Studying regulatory elements in yeast and Dictyostelium discoideum

Shoshana Elgart, Taissija Rychkova

Under the guidance of Irina Zhegalova

Background

Goal - To locate enhancers in both the yeast and Dictyostelium discoideum genomes by evaluating patterns of histone modifications in the chromatin that indicate enhancer presence

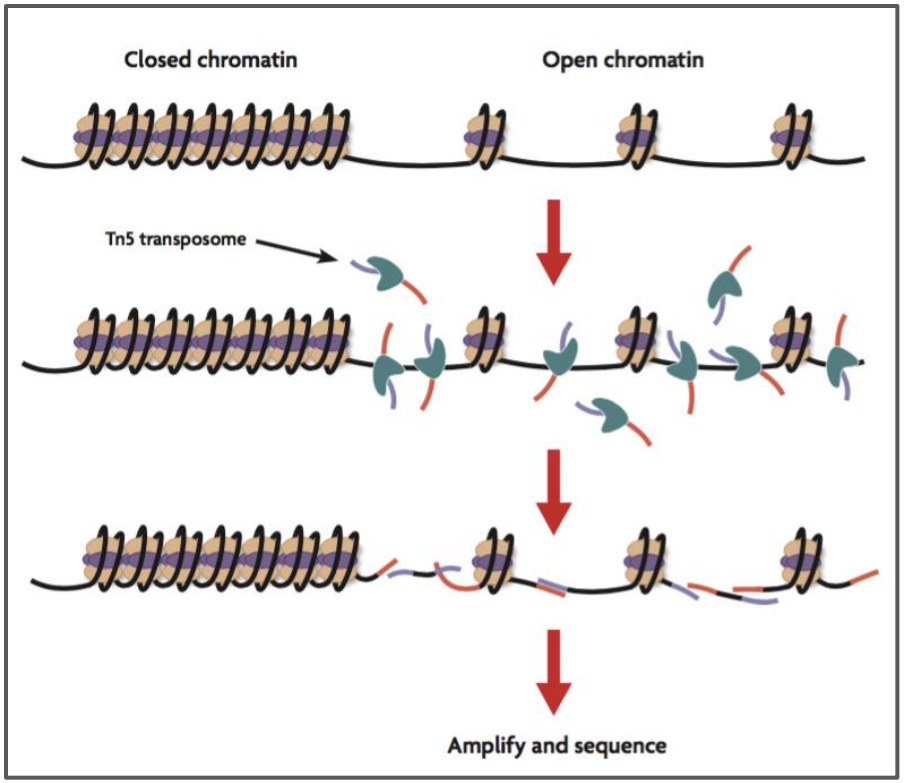
- A histone modification is a covalent post-translational modification (PTM) to histone proteins. Types of histone modifications include:
  - methylation,
  - phosphorylation
  - acetylation,
  - ubiquitylation,
  - sumoylation.

- An enhancer is a cis-acting regulatory region of DNA that can be bound with proteins known as transcription factors to increase the likelihood that its corresponding gene will be transcribed.

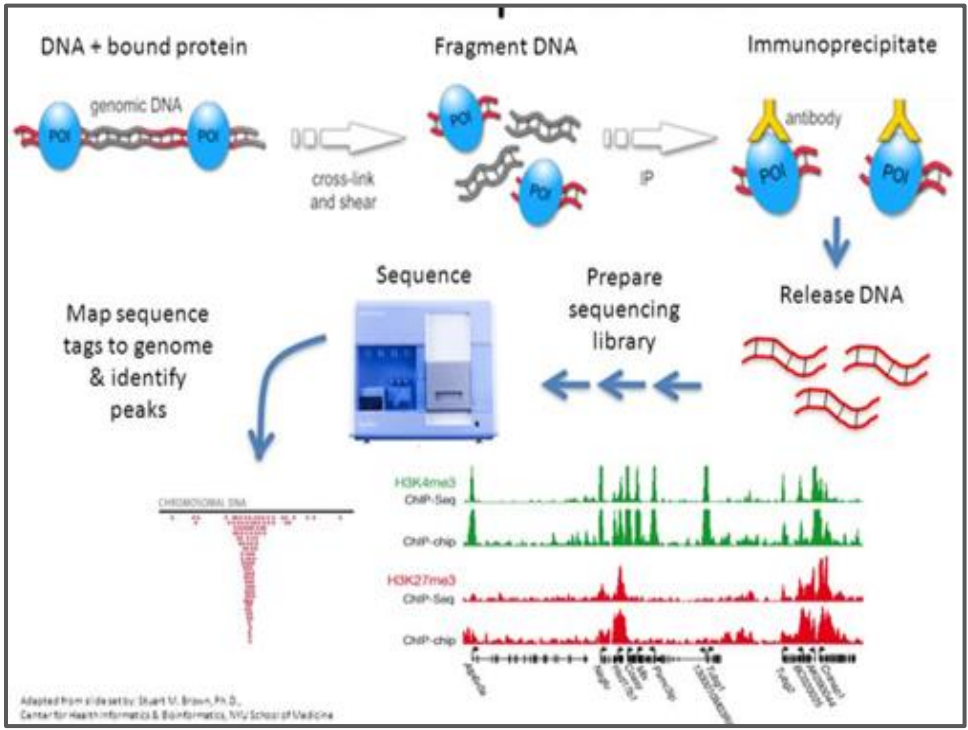
Methods used in getting genome data:

1. **ATAC-seq** (**Assay for Transposase-Accessible Chromatin using sequencing**) is a technique used to assess chromatin accessibility throughout the genome with the mutant enzyme Tn5 Transposase that inserts sequencing adapters into open regions.

2. **ChIP-sequencing**, also known as **ChIP-seq**, is a method used to analyze protein interactions with DNA. ChIP-seq combines chromatin immunoprecipitation (ChIP) with DNA sequencing to identify the binding sites of DNA-associated proteins.



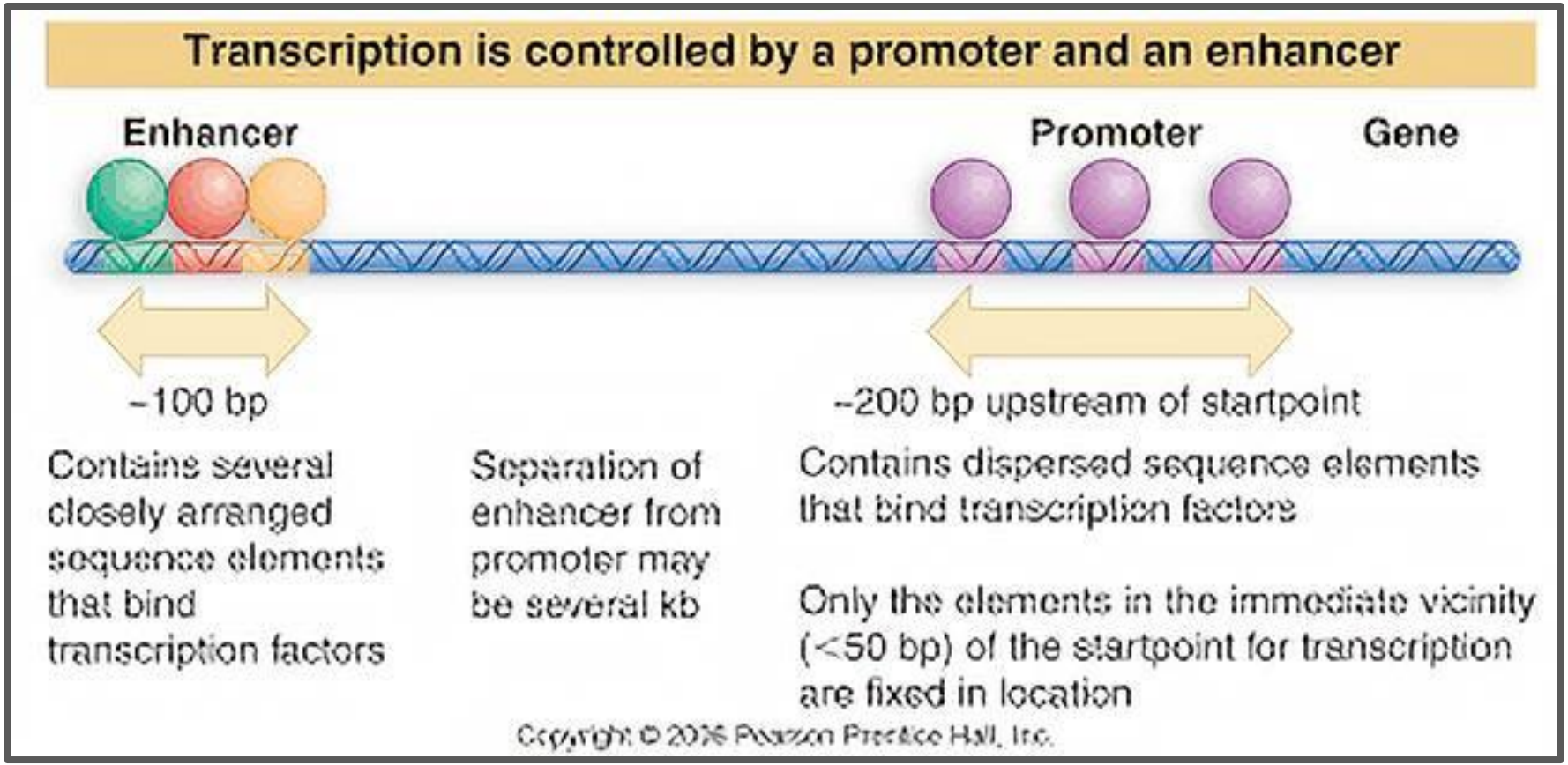
ATAC-seq



ChIP-seq

Diagrams

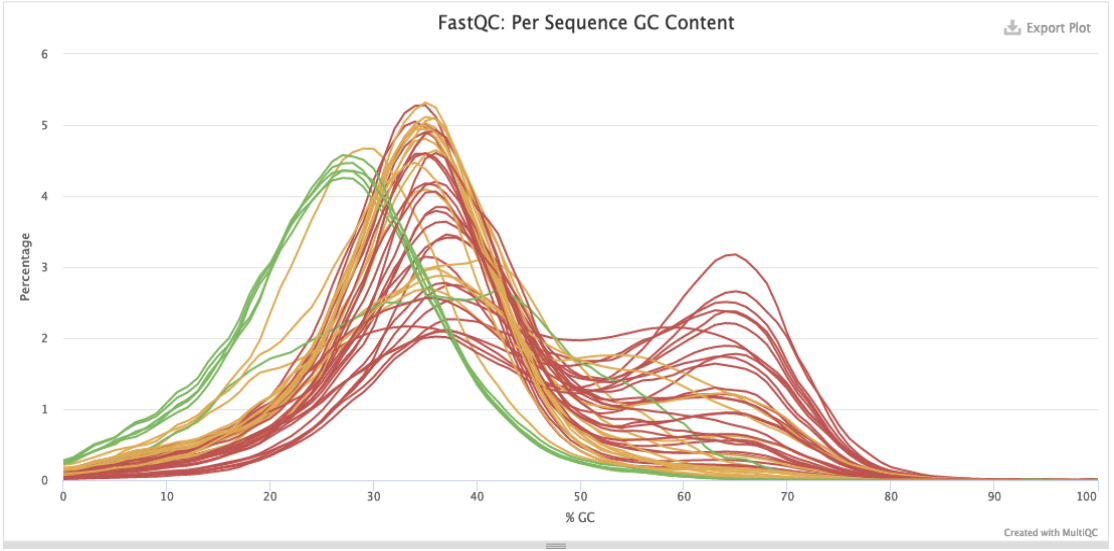
Promoters vs. Enhancers

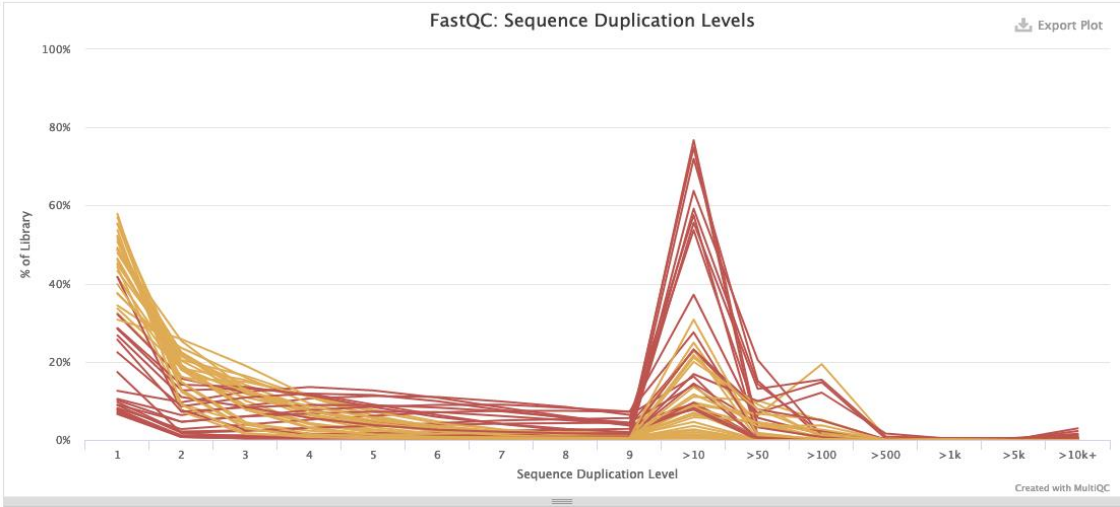


Steps

Step 1

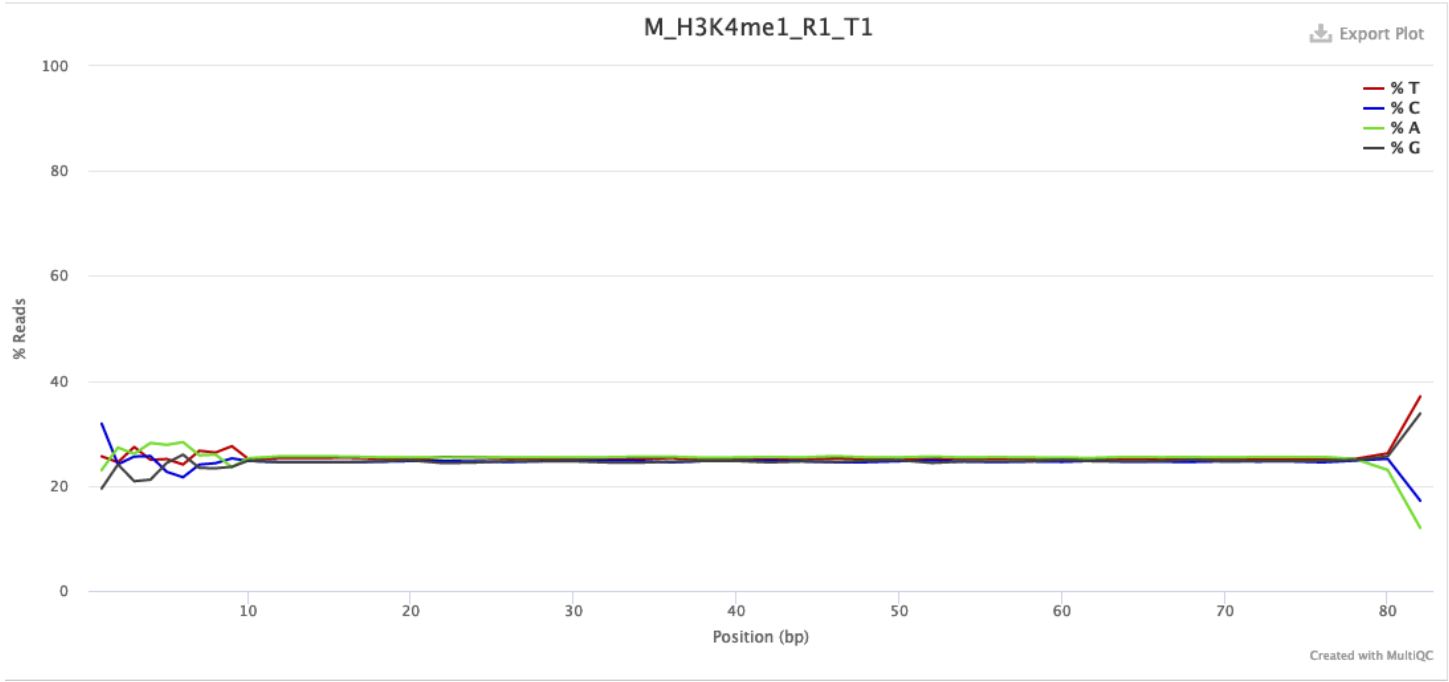
- Upon obtaining and considering a FastQC report on the ChIP-seq for the Dictyostelium Discoideum genome, we noted several considerable issues with the genome.
  - GC counts were abnormal - see upper image.
