Biomedical Research and Therapy, 2022; 9(3):4971-4985

Open Access Full Text Article

Report

Motif prediction of abemaciclib in a breast cancer cell line using ChIP-Seq data analysis

Utkarsh Deep, Ruchi Yadav*

Use your smartphone to scan this QR code and download this article

ABSTRACT

Introduction: Chromatin immunoprecipitation sequencing (ChIP-Seq) is a DNA sequencing technique for the identification of binding sites in genomic sequences. ChIP-Seq experiments are a combination of immunoprecipitation and sequencing techniques that are used for the identification of chromatin regions that bind molecules such as transcription factors (TFs), histones, and drugs. In this study, computational analysis of ChIP-Seq data was performed to predict the binding sites in breast cancer cells and their association with several molecules, such as TFs and drugs. A complete and comprehensive computational study has been performed to predict the binding sites of abemaciclib. Functional enrichment of selected motifs was performed to identify important motifs that function in breast cancer and show binding with the drug abemaciclib. **Methods and Materials:** The ChIP-Seq analysis protocol was performed using the Galaxy server (https://usegalaxy.org/). The abemaciclib binding motif was identified using MEME tools. For this research, ChIP-Seq data from a breast cancer cell line was retrieved from the GEO database, accession number GSM4763932. This dataset includes ChIP-Seq of MCF-7 cells exposed to abemaciclib. The ENA browser was used to retrieve the data. Statistical analysis was performed using the default parameters of Fast-QC, Multi-QC, Map with BWA, Filter SAM and BAM, MACS2, and ChIP Seeker tools on the Galaxy server. **Results:** Computational analysis identifies the abemaciclib consensus binding sequence as TGGCTCACGCCTGTAATCCCAGCACTTT, and this motif occurs 2980 times in the *Homo sapiens* reference genome hg19. **Conclusions:** This study identifies the binding sites and affinity of abemaciclib in a breast cancer cell line.

Key words: ChIP-Seq, Motif, Binding site, Abemaciclib, Galaxy, MCF-7

INTRODUCTION

Cancer of the breast is uncontrolled production of breast cells[1]. Breast cancer is the second most common cancer diagnosed in women in the United States after skin cancer. Breast cancer may occur in both men and women, but it is much more prevalent in women. It is the most prevalent cause of death from cancer among women worldwide[2]. In more developed countries, incidence rates are high, while rates in less developed countries are low but rising rapidly. Genomics studies have revealed genes associated with breast cancer. The identification and functional understanding of these genes are very important for the molecular diagnostics and drug development industries. Breast cancer genes such as *BRCA1*, *BRCA2*, *CHEK2*, *ATM* and *PALB2* have been widely studied since these genes play a major role in breast cancer development[3]. Cyclin-dependent kinases 4 and 6 (CDK4/6) are pharmacological inhibitors and have a significant effect on the practice of oncology[4]. They are routinely recommended for estrogen receptor positive breast cancer therapy, and many studies are underway to assess their activity against other forms of cancer[5]. CDK4/6 mediates the cell cycle transition from G1 to S phase, and CDK4/6 inhibitors cause arrest of the G1 phase in tumor cells[6]. Other phenotypes in cancer cells could also be activated, including enhanced immunogenicity, apoptotic evasion, histologic tumor differentiation, and increased dependence on receptor tyrosine kinase signaling[7]. Abemaciclib is a CDK4- and CDK6-selective ATP-competitive, reversible kinase inhibitor that has demonstrated antitumor activity in clinical trials as a single agent in hormone receptor-positive (HR+) metastatic breast cancer[8].

With advancements in next-generation sequencing technologies, genome-wide screening of binding sites can be performed by sequencing chromatin associated with proteins. Chromatin immunoprecipitation (ChIP-Seq) is used to identify the binding sites of proteins of interest at the genomic scale. ChIP-Seq analysis is used to identify the binding sequence for any drug or transcription factor[9]. Transcription factors are proteins that bind upstream of genes to enhance

Amity Institute of Biotechnology, Amity University Uttar Pradesh, Lucknow Campus, Lucknow, UP, India

Correspondence

Ruchi Yadav, Amity Institute of Biotechnology, Amity University Uttar Pradesh, Lucknow Campus, Lucknow, UP, India

Email: ryadav@lko.amity.edu

History

- Received: Nov 26, 2021
- Accepted: Mar 20, 2022
- Published: Mar 31, 2022

DOI : 10.15419/bmrat.v9i3.732

Check for updates

BioMedPress
The Open Access Publisher

Cite this article : Deep U, Yadav R. **Motif prediction of abemaciclib in a breast cancer cell line using ChIP-Seq data analysis**. *Biomed. Res. Ther.,* 2022; 9(3):4971-4985.

or initiate transcription [10]. Many genes are responsible for causing breast cancer, as described above [11]. As the main genes responsible for tumor formation in breast cancer are *BRCA1* and *BRCA2*, ChIP-Seq can be used to identify other transcription factors that initiate the activation of BRCA1 or BRCA2 and facilitate the formation of tumors, causing cancer [12].

ChIP-Seq is an effective technique for defining genome-wide DNA-binding sites for transcription factors and other proteins by combining chromatin immunoprecipitation (ChIP) with sequencing [13]. The DNA-bound protein is immunoprecipitated using a specific antibody [14]. Bound DNA is co-precipitated, purified, and sequenced. ChIP-Seq can be used to identify the binding sites of proteins associated with DNA and can be used to map a given protein's genome-wide binding sites [15]. Usually, ChIP-Seq begins with DNA–protein complex crosslinking. The samples are fragmented and processed to trim unbound oligonucleotides with an exonuclease. To immunoprecipitate a DNA–protein complex, protein-specific antibodies are used [16]. The DNA is extracted and sequenced, thereby identifying the sequences of the protein-binding sites with high resolution [17]. Insights into gene regulation events that play a role in different diseases and biological pathways, such as development and cancer progression, have been uncovered by the application of next-generation sequencing (NGS) to ChIP. ChIP-Seq allows the interactions between proteins and nucleic acids to be thoroughly studied on a genome-wide scale [18].

Michigan Cancer Foundation-7 (MCF-7) is a breast cancer cell line that was first isolated from a 69-year-old Caucasian women in 1970 [19]. It is a well-known cell line used worldwide. It has features of mammary epithelial differentiation and is positive for many epithelial markers, such as β-catenin, cytokeratin 18 (CK18) and E-cadherin [20]. It is also negative signal for mesenchymal markers such as smooth muscle actin and vimentin [21]. MCF-7 cells maintain the expression of other specific molecular markers of epithelium and do not express CD-44 ligand.

Abemaciclib is an antitumor agent and a dual cyclin-dependent kinase 4 (CDK4) and 6 (CDK6) inhibitor. CDK4/6 are involved in the cell cycle and in the event of unregulated activity, promote cancer development [22]. On September 28, 2017, the FDA approved abemaciclib for the treatment of HR-positive and HER2-negative advanced or metastatic breast cancer that has progressed following failed endocrine therapy, under the brand name Verzenio [23]. It is an oral inhibitor of cyclin-dependent kinase (CDK) that targets the cell cycle pathways of CDK4 (cyclin D1) and CDK6 (cyclin D3) with potential antineoplastic activity [24]. Abemaciclib inhibits CDK4 and 6 directly, thus inhibiting the phosphorylation of retinoblastoma (Rb) protein in early G1 [25]. Inhibition of Rb phosphorylation prevents the transition of the CDK-mediated G1-S process, thereby arresting the cell cycle in G1 phase, suppressing DNA synthesis and inhibiting the growth of cancer cells [26]. The interaction of abemaciclib with different genes such as *AKT1*, *BARD1*, and *CDK4*, and the functional consequences thereof, is shown in Table 1. Identification of binding sites for drugs is important to understand their function. Any mutation in the genome of the cell that disrupts a binding sites will result in loss of binding of the drug. In order for a drug to be effective in a mutated cell, such as a cancer cell, it is essential for the drug to bind its target sequence. Keeping this in mind, breast cancer data were used in the current study to identify the core binding sites in the genome of a breast cancer cell line. Functional analysis was also performed to identify functional sequences and core binding sites.

