outfile
#
#
# system(cmd)
# cmd <- paste("bedtools closest",
#              "-D",           # Gives distance between gene start and
peaks, gives negative distance for upstream distance
#              "ref",          # Report distance with respect to reference
genome
#              "-a",
#              A.in,
#              "-b",
#              B3.in,
#              "> /mnt/users/ckristia/R/markdown/Data/distC.bed") #
defining outfile
#
#
# system(cmd)
```

Density plot

```
#Naive
# Opening bedtools output
distN <- fread("/mnt/users/ckristia/R/markdown/Data/distN.bed", data.table
```

```
= F)
# Distance=0 indicates overlapping features, distance= -1 indicates no peak
on the genes chromosome/scaffold, distances are in column 8
# Creating an integer with the distances for peaks that do not overlap a
gene
dist_N <- distN$V8[which(distN$V8!= -1)]
Ndist <- dist_N[which(dist_N!=0)]

# Optimal
distO <- fread("/mnt/users/ckristia/R/markdown/Data/distO.bed", data.table
= F)
dist_O <- distO$V8[which(distO$V8!= -1)]
Odist <- dist_O[which(dist_O !=0)]

# Conservative
distC <- fread("/mnt/users/ckristia/R/markdown/Data/distC.bed", data.table
= F)
dist_C <- distC$V8[which(distC$V8!= -1)]

Cdist <- dist_C[which(dist_C!=0)]

plot(density(Ndist[12000:13514]), xlim = c(-100000, 100000),col="blue",
main="Density plot of all the data sets", xlab = "distance frome gene start
in bp", ylab="Peak density")
lines(density(Odist[28000:29410]), col="green")
lines(density(Cdist[17000:19000]), col="red")
legend(x="topright", legend=c("Naive", "Optimal", "Conservative"),lwd=0.8,
col=c("blue", "green", "red"))
```



Bedtools coverage

Using all the peaks from the optimal data set and all genes with their geneID's

```
# all genes
all_genes[1:3,]
```

```
##               seqid start  end GeneID        ID strand
## 2958551 NC_001960.1  3846 4820 808311 gene81573      +
## 2958559 NC_001960.1  5036 6085 808316 gene81574      +
## 2958572 NC_001960.1  6476 8026 808314 gene81575      +
```

```
#Peaks from optimal data set

OptimalPeaks <- fread("/mnt/users/ckristia/R/Data/optimal.bed",
data.table=F)
names(OptimalPeaks) <- c("Chrom", "Start", "Stop", "Name", "Score",
"Strand", "SignalVal","pVal", "qVal","Peak")#Peak=point source
OptimalPeaks[1:3,]
```

```
##         Chrom Start Stop Name Score Strand SignalVal      pVal      qVal
## 1 NC_001960.1     2 4714    .  1000      .   1.07483  82.79078  80.75653
## 2 NC_001960.1     2 4714    .  1000      .   1.15535 314.40884 311.99133
## 3 NC_001960.1     2 4714    .  1000      .   1.17755 428.12784 425.58817
##   Peak
## 1 4457
## 2 3871
## 3 4106
```

Creating input files

Creating the files with the wanted areas to find out more about the peak distribution compared to gene start

```
# Creating file A based on all genes
# 100 bp downstream to 1000 bp upstream of gene start, did this for all the
different distances (are1-area10)
stops <- all_genes$start+100
starts <- all_genes$start - 1000
for(i in 1:length(starts)){
  if (starts[i]<0){
  starts[i]<- 0 #changing negative values to 0
}
}

a <- data.frame(all_genes$seqid, # Chromosome
                starts, # promoter start
                stops, # promoter end
                all_genes$GeneID, # GeneID
                all_genes$ID, # ID
                all_genes$start, # gene start
                all_genes$end, # gene end
                all_genes$strand)
#write.table(a,file="/mnt/users/ckristia/R/Data/Area1.bed",quote=F,sep="\t"
,col.names=F,row.names=F)
```

```
a[1:3,]
```

```
##   all_genes.seqid starts stops all_genes.GeneID all_genes.ID
## 1     NC_001960.1   2846  3946           808311    gene81573
## 2     NC_001960.1   4036  5136           808316    gene81574
## 3     NC_001960.1   5476  6576           808314    gene81575
##   all_genes.start all_genes.end all_genes.strand
## 1            3846          4820                +
## 2            5036          6085                +
## 3            6476          8026                +
```

```
#Creating file B (all "optimal" peaks)
b <- data.frame(OptimalPeaks[,c(1:3,5,7:10)])
#write.table(b,file="/mnt/users/ckristia/R/Data/AllOpeaks.bed",quote=F,sep=
"\t",col.names=F,row.names=F)
b[1:3,]
```

```
##         Chrom Start Stop Score SignalVal      pVal      qVal Peak
## 1 NC_001960.1     2 4714  1000   1.07483  82.79078  80.75653 4457
## 2 NC_001960.1     2 4714  1000   1.15535 314.40884 311.99133 3871
## 3 NC_001960.1     2 4714  1000   1.17755 428.12784 425.58817 4106
```

Running bedtools for all the areas, and making tables with relevant info. Bedtools coverage gives the additional columns: 1.The number of features in B that overlapped (by at least one base pair) the A interval. 2.The number of bases in A that had non-zero coverage from features in B. 3.The length of the entry in A. 4.The fraction of bases in A that had non-zero coverage from features in B. (source: https://bedtools.readthedocs.io/en/latest/content/tools/coverage.html)

```
#Did this for all areas, example of area1
#bedtools coverage

# A.in <-"/mnt/users/ckristia/R/Data/Area1.bed"
# B.in <-"/mnt/users/ckristia/R/Data/AllOpeaks.bed"
#
#
# ## Building the command line
# cmd <- paste("bedtools coverage",
#              "-a",
#              A.in,
#              "-b",
#              B.in,
#              "> /mnt/users/ckristia/R/Data/covArea1.bed")# defining
outfile
#
#
# system(cmd)

coverage <- fread("/mnt/users/ckristia/R/Data/covArea1.bed", data.table =
F)

tab_area1 <- data.frame(coverage[,c(1,4:9,12)])
names(tab_area1) <- c("Chromosome", "GeneID","ID", "Start",
"Stop","Strand"
```

