**INTRODUCTION**

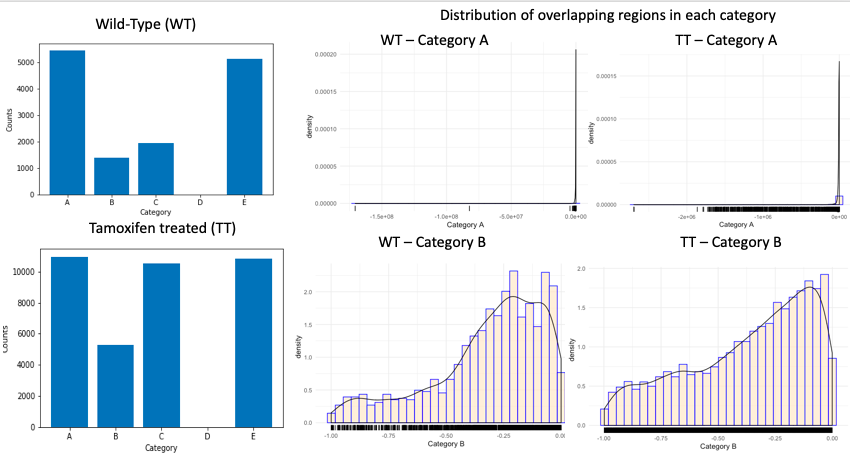
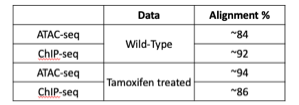
In recent years new genetic techniques have been developed and paired with high-throughput next-generation sequencing to obtain new information about chromatin structure and DNA-protein binding. ATAC-seq, assay for transposase accessible chromatin, utilizes a hyperactive tn5 prokaryotic transposase to chop out areas of open chromatin which can then be sequenced and aligned to a reference genome [1]. ChIP-seq uses cross-links between DNA and bound proteins, sonically fragments the DNA, then makes use of immunoprecipitation, specific for the protein of interest, to pull down those DNA-protein fragments of interest. After releasing the bound protein, sequencing, and mapping can take place. ChIP-seq is often used to study transcription factor binding and transcription network regulation [2]. These techniques used individually provide useful scientific knowledge, however even more detailed information can be obtained by analyzing the intersection of data from multiple techniques. The relationship between protein binding and chromatin structure may begin to be elucidated by study of a cross-section of data from correlated ATAC-seq and ChIP-seq experiments. Our study aims to do just this for the MCF-7 breast cancer cell line, utilizing estrogen receptor as the protein-of-interest, under two conditions, wild-type and estrogen receptor ligand treated. Merging the ATAC-seq and ChIP-seq datasets for each condition enables creation of an accessible-chromatin DNA-element binding profile to facilitate comparison across the two conditions. We expect for the ligand treatment to modulate the accessible-chromatin DNA-element binding profile for the MCF-7 cell line, and modulate thus alter the DNA binding of the estrogen receptor in the cell.

**METHODS**

1. Data: The data for this study was obtained from the Gene Expression Omnibus (GEO) and the analysis were performed using the tools on Carbonate. For this study, four samples sequenced using the Illumina NextSeq 500 (Homo Sapiens) were considered – ChIP-seq (2 samples) and ATAC-seq (2 samples). The FASTQ files of wild-type and tamoxifen treated samples of ChIP-seq and ATAC-seq were downloaded from the European Nucleotide Archive (ENA) [3]. The GEO Accession ID of ChIP-seq and ATAC-seq dataset are GSE117941 and GSE117940 respectively. The sample ID’s are: ChIP-seq wild-type data - [GSM3315622](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM3315622), treated set: [GSM3315625](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM3315625) and ATAC-seq wild-type data: [GSM3315603](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM3315603), treated set: [GSM3315614](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM3315614). Also, the human reference genome (hg38) was downloaded from Ensemble and indexed using the HISAT2 tool to be used for alignment.
2. Quality Control and Alignment: The first step, after downloading the data, was to check the quality of the FASTQ files. The quality control check was performed using the FastQC tool [4] and the adapters were trimmed if the quality score was less than 28. Alignment of FASTQ files with the indexed reference genome using the HISAT2 tool was performed and the BAM files of four samples were obtained [5].
3. Peak Calling: In order to find the transcription factor binding sites, peak calling of the aligned files was performed using MACS2 [6]. The input file to MACS2 was a BAM file and the output is an .xls file which has the chromosome co-ordinates, peak length, peak ID, and fold enrichment along with a BED file. Based on input given by the panel, broad peak calling was also performed for the ATAC-seq data to account for data representing nucleosome free regions.
4. Upon completion of the two sets of ATAC-seq and ChIP-seq analysis a comparative analysis between the corresponding peak-called files was performed. This was done by using an ad hoc Python script. The script takes a pair of peak-called files (ChIP and ATAC seq), in the .xls format, as input and computes the shortest distance between each ChIP peak and the closest ATAC peak. Instances of overlapping coordinate pairs, were given special consideration: with completely contained pairs assigned as 0, partially contained pairs set equal to ±1 ± the number of overlapping nucleotides divided by the length of the ChIP peak. In addition to the distance calculation, the script also classified the peak into one of 5 categories (< -1, -1 < x < 0, 0, 0 < x < 1, > 1).
5. Homer[7] is a popular tool used for finding motifs enriched in input (also called target) against a reference background generated by the tool itself. It expects to find motifs with patterns that occur more than by chance in the target set. The input given to this tool is a bed file from Macs2. The tool outputs a consolidated .html report and motifs files for each motif found.

