Epigenetic Regulations, Motif and Pathway Identification of Hodgkin Lymphoma Chip Sequence

Priyanka Stephen Mukhia[1,2](✉), I. A. Shylesh Murthy[2], and Preenon Bagchi[1,2,3]

[1] Padmashree Institute of Management and Science, Bangalore, India
priyankamukhia19@gmail.com
[2] Vasishth Academy of Advanced Studies and Research (Sarvasumana Association), Bangalore, India
[3] MGM Institute of Biosciences and Technology, Aurangabad, India

Abstract. ChIP-sequencing (Chromatin immunoprecipitation) is a new method for genome wide mapping of protein binding sites on DNA. ChIP-sequencing helps to understand better the mechanisms of transcription factors, cofactors and histone modifications as well the regulation of gene expressions. The high degree of flow rate of sequencing data can be retrieved through international data repositories that is NCBI Sequence Read Archive (SRA). The BCL6 proto- oncogene encodes a nuclear transcriptional repressor. BCL6 suppresses p53 in germinal center, where the deregulation of BCL6 is located contributes to the malignant transformation in germinal center derived B cells. Hodgkin lymphoma arises from germinal center derived B cell. Epigenetics is the study of genetic changes in gene activity which are generally caused by mechanisms other than DNA sequence changes. For studying epigenomics data, bioinformatics is a successful approach in the field of molecular biology.

Keywords: Classical Hodgkin lymphoma · Epstein barr virus · Hodgkin lymphoma · Insulin-like growth factor · Lymphocyte-rich Hodgkin lymphoma · Reed-Sternberg cells · Chromatin immunoprecipitation sequencing · National Centre for Biotechnology Information · Binary Alignment Map format · FASTA/BLAST format · Sequence Read Archive · B-cell lymphoma 6 protein · Fast Alignment sequence test for application · European Bioinformatics Institute · Gene Expression Omnibus Genome Analyzer · Kyoto Encyclopedia of Genes and Genomes pathway database

1 Introduction

Chromatin immunoprecipitation (ChIP) is followed by a great yield of DNA sequencing (also known as ChIP-sequencing), enhances the protein target and Illumina GA 2 is the widely used platform of sequencing of a genomic region multiple times for Chip sequencing [1, 2]. The sequencing data can be retrieved from NCBI (SRA) accessible through Gene Expression Omnibus (GEO) (ncbi.nlm.nih.gov/geo) and also at times EBI Sequence Read Archive.

R. Somashekhar et al. (Eds.): ICBDS 2022, AHSR 58, pp. 132–146, 2023.
https://doi.org/10.2991/978-94-6463-164-7_11

Raw data is stored on SRA and GEO, metadata about the sample can be found from the SRA link (e.g. SRX11465132, SRX11465131) which was taken into account from NCBI page. It contains information about the samples such as: layout (e.g. pairout), sequencing instrument (Illumina HiSeq 4000), source name (Bone marrow:AML), experimental strategy (e.g. RNA-seq), material source (e.g. Genomic DNA), experimental protocol as well as the identifiers of the sequencing runs generated for this sample (that is SRR15157768 and SRR15157767) which was further used to extract data from Short Read Archive at the NCBI and this was done by the tool of Download and Extract Reads in FASTA/Q in the galaxy web page (www.usegalaxy.eu). Also the Gene Expression Omnibus (GEO) accession attributes number is GSM5454820 of SRR15157768 and the GEO accession attributes number of SRR15157767 is GSM5454819.

The common symptoms of Hodgkin lymphoma is swelling of lymph nodes resulting into formation of lump (mostly lumps usually isn't painful) under the skin like on the neck, in the armpit, around groin area [3]. There are few other symptoms like night sweats, itchy skin, fever, fatigue, unintended weight loss, persistent cough, trouble breathing, and chest pain and sometimes after consumption of alcohol pain may arise in the lymph nodes and enlarged spleen.

B cell lymphoma 6 gene (BCL6) encodes a transcription factor [4] which is critical for normal germinal center reaction B cell development and also maintains the translational state and epigenetic which is non-restrictive for cellular proliferation and DNA damage. [5, 6] The relatable pathways of BCL6 are B-cell receptors signalling pathways (KEGG) and B-cell signalling pathway (sino). The protein encoded by BCL6 gene is a zinc finger transcription factor and contains an N-terminal POZ domain. Thus, in response to antigen receptor activation [7, 8].

2 Materials and Methods

The requirements for performing chip sequencing were*:*

NCBI, Galaxy European page, SRR sequences, Galaxy cistrome page, Swiss prot model page and KEGG pathway database.

Two SRR sequences of BCL6 gene in Hodgkin lymphoma disease was taken into account i.e. SRR15157767 and SRR15157768 from the NCBI page.

In galaxy web page, the tool Download an extract reads in FASTA/Q was used, whose function is to extract data in FASTAQ format from the SRA. The raw reads from sequencing are run first quality control by program FastQC (www.usegalaxy.eu) to remove the adapt and other contaminative sequences, FastQC intends to make quality control checks on raw sequence data from high-throughput sequencing pipelines as simple as possible [9].

The reference genome is mapped using short read mapping program such as Bowtie2, which is a tool for aligning sequencing reads to large reference sequences which is both rapid and memory-efficient. [10] It specializes in aligning reads which ranges from 50 to 100s or 1,000s of characters to genomes that are long (e.g. mammalian). Bowtie 2 includes gapped, local, and paired-end alignment solutions. [11] This tool constructs a summary alignment from a SAM or Bam file; [12] it takes a SAM or BAM file as input and gives metrics that detail the read alignment quality as well as the proportion of reads that passed machine signal to threshold quality filters [13].