```
                    ,"No.peaks" ,"Coverage")
tab_area1[1:3,]
```

```
##    Chromosome GeneID        ID Start Stop Strand No.peaks  Coverage
## 1 NC_001960.1 808311 gene81573  3846 4820      +       14 1.0000000
## 2 NC_001960.1 808316 gene81574  5036 6085      +       17 0.9427273
## 3 NC_001960.1 808314 gene81575  6476 8026      +        2 0.9327273
```

A table was created for all the areas and then merged together to one table with coverage and peak number information for all the areas

Putting together the coverage table

```
#tab_cov <- list(tab_area1,tab_area2, tab_area3, tab_area4,
tab_area5,tab_area6, tab_area7,tab_area8,tab_area9, tab_area10) %>%
  # reduce(left_join,by=c("Chromosome"="Chromosome", "GeneID"="GeneID",
"ID"="ID", "Start"="Start", "Stop"="Stop", "Strand"="Strand"))

#names(tab_cov) <- c("Chromosome", "GeneID","ID", "Start", "Stop",
                    # "Strand", "A1Peaks", "A1Cov", "A2Peaks",
                    #
"A2Cov","A3Peaks","A3Cov","A4Peaks","A4Cov","A5Peaks",
                    # "A5Cov","A6Peaks","A6Cov", "A7Peaks",
"A7Cov","A8Peaks",
                    # "A8Cov", "A9Peaks", "A9Cov", "A10Peaks", "A10Cov")
#save(tab_cov, file="/mnt/users/ckristia/R/Data/coveragetable.Rdata")
load("/mnt/users/ckristia/R/Data/coveragetable.Rdata")
tab_cov[1:3,]
```

```
##    Chromosome GeneID        ID Start Stop Strand A1Peaks     A1Cov
A2Peaks
## 1 NC_001960.1 808311 gene81573  3846 4820      +      14 1.0000000
14
## 2 NC_001960.1 808316 gene81574  5036 6085      +      17 0.9427273
14
## 3 NC_001960.1 808314 gene81575  6476 8026      +       2 0.9327273
17
##       A2Cov A3Peaks A3Cov A4Peaks     A4Cov A5Peaks A5Cov A6Peaks
## 1 1.0000000      14     1      14 0.9976332       0     0       0
## 2 1.0000000      14     1      14 1.0000000      14     1      14
## 3 0.8968969      14     1      14 1.0000000      14     1      14
##       A6Cov A7Peaks     A7Cov A8Peaks A8Cov A9Peaks A9Cov A10Peaks
A10Cov
## 1 0.0000000       0 0.0000000       0     0       0     0        0
0
## 2 0.9428571       0 0.0000000       0     0       0     0        0
0
## 3 1.0000000      14 0.9957895       0     0       0     0        0
0
```

Merging coverage table with gene expression

The RNA-seq count data from the four samples were TPM normalized and log2 adjusted before adding the mean expression of the four fish. Creating a table with all the genes that have expression value and their peak data

```
# Expression table
tab_expr[1:3,]
```

```
##      GeneID        ID Fish1 Fish2 Fish3 Fish4 meansof4
## 1 106560212     gene0     0     0     0     0        0
## 2 106603566 gene10000     0     0     0     0        0
## 3 106603565 gene10001     0     0     0     0        0
```

```
# # merging the data.frames into one
# tab <- merge(tab_cov, tab_expr, by = c("ID"="ID","GeneID"="GeneID"))
#
# # sum peaks and coverage in +100 --> - 10000
# tab$totpeaks <- rowSums(tab[,grep('Peaks', colnames(tab))], na.rm = T)
# tab$totcov <- rowMeans(tab[,grep('Cov', colnames(tab))], na.rm = T)
#
# save(tab, file="/mnt/users/ckristia/R/markdown/Data/tab.Rdata")
load("/mnt/users/ckristia/R/markdown/Data/tab.Rdata")
tab[1:3,]
```

```
##       ID    GeneID Chromosome  Start   Stop Strand A1Peaks     A1Cov
## 1  gene0 106560212      ssa01   5501  62139      -       1 0.5345455
## 2  gene1 106607996      ssa01 160437 198815      -       0 0.0000000
## 3 gene10 106599499      ssa01 516060 519262      +       0 0.0000000
##   A2Peaks A2Cov A3Peaks A3Cov A4Peaks A4Cov A5Peaks A5Cov A6Peaks A6Cov
## 1       0     0       0     0       0     0       0     0       0     0
## 2       0     0       0     0       0     0       0     0       0     0
## 3       0     0       0     0       0     0       0     0       0     0
##   A7Peaks A7Cov A8Peaks A8Cov A9Peaks A9Cov A10Peaks A10Cov      Fish1
## 1       0     0       0     0       0     0        0      0 0.00000000
## 2       0     0       0     0       0     0        0      0 0.47508136
## 3       0     0       0     0       0     0        0      0 0.04840276
##       Fish2     Fish3      Fish4  meansof4 totpeaks     totcov
## 1 0.0000000 0.0000000 0.00000000 0.0000000        1 0.05345455
## 2 0.1151858 0.0487502 0.73042042 0.3423595        0 0.00000000
## 3 0.2621157 0.1142145 0.07919524 0.1259821        0 0.00000000
```

Correlation

Correlation test between the mean expression of the four sample fish and different peak data

```
cor.test(tab$meansof4, tab$A1Peaks) # Number of peaks in the promoterregion
```

```
##
##  Pearson's product-moment correlation
##
## data:  tab$meansof4 and tab$A1Peaks
## t = 104.9, df = 79016, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
```

```
## 95 percent confidence interval:
##  0.3435015 0.3557418
## sample estimates:
##       cor
## 0.3496366
```