  - Sequence duplication levels were too high - see lower image.





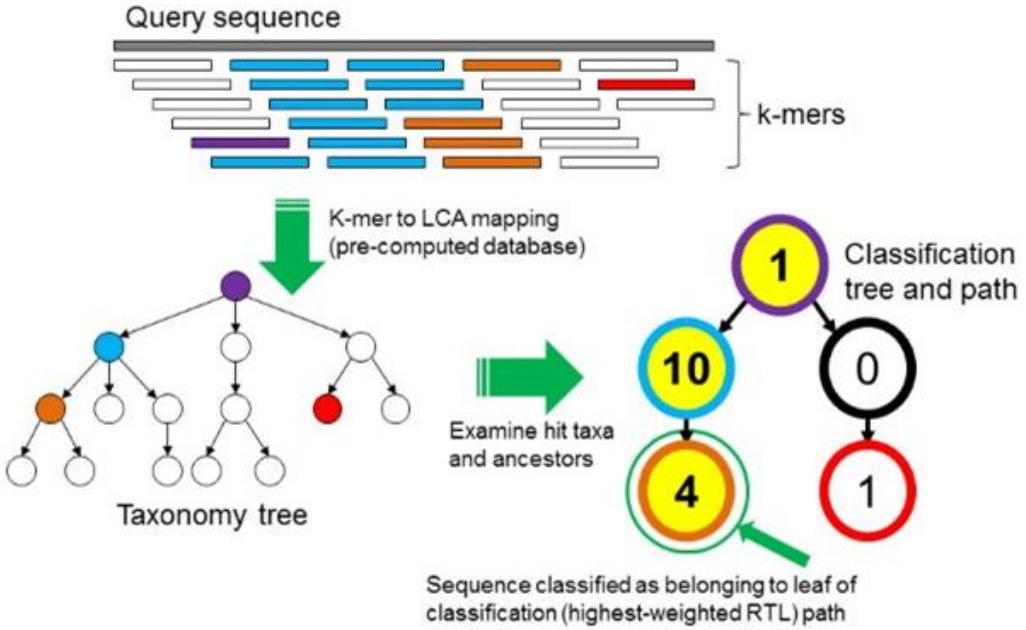
Step 2

- ❑ Further analysis of the data on the GC counts showed that the majority of errors occurred on the base pairs at the very ends of each sequence fragment. (see example for sequence M_H3K4me1_R1_T1 at bottom).
- ❑ Trimming these (relatively small amounts) of base pairs could thus greatly improve the quality of the data.



