R Notebook

Code

ChIP-Seq analysis of the transcription factor HOMEOBOX GENE 1 from *Arabidopsis thaliana*

Authors:

Julio Ramírez Guerrero, Julián Román Camacho, Silvestre Ruano Rodríguez, Manuel Racero de la Rosa, Rafael Rubio Ramos.

Introduction

Background

The motivation for this study lies in the need to understand how genetic mechanisms and environmental signals are linked to regulate some vital functions of plants, such as their growth rate.

Objective

The aim of the study is to identify the regions of the genome where the transcription factor will bind, in order to obtain information on the gene regulation of the *ARABIDOPSIS THALIANA HOMEOBOX GENE 1* (*ATH1*) gene.

Materials y Methods

Experimenta Design

In this task, data deposited in GEO under accession number GSE157332 will be used. The article samples are as follows:

- Chip
    - ATH1-GFP_chipseq_rep1
    - ATH1-GFP_chipseq_rep2
    - ATH1-GFP_chipseq_rep3
- Mock
    - WT_control_Ler_chipseq_rep1
    - WT_control_Ler_chipseq_rep2
    - WT_control_Ler_chipseq_rep3

For Chip samples, an anti-GFP antibody was used on samples of transgenic *Arabidopsis thaliana* plants expressing ATH1 fused to GFP. In the Mock samples, the same protocol was followed but using wild-type plants, i.e. the antibody is used without GFP being present.

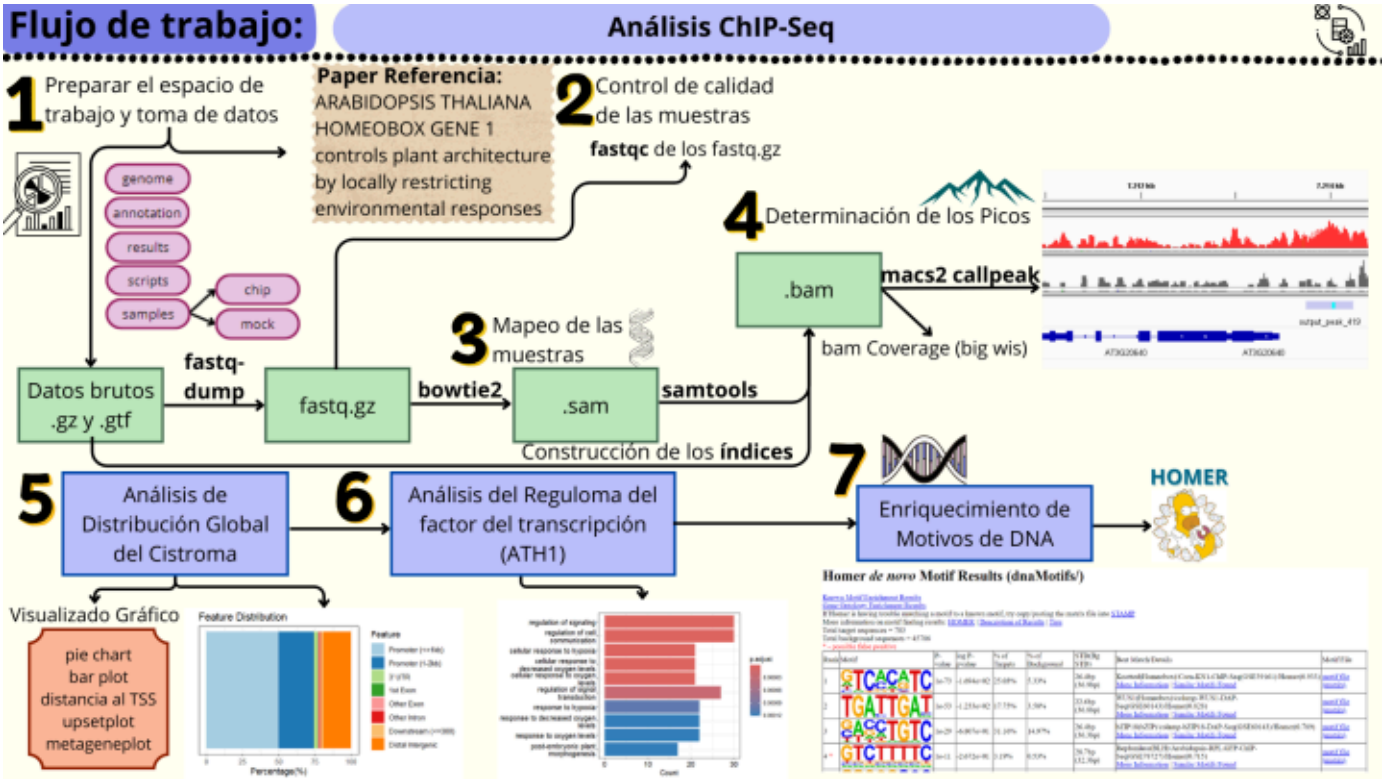
Workflow

To identify the target genes of the ATH1 transcription factor, the following workflow has been followed.

The starting data is a fastq file for each replicate, containing the sequencing results (after immunoprecipitating the chromatin according to the experimental design explained). First, we prepare the workspace and check the quality of the reads from the fastq file.

Next, the reads are mapped to the reference genome, using bowtie2. Using the result of this alignment, peak calling is performed with macs2. The aim of this step is to identify sites in the genome where a large number of reads accumulate in the ChIP samples and not in the control samples. We will consider these to be regions to which ATH1 binds.

Once the list of sites in the genome is obtained, we try to assign to each peak found the gene whose activity is regulated, using the Nearest Downstream Gene method (we look for the nearest downstream gene). Finally, we use HOMER to search for DNA motifs in the identified peaks.



Workflow for ChIP-Seq data analysis

To download the *Arabidopsis thaliana* genome and the corresponding database annotation, the following scripts were used:

```
#!/bin/bash

#SBATCH --job-name=index
#SBATCH --export=ALL
#SBATCH --output=genome_index

## Download reference genome
cd /home/omicas/grupo5/tarea2/genome
wget -O genome.fa.gz https://ftp.ebi.ac.uk/ensemblgenomes/pub/release-58/plants/fasta/arabido
psis_thaliana/dna/Arabidopsis_thaliana.TAIR10.dna.toplevel.fa.gz
gunzip genome.fa.gz

# Download annotation
cd /home/omicas/grupo5/tarea2/annotation
wget -O annotation.gtf.gz https://ftp.ebi.ac.uk/ensemblgenomes/pub/release-58/plants/gtf/arab
idopsis_thaliana/Arabidopsis_thaliana.TAIR10.58.gtf.gz
gunzip annotation.gtf.gz

## Construction of the genome index
cd /home/omicas/grupo5/tarea2/genome
bowtie2-build genome.fa index
```

For the processing of each sample, the following script was used.

```
#!/bin/bash

#SBATCH --export=ALL

SAMPLE_DIR=$1
SRR=$2
NAME=$3

## Download fastq files
cd $SAMPLE_DIR
fastq-dump --gzip --split-files $SRR

## Sample quality control and mapping of readings
if [ -f ${SRR}_2.fastq.gz ]
then
   fastqc ${SRR}_1.fastq.gz
   fastqc ${SRR}_2.fastq.gz

   bowtie2 -x ../../genome/index -1 ${SRR}_1.fastq.gz -2 ${SRR}_2.fastq.gz -S $NAME.sam
else
   fastqc ${SRR}_1.fastq.gz

   bowtie2 -x ../../genome/index -U ${SRR}_1.fastq.gz -S $NAME.sam
fi

## The generated sam is transformed into a bam (takes less space)
samtools sort -o $NAME.bam $NAME.sam
## The sam file and the original fastq are deleted
rm $NAME.sam
rm *.fastq.gz
## An index is built for the .bam file
samtools index $NAME.bam
bamCoverage -bs 5 --normalizeUsing CPM --bam $NAME.bam -o $NAME.bw
```

Finally, the script used for the call to peaks is included.

```
#!/bin/bash

#SBATCH --export=ALL

SAMPLE_DIR=$1
IP=$2
CTRL=$3
NAME=$4

cd $SAMPLE_DIR

macs2 callpeak -t $IP -c $CTRL -f BAM --outdir . -n $NAME
```

In the peak call, IP refers to the “Chip” samples, and “CTRL” refers to the “mock” samples. Thus, the mock is used to estimate the background noise. By using three ChIP files and three mock files, this tool generates a pool with the data from the different replicates before making the peak call. A peak will be considered to exist when a certain area accumulates many readings in the ChIP samples but not in the mock samples. In this step, the default parameters (q-value < 0.05 and fold-enrichment between 5 and 50) are used.