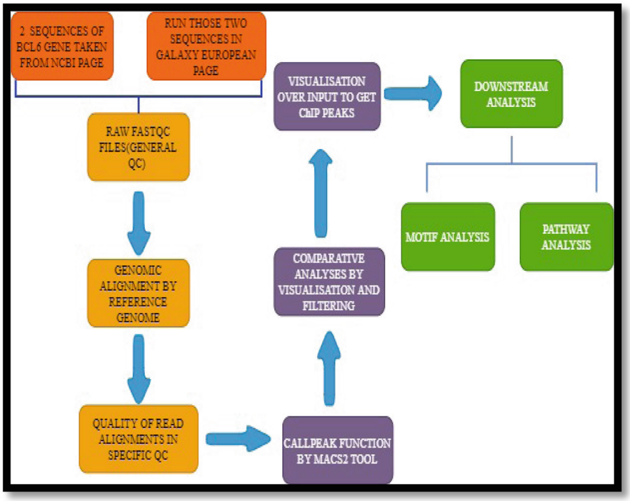

2 SEQUENCES OF BCL6 GENE TAKEN FROM NCBI PAGE
RUN THOSE TWO SEQUENCES IN GALAXY EUROPEAN PAGE
RAW FASTQC FILES(GENERAL QC)
GENOMIC ALIGNMENT BY REFERENCE GENOME
QUALITY OF READ ALIGNMENTS IN SPECIFIC QC
CALLPEAK FUNCTION BY MACS2 TOOL
COMPARATIVE ANALYSES BY VISUALISATION AND FILTERING
VISUALISATION OVER INPUT TO GET ChIP PEAKS
DOWNSTREAM ANALYSIS
MOTIF ANALYSIS
PATHWAY ANALYSIS

Fig. 1. ChIP-seq pipeline of Hodgkin lymphoma

Then the step to remove potential PCR duplicates and to retain the highest mapping quality is called as "RmDup". [14] Only the read pair is kept with the highest mapping quality and also if multiple read pairs have identical external coordinates. This command only works in paired-end mode with FR alignment and requires size to be correctly set. [15, 16] Then again collect alignment metrics step was done again; so as to produce metrics detailing of the quality read alignments [17].

Collect Alignment Summary Metrics is followed by callpeak function which is the vital part of MACS2 package. MACS helps identifies enriched binding sites in ChIP seq, it captures the influence of genome complexity to evaluate the significance of enriched ChIP regions and improves spatial resolution of binding sites through combining the information of both sequencing tag position and orientation [18]. MACS can be used for ChIP seq data (treatment) alone or with a control sample with the increase of specificity. A treatment file is the only required parameter for MACS. The file can be BAM or BED format and this tool will auto detect the format using the first treatment file provided as input. MACS can pool files together that are control file [19].

Sort tool arranges the dataset in any number of columns in either ascending or descending order. In filter tool, the data is restricted using simple conditional statements; for example only the first column of a tab-delimited file is taken when the filtering condition of the logical operators used was 'c1! = #'.

Then the text manipulation tool was clicked and under that option the select first tool gives output of a specified number of lines from the beginning of a dataset and thus only the first 100 lines were taken into account which was renamed as 'Top 100 CTCF peaks'.

The UCSC Main table browser is the tool used, also provides content access to the Genomic Browser Database's vast collection of genome assemblies and annotation data. (http://genome.ucsc.edu) [20, 21] (Fig. 1).

Bedtools allows to intersect, merge, count, complement, and shuffle genomic intervals from multiple files in widely used genomic file formats i.e. when columns are specified (for e.g. c14); and which selects out the specified columns from the dataset (Bjoern A. Gruening (2014), Galaxy wrapper) [22] The SeqPos tool will find motifs enriched in a set of regions. SeqPos uses the distances from the motif positions to the peak summits as to find the most enriched motifs near peak summits [23].

In Swiss Prot, genes were selected from the motif analysis and processed to obtain models using with ramachandran plot. [24–26] KEGG is the short form for Kyoto Encyclopedia of Genes and Genomes. KEGG is the source of database for understanding high level functions of biological system (https://www.genome.jp/kegg/module.html) KEGG PATHWAY is a set of hand-drawn pathway diagrams that represent our knowledge of molecular interaction, response, and link networks for:

1. Metabolism
2. Processing of Genetic Input
3. Processing of Environmental Information
4. Cellular Functions
5. Organismal Systems
6. Human Diseases
7. Drug Research and Development

3 Results and Discussion

3.1 FASTQC Quality Report of SRR15157767 and SRR15157768

The GC % of sequences SRR15157767 and SRR15157768 is 49. The sequence length of both SRR sequences is about 35–151. There were 0 sequences that were flagged as of poor quality in both the sequences SRR15157767 and SRR15157768. The encoding measure was by using Sanger/Illumina 1.9. Total sequences that were found was 52938296 in SRR15157768 and in sequence SRR15157767 was 59763960 (Figs. 2 and 3).

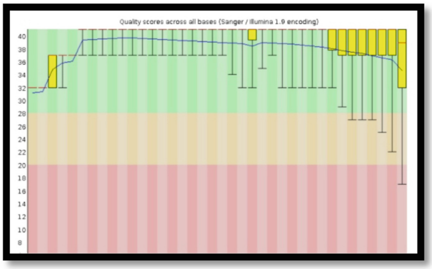


Fig. 2. SRR15157767 of Fast QC report

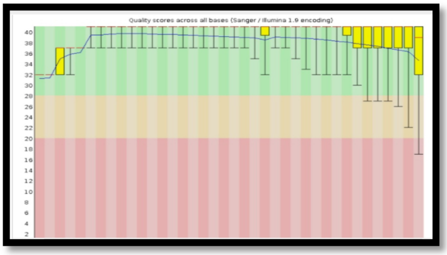
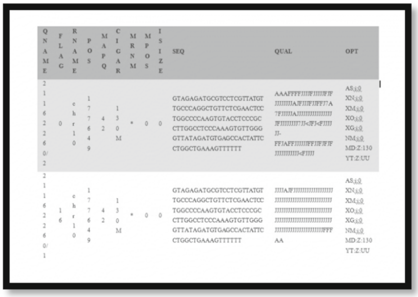
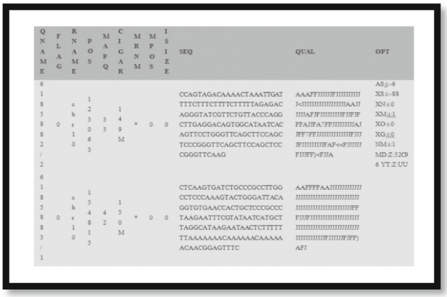


