PhD degree in Molecular Medicine (curriculum in Computational Biology)

European School of Molecular Medicine (SEMM),

University of Milan and University of Naples "Federico II"

Settore disciplinare: Bio/11

ChIP_QC, computational platform for multivariate epigenetic studies and its application in uncovering role of polycomb dependent methylations states

SRIGANESH JAMMULA

IEO, Milan

Matricola n. R09869

Supervisor: Dr. Diego Pasini

European Institute of Oncology, IEO, Milan

Added Supervisor: Dr. Mattia Pelizzola

Italian Institute of Technology, IIT, Milan

Anno accademico 2014-2015

Table of Contents

LIST OF ABBREVIATIONS

Acetylated Lysine 9 of histone H3 (H3K9ac)

Acetylated Lysine 27 of histone H3 (H3K27ac)

Analysis of Variance (ANOVA)

Area Under Curve (AUC)

Bayesian Network (BN)

Cervical carcinoma (HeLa-S3)

Chromatin immunoprecipitation (ChIP)

CpG island (CpGi)

Complementary DNA (cDNA)

CREB binding protein (CBP)

Cytosine-guanine dinucleotides (CpG)

Deacetylase enzymes (HDAC).

Dimethylated Lysine 27 of histone H3 (H3K27me2)

Dimethylated Lysine 79 of histone H3 (H3K79me2)

Embryoid bodies (EBs)

Embryonic ectoderm development (Eed)

Enhancer of Zeste homolog (Ezh1/2)

Epidermal Kertainocytes (NHEK)

False Positive Rate (FPR)

Fold Change (FC)

Graphical User Interface (GUI)

H1 human embryonic stem cell (H1hESC)

Histone Modifications (HMs)

Histone Acetyl Transferase (HAT)

High Throughput Sequencing (HTS)

Human Lung Fibroblast (NHLF)

Immunoprecipitation (IP)

Induced pluripotent stem cells (iPSCs)

Keratinocytes (Nhek)

Knockout (KO)

Liver carcinoma (Hepg2)

Leukemia (K562)

Lymphblastoid (Gm12878)

Lysine Methyl Transferase (KTM)

Mass spectrometry (MS)

Monomethylated Lysine 27 of histone H3 (H3K27me1)

Mono-ubiquitination of the lysine 119 on the histone H2a (H2aK119Ubq)

Mouse embryonic fibroblast (MEF)

Mouse embryonic stem cells (mESC)

Mouse Reference Genome (mm9)

Next Generation Sequencing (NGS)

Polycomb group protein (PcG)

Polycomb Repressive Complex 1 (PRC1)

Polycomb Repressive Complex 2 (PRC2)

Post translational modification (PTM)

Principal component analysis (PCA)

Quality Control (QC)

Reads Per Kb per Million (RPKM)

Real time quantitative PCR (qRT-PCR)

Regions Of Interest (ROI)

Retinoblastoma protein associated protein 46/48 (RbAp46/48)

RNA sequencing (RNA-seq)

Receiver Operating Characteristic (ROC)

Short hairpin RNA (shRNA)

Skeletal Muscle Fibroblast (Hsmm)

Support Vector Machine (SVM)

Suppressor of zeste (Suz12)

Transcription end site (TES)

Transcription factors (TFs)

Transcription start site (TSS)

Trimethylated Lysine 4 of histone H3 (H3K4me3)

Trimethylated Lysine 27 of histone H3 (H3K27me3)

Trimethylated Lysine 36 of histone H3 (H3K36me3)

True Positive Rate (TPR)

Umbilical Vein Endothelial Cells (HUVEC)

Western blot (WB)

Wild type (WT)

LIST OF FIGURES

genes in each cluster as identified in A. Top panel represents expression levels for target gene clusters for H3K27ac regions where as lower panel represents for H3K27me3 positive regions. (C) Same as A (D) Same as A, but quantification, normalization and scaling are restricted only to ROI.

Fig. 2.1.3.2 A,B. Quantification (A) Intensities of H3K27ac ChIP around ±5kb region surrounding the centre of enhancer regions across five different cell lines. (B) Expression levels of target genes in Gm12878 in clusters as identified in A.

Fig. 2.1.3.2 C,D. Quantification (C) Intensities of H3K27ac ChIP around 5kb region surrounding from the centre of enhancer regions across five different cell lines. (D) Same as A, where intensities are scaled globally over all samples.

Fig. 2.1.3.3 A-D. Profiling (A) Average profile of H3K4me3 with confidence interval in promoters regions of genes classified based on expression levels (high to low). (B) Average profile of H3K36me3 with confidence interval in gene bodies of genes classified based on expression levels (high to low). (C) Same as A, but without using strand information. (D) Same as B, but without using strand information.

Fig. 2.1.3.4 A-D. Spike-In Quantification (A) Normalized intensities of H3K79me2 around 10kb surrounding TSS (both up and downstream) in regions possessing H3K79me2 in WT samples and its fate in other samples induced with different levels of inhibitor harbouring no reference genome. (B) Same as A, in these intensities are spike-in normalized intensities. (C) Average normalized profile of H3K79me2 around 10kb surrounding TSS (both up and downstream) in regions possessing H3K79me2 in WT samples and its fate in other samples induced with different levels of inhibitor

harbouring no reference genome. (D) Same as C, in these intensities are spike-in normalized intensities.

Fig. 2.1.4 A-E. Differentially regulated regions. (A) Volcano plot representing significantly enriched promoters (marked in cyan) harbouring different levels of H3K4me3 methylation in skeletal muscle when compared keratinocytes. (B) Distribution of expression levels of genes where their promoters show significantly higher levels of H3K4me3 in skeletal muscle as compared to that of keratinocytes. (C) Distribution of expression levels of genes where their promoters show significantly higher levels of H3K4me3 in keratinocytes as compared to that of skeletal muscle. (D) Tissue specificity of genes whose promoters were differentially regulated skeletal muscle as identified in A. (E) Tissue specificity of genes whose promoters were differentially regulated keratinocytes as identified in A.

Fig. 2.1.4 F,G. Differentially regulated regions. (F) Significantly enriched promoters on the basis of K4me3 across 9 different cell lines. Represented here are their intensities in standard z-score form. (G) Expression level of target genes in each cluster across 9 different cell lines identified in F.

Fig. 2.1.5 A-C. Probabilistic relationships (A) Bayesian network showing dependency between different factors in compact chromatin regions of genome presided by Suz12. (B) Bayesian network showing dependency between different factors in random regions of genome. (C) Normalized intensities of Suz12, Ezh2 and Ctbp2 in Suz12 binding regions.

Fig. 2.1.6 A,B. Variable selection and classification. (A) Plot signifying the accuracy of different set of variables for characterizing active enhancers and promoters. (B) Sensitivity over specificity of SVM trained model for classifying active enhancers and promoters using variables with high accuracy level identified in A.

Fig.2.2.1 A,B. PTMs on H3K27 in mESC and its regulation by PRC2. (A) Larger pie graph show relative abundance of different PTMs on lysine 27 of Histone H3. Smaller pies show the same PTMs but in different Histone variants H3.2 and H3.3. (B) Western blot analysis showing loss of all forms of methylations on H3K27 using in and *Eed*, *Ezh2* and *Suz12* KO (-/-) as compared to that of indicated antibodies of protein extracts obtained from WT (+/+) mESC line. Similar trend was observed on knock down of Eed and Suz12 using shRNA in E14. Histone H3 served as loading control.

Fig. 2.2.2 A. Localization of different forms of H3K27 methylation Genomic regions showing enrichment for different forms of methylations on H3K27. H3K27me1 and H3K27me3 enriched domains in genome are depicted in blue and red.

Fig. 2.2.2 B,C. Correlation between PTMs. (B) Scatter plots showing the correlation of enrichments normalized to the histone H3 density of between K27 and K36 PTMs in gene bodies of all annotated genes. Pearson correlation values are indicated on top of the plot. (C) Variable plot from Principal component analysis (PCA) representing degree of correlation between PTMs in gene bodies of all annotated genes.

Fig. 2.2.3 A-C. Correlation between levels of K27 methylation and gene transcription. (A) Expression levels of all RefSeq genes grouped in three categories relative to H3K27me2 and H3K27me1 enrichments within their gene bodies. (B) Proportion of

K27me1 and K27me2 enriched genes within each group of expression. (C) Composite profiles of H3K27me1 and H3K27me2 over gene bodies for all the three groups of gene sets classified on the basis of their expression level.

Fig. 2.2.4 A-D. PRC2 dependent H3K27 methylation. (A) qRT-PCR of K27me1/2 ChIP in WT and Eed KO samples in the selected genomic regions. Black boxes indicate primers position within genomic loci. ChIP enrichments are normalized to histone H3 density. IgG ChIPs from rabbit were used as negative control. (B) Genomic snapshots of H3K27me1/2/3 in WT (Eed +/+) and Eed KO (Eed -/-) in mESC along with H3K36me3 from E14 mESC. H3K27me1 domains are highlighted in blue while H3K27me3 domains are highlighted in red. (C) Heat map of H3K27me1 enrichment in WT (Eed +/+) and Eed KO (Eed −/−) for genes enriched for H3K27me1 in WT condition (−10log10 p value ≥ 10 scored from chi-square test between H3K27me1 and H3). (D) Box plot analysis of H3K27me1 ChIP-seq enrichment intensities between WT (+/+) and Eed KO (-/-) mESC for all the annotated RefSeq genes that were divided in two groups based on their H3K27me1 levels in WT mESC (-Log10p-value cut off = 10).

Fig. 2.2.4 E-G. Changes of genes expression upon loss of PRC2 activity. (E) Box plot of fold change in expression levels of differentially regulated genes between WT and Eed KO mESC for H3K27me2 and H3K27me1. For the analysis, the top 15% enriched genes (N~1000) were considered. (F) Relative differences in expression levels between WT and Eed KO mESC of the selected target genes determined by qRT-PCR analysis. (G) qRT-PCR analysis for the indicated intragenic regions of H3K27me1 and H3K27me2 ChIP assays performed in WT and Eed KO mESC using. ChIPs with IgG

rabbit were performed as negative control. ChIP enrichments were normalized to histone H3 density.

Fig. 2.2.5 A. mESC deficient for PRC2 fail to differentiate. Relative expression of the indicated differentiation markers determined by qRT-PCR in WT and *Eed* KO mESC before (ES) and after 9 days of differentiation (EB).

Fig. 2.2.5 B,C. H3K27me1 is gained in genes which are up-regulated in the process of differentiation. (B) Expression levels of up-regulated genes during differentiation process in WT and Eed KO samples (N=844). (C) Average profiles of H3K27me1 and H3K36me3 through the intragenic regions of genes activated upon EB differentiation.

Fig. 2.2.6 A. Global levels of H3K27ac increase upon loss of PRC2 activity. WB analysis of different modifications upon loss of different components of PRC2 in mESC.

Fig. 2.2.6 B,C. Distribution of H3K27ac enriched regions upon loss of PRC2 activity. (B) Overlap of H3K27ac peaks between WT (Eed +/+) and Eed KO (Eed -/-) mESC. The pies depict the percentage distribution of the different groups of H3K27ac peaks relative to promoter region of all genes. Promoters regions are defined as a ±. 2.5kb region around centered the TSS. (C) Snapshots representing different PTMs in regions where H3K27ac is lost and gained in WT (Eed +/+) and Eed KO (Eed -/-) mESC highlighted in yellow.

Fig. 2.2.6 D,E. Mapping enhancer elements upon loss of H3K27me2. (D) Snapshots representing different PTMs in regions where H3K27ac is lost and gained in WT (Eed +/+) and Eed KO (Eed -/-) mESC highlighted in yellow. (E) Box plot showing levels of H3K4me1 signal in the unique H3K27ac distal peaks of Eed WT and Eed KO samples. Number of Eed WT unique peaks = 12341; Eed KO unique peaks = 9210

Fig. 2.2.6 F, G. Regulation of enhancers upon loss of H3K27me2 in mESC. (F) Heatmap of normalized intensities of H3K27ac, H3K4me1, H3K4me3, H3K27me3, H3K27me2 in WT and Eed KO mESC for all distal H3K27ac peaks found in either WT (Eed WT unique peaks) or Eed KO (Eed KO unique peaks). Classification of H3K27ac peaks found only in Eed KO into two classes, Class I (n = 4,391) and Class II (n = 4,819) was applied on the basis of pre-existence of H3K4me1 in Eed WT sample. Grouping was based on k mean clustering (k = 2) with respect to the H3K4me1 normalized intensities in Eed WT ESCs. (G) Boxplot analyses quantifying the data shown in figure 2.2.6 F. p value was calculated by Wilcoxon rank test.

Fig. 2.2.6 H. Validation of lost and gained enhancer elements. (H) qRT-PCR analyses of DNA purified from H3K27ac ChIP in WT and Eed KO mESC using primers amplifying the indicated genomic loci.

Fig. 2.2.6 I,J. Unique enhancers are not enriched for H3K27me3 and do not reside on CpG islands. (I) Box plot showing the quantification of H3K27me3 signal in the unique H3K27ac distal peaks of Eed WT and Eed KO samples. (J) Percentage of Ezh2 peaks occupancy (determined by ChIP-seq analysis in mouse E14 ES cells) and of CpG islands respect to genomic regions corresponding to H3K27ac peaks uniquely found in Eed KO mESC.

Fig. 2.2.6 K,L. Activation of enhancers upon loss of PRC2 correlates with closest gene activation (K) Box plot representing distance between enhancers (for WT and Eed KO samples) and the up-regulated genes in Eed KO ES cells. All identified enhancers are included in the analysis. (L) Same as K, but in this case enhancers associated to a H3K27me3 positive gene in WT ES cells were excluded from the analysis. H3K27me3

enriched genes were defined by the presence of a H3K27me3 peak within +/- 2.5kb from the TSS. p-values were calculated by Mann-Whitney Test.

Fig.2.2.8 M-P. Anti-correlation between H3K27ac and H3K27me2 at unique enhancers sites and loss of H3K27ac at enhancer sites is replaced by H3K27me2. (M) Scatter plot showing correlation between H3K27ac and H3K27me2 levels in WT ESCs for all unique enhancers regions identified in WT and Eed KO samples. Left panel shows whole density distributions. Right panel distinguishes Eed WT unique (red) and Eed KO unique (blue) enhancers. The Spearman correlation value is indicated (rs = 0.5106). p value was calculated by asymptotic t approximation (N) Immunoblot analysis for H3K27ac antibody of histones extracted from mouse E14 mESC treated with 35 μM C646 p300 inhibitor for 48 h. DMSO was used as vehicle control. Histone H3 was used as loading control. (O) Average profiles of H3K27ac and H3K27me2 deposition around 2500 bp up and downstream from centered H3K27ac peak summit of regions that loose H3K27ac upon treatment with C646 compound for 48 h (N=4838). (P) Box plots with quantification levels of H3K27ac and H3K27me2 at the same enhancer sites of figure 2.2.8 O upon treatment with C646 for the complete H3K27ac peak region or for a 1kb genomic region surrounding the summit of peak.

Fig 3.1.5.1 Workflow of quantification within ROIs.

Fig 3.1.5.2 Workflow for genome wide based quantification for selected ROI.

Fig. 4.2. Our proposed model on different functionalities of PRC2 dependent methylation forms.

ABSTRACT

During my PhD tenure, I have been involved in developing a user-friendly cross-platform system capable of analyzing epigenomic data and further use it in understanding the role of the Polycomb Repressive Complex 2 (PRC2) in genome regulation.

From current trending in epigenetics research, we can sense increasing ease of high throughput sequencing and greater interest towards genome wide epigenomic studies. As a result of which we experience an exponential flooding of epigenetic related data such as Chromatin immunoprecipitation followed by sequencing (ChIP-seq), and RNA sequencing (RNA-seq) in public domain. This creates an opportunity for crowd sourcing and exploring data outside the boundaries of specific query centered studies. Such data has to undergo standard primary analysis, which with the aid of multiple programs has been stabilized courtesy to the scientific community. Further downstream, out of many, genome wide comparative, correlative and quantitative studies have proven to be critical and helpful in deciphering key biological features. For such studies we lack platforms, which can be capable of handling, analyzing and linking multiple interdisciplinary (ChIP-seq/RNA-seq) datasets with efficient analytical methods. With this aim we developed ChIP_QC, a user-friendly standalone computational program with an ability to support numerous datasets with high/moderate sequencing depth for performing genome wide analysis. First, using ENCODE project (Consortium, 2012) data, we illustrated few applications of the program by posing different biological scenarios and showed the comfort with which some known observations can be verified and also how it can be helpful in deducing some other novel observations.

Second, we were interested in understanding the functionality of the products generated through catalytic activity of PRC2. It is known that Lysine 27 of histone H3 (H3K27) undergoes posttranslational modification (PTM) and methylation is one such dominant PTM. Methylation on H3K27 can be either mono/di/tri-methylation form. Out of all three forms, it is very well demonstrated that trimethylation of H3K27 (H3K27me3) is PRC2 dependent and at the same time its role in gene repression is well characterized, but functional roles of other forms of methylation on H3K27 are still poorly characterized. For understanding this, we used mouse embryonic stem cells (mESC) as model system of our study and we were able to provide an extensive characterization of other forms of methylation, highlighting their differential deposition along the genome, their fundamental role in transcriptional regulation, and their indispensability during differentiation program. Using ChIP_QC and with other computational methods along with experimental evidences, our data demonstrated that the monomethylation of Lys27 (H3K27me1) is required for correct transcription of genes and positively correlates with trimethylated Lys36 (H3K36me3); on the other hand dimethylated Lys27 (H3K27me2), that we identified to be the principal activity of PRC2, prevents firing of non cell type specific enhancers.

Chapter 1 - INTRODUCTION

1.1. Epigenetics

Cell is the basic unit of life and the blue print of every cell is in its DNA. Every cell maintains its identity through robust genome organization. It differs from single celluar to multicellular organism. As compared to prokaryotes, eukaryotic genome undergoes much complex organization and is regulated at many different layers. One major contributing factor for this property is Epigenetics. It plays vital role by its influence on genome at different levels starting from changes at single nucleotide to higher order of its organization. Epigenetic behavior differs from one cell type to other. It maintains unique behavior in individual cell type by regulating expression of set of genes required for maintaining that specific system. Transcription can be affected by chromatin organization providing access to chromatin modifiers/transcription factors for driving or repressing transcription. With these consequences, epigenetics becomes driving factor for deciding the fate of a cell.

1.1.1. Chromatin and its structural organization

In eukaryotes, both genomic content and volume of nucleus of a cell vary indefinitely. In order to fit complete genome into the small volume of nucleus, it has to undergo a high degree of compaction (Woodcock and Ghosh, 2010). This is achieved by set of proteins, which are associated with genomic DNA forming bead like structures called nucleosomes, which are further organized forming chromatin fibers, which are folded hierarchically within the nucleus ultimately condensing size of DNA. Nucleosomes are core functional units of chromatin fibers, consisting of 147 bp DNA wrapped around an eight histone proteins forming octamer complex (Luger et al., 1997). Each octamer

complex consists of four histones: H2A, H2B, H3, and H4 and are organized in a manner where H3–H4 tetramer binds to two adjacent H2A–H2B dimers (Luger et al., 1997). Histone octamer complexes are connected through linker DNA thus resulting in string of nucleosomes. Linker histones H1 bind to these regions and are situated at the sites of DNA entry and exit to the nucleosome core (Luger, 2003).

All histones, except H4, exist in different variants, which differ in their amino acid sequence and are expressed at very low levels as compared to that with canonical histones. Histone variants change chromatin dynamics and are incorporated into the nucleosome as a footprint that guides the cell to regulate transcription, repair, chromosome assembly and segregation. According to their function, histone variants can be of two kinds: replicative and replacement. Replicative histones are encoded by multiple gene copies which are expressed in S phase and their incorporation into chromatin is coupled to DNA synthesis; in humans they are represented by H3.1 and H3.2 (Szenker et al., 2011). On the other hand, replacement histone variants are encoded by single gene, often in a tissue specific manner, which are transcribed throughout the cell cycle. Histone H3.3 is one of the known replacement histone variants in humans and is very well characterized, it is largely considered as marker of transcriptional activity (Szenker et al., 2011).

Nucleosomes, the core functional unit of chromatin are the main determinant to DNA accessibility. The characteristics of an individual nucleosome depend on the DNA sequence it is wrapped around (Wallrath et al., 1994) and also its stability and positioning is governed on the basis of different PTMs it resides on its histone tails (Luger et al., 1997). Nucleosomes can be arranged very close to each other or can be arranged distantly resulting in compaction or loosening of chromatin, such

arrangement becomes deciding factor for gene transcription. Taking into consideration such arrangement of nucleosomes and corresponding transcriptional status, chromatin exits in two different forms namely euchromatin and heterochromatin. Euchromatin can be described as loosely packed chromatin, providing accessibility to transcription factors and other regulators thus favoring transcription. Due to this active transcription euchromatin can also be known as active chromatin. On the other hand, heterochromatin is known as inactive chromatin where chromatin is tightly packed and compressed, not accessible to any transcription factor or other regulators resulting in transcriptional repression.

1.1.2. Histone Modifications

From early studies since 60's, it was evident that histone proteins are subject of PTM at their N-terminal tails (Allfrey et al., 1964). Since then, over 100 different histone modifications have been discovered and studies in deep detail. From X-ray structure of the nucleosome (Luger et al., 1997) it was demonstrated that PTMs are able to influence the chromatin structure. Histone modifications play vital role in nucleosome stability. They are the principle factors for recruiting chromatin modelers and other regulatory proteins for positioning nucleosome and regulating transcription accordingly. Among all, some well known and characterized PTMs are methylation, acetylation and ubiquitination on lysine, phoshorilation on threonine and serine; newly identified, but less abundant PTMs are serine/threonine O-GlcNAcylation (Arnaudo and Garcia, 2013) and crotonilation (Tan et al., 2011), lysine butyrylation and proprionylation (Chen et al., 2007), lysyl 5-hydroxylation (Unoki et al., 2013). Most PTMs occurs at the N- and C-terminal "tail" domains protruding from the nucleosome core particle, but a significant fraction of modification takes place

also in the globular domain of the histones, which regulates histone-histone and histone-DNA interactions (Cosgrove et al., 2004). From the findings of different experiments it was possible in relating different histone modifications with gene activation or repression and also in better defining various regulatory regions of genome such as active or repressed promoters of a gene, active/poised enhancers, and transcribed regions. In brief, repressed promoters of a gene is marked by H3K27me3 (Morey and Helin, 2010) where as active promoter of a gene is marked by H3K4me3 (Bernstein et al., 2005). Similarly, active enhancers are marked by H3K27ac, H3K4me1 and with very minimal H3K4me3 where as poised enhancers are marked by H3K4me1 with very minimal H3K4me3 (Creyghton et al., 2010a).

1.1.3. Transcriptomics

In eukaryotes, every cell contains the same genome and thus the same genes. However, not every gene is transcriptionally active in every cell, different cells show different patterns of gene expression. These differences in expression are governed by wide range of physical and biochemical factors of a cell. Within a cell, not complete genome is transcribed only small proportion of genome is transcribed which is then translated into functional protein. This small proportion of transcribed genome is estimated to nearly 5 percent of whole genome (Frith et al., 2005). In recent years, it was discovered that fraction of genome to be transcribed but not further coded into functional protein. Such transcribed RNA is termed as non-coding RNA. With the aid of transcriptomics studies, we can understand which genes are active or inactive in various types of cells and tissues. This information can further be useful in understanding the dynamics behind different expression patterns both within and across similar or different cell types. With support of high throughput technology

today we can measure levels of expression individual gene. Using this one can be able to quantify transcripts and get to know which set of genes are highly expressed and contribute in defining unique properties of that cell. Furthermore, transcriptomics can also be used in understanding alterations in expression patterns in cancer cells as compared its normal counterpart.

1.2. Next Generation Sequencing

With advance in technology and the search for better scientific evidence, current epigenomic studies don't restrict their observations to any local environment of genome but are interested in characterizing the phenomena at genome wide level. This resulted in designing experiments coupled with Next Generation Sequencing (NGS), also known as high throughput sequencing (HTS). It allows massive parallel sequencing during which millions of fragments of DNA are sequenced in relatively very small amount of time at appreciable cost. In practice, HTS is applied in broad spectrum of disciplines. For instance, it is used for genome assembly, mutational studies, transcription profiling and also for epigenetic studies. Apart from these major types, there are many other supplement sequencing methods for different purposes. Here as part of epigenomic study, we stress on chromatin immunoprecipitation and RNA based sequencing methodologies.

In recent years, the development of chromatin immunoprecipitation (ChIP) assay, coupled to HTS technology (ChIP-seq) has provided a powerful tool for investigating the nature and patterns of deposition of several histone PTMs/target protein binding events in the genome of different organisms and has helped to unravel their functions through unbiased approaches. This allowed a greater understanding of the molecular mechanisms behind transcriptional regulation.

Similarly, RNA-Seq is another deep-sequencing technology in which mRNA molecules from population of cells of same kind are extracted and sequenced in any NGS platforms, allowing measuring the levels of expression of different genes. This information has been used extensively in many research studies for characterizing respective biological phenomena.

Taking advantage of above two sequencing techniques and linking their results help us in better understanding transcriptional regulation.

1.2.1. Primary Data Analysis

Before making any biological inferences from the data generated through HTS, it has to undergo a series of computational analysis steps. Once sequencing data is generated through ChIP-seq/RNA-seq, it is initially checked for its quality. Once satisfied, data is aligned to known reference genome. This constitutes primary data analysis, which is common for both ChIP-seq/RNA-seq. These steps are explained below in much more detail.

1.2.1.1. Quality Control

It is always recommended to check the quality of data generated through NGS platforms. At times quality can be poor due to bad library preparation or due to sequencing platform specific errors or both. Some possible issues during library preparation can be amplification biasness, contamination of sample with some other unkown samples, high level of duplication and many others. Similarly, sequencing platform specific issues can be variation in reading and scoring quality of bases (Schmieder and Edwards, 2011; Zhou et al., 2013). For such reasons, it is certainly important to test the quality of data generated through these platforms. In terms of

quality control (QC), every data should be assessed for the quality of raw reads using metrics generated by the sequencing platform. FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), is one of the most popular tools used for initial QC of sequencing data. It generates report containing base quality information, duplication levels, GC content per sequence, duplication levels and other necessary information. Similarly, FaQC (Lo and Chain, 2014), NGS QC Toolkit (Patel and Jain, 2012) and QC-Chain (Zhou et al., 2013) are some other open source tools that can be used for similar purpose.

1.2.1.2. Alignment

Once sequencing data passes through quality control, next analytical step is alignment of reads to reference genome. Scanning huge number of sequenced reads through reference genome for alignment is inefficient, time taking and computationally expensive. To avoid this, reference genome is indexed prior to alignment process. Once indexed, multiple sequenced samples can be aligned easily. Different software index genome differently. But most software use either suffix/prefix tries or hash tables based index algorithms. Some popular software designed on the basis of suffix/prefix tries are BWT(Li and Durbin, 2009), Bowtie (Langmead et al., 2009) and many others. Similarly, software designed on the basis of hash tables are BFAST (Homer et al., 2009), MAQ (Li et al., 2008a) and others. Once reference genome is indexed, actual alignment of sequenced reads to genome is initiated. Irrespective of indexing method, alignment of sequenced reads with reference genome is done either by Smith–Waterman (Smith and Waterman, 1981) or the Needle–Wunsch (Needleman and Wunsch, 1970) algorithms. Depending on the alignment tool used, a

resulting alignment is either gapped or ungapped. ChIP-seq data can be directly aligned to reference genome through above-mentioned approach.

As compared to ChIP-seq data, RNA-seq data are not genomic DNA but are processed complementary DNA (cDNA) generated from mRNA. Usually, mRNA is transcribed from genomic DNA, which then undergoes splicing process removing introns and joining exons. For this reason RNA-seq reads cannot be aligned directly to reference genome. For aligning such data special considerations are to be considered. Two approaches are followed for aligning RNA-seq data. One is guided approach, in which gene models are used for constructing possible splice junctions; these junctions and exon genomic regions are then used directly for aligning sequenced reads. And the other is unguided approach, where reads are aligned directly to the genome, identifying potential exons. Junctions are then constructed on the basis of exons derived from mapped reads. Other remaining reads are mapped to these newly constructed junctions. These aligned junctions are further used for constructing gene model. TopHat (Trapnell et al., 2009), PALMapper (Jean et al., 2010) and STAR (Dobin et al., 2013) are some software that follows guided approach for aligning reads where as MapSplice (Wang et al., 2010) and TopHat (Trapnell et al., 2009) are other software that uses non-guided approach for aligning reads. Most RNA-seq aligners are developed on top of short read alignment tools such as bowtie and SOAP (Li et al., 2008b).

1.2.2. Secondary Data Analysis

In this stage, ChIP-seq aligned data is further processed with the aim at finding enriched regions signifying potential binding of target protein or presence of PTMs.

At the same time, RNA-seq aligned data is processed to measure the level of expression of different genes.

1.2.2.1. Peak Calling

Once ChIP-seq data is aligned to reference genome, next critical step is to identify enriched regions across whole genome. This processing is commonly known as Peaking Calling. Simplest method for identifying such enriched regions will be sliding window approach; in which read density in fixed length of window is computed across whole genome (Nix et al., 2008; Qin et al., 2010; Spyrou et al., 2009; Zhang et al., 2008). This approach is further improved by considering Gaussian kernel estimator, which results in continuous signal density and avoids dependency of window size (Boyle et al., 2008; Lun et al., 2009). In these approaches, certain height of read density is considered as criteria for identifying significantly enriched regions and in cases enriched regions lie with in some minimal distance are merged together. To further know the statistical confidence of enriched regions background models are taken into consideration. These background models are statistical models that are constructed on the basis of low coverage regions in genome. Later, looking for more precise results, many software take into consideration sequencing data generated from control data this can be ChIP from any non specific antibody like IgG or input of fixed chromatin generated without any specificity for any antibody. With the support of control datasets peak calling can be improved in many ways either by subtracting background signal of control data from target ChIP removing noise or by considering fold changes differences between target ChIP and control data (Chen et al., 2008; Johnson et al., 2007). This way many false positive peaks can be controlled.

Many software implement different statistical models for calling enriched regions. Models like Poisson (Valouev et al., 2008), local Poisson (Zhang et al., 2008), t-distribution (Blahnik et al., 2010), conditional binomial (Nix et al., 2008; Rozowsky et al., 2009) and hidden Markov (Qin et al., 2010; Spyrou et al., 2009) models are being used by different software. Using these models each enriched region is assigned with some significant value and with the support of control dataset false discovery rate is computed. All these methods ultimately aim at calling enriched regions with higher level of confidence and reduce false positive rate.

1.2.2.2. mRNA quantification

On the basis of annotated gene model and aligned RNA-seq reads, reads lying within exons of a gene are summed and can be used as a measure of level of expression for that gene. Comparison of these expression levels between genes of different lengths can be misleading, as shorter genes will have greater count than lengthy genes with few count but their concentration in sample is same. Such biasness can be avoided by normalizing read count with the length of mRNA and sequencing depth to obtain expression level in terms of Reads Per Kb per Million (RPKM) values (Mortazavi et al., 2008). These RPKMs can now be used for comparing expressing levels between different genes within a sample. Several tools like ERANGE (Mortazavi et al., 2008), Tophat (Trapnell et al., 2009) and RSAT (Medina-Rivera et al., 2015) provide provision for computing RPKMs. For computing RPKM for a locus, ERANGE takes into consideration read counts of all known and novel exons where as TopHat (Trapnell et al., 2009) and RSAT (Medina-Rivera et al., 2015) make use of only specified exons.

