International Journal of Biology and Life Sciences
ISSN: 2957-9511 | Vol. 3, No. 3, 2023

Deciphering the Gene Regulatory Networks during Embryonic Stem Cell Differentiation

Yisong Zhang, Kai Zhu

Faculty of Biology, Medicine and Health, The University of Manchester, UK

Abstract: With the development of sequencing technologies, genomics, and the advent of the era of big data, bioinformatics has become more and more important. One very important research area is the regulation of gene expression, which has become one of the hot research fields in bioinformatics. Transcription factors are important transcriptional regulators. In the process of gene expression, A key step is binding of transcription factors in gene expression by combining with a specific DNA sequence to regulate gene expression and inhibit or enhance its role. The difference between these specific DNA sequences is important for understanding gene regulation. With the rapid development of high-throughput sequencing technology, ChIP-seq, which combines chromatin immunoprecipitation technology and next-generation sequencing, provides massive data for transcription factor binding site across the genome. HiChIP, as a protein-centric chromatin conformation analysis method, is concerned because of its sensitive and efficient characteristics. In this project, we are working on the early demobilization of embryonic stem cells, in which ZIC3, Otx2, etc. play an important regulatory role. We are starting from analyzing HiChIP data on ZIC3 but mainly on calling peaks on HiChIP data, and ground tools to analysis HiChIP data. To integrate the chromatin binding properties of transcription factors with other chromatin markers. This report is based on the work done on HiChIP data during the early differentiation of stem cells, and has compared the methods required to reach the peak of this data. In addition, this project verified the output by using ground truth from histone mark ChIP-seq data (H3K27ac) and did the motif analysis from some of the output peaks.

Keywords: Embryonic Stem Cell; Gene Regulatory Networks; HiChIP; ChIP-seq.

1. Introduction

1.1. General Introduction

Stem cells are derived from embryos, fetuses or adults, and have certain self-renewal, proliferation and differentiation capabilities under specific conditions(Alison et al., 2002). These cells can not only produce daughter cells with the same phenotype and genotype as themselves, but also produce specialized cells that make up the body's tissues and organs. According to the origin of stem cells, they can be divided into two categories: embryonic stem cell (ESC) and adult stem cell (ASC)(Lanza and Rosenthal, 2004). Compared with ASC, ESC has unique biological characteristics: (1) developmental totipotency. Under certain conditions, ESC has the ability to differentiate into endoderm, mesoderm, and ectoderm cells. In theory, ESC can induce differentiation into all types of cells in the body. (2) Ability of germline transmission. ESCs contribute to the reproductive cells and are genetically passed to its offspring(van de Lavoir et al., 2006). (3) The convenience of genetic modification. The combination of ESC technology and gene targeting technology, such as gene knockout and transfections, has become an important means of studying gene function.

In the research and application of stem cells, ESC has always been a focus of attention (Fijnvandraat et al., 2003). ESC has the potential to self-differentiate both in vivo and in vitro, and is very easy to differentiate into other cells. Although significant progress has been made in inhibiting the differentiation of stem cells in vitro, their culture conditions still need to be optimized to maximize the potential for multi-directional differentiation of stem cells during expansion. One of the major bottlenecks in stem cell research is to clarify the development of ESC Regulation mechanism to achieve its directed induction differentiation. (Clements and Traver, 2013). At present, the research on its regulation mechanism is still very limited, which can be divided into two aspects: endogenous and exogenous regulation. Among them, endogenous regulation mainly involves structural proteins, structural factors, transcription factors, etc. in stem cells, which can affect the asymmetric mitosis of stem cells and the secretion of cytokines, so as to achieve the regulation of stem cell differentiation and development (Betschinger et al., 2003). Transcription factors form a network and can establish complex regulatory mechanisms. These networks change dynamically over time, so combining them through differential studies can infer key regulatory steps that affect cell differentiation. This report will combine these different types of data to characterize the regulation of chromatin and infer the control network in ESC differentiation. Three aspects of embryonic stem cell background will be introduced below.

1.2. ESC Role in Embryonic Development

Embryonic stem cells are called pluripotent stem cells because they have the potential for multi-directional differentiation. Their self-renewal and pluripotency are regulated by various transcription factors, epigenetic regulatory factors, and related signaling pathways. ESC originates from early embryos or primordial gonads, such as inner cell mass in blastocysts, early ectodermal leaves, etc. In 1981, two independent research groups successfully obtained ESC in mouse blastocysts(Evans and Kaufman, 1981). In 1988, human ESCs were also successfully separated (Thomson et al., 1998). Subsequently, the pluripotent embryonic stem cells of other species such as primates and mice were also isolated and cultured(Mitalipov et al., 2006, Buehr et al., 2008). The study of ESC has important

theoretical guiding significance for understanding the embryonic development process and its molecular mechanism. Therefore, ESC has become an important in vitro model for early embryo development, and is also widely used in basic biomedical research and clinical medicine.

1.3. ESC Differentiation

When the animal's embryonic development enters the gastrointestinal embryo stage, it marks the formation of a three-germ layer structure (endoderm, mesoderm, ectoderm), after which the cells derived from these three germ layers will continue to differentiate into various layers of tissues and organs. The research and utilization of the early molecular mechanisms of the three germ cells have laid the foundation for the rapid development of regenerative medicine, and the realization of artificially guided pluripotent stem cells to differentiate into specific germ layers to form the cell types needed in the clinic will be one of the clinical medicines. An important milestone. In addition, we can use the in vitro differentiation of mouse embryonic stem cells to focus on the development of early embryos (Murry and Keller, 2008). The specific medium can maintain the mouse ESCs in the naive ground state. According to research, inhibitors of extracellular signal-regulated kinase (Erk) and glycogen synthase kinase 3 (Gsk3) pathways can be used to maintain pluripotency in mouse ES cells. (called "2i condition"). In previous studies, it was discovered that the Erk and Gsk3 pathways affect the differentiation and development process and the regulation of gene expression. These observations indicate that Erk and Gsk3 pathways play an important role in the decision-making process of ES cells. Therefore, when the culture conditions are changed, the cells are changed. Pluripotent stem cells will be developed as EpiLCs or EpiSCs when 2i is reduced and then replaced with alternative media in their growing environment. In general, the mouse ESC we refer to is derived from the inner cell mass of the blastocyst stage, and there is another type of cells that are also considered to be mouse embryonic stem cells. They are derived from ectodermal cells after implantation of the embryo and become EpiSCs. Mouse embryonic stem cells (ESCs) and epiblast stem cells (EpiSCs) are two types of pluripotent stem cells with different epigenetic states. The former is derived from the inner cell mass of the blastocyst and the latter is derived from the ectoderm cells of the embryo after implantation. The cells in these two states are very different in terms of transcription resistance expression profile, clonal morphology, self-renewal capacity and multi-directional differentiation potential(Nichols and Smith, 2009).