METHODS

Data Retrieval

The SRA (Short Read Archive) database was used for ChIP-Seq data retrieval. GSM4763932 ID was seen relevant because the study was on the MCF-7-cell line treated with abemaciclib. The SRA ID SRA1120692 was used. Two samples with accession numbers SRR12576544 and SRR12576545 were selected for motif enrichment and analysis. The ENA browser was used to retrieve the data (https://www.ebi.ac.uk/ena/browser/text-search?query=SRA1120692). The SRA IDs of MCF-7 DMSO and MCF-7 abemaciclib samples are shown in **Table 2**. FASTQ files were uploaded to the Galaxy server https://usegalaxy.org/ for analysis and annotation.

The Galaxy (https://usegalaxy.org/) web platform was used for the analysis and annotations of ChIP-Seq data. The methodology is shown in **Figure 1**. The tools used in Galaxy were Fast-QC, Multi-QC, Map with BWA, Filter SAM and BAM, MACS2, and ChIP Seeker. MEME Multiple EM for motif elicitation was used for motif identification at http://meme-suite.org/tools/meme. The steps followed for the identification of motifs are described below.

Step I: Fast-QC [34]

The Fast-QC tool was used to perform a quality check on raw sequence data. This offers a modular collection of tests that can be used to provide an indication of whether data have any issues that should be

Table 1: Genes accosiated with Abemaciclib Drug

S.No	PubChem Gene	Interaction with Abemaciclib drug	Evidence PMID
1	ABCB1 and ABCB2	Abemaciclib results in decreasing the activity of ABCB1 and ABCB2 protein [27]	27816545
2	AKT1	Abemaciclib results in decreased phosphorylation of and results in decreased activity of AKT1 protein [28]	26909611
3	BARD1	Abemaciclib results in decreased expression of BARD1 mRNA [29]	28620137
4	CDK4	Abemaciclib results in decreased activity of [CDK4 protein binds to CCND1 protein] [30]	24919854
5	RB1	Abemaciclib results in decreased phosphorylation of RB1 protein [31]	26909611
6	SPARC	Abemaciclib results in increased expression of SPARC mRNA [32]	28620137
7	WNT4	Abemaciclib results in decreased expression of WNT4 mRNA [33]	28620137

resolved before further processing. SRR12576544 and SRR12576545 were used as inputs for the Fast-QC analysis. The outputs were in HTML and .txt format. Fast-QC tools produce statistical values that can be interpreted to identify filtering or preprocessing steps.

Step II: Multi-QC [35]

Multi-QC aggregates data from different samples into a single report. It compiles all results into a single HTML report that can be used for comparative study. The inputs for the Multi-QC tool were the Fast-QC reports, and the compiled results were in the form of plots.

Step III: Map with BWA [36]

This tool is used to align short sequences to the large sequence database. SRR12576544 and SRR12576545 in fasta.qz file format were used as inputs for the map with BWA tool. The output was in BAM format, which can be visualized in the UCSC, bam.iobib.io and IGV genome browsers.

Step IV: Filter SAM or BAM [37]

This tool uses the samtools view commands in SAM Tools to filter SAM and BAM files on the mapping quality. This is a unique way of mapping by removing the non-required sequence. The input file is the result of the map with the bam tool, and the output is in BAM format.

Step V: Multi Bam Summary [38]

For two or more BAM files, multiBamSummary calculates the read coverage for genomic regions. The input was the result from the Filter SAM or BAM tool, and output was in the deeptools_coverage_matrix format. This result can be used for comparison of genomic alignments in different samples.

Step VI: plot Correlation [39]

This tool was used to analyze and visualize sample correlations based on multiBamSummary or multiBigwigSummary. The input was the result of multiBamSummary, and the output was in png format. This plot was used for the comparison of variations in sample files.

Step VII: plot fingerprint [40]

This tool samples indexed BAM files and graphs a cumulative read coverage profile for each file. The input was the same as for plot correlation, and the output was also in png format. This graph gives detailed information about the location of binding sites in genomic regions.

Step VIII: bamCoverage [41]

This method takes an alignment of reads or fragments as input (BAM file) and produces as output a coverage track (bigWig or bedGraph).

Step IX: MergeSamFiles [42]

Similar to the ”merge” feature of Samtools. This tool is used to combine SAM or BAM files from separate

runs and read groups into a single file.

Step X: Split[43]

This is used to separate a combined file from a single BAM file. Aligned reads corresponding to the original four datasets are included in every subsequent BAM file.

Step XI: MACS2[44]

ChIP-Seq model-based analysis (MACS) is a widely used method for identification of transcription factor binding sites. To determine the importance of enriched ChIP regions, the MACS algorithm captures the effect of genome complexity. It is also suitable for wider areas, although it was developed for the detection of transcription factor binding sites. The input was the result of Split, and the output was in the form of bed format for submit peaks and narrow peaks.

Step XII: ChIPSeeker[45]

ChIPseeker is a Bioconductor package in Galaxy for annotating ChIP-Seq data. The annotatePeak feature performs peak annotation. In addition to the distance from the peak to the TSS of its closest gene, the location and strand information of the nearest genes were recorded. The input file is obtained from the gencode tool, and the output is in the form of plots.

Step XIII: MEME-ChIP[46]

MEME-ChIP performs motif analysis on large sets of sequences, such as those identified by ChIP-Seq experiments. The input was obtained from the Extract Genomic DNA tool, and the results were in HTML format.

Step XIV: plot Heatmap[47]

This tool creates a heatmap for scores associated with genomic regions. This plot is used to visualize the genomic locations of predicted binding sites.

RESULTS

Following the protocol described in **Figure 1**, Fast-QC was performed for SRR12576544 and SRR12576545 samples in Galaxy, followed by Multi-QC. The input file used for this workflow was in fastq format, and the output file generated from Multi-QC was in HTML format. The results of Multi-QC are shown in **Table 3**. The comparison between sample files in terms of the number of duplicate reads, GC content and maximum length (M Seq) shows that both files can be used for further study. **Figure 2** shows the number of duplicate and unique reads in both samples.

The BAM file was generated from the fastq.gz file by using Map with the BWA tool of Galaxy. The alignment between the sample file and the reference genome, hg19, was performed with the tool Map with BWA to identify the similarity between read sequences and genomic sequences. The results were in bam.iobio, IGV and UCSC formats, which were visualized using https://bam.iobio.io/home, Integrative Genomics Viewer (IGV) (https://software.broadinstitute.org/software/igv/) and the UCSC Genome Browser (https://genome.ucsc.edu/), respectively. The data used for this study satisfied all the parameters of quality control and were of very high quality, so the data were not trimmed or postprocessed in any way before mapping. **Figure 3** and **Figure 4** show the result of alignment of the read sequences with the hg19 reference genome. **Figure 3** shows the alignment between the MCF-7 vehicle control reads and the reference human genome (SRR12576544 to hg19). **Figure 4** shows the alignment between the MCF-7 abemaciclib sample and the reference genome (SRR12576545 to hg19). Both samples had more than 95% alignment with the reference genome.

All non-uniquely mapped reads were identified. This was accomplished by simply filtering out all reads with a mapping quality of less than 20 using NGS:SAMtools (SAM or BAM filter). Then, the mapping quality of the mapped reads was analyzed. The multiBamSummary tool was used to calculate the average read coverage for a list of two Bam files. Then, the plotCorrelation tool was used to create a heatmap of correlation scores between the samples, as shown in **Figure 5**, demonstrating that both files have approximately the same number of aligned regions.

plotFingerprint was used to identify the signal strength of the ChIP sample. **Figure 6** shows the comparison of MCF-7 DMSO (SRR12576544, red) and MCF-7 abemaciclib (SRR12576545, green), which shows that there might be some variation in the aligned regions that can be explored in further studies.

The input for the MACS2 peak calling tool was obtained by merging the BAM file with the tool MergeSamFiles, and then the split tool was used to separate the merged file into the individual BAM files. Aligned reads corresponding to the original two datasets were included in each subsequent BAM file. Peak calling was performed using the MACS2 package, the results of which are shown in **Figure 7**. This tool finds the best parameters for the MACS2 peak calling.