1.2.3. Tertiary Data Analysis

Using data from secondary analysis further downstream analysis is planned. Depending upon biological questions to be answered tertiary analysis differs from one experiment to another. In particular ChIP-seq data can be further processed in many different ways like ChIP enriched regions can be mapped to closest gene and complete gene set can be annotated to know which biological process/molecular functions or biochemical pathways are enriched. GREAT (McLean et al., 2010) is one such application, which is specifically designed for annotating such regions of interest. Other common task is to identify highly represented motifs in enriched regions. MEME-ChIP (Machanick and Bailey, 2011), Pscan (Zambelli et al., 2009) are few such programs, which are commonly used for such purposes. Similarly RNA-seq data from different experiments can be further processed specifically for identifying differential expressed genes. Cuffdiff2 (Trapnell et al., 2013), DESeq2 (Love et al., 2014) and edgeR (Zhou et al., 2014) are most commonly used tools for identifying differentially regulated genes.

In terms of usage, different programs are designed with the perspective to make all users easy to carry out computational analysis. The only systems that render specific programs available in a linked pipeline are Galaxy (Goecks et al., 2010) and Cistrome (Liu et al., 2011). With the recent understanding about the level of complexity of epigenomic dynamics such minimal analysis remain insufficient. Taking together the data from different studies creates an opportunity for exploring their relationship at a much deeper level which can contributes to a better characterization of the genome wide dynamics and solve hidden layers of regulation. For doing so programs with in-build analytical and data mining methods, power for supporting bulk-processed data from different disciplines are needed. While primary and its extended analysis (i.e.

ChIP-seq peak calling) are mature and broadly available, programs with such extended capabilities are still not available; restricting the analytical power to highly experienced computational biologists. With such approach tools like SeqMINER (Ye et al., 2011), and some modules of HOMER (Heinz et al., 2010), Cistrome (Liu et al., 2011) provide provisions for quantitative and correlative analysis, which remain restricted to a limited framework. However, the above mentioned tools do not provide provision for handling multiple samples, link changes within a sample with transcription, predicting dependencies, filtering datasets on the basis of their relevance, identifying features to characterize samples and to perform differential analysis. Moreover programs like HOMER (Heinz et al., 2010) and Cistrome (Liu et al., 2011) cannot deal with raw aligned data as they require processed aligned data, which add a further layer of complexity. Apart from these standalone applications and command line tools there several packages designed in R for similar purpose. RepiTools (Statham et al., 2010) for analysis of enrichment based epigenomic data, ChIPpeakAnno (Zhu et al., 2010) for annotating enriched regions, DiffBind (Ross-Innes et al., 2012) for identifying differentially regulated regions between experiments and many others are available in binconductor. Main disadvantage of using R packages and other command line tools is that it requires prior knowledge of programming. From a biologist point of view this would be very tedious job. Even for small task this would be painful.

Taking into consideration these concerns, we were interested in developing a light weighted standalone open source application with genome wide analytical features. The application is designed in a way that it should be easy to use with minimal input files and can be used by any biologist with minimal computational background thus reducing dependencies on others. Such application should be made available with

Graphical User Interface (GUI) and command line facility. Providing GUI option makes biologist more easy to use. With these considerations we designed ChIP_QC, a computational program for Quantitative and Correlative analysis of ChIP-seq data.

1.3. Polycomb Group Proteins

1.3.1. Overview

Polycomb group proteins (PcGs) are family of proteins, which were initially discovered in *D. Melanogaster* (Lewis, 1978). They are regarded as key players in the process of development and tissue morphogenesis. As research continued on these proteins, their homologs were identified in mammalians too (Brunk et al., 1991; van Lohuizen et al., 1991a). Bmi was the very initial homolog identified in mammalian system whose activity was linked with Myc in inducing lymphomagenesis (Haupt et al., 1991; van Lohuizen et al., 1991b). And slowly other homologs in mammalian system were identified (Schumacher and Magnuson, 1997). These developments led to in-depth characterization of these proteins in mammalian system and it resulted that they exist as multi-protein complex in cell nuclei (Piunti and Pasini, 2011). PcGs exist mainly in two complexes namely Polycomb Repressive Complex 1 (PRC1) and Polycomb Repressive Complex 2 (PRC2). Each complex is made of several proteins and their functional role is not fully understood (Schwartz and Pirrotta, 2013). PRC1 is the bigger complex and very large in size as compared to the other. It catalyses mono-ubiquitination of the lysine 119 on the histone H2A (H2aK119Ubq) (Wang et al., 2004) and this enzymatic activity is carried out by two main proteins of the complex Ring1a and Ring1b. Ring1a/b ubiquitin-ligase activity is highly dependent on the presence of Pcgf2 and Pcgf4 (Cao et al., 2005; Elderkin et al., 2007). In recent years, it has been found that PRC1 complex exists in different forms in different

tissues with different functional roles (Gao et al., 2012). On the other hand PRC2 is smaller complex and it catalyses methylation on Lysine 27 of Histone H3. This activity of methylation of PRC2 is carried out by Ezh2 and Ezh1 proteins of the complex (Margueron and Reinberg, 2011). Ezh2/1 methyltransferase activity is fully dependent on other proteins of the complex namely Suz12 and Eed (Cao et al., 2002; Cao and Zhang, 2004; Pasini et al., 2004a). Both PRC1/2 localize at genes promoter, which are involved in the process of differentiation and proliferation (Orlando and Paro, 1993; Simon and Kingston, 2013). In terms of their functional role, both these complexes are associated with transcriptional repression of target genes (Laugesen and Helin, 2014). Mechanism by which both complex regulate transcriptional repression can be explained by initial methylation on lysine 27 of histone H3 by PRC2 at target gene promoters which acts as a recruiting factor for PRC1 which then ubiquitinilates lysine 119 on the histone H2A this results in chromatin compaction and transcriptional repression (Cao et al., 2002). Recent studies have shown that the vice versa mechanism also exists, where PRC1 initially ubiquitinilates which act as recruiting factor PRC2 complex (Blackledge et al., 2014). Functional characterization of PRC1/2 has been studied in depth, but the mechanism by which they are recruited to chromatin is not fully understood in higher order organisms. In Dorsophila, it has been clearly shown that cis-regulatory PRE elements are responsible for recruiting polycomb (Kassis and Brown, 2013) to their target promoter but this does not hold true in mammalian system. Many mechanisms has been proposed like Rybp, Kdm2b and Jarid2 proteins are shown as recruiting factors for polycomb (He et al., 2013; Pasini et al., 2010; Tavares et al., 2012; Wu et al., 2013) and some other studies show long non coding RNA as the mediators for recruiting polycomb to their target gene promoters. Besides these evidences, clear understanding on PRC1/2 recruiting

mechanism to chromatin is still being debated. PcGs proteins are always studied in terms of differentiation, development, proliferation and their tumorigenesis (Piunti and Pasini, 2011; Sparmann and van Lohuizen, 2006). Their importance in cell viability is shown through knock out studies of different components of the complex in mouse embryonic stem cells, this resulted in embryonic lethality or developmental defects. In many tumours, it has been reported that PcGs overexpression is a negative prognostic factor. In such cases its inhibition is regarded as potential strategy for tumour treatment (Piunti and Pasini, 2011).

1.3.2. Polycomb Repressive Complex 2

PRC2 is one of the two complexes of PcGs and is regarded as key regulator of gene expression. It is associated with transcriptional repression of target genes, which are mainly required for differentiation and developmental process. Primary activity of PRC2 is to trimethylate Lys27 on histone H3 (H3K27me3) (Cao et al., 2002; Czermin et al., 2002; Kuzmichev et al., 2002) contributing to chromatin compaction, which ultimately results in transcriptional repression. H3K27me3 is deposited in genomic regions with high density of CpG nucleotides mainly comprising promoters. PRC2 is composed of four key proteins namely Enhancer of Zeste homolog (Ezh1/2), Suppressor of zeste (Suz12), Retinoblastoma protein associated protein 46/48 (RbAp46/48), and Embryonic ectoderm development (Eed). Individual core protein of PRC2 is fundamental for proper functioning of complex on histone substrate (Cao and Zhang, 2004; Ketel et al., 2005). This has been very well demonstrated by early embryonic lethality of mice deficient for Eed, Suz12 or Ezh2 (Faust et al., 1995; O'Carroll et al., 2001; Pasini et al., 2004b). This observation is consistent with the functional activity of PRC2 in repressing genes that are involved in lineage

specification (Boyer et al., 2006; Bracken et al., 2006; Lee et al., 2006; Mohn et al., 2008). Out of four core proteins, Ezh1/2 is the one, which carries out enzymatic activity of methylation. These proteins contain SET domain responsible for methyil-transferase activity. Eed protein other component of the complex has been shown to have the ability to recognize and bind to H3K27me3 modification and enhance lysine methyl transferase (KTM) activity by the complex there by establishing a positive feedback loop (Margueron et al., 2009). It was shown that in presence of H3K27me3 peptide, KTM activity is boosted up by several folds. There by it can be concluded that, PRC2 is more efficient in trimethylating K27 in presence of pre-existing H3K27me3. This suggests a way to maintain some minimal levels of K27me3 through cell cycle progression (Hansen et al., 2008; Margueron et al., 2009). EZH1/2 has also the ability to methylate K27 residue on histone H3 in a stepwise manner from mono to tri form (H3K27me1/2/3), with different functionalities. It would be interesting to know how these three different modifications localize throughout the genome and understand functional role of each individual form of methylation. My work is focused in this area and we recently published our findings (Ferrari et al., 2014), where we showed that H3K27me1 localize in the actively expressing gene bodies and correlate well with H3K36me3, where as H3K27me2 is diffused through the genome with an preventive aim to not open non cell type specific enhancers.

1.3.3. Polycomb Repressive Complex 1

PRC1 is other complex of PcGs whose main enzymatic activity is to ubiquitinilate lysine 119 on the histone H2A. In mammalian cells, PRC1 exists in many forms each with its own functional importance (Gao et al., 2012; Vandamme et al., 2011). One common thing among all variants of PRC1 is that they all contain Ring1b/a as core

protein, with the enzymatic activity to ubiquitinilate lysine 119 on the histone H2A (Wang et al., 2004). All variants of PRC1 sub-complexes can be defined by the presence of one of the six Pcgf proteins mutually exclusive manner along with other proteins like Phc and Cbx (Gao et al., 2012). Canonical PRC1 can be nomenclature as PRC1.2 or PRC1.4 depending on the presence of either Pcgf2 or Pcgf4 (Gao et al., 2012, 24217316), along with Phc1 and Cbx proteins, which were initially discovered in human (Levine et al., 2002). Other non-canonical forms of PRC1 are namely PRC1.1 also known as Ring2-KDM2B complex comprising of Pcgf1 along with KDM2B, BCOR and USP7; PRC1.6 also known as Ring2-L3MBTL2 complex comprising of Pcgf6 along with L3MBTL2, MGA and Cbx3; PRC1.3 and PRC1.5 variant comprising of either Pcgf3/5 along with FBRS and CSNK2A1 (Schwartz and Pirrotta, 2013). In functional terms Cbx protein in canonical PRC1 has the ability to recognize H3K27me3 modification deposited by PRC2 and acts as recruiting factor for PRC1 on to the chromatin. This dependency can be proved by the loss of Ring1b binding on common targets genes of both PRC1 and PRC2 upon depletion of PRC2 (Tavares et al., 2012). But this didn't affect global levels of H2AK119ubq suggesting a PRC2 independent PRC1 activity on chromatin (Tavares et al., 2012). All variants of PRC1 contain RYBP explaining its crucial role in PRC1 functions (Gao et al., 2012). Presence of RYBP and CBX in PRC1 is mutually exclusive (Gao et al., 2012; Tavares et al., 2012). PRC1 containing CBX and RYBP share common target genes (Gao et al., 2012; Morey et al., 2013) but they bind adjacent in regions (Gao et al., 2012). PRC1 variants PRC1.2 and PRC1.4 were unable to ubiquintilate H2AK119 when forced to recruit to chromatin (Blackledge et al., 2014), suggesting pre-deposition of H3K27me3 might be a requirement for their activity. In recent findings it was observed that Kdm2b component of PRC1.1 variant is important for recruiting PRC1 onto chromatin

(Barrero and Izpisua Belmonte, 2013). Variant PRC1.6 might have some role in cell identity (Gao et al., 2012; Trojer et al., 2011). Depletion of L3mbtl2 and Wdr5 proteins from the complex has lead to mESC premature differentiation (Ang et al., 2011; Qin et al., 2012). Other variants, PRC1.3 and PRC1.5 are not fully characterized and functional role of their protein components are undetermined. But it is known on forced recruitment of Pcgf3 and Pcgf5 onto chromatin leads to monoubiquitination of H2AK119 and recruitment of PRC2 (Blackledge et al., 2014).

1.3.4. Role of Polycomb in stem cells and cellular differentiation.

Stem cells are of unique kind, which have capabilities of self-renewal and differentiation. Balance between these two states is regulated by means of integrated signaling and strong transcriptional regulation. A single genome can give multiple cellular identities through its robust mechanism of reorganization and modification at different levels; out of many underlying mechanisms that govern in stem cell identity and cell fate determination, epigenetic modifications driven by Tritorax and Polycomb proteins having greater importance. They establish patterns of gene silencing which are heritable and are preserved in a cell type specific manner.

Evidences have shown that Polycomb proteins can mediate gene repression through different mechanisms. Two main mechanisms by which PcG-mediate gene repression is: through chromatin compaction and impairment of transcription machinery. Chromatin compaction ability by PRC1 was described in *Drosophila*, which resulted in repression of Posterior Sex Combs region. Polycomb induced chromatin compaction makes itself inaccessible to modellers like SWI/SNI complex and other transcription factors, which would otherwise have triggered transcription (Francis et al., 2001).

Moreover, densely packed nucleosomes have been shown to stimulate PRC2 activity on H3K27, thus generating a positive feedback loop (Yuan et al., 2012).

Another mechanism by polycomb induce repression is through inhibition of transcription machinery. In *Drosophila* it was shown that polycomb binding does not block binding of RNA Pol II to their target promoters (Breiling et al., 2001). Another study through genome wide approach showed that bivalent promoters posses reduced levels of Pol II occupancy (Min et al., 2011). Promoters harbouring polycomb have RNA Pol II in paused form. On removal of polyomb from target promoters results in switch from paused to elongating form of Pol II, this was demonstrated by deletion of ligases in Ring1a and Ring1b causes the switch (Brookes et al., 2012; Stock et al., 2007). This suggests that PcG occupancy at bivalent promoters is able to block Pol II at their transcription start sites.

PcG proteins play very important role in reprogramming terminally differentiated cells into induced pluripotent stem cells (iPSCs). This process involves re-organization of chromatin state exerted by different factors. Through different approaches, it was shown that both PRC1 and PRC2 complexes are essential for reprogramming of human fibroblast (Onder et al., 2012) and B cells (Pereira et al., 2010). It has been found that on induction of pluripotency in mouse embryonic fibroblast (MEF) does not require Ezh2 activity, this might be due to Ezh1 compensation, and levels of H3K27me3 are decreased at developmental genes on co-deletion of Eed with impaired reprogramming (Fragola et al., 2013).

Chapter 2 - RESULTS

2.1. ChIP_QC

ChIP_QC, is a standalone application which can be used either through Graphical User Interface (GUI) or through command line. Figure 2.1 represents GUI of ChIP_QC. It has 15 different modules each with specific analysis of interest. Navigation from one module to another can be done by clicking on the tabs provided on the top menu. To highlight different features of ChIP_QC and demonstrate its capabilities, we took advantage of processed ChIP-seq and RNA-seq data generated by the ENCODE consortium (Consortium, 2012) from different human cell lines to postulate different biological scenarios and analysed obtained results.

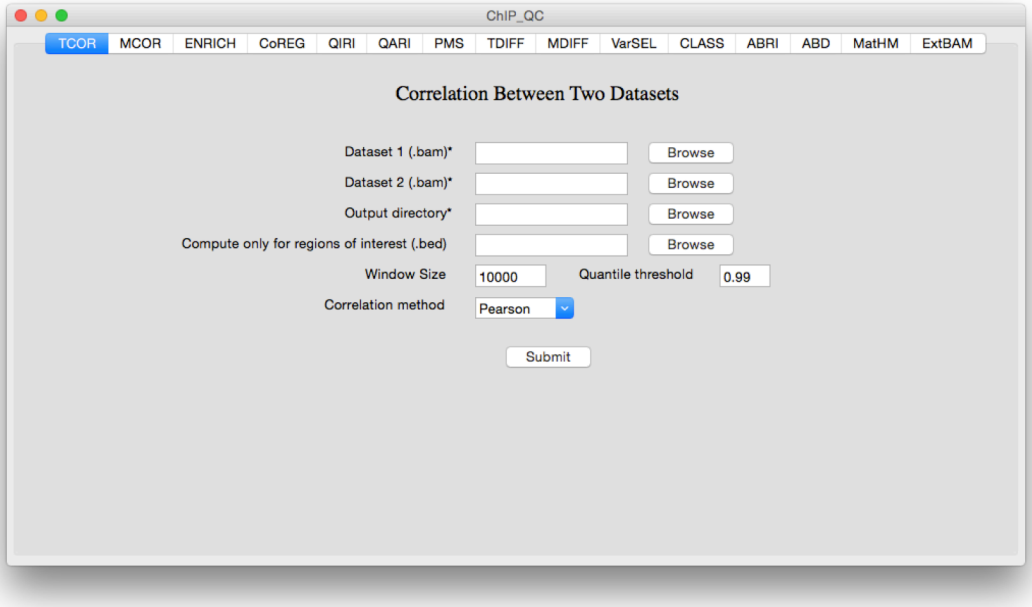


Fig. 2.1 ChIP_QC GUI This tool is composed of 15 different modules with each module designed with specific analysis of interest. Different modules can be navigated from top menu bar.

2.1.1. Enrichment and coexistence

The common question behind most epigenomic analysis is the requirement to determine whether our ROI show any preferential enrichment towards any known set of annotated regions. In such situations, this section of the program called ENRICH becomes useful. For instance, we were interested in determining if a set of different factors, for which we have obtained ChIP-seq location data, can bind preferentially active promoter or enhancer elements in human embryonic cells (H1hESC). The genomic location of active promoters or enhancers can be easily determined by the accumulation of H3K27 acetylation (H3K27ac) respect to a mapped TSS. Thus, we took into consideration H3K27ac enriched regions in H1hESC and separated these regions into two broad categories: 1) regions residing in close proximity of promoters (± 2.5kb from TSS) and 2) regions lying away from promoters. This analysis identified bona fide active promoters (n=4600) and enhancers (n=2033) in H1hESC. These two sets of regions were used to determine the levels of association of 49 different factors for which ChIP-seq results were generated by the ENCODE consortium (Consortium, 2012) in H1hESC cells.

The results of the analysis are shown in Fig 2.1.1A which represents the proportion of ROI being bound by each individual factor. Apart from analysis with input files, this program provides option to introduce random regions in the analysis (figure 2.1.1A, blue bars). Program generates random regions by shuffling coordinates of input set of regions. Introducing random regions allows determining the extent of significance of this comparative analysis respect to random occurrence. This analysis clearly shows that these active regions are all completely devoid of repressive factors such as Ezh2 and Suz12. Importantly, it can be noted that the transcription machinery such as

RNA-POLII, TBP, TAF1 are present in both active promoters and enhancers, while factors like Gabp, Brca1, Nrf1, Six5, Sp2, Cmyc and Gtf2f1 are specifically enriched in active promoters. Differently, transcription factors (TFs) like Oct4 (Pou5f1) and Nanog resulted preferentially enriched at active enhancers as previously reported (Chen et al., 2008; Whyte et al., 2013). This analysis clearly showed that a peculiar set of factors help in defining enhancers from promoters. Interestingly, this unbiased analysis allowed the identification of Bcl11a and Tcf12 as novel factors specifically associated at active enhancer regions.

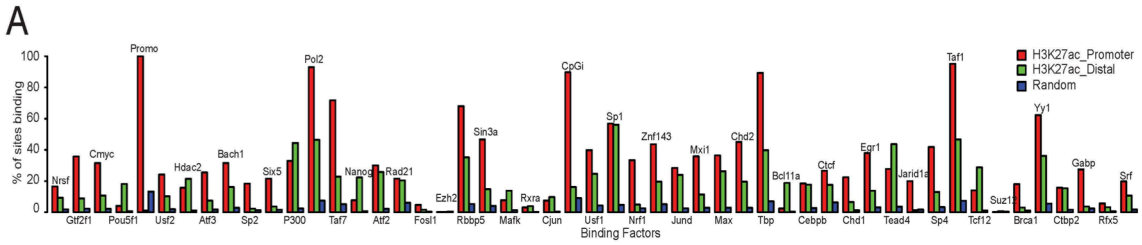


Fig. 2.1.1 A. Enrichment (A) Barplot representing proportion of H3K27ac positive promoters (in red), H3K27ac positive enhancers (in green) and random regions (in blue) bound by different factors.

Using CoREG, we further investigated whether the factors that are specifically enriched at enhancers co-exist or not. This tool helps in dissecting the extent of co-regulation between different factors based on absence or presence of factor in each ROI. Taking into consideration all Bcl11a enriched regions as reference, we found that Bcl11a frequently co-localized with the enhancer specific TFs Nanog, Pou5f1, Tead4, Tcf12 as well as with more promiscuous factors such as P300 and Sp1 (figure 2.1.1B). When the same analysis was performed using a set of promoters corresponding to the top 3000 highest expressed genes in H1hESC, this set of factors was not enriched (figure 2.1.1C). Hence, this analysis strongly suggest that the novel enhancer associated factors Bcl11a and Tcf12 co-exsist at *cis*-regulatory regions together with

Nanog, Tead4 and Pou5f1, highlighting the power of our new analytical tools. This observation can further be taken into consideration for experimental studies to see for any co-regulation between Bcl11a and other enhancer specific transcription factors.

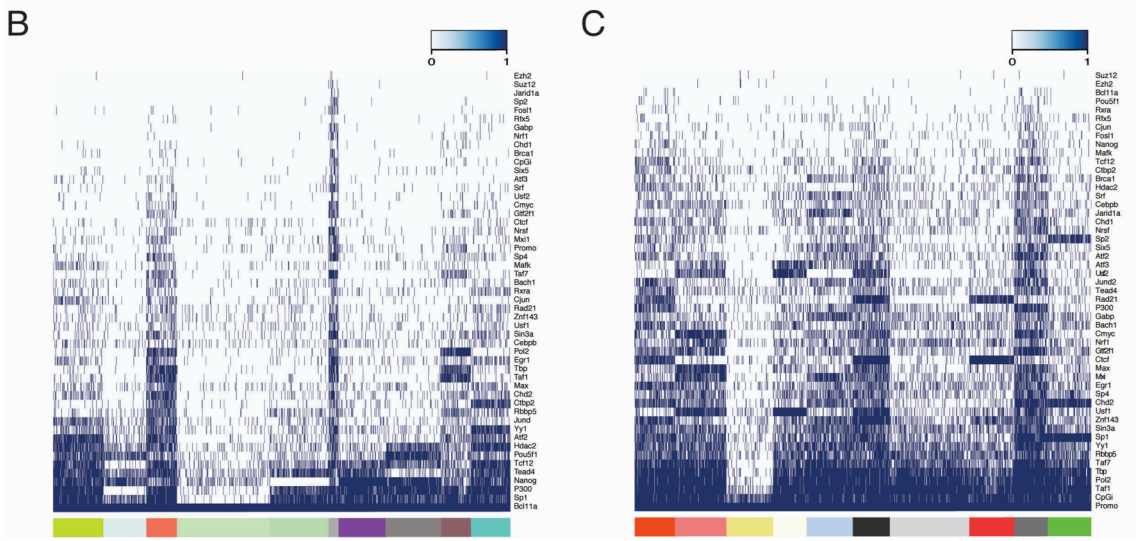


Fig. 2.1.1 B-C. Coexistence (B) Heatmap showing presence (dark blue) or absence (light blue) of different factors in Bcl11a regions. Closer the presence of any factor to Bcl11a greater the co-presence. Colorbar at the bottom represents different clusters generated by kmeans clustering, where k=10. (C) Heatmap showing presence (dark blue) or absence (light blue) of different factors in promoters of first 3000 highly expressed genes. Closer the presence of any factor to promoters greater the co-presence. Colorbar at the bottom represents different clusters generated by kmeans clustering, where k=10.

2.1.2. Quantification and Correlation

Most genome wide location studies generate multiple large ChIP-seq datasets, for which a major task is determining the extent of correlation among multiple datasets to identify closely related datasets clustering together to determine convergent or divergent biological behaviours. This type of analysis is facilitated with the section of the program called MCOR, which can take multiple datasets and perform correlation at a genome wide level or along specific ROIs. To illustrate this tool, we have scanned the behaviour of 27 different factors from H1hESCs respect to all human promoters.

We subjected the datasets to two independent correlation methods: pearson correlation (figure 2.1.2A) and PCA (figure 2.1.2B). The results were consistent with both types of analysis. Mainly, they have identified two type of clusters: one representing a repressive cluster marked by a strong correlation between Polycomb proteins (Suz12 and Ezh2) and their related histone post-translational modifications (PTMs; H3K27me3); and a second cluster represented by factors and histone PTMs associated with active transcription (H3K27ac, H3K9ac H3K4me3, Pol2). With respect to Pearson correlation, PCA analysis provided much more extended information. First, the angle of separation allows depicting the lack of any relationship between datasets representing active vs. repressive features. Second, the profiles of H3K9me3 deposition strongly diverge form all other datasets consistent with its well-established deposition in constitutive heterochromatin. Third, the arrow length for each dataset provides information related to the contribution of each factor. For instance, the limited length of H2AZ, Ctcf and Jarid1a highlights its minimal

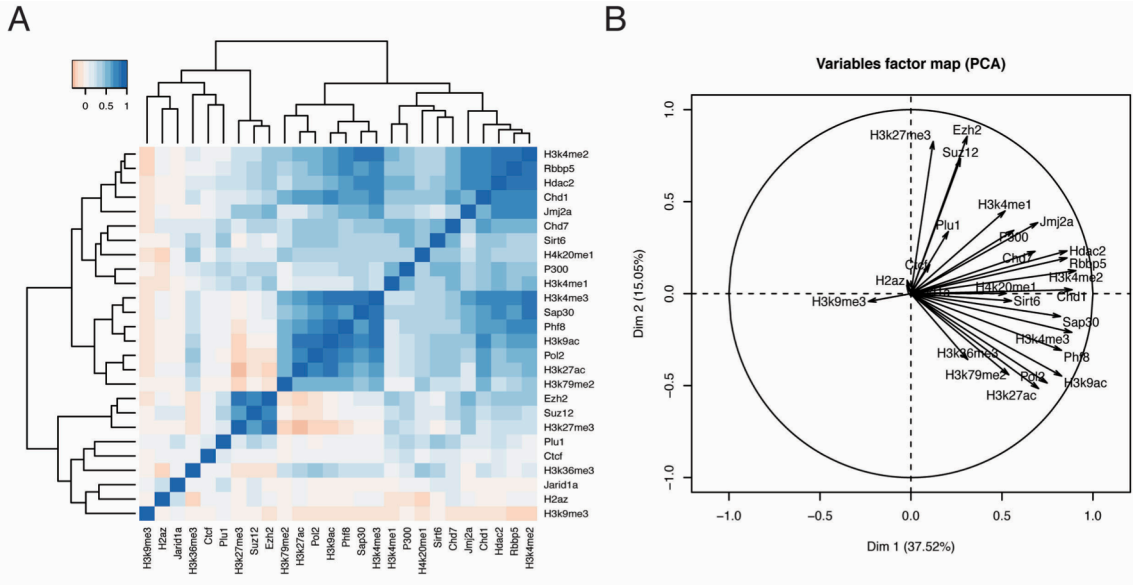


Fig. 2.1.2 A,B. Correlation (A) Genome wide correlation between different factors along all promoters of human genome. (B) Variable plot with different factors and their degree of correlation with others along all promoters of human genome across first two principal components.

role in defining promoter elements.

2.1.3. Comparative quantification and its effects

A great challenge of ChIP-seq analysis is to move from qualitative information not just knowing the presence or absence of different HMs/TFs to quantitative information, where one can study how different levels of modification can affect the biological system. These quantitative changes can further correlated with transcription levels helping us in better understanding the underlying mechanism of transcription. This implies more complex computation and to take into consideration intrinsic biases of the sequencing procedure. To allow capturing of such changes we designed quantitative methods that can identify such changes among multiple datasets and relate them with expression information (when provided). To exemplify our tool, we portrayed different scenarios to show how different ways of quantification can be experimentally meaningful.

2.1.3.1. Quantification within ROI

We took two samples of H1hESC, one representing a set of H3K27ac enriched regions (active transcription; n= 6633) and the other represents a set of H3K27me3 enriched regions (repressed transcription; n= 5406) to quantify the extent of deposition of other histone PTMs between these two functionally different sets of genomic region. For this, we provided the program with ChIP-seq datasets, comprising 10 different histone PTMs together with RNA polymerase II and complemented with H1hESC gene expression data for all genes with their respective FPKM values in log2 form. The program processes the data, computes the levels of RNA polymerase II and other different PTMs within the ROI and present results in the form, which can be visualized as heatmap (figure 2.1.3.1A). Two samples are separated by red line; the

upper panel represents the H3K27ac cluster and the lower panel represents the H3K27me3 enriched regions. Each row in heatmap is one ROI. For uncovering specific patterns within each cluster, the data can be subjected to either hierarchical or k-means clustering. For the presented analysis, the data from each sample was subjected to k-means with nine clusters. The clustered data were further explored for specific expression patterns. The program associates each ROI to the closest gene and represents the distribution of expression of all genes associated within each cluster as boxplots (figure 2.1.3.1B) allowing easy visual comparisons of results. This analysis clearly showed that in general all H3K27ac target genes presents a higher level of expression as compared to H3K27me3 target genes consistent with their respective roles in activating and repressing transcription. Within H3K27ac enriched regions, clusters 5/7 identified active enhancers marked by the presence of H3K27ac, H3K4me1 and by the absence of H3K4me3 deposition.

The other clusters identified active promoters, which are marked by presence of H3K27ac and high levels of H3K4me3. Interestingly, the closest genes to cluster 7 enhancers, which contain higher levels of H4K20me1, H3K79me2 and RNA polymerase II displayed higher level of expression respect to cluster 5 enhancers which contains lower level of deposition for these modifications. Finally, this analysis strikingly identified set of H3K27ac enriched genomic regions cluster 9 devoid of other histone marks suggesting a different regulatory function. The results related to H3K27me3 enriched regions identified clusters 2/3/5/7 to represent bivalent domains (Ku et al., 2008), which are marked by the presence of both H3K27me3 and H3K4me3. Within these clusters, cluster 2 present high levels of RNA-PolII association respect to clusters 3/5/7 that reflect in a higher level of expression of the

associated genes showing that deposition of repressive marks is not sufficient to exclude transcription highlighting the requirement of to quantify in parallel multiple type of datasets to stratify the functional status of transcriptional regulatory regions. In agreement with this, is important to note that the small set of genes linked to cluster 6, which are marked by co-deposition of both H3K27me3 and H3K9me3, basically undetectable transcription, showing that acquisition of H3K9me3 locks H3K27me3 repressed genes in a transcriptionally non permissive status.

Quantification can be based on a genome wide approach or it can be specifically applied to ROIs. While the genome wide approaches normalize, quantify and scales the ChIP-seq signals along the entire genome where as the ROI-selection performs the same quantification taking only the ROIs genomic regions into account. In genome wide approach, quantification is processed in small bins, then the bins representing each ROI are merged and the mean signal is reported. If the analysis is restricted to a set of ROI, the quantification and scaling will be specifically applied to this frame. It is important to note that for capturing true intensities, it is always advisable to perform genome wide analysis since it is possible that by quantifying signals respect to a restricted set of genomic regions (ROI option) the intensities could result over or under represented respect to a quantification that takes into account the entire range of signals along the genome for a specific ChIP analysis. This effect can be appreciated in Figure 2.1.3.1C,D were ROI-specific normalization result in the overrepresentation of specific signals (i.e. H3K9me3) or the under-representation of others (i.e. H3K9ac or RNA-PolII).

