Journal of
Stem Cell Research & Therapy

Wang et al., J Stem Cell Res Ther 2015, 5:8
DOI: 10.4172/2157-7633.1000296

Commentary Open Access

Genome-wide Massive Sequencing in Embryonic Stem Cell Biology: Recent Insights and Challenges

Yaofeng Wang[1,2], Alex Chun Cheung[1], Jun-Tao Guo[3] and Bo Feng[1,2]*

[1]Key Laboratory for Regenerative Medicine, Ministry of Education, School of Biomedical Sciences, Faculty of Medicine, The Chinese University of Hong Kong, Hong Kong SAR, China
[2]SBS Core Laboratory, Shenzhen Research Institute, The Chinese University of Hong Kong, Shenzhen, China
[3]Department of Bioinformatics and Genomics, the University of North Carolina at Charlotte, Charlotte, NC, USA

Abstract

The discovery of pluripotent Embryonic Stem Cells (ESCs) in mammals has been vastly transforming stem cell research and regenerative medicine. To understand the molecular mechanism of pluripotency, massive sequencing technologies have been adopted with intense scientific interest due to their advantages, including high resolution, low noise, as well as their extensive coverage across the entire genome. Here we review the principles of genome wide massive sequencing technologies widely performed in ESCs studies, including ChIP-Seq, RNA-Seq and methylC-Seq. Recent improvements and applications of these technologies will also be discussed. In addition, a summary of various methodologies used to integrate the massive genome wide sequencing data will be presented. Integrating the massive data that delineate different aspects of ESCs can prompt numerous innovations for understanding the transcription networks in maintaining pluripotency as well as gene regulations and epigenetic modifications in ESCs, which are important for research and clinical applications. Furthermore, we highlight the features that are worthy to pay attention from biologists due to current challenges in massive sequencing data analysis in bioinformatics and biostatistics.

Keywords: Pluripotent; Embryonic; Stem; Endoderm

Introduction

Since mouse Embryonic Stem Cells (ESCs) were successfully established and cultured *in vitro* in the early 1980s [1, 2], research on pluripotent stem cells has become one of the most exciting areas in life sciences. ESCs are derived from the inner cell mass of mammalian blastocysts. They have the ability to indefinitely self-renew while maintaining pluripotency, which means that they can differentiate into all three germ layers (ectoderm, endoderm, and mesoderm). Somatic cells differentiated from ESCs *in vitro* are shown to possess the morphology and function similar to their counterparts isolated from adult tissues (reviewed in [3-5]). Thus, ESCs have been prospected as a novel source for cell replacement therapies in clinical applications, including organ transplantation and the treatment of debilitating diseases such as diabetes, Parkinson's, and Huntington's disease [6].

Furthermore, mammalian somatic cells were successfully reprogrammed to ESC-like pluripotent cells, referred to as induced Pluripotent Stem Cells (iPSCs), by Yamanaka and his colleagues in 2006 and 2007 [7,8]. Four pluripotency transcription factors Oct4, Sox2, Klf4, and c-Myc were first used to obtain iPSCs, and subsequent studies have found other factors could also facilitate this reprogramming process. Successful derivations of iPSCs allow us to get access to the pluripotent stem cell without leading to ethnic concern. Patient-specific iPSCs provide a valuable platform for autologous cell therapy and the modelling of human diseases.

The unique properties and unprecedented potential of these pluripotent ESCs have attracted much attention towards its underlying molecular network. Transcription factors and epigenetic modulation complexes specific to pluripotent stem cells have been extensively studied to investigate the molecular regulations of pluripotency in ESCs and iPSCs as illustrated in Figure 1A [9,10].

Emerging massive sequencing technologies, also referred to as next generation sequencing (NGS), have played crucial roles in unraveling genome-wide epigenetic landscapes, DNA binding profiles of transcription factors, as well as transcriptome discoveries. Genome-wide NGS datasets provide abundant information with ultra-high resolution (single base pair level) for depicting molecular mechanisms of transcription factor regulations, gene expressions, and epigenetic regulation. So far, three categories of genome-wide deep sequencing technologies have been applied in ESCs research: Chromatin immunoprecipitation followed by deep sequencing (ChIP-Seq), whole-genome RNA sequencing (RNA-Seq) and whole-genome bisulfite sequencing (MethylC-Seq) (Figure 1B). ChIP-seq has become one of the most popular techniques in demonstrating histone modifications and transcription factor (TF)-DNA binding profiles in ESCs studies. ChIP-Seq offers the opportunity for researchers to study gene regulation and epigenetic regulation conveniently due to its advantages including its high resolution, low noise performance and wide coverage [11,12]. RNA-Seq can be applied to quantify gene expression levels in transcriptome-wide levels and determine exon/intron boundaries [13]. DNA methylation, a major epigenetic regulatory mechanism for gene expression and cell differentiation, plays a critical role in functioning and regulating pluripotency networks in ESCs [14,15]. Emerging MethylC-Seq data in ESCs studies provide a new insight into the dynamic nature of DNA methylation and demethylation during cell reprogramming and differentiation, which is fundamental to the knowledge of epigenomics in ESCs. As genome-wide deep sequencing data in ESCs research rapidly expands, it is important and worthy to

***Corresponding author:** School of Biomedical Sciences, The Chinese University of Hong Kong, Lo Kwee-Seong Integrated Biomedical Sciences Building, Hong Kong, Tel: 852- 3943 1455; Fax: 852-2603 5123; E-mail: fengbo@cuhk.edu.hk

Received May 30, 2015; **Accepted** July 24, 2015; **Published** July 26, 2015

Citation: Wang Y, Cheung AC, Guo JT, Feng B (2015) Genome-wide Massive Sequencing in Embryonic Stem Cell Biology: Recent Insights and Challenges. J Stem Cell Res Ther 5: 296. doi:10.4172/2157-7633.1000296

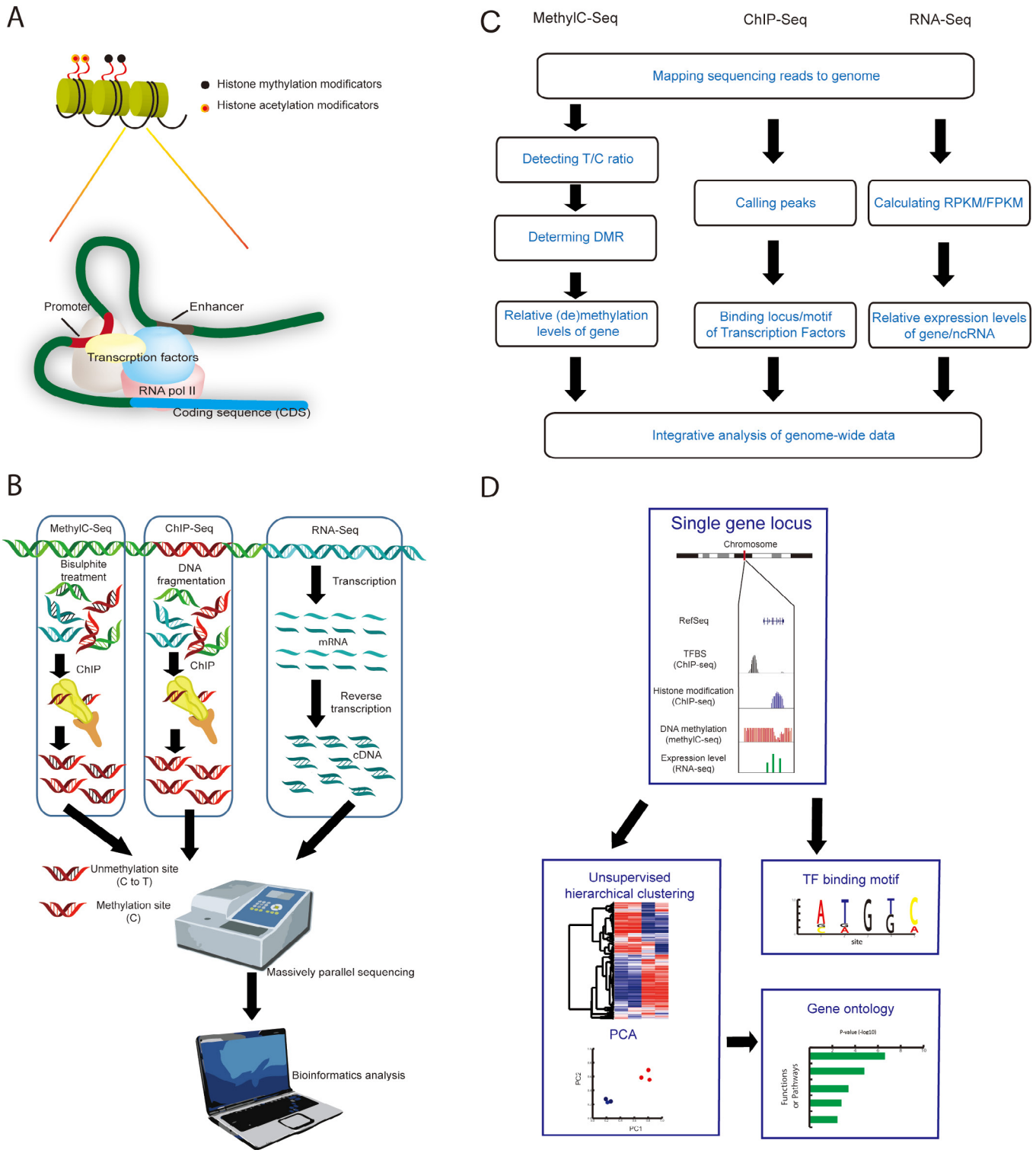

Figure 1: Overview of genome-wide massive sequencing in pluripotency studies.
(A) Histone modifications and transcription regulations in ESCs. Histone modifications cause chromosome conformational arrangement. Transcription factors bind to promotor and enhancer regions. Histone modifications and transcription factors regulate gene expression co-orperatively. (B) Library constructions for genome wide massive sequencing. Genomic DNA are fragmented in ChIP-Seq, and bisulfite-treated in MethylC-Seq. Short DNA fragments which bind to target protein are pulled down by ChIP. In RNA-Seq, mRNA transcriptome are converted to cDNA library. Millions of short DNA readings in ChIP-Seq, RNA-Seq or MethylC-Seq library are sequenced in NGS platform, and are further analyzed using bioinformatics tools. (C) Workflow of bioinformatics processing; DMR: differentially methylated regions. (D) Integrative analysis of genome wide sequencing data. ChIP-Seq, RNA-Seq and MethylC-Seq data can be integrated at a single locus, and further analyzed systematically genome-wide by unsupervised hierarchical clustering, PCA, and gene ontology.

have a clear insight of genome-wide deep sequencing techniques and their data processing methods for the researchers in the field. In this review, besides addressing the current "state of the art" in ChIP-Seq, RNA-Seq and MethylC-Seq research in ESCs studies, we also discuss the recent progress of technical optimizations and summarize a framework of computational methods and software packages for the processing of giant sequencing data. More importantly, due to the limitations of bioinformatics and biostatistics, we will highlight issues that need attention from biologists in this quickly expanding field.