```
cor.test(tab$meansof4, tab$A1Cov) # Coverage in the promoterregion
```

```
##
##  Pearson's product-moment correlation
##
## data:  tab$meansof4 and tab$A1Cov
## t = 104.4, df = 79016, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3420276 0.3542822
## sample estimates:
##       cor
## 0.3481698
```

```
cor.test(tab$meansof4, tab$totpeaks) # Peaks from +100/ -10000
```

```
##
##  Pearson's product-moment correlation
##
## data:  tab$meansof4 and tab$totpeaks
## t = 101.37, df = 79016, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3330575 0.3453976
## sample estimates:
##       cor
## 0.3392421
```

```
cor.test(tab$meansof4, tab$totcov) # Coverage from +100/ -10000
```

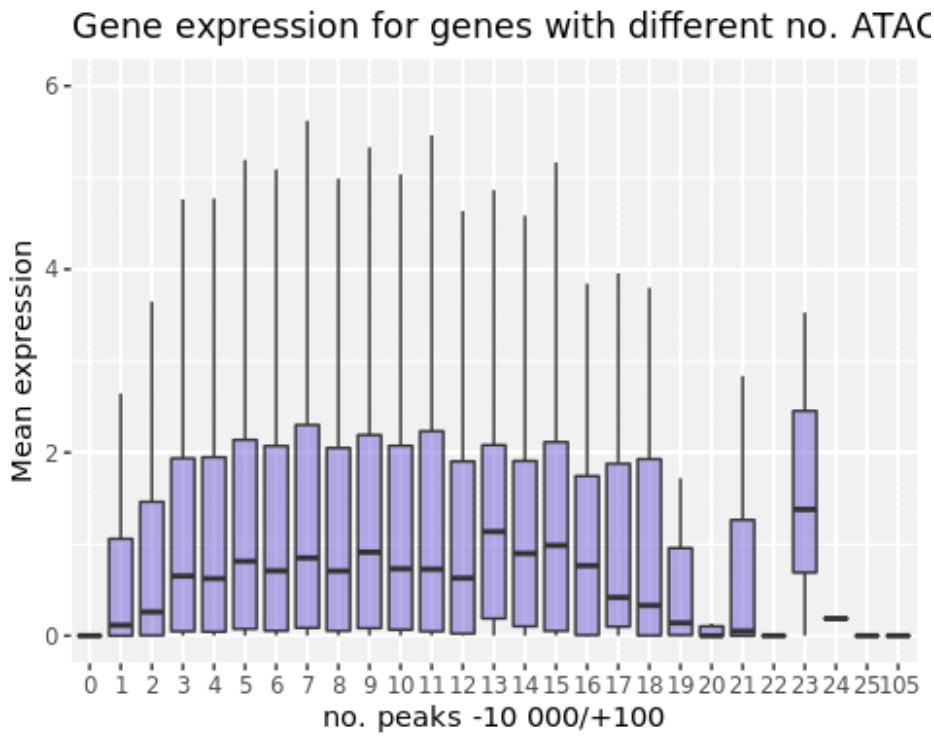
```
##
##  Pearson's product-moment correlation
##
## data:  tab$meansof4 and tab$totcov
## t = 107.44, df = 79016, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3509404 0.3631077
## sample estimates:
##       cor
## 0.3570392
```

Boxplot

```
# The mean expression of the four fish is the continuous variable, while
the
# no. peaks and coverage is the quantitative variable in the boxplots

# no.peaks
ggplot(tab,aes(x=as.factor(totpeaks),y=meansof4)) +
```

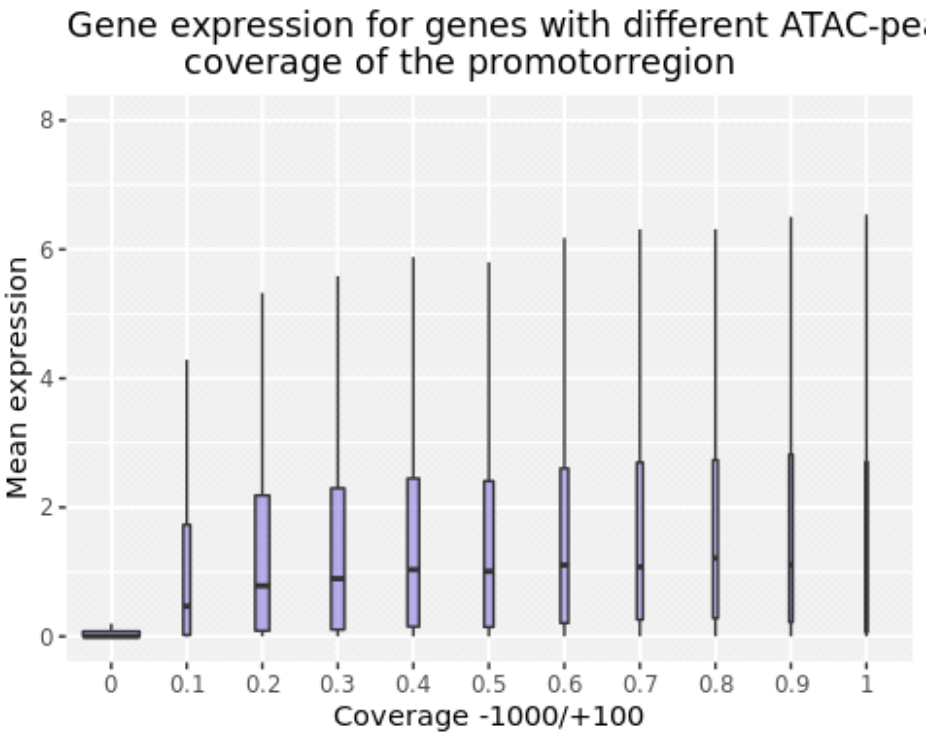
```
  geom_boxplot(fill="slateblue",alpha=0.5,outlier.shape = NA) +
  ggtitle("Gene expression for genes with different no. ATAC-peaks") +
  xlab("no. peaks -10 000/+100")+
  ylab("Mean expression") +
  ylim(0,6)
```

```
## Warning: Removed 928 rows containing non-finite values (stat_boxplot).
```