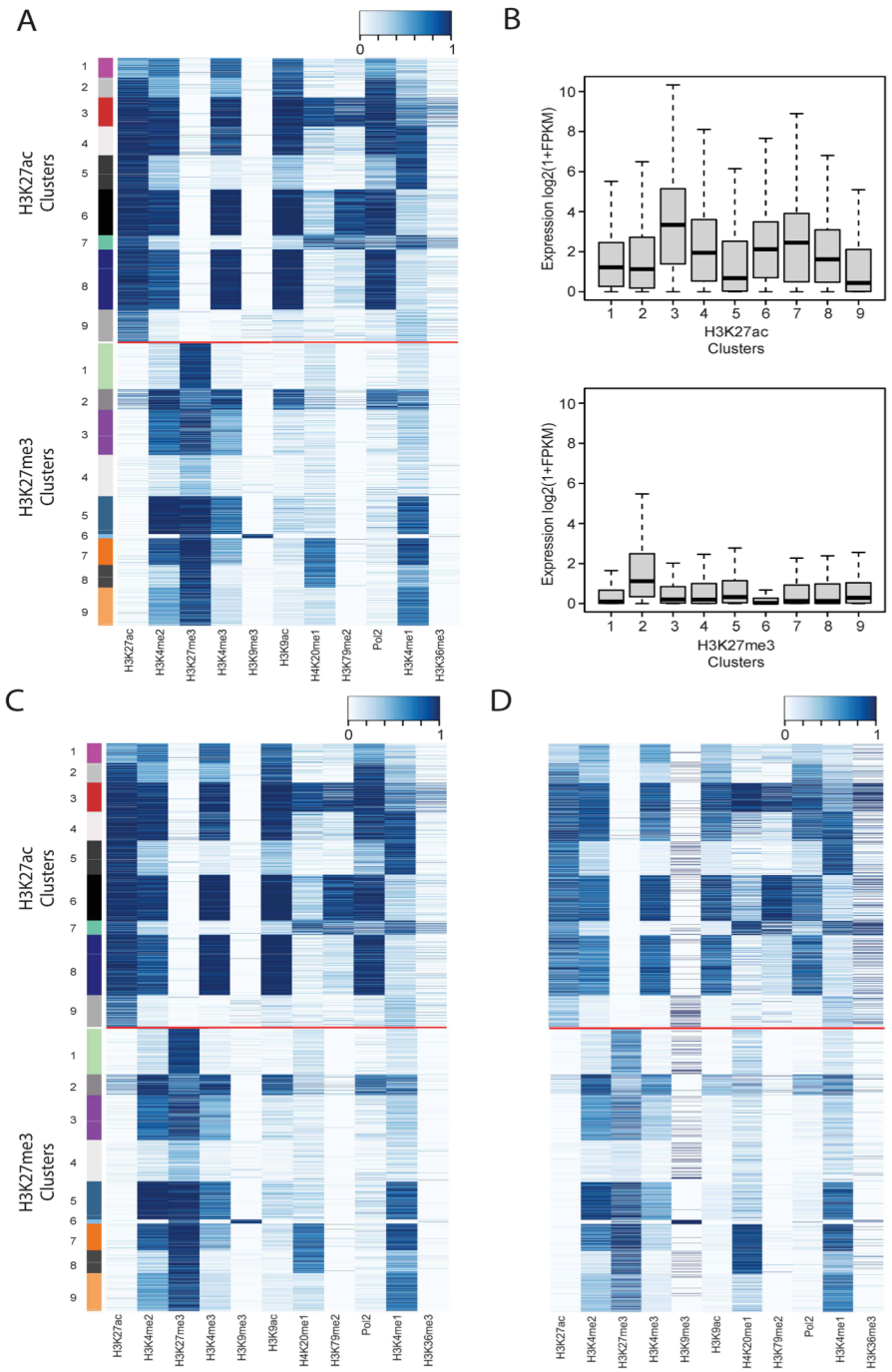


Fig. 2.1.3.1 Quantification within ROI (A) Heatmap with genome wide based normalized intensities for different histone modifications and RNA polII in H3K27ac and H3K27me3 binding regions separated by red line. (B) Expression level of target genes in each cluster as identified in A. Top panel represents expression levels for target gene clusters for H3K27ac regions where as lower panel represents for H3K27me3 positive regions. (C) Same as A (D) Same as A, but quantification, normalization and scaling are restricted only to ROI.

2.1.3.2. Quantification around ROI

To extract more information, signal quantification at specific loci can be analysed in relation to its surrounding genomic region to determine the extent of spreading of the signal respect to each ROI. To elucidate this option, we took into consideration active enhancers from four different tissues: lymphblastoid (Gm12878), leukemia (K562), liver carcinoma (Hepg2), cervical carcinoma (HeLa-S3) and determined the spreading of the H3K27ac signal over a 10kb region. All enhancers from individual tissues were merged together and submitted to the program along with gene expression data specific to Gm12878. The program extends to fixed length (default 5kb up and down stream) from the centre of each region and further segment these regions into small bins of 50bp length (set as default). Levels of H3K27ac from the four different tissues was quantified, scaled and subjected to K-means clustering. The results of this analysis can be visualized as heatmap (Figure 2.1.3.2A). This analysis clearly segregated tissue specific enhancers like cluster 1,6 highly specific to Gm12878, clusters 5/7 specific to K562, cluster 2/4 specific to Hela-S3, cluster 3 specific to Hepg2 plus a cluster (cluster 8) that seems to represent a small set of constitutive enhancers present in all four tissues (figure 2.1.3.2A). It's interesting to note within each tissue there are set of enhancers with higher levels of H3K27ac respect to others. In G12878, the levels of H3K27ac in cluster 6 are much higher than cluster 1. The same applies to K562 and Hela-S3. It is possible that clusters with highest levels of H3K27ac may represent super enhancer region as reported previously (Whyte et al., 2013). Consistent with this, when the results are related to expression of genes in Gm12878 it can be seen that (figure 2.1.3.2B) the expression of target genes associated with clusters 1/6/8 result significantly higher than other clusters

representing active enhancers in other tissues further supporting the active role in promoting tissue specific transcription for the identified enhancers.

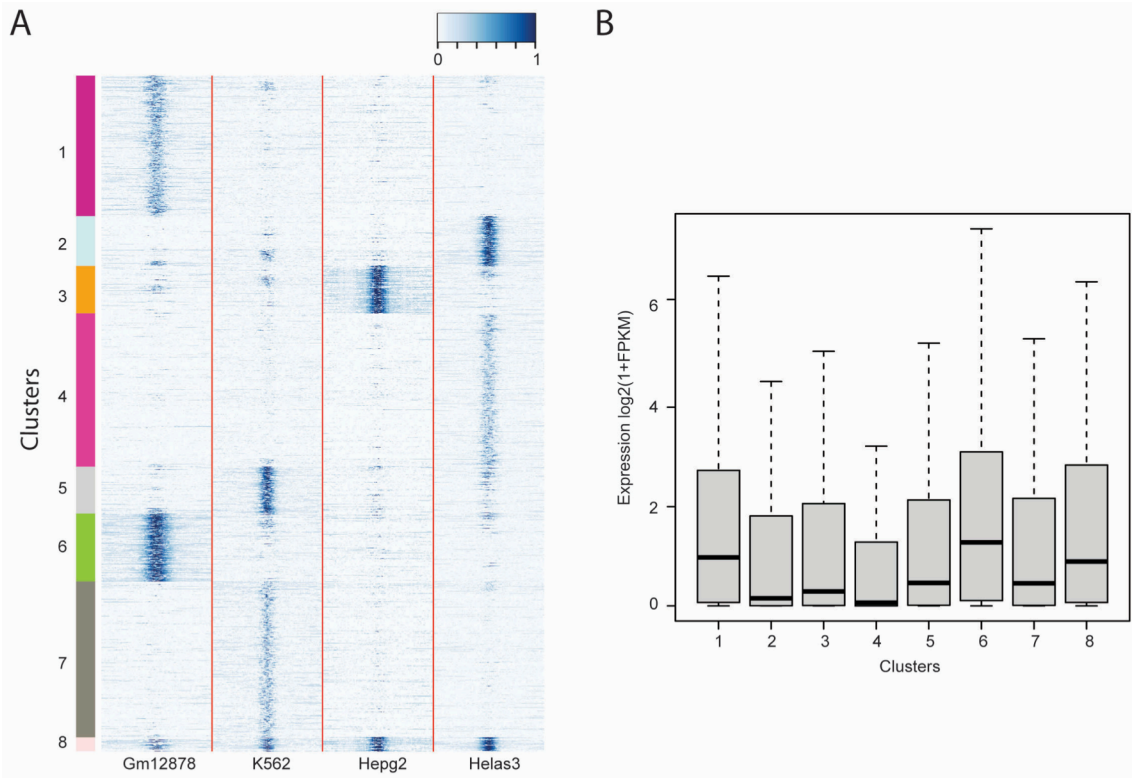


Fig. 2.1.3.2 A,B. Quantification (A) Intensities of H3K27ac ChIP around ±5kb region surrounding the center of enhancer regions across five different cell lines. (B) Expression levels of target genes in Gm12878 in clusters as identified in A.

Further advantage that this tool provides is the possibility to combine multiple experimental conditions scaling. This becomes particularly useful when the same factor or modification (i.e. the same antibody is used) is used in different experimental conditions. In such case, the program applies scaling globally over all datasets; else scaling is applied on individual dataset. The pros and cons of these approaches can be appreciated in Figure 2.1.3.2C,D. Figure 2.1.3.2C is same as 2.1.3.2A, where the programme by defaults assumes all datasets are handled independently irrespective of which antibody was used while in figure 2.1.3.2D this option is turned on and all datasets are scaled together.

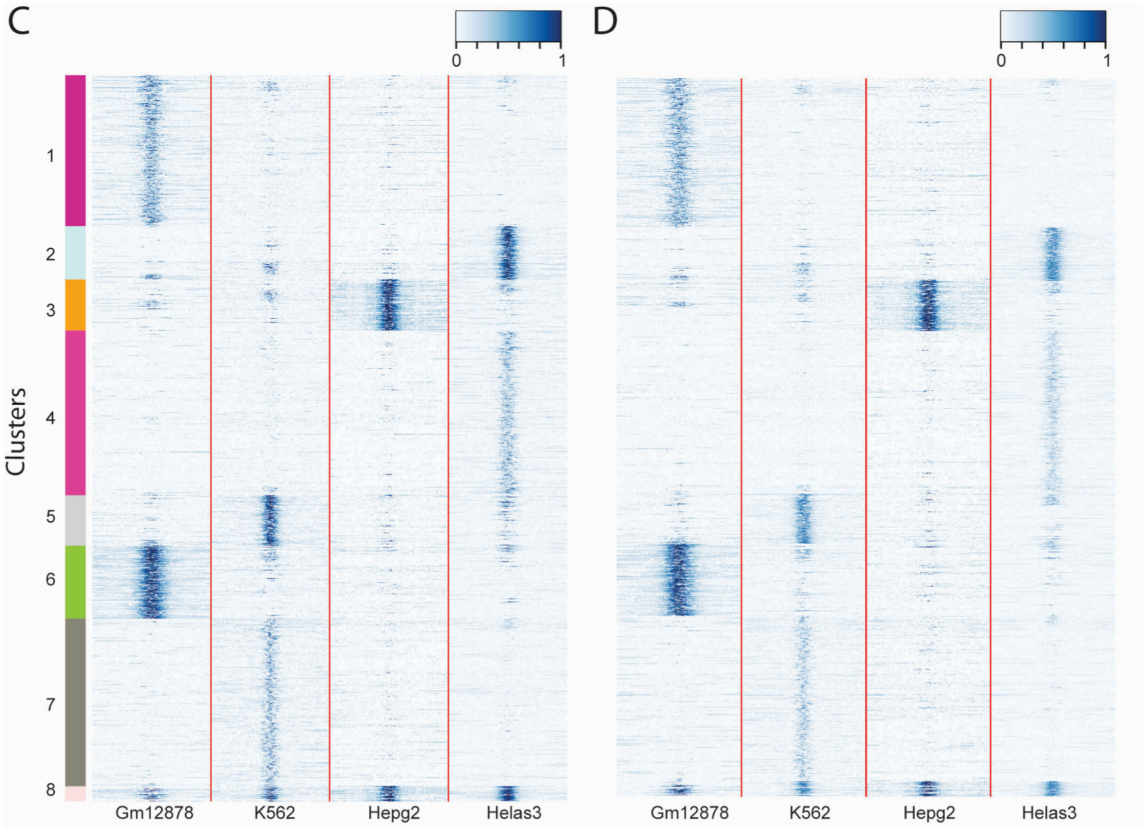


Fig. 2.1.3.2 C,D. Quantification (C) Intensities of H3K27ac ChIP around 5kb region surrounding from the center of enhancer regions across five different cell lines. (D) Same as A, where intensities are scaled globally over all samples.

2.1.3.3. Profiling

Previous section of program allows to quantify levels of different ChIP analysis over each and individual regions, we were also interested in determining the general genome wide behaviour of a specific factor/modification along different set of regions or experimental conditions. In such cases composite profiles of different ChIP's become simple and highly informative. In our program, we have supported such analysis with provision of quantifying target(s) over complete ROI with invariable length or over constant region surrounding from the centre. We took advantage of expression data from H1hESC and sorted the genes on the basis of their expression levels (high to low), which were further partitioned into quarters. The first quarter

represents highly expressed genes and lower quarter represents low or no expressed genes. We then quantified the levels of H3K4me3 and H3K36me3 in the promoters and gene bodies of each individual quarter respectively (figure 2.1.3.3A,B). In the case of H3K4me3, we quantified its levels for each individual gene within each quarter respect to the centre of promoters (i.e. centred on TSS and extended surrounding region by constant length of 5kb up and down stream). Quantification was measured over this constant region, where each region was further broken in smaller bins of 50bp in size. Similarly, we quantified H3K36me3 levels within gene bodies of individual gene. In this case, due to invariable gene length each gene is subdivided into certain finite blocks where each block represents a fixed proportion of the total length of each gene body. This data were averaged within each quarter and plotted together (figure 2.1.3.3A,B). It is well established that gene expression and both the levels of H3K4me3 deposition at promoters and the accumulation of H3K36me3 within gene bodies shows a positive correlation (Li et al., 2002b). Indeed, our analysis perfectly validated such behaviour with the genes belonging to quarter1 displayed higher levels of both H3K4me3 at promoters and H3K36me3 within gene bodies respect to the genes with lower expression levels (quarter 2/3/4; figure 2.1.3.3A,B). One of the advantages of this analysis is that the program makes use of strand information (when provided), which gives more sense to the data. For instance, from our results can be easily observed that the deposition of H3K4me3 preferentially occurs towards the +1 nucleosome, aiding proper positioning and active transcription. In same way, the levels of H3K36me3 are higher towards the gene terminal portion. The differences of not using strand information can seen in figure 2.1.3.3C,D.

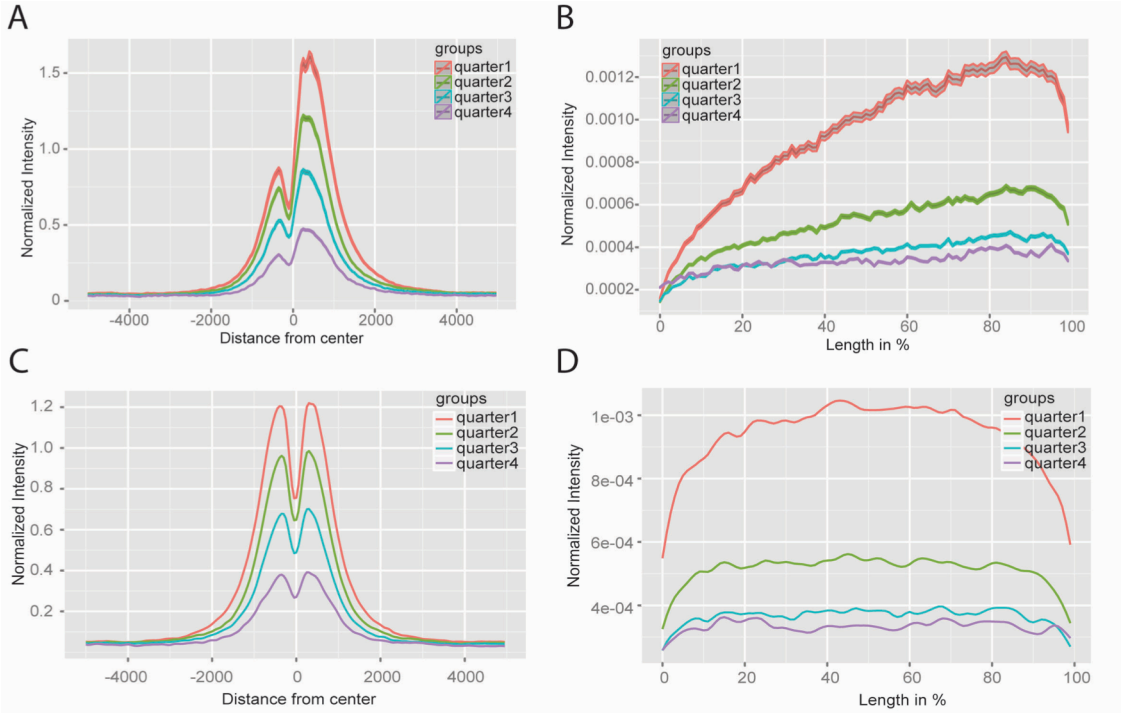


Fig. 2.1.3.3 A-D. Profiling (A) Average profile of H3K4me3 with confidence interval in promoters regions of genes classified based on expression levels (high to low). (B) Average profile of H3K36me3 with with confidence interval in gene bodies of genes classified based on expression levels (high to low). (C) Same as A, but without using strand information. (D) Same as B, but without using strand information.

2.1.3.4. Spike-In Normalization

In addition, this tool also supports quantification based on spike-in data. Recent reports have shown that the data generated through standard ChIP-Seq procedures are not able to capture the real changes in histone PTM deposition particularly when the overall global levels of a specific modification change between experimental conditions due to a flattening of the signal by the ChIP-seq procedure (Orlando et al., 2014). To circumvent this technical problem, a standard ChIP-Seq process can be combined with a spike-in chromatin from other reference genomes. This new procedure is able to quantify the true levels of a target protein/PTM among different experimental conditions at each specific ROI. Our quantification tool also supports this type of analysis. To prove the power of this option, we analysed data generated for H3K79me2 where different amount of chromatin in which H3K79me2 is either

present or absent were mixed in different proportions (0%, 25%, 50%, 75%, 100%) to mimic a linear reduction in global H3K79me2 levels and analysed in the presence of an equal amount of reference Drosophila chromatin. (Orlando et al., 2014). This analysis clearly shows the lack of linearity of the ChIP-seq signals among the samples when the spike-in normalization is not used in the analysis (figure 2.1.3.4A-D)

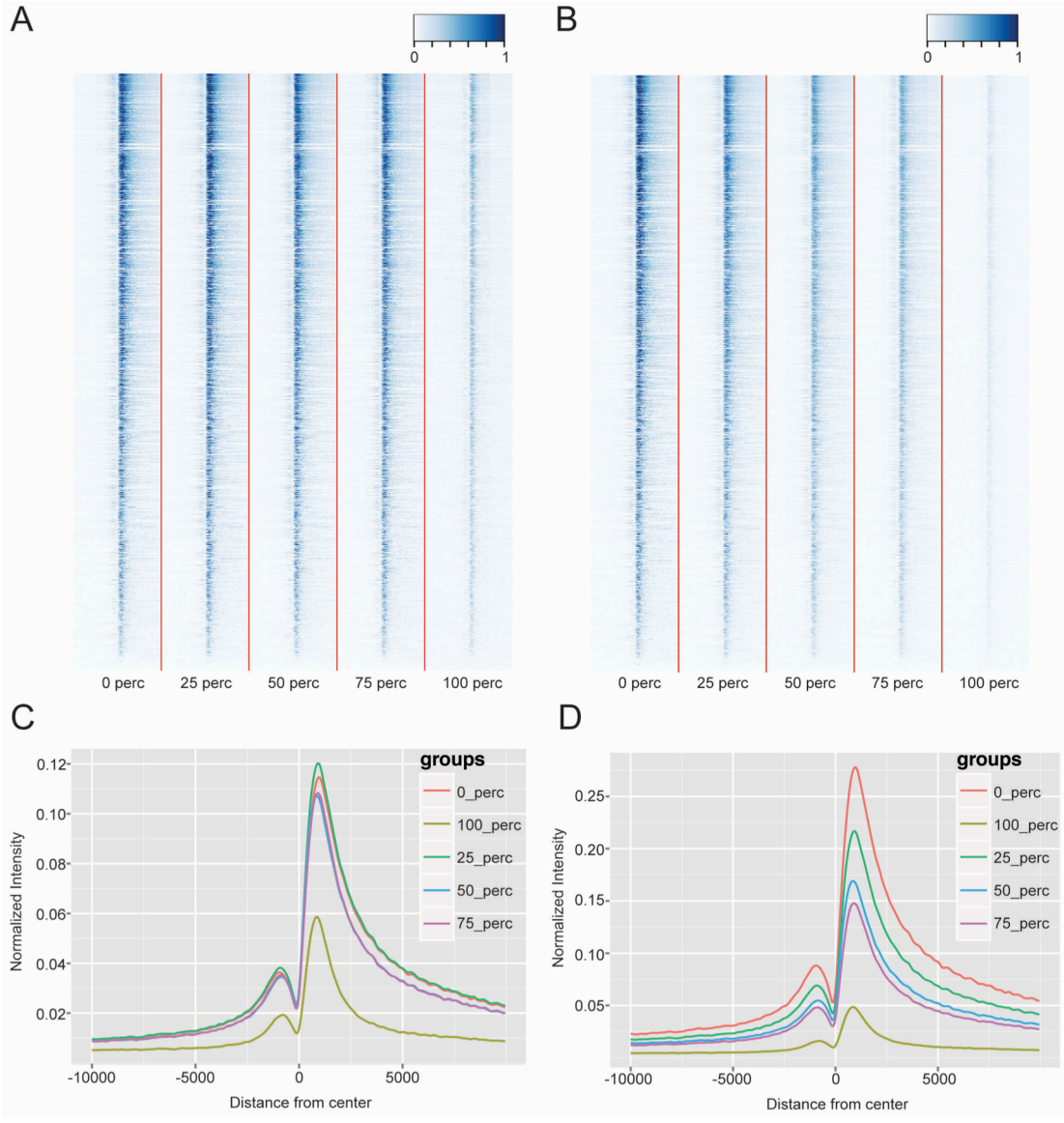


Fig. 2.1.3.4 A-D. Spike-In Quantification (A) Normalized intensities of H3K79me2 around 10kb surrounding TSS (both up and downstream) in regions possessing H3K79me2 in WT samples and its fate in other samples induced with different levels of inhibitor harboring no reference genome. (B) Same as A, in these intensities are spike-in normalized intensities. (C) Average normalized profile of H3K79me2 around 10kb surrounding TSS (both up and downstream) in regions possessing H3K79me2 in WT samples and its fate in other samples induced with different levels of inhibitor harboring no reference genome. (D) Same as C, in these intensities are spike-in normalized intensities.

2.1.4. Differential quantification

Quantification based differential studies can be helpful in identifying markers that allows differentiating two or more biological systems. For instance, we questioned whether the deposition pattern of the same histone PTM in two different tissues could be used to distinguish one tissue from the other. To test this, we have chosen ChIP-seq data for H3K4me3 (marker of active transcription in gene promoters) from SkeletalMuscle (Hsmm) and Keratinocytes (Nhek). We applied differential analysis over all promoters for H3K4me3 deposition. Using computed normalized read intensities the program identified a set of genes that were significantly enriched for H3K4me3 through fisher's test in SkeletalMuscle (n= 875) over Keratinocyte (n= 409) (figure 2.1.4A).

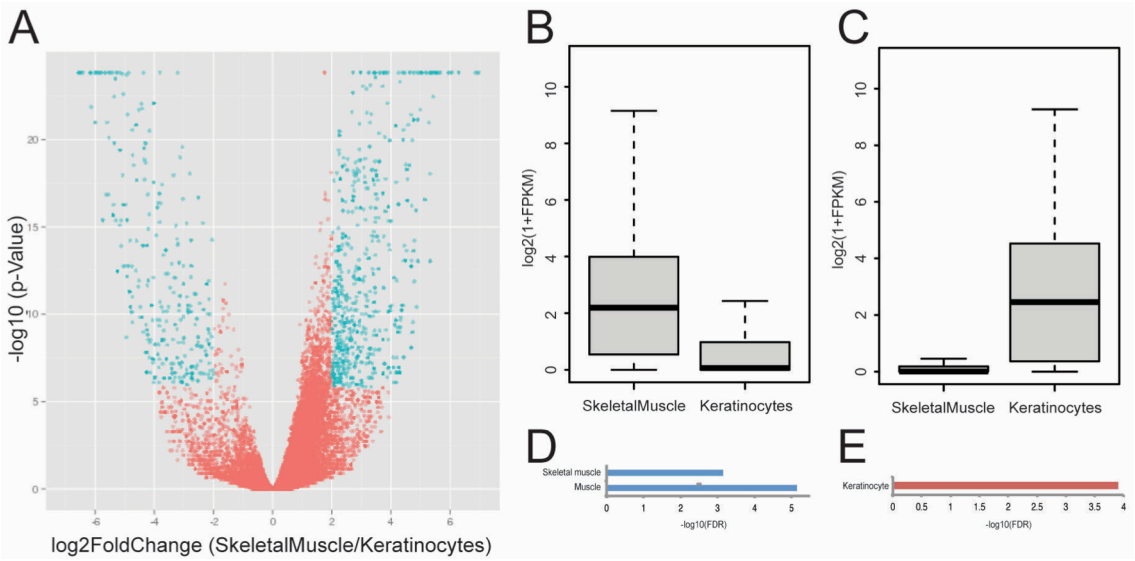


Fig. 2.1.4 A-E. Differentially regulated regions. (A) Volcano plot representing significantly enriched promoters (marked in cyan) harboring different levels of H3K4me3 methylation in skeletal muscle when compared keratinocytes. (B) Distribution of expression levels of genes where their promoters show significantly higher levels of H3K4me3 in skeletal muscle as compared to that of keratinocytes. (C) Distribution of expression levels of genes where their promoters show significantly higher levels of H3K4me3 in keratinocytes as compared to that of skeletal muscle. (D) Tissue specificity of genes whose promoters were differentially regulated skeletal muscle as identified in A. (E) Tissue specificity of genes whose promoters were differentially regulated keratinocytes as identified in A.

When we provided expression data for all genes in these two tissues, the program linked all enriched promoters to their respective target genes. This analysis clearly showed that the expression levels of H3K4me3 target genes in their respective tissue were significantly higher than others (figure 2.1.4B,C). This was further confirmed by performing tissue specificity with DAVID (Huang et al., 2009) with the output files containing the list of promoters significantly enriched in either SkeletalMuscle/Keratinocyte. Both lists showed higher specificity toward their respective tissue validating the tissue specificity of our results (Figure 2.1.4 D-E).

We come across situations were differentially enriched regions needed to be detected not only between two independent systems but also across multiple systems. This possibility is included in our tool and to show its functionality we extended biological logic of figure 2.1.4A over multiple tissues with the aim of identifying tissue specific markers. On the basis of normalized read intensities, the program identifies differentially enriched H3K4me3 promoters across all datasets through chosen statistical test (ANOVA or kruskal-wallis). All statistically significant regions are represented in the form of heatmap where normalized read intensities are transformed to standard z-score (figure 2.1.4F). To identify tissue specific patterns, the results were subjected to k-means clustering where k was set to 10. This analysis clearly showed that all tissue specific differentially enriched H3K4me3 promoters were clustered together (figure 2.1.4F). To further cross validate that these promoters are true markers of tissue specificity, the program also linked all promoters with the expression of their respective genes across all tissues within individual clusters, which resulted in the distribution of the expression of all target genes within individual clusters (figure 2.1.4G). Comparing the results side by side, we confirmed that genes associated with tissue specific promoter indeed displayed

tissue specific expression (figure 2.1.4F,G). For example, cluster1 represents the promoters that are specific for liver carcinoma (Hepg2), which indeed display greater expression levels respect to all other tissues. Similar conclusions can be applied for all other clusters.

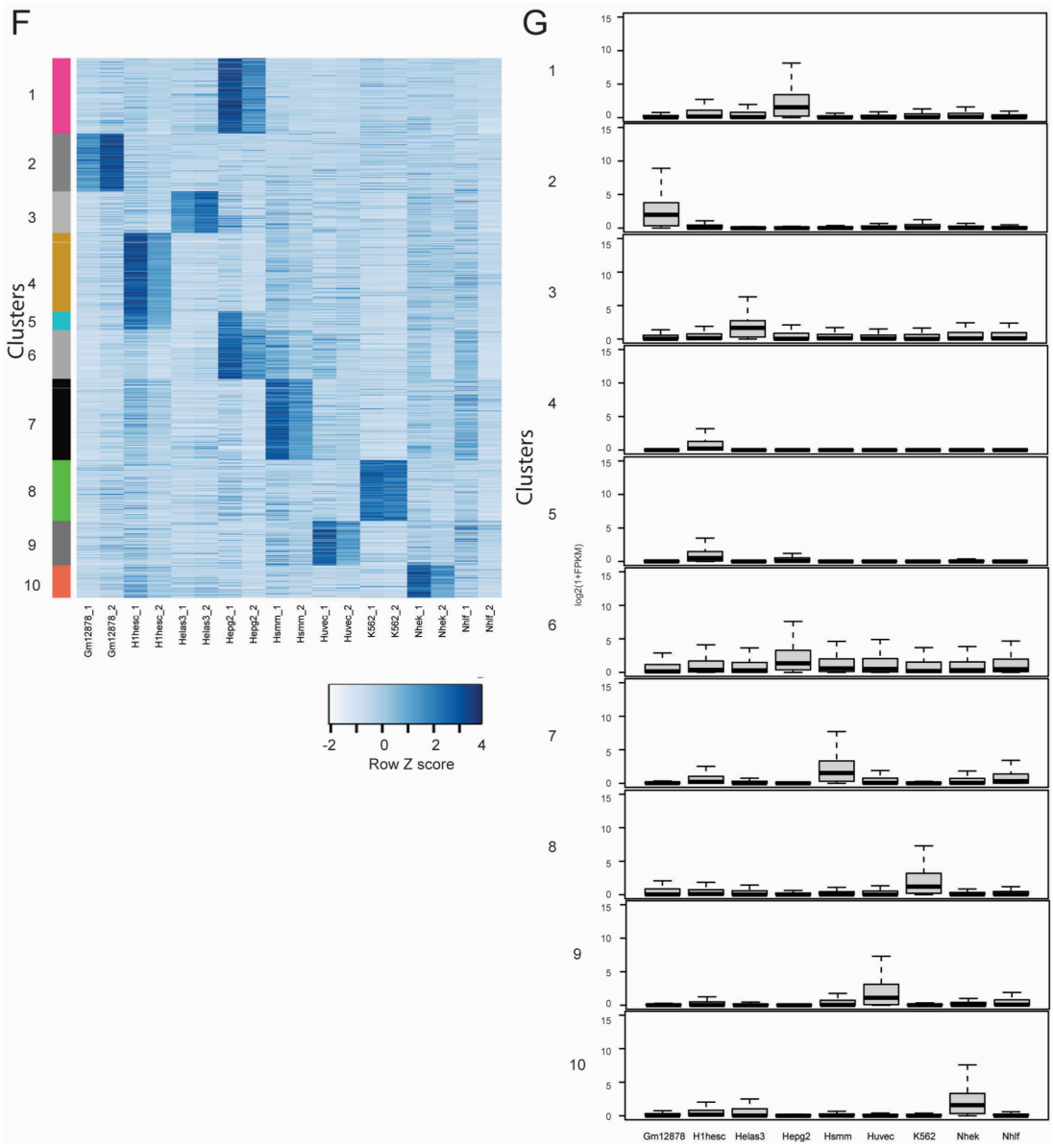


Fig. 2.1.4 F,G. Differentially regulated regions. (F) Significantly enriched promoters on the basis of K4me3 across 9 different cell lines. Represented here are their intensities in standard z-score form. (G) Expression level of target genes in each cluster across 9 different cell lines identified in F.

