Differential ChIP-Seq Analysis of O-GlcNAc and Adriamycin bound Chromatin Loci in Human Breast Cancer Cells

Nigam M. and Awasthi G.*
Amity Institute of Biotechnology, Amity University Uttar Pradesh, Lucknow Campus, Lucknow, Uttar Pradesh, INDIA
*garima.baj@gmail.com

Abstract

GlcNAcylation is a dynamic post-translational modification that is involved in human diseases and a wide range of biological processes. There is evidence that GlcNAcylation modifies some tumor-associated proteins, but in tumor progression, the role of GlcNAcylation remains unclear. The decrease in cell surface E-cadherin is the molecular mechanism underlying GlcNAcylation-induced breast cancer metastasis. p120 and β-catenin, but not E-cadherin, are GlcNAcylated; the GlcNAcylation of p120 and β-catenin might play roles in the decrease of cell surface E-cadherin. Moreover, immunohistochemistry analysis indicated that the global GlcNAcylation level in breast tumor tissues is elevated significantly as compared to corresponding adjacent tissues; further, GlcNAcylation was significantly enhanced in metastatic lymph nodes compared to their corresponding primary tumor tissues.

This study shows that GlcNAcylation enhances the migration/invasion of tumor cells through the analysis of O-GlcNAc protein-bound chromatin loci in human breast cancer cells. This is an important step towards understanding the role of this protein in inducing breast cancer. The main aim of this study is to examine the O-GlcNAc protein-bound chromatin loci in the human breast cancer cells MCF-7 and MCF-7/ADR through differential ChIP-Seq analysis. Also, the prediction and identification of overall binding sites of MCF-7 and MCF-7/ADR cells were performed by the ChIP-seq strategy.

Keywords: GlcNAcylation, O-GlcNAc protein, breast cancer, ChIP-seq, differential binding sites.

Introduction

Breast cancer is the most prevalent cancer type and the second-leading cancer-related cause of death in women.[15] O-linked β-N-Acetylglucosamine (O-GlcNAc), the monosaccharide modification of serines and threonines in nuclear and cytosolic proteins, was first reported more than 30 years ago.[16] O-GlcNAcylation is the only type of glycosylation that occurs in the nucleus and cytosol and is catalyzed by O-GlcNAc transferase (Ogt), using uridine diphosphate (UDP)-GlcNAc as a donor of the GlcNAc moiety.[10] Animals contain a single Ogt enzyme and also a single enzyme, O-GlcNAcase (Oga), that removes the modification from nucleocytosolic proteins.[6]

O-GlcNAc has been proposed to be linked to thousands of proteins that are involved in various distinct cellular processes. Structural studies on Ogt provided insight into how this enzyme modifies this baffling diversity of substrates: Ogt primarily binds to the peptide backbone of substrates and shows no clear specificity for the modification of specific serines or threonines.[8] Several previous studies have implicated protein O-GlcNAcylation in the promotion of cancer hallmarks by sustaining growth and invasion[7], regulating DNA damage- and stress-responses[3,21] and controlling cell cycle progression.[2,12,19] O-GlcNAcylation is increased in most malignant tumors including breast cancer where it positively correlates with tumor progression.[2,17]

It has been shown that both ER[4,9] and PR[18] are O-GlcNAcylated. ChIP-seq (Chromatin immunoprecipitation followed by sequencing) is widely used in studying protein-DNA binding on a genome-wide scale. After cross-linking, immunoprecipitation and shearing, millions of sequenced DNA fragments (reads) are mapped to a reference genome and sites with over-abundant reads are declared putative binding sites. We focus on an important class of binding sites that have similar read profiles throughout the genome.

Specifically, the lengths of the binding sites are similar across the genome and their centers are well defined. These binding sites tend to have read profiles that look like sharp peaks. Adriamycin is commonly used to treat both early-stage and metastatic breast cancer, usually in combination with other drugs. This class includes transcription factor binding and some histone modifications measured by ChIP-seq. Thus, a program that can detect differential binding of transcription factors across multiple conditions is much needed. The main aim of this study is to examine the O-GlcNAc protein-bound chromatin loci in human breast cancer cells MCF-7 and adriamycin treated MCF-7/ADR cells through ChIP-Seq analysis.

Adriamycin is a chemotherapy drug that can slow or stop the growth of cancer cells. Also, the prediction and identification of overall and differential binding sites of MCF-7 and MCF-7/ADR cells were performed by the ChIP-seq strategy. To study the differential binding sites of O-GlcNAcylated protein and adriamycin linked to breast cancer chromatin, this hypothesis will identify the binding site location of this protein and the adriamycin drug. This

research can help us to identify the function of this protein and drug in sustaining the growth of tumors in breast cancer and in further invasion of healthy cells.

