

BMEG 400F Final Project

- [Introduction](#)
- [Methods](#)
- [Results](#)
- [Conclusion](#)
- [References](#)
- [Code](#)
- [Appendix](#)

Introduction

Wetland plants such as rice often experience varying degrees of hypoxia: a type of abiotic stress due to partial depletion of oxygen from reduced oxygen diffusion in water (1). Thus, these crops' ability to germinate and grow when submerged in water is important for their survival. Many studies have been conducted to investigate the genetic basis of the anaerobic development of various plants (2–4). One such study identified *Arabidopsis* genes related to the plant's response to submergence stress. In particular, it looked at genes regulated by the WRKY33 transcription factor through chromatin immunoprecipitation (ChIP)-Seq and attempted to identify genes in *Arabidopsis thaliana* that are regulated by WRKY33 while experiencing stress from submergence in water (5).

We chose to conduct a reanalysis on this study using the computational methods we have become familiar with through our coursework thus far. In the study, WRKY33 overexpression transgenic *Arabidopsis thaliana* plants in the Columbia ecotype background were submerged underwater and compared to the wild-type Columbia ecotype of *Arabidopsis thaliana* to confirm their higher tolerance to submergence, which was a finding made in a previous paper. The WRKY33 cDNA was used for constructing the overexpression vector. It was also cloned under the 35S promoter for overexpression, with an in-frame FLAG epitope tag to facilitate future ChIP experiment. The submergence response of these plants was confirmed by several hypoxia response markers. These FLAG-WRKY33 overexpression plants had higher submergence tolerance than the control Columbia ecotype, as well as lower levels of Malondialdehyde (MDA), a marker of oxidative stress which is expressed when plants are under submergence stress. In addition, RNA extracted from the FLAG-WRKY33 overexpression plants had higher relative expression levels of WRKY33 than the control. Overall, the 35S-FLAG-WRKY33 transgenic plants

behave the same as the WRKY overexpression ones, suggesting that the FLAG tag is not affecting the function of native WRKY33.

Following this, the 35S-FLAG-WRKY33 plants were grown and submerged in water, and the leaves of these plants were used to conduct Chromatin Immunoprecipitation (ChIP) experiments. After cross-linking the DNA to the proteins, cellular extraction was treated with sonication to break up the DNA. The anti:FLAG antibody was then used to pull down DNA fragments that are cross-linked to FLAG-WRKY33.

The ChIP sequencing was then conducted by an external company, and the raw data produced is publicly available. Due to course limitations, we did not conduct the experimental portions of the study and instead used the data that the researchers obtained for data analysis. Experimental decisions that may have impacted downstream analysis are described in the results of this paper.

The goal of this reanalysis was to identify any flaws that may have been looked over during the original analysis, and to learn more about genomics analysis in a research setting. Through performing reanalysis, researching, and adjusting certain methodologies, we hoped to provide a comprehensive review of the study's analysis, and along the way gain a better basis in genomics.

Methods

To confirm the study's analysis and results, a reanalysis was first conducted, following the paper's methods. QC was performed using FASTQC to confirm the quality of the raw sequence data provided by the authors. Following this, initial analyses involved aligning ChIP-Seq reads with Bowtie2 and inspecting the alignments further using the Integrative Genomics Viewer (IGV) tool and the MACS2 toolbox we used in Assignment 5. The *Arabidopsis thaliana* TAIR10 genome was used as the reference for alignment, and the bowtie index files are available from Illumina (6).

Next, The alignments were sorted and filtered to only contain uniquely mapped reads to remove ambiguously mapped reads. Bigwig files were created to visualize read coverage within IGV. The distribution and coverage of the peaks was also analyzed in R. Using the peaks identified using MACS2, motif analysis was conducted to identify the genomic sequences that bind to WRKY33. Both MEME (Multiple EM for Motif Elicitation)- ChIP (7) and Regulatory Sequence Analysis Tools (RSAT) (8) were used to identify motifs. The peaks identified previously were also used to categorize ontological genes and gain a better understanding of the gene and gene product attributes related to WRKY33. We

performed this using the online gene ontology enrichment analysis tool GOEAST (9). The study did not specify how they performed gene ontology, but used p-values and false discovery rates for the annotation were generated using the default Fisher's Exact Test and Benjamini-Yeuktieli methods.