1.4. Factors that Regulate ESC Differentiation

A number of factors are involved in ESC differentiation. According to Yang's research, when ESC differentiated into EpiLCs, ZIC3 showed a combination of transient expression kinetics and chromatin(Yang et al., 2019). ZIC3 regulates cell differentiation by regulating gene expression, especially a large number of genes encoding signal molecules and transcription factors. ZIC3 can also maintain the specificity of stem cells by regulating the transcription factor GRHL 2(Yang et al., 2019), regulating Tcf 15 to initiate EpiSCs differentiation. This shows that ZIC3 can activate the expression of transcription factor-encoding genes, which determines the phenotype of EpiLCs, and thus plays a key role in the transcription cascade. What's more, ZIC3 is also involved in the maintenance of pluripotency of ESCs (Lim et al., 2007). Declercq found that knocking out the *ZIC3 gene* will significantly reduce the number of clones formed in mouse iPS cells, and overexpressing the *ZIC3 gene* can significantly improve the induction efficiency of iPS cells, which provides an extremely advantageous tool for the generation of animal iPS cells. Additionally, the deletion of the *ZIC3 gene* also causes mice to exhibit early embryonic development defects, leading to defects in the left-right pattern (Ware et al., 2006). There are similar mutations in humans, such as the X-linked heterogeneous syndrome (Bellchambers and Ware, 2018).

Otx2 (Orthodenticle homeobox 2) is a member of the Otx transcription factor superfamily (Finkelstein et al., 1990). The mouse *Otx2 gene* is composed of 3 exons and is located on chromosome 1 between the exons and introns the cutting site conforms to the GT-AC principle. The *Otx2 gene* was first expressed in the endoderm and ectoderm at the early stage of embryo formation. As the development progressed, the expression of this gene was gradually restricted to the ectoderm at the front of the embryo. In the later stages of development, the expression of Otx2 was mainly concentrated in the forebrain and midbrain (Beby and Lamonerie, 2013). Otx2 has an important role in the development of organs and nervous system(Acampora et al., 2013). When OTX2 is knocked out, the embryonic lethality of the mouse is significantly increased, and the brain structure cannot develop normally. Studies have shown that Otx2 is required for ESC to transition to EpiSCs and maintain EpiSCs status. Without Otx2, ESC cannot transition to EpiSCs and can only differentiate into nerve cells prematurely. By inhibiting the gradual conversion of mesoderm to neural fate, Otx2 serves as an internal determinant to stabilize the EpiSCs state(Acampora et al., 2013).

In mouse embryonic stem cells, the *Otx2 gene* promotes the activation of FGF signaling pathways that are dependent on excited state pluripotent stem cells. The FGF activity of mouse embryonic stem cells were down-regulated with the knockdown of Otx2 and up-regulated with the overexpression of Otx2. After bFGF was added to the culture system of Otx2 knockout embryonic stem cells, the expression of p-Erk1 / 2 was significantly down-regulated relative to normal embryonic stem cells, at this time endogenous FGF was not activated; while the culture of embryonic stem cells overexpressing Otx2 After adding bFGF to the system, the expression of p-Erk 1/2 was rapidly up-regulated, and the endogenous FGF was continuously activated at this time. At the same time, activating the FGF signaling pathway can upregulate the *Otx2 gene* expression, while inhibiting the FGF signaling pathway can also suppress the Otx2 expression(Kunath et al., 2007).

The *Otx2 gene* can suppress the LIF / STAT3 signaling pathway of initial state pluripotent stem cells. P-Stat3 in LIF / STAT3 signaling pathway will be down-regulated with the knockdown of Otx2, at this time the endogenous LIF activity of embryonic stem cells will increase; in embryonic stem cells overexpressing Otx2, p-Stat3 expression will be up-regulated and endogenous to the cells LF activity is reduced. Remove the LF factor in the culture system of Otx2 knockout embryonic stem cells. After a period of culture, add the LIF factor to the culture system again. Compared with normal embryonic stem cells, the expression level of p-Stat3 in Otx2 knockout cells is rapidly increased. At this time the endogenous LIF is activated. In embryonic stem cells overexpressing Otx2, p-Sat3 expression is down-regulated,

and endogenous LIF is not activated at this time (Kunath et al., 2007, Acampora et al., 2013). When the LIF in the culture system was removed and the JAK inhibitor was added, the Otx2 knockout cells were still in an undifferentiated state (Acampora et al., 2013).

Related literature reports that Otx2 plays a leading role in the transformation of embryonic stem cells into ectoderm stem cells (EpiSCs) (Acampora et al., 2013). The continuous expression of *Otx2 gene* will affect the morphology and molecular biological characteristics of embryonic stem cells. Overexpression of Otx2 in mouse ESCs, pluripotent stem cells tend to EpiSCs, showing loose clone morphology, uneven AP staining, downregulation of pluripotency related genes Sox2, Nanog, Oct4 and Rex1, Klf family genes (Klf214 / 5) Silencing and differentiation-related genes T-brachyury; and Fgf5 were up-regulated. Currently, cells mainly depend on FGF and Activin A signaling pathways ((Hanna et al., 2010, Najm et al., 2011). When knocking down the expression of Otx2, the state of pluripotent stem cells tends to be more toward the initial state of embryonic stem cells. It showed compact clone morphology, uniform AP staining, upregulation of pluripotency-related genes such as Sox2, Nanog, Oct4, KIf family and Rex1, and down-regulation of differentiation-related genes T-brachyury and Fgf5. At this time, the maintenance of cell pluripotency mainly depends on the LF signaling pathway, and the dependence on the FGF signaling pathway is weakened(Fujishiro et al., 2013).

OCT4 is a core transcription factor during embryonic development. It can regulate the pluripotency of cells and participate in the regulation of the formation of the three germ layers during embryonic development. The function of OCT4 depends to a large extent on its protein level. Changes in the level of OCT4 protein will affect its interaction with other proteins and its DNA binding. Some studies have reported that post-translational modifications of specific sites on the OCT4 protein (such as SUMOylation, phosphorylation, and ubiquitination) will significantly affect the biological functions of the OCT4 protein, but which signal pathways are regulated by the post-translational modification of the OCT4 protein. And how these modifications affect the mechanism by which OCT4 selectively regulates target genes remains unclear(Buecker et al., 2014).

1.5. HiChIP Data Analysis

HiChIP (Hi-C chromatin immunoprecipitation) is becoming more and more popular in analyzing the three-dimensional chromatin contact between regulatory elements and annotating gene mutation functions. HiChIP pipeline can analyze single-ended and double-ended data, and read the map based on the repeat level for filtering, as well as calling peaks through MACS(Yan et al., 2014). To generate binding profiles over key genomic features, we use HiChIP data to annotate peaks and calculating enrichment for peak-associated genes

2. Methods

In order to get the highest confident set of peaks, different kinds of bioinformatic tools are used to call and compare peak summits.

2.1. ChIP-seq for ZIC3 and H3K27ac Data

ChIP-Seq is a combination of chromatin immunoprecipitation and deep sequencing technology. Recognizing the interaction between protein and DNA plays a key role in understanding gene transcription regulation. Recently, many transcription factor studies using ChIP-seq technology have found that the recognition sites of transcription factors are obviously enriched near the peak summit of ChIP-seq. Recently, many transcription factor studies using ChIP-seq technology have found that the recognition sites of transcription factors are obviously enriched near the peak summit of ChIP-seq(Moqtaderi et al., 2010).