Principles of ChIP-Seq, RNA-Seq and MethylC-Seq

Chromatin immune-precipitation followed by deep sequencing (ChIP-Seq)

Chromatin immune-precipitation (ChIP) is a technology for assaying histone modifications or protein–DNA binding *in vivo* [11,16]. In ChIP, antibodies are applied to pull down target proteins or nucleosomes, which bind to specific DNA fragments. Due to rapid development of NGS, CHIP-Seq has been successfully applied since 2007 [11,16]. The DNA fragments pulled down by ChIP are sequenced directly with tens or hundreds of millions of readings and then mapped to genome. The enrichments of DNA fragments on genome pinpoint the binding loci of target proteins. Compared to previous ChIP assay technology (ChIP-chip), ChIP–Seq has advantages such as ultra-high resolution and low noise owing to the single base-pair resolution and high accuracy of NGS. The coverage of ChIP–Seq is "real" genome-wide. Furthermore, the prospect of ChIP–Seq is promising since the cost of NGS is sustainably decreasing and the development of "third or fourth" generation sequencing techniques is with keen anticipation. Currently, ChIP–Seq is widely used for the genome-wide profiling of histone modifications, DNA-binding proteins, and nucleosomes in ESCs studies. More importantly, ChIP–Seq data have been massively mined to analyze direct binding or co-factor effects between transcription factors in pluripotency network in ESCs [16,17].

Sequencing of mRNA for gene expression profiling (RNA–Seq)

Determining the relative abundance of mRNA which reflects gene expression level is a significant topic in cell biology. Since the DNA microarray was developed through hybridization with labeled cDNA probes in 1996 [18], it has been widely used to detect relative gene expression levels in the past two decades [19-21]. However, the process of nucleic acid hybridization may lead to unavoidable noise [22]. Therefore, the NGS technique, referred to as RNA-Seq, was quickly adopted for genome-wide transcriptome analysis. Compared to microarray technology, RNA-Seq has similar advantages to ChIP-Seq, including higher resolution, lower noise and "real" genome-wide coverage. More importantly, RNA-Seq opens the gate to study noncoding RNAs, which were hardly covered previously in microarray despite their importance in biological research [23-25]. The principle of conducting RNA-Seq is simple: RNA samples, such as whole transcriptome of cells, are reversely transcribed to construct a cDNA library and then sequenced by NGS platforms. The readings are mapped to the genome and enriched regions will be picked and represented as FPKM (Fragments Per Kilo base of exon per Million fragments mapped) values. In this way, relative gene expression levels and exon/intron boundaries can be discovered. It is worth noting that the quality of RNA sample is very important. The contamination of DNA, ribosome or tRNA should be prevented. At the current stage, RNA-Seq data are cohesively analyzed with ChIP-Seq data on genome map to investigate the molecular mechanisms of transcription factors and their functions of gene regulation in ESCs [26].

Whole-genome bisulfite sequencing (MethylC-Seq)

DNA methylation and covalent modifications of histone proteins are regarded as epigenetic modifications, and are crucial in controlling transcriptions. Additions of methyl groups to the adenine or cytosine bases of DNA are stable during different states in cell fate process. A bivalent state of DNA demethylation formed by active H3K4 trimethylation (H3K4me3) and repressive H3K27 trimethylation (H3K27me3) [27,28] was identified in ESCs. Genome-scale mapping of DNA methylations and histone modifications can be carried out by NGS techniques. MethylC-Seq, also known as bisulphite conversion followed by sequencing (BS–seq), is a technique using bisulphite treatment of DNA [29]. The bisulphite treatment of DNA can lead to deamination of cytosine to uracil, which is subsequently converted to thymine following PCR amplification, whereas methylated cytosines are resistant to deamination and remain as cytosines. Treated DNA samples are analyzed by reading thymidine as indicators of de-methylated cytosine positions and cytosine as indicators of methylated cytosine positions. By mapping sequence readings to the genome and calculating the ratio between thymidine and cytosine at base pair resolution, the methylation levels can be compared. MethylC-Seq has been widely applied in ESCs integratively with ChIP-Seq and RNA-Seq data on genome mapping [30-32]. Moreover, genome-scale mapping of DNA methylations plays a critical role in cancer diagnosis and therapies [33-35]. Recently MethylC-Seq technology coupled with ChIP, referred to as BisChIP-seq [36,37], has been developed and performed to investigate cross-talk between chromatin and DNA methylation.

Recent improvements in genome-wide sequencing techniques

Since other recent reviews have covered the details of conventional protocols of ChIP-Seq, RNA-Seq and MethylC-Seq [11,13,29], here we only discuss the protocols of these techniques briefly and focus on the recent progress in technical improvements in terms of four aspects: lowering sample input, increasing throughput, improving accuracy, and reducing costs [29]. Reduction in costs can be optimized by engineering commercial improvements in NGS platforms. However, there are other ways that can make genome-wide sequencing technologies more cost-effective, such as lowering sample input, increasing process efficiency, and improving experimental and computational accuracy. Thus, in this review we focus on possible approaches for lowering sample input, increasing throughput, and improving accuracy.

Approaches in sequencing library construction

It has been shown that heterogeneity of cells populations may exist in biological samples from *in vitro* derivation of ESCs and reprogramming of iPSCs [30,38]. Such variations may result in different cellular compositions and complicate contaminations of target cell samples. Therefore, only a limited amount of homogeneous cells is obtained in ESCs studies with fluorescence-activated cell sorting (FACS). In RNA-Seq and MethylC-Seq experiments, the sequencing library is constructed from whole transcriptomes and the genome of cell samples respectively. Therefore, the amount of mRNA or DNA samples easily meet the sequencing requirement. In a recent RNA-Seq study, the number of cells used for library construction was reduced to 10 cells [39]. However, in ChIP-Seq, the quality of a library is mainly dependent on the efficiency of immunoprecipitation (IP) antibodies. The steps of DNA purification and fragmentation during library construction apparently result in sample loss. Thus, to obtain a sufficient starting amount of DNA fragments (1–10 ng) following several cycles of PCR, a

large number of cells are required in ChIP-Seq experiments. Currently, 10^7 of homogeneous cells are commonly applied in most of ChIP-Seq experiments in ESCs studies. Therefore, the need to lower sample input and increase throughput is urgent for ChIP-Seq.

To lower the sample input, several attempts have been applied in ChIP-Seq. Native ChIP (N-ChIP), which avoided the formaldehyde crosslinking in library construction, was developed prior to the applications of ChIP-Seq [40,41]. N-ChIP has a higher resolution than cross-linked ChIP (X-ChIP), and lacks unspecific interaction due to formaldehyde cross-linking. N-ChIP is also more sensitive than X-ChIP, making N-ChIP suitable for studies with low cell numbers, such as ESCs. Native ChIP-Seq protocol was developed by Zhao et al. [42,43], and optimized by Lyle et al. [44]. The input number of cells was successfully reduced 200 folds to 100k per IP. However, low cell numbers in Lyle's group study led to increasing levels of unmapped sequence readings and PCR-generated duplicated readings. Further improvements are needed to overcome this challenge. Furthermore, an ultra-low-input micrococcal nuclease-based native ChIP (ULI-NChIP) sequencing method was developed by Lorincz et al. [45]. In this study, genome-wide histone mark H3K27me3 profiles was demonstrated from as few as 1K ESCs. Thereby, high quality libraries from rare cell populations were proved successfully and illustrated by this method. In addition, several other approaches have reduced the number of cell samples to 10K or even 5K in ChIP-Seq [46-49]. However, pre-amplifications of ChIP DNA fragments were required in these experiments before sequencing by *in vitro* transcription or PCR. These comprehensive pre-amplifications introduced potential bias significantly. To reduce this bias, a small number of 10K cells was successfully used as starting material without pre-amplifications in ChIP-Seq [50]. To further reduce the sample input without pre-amplifications, Huang et al. applied automated microfluidic ChIP technique to obtain the high-quality ChIP-Seq data from only 1K ESCs in 2015 [51]. More importantly, by developing and applying the semi-automatic microfluidic devices in this experiment, the whole ChIP process has been greatly shortened to 8 h. As a result, this protocol shows a great potential to have high throughput in ChIP-Seq applications commercially.

Approaches in processing large data sets

The most impressive feature of NGS is the unprecedented amount of data. Usually, raw data and images are measured in the scale of terabytes per run. Processing such a large amount of raw data from ChIP-Seq, RNA-Seq and methyC-Seq presents a great challenge. Computationally, data analysis performed using reasonable computer time and resources should be of high accuracy. Here we review the data analysis for genome-wide NGS data as a bottom-up process as shown in Figure 1C, which starts with mapping sequence readings to the genome. The recent optimizations of data processing techniques will be discussed. All discussed software packages in this section are illustrated in Table 1.

