EPIGENETIC REGULATIONS, MOTIF AND PATHWAY IDENTIFICATION OF MYELOFIBROSIS CHIP SEQUENCES

Sravani Sane[1*], Shylesh Murthy IA[2] and Preenon Bagchi[3]

[1*] Padmashree Institute of Management and Sciences, Bengaluru, India. Corresponding author: Email: sanesravani836@gmail.com
[2]Junior Researcher, Vasishth Academy of Advanced Studies and Research, Bengaluru, India.
[3]Institute of Biosciences and Technology, MGM University, India.

Abstract: Next generation sequence technique that is Chromatin immunoprecipitation (ChIP) on the ChIP–Seq of Myelofibrosis. Myelofibrosis is a type of blood cancer which is also considered as a form of chronic leukemia. It is two types primary myelofibrosis and secondary myelofibrosis its leads to complication of an autoimmune disease, additional disease features include hepatosplenomegaly, extramedullary hematopoiesis (EMH) etc., In this work, I have retrieved its chip-Seq from SRA database, removed the PCR duplicates and finally identified the genome enriched region using the MAC2 call peak tool This annotated peak table was used to identify the motifs present in the chip-Seq of Myelofibrosis. Peak table was annotated to identify the genes and the corresponding pathways was identified from the KEGG pathway.

Keywords: Chromatin immunoprecipitation (ChIP), Myelofibrosis, Next Generation sequencing, PCR duplicates, MACS2 Callpeak, Motif, KEGG pathway.

1 INTRODUCTION

ChIP sequencing (ChIP-Seq) is a powerful method for identifying genome-wide DNA binding sites for transcription factors and other proteins. “Next generation” genome sequences technology can provide one – two orders of magnitude increase in the amount of sequence that will be cost effective and new technology ChIP – Seq directly provide whole genome of mammalian Protein – DNA interactions [1, 2]. It starts with crosslinking of DNA-protein complexes and then fragmented, they are treated with an exonuclease to cut unbound oligonucleotides. Protein-specific antibodies are used to immunoprecipitation the DNA-protein complex. DNA is extracted and sequenced, giving high-resolution sequences of the protein-binding sites. ChIP Seq has many advantages such as, it captures DNA targets for transcription factors or histone modifications across the entire genome of any organism and also it defines transcription factor binding sites [3]. ChIP–Seq could be the way for genome-wide profiling of DNA-binding proteins, or nucleosomes and attributable to the tremendous progress in next-generation sequencing technology [4].

Myelofibrosis is a type of blood cancer which is also considered as a form of chronic leukemia [5, 6]. In this disease, the bone marrow is replaced by fibrous scar tissue. It is two types primary myelofibrosis (occurs on its own) and secondary myelofibrosis (oc-

R. Somashekhar et al. (eds.), *Proceedings of the International Conference on Advances in Nano-Neuro-Bio-Quantum (ICAN 2023)*,
Advances in Intelligent Systems Research 69,
https://doi.org/10.2991/978-94-6463-294-1_15

curs as the result of separate disease) its leads to complication of an autoimmune disease. The bone marrow develops three types of cells white blood cells, red blood cells and platelets [7]. When mutation occurs in a single cell of DNA, it passes on to new cells and leading to defective of cell divides. Characteristic of myelofibrosis like megakaryocytes (over production of giant cells) [8], decreases production of red blood cells – may occur anemia and thrombocytopenia (deficiency of platelet production). Overall world in about 12% cases, primary myelofibrosis will progress to acute myeloid leukemia fastly growing of blood cancer [9, 10].

2. MATERIALS AND METHODS

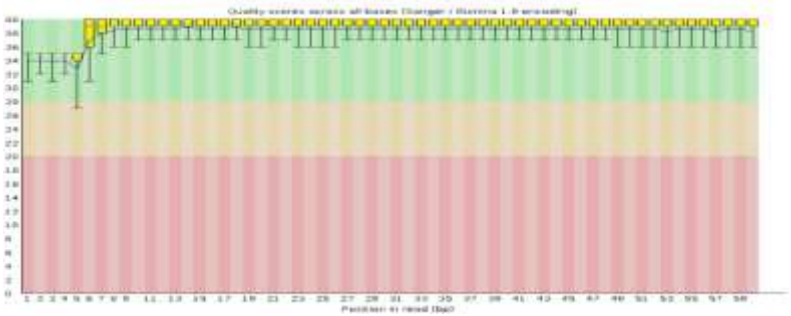
The primary aim of the current research is to provide a detailed overview of epigenetic regulations of Myelofibrosis and throw a light on the pathways involved [11]. Galaxy tutorial by Lauren Mills, Analyzing ChIP-Seq Data in Galaxy is used to analyze *Mus musculus* Myelofibrosis fastq sequences having SRA accession number DRR311742 and DRR311743 were retrived from SRA database. We mapped the reads using Bowtie2 [12]. Using Collect Alignment Summary Metrics tool we take the summary of our alignment. Next using RmDup we remove PCR duplicates [13] and Collect Alignment Summary Metrics tool was re-run. Finally, using MACS2 Callpeak we identify peaks from alignment results [14, 15]. Using Peak calling we identify areas in our genome that have been enriched with our aligned reads, these areas are those where protein interacts with DNA.