Fig. 3. SRR15157768 of Fast QC reports



QNAME	FLAG	RNAME	POS	MAPQ	CIGAR	MRNM	MPOS	ISIZE	SEQ	QUAL	OPT
21162260/2	0	chr10	177649	42	130M	*	0	0	GTAGAGATGCGTCCTCGTTATGTTGCCCAGGCTGTTCTCGAACTCCTGGCCCCAAGTGTACCTCCCGCCTTGGCCTCCCAAAGTGTTGGGGTTATAGATGTGAGCCACTATTCCTGGCTGAAAGTTTTTT	AAAFFFFJJJJFJJJJJFJFJJJJJJJAJFJJJFJJFFJ7A7FJJJJAJJJJJJJJJJJJJJJJFJJJJJJJJ7JJ<JFJ<FJJJJJJ-FFJAFFJJJJJFFJJFJFJFJJJJJJJJJJ<FJJJJ	AS:i:0 XN:i:0 XM:i:0 XO:i:0 XG:i:0 NM:i:0 MD:Z:130 YT:Z:UU
21162260/1	16	chr10	177649	42	130M	*	0	0	GTAGAGATGCGTCCTCGTTATGTTGCCCAGGCTGTTCTCGAACTCCTGGCCCCAAGTGTACCTCCCGCCTTGGCCTCCCAAAGTGTTGGGGTTATAGATGTGAGCCACTATTCCTGGCTGAAAGTTTTTT	JJJJAJFJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJFFFAA	AS:i:0 XN:i:0 XM:i:0 XO:i:0 XG:i:0 NM:i:0 MD:Z:130 YT:Z:UU

Fig. 4. Bowtie 2 result of sequence SRR15157767



QNAME	FLAG	RNAME	POS	MAPQ	CIGAR	MRNM	MPOS	ISIZE	SEQ	QUAL	OPT
6185882/2	0	chr10	123063	33	149M	*	0	0	CCAGTAGACAAAACTAAATTGATTTTCTTTCTTTTCTTTTTAGAGACAGGGTATCGTTCTGTTACCCAGGCTTGAGGACAGTGGCATAATCACAGTTCCTGGGTTCAGCTTCCAGCTCCCGGGTTCAGCTTCCAGCTCCCGGGTTCAAG	AAAFFJJJJJJFJJJJJJJJJJJ<JJJJJJJJJJJJJJJJJJAAJJJJJJAFJFJJJJJJJJJFJJFJFFFAJJFA7FFJJJJJJJJJAJJFF7FFJJJJJJJJJJJJJJFJJJJFJJJJJJJJJFAF<<FJJJJJJFJJJFF)<FJJA	AS:i:-6 XS:i:-88 XN:i:0 XM:i:1 XO:i:0 XG:i:0 NM:i:1 MD:Z:52C9 6 YT:Z:UU
6185883/1	0	chr10	154815	42	150M	*	0	0	CTCAAGTGATCTGCCCGCCTTGGCCTCCCAAAGTACTGGGATTACAGGTGTGAACCACTGCTCCCGCCCTAAGAATTTCGTATAATCATGCTTAGGCATAAGAATAACTCTTTTTTTAAAAAAACAAAAAACAAAAAACAACGGAGTTTC	AAFFFFAAJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJFFFJJJFJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJFJJJJJJFJFF)AFJ	

Fig. 5. Bowtie 2 result of sequence SRR15157768

BOWTIE 2: For aligning sequences to long reference sequences

We used data from a ChIPseq experiment performed on the BCL6 gene (SRA accession numbers SRR15157767 and SRR15157768).The readings are mapped to the human genome of Hodgkin disease, mapping statics are obtained, and duplicated reads produced by PCR mistakes are removed. In SRR15157768, The Dbkey was taken from hg19 and the reference genome is hg19, with bam format and the single parameter library and the dataset peek is binary bam alignment file. This dataset is stored in a Galaxy object store named University of Freiburg Isilon storage (id = files 10). In SRR15157767, the Dbkey and reference genome is hg19, is bam format and the value is single as of input parameter library. The dataset is stored in a Galaxy object store named University of Freiburg Isilon storage (id = files 10), also the dataset peek is binary bam alignment file (Figs. 4 and 5).

3.2 Collect Alignment Summary Metrics

This Galaxy application uses Picard to report high-level alignment measures based on a SAM or BAM file provided. The following columns appeared in the tool's output (Figs. 6 and 7).

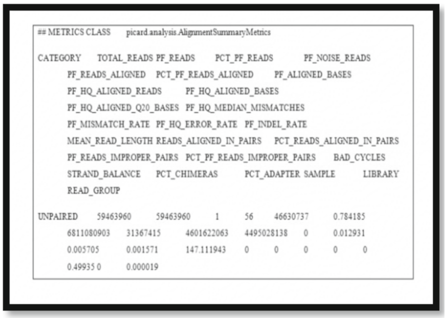
The following are the results obtained for SRR15157767 and SRR15157768:-

3.3 After RmDup

The PCR duplicates were removed by RmDup, and were done on SRA accession number of BCL6 gene of Hodgkin lymphoma:-Post RmDup (Figs. 8 and 9).

3.4 Peak Analysis Obtained in Macs2callpeak of Sequences SRR15157767 and SRR15157768

MACS produces four files: peaks: bed, peaks: interval, negative peaks, and an HTML report. It also contains links that will let us download the other files in Galaxy history as well as additional files that gives information about the model (Figs. 10 and 11).

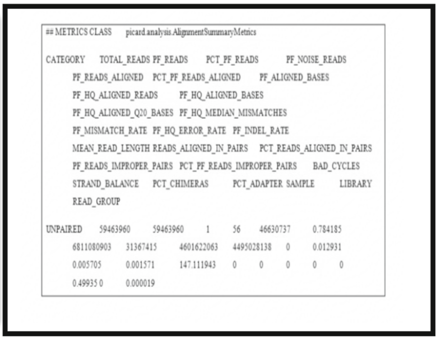


```
## METRICS CLASS      picard.analysis.AlignmentSummaryMetrics

CATEGORY     TOTAL_READS PF_READS      PCT_PF_READS        PF_NOISE_READS
        PF_READS_ALIGNED   PCT_PF_READS_ALIGNED      PF_ALIGNED_BASES
        PF_HQ_ALIGNED_READS       PF_HQ_ALIGNED_BASES
        PF_HQ_ALIGNED_Q20_BASES  PF_HQ_MEDIAN_MISMATCHES
        PF_MISMATCH_RATE  PF_HQ_ERROR_RATE  PF_INDEL_RATE
        MEAN_READ_LENGTH READS_ALIGNED_IN_PAIRS    PCT_READS_ALIGNED_IN_PAIRS
        PF_READS_IMPROPER_PAIRS  PCT_PF_READS_IMPROPER_PAIRS     BAD_CYCLES
        STRAND_BALANCE     PCT_CHIMERAS        PCT_ADAPTER SAMPLE       LIBRARY
        READ_GROUP

UNPAIRED      59463960       59463960       1        56       46630737       0.784185
        6811080903      31367415       4601622063      4495028138     0       0.012931
        0.005705        0.001571       147.111943      0        0       0       0       0
        0.49935 0       0.000019
```