2.1.5. Probabilistic Relationships

The increasing availability as well as the capability of generating large set of genome-wide location analysis, implies the interest in exploring relationship between chromatin associated factors among a large number of datasets, with the aim to better dissect the role of each entity and its functional contribution in a given biological process. Thus, we designed a module that helps in predicting the probabilistic relation between different factors, either at a genome wide level or specifically within ROIs, taking advantage of a Bayesian Network approach. To introduce this analysis, we wanted to determine which factors localize at genomic regions of compact chromatin and, among these, which factors show dependency with each other. We selected regions enriched for Suz12 (n=4789, a component of the Polycomb repressive complex 2 (PRC2) as established marker of compact chromatin) and test where does this protein localize in respect to genomic features and which other DNA binding factors could contribute to such functionality. For this, we took the binding sites of 51 different DNA binding factors along with 2 sets of annotated genomic regions (CpGi, and gene promoters). When this data are provided to the program, they are processed by implementing a learning algorithm where either a constraint or scoring analysis can be performed depending on the users interest (see methods for further details). In our analysis, we used a constraint based grow shrink algorithm in an iterative bootstrap process where 70 percent of the total data were selected randomly and Bayesian network was constructed from this. This step is repeated 500 times and only the dependency factors that were identified in 95% of the networks were retained generating a final (figure 2.1.5A). We also wanted to further determine the validity of the generated network, for this, we repeated the above process using random regions of the same input size (n=4789) and generated a

"control" network (figure 2.1.5B). On comparing both the networks it can be clearly identified that the dependency between Ctcf and Rad21 was not related to Suz12 bound regions, while the rest of the dependencies resulted specific (figure 2.1.5A,B). Indeed, dependency between Suz12 and Ezh2 is well known from literature as they are components of polycomb and also its preferential localization Suz12 at CpG rich genomic regions at gene promoters, all these observation can be captured in this analysis and can be seen in (figure 2.1.5A). In addition, this analysis identified novel specific dependencies between Ezh2, Ctbp2 and Egr1 that were never reported. Thus, to test whether the dependency between Ezh2/Ctbp2 is valid, we used ChIP-seq profiles of these proteins and checked for their association among all Suz12 binding sites. Indeed, we found that Ctbp2 occupies nearly more than half of the Suz12 binding sites (figure 2.1.5C).

2.1.6. Classification

Recent studies have shown that different loci of the genome display precise epigenetic characteristics. In terms of regulation, classifying genomic regions on the basis of their epigenetic characteristics is helpful for classifying distinct roles. Here we present a system where different classes can be segregated on the basis of the quantification of different factors (which can be either TFs/histone PTMs). In this tool we support classification based on Support Vector Machine (SVM) approach. It employs both linear and non-linear models of classification. This system can be used either for training or as a combined training and prediction processes. In the training scheme, the program takes into consideration all provided datasets and lists out the performance as an area under ROC with accuracy scores. If the user is satisfied with the classification on the training data, the classification model can be further applied

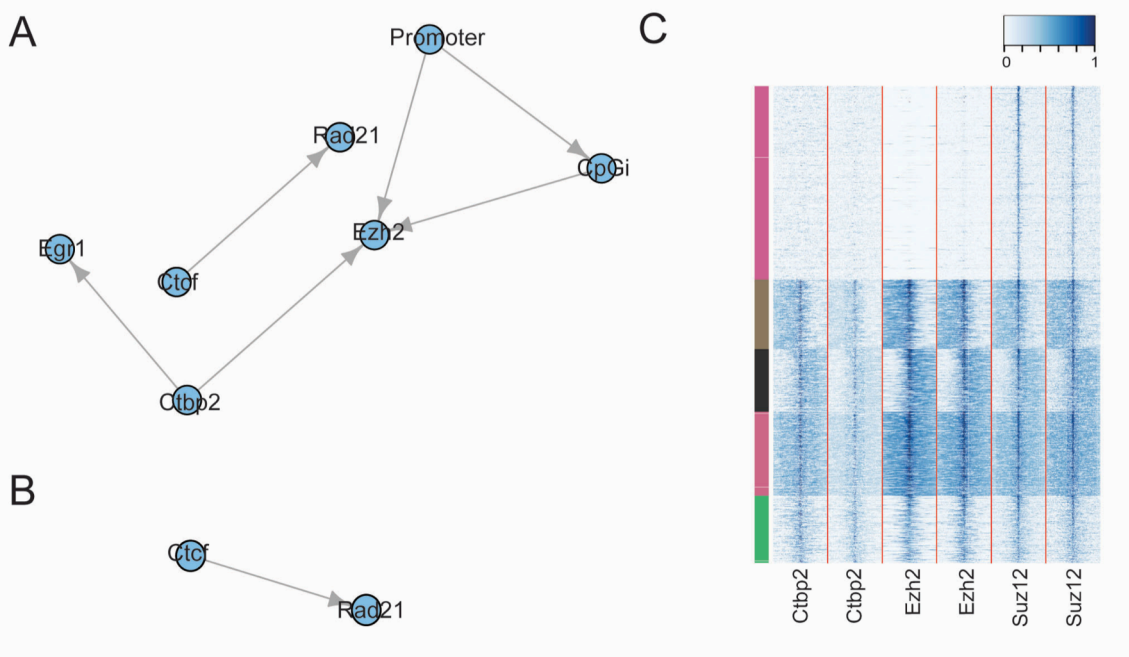


Fig. 2.1.5 A-C. Probabilistic relationships (A) Bayesian network showing dependency between different factors in compact chromatin regions of genome presided by Suz12. (B) Bayesian network showing dependency between different factors in random regions of genome. (C) Normalized intensities of Suz12, Ezh2 and Ctbp2 in Suz12 binding regions.

on new set of ROIs. In addition to the classification, program also supports pre data analysis (when required). Such approach is however not wise in the case of large number of data sets. The user has to judge which datasets contribute the most for the classification. In such situations user can choose pre-selection analysis, where all datasets are subjected to recursive feature selection process in which all-possible subsets are considered, and accuracies for all variable sizes are reported. The program also generates a list of predictors with the highest accuracies, which can be further used for the classification.

As example, we showed the process of characterizing active enhancers and active promoters marked by H3K27ac in ES cells on the basis of 42 different TFs. Initially all 42 TFs were subjected to a variable pre-selection process. This process helps out in eliminating the less contributing datasets. From the results (figure 2.1.6A), we observed that the combination of 18 datasets was sufficient for classifying active

enhancers and active promoters. Including additional datasets captured no major improvements.

Therefore, using a pre-selection analysis, we were able to down size the number of datasets for the further analysis, thus removing noise and reducing computing power. From the known literature, we can easily pick that all critical factors known to characterize active promoters and enhancers were selected (figure 2.1.1A). Interestingly, we also found that Bcl11a was one of the new selected datasets, which supports our initial findings (figure 2.1.1A) showing that Bcl11a is a TF specifically associated with enhancers. As final validation, when all these 18 datasets were fed into the classification process using SVM, we achieved an ROC of 0.96 (figure 2.1.6B).

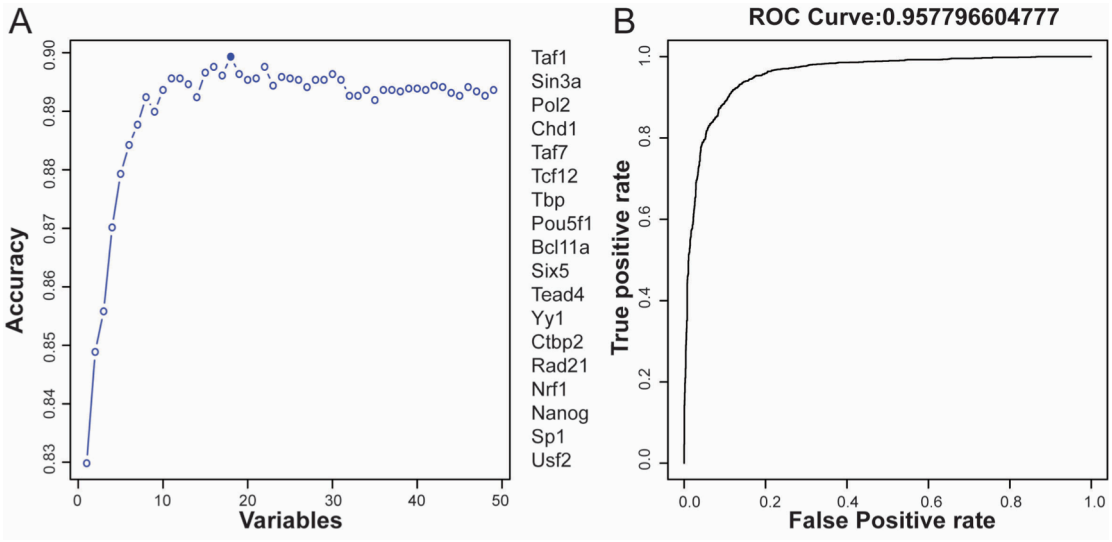


Fig. 2.1.6 A,B. Variable selection and classification. (A) Plot signifying the accuracy of different set of variables for characterizing active enhancers and promoters. (B) Sensitivity over specificity of SVM trained model for classifying active enhancers and promoters using variables with high accuracy level identified in A.

All above analysis demonstrates how the program can be helpful in many different ways for epigenomic studies. Apart from these main features our program is packed with additional add on features. These are:

1) tool for generating heatmaps from already generated results in different forms without rerunning complete analysis this saves time,

2) extending aligned reads to certain fixed length and

3) tool for checking correlation between replicates.

2.2. Polycomb dependend H3K27me1 and H3K27me2 regulate active transcription and enhancer fidelity.

2.2.1. PRC2 controls three different forms of methylation on H3K27

The initial aim of this study was to assess the distribution of different PTMs on Lysine 27 of histone H3 (H3K27) tails in mouse embryonic stem cells. For understanding this, we carried out a tandem Mass Spectrometry study (Jung et al., 2010), and this helped us in measuring relative abundance of PTMs of K27 in histone variants H3.3 and H3.2. From the analysis, we noticed that, different forms of methylation on H3K27 account for more than 80% of the total histone H3, H3K27 without any PTM accounted for on average 16% of H3 modifications, lastly H3K27ac account for approximately 2% of the total H3. Among different forms of H3K27 methylation, H3K27me2 was observed to be most dominant accounting for more than 70%, while H3K27me3 and H3K27me1 account for 7% and 4% of total H3, respectively (figure 2.2.1A). On comparing distributions of these modifications with other H3 isoforms, we saw that H3K27me3 is preferentially accumulated in H3.3 variant: this data confirms that H3.3 is found mostly regions of genes promoter, which are both silent and expressed (Goldberg et al., 2010). From this data we can conclude that in mESC the vast majority of histone H3K27 is post-translationally modified, among all modifications H3K27me2 is the most abundant form of modification.

In the scientific field, there was speculation about H3K27me1 modification. It was not known clear whether the modification is the by-product of de-methylation process or is the modification dependent on PRC2 catalytic activity (Margueron and Reinberg, 2011). To address this, Western Blot (WB) analysis were performed on histones extracted from mESC lysates, both wild type (WT) and knock out (KO) for different core members (Ezh, Eed and Suz12) of PRC2. For further cross validation, similar

analysis was conducted in mESC lysates acutely knocked down for the protein Eed by means of lentiviral transduction of sequence specific short hairpin RNA (shRNA). On doing so, we observed that levels of both di and tri-methylation were completely reduced, where as levels of monomethylation were significantly reduced (figure 2.2.1B), demonstrating that H3K27me1 and H3K27me2 depositions are dependent on PRC2 activity.

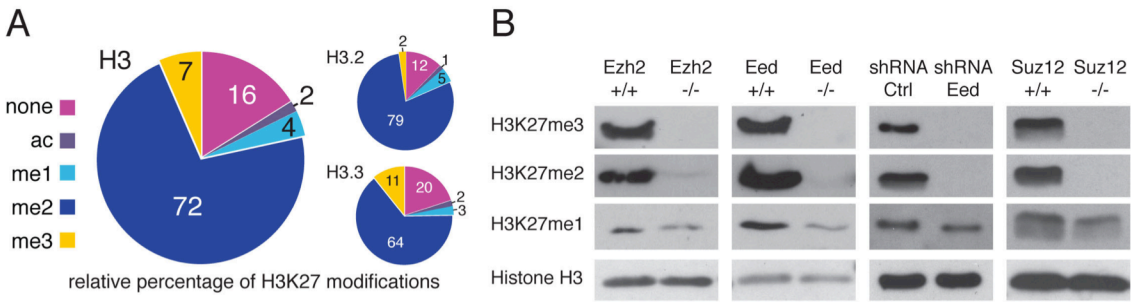


Fig.2.2.1 A,B. PTMs on H3K27 in mESC and its regulation by PRC2. (A) Larger pie graph show relative abundance of different PTMs on lysine 27 of Histone H3. Smaller pies show the same PTMs but in different Histone variants H3.2 and H3.3. (B) Western blot analysis showing loss of all forms of methylations on H3K27 using in and *Eed*, *Ezh2* and *Suz12* KO (-/-) as compared to that of indicated antibodies of protein extracts obtained from WT (+/+) mESC line. Similar trend was observed on knock down of Eed and Suz12 using shRNA in E14. Histone H3 served as loading control.

2.2.2. PRC2 dependent methylation states on H3K27 form distinct domains in genome.

After knowing relative abundance of three forms of methylation, as next step we were interested in understanding its genome-wide distribution. For this, we performed ChIP-seq for all three forms of methylation and Histone H3 using specific antibodies against it. Interestingly, from the analysis we saw that all three forms of methylation of H3K27 are deposited in genome in mutually exclusive manner (figure 2.2.2A). It can be noticed that, H3K27me2 is spread across both intergenic and intragenic regions, while H3K27me1 is preferentially enriched within gene bodies (figure 2.2.2A). Localization of H3K27me1 in genome drove us to compare the data with

H3K36me3, which is also known to accumulate in intragenic regions. For this, we made use of publically available H3K36me3 dataset (Mikkelsen et al., 2007) and saw that both H3K27me1 and H3K36me3 both co-localize in gene bodies (figure 2.2.2A).

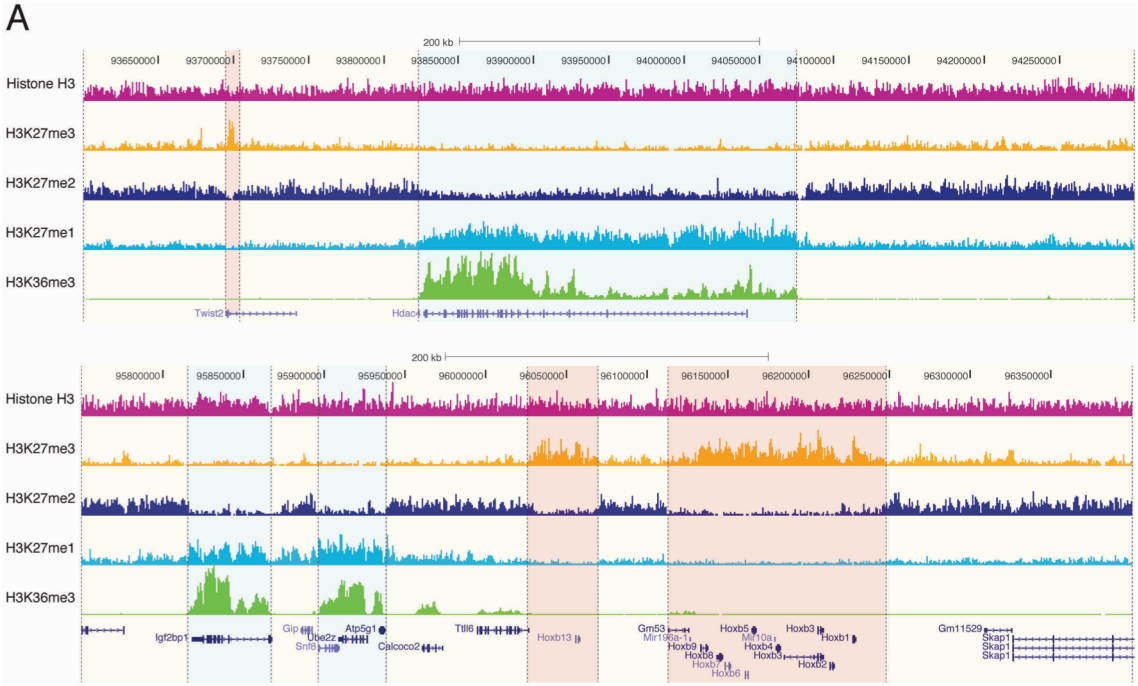


Fig. 2.2.2 A. Localization of different forms of H3K27 methylation Genomic regions showing enrichment for different forms of methylations on H3K27. H3K27me1 enrichment domains are highlighted in blue, while H3K27me3 domains are depicted in red.

To see whether above observation holds true over genome wide scale, correlative analysis was performed between all datasets. On comparing enrichment scores of H3K27me1, H3K27me2 and H3K36me3 within all annotated intragenic regions we observe that H3K27me1 positively correlate with H3K36me3 deposition whereas H3K27me2 negatively correlate with that of H3K27me1 and H3K36me3 deposition (figures 2.2.2B,C).

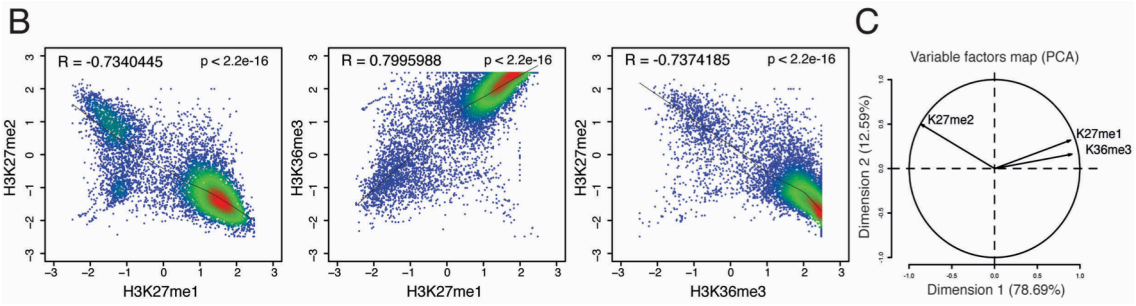


Fig. 2.2.2 B,C. Correlation between PTMs. (B) Scatter plots showing the correlation of enrichments normalized to the histone H3 density between K27 and K36 PTMs in gene bodies of all annotated genes. Pearson correlation values are indicated on top of the plot. (C) Variable plot from Principal component analysis (PCA) representing drgree of correlation between PTMs in gene bodies of all annotated genes.

2.2.3. Distinct H3K27 methylation domains correlate with transcription status.

From literature, it is known that histone PTM H3K36me3 is shown to have their presence in gene bodies of genes undergoing active transcription (Kizer et al., 2005; Li et al., 2003; Li et al., 2002a; Xiao et al., 2003). We were interested in exploring further correlation between histone marks accumulation and transcriptional status of genes. Based on the levels of H3K27me1/2 in intragenic regions we divided all genes into three main groups, and genes within each group were linked to their expression levels. For this analysis we took advantage of publicly available microarray data (Leeb et al., 2010). From the results it can be clearly observed that, highly transcribed genes posses high levels of H3K27me1 in their intragenic regions, while genes with lower transcriptional levels accumulate H3K27me2 in their gene bodies (figure 2.2.3A). This observation holds true, even if we do analysis in reverse order. Like ranking all genes on the basis of their expression levels, and then correlating with levels of enrichment for both H3K27me1 and H3K27me2 along gene bodies. This analysis too inferred same results as above where nearly 90% of all highly expressed genes were enriched

for H3K27me1, whereas 90% of low expressed genes showed enrichment for H3K27me2 in their gene bodies (figure 2.2.3B, C).

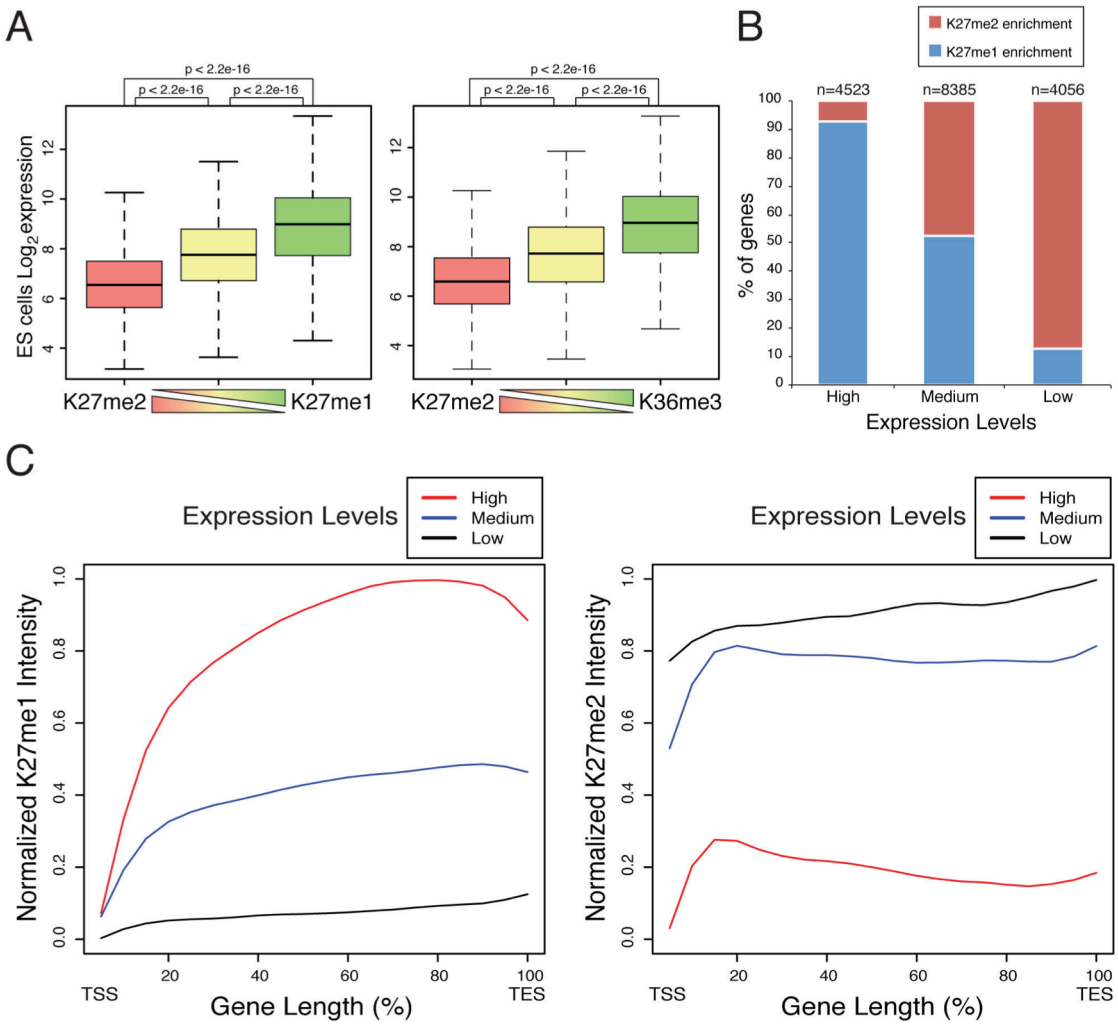


Fig. 2.2.3 A-C. Correlation between levels of K27 methylation and gene transcription. (A) Expression levels of all RefSeq genes grouped in three categories relative to H3K27me2 and H3K27me1 enrichments within their gene bodies. (B) Proportion of K27me1 and K27me2 enriched genes within each group of expression. (C) Composite profiles of H3K27me1 and H3K27me2 over gene bodies for all the three groups of gene sets classified on the basis of their expression level.

Taking together, above data indicate that H3K27 methylation domains are deposited throughout the genome in a mutually exclusively manner and correlate with genes transcriptional activity.

2.2.4. Intragenic H3K27me1 deposition is PRC2 dependent and is linked to active transcription.

To show that H3K27me1 and H3K27me2 depositions are completely driven by the enzymatic activity of PRC2, we performed ChIP qPCR analysis in both intergenic and intragenic regions corresponding to H3K27me1 and H3K27me2 enrichment. This experiment showed that both intergenic H3K27me2 and intragenic H3K27me1 were lost in Eed KO cells (figure 2.2.4A). Similar results were observed on other targets too. To see whether same phenomena can be replicated on genome wide scale, we performed ChIP-seq for both H3K7me1 and H3K27me2 in WT and EedKO mESC. From the analysis of data, a difference in the levels of H3K27me1 modification in WT and Eed KO samples can be easily captured (figure 2.2.4B). We demonstrated that in K27 monomethylation is lost from genes harboring the modification in WT cells (figure 2.2.4C,D). With all above-mentioned evidences, we can confirm that H3K27me1 deposition in gene bodies of genes with active transcription is dependent on the enzymatic activity of the PRC2 complex. As mentioned earlier in 2.2.3, H3K27me1 and H3K27me2 correlate with the transcriptional status of the genes in which they are deposited. For this reason we conceive of a possible role for intragenic H3K27me1 in promoting transcription of gene.

To prove above hypothesis, we analyzed published transcription data for WT and Eed KO mESC (Leeb et al., 2010) correlating differentially expressed genes with the enrichment of different forms of H3K27 methylation at intragenic regions in WT mESC. From the results it is evident that genes enriched for H3K27me2 had increased expression in EedKO sample, while genes enriched for H3K27me1 show low levels of expression in Eed KO mESC (figure 2.2.4E). These results were further confirmed by expression and ChIP qRT-PCR analysis for certain selected genes (figure 2.2.4F, G).

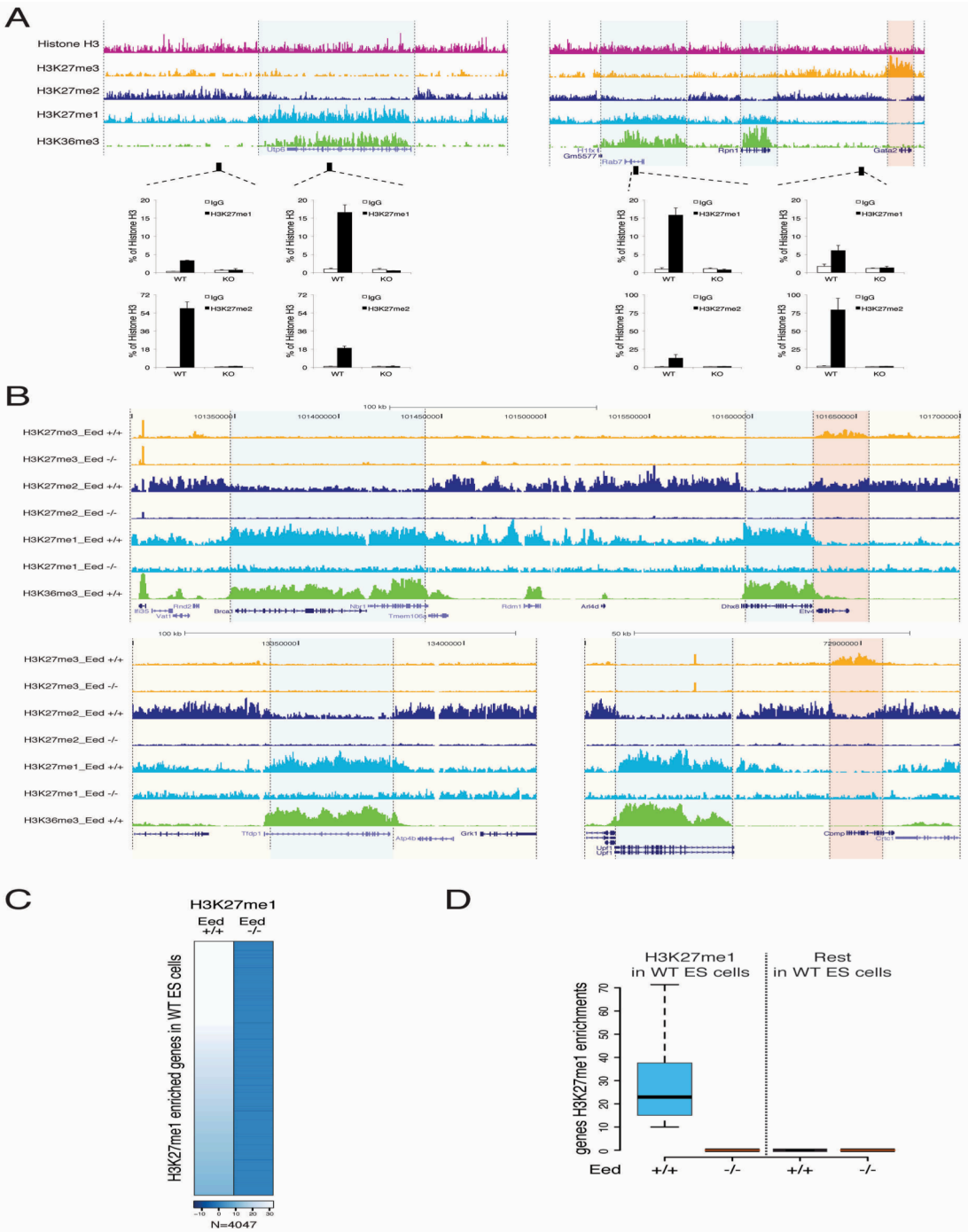


Fig. 2.2.4 A-D. PRC2 dependent H3K27 methylation. (A) qRT-PCR of K27me1/2 ChIP in WT and Eed KO samples in the selected genomic regions. Black boxes indicate primers position within genomic loci. ChIP enrichments are normalized to histone H3 density. IgG ChIPs from rabbit were used as negative control. (B) Genomic snapshots of H3K27me1/2/3 in WT (Eed +/+) and Eed KO (Eed -/-) in mESC along with H3K36me3 from E14 mESC. H3K27me1 domains are highlighted in blue while H3K27me3 domains are highlighted in red. (C) Heat map of H3K27me1 enrichment in WT (Eed +/+) and Eed KO (Eed −/−) for genes enriched for H3K27me1 in WT condition (−10log10 p value ≥ 10 scored from chi-square test between H3K27me1 and H3). (D) Box plot analysis of H3K27me1 ChIP-seq enrichment intensities between WT (+/+) and Eed KO (-/-) mESC for all the annotated RefSeq genes that were divided in two groups based on their H3K27me1 levels in WT mESC (-Log10p-value cut off = 10).

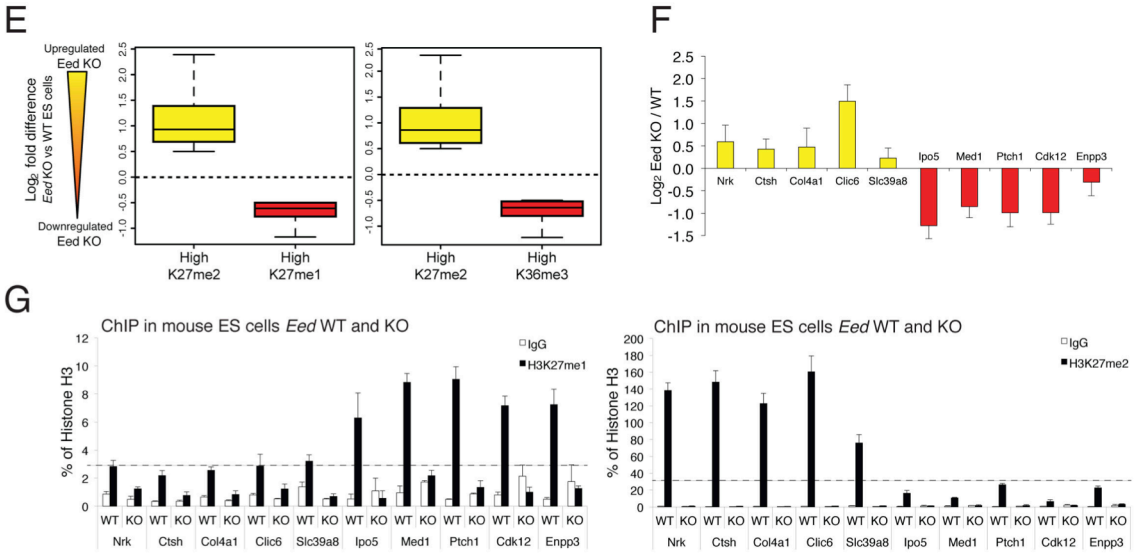


Fig. 2.2.4 E-G. Changes of genes expression upon loss of PRC2 activity. (E) Box plot of fold change in expression levels of differentially regulated genes between WT and Eed KO mESC for H3K27me2 and H3K27me1. For the analysis, the top 15% enriched genes (N~1000) were considered. (F) Relative differences in expression levels between WT and Eed KO mESC of the selected target genes determined by qRT-PCR analysis. (G) qRT-PCR analysis for the indicated intragenic regions of H3K27me1 and H3K27me2 ChIP assays performed in WT and Eed KO mESC using. ChIPs with IgG rabbit were performed as negative control. ChIP enrichments were normalized to histone H3 density.