Material and Methods

ChIP-seq Data Retrieval from ENA database (European Nucleotide Archive database): In this research, ChIP-seq data was retrieved from ENA database (https://www.ebi.ac.uk/ena/browser/view/PRJNA594402**)** as shown in table 1.

Tools and Software's used: For ChIP-seq data analysis, Galaxy server (usegalaxy.org) was used. Table 2 shows the list of tools and packages that were used for different analysis steps.

Overall binding sites prediction: Galaxy is an open source, web-based platform for data intensive biomedical research. Following tools and packages were used from the usegalaxy.org as shown in figure 1. FastQC was performed using the four downloaded fastq.gz files of the raw sample data.[1] It aims to provide a simple way to do some quality control checks on raw sequence data coming from high-throughput sequencing pipelines. MultiQC aggregates result from FastQC analyses across our four samples into a single report.

Table 1
ChIP-seq data samples Accession No. SRP235291 from ENA database

S.N.	Sample Accession	Sample Name	Treatment
1.	SAMN16054559	MCF-7/ADR_1	Adriamycin treated
2.	SAMN16054558	MCF-7/ADR_2	Adriamycin treated
3.	SAMN16054547	MCF-7_1	O-GlcNAc
4.	SAMN16054546	MCF-7_2	O-GlcNAc

Table 2
List of tools and databases used for differential ChIP-Seq analysis.

S.N.	Tools	Version	Description
1.	FastQC and MultiQC	0.72+galaxy1	quality control checks on raw sequence data, aggregates result from FastQC[1]
3.	Bowtie2	2.3.4.3+galaxy0	aligns sequencing reads to long reference sequences[11]
4.	MACS2 callpeak	2.1.1.20160309.6	enriched binding sites in ChIP-seq experiments[5]
5.	ChIPseeker	1.18.0+galaxy1	annotates ChIP-seq data analysis[20]
6.	MACS2 bdgdiff	2.1.1.20160309.1	performs differential peak detection[5]
8.	Motif Analysis of Large Nucleotide Datasets (MEME-ChIP)	5.3.2	performs motif discovery, motif enrichment analysis and clustering on large nucleotide datasets[14]
9.	Gene Ontology for Motifs (GOMo)	5.3.2	determine if any motif is significantly associated with genes linked to one or more Genome Ontology (GO) terms[13]
10.	SeqMonk	1.47.1	enables the visualization and analysis of mapped sequence data[20]

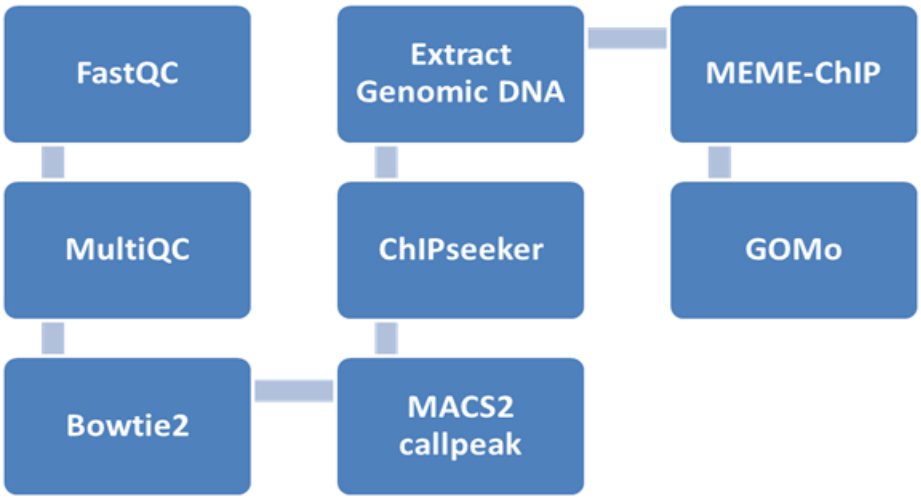


Figure 1: ChIP-seq analysis work flow to identify differential binding sites.