Mapping: The first step to handling the genome-wide NGS data is mapping the short sequence readings to the genome. The reading lengths of ChIP-Seq, RNA-Seq and MethylC-Seq are 30-50bp, 200-300bp and 200-300bp respectively for high resolution readings. In ESCs studies, the typical mammalian genome sizes are in scale of several gigabytes [52]. Thus, mapping of millions of short readings to a mammalian genome is one of the most intensive steps in the entire process. The mapping of ChIP-Seq and MethylC-Seq data is simpler than RNA-seq data that contains large gaps corresponding to introns that must be considered. Popular aligner software for ChIP-Seq data include Eland of Illumina platform, MAQ [53], Bowtie/Bowtie2 [54,55], and BWA [56,57]. For MethylC-Seq data mapping, the additional step is to detect the ratio of thymidine/cytosine at methylation sites, which can be achieved by specific aligner software such as BSMap [58], Bismark [59], BS-Seeker [60] and MethylCoder [61]. Aligners for RNA-seq data include TopHat [62], ERANGE [63], QPALMA [64], as well as Subread [65] and STAR [66] (for both ChIP-Seq and RNA-Seq data). First generation RNA-seq aligners such as TopHat were based on mapping algorithms for ChIP-Seq readings such as Bowtie, and then the addition of splicing steps of transcriptome fragments were handled in loops. However, this method needed enormous amounts of memory and computational time to run. Later on, optimizations of mapping algorithms of RNA-seq were applied in Subread and STAR. Reports on these improved methods describe them to be highly accurate and ultra-fast [65,66]. The only limitation of Subread and STAR is the excessive usage of memory (over 30GB), which makes them impossible for a typical desktop computer to run. Recently, a new RNA-Seq aligner, HISAT [67], was developed by Salzberg et al. who are also the developers of Bowtie and TopHat. HISAT requires much less memory than previous RNA-Seq aligners while maintaining high accuracy and ultra-fast speed.

It's worthy to note that it is more challenging to map RNA-Seq readings than ChIP-Seq and MethylC-Seq readings. The challenges in mapping RNA-seq readings are caused by splice junctions, paralogous gene families and pseudogenes. For instance, some readings from one paralog may be mapped to other paralogs or pseudogenes due to sequencing errors, which vary around 1% so far. On the other hand, pseudogenes can be masked when the differences between pseudogenes and encoding genes are greater than the sequencing error, which is expected to improve with the development of new NGS platforms. It is an advantage for RNA-Seq since some of the pseudogenes are hardly masked in traditional RNA experiments [68].

Peak calling of ChIP-Seq signals: In ChIP-Seq, once alignment is completed, the results of mapped readings can be visualized on genome browsers such as UCSC (https://genome.ucsc.edu) or Ensembl (http://www.ensembl.org). The visualizations can provide a semi-quantitative observation of informative impressions of enriched regions at a genome loci. However, this visualization cannot quantitatively identify the binding and transcription events or detect the global protein/DNA binding patterns across the entire genome. Therefore, enriched DNA fragments (peaks) at target protein binding locus need to be selected based on statistics. As [68], the current peak calling software packages for ChIP-Seq signals generally covers following basic steps: (i) detect signal profiles from experimental group, (ii) collect background profiles from control group, (iii) peaks calling criteria, (iv) post-call filtering of artificial peaks and (v) significance ranking of called peaks.

For ChIP-Seq reading signals, additional adjustment will be applied to discriminate the artifacts of single-end readings, which are typically performed nowadays. Single-end sequencing of DNA fragments reads from one of the two strands in the 5' to 3' direction, which results in two related distributions besides the expected read upstream and downstream. These "shifts" will be normalized to the standard signal tags input to peak calling criteria. Current peak calling software apply various models to handle the shifts. For instance, in FindPeaks [69] and peakSeq [70], shift distances can be input by user; in cisGenome [71] and SiSSRs [72], the average of paired tags is applied; in QuEST [16], shifts are applied locally by cross-correlation. In MACS [73], the most widely used software package, a global shift is applied by evaluating 1000 high quality pairs on genome wide.

Next, the enriched regions of experimental data will be compared to control data. A region will be identified as a "peak" if the fold enrichment between them is statistically significant. Generally the

Function	Name	Full Term	Contributors	Ref.
ChIP-Seq				
Mapping	MAQ	Mapping and Assembly with Quality	Durbin *et al.*, The Wellcome Trust Sanger Institute, UK	[53]
	Bowtie/Bowtie2	---	Langmead *et al.*, University of Maryland, USA	[54, 55]
	BWA	Burrows-Wheeler Aligner	Durbin *et al.*, The Wellcome Trust Sanger Institute, UK	[56, 57]
Calling Peaks	FindPeaks	---	Fejes *et al.*, BC Cancer Agency, Canada	[69]
	peakSeq	---	Rozowsky *et al.*, Yale University, USA	[70]
	cisGenome	---	Wong *et al.*, Stanford University, USA	[71]
	SiSSRs	Site Identification from Short Sequence Reads	Zhao *et al.*, National Institutes of Health, USA	[72]
	QuEST	Quantitative Enrichment of Sequence Tags	Sidow *et al.*, Stanford University, USA	[16]
	MACS	Model-based Analysis of ChIP-Seq	Liu *et al.*, Harvard University, USA	[73]
	USeq	---	Nix *et al.*, University of Utah, USA	[78]
RNA-Seq				
Mapping	TopHat	---	Salzberg *et al.*, The Johns Hopkins University, USA	[62]
	ERANGE	Enhanced Read Analysis of Gene Expression	Wold *et al.*, California Institute of Technology, USA	[63]
	QPALMA	Optimal Spliced Alignments of Short Sequence Reads	Rätsch *et al.*, Max Planck Society, Germany	[64]
	Subread	---	Shi *et al.*, The University of Melbourne, Australia	[65]
	STAR	Spliced Transcripts Alignment to a Reference	Dobin *et al.*, Cold Spring Harbor Laboratory, USA	[66]
	HISAT	Hierarchical Indexing for Spliced Alignment of Transcripts	Salzberg *et al.*, The Johns Hopkins University, USA	[67]
RPKM/FPKM calculations	IsoInfer	Inference of isoforms	Feng *et al.*, Tongji University, China	[79]
	Scripture	---	Guttman *et al.*, Massachusetts Institute of Technology, USA	[80]
	SLIDE	Sparse linear modeling of RNA-Seq data for isoform discovery and abundance estimation	Huang *et al.*, University of California, Berkeley, USA	[81]
	IsoLasso	Isoforms of Least Absolute Shrinkage and Selection Operator	Li *et al.*, University of California, Riverside, USA	[82]
	iReckon	Isoform reconstruction and abundance estimation	Brudno *et al.*, University of Toronto, Canada	[83]
	Traph	Transcripts in gRAPHs	Tomescu *et al.*, University of Helsinki, Finland	[84]
	Cufflinks	---	Pachter *et al.*, University of California, Berkeley, USA	[85]
	MiTie	Mixed Integer Transcript IdEntification	Behr *et al.*, Sloan-Kettering Institute, USA	[86]
	StringTie	---	Salzberg *et al.*, The Johns Hopkins University, USA	[87]
MethylC-Seq				
Mapping and determine of T/C ratio	BSMap	Bisulfite sequence MAPping	Li *et al.*, Baylor College of Medicine, USA	[58]
	Bismark	---	Krueger *et al.*, The Babraham Institute, UK	[59]
	BS-Seeker	Precise mapping for bisulfite sequencing	Pellegrini *et al.*, University of California, Los Angeles, USA	[60]
	MethylCoder	---	Pedersen *et al.*, University of California, Berkeley, USA	[61]

Table 1: Popular software packages in data processing of genome-wide massively parallel sequencing data.

cutoff of fold enrichment is defined by user. In popular peak calling software such as cisGenome, QuEST, SiSSRs, peakSeq and MACS, control data can be considered as the background signal of peak calling. In recent publications of ChIP-Seq experiments, using control data such as ChIP-Seq signals of protein G or GFP is a popular and confident way of verifying the background. Recently, a novel strategy, KOIN (knockout implemented normalization), was developed to increase signal specificity and reduce noise by knocking-out the target transcription factor as control [74]. Also, fold enrichment of control signals over experimental signals can be used to calculate the False Discovery Rate (FDR) in MACS and QuEST. If control data is not available, Poisson distribution can be applied to calculate the background profiles based on experimental signals. Finally, ChIP-Seq peaks can be ranked by *p-values* (the significance of fold enrichment) in most software (cisGenome, QuEST, SiSSRs, MACS), or fold enrichment values if *p-values* are not provided.

Comparisons of peak calling tools have been investigated before [73,75,76]. Among these tools, MACS have shown remarkably lower FDR, better motif occurrence, and better spatial resolution of FoxA1 and NRSF ChIP-Seq data compared to Peak Finder [77], Findpeaks and QuEST. On the other hand, in ChIP-Seq analysis of histone modification marker profile of H3K27me3, although FindPeaks, PeakSeq, USeq [78], and MACS identified various peaks in terms of peak size, number, and position relative to genes, similar conclusions about the effect of H3K27me3 on gene expression were reached. Although the calling of each genome-wide peak was very different in a comparative analysis

between 14 peak calling algorithms with extensive ChIP-Seq data of NRSF, FoxA1 and STAT6, the top 1000 and 2000 highest-quality peaks were very stable with high accuracy. Taken altogether, the efficiency of the peak calling tools depends largely on the experimental dataset.

Gene expression calculations of RNA-Seq data: In RNA-Seq experiments, Reads Per Kilobase per Million (RPKM) and Fragments Per Kilobase of exon per Million fragments mapped (FPKM) values are used to represent gene expression levels and quantify the enrichment of RNA fragments located on certain gene exons. Software packages which can be used to calculate the RPKM and FPKM include IsoInfer [79], Scripture [80], Slide [81], IsoLasso [82], iReckon [83], Traph [84] and Cufflinks [85] and MiTie [86]. Cufflinks had been widely used in past ESCs research. However, it requires massive computational memory and time. Recently, a software running on less memory and time known as StringTie [87] was developed by Salzberg et al. who previously developed Bowtie, TopHat and HISAT. A comparison of Cufflinks, Traph, Scripture, IsoLasso and StringTie [87] showed that StringTie performs significantly better on transcriptome assemblies in both simulated and experimental data. Generally StringTie recognizes 40% more mRNA assemblies than Cufflinks, despite only needing 10% of Cufflink's total computational time.

In summary, recent optimizations of algorithm in the processing of genome-wide NGS data have improved accuracy with reduced memory and computing time.