Shen-hsi et al., cultivated EpiLc cells on a specific medium for one day and two days respectively, and after a series of operations, the original ChIP data of d1EpiLC and d2EpiLC were generated. At the beginning, I did the same ChIP-seq analysis in ZIC3 paper(Yang et al., 2019), and then processed the following H3K27ac data analysis. To start with, SRR data were download from Buecker (2014). Then, fastq-dump tool was used to convert SRR data from H3K27ac_ESC_rep1, H3K27ac_ESC_rep2, H3K27ac_EpiLC_plusActivin and H3K27ac_EpiLC_minusActivin (single-end) to fastq format. After generating the fastq file, trimoraic v0.39 was used to trim and filter these reads with single-end mode to remove adapters, truncated with <25 nucleotides (TRAILING: 5 SLIDINGWINDOW: 4: 15 MINLEN: 25; Bolger et al., 2014). Filtered reads were mapped against National Center for Biotechnology Information build 37 / mm9 of mouse genome using Bowtie2 v2.3.0(-X 2000 and –dovetail)(Gurtowski et al., 2012). Before this step, the genome sequence should be compressed and indexed according to certain rules. Unmapped pairs (-F 4) were discarded using SAMtools v1.9(Li et al., 2009). In order to read the unique mapping, reads were then de-duplicated using the MarkDuplicates function of the Picard tools. (http://broadinstitute. github. io/ picard/). Then MACS2 v2.2.5 was used to find peaks enriched in transcription factor binding sites using default parameters(Feng et al., 2011). In ChIP-seq analysis, the ratio of the number of unique sequences to the total number of sequences is a focus of attention. The subsequent Peak Calling process uses unique aligned and non-repeating sequences for Peak Calling.

The version of the mouse reference genome used in this article is mm9, and the mouse reference genome download address is ‘http://hgdownload.soe.ucsc.edu/downloads.html’

2.2. HiChIP Calling

2.2.1. Hichipper

Hichipper is used to call peaks from HiChIP data directly, Hichipper aggregates read density from either all samples or each sample individually. To specify these options, a total of four parameters can be selected, respectively ‘combined’ or ‘all’ are matched to ‘each’ or ‘self’. So Hichipper can run in four modes. Output from HiC-Pro and a sample manifest file (.yaml) that coordinates Optional high-quality peaks which are identified through ChIP-Seq and restriction fragment locations were used as input. Command line is below (Caleb and Martin 2018).

Hichipper --out output input.yaml

Each time the user runs Hichipper, a *.Hichipper.log file containing information pertaining to the flow of the software execution is placed in the out directory (specified by the --out flag). Unless otherwise specified, a file ending in Hichipper-qcReport.html provides an interactive quality control report for all samples(Lareau and Aryee, 2018).

2.2.2. HiChIP-Peaks

This package is used to find enriched peak regions from the

same d1EpiLC and d2EpiLC data as in Hichipper. The package takes the output of HiC-Pro as an input file and converts it to a restriction site level resolution map. Then reads within a specified number of restriction sites are selected from the diagonal (default = 2) and the background is modeled as a negative binomial. After running this packages, a list of peaks with their properties is generated and we can get a bedgraph at a restriction site level resolution that describes the reads per site(Shi et al., 2020).

It is worth noting that in order to ensure that the package can run correctly there are three files which are .REPairs, SCPairs and DEPairs files must be in the target folder, which is HICPRO_RESULTS / hic_results / data / sample / output folder(Shi et al., 2020).

2.3. Visualize Output Files and Functional Annotations

ChIPseeker uses the nearest gene and the genomic region where the peak is located to annotate the peak function(Yu et al., 2015). As one of the most widely used peak annotation software, ChIPseeker provides the following multiple functions: 1. peak on the chromosome Visualize the distribution near the TSS site. 2.Peak related gene annotations and distribution on various elements of the genome. 3. Obtain the peak bed file in the GEO database. 4) Comparison and overlap analysis of multiple peak files

2.4. UCSC Browser Tracks

In the research of biological information, the visualization of data is often a very important step. In any genome-wide epigenomic analysis, it is often interesting to check the enrichment at certain genomic loci, such as various Histone modifications that define the chromatin state or promoter transcription factor co-binding or enhancer elements. The classic way to accomplish this task is to load all bigbed files into UCSC and then navigate to the area of interest(Kent et al., 2002). UCSC genome browser is used to visualize data in order to track the regions where our readings are mapped to the genome, display all types of position-specific annotation information on the genomes of humans and model organisms(Brown, 2013), or track the coverage of each region of the reference genome and the depth of sequencing. Furthermore, UCSC Genome Browser provides a very confidential Custom Tracks to support many types of file formats, including bedGragh, wig, bigwig, bed, bigBed, bam, etc. Before the display annotation track, the peak file needs to be uploaded first .Since sequencing data is generally in the tens of tens of thousands of bases, the UCSC Genome Browser generally uses compressed binary file formats, namely BigBed and BigWig, to ensure long-distance transmission of big data(Brown, 2013), bed files need to be converted into bigBed format by the bedToBigBed tool for recognition of the UCSC genome browser(Kent et al., 2010). Before converting the format, we need to sort the files by chromosome and then by feature size. I used the sort command recommended by UCSC browser command 'sort -k1, 1 k2,2n '. In addition, Color, name, etc. can be modified based on the need to define the annotation track display characteristics

2.5. Motif Enrichment Analysis

HOMER is a set of tools for Motif search and second-generation data analysis. The tools in HOMER are written in Perl and C ++ and are Linux command line based. The software of HOMER is very comprehensive and can solve almost all the analysis of high-throughput sequencing data(Heinz et al., 2010).

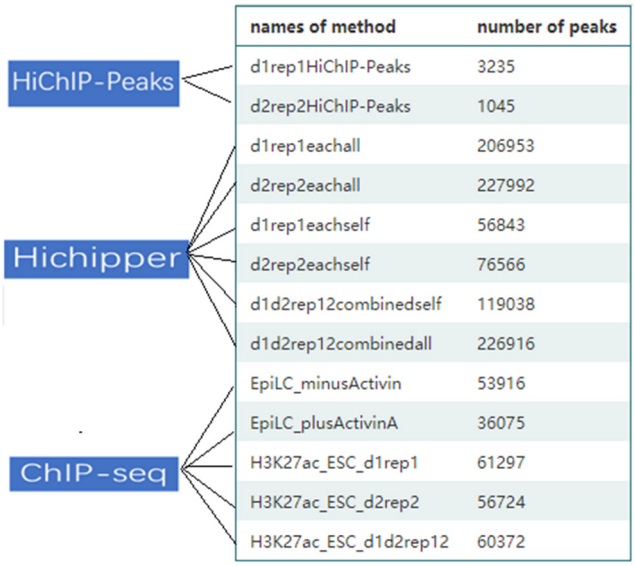
In the analysis of genome regulatory elements, HOMER can be used to discover new motifs. HOMER performs enrichment analysis by comparing two sequence sets, and then uses ZOOPS scoring (zero or one occurrence per sequence) and hypergeometric test. Through the difference analysis algorithm, the two sequences can be compared as parameters. In fact, the genomic sequence such as hg19, mm9, promoter sequence, custom FASTA sequence, etc. are used as reference sequences to check the enrichment of the target sequence on it to explore its function.

