Computer and Statistical Analysis of Transcription Factor Binding and Chromatin Modifications by ChIP-seq data in Embryonic Stem Cell

Yuriy Orlov[1,2,4,*], Han Xu[1], Dmitri Afonnikov[1,4], Bing Lim[1], Jian-Chien Heng[1], Ping Yuan[1], Ming Chen[5], Junli Yan[1], Neil Clarke[1], Nina Orlova[3], Mikael Huss[1], Konstantin Gunbin[1], Nikolay Podkolodnyy[2,4], Huck-Hui Ng[1]

[1]Genome Institute of Singapore, Singapore 138672, Singapore
[2]Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
[3]Siberian University of Consumer Cooperatives, Novosibirsk, Russia
[4]Novosibirsk State University, Novosibirsk, Russia
[5]Zhejiang University, Hangzhou, China

Summary

Advances in high throughput sequencing technology have enabled the identification of transcription factor (TF) binding sites in genome scale. TF binding studies are important for medical applications and stem cell research. Somatic cells can be reprogrammed to a pluripotent state by the combined introduction of factors such as Oct4, Sox2, c-Myc, Klf4. These reprogrammed cells share many characteristics with embryonic stem cells (ESCs) and are known as induced pluripotent stem cells (iPSCs). The signaling requirements for maintenance of human and murine embryonic stem cells (ESCs) differ considerably. Genome wide ChIP-seq TF binding maps in mouse stem cells include Oct4, Sox2, Nanog, Tbx3, Smad2 as well as group of other factors. ChIP-seq allows study of new candidate transcription factors for reprogramming. It was shown that Nr5a2 could replace Oct4 for reprogramming. Epigenetic modifications play important role in regulation of gene expression adding additional complexity to transcription network functioning. We have studied associations between different histone modification using published data together with RNA Pol II sites. We found strong associations between activation marks and TF binding sites and present it qualitatively. To meet issues of statistical analysis of genome ChIP-sequencing maps we developed computer program to filter out noise signals and find significant association between binding site affinity and number of sequence reads. The data provide new insights into the function of chromatin organization and regulation in stem cells.

1 Introduction

Combination of chromatin immune precipitation and high throughput sequencing (ChIP-seq) has been used extensively to determine chromosome binding patterns of DNA-associated proteins as well as chromatin epigenetics modification marks [1-4]. The new generation of sequencing platforms provides orders of magnitude increase in the number of generated sequences and also raises challenges in the analysis and integration of genome scale data [5,6]. Such ChIP-seq genome wide TF binding maps in human include Oct4, Sox2 and Nanog in stem cells and related transcription factors in human (OCT4, MYC) [1-4,7]. Key problem of gene expression regulation analysis is detection of functional binding sites responsible for gene activation. TF binds in vivo to only a small fraction of sequence motifs or eligible (computationally predicted) binding sites in the genome to be defined experimentally.

* To whom correspondence should be addressed. Email: orlov@bionet.nsc.ru

Stem cell research is important application for molecular genetics, genomics and fundamental medicine. Embryonic stem cells (ESCs) were first derived from the mouse blastocysts [8]. These cells have the capacity for extensive self-renewal under in vitro culture conditions. Another hallmark of these cells is the ability to undergo lineage-specific differentiation to give rise to all somatic cell-types [1]. Mouse ESCs (mESCs) are able to maintain genetic stability, show high rate of homologous recombination and do not exhibit senescence. These properties allow them to be used for gene targeting for the production of genetically modified animals. Human embryonic stem cells (hESCs) provide the opportunities to study processes implicated in human developmental biology. [9]. Like mESCs, the hESCs can be propagated stably in culture and they can also differentiate into all three germ lineages. Robust self-renewal capability of these pluripotent hESCs makes them a renewable source for the generation of functional cells or tissues for future therapeutic applications and drug discovery. Although human and mouse embryonic stem cells (ESCs) share similar fundamental properties such as pluripotency and unique transcriptional network, they differ significantly in signalling pathways [1,7]. For example, LIF/STAT3 signaling that is critical for mESC self-renewal is instead dispensable for self-renewal of undifferentiated hESCs. BMP4 together with LIF supports expansion of undifferentiated mESCs, while BMP4 induces trophoblastic differentiation of hESCs [10].

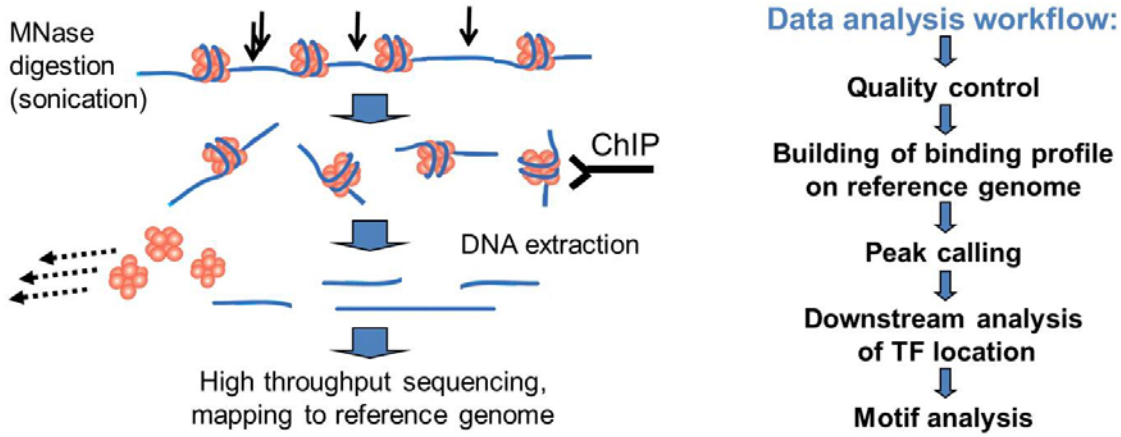

Figure 1: ChIP-seq workflow and data analysis.