Next, we annotate our peaks table to take top 100 most significant peaks and identify the genes overlapping with these peaks and Genes were selected model using Swiss prot with Ramachandran plot [16,17,18] was done. We identify the motifs using SeqPos motif analysis tool. Biological sequence motifs are short conserved sequence pattern associated with distinct functions that usually represents important structural or functional features [19]. Finally, the pathways of these genes were identified from KEGG pathways [20, 21].

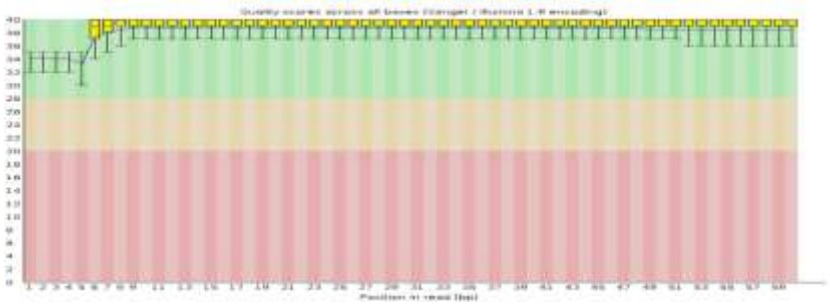
3. RESULTS AND DISCUSSION

3.1 FASTQC quality reports

FASTQC quality reports was given quality control to *Mus musculus* chip-sequences with SRA accession number DRR311742 and DRR311743. FASTQC results were given in fig. 2 DRR311742 and fig.3 DRR311743.



(Fig. 1) FASTQC DRR311742

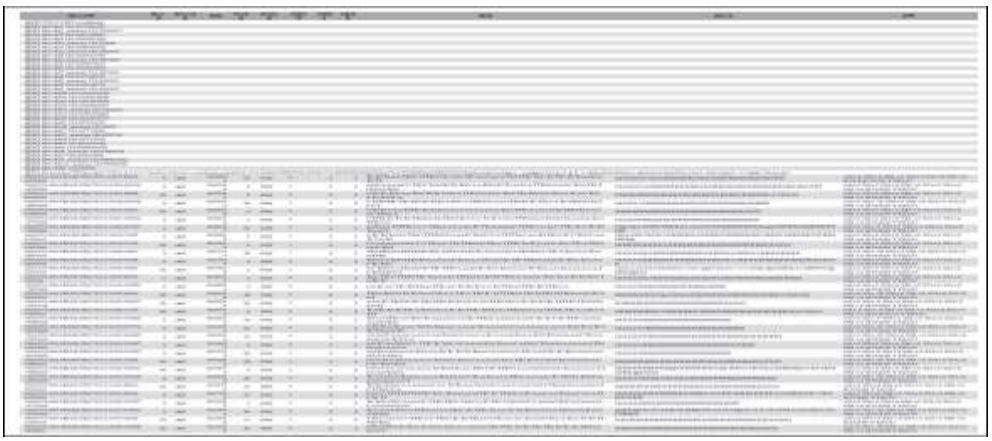


(Fig. 2) FASTQC DRR311743

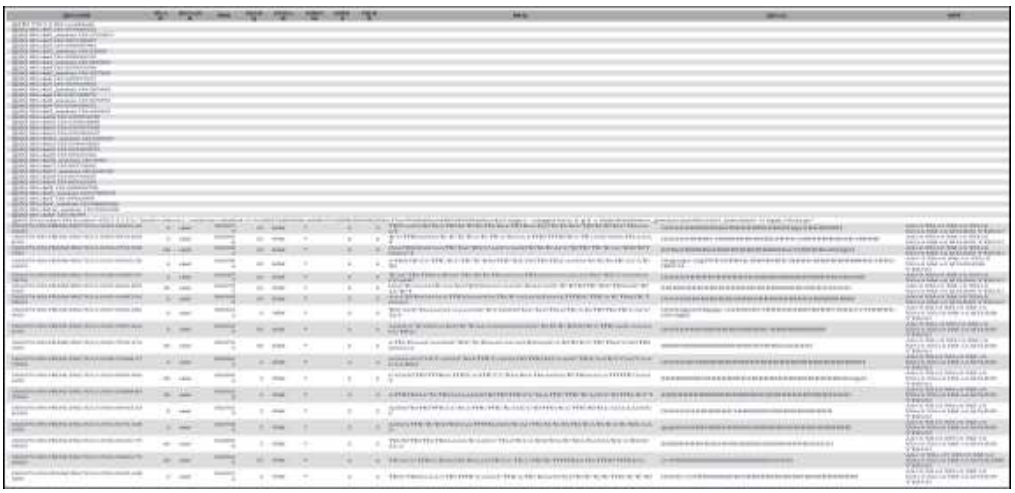
3.2 Bowtie2

Mus musculus chip-sequences with SRA accession number DRR311742 and DRR311743 were mapped using *Mus musculus* ref Seq using Bowtie2.