In this project, two sequence sets need to be provided before HOMER is used to predict the motif. One is the peak file generated before and the other is the background sequence set, mm9. FindMotifsGenome.pl is performed According to my specific settings (-size 250 -len 8,20) HOMER will retrieve the enrichment of known motifs in the target peak file and background genes, and output the results to the file knownResults.html.

3. Results

3.1. Comparative Analysis of Acetylation Peak Identification by Different Peak Callers

In order to generate the highest confident set of peaks, all peak files are put into tables for comparison. As can be seen from the table (Figure 1), the first eight files are generated from HiChIP data of d1EpiLC and d2EpiLC. Among them, the first two files are generated by HiChIP-peaks, and the last six files are generated by Hichipper. D1rep1hichippeaks and d2rep2hichippeaks are generated by processing the HiC-Pro output of d1EpiLC and d2EpiLC through the hichippeaks package. The original design of Hichip-peaks was based on identifying peaks in ChIP data centered on reconnection sites, and then used it for ring discovery and differential peak analysis(Shi et al., 2020). So, for data d1 and d2, Hichip-peaks may not be the best peak calling tool. Furthermore, d1rep1eachall, d2rep2eachall, d1rep1eachself, d2rep2eachself, d12rep12combinedself, rep12combinedall, which is generated by the HiC-Pro output of the d1chEpiLC and d2EpiLC through the package Hichipper. In addition, at the bottom of the table, EpiLC_minusActin, EpiLC_plusActin, H3K27ac_ESC_rep1 and H3K27ac_ESC_rep2 are ChIP-seq results generated through ChIP-seq workflow, using fastq-dump, Bowtie2, SAMtools, MACS2 to process SRR files step by step. The last file H3K27ac_rep12 is generated by merging H3K27ac_rep12, which uses the merge operation of bedtools. Since the two files H3K27ac_rep1 and H3K27ac_rep2 are very similar, they are merged. In the following result interpretation, Rep1 and Rep2 are the peak files generated by Hichipper for d1EpiLCs and d2EpiLCs, respectively. Rep12 is the file generated by Rep1 and Rep through the bedtools merge operation. Compared with HiChIP-Peaks, Hichipper got more peaks than HiChIP-Peaks. Because the number of peaks generated by HiChIP-Peaks is too small, I guess that the HiChIP data of d1EpiLC and d2EpiLC does not apply to HiChIP-Peaks.



names of method	number of peaks
d1rep1HiChIP-Peaks	3235
d2rep2HiChIP-Peaks	1045
d1rep1eachall	206953
d2rep2eachall	227992
d1rep1eachself	56843
d2rep2eachself	76566
d1d2rep12combinedself	119038
d1d2rep12combinedall	226916
EpiLC_minusActivin	53916
EpiLC_plusActivinA	36075
H3K27ac_ESC_d1rep1	61297
H3K27ac_ESC_d2rep2	56724
H3K27ac_ESC_d1d2rep12	60372

Figure 1. The number of peaks in each method. The first column represents the name of each method, from HiChIP-Peaks to ChIP-seq. The second column represents the number of peaks called by each corresponding method.

3.2. Comparison of Overlap between Pairs of 11 Samples

The heatmap in the table (Figure 2) represents the overlap of the matrix. The package overlapRateMat contains the rate of the overlaps between any two of the peak sets, which calculate the specific percentage of overlap between each other. Due to the limitation of the number of figures (up to 8), I hide the matrix. In order to ensure that the subsequent analysis can get more accurate results, biological repetition is very important in the analysis of biological data. Each square in the figure represents a sample, and the color of the square represents the correlation coefficient between samples. The brighter the color, the closer the reads distribution pattern in the sample. The correlation coefficient within a successful biological repeat sample is high. A package called GenomicRanges is loaded in R library, and then statistical methods are used to process these peak files. The average degree of the overlap covered by other samples is 31%. The heatmap below intuitively reflects the similarity of each two peak sets, which is made from the data in figure 2.

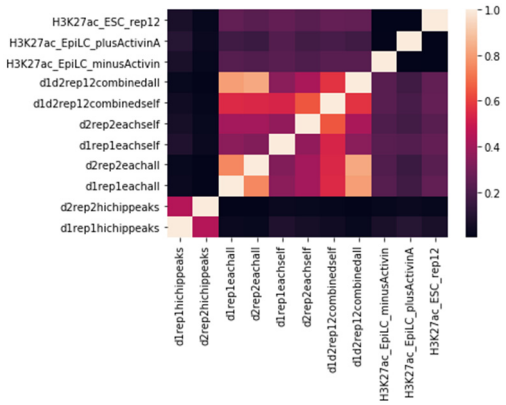


Figure 2. The heatmap of overlapRateMat. Datasets x, y, and z are depicted, while values represent the rate of overlap according to the color scale bar.

As can be seen from the heatmap, compared to several other peak files, the rep12combinedself generated by the Hichipper has a higher correlation to the data of EpiLC_minusActivin, EpiLC_plusActivin and H3K27ac_ESC_rep12, which is 23.43%, 19.23%, 26.36% respectively. So rep12combinedself is used for later analysis.

3.3. ChIP Peak Annotation, Comparison and Visualization

3.3.1. Distribution of Transcription Factor-binding Loci Relative to TSS

The distribution of unique alignment reads near the transcription start site (TSS) is affected by special gene regulation mechanisms. In order to visualize peak location and intensity from Rep12combinedself, Rep12_ESCrep12, Rep12_EpiLCminus, Rep12_EpiLCplus, ChIPseeker provides a powerful function plotAvgProf2 to visualize the average contour of the combined ChIP peak and TSS(Yu et al., 2015). X value is the transcription start site centered on TSS, and a negative value represents the upstream of the start site. The range of the abscissa is 3000bp upstream and downstream of the transcription initiation area, and the ordinate is the count of peaks. Transcription factors are usually combined near TSS to regulate gene expression. This figure indicates quality and distribution of the data that most of the reads of H3K27ac are distributed near TSS (Convex distribution), while a small part of Rep12combinedself is more scattered, and is also distributed at the far end of TSS.