Results

Hide

```
if (!require("BiocManager", quietly = TRUE))
    install.packages("BiocManager")

BiocManager::install("TxDb.Athaliana.BioMart.plantsmart28")
BiocManager::install("org.At.tair.db")
BiocManager::install("pathview")
```

Analysis of the global distribution of the Cystrome

Hide

```
library(ChIPseeker)
library(BiocManager)
library(GenomicFeatures)
library(TxDb.Athaliana.BioMart.plantsmart28)
library(ggupset)
txdb <- TxDb.Athaliana.BioMart.plantsmart28
chip.peaks <- readPeakFile(peakfile = "output_peaks.narrowPeak",header=FALSE)
length(chip.peaks)
```

```
[1] 783
```

Peaks represent the regions of the genome where the highest binding of the protein of interest has been found. In this case we found 783 peaks where the transcription factor binds.

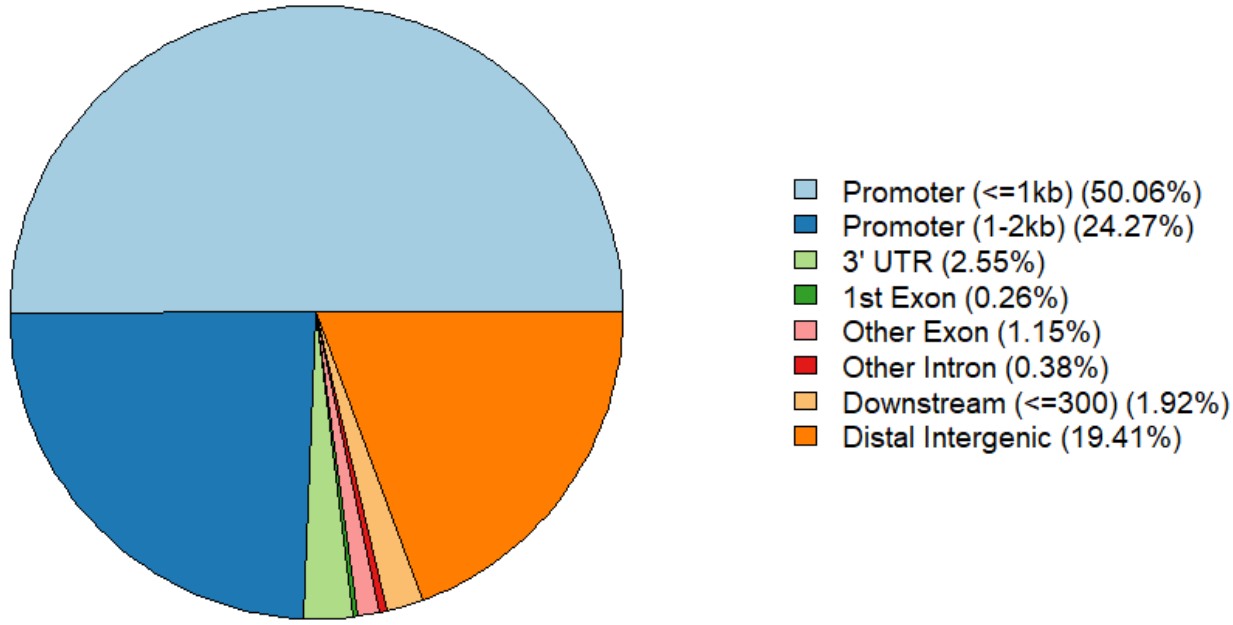
Hide

```
promoter <- getPromoters(TxDb=txdb,
                         upstream=2000,
                         downstream=2000)
chip.peakAnno <- annotatePeak(peak = chip.peaks,
                              tssRegion=c(-2000, 2000),
                              TxDb=txdb)
```

```
>> preparing features information...         2024-07-24 20:05:07
>> identifying nearest features...       2024-07-24 20:05:07
>> calculating distance from peak to TSS...  2024-07-24 20:05:08
>> assigning genomic annotation...       2024-07-24 20:05:08
>> assigning chromosome lengths          2024-07-24 20:05:14
>> done...                   2024-07-24 20:05:14
```

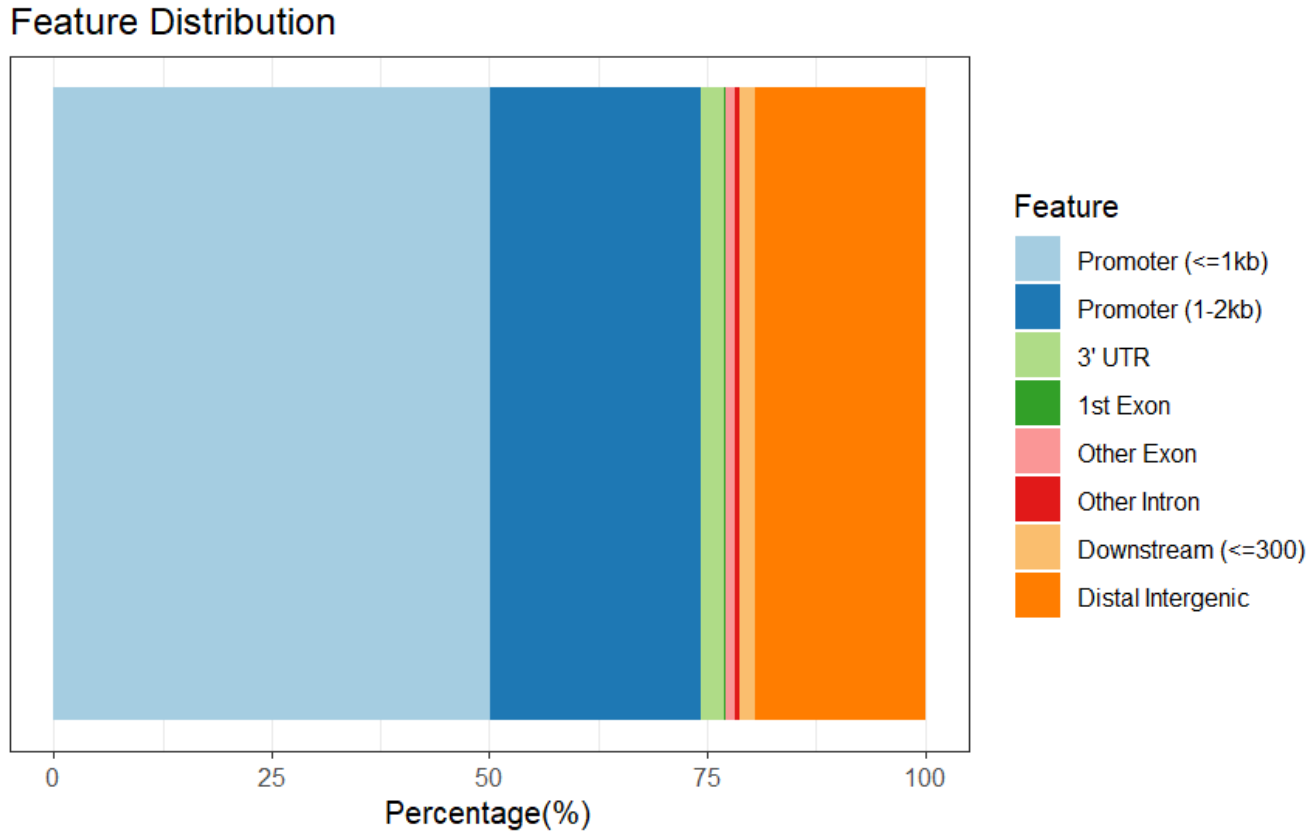
Hide

```
plotAnnoPie(chip.peakAnno)
```