Fig. 6. SRR15157767 sequence obtained in alignment metrics

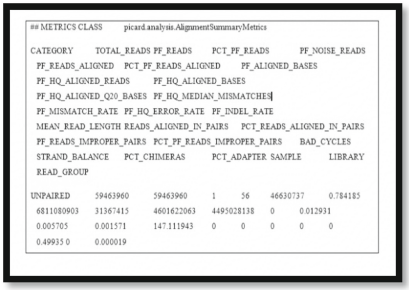


```
## METRICS CLASS      picard.analysis.AlignmentSummaryMetrics

CATEGORY     TOTAL_READS PF_READS      PCT_PF_READS        PF_NOISE_READS
        PF_READS_ALIGNED   PCT_PF_READS_ALIGNED      PF_ALIGNED_BASES
        PF_HQ_ALIGNED_READS       PF_HQ_ALIGNED_BASES
        PF_HQ_ALIGNED_Q20_BASES  PF_HQ_MEDIAN_MISMATCHES
        PF_MISMATCH_RATE  PF_HQ_ERROR_RATE  PF_INDEL_RATE
        MEAN_READ_LENGTH READS_ALIGNED_IN_PAIRS    PCT_READS_ALIGNED_IN_PAIRS
        PF_READS_IMPROPER_PAIRS  PCT_PF_READS_IMPROPER_PAIRS     BAD_CYCLES
        STRAND_BALANCE     PCT_CHIMERAS        PCT_ADAPTER SAMPLE       LIBRARY
        READ_GROUP

UNPAIRED      59463960       59463960       1        56       46630737       0.784185
        6811080903      31367415       4601622063      4495028138     0       0.012931
        0.005705        0.001571       147.111943      0        0       0       0       0
        0.49935 0       0.000019
```

Fig. 7. SRR15157768 sequence of alignment metrics

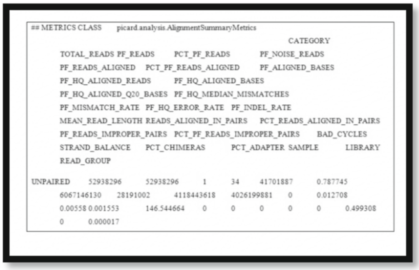


```
## METRICS CLASS      picard.analysis.AlignmentSummaryMetrics

CATEGORY     TOTAL_READS PF_READS      PCT_PF_READS        PF_NOISE_READS
  PF_READS_ALIGNED   PCT_PF_READS_ALIGNED      PF_ALIGNED_BASES
  PF_HQ_ALIGNED_READS       PF_HQ_ALIGNED_BASES
  PF_HQ_ALIGNED_Q20_BASES  PF_HQ_MEDIAN_MISMATCHES
  PF_MISMATCH_RATE  PF_HQ_ERROR_RATE  PF_INDEL_RATE
  MEAN_READ_LENGTH READS_ALIGNED_IN_PAIRS    PCT_READS_ALIGNED_IN_PAIRS
  PF_READS_IMPROPER_PAIRS  PCT_PF_READS_IMPROPER_PAIRS     BAD_CYCLES
  STRAND_BALANCE     PCT_CHIMERAS        PCT_ADAPTER SAMPLE       LIBRARY
  READ_GROUP

UNPAIRED      59463960       59463960       1        56       46630737       0.784185
  6811080903      31367415       4601622063      4495028138     0       0.012931
  0.005705        0.001571       147.111943      0        0       0       0       0
  0.49935 0       0.000019
```

Fig. 8. The alignment summary result of Sequence SRR15157767



```
## METRICS CLASS      picard.analysis.AlignmentSummaryMetrics
                                                                     CATEGORY
        TOTAL_READS PF_READS      PCT_PF_READS        PF_NOISE_READS
        PF_READS_ALIGNED   PCT_PF_READS_ALIGNED      PF_ALIGNED_BASES
        PF_HQ_ALIGNED_READS       PF_HQ_ALIGNED_BASES
        PF_HQ_ALIGNED_Q20_BASES  PF_HQ_MEDIAN_MISMATCHES
        PF_MISMATCH_RATE  PF_HQ_ERROR_RATE  PF_INDEL_RATE
        MEAN_READ_LENGTH READS_ALIGNED_IN_PAIRS    PCT_READS_ALIGNED_IN_PAIRS
        PF_READS_IMPROPER_PAIRS  PCT_PF_READS_IMPROPER_PAIRS     BAD_CYCLES
        STRAND_BALANCE     PCT_CHIMERAS        PCT_ADAPTER SAMPLE       LIBRARY
        READ_GROUP

UNPAIRED      52938296       52938296       1        34       41701887       0.787745
        6067146130      28191002       4118443618      4026199881     0       0.012708
        0.00558 0.001553       146.544664      0        0       0       0       0       0.499308
        0       0.000017
```

Fig. 9. The alignment summary result of sequence SRR15157768

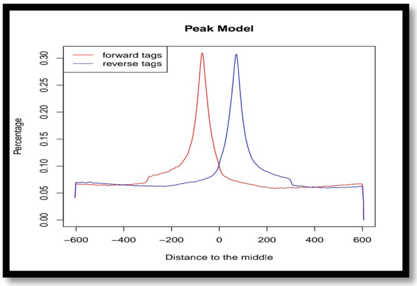


Fig. 10. Peak result of two sequences SRR15157767 and SRR15157768

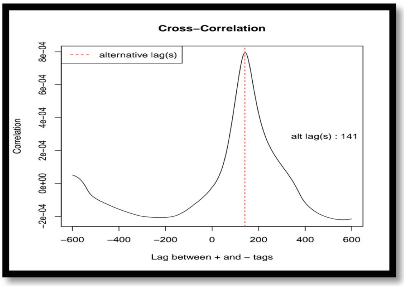
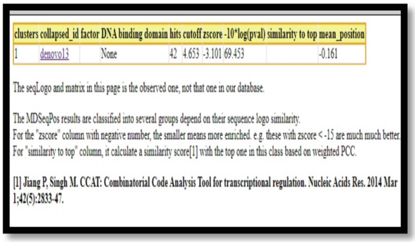


Fig. 11. Cross correlation of two sequences



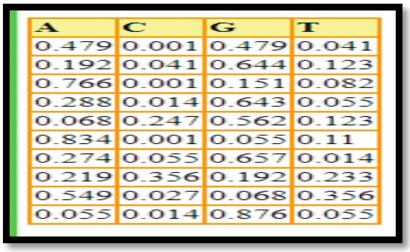
clusters	collapsed_id	factor	DNA binding domain	hits	cutoff	zscore	-10*log(pval)	similarity to top	mean_position
1	denovo13		None	42	4.653	-3.101	69.453		-0.161

The seqLogo and matrix in this page is the observed one, not that one in our database.