Bowtie2 was performed using the four downloaded fastq.gz files of the raw sample data. It is an ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences. Galaxy wrapper for Bowtie2 outputs alignments in BAM format, enabling interoperation with a large number of other tools available at this site.[11] Using these .bam files generated by Bowtie2 as input, MACS2 callpeak was performed. A MAC identifies enriched binding sites in ChIP-seq experiments. This generates the output in BED format.[5] The .bed file of the narrow peak generated by MACS2 callpeak was taken as input for ChIPseeker.[20]

It is a bioconductor package for annotating ChIP-seq data analysis. Peak annotation was performed by the annotatePeak function. It gives information about the positions of exons and introns in the detected narrow peaks. It requires an annotation source file in GTF format and in this case, comprehensive gene annotation on the reference chromosomes of humans was taken from gencodegenes.org.

Again, the .bed file of the narrow peak generated by MACS2 callpeak was taken as input for fetching genomic DNA in FASTA or interval format using the Extract Genomic DNA tool. This tool also requires a reference genome file in FASTA format and in this study, the reference taken was the human genome FASTA sequence (hg38.ga) from hgdownload.soe.ucsc.edu. Motif analysis of large nucleotide datasets (MEME-ChIP) used these FASTA files of extracted genomic DNA and performed motif discovery, motif enrichment analysis and clustering on large nucleotide datasets. MEME discovers novel, ungapped motifs (recurring, fixed-length patterns) in our sequences (sample output from sequences).[17]

The motifs generated by MEME-ChIP were taken as input for GOMo. It scanned all promoters using nucleotide motifs, we provided to determine if any motif is significantly associated with genes linked to one or more genome ontology (GO) terms (sample output from motifs and the E. *coli* K12 database).[13] The significant GO terms can suggest the biological roles of the motifs. Finally, all the results were visualized using SeqMonk, which is a program to enable the visualization and analysis of mapped sequence data.[20]

Differential binding sites prediction: Two groups were made of samples as shown in table 3 for the prediction of differential binding sites.

Table 3
Samples divided in two conditions.

S.N.	Conditions	Cell lines
1.	Condition-1	MCF-7/ADR_1
		MCF-7/ADR_2
2.	Condition-2	MCF-7_1
		MCF-7_2

For identifying the differential binding sites also, we used the same initial steps as in identifying the overall binding sites until MACS2 Callpeak, as shown in figure 2. After this step, we performed MACS2 bdgdiff using the already-generated bedgraph files from MACS2 callpeak. MACS2 bdgdiff was performed for both condition-1 and condition-2. It performs differential peak detection based on paired four-bedgraph files.[18] The interval data generated by MACS2 bgdiff was used for extracting genomic sequences. Finally, motif was done using MEME-ChIP for motif discovery and motif enrichment analysis.[20] Further motifs generated by MEME-ChIP were used for gene ontology analysis using the GOMo tool. Visualization of all the files was also done using the SeqMonk tool.[19]

Results and Discussion

Identification of overall binding sites

FASTQC result analysis: FASTQC tool gives the percentage of reads which were duplicate and also shows the total number of read sequences in our samples.