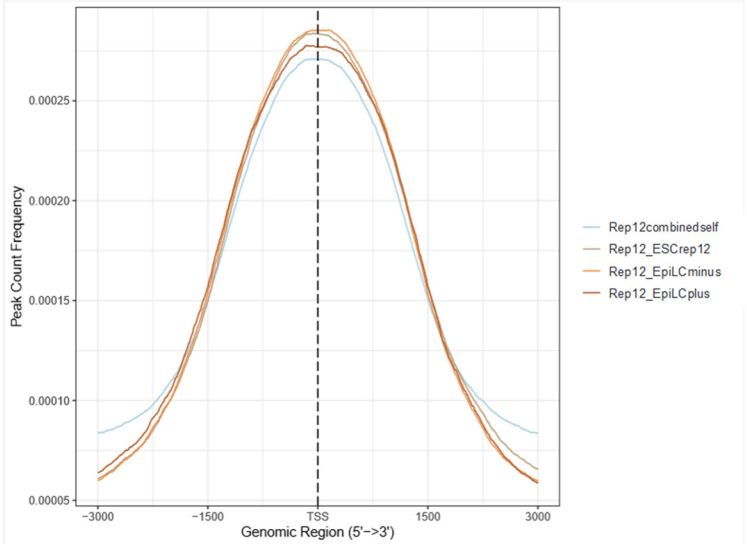


Figure 3. Distribution of transcription factor-binding loci relative to TSS. The x value represents the distance between reads and TSS, and the y value represents the frequency of reads in the interval.

3.3.2. Distribution of Acetylation Signals Across REP12 Peak Dataset

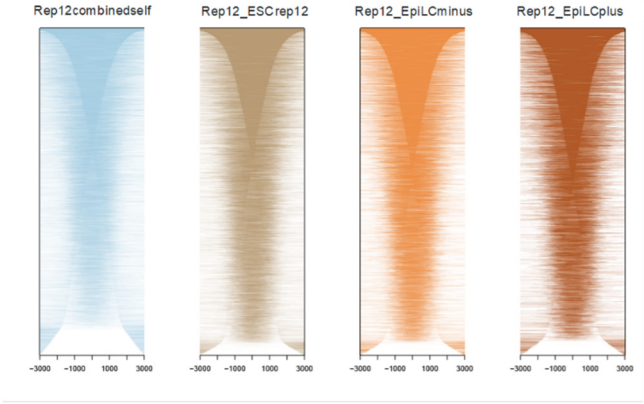


Figure 4. Heatmap of overlap peak region. Rep12combinedself is used to intersect with other 3 H3K27ac peak files which performed on the right. X value is the transcription start site centered on TSS, the negative value represents the upstream of the start site, and they value is the gene distribution.

In order to further observe the distribution of peaks on the basis of Figure 3, TagHeatmap (Figure 4) is provided by ChIPseeker to generate a heatmap to visualize the average profile of the combined ChIP peak and TSS area(Yu et al., 2015). By entering a series of bed files, multiple heat maps based on the combined peak of the TSS area can be compared. As can be seen from the figure, most of the peaks are concentrated around TSS. Part of the binding area of Rep12combinedself is far away from the TSS, and the peaks of the overlap are mainly the peak regions overlapping with H3K27ac data. In addition, Rep12combinedself incorporates many non-transcribed active regions of H3K27ac data.

3.3.3. Feature Distribution of Active Region

In order to analyze the proportion of peaks in each functional area near TSS, the function plotAnnoBar makes a clear comparison by drawing a percentage graph(Yu et al., 2015) (Figure 5). By annotating the peak with ChIPseeker, the function of the peak area can be viewed. It can be seen from the figure that most of the rep12combinedself peaks are distal Intergenic, accounting for about 40%. After H3K27ac and Rep12combinedself were intersected, the peak content of the promoter region inside increased, because H3K27ac is mainly the combined enhancer region. This is also consistent with the results shown in Figure 4. Most of the peaks in rep12 are combined with the remote location of TSS.

3.3.4. The Results of Pathway Annotations

In order to study what genes these tens of thousands of peaks are regulating, and what functions these genes have. The package ReactomePA in clusterProfiler is used to enrich the annotations by combining adjacent genes in peaks(Yu et al., 2015). After generating the peak files of overlap Rep12combinedself and H3K27ac, their peaks are annotated by using the enrichPathway function (Figure 6). According to the nearby genes of each peak, functional enrichment analysis is performed on these genes. We can see from the figure that these genes mainly regulate a variety of cellular responses

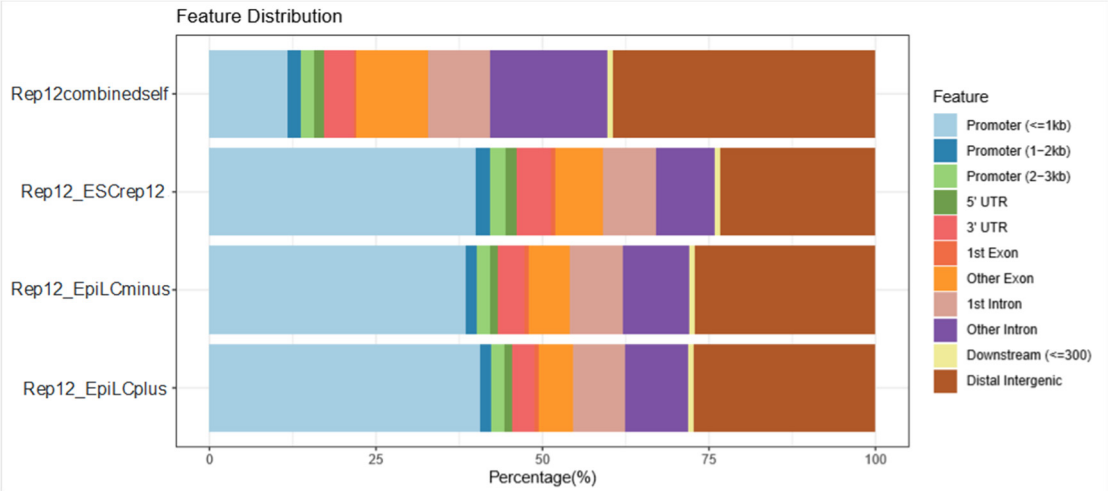


Figure 5. Feature distribution of Active region. The dataset comes from Rep12combinedself and three intersection files, and the names of each file are listed on the left. X value represents the percentage of each feature.

3.3.5. UCSC Browser Visualization

Fully functional and highly flexible visualization of gene annotations is provided by the USC Genome Browser through an efficient and dynamic web interface (Figure 7). All annotation data sets in the browser have been formatted and performed visually at different resolutions(Rosenbloom et al., 2015). After selecting the appropriate zoom factor on the display navigation, the annotation track image performed by UCSC genome browser is shown in Figure 7. The red marker shows exactly which part of the chromosome is enlarged in the trajectory image. The leftmost column is 14 customized bed file names, and the genome annotation track image displays the annotated genome in the horizontal direction. As can be seen from Figure 7, the bottom two tracks are generated by the HiChIP-Peaks method. Compared with Hichipper, the number of visualized peaks is significantly less, which also corresponds to the previous table (figure 1).