Mapping output of DRR311742:



Mapping output of DRR311743:



3.3 Collect alignment before:

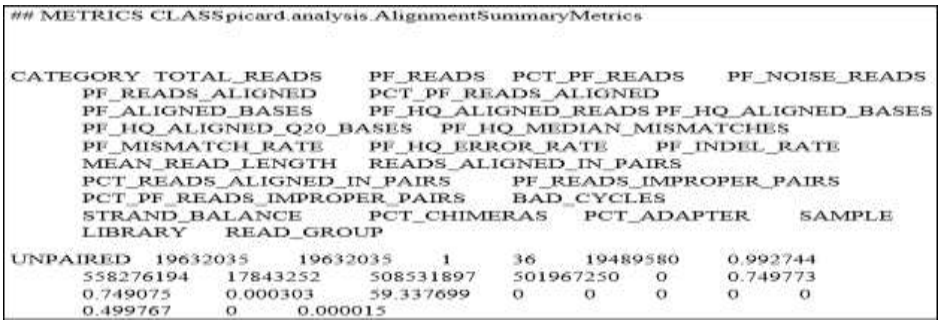
We use the tool Collect Alignment Summary Metrics tool take the summary of our mapping done above. Table 1 contains the alignment summary for DRR311742and Table 2 for DRR311743.

Table: 1 Alignment summary for DRR311742

METRICS CLASSpicard.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	20249149	20249149	1	41	20088298	0.992056	572908355	18350815	519991244	512894155	0	0.750071	0.74941	0.000289	59.260725	0	0	0	0	0	0.499535	0	0.000015			

Table: 2 Alignment summary for DRR311743

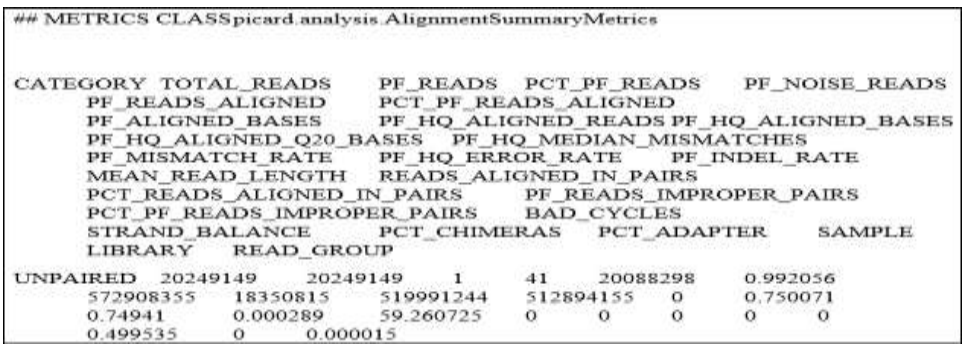


METRICS CLASS picard.analysis.AlignmentSummaryMetrics

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	19632035	19632035	1	36	19489580	0.992744
558276194	17843252	508531897	501967250	0	0.749773	
0.749075	0.000303	59.337699	0	0	0	0	0
0.499767	0	0.000015				

Next, we removed the PCR duplicates using RmDup. Table 3 & 4 contains the Alignment summary for DRR311742 and DRR311743 and removing PCR duplicates.

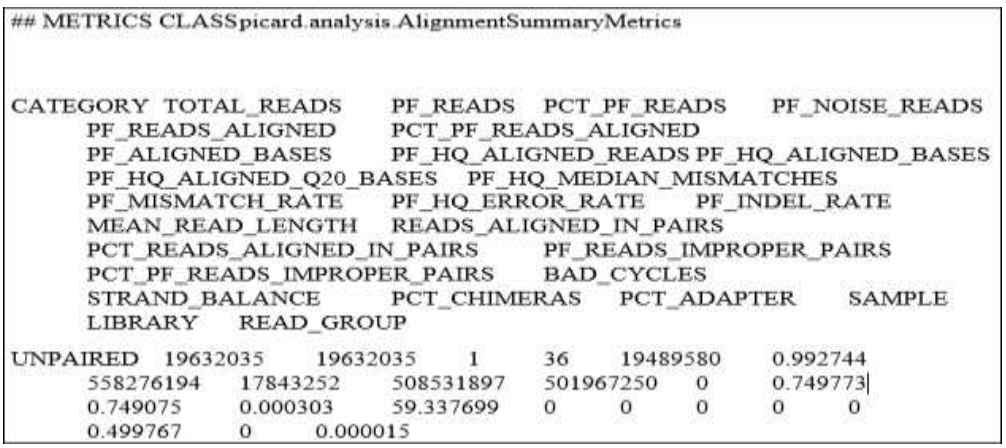
Table: 3 Alignment summary for DRR311742 post RmDup



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	20249149	20249149	1	41	20088298	0.992056
572908355	18350815	519991244	512894155	0	0.750071	
0.74941	0.000289	59.260725	0	0	0	0	0
0.499535	0	0.000015				