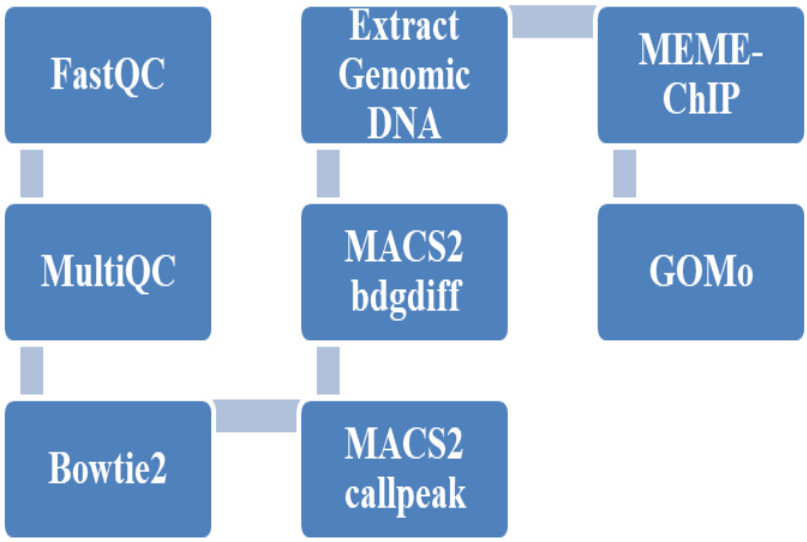


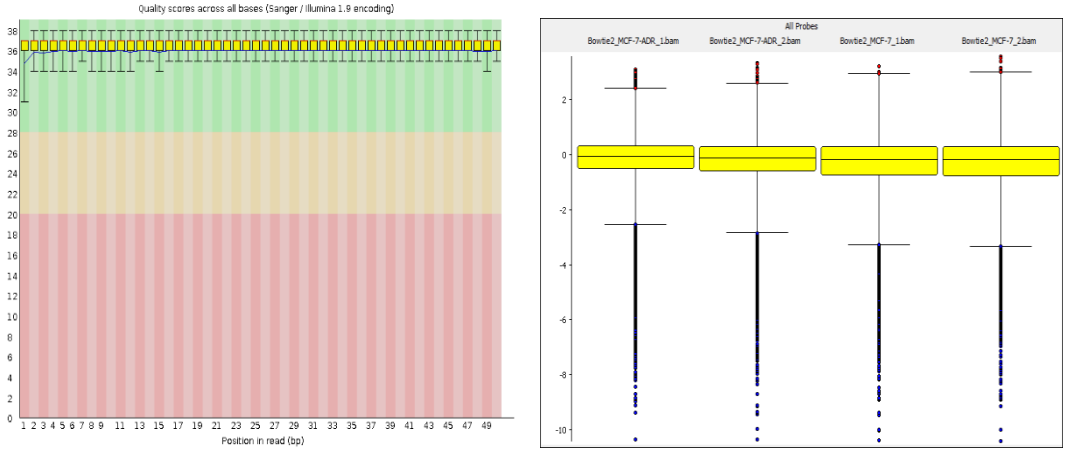
Figure 2: ChIP-seq analysis work flow to identify differential binding sites.

The Box Whisker plot shows the distribution of probe values across our several data as shown in figure 3(b).

Alignment to Reference Genome: Bowtie2 was done for the alignment of sequence reads with hg19 reference genome. The sequence alignment gave the percentage of alignment as shown in table 4. Sample MCF-7/ADR_1 gave the highest percentage 98.7% of alignment against a reference sequence. Visualization and analysis of mapped sequence data were performed using SeqMonk. The MA plots here allow us to look at the relationship between intensity and difference between two data stores at a time from our selected probe list of alignment outputs of overall binding sites as shown in figure 4. The X-axis represents the average quantitated value across the data stores and the Y-axis shows the difference between them.

MACS Peak Call: MACS2 callpeak was used here to identify enriched binding sites in ChIP-seq experiments for our 4 sample sequences.

MACS2 tool identifies the enriched binding sites as peaks as shown in figure 5. For each studied sample, MACS Peak was annotated using ChIPseeker tool. ChIPseeker package was used here for annotating ChIP-seq data of our 4 samples. Annotation of the generated narrow peaks gave the information about the positions of exons and introns in them. Figure 6 shows the percentage of different types of introns and exons present on our sample sequences.

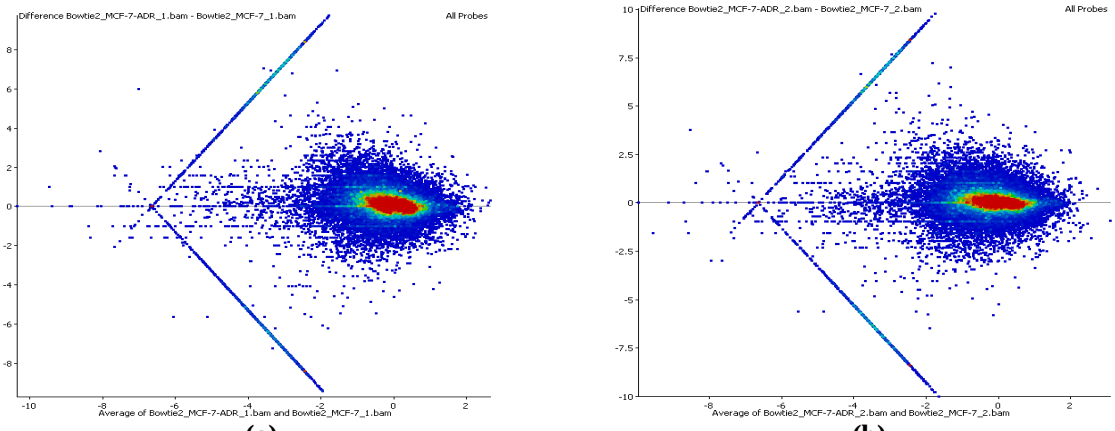


(a) **(b)**

Figure 3: (a) Quality scores per raw sequence data using FastQC (b) Box Whisker plot showing alignment for each sample.