Bioinformatics Integrative Analysis

As discussed above, increasing genome-wide sequencing data are available in ESCs studies. In addition to the generation of genome-wide data in genomics, epigenomics and transcriptomics as discussed above, we also witnessed the rapid increase of proteomics data from ESCs studies [88]. While insights can be provided from each data source, an integrative analysis of multiple data systems can offer a holistic view of gene functions. Data integration of ChIP-Seq, RNA-Seq and MethylC-Seq data in ESCs can provide valuable information about a) annotating functional features of the genome; b) inferring the functions of genetic variants; and c) understanding the mechanisms of gene regulation [26]. However, the large amount of NGS data from diverse technology platforms presents challenges in integrated data analysis. Better strategies need to be developed and optimized in data access and processing.

Tips for integrative analysis of genome-wide NGS data

The key to integrate ChIP-Seq, RNA-Seq and MethylC-Seq data is to reduce data complexity. By calling peaks, data complexity of ChIP-Seq is reduced from tens of millions of sequence readings to only thousands of peaks that encompass the histone modification or transcription binding sites on a genome. Similarly in RNA-Seq data analysis, FPKM calculations and rankings are carried out, significantly expressed genes are ranked and mapped on the genome locus. In MethylC-Seq, the data complexity is reduced through identifying differentially methylated regions (DMRs). After reducing complexity, sequence readings of ChIP-Seq, RNA-Seq and MethylC-Seq data are reorganized as thousands of marked regions on the genome. The genomic annotations of these regions can be further analyzed by clustering, principal component analysis (PCA), gene ontology as well as other approaches (Figure 1D).

Unsupervised hierarchical clustering and PCA analysis

Although the data complexity can be reduced in terms of gene annotation, it is still hard to represent the biological importance. To integrate high-throughput data, especially with multiple sample groups, the more scalable way is to apply unsupervised clustering approaches. Clustering is an efficient tool for partitioning a large data set into subsets based on their similarity. Unsupervised approaches do not use any prior knowledge of the samples. Unsupervised hierarchical clustering has been widely used in gene-expression profiles such as microarray and RNA-seq data. In ESCs studies, the gene expression profiles were commonly compared by hierarchical clustering. For example, gene expression values are calculated in various cell types or conditions, and clustering identifies sets of co-expressed genes. In hierarchical clustering, relationships among objects are represented by a tree (also referred to as dendogram) with similar objects being grouped into "clusters". The advantage of hierarchical methods over non-hierarchical clustering methods is that the relationships found between or within clusters can be visualized directly.

One important step in hierarchical clustering is the way to measure the similarity or distance between any two objects or clusters. The pairwise distance can be calculated as Euclidean distance which focuses on the absolute expression value, or Pearson correlation coefficient and Spearman correlation coefficient which rely on the expression profile shape. There are three major methods for calculating the distance/similarity between any two clusters. Single linkage method defines the distance as the smallest distance of all pairwise distances between members of the two clusters. Complete linkage method calculates the maximum distance of all pairwise distances between members of the two clusters. Average linkage computes the average distance of all pairwise distances between two clusters.

PCA [89] is an alternative way to cluster information among samples. As a statistical technique for determining key features of a high dimensional dataset, PCA can simplify data complexity and help to visualize high dimensional data [90]. The goal of PCA is to reduce high dimensional data to a few sets (usually two or three) of new orthogonal variables called Principal Components (PCs) that capture most of the variances in the original data set. Whereas the last few PCs are often assumed to capture only the residual 'noise' in the data [91]. In ESCs studies, PCA can be applied to clarify the different groups of transcription factors in various expression and regulation levels [92,93].

Gene ontology

A number of annotated gene sets can be obtained by the integration of ChIP-Seq, RNA-Seq and MethylC-Seq data. The biological roles of these gene sets, including cellular components, molecular function and biological processes, can be mapped by gene ontology analysis. For instance, over-expressed gene sets in a certain condition can be selected to investigate the pathways involved. GO enrichment analysis highly depends on the GO database when carrying out the cross-relationship test. Briefly, the principle of GO enrichment analysis is testing sample frequency and background frequency. Sample frequency is the ratio between the number of genes in a certain GO term and total number of genes in the sample. Background frequency is the ratio between the number of genes in this GO term in the database and total gene number of whole database [94]. The P-value of sample frequency vs. background frequency obtained from a statistical test, such as Fisher's exact or chi-square, represents the significance of this GO term in the sample.

In ESCs studies, DAVID [95,96] and PANTHER [97] are online tools that have been most widely used for GO analysis. As discussed above, the principle of GO test is quite straightforward. The accuracy of GO analysis highly depends on the annotations of genes in GO database. The number of functional annotation tools and the knowledge

bases of human, mouse and rat genes in GO terms make DAVID and PANTHER efficient tools in ESC studies.

Protein–nucleotide binding motif discovery: Bayesian approach and hidden Markov model

Besides clustering analysis of gene expression profiles and GO enrichment analysis of cellular components, molecular functions, and biological processes as discussed above, the binding motif between protein and DNA (or RNA) on genome-scale is also an important topic in ESCs studies. Predictions of transcription factor binding sites (TFBSs) and motifs are crucial for biologists to understand gene regulation and pluripotent network. TFs bind to DNA in a sequence specific manner. However, TFBSs are generally short and have sequence variations at various positions, which make it a challenging task for predicting TFBS on a genome-wide scale. NGS genome-wide data, especially ChIP-Seq, provide extensive and high resolution information for TFBS data mining and bioinformatics prediction. Bayesian approaches [98], especially the hidden Markov models (HMM) [99], are applied in the investigations of protein-nucleotide interaction patterns. Bayesian approach and HMM are machine learning methods and are powerful to recognize the pattern from large data sets. In the modeling of DNA sequences, two sets of probabilities are needed in an HMM model. One is the hidden states emitting probabilities of nucleotides, and the other is the state to state transition probabilities [100].

Recently, a series of bioinformatics algorithms that identify protein-nucleoid binding motifs using genome-wide NGS data have been reported. In 2009, Choi et al. applied a hierarchical hidden Markov model to analyze simulation data and ChIP data of NRSF and CTCT. They demonstrated an improved TFBS identification with integrative resource usages over single data sources or a simple combination of two [101]. Furthermore, efforts to improve and optimize TF binding motifs by Bayesian approaches had been made [100,102-107]. Bayesian mixture models were also used to perform the integrative analysis of ChIP-Seq and RNA-Seq data in TF binding motif and expression levels analysis [108]. Twelve ECS-specific transcription factors were identified using the web-based TFBS predictor RSAT [109]. RSAT, which can process several thousand peaks within minutes using less memory, makes it easy for new bioinformatics users. Other online databases of TFBS motif including TRANSFEC [110] and JASPER [111] allow biologists to conveniently access the TFBS information as well. Besides TF binding motifs, recently HMM was applied to identify RNA sequence motifs of RBM10 binding in 2014 [112].

Recent applications of genome-wide deep sequencing data in mammalian ESCs studies

Transcription factors discovery

Accumulating discoveries about the functions of transcription factors in ESCs have been reviewed in greater details by Ng and Surani [10] and Lee et al. [9]. Most of the studies were done using ChIP-Seq technology. For instance, Chen et al. performed an integrative analysis of ChIP-Seq data in mouse and identified an 'Oct4-centric' module of core pluripotency factors [17]. Oct4, Sox2 and Nanog as well as Smad1, Stat3 and Tcf3 are the downstream effectors of signaling pathways controlled by BMP, LIF and Wnt respectively [113]. Oct4, Sox2 and Nanog form the primary network that governs the robust pluripotent state by binding to their own promoters and forming an auto-regulatory circuitry. They play critical roles in controlling the self-renewal and differentiation of ESCs (reviewed in [114,115]). The associated transcription factors linked to the 'Oct4-centric' module identified by ChIP-Seq include Esrrb, Nr5a2, Tcfcp2l1 and Klf4 [17,116]. Besides, a second 'Myc-centric' module, also identified by ChIP-Seq, includes c-Myc, n-Myc, E2f1, Zfx, Rex1 and Ronin, which target the genes involved in protein metabolism [17, 117-119]. 'Oct4-centric' and 'Myc-centric' modules have been reported to auto-regulate their own expression and therefore be the central pluripotent network [9,10]. Additionally, other TFs have been identified by ChIP-seq, which is associated with the core pluripotent network including Prdm14 [120], SetDB1 [121], Chd7 [122], p300 [123], esBAF [124], E2F4 [125], Smad2/3/4, Foxh1 [126], YY1 [127], Mediator (Med1 and Med12) and Cohesin (Smc1a, Smc3) [128], PCL2 [129], Mbd3 [130], KAP1 and ZNF486 [131], GATA1 [132], BRD2/3/4 [133], KAP1 [134], TEAD4 [135], FOXO3 [136], p53 [137], NUP98 [138], FOXA1/2 [139] and Tbx3 [140]. Recently, a systematic analysis of 38 transcription factors with extensive ChIP-Seq data across the differentiation of hESCs in three germ layers were reported [141]. More importantly, co-occupied transcription factors and their binding sites can be revealed with high resolution by studying the overlapping peaks of ChIP-Seq data. For example, so far at least 14 transcription factors have been found to bind at the enhancer region of Oct4 [10] and eleven of them were identified by ChIP-Seq (Oct4, Sox2, Nanog , Stat3, Smad, Esrrb, Klf4, Tcf3, E2f1, n-Myc and Zfx) [17,113,126,142]. Similarly, at least nine transcription factors have been shown to co-occupy the enhancer region of Nanog gene [10] and five of them were identified by ChIP-Seq (Nanog, E2f1, Esrrb, Stat3 and Tcfcp2l1) [17].

Genome-wide NGS datasets can also be used to directly compare the genomic binding sites between species. Taking endogenous retroviral sequence as an example, human OCT4 and NANOG bind to human-specific ERV1 (endogenous retroviral sequence 1)-repeat transposable elements, whereas mouse Oct4 and Nanog bind to murine-specific ERVK (endogenous retrovirus K)-repeat elements [143,144]. These comparative studies provide valuable information about sequence conservations of TF binding sites between species and demonstrate the diversity of the transcriptional circuitries.