The ChIPseeker tool was used to annotate the ChIP-Seq data. It annotates ChIP peaks and provides functionality to visualize the coverage of ChIP peaks over

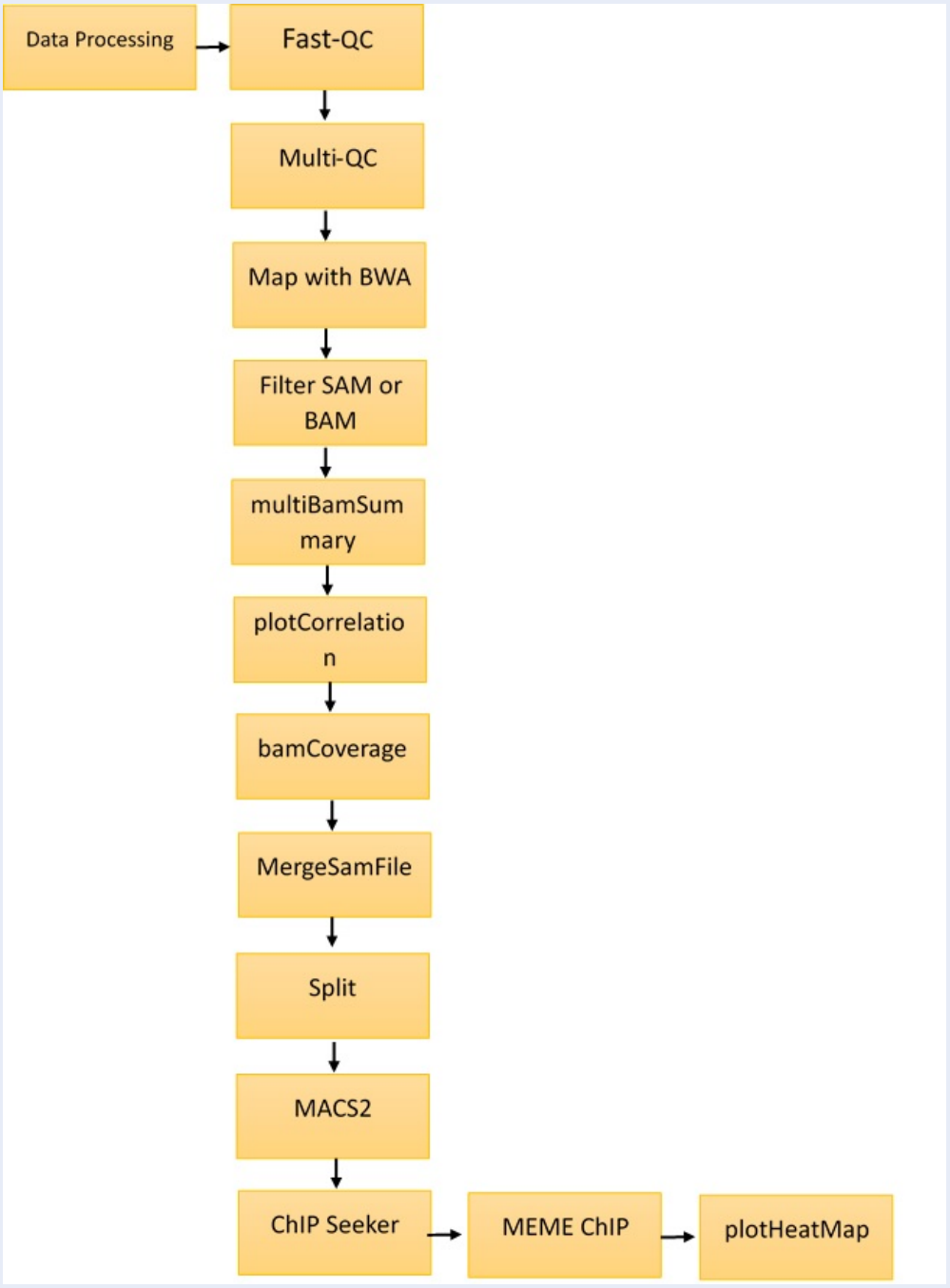

Figure 1: **Workflow used for ChiP-Seq data analysis and motif identification**. This workflow shows the data pre-processing steps, tools used for genomic alignment, prediction of motifs, and functional enrichment of motifs.

Table 2: SRR ids with their GSM number used for this study

S.No	Name	SRR number	GSM number
1	MCF-7 DMSO	SRR12576544	GSM4763931
2	MCF-7 Abemaciclib	SRR12576545	GSM4763932

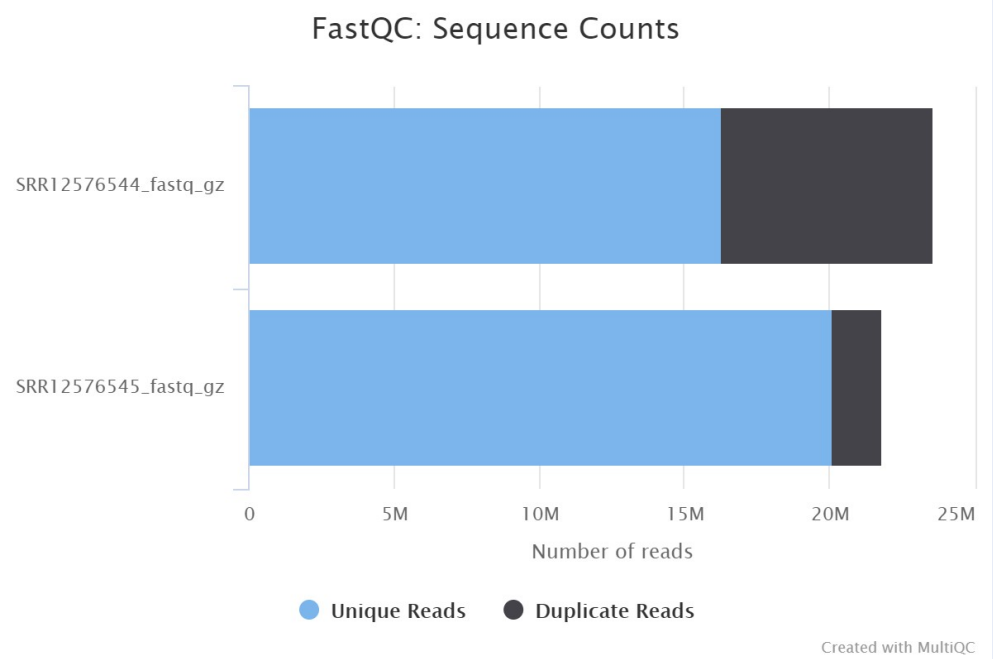

Figure 2: **Number of duplicate and unique reads in ChiP-Seq files of MCF-7 DMSO (SRR12576544) and MCF-7 Abemaciclib (SRR12576545)**. This comparison shows the number of duplicated reads in both the samples, that is 15 M in (SRR12576544) and 20 M in (SRR12576545).

Table 3: Multi-QC result showing Duplicates, GC content and Maximum length sequence

Sample Name	% Dups	% GC	M Seqs
SRR12576544_fastq_gz	31.0%	49%	23.6
SRR12576545_fastq_gz	8.0%	46%	21.8

chromosomes and peak profiles that bind to TSS regions. ChIP peak profile comparison and annotation are also provided. In addition, it supports the evaluation of substantial overlap between ChIP-Seq datasets. Currently, ChIPseeker includes 17,000 GEO database bed file details. The input was given after performing the gene pool, which helped identify promoter, exon, intron, UTR, and downstream regions.

The output was in the form of plots and Excel files, which give all introns and exons with chromosome numbers. The result of the peak annotation is shown in **Figure 8**, which shows the percentage of identified peaks in promoter, exon, intron, and downstream regions. This is important to identify functional regions that show alignment with sample files. The reference genome result shows that there are alignment regions in almost all functional regions of the genome, such as promoters, genes, and exons.

De novo motif discovery, motif enrichment analysis, motif position analysis and motif clustering were performed using MEME-ChIP, providing a detailed image of the motifs that are enriched in the input sequences. Two complementary types of de novo motif discovery are performed by MEME-ChIP: weight matrix-based discovery for high accuracy and word-based discovery for high sensitivity. The study of motif enrichment using human, mouse, worm, fly and other model organism motifs offers even greater sensitivity.