Step 3: Identifying Contamination With Kraken

- Kraken is a sequence classification tool used for assigning taxonomic labels to metagenomic DNA sequences.
- Using Kraken on H3K4me1_R1 enabled us to locate the sources of contamination: Human DNA and various associated bacteria.



Step 4: Trimming Data

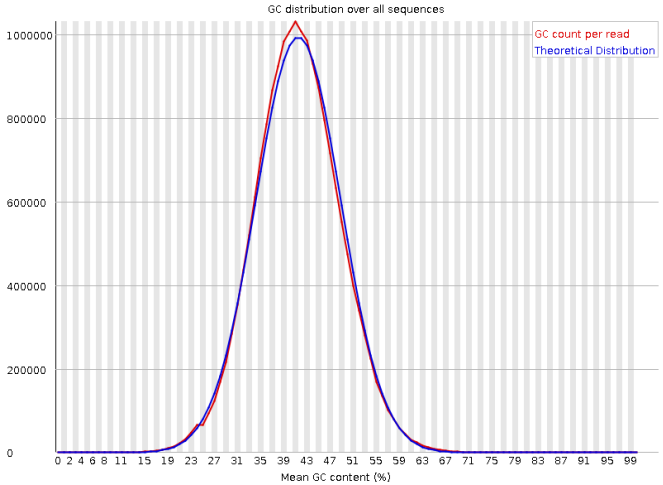
- ❏ Reads were first trimmed using the software Trimgalore, which is capable of removing a certain number of bases at either end of a fragment.
- ❏ The newly trimmed reads were then aligned.

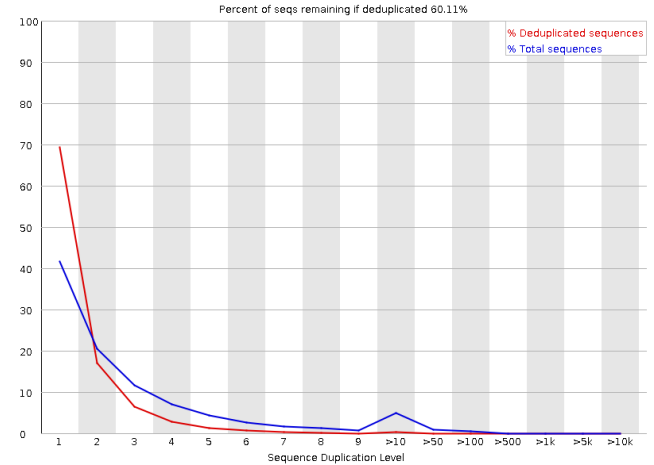
Step 5: Creating the Yeast Pipeline

- ❏ As we had a FastQC report for only Dictyostelium discoideum, we needed to obtain similar data for yeast, so a pipeline was created to see this report.
- ❏ This way, we obtained FastQC data of H3K4me1_R1_T1, H3K4me2_R1_T1, H3K4me3_R1_T1 and H3K27ac_R1_T1.

Step 6: More Quality Control

- After looking through the yeast FastQC report, we noticed that, in comparison to the Dictyostelium discoideum genome, GC content in all of the regions was good (upper image: H3K4me1) and we had no overrepresented sequences, but there were relatively high levels of duplication in every region (see the image below: H3K4me1).





Step 7: Bedtools Intersect

- ❑ Bedtools intersect allows to identify whether or not the analysed sets of genomic features have “overlap” with each other.
- ❑ It was used to find enhancer regions in Dictyostelium discoideum and yeast genomes in the areas of histone modifications H3K4me1 and H3K27ac, as they are common for the enhancers.