Common scheme of ChIP-seq workflow and data analysis is given in Figure 1. For example, to gain insights into the transcriptional regulatory networks in embryonic stem (ES) cells, we used ChIP-seq to map the locations of 15 sequence specific TFs (Nanog, Oct4, STAT3 and others) [1-4] and transcription regulators (p300 and Suz12) in mouse [1]. These factors are known to play different roles in ES-cell biology as components of the cell development signaling pathways, self-renewal regulators, and key reprogramming factors [1,11]. The ChIP-seq data are available at http://t2g.bii.a-star.edu.sg, latest raw data are at GEO NCBI archive.

Induced pluripotent stem (iPS) cells can be obtained through the introduction of defined factors into somatic cells [11]. The reversion of somatic cells to pluripotent cells is commonly referred to as reprogramming. The combination of Oct4, Sox2 and Klf4 constitutes the minimal requirement for generating iPS cells from mouse embryonic fibroblasts (MEF). These cells are thought to resemble embryonic stem cells based on global gene expression analyses. So, ChIP-seq binding profiles and microarray expression experiments are necessary to reveal transcription regulation network and found candidate genes for reprogramming [1-4]. The ability to self-renew and differentiate is common for hESCs and mESCs. Both express genes which are associated with pluripotency. POU5F1 (encoding for OCT4) and NANOG are specifically up-regulated in undifferentiated ESCs. Upon differentiation, the expression of

these genes is down-regulated. OCT4 and NANOG are key components of the core transcriptional regulatory network in both mESCs and hESCs. These and other transcription regulators, including co-activator p300, show extensive co-binding at genomic sites and this binding configuration may be important for the expression of pluripotency-specific genes [1].

As example of candidate genes that have roles in pluripotency and fusion-mediated somatic cell reprogramming, Tbx3 was identified as a transcription factor that significantly improves the quality of iPS cells [4]. Genome-wide ChIP-seq analysis of Tbx3 binding sites in ESCs shows that Tbx3 regulates pluripotency-associated and reprogramming factors, in addition to sharing many common downstream regulatory targets with Oct4, Sox2, Nanog and Smad1. The study [4] underscores the intrinsic qualitative differences between iPS cells generated by different methods.

Extensive efforts were made in the identification of regulators for mESCs through the use of loss-of-function genetic approach. The profound effect of transcription factors is exemplified by the conversion of somatic cells into induced pluripotent stem cells (iPSCs) through the co-expression of four transcription factors, OCT4, SOX2, KLF4 and c-MYC [1]. Reprogramming of somatic cells provide unique opportunities for generating patient-specific pluripotent cells which may be used as in vitro models for studying and treating human diseases [7].

Using genome-wide RNAi screen to identify candidates required for self-renewing of hESCs PRDM14 was identified as a novel regulator of hESC and it is required to maintain the expression of POU5F1 [7]. Prdm14 is essential for the germ cell specification in mouse. Genome-wide location analysis revealed that PRMD14 binds to the proximal enhancer of POU5F1. In a gain-of-function assay, it was found that PRDM14 can enhance the efficiency of generating human iPSCs.

In general, close family members of reprogramming factors are also capable of replacing their counterparts, for example, Klf2 and Klf5 can replace Klf4, Sox1 and Sox5 can substitute for Sox2 while c-Myc can be replaced by N-myc and L-myc. Similarly, Esrrg can replace Esrrb in the reprogramming of MEFs [12]. However, Oct4 remains irreplaceable by other transcription factors including its close family members such as Oct1 and Oct6 [13]. Regulatory targeting of TFs could help in candidate selection for reprogramming genes. Close network connection was shown for key regulators in mESC (Figure 1) [1].

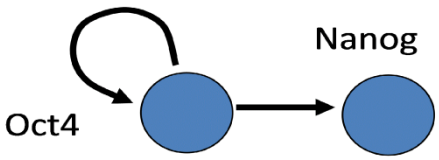

Figure 2: Example of TF regulatory events defined by ChIP-seq data (gene targeting by TF binding in promoter regions).

Protein kinases are key elements for intracellular signalling networks that can modulate gene expression in response to specific extracellular signals. An apparently interaction between the protein kinase and chromatin has been detected by Chromatin immunoprecipitation (ChIP) analysis. ERK2 is reported to act as a transcriptional repressor regulating interferon gamma signaling in mammalian cells [14]. The best-characterized transcription factor substrates of ERKs might be ternary complex factors (TCFs), including ELK1.

Epigenetic modifications play important role in regulation of gene expression adding additional complexity to transcription network functioning. We have studied associations between different histone modification using data for activation histone marks H3K4me3,

H3K4me1, H3K9ac and repressive histone marks H3K27me3 and H3K9me3 together with RNA Pol II sites in human [15]. We found strong associations between activation marks and TF binding sites and present it qualitatively, both for mouse [1] and human ChIP-seq data [15]. To meet issues of statistical analysis of genome ChIP-sequencing maps we developed computer program to filter out noise signals and find significant association between binding affinity and number of sequence tags. Moreover, TF regulation may have dose dependent effect targeting different group of genes, as it was shown on example of Smad2 binding in mouse [16].

Computer models of transcription factor binding based on chromatin modification sequencing allows better prediction of TF binding sites in genome scale. For example of ERalpha TF binding prediction with high accuracy could be done using chromatin modification data [15]. Our data and studies provide new insights into the function of chromatin organization in human genome. Recent works on special contacts of chromosome in nucleus confirmed interplay between chromatin modification and transcription factor binding in general in genome scale [17].