Results

Initial quality control measures were determined from the fastQC (10) of the two input sequence files. The average quality per read for both files was high (around 36 for both), and the sequence length distribution, adapter content, and per base N contents both passed the QC check. However, the per sequence GC content did not pass, and neither did the per-base sequence content for one of the input files. This could be due to the overrepresentation of certain sequences. For the input file, two sequences were overrepresented, and for the ChIP-SEQ fq file six sequences were overrepresented. This could be a signal of contamination, or the overexpression of certain sequences related to WRKY33. The fastqc html files can be found in our github repository. To improve the data quality, experiments would need to be redone and methods might need to be adjusted. The full fastQC files can be found in our github repository, and the summaries can be found in Appendix A.

Figure 1 shows the aligned peaks of the data. The top bar shows the filtered and aligned input file using the whole genome obtained from the transgenic plants. The second bar shows the filtered ChIP peaks, and the third bar shows the aligned peaks' peak regions obtained using the input file as control. From the peak visualization it can be seen that the transcription factor binding sites are quite limited, and thus the number of target genes for the WRKY33 transcription factor is small under our analysis. This may have occurred because our filtering differed from the methods used by the study.

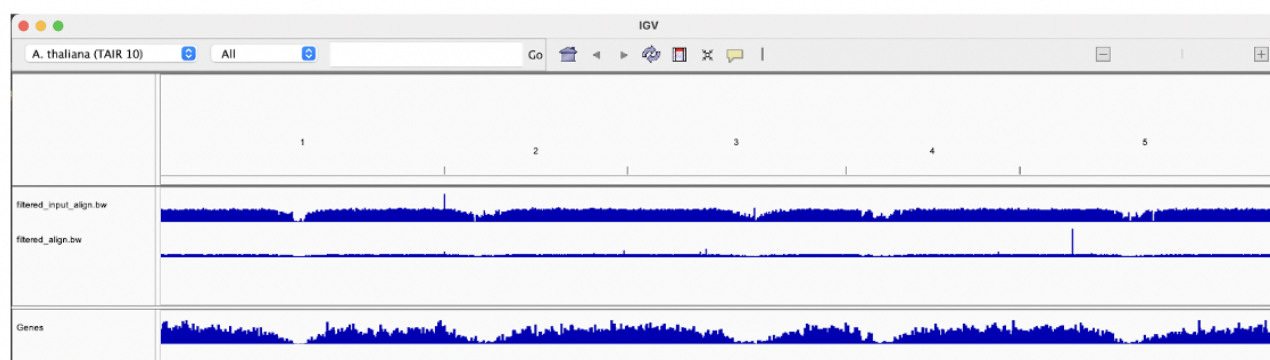


Figure 1. Peak visualization using IGV

The motifs should show the transcription factor target sequences which may be related to WRKY33. Figure 2 shows the first pair of motifs obtained using RSAT. These differ from the top motifs identified in the study. This could be due to differences in software, or additionally our upstream filtering may have impacted the input data. We also tried obtaining results from MEME-ChIP which inherently also includes results from a similar technique known as STREME. The results from MEME-ChIP were not coherent with the paper or RSAT because it was run in “classic” mode where it just tries to find recurring fixed length patterns without differential enrichment. To be able to perform differential enrichment we were required to provide a control sequence, similar length as input, to which it would compare