2.2.5. H3K27me1 PTM is required for correct gene transcription.

From above results, we have shown that transcriptional changes on loss of PRC2 correlated with H3K27me1 deposition, suggesting that H3K27me1 is required for proper activation of transcription of target genes. To further validate this link, we extended our analysis to see if it also occurred during the process of differentiation of WT and Eed KO mESC. For this study we made use of cell culture differentiation into embryoid bodies (EBs) as our model. Eed KO cells were failed to differentiate properly as compared to that of WT cells, this could be confirmed by inactivation of canonical differentiation genes (figure 2.2.5A).

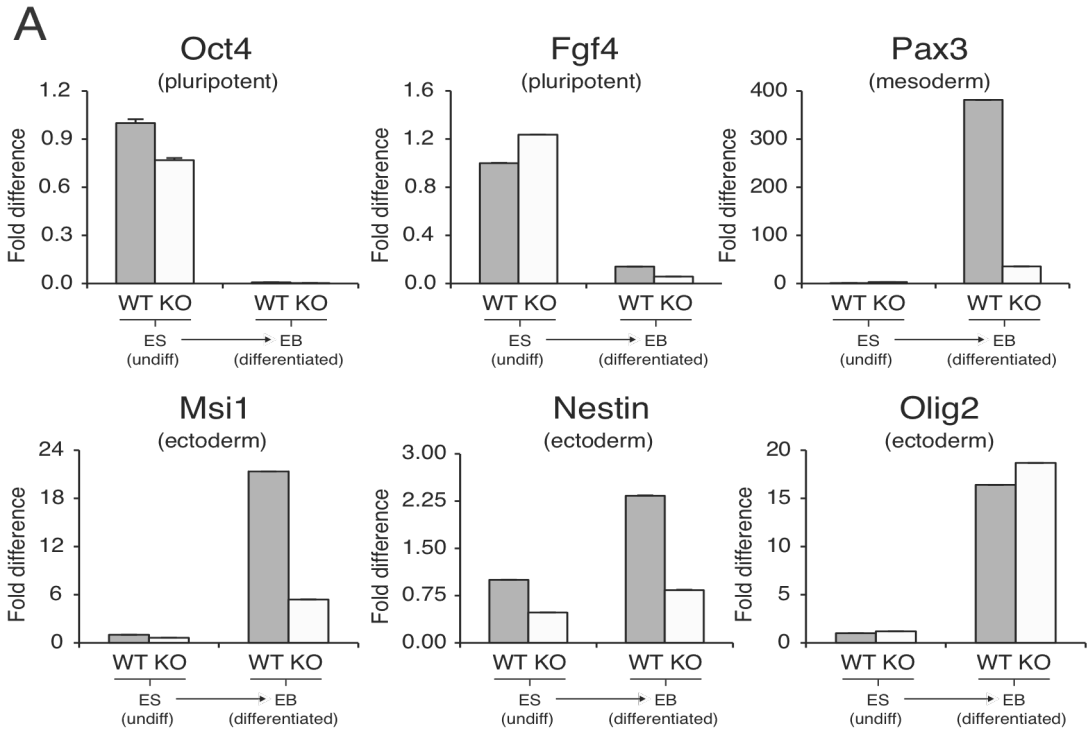


Fig. 2.2.5 A. mESC deficient for PRC2 fail to differentiate. Relative expression of the indicated differentiation markers determined by qRT-PCR in WT and *Eed* KO mESC before (ES) and after 9 days of differentiation (EB).

We carried out transcriptome profiling on WT and Eed KO EBs, as well as on un-differentiated samples. From the analysis, we observed impairment in the activation of genes upon differentiation in PRC2 KO mESC, and this correlated with lack of deposition of H3K27me1 at their gene bodies. The deposition of H3K36me3 is unaltered (figure 2.2.5B,C).

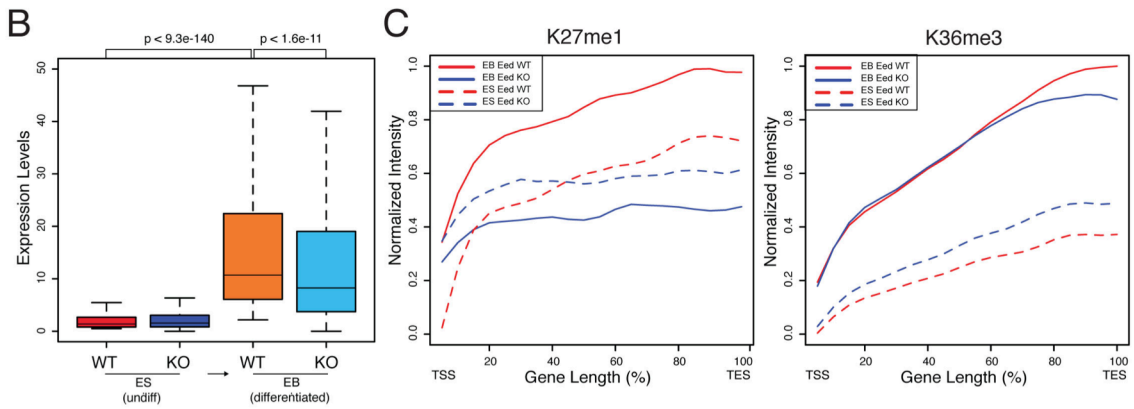


Fig. 2.2.5 B,C. H3K27me1 is gained in genes which are up-regulated in the process of differentiation. (B) Expression levels of up-regulated genes during differentiation process in WT and Eed KO samples (N=844). (C) Average profiles of H3K27me1 and H3K36me3 through the intragenic regions of genes activated upon EB differentiation.

From all these evidences, we can say that, during differentiation process also, H3K27me1 is deposited by PRC2 enzymatic activity and it is required for proper activation of PRC2 target genes. These results also showed that H3K36me3 modification is unaltered in absence of PRC2 that mean that the modification precedes H3K27me1 deposition.

2.2.6. H3K27me2 deposition in genome protects non-cell type specific enhancers.

From our initial mass spectrometry results we observed that the widely deposited H3K27me2 histone mark in the genome accounts for approximately the 70% of total K27 PTMs of histone H3, this led us to speculate about its functional importance in the genome. Instead of exploring role of H3K27me2 as any other histone mark, which

is present between functional domains marked by H3K27me1 and H3K27me3, we speculated its role in the context of genomic control. From literature we know that upon loss of PRC2 enzymatic activity levels of H3K27ac PTM are significantly increased (Pasini et al., 2010; Tie et al., 2009); we too observed identical phenomenon on performing similar experiment in mESC lines KO or interfered for PRC2 component compared to wild type or control mESC (figure 2.2.6A). It has been reported that H3K27ac is enriched, along with other PTMs.

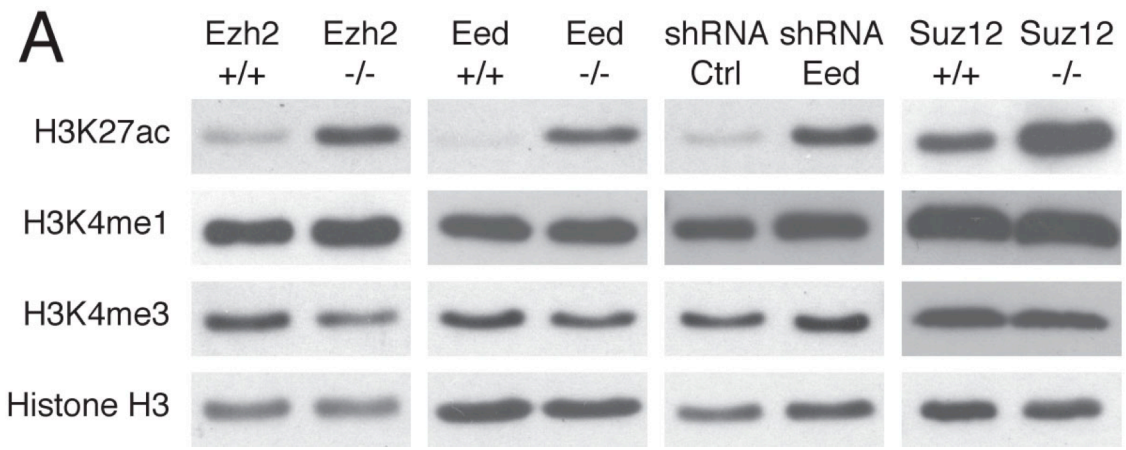


Fig. 2.2.6 A. Global levels of H3K27ac increase upon loss of PRC2 activity. WB analyses of different modifications upon loss of different components of PRC2 in mESC.

Has H3K27me2 is broadly distributed in intergenic, this derived us to suspect that H3K27me2 functionality might be linked with regulatory enhancer elements. On the basis of H3K27ac levels, enhancers can be either in active or poised. Active enhancers are characterized by high levels of H3K27ac where as poised enhancers are characterized by low/no H3K27ac (Creyghton et al., 2010b; Rada-Iglesias et al., 2011). To explore role of H3K27me2 in enhancer control, we performed ChIP-seq analysis to assess the genome wide changes of H3K27ac both in presence and absence of PRC2 activity in mESC cells. On analyzing data we observe differential

enrichment of H3K27ac between Eed KO and WT mESC, which is consistent with increase in levels of H3K27ac from immunoblot analysis. Venn diagram in figure 2.2.6B, clearly shows that number of H3K27ac peaks are preferentially gained in Eed KO (called as "Eed KO unique"), while other peaks were present only in Eed wt mESC (called as "Eed wt unique"). Figure 2.2.6C represents few genomic screenshots of "unique" regions both in WT and Eed KO samples. H3K27ac peaks were annotated with respect to promoters (± 2.5 Kb region from TSS genes); on doing so we observed that the peaks which are shared by both Eed WT and Eed KO are equally distributed among TSS and not TSS regions, where as "unique" peaks in both Eed WT and Eed KO showed a preferential enrichment towards not TSS regions (figure 2.2.6B). This suggested that unique peaks, which are influenced by PRC2 activity, could reside at regulatory elements throughout the genome.

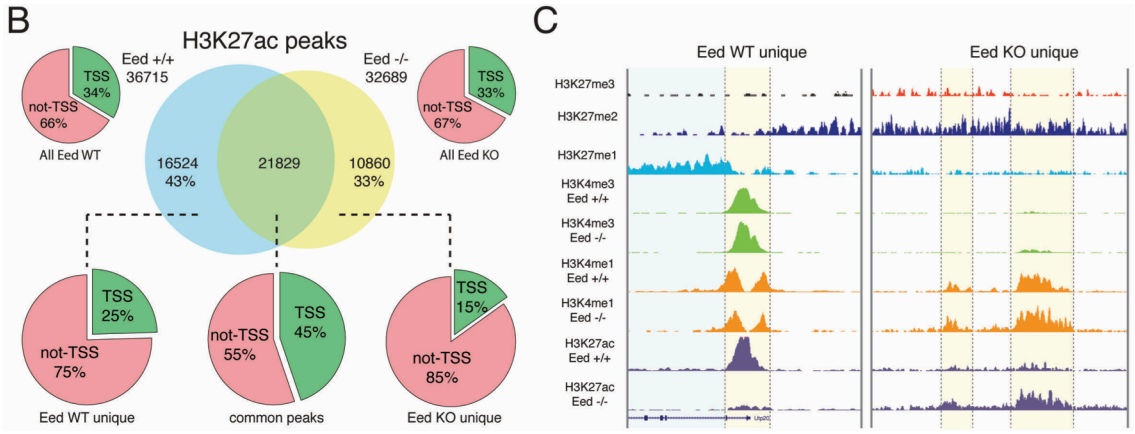


Fig. 2.2.6 B,C. Distribution of H3K27ac enriched regions upon loss of PRC2 activity. (B) Overlap of H3K27ac peaks between WT (Eed +/+) and Eed KO (Eed -/-) mESC. The pies depict the percentage distribution of the different groups of H3K27ac peaks relative to promoter region of all genes. Promoters regions are defined as a ±. 2.5kb region around centered the TSS. (C) Snapshots representing different PTMs in regions where H3K27ac is lost and gained in WT (Eed +/+) and Eed KO (Eed -/-) mESC highlighted in yellow.

In order to understand whether these unique regions are true enhancers we performed ChIP-seq for H3K4me1 and H3K4me3 in Eed WT and Eed KO mESC. From the snapshots (figure 2.2.6D) of unique acetylation regions in Eed WT and Eed KO samples it can be noticed that these are enriched for H3K4me1 but show no signs of H3K4me3. Even through genome wide quantification analysis we observe minimal levels of H3K4me3 as reported by box plot distribution (figure 2.2.6E).

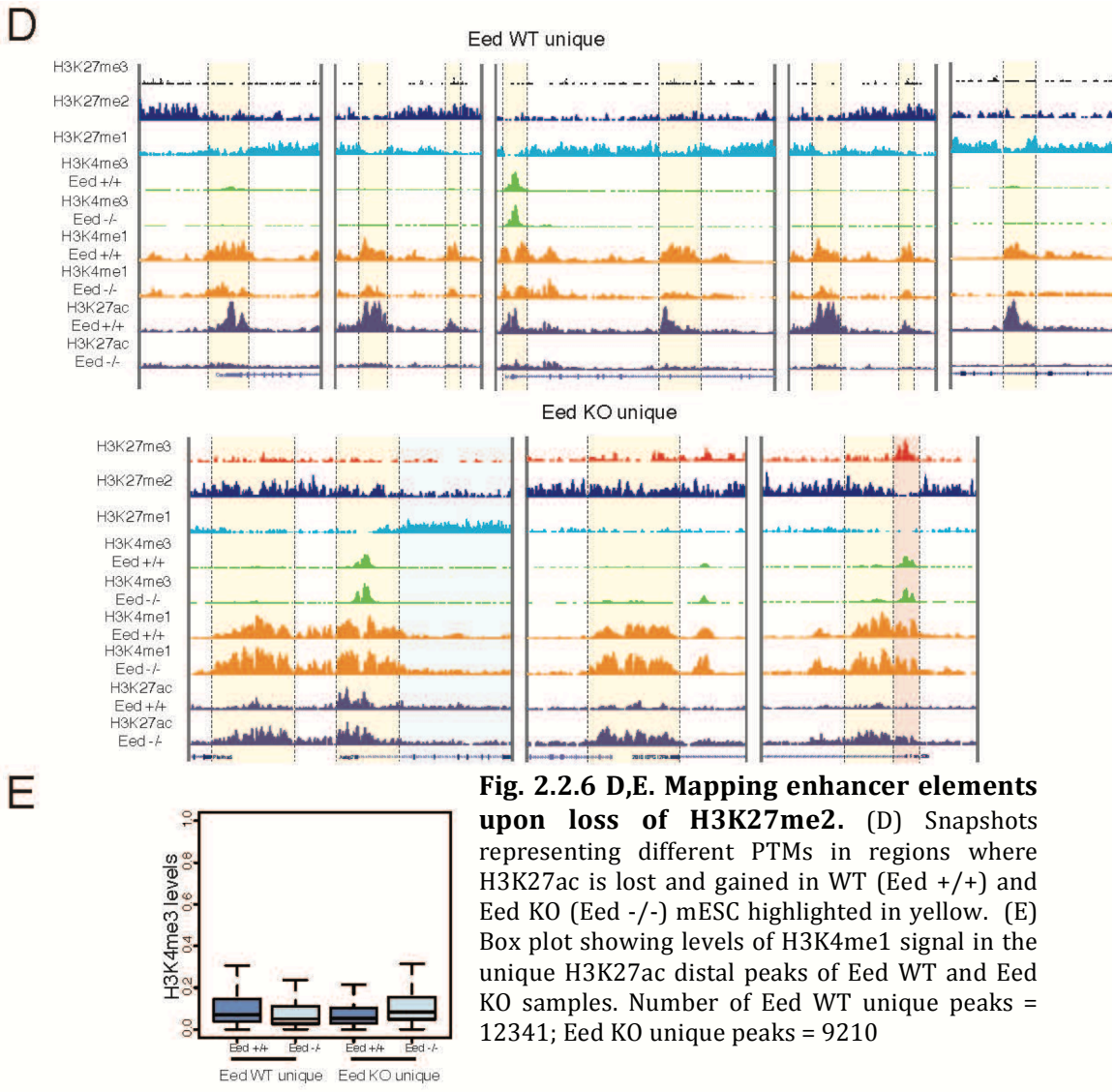


Fig. 2.2.6 D,E. Mapping enhancer elements upon loss of H3K27me2. (D) Snapshots representing different PTMs in regions where H3K27ac is lost and gained in WT (Eed +/+) and Eed KO (Eed -/-) mESC highlighted in yellow. (E) Box plot showing levels of H3K4me1 signal in the unique H3K27ac distal peaks of Eed WT and Eed KO samples. Number of Eed WT unique peaks = 12341; Eed KO unique peaks = 9210

To get global picture of changes of different histone modification in these unique regions of acetylation in both Eed WT and Eed KO samples we quantified all modifications and represented their normalized intensities in form of heat maps (figure 2.2.6F). From this analysis we noticed that on the basis of H3K4me1 levels, activated enhancers in Eed KO cluster fall into two different groups as shown in figure 2.2.6F. It can seen that, in class I enhancers H3K4me1 was pre-existing in Eed WT, while in class II enhancers H3K4me1 is gained along with H3K27ac deposition (figure 2.2.6F,G). Class I enhancers are accounted for 60% of the total enhancers in Eed KO sample, whole class II enhancers accounted for approximately 40% of total. Distributions of different normalized ChIP intensities of different PTMs, which are

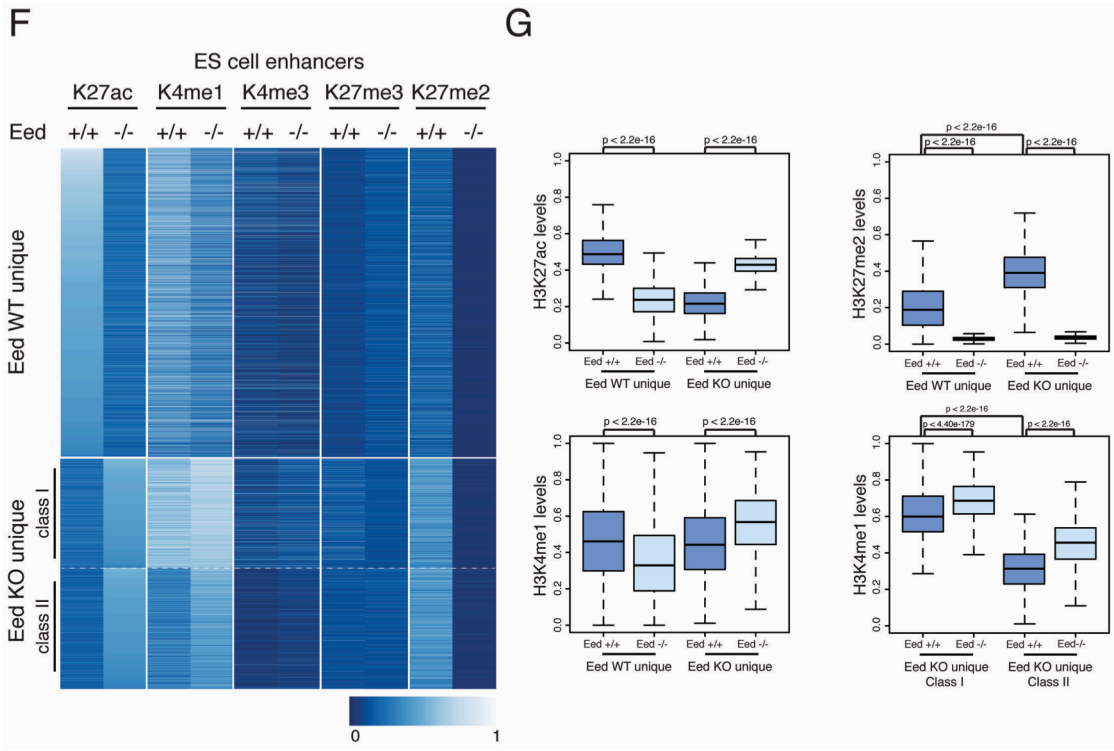


Fig. 2.2.6 F, G. Regulation of enhancers upon loss of H3K27me2 in mESC. (F) Heatmap of normalized intensities of H3K27ac, H3K4me1, H3K4me3, H3K27me3, H3K27me2 in WT and Eed KO mESC for all distal H3K27ac peaks found in either WT (Eed WT unique peaks) or Eed KO (Eed KO unique peaks). Classification of H3K27ac peaks found only in Eed KO into two classes, Class I (n = 4,391) and Class II (n = 4,819) was applied on the basis of pre-existence of H3K4me1 in Eed WT sample. Grouping was based on k mean clustering (k = 2) with respect to the H3K4me1 normalized intensities in Eed WT ESCs. (G) Boxplot analyses quantifying the data shown in figure 2.2.6 F. p value was calculated by Wilcoxon rank test.

represented as heatmap, are also presented in box plot from (figure 2.2.6G). To show that above quantification holds true for both gain and loss of H3K27ac in Eed WT and KO cells, qPCR experiments were carried out in K27ac WT unique and KO unique regions. From the results (figure 2.2.6H), it is evident that the trend holds true and is very similar to that of genome wide phenomena (2.2.6F,G).

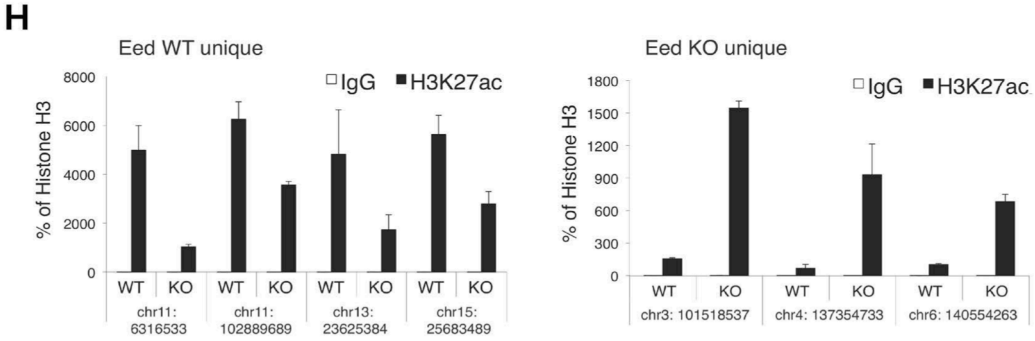


Fig. 2.2.6 H. Validation of lost and gained enhancer elements. (H) qRT-PCR analyses of DNA purified from H3K27ac ChIP in WT and Eed KO mESC using primers amplifying the indicated genomic loci.

Previous work in human ESC described a class of enhancers enriched for H3K27me3 (Rada-Iglesias et al., 2011), we tested for its levels in our unique enhancers in Eed WT and Eed KO, and showed that H3K27me3 is present at very minimal levels (figure 2.2.6I). In accordance with this, we also show that Eed KO unique enhancer region had no Ezh2 association and did not overlap with CpGi, which are well known determinants for targeting PRC2 and sites of H3K27me3 (figure 2.2.6J).

To see if have any relation between activated enhancers in Eed KO mESC and its affect on gene activation, we carried out analysis linking enhancer sites with the closest upregulated genes in Eed KO mESC. As shown by box plot in figure 2.2.6K, the distance between acquired enhancers in Eed KO mESC and the closest upregulated

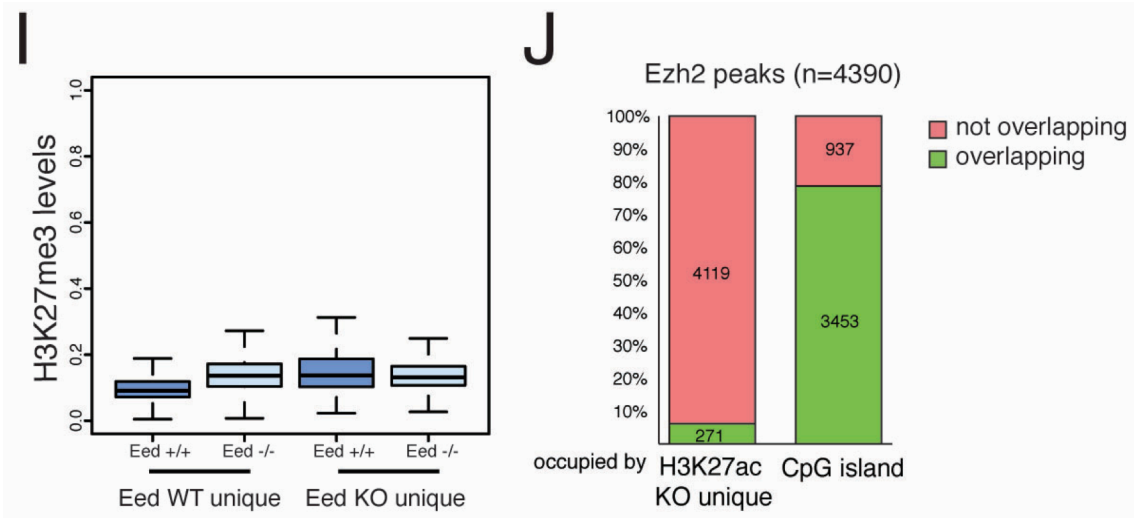


Fig. 2.2.6 I,J. Unique enhancers are not enriched for H3K27me3 and do not reside on CpG islands. (I) Box plot showing the quantification of H3K27me3 signal in the unique H3K27ac distal peaks of Eed WT and Eed KO samples. (J) Percentage of Ezh2 peaks occupancy (determined by ChIP-seq analysis in mouse E14 ES cells) and of CpG islands respect to genomic regions corresponding to H3K27ac peaks uniquely found in Eed KO mESC.

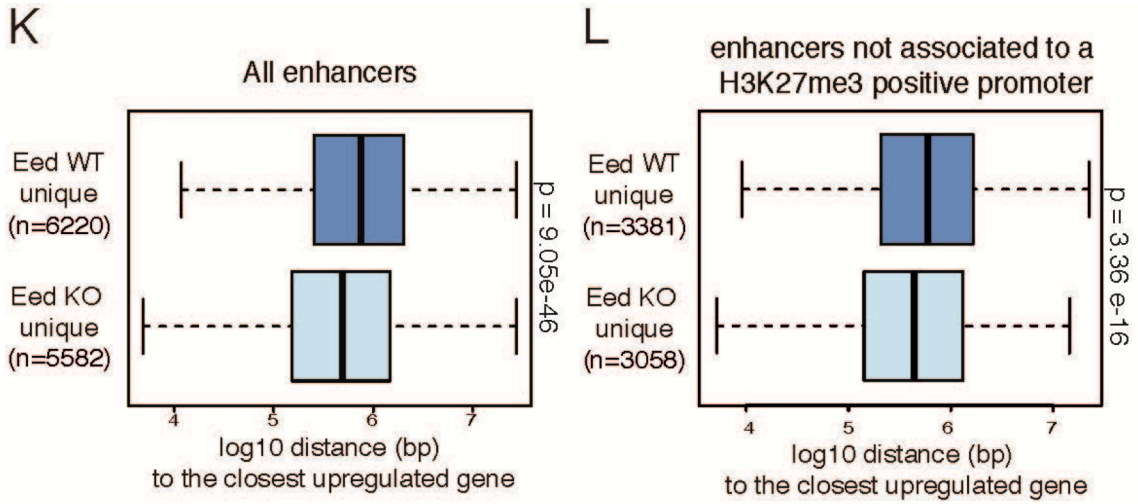


Fig. 2.2.6 K,L. Activation of enhancers upon loss of PRC2 correlates with closest gene activation (K) Box plot representating distance between enhancers (for WT and Eed KO samples) and the up-regulated genes in Eed KO ES cells. All identified enhancers are included in the analysis. (L) Same as K, but in this case enhancers associated to a H3K27me3 positive gene in WT ES cells were excluded from the analysis. H3K27me3 enriched genes were defined by the presence of a H3K27me3 peak within +/- 2.5kb from the TSS. p-values were calculated by Mann-Whitney Test.

genes was reduced respect to the distance observed analyzing enhancers that lost H3K27ac upon PRC2 depletion (Eed WT unique). This result did not depend on H3K27me3 levels deposition at TSS of analyzed genes in WT mESC (figure 2.2.6L).

This observation further endorses our protective model by which H3K27me2 controls enhancer activation by preventing aberrant H3K27ac deposition at these regulatory elements.

If we assume that H3K27me2 act as protective model then in that case it would be right to argue that both classes of enhancers regions in Eed WT and Eed KO should be enriched for H3K27me2 on losing H3K27ac. Using, unique enhancer sites in Eed WT mESC we show that H3K27ac is negatively correlated with H3K27me2 deposition (figure 2.2.6M). As additional proof supporting our mechanism, we set up experiment where we inhibited the enzymatic activity of histone acetyl transferase (HAT) Cbp and p300 by treating WT mESC with the chemical compound C646 for 48 hours. Levels of global H3K27ac with respect to control cells were reduced (figure 2.2.6N), and then we performed ChIP-seq for H3K27me2 and H3K27ac in treated and control mESC to test the behavior of such histone PTMs. From composite profile and from relative quantification by box plots, we found that upon treatment, there is a loss of H3K27ac at 4800 enhancers, which instead is accumulated by H3K27me2 (figure 2.2.6O,P). These results validate our model, proving that loss of H3K27ac from enhancer elements is replaced by H3K27me2 deposition.