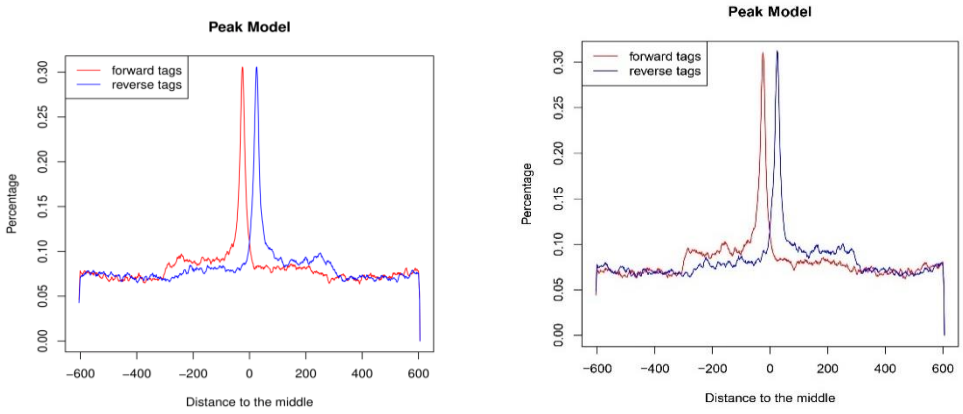
Table 4
The percentage of aligned sequence reads for each sample.

S.N.	Sample Name	%Aligned
1.	Bowtie2 MCF-7/ADR_1	**98.7%**
2.	Bowtie2 MCF-7/ADR_2	**98.6%**
3.	Bowtie2 MCF-7_1	**98.5%**
4.	Bowtie2 MCF-7_2	**98.5%**



(a) **(b)**

Figure 4: Figures (a) and (b) showing MA plots for relationship between intensity and difference between two data stores at a time using SeqMonk



(a) **(b)**

Figure 5: (a) Narrow peaks in the MACS2 callpeak plot for MCF-7 cell line (b) Narrow peaks in the MACS2 callpeak plot for Adriamycin treated MCF-7 cell line

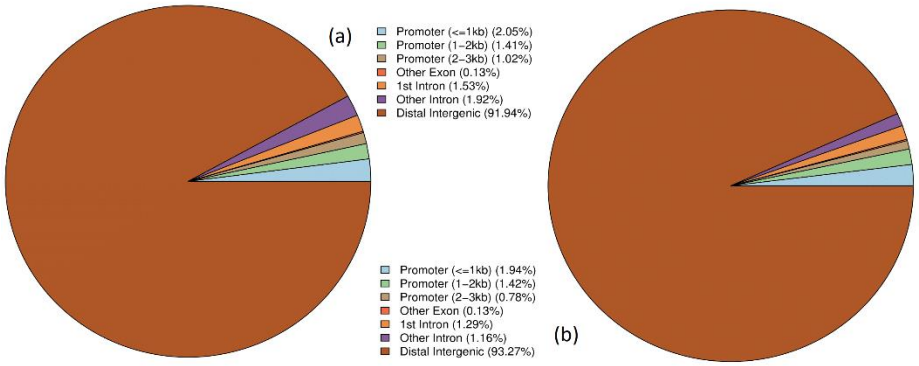


Figure 6: Percentage of different annotations of peaks in the MACS2 callpeak plot for (a) MCF-7 cell line and (b) Adriamycin treated MCF-7 cell line

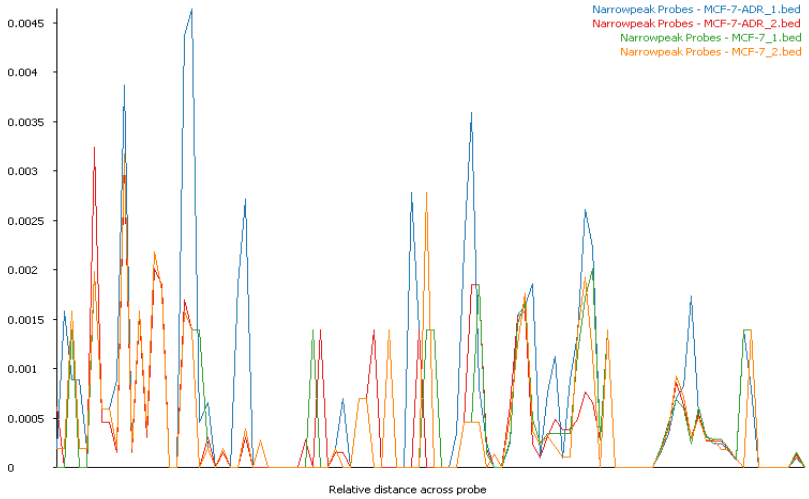


Figure 7: Plot showing the trend of narrow peaks indicating the identified enriched binding sites.