3.4. Motif Results

Motif is a structural component with a specific spatial conformation and a specific function in a protein molecule. A motif has its characteristic amino acid sequence, which binds to transcription factors and histones and performs special functions. Therefore, the combination of transcription factors, histones and DNA It is not random, but has a certain sequence preference. The first row of data in the table reveals the dynamic changes of ZIC3 during the transition from ESC to EpiLCs. ZIC3 transcription factors are present in embryonic stem cells but not in epiblast stem cells. The genes regulated

by ZIC3 are related to early mouse embryonic development. ZIC3 depletion and chromatin binding and genome-wide transcriptional profiling analysis confirm that ZIC3 is an important regulatory transcription factor (Yang et al., 2019). OCT4 is considered to be an interacting factor between ZIC2 and ZIC3, and it plays a regulatory role when embryonic stem cells differentiate into embryonic stem cells (Buecker et al., 2014). Transcription factor enrichment results show that, compared with EpiLCs, OCT4 accounts for a greater proportion of ESCs. The research by Buecker et al., Showed that the combination of OCT 4 with the genome is dynamic and regulated by the presence or absence of Otx2. When ESCs transition to EpiLCs, Otx2 will promote the binding of Oct4.

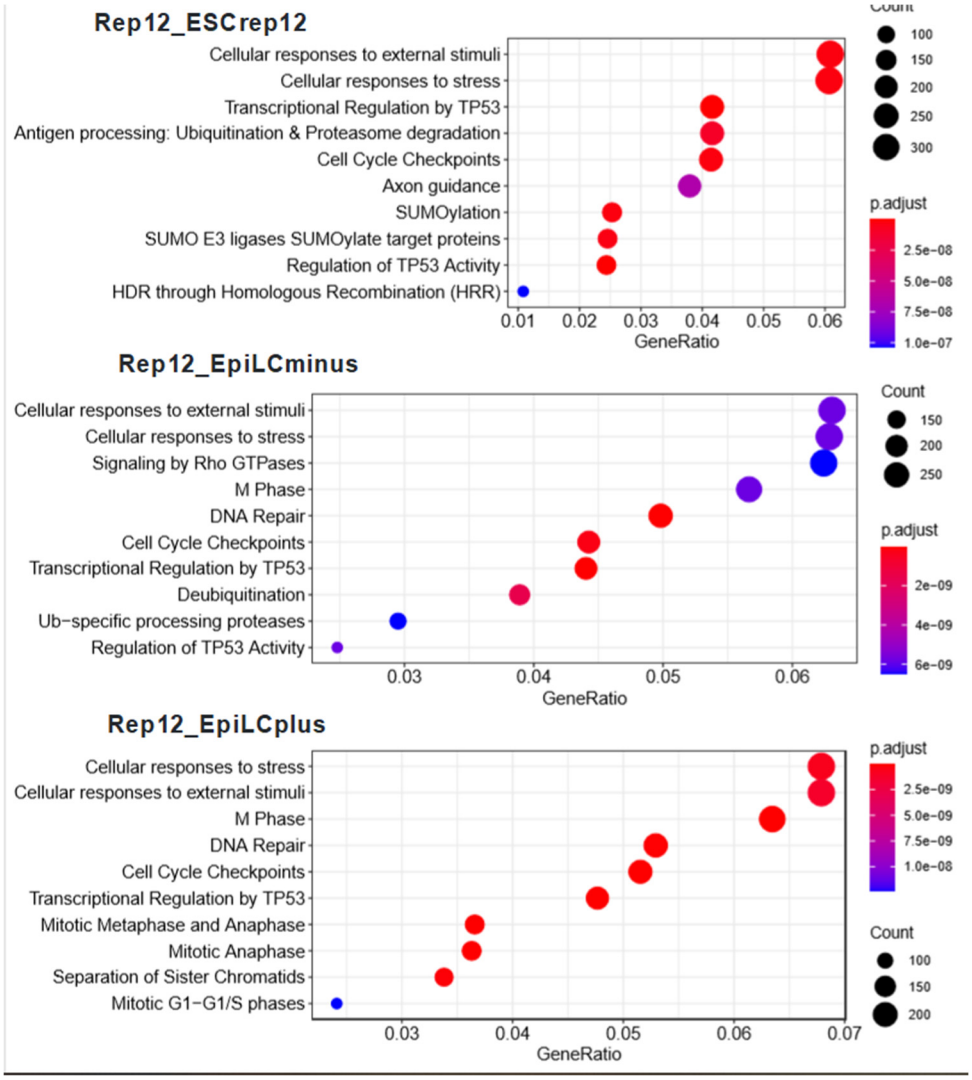


Figure 6. The results of pathway annotations. The top ten pathways are listed in the picture. The color represents the P value, and the circle size represents the number of genes enriched in this pathway.

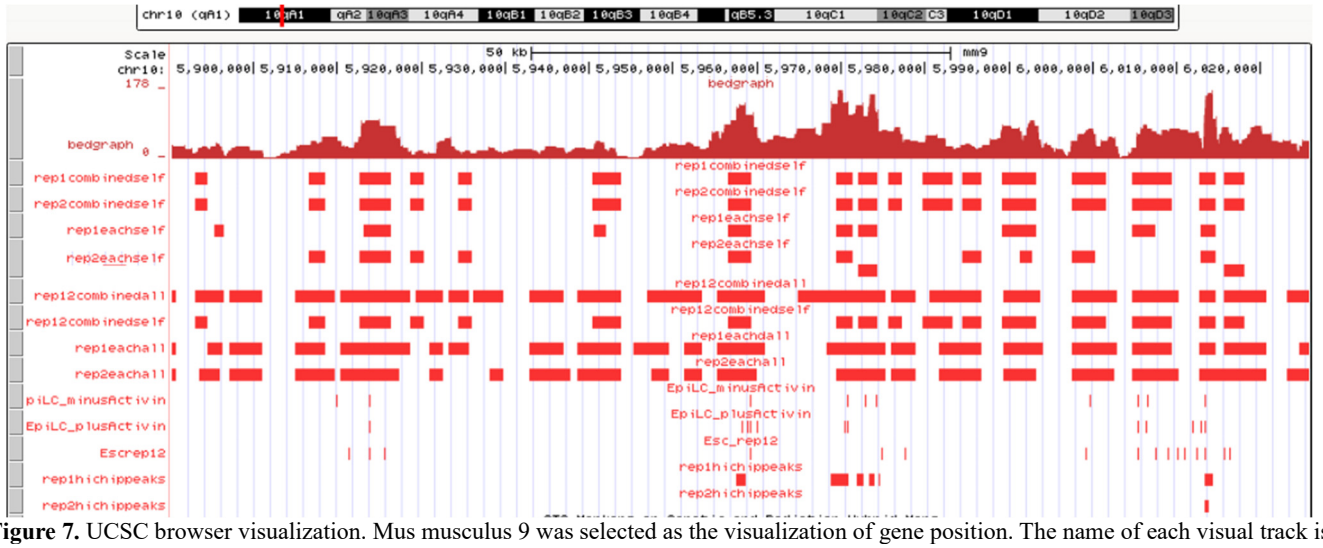


Figure 7. UCSC browser visualization. Mus musculus 9 was selected as the visualization of gene position. The name of each visual track is marked on the left, arranged from top to bottom. Compared with mm9 bedgraph, the distribution of reads in each track will be visualized.