The MDSeqPos results are classified into several groups depend on their sequence logo similarity.
For the "zscore" column with negative number, the smaller means more enriched. e.g. these with zscore < -15 are much much better.
For "similarity to top" column, it calculate a similarity score[1] with the top one in this class based on weighted PCC.

[1] Jiang P, Singh M. CCAT: Combinatorial Code Analysis Tool for transcriptional regulation. Nucleic Acids Res. 2014 Mar 1;42(5):2833-47.

Fig. 12. DNA binding domain and log information of motif



A	C	G	T
0.479	0.001	0.479	0.041
0.192	0.041	0.644	0.123
0.766	0.001	0.151	0.082
0.288	0.014	0.643	0.055
0.068	0.247	0.562	0.123
0.834	0.001	0.055	0.11
0.274	0.055	0.657	0.014
0.219	0.356	0.192	0.233
0.549	0.027	0.068	0.356
0.055	0.014	0.876	0.055

Fig. 13. ACGT analysis

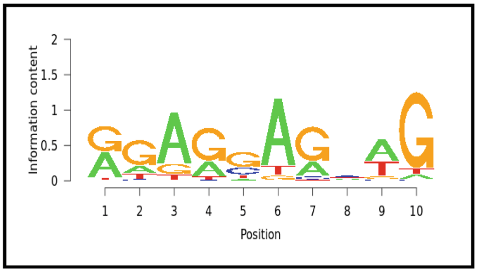


Fig. 14. Motif analysis

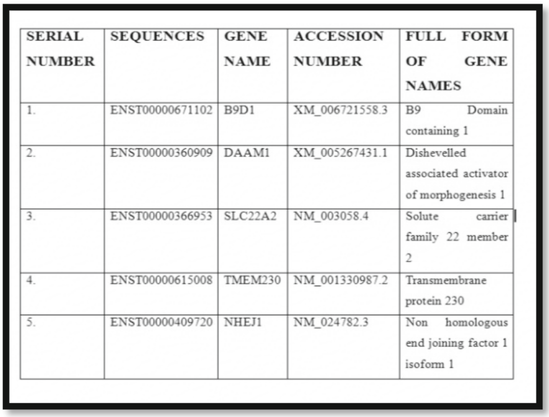
After aligning de-duplicated reads to the genome, peak caller can be used to find regions of the genome with an enrichment of reads, or peaks. Peak callers tend to generate two types of files, discrete and continuous. Discrete files tend to be in BED format while continuous files tend to be in WIG or BigWIG format.

3.5 Motif Analysis Obtained in Galaxy Cistrome Browser

There are numerous distinct motif identification applications accessible both in Galaxy and at the task line. We use the Cistrome consortium's SeqPos motif analysis tool to find recognised motifs in the data (Figs. 12, 13, 14 and 15).

After motif model, genes were selected from hodgkin lymphoma genome, followed by making of swiss prot model along with ramachandran plot:-

This step is carried out by SWISS-MODEL by using Open Structure computation structural biology system and the ProMod3 modelling engine. Colors were adjusted after the plot was created, and the plot is a visualisation produced by Swiss-PDB viewer. In the plot, the dihedral angles of amino acid residues appear as crosses. The desired and allowed regions are depicted by blue and red, accordingly [27, 28].



SERIAL NUMBER	SEQUENCES	GENE NAME	ACCESSION NUMBER	FULL FORM OF GENE NAMES
1.	ENST00000671102	B9D1	XM_006721558.3	B9 Domain containing 1
2.	ENST00000360909	DAAM1	XM_005267431.1	Dishevelled associated activator of morphogenesis 1
3.	ENST00000366953	SLC22A2	NM_003058.4	Solute carrier family 22 member 2
4.	ENST00000615008	TMEM230	NM_001330987.2	Transmembrane protein 230
5.	ENST00000409720	NHEJ1	NM_024782.3	Non homologous end joining factor 1 isoform 1

Fig. 15. Table of Swiss prot model taken into account with gene names

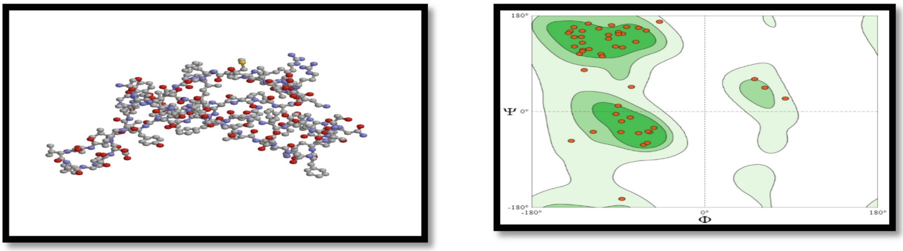


Fig. 16. Swiss model and ramachandran plot of Gene B9 Domain Containing 1

3.6 Following are the Results Obtained in: SWISS PROT Model and Ramachandran Plot of ENST00000671102 (GENE B9 Domain Containing 1):

The Molprobity score is 2.83. The ramachandran favoured is of percentage - 83.87%. The ramachandran outliers are 6.45% (A113GLY, A69PRO, A97ASN, A112PRO). The C-Beta deviations is 4(A111SER, A79PRO), A103TYR, A74PHE) (Fig. 16).

- **THE SWISS PROT MODEL AND RAMACHANDRAN PLOT OF ENST00000360909 (GENE Dishevelled associated activator of morphogenesis):-**

The Molprobity score is 1.66. The ramachandran favoured is of percentage – 93.63%. The ramachandran outliers are 1.84% (B103 PRO, B391 PRO, A391 PRO, B417 ASP, B235 GLY, A113 MET, A269 GLY, B118 SER, B121 ALA, A426 PRO, B269 GLY, A417 ASP, A127 GLU) The C-Beta deviations is 5(B116 ARG, A411 GLN, A156 ASP, B180 SER, A313 ILE) (Fig. 17).