Table: 4 Alignment summary for DRR311743 post RmDup

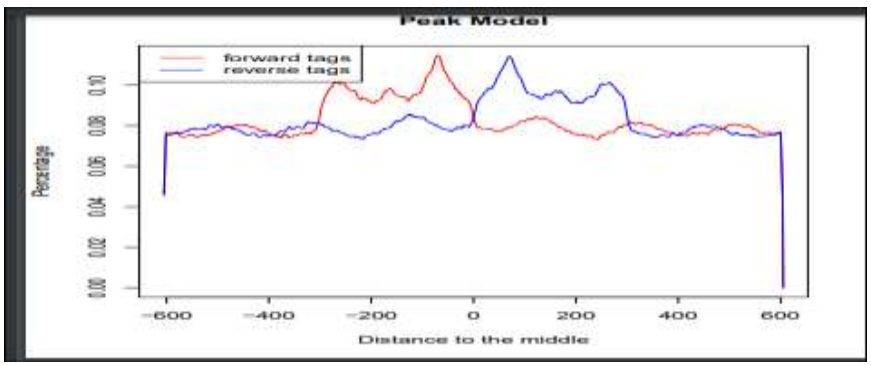


METRICS CLASS picard.analysis.AlignmentSummaryMetrics

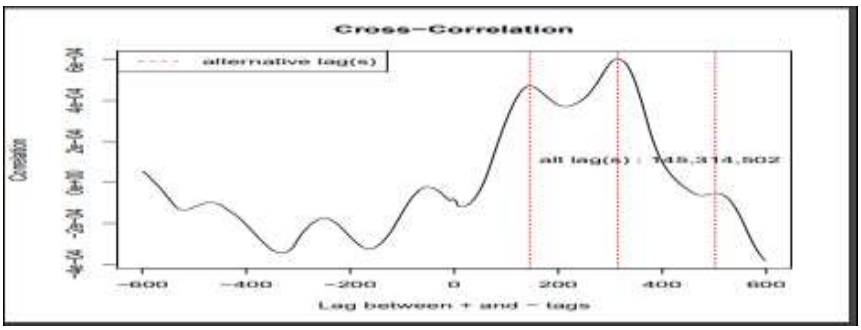
CATEGORY	TOTAL_READS	PF_READS	PCT_PF_READS	PF_NOISE_READS
PF_READS_ALIGNED	PCT_PF_READS_ALIGNED			
PF_ALIGNED_BASES	PF_HQ_ALIGNED_READS	PF_HQ_ALIGNED_BASES		
PF_HQ_ALIGNED_Q20_BASES	PF_HQ_MEDIAN_MISMATCHES			
PF_MISMATCH_RATE	PF_HQ_ERROR_RATE	PF_INDEL_RATE		
MEAN_READ_LENGTH	READS_ALIGNED_IN_PAIRS			
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	19632035	19632035	1	36	19489580	0.992744
558276194	17843252	508531897	501967250	0	0.749773	
0.749075	0.000303	59.337699	0	0	0	0	0
0.499767	0	0.000015				

3.4 MACS2 Callpeak:

As per the alignment summary (Table 3, 4), we see that the reads are less post RmDup which implies that the duplicate reads are removed. Next, we use MAC2 call peak tool to identify areas in the genome that are enriched with the aligned reads. Model-based Analysis of ChIP-Seq (MACS) is a commonly used tool for identifying transcription factor binding sites. The algorithm confines the influence of genome complexity to evaluate the significance of enriched ChIP regions. This tool improves the spatial resolution of binding sites by combining the information of both sequencing tag position and orientation. Here, MACS is used along with a control sample (DRR311742) which increases specificity of the peak calls (Fig. iii). MACS2 models the distance between the paired forward and reverse strand peaks and uses 1000 enriched regions to model the distance between the forward and reverse strand peaks.



(Fig. 4) peak model in graphical format



(Fig. 5) Correlation metric of Peak

3.5 Motif analysis:

We identify the motifs present in our Myelofibrosis genome ChIP-Seq. We used SeqPos motif analysis tool. The file, top 100 most significant peaks in bed format was selected for motif identification (Table 5).

Table – 5 Motif Analysis:

clusters	collapsed_id	Factor	DNA binding domain	hits	cutoff	zscore	-10*log(pval)	similarity to top	mean_position
1	M01231	GLIS3		46	2.86	-3.046	67.608		-0.152
2	MA0024	E2F1	E2F	48	0.965	-2.849	61.239		-0.143
	M00050	E2F1/E2f1	Fork head / winged helix	54	0.858	-2.598	53.613	0.999	-0.127
	M00740	E2F1		36	1.1.02	-2.395	47.896	0.983	-0.146
	M00738	E2F4::TFDP1		51	0.78	-2.338	46.367	0.976	-0.122
3.	MA0145	Tcfcp2l1	CP2	52	2.579	-2.712	56.997		-0.133
4.	hPDI021	RFXANK		38	4.527	-2.664	55.56		-0.152
5.	M01715	MAC1		27	2.364	-2.48	50.257		-0.171
6.	M00469	AP-2alpha/TFAP2A		31	4.772	-2.35	46.672		-0.155
	MA0003	TFAP2A	Helix-Loop-Helix	31	4.772	-2.35	46.672	1.0	-0.155

3.6 SWISS PORT

Genes were selected from the Myelofibrosis genome, model using Swiss prot with Ramachandran plot.