2 Analysis of TF binding in stem cells by ChIP-seq

2.1 ChIP-seq technology and data analysis

ChIP-seq technology provides a new powerful technique for localization of the most physically specific mammalian TF binding regions at a resolution of up to a few base pairs [1]. Immunoprecipitated DNA fragments could be sequenced directly using Roche 454, Illumina Solexa or SOLiD technologies. Disregarding the technological platform one can see set of statistical problems to be solved in the process of data analysis [5]. Most unexpectedly, all studies using this method have shown that TFs bind specifically to a surprisingly large number of genomic regions (extrapolated to 3,000-40,000 depending on the protein) [1]. The major fraction of these BS would not be validated by traditional methods. The application of this technology to mammalian genomes have been described in growing number of publications; and new statistical problems of peak calling and signal intensity normalization have been issued [5,6].

Typically, the extracted DNA (in [1-4]) are quantified and subjected to Illumina Solexa sequencing according to the manufacturer's instruction. The processed ChIP or FAIRE-enriched DNA fragments then are used for single read sequencing analysis. We used manufacture's software and in-house computational tools for mapping the sequence tags to the reference genome and clustering short sequences. In order to avoid potential PCR amplification bias, tags that shared the same mapping location on the same strand were removed. The uniquely-mapped reads with at most 2-mismatched were kept for further processing. The oriented 25-36 bp DNA reads (depending on sequencing library) were then extended to 200 bp regions to count clusters of overlapping sequences [1-4, 7, 15, 16]. The identified peaks were filtered in three steps (Figure 3).

First, an estimated false discovery rate (FDR) based on a random distribution of tags over the genome was used to remove random low-intensity peaks. Next, we further filtered the peaks based on the fold-change of peak intensity against an input DNA control library for same cell line. Downstream analysis of peak location relative to genes was fulfilled using custom-made software (Figure 3, bottom).

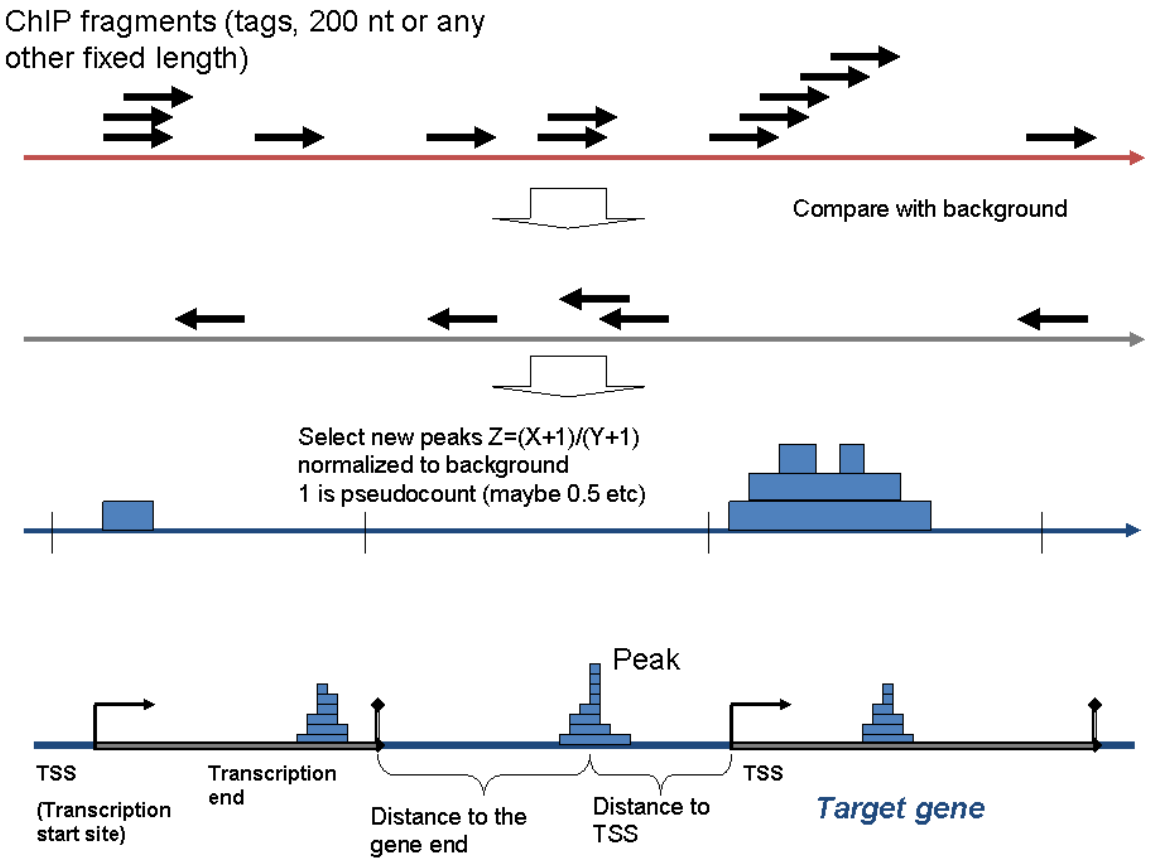

Figure 3: (top panels): Construction of binding profiles (piled DNA fragments) on chromosome coordinates and peak selection. (bottom): Peak location analysis. Selection of closest (nearest) gene for every cluster either to 5' or to 3'-end.