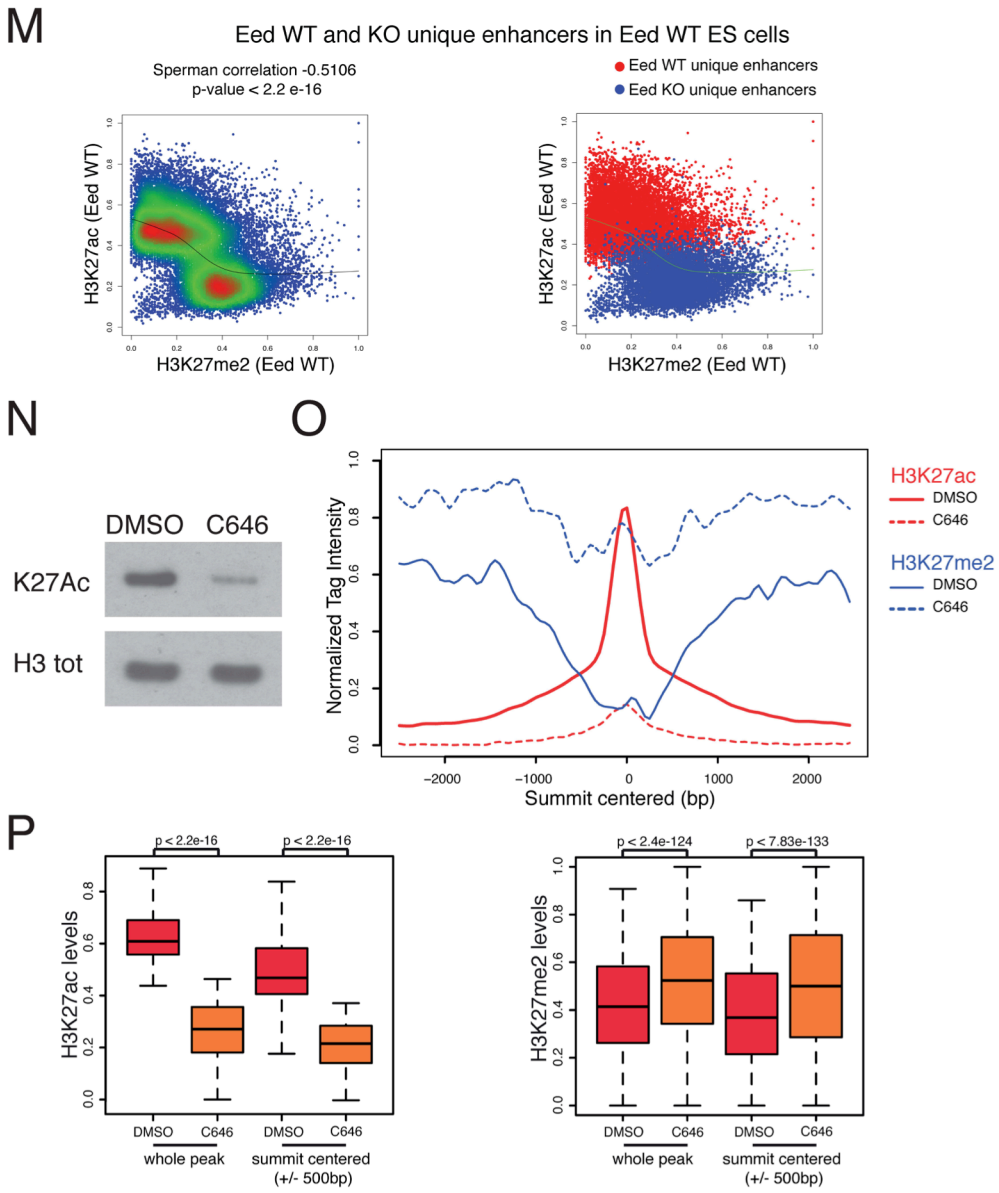


Fig.2.2.8 M-P. Anti-correlation between H3K27ac and H3K27me2 at unique enhancers sites and loss of H3K27ac at enhancer sites is replaced by H3K27me2. (M) Scatter plot showing correlation between H3K27ac and H3K27me2 levels in WT ESCs for all unique enhancers regions identified in WT and Eed KO samples. Left panel shows whole density distributions. Right panel distinguishes Eed WT unique (red) and Eed KO unique (blue) enhancers. The Spearman correlation value is indicated (rs = 0.5106). p value was calculated by asymptotic t approximation (N) Immunoblot analysis for H3K27ac antibody of histones extracted from mouse E14 mESC treated with 35 μM C646 p300 inhibitor for 48 h. DMSO was used as vehicle control. Histone H3 was used as loading control. (O) Average profiles of H3K27ac and H3K27me2 deposition around 2500 bp up and downstream from centered H3K27ac peak summit of regions that loose H3K27ac upon treatment with C646 compound for 48 h (N=4838). (P) Box plots with quantification levels of H3K27ac and H3K27me2 at the same enhancer sites of figure 2.2.8 O upon treatment with C646 for the complete H3K27ac peak region or for a 1kb genomic region surrounding the summit of peak.

Overall, from our findings we can conclude that H3K27me2 by PRC2 ensures correct activation of enhancers, preventing aberrant deposition of H3K27ac at these regulatory elements.

Chapter 3 – MATHERIAL AND METHODS

3.1. ChIP_QC

This program constitutes several modules, each with its own capabilities. They are dedicated for quantification, correlation, differential, enrichment, or classification studies. Here we describe about its design, file formats and different methods it uses for analysis. ChIP_QC is available online and can be downloaded from https://sourceforge.net/projects/epimine/

3.1.1. Input Data

Each module requires input data, which can be any of the following three types: bed, bam or genome file.

- bed file(s) are tab separated file(s) containing information about certain locus of genome. More details about the format can be kown from https://genome.ucsc.edu/FAQ/FAQformat.html#format1.
- bam files are standard binary format file containing details about the alignment of sequencing data with reference genome.
- genome file is also tab-separated file with two columns listing chromosomes and their length.

3.1.2. Samples and Datasets

Samples are referred as multiple bed files containing ROI where one bed file implies one sample. Datasets are referred to as aligned files against which multiple samples are to be analysed.

3.1.3. Overlap

For computing overlapping regions between any two given samples we implemented fast and efficient BITS algorithm (Layer et al., 2013).

3.1.4. Random regions

Irrespective of analysis it is always important to know whether the behaviour shown by data of our interest is something meaningful or is just random by chance. For such comparison, we provide option in our program for generating random data (if enabled). Random regions are generated with a similar size as that of input data size by shuffling the genomic coordinates and chromosomes of input data but within the framework of the reference genome. These regions are analysed in parallel to the main input data and results are generated correspondingly. On comparing results between main input data and random data one can easily judge whether both look similar or different. If they look completely different that means that the results of main input data is not random by chance. Below is the table representing ROI in left column and right column shows random data generated by shuffling coordinates from ROI.

Regions of Interest	**Random Data**
Chr1:238374-38987	Chr5:238374-38987
Chr2:84845-98452	Chr4:84845-98452
Chr4:652873-782383	Chr10:652873-782383
Chr5:187384-198472	Chr2:187384-198472

3.1.5. Quantification

For quantifying different ChIP-seq datasets ROI, program counts total number of reads within each ROI and then normalized to the sequencing depth, length (if the length of different regions are varying). In cases where an input/control sample is provided, normalized reads for input dataset are computed for individual bin and is subtracted from normalized reads of target datasets. If spike-in data is provided then the normalization is carried out in a similar manner as explained in publication (Orlando et al., 2014). For avoiding any skewness in data distribution normalized intensities are log transformed. Program provides option for quantification either on genome wide basis (fig. 3.1.5.2) or quantification restricted to only ROI (fig. 3.1.5.1). In case of genome wide computation, genome is fragmented into small bins on the basis of average length size of user provided ROI and then above mentioned quantification is followed in these regions and later only the bins representing ROI are retrieved. For efficient comparison between different datasets derived either with similar and or different antibodies, quantification is subjected to scaling. If datasets are generated with the same antibody (option provided in program), the whole quantification will be scaled to 0-1. On the other hand, if datasets are generated by different antibodies, then in that case individual datasets are scaled to 0-1 separately allowing liable comparison between datasets. Scaling can be explained better by considering matrix (X) containing n rows and m columns where each row represents one ROI and each column represents each dataset. If all m datasets are generated with same antibody, then scaling is performed in such way that minimum and maximum value of matrix is set to 0 and 1. On the other if all m datasets are generated through

different antibodies then scaling in performed in such way that minimum and maximum value for each column of matrix are set to 0 and 1.

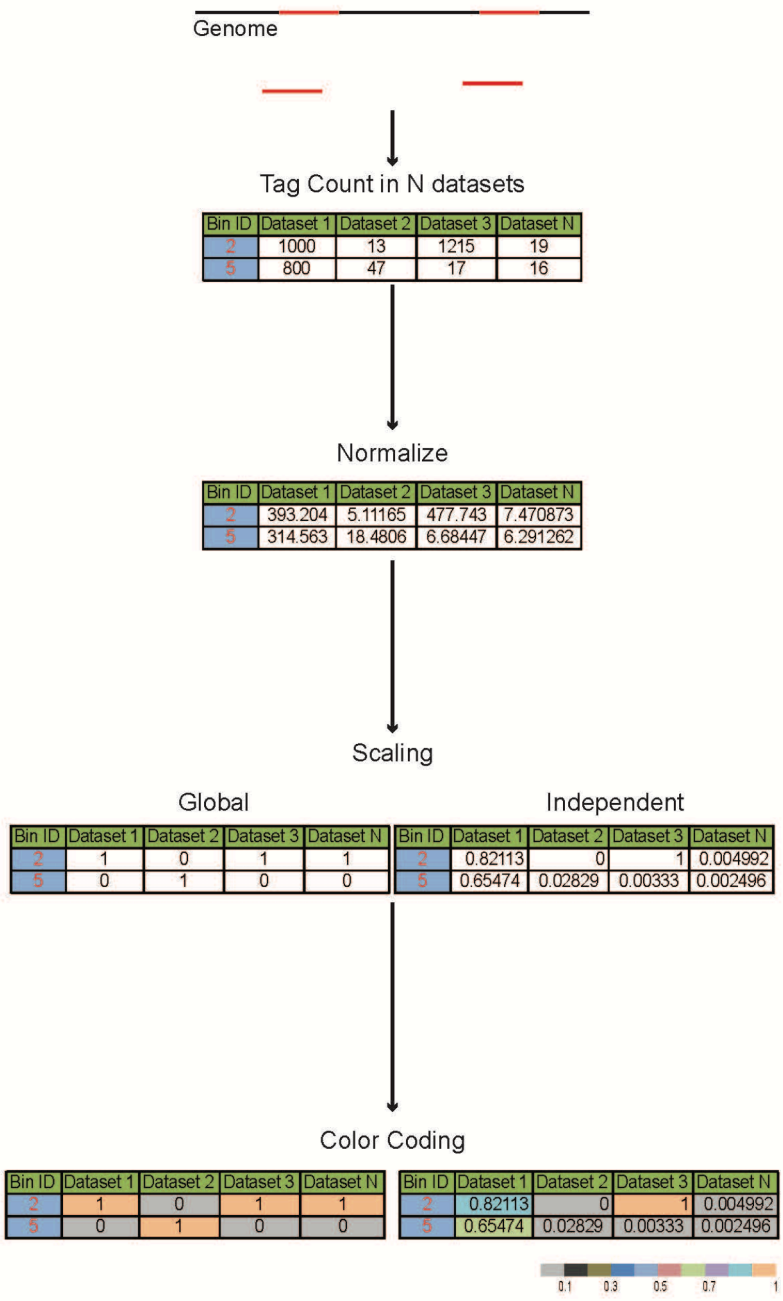
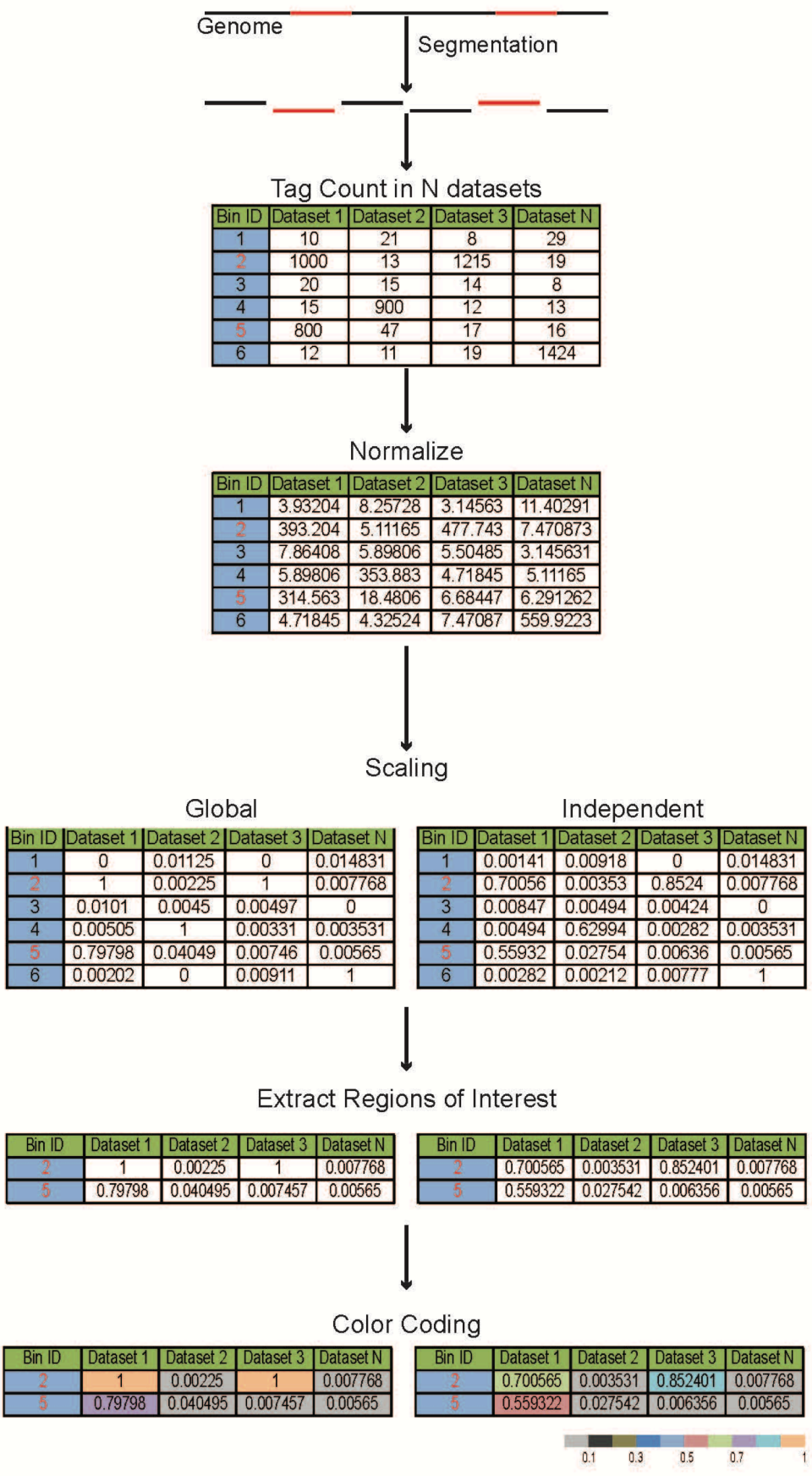


Fig 3.1.5.1 Workflow of quantification within ROIs.



Bin ID	Dataset 1	Dataset 2	Dataset 3	Dataset N
1	10	21	8	29
2	1000	13	1215	19
3	20	15	14	8
4	15	900	12	13
5	800	47	17	16
6	12	11	19	1424

Bin ID	Dataset 1	Dataset 2	Dataset 3	Dataset N
1	3.93204	8.25728	3.14563	11.40291
2	393.204	5.11165	477.743	7.470873
3	7.86408	5.89806	5.50485	3.145631
4	5.89806	353.883	4.71845	5.11165
5	314.563	18.4806	6.68447	6.291262
6	4.71845	4.32524	7.47087	559.9223

Bin ID	Dataset 1	Dataset 2	Dataset 3	Dataset N
1	0	0.01125	0	0.014831
2	1	0.00225	1	0.007768
3	0.0101	0.0045	0.00497	0
4	0.00505	1	0.00331	0.003531
5	0.79798	0.04049	0.00746	0.00565
6	0.00202	0	0.00911	1

Bin ID	Dataset 1	Dataset 2	Dataset 3	Dataset N
1	0.00141	0.00918	0	0.014831
2	0.70056	0.00353	0.8524	0.007768
3	0.00847	0.00494	0.00424	0
4	0.00494	0.62994	0.00282	0.003531
5	0.55932	0.02754	0.00636	0.00565
6	0.00282	0.00212	0.00777	1

Bin ID	Dataset 1	Dataset 2	Dataset 3	Dataset N
2	1	0.00225	1	0.007768
5	0.79798	0.040495	0.007457	0.00565

Bin ID	Dataset 1	Dataset 2	Dataset 3	Dataset N
2	0.700565	0.003531	0.852401	0.007768
5	0.559322	0.027542	0.006356	0.00565

Bin ID	Dataset 1	Dataset 2	Dataset 3	Dataset N
2	1	0.00225	1	0.007768
5	0.79798	0.040495	0.007457	0.00565

Bin ID	Dataset 1	Dataset 2	Dataset 3	Dataset N
2	0.700565	0.003531	0.852401	0.007768
5	0.559322	0.027542	0.006356	0.00565

Fig 3.1.5.2 Workflow for genome wide based quantification for selected ROI.

3.1.6. Differential regulated regions

For any two given datasets, differentially enriched ROI are discovered by computing reads density within and outside the ROI which are then subjected to fisher test, followed by bonferroni correction. In many scenarios we come across situations where we lack replicates for our analysis. In such cases to support identifying enriched regions between two datasets without replicates we implemented Fisher's test. In case of multiple datasets with replicates, differential regions are identified either by kruskal-wallis or by ANOVA statistical test. Reads density within the regions identified as differential are further converted into standard z-score which are represented as heatmap. To segregate differential regions specific to any dataset results can be subjected to clustering. If expression data across multiple systems is provided, then each differential ROI is assigned to the closest gene and the distribution of expression across different clusters is presented as boxplot. In this version of the application we support kruskal-wallis or by ANOVA statistical test for identifying differentially enriched regions among different datasets. Apart from these statistical test much more powerful test are also available. In coming version, we plan to support such statistical analysis in our applications. For this, in coming version we aim to implement limma (Ritchie et al., 2015), which uses linear models for analysing data, and identify differentially enriched regions accordingly.

3.1.7. Correlation

Using genome wide or regions specific quantification (as explained above) correlation between different datasets can be computed either by Pearson, Spearman, Kendall or Principal Component Analysis (PCA) methods. In case where any of thee three initial methods is chosen, the correlation between all possible dataset pairs is computed and

transformed into a correlation matrix. This matrix is then represented as a heatmap where the degree of correlation is associated with a colour code. In case PCA is chosen, the program generates a variable graph with circle of correlation across the first two principal components capturing maximum variance from the data. Variable graph signifies the degree of closeness/relatedness between multiple datasets, where each dataset is represented as an arrow. Variable graph can be interpreted at different levels. First, the amplitude of the angle between two arrows is directly linked to the degree of correlation. The smaller is the angle between two datasets the higher is their correlation. A 90° angle signifies no correlation, while an opposite angle (>90°) reflects an anti-correlation between two datasets. Second, the length of the arrow represent how important is that dataset in representing whole data. Longer the length greater the importance of that variable and the opposite holds true.

3.1.8. Selection and Classification

Many studies in epigenetics involve characterizing/classifying set of ROIs on the basis of some known properties. For such studies we implemented Support Vector Machine (SVM) in our program. It can be used for characterizing two sets of ROI on the basis of given datasets to see whether the given datasets are capable of differentiating it. If the datasets are fruitful in characterization of ROI through above process, analysis can be extended further in classifying new set of ROIs using the constructed model. We choose SVM for classification purposes because of its advantages over others. SVM provides unique and accurate classifiers, it avoids over-fitting of classifier with proper choice of parameters and is robust in classifying noisy data. With biological data we always tend to have noise and data not necessarily is regularly distributed in

such cases we need to make choice of such classifier, which can perform better under such circumstances. In such situations SVM provides right choice of classifier.

In cases where number of datasets used for characterizing two sets ROIs is too large, program provides option for pre-selection of meaningful datasets. Advantage of pre-selection is that it tries to filter out datasets, which contribute very minimal, or none in classification. For filtering out non-contributing factors, this program uses recursive feature selection approach where all-possible subsets are considered and accuracy score for each best combination is reported. Out of these, variables with best combination scoring high accuracy are reported. This combination of datasets can now be further used for building SVM model.

For building classification model, program provides the facility of choosing either linear/no-linear method of classification. Given positive and negative dataset, in case of characterizing/training, the program quantifies provided datasets within all ROIs and employs classification. Quality of analysis can be improved by running k-fold cross validation. Using this approach the training set is split into k groups of approximately the same size, then iteratively train a SVM using k−1 groups and make prediction on the group which was left aside. By default this is set to 10. For a given combination of datasets, analysis presents the performance as receiver operating characteristic (ROC) curve where True Positive Rate (TPR) is plotted against False Positive Rate (FPR). The program lists out the area under the ROC curve (AUC) for the classification. Higher the AUC, greater the possibility of classifying two sets of ROIs. Once satisfied with classification model on training data, the analysis can be further extended in predicting a similar classification on new set of ROIs either in same system or in different system.

3.1.9. Bayesian Network

In any given cell type, different histone modifications and TFs are enriched/bound through the genome. These factors (HMs/TFs) together regulate transcription. Using ChIP-seq approach, for many cell types we have mapped the localization of these different factors through the genome. From these data, one can study which different factors function dependently/independently either genome wide or within ROI. Such studies can be explored in this program using Bayesian Netowrk (BN), which helps us in predicting probabilistic relationships between a set of different factors. In general terms, for a given finite set of random discrete variables X=(x1, x2, x3 ... xn), BN is an directed acyclic graph that signifies joint probability distribution over X. Where nodes correspond to variables and edge correspond to influence of one variable on other. A unique joint probability distribution P of X can be written as:

$$P(X) = \prod_{i=1}^{n} P(\mathrm{Xi}|\Pi\mathrm{Xi})$$

Similarly, for continuous variables, it can be written in terms of global density function as:

$$f(X) = \prod_{i=1}^{n} f(\mathrm{Xi}|\Pi\mathrm{Xi})$$

where Πx_i represent parent(s) of Xi

ChIP_QC supports both discrete and continuous data formats. In discrete based method, the program is fed with n different bed files where each bed file represents one factor. In the case of continuous method, the program is instead fed with n different aligned (bam) files. Depending on the type of analysis selected (genome

wide/only within ROI), for the discrete method, a matrix is constructed signifying the presence or absence of the factor within the region of analysis. In case the continuous method is used, a matrix is constructed with normalized read counts. Using such constructed data, a joint distribution model is learned either by constraint/score/hybrid-based methods. For generating networks with high predictive power, a selected learning method is applied iteratively on randomly selected data (default 90% percent) from original data for 100 times (default). Based on a selected threshold, a probabilistic network is generated with only those edges that are identified at least in 95% (default) of networks.

3.1.10. Datasets

All presented results are generated with human ENCODE (Consortium, 2012) data. We used histone modification (HM), transcription factor and expression datasets of human embryonic stem cells (H1hESC), lymphblastoid (Gm12878), umbilical vein endothelial cells (HUVEC), cervical carcinoma (HeLa-S3), liver carcinoma (HepG2), leukemia (K562), skeletal muscle fibroblast (HSMM), human lung fibroblast (NHLF) and epidermal kertainocytes (NHEK).

3.1.11. Design and Dependencies

Complete program is developed in python platform. Different components wxpython, numpy, pysam and rpy2 are integrated together. R (http://www.r-project.org/) with following packages gplots, RColorBrewer, FactoMineR, ROCR, kernlab, bnlearn, fastcluster, igraph. wxpython is used for graphical user interface, numpy for handling

numerical data in form of matrix, pysam for processing alignment files and rpy2 for statistical analysis through R. This program is supported in Mac OS and Linux.

3.1.12. Aligned Datasets Structure

Each target dataset can be processed individually or in support of input and/or spike in data (if provided).

If input data is provided then it should be provided in either of the following ways:

- each target dataset can have corresponding input with same filename (Table1 Scenario 1),
- or all target datasets can have one input (Table1 Scenario 2).

Scenario 1

Target Directory	Input Directory
X1.bam	X1.bam
X2.bam	X2.bam
X3.bam	X3.bam
X4.bam	X4.bam
Xn.bam	Xn.bam

Scenario 2

Target Directory	Input Directory
X1.bam	X.bam
X2.bam	
X3.bam	
X4.bam	
Xn.bam	

Table1: ChIP_QC supports aligned datasets with input data, which can be provided in following folder structure. Scenario 1: representing folder structure where each input file corresponding to individual target dataset. Scenario 2: representing folder structure where single input file corresponding to all target datasets.

Similarly, if spike-in data is provided, then each target dataset should have corresponding spike-in with same filename (Table 2).

Target Directory	Spike-In Directory
X1.bam	X1.bam
X2.bam	X2.bam
X3.bam	X3.bam

X4.bam	X4.bam
Xn.bam	Xn.bam

Table2: ChIP_QC supports aligned datasets with spike-in data, where program expects each Spike-In file corresponding to individual target dataset.

3.1.13. Modules

ChIP_QC has different modules, each module with its paramters are listed below:

3.1.13.1. TCOR

Program to compute correlation between two different datasets or replicates.

Command line usage

usage: python TCOR.py [options] roi afiles

Mandatory arguments:

taf1	target aligned file 1 (bam file format)
taf2	target aligned file 2 (bam file format)

Optional arguments:

-h, --help	show this help message and exit
--roi	file with list of Regions Of Interest to be analyzed (bed file format). Analysis will be restricted to this set of regions.
-w, --window	window Size. Allowed values. (default 10000)
-o, --output	output directory (default: execution directory)
--qn	set values greater than quantile(x) to quantile(x). Allowed values between 0-1 (default value 0.995)
--method	correlation method (default Pearson). Allowed

values:['Pearson', 'Spearman', 'Kendall'].

Output

- Generates scatter plot (*Correlation.pdf*) with color density distribution where low, medium and high dense population are colored with blue, green and red. Correlation coefficient and p-value are listed as title of the plot.

3.1.13.2. MCOR

Program to compute correlation between more than two different datasets.

Command line usage

usage: python MCOR.py [options] taf genome

Mandatory arguments:

taf	directory with target aligned files (bam file format)
genome	genome file

Optional arguments:

-h, --help	show this help message and exit
--roi	file with list of Regions Of Interest to be analyzed (bed file format). Analysis will be restricted to this set of regions.
--caf	directory with control aligned files (bam file format)
--spikein	directory with spikein aligned files for each corresponding dataset (bam file format)
-p	pearson correlation (default: True)
-s	spearman correlation (default: False)

-k	kendall Correlation (default: False)
--pca	principal component analysis (default: True)
-w, --window	window size. (default 10000)
--qn	set values greater than quantile(x) to quantile(x). Allowed values between 0-1 (default value 0.995)
--color	heatmap color (default Blues). Allowed values:['Blues', 'BuGn', 'BuPu', 'GnBu', 'Greens', 'Greys', 'Oranges', 'OrRd', 'PuBu', 'PuBuGn', 'PuRd','Purples', 'RdPu', 'Reds', 'YlGn', 'YlGnBu', 'YlOrBr','YlOrRd', 'BrBG', 'PiYG', 'PRGn', 'PuOr', 'RdBu', 'RdGy','RdYlBu', 'RdYlGn', 'Spectral'].
-o, --output	output directory (default: execution directory)

Output

- On choosing pearson/spearman/kendall as method of correlation, program generates two *files Correlation_X.pdf* and *Correlation.txt*.
- *Correlation_X.pdf* represents the correlation between factors in form of heatmap where the degree of correlation is signified through color code. X in file name refers to correlation method chosen,
- *Correlation.txt* later file contains the matrix with correlation coefficient between factors.
- On choosing PCA, program generates two files *Correlation_PCA.pdf and PCA.txt*.
- *Correlation_PCA.pdf* represents variable map with circle of correlation through first two principal components capturing maximum variance from the data. All datasets analyzed through PCA are represented as arrows. It can be interpreted at different levels: lower the angle between two datasets higher

the correlation; if they are 90 degree apart from each other signify no correlation, if they are opposite to each other signifying anti-correlation between two datasets. Similarly, length of the arrow represent how important is that dataset in representing whole data. Longer the length greater the importance, lower the length lesser the importance.

- *PCA.txt* file contains details about correlation coefficient of each dataset with two components along with their level of significance.

3.1.13.3. ENRICH

Program to compute to find preferential enrichment of different factors in given ROI.

Command line usage

usage: python ENRICH.py [options] roi afiles

Mandatory arguments:

roi	directory with single/multiple files with Regions Of Interest to be analyzed (bed file format)
afiles	directory with single/multiple files with annotated Regions Of Interest to be analyzed (bed file format)

Optional arguments:

-h, --help	show this help message and exit
-g , --genome	genome file
-r, --random	generate and analyze random regions (default: False)
--stringency	Stringency
-c , --cutoff	proportion/extent of overlap to be considered (default 0.1, this means at least 10 percent region of interest

	should overlap with target region). Allowed values 0-1.
-o , --output	output directory (default: execution directory)

Output

This program generates following output files:

- *<roi filename>_<afile filename>.txt* file(s) containing list of ROI being bound by provided annotated regions (afile),
- *Factors_Plot.pdf* consists of barplot representing the proportion by which ROI is bound by different sets of annotated regions (afile).

3.1.13.4. CoREG

Program to compute to multiple factors co-localization and expression studies.

Command line usage

usage: python CoREG.py [options] roi afiles

Mandatory arguments:

roi	directory with single/multiple files with Regions Of Interest to be analyzed (bed file format)
Afiles	directory with single/multiple files with annotated Regions Of Interest to be analyzed (bed file format)

Optional arguments:

-h, --help	show this help message and exit
-g , --genome	genome file
-e, --expression	expression file (bed file format)
--hc	hierarchical clustering. (default: False)

--km	kmeans clustering. (default: False)
--cut	cut tree at height. Goes with --hc (default value 1.5)
--clusters	number of clusters. Goes with --km (default value 10)
-r, --random	generate and analyze random regions (default: False)
--color	heatmap color (default Blues). Allowed values:['Blues','BuGn', 'BuPu', 'GnBu', 'Greens', 'Greys', 'Oranges', 'OrRd', 'PuBu', 'PuBuGn', 'PuRd', 'Purples', 'RdPu','Reds', 'YlGn', 'YlGnBu', 'YlOrBr', 'YlOrRd', 'BrBG', 'PiYG', 'PRGn', 'PuOr', 'RdBu', 'RdGy', 'RdYlBu', 'RdYlGn', 'Spectral']
--stringency	stringency for overlap
-c , --cutoff	proportion/extent of overlap to be considered (default 0.1, this means at least 10 percent region of interest should overlap with target region). Allowed values 0-1
-o , --output	output directory (default: execution directory)

Output

This program generates following output files:

- *<roi filename>_<hm/hc/km>.png* heatmap showing presence or absence of different sets of annotated regions (afiles) in each individual ROI. File with suffix hm is generated when no clustering is turned on. Similarly file with suffix hc/km refers to the analysis subjected to either hierarchical or kmeans clustering,
- *<roi filename>.txt* matrix file representing presence or absence of different sets of annotated regions in each individual ROI. Presence of any annotated regions

in ROI is denoted by value greater than or equal to 0.99999999999. Similarly, absence is denoted by value less than or equal to 0.00000000001.

- *All_<hm/hc/km>.png* this file is generated only when the number of sets of ROI is more than 1, in that case this file contains combined heatmap view of all sets of ROI.
- *<roi filename>_<hc/km>_cluster_<cluster number>_<color represented in heatmap>.txt* these file(s) are generated only when either (hierarchical/kmeans) clustering options is turned on. Depending on clustering options this file will contain list of ROI being part of that particular selection.
- *<roi filename>_expDist_<hc/km>.pdf* this boxplot file is generated when clustering option is turned on and expression data is provided for the analysis. This represents distribution of target genes expression among different clusters.

3.1.13.5. QIRI

Program to quantify different chip-seq datasets in different sets of ROI.