The Extract Genomic DNA tool was used to generate genomic sequences corresponding to the ChIP peaks. **Figure 9** shows the motif that was predicted using MEME ChIP, and **Figure 10** shows the motif TGGCTCACGCCTGTAATCCCAGCACTTT, which was identified in 2980 positions in the *Homo sapiens* reference genome hg19.

A heatmap was generated to visualize the position of the motif in the whole genome (**Figure 11**). The blue region in the plot depicts the region of all motifs in the whole genome, showing the regions across

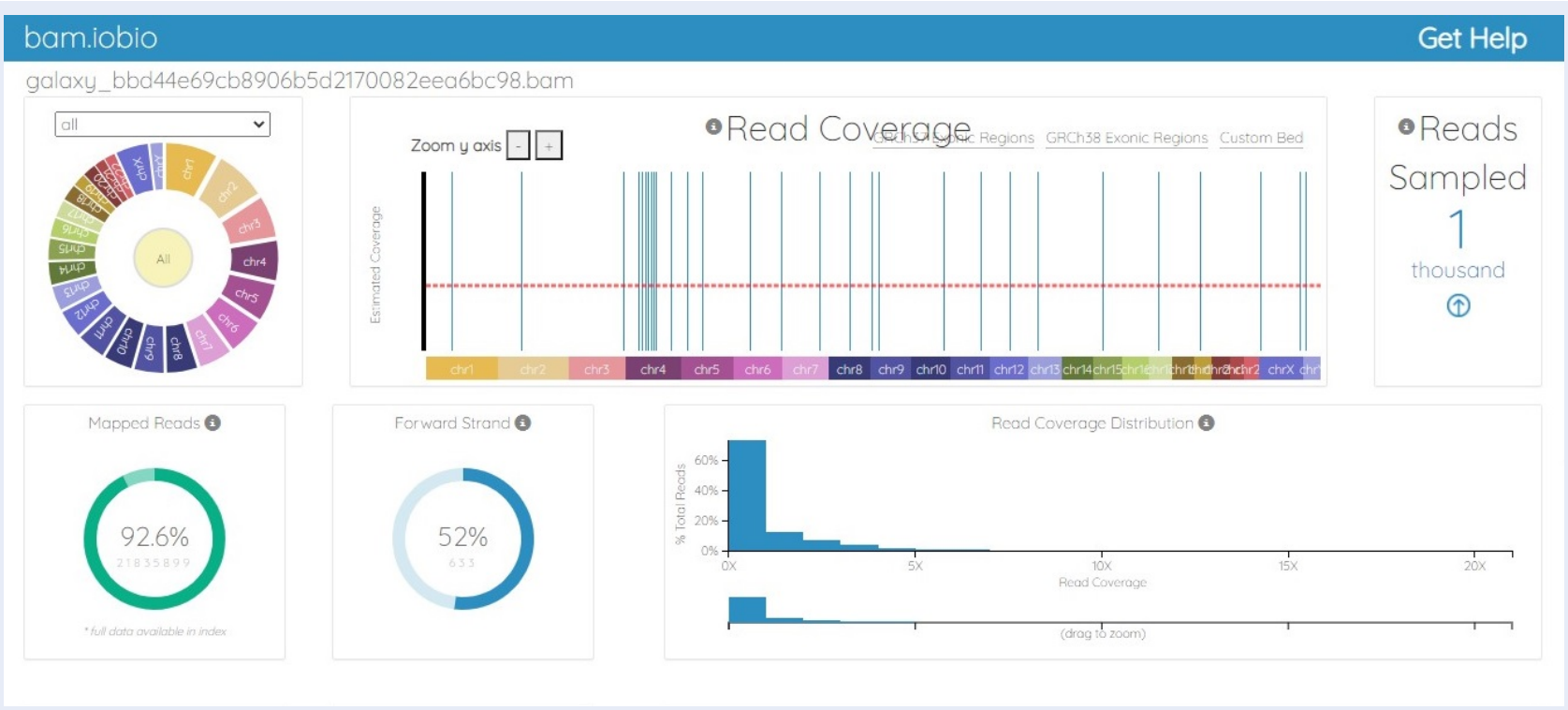

Figure 3: **Genomic alignment of SRR12576544 to reference genome hg19.** This data details the number of read alignments in each chromosome and for the visualization of genomic alignment. The result shows 92.6% mapped reads across the genome.

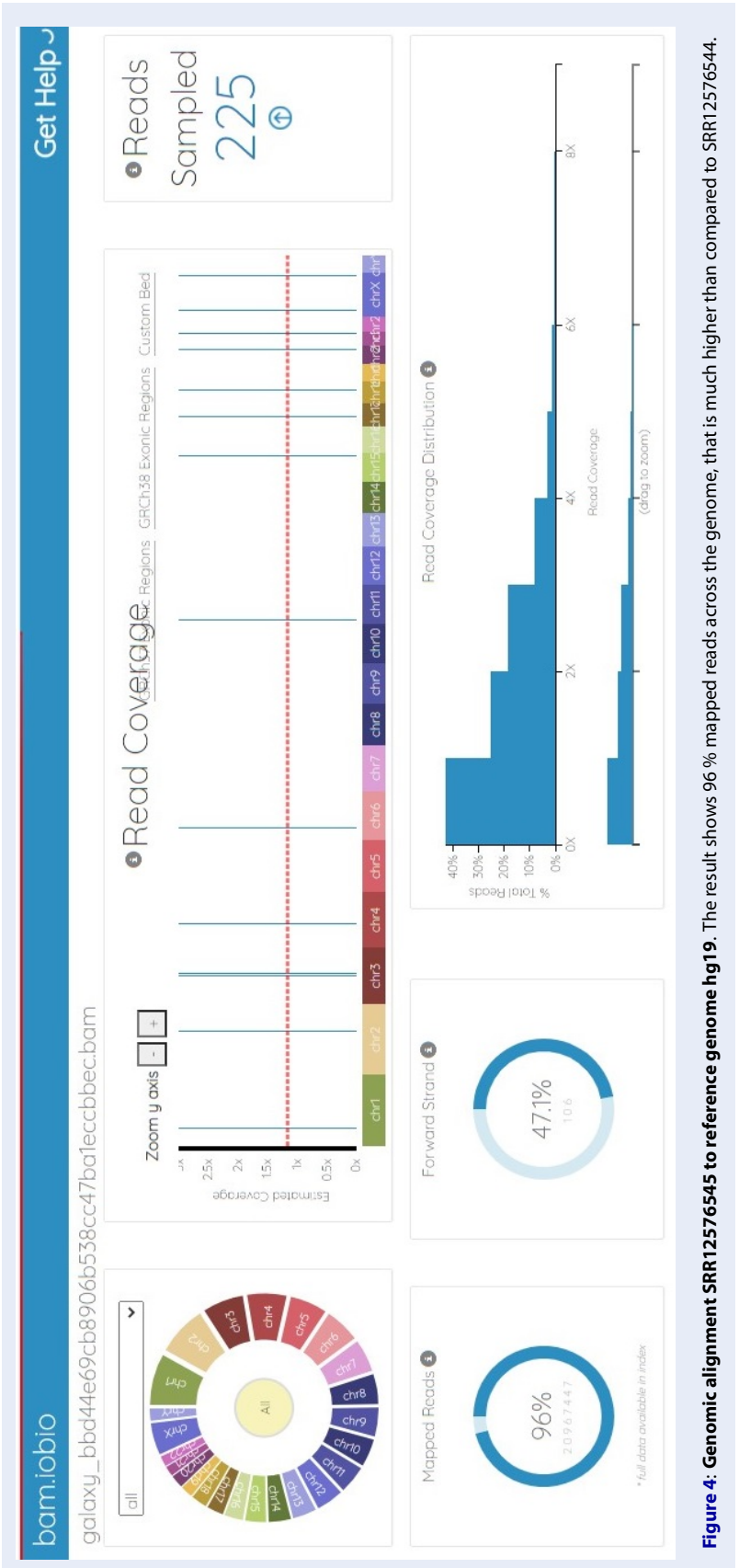

Figure 4: Genomic alignment SRR12576545 to reference genome hg19. The result shows 96 % mapped reads across the genome, that is much higher than compared to SRR12576544.