This is how it works:

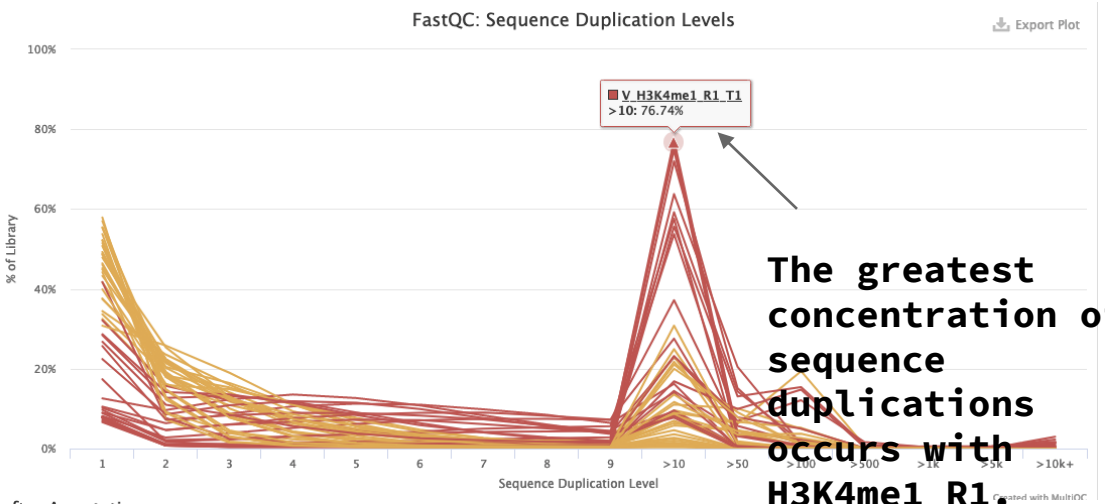
Step 8: More Steps

- **The first thing we did** was to compare the regions of H3K4me1 and H3K27ac. As we had two replicates for Dictyostelium discoideum, we had to first find overlaps between H3K4me1_R1 and H3K4me1_R2, then H3K27ac_R1 and H3K27ac_R2.
  - After a deeper analysis of the Dictyostelium discoideum MultiQC report, the decision was made to exclude H3K4me1_R1 from intersections, as it was contaminated with extra human DNA (see the next slide).

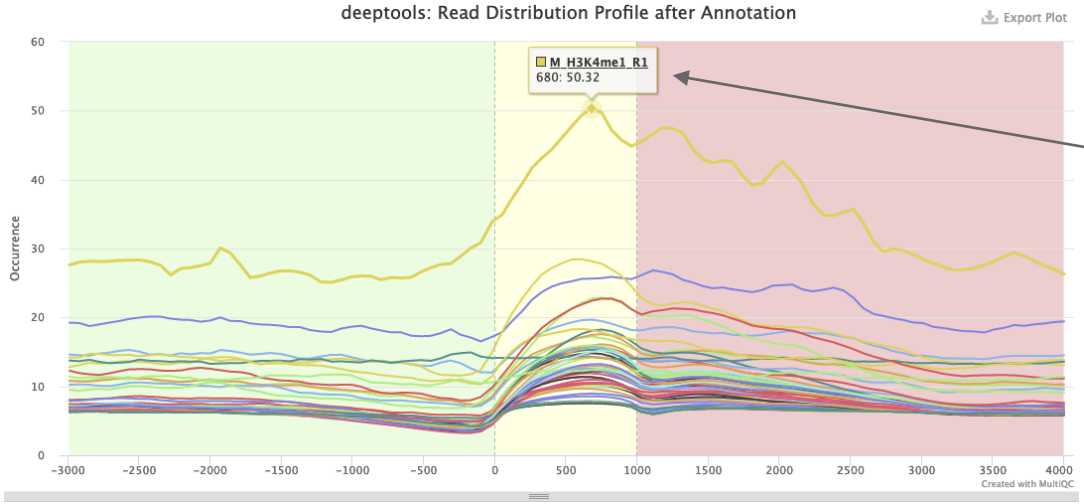
Only after this procedure were we able to begin looking for the overlaps in the output files to get more accurate results.

- **Secondly,** we compared our results with the reference genes to exclude such overlaps in genes, as they would be not relevant in terms of promoters and enhancers.
- **The third step** was to to separate the result by the principle of presence of overlaps with promoter regions, which enabled us to find the enhancers in the Dictyostelium discoideum

Issues with H3K4me1_R1 in Dictyostelium Discoideum



The greatest concentration of sequence duplications occurs with H3K4me1_R1.



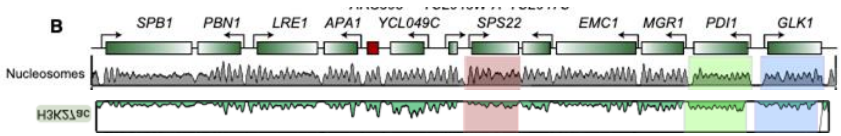
The most abnormal read distribution occurs with H3K4me1_R1.