Hide

```
plotAnnoBar(chip.peakAnno)
```

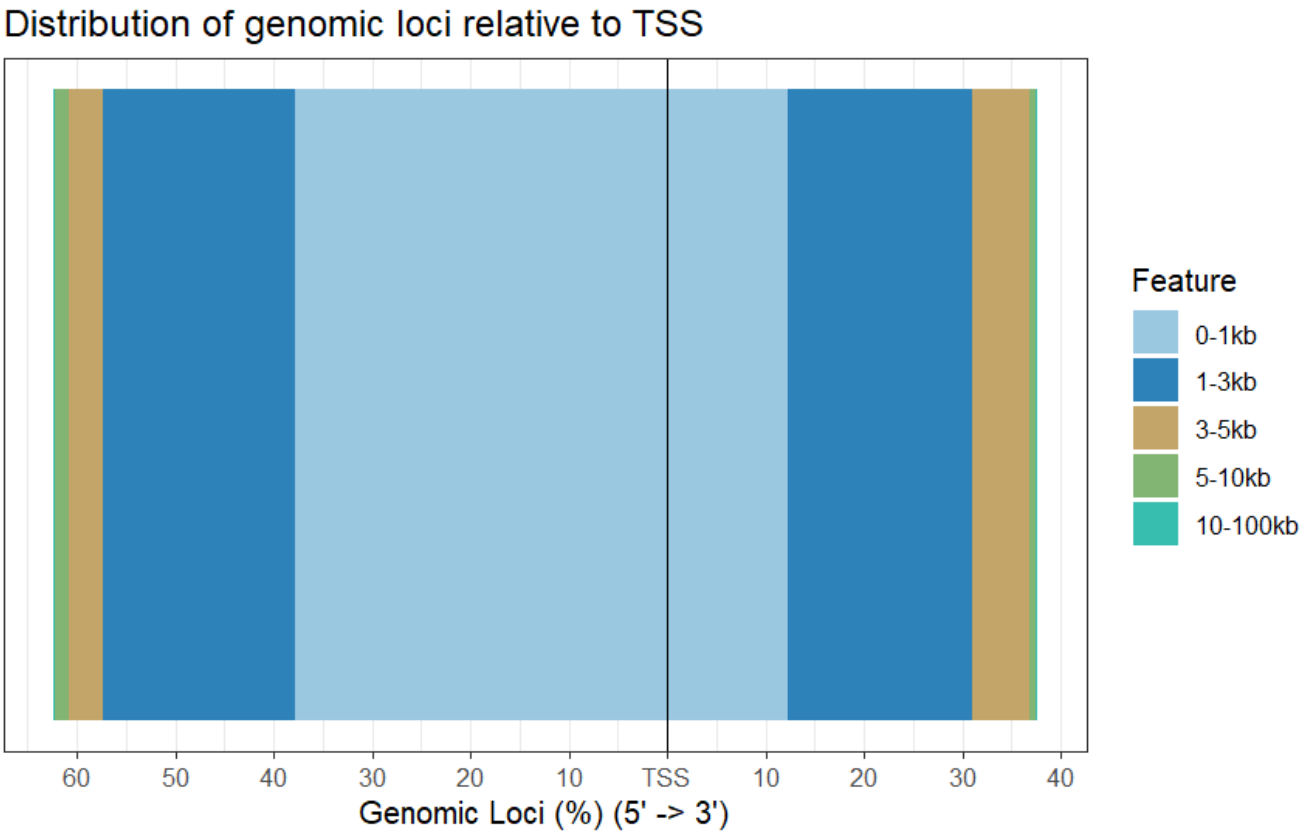


In this series of graphs we will look at how the peaks are distributed in the different gene structures, to see where the proteins of interest bind. The first graph is a pie chart showing the distribution of the peaks, where we can see that they will mainly bind to promoters at a distance less than or equal to 1kb (50.06%), then those promoters at a distance of 1-2 kb (24.27%). The rest of the elements that are at a greater distance should not

be taken into account, as we assume that it is not very reliable and there is not much certainty that they regulate the corresponding gene. The bar graph is more of the same, it shows the distributions of the distances and as we can see again, most of the peaks are found in the promoters.

Hide

```
plotDistToTSS(chip.peakAnno,
              title="Distribution of genomic loci relative to TSS",
              ylab = "Genomic Loci (%) (5' -> 3')")
```