Co-occurrence analysis to study overlap of PRDM14 with other transcription factors binding sites was performed as described previously [1,7]. CTCF, OCT4 and NANOG ChIP-seq data sets were generated at GIS; KLF4, MYC, p300 and SOX2 ChIP-seq data were obtained from GEO (GSE18292 and GSE17917). We have processed external ChIP-seq data by MACS [18] program with the same parameters using control sequencing data. Heatmaps for co-localization were obtained from table data using R program environment.

Gene ontology analysis was done using PANTHER DB [19]. Extended GO categories were analyzed by DAVID (http://david.abcc.ncifcrf.gov/) and PANTHER software (data not shown).

2.2 Multiple binding of TFs in mESC

Analysis of TF binding peaks revealed close (in +/- 100 bp) co-localization of binding peaks from different factors. Such co-localization has preferences related to nature of TFs [1].

To address the problem of data integration in course of gene expression regulation we developed software for definition of potential target genes having TF binding sites in proximity of the transcription start site. Algorithm defines location of nearest RefSeq gene in the genome (either to 5' or 3' end), intronic/exonic location if inside gene borders and distance to from the TSS (transcription start site) for each TF binding site. The same approach was applied for classification of MTL (multiple binding loci) (Figure 4, 5) [1].

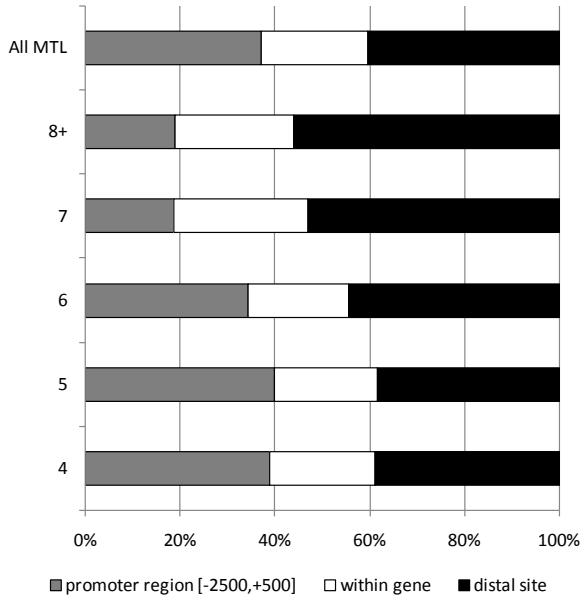

Figure 4: Distribution of TFBS in multiple regulatory loci relative to RefSeq genes depending on number of sites in the cluster.

Significant fractions of MTL are located in proximal promoter (2.5Kb upstream to TSS, or 0.5Kb downstream) depending on size of MTL (Figure 4). In MTL we have selected two groups associated preferably to Nanog/Oct4/Sox2 or Myc-related TF [1]. Relative location of Nanog-like MTL tend to be in distal regions from TSS indicating distal regulation enhancer, while MTL containing Myc TF tend to be in proximal gene promoters.

Pattern of TF binding location relative to gene correlates also with preferences to promoter/distal binding. Nanog and Oct4 locate preferably in distal sites [1], that later was confirmed by co-localization studies from independent experiments [2-4].

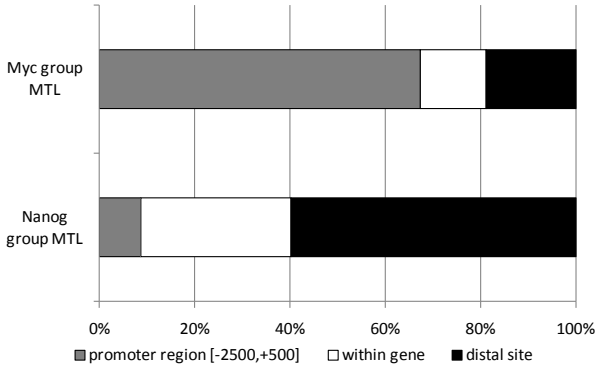

Figure 5: Pattern of promoter location of MTL depending on type of transcription factors in the cluster – Myc and Nanog groups.

To validate potential target we used microarray expression data. We have shown enrichment of differentially expressed genes in sets of target genes. We analyzed the chromosomal profile of tags and found correlation with chromatin structure and histone methylation patterns [1].

Figure 6: Binding motifs (logo) found in ChIP-seq peaks.

Figure 6 contains example of binding motifs obtained by sequence analysis of over-represented oligonucleotides in ChIP-seq data in mESC. Large samples of experimentally verified genome sequences from ChIP-seq experiments allow confirm and update known binding matrices and reveal some features characterising for in vivo binding. Important feature is symmetry of motifs for some TFs known to bind as dimmers. At the same time motifs in genome regions may not overlap and form ordered arrays named MTL (multiple binding loci) [1].