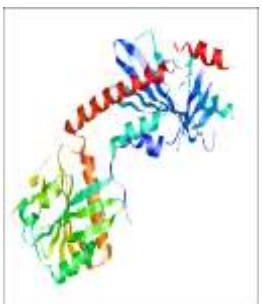
Table: 6 Genes are involved in Myelofibrosis with Accession number:

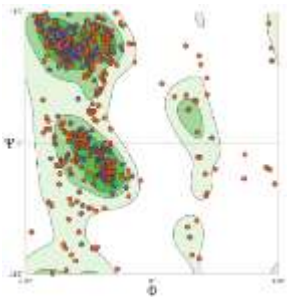
Sl. No	Genes	Name	Accession number
1.	ENSMUST00000116495	Rap1gap2	NP_001015046.1
2.	ENSMUST00000183873	Lrrc1	NP_001139520.1
3.	ENSMUST0000058725	Antxr1	NP_766396.2
4.	ENSMUST00000025862	Smarca2	NP_035546.2
5.	ENSMUST00000113531	Lnx1	NP_001346003.1

Abbreviations of genes:

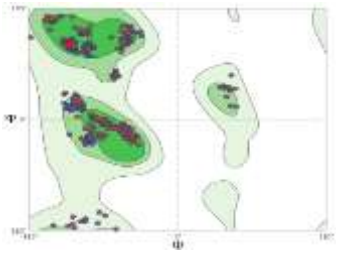
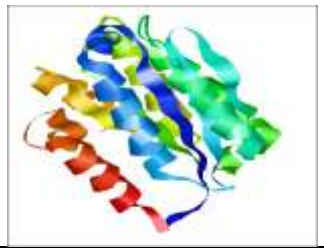
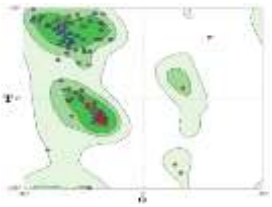
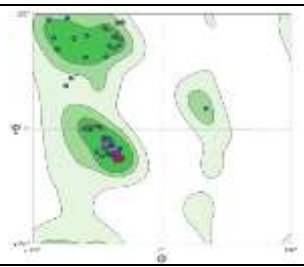
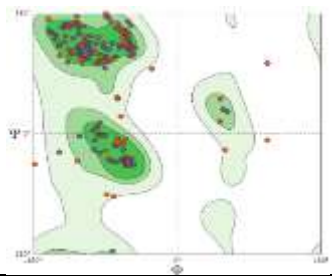
1. Rap1gap2: Rap1 GTPase activating protein-2
2. Lrrc1: Leucine rich repeat containing 1
3. Antxr1: Anthrax toxin receptor – like
4. Smarca2: Sw1/SNF related, matrix associated actin depended of chromatin subfamily
5. Lnx1: Ligand of numb protein x1

The receptor model and corresponding Ramachandran plot results are given in Fig.6. Accession number used for modeling is given in table 6.





Rap1 gap2

	
Lrcc1	
	
Antxr1	
	
Smarca2	
	
Lnx1	

(Fig.6) Swiss-model generated receptor models with their Ramachandran plot.

4. KEGG pathway analysis:

Finally, from the annotation results our peaks table by taking top 100 most significant peaks, we identify the genes overlapping with these peaks which are given in Table 7. The pathways of these genes were identified from KEGG pathways names in Table: 7

Table: 7 Identified Genes and pathways

Sl. No.	GENES	PATHWAYS
1.	NP_001139520 to NM_001146048.1	Tight junctions (TJs)
2.	NP_001015046.1 to NM_001015046.	RAP1 signaling
3.	NP_766396.2 to NM_172808	cGMP - PKG signaling
4.	NP_035546.2 to XM_036161667.1	Thermogenesis
5.	NP_001346003.1 to XM_036164835.1	Tight junctions (TJs)

4.1 Tight Junctions (TJs) Pathway:

Tight junctions (TJs) are essential for establishing a selectively permeable barrier to diffusion through the paracellular space between neighboring cells. TJs are composed of at least three types of transmembrane protein -occluding, claudin and junctional adhesion molecules (JAMs)- and a cytoplasmic 'plaque' consisting of many different proteins that form large complexes. These are proposed to be involved in junction assembly, barrier regulation, cell polarity, gene transcription, and other pathways.