Probe Trend plot for Identified Enriched Binding Sites: The probe trend plot here as shown in figure 7 allows us to look at differences in read density over our probes. It has taken all the probes in our probe list and worked out how many reads overlap with each position in each probe and then plotted this out as an average over all of the probes showing the trend of narrow peaks indicating the identified enriched binding sites.

Motif Discovery performed using MEME-ChIP: The FASTA files of extracted genomic DNA were used to generate the exact common binding sites also known as motifs or consensus region. MEME discovered novel, ungapped motifs (recurring, fixed-length patterns) in our 4 sample sequences as shown in figure 8. For the two samples (MCF-7_1 and MCF-7_2) of O-GlcNAc protein, motif identified are "GAGCAGGAATTGAATCCTACTTCT" and "GAATCTGCAAGTGGATATTTG" whereas for the

other two samples (MCF-7/ADR_1 and MCF-7/ADR_2), ADR drug motif sequence is "AGCTAGTTGGAAAC ACTTTTT" and "AAGGTAGACTTTGGAAATGGAGA TTTCCA".

Gene Ontology for Motifs: Gene ontology was performed for all four samples. Significant GO terms can be further studied to understand the biological roles of the identified motifs in all four samples. GO analysis shows that these O-GlcNAc protein motifs have olfactory receptor activity (MF), sensory perception of smell (BP) and are involved in the G-protein-coupled receptor protein signalling pathway (BP). ADR bound chromatin motif has olfactory receptor activity (MF), sensory perception of smell (BP) and is involved in the G-protein-coupled receptor protein signalling pathway (BP), as well as a dense response to the bacterium and extracellular region (CC).

Differential binding sites prediction using MACS2 bdgdiff: MACS2 bdgdiff was used for differential peak detection based on paired four bedgraph files for both condition 1 and 2. BED file shows the chromosomal view of different reads of identified differential binding sites of our two condition samples. The probe trend plot here, as shown in figure 9, allows us to look at differences in read density over our probes for two different conditions. It shows the overlap between reads of all samples across each position and it shows the trend of narrow peaks identified as differential binding sites.

Motif Discovery performed using MEME-ChIP: The FASTA files of extracted genomic DNA were used to generate the exact differential binding sites, also known as motifs or consensus regions for both conditions 1 and 2. MEME discovered novel, ungapped motifs (recurring, fixed-length patterns) in our two condition sequences as shown in figure 10. Condition 1 O-GlcNAc protein motif identified is "GGGAGGCCGAGACGG" and for condition 2, ADR drug motif sequence is "TGTTGGCCAGGCTGG".

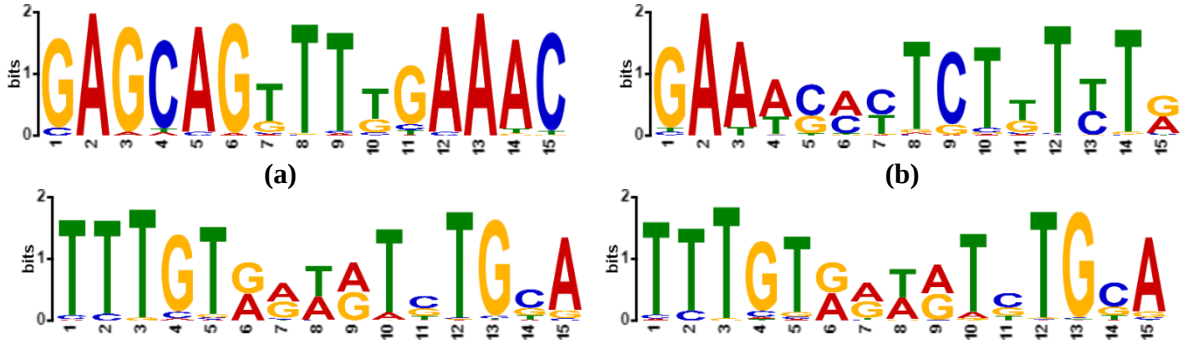


Figure 8: (a), (b), (c) and (d) showing the discovered motifs from each sample sequence respectively.