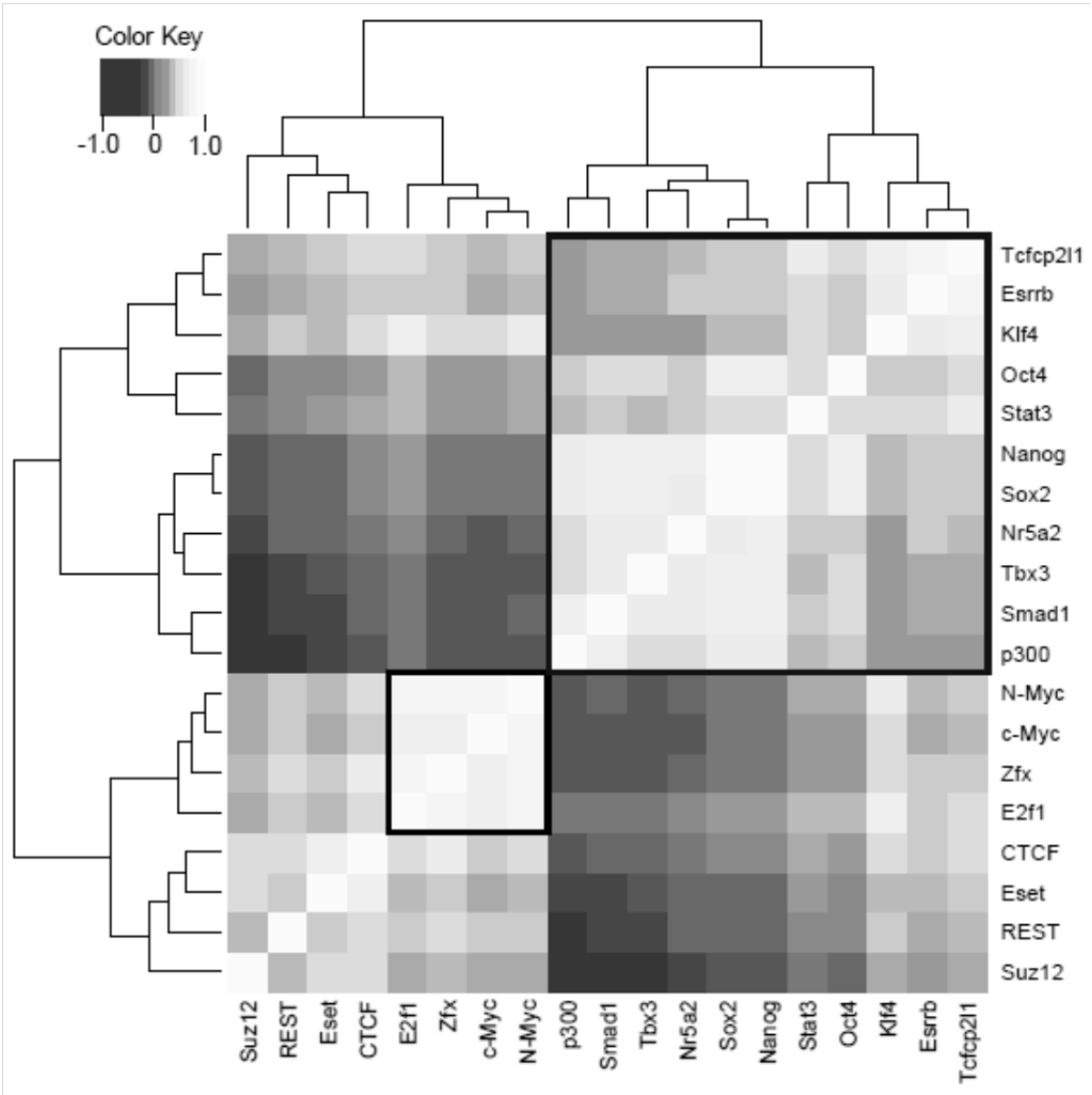

Figure 7: Co-occurrence of transcription factor groups in mESCs.

Figure 7 shows combined figure of co-localization of TF binding using ChIP-seq data in mESC from different experiments [1-4] including unrelated TF such as REST and CTCF (GIS sequencing data). TFs have been clustered along both axes based on the similarity in their co-localization with other factors. Colours in the heat map reflect the co-localization frequency of each pair of TFs (bright means more frequently co-localized, dark means less).

2.3 Key TFs for maintenance and reprogramming of ESC

2.3.1 Genome-wide binding analysis of Nr5a2 in ESC

Nuclear receptors Nr5a2 is important for reprogramming [3]. It was shown via mutation analysis that the DNA-binding capability of Nr5a2 is important for proper binding of the nuclear receptor to promoter/enhancer regions of target genes to initiate the reprogramming process in MEFs [3]. Unlike most nuclear receptors which function as dimers, Nr5a2 is able to bind DNA in its monomeric state The DNA binding is crucial for the reprogramming function of Nr5a2 while ligand binding is dispensable for its role in reprogramming.

Peak calling of the Nr5a2 ChIP-seq data (8,023 427 uniquely mapped tags) was carried out using MACS with a p value cut-off of 1e-9 and 3,346 peaks were generated. The control anti-HA ChIP-seq library contained 13,001,272 uniquely mapped tags. Enriched motifs were identified by the *de novo* motif discovery tool MEME using 200-bp sequences centered on the ChIP-seq peaks. Co-occurrence analysis to study overlap of Nr5a2 binding sites with binding sites of other important transcription factors was performed with Nr5a2 ChIP-seq data and data set generated from previous study [1] (see Figure 7).

We used a *de novo* motif discovery algorithm MEME and uncovered a known Nr5a2 motif enriched in the dataset. More importantly, from our pairwise co-occurrence analyses we find that Nr5a2 tends to co-localize with Nanog, Oct4, Sox2, Smad1 and Esrrb (Figure 7). This result associates Nr5a2 with the previously reported Nanog-Oct4-Sox2 cluster [1,3]. In addition, this high degree of co-localizations suggests that Nr5a2 share many common target genes with important pluripotent and self-renewal factors Oct4, Sox2 and Nanog [3]. Combining Nr5a2 ChIP-seq data with microarray analysis of Nr5a2 knockdown shows that several genes bound by Nr5a2, such as Nanog, were also regulated by it [3].

Table 1: Number of genes-Nr5a2 targets depending on distance to TSS.

Distance to TSS, bp	#Nr5a2 targets gene symbols	# confirmed by microarray (knockdown)	Fraction of targets confirmed by microarray
<50000	3507	247	0.070
<40000	2989	229	0.077
<30000	2397	190	0.079
<20000	1740	144	0.083
<10000	1035	82	0.079
<5000	600	49	0.082
<1000	181	13	0.072

Using Table 1 we can see that reasonable number of gene-targets of Nr5a2 binding could be selected using 50Kbp or 20Kbp threshold. Number of genes confirmed by microarray is up to 2 hundreds. Relative fraction of confirmed targets is not big, about 7%. But all gene numbers confirmed by the microarray are statistically significant.