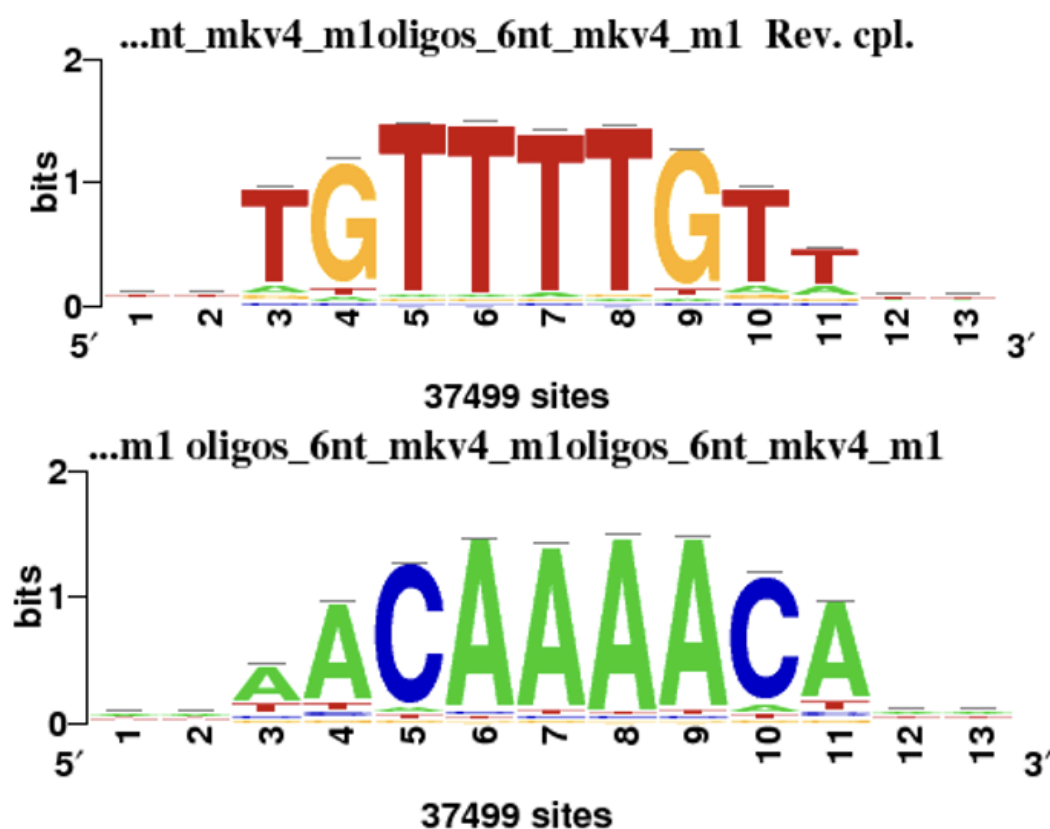


Figure 2. The first pair of motifs obtained using RSAT

The motifs should show the transcription factor target sequences which may be related to WRKY33. Figure 2 shows the first pair of motifs obtained using RSAT. These differ from the top motifs identified in the study. This could be due to differences in software, or additionally our upstream filtering may have impacted the input data. We also tried obtaining results from MEME-ChIP which inherently also includes results from a similar technique known as STREME. The results from MEME-ChIP were not coherent with the paper or RSAT because it was run in “classic” mode where it just tries to find recurring

fixed length patterns without differential enrichment. To be able to perform differential enrichment we were required to provide a control sequence, similar length as input, to which it would try to find the motifs which are enriched in input relative to control sequence. We simply did not have access to the control sequence in this case. Furthermore, in efforts to create our own control sequence we used samtools amplicon clip to get fasta sequence using peaks obtained from macs2, however, we were not able to get significant results when the ChIP-seq, input whole genome and the reference genome was used to obtain the fasta sequence.

Figure 3 shows the distribution of the ChIP peaks we obtained from the data. Most of the peaks fall close to start codons ($\leq 1\text{kb}$), which aligns with what we would expect as well as the results from the study. This implies that the transcription factor binds to these regulatory components, and since they are close to the start codon, they are more likely to initiate its transcription and contribute to the regulation of the gene (11).