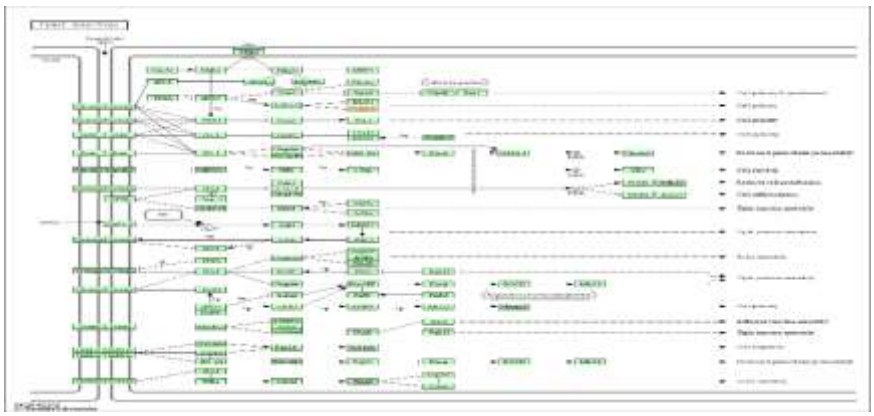


Fig. 7 Tight Junction Pathway

4.2 RAP1 signaling Pathway:

Rap1 is a small GTPase that controls diverse processes, such as cell adhesion, cell-cell junction formation and cell polarity. Like all G proteins, Rap1 cycles between an inactive GDP-bound and an active GTP-bound conformation. A variety of extracellular signals control this cycle through the regulation of several unique guanine nucleotide exchange factors (GEFs) and GTPase activating proteins (GAPs). Rap1 plays a dominant role in the control of cell-cell and cell-matrix interactions by regulating the function of integrin's and other adhesion molecules in various cell types. Rap1 also regulates MAP kinase (MAPK) activity in a manner highly dependent on the context of cell types.

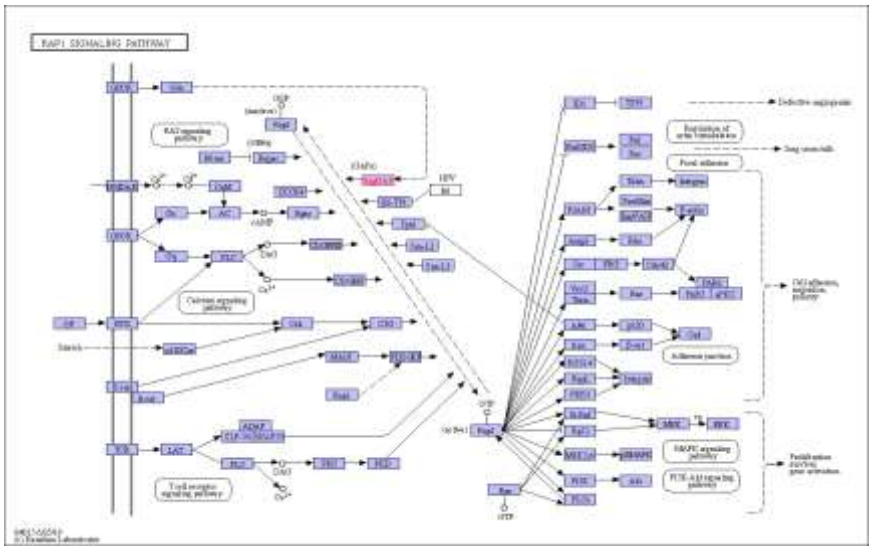


Fig.8 RAP1 signaling Pathway

4.3 cGMP - PKG signaling Pathway:

Cyclic GMP (cGMP) is the intracellular second messenger that mediates the action of nitric oxide (NO) and natriuretic peptides (NPs), regulating a broad array of physiologic processes. The elevated intracellular cGMP level exerts its physiological action through two forms of cGMP-dependent protein kinase (PKG), cGMP-regulated phosphodiesterase's (PDE2, PDE3) and cGMP-gated cation channels, among which PKGs might be the primary mediator. PKG1 isoform-specific activation of established substrates leads to reduction of cytosolic calcium concentration and/or decrease in the sensitivity of myofilaments to Ca2+ (Ca2+-desensitization), resulting in smooth muscle relaxation. In cardiac myocyte, PKG directly phosphorylates a member of the transient potential receptor canonical channel family, TRPC6, suppressing this nonselective ion channel's Ca2+ conductance, G-alpha-q agonist-induced NFAT activation, and myocyte hypertrophic responses. PKG also opens mitochondrial ATP-sensitive K+ (mitoKATP) channels and subsequent release of ROS triggers cardio protection.