```
#coverage, promoterregion
ggplot(tab,aes(x=as.factor(round(A1Cov,1)),y=meansof4)) +
  geom_boxplot(alpha=0.5, fill="slateblue", outlier.shape = NA, varwidth =
T) +
  ggtitle("Gene expression for genes with different ATAC-peak
          coverage of the promotorregion") +
  xlab("Coverage -1000/+100") +
  ylab("Mean expression") +
  ylim(0,8)
```

```
## Warning: Removed 359 rows containing non-finite values (stat_boxplot).
```



Expression divergence

EVE data for expression divergence

Functions to use with the EVE-data

```
# function to get the up-/downregulated duplicates geneID
# also adds column with product description of the gene
load('/mnt/users/ckristia/R/Data/EVE.clan.tables.RData') #used to get IDs
and gene pair relationship

all.genes <- get.id('*')

add.info <- function(EVE){

  idx <- c()
  for (i in 1: nrow(EVE)) {

    n <- if (EVE$gene.type[i] == "dupA"){
      match(EVE$Ssal.dupA[i], sub('\\..*', '', all.genes$protein_id))
    } else {
      match(EVE$Ssal.dupB[i], sub('\\..*', '', all.genes$protein_id))
    }
    idx <- c(idx,n)
  }


  EVE$geneID <- all.genes$gene_id[idx] #adds the GeneID of the up-
/downregulated gene
  EVE$product <- get.genes(EVE$geneID, match = T)$product
```

```
  return(EVE)
}


# The function takes a tbl_df table containing EVE results and merges it
with
# the tab table which contains geneIDs, peak info and expression info etc
# dup_data as input, first gets the geneIDs of both duplicates,
# returns a list containing dupe_data table and a few vectors

# load table with peak coverage info
load("/mnt/users/ckristia/R/markdown/Data/tab.Rdata")
get.dups <- function(dup_data) {
  # getting the geneid
  dup_data$ID <- sapply(strsplit(dup_data$geneID, split = '\\:'), '[[', 1)
  # merging atac data with EVE dupe data
  dup_data <- merge(dup_data, tab, by="ID")

  ## getting the ID of both duplicates
  all = get.id('*')
  all$protein_id_v2 <- gsub('\\..*', '', all$protein_id)
  dup_data$dupAID = gsub( '\\:.*', '',
all$gene_id[match(dup_data$Ssal.dupA, all$protein_id_v2)])
  dup_data$dupBID = gsub( '\\:.*', '',
all$gene_id[match(dup_data$Ssal.dupB, all$protein_id_v2)])

  dup_data <- dup_data[!duplicated(dup_data$ID),]
  A <- NULL
  B <- NULL
  diff.peaks <- rep(NA, nrow(dup_data))
  diff.cov = rep(NA, nrow(dup_data))
  diff.expr = rep(NA, nrow(dup_data))
  diff.theta = rep(NA, nrow(dup_data))
  diff.cov.all = rep(NA, nrow(dup_data))

  for(i in 1:nrow(dup_data)){
  # defining significant shift geneID collumn
    A = grep(dup_data[i, 'gene.type'], colnames(dup_data))[2]
    if(gsub('dup', '', dup_data[i, 'gene.type']) == 'A') {
      B = grep(dup_data[i, 'gene.type'], colnames(dup_data))[2]+1
      } else { B = grep(dup_data[i, 'gene.type'], colnames(dup_data))[2]-1
}

  # get the absolute difference between shift gene and non-shift gene

    diff.cov[i] <- abs(tab$A1Cov[match(dup_data[i,A], tab$ID)] -
tab$A1Cov[match(dup_data[i,B], tab$ID)])
    diff.peaks[i] <- abs(tab$totpeaks[match(dup_data[i,A], tab$ID)] -
tab$totpeaks[match(dup_data[i,B], tab$ID)])
    diff.expr[i] <- abs(tab$meansof4[match(dup_data[i,A], tab$ID)] -
tab$meansof4[match(dup_data[i,B], tab$ID)])
    diff.theta[i] <- abs(dup_data[i,'thetaShift'])
    diff.cov.all[i] <- abs(tab$totcov[match(dup_data[i,A], tab$ID)] -
tab$totcov[match(dup_data[i,B], tab$ID)])
```

```
  }
  # Creating list to be able to return all the relevant variables
  lst <- list("dup_data"=dup_data, "diff.cov"=diff.cov,
              "diff.peaks"=diff.peaks, "diff.expr"=diff.expr,
              "diff.theta"=diff.theta, "diff.cov.all"=diff.cov.all)
  return(lst)
}
```

Getting upregulated and downregulated genes in one table each

```
# Loading EVE results tables with up-/downregulated duplicates
load('/mnt/users/ckristia/R/Data/EVE.results.table.31.10.RData')
load('/mnt/users/ckristia/R/Data/EVE.clan.tables.RData')

library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse
1.2.1 ──
```

```
## ✔ tibble  2.0.0     ✔ readr   1.3.1
## ✔ tidyr   0.8.2     ✔ stringr 1.3.1
## ✔ tibble  2.0.0     ✔ forcats 0.3.0
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::between()   masks data.table::between()
## ✖ dplyr::filter()    masks stats::filter()
## ✖ dplyr::first()     masks data.table::first()
## ✖ dplyr::lag()       masks stats::lag()
## ✖ dplyr::last()      masks data.table::last()
## ✖ purrr::transpose() masks data.table::transpose()
```

```
# LRT>4 gives only signficant dupes
EVE.up <- EVE.results.table %>% filter(shift.direction == 'up' & LRT > 4 &
test.type == 'Ss4R' & gene.type %in% c('dupA', 'dupB') & clan.maxLRT ==
'TRUE' & data.type == 'BSNsgl')
EVE.down <- EVE.results.table %>% filter(shift.direction == 'down' & LRT >
4 & test.type == 'Ss4R' & gene.type %in% c('dupA', 'dupB') & clan.maxLRT ==
'TRUE' & data.type == 'BSNsgl')
head(EVE.up)
```