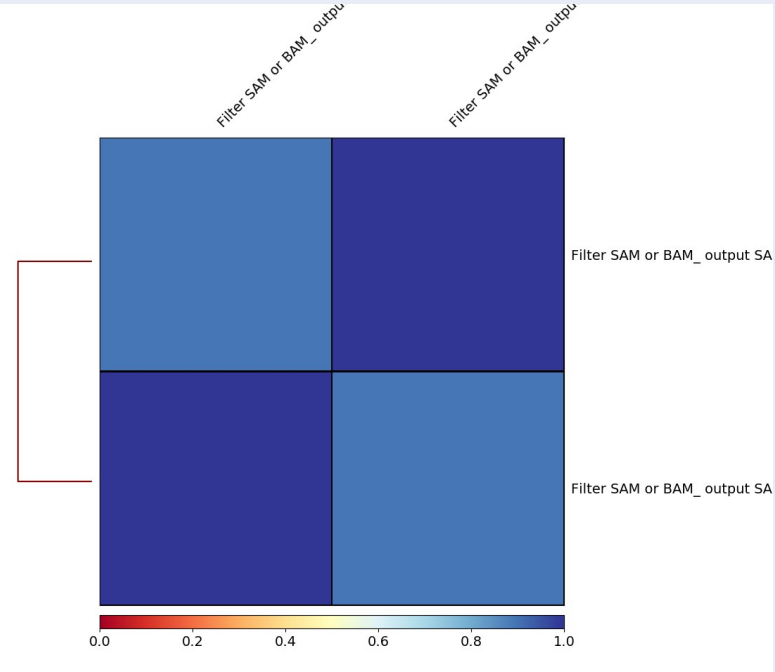

Figure 5: **Heatmap Correlation between aligned reads of both the files MCF-7 DMSO (SRR12576544) and MCF-7 Abemaciclib (SRR12576545)**. This data compares the percentage of aligned reads in approximately same sample file.

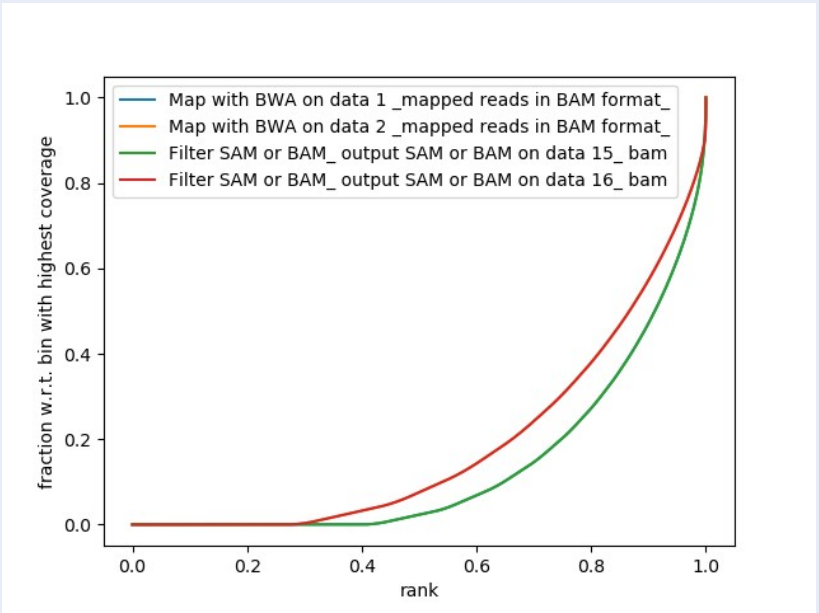

Figure 6: **Fingerprint plot of both the samples produced using plot fingerprint tool**. It samples indexed BAM files and graphs a cumulative read coverage profile for each sample. It shows (SRR12576545, red line) has more read coverage.

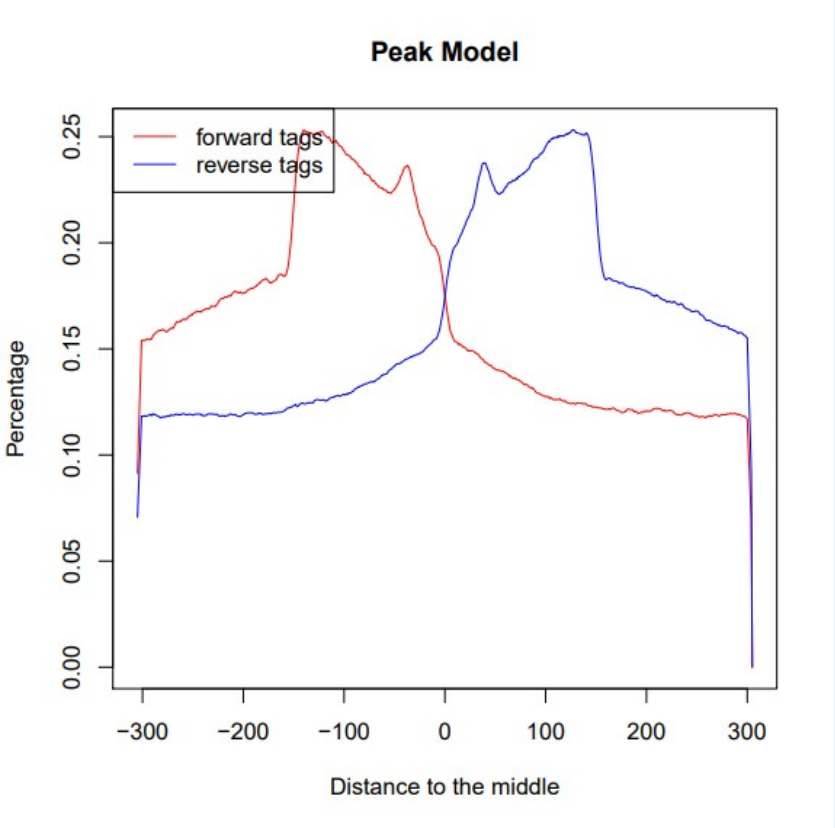

Figure 7: **MACS2 peak result shows the percentage of DNA-Protein interaction in both the strands forward and reverse strands**. Enrichment of binding sites in both the strands is compared to verify the predicted binding sites.

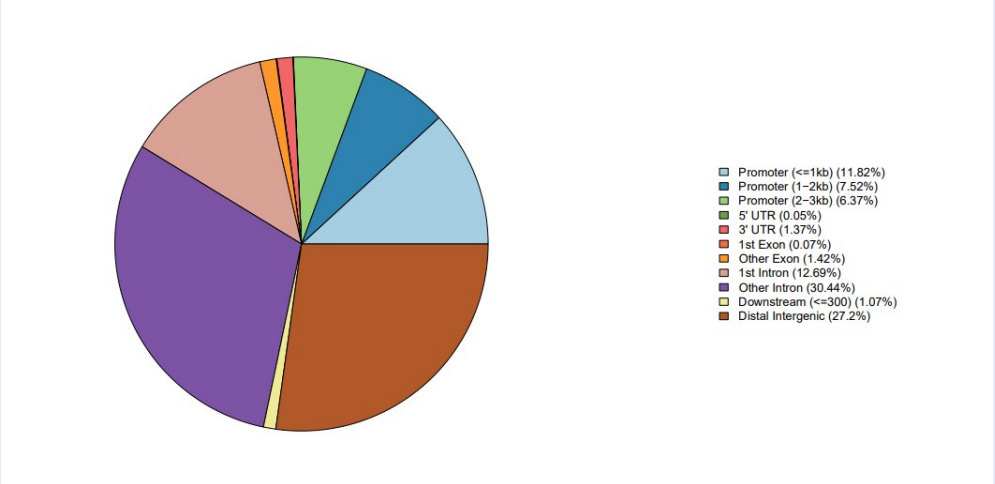

Figure 8: **Pie graph of identified peaks in MCF-7 Abemaciclib (SRR12576545)**. Across different genomic functional regions with percentages identified using ChIPSeeker. It shows binding sites are present in functional regions like promoter 11.82%, Exon 0.07%, and others are in the intergenic region 27.2%.