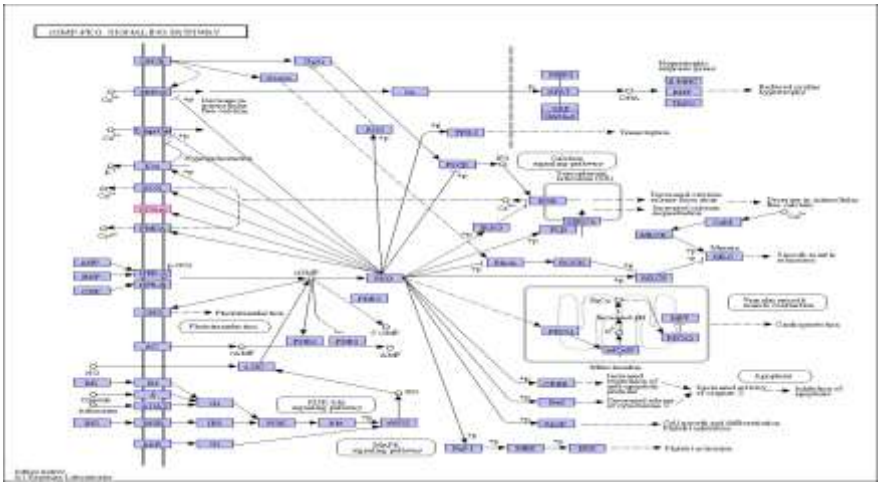


Fig. 10 cGMP - PKG signaling Pathway

4.4 Thermogenesis Pathway:

Thermogenesis is essential for warm-blooded animals, ensuring normal cellular and physiological function under conditions of environmental challenge. Thermogenesis in brown and beige adipose tissue is mainly controlled by norepinephrine, which is released from sympathetic nervous system in response to cold or dietary stimuli. The mitochondrial uncoupling protein 1 (UCP1) is responsible for the process whereby chemical energy is converted into heat in these adipocytes. Activation of these adipocytes leads to an increase in calorie consumption and is expected to improve overweight conditions, providing a potential strategy for treating obesity and its related metabolic disorders

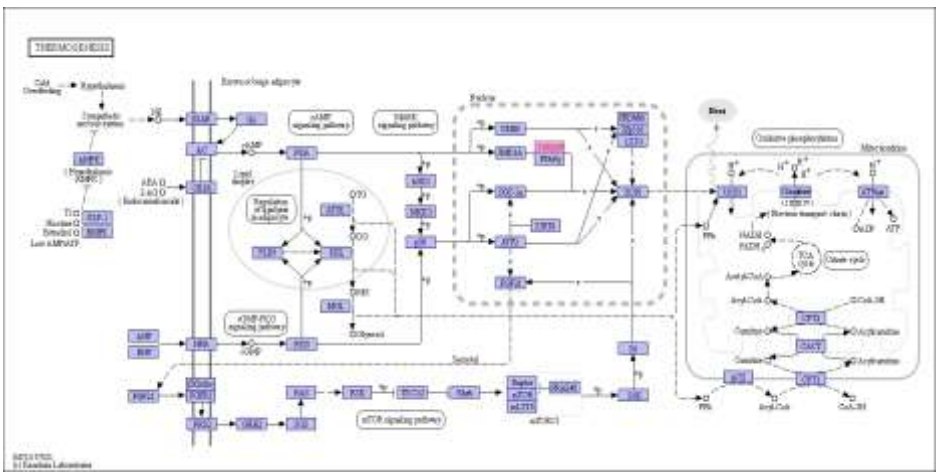


Fig. 10 Thermogenesis Pathway

4.5 Tight Junction (TJs) Pathway:

Tight junctions (TJs) are essential for establishing a selectively permeable barrier to diffusion through the paracellular space between neighboring cells. TJs are composed of at least three types of transmembrane protein -occluding, claudin and junctional adhesion molecules (JAMs)- and a cytoplasmic 'plaque' consisting of many different proteins that form large complexes. These are proposed to be involved in junction assembly, barrier regulation, cell polarity, gene transcription, and other pathways.

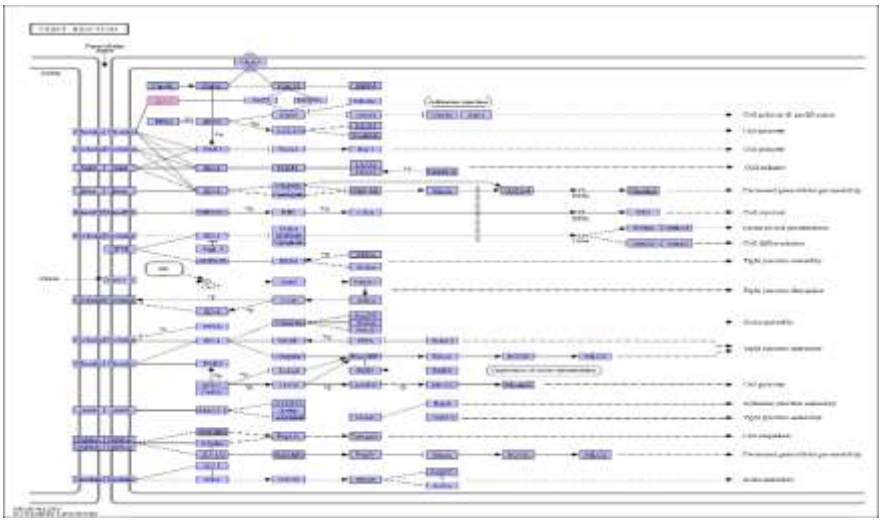


Fig.12 Tight Junction Pathway

1. CONCLUSION

A high-quality immunoprecipitation ChIP peak is detected in our results which imply high concentration of DNA binding proteins. Further, motifs were detected from the ChIP peak. Further, top DNA binding protein genes were identified and corresponding pathways was identified from KEGG pathway. Tight junctions (TJs) pathway, RAP1 signaling pathway, cGMP - PKG signaling pathway, Thermogenesis pathway and Tight junctions (TJs).