2.3.2 Interplay between TF binding: ERK2 ChIP-Seq identified ELK1 as a functionally important co-motif in hESC

The self-renewal of pluripotent human ES cells has been shown to require extrinsic stimulation by the FGF, Activin and IGF signaling pathways and the expression of the transcription factors Oct4 and NANOG. However, the network of interactions among extrinsic and intrinsic determinants of ES cell pluripotency is currently poorly understood. ERK activity has been shown to be required for the maintenance of pluripotency. Furthermore, ERK kinase was recently shown to occupy gene promoters suggesting that ERK2 may be frequent occupants of signal-regulated gene promoters [20]. To gain insights into how signaling control the transcriptional regulatory networks in hES cells, we performed chromatin immuneprecipitation combined with high-throughput sequencing (ChIP-Seq) to map the locations of ERK2 in the ES-cell genome using chromatin prepared from H1 hESCs cultured in CM. Interesting that genome-wide screen for binding sites for ERK2 in hESC reveals that ERK2 binds to pluripotency genes and lineage specific transcription factors besides cell proliferation control genes. Phosphorylated ERK can activate a number of transcription factors, including ETS, AP-1, MYC and CREB. Consistent with this, bioinformatics analysis of ERK2 binding peak regions revealed the presence of multiple conserved transcription factor-binding sites, including binding sites for ELK1, CREB, E2F1, SP1 and TEAD1.

2.3.3 Tbx3 analysis

To better understand how Tbx3 may contribute to improving iPS cell quality, we performed Solexa ChIP-sequencing to uncover the direct regulatory targets of Tbx3 in mouse ESCs [4]. Strikingly, hierarchical clustering of Tbx3 with the previously mapped ESC factors revealed that it shares a large number of common binding sites with the classic pluripotency-associated transcription factors Oct4, Sox2, Nanog and Smad1 (Figure 7). Tbx3 is also found to target the ESC factors *Oct4*, *Sox2*, *Sall4*, *Lefty1*, *Lefty2*, *Zfp42*, as well as reprogramming factors *Klf2*, *Klf4*, *n-Myc* and *c-Myc* [4].

2.3.4 Eset ChIP-seq studies

Peak calling based on the Eset ChIP-seq data [2] was performed using MACS [18] with a p-value cutoff of 1e-12 resulted in 4,633 peaks. To determine regions that are marked by H3K9me3 and significantly affected by Eset knockdown, we search for regions that were significantly depleted in H3K9me3 after *Eset* RNAi. The program is suitable for histone modification enriched regions which detects changes over broader genomic regions, rather than the localized peaks detected by most peak caller programs. To define target genes we counted all the RefSeq genes having at least one ChIP-seq peak in +/-50Kbp from TSS. In order to define a core set of genes regulated by Eset, we selected all RefSeq genes that had at least one Eset binding peak as well as an Eset-dependent H3K9me3-enriched region. This resulted in a list of 1283 genes [2]. If Eset binding sites was not overlap directly with H3K9me regions, but resided in same gene promoter, then gene was considered as target gene.

We assessed the overlap of the Eset ChIP-seq peaks with other transcription factors (Oct4, Sox2, Nanog, Suz12) in mouse ES cells by intersecting the peak list with data from a previous study [1,2] (Figure 7). We allowed up to 200 bp between the borders of two peaks. Instead of assessing overrepresentation by comparing the observed overlaps to overlaps with random regions, we used a control library generated from sequencing input DNA. To construct control sites we used low threshold in the MACS program, and then select randomly 40,000 sites in

total. This was done to partially correct possible bias due to uneven fragmentation and read mapping. The observed number of overlaps between TFs and Eset was then compared to this baseline, and statistical significance of Eset enrichment for each TF was estimated (Table 1).

Table 2: Binding of Eset and histone methylation targets by ChIP-seq data.

	Eset	H3K9me3	Core Eset+H3K9me3
ChIP-seq peaks	4633	10798	1890
Gene Targets	2353	4169	1283

2.3.5 Co-binding analysis of TFs in human genome

Using available ChIP-seq libraries we compared co-occupation of binding sites in genome. We did it for mouse ESC previously and extended the work to new available TFs [7,13].

Figure 8 shows co-localization of TF binding in hESC in genome scale. The same observation for distinct clusters of OCT4-related and MYC-related groups is true for both mouse and human ESC.

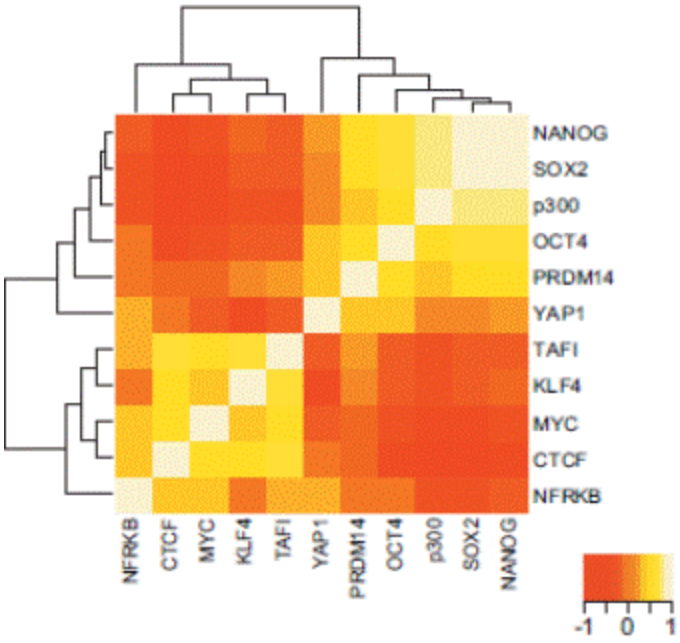

Figure 8: Co-occupation map of TF in hESC.