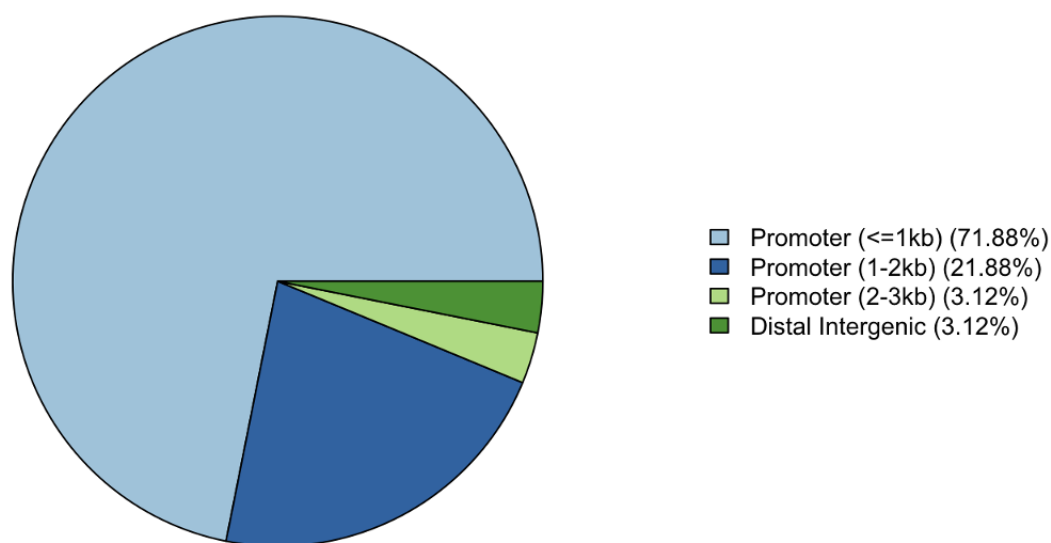


Figure 3. Distribution of the ChIP peaks

Figure 4. Shows the coverage of the ChIP peaks identified. As previously mentioned, data processing differences such as in filtering could have contributed to the differences in the number and distribution of the ChIP peaks. More peaks were filtered out during our analyses, but the locations of the peaks in our analysis are similar to those found in the study.

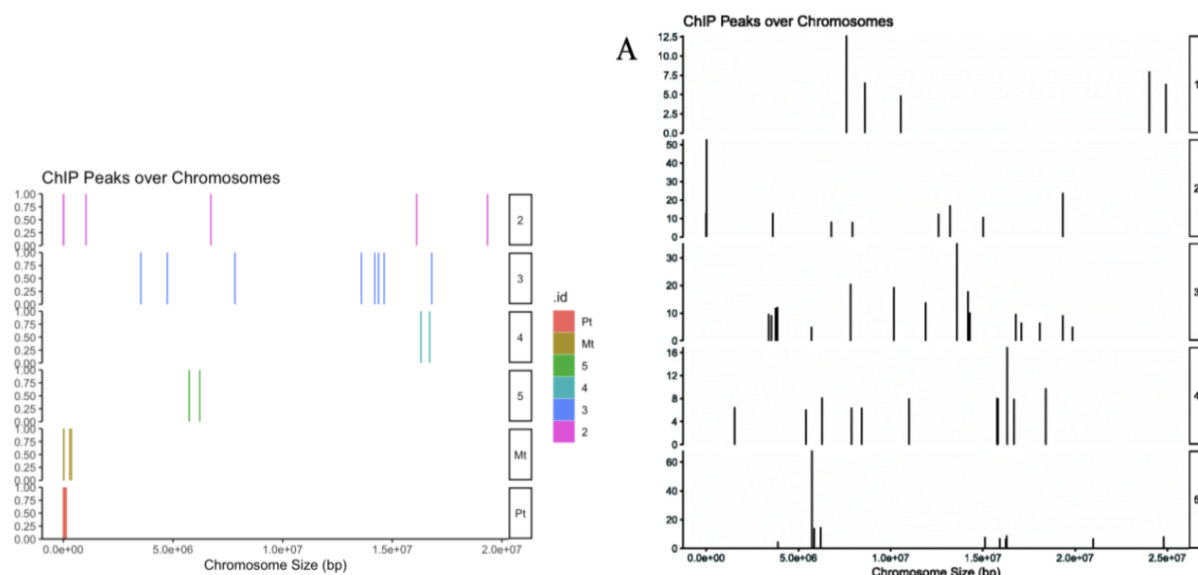


Figure 4. ChIP peaks coverage plot: left shows our results, right shows results from the study

Upon manually performing gene ontology, we were unable to identify significant genes. This differed from the study, which discovered 61 GO categories. This could be due to an upstream mistake we made that filtered out important information. In addition, the experiment data did not pass all quality control metrics, and thus may not have been representative of the case being studied.