Histone modification and DNA methylation of epigenetics

Histone modifications have been proposed to be essential for the maintenance of pluripotency of ESCs. ChIP-Seq is widely used in ESCs studies to probe histone modifications. It has been shown that histone demethylases can prevent the accumulation of repressive methylation at the promoters of genes that maintain pluripotency of ESCs [10] and hyperacetylated chromatin in ESCs is proposed to adopt an open structure and reduce repressive methylation [145]. Therefore, histone methylation modifications, such as of H3 lysine 4 (H3K4), H3 lysine 9 (H3K9) and H3 lysine 27 (H3K27); and acetylation modifications, such as H3 acetyl lysine 9 (H3K9ac) and H3 acetyl lysine 27 (H3K27ac), are critical histone modification markers of cell pluripotent states [146]. Analysis of ChIP-Seq data demonstrated that the expressions of *Jmjd1a* and *Jmjd2c* genes, which encode histone H3 lysine 9 demethylases, are positively regulated by Oct4 [17]. Jarid2 and Mtf2, components of the Polycomb Repressive Complex 2 (PRC2) that mediate H3K27me3, are downstream targets of Oct4 [17,147]. Moreover, analyzing the ChIP-Seq data of these histone markers in ESCs is an important way to depict the pluripotent gene network in ESCs, since the binding of histone modification are expected to enrich in the gene enhancer regions [113,136,139,148-154].

More importantly, massively epigenetic studies of ESCs or iPSCs have been carried out by the combination of ChIP-Seq, RNA-Seq and MethylC-Seq data recently [31,155-163]. DNA methylation plays the crucial role as the epigenetic switch driving somatic cells

to pluripotent [164]. Taking a recent epigenetic study of mouse iPSC reprogramming process as an example [30], by combining RNA-Seq (gene expression profiles), MethylC-Seq (DNA methylation profiles at TF promoter regions), and ChIP-Seq (histone modification profiles) data, the epigenomic mechanism of pluripotency in a roadmap of the reprogramming process was demonstrated. Genes with CpG-rich promoters demonstrate consistent low methylation levels and strong engagement of histone markers, whereas genes with CpG-poor promoters are safeguarded by methylation.

non-coding RNA (ncRNA) studies in ESCs

RNA-Seq data can be used to systematically analyze the transcriptomes of ESCs and iPSCs with single-base resolution. Several transcriptome RNA-Seq datasets are available in mESC [148], mouse iPSC [149], hESC [150,151] and induced naïve-state hESC [152]. One crucial molecular mechanism of pluripotency in ESCs is the function of non-coding RNAs (ncRNA), which inhibit gene expression by binding to mRNAs. Crosstalk bewteen ChIP-Seq and RNA-Seq data provides insight into the details of ncRNA-mRNA binding events [10,24]. For instance, microRNA (miRNA) such as mir302 and mir290 clusters, which are involved in regulations of the G1 phase of ESCs, were reported to be regulated by Oct4, Sox2 and Nanog with ChIP-Seq data in 2008 [113]. Moreover, analysis of ChIP-Seq data revealed that a 30-amino-acid region of JARID2 mediated interactions with long noncoding RNAs (lncRNAs) and the presence of lncRNAs stimulated JARID2-EZH2 interactions *in vitro* as well as JARID2-mediated recruitment of PRC2 to chromatin *in vivo* [165]. Recently, analysis of ChIP-Seq and RNA-Seq data revealed that pluripotency factors OCT3/4, SOX2, and KLF4 transiently activated LTR7, long-terminal repeats of HERV type-H (HERV-H), to maintain the gene expression profile required for the pluripotent state during the reprogramming [166,167]. In addition, ncRNA-mRNA gene pairs were identified through systematic analysis of RNA-Seq data in ESCs [80,168,169]. X chromosome inactivation (XCI) is another key feature of ESCs in pluripotent states. Previously XIST, a long noncoding RNA, was suspected to be crucial in events of XCI. Recently, ChIP-Seq and RNA-Seq experiments showed that loss of *XIST* expression is not the primary cause of XCI instability and that gene reactivation from the inactive X precedes loss of *XIST* coating in hPSC [170]. Expression and coating by the long non-coding RNA *XACT* are early events in XCI erosion and may therefore play a role in mediating this process.

Current Challenges and Prospective

Up until now, alignment and analysis tools have been designed for the short sequence readings of NGS platforms. With the improvement of long sequence accuracy, software programs need to be updated to deal with the long readings of raw data from ChIP-Seq, RNA-Seq and MethylC-seq. Currently, the unmapped readings in raw data are commonly removed in analyses. With improvements of algorithms, the unmapped readings can be re-analyzed to gain further information including single-nucleotide polymorphism annotations and updated genome references. In addition, combination of ChIP-Seq data and chromatin conformation capture methods [171] can provide valuable information about distal regulatory elements and transchromosomal gene regulation networks. These questions may lead to critical information of hallmarks in ESCs study. Robust software tools for data analysis and closer interaction between laboratory scientists and bioinformaticians are clearly needed.

Besides ChIP-Seq, RNA-Seq and MethylC-Seq, other types of genome-wide NGS data have been carried out in ESCs studies. For instance, reduced representation bisulfite sequencing (RRBS-Seq) [172], methylated DNA immunoprecipitation (MeDIP-Seq) [173], as well as methyl-CpG binding domain (MBD-seq) [174] have been developed to detect DNA methylations. A comparison of these three technologies as well as MethylC-seq suggested the advantages and disadvantages among them [175]. In addition, circular chromosome conformation capture coupled with NGS (4C-seq) [176], a technique of NGS application in chromosome conformation capture (3C), has been carried out to demonstrate the organization of chromosomes and the physical interactions that occur within and between chromosomes [177-179]. Another example is the recent deep sequencing data, which revealed the genome-wide profiling of Clustered Regularly Interspaced Short Palindromic Repeats-associated protein-9 nuclease (CRISPR-Cas9) off-target effects [180,181]. By mapping NGS readings of CRISPR-Cas9 target fragments to the human genome, CRISPR-Cas9 off-target rate has been analyzed in a very high resolution, which significantly improves the computational accuracy. In the foreseeable future, increasing novel applications of deep sequencing data in ESCs studies will be desirable.

Funding

This work was supported by grants from the Research Grants Council of Hong Kong [CUHK 478812, CUHK 14102214 and CUHK 14104614 to B.F.], the National Science Foundation [DBI0844749 and DBI1356459 to J.G], and in part by funds from the National Natural Science Foundation of China [NSFC 31171433 to B.F.] and Guangdong Science and Technology Bureau International Science and from the Technology Collaboration Program [20130501c to W.Y.C.].

References

1. Martin GR (1981) Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells. Proc Natl Acad Sci U S A 78: 7634-7638. [PubMed]
2. Evans MJ, Kaufman MH (1981) Establishment in culture of pluripotential cells from mouse embryos. Nature 292: 154-156. [PubMed]
3. Spradling A, Drummond-Barbosa D, Kai T (2001) Stem cells find their niche. Nature 414: 98-104. [PubMed]
4. Surani MA (2001) Reprogramming of genome function through epigenetic inheritance. Nature 414: 122-128. [PubMed]
5. Donovan PJ, Gearhart J (2001) The end of the beginning for pluripotent stem cells. Nature 414: 92-97. [PubMed]
6. Lovell-Badge R (2001) The future for stem cell research. Nature 414: 88-91. [PubMed]
7. Takahashi K, Yamanaka S (2006) Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell 126: 663-676. [PubMed]
8. Takahashi K, Tanabe K, Ohnuki M, Narita M, Ichisaka T, et al. (2007) Induction of Pluripotent Stem Cells from Adult Human Fibroblasts by Defined Factors. Cell 131: 861-872.
9. Lee K, Wong W, Feng B (2013) Decoding the Pluripotency Network: The Emergence of New Transcription Factors. Biomedicines 1: 49-78.
10. Ng HH, Surani MA (2011) The transcriptional and signalling networks of pluripotency. Nat Cell Biol 13: 490-496. [PubMed]
11. Park PJ (2009) ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet 10: 669-680. [PubMed]
12. Mundade R, Ozer HG, Wei H, Prabhu L, Lu T (2014) Role of ChIP-seq in the discovery of transcription factor binding sites, differential gene regulation mechanism, epigenetic marks and beyond. Cell Cycle 13: 2847-2852. [PubMed]
13. Ares M Jr (2014) Methods for processing high-throughput RNA sequencing data. Cold Spring Harb Protoc 2014: 1139-1148. [PubMed]
14. Hackett JA, Surani MA (2013) DNA methylation dynamics during the mammalian life cycle. Philos Trans R Soc Lond B Biol Sci 368: 20110328. [PubMed]
15. Li CJ (2013) DNA demethylation pathways: recent insights. Genet Epigenet 5: 43-49. [PubMed]

16. Valouev A, Johnson DS, Sundquist A, Medina C, Anton E, et al. (2008) Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. Nat Methods 5: 829-834. [PubMed]

17. Chen X, Xu H, Yuan P, Fang F, Huss M, et al. (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell 133: 1106-1117. [PubMed]

18. Shalon D, Smith SJ, Brown PO (1996) A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. Genome Res 6: 639-645. [PubMed]

19. Schulze A, Downward J (2001) Navigating gene expression using microarrays--a technology review. Nat Cell Biol 3: E190-195. [PubMed]

20. Macgregor PF, Squire JA (2002) Application of microarrays to the analysis of gene expression in cancer. Clin Chem 48: 1170-1177. [PubMed]

21. Slonim DK, Yanai I (2009) Getting started in gene expression microarray analysis. PLoS Comput Biol 5: e1000543. [PubMed]

22. Hurd PJ, Nelson CJ (2009) Advantages of next-generation sequencing versus the microarray in epigenetic research. Brief Funct Genomic Proteomic 8: 174-183. [PubMed]

23. Sun L, Zhang Z, Bailey TL, Perkins AC, Tallack MR, et al. (2012) Prediction of novel long non-coding RNAs based on RNA-Seq data of mouse Klf1 knockout study. BMC Bioinformatics 13: 331. [PubMed]

24. Ilott NE, Ponting CP (2013) Predicting long non-coding RNAs using RNA sequencing. Methods 63: 50-59. [PubMed]

25. Soreq L, Guffanti A, Salomonis N, Simchovitz A, Israel Z, et al. (2014) Long non-coding RNA and alternative splicing modulations in Parkinson's leukocytes identified by RNA sequencing. PLoS Comput Biol 10: e1003517. [PubMed]