Command line usage

usage: python QIRI.py [options] roi taf

Mandatory arguments:

Roi	directory with single/multiple files with regions of interest to be analyzed (bed file format)
taf	directory with target aligned files (bam file format)

Optional arguments:

-h, --help	show this help message and exit
--caf	directory with control aligned files (bam file format)
--spikein	directory with spikein aligned files for each corresponding target aligned file (bam file format)
-g, --genome	genome file
--expression	expression file (bed file format)
--cg	compute genome wide analysis (default: False)
--stat	generate boxplots and statistics. (default: False)
--hc	hierarchical clustering. (default: False)
--km	kmeans clustering. (default: False)
--cut	cut tree at height. Goes with --hc (default value 1.5)
--clusters	number of clusters. Goes with --km (default value 10)
--qn	set values greater than quantile(x) to quantile(x). Allowed values between 0-1 (default value 0.99)
--sameAb	used when all datasets are generated against same antibody (default: False)
-r, --random	generate and analyze random regions (default: False)
--color	heatmap color (default Blues). Allowed values:['Blues', 'BuGn', 'BuPu', 'GnBu', 'Greens', 'Greys', 'Oranges', 'OrRd', 'PuBu', 'PuBuGn', 'PuRd', 'Purples', 'RdPu', 'Reds', 'YlGn', 'YlGnBu', 'YlOrBr', 'YlOrRd']
-o, --output	output directory (default: execution directory)

Output

This program generates following output files:

- *<roi filename>_<hm/hc/km>.png* heatmap showing normalized intensities of different datasets (--taf) in each individual ROI. File with suffix hm is generated when no clustering is turned on. Similarly file with suffix hc/km refers to the analysis subjected to either hierarchical or kmeans clustering,
- *<roi filename>.txt* matrix file quantifying different datasets (--taf) in each individual ROI.
- *All_<hm/hc/km>.png* this file is generated only when the number of sets of ROI is more than 1, in that case this file contains combined heatmap view of all sets of ROI.
- *<roi filename>_<hc/km>_cluster_<cluster number>_<color represented in heatmap>.txt* these file(s) are generated only when either (hierarchical/kmeans) clustering options is turned on. Depending on clustering options this file will contain list of ROI being part of that particular selection.
- *<roi filename>_expDist_<hc/km>.pdf* this boxplot file is generated when clustering option is turned on and expression data is provided for the analysis. This represents distribution of target genes expression among different clusters.
- *<taf filename>_boxplot.pdf* this boxplot file is generated on turning on statistics option. It represents the distribution of quantification of that specific dataset among different sets of ROI and the level of significance between all two combinations of ROI is listed in Statboxplot.txt.

3.1.13.6. QARI

Program to quantify different chip-seq datasets around different sets of ROI.

Command line usage

usage: python QARI.py [options] roi taf

Mandatory arguments:

roi	directory with single/multiple files with regions of interest to be analyzed (bed file format)
taf	directory with target aligned files (bam file format)

Optional arguments:

-h, --help	show this help message and exit
--caf	directory with control aligned files (bam file format)
--spikein	directory with spikein aligned files for each corresponding target aligned file (bam file format)
-g, --genome	genome file
--expression	expression file (bed file format)
--extension	number of base pairs to extend from center of region of interest (default 2500bp)
--binSize	bin size in base pair (default 50bp)
--hc	hierarchical clustering. (default: False)
--km	kmeans clustering. (default: False)
--cut	cut tree at height. Goes with --hc (default value 1.5)
--clusters	number of clusters. Goes with --km (default value 10)
--qn	set values greater than quantile(x) to quantile(x). Allowed values between 0-1 (default value 0.95)

--sameAb	used when all datasets are generated against same antibody (default: False)
--smooth	smooth profiles. (default: False)
--strand	use strand information. (default: False)
-r, --random	generate and analyze random regions (default: False)
--color	color of heatmap for correlation plots (default Blues). Allowed values:['Blues', 'BuGn', 'BuPu', 'GnBu', 'Greens', 'Greys', 'Oranges', 'OrRd', 'PuBu', 'PuBuGn', 'PuRd', 'Purples', 'RdPu', 'Reds', 'YlGn', 'YlGnBu', 'YlOrBr', 'YlOrRd']
-o, --output	output directory (default: execution directory)

Output

This program generates following output files:

- *<roi filename>_<hm/hc/km>.png* heatmap showing normalized intensities of different datasets (--taf) in each individual ROI. File with suffix hm is generated when no clustering is turned on. Similarly file with suffix hc/km refers to the analysis subjected to either hierarchical or kmeans clustering,
- *<roi filename>.txt* matrix file quantifying different datasets (taf) in each individual ROI.
- *<taf filename>.pdf* profile of target dataset over all sets of ROI.
- *<taf filename>_Smooth.pdf* same as above but the profile is smoothed. This is generated on turning on –-smooth option.

- *All_<hm/hc/km>.png* this file is generated only when the number of sets of ROI is more than 1, in that case this file contains combined heatmap view of all sets of ROI.
- *Composite_hm.png* composite heatmap of all datasets over all ROI.
- *<roi filename>_<hc/km>_cluster_<cluster number>_<color represented in heatmap>.txt* these file(s) are generated only when either (hierarchical/kmeans) clustering options is turned on. Depending on clustering options this file will contain list of ROI being part of that particular selection.

<roi filename>_expDist_<hc/km>.pdf this boxplot file is generated when clustering option is turned on and expression data is provided for the analysis. This represents distribution of target genes expression among different clusters.

3.1.13.7. PMS

Program to profile chip-seq datasets over different sets of ROI.

Command line usage

usage: python PMS.py [options] --midpoint/--complete roi taf

Mandatory arguments:

Roi	directory with single/multiple files with regions of interest to be analyzed (bed file format).
taf	directory with target aligned files (bam file format).

Optional arguments:

-h, --help	show this help message and exit.
--caf	directory with control aligned files (bam file format).

--spikein	directory with spikein aligned files for each corresponding target aligned file (bam file format)
-g, --genome	genome file.
--midpoint	compute analysis from the midpoint of region of interest (default: False)
--complete	compute analysis over complete region of interest (default: False).
--strand	use strand information (default: False).
--smooth	apply smoothing to profiles (default: False).
--scale	apply scaling (default: False).
--sameAb	used when all datasets are generated against same antibody. Goes with --scale option (default: False).
-r, --random	generate and analyze random regions (default: False).
--extension	number of base pair to be extended from center. Goes with --midpoint option. (default 2500bp)
--binSize	bin Size. Goes with --midpoint option. (default 50bp)
--bins	number of bins. Goes with --complete option. (default 20)
--type	profiling type (default mean). Allowed values: mean/median
--qn	set values greater than quantile(x) to quantile(x). Allowed values between 0-1 (default value 1)
-o, --output	output directory (default: execution directory)

Output

This program generates following output files:

- *Profile.pdf* profile of target datasets (taf) over all sets of ROI.
- *Profile_Smooth.pdf* same as above but the profile is smoothed. This is generated on turning on –smooth option.
- *Profile_CI.pdf* profile of target datasets (taf) wilth confidence intervals over all sets of ROI (--roi).
- *scaledProfile.pdf* scaled profile of target datasets (taf) over all sets of ROI.
- *scaledProfile_Smooth.pdf* same as above but the profile is smoothed. This is generated on turning on –smooth option.

3.1.13.8. TDIFF

Program to identify differentially enriched regions between two datasets

Command line usage

usage: python TDIFF.py [options] roi taf1 taf2

Mandatory arguments:

Roi	file with list of Regions Of Interest to be analyzed (bed file format)
taf1	target aligned file 1 (bam file format).
taf2	target aligned file 2 (bam file format).

Optional arguments:

-h, --help	show this help message and exit.
--spikein	directory with spikein aligned files for each corresponding target aligned file (bam file format)
--fc	fold change cutoff . (default 4)
--qn	set values greater than quantile(x) to quantile(x).

	Allowed values between 0-1 (default value 0.99)
--sig	level of signficance. (default 0.05)
--expression	expression file (bed file format)
--discard	discard lower distribution of data. (default 0.10)
-o, --output	output directory (default: Current Directory)

Output

This program generates following output files:

- *Results.txt* file containing list of all ROI along with their intensities in two different datasets, fold change, p-value and bonferroni corrected p-values.
- *DiffEnhRegions_pValue_Up.txt* file containing list of ROI, which are significantly enriched in taf1 as compared to that of taf2 with their level of significance less than --sig.
- *DiffEnhRegions_pValue_Down.txt* file containing list of ROI, which are significantly enriched in taf2 as compared to that of taf1 with their level of significance less than --sig.
- *VolcanoPlot_pValue.png* plot representing all significantly enriched ROI on the basis of p-value in taf1 and taf2 on right and left side in cyan color.
- *DiffEnhRegions_qValue_Up.txt* file containing list of ROI, which are significantly enriched in taf1 as compared to that of taf2 with qValue is less than --sig.
- *DiffEnhRegions_qValue_Down.txt* file containing list of ROI, which are significantly enriched in taf2 as compared to that of taf1 with qValue is less than --sig.
- *VolcanoPlot_qValue.png* plot representing all significantly enriched ROI on the basis of q-value in taf1 and taf2 on right and left side in cyan color.

- *expDist_DiffEnhRegions_pValue_Up.pdf* boxplot representing distribution of expression in taf1 and taf2 datasets of target regions significantly enriched in taf1 dataset on the basis of p-value. This file is only generated on providing expression data using --expression.
- *expDist_DiffEnhRegions_pValue_Down.pdf* boxplot representing distribution of expression in taf1 and taf2 datasets of target regions significantly enriched in taf2 dataset on the basis of p-value. This file is only generated on providing expression data using --expression.
- *expDist_DiffEnhRegions_qValue_Up.pdf* boxplot representing distribution of expression in taf1 and taf2 datasets of target regions significantly enriched in taf1 dataset on the basis of q-value. This file is only generated on providing expression data using --expression.
- *expDist_DiffEnhRegions_qValue_Down.pdf* boxplot representing distribution of expression in taf1 and taf2 datasets of target regions significantly enriched in taf2 dataset on the basis of q-value. This file is only generated on providing expression data using --expression.

3.1.13.9. MDIFF

Program to identify differentially enriched regions between multiple datasets

Command line usage

usage: python MDIFF.py [options] roi taf

Mandatory arguments:

Roi	file with list of regions of interest to be analyzed (bed file format)

Taf	directory with target aligned files with replicates (bam file format).
Optional arguments:	
-h, --help	show this help message and exit.
--spikein	directory with spikein aligned files for each corresponding target aligned file (bam file format)
--expression	expression file (bed file format)
--hc	hierarchical clustering. (default: False)
--km	kmeans clustering. (default: False)
--cut	cut tree at height. Goes with --hc (default value 1.5).
--clusters	number of clusters. Goes with --km (default value 10).
--color	Heatmap color (default Blues). Allowed values:['Blues', 'BuGn', 'BuPu', 'GnBu', 'Greens', 'Greys', 'Oranges', 'OrRd', 'PuBu', 'PuBuGn', 'PuRd', 'Purples', 'RdPu', 'Reds', 'YlGn', 'YlGnBu', 'YlOrBr', 'YlOrRd'].
--discard	discard lower distribution of data. (default 0.10)
-p	level of signficance. (default 0.05)
-o, --output	output directory (default: Current Directory)
--test	statistical test to be used for analysis. Allowed values: Kruskal-Wallis/ANOVA (default: Kruskal-Wallis)

Output

This program generates following output files:

- *Results.txt* file containing list of all ROI along with their intensities in different datasets, p-value and bonferroni corrected p-values.

- zscore.txt file containing intensities transformed zscore for all list of ROI.
- *pValueBased_Diff_<hm/hc/km>.png* heatmap showing intensities transformed z-score for all differentially regulated regions filtered on the basis of p-value. File with suffix hm is generated when no clustering is turned on. Similarly file with suffix hc/km refers to the analysis subjected to either hierarchical or kmeans clustering.
- *qValueBased_Diff_<hm/hc/km>.png* heatmap showing intensities transformed z-score for all differentially regulated regions filtered on the basis of q-value. File with suffix hm is generated when no clustering is turned on. Similarly file with suffix hc/km refers to the analysis subjected to either hierarchical or kmeans clustering.
- *pValueBased_Diff_<hm/hc/km>_cluster_<cluster number>_<color represented in heatmap>.txt* these file(s) containing list of differentially regulated ROI on basis of p-value are generated only when either (hierarchical/kmeans) clustering options is turned on. Depending on clustering options this file will contain list of ROI being part of that particular selection.
- *qValueBased_Diff_<hm/hc/km>_cluster_<cluster number>_<color represented in heatmap>.txt* these file(s) containing list of differentially regulated ROI on basis of q-value are generated only when either (hierarchical/kmeans) clustering options is turned on. Depending on clustering options this file will contain list of ROI being part of that particular selection.
- *pValueBased_Diff_<hc/km>_cluster_<cluster number>_<color represented in heatmap>_expDist.pdf* this boxplot file is generated when clustering option is turned on and expression data is provided for the analysis. This represents distribution of target genes expression among different datasets (taf).

3.1.13.10. ABRI

Program to predict probabilistic dependencies between different datasets (in bed file format)

Command line usage

usage: python ABRI.py [options] roi datasets

Mandatory arguments:

roi	file with list of Regions Of Interest to be analyzed (bed file format)
datasets	directory with single/multiple files with regions of interest to be analyzed (bed file format)

Optional arguments:

-h, --help	show this help message and exit
--cgw	perform genome wide analysis. Provide genome file using –g
-g , --genome	genome file
--bs	bin Size. (default 500bp)
--stringency	stringency for overlap
--cutoff	cutoff for overlap (default 10 percent)
--boots	number of bootstraps. (default 100)
--size	datasize for each bootstrap. (default 0.90)
--strength	consider edges which occur in all networks (default 0.80)
-w , --white	white list
-b , --black	black list
--algo	learning algorithms (default Grow-Shrink). Allowed values: Hill-Climbing, Grow-Shrink, Incremental_Association, Fast_Incremental_Association, Interleaved_Incremental_Association

-o , --output output directory (default: execution directory)

Output

This program generates following output files:

- *data.txt* file containing information of presence or absence of different sets of annotated regions in each individual ROI.
- *arc_strength_direction.txt* file containing list of edges between two datasets along with its their strength and direction.
- *Network.pdf* file representing connectivity between different datasets with their strength fulfilled.

3.1.13.11. ABD

Program to predict probabilistic dependencies between different datasets (in aligned bam format)

Command line usage

usage: python ABD.py [options] genome datasets

Mandatory arguments:

roi file with list of Regions Of Interest to be analyzed (bed file format)

taf directory with target aligned files (bam file format)

Optional arguments:

-h, --help show this help message and exit

--caf directory with control aligned files (bam file format)

--spikein directory with spikein aligned files for each corresponding dataset (bam file format)

-g, --genome	genome file
--cgw	perform genome wide analysis. Provide genome file using –g
--bs	bin Size. (default 500bp)
--boots	number of bootstraps. (default 100)
--size	datasize for each bootstrap. (default 0.90)
--strength	consider edges which occur in all networks (default 0.80)
-w , --white	white list
-b , --black	black list
--algo	Learning algorithms (default Grow-Shrink). Allowed values: Hill-Climbing, Grow-Shrink, Incremental_Association, Fast_Incremental_Association, Interleaved_Incremental_Association
-o , --output	output directory (default: execution directory)
--qn	set values greater than quantile(x) to quantile(x). Allowed values between 0-1 (default value 0.99)

Output

This program generates following output files:

- *Profile.txt* matrix file quantifying different datasets (taf) in each individual ROI.
- *arc_strength_direction.txt* file containing list of edges between two datasets along with its their strength and direction.
- *Network.pdf* file representing connectivity between different datasets with their strength fulfilled.

3.1.13.12. VarSEL

Program to filter meaningful datasets best describing two sets of ROI.

Command line usage

usage: python VarSEL.py [options] genome class1 class2 taf

Mandatory arguments:

genome	genome file
class1	class 1 regions of interest (bed file format)
class2	class 2 regions of interest (bed file format)
taf	directory with target aligned files (bam file format)

Optional arguments:

-h, --help	show this help message and exit
--caf	directory with control aligned files (bam file format)
--spikein	directory with spikein aligned files for each corresponding dataset (bam file format)
--qn	set values greater than quantile(x) to quantile(x). Allowed values between 0-1 (default value 0.999)
--method	type of analysis (default CrossValidation). Allowed values:['Bootstrap', 'CrossValidation']
--number	either the number of folds or number of resampling iterations
-o, --output	output directory (default: execution directory)

Output

This program generates following output files:

- *Class1.txt* matrix file quantifying different datasets (taf) in each individual class1 ROI.
- *Class2.txt* matrix file quantifying different datasets (taf) in each individual class2 ROI.
- *Predictors.txt* file containing list of shortlisted datasets capable of explaining two sets of ROI.
- *Stats.txt* file containing other statistics.
- *Score.pdf* plot representing score on considering different size of datasets.

3.1.13.13. CLASS

Program to generate and apply classification model set of ROI on the basis of provided datasets.

Command line usage

usage: python CLASS.py [options] genome class1 class2 taf

Mandatory arguments:

genome	genome file
class1	class 1 regions of interest (bed file format)
class2	class 2 regions of interest (bed file format)
taf	directory with target aligned files (bam file format)

Optional arguments:

-h, --help	show this help message and exit
--caf	directory with control aligned files (bam file format)
--spikein	directory with spikein aligned files for each corresponding dataset (bam file format)

--test	test regions of interest to be classified (bed file format)
--sigma	sigma. (default value 1)
--cost	cost of constriant violation. (default value 100)
--kcross	k-means cross validation. (default value 10)
--kernal	type of kernal to be used for classification (default Linear). Allowed values:['Radial_Basis', 'Linear', 'Laplacian']
--ctype	classification type (default C classification). Allowed values:['C classification', 'nu classification']
--type	type of analysis (default Train). Allowed values:['Train', 'TrainPredict']
--method	classification method (default Support_Vector_Machine). Allowed values:['Support_Vector_Machine']
--qn	set values greater than quantile(x) to quantile(x). Allowed values between 0-1 (default value 0.999)
-o, --output	output directory (default: execution directory)

Output

This program generates following output files:

- *Class1.txt* matrix file quantifying different datasets (taf) in each individual class1 ROI.
- *Class2.txt* matrix file quantifying different datasets (taf) in each individual class2 ROI.
- *ROC.pdf* file depicting area under ROC.

3.1.13.14. MatHM

Program to generate heatmap.

Command line usage

usage: python VarSEL.py [options] genome class1 class2 taf

Mandatory arguments:

dmf	directory with matrix file(s) with .txt extension
nsamples	number of samples

Optional arguments:

-h, --help	show this help message and exit
-e, --expression	expression file (bed file format). This option should be supported with ROI (bed file format with .bed extension) for each matrix file with same filename.
--hc	hierarchical clustering. (default: False)
--km	kmeans clustering. (default: False)
--cut	cut tree at height. Goes with --hc (default value 1.5)
--clusters	number of clusters. Goes with --km (default value 10)
--qn	set values greater than quantile x to x. Allowed values between 0-1 (default value 0.99)
--sortd	sort on the basis of intensities of dataset (default value 0).
--rownames	data contains row names (default: False)
--colnames	data contains coloumn names (default: False)
--zscore	transform data to z-score (default: False)
--gs	global sample scaling (default: False)
--ss	individual sample scaling (default: False)

--color	heatmap color (default Blues). Allowed values:['Blues', 'BuGn', 'BuPu', 'GnBu', 'Greens', 'Greys', 'Oranges', 'OrRd', 'PuBu', 'PuBuGn', 'PuRd', 'Purples', 'RdPu', 'Reds', 'YlGn', 'YlGnBu', 'YlOrBr', 'YlOrRd', 'BrBG', 'PiYG', 'PRGn', 'PuOr', 'RdBu', 'RdGy', 'RdYlBu', 'RdYlGn', 'Spectral']
-o, --output	output directory (default: execution directory)
--height	image height (default: 7)
--width	image width (default: 7)
--res	image resolution (default: 300)

Output

This program generates following output files:

- *<dmf filename>_<hm/hc/km>.png* heatmap showing intensities of different datasets in each individual matrix file (dmf). File with suffix hm is generated when no clustering is turned on. Similarly file with suffix hc/km refers to the analysis subjected to either hierarchical or kmeans clustering.
- *<dmf filename>_<hc/km>_cluster_<cluster number>_<color represented in heatmap>.txt* these file(s) are generated only when either (hierarchical/kmeans) clustering options is turned on. Depending on clustering options this file will contain list of records being part of that particular selection.
- *<dmf filename>_expDist_<hc/km>.pdf* this boxplot file is generated when clustering option is turned on and expression data is provided for the analysis. This represents distribution of target genes expression among different clusters.

3.1.13.15. ExtBAM

Program for extending aligned reads

Command line usage

usage: python ExtBAM.py [options] taf

Mandatory arguments:

taf	directory with aligned files (bam file format)

Optional arguments:

-h, --help	show this help message and exit
--extend	number of base pairs by which each should be extended. (default value 200bp).
-o, --output	output directory (default: execution directory)

Output

This program generates following output files:

- *<taf filename>_ext.bam* extended aligned reads in bam file.

3.2. Data Analysis for characterizing polycomb dependent methylation forms

3.2.1. ChIP sequencing data analysis.

Sequencing data generated from the Illumina platforms related to the second project described, were aligned to mouse reference genome (mm9) using Bowtie version 0.12.7 (Langmead et al., 2009). Only reads with unique alignment were retained for downstream analysis. Peak calling and bigWig files were generated using MACS version 1.4 (Zhang et al., 2008). Only peaks with 10x-Log p-value ≥70 are considered

for further processing. bigWig files were visualized using the UCSC browser (http://genome.ucsc.edu). The list of mm9 annotated RefSeq genes used for the different analysis was downloaded from the UCSC database. Intragenic reads density for histone H3, H3K27me1, H3K27me2 and H3K36me3 were determined by computing the aligned reads within each RefSeq genes normalized for sequencing depth. PTM enrichments relative to histone H3 density were determined for each gene as the –Log10p-value computed using a chi-square test (PTM vs. H3) and adjusted using Bonferroni correction. The corrected p-values between different PTMs were compared using Pearson correlation test. Genome wide correlation among H3K27me1, H3K27me2 and H3K36me3 modifications with the read intensities in gene bodies was computed using PCA method in R factorMineR (http://factominer.free.fr/) package. TSS vs. non-TSS location of H3K27ac peaks was determined by overlapping H3K27ac peaks with a 5 kb region centered on TSS for each mm9 RefSeq annotated gene

Each H3K27ac KO distal peak was assigned to the closest TSS RefSeq gene. These genes were then classified accordingly to their expression levels between WT and *Eed* KO and classified as up regulated (FC > 1.5) or down-regulated (FC < -1.5). For the genes belonging to each class as well as in the entire RNA-seq dataset, we determined if the observed frequencies of up-regulated and down-regulated genes under putative control of the H3K27ac distal peaks were significantly different respect to the expected frequencies determined by analyzing the whole RNA-seq dataset. Accordingly, we determined the relative distance of each H3K27ac distal peak identified in either WT or *Eed* KO samples respect to the closest up-regulated gene in Eed KO ES cells.

Active enhancers were classified on the basis of presence of both H3K27ac and H3K4me1 peaks, the absence of H3K4me3 and a minimal distance of 2.5Kb from annotated TSSs. Poised enhancers were defined by the absence of H3K27ac using the same criteria. The relative intensities of all the indicated histone marks were determined at H3K27me3 positive promoters, at poised and at active enhancers in mESC.

3.2.2. RNA sequencing data analysis.

RNA-seq data generated for ES WT, ES *Eed* KO, Ebs WT, Ebs *Eed* KO samples were aligned to mouse reference genome using TopHat (Trapnell et al., 2009). Differentially expressed genes were identified with cuffdiff (Trapnell et al., 2010). Microarray raw data were retrieved from the Gene Omnibus Database (http://www.ncbi.nlm.nih.gov/geo/) at the accession number GSE19076 and were processed using affy (Gautier et al., 2004) package in R.

Chapter 4 - DISCUSSION

4.1. ChIP_QC

The fast development of NGS technologies has radically changed the experimental approaches in “wet labs” leading to the generation of a surplus of high quality data, which are also available as public resources. Due to this increasing availability of data both in terms of size and complexity, we need a platform with efficient analytical methods. Aiming this, we designed ChIP_QC and showed with what flexibility it can be applied for different epigenomic studies. This development is an attempt to open new window for high throughput data analysis, where we provide platform with methods, which are at most helpful for genome wide studies. The uniqueness of the program lies in handling and analyzing the changes within and/or across multiple samples against different datasets and their flexible linkage to expression data. Each module of the program generates all necessary results, different plots of good resolution and many other supplementary files. Supplementary files can be helpful for further downstream analysis, which can be used as input to other modules of the ChIP_QC program, thus increasing its flexibility without any major computational skills. Depending on the module, the program offers some additional features for enhancing the results. This includes, option for smoothing the data, making use of strand information for analysis, plotting confidence intervals, selecting color for heatmaps etc. Data are kept in a format that allows its usage by a wide range of users. This program comes with both graphical and command line utility allowing its usage by all type of users mainly experimental biologists with very minimal computational background. It can be executed in both Mac and Linux operating systems.

Taking advantage of publicly available human ENCODE (Consortium, 2012) datasets and analyzing them using ChIP_QC, we have cross-verified some known observations to show the power and accuracy of ChIP_QC and at the same time we have generated some novel findings such as the preferential association of the Bcl11a transcription factor at active enhancers with respect to that of promoters and the association between Suz12 and Ctbp2 in chromatin compact regions of human embryonic stem cells.

We have specifically chosen to restrict the program on tertiary analysis. We tried to avoid redundancy of running alignment, peaking calling and others. These steps are now very much standardized and many pipelines have been well established for doing such tasks. Apart from this many sequencing facilities by default provide support for both primary and secondary analysis. Most crucial step is to handle further downstream analysis on the basis of experimental design. Except very few, we didn't see many programs, which are capable for performing comprehensive genome wide analysis with multiple datasets as this program does. For instance, both HOMER's (Heinz et al., 2010) ChIP-seq functions, seqMINER (Ye et al., 2011) perform quantifications within ROI, but doesn't support multiple class of ROI, performs only kmeans clustering, lacks facility to link results with expression data. Apart from such advantages, this program is first of its kind, which provides much wider scope for performing genome wide analysis with new approaches. Table3 list out advantages and disadvantages of ChIP_QC in comparison with other GUI and command line based applications. Here we compared ChIP_QC with seqMINER (Ye et al., 2011), HOMER (Heinz et al., 2010), ChIPSeeqer (Giannopoulou and Elemento, 2011), CIstrome (Liu et al., 2011), macs2 bdgdiff (https://github.com/taoliu/MACS) and diffReps (Shen et al., 2013). In this version of ChIP_QC, we aimed at analyzing datasets generated through

ChIP-seq or related approach, RNA-seq and annotated datasets. In coming versions, we want to integrate other statistical methods for analysis like incorporating limma for differential analysis, support other classification methods and extend its capabilities to integrate and support methylation and nucleosome-based studies.

	seqMINER	HOMER	ChIPseeqer	Cistrome	ChIP_QC
User Interface	**GUI**	**CL**	**GUI**	**WEB GUI**	**GUI/CL**
Standard Analyses	**seqMINER**	**HOMER**	**ChIPseeqer**	**Cistrome**	**ChIP_QC**
Peak Calling	FALSE	TRUE	TRUE	TRUE	FALSE
Gene Ontology	FALSE	FALSE	TRUE	FALSE	FALSE
Motif Analysis	FALSE	TRUE	TRUE	TRUE	FALSE
Peaks based analysis	**seqMINER**	**HOMER**	**ChIPseeqer**	**Cistrome**	**ChIP_QC**
Enrichment of different samples in ROI	FALSE	FALSE	FALSE	FALSE	TRUE
Coexistence of different samples in ROI	FALSE	FALSE	FALSE	FALSE	TRUE
Correlation based on peaks	FALSE	FALSE	TRUE	FALSE	FALSE
Predeicting dependencies based on peaks	FALSE	FALSE	FALSE	FALSE	TRUE
Introducing and analyzing random regions	FALSE	FALSE	FALSE	FALSE	TRUE
Quantification Based Analysis	**seqMINER**	**HOMER**	**ChIPseeqer**	**Cistrome**	**ChIP_QC**
Number of samples can be processed in a run	1	1	NA	>=1 (applicable for few cases)	>=1 (applicable for all cases)
Works with raw data (no processing required)	TRUE	FALSE	FALSE	FALSE	TRUE
Signal within ROI	FALSE	FALSE	FALSE	FALSE	TRUE
Spread of Signal around ROI	TRUE	TRUE	FALSE	TRUE	TRUE
Profile generator	TRUE	TRUE	FALSE	TRUE	TRUE
Genome wide/ROI specific correlation	FALSE	FALSE	FALSE	TRUE	TRUE
Introducing and analyzing random regions	FALSE	FALSE	FALSE	FALSE	TRUE
provision of kmeans clustering on quantified data	TRUE	Not in build	FALSE	TRUE	TRUE
provision of hierarchical clustering on quantified data	FALSE	Not in build	FALSE	FALSE	TRUE
Corelating clustered data with expression	FALSE	FALSE	FALSE	FALSE	TRUE
SpiKein normalization	FALSE	FALSE	FALSE	FALSE	TRUE
Predeicting dependencies based on peaks	FALSE	FALSE	FALSE	FALSE	TRUE
Filtering datasets on the basis of their importance wrt ROI	FALSE	FALSE	FALSE	FALSE	TRUE
Classification studies	FALSE	FALSE	FALSE	FALSE	TRUE

	macs2 bdgdiff	diffReps	EpiMINE
User Interface	**CL**	**CL**	**GUI/CL**
Differential Studies	**macs2 bdgdiff**	**diffReps**	**EpiMINE**
Identifying differential regions between two conditions without replicates	FALSE	TRUE	TRUE
Identifying differential regions between two conditions with replicates	TRUE	TRUE	TRUE
Identifying differential regions between more than two conditions with replicates	FALSE	FALSE	TRUE
Corelating differentially enriched regions with expression	FALSE	FALSE	TRUE

Legend

GUI	Graphical User Interface
CL	Command line
TRUE	Feature is available in tool
FALSE	Feature is not available in tool
NA	Feature is not applicable for tool
Not in build	Feature should be excuted outside the tool

Table3: Comparison of features of ChIP_QC with other GUI and command line tools

4.2. Polycomb dependent H3K27me1 and H3K27me2 regulate active transcription and enhancer fidelity.

On the basis MS data we showed here that most abundant PTM on H3K27 is methylation accounting for 80% of total histone and all three forms of methylations are PRC2 dependent. In our study we used genome wide approaches and showed that all three forms of methylation accumulate in mutual exclusive manner through out the genome. Monomethylation is enriched in intragenic regions and this behavior seems to be conserved in other species, this was already reported in correlation with human $CD4^+$ T cells (Barski et al., 2007). Reduced expression of H3K27me1 enriched genes in PRC2 deficient mESC, and its strong positive correlation with H3K36me3, strongly suggests importance of PRC2 activity for proper expression. Various mechanisms underlying this correlation can be envisaged. For instance, both H3K27me1 and H3K36me3 modifications deposited in intragenic regions of active genes could lead to nucleosome mobility, or H3K27me1 could be involved in the RNA splicing-dependent recruitment of Setd2 (de Almeida et al., 2011), thus acting as a permissive modification for elongation or splicing, while H3K27me2 could inhibit this process.

We also show that upon loss of PRC2 activity in mESC global levels of H3K27ac are increased suggesting that these conditions favor histone acetyltransferases (HATs) to access to chromatin. H3K27ac is regarded as a marker of active enhancer and discriminates it from poised enhancer (Creyghton et al., 2010a). Thus, in our work we showed that loss of PRC2 activity triggers activation of poised enhancers, possibly through greater accessibility of chromatin to HATs; this makes possible reasoning that the broad unspecific deposition of H3K27me2 protects H3K27 from HAT activity. This logic is supported by the rapid accumulation of H3K27me2 at sites that lose

H3K27ac upon inhibition of global HAT activity. We also observed that, 60% of newly marked H3K27ac occurs at regions that were already marked by H3K4me1, suggesting a possible link between HATs recruitment and H3K4me1. Aberrantly activated enhancers upon PRC2 loss could contribute to the several defects in development and lineage specification. In general Figure 4.2 summarizes the model that we propose.