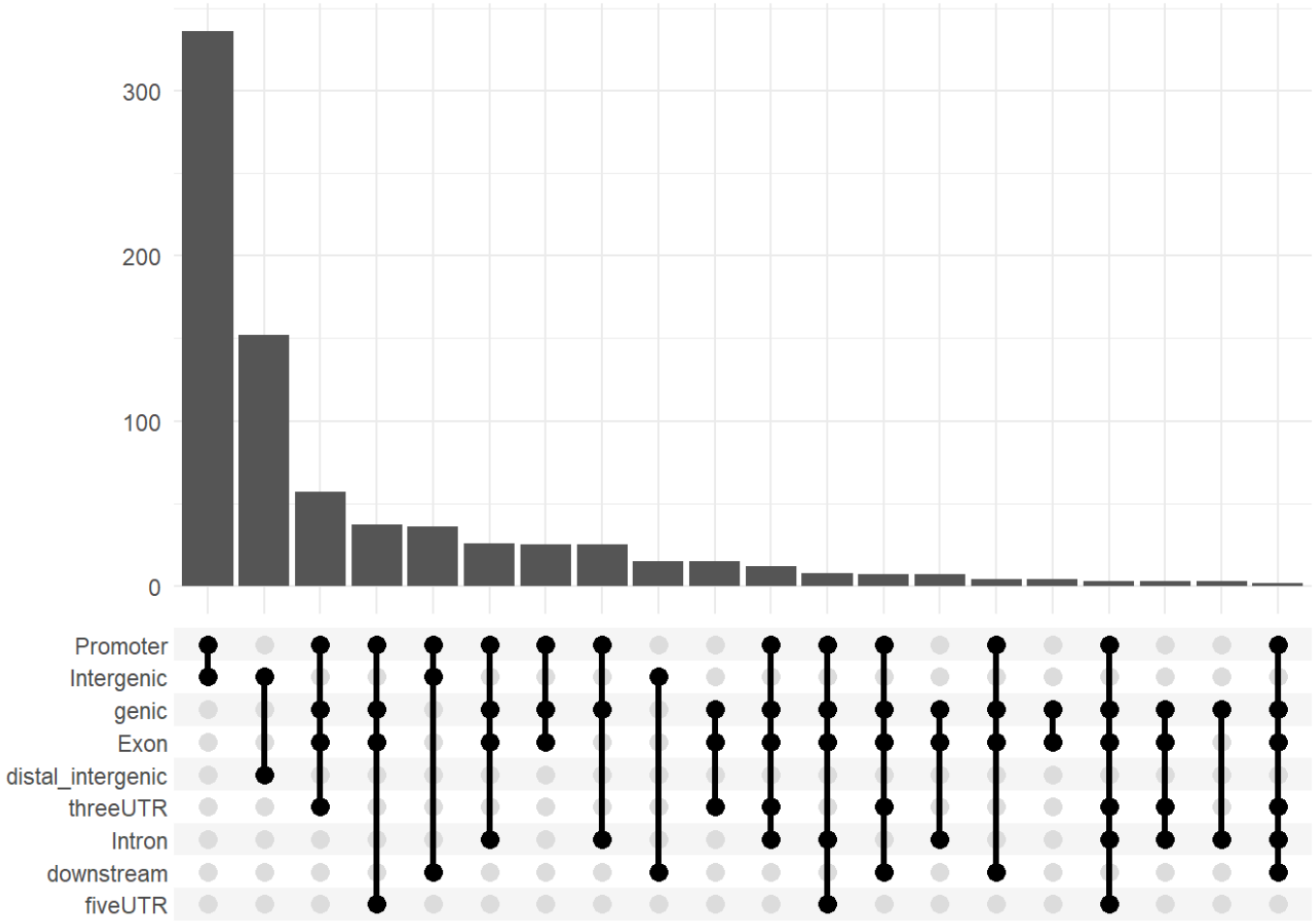


This can again be corroborated by the **distance to TSS** graph, which shows the distribution of ChIP-Seq peaks with respect to the position relative to the transcription start site (TSS). About 40% of the peaks are upstream of the TSS, between 0-1kb. The same is true for downstream, but to a lesser extent. There is also a large percentage of peaks that are 1-3kb away from the TSS.

Hide

```
upsetplot(chip.peakAnno)
```

```
Error en base::nchar(wide_chars$test, type = "width"):
  entrada en evaluacion: recursivo por defecto o problemas anteriores?
```

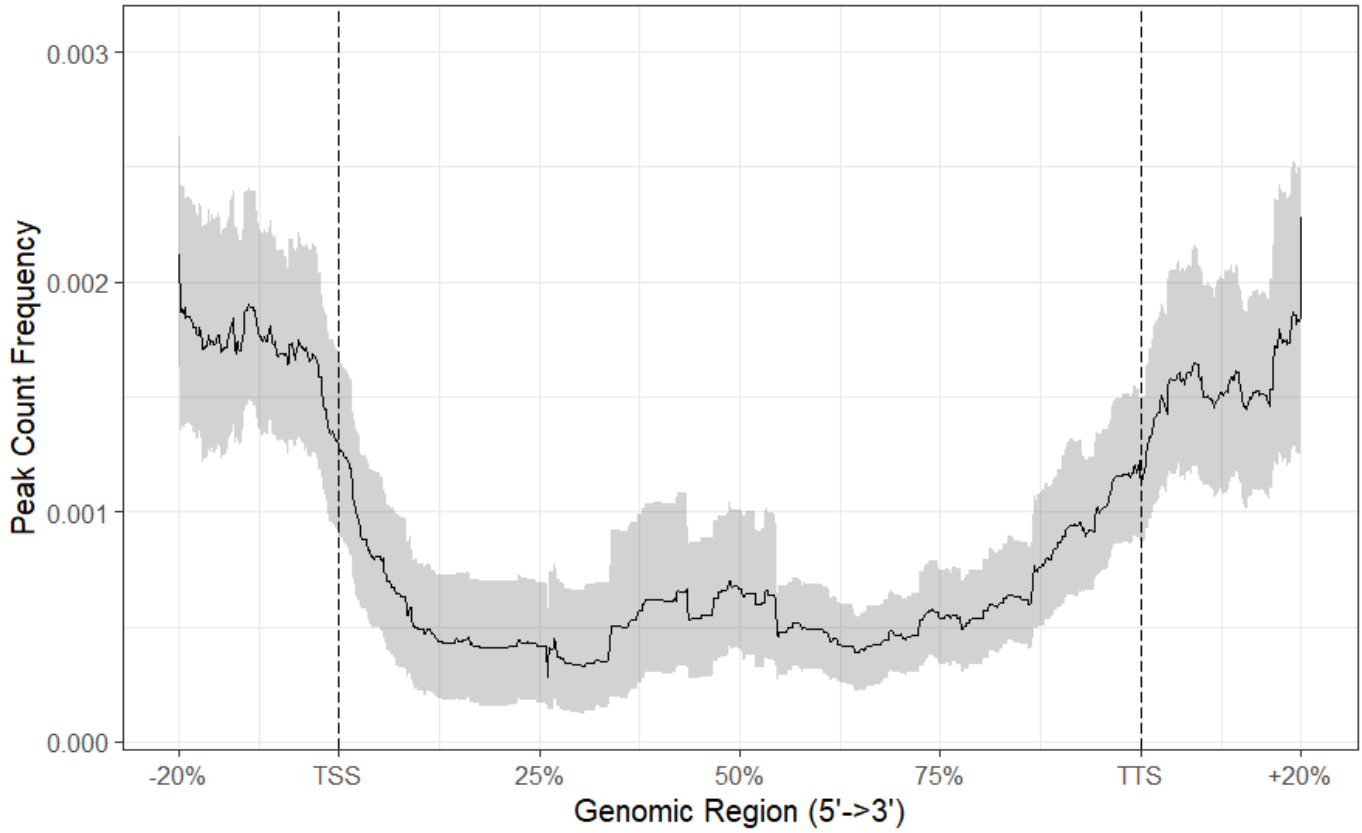


The **upsetplot** graph shows where the peaks overlap and their frequency. It is observed that in most peaks (around 300) the promoters are located in areas rich in intergenic regions, or for example, the promoters also overlap with some gene areas, such as exons and 5'UTRs (around 50).

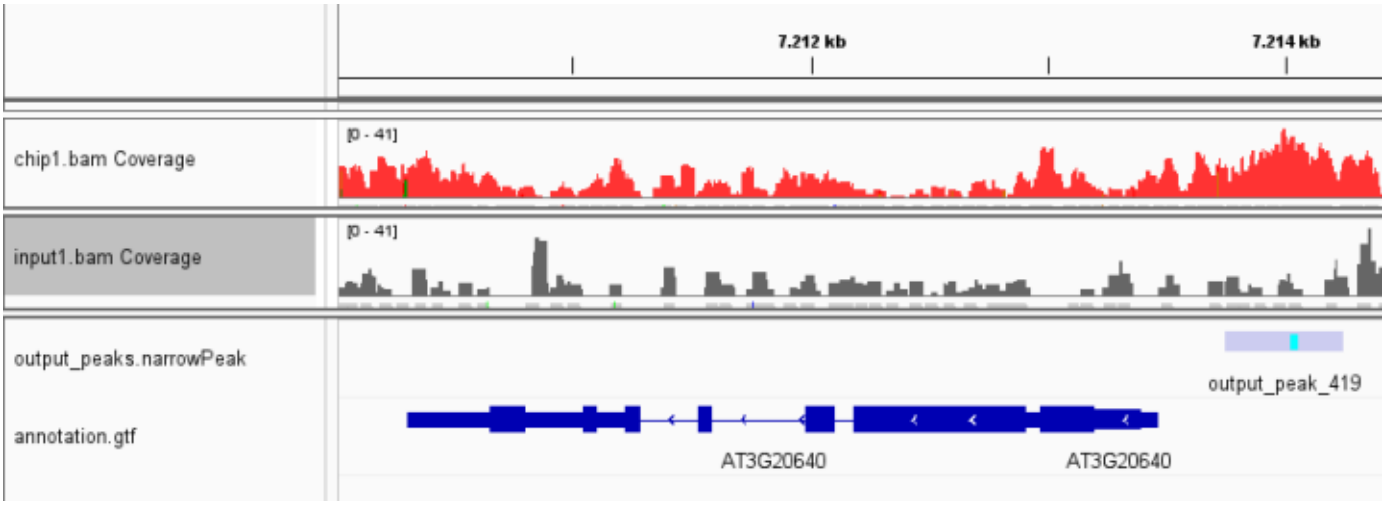
Hide

```
plotPeakProf2(peak = chip.peaks, upstream = rel(0.2), downstream = rel(0.2),
              conf = 0.95, by = "gene", type = "body", nbin = 800,
              TxDb = txdb, weightCol = "V5",ignore_strand = F)
```