Note similar clustering of TF binding in mouse and human ESCs (Figure 7 and 8).

De novo computational method identified an over-represented PRDM14 motif [7]. Interestingly, PRDM14 shows co-binding with OCT4, SOX2, NANOG and co-activator p300, indicating that the PRDM14 circuitry is integrated to the core hESC transcriptional regulatory network.

3 Discussion

Important problem of biomedical studies is transcription networks governing stem cell maintenance. Despite the species-specific differences in the wiring of key transcription factors to the genome, certain ESC transcription factors can exert dominant effects on pluripotency-associated cellular identity in both mouse and human cells [13]. Recently, the repertoire of

transcription factors associated with reprogramming was updated by Tbx3, a T-box factor that could significantly improve the germline competency of murine iPSCs [4].

Analysis of genomic sequences surrounding highest peaks (mapped ChIP-seq clusters) yields de novo motifs, sometimes not annotated in published manually curated databases. Some genomic regions have multiple TF occupancy (bound by several different transcription factors) challenging new bioinformatics problems. We found more than 3 thousands such multiple transcription loci (MTL) formed by 4 or more different TF sites in the mouse genome [1]. To distinguish regions enriched beyond random expectation from noise, we developed an algorithm that takes into account both the abundance of and the signal intensities of the bound regions for each transcription factor.

The process of DNA reads mapping to the reference genome can bias the analysis toward genomic regions with unique and complex sequence patterns, requiring adjustment of the expected chance to observe moderate peaks in ChIP sequence density [6]. After mapping unique DNA sequence reads onto genome we have obtained 3-10 millions positions and 3,000-50,000 peaks for each TF [1-4]. Difference in site numbers for TFs from different experiments demands development of new methods for statistical normalization and comparison of ChIP-seq peak calling. We made statistical estimates of genome coverage and false discovery rate of the total number of binding sites in genome based on ChIP-seq data that prove high quality of the data. Method of simulation backward from observed number of ChIP-seq clusters (peaks) and extrapolation forward to estimate total number of specific TF binding sites could be used for any TF sequencing data. Statistical estimations were proved by independent computer prediction methods and sequencing experiments [1]. We found strong correlations between binding motif and ChIP-seq peaks independently for ChIP-PET and ChIP-seq experiments [5,15].

We developed statistical approaches for ChIP-seq transcription factor binding data and for definition of multiple TF binding events. Such combinations of different TFs on the same genome assembly (multiple transcription loci) allowed us to describe potential enhancers. Analysis of co-localization of transcription factor binding sites at genome scale revealed distinct patterns of sites related to proximal/distal location with respect to gene borders. Large fraction (up to 40%) of multiple TF binding loci are far from gene borders indicating to distal type of gene expression regulation.

Integration of genome annotations of regulatory regions for TF binding and overlapping non-coding transcripts allows define special classes of regulatory events [21] and looping of chromosome sequences. Tightly enclosed chromatin interaction centres could help achieve and maintain high local concentration of transcription components for efficient cycling of transcriptional machinery on target gene templates, as it suggested by expanded studies [17].

Acknowledgements

The statistics of TF binding using publicly available ChIP-seq data was updated using high-throughput computer cluster “Bioinformatics” SB RAS. The implementation of the programs for genomics data analysis was supported by the Russian Ministry of Education and Science (project No. 07.514.11.4003). Y.O. is grateful to RFBR (11-04-01888).

References

[1] X. Chen, H. Xu, P. Yuan, F. Fang, M. Huss, V.B. Vega, E. Wong, Y. L. Orlov, W. Zhang, J. Jiang, Y. H. Loh, H. C. Yeo, Z. X. Yeo, V. Narang, K. R. Govindarajan, B. Leong, A. Shahab, Y. Ruan, G. Bourque, W. K. Sung, N. D. Clarke, C. L. Wei and

H. H. Ng. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell, 133(6):1106-17, 2008.

[2] P. Yuan, J. Han, G. Guo, Y. L. Orlov, M. Huss, Y. H. Loh, L. P. Yaw, P. Robson, B. Lim and H. H. Ng. Eset partners with Oct4 to restrict extraembryonic trophoblast lineage potential in embryonic stem cells. Genes and development, 23(21):2507-20, 2009.

[3] J. C. Heng, B. Feng, J. Han, J. Jiang, P. Kraus, J. H. Ng, Y. L. Orlov, M. Huss, L. Yang, T. Lufkin, B. Lim and H.H. Ng. The nuclear receptor Nr5a2 can replace Oct4 in the reprogramming of murine somatic cells to pluripotent cells. Cell Stem Cell, 6(2):167-74, 2010.

[4] J. Han, P. Yuan, H. Yang, J. Zhang, B. S. Soh, P. Li, S. L. Lim, S. Cao, J. Tay, Y. L. Orlov, T. Lufkin, H. H. Ng, W. L. Tam and B. Lim. Tbx3 improves the germ-line competency of induced pluripotent stem cells. Nature, 463(7284):1096-100, 2010.

[5] V. A. Kuznetsov, Y. L. Orlov, C. L. Wei and Y. Ruan. Computational analysis and modeling of genome-scale avidity distribution of transcription factor binding sites in chip-pet experiments. Genome informatics, 19:83-94, 2007.