6. REFERENCES

1. Deliard S, Zhao J, Xia Q, Grant SF. Generation of high quality chromatin immunoprecipitation DNA template for high-throughput sequencing (ChIP-seq). J Vis Exp. 2013 Apr 19;(74):50286.
2. Furey TS. ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. Nat Rev Genet. 2012;13(12):840-852.

3. Pavesi, Giulio (2016). [Advances in Biochemical Engineering/Biotechnology] || ChIP-Seq Data Analysis to Define Transcriptional Regulatory Networks. , (Chapter 43), –. doi:10.1007/10_2016_43 .
4. Meyer CA, Liu XS. Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. Nat Rev Genet. 2014;15(11):709-721.
5. Rumi E, Trotti C, Vanni D, Casetti IC, Pietra D, Sant'Antonio E. The Genetic Basis of Primary Myelofibrosis and Its Clinical Relevance. Int J Mol Sci. 2020 Nov 24;21(23):8885. doi: 10.3390/ijms21238885. PMID: 33255170; PMCID: PMC7727658.
6. Mughal TI, Vaddi K, Sarlis NJ, Verstovsek S. Myelofibrosis-associated complications: pathogenesis, clinical manifestations, and effects on outcomes. Int J Gen Med. 2014;7:89-101. Published 2014 Jan 29. doi:10.2147/IJGM.S51800.
7. Zahr AA, Salama ME, Carreau N, Tremblay D, Verstovsek S, Mesa R, Hoffman R, Mascarenhas J. Bone marrow fibrosis in myelofibrosis: pathogenesis, prognosis and targeted strategies. Haematologica. 2016 Jun;101(6):660-71. doi: 10.3324/haematol.2015.141283. PMID: 27252511; PMCID: PMC5013940.
8. J. T. Reilly; D. Barnett; G. Dolan; P. Forrest; J. Eastham; A. Smith (1993). Characterization of an acute micromegakaryocytic leukaemia: evidence for the pathogenesis of myelofibrosis. , 83(1), 58–62. doi:10.1111/j.1365-2141.1993.tb04631.x
9. Tefferi A. Primary myelofibrosis: 2019 update on diagnosis, risk-stratification and management. Am J Hematol. 2018 Dec;93(12):1551-1560. doi: 10.1002/ajh.25230. Epub 2018 Oct 26. PMID: 30039550.
10. ().The Korean Journal of Internal Medicine, 33(4), –. doi:10.3904/kjim.2018.033
11. Ambrosini, G., Dreos, R., Kumar, S. et al. The ChIP-Seq tools and web server: a resource for analyzing ChIP-seq and other types of genomic data. BMC Genomics 2016, 17, 938.
12. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012 Mar 4;9(4):357-9. doi: 10.1038/nmeth.1923. PMID: 22388286; PMCID: PMC3322381.
13. Mark T. W. Ebbert,Mark E. Wadsworth,Lyndsay A. Staley… (2016). Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches. BMC Bioinformatics, 17(7 Supplement), –. doi:10.1186/s12859-016-1097-3
14. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS. (2008) Model-based Analysis of ChIP-Seq (MACS), Genome Biology, 2008;9 (9):R137.
15. Baxevanis, Andreas D.; Petsko, Gregory A.; Stein, Lincoln D.; Stormo, Gary D. (2002). Current Protocols in Bioinformatics || Using MACS to Identify Peaks from ChIP-Seq Data. , (), –. doi:10.1002/0471250953.bi0214s34
16. Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res. 2000 Jan 1;28(1):45-8. doi: 10.1093/nar/28.1.45. PMID: 10592178; PMCID: PMC102476.
17. Hollingsworth SA, Karplus PA. A fresh look at the Ramachandran plot and the occurrence of standard structures in proteins. Biomol Concepts. 2010 Oct;1(3-4):271-283. doi: 10.1515/BMC.2010.022. PMID: 21436958; PMCID: PMC3061398.

18. Mannige, R. V., Kundu, J., & Whitelam, S. (2016). *The Ramachandran Number: An Order Parameter for Protein Geometry. PLOS ONE, 11(8), e0160023.* doi:10.1371/journal.pone.0160023
19. Salma Aouled El Haj Mohamed, Mourad Elloumi and Julie D. Thompson (December 14th 2016). Motif Discovery in Protein Sequences, Pattern Recognition - Analysis and Applications, S. Ramakrishnan, IntechOpen, DOI: 10.5772/65441
20. Altermann, E., Klaenhammer, T.R. PathwayVoyager: pathway mapping using the Kyoto Encyclopedia of Genes and Genomes (KEGG) database. BMC Genomics 6, 60 (2005). https://doi.org/10.1186/1471-2164-6-60.
21. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28(1):27-30. doi:10.1093/nar/28.1.27