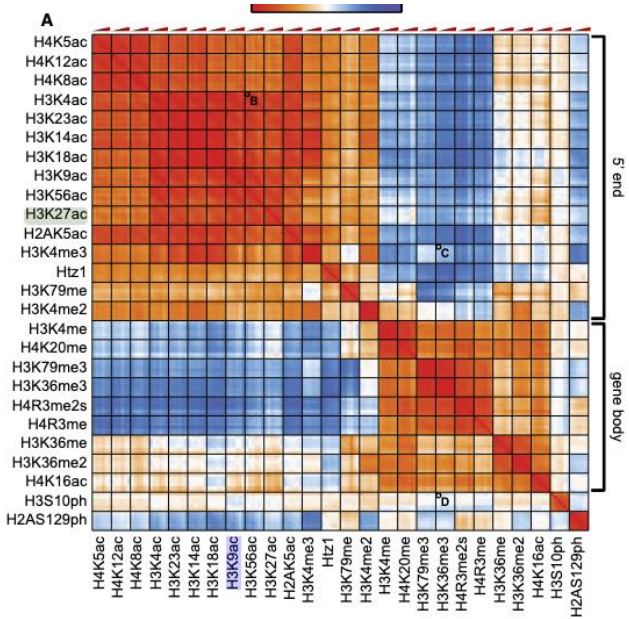
When the Problem Appeared -

- While looking for overlaps between H3K4me1 and H3K27ac in yeast, we found only two of them. This meant that:
  - Yeast has only two promoters/enhancers, which is hardly possible.
  - One of these histone modifications just doesn't work for yeast.

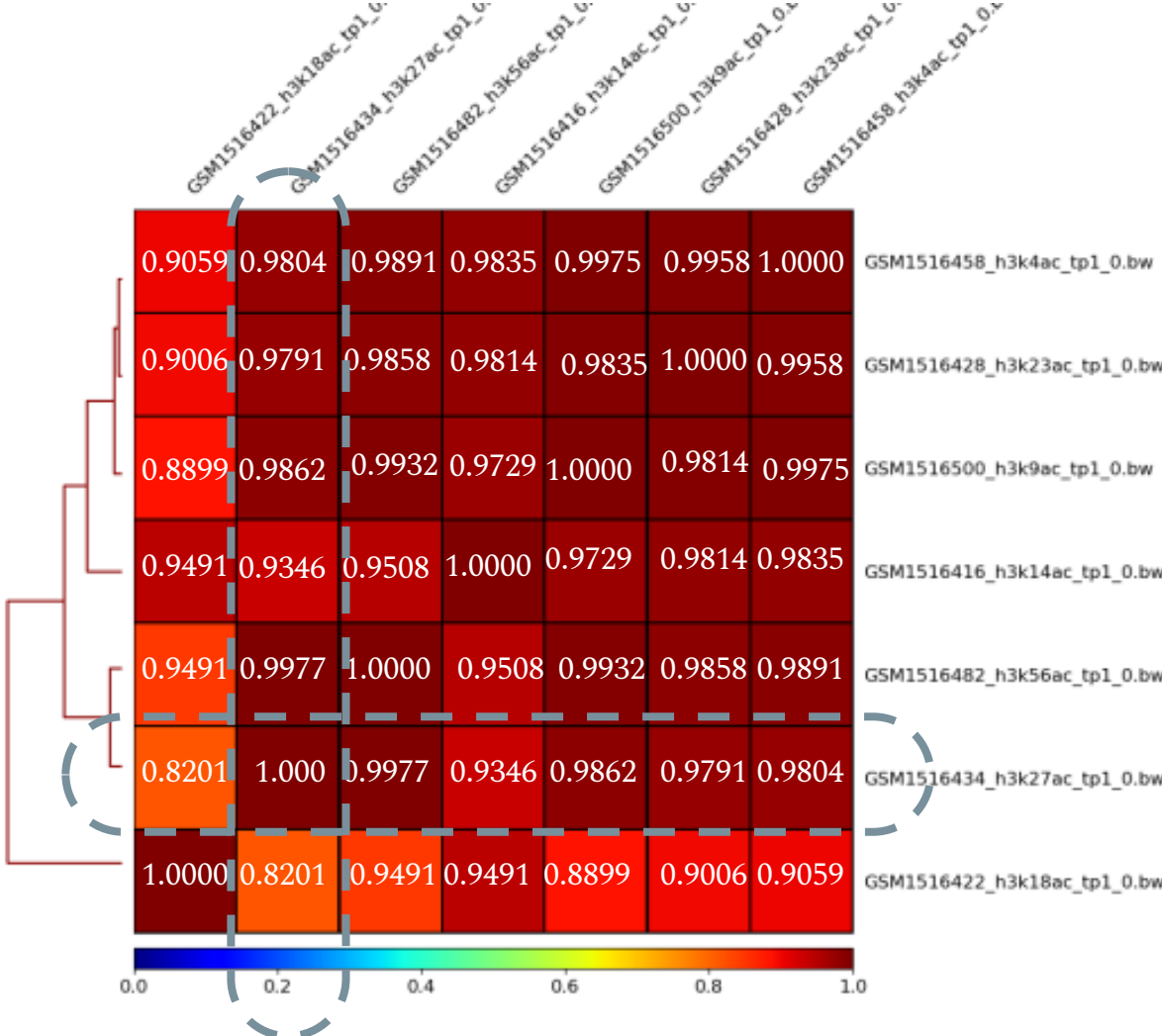
We decided to check both histone modifications in the IGV and found out the that peaks for H3K27ac look just like those in the input file, so there are no peaks for yeast here. Because of that we started looking for replacement acetylations, which might be suitable for our enhancer-focused research and would have a high functional correlation with H3K27ac.