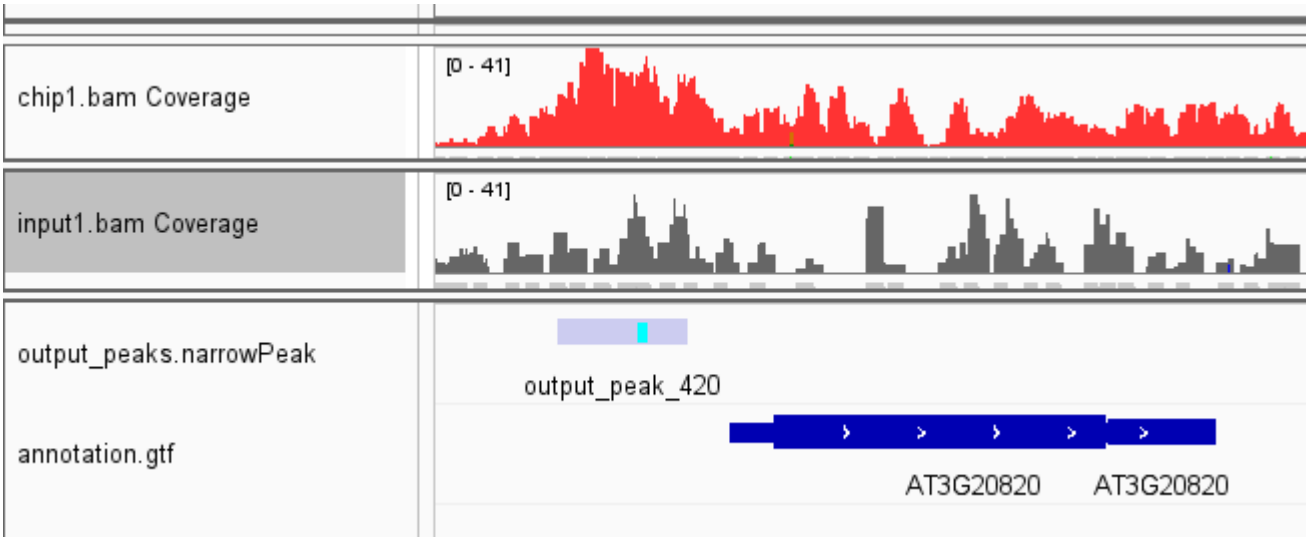
```
>> binning method is used...2024-07-24 20:05:31
>> preparing body regions by gene... 2024-07-24 20:05:31
>> preparing tag matrix by binning...  2024-07-24 20:05:31
>> preparing matrix with extension from (TSS-20%)~(TTS+20%)... 2024-07-24 20:05:31
>> 40 peaks(7.155635%), having lengths smaller than 800bp, are filtered... 2024-07-24 20:05:3
1
>> Running bootstrapping for tag matrix...       2024-07-24 20:05:35
```



Finally, in the **metageneplot**, the peak upstream of the TSS could indicate a strong binding of the protein to the promoter region of the gene, which is what is observed in the other plots. We also found other peaks downstream of the TTS at the end of transcription, perhaps related to post-transcriptional regulation. And the intermediate zone between TSS and TTS has no prominent peaks indicating the absence of the protein in this transcriptional region.



First IGV screenshot.



Second IGV screenshot.

In the following images we see a series of peaks, which correspond to our output_peak. If we look at the sequence we see that there is a high concentration of reads in the Chip1 sample that does not appear in the Mock1. These peaks are indicative that our transcription factor (ATH1) binds to that region. At the bottom of the image we have our genes, the lines correspond to the introns, the squarer areas are the exons and the arrow indicates the direction in which it is transcribed.

It is important to mention that this peak is located upstream of the gene, so it would be the promoter of the gene.

Regulome analysis of the corresponding transcription factor or epigenetic mark.

After the graphical analysis, we saved in a txt file called “target_genes.txt” the genes found in the promoters, which are the ones we considered important for the distribution of peaks.

Hide

```
chip.annotation <- as.data.frame(chip.peakAnno)

target.genes <- c(chip.annotation$geneId[chip.annotation$annotation == "Promoter (<=1kb)"],
                  chip.annotation$geneId[chip.annotation$annotation == "Promoter (1-2kb)"])
length(target.genes)
```

```
[1] 582
```

Hide

```
write(x = target.genes,file = "target_genes.txt")
```

Of the 783 peaks found in the IGV file, we will consider those that are at least 2 kb away from the promoter of a downstream gene. The percentage of peaks that fulfil this characteristic corresponds to 74.31% of all peaks found. Thus, we consider a total of 582 genes that correspond to DNA reads where transcription factors bind.

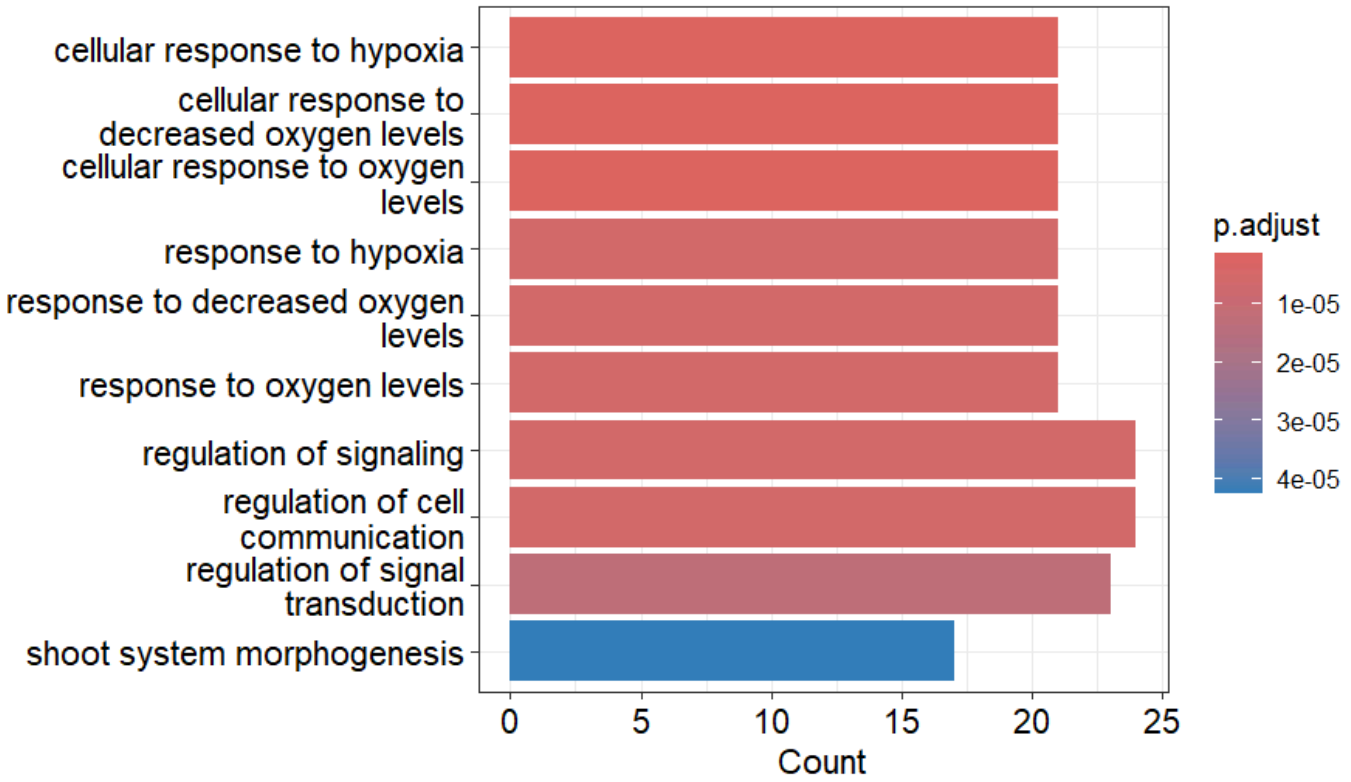
ATH1 target genes. (target_genes.txt)