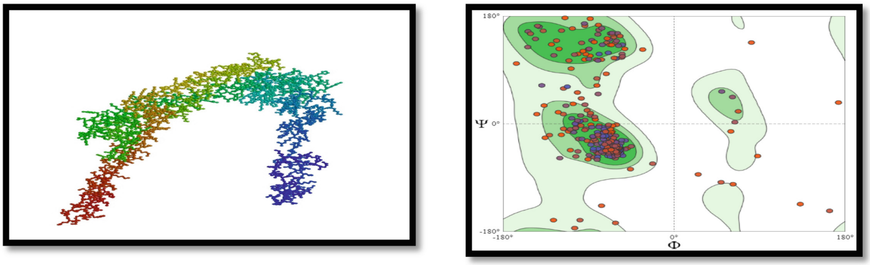


Fig. 17. Swiss model and ramachandran plot of Gene Dishevelled associated activator of morphogenesis

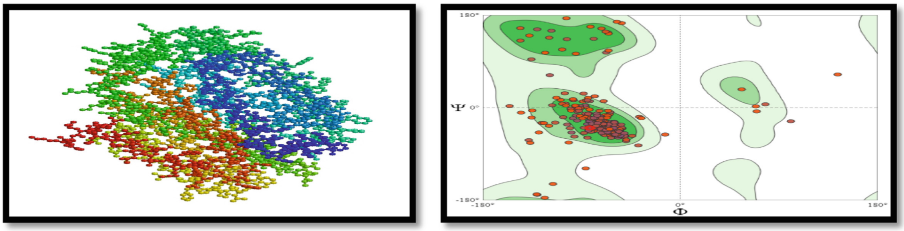


Fig. 18. Swiss model and ramachandran plot of Solute carrier family 22 member 2

- **THE SWISS MODEL AND RAMACHANDRAN PLOT OBTAINED OF ENST00000366953 (GENE Solute carrier family 22 member 2):**

The Molprobity score is 2.34. The ramachandran favoured is of percentage – 90.65%. The ramachandran outliers are 3.38% (B284 PRO, B235 ARG, B342 PRO, B146 SER, B433 ILE, B296 ASN, B321 LEU, B341 THR, B198 PRO, B425 PRO, B515 PRO, B343 GLN, B492 TRP) The C-Beta deviations is 18(B436 SER, B146 SER, B378 LEU, B166 SER, B276 PHE, B223 ILE, B488 LEU, B363 GLN, B503 LEU, B491 ILE, B327 THR, B341 THR, B453 VAL, B136 VAL, B506 VAL, B492 TRP, B387 GLU, B288 ARG) (Fig. 18).

- **THE SWISS MODEL AND RAMACHANDRAN PLOT OF ENST00000409720 (GENE Non homologous end joining factor 1 isoform 1):**

The Molprobity score is 1.91. The ramachandran favoured is of percentage – 90.57%. The ramachandran outliers are 1.68% (I293 LYS, I208 LYS, I280 PRO, I281 LEU, I244 ALA) The C-Beta deviations is 3(I17 GLN, I18 LEU, I155 THR) (Fig. 19).

- **THE SWISS PROT AND RAMACHANDRAN PLOT OF ENST00000615008 (GENE Transmembrane 230):**

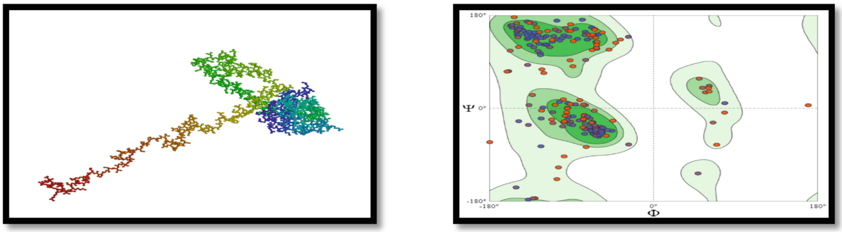


Fig. 19. Swiss model and ramachandran plot of Non homologous end joining factor 1 isoform 1

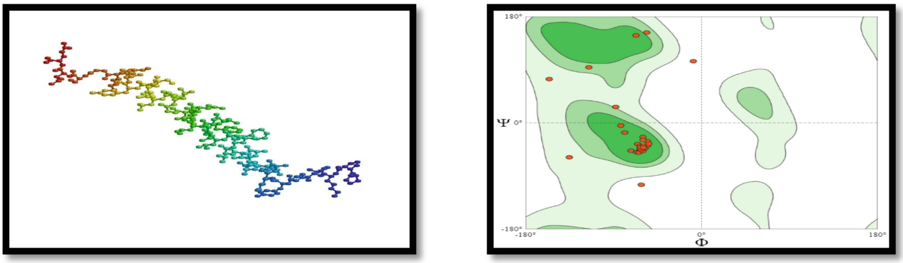
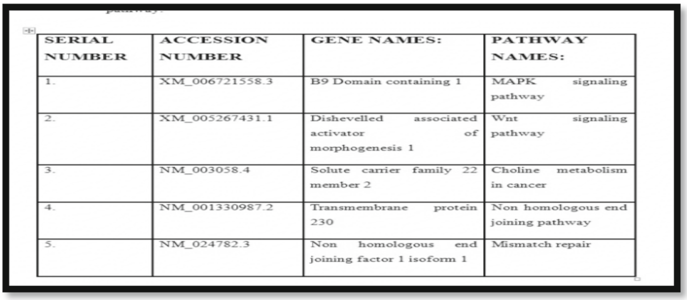


Fig. 20. Swiss model and ramachandran plot of Gene Transmembrane 230



SERIAL NUMBER	ACCESSION NUMBER	GENE NAMES:	PATHWAY NAMES:
1.	XM_006721558.3	B9 Domain containing 1	MAPK signaling pathway
2.	XM_005267431.1	Dishevelled associated activator of morphogenesis 1	Wnt signaling pathway
3.	NM_003058.4	Solute carrier family 22 member 2	Choline metabolism in cancer
4.	NM_001330987.2	Transmembrane protein 230	Non homologous end joining pathway
5.	NM_024782.3	Non homologous end joining factor 1 isoform 1	Mismatch repair

Fig. 21. Table of pathway obtained from selected five genes

The Molprobity score is 2.22. The ramachandran favoured is of percentage – 81.58%. The ramachandran outliers are 7.89% (I A139 LEU, A118 ILE, A102 PRO) The C-Beta deviations is 0 (Fig. 20).