Table 1. Motif results. The table shows the proportion of four transcription factors in three intersected files

Transcription factors		EpiLCplus_rep12 (Percentage/ p-value)	EpiLCminus_rep12 (Percentage/ p-value)	ESCrep12_rep12 (Percentage/ p-value)
	ZIC3	0	0	8.2%/1e-9
	Otx2	11.09%/1e-7	0	9.88%/1e-4
	OCT4	2.66%/1e-26	2.37%/1e-12	1.97%/1e-15
	ZIC2	0	0	6.69%/1e-9

4. Discussion and Conclusion

To sum up, this report studies the active areas of chromatin in embryonic stem cells and embryonic epiblast stem cells by using a variety of bioinformatics analysis tools. The results show that Otx2 and ZIC3 is involved in the regulation, replication and differentiation of embryonic stem cells. ZIC3 controls the phenotype of EpiLCs by activating the expression of genes encoding transcription factors. Although Luo's research indicates that ZIC2 has an inhibitory effect, 65% of the regulated ZIC3 target genes are still activated by ZIC3. With the activation of the target gene, ZIC3 will play a key role in the differentiation of ESCs to EpiLCs. The diversity of differentiation potential possessed by ESC itself reflects the complexity of the regulation of stem cell differentiation and development. At present, the research on its regulation mechanism is still very limited, and it can be divided into two aspects: endogenous and exogenous regulation. Among them, exogenous regulation mainly involves the microenvironment of stem cell growth(Watt and Hogan, 2000). Endogenous regulation mainly involves structural proteins, structural factors, transcription factors, etc. in stem cells, which can affect the asymmetric mitosis of stem cells and the secretion of cytokines, thus regulating the differentiation and development of stem cells(Chen and McKearin, 2003). The exogenous regulation of stem cell differentiation has also received increasing attention in recent years. All the exogenous factors involved in the regulation constitute the stem cell microenvironment.

However, the results of motif annotations still have limitations because using Hichipper and HiChIP-Peaks is challenging. Due to the different peak calling packages, the peak files generated are also different. To get more accurate results still need to consider other factors for analysis. At present, the process of naive ESCs starting to differentiate through the status of EpiLCs is still unclear(Yang et al., 2019). Although embryonic stem cells have shown ideal application prospects in basic science and clinical application research, it must be clear that as a new application element, stem cells still have some unavoidable problems, such as the directed induction of embryonic stem cell differentiation.

Declaration

No portion of the work referred to in this wthesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Intellectual Property Statement

1. The author of this thesis (including any appendices and/ or schedules to this thesis) owns certain copyright or related rights in it (the "Copyright") and s/he has given The University of Manchester certain rights to use Copyright, including for administrative purposes.

2. Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

3. The ownership of certain Copyright, patents, designs, trademarks and other intellectual property (the "Intellectual Property") and any reproductions of copyright works in the thesis, for example graphs and tables ("Reproductions"), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/ or Reproductions.

4. Further information on the conditions under which disclosure, publication and commercialization of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see http://documents. manchester.ac.uk/DocuInfo. aspx? DocID=2442 0), in any relevant Thesis restriction declarations deposited in the University Library, The University Library's regulations (see http://www.library.manchester.ac.uk/about/regulations/) and in The University's policy on Presentation of Theses.

Acknowledgments

I wish to thank all of my supervisors, Prof. Andrew Sharrocks, Dr. Mudassar Iqbal and Dr. Shen-Hsi Yang, for all their help and support throughout semester 2. I would also like to thank my Lab colleagues, Dr. Yaoyong Li, for his helpful Bioinformatics method suggestions and discussions. I am also thankful to all members, past and present, of the Andrew Sharrocks's Lab for their help.

I would also like to thank PhD Chenfu Shi in Gisela Orozco's lab from the University of Manchester for his help

in solving the problem of software operation.

Finally, I would like to thank my friends and family who have supported me throughout my postgraduate and have always been there when needed.

References

[1] ACAMPORA, D., DI GIOVANNANTONIO, L. G. & SIMEONE, A. 2013. Otx2 is an intrinsic determinant of the embryonic stem cell state and is required for transition to a stable epiblast stem cell condition. Development, 140, 43-55.

[2] ALISON, M. R., POULSOM, R., FORBES, S. & WRIGHT, N. A. 2002. An introduction to stem cells. J Pathol, 197, 419-23.

[3] BEBY, F. & LAMONERIE, T. 2013. The homeobox gene Otx2 in development and disease. Exp Eye Res, 111, 9-16.

[4] BELLCHAMBERS, H. M. & WARE, S. M. 2018. ZIC3 in Heterotaxy. Adv Exp Med Biol, 1046, 301-327.

[5] BETSCHINGER, J., MECHTLER, K. & KNOBLICH, J. A. 2003. The Par complex directs asymmetric cell division by phosphorylating the cytoskeletal protein Lgl. Nature, 422, 326-30.

[6] BROWN, S. M. 2013. Next-generation DNA sequencing informatics, Cold Spring Harbor, New York, Cold Spring Harbor Laboratory Press.

[7] BUECKER, C., SRINIVASAN, R., WU, Z., CALO, E., ACAMPORA, D., FAIAL, T., SIMEONE, A., TAN, M., SWIGUT, T. & WYSOCKA, J. 2014. Reorganization of enhancer patterns in transition from naive to primed pluripotency. Cell Stem Cell, 14, 838-53.

[8] BUEHR, M., MEEK, S., BLAIR, K., YANG, J., URE, J., SILVA, J., MCLAY, R., HALL, J., YING, Q. L. & SMITH, A. 2008. Capture of authentic embryonic stem cells from rat blastocysts. Cell, 135, 1287-98.

[9] CHEN, D. & MCKEARIN, D. 2003. Dpp signaling silences bam transcription directly to establish asymmetric divisions of germline stem cells. Curr Biol, 13, 1786-91.

[10] CLEMENTS, W. K. & TRAVER, D. 2013. Signalling pathways that control vertebrate haematopoietic stem cell specification. Nat Rev Immunol, 13, 336-48.

[11] EVANS, M. J. & KAUFMAN, M. H. 1981. Establishment in culture of pluripotential cells from mouse embryos. Nature, 292, 154-6.

[12] FENG, J., LIU, T. & ZHANG, Y. 2011. Using MACS to identify peaks from ChIP-Seq data. Curr Protoc Bioinformatics, Chapter 2, Unit 2 14.

[13] FIJNVANDRAAT, A. C., VAN GINNEKEN, A. C., SCHUMACHER, C. A., BOHELER, K. R., LEKANNE DEPREZ, R. H., CHRISTOFFELS, V. M. & MOORMAN, A. F. 2003. Cardiomyocytes purified from differentiated embryonic stem cells exhibit characteristics of early chamber myocardium. J Mol Cell Cardiol, 35, 1461-72.

[14] FINKELSTEIN, R., SMOUSE, D., CAPACI, T. M., SPRADLING, A. C. & PERRIMON, N. 1990. The orthodenticle gene encodes a novel homeo domain protein involved in the development of the Drosophila nervous system and ocellar visual structures. Genes Dev, 4, 1516-27.