Hide

```
library(clusterProfiler)
library(org.At.tair.db)
library(enrichplot)

chip.enrich.go <- enrichGO(gene = target.genes,
                           OrgDb         = org.At.tair.db,
                           ont           = "BP",
                           pAdjustMethod = "BH",
                           pvalueCutoff  = 0.05,
                           readable      = FALSE,
                           keyType = "TAIR")




chip.enrich.kegg <- enrichKEGG(gene  = target.genes,
                               organism = "ath",
                               pAdjustMethod = "BH",
                               pvalueCutoff  = 0.05)
```

Hide

```
barplot(chip.enrich.go,showCategory = 10)
```

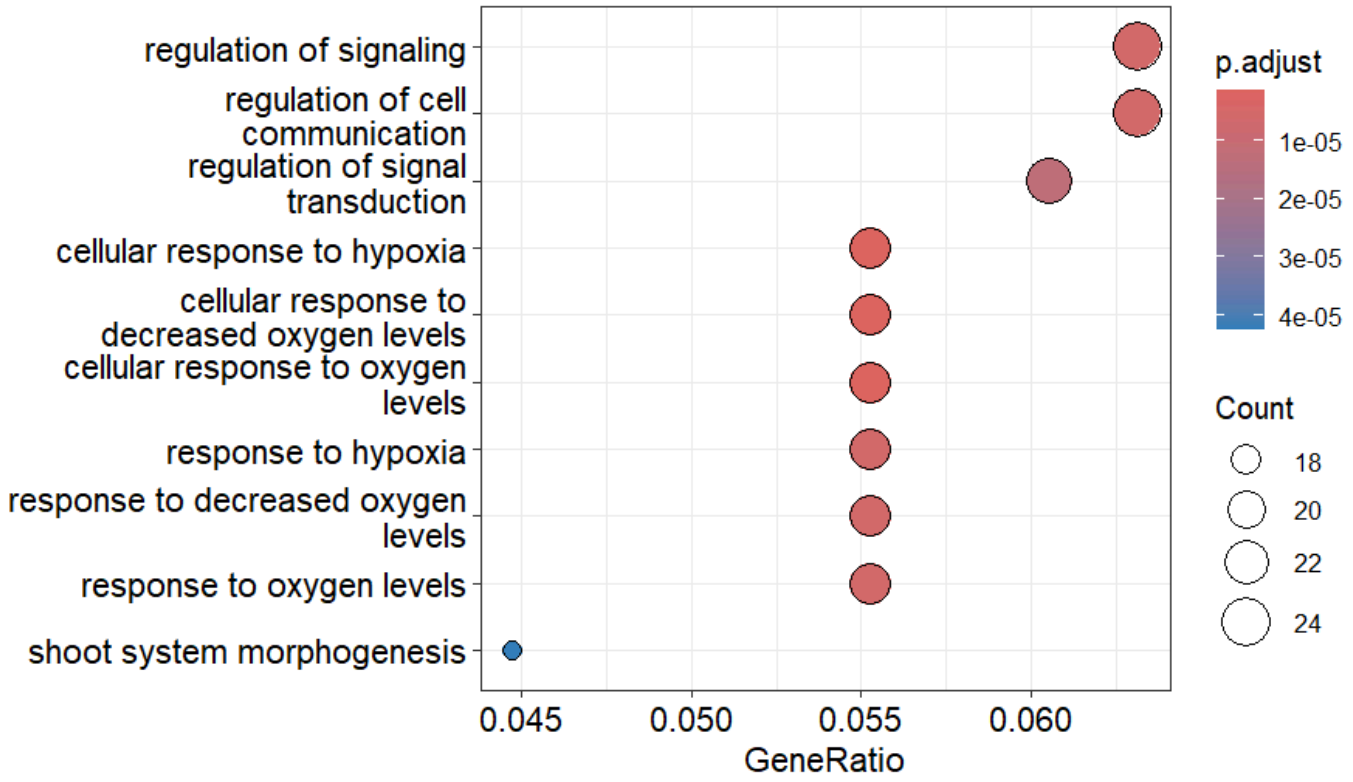


In the first barplot, the length of the bars indicates the magnitude of the enrichment for the associated function of each group of genes, so that the longer the bar length, the stronger the enrichment of the genes for that process or function. The p-value of each bar indicates whether the enrichment is significant or not; in this case, red indicates higher statistical significance (lower p-value).

In this context, when analysing the barplot graph shown, there is high statistical significance that the selected peaks are transcription factor binding sites for genes that are involved in cell signalling and communication processes.