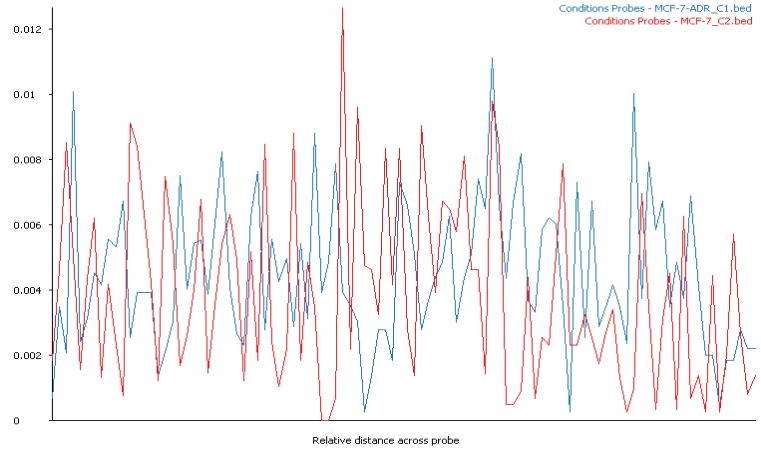


Figure 9: The trend of narrow peaks indicating the identified enriched binding sites of condition 1 and 2

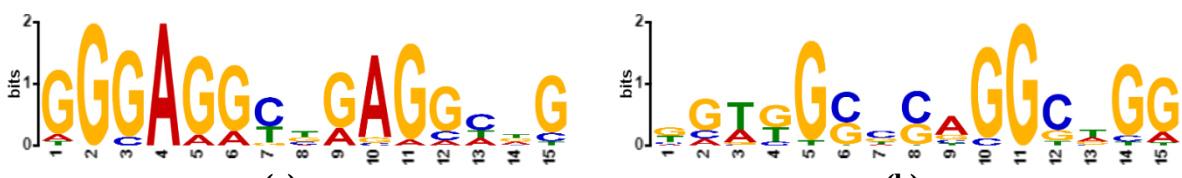


Figure 10: Figures (a) and (b) showing the discovered motifs from conditions 1 and 2 sequences respectively.

Gene Ontology for Motifs: Gene ontology was performed for both conditions. Significant GO terms can be further studied to understand the biological roles of the identified motifs for conditions 1 and 2. GO analysis shows that this motif has transcription factor activity, sequence-specific DNA binding (MF), cell development, the BP pattern specification process (BP) and MF histone binding (MF). Functional analysis shows that the condition 1 O-GlcNAc motif has functions related to TFAP2A_DBD_4 (Transcription factor AP-2-alpha) and KLF4 (Krueppel-like factor 4 protein).

ADR-bound chromatin motif has function transcription factor activity, sequence-specific DNA binding (MF), positive regulation of transcription RNA polymerase II promoter, negative regulation of signal transduction (BP) and transcription activator activity (MF).

Functional enrichment of the ADR binding site shows that the motif has functions related to Hic1_DBD (Hypermethylated in cancer 1 protein), TFAP2C_full_3 (Transcription factor AP-2 gamma) and PLAG1 (Zinc finger protein PLAG1).

Conclusion

ChIP-Seq Analysis of O-GlcNAc protein bound chromatin loci in human breast cancer cells is an important step towards understanding the role of this protein in inducing breast cancer. The O-GlcNAc protein binding site motif "GGGAGGCCGAGACGG" shows that it has functions related to TFAP2A_DBD_4 and KLF4**.** ADR bound chromatin has the motif "TGTTGGCCAGGCTGG," and functional enrichment of this motif shows that it has functions related to Hic1_DBD_1, TFAP2C_full_3 and PLAG1.

Acknowledgement

We would like to acknowledge the Amity Institute of Biotechnology, Amity University Uttar Pradesh, Lucknow campus, for providing us facilities to conduct this study.

References

1. Albrecht S., Sprang M., Andrade-Navarro M.A. and Fontaine J.F., SeqQscorer: automated quality control of next-generation sequencing data using machine learning, *Genome Biology*, **22**, 1-20 **(2021)**

2. Caldwell S.A., Jackson S.R., Shahriari K.S., Lynch T.P., Sethi G., Walker S., Vosseller K. and Reginato M.J., Nutrient sensor O-GlcNAc transferase regulates breast cancer tumorigenesis through targeting of the oncogenic transcription factor FoxM1, *Oncogene*, **29(19)**, 2831-2842 **(2010)**

3. Chen Q. and Yu X., OGT restrains the expansion of DNA damage signaling, *Nucleic Acids Research*, **44(19)**, 9266-9278 **(2016)**