I. KEGG pathway analysis:-

The KEGG pathway map is a cellular interaction/reaction network diagram stated in terms of the KEGG Orthology (KO) groups, which enables experimental results from one organism to be generalised to other organisms utilizing genomic data [28, 29]. The pathways of the genes overlapping with peaks were identified from KEGG pathway (Fig. 21):-

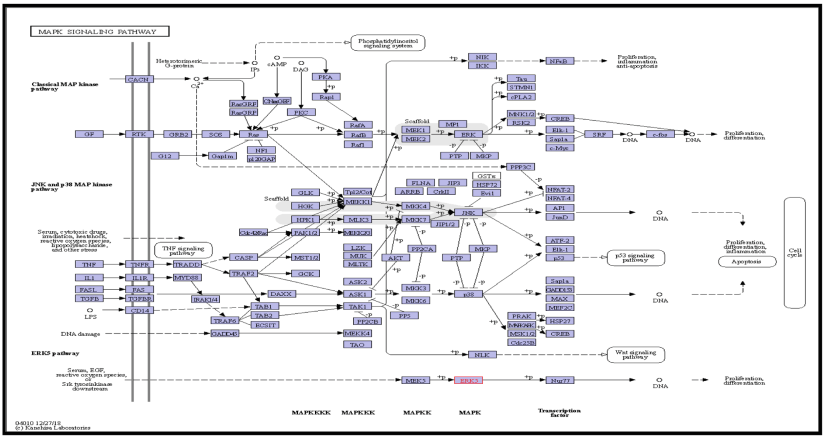


Fig. 22. MAPK signalling pathway diagram

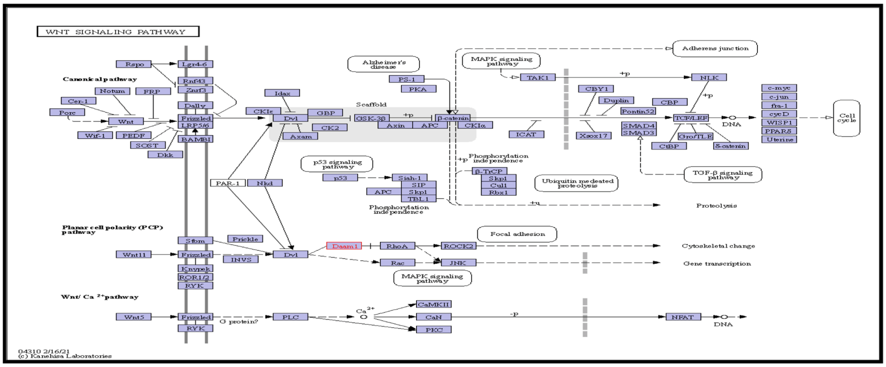


Fig. 23. WNT signaling pathway

(a) PATHWAY 1- XM_006721558.3 (B9 Domain containing 1)

The mitogen activated protein kinase (MAPK) pathway is a highly conserved mechanism that plays an important role in cellular proliferation, development, and motility. However, each MAPKK can be activated by a significant number of MAPKKKs, increasing the complexity of MAPK signaling (Fig. 22).

(b) PATHWAY 2- XM_005262368.4 (Dishevelled associated activator of Morphogenesis 1):-

WNT proteins are secreted morphogens that are required for basic developmental processes in a range of species and organs, involving cell fate determination, progenitor proliferation and differentiation, and asymmetric cell division control (Fig. 23).

(c) PATHWAY 3- NM_003058.4 (Solute Carrier Family 22 Member 2):

Oncogenesis and tumour growth are associated to abnormal choline metabolism that is emerging as a metabolic characteristic. Phosphocholine (PCho), diacylglycerol (DAG),

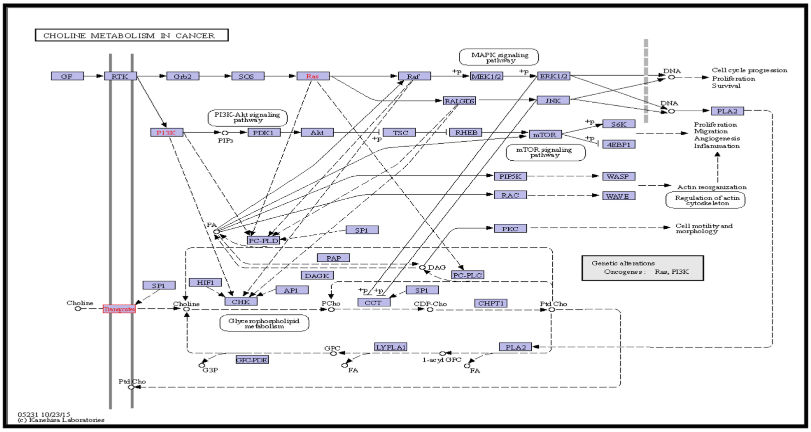


Fig. 24. Choline metabolism in cancer

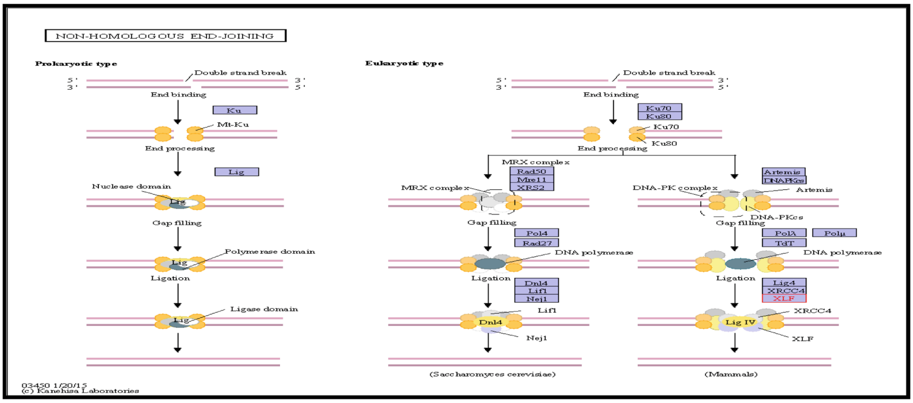


Fig. 25. Non homologous end joining pathway

and phosphatidic acid constitute choline phospholipid metabolism products which may work as signalling molecules (Fig. 24).

(d) PATHWAY 4- NM_001330987.2 (Non homologous end joining pathway):
By direct ligation, nonhomologous end joining (NHEJ) eliminates DNA double-strand breaks (DSBs). NHEJ is error-prone since it repairs DSBs at all stages of the cell cycle, leading in the ligation of two DNA DSBs without the requirement for sequence homology (Fig. 25).

(e) PATHWAY 5- NM_024782.3 (Mismatch repair):
DNA mismatches that arise during DNA replication, preventing mutations in dividing cells from becoming permanent. MMR slows homologous recombination and it has recently been connected to DNA damage signalling (Fig. 26).