**RESULTS**

From the FastQC reports generated by the FastQC tool, 3 out of 4 samples had a good quality score (>28) whereas the ChIP-seq Wild-type sample contained adapter contamination and it affected the alignment percentage. Hence, the reads with phred scores less than 28 were trimmed and the resulting FASTQ file was used for further analysis. The FASTQ files were aligned to the reference genome (hg38) and the alignment percentage is tabulated below:



The Python script analysis revealed a stark difference in the binding behaviors of transcription factors from wild type to tamoxifen treated. In all present categories, (D excluded) the number of peaks in each category at least doubled, with category B and category C going up by 3x and 5x, respectively. A closer examination of the distribution of distances in each category shows further distortion between the wild type and the tamoxifen treated cells, especially among category A and category B. While category D does not appear to be present in the analysis, it is uncertain result is biologically relevant and not as an artifact of the methodology, or a result of the specific experimental conditions.

From the HOMER analysis, we were able to find more motifs in the tamoxifen treated overlapping peaks of ChiP-seq and ATAC-seq compared to wild type. It was also observed that ratio of number of transcription factor motifs found in the targeted sequences compared to background was higher in the wild type overlapping peaks. Transcription factor families such as Fosl and Jun, important for cell cycle regulation and signaling[8,9,10] showed up in the homer results for both the sets. More specific estrogen receptor motifs factors like ERE and ERR were found only in the tamoxifen treated set.

**DISCUSSION**

From the Python analysis for both the treated and wild type overlapping peaks, it was expected to find least number of overlapping peaks in category B, as most transcription factor and promoter binding sites are at least 2k bases upstream of the gene to be transcribed. This explains the greater number of peaks found in category A and E as upstream of a gene can be the downstream of another gene. In the wild type set, category C has fewer peaks (than A and E), while in the treated set, category C has a similar number of peaks compared to category A and E. This is most likely due to tamoxifen altering the chromatin structure leaving it more accessible for transcription factors to bind[3]. There is further evidence for this, when one compares the peak distribution for category A; the wild type peaks are more densely located at a specific region, while in the treated set they can be seen to be distributed throughout. The Homer analysis showed a lower ratio of targeted sequences in the tamoxifen treated samples than in the wild type. Given tamoxifen is known to act as an SERM[11] (selective estrogen receptor modulator) is a partial agonist[12] of the estrogen receptor and is used treat breast cancer. This acts as additional support that tamoxifen may interfere with certain transcription factors important for the cell signaling cascade.

**CONTRIBUTION**

Xander came up with the idea for the project, found the wild type studies, wrote the Python script, and performed the ChIP sequencing for the wild type. Sharmila performed the ATAC sequencing for the treated samples, researched and designed the workflow pipeline, and created the figures. Gayathri performed the Homer analysis, performed the ATAC sequencing for the wild type, and performed the research to analyze the Python and Homer results. Trever did the bulk of the research into the methodology, provided the list of references, found the treated study, and performed the ChIP sequencing for the treated samples.

**REFERENCES**

1. Buenrostro JD, Wu B, Chang HY, Greenleaf WJ. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. Curr Protoc Mol Biol. 2015;109:21.29.1–21.29.9. Published 2015 Jan 5. doi:10.1002/0471142727.mb2129s109 <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4374986/pdf/nihms653929.pdf>
2. Pavesi G. ChIP-Seq Data Analysis to Define Transcriptional Regulatory Networks. Adv. Biochem. Eng. Biotechnol. 2017;160:114. [https://air.unimi.it/retrieve/handle/2434/528076/915386/chp%253A10.1007%252F10\_2016\_43%20%281%29.pdf](https://air.unimi.it/retrieve/handle/2434/528076/915386/chp%253A10.1007%252F10_2016_43 (1).pdf).
3. Guan J, Zhou W, Hafner M, Blake RA et al. Therapeutic Ligands Antagonize Estrogen Receptor Function by Impairing Its Mobility. Cell 2019 Aug 8;178(4):949-963.e18. PMID: 31353221
4. Andrews, S. (2010). FASTQC. A quality control tool for high throughput sequence data.
5. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019Aug;37(8):907-915. doi: 10.1038/s41587-019-0201-4. Epub 2019 Aug 2. PubMed PMID: 31375807.
6. Zhang, Y., Liu, T., Meyer, C.A. et al. Model-based Analysis of ChIP-Seq (MACS). Genome Biol 9, R137 (2008) doi:10.1186/gb-2008-9-9-r137.
7. Heinz S, Benner C, Spann N, Bertolino E et al. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Mol Cell 2010 May 28;38(4):576-589.
8. Murphy, L. O. & Blenis, J. MAPK signal specificity: the right place at the right time. Trends Biochem. Sci. 31, 268–75 (2006).
9. Mina, M., Magi, S., Jurman, G. et al. Promoter-level expression clustering identifies time development of transcriptional regulatory cascades initiated by ErbB receptors in breast cancer cells. Sci Rep 5, 11999 (2015) doi:10.1038/srep11999
10. He, H., Sinha, I., Fan, R., Haldosen, L.A., Yan, F., Zhao, C., Dahlman-Wright, K. c-Jun/AP-1 overexpression reprograms ERα signaling related to tamoxifen response in ERα-positive breast cancer. Oncogene. 2018 May;37(19):2586-2600
11. Lee, WL., Cheng, M.H., Chao, H.T., Wang PH. The role of selective estrogen receptor modulators on breast cancer: from tamoxifen to raloxifene. Taiwan J Obstet Gynecol. 2008 Mar;47(1):24-31
12. Deroo, B. J., & Korach, K. S. (2006). Estrogen receptors and human disease. The Journal of clinical investigation, 116(3), 561–570. doi:10.1172/JCI27987