After reading some articles and looking for correlations, we found that H3K27ac has correlations with several other H3 acetylations.



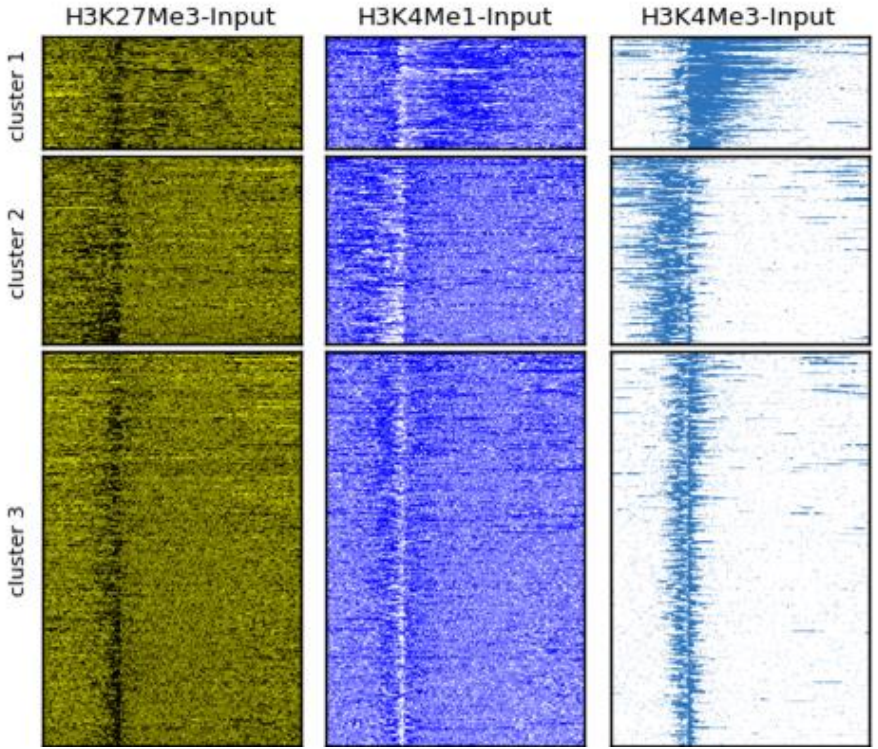


- We chose the most relevant acetylations, downloaded some data about them from NCBI, and started looking for the one with the greatest correlational score.
- We then used the *multiBigwigSummary* tool to compute the average scores for each of the files in every genomic region.
- Finally, we created a correlation heatmap and a scatterplot using the *plotCorrelation* tool. (see image on the right: heatmap and scatterplot data combined)



Further Steps

- Our next goal is to successfully rerun the yeast ChIP-seq pipeline using the new histone modification H3K56ac (with a correlation of 0.9977 with H3K27ac).
- We also hope to be able to further analyze the Kraken data for the faulty H3K4me1.
- Finally, we aim to study the epigenetics features of the genome surrounding the enhancers and promoters we find, using a deepTools heatmap.



Links

1. ATAC-seq
    a. https://youtu.be/uuxpyhGNDsk
    b. https://www.encodeproject.org/atac-seq/
    c. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
2. ChIP-seq
    a. https://www.youtube.com/watch?v=nkWGmaYRues
    b. https://www.encodeproject.org/chip-seq/histone/
    c. https://nf-co.re/chipseq/1.2.2
    d. https://drive.google.com/file/d/1Q473kRWCxpLuBYXCOmCw4S8s68FtpIVV/view?usp=sharing
    e. https://www.ncbi.nlm.nih.gov/Traces/study/?acc=PRJNA262623&o=acc_s%3Aa&s=SRR1593092,SRR1593098,SRR1593104,SRR1593122,SRR1593128,SRR1593134,SRR1593214,SRR1593217,SRR1593225,SRR1593228,SRR1593231,SRR1593247,SRR1593251,SRR1593252
    f. https://ccb.jhu.edu/software/kraken2/index.shtml
    g. https://broadinstitute.github.io/picard/explain-flags.html
    h. https://deeptools.readthedocs.io/en/develop/content/feature/effectiveGenomeSize.html - about Effective genome Size
    i. https://bedtools.readthedocs.io/en/latest/content/tools/intersect.html - how to intersect