```
## # A tibble: 6 x 12
##   clan  data.type gene.type test.type   LRT clan.maxLRT   theta
thetaShift
##   <chr> <fct>     <fct>     <fct>     <dbl> <lgl>         <dbl>
<dbl>
## 1 OG00… BSNsgl    dupB      Ss4R       4.71 TRUE        0.0161
28.7
## 2 OG00… BSNsgl    dupA      Ss4R       5.21 TRUE        0.0874
4.28
## 3 OG00… BSNsgl    dupB      Ss4R      13.3  TRUE        1.07
2.92
## 4 OG00… BSNsgl    dupB      Ss4R       5.56 TRUE        0.722
1.87
```

```
## 5 OG00… BSNsgl    dupA      Ss4R       4.28 TRUE        0.00521
1.71
## 6 OG00… BSNsgl    dupA      Ss4R       8.86 TRUE        0.340
2.63
## # … with 4 more variables: shift.direction <fct>, alpha <dbl>, beta
<dbl>,
## #   sigma.sq <dbl>
```

```
#Adding column with the ID of duplicate A and B
EVE.up$Ssal.dupA <- duplicate.clan.table$Ssal.a[match(EVE.up$clan,
duplicate.clan.table$clan)]
EVE.up$Ssal.dupB <- duplicate.clan.table$Ssal.b[match(EVE.up$clan,
duplicate.clan.table$clan)]
EVE.up[1:3,]
```

```
## # A tibble: 3 x 14
##   clan  data.type gene.type test.type   LRT clan.maxLRT  theta
thetaShift
##   <chr> <fct>     <fct>     <fct>     <dbl> <lgl>        <dbl>
<dbl>
## 1 OG00… BSNsgl    dupB      Ss4R       4.71 TRUE        0.0161      28.7
## 2 OG00… BSNsgl    dupA      Ss4R       5.21 TRUE        0.0874
4.28
## 3 OG00… BSNsgl    dupB      Ss4R      13.3  TRUE        1.07
2.92
## # … with 6 more variables: shift.direction <fct>, alpha <dbl>, beta
<dbl>,
## #   sigma.sq <dbl>, Ssal.dupA <chr>, Ssal.dupB <chr>
```

```
#Using a function that gets the ID of the up-/downregulated duplicate
indicated by gene.type
EVE.up <- add.info(EVE.up)
EVE.up[1:3,]
```

```
## # A tibble: 3 x 16
##   clan  data.type gene.type test.type   LRT clan.maxLRT  theta
thetaShift
##   <chr> <fct>     <fct>     <fct>     <dbl> <lgl>        <dbl>
<dbl>
## 1 OG00… BSNsgl    dupB      Ss4R       4.71 TRUE        0.0161      28.7
## 2 OG00… BSNsgl    dupA      Ss4R       5.21 TRUE        0.0874
4.28
## 3 OG00… BSNsgl    dupB      Ss4R      13.3  TRUE        1.07
2.92
## # … with 8 more variables: shift.direction <fct>, alpha <dbl>, beta
<dbl>,
## #   sigma.sq <dbl>, Ssal.dupA <chr>, Ssal.dupB <chr>, geneID <chr>,
## #   product <chr>
```

```
#Same for downregulated duplicates
#Adding column with the id of duplicate A and B
EVE.down$Ssal.dupA <- duplicate.clan.table$Ssal.a[match(EVE.down$clan,
duplicate.clan.table$clan)]
EVE.down$Ssal.dupB <- duplicate.clan.table$Ssal.b[match(EVE.down$clan,
duplicate.clan.table$clan)]
```

```
EVE.down <- add.info(EVE.down)


# The GeneID column in EVE.up and EVE.down should now be the GeneID of the
duplicate with the change in expression
#save(EVE.up,
file="/mnt/users/ckristia/R/markdown/Data/EVE.up.dupes.Rdata")
#save(EVE.down,
file="/mnt/users/ckristia/R/markdown/Data/EVE.down.dupes.Rdata")
```

Upregulated genes

```
load("/mnt/users/ckristia/R/markdown/Data/EVE.up.dupes.Rdata")

# Using the get.dups function and extracting the variables from the list
# output

dup_lst.up <- get.dups(EVE.up)
dup_data.up <- dup_lst.up$dup_data
diff.cov.up <- dup_lst.up$diff.cov
diff.peaks.up <- dup_lst.up$diff.peaks
diff.expr.up <- dup_lst.up$diff.expr
diff.theta.up <- dup_lst.up$diff.theta
diff.cov.all.up <- dup_lst.up$diff.cov.all

dup_data.up[1:3,]
```

```
##          ID        clan data.type gene.type test.type      LRT
clan.maxLRT
## 1 gene10040 OG0007686_1    BSNsgl      dupB      Ss4R 4.355495
TRUE
## 2 gene10138 OG0003954_1    BSNsgl      dupB      Ss4R 6.581602
TRUE
## 3 gene10284 OG0006856_1    BSNsgl      dupB      Ss4R 5.922879
TRUE
##       theta thetaShift shift.direction        alpha     beta
sigma.sq
## 1 0.1285349   1.222721              up 6.452891e+01 2.038816
1.380691e+01
## 2 1.4814984   2.494448              up 1.140850e+07 1.770308
1.854098e+06
## 3 2.0401273   2.865670              up 5.646068e+01 3.246431
4.011567e+00
##      Ssal.dupA    Ssal.dupB              geneID
## 1 XP_013993348 XP_014052829 gene10040:106603530
## 2 XP_013992709 XP_014052627 gene10138:106603450
## 3 NP_001135039 XP_014053249 gene10284:106603731
##                                     product    GeneID Chromosome
Start
## 1   C2 calcium-dependent domain containing 2 106603530      ssa04
49386425
## 2 rab GTPase-binding effector protein 1-like 106603450      ssa04
55131992
## 3                         TIP41-like protein 106603731      ssa04
```