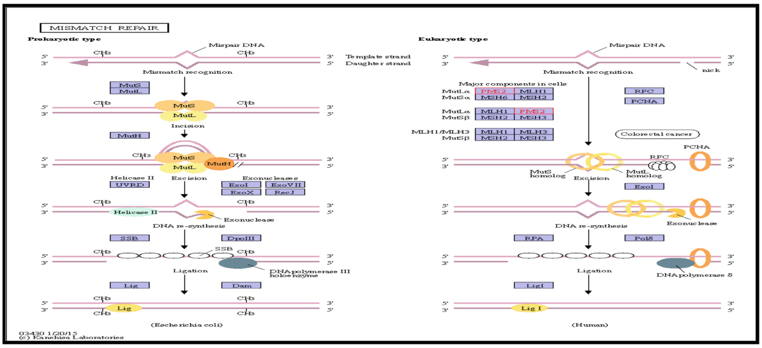


Fig. 26. Mismatch repair pathway

4 Conclusion

Hodgkin lymphoma is the disease that was taken to wade through the process of ChIP-seq. The BCL6 gene is the major regulator of GCB cell development and the function is through the recruitment of corepressor complexes that catalyzes broad epigenetic changes. To understand genetic regulatory system, in living beings thus identification of short repeating patterns is done by motif analysis. ChIP-seq analysis was made using the Galaxy platform.

References

1. Park P. J. (2009). ChIP-seq: advantages and challenges of a maturing technology. Nature reviews. Genetics, 10(10), 669–680. https://doi.org/10.1038/nrg2641
2. Schmidt, D., Wilson, M. D., Spyrou, C., Brown, G. D., Hadfield, J., & Odom, D. T. (2009). ChIP-seq: using high-throughput sequencing to discover protein-DNA interactions. Methods (San Diego, Calif.), 48(3), 240–248. https://doi.org/10.1016/j.ymeth.2009.03.001
3. Shanbhag S, Ambinder RF. (2018) Hodgkin lymphoma: A review and update on recent progress. CA Cancer J Clin. 68(2):116-132.
4. Jardin F, Ruminy P, Bastard C, Tilly H. The BCL6 proto-oncogene: a leading role during germinal center development and lymphomagenesis. Pathol Biol (Paris). 2007 Feb;55(1):73-83. doi: https://doi.org/10.1016/j.patbio.2006.04.001. Epub 2006 Jul 3. PMID: 16815642.
5. Yang, H., & Green, M. R. (2019). Epigenetic Programing of B-Cell Lymphoma by BCL6 and Its Genetic Deregulation. Frontiers in cell and developmental biology, 7, 272. https://doi.org/10.3389/fcell.2019.00272
6. Hatzi, K., & Melnick, A. (2014). Breaking bad in the germinal center: how deregulation of BCL6 contributes to lymphomagenesis. Trends in molecular medicine, 20(6), 343–352. https://doi.org/10.1016/j.molmed.2014.03.001
7. Proc Natl Acad Sci U S A, C C Chang et al (1996) Jul 9;93(14):6947–52. BCL-6, a POZ/zinc-finger protein, is a sequence-specific transcriptional repressor doi: https://doi.org/10.1073/pnas.93.14.6947
8. Hcjnjsanc
9. Andrews, S. (n.d.). FastQC A Quality Control tool for High Throughput Sequence Data. Retrieved from http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

10. Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology, 10(3), R25. https://doi.org/10.1186/gb-2009-10-3-r25
11. Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. Nature Methods, 9(4), 357–359. https://doi.org/10.1038/nmeth.192
12. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … and, R. D. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics, 25(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352
13. Definition of SAM/BAM format. (n.d.). Retrieved from https://samtools.github.io/hts-specs/
14. Segregation based metric for variant call QC. (n.d.). Retrieved from http://samtools.github.io/bcftools/rd-SegBias.pdf
15. Li, H. (2011). Improving SNP discovery by base alignment quality. Bioinformatics, 27(8), 1157–1158. https://doi.org/10.1093/bioinformatics/btr076
16. Multiallelic calling model in bcftools (-m). (n.d.). Retrieved from http://samtools.github.io/bcftools/call-m.pdf
17. Institute, B. (n.d.). Picard. Broad Institute, GitHub repository. GitHub. Retrieved from http://broadinstitute.github.io/picard/
18. Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., … Liu, X. S. (2008). Model-based Analysis of ChIP-Seq (MACS). Genome Biology, 9(9), R137. https://doi.org/10.1186/gb-2008-9-9-r137
19. Feng, J., Liu, T., Qin, B., Zhang, Y., & Liu, X. S. (2012). Identifying ChIP-seq enrichment using MACS. Nature Protocols, 7(9), 1728–1740. https://doi.org/10.1038/nprot.2012.101
20. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ.(2004) The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* Jan 1;32(Database issue):D493–6.
21. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ(2004) The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* Jan 1;32(Database issue):D493–6.
22. Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics, 26(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033
23. Liu T, Ortiz JA, Taing L, Brown M, Turpaz Y, Liu XS et al.(2011) *Genome Biol* ; 12(8):R83
24. Andrew Waterhouse, Martino Bertoni,Torsten Schwede et al *Nucleic Acids Research*, Volume 46, Issue W1, 2 July 2018, Pages W296–W303, https://doi.org/10.1093/nar/gky427
25. Bertoni, M., Kiefer, F., Biasini, M. et al. Modeling protein quaternary structure of homo- and hetero-oligomers beyond binary interactions by homology. Sci Rep 7, 10480 (2017). https://doi.org/10.1038/s41598-017-09654-8
26. Gabriel Studer, Christine Rempfer, Andrew M Waterhouse, Rafal Gumienny *Bioinformatics*, Volume 36, Issue 6, 15 March 2020, Pages 1765-1771, https://doi.org/10.1093/bioinformatics/btz828
27. Nicolas Guex,Manuel C. Peitsch,Torsten Schwede(2009) Automated comparative protein structure modeling with SWISS-MODEL and Swiss-PdbViewer: A historical perspective https://doi.org/10.1002/elps.200900140
28. Stefan Bienert, Andrew Waterhouse, Tjaart A. P. de Beer, Gerardo Tauriello, Gabriel Studer, Lorenza Bordoli, Torsten Schwede *Nucleic Acids Research*, Volume 45, Issue D1, January 2017, Pages D313–D319, https://doi.org/10.1093/nar/gkw1132
29. Kanehisa, M. and Sato, Y.(2020); KEGG Mapper for inferring cellular functions from protein sequences. Protein Sci. 29, 28–35
30. Kanehisa, M., Sato, Y., and Kawashima, M.(2022); KEGG mapping tools for uncovering hidden features in biological data. Protein Sci. 31, 47–53