To verify the significance of the TC box that the study identified during motif analysis, a qPCR test was performed. The researchers then used the results to conclude that the WRKY33 transcription factor regulated four of the genes containing the TC box. However, upon inspection of the results, the difference in expression levels between the control and the WRKY33 transgenic plants was not large (less than twofold) for three of the four genes. Typically, when analyzing qPCR results, a difference of more than two-fold is needed to conclude significance (12). In the future, experiments can be designed to affirm that the analysis performed by us and in the paper were accurate. RNA-SEQ could be performed on the transgenic plants, and expression of all the target genes from both sets of analysis could be compared.

Conclusion

Though we were unable to make any conclusions based on the analyses performed, we were able to learn more about the process of conducting genomic analysis unguided by the prewritten templates provided for assignments. This self-directed learning was very beneficial in learning the logic behind each of the steps performed throughout analysis. Quality of data, significance metrics, and filtering parameters are important considerations

that have a large impact on a study's results, as demonstrated by our failure to replicate the analyses performed by the researchers of the study being analyzed. It is also to be mentioned that providing us with all the data would have been helpful, in terms of performing sanity checks after each step. By more data we mean by providing us specific details as to what files were used where, and what specific flags were used. Providing more details on running MEME-ChIP would have been most helpful as it would have given us direction as to how to create our own control sequence for differential analysis. Though we attempted to follow their methods, our gene ontology was also different from theirs.

Next time, it could be useful to reattempt analysis using different metrics for certain steps. The authors may have failed to include certain metrics or steps, which yielded different ChIP peak results after filtering. This project was a good demonstration of how informatics studies can begin at the same raw data but differ in their results.

References

1. Lerant AA, Hester RL, Coleman TG, Phillips WJ, Orledge JD, Murray WB. Preventing and treating hypoxia: Using a physiology simulator to demonstrate the value of pre-oxygenation and the futility of hyperventilation. *International Journal of Medical Sciences*. 2015;12(8).
2. Tang H, Bi H, Liu B, Lou S, Song Y, Tong S, et al. WRKY33 interacts with WRKY12 protein to up-regulate RAP2.2 during submergence induced hypoxia response in *Arabidopsis thaliana*. *New Phytologist*. 2021;229(1).
3. Kennedy RA, Rumpho ME, Fox TC. Anaerobic metabolism in plants. *Plant Physiology*. 1992;100(1).
4. Arango-Osorio S, Vasco-Echeverri O, López-Jiménez G, González-Sánchez J, Isaac-Millán I. Methodology for the design and economic assessment of anaerobic digestion plants to produce energy and biofertilizer from livestock waste. *Science of the Total Environment*. 2019;685.
5. Zhang J, Liu B, Song Y, Chen Y, Fu J, Liu J, et al. Genome-wide (ChIP-seq) identification of target genes regulated by WRKY33 during submergence stress in *Arabidopsis*. *BMC Genomic Data*. 2021;22(1).
6. Du P, Kibbe WA, Lin SM. lumi: A pipeline for processing Illumina microarray. *Bioinformatics*. 2008;24(13).

7. Machanick P, Bailey TL. MEME-ChIP: Motif analysis of large DNA datasets. *Bioinformatics*. 2011;27(12).
8. Medina-Rivera A, Defrance M, Sand O, Herrmann C, Castro-Mondragon JA, Delerce J, et al. RSAT 2015: Regulatory sequence analysis tools. *Nucleic Acids Research*. 2015;43(W1).
9. Zheng Q, Wang XJ. GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis. *Nucleic Acids Res*. 2008;36(Web Server issue).
10. Andrews S. FastQC. *Babraham Bioinformatics*. 2010;
11. Twell D, Yamaguchi J, McCormick S. Pollen-specific gene expression in transgenic plants: Coordinate regulation of two different tomato gene promoters during microsporogenesis. *Development*. 1990;109(3).
12. Liu X, Chu KM. α -Actinin-4 promotes metastasis in gastric cancer. *Laboratory Investigation*. 2017;97(9).