26. Hawkins RD, Hon GC, Ren B (2010) Next-generation genomics: an integrative approach. Nat Rev Genet 11: 476-486. [PubMed]

27. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, et al. (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 448: 553-560. [PubMed]

28. Mohn F, Weber M, Rebhan M, Roloff TC, Richter J, et al. (2008) Lineage-specific polycomb targets and de novo DNA methylation define restriction and potential of neuronal progenitors. Mol Cell 30: 755-766. [PubMed]

29. Plongthongkum N, Diep DH, Zhang K (2014) Advances in the profiling of DNA modifications: cytosine methylation and beyond. Nat Rev Genet 15: 647-661. [PubMed]

30. Lee DS, Shin JY, Tonge PD, Puri MC, Lee S, et al. (2014) An epigenomic roadmap to induced pluripotency reveals DNA methylation as a reprogramming modulator. Nat Commun 5: 5619. [PubMed]

31. Banovich NE, Lan X, McVicker G, van de Geijn B, Degner JF, et al. (2014) Methylation QTLs are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels. PLoS Genet 10: e1004663. [PubMed]

32. Bahar Halpern K, Vana T, Walker MD2 (2014) Paradoxical role of DNA methylation in activation of FoxA2 gene expression during endoderm development. J Biol Chem 289: 23882-23892. [PubMed]

33. Kalia M (2015) Biomarkers for personalized oncology: recent advances and future challenges. Metabolism 64: S16-21. [PubMed]

34. Marzese DM, Hoon DS (2015) Emerging technologies for studying DNA methylation for the molecular diagnosis of cancer. Expert Rev Mol Diagn 15: 647-664. [PubMed]

35. Shull AY, Noonepalle SK, Lee EJ, Choi JH, Shi H (2015) Sequencing the cancer methylome. Methods Mol Biol 1238: 627-651. [PubMed]

36. Statham AL, Robinson MD, Song JZ, Coolen MW, Stirzaker C, et al. (2012) Bisulfite sequencing of chromatin immunoprecipitated DNA (BisChIP-seq) directly informs methylation status of histone-modified DNA. Genome Res 22: 1120-1127. [PubMed]

37. Brinkman AB, Gu H, Bartels SJ, Zhang Y, Matarese F, et al. (2012) Sequential ChIP-bisulfite sequencing enables direct genome-scale investigation of chromatin and DNA methylation cross-talk. Genome Res 22: 1128-1138. [PubMed]

38. Mikkelsen TS, Hanna J, Zhang X, Ku M, Wernig M, et al. (2008) Dissecting direct reprogramming through integrative genomic analysis. Nature 454: 49-55. [PubMed]

39. Streets AM, Zhang X, Cao C, Pang Y, Wu X, et al. (2014) Microfluidic single-cell whole-transcriptome sequencing. Proc Natl Acad Sci U S A 111: 7048-7053. [PubMed]

40. O'Neill LP, Turner BM (2003) Immunoprecipitation of native chromatin: NChIP. Methods 31: 76-82. [PubMed]

41. Thorne AW, Myers FA, Hebbes TR (2004) Native chromatin immunoprecipitation. Methods Mol Biol 287: 21-44. [PubMed]

42. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, et al. (2007) High-resolution profiling of histone methylations in the human genome. Cell 129: 823-837. [PubMed]

43. Cuddapah S, Barski A, Cui K, Schones DE, Wang Z, et al. (2009) Native chromatin preparation and Illumina/Solexa library construction. Cold Spring Harb Protoc 2009: pdb. [PubMed]

44. Gilfillan GD, Hughes T, Sheng Y, Hjorthaug HS, Straub T, et al. (2012) Limitations and possibilities of low cell number ChIP-seq. BMC Genomics 13: 645. [PubMed]

45. Brind'Amour J, Liu S, Hudson M, Chen C, Karimi MM, et al. (2015) An ultra-low-input native ChIP-seq protocol for genome-wide profiling of rare cell populations. Nat Commun 6: 6033. [PubMed]

46. Adli M, Zhu J, Bernstein BE (2010) Genome-wide chromatin maps derived from limited numbers of hematopoietic progenitors. Nat Methods 7: 615-618. [PubMed]

47. Shankaranarayanan P, Mendoza-Parra MA, Walia M, Wang L, Li N, et al. (2011) Single-tube linear DNA amplification (LinDA) for robust ChIP-seq. Nat Methods 8: 565-567. [PubMed]

48. Shankaranarayanan P, Mendoza-Parra MA, van Gool W, Trindade LM, Gronemeyer H (2012) Single-tube linear DNA amplification for genome-wide studies using a few thousand cells. Nat Protoc 7: 328-338. [PubMed]

49. Sachs M, Onodera C, Blaschke K, Ebata KT, Song JS, et al. (2013) Bivalent chromatin marks developmental regulatory genes in the mouse embryonic germline in vivo. Cell Rep 3: 1777-1784. [PubMed]

50. Ng JH, Kumar V, Muratani M, Kraus P, Yeo JC, et al. (2013) In vivo epigenomic profiling of germ cells reveals germ cell molecular signatures. Dev Cell 24: 324-333. [PubMed]

51. Shen J, Jiang D, Fu Y, Wu X, Guo H, et al. (2015) H3K4me3 epigenomic landscape derived from ChIP-Seq of 1,000 mouse early embryonic cells. Cell Res 25: 143-147. [PubMed]

52. Trapnell C, Salzberg SL (2009) How to map billions of short reads onto genomes. Nat Biotechnol 27: 455-457. [PubMed]

53. Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res 18: 1851-1858. [PubMed]

54. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10: R25. [PubMed]

55. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9: 357-359. [PubMed]

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

57. Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26: 589-595. [PubMed]

58. Xi Y, Li W (2009) BSMAP: whole genome bisulfite sequence MAPping program. BMC Bioinformatics 10: 232. [PubMed]

59. Krueger F, Andrews SR (2011) Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics 27: 1571-1572. [PubMed]

60. Guo W, Fiziev P, Yan W, Cokus S, Sun X, et al. (2013) BS-Seeker2: a versatile aligning pipeline for bisulfite sequencing data. BMC Genomics 14: 774. [PubMed]

61. Pedersen B, Hsieh TF, Ibarra C, Fischer RL (2011) MethylCoder: software pipeline for bisulfite-treated sequences. Bioinformatics 27: 2435-2436. [PubMed]

62. Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25: 1105-1111. [PubMed]

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

64. De Bona F, Ossowski S, Schneeberger K, Rätsch G (2008) Optimal spliced alignments of short sequence reads. Bioinformatics 24: i174-180. [PubMed]

65. Liao Y, Smyth GK, Shi W (2013) The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. Nucleic Acids Res 41: e108. [PubMed]

66. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, et al. (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29: 15-21. [PubMed]

67. Kim D, Langmead B, Salzberg SL (2015) HISAT: a fast spliced aligner with low memory requirements. Nat Methods 12: 357-360. [PubMed]

68. Pepke S, Wold B, Mortazavi A (2009) Computation for ChIP-seq and RNA-seq studies. Nat Methods 6: S22-32. [PubMed]

69. Fejes AP, Robertson G, Bilenky M, Varhol R, Bainbridge M, et al. (2008) FindPeaks 3. 1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. Bioinformatics 24: 1729-1730. [PubMed]

70. Rozowsky J, Euskirchen G, Auerbach RK, Zhang ZD, Gibson T, et al. (2009) PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. Nat Biotechnol 27: 66-75. [PubMed]

71. Ji H, Jiang H, Ma W, Johnson DS, Myers RM, et al. (2008) An integrated software system for analyzing ChIP-chip and ChIP-seq data. Nat Biotechnol 26: 1293-1300. [PubMed]

72. Jothi R, Cuddapah S, Barski A, Cui K, Zhao K (2008) Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. Nucleic Acids Res 36: 5221-5231. [PubMed]

73. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, et al. (2008) Model-based analysis of ChIP-Seq (MACS). Genome Biol 9: R137. [PubMed]

74. Krebs W, Schmidt SV, Goren A, De Nardo D, Labzin L, et al. (2014) Optimization of transcription factor binding map accuracy utilizing knockout-mouse models. Nucleic Acids Res 42: 13051-13060. [PubMed]

75. Malone BM, Tan F, Bridges SM, Peng Z (2011) Comparison of four ChIP-Seq analytical algorithms using rice endosperm H3K27 trimethylation profiling data. PLoS One 6: e25260. [PubMed]

76. Laajala TD, Raghav S, Tuomela S, Lahesmaa R, Aittokallio T, et al. (2009) A practical comparison of methods for detecting transcription factor binding sites in ChIP-seq experiments. BMC Genomics 10: 618. [PubMed]

77. Johnson DS, Mortazavi A, Myers RM, Wold B (2007) Genome-wide mapping of in vivo protein-DNA interactions. Science 316: 1497-1502. [PubMed]

78. Nix DA, Courdy SJ, Boucher KM (2008) Empirical methods for controlling false positives and estimating confidence in ChIP-Seq peaks. BMC Bioinformatics 9: 523. [PubMed]

79. Feng J, Li W, Jiang T (2011) Inference of isoforms from short sequence reads. J Comput Biol 18: 305-321. [PubMed]

80. Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, et al. (2010) Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nat Biotechnol 28: 503-510. [PubMed]

81. Li JJ, Jiang CR, Brown JB, Huang H, Bickel PJ (2011) Sparse linear modeling of next-generation mRNA sequencing (RNA-Seq) data for isoform discovery and abundance estimation. Proc Natl Acad Sci U S A 108: 19867-19872. [PubMed]

82. Li W, Feng J, Jiang T (2011) IsoLasso: a LASSO regression approach to RNA-Seq based transcriptome assembly. J Comput Biol 18: 1693-1707. [PubMed]

83. Mezlini AM, Smith EJ, Fiume M, Buske O, Savich GL, et al. (2013) iReckon: simultaneous isoform discovery and abundance estimation from RNA-seq data. Genome Res 23: 519-529. [PubMed]

84. Tomescu AI, Kuosmanen A, Rizzi R, Mäkinen V (2013) A novel min-cost flow method for estimating transcript expression with RNA-Seq. BMC Bioinformatics 14 Suppl 5: S15. [PubMed]

85. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, et al. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28: 511-515. [PubMed]

86. Behr J, Kahles A, Zhong Y, Sreedharan VT, Drewe P, et al. (2013) MITIE: Simultaneous RNA-Seq-based transcript identification and quantification in multiple samples. Bioinformatics 29: 2529-2538. [PubMed]

87. Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, et al. (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol 33: 290-295. [PubMed]

88. Benevento M, Tonge PD, Puri MC, Hussein SM, Cloonan N, et al. (2014) Proteome adaptation in cell reprogramming proceeds via distinct transcriptional networks. Nat Commun 5: 5613. [PubMed]

89. Pearson K (1901) On lines and planes of closest fit to systems of points in space. Philosophical Magazine 2: 559-572. [PubMed]

90. Abdi H,Williams LJ (2010) Principal component analysis. Wiley Interdisciplinary Reviews: Computational Statistics 2: 433-459. [PubMed]

91. Yeung KY, Ruzzo WL (2001) Principal component analysis for clustering gene expression data. Bioinformatics 17: 763-774. [PubMed]

92. Ji H, Li X, Wang QF, Ning Y (2013) Differential principal component analysis of ChIP-seq. Proc Natl Acad Sci U S A 110: 6789-6794. [PubMed]

93. Huff JT, Plocik AM, Guthrie C, Yamamoto KR (2010) Reciprocal intronic and exonic histone modification regions in humans. Nat Struct Mol Biol 17: 1495-1499. [PubMed]

94. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25-29. [PubMed]

95. Huang da W, Sherman BT, Lempicki RA (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 37: 1-13. [PubMed]

96. Huang da W, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4: 44-57. [PubMed]

97. Mi H, Muruganujan A, Casagrande JT, Thomas PD (2013) Large-scale gene function analysis with the PANTHER classification system. Nat Protoc 8: 1551-1566. [PubMed]

98. Ghosh SK (2010) Basics of Bayesian methods. Methods Mol Biol 620: 155-178. [PubMed]

99. Wu J, Xie J (2010) Hidden Markov model and its applications in motif findings. Methods Mol Biol 620: 405-416. [PubMed]

100. Mathelier A, Wasserman WW (2013) The next generation of transcription factor binding site prediction. PLoS Comput Biol 9: e1003214. [PubMed]

101. Choi H, Nesvizhskii AI, Ghosh D, Qin ZS (2009) Hierarchical hidden Markov model with application to joint analysis of ChIP-chip and ChIP-seq data. Bioinformatics 25: 1715-1721. [PubMed]

102. Hu M, Yu J, Taylor JM, Chinnaiyan AM, Qin ZS (2010) On the detection and refinement of transcription factor binding sites using ChIP-Seq data. Nucleic Acids Res 38: 2154-2167. [PubMed]

103. Levinson M, Zhou Q (2014) A penalized Bayesian approach to predicting sparse protein-DNA binding landscapes. Bioinformatics 30: 636-643. [PubMed]

104. Yan H, Evans J, Kalmbach M, Moore R, Middha S, et al. (2014) HiChIP: a high-throughput pipeline for integrative analysis of ChIP-Seq data. BMC Bioinformatics 15: 280. [PubMed]

105. Ha N, Polychronidou M, Lohmann I (2012) COPS: detecting co-occurrence and spatial arrangement of transcription factor binding motifs in genome-wide datasets. PLoS One 7: e52055. [PubMed]

106. Cielalik M, Bekiranov S (2014) Combinatorial epigenetic patterns as quantitative predictors of chromatin biology. BMC Genomics 15: 76. [PubMed]

107. Boeva V, Surdez D, Guillon N, Tirode F, Fejes AP, et al. (2010) De novo motif identification improves the accuracy of predicting transcription factor binding sites in ChIP-Seq data analysis. Nucleic Acids Res 38: e126. [PubMed]

108. Klein HU, Schäfer M, Porse BT, Hasemann MS, Ickstadt K, et al. (2014) Integrative analysis of histone ChIP-seq and transcription data using Bayesian mixture models. Bioinformatics . [PubMed]

109. Thomas-Chollier M, Herrmann C, Defrance M, Sand O, Thieffry D, et al. (2012) RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets. Nucleic Acids Res 40: e31. [PubMed]

110. Wingender E, Dietze P, Karas H, Knüppel R (1996) TRANSFAC: A Database on Transcription Factors and Their DNA Binding Sites. Nucleic Acids Research 24: 238-241. [PubMed]

111. Mathelier A, Zhao X, Zhang AW, Parcy F, Worsley-Hunt R, et al. (2014) JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. Nucleic Acids Res 42: D142-147. [PubMed]

112. Maaskola J, Rajewsky N2 (2014) Binding site discovery from nucleic acid sequences by discriminative learning of hidden Markov models. Nucleic Acids Res 42: 12995-13011. [PubMed]

113. Marson A, Levine SS, Cole MF, Frampton GM, Brambrink T, et al. (2008) Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. Cell 134: 521-533. [PubMed]

114. Jaenisch R, Young R (2008) Stem cells, the molecular circuitry of pluripotency and nuclear reprogramming. Cell 132: 567-582. [PubMed]

115. Young RA (2011) Control of the embryonic stem cell state. Cell 144: 940-954. [PubMed]

116. Heng JC, Feng B, Han J, Jiang J, Kraus P, et al. (2010) The nuclear receptor Nr5a2 can replace Oct4 in the reprogramming of murine somatic cells to pluripotent cells. Cell Stem Cell 6: 167-174. [PubMed]

117. Dejosez M, Levine SS, Frampton GM, Whyte WA, Stratton SA, et al. (2010) Ronin/Hcf-1 binds to a hyperconserved enhancer element and regulates genes involved in the growth of embryonic stem cells. Genes Dev 24: 1479-1484. [PubMed]

118. Gontan C, Achame EM, Demmers J, Barakat TS, Rentmeester E, et al. (2012) RNF12 initiates X-chromosome inactivation by targeting REX1 for degradation. Nature 485: 386-390. [PubMed]

119. Chen PY, Feng S, Joo JW, Jacobsen SE, Pellegrini M (2011) A comparative analysis of DNA methylation across human embryonic stem cell lines. Genome Biol 12: R62. [PubMed]

120. Ma Z, Swigut T, Valouev A, Rada-Iglesias A, Wysocka J (2011) Sequence-specific regulator Prdm14 safeguards mouse ESCs from entering extraembryonic endoderm fates. Nat Struct Mol Biol 18: 120-127. [PubMed]

121. Bilodeau S, Kagey MH, Frampton GM, Rahl PB, Young RA (2009) SetDB1 contributes to repression of genes encoding developmental regulators and maintenance of ES cell state. Genes Dev 23: 2484-2489. [PubMed]

122. Schnetz MP, Handoko L, Akhtar-Zaidi B, Bartels CF, Pereira CF, et al. (2010) CHD7 targets active gene enhancer elements to modulate ES cell-specific gene expression. PLoS Genet 6: e1001023. [PubMed]

123. Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, et al. (2009) ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature 457: 854-858. [PubMed]

124. Ho L, Jothi R, Ronan JL, Cui K, Zhao K, et al. (2009) An embryonic stem cell chromatin remodeling complex, esBAF, is an essential component of the core pluripotency transcriptional network. Proceedings of the National Academy of Sciences of the United States of America 106: 5187-5191.

125. Lee BK, Bhinge AA, Iyer VR (2011) Wide-ranging functions of E2F4 in transcriptional activation and repression revealed by genome-wide analysis. Nucleic Acids Res 39: 3558-3573. [PubMed]

126. Kim SW, Yoon SJ, Chuong E, Oyolu C, Wills AE, et al. (2011) Chromatin and transcriptional signatures for Nodal signaling during endoderm formation in hESCs. Dev Biol 357: 492-504. [PubMed]

127. Mendenhall EM, Koche RP, Truong T, Zhou VW, Issac B, et al. (2010) GC-rich sequence elements recruit PRC2 in mammalian ES cells. PLoS Genet 6: e1001244. [PubMed]

128. Kagey MH, Newman JJ, Bilodeau S, Zhan Y, Orlando DA, et al. (2010) Mediator and cohesin connect gene expression and chromatin architecture. Nature 467: 430-435. [PubMed]

129. Walker E, Chang WY, Hunkapiller J, Cagney G, Garcha K, et al. (2010) Polycomb-like 2 Associates with PRC2 and Regulates Transcriptional Networks during Mouse Embryonic Stem Cell Self-Renewal and Differentiation. Cell stem cell 6: 153-166. [PubMed]

130. Yildirim O, Li R, Hung JH, Chen PB, Dong X, et al. (2011) Mbd3/NURD complex regulates expression of 5-hydroxymethylcytosine marked genes in embryonic stem cells. Cell 147: 1498-1510. [PubMed]

131. Jacobs FM, Greenberg D, Nguyen N, Haeussler M, Ewing AD, et al. (2014) An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons. Nature 516: 242-245. [PubMed]

132. Byrska-Bishop M, VanDorn D, Campbell AE, Betensky M, Arca PR, et al. (2015) Pluripotent stem cells reveal erythroid-specific activities of the GATA1 N-terminus. J Clin Invest 125: 993-1005. [PubMed]

133. Di Micco R, Fontanals-Cirera B, Low V, Ntziachristos P, Yuen SK, et al. (2014) Control of embryonic stem cell identity by BRD4-dependent transcriptional elongation of super-enhancer-associated pluripotency genes. Cell Rep 9: 234-247. [PubMed]

134. Castro-Diaz N, Ecco G, Coluccio A, Kapopoulou A, Yazdanpanah B, et al. (2014) Evolutionally dynamic L1 regulation in embryonic stem cells. Genes Dev 28: 1397-1409. [PubMed]

135. Beyer TA, Weiss A, Khomchuk Y, Huang K, Ogunjimi AA, et al. (2013) Switch enhancers interpret TGF-Î² and Hippo signaling to control cell fate in human embryonic stem cells. Cell Rep 5: 1611-1624. [PubMed]

136. Eijkelenboom A, Mokry M, Smits LM, Nieuwenhuis EE, Burgering BM (2013) FOXO3 selectively amplifies enhancer activity to establish target gene regulation. Cell Rep 5: 1664-1678. [PubMed]