Figure 9: **MEME motifs found in 2980 sequences corresponding to common peak regions in MCF-7 Abemaciclib (SRR12576545)**. This motif was predicted using the MEME ChIP tool, and it shows that the Abemaciclib drug has a binding affinity toward this predicted motif.

Motif	Sequence Name	Strand	Start	End	p-value	q-value	Matched Sequence
1	chr17	-	1466588	1466615	9.31e-18	3.75e-13	TGGCTCACGCCTGTAATCCCAGCACTTT
1	chr11	-	1580748	1580775	9.31e-18	3.75e-13	TGGCTCACGCCTGTAATCCCAGCACTTT
1	chr11	-	1793869	1793896	9.31e-18	3.75e-13	TGGCTCACGCCTGTAATCCCAGCACTTT
1	chr4	-	6690971	6690998	9.31e-18	3.75e-13	TGGCTCACGCCTGTAATCCCAGCACTTT
1	chr5	-	10962574	10962601	9.31e-18	3.75e-13	TGGCTCACGCCTGTAATCCCAGCACTTT
1	chr17	-	15602887	15602914	9.31e-18	3.75e-13	TGGCTCACGCCTGTAATCCCAGCACTTT
1	chr16	-	16116164	16116191	9.31e-18	3.75e-13	TGGCTCACGCCTGTAATCCCAGCACTTT
1	chr7	-	16793309	16793336	9.31e-18	3.75e-13	TGGCTCACGCCTGTAATCCCAGCACTTT
1	chr2	-	27851799	27851826	9.31e-18	3.75e-13	TGGCTCACGCCTGTAATCCCAGCACTTT
1	chr13	-	28000284	28000311	9.31e-18	3.75e-13	TGGCTCACGCCTGTAATCCCAGCACTTT
1	chr1	-	28764533	28764560	9.31e-18	3.75e-13	TGGCTCACGCCTGTAATCCCAGCACTTT
1	chr22	-	30375217	30375244	9.31e-18	3.75e-13	TGGCTCACGCCTGTAATCCCAGCACTTT
1	chr16	-	30441193	30441220	9.31e-18	3.75e-13	TGGCTCACGCCTGTAATCCCAGCACTTT
1	chr19	-	33532299	33532326	9.31e-18	3.75e-13	TGGCTCACGCCTGTAATCCCAGCACTTT
1	chr20	-	34166523	34166550	9.31e-18	3.75e-13	TGGCTCACGCCTGTAATCCCAGCACTTT
1	chr6	-	36647323	36647350	9.31e-18	3.75e-13	TGGCTCACGCCTGTAATCCCAGCACTTT
1	chr2	-	38374152	38374179	9.31e-18	3.75e-13	TGGCTCACGCCTGTAATCCCAGCACTTT
1	chr20	-	39695416	39695443	9.31e-18	3.75e-13	TGGCTCACGCCTGTAATCCCAGCACTTT
1	chr15	-	41138234	41138261	9.31e-18	3.75e-13	TGGCTCACGCCTGTAATCCCAGCACTTT
1	chr11	-	47256301	47256328	9.31e-18	3.75e-13	TGGCTCACGCCTGTAATCCCAGCACTTT
1	chr17	-	47755519	47755546	9.31e-18	3.75e-13	TGGCTCACGCCTGTAATCCCAGCACTTT
1	chr20	-	50180455	50180482	9.31e-18	3.75e-13	TGGCTCACGCCTGTAATCCCAGCACTTT

Figure 10: **Identification of motif "TGGCTCACGCCTGTAATCCCAGCACTTT", which was identified in the 2980 positions in the *Homo sapiens* reference genome version hg19**. This details about presence of predicted motif in different chromosomes along with their start and end positions.

the genome that represent motifs and binding sites. Heatmap results verify that abemaciclib has binding regions in this breast cancer cell line, and motifs identified are true motifs that have function in cells.

DISCUSSION

ChIP-Seq is a next-generation sequencing technology that is used to identify binding sites genome-wide. This technique has been explored in the drug development industry to characterize the effects of drugs and identify conserved regions of their binding sites. Since drug binding site identification is important to study the functionality and effectiveness of drugs, ChIP-Seq is used to study novel drugs and their effectiveness. It can also be used to screen ligand libraries for binding efficiency in cancerous cells[48]. Studies of various diseases and cancers have identified variations in binding sites and differential binding sites for drugs[49]. ChIP-Seq technology has been explored in different dimensions to study breast cancer cells and to identify novel drugs that can stop breast cancer[50]. Studies have been carried out mainly to identify drugs against breast cancer[51]. In this study, computational analysis of ChIP-Seq data of breast cancer cells was performed to identify binding sites of the drug abemaciclib. The binding motif for abemaci-

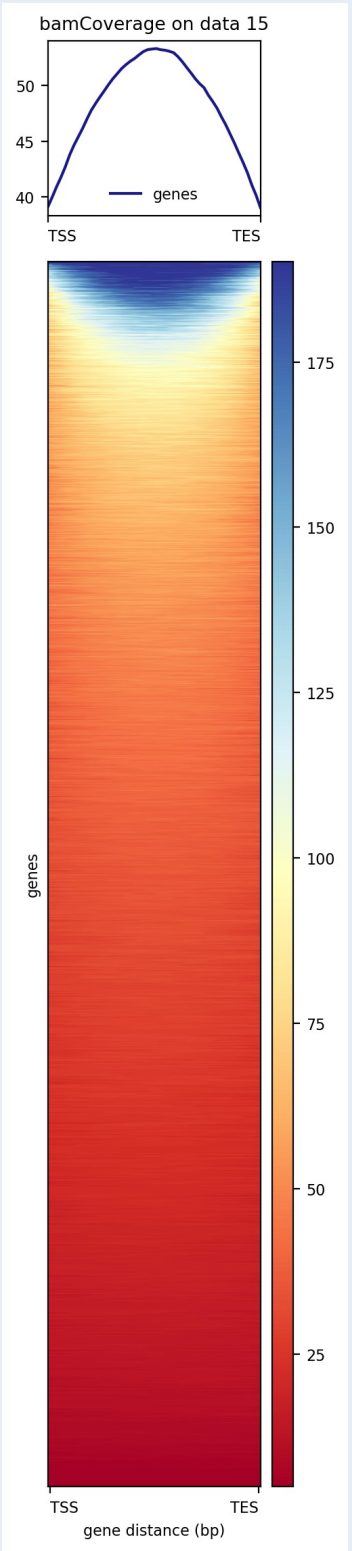

Figure 11: **Heatmap plot shows the percentage of predicted motifs across the genome**. Blue region in the plot depicts the region of unique motifs predicted in the sample files. Heatmap results confirms that Abemaciclib has binding regions in this breast cancer cell line and predicted motifs have function in cells.

clib was identified as TGGCTCACGCCTGTAATCCCAGCACTTT. Mapping of the identified motif to the human reference genome shows that the predicted motif occurs in 2980 positions in the hg19 reference genome. The results identify the binding sites for abemaciclib in the MCF-7 genome. This study can be used to better understand the function of abemaciclib against breast cancer and provide insight into its conserved and functional region in the human genome.

CONCLUSIONS

Breast cancer is the most common type of cancer in women. Breast cancers can become metastatic when breast cancer cells spread to different organs of the body through blood vessels or nodes. Currently, abemaciclib is used to treat breast cancer, and studies have been performed to identify the role of abemaciclib in breast cancer. In this research, ChIP-Seq data were retrieved from the SRA database, and the interaction between abemaciclib and chromatin was studied. These data were selected to identify the motif that abemaciclib binds to, which can be explored in different dimensions to understand the function of the motif and the drug.It can be used in toxicology studies, or in studies of the effect of abemaciclib in treating breast cancer. Further verification and wet lab study are required to study the function of the identified motif and its interaction with abemaciclib.

ABBREVIATIONS

ChIP-Seq: Chromatin immunoprecipitation sequencing, **TFs**: transcription factors, **MCF**: Michigan Cancer Foundation-7, **SRA**: Short Read Archive, **QC**: Quality Control, **SAM**: Sequence Alignment Map, **BAM**: Binary Alignment Map, **MACS2**: Model-based Analysis of ChIP-Seq, **MEME**: Multiple Expectation maximizations for Motif Elicitation

ACKNOWLEDGMENTS

We would like to acknowledge Amity Institute of biotechnology, Amity University Uttar Pradesh, Lucknow campus for providing us facilities to conducting this study. This research project is not funded by any specific grant from funding agencies in the public, commercial, or non-profit sectors.