Code

Downloaded reference TAIR-10 genome from:

https://support.illumina.com/sequencing/sequencing_software/igenome.html

Downloaded data from study using accession number CRA003775 from:

<https://bigd.big.ac.cn/gsa>

```
# upload files to remote server
scp /Users/alicezhang/Documents/400E_Assignments/Project/* alzhang_bmeg22@orca

scp -r /Users/alicezhang/Downloads/Arabidopsis_thaliana_Ensembl_TAIR10/Arabidc

#align the input file
bowtie2 -x genome/genome -U CRR242641.fq -S input_align.sam
#align the Chip seq file
bowtie2 -x genome/genome -U CRR242640.fq -S align.sam

#convert sam files to bam and also sort them
samtools view -S -b -h -@ 32 align.sam > align.bam
samtools view -S -b -h -@ 32 input_align.sam > input_align.bam
```



```
#filter only uniquely mapped reads with at most 2 mismatches
sambamba view -F "[XS] == null and not unmapped and [NM] <=2)" -h -t 32 --form
sambamba view -F "[XS] == null and not unmapped and [NM] <=2)" -h -t 32 --form

#gotta sort for the index thing
sort filtered_align.bam -o sorted_filtered_align.bam
sort filtered_align.bam -o sorted_filtered_input_align.bam

#Create index to create bigwig files
samtools index -b -@ 32 sorted_filtered_align.bam
samtools index -b -@ 32 sorted_filtered_input_align.bam

#creating the bw files
bamCoverage -b sorted_filtered_align.bam -o filtered_align.bw --numberOfProces
bamCoverage -b sorted_filtered_input_align.bam -o filtered_input_align.bw --nu

#move files to local to view in IGV
scp adhir_bmeg22@orca1.bcgsc.ca:/home/adhir_bmeg22/project/filtered_align.bw
scp adhir_bmeg22@orca1.bcgsc.ca:/home/adhir_bmeg22/project/filtered_input_ali

#install macs2 and call peaks
conda install -c bioconda macs2
macs2 callpeak -q 0.05 -t sorted_filtered_align.bam -c sorted_filtered_input_a

#fasta for motif analysis, indexed reference genome to itself to obtain whole g
bedtools getfasta -fi ../project/whole_genome.fa -bed peaks_align_peaks.narrow

#changing bed file and bam to fasta for meme
samtools ampliconclip -b peaks_align_summits.bed sorted_filtered_align.bam -o
#ran meme locally and on meme server, results in repository

# install.packages("igraph", type = "binary") (nonbinary didn't work for me fo
library(dplyr)
library(igraph)

#install these with Bioconductor, don't update igraph
library(ChIPseeker)
library(clusterProfiler)
library(rtracklayer)

## Warning: package 'S4Vectors' was built under R version 4.1.3

library(GenomicRanges)
library(regioner)
```

```
library(VennDetail)
```

```
# calculate peak coverage using gRanges
bed2 <- as.data.frame(read.table("peaks_broad.bed", header = FALSE, sep="\t"))
res2 <-
  lapply(split(bed2, bed2$V1), function(i){
    GRanges(seqnames = i$V1,
            ranges = IRanges(start = i$V2,
                             end = i$V3,
                             names = i$V4))
  })
# covplot(res2)
```

```
# find overlaps
library(TxDb.Athaliana.BioMart.plantsmart28)
all_gene <- genes(TxDb.Athaliana.BioMart.plantsmart28)
resUnlisted <- unlist(GRangesList(res2))
hits <- findOverlaps(resUnlisted, all_gene)
```

Obtained EnsDb database from:

<https://github.com/wyguo/EnsDb.Athaliana.vAtRTD2/blob/master/inst/extdata/EnsDb.Athaliana.vAtRTD2.sqlite>

```
# peak annotation
annoData <- GRanges(all_gene)
library(ChIPpeakAnno)
library(TxDb.Athaliana.BioMart.plantsmart28)
library(org.At.tair.db)
library(ensemldb)
```

```
## Warning: package 'ensemldb' was built under R version 4.1.3
```

```
txdb <- TxDb.Athaliana.BioMart.plantsmart28
anno <- annotatePeakInBatch(resUnlisted, AnnotationData=annoData,
                           output="overlapping",
                           FeatureLocForDistance="TSS",
                           bindingRegion=c(-2000, 300))
anno$symbol <- xget(anno$feature, org.At.tairSYMBOL)
edb <- EnsDb("EnsDb.Athaliana.vAtRTD2.sqlite")
peakAnno <- annotatePeak(resUnlisted, tssRegion=c(-3000, 3000),
                        TxDb=txdb, annoDb="org.At.tair.db")
```

```
## >> preparing features information...      2022-04-14 22:12:02
## >> identifying nearest features...        2022-04-14 22:12:03
## >> calculating distance from peak to TSS... 2022-04-14 22:12:03
## >> assigning genomic annotation...        2022-04-14 22:12:03
## >> adding gene annotation...             2022-04-14 22:12:08

## Warning in annotatePeak(resUnlisted, tssRegion = c(-3000, 3000), TxDb = txc
## Unknown ID type, gene annotation will not be added...

## >> assigning chromosome lengths          2022-04-14 22:12:08
## >> done...                             2022-04-14 22:12:08
```

```
peakAnno.edb <- annotatePeak(resUnlisted, tssRegion=c(-3000, 3000),
                           TxDb=txdb, annoDb="org.At.tair.db")
```

```
## >> preparing features information...      2022-04-14 22:12:08
## >> identifying nearest features...        2022-04-14 22:12:08
## >> calculating distance from peak to TSS... 2022-04-14 22:12:08
## >> assigning genomic annotation...        2022-04-14 22:12:08
## >> adding gene annotation...             2022-04-14 22:12:09

## Warning in annotatePeak(resUnlisted, tssRegion = c(-3000, 3000), TxDb = txc
## Unknown ID type, gene annotation will not be added...

## >> assigning chromosome lengths          2022-04-14 22:12:09
## >> done...                             2022-04-14 22:12:09
```

```
# plotAnnoPie(peakAnno)
```

```
# attempted to get sequence from summits but this did not yield good results a
library(tidyverse)
library(microseq)
```

```
BiocManager::install("BSgenome.Athaliana.TAIR.TAIR9")
```

```
## Warning: package(s) not installed when version(s) same as current; use `for
## re-install: 'BSgenome.Athaliana.TAIR.TAIR9'
```

```
library(BSgenome.Athaliana.TAIR.TAIR9)
genome <- BSgenome.Athaliana.TAIR.TAIR9
#summitSeq = get_peak_summit_seq(file = "peaks_align_peaks.narrowPeak", peakFc

# https://rdr.io/github/lakhanp1/chipmine/man/get_peak_summit_seq.html

#install.packages("remotes")
#remotes::install_github("lakhanp1/chipmine")
```

Appendix

FastQC Report

Summary



Basic Statistics



Per base sequence quality








Per sequence quality scores



Per base sequence content






Per sequence GC content

-  Per base N content
-  Sequence Length Distribution
-  Sequence Duplication Levels
-  Overrepresented sequences
-  Adapter Content

1.

FastQC Report

Summary

-  Basic Statistics
-  Per base sequence quality
-  Per sequence quality scores



Per base sequence content



Per sequence GC content



Per base N content



Sequence Length Distribution



Sequence Duplication Levels



Overrepresented sequences



Adapter Content

Authors and contributions

Authors: Alice Zhang (52721552) and Abhishek Dhir (87603866)

Contributions: Alice and Abhishek worked together synchronously and asynchronously on the project.