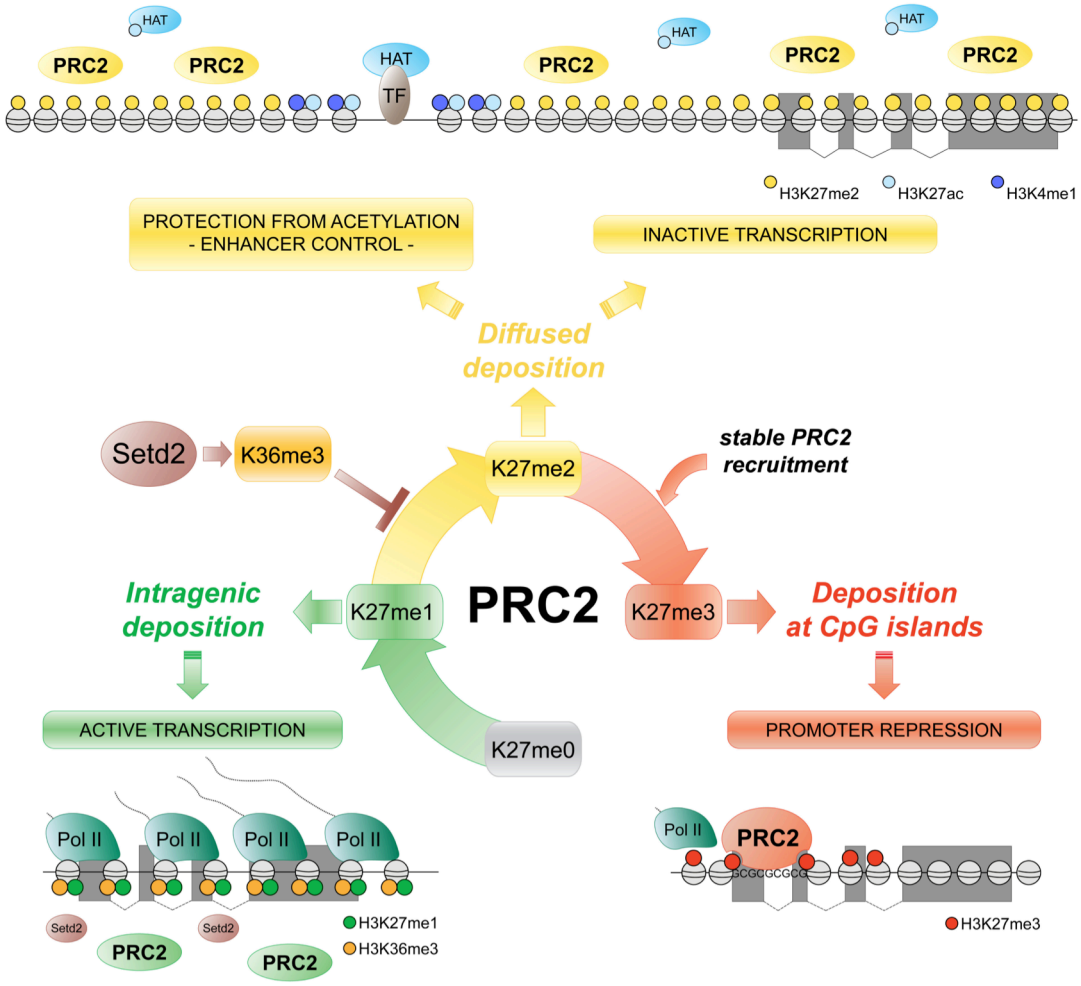


Fig. 4.2. Our proposed model on different functionalities of PRC2 dependent methylation forms.

From our results further questions about the roles of methylation forms of H3K27, and mechanisms underlying their deposition arise. As mentioned above about the

conserved phenomena of H3K27me1 enrichment at intragenic regions of transcribed genes seems; it would also be interesting to know if this mutual exclusivity between different forms of methylation is maintained in other cellular models, terminal differentiated cells, and see if there are any differences in cancer cells too.

REFERENCES

Allfrey, V.G., Faulkner, R., and Mirsky, A.E. (1964). Acetylation and Methylation of Histones and Their Possible Role in the Regulation of Rna Synthesis. Proceedings of the National Academy of Sciences of the United States of America *51*, 786-794.

Ang, Y.S., Tsai, S.Y., Lee, D.F., Monk, J., Su, J., Ratnakumar, K., Ding, J., Ge, Y., Darr, H., Chang, B.*, et al.* (2011). Wdr5 mediates self-renewal and reprogramming via the embryonic stem cell core transcriptional network. Cell *145*, 183-197.

Arnaudo, A.M., and Garcia, B.A. (2013). Proteomic characterization of novel histone post-translational modifications. Epigenetics & chromatin *6*, 24.

Barrero, M.J., and Izpisua Belmonte, J.C. (2013). Polycomb complex recruitment in pluripotent stem cells. Nature cell biology *15*, 348-350.

Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. Cell *129*, 823-837.

Bernstein, B.E., Kamal, M., Lindblad-Toh, K., Bekiranov, S., Bailey, D.K., Huebert, D.J., McMahon, S., Karlsson, E.K., Kulbokas, E.J., 3rd, Gingeras, T.R.*, et al.* (2005). Genomic maps and comparative analysis of histone modifications in human and mouse. Cell *120*, 169-181.

Blackledge, N.P., Farcas, A.M., Kondo, T., King, H.W., McGouran, J.F., Hanssen, L.L., Ito, S., Cooper, S., Kondo, K., Koseki, Y.*, et al.* (2014). Variant PRC1 complex-dependent H2A ubiquitylation drives PRC2 recruitment and polycomb domain formation. Cell *157*, 1445-1459.

Blahnik, K.R., Dou, L., O'Geen, H., McPhillips, T., Xu, X., Cao, A.R., Iyengar, S., Nicolet, C.M., Ludascher, B., Korf, I.*, et al.* (2010). Sole-Search: an integrated analysis program for peak detection and functional annotation using ChIP-seq data. Nucleic acids research *38*, e13.

Boyer, L.A., Plath, K., Zeitlinger, J., Brambrink, T., Medeiros, L.A., Lee, T.I., Levine, S.S., Wernig, M., Tajonar, A., Ray, M.K.*, et al.* (2006). Polycomb complexes repress developmental regulators in murine embryonic stem cells. Nature *441*, 349-353.

Boyle, A.P., Guinney, J., Crawford, G.E., and Furey, T.S. (2008). F-Seq: a feature density estimator for high-throughput sequence tags. Bioinformatics *24*, 2537-2538.

Bracken, A.P., Dietrich, N., Pasini, D., Hansen, K.H., and Helin, K. (2006). Genome-wide mapping of Polycomb target genes unravels their roles in cell fate transitions. Genes & development *20*, 1123-1136.

Breiling, A., Turner, B.M., Bianchi, M.E., and Orlando, V. (2001). General transcription factors bind promoters repressed by Polycomb group proteins. Nature *412*, 651-655.

Brookes, E., de Santiago, I., Hebenstreit, D., Morris, K.J., Carroll, T., Xie, S.Q., Stock, J.K., Heidemann, M., Eick, D., Nozaki, N.*, et al.* (2012). Polycomb associates genome-wide with a specific RNA polymerase II variant, and regulates metabolic genes in ESCs. Cell Stem Cell *10*, 157-170.

Brunk, B.P., Martin, E.C., and Adler, P.N. (1991). Drosophila genes Posterior Sex Combs and Suppressor two of zeste encode proteins with homology to the murine bmi-1 oncogene. Nature *353*, 351-353.

Cao, R., Tsukada, Y., and Zhang, Y. (2005). Role of Bmi-1 and Ring1A in H2A ubiquitylation and Hox gene silencing. Molecular cell *20*, 845-854.

Cao, R., Wang, L., Wang, H., Xia, L., Erdjument-Bromage, H., Tempst, P., Jones, R.S., and Zhang, Y. (2002). Role of histone H3 lysine 27 methylation in Polycomb-group silencing. Science *298*, 1039-1043.

Cao, R., and Zhang, Y. (2004). SUZ12 is required for both the histone methyltransferase activity and the silencing function of the EED-EZH2 complex. Molecular cell *15*, 57-67.

Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V.B., Wong, E., Orlov, Y.L., Zhang, W., Jiang, J.*, et al.* (2008). Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell *133*, 1106-1117.

Chen, Y., Sprung, R., Tang, Y., Ball, H., Sangras, B., Kim, S.C., Falck, J.R., Peng, J., Gu, W., and Zhao, Y. (2007). Lysine propionylation and butyrylation are novel post-translational modifications in histones. Molecular & cellular proteomics : MCP *6*, 812-819.

Consortium, E.P. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57-74.

Cosgrove, M.S., Boeke, J.D., and Wolberger, C. (2004). Regulated nucleosome mobility and the histone code. Nature structural & molecular biology *11*, 1037-1043.

Creyghton, M.P., Cheng, A.W., Welstead, G.G., Kooistra, T., Carey, B.W., Steine, E.J., Hanna, J., Lodato, M.A., Frampton, G.M., Sharp, P.A.*, et al.* (2010a). Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proceedings of the National Academy of Sciences of the United States of America *107*, 21931-21936.

Creyghton, M.P., Cheng, A.W., Welstead, G.G., Kooistra, T., Carey, B.W., Steine, E.J., Hanna, J., Lodato, M.A., Frampton, G.M., Sharp, P.A.*, et al.* (2010b). Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proceedings of the National Academy of Sciences of the United States of America *107*, 21931-21936.

Czermin, B., Melfi, R., McCabe, D., Seitz, V., Imhof, A., and Pirrotta, V. (2002). Drosophila enhancer of Zeste/ESC complexes have a histone H3 methyltransferase activity that marks chromosomal Polycomb sites. Cell *111*, 185-196.

de Almeida, S.F., Grosso, A.R., Koch, F., Fenouil, R., Carvalho, S., Andrade, J., Levezinho, H., Gut, M., Eick, D., Gut, I.*, et al.* (2011). Splicing enhances recruitment of methyltransferase HYPB/Setd2 and methylation of histone H3 Lys36. Nature structural & molecular biology *18*, 977-983.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15-21.

Elderkin, S., Maertens, G.N., Endoh, M., Mallery, D.L., Morrice, N., Koseki, H., Peters, G., Brockdorff, N., and Hiom, K. (2007). A phosphorylated form of Mel-18 targets the Ring1B histone H2A ubiquitin ligase to chromatin. Molecular cell *28*, 107-120.

Faust, C., Schumacher, A., Holdener, B., and Magnuson, T. (1995). The eed mutation disrupts anterior mesoderm production in mice. Development *121*, 273-285.

Ferrari, K.J., Scelfo, A., Jammula, S., Cuomo, A., Barozzi, I., Stutzer, A., Fischle, W., Bonaldi, T., and Pasini, D. (2014). Polycomb-dependent H3K27me1 and H3K27me2 regulate active transcription and enhancer fidelity. Molecular cell *53*, 49-62.

Fragola, G., Germain, P.L., Laise, P., Cuomo, A., Blasimme, A., Gross, F., Signaroldi, E., Bucci, G., Sommer, C., Pruneri, G.*, et al.* (2013). Cell reprogramming requires silencing of a core subset of polycomb targets. PLoS Genet *9*, e1003292.

Francis, N.J., Saurin, A.J., Shao, Z., and Kingston, R.E. (2001). Reconstitution of a functional core polycomb repressive complex. Molecular cell *8*, 545-556.

Gao, Z., Zhang, J., Bonasio, R., Strino, F., Sawai, A., Parisi, F., Kluger, Y., and Reinberg, D. (2012). PCGF homologs, CBX proteins, and RYBP define functionally distinct PRC1 family complexes. Molecular cell *45*, 344-356.

Gautier, L., Cope, L., Bolstad, B.M., and Irizarry, R.A. (2004). affy--analysis of Affymetrix GeneChip data at the probe level. Bioinformatics *20*, 307-315.

Giannopoulou, E.G., and Elemento, O. (2011). An integrated ChIP-seq analysis platform with customizable workflows. BMC bioinformatics *12*, 277.

Goecks, J., Nekrutenko, A., Taylor, J., and Galaxy, T. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome biology *11*, R86.

Goldberg, A.D., Banaszynski, L.A., Noh, K.M., Lewis, P.W., Elsaesser, S.J., Stadler, S., Dewell, S., Law, M., Guo, X., Li, X.*, et al.* (2010). Distinct factors control histone variant H3.3 localization at specific genomic regions. Cell *140*, 678-691.

Hansen, K.H., Bracken, A.P., Pasini, D., Dietrich, N., Gehani, S.S., Monrad, A., Rappsilber, J., Lerdrup, M., and Helin, K. (2008). A model for transmission of the H3K27me3 epigenetic mark. Nature cell biology *10*, 1291-1300.

Haupt, Y., Alexander, W.S., Barri, G., Klinken, S.P., and Adams, J.M. (1991). Novel zinc finger gene implicated as myc collaborator by retrovirally accelerated lymphomagenesis in E mu-myc transgenic mice. Cell *65*, 753-763.

He, J., Shen, L., Wan, M., Taranova, O., Wu, H., and Zhang, Y. (2013). Kdm2b maintains murine embryonic stem cell status by recruiting PRC1 complex to CpG islands of developmental genes. Nature cell biology *15*, 373-384.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Molecular cell *38*, 576-589.

Homer, N., Merriman, B., and Nelson, S.F. (2009). BFAST: an alignment tool for large scale genome resequencing. PLoS One *4*, e7767.

Huang, D.W., Sherman, B.T., and Lempicki, R.A. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res *37*, 1-13.

Jean, G., Kahles, A., Sreedharan, V.T., De Bona, F., and Ratsch, G. (2010). RNA-Seq read alignments with PALMapper. Curr Protoc Bioinformatics *Chapter 11*, Unit 11 16.

Johnson, D.S., Mortazavi, A., Myers, R.M., and Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. Science *316*, 1497-1502.

Jung, H.R., Pasini, D., Helin, K., and Jensen, O.N. (2010). Quantitative mass spectrometry of histones H3.2 and H3.3 in Suz12-deficient mouse embryonic stem cells reveals distinct, dynamic post-translational modifications at Lys-27 and Lys-36. Molecular & cellular proteomics : MCP *9*, 838-850.

Kassis, J.A., and Brown, J.L. (2013). Polycomb group response elements in Drosophila and vertebrates. Adv Genet *81*, 83-118.

Ketel, C.S., Andersen, E.F., Vargas, M.L., Suh, J., Strome, S., and Simon, J.A. (2005). Subunit contributions to histone methyltransferase activities of fly and worm polycomb group complexes. Mol Cell Biol *25*, 6857-6868.

Kizer, K.O., Phatnani, H.P., Shibata, Y., Hall, H., Greenleaf, A.L., and Strahl, B.D. (2005). A novel domain in Set2 mediates RNA polymerase II interaction and couples histone H3 K36 methylation with transcript elongation. Mol Cell Biol *25*, 3305-3316.

Ku, M., Koche, R.P., Rheinbay, E., Mendenhall, E.M., Endoh, M., Mikkelsen, T.S., Presser, A., Nusbaum, C., Xie, X.H., Chi, A.S.*, et al.* (2008). Genomewide Analysis of PRC1 and PRC2 Occupancy Identifies Two Classes of Bivalent Domains. Plos Genet *4*.

Kuzmichev, A., Nishioka, K., Erdjument-Bromage, H., Tempst, P., and Reinberg, D. (2002). Histone methyltransferase activity associated with a human multiprotein complex containing the Enhancer of Zeste protein. Genes & development *16*, 2893-2905.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome biology *10*, R25.

Laugesen, A., and Helin, K. (2014). Chromatin Repressive Complexes in Stem Cells, Development, and Cancer. Cell Stem Cell *14*, 735-751.

Layer, R.M., Skadron, K., Robins, G., Hall, I.M., and Quinlan, A.R. (2013). Binary Interval Search: a scalable algorithm for counting interval intersections. Bioinformatics *29*, 1-7.

Lee, T.I., Jenner, R.G., Boyer, L.A., Guenther, M.G., Levine, S.S., Kumar, R.M., Chevalier, B., Johnstone, S.E., Cole, M.F., Isono, K.*, et al.* (2006). Control of developmental regulators by Polycomb in human embryonic stem cells. Cell *125*, 301-313.

Leeb, M., Pasini, D., Novatchkova, M., Jaritz, M., Helin, K., and Wutz, A. (2010). Polycomb complexes act redundantly to repress genomic repeats and genes. Genes & development *24*, 265-276.

Levine, S.S., Weiss, A., Erdjument-Bromage, H., Shao, Z., Tempst, P., and Kingston, R.E. (2002). The core of the polycomb repressive complex is compositionally and functionally conserved in flies and humans. Mol Cell Biol *22*, 6070-6078.

Lewis, E.B. (1978). A gene complex controlling segmentation in Drosophila. Nature *276*, 565-570.

Li, B., Howe, L., Anderson, S., Yates, J.R., 3rd, and Workman, J.L. (2003). The Set2 histone methyltransferase functions through the phosphorylated carboxyl-terminal domain of RNA polymerase II. The Journal of biological chemistry *278*, 8897-8903.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754-1760.

Li, H., Ruan, J., and Durbin, R. (2008a). Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome research *18*, 1851-1858.

Li, J., Moazed, D., and Gygi, S.P. (2002a). Association of the histone methyltransferase Set2 with RNA polymerase II plays a role in transcription elongation. The Journal of biological chemistry *277*, 49383-49388.

Li, J.X., Moazed, D., and Gygi, S.P. (2002b). Association of the histone methyltransferase Set2 with RNA polymerase II plays a role in transcription elongation. J Biol Chem *277*, 49383-49388.

Li, R., Li, Y., Kristiansen, K., and Wang, J. (2008b). SOAP: short oligonucleotide alignment program. Bioinformatics *24*, 713-714.

Liu, T., Ortiz, J.A., Taing, L., Meyer, C.A., Lee, B., Zhang, Y., Shin, H., Wong, S.S., Ma, J., Lei, Y.*, et al.* (2011). Cistrome: an integrative platform for transcriptional regulation studies. Genome biology *12*, R83.

Lo, C.C., and Chain, P.S. (2014). Rapid evaluation and quality control of next generation sequencing data with FaQCs. BMC bioinformatics *15*, 366.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome biology *15*, 550.

Luger, K. (2003). Structure and dynamic behavior of nucleosomes. Current opinion in genetics & development *13*, 127-135.

Luger, K., Mader, A.W., Richmond, R.K., Sargent, D.F., and Richmond, T.J. (1997). Crystal structure of the nucleosome core particle at 2.8 A resolution. Nature *389*, 251-260.

Lun, D.S., Sherrid, A., Weiner, B., Sherman, D.R., and Galagan, J.E. (2009). A blind deconvolution approach to high-resolution mapping of transcription factor binding sites from ChIP-seq data. Genome biology *10*, R142.

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

Margueron, R., Justin, N., Ohno, K., Sharpe, M.L., Son, J., Drury, W.J., 3rd, Voigt, P., Martin, S.R., Taylor, W.R., De Marco, V.*, et al.* (2009). Role of the polycomb protein EED in the propagation of repressive histone marks. Nature *461*, 762-767.

Margueron, R., and Reinberg, D. (2011). The Polycomb complex PRC2 and its mark in life. Nature *469*, 343-349.

McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., Wenger, A.M., and Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol *28*, 495-501.

Medina-Rivera, A., Defrance, M., Sand, O., Herrmann, C., Castro-Mondragon, J.A., Delerce, J., Jaeger, S., Blanchet, C., Vincens, P., Caron, C.*, et al.* (2015). RSAT 2015: Regulatory Sequence Analysis Tools. Nucleic acids research *43*, W50-56.

Mikkelsen, T.S., Ku, M., Jaffe, D.B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.K., Koche, R.P.*, et al.* (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature *448*, 553-560.

Min, I.M., Waterfall, J.J., Core, L.J., Munroe, R.J., Schimenti, J., and Lis, J.T. (2011). Regulating RNA polymerase pausing and transcription elongation in embryonic stem cells. Genes & development *25*, 742-754.

Mohn, F., Weber, M., Rebhan, M., Roloff, T.C., Richter, J., Stadler, M.B., Bibel, M., and Schubeler, D. (2008). Lineage-specific polycomb targets and de novo DNA methylation define restriction and potential of neuronal progenitors. Molecular cell *30*, 755-766.

Morey, L., Aloia, L., Cozzuto, L., Benitah, S.A., and Di Croce, L. (2013). RYBP and Cbx7 define specific biological functions of polycomb complexes in mouse embryonic stem cells. Cell reports *3*, 60-69.

Morey, L., and Helin, K. (2010). Polycomb group protein-mediated repression of transcription. Trends in biochemical sciences *35*, 323-332.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods *5*, 621-628.

Needleman, S.B., and Wunsch, C.D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol *48*, 443-453.

Nix, D.A., Courdy, S.J., and Boucher, K.M. (2008). Empirical methods for controlling false positives and estimating confidence in ChIP-Seq peaks. BMC bioinformatics *9*, 523.

O'Carroll, D., Erhardt, S., Pagani, M., Barton, S.C., Surani, M.A., and Jenuwein, T. (2001). The polycomb-group gene Ezh2 is required for early mouse development. Mol Cell Biol *21*, 4330-4336.

Onder, T.T., Kara, N., Cherry, A., Sinha, A.U., Zhu, N., Bernt, K.M., Cahan, P., Marcarci, B.O., Unternaehrer, J., Gupta, P.B.*, et al.* (2012). Chromatin-modifying enzymes as modulators of reprogramming. Nature *483*, 598-602.

Orlando, D.A., Chen, M.W., Brown, V.E., Solanki, S., Choi, Y.J., Olson, E.R., Fritz, C.C., Bradner, J.E., and Guenther, M.G. (2014). Quantitative ChIP-Seq normalization reveals global modulation of the epigenome. Cell reports *9*, 1163-1170.

Orlando, V., and Paro, R. (1993). Mapping Polycomb-repressed domains in the bithorax complex using in vivo formaldehyde cross-linked chromatin. Cell *75*, 1187-1198.

Pasini, D., Bracken, A.P., Jensen, M.R., Lazzerini Denchi, E., and Helin, K. (2004a). Suz12 is essential for mouse development and for EZH2 histone methyltransferase activity. Embo J *23*, 4061-4071.

Pasini, D., Bracken, A.P., Jensen, M.R., Lazzerini Denchi, E., and Helin, K. (2004b). Suz12 is essential for mouse development and for EZH2 histone methyltransferase activity. The EMBO journal *23*, 4061-4071.

Pasini, D., Cloos, P.A., Walfridsson, J., Olsson, L., Bukowski, J.P., Johansen, J.V., Bak, M., Tommerup, N., Rappsilber, J., and Helin, K. (2010). JARID2 regulates binding of the Polycomb repressive complex 2 to target genes in ES cells. Nature *464*, 306-310.

Patel, R.K., and Jain, M. (2012). NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. PLoS One *7*, e30619.

Pereira, C.F., Piccolo, F.M., Tsubouchi, T., Sauer, S., Ryan, N.K., Bruno, L., Landeira, D., Santos, J., Banito, A., Gil, J.*, et al.* (2010). ESCs require PRC2 to direct the successful reprogramming of differentiated cells toward pluripotency. Cell Stem Cell *6*, 547-556.

Piunti, A., and Pasini, D. (2011). Epigenetic factors in cancer development: polycomb group proteins. Future oncology *7*, 57-75.

Qin, J., Whyte, W.A., Anderssen, E., Apostolou, E., Chen, H.H., Akbarian, S., Bronson, R.T., Hochedlinger, K., Ramaswamy, S., Young, R.A.*, et al.* (2012). The polycomb group protein L3mbtl2 assembles an atypical PRC1-family complex that is essential in pluripotent stem cells and early development. Cell Stem Cell *11*, 319-332.

Qin, Z.S., Yu, J., Shen, J., Maher, C.A., Hu, M., Kalyana-Sundaram, S., and Chinnaiyan, A.M. (2010). HPeak: an HMM-based algorithm for defining read-enriched regions in ChIP-Seq data. BMC bioinformatics *11*, 369.

Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S.A., Flynn, R.A., and Wysocka, J. (2011). A unique chromatin signature uncovers early developmental enhancers in humans. Nature *470*, 279-283.

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic acids research *43*, e47.

Ross-Innes, C.S., Stark, R., Teschendorff, A.E., Holmes, K.A., Ali, H.R., Dunning, M.J., Brown, G.D., Gojis, O., Ellis, I.O., Green, A.R.*, et al.* (2012). Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. Nature *481*, 389-393.

Rozowsky, J., Euskirchen, G., Auerbach, R.K., Zhang, Z.D., Gibson, T., Bjornson, R., Carriero, N., Snyder, M., and Gerstein, M.B. (2009). PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. Nat Biotechnol *27*, 66-75.

Schmieder, R., and Edwards, R. (2011). Fast identification and removal of sequence contamination from genomic and metagenomic datasets. PLoS One *6*, e17288.

Schumacher, A., and Magnuson, T. (1997). Murine Polycomb- and trithorax-group genes regulate homeotic pathways and beyond. Trends in genetics : TIG *13*, 167-170.

Schwartz, Y.B., and Pirrotta, V. (2013). A new world of Polycombs: unexpected partnerships and emerging functions. Nature reviews Genetics *14*, 853-864.

Shen, L., Shao, N.Y., Liu, X., Maze, I., Feng, J., and Nestler, E.J. (2013). diffReps: detecting differential chromatin modification sites from ChIP-seq data with biological replicates. PLoS One *8*, e65598.

Simon, J.A., and Kingston, R.E. (2013). Occupying chromatin: Polycomb mechanisms for getting to genomic targets, stopping transcriptional traffic, and staying put. Molecular cell *49*, 808-824.

Smith, T.F., and Waterman, M.S. (1981). Identification of common molecular subsequences. J Mol Biol *147*, 195-197.

Sparmann, A., and van Lohuizen, M. (2006). Polycomb silencers control cell fate, development and cancer. Nature reviews Cancer *6*, 846-856.

Spyrou, C., Stark, R., Lynch, A.G., and Tavare, S. (2009). BayesPeak: Bayesian analysis of ChIP-seq data. BMC bioinformatics *10*, 299.

Statham, A.L., Strbenac, D., Coolen, M.W., Stirzaker, C., Clark, S.J., and Robinson, M.D. (2010). Repitools: an R package for the analysis of enrichment-based epigenomic data. Bioinformatics *26*, 1662-1663.

Stock, J.K., Giadrossi, S., Casanova, M., Brookes, E., Vidal, M., Koseki, H., Brockdorff, N., Fisher, A.G., and Pombo, A. (2007). Ring1-mediated ubiquitination of H2A restrains poised RNA polymerase II at bivalent genes in mouse ES cells. Nature cell biology *9*, 1428-1435.

Szenker, E., Ray-Gallet, D., and Almouzni, G. (2011). The double face of the histone variant H3.3. Cell research *21*, 421-434.

Tan, M., Luo, H., Lee, S., Jin, F., Yang, J.S., Montellier, E., Buchou, T., Cheng, Z., Rousseaux, S., Rajagopal, N.*, et al.* (2011). Identification of 67 histone marks and histone lysine crotonylation as a new type of histone modification. Cell *146*, 1016-1028.

Tavares, L., Dimitrova, E., Oxley, D., Webster, J., Poot, R., Demmers, J., Bezstarosti, K., Taylor, S., Ura, H., Koide, H.*, et al.* (2012). RYBP-PRC1 complexes mediate H2A ubiquitylation at polycomb target sites independently of PRC2 and H3K27me3. Cell *148*, 664-678.

Tie, F., Banerjee, R., Stratton, C.A., Prasad-Sinha, J., Stepanik, V., Zlobin, A., Diaz, M.O., Scacheri, P.C., and Harte, P.J. (2009). CBP-mediated acetylation of histone H3 lysine 27 antagonizes Drosophila Polycomb silencing. Development *136*, 3131-3141.

Trapnell, C., Hendrickson, D.G., Sauvageau, M., Goff, L., Rinn, J.L., and Pachter, L. (2013). Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol *31*, 46-53.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105-1111.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol *28*, 511-515.

Trojer, P., Cao, A.R., Gao, Z., Li, Y., Zhang, J., Xu, X., Li, G., Losson, R., Erdjument-Bromage, H., Tempst, P.*, et al.* (2011). L3MBTL2 protein acts in concert with PcG protein-mediated monoubiquitination of H2A to establish a repressive chromatin structure. Molecular cell *42*, 438-450.

Unoki, M., Masuda, A., Dohmae, N., Arita, K., Yoshimatsu, M., Iwai, Y., Fukui, Y., Ueda, K., Hamamoto, R., Shirakawa, M.*, et al.* (2013). Lysyl 5-hydroxylation, a novel histone modification, by Jumonji domain containing 6 (JMJD6). The Journal of biological chemistry *288*, 6053-6062.

Valouev, A., Johnson, D.S., Sundquist, A., Medina, C., Anton, E., Batzoglou, S., Myers, R.M., and Sidow, A. (2008). Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. Nat Methods *5*, 829-834.

van Lohuizen, M., Frasch, M., Wientjens, E., and Berns, A. (1991a). Sequence similarity between the mammalian bmi-1 proto-oncogene and the Drosophila regulatory genes Psc and Su(z)2. Nature *353*, 353-355.

van Lohuizen, M., Verbeek, S., Scheijen, B., Wientjens, E., van der Gulden, H., and Berns, A. (1991b). Identification of cooperating oncogenes in E mu-myc transgenic mice by provirus tagging. Cell *65*, 737-752.

Vandamme, J., Volkel, P., Rosnoblet, C., Le Faou, P., and Angrand, P.O. (2011). Interaction proteomics analysis of polycomb proteins defines distinct PRC1 complexes in mammalian cells. Molecular & cellular proteomics : MCP *10*, M110 002642.

Wallrath, L.L., Lu, Q., Granok, H., and Elgin, S.C. (1994). Architectural variations of inducible eukaryotic promoters: preset and remodeling chromatin structures. BioEssays : news and reviews in molecular, cellular and developmental biology *16*, 165-170.

Wang, H., Wang, L., Erdjument-Bromage, H., Vidal, M., Tempst, P., Jones, R.S., and Zhang, Y. (2004). Role of histone H2A ubiquitination in Polycomb silencing. Nature *431*, 873-878.

Wang, K., Singh, D., Zeng, Z., Coleman, S.J., Huang, Y., Savich, G.L., He, X., Mieczkowski, P., Grimm, S.A., Perou, C.M.*, et al.* (2010). MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. Nucleic acids research *38*, e178.

Whyte, W.A., Orlando, D.A., Hnisz, D., Abraham, B.J., Lin, C.Y., Kagey, M.H., Rahl, P.B., Lee, T.I., and Young, R.A. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell *153*, 307-319.

Woodcock, C.L., and Ghosh, R.P. (2010). Chromatin higher-order structure and dynamics. Cold Spring Harbor perspectives in biology *2*, a000596.

Wu, X., Johansen, J.V., and Helin, K. (2013). Fbxl10/Kdm2b recruits polycomb repressive complex 1 to CpG islands and regulates H2A ubiquitylation. Molecular cell *49*, 1134-1146.

Xiao, T., Hall, H., Kizer, K.O., Shibata, Y., Hall, M.C., Borchers, C.H., and Strahl, B.D. (2003). Phosphorylation of RNA polymerase II CTD regulates H3 methylation in yeast. Genes & development *17*, 654-663.

Ye, T., Krebs, A.R., Choukrallah, M.A., Keime, C., Plewniak, F., Davidson, I., and Tora, L. (2011). seqMINER: an integrated ChIP-seq data interpretation platform. Nucleic acids research *39*, e35.

Yuan, W., Wu, T., Fu, H., Dai, C., Wu, H., Liu, N., Li, X., Xu, M., Zhang, Z., Niu, T.*, et al.* (2012). Dense chromatin activates Polycomb repressive complex 2 to regulate H3 lysine 27 methylation. Science *337*, 971-975.

Zambelli, F., Pesole, G., and Pavesi, G. (2009). Pscan: finding over-represented transcription factor binding site motifs in sequences from co-regulated or co-expressed genes. Nucleic acids research *37*, W247-252.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W.*, et al.* (2008). Model-based analysis of ChIP-Seq (MACS). Genome biology *9*, R137.

Zhou, Q., Su, X., Wang, A., Xu, J., and Ning, K. (2013). QC-Chain: fast and holistic quality control method for next-generation sequencing data. PLoS One *8*, e60234.

Zhou, X., Lindsay, H., and Robinson, M.D. (2014). Robustly detecting differential expression in RNA sequencing data using observation weights. Nucleic acids research *42*, e91.

Zhu, L.J., Gazin, C., Lawson, N.D., Pages, H., Lin, S.M., Lapointe, D.S., and Green, M.R. (2010). ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. BMC bioinformatics *11*, 237.