[15] FUJISHIRO, S. H., NAKANO, K., MIZUKAMI, Y., AZAMI, T., ARAI, Y., MATSUNARI, H., ISHINO, R., NISHIMURA, T., WATANABE, M., ABE, T., FURUKAWA, Y., UMEYAMA, K., YAMANAKA, S., EMA, M., NAGASHIMA, H. & HANAZONO, Y. 2013. Generation of naive-like porcine-induced pluripotent stem cells capable of contributing to embryonic and fetal development. Stem Cells Dev, 22, 473-82.

[16] GURTOWSKI, J., SCHATZ, M. C. & LANGMEAD, B. 2012. Genotyping in the cloud with Crossbow. Curr Protoc Bioinformatics, Chapter 15, Unit15 3.

[17] HANNA, J. H., SAHA, K. & JAENISCH, R. 2010. Pluripotency and cellular reprogramming: facts, hypotheses, unresolved issues. Cell, 143, 508-25.

[18] HEINZ, S., BENNER, C., SPANN, N., BERTOLINO, E., LIN, Y. C., LASLO, P., CHENG, J. X., MURRE, C., SINGH, H. & GLASS, C. K. 2010. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell, 38, 576-89.

[19] KENT, W. J., SUGNET, C. W., FUREY, T. S., ROSKIN, K. M., PRINGLE, T. H., ZAHLER, A. M. & HAUSSLER, D. 2002. The human genome browser at UCSC. Genome Res, 12, 996-1006.

[20] KENT, W. J., ZWEIG, A. S., BARBER, G., HINRICHS, A. S. & KAROLCHIK, D. 2010. BigWig and BigBed: enabling browsing of large distributed datasets. Bioinformatics, 26, 2204-7.

[21] KUNATH, T., SABA-EL-LEIL, M. K., ALMOUSAILLEAKH, M., WRAY, J., MELOCHE, S. & SMITH, A. 2007. FGF stimulation of the Erk1/2 signalling cascade triggers transition of pluripotent embryonic stem cells from self-renewal to lineage commitment. Development, 134, 2895-902.

[22] LANZA, R. & ROSENTHAL, N. 2004. The stem cell challenge. Sci Am, 290, 92-9.

[23] LAREAU, C. A. & ARYEE, M. J. 2018. Hichipper: a preprocessing pipeline for calling DNA loops from HiChIP data. Nat Methods, 15, 155-156.

[24] LI, H., HANDSAKER, B., WYSOKER, A., FENNELL, T., RUAN, J., HOMER, N., MARTH, G., ABECASIS, G., DURBIN, R. & GENOME PROJECT DATA PROCESSING, S. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics, 25, 2078-9.

[25] LIM, L. S., LOH, Y. H., ZHANG, W., LI, Y., CHEN, X., WANG, Y., BAKRE, M., NG, H. H. & STANTON, L. W. 2007. Zic3 is required for maintenance of pluripotency in embryonic stem cells. Mol Biol Cell, 18, 1348-58.

[26] MITALIPOV, S., KUO, H. C., BYRNE, J., CLEPPER, L., MEISNER, L., JOHNSON, J., ZEIER, R. & WOLF, D. 2006. Isolation and characterization of novel rhesus monkey embryonic stem cell lines. Stem Cells, 24, 2177-86.

[27] MOQTADERI, Z., WANG, J., RAHA, D., WHITE, R. J., SNYDER, M., WENG, Z. & STRUHL, K. 2010. Genomic binding profiles of functionally distinct RNA polymerase III transcription complexes in human cells. Nat Struct Mol Biol, 17, 635-40.

[28] MURRY, C. E. & KELLER, G. 2008. Differentiation of embryonic stem cells to clinically relevant populations: lessons from embryonic development. Cell, 132, 661-80.

[29] NAJM, F. J., CHENOWETH, J. G., ANDERSON, P. D., NADEAU, J. H., REDLINE, R. W., MCKAY, R. D. & TESAR, P. J. 2011. Isolation of epiblast stem cells from preimplantation mouse embryos. Cell Stem Cell, 8, 318-25.

[30] NICHOLS, J. & SMITH, A. 2009. Naive and primed pluripotent states. Cell Stem Cell, 4, 487-92.

[31] ROSENBLOOM, K. R., ARMSTRONG, J., BARBER, G. P., CASPER, J., CLAWSON, H., DIEKHANS, M., DRESZER, T. R., FUJITA, P. A., GURUVADOO, L., HAEUSSLER, M., HARTE, R. A., HEITNER, S., HICKEY, G., HINRICHS, A. S., HUBLEY, R., KAROLCHIK, D., LEARNED, K., LEE, B. T., LI, C. H., MIGA, K. H., NGUYEN, N., PATEN, B., RANEY, B. J., SMIT, A. F., SPEIR, M. L., ZWEIG, A. S.,

HAUSSLER, D., KUHN, R. M. & KENT, W. J. 2015. The UCSC Genome Browser database: 2015 update. Nucleic Acids Res, 43, D670-81.

[32] SHI, C., RATTRAY, M. & OROZCO, G. 2020. HiChIP-Peaks: A HiChIP peak calling algorithm. Bioinformatics.

[33] THOMSON, J. A., ITSKOVITZ-ELDOR, J., SHAPIRO, S. S., WAKNITZ, M. A., SWIERGIEL, J. J., MARSHALL, V. S. & JONES, J. M. 1998. Embryonic stem cell lines derived from human blastocysts. Science, 282, 1145-7.

[34] VAN DE LAVOIR, M. C., DIAMOND, J. H., LEIGHTON, P. A., MATHER-LOVE, C., HEYER, B. S., BRADSHAW, R., KERCHNER, A., HOOI, L. T., GESSARO, T. M., SWANBERG, S. E., DELANY, M. E. & ETCHES, R. J. 2006. Germline transmission of genetically modified primordial germ cells. Nature, 441, 766-9.

[35] WARE, S. M., HARUTYUNYAN, K. G. & BELMONT, J. W. 2006. Zic3 is critical for early embryonic patterning during gastrulation. Dev Dyn, 235, 776-85.

[36] WATT, F. M. & HOGAN, B. L. 2000. Out of Eden: stem cells and their niches. Science, 287, 1427-30.

[37] YAN, H., EVANS, J., KALMBACH, M., MOORE, R., MIDDHA, S., LUBAN, S., WANG, L., BHAGWATE, A., LI, Y, SUN, Z., CHEN, X. & KOCHER, J. P. 2014. HiChIP: a high-throughput pipeline for integrative analysis of ChIP-Seq data. BMC Bioinformatics, 15, 280.

[38] YANG, S. H., ANDRABI, M., BISS, R., MURTUZA BAKER, S., IQBAL, M. & SHARROCKS, A. D. 2019. ZIC3 Controls the Transition from Naive to Primed Pluripotency. Cell Rep, 27, 3215-3227 e6.

[39] YU, G., WANG, L. G. & HE, Q. Y. 2015. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. Bioinformatics, 31, 2382-3.