Hide

```
dotplot(chip.enrich.go,showCategory = 10)
```

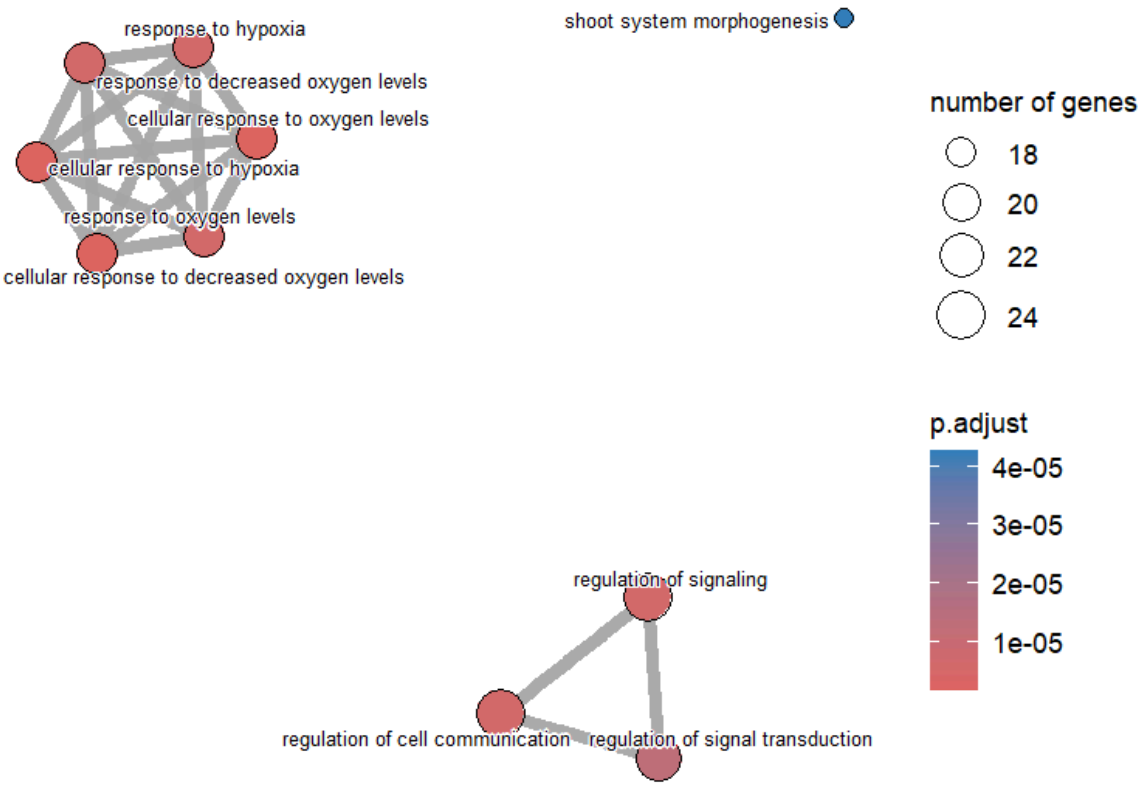


This dotplot indicates the ratio of enriched genes to the total number of genes associated with a specific function. A high GeneRatio indicates that a significant proportion of genes associated with a specific function are present in the set of genes close to the peaks.

In our case, the genes with the highest GeneRatio are those associated with **cell signalling and cell communication regulation** functions, as well as those related to **regulation of oxygen levels** as in the previous graph. This set of genes has a low p-value, indicating that the observation is unlikely to be due to chance. In summary, the combination of a high GeneRatio and a low p-value reinforces the evidence that these genes related to cell signalling and communication are biologically relevant to our gene set, and that this association is not simply the result of random fluctuations.

Hide

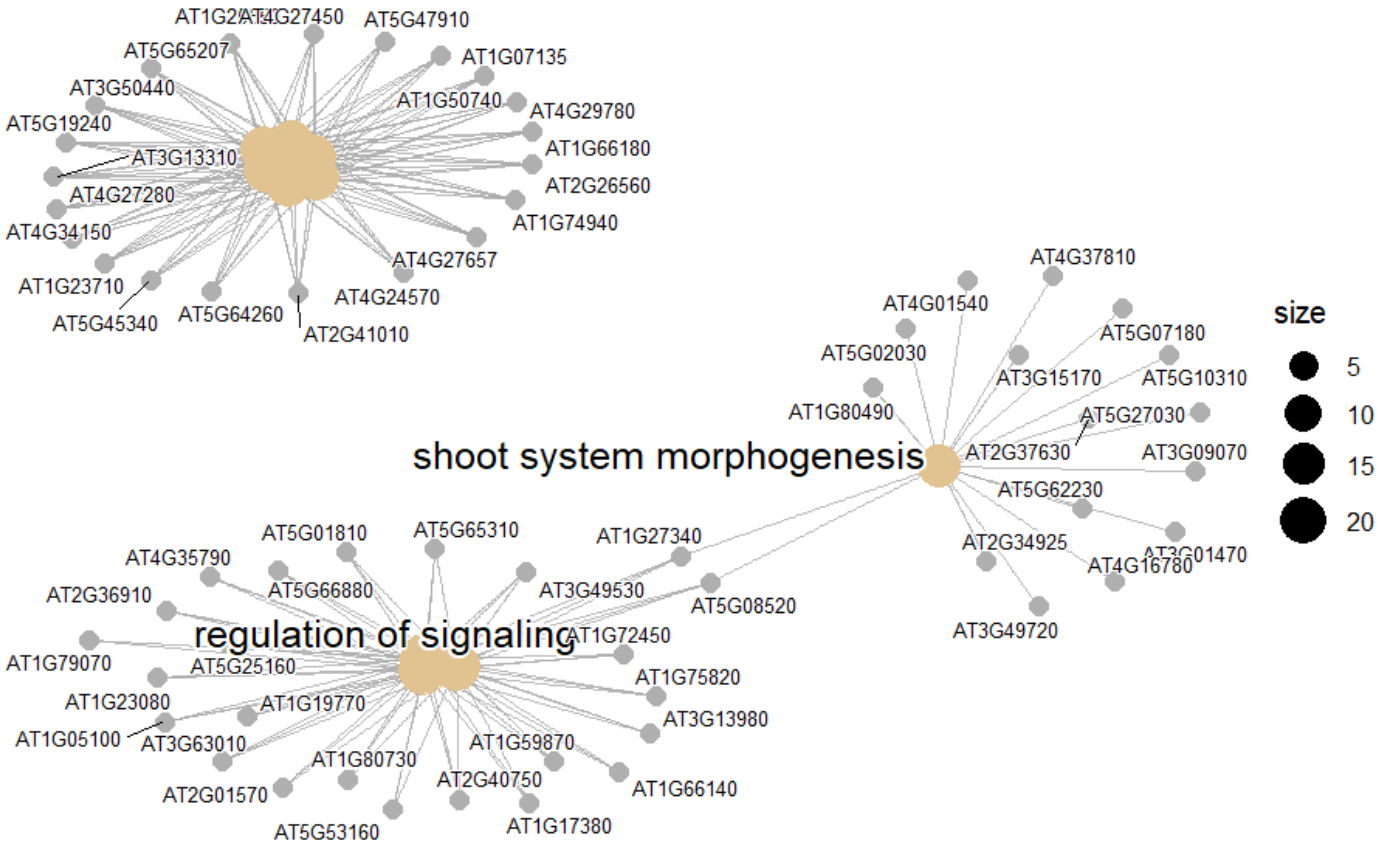
```
emapplot(pairwise_termsim(chip.enrich.go),showCategory = 10,cex_label_category=0.5)
```



Thirdly, the emapplot graph is used to visualise similarity relationships between the different genes under study. The graph shows relationships between cell regulation and cell communication genes, forming a triangular figure, and a hexagon is also displayed, indicating similarity relationships between genes involved in cell response as a function of oxygen levels in the cell. Finally, there is a set of genes set apart, represented as an isolated point, related to developmental functions of the plant after the embryonic stage. Notably, this single gene contains the highest p-value and the lowest number of genes.

Hide

```
cnetplot(chip.enrich.go,showCategory = 10,cex.params=list(gene_label=0.5))
```



In this Cnetplot graph we can see the grouping of the enriched genes in 3 very well differentiated clusters, which coincide with the previous graph. The cluster in the upper right-hand zone is related to the post-embryonic morphogenesis functions of the plant. Meanwhile, the central cluster shows genes that influence cell regulation and communication. Finally, the bottom cluster reflects those genes associated with the regulation of oxygen levels, such as response to hypoxia, cellular response to oxygen depletion, etc.