AUTHOR'S CONTRIBUTIONS

Author Utkarsh Deep has done the work on methodology and interpretation of results and Ruchi Yadav contributed as framing methodology and analysis of the results. All authors read and approved the final manuscript.

FUNDING

None.

AVAILABILITY OF DATA AND MATERIALS

Data and materials used and/or analyzed during the current study are available from the corresponding author on reasonable request.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

CONSENT FOR PUBLICATION

Not applicable.

COMPETING INTERESTS

The authors declare that they have no competing interests.

REFERENCES

1. Vasconcelos AG, Valim MO, Amorim AG, do Amaral CP, de Almeida MP, Borges TK, et al. Cytotoxic activity of poly-ε-caprolactone lipid-core nanocapsules loaded with lycopene-rich extract from red guava (Psidium guajava L.) on breast cancer cells. Food Research International. 2020;136:109548. PMID: 32846600. Available from: 10.1016/j.foodres.2020.109548.
2. Wu C, Wei Q, Utomo V, Nadesan P, Whetstone H, Kandel R. Side population cells isolated from mesenchymal neoplasms have tumor initiating potential. Cancer Research. 2007;67(17):8216–22. PMID: 17804735. Available from: 10.1158/0008-5472.CAN-07-0999.
3. Coppi PD, Callegari A, Chiavegato A, Gasparotto L, Piccoli M, Taiani J. Amniotic fluid and bone marrow derived mesenchymal stem cells can be converted to smooth muscle cells in the cryo-injured rat bladder and prevent compensatory hypertrophy of surviving smooth muscle cells. The Journal of Urology. 2007;177(1):369–76. PMID: 17162093. Available from: 10.1016/j.juro.2006.09.103.
4. Singh-Ranger G, Mokbel K. Current concepts in cyclooxygenase inhibition in breast cancer. Journal of Clinical Pharmacy and Therapeutics. 2002;27(5):321–7. PMID: 12383132. Available from: 10.1046/j.1365-2710.2002.00435.x.
5. Goel S, DeCristo MJ, McAllister SS, Zhao JJ. CDK4/6 inhibition in cancer: beyond cell cycle arrest. Trends in Cell Biology. 2018;28(11):911–25. PMID: 30061045. Available from: 10.1016/j.tcb.2018.07.002.
6. Robert M, Frenel JS, Bourbouloux E, Rigaud DB, Patsouris A, Augereau P. An update on the clinical use of CDK4/6 inhibitors in breast Cancer. Drugs. 2018;78(13):1353–62. PMID: 30143968. Available from: 10.1007/s40265-018-0972-9.
7. Sánchez-Martínez C, Gelbert LM, Lallena MJ, de Dios A. Cyclin dependent kinase (CDK) inhibitors as anticancer drugs. Bioorganic & Medicinal Chemistry Letters. 2015;25(17):3420–35. PMID: 26115571. Available from: 10.1016/j.bmcl.2015.05.100.
8. Schaer DA, Beckmann RP, Dempsey JA, Huber L, Forest A, Amaladas N. The CDK4/6 inhibitor abemaciclib induces a T cell inflamed tumor microenvironment and enhances the efficacy of PD-L1 checkpoint blockade. Cell Reports. 2018;22(11):2978–94. PMID: 29539425. Available from: 10.1016/j.celrep.2018.02.053.

9. Harper JW, Elledge SJ, Keyomarsi K, Dynlacht B, Tsai LH, Zhang P. Inhibition of cyclin-dependent kinases by p21. Molecular Biology of the Cell. 1995;6(4):387–400. PMID: 7626805. Available from: 10.1091/mbc.6.4.387.
10. Wu T, Chen Z, To KK, Fang X, Wang F, Cheng B. Effect of abemaciclib (LY2835219) on enhancement of chemotherapeutic agents in ABCB1 and ABCG2 overexpressing cells in vitro and in vivo. Biochemical Pharmacology. 2017;124:29–42. PMID: 27816545. Available from: 10.1016/j.bcp.2016.10.015.
11. Chen K, Jiao X, Rocco AD, Shen D, Xu S, Ertel A. Endogenous cyclin D1 promotes the rate of onset and magnitude of mitogenic signaling via Akt1 Ser473 phosphorylation. Cell Reports. 2020;32(11):108151. PMID: 32937140. Available from: 10.1016/j.celrep.2020.108151.
12. Knudsen ES, Hutcheson J, Vail P, Witkiewicz AK. Biological specificity of CDK4/6 inhibitors: dose response relationship, in vivo signaling, and composite response signature. Oncotarget. 2017;8(27):43678–91. PMID: 28620137. Available from: 10.18632/oncotarget.18435.
13. Torres-Guzmán R, Calsina B, Hermoso A, Baquero C, Alvarez B, Amat J. Preclinical characterization of abemaciclib in hormone receptor positive breast cancer. Oncotarget. 2017;8(41):69493–507. PMID: 29050219. Available from: 10.18632/oncotarget.17778.
14. Beck TN, Georgopoulos R, Shagisultanova EI, Sarcu D, Handorf EA, Dubyk C. EGFR and RB1 as dual biomarkers in HPV-negative head and neck cancer. Molecular Cancer Therapeutics. 2016;15(10):2486–97. PMID: 27507850. Available from: 10.1158/1535-7163.MCT-16-0243.
15. Jin J, Fang H, Yang F, Ji W, Guan N, Sun Z. Combined inhibition of ATR and WEE1 as a novel therapeutic strategy in triple-negative breast cancer. Neoplasia (New York, NY). 2018;20(5):478–88. PMID: 29605721. Available from: 10.1016/j.neo.2018.03.003.
16. Testa U, Castelli G, Pelosi E. Breast cancer: a molecularly heterogenous disease needing subtype-specific treatments. Medical Sciences : Open Access Journal. 2020;8(1):18. PMID: 32210163. Available from: 10.3390/medsci8010018.
17. Kaufmann K, Muiño JM, Osteras M, Farinelli L, Krajewski P, Angenent GC. Chromatin immunoprecipitation (ChIP) of plant transcription factors followed by sequencing (ChIP-SEQ) or hybridization to whole genome arrays (ChIP-CHIP). Nature Protocols. 2010;5(3):457–72. PMID: 20203663. Available from: 10.1038/nprot.2009.244.
18. Lee TI, Johnstone SE, Young RA. Chromatin immunoprecipitation and microarray-based analysis of protein location. Nature Protocols. 2006;1(2):729–48. PMID: 17406303. Available from: 10.1038/nprot.2006.98.
19. Choul-Li S, Legrand AJ, Bidon B, Vicogne D, Villeret V, Aumercier M. Ets-1 interacts through a similar binding interface with Ku70 and Poly (ADP-Ribose) Polymerase-1. Bioscience, Biotechnology, and Biochemistry. 2018;82(10):1753–9. PMID: 29912634. Available from: 10.1080/09168451.2018.1484276.
20. Buck MJ, Lieb JD. ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. Genomics. 2004;83(3):349–60. PMID: 14986705. Available from: 10.1016/j.ygeno.2003.11.004.
21. Stanojević D, Hoey T, Levine M. Sequence-specific DNA-binding activities of the gap proteins encoded by hunchback and Krüppel in Drosophila. Nature. 1989;341(6240):331–5. PMID: 2507923. Available from: 10.1038/341331a0.
22. Kulski JK. Next-Generation Sequencing — An Overview of the History, Tools, and "Omic" Applications. In: Kulski, J. , editor. Next Generation Sequencing - Advances, Applications and Challenges [Internet]. London: IntechOpen; 2016 [cited 2022 Mar 31]. Available from: 10.5772/61964.
23. Su Z, Ning B, Fang H, Hong H, Perkins R, Tong W. Next-generation sequencing and its applications in molecular diagnostics. Expert Review of Molecular Diagnostics. 2011;11(3):333–43. PMID: 21463242. Available from: 10.1586/erm.11.3.
24. Forde BM, O'Toole PW. Next-generation sequencing technologies and their impact on microbial genomics. Briefings in Functional Genomics. 2013;12(5):440–53. PMID: 23314033. Available from: 10.1093/bfgp/els062.
25. Olkhov-Mitsel E, Bapat B. Strategies for discovery and validation of methylated and hydroxymethylated DNA biomarkers. Cancer Medicine. 2012;1(2):237–60. PMID: 23342273. Available from: 10.1002/cam4.22.
26. Serrati S, De Summa S, Pilato B, Petriella D, Lacalamita R, Tommasi S, et al. Next-generation sequencing: advances and applications in cancer diagnosis. OncoTargets and Therapy. 2016;9:7355–65. PMID: 27980425. Available from: 10.2147/OTT.S99807.
27. Mardis ER. Next-generation sequencing platforms. Annual Review of Analytical Chemistry (Palo Alto, Calif). 2013;6(1):287–303. PMID: 23560931. Available from: 10.1146/annurev-anchem-062012-092628.
28. Fujioka T, Kubota K, Mori M, Kikuchi Y, Katsuta L, Kasahara M. Distinction between benign and malignant breast masses at breast ultrasound using deep learning method with convolutional neural network. Japanese Journal of Radiology. 2019;37(6):466–72. PMID: 30888570. Available from: 10.1007/s11604-019-00831-5.
29. Freedman DM, Dosemeci M, McGlynn K. Sunlight and mortality from breast, ovarian, colon, prostate, and non-melanoma skin cancer: a composite death certificate based case-control study. Occupational and Environmental Medicine. 2002;59(4):257–62. PMID: 11934953. Available from: 10.1136/oem.59.4.257.
30. Lee AJ, Cunningham AP, Tischkowitz M, Simard J, Pharoah PD, Easton DF. Incorporating truncating variants in PALB2, CHEK2, and ATM into the BOADICEA breast cancer risk model. Genetics in Medicine. 2016;18(12):1190–8. PMID: 27464310. Available from: 10.1038/gim.2016.31.
31. Spring LM, Wander SA, Andre F, Moy B, Turner NC, Bardia A. Cyclin-dependent kinase 4 and 6 inhibitors for hormone receptor-positive breast cancer: past, present, and future. Lancet. 2020;395(10226):817–27. PMID: 32145796. Available from: 10.1016/S0140-6736(20)30165-3.
32. Wildiers H, de Glas NA. Anticancer drugs are not well tolerated in all older patients with cancer. The Lancet Healthy Longevity. 2020;1(1):e43–7. Available from: 10.1016/S2666-7568(20)30001-5.
33. Franco J, Witkiewicz AK, Knudsen ES. CDK4/6 inhibitors have potent activity in combination with pathway selective therapeutic agents in models of pancreatic cancer. Oncotarget. 2014;5(15):6512–25. PMID: 25156567. Available from: 10.18632/oncotarget.2270.
34. Nigro A, Ricciardi L, Salvato I, Sabbatino F, Vitale M, Crescenzi MA. Enhanced expression of CD47 is associated with off-target resistance to tyrosine kinase inhibitor gefitinib in NSCLC. Frontiers in Immunology. 2020;10:3135. PMID: 32082304. Available from: 10.3389/fimmu.2019.03135.
35. Dey N, De P, Leyland-Jones B. PI3K-AKT-mTOR inhibitors in breast cancers: from tumor cell signaling to clinical trials. Pharmacology & Therapeutics. 2017;175:91–106. PMID: 28216025. Available from: 10.1016/j.pharmthera.2017.02.037.
36. He X, Chen CC, Hong F, Fang F, Sinha S, Ng HH. A biophysical model for analysis of transcription factor interaction and binding site arrangement from genome-wide binding data. PLoS One. 2009;4(12):e8155. PMID: 19956545. Available from: 10.1371/journal.pone.0008155.
37. Sawadogo M, Roeder RG. Interaction of a gene-specific transcription factor with the adenovirus major late promoter upstream of the TATA box region. Cell. 1985;43(1):165–75. PMID: 4075392. Available from: 10.1016/0092-8674(85)90021-2.
38. Jatoi I, Anderson WF, Jeong JH, Redmond CK. Breast cancer adjuvant therapy: time to consider its time-dependent effects. Journal of Clinical Oncology. 2011;29(17):2301–4. PMID: 21555693. Available from: 10.1200/JCO.2010.32.3550.