```
63228496
##       Stop Strand A1Peaks     A1Cov A2Peaks A2Cov A3Peaks A3Cov A4Peaks
## 1 49415417      +       5 0.8854545       0     0       0     0       0
## 2 55169353      +       4 0.3018182       0     0       0     0       0
## 3 63233956      -       4 0.8854545       0     0       0     0       0
##   A4Cov A5Peaks A5Cov A6Peaks     A6Cov A7Peaks A7Cov A8Peaks A8Cov
## 1     0       0     0       0 0.0000000       0     0       0     0
## 2     0       0     0       1 0.2632633       0     0       0     0
## 3     0       0     0       0 0.0000000       0     0       0     0
##   A9Peaks A9Cov A10Peaks A10Cov     Fish1     Fish2     Fish3     Fish4
## 1       0     0        0      0 0.2213262 0.1362179 0.1256381 0.1371302
## 2       0     0        0      0 2.1348941 2.5629574 2.9263505 1.7084812
## 3       0     0        0      0 2.7618516 3.1325692 3.1497970 2.4244363
##    meansof4 totpeaks     totcov    dupAID    dupBID
## 1 0.1550781        5 0.08854545 gene28935 gene10040
## 2 2.3331708        5 0.05650815 gene28825 gene10138
## 3 2.8671635        4 0.08854545 gene28686 gene10284
```

```
#correlation tests
#expression, difference
cor.test(diff.peaks.up, diff.expr.up)
```

```
##
##  Pearson's product-moment correlation
##
## data:  diff.peaks.up and diff.expr.up
## t = 0.25287, df = 115, p-value = 0.8008
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1586385  0.2042323
## sample estimates:
##        cor
## 0.02357334
```

```
cor.test(diff.cov.up, diff.expr.up)
```

```
##
##  Pearson's product-moment correlation
##
## data:  diff.cov.up and diff.expr.up
## t = 0.99044, df = 115, p-value = 0.324
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.09108571  0.26900975
## sample estimates:
##        cor
## 0.09196796
```

```
cor.test(diff.cov.all.up, diff.expr.up)
```

```
##
##  Pearson's product-moment correlation
##
## data:  diff.cov.all.up and diff.expr.up
## t = 0.99682, df = 115, p-value = 0.3209
```

```
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.09049862  0.26955880
## sample estimates:
##       cor
## 0.09255489

#no difference
cor.test(dup_data.up$meansof4, dup_data.up$A1Peaks)

##
##  Pearson's product-moment correlation
##
## data:  dup_data.up$meansof4 and dup_data.up$A1Peaks
## t = 1.8107, df = 115, p-value = 0.0728
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.01551316  0.33781189
## sample estimates:
##       cor
## 0.1664887

cor.test(dup_data.up$meansof4, dup_data.up$A1Cov)

##
##  Pearson's product-moment correlation
##
## data:  dup_data.up$meansof4 and dup_data.up$A1Cov
## t = 2.8021, df = 115, p-value = 0.005959
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.07470449 0.41528334
## sample estimates:
##      cor
## 0.252809

cor.test(dup_data.up$meansof4, dup_data.up$totpeaks)

##
##  Pearson's product-moment correlation
##
## data:  dup_data.up$meansof4 and dup_data.up$totpeaks
## t = 1.9142, df = 115, p-value = 0.05808
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.005999127  0.346213991
## sample estimates:
##       cor
## 0.1757252

cor.test(dup_data.up$meansof4, dup_data.up$totcov)

##
##  Pearson's product-moment correlation
##
## data:  dup_data.up$meansof4 and dup_data.up$totcov
```

```
## t = 1.9877, df = 115, p-value = 0.04922
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.0007410844 0.3521324554
## sample estimates:
##       cor
## 0.1822495
```

Downregulated genes

```
load("/mnt/users/ckristia/R/Data/EVE.down.dupes.Rdata")

# Using the get.dups function and extracting the variables from the list
# output
dup_lst.down <- get.dups(EVE.down)
dup_data.down <- dup_lst.down$dup_data
diff.cov.down <- dup_lst.down$diff.cov
diff.peaks.down <- dup_lst.down$diff.peaks
diff.expr.down <- dup_lst.down$diff.expr
diff.theta.down <- dup_lst.down$diff.theta
diff.cov.all.down <- dup_lst.down$diff.cov.all

dup_data.down[1:3,]
```

```
##          ID        clan data.type gene.type test.type       LRT
## 1 gene10016 OG0005206_1    BSNsgl      dupB      Ss4R 14.307513
## 2 gene10042 OG0008247_1    BSNsgl      dupB      Ss4R  5.662542
## 3 gene10059 OG0006185_1    BSNsgl      dupA      Ss4R  5.226558
##   clan.maxLRT    theta  thetaShift shift.direction       alpha
## 1        TRUE 2.980578 1.391384498            down 4086237.768
## 2        TRUE 2.326471 0.008267389            down    2015.207
## 3        TRUE 0.780068 0.003023178            down    2331.408
##           beta     sigma.sq    Ssal.dupA    Ssal.dupB
geneID
## 1 6.432688e+08 3.706047e-03 XP_013993315 XP_014052863
gene10016:106603551
## 2 7.595542e-01 3.863408e+03 XP_013993350 XP_014052821
gene10042:106603528
## 3 3.913333e+00 3.568146e+02 XP_014052781 XP_013993372
gene10059:106603515
##                                                  product    GeneID
## 1               ubiquitin carboxyl-terminal hydrolase 25-like 106603551
## 2 receptor-interacting serine/threonine-protein kinase 4-like 106603528
## 3            A-kinase anchor protein 10%2C mitochondrial-like 106603515
##   Chromosome    Start     Stop Strand A1Peaks     A1Cov A2Peaks
A2Cov
## 1      ssa04 48543604 48608795      +       2 0.2209091       0
0.0000000
## 2      ssa04 49457439 49471649      +       0 0.0000000       1
0.2342342
## 3      ssa04 51145399 51176417      -       0 0.0000000       0
0.0000000
##   A3Peaks A3Cov A4Peaks     A4Cov A5Peaks A5Cov A6Peaks A6Cov A7Peaks
## 1       0     0       1 0.2152152       0     0       0     0       0
## 2       0     0       0 0.0000000       0     0       0     0       0
```