4. Cheng X. and Hart G.W., Glycosylation of the murine estrogen receptor-α, *The Journal of Steroid Biochemistry and Molecular Biology*, **75(2-3)**, 147-158 **(2000)**

5. Feng J., Liu T., Qin B., Zhang Y. and Liu X.S., Identifying ChIP-seq enrichment using MACS, *Nature Protocols*, **7(9)**, 1728-1740 **(2012)**

6. Gao Y., Wells L., Comer F.I., Parker G.J. and Hart G.W., Dynamic O-glycosylation of nuclear and cytosolic proteins: cloning and characterization of a neutral, cytosolic β-N-acetylglucosaminidase from human brain, *Journal of Biological Chemistry*, **276(13)**, 9838-9845 **(2001)**

7. Gu Y., Mi W., Ge Y., Liu H., Fan Q., Han C., Yang J., Han F., Lu X. and Yu W., GlcNAcylation plays an essential role in breast cancer metastasis, *Cancer Research*, **70(15)**, 6344-6351 **(2010)**

8. Janetzko J. and Walker S., The making of a sweet modification: structure and function of O-GlcNAc transferase, *Journal of Biological Chemistry*, **289(50)**, 34424-34432 **(2014)**

9. Jiang M.S. and Hart G.W., A subpopulation of estrogen receptors are modified by O-linked N-acetylglucosamine, *Journal of Biological Chemistry*, **272(4)**, 2421-2428 **(1997)**

10. Kreppel L.K., Blomberg M.A. and Hart G.W., Dynamic glycosylation of nuclear and cytosolic proteins: cloning and characterization of a unique O-GlcNAc transferase with multiple tetratricopeptide repeats, *Journal of Biological Chemistry*, **272(14)**, 9308-9315 **(1997)**

11. Langmead B. and Salzberg S.L., Fast gapped-read alignment with Bowtie 2, *Nature Methods*, **9(4)**, 357-359 **(2012)**

12. Lazarus M.B., Jiang J., Kapuria V., Bhuiyan T., Janetzko J., Zandberg W.F., Vocadlo D.J., Herr W. and Walker S., HCF-1 is cleaved in the active site of O-GlcNAc transferase, *Science*, **342(6163)**, 1235-1239 **(2013)**

13. Machanick P. and Bailey T.L., MEME-ChIP: motif analysis of large DNA datasets, *Bioinformatics*, **27(12)**, 1696-1697 **(2011)**

14. Meienberg J., Zerjavic K., Keller I., Okoniewski M., Patrignani A., Ludin K., Xu Z., Steinmann B., Carrel T., Röthlisberger B. and Schlapbach R., New insights into the performance of human whole-exome capture platforms, *Nucleic Acids Research*, **43(11)**, 76 **(2015)**

15. Siegel R., Naishadham D. and Jemal A., Cancer statistics for hispanics/latinos, *CA: A Cancer Journal for Clinicians*, **62(5)**, 283-298 **(2012)**

16. Torres C.R. and Hart G.W., Topography and polypeptide distribution of terminal N-acetylglucosamine residues on the surfaces of intact lymphocytes. Evidence for O-linked GlcNAc, *Journal of Biological Chemistry*, **259(5)**, 3308-3317 **(1984)**

17. Trinca G.M. and Hagan C.R., O-GlcNAcylation in women's cancers: breast, endometrial and ovarian, *Journal of Bioenergetics and Biomembranes*, **50**, 199-204 **(2018)**

18. Trinca G.M., Goodman M.L., Papachristou E.K., D'Santos C.S., Chalise P., Madan R., Slawson C. and Hagan C.R., O-GlcNAc-dependent regulation of progesterone receptor function in breast cancer, *Hormones and Cancer*, **9**, 12-21 **(2018)**

19. Vickers N.J., Animal communication: when i'm calling you, will you answer too?, *Current Biology*, **27(14)**, 713-715 **(2017)**

20. Yu G., Wang L.G. and He Q.Y., ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization, *Bioinformatics*, **31(14)**, 2382-2383 **(2015)**

21. Zhong J., Martinez M., Sengupta S., Lee, A., Wu X., Chaerkady R., Chatterjee A., O'meally R.N., Cole R.N., Pandey A. and Zachara N.E., Quantitative phosphoproteomics reveals crosstalk between phosphorylation and O-GlcNAc in the DNA damage response pathway, *Proteomics*, **15(2-3)**, 591-607 **(2015)**.

(Received 20th February 2024, accepted 13th April 2024)