137. Akdemir KC, Jain AK, Allton K, Aronow B, Xu X, et al. (2014) Genome-wide profiling reveals stimulus-specific functions of p53 during differentiation and DNA damage of human embryonic stem cells. Nucleic Acids Research 42: 205-223. [PubMed]

138. Liang Y, Franks TM, Marchetto MC, Gage FH, Hetzer MW (2013) Dynamic association of NUP98 with the human genome. PLoS Genet 9: e1003308. [PubMed]

139. Wang A, Yue F, Li Y, Xie R, Harper T, et al. (2015) Epigenetic priming of enhancers predicts developmental competence of hESC-derived endodermal lineage intermediates. Cell Stem Cell 16: 386-399. [PubMed]

140. Han J, Yuan P, Yang H, Zhang J, Soh BS, et al. (2010) Tbx3 improves the germ-line competency of induced pluripotent stem cells. Nature 463: 1096-1100. [PubMed]

141. Tsankov AM, Gu H, Akopian V, Ziller MJ, Donaghey J, et al. (2015) Transcription factor binding dynamics during human ES cell differentiation. Nature 518: 344-349. [PubMed]

142. Tanimura N, Saito M, Ebisuya M, Nishida E, Ishikawa F (2013) Stemness-related factor Sall4 interacts with transcription factors Oct-3/4 and Sox2 and occupies Oct-Sox elements in mouse embryonic stem cells. J Biol Chem 288: 5027-5038. [PubMed]

143. Bourque G, Leong B, Vega VB, Chen X, Lee YL, et al. (2008) Evolution of the mammalian transcription factor binding repertoire via transposable elements. Genome Res 18: 1752-1762. [PubMed]

144. Kunarso G, Chia NY, Jeyakani J, Hwang C, Lu X, et al. (2010) Transposable elements have rewired the core regulatory network of human embryonic stem cells. Nat Genet 42: 631-634. [PubMed]

145. Meshorer E, Misteli T (2006) Chromatin in pluripotent embryonic stem cells and differentiation. Nat Rev Mol Cell Biol 7: 540-546. [PubMed]

146. Bannister AJ, Kouzarides T (2011) Regulation of chromatin by histone modifications. Cell Res 21: 381-395. [PubMed]

147. Peng JC, Valouev A, Swigut T, Zhang J, Zhao Y, et al. (2009) Jarid2/Jumonji coordinates control of PRC2 enzymatic activity and target gene occupancy in pluripotent cells. Cell 139: 1290-1302. [PubMed]

148. Qiao Y, Wang R, Yang X, Tang K, Jing N (2015) Dual Roles of Histone H3 Lysine 9 Acetylation in Human Embryonic Stem Cell Pluripotency and Neural Differentiation. Journal of Biological Chemistry 290: 2508-2520. [PubMed]

149. Theunissen TW, Powell BE, Wang H, Mitalipova M, Faddah DA, et al. (2014) Systematic identification of culture conditions for induction and maintenance of naive human pluripotency. Cell Stem Cell 15: 471-487. [PubMed]

150. Loh KM, Ang LT, Zhang J, Kumar V, Ang J, et al. (2014) Efficient endoderm induction from human pluripotent stem cells by logically directing signals controlling lineage bifurcations. Cell Stem Cell 14: 237-252. [PubMed]

151. Gafni O, Weinberger L, Mansour AA, Manor YS, Chomsky E, et al. (2013) Derivation of novel human ground state naive pluripotent stem cells. Nature 504: 282-286. [PubMed]

152. Li B, Su T, Ferrari R, Li JY, Kurdistani SK (2014) A unique epigenetic signature is associated with active DNA replication loci in human embryonic stem cells. Epigenetics 9: 257-267. [PubMed]

153. Gifford CA, Ziller MJ, Gu H, Trapnell C, Donaghey J, et al. (2013) Transcriptional and epigenetic dynamics during specification of human embryonic stem cells. Cell 153: 1149-1163. [PubMed]

154. Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, et al. (2010) 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.

155. Phanstiel DH, Brumbaugh J, Wenger CD, Tian S, Probasco MD, et al. (2011) Proteomic and phosphoproteomic comparison of human ES and iPS cells. Nat Methods 8: 821-827. [PubMed]

156. Beerman I, Bock C, Garrison BS, Smith ZD, Gu H, et al. (2013) Proliferation-dependent alterations of the DNA methylation landscape underlie hematopoietic stem cell aging. Cell Stem Cell 12: 413-425. [PubMed]

157. Hon GC, Rajagopal N, Shen Y, McCleary DF, Yue F, et al. (2013) Epigenetic memory at embryonic enhancers identified in DNA methylation maps from adult mouse tissues. Nat Genet 45: 1198-1206. [PubMed]

158. Hosogane M, Funayama R, Nishida Y, Nagashima T, Nakayama K (2013) Ras-induced changes in H3K27me3 occur after those in transcriptional activity. PLoS Genet 9: e1003698. [PubMed]

159. Liu H, Chen Y, Lv J, Liu H, Zhu R, et al. (2013) Quantitative epigenetic co-variation in CpG islands and co-regulation of developmental genes. Sci Rep 3: 2576. [PubMed]

160. Moison C, Senamaud-Beaufort C, Fourrière L, Champion C, Ceccaldi A, et al. (2013) DNA methylation associated with polycomb repression in retinoic acid receptor Î² silencing. FASEB J 27: 1468-1478. [PubMed]

161. Neri F, Krepelova A, Incarnato D, Maldotti M, Parlato C, et al. (2013) Dnmt3L antagonizes DNA methylation at bivalent promoters and favors DNA methylation at gene bodies in ESCs. Cell 155: 121-134. [PubMed]

162. Wijetunga NA, Delahaye F, Zhao YM, Golden A, Mar JC, et al. (2014) The meta-epigenomic structure of purified human stem cell populations is defined at cis-regulatory sequences. Nat Commun 5: 5195. [PubMed]

163. Yan H, Zhang D, Liu H, Wei Y, Lv J, et al. (2015) Chromatin modifications and genomic contexts linked to dynamic DNA methylation patterns across human cell types. Sci Rep 5: 8410. [PubMed]

164. Smith ZD, Meissner A (2013) DNA methylation: roles in mammalian development. Nat Rev Genet 14: 204-220. [PubMed]

165. Kaneko S, Bonasio R, Saldaña-Meyer R, Yoshida T, Son J, et al. (2014) Interactions between JARID2 and noncoding RNAs regulate PRC2 recruitment to chromatin. Mol Cell 53: 290-300. [PubMed]

166. Ohnuki M, Tanabe K, Sutou K, Teramoto I, Sawamura Y, et al. (2014) Dynamic regulation of human endogenous retroviruses mediates factor-induced reprogramming and differentiation potential. Proceedings of the National Academy of Sciences 111: 12426-12431. [PubMed]

167. Santoni FA, Guerra J, Luban J (2012) HERV-H RNA is abundant in human embryonic stem cells and a precise marker for pluripotency. Retrovirology 9: 111. [PubMed]

168. Hamazaki N, Uesaka M, Nakashima K, Agata K, Imamura T (2015) Gene activation-associated long noncoding RNAs function in mouse preimplantation development. Development 142: 910-920. [PubMed]

169. Fort A, Yamada D, Hashimoto K, Koseki H, Carninci P (2015) Nuclear transcriptome profiling of induced pluripotent stem cells and embryonic stem cells identify non-coding loci resistant to reprogramming. Cell Cycle 14: 1148-1155. [PubMed]

170. Vallot C, Ouimette JF, Makhlouf M, Féraud O, Pontis J, et al. (2015) Erosion of X Chromosome Inactivation in Human Pluripotent Cells Initiates with XACT Coating and Depends on a Specific Heterochromatin Landscape. Cell Stem Cell 16: 533-546. [PubMed]

171. Dostie J, Richmond TA, Arnaout RA, Selzer RR, Lee WL, et al. (2006) Chromosome Conformation Capture Carbon Copy (5C): A massively parallel solution for mapping interactions between genomic elements. Genome Research 16: 1299-1309. [PubMed]

172. Meissner A, Gnirke A, Bell GW, Ramsahoye B, Lander ES, et al. (2005) Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. Nucleic Acids Res 33: 5868-5877. [PubMed]

173. Jacinto FV, Ballestar E, Esteller M (2008) Methyl-DNA immunoprecipitation (MeDIP): hunting down the DNA methylome. Biotechniques 44: 35, 37, 39 passim. [PubMed]

174. Serre D, Lee BH, Ting AH (2010) MBD-isolated Genome Sequencing provides a high-throughput and comprehensive survey of DNA methylation in the human genome. Nucleic Acids Res 38: 391-399. [PubMed]

175. Harris RA, Wang T, Coarfa C, Nagarajan RP, Hong C, et al. (2010) Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. Nat Biotechnol 28: 1097-1105. [PubMed]

176. van de Werken HJ, Landan G, Holwerda SJ, Hoichman M, Klous P, et al. (2012) Robust 4C-seq data analysis to screen for regulatory DNA interactions. Nat Methods 9: 969-972. [PubMed]

177. Raviram R, Rocha PP, Bonneau R, Skok JA (2014) Interpreting 4C-Seq data: how far can we go? Epigenomics 6: 455-457. [PubMed]

178. Gao F, Wei Z, Lu W, Wang K (2013) Comparative analysis of 4C-Seq data generated from enzyme-based and sonication-based methods. BMC Genomics 14: 345. [PubMed]

179. Splinter E, de Wit E, van de Werken HJ, Klous P, de Laat W (2012) Determining long-range chromatin interactions for selected genomic sites using 4C-seq technology: from fixation to computation. Methods 58: 221-230. [PubMed]

180. Kim D, Bae S, Park J, Kim E, Kim S, et al. (2015) Digenome-seq: genome-wide profiling of CRISPR-Cas9 off-target effects in human cells. Nat Methods 12: 237-243, 1 p following 243. [PubMed]

181. Wang X, Wang Y, Wu X, Wang J, Wang Y, et al. (2015) Unbiased detection of off-target cleavage by CRISPR-Cas9 and TALENs using integrase-defective lentiviral vectors. Nat Biotechnol 33: 175-178. [PubMed]