Hide

```
library(pathview)
df.chip.enrich.kegg <- as.data.frame(chip.enrich.kegg)
head(df.chip.enrich.kegg)
```

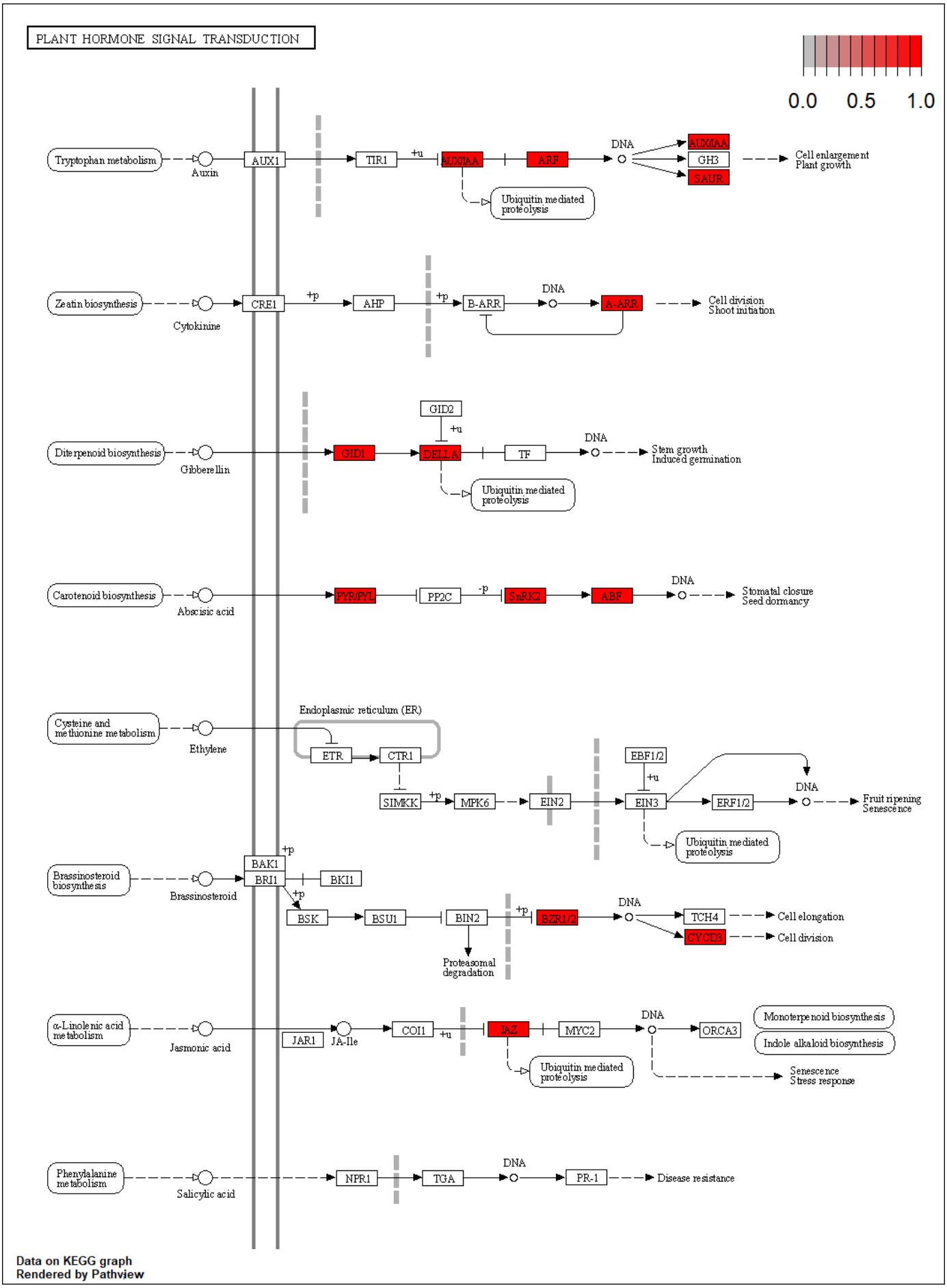
	category <chr>	subcategory <chr>	ID <chr>
ath04075	Environmental Information Processing	Signal transduction	ath04075
ath04016	Environmental Information Processing	Signal transduction	ath04016

2 rows | 1-4 of 11 columns

Hide

```
pathview(gene.data = target.genes,
         pathway.id = "ath04075",
         species = "ath",gene.idtype = "TAIR")
```

```
[1] "Note: 57 of 551 unique input IDs unmapped."
```



Hormone signalling pathway in Arabidopsis.

It is important to note from the image above that the transcription factor under study regulates those genes coloured red, which is not the case with uncoloured genes, where it shows no regulation at all.

Enrichment of DNA motifs at binding sites

Among the known motifs, the most enriched is related to Replumless (BLH), which is a BELL (plant developmental regulation) transcription factor related to ATH1.

It can be seen that among the most enriched motifs, the most frequently repeated motifs belong to the Homeobox family, which is characteristic of transcription factors involved in the regulation of cell development and differentiation. Within the motifs associated with the Homeobox (HB) family, a rather repeated motif, AATGATTG, associated with the ATHB transcription factors that are more specifically involved in **leaf and root development, stress response and growth regulation,** is clearly visible.

Among the unknown motifs, the second most enriched is TGATTGAT, largely matching the sequence of the most repeated one seen above. Furthermore, according to the Homer tool it has the best match to the Homeobox family, suggesting that ATH1 , binds to this motif.

The known motifs with the highest enrichment are shown below.

Clasifcation	Known Motifs	Name	P-value
1	-GTC–ATCA	Replumless(BLH)/Arabidopsis-RPL.GFP	1e-88
2	GA-G-GAC-GG-	Knotted(Homeobox)/Corn-KN1	1e-71
3	TGACGTCAC-	FEA4(bZIP)/Corn-FEA4	1e-23
4	CA-TCATTCA	WUS1(Homeobox)/colamp-WUS1	1e-21
5	AATGATTG	ATHB5(HB)/colamp-ATHB5	1e-18
6	AATGATTG	ATHB7(Homeobox)/col-ATHB7	1e-17
13	AATGATTG	ATHB6(Homeobox)/col-ATHB6	1e-8

En la tabla se muestran los motivos desconocidos más enriquecidos.

Clasification	Unknown Motifs	Best Match	P-value
1	GTCACATC	Knotted(Homeobox)/Corn-KN1	1e-73
2	TGATTGAT	WUS1(Homeobox)/colamp-WUS1	1e-53
3	-ACCTGTC	bZIP18(bZIP)/colamp-bZIP18	1e-29
4 (posible falso positivo)	GTCTTTTC	Replumless(BLH)/Arabidopsis-RPL.GFP	1e-11

Conclusions

The most relevant conclusions of our study are the following:

- Transcription factors of the Homeobox family (including ATH1) repeatedly bind to a well-known TGATTG motif, as shown by Homer analysis.
- The ATH1 transcription factor regulates a multitude of vital plant processes, such as regulation of cell signalling and cell communication, regulation of oxygen levels and leaf and root development, post-embryonic morphogenesis, among others.

Below are the common conclusions drawn from our study and the article:

- Both the findings in the article and our results agree that ATH1 regulates genes involved in cell signalling and communication.
- ATH1 regulates the DELLA protein (which is regulated by gibberellins), which is closely related to plant morphology and its regulation.
- The most repeated motif among the enriched motifs, both in our study and in the paper, is TGATTG , which is in agreement with the results extracted from the analysis with the Homer software.

A possible experiment to validate our results would be to compare a control plant and an ath1- mutant to check the type of plant growth in response to hormones, such as auxins. The control plants without hormone treatment grow following a rosette-like structure. However, the ath1- mutant plants grow following a default caulescent morphology. To our astonishment, after several days in which the plant structure was already established, auxins were inoculated to our mutant, resulting in a morphological change that led to a shortening of the plant internodes, resulting in a rosette shape.

References

Bencivenga, S. et al. (2021). ARABIDOPSIS THALIANA HOMEOBOX GENE 1 controls plant architecture by locally restricting environmental responses. *Proc Natl Acad Sci U S A.118(17)*. PMID: 33888582.