```
## 3       0     0       0 0.0000000       0     0       0     0       0
##   A7Cov A8Peaks A8Cov A9Peaks A9Cov A10Peaks A10Cov     Fish1     Fish2
## 1     0       0     0       0     0        0      0 0.9170117 0.8282533
## 2     0       0     0       0     0        0      0 0.5242957 0.5571842
## 3     0       0     0       0     0        0      0 0.2639514 0.2324414
##       Fish3     Fish4  meansof4 totpeaks     totcov    dupAID    dupBID
## 1 0.9123800 0.4468718 0.7761292        3 0.04361243 gene28956 gene10016
## 2 0.6894081 0.3982216 0.5422774        1 0.02342342 gene28933 gene10042
## 3 0.2474455 0.3945775 0.2846040        0 0.00000000 gene10059 gene28914
```

```
#correlation tests
#mean expression, difference
cor.test(diff.peaks.down, diff.expr.down)
```

```
##
##  Pearson's product-moment correlation
##
## data:  diff.peaks.down and diff.expr.down
## t = -0.14274, df = 619, p-value = 0.8865
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.08437749  0.07297408
## sample estimates:
##          cor
## -0.005737221
```

```
cor.test(diff.cov.down, diff.expr.down)
```

```
##
##  Pearson's product-moment correlation
##
## data:  diff.cov.down and diff.expr.down
## t = -0.79941, df = 619, p-value = 0.4244
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.11051346  0.04668198
## sample estimates:
##         cor
## -0.03211433
```

```
cor.test(diff.cov.all.down, diff.expr.down)
```

```
##
##  Pearson's product-moment correlation
##
## data:  diff.cov.all.down and diff.expr.down
## t = 0.98831, df = 619, p-value = 0.3234
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.0391084  0.1180019
## sample estimates:
##       cor
## 0.0396921
```

```r
#no difference
cor.test(dup_data.down$meansof4, dup_data.down$A1Peaks)
```

```
##
##  Pearson's product-moment correlation
##
## data:  dup_data.down$meansof4 and dup_data.down$A1Peaks
## t = 1.5557, df = 619, p-value = 0.1203
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.01635189  0.14039577
## sample estimates:
##       cor
## 0.06240676
```

```r
cor.test(dup_data.down$meansof4, dup_data.down$A1Cov)
```

```
##
##  Pearson's product-moment correlation
##
## data:  dup_data.down$meansof4 and dup_data.down$A1Cov
## t = 0.53199, df = 619, p-value = 0.5949
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.05739724  0.09988802
## sample estimates:
##       cor
## 0.02137766
```

```r
cor.test(dup_data.down$meansof4, dup_data.down$totpeaks)
```

```
##
##  Pearson's product-moment correlation
##
## data:  dup_data.down$meansof4 and dup_data.down$totpeaks
## t = 2.4649, df = 619, p-value = 0.01398
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.02006682 0.17590342
## sample estimates:
##       cor
## 0.09858952
```

```r
cor.test(dup_data.down$meansof4, dup_data.down$totcov)
```

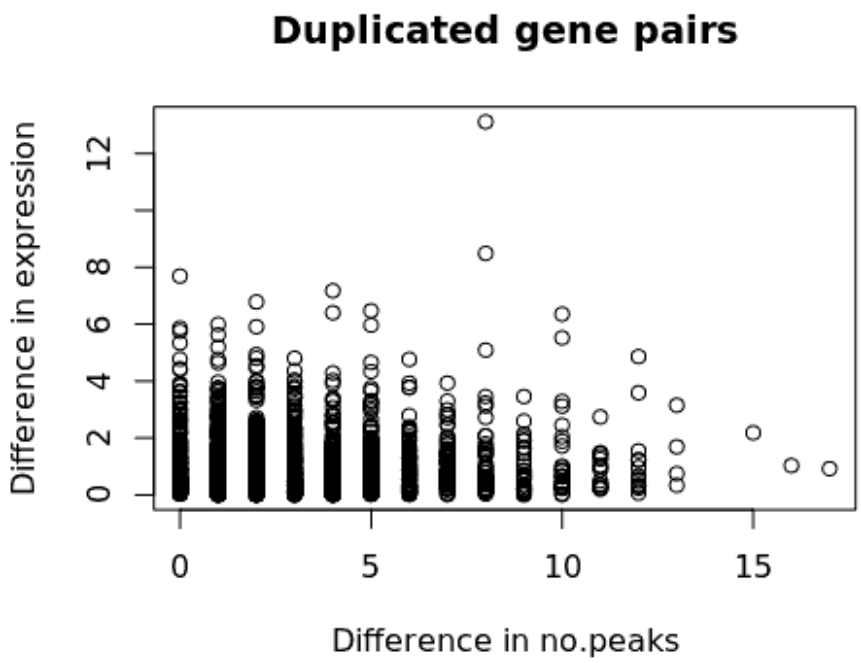
```
##
##  Pearson's product-moment correlation
##
## data:  dup_data.down$meansof4 and dup_data.down$totcov
## t = 2.0557, df = 619, p-value = 0.04023
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.003691444 0.159987716
## sample estimates:
```

```
##        cor
## 0.08234589
```

All duplicates

```
#getting all duplicates (also non-significant) with one test type and data
type


# gene.type is either dupA or dupB to only include the duplicated genes
EVE <- EVE.results.table %>% filter( test.type == 'Ss4R' & gene.type %in%
c('dupA', 'dupB') & clan.maxLRT == 'TRUE' & data.type == 'BSNsgl')
head(EVE)