39. Hedenfalk I, Ringnér M, Ben-Dor A, Yakhini Z, Chen Y, Chebil G. Molecular classification of familial non-BRCA1/BRCA2 breast cancer. Proceedings of the National Academy of Sciences of the United States of America. 2003;100(5):2532–7. PMID: 12610208. Available from: 10.1073/pnas.0533805100.
40. Zhang W, Cui H, Wong LJ. Application of next generation sequencing to molecular diagnosis of inherited diseases. Top Curr Chem. 2014;336:19–45. Available from: 10.1007/128_2012_325.
41. Naidoo N, Pawitan Y, Soong R, Cooper DN, Ku CS. Human genetics and genomics a decade after the release of the draft sequence of the human genome. Human Genomics. 2011;5(6):577–622. PMID: 22155605. Available from: 10.1186/1479-7364-5-6-577.
42. Ziogas DE. Genome-based approaches for the diagnosis of breast cancer: a review with perspective. Breast Cancer Management. 2014;3(2):173–93. Available from: 10.2217/bmt.13.81.
43. Watt AC, Cejas P, DeCristo MJ, Metzger-Filho O, Lam EY, Qiu X. CDK4/6 inhibition reprograms the breast cancer enhancer landscape by stimulating AP-1 transcriptional activity. Naturaliste Canadien. 2021;2(1):34–48. PMID: 33997789. Available from: 10.1038/s43018-020-00135-y.
44. Newell R, Pienaar R, Balderson B, Piper M, Essebier A, Bodén M. ChIP-R: assembling reproducible sets of ChIP-seq and ATAC-seq peaks from multiple replicates. Genomics. 2021;113(4):1855–66. PMID: 33878366. Available from: 10.1016/j.ygeno.2021.04.026.
45. Han Z, Yang B, Wang Q, Hu Y, Wu Y, Tian Z. Comprehensive analysis of the transcriptome-wide m6A methylome in invasive malignant pleomorphic adenoma. Cancer Cell International. 2021;21(1):142. PMID: 33653351. Available from: 10.1186/s12935-021-01839-6.
46. Yu CP, Kuo CH, Nelson CW, Chen CA, Soh ZT, Lin JJ, et al. Discovering unknown human and mouse transcription factor binding sites and their characteristics from ChIP-seq data. Proceedings of the National Academy of Sciences. 2021;118(20):e2026754118. Available from: 10.1073/pnas.2026754118.
47. Kang Y, Kang J, Kim YW, Kim A. ChIP-seq Library Preparation and NGS Data Analysis Using the Galaxy Platform. Journal of Life Science. 2021;31:410–7. Available from: 10.5352/JLS.2021.31.4.410.
48. Li QL, Lin X, Yu YL, Chen L, Hu QX, Chen M. Genome-wide profiling in colorectal cancer identifies PHF19 and TBC1D16 as oncogenic super enhancers. Nature Communications. 2021;12(1):6407. PMID: 34737287. Available from: 10.1038/s41467-021-26600-5.
49. Polit L, Kerdivel G, Gregoricchio S, Esposito M, Guillouf C, Boeva V. CHIPIN: ChIP-seq inter-sample normalization based on signal invariance across transcriptionally constant genes. BMC Bioinformatics. 2021;22(1):407. PMID: 34404353. Available from: 10.1186/s12859-021-04320-3.
50. Qin HL, Wang XJ, Yang BX, Du B, Yun XL. Notoginsenoside R1 attenuates breast cancer progression by targeting CCND2 and YBX3. Chinese Medical Journal. 2021;134(5):546–54. PMID: 33480613. Available from: 10.1097/CM9.0000000000001328.
51. Vishnubalaji R, Alajez NM. Epigenetic regulation of triple negative breast cancer (TNBC) by TGF-β signaling. Scientific Reports. 2021;11(1):15410. PMID: 34326372. Available from: 10.1038/s41598-021-94514-9.