[6] Y. L. Orlov, M. E. Huss, R. Joseph, H. Xu, V. B. Vega, Y. K. Lee, W. S. Goh, J. S. Thomsen, E. C. Cheung, N. D. Clarke and H. H. Ng. Genome-wide statistical analysis of multiple transcription factor binding sites obtained by ChIP-seq technologies. In: Proceedings of the 1st ACM Workshop on Breaking Frontiers of Computational Biology (CompBio '09). ACM, New York, NY, 11-18, 2009.

[7] N.-Y. Chia, Y.-S. Chan, B. Feng, X. Lu, Y. L. Orlov, D. Moreau, P. Kumar, L. Yang, J. Jiang, M.-S. Lau, M. Huss, B.-S. Soh, B.-S. Kraus, T. Lufkin, B. Lim, N. Clarke, F. Bard and H. H. Ng. A genome-wide RNAi screen identifies PRDM14 as a regulator of POU5F1 and human embryonic stem cell identity. Nature, 468(7321): 316-20, 2010.

[8] A. G. Smith. Embryo-derived stem cells: of mice and men. Annual review of cell and developmental biology, 17, 435–462, 2001.

[9] J. A. Thomson, J. Itskovitz-Eldor, S. S. Shapiro, M. A. Waknitz, J. J. Swiergiel, V. S. Marshall and J. M. Jones. Embryonic stem cell lines derived from human blastocysts. Science, 282, 1145-7, 1998.

[10] R. H. Xu, T. L. Sampsell-Barron, F. Gu, S. Root, R. Peck M., G. Pan, J. Yu, J. Antosiewicz-Bourget, S. Tian, R. Stewart and J. A. Thomson. NANOG is a direct target of TGFbeta/activin-mediated SMAD signaling in human ESCs. Cell Stem Cell, 3, 196-206, 2008.

[11] K. Takahashi and S. Yamanaka. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell, 126, 663–676, 2006.

[12] B. Feng, J. Jiang, P. Kraus, J. H. Ng, J. C. Heng, Y. S. Chan, L. P. Yaw, W. Zhang, Y. H. Loh, J. Han, V. B. Vega, V. Cacheux-Rataboul, B. Lim, T. Lufkin and H.H. Ng. Reprogramming of fibroblasts into induced pluripotent stem cells with orphan nuclear receptor Esrrb. Nature Cell Biology, 11: 197-203, 2009.

[13] J. C. Heng, Y. L. Orlov and H. H. Ng. Transcription Factors for the Modulation of Pluripotency and Reprogramming. In: Cold Spring Harbor Symposia on Quantitative Biology, 75:237-44, 2010.

[14] G. Hu, J. Kim, Q. Xu, Y. Leng, S. H. Orkin and S. J. Elledge. A genome-wide RNAi screen identifies a new transcriptional module required for self-renewal. Genes Development, 23, 837-48, 2009.

[15] R. Joseph, Y. L. Orlov, M. Huss, W. Sun, S. L. Kong, L. Ukil, Y. F. Pan, G. Li, M. Lim, J. S. Thomsen, Y. Ruan, N. D. Clarke, S. Prabhakar, E. Cheung and E. T. Liu. Integrative model of genomic factors for determining binding site selection by estrogen receptor α. Molecular Systems Biology, 6:456, 2010.

[16] K. L. Lee, S. K. Lim, Y. L. Orlov, Y. Yit le, H. Yang, L. T. Ang, L. Poellinger and B. Lim. Graded Nodal/Activin signaling titrates conversion of quantitative phospho-Smad2 levels into qualitative embryonic stem cell fate decisions. PLoS Genetics, 7(6):e1002130, 2011.

[17] G. Li, X. Ruan, R. K. Auerbach, K. S. Sandhu, M. Zheng, P. Wang, H. M. Poh, Y. Goh, J. Lim, J. Zhang, H. S. Sim, S. Q. Peh, F. H. Mulawadi, C. T. Ong, Y. L. Orlov, S. Hong, Z. Zhang, S. Landt, D. Raha, G. Euskirchen, C. L. Wei, W. Ge, H. Wang, C. Davis, K. I. Fisher-Aylor, A. Mortazavi, M. Gerstein, T. Gingeras, B. Wold, Y. Sun, M. J. Fullwood, E. Cheung, E. Liu, W. K. Sung, M. Snyder and Y. Ruan. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell, 148(1-2):84-98, 2012.

[18] Y. Zhang, T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nusbaum, R. M. Myers, M. Brown, W. Li and X. S. Liu. Model-based analysis of ChIP-Seq (MACS). Genome Biology, 9(9): R137, 2008.

[19] H. Mi, B. Lazareva-Ulitsky, R. Loo, A. Kejariwal, J. Vandergriff, S. Rabkin, N. Guo, A. Muruganujan, O. Doremieux, M. J. Campbell, H. Kitano and P. D. Thomas. The PANTHER database of protein families, subfamilies, functions and pathways. Nucleic Acids Research, 31:334-341, 2005.

[20] H. M. Zhang, L. Li, N. Papadopoulou, G. Hodgson, E. Evans, M. Galbraith, M. Dear, S. Vougier, J. Saxton and P. E. Shaw. Mitogen-induced recruitment of ERK and MSK to SRE promoter complexes by ternary complex factor Elk-1. Nucleic Acids Research, 36, 2594-2607, 2008.

[21] O. V. Grinchuk, P. Jenjaroenpun, Y. L. Orlov, J. Zhou and V. A. Kuznetsov. Integrative analysis of the human cis-antisense gene pairs, miRNAs and their transcription regulation patterns. Nucleic Acids Research, 38(2):534-47, 2010.