## # A tibble: 6 x 12
##   clan  data.type gene.type test.type    LRT clan.maxLRT theta
thetaShift
##   <chr> <fct>     <fct>     <fct>      <dbl> <lgl>       <dbl>
<dbl>
## 1 OG00… BSNsgl    dupB      Ss4R       6.47  TRUE         5.33
0.00517
## 2 OG00… BSNsgl    dupB      Ss4R       0.869 TRUE         3.56    0.0333
## 3 OG00… BSNsgl    dupB      Ss4R       3.42  TRUE         2.04    0.942
## 4 OG00… BSNsgl    dupA      Ss4R       0.353 TRUE         3.22    4.33
## 5 OG00… BSNsgl    dupA      Ss4R       1.83  TRUE         3.41    3.75
## 6 OG00… BSNsgl    dupB      Ss4R      14.1   TRUE         2.75    0.0662
## # … with 4 more variables: shift.direction <fct>, alpha <dbl>, beta
<dbl>,
## #   sigma.sq <dbl>

#Adding column with the id of duplicate A and B
EVE$Ssal.dupA <- duplicate.clan.table$Ssal.a[match(EVE$clan,
duplicate.clan.table$clan)]
EVE$Ssal.dupB <- duplicate.clan.table$Ssal.b[match(EVE$clan,
duplicate.clan.table$clan)]

#EVE <- add.info(EVE)
#save(EVE, file="/mnt/users/ckristia/R/markdown/Data/EVE_all_dups.Rdata")
load("/mnt/users/ckristia/R/markdown/Data/EVE_all_dups.Rdata")

dup_lst <- get.dups(EVE)
dup_data <- dup_lst$dup_data
diff.cov <- dup_lst$diff.cov
diff.peaks <- dup_lst$diff.peaks
diff.expr <- dup_lst$diff.expr #difference in expression for the mean of
the four fish
diff.theta <- dup_lst$diff.theta
diff.cov.all <- dup_lst$diff.cov.all
save(dup_data,file="/mnt/users/ckristia/R/markdown/Data/duplicate_data.Rdat
a")
load("/mnt/users/ckristia/R/markdown/Data/duplicate_data.Rdata")

 plot(diff.peaks,diff.expr, ylab="Difference in expression", xlab =
"Difference in no.peaks", main="Duplicated gene pairs")
```

```
Cor.test(diff.peaks,diff.expr)
```



Boxplot

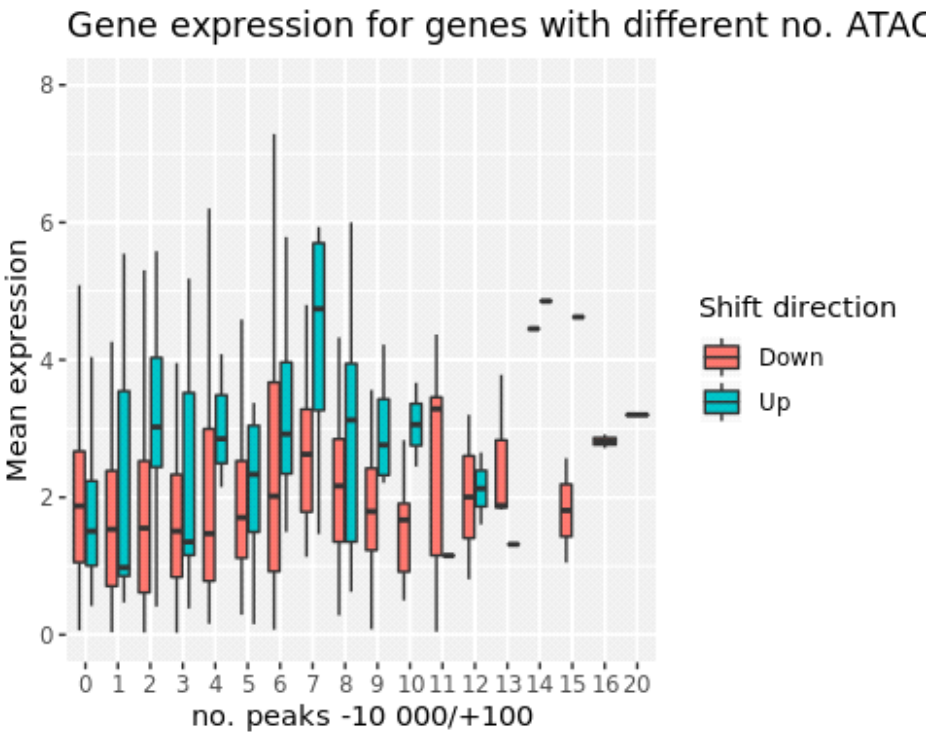
Upreguleated gene duplicates and downredulated gene duplicates

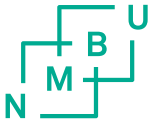
```
# Upregulated genes, downregulated genes
duplicates <- full_join(dup_data.down,dup_data.up)
```

```
## Joining, by = c("ID", "clan", "data.type", "gene.type", "test.type",
"LRT", "clan.maxLRT", "theta", "thetaShift", "shift.direction", "alpha",
"beta", "sigma.sq", "Ssal.dupA", "Ssal.dupB", "geneID", "product",
"GeneID", "Chromosome", "Start", "Stop", "Strand", "A1Peaks", "A1Cov",
"A2Peaks", "A2Cov", "A3Peaks", "A3Cov", "A4Peaks", "A4Cov", "A5Peaks",
"A5Cov", "A6Peaks", "A6Cov", "A7Peaks", "A7Cov", "A8Peaks", "A8Cov",
"A9Peaks", "A9Cov", "A10Peaks", "A10Cov", "Fish1", "Fish2", "Fish3",
"Fish4", "meansof4", "totpeaks", "totcov", "dupAID", "dupBID")
```

```
ggplot(duplicates,aes(x=as.factor(totpeaks),y=meansof4,fill=shift.direction
)) +
  geom_boxplot(outlier.shape = NA) +
  ggtitle("Gene expression for genes with different no. ATAC-peaks") +
  xlab("no. peaks -10 000/+100") +
  ylab("Mean expression") +
  scale_fill_discrete(name="Shift direction", labels=c("Down", "Up")) +
  ylim(0,8)
```

```
## Warning: Removed 14 rows containing non-finite values (stat_boxplot).
```


Gene expression for genes with different no. ATAC
Mean expression
0
2
4
6
8
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 20
no. peaks -10 000/+100
Shift direction
Down
Up



Norges miljø- og biovitenskapelige universitet
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway