BIOINFORMATICS ORIGINAL PAPER Vol. 25 no. 15 2009, pages 1952–1958
doi:10.1093/bioinformatics/btp340

Data and text mining

A clustering approach for identification of enriched domains from histone modification ChIP-Seq data

Chongzhi Zang[1], Dustin E. Schones[2], Chen Zeng[1], Kairong Cui[2], Keji Zhao[2] and Weiqun Peng[1,*]

[1]Department of Physics, The George Washington University, Washington, DC 20052 and [2]Laboratory of Molecular Immunology, National Heart Lung and Blood Institute, NIH, Bethesda, MD 20892, USA

Received on March 3, 2009; revised on May 7, 2009; accepted on May 27, 2009

Advance Access publication June 8, 2009

Associate Editor: Joaquin Dopazo

ABSTRACT

Motivation: Chromatin states are the key to gene regulation and cell identity. Chromatin immunoprecipitation (ChIP) coupled with high-throughput sequencing (ChIP-Seq) is increasingly being used to map epigenetic states across genomes of diverse species. Chromatin modification profiles are frequently noisy and diffuse, spanning regions ranging from several nucleosomes to large domains of multiple genes. Much of the early work on the identification of ChIP-enriched regions for ChIP-Seq data has focused on identifying localized regions, such as transcription factor binding sites. Bioinformatic tools to identify diffuse domains of ChIP-enriched regions have been lacking.

Results: Based on the biological observation that histone modifications tend to cluster to form domains, we present a method that identifies spatial clusters of signals unlikely to appear by chance. This method pools together enrichment information from neighboring nucleosomes to increase sensitivity and specificity. By using genomic-scale analysis, as well as the examination of loci with validated epigenetic states, we demonstrate that this method outperforms existing methods in the identification of ChIP-enriched signals for histone modification profiles. We demonstrate the application of this unbiased method in important issues in ChIP-Seq data analysis, such as data normalization for quantitative comparison of levels of epigenetic modifications across cell types and growth conditions.

Availability: http://home.gwu.edu/∼wpeng/Software.htm

Contact: wpeng@gwu.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Covalent modifications of chromatin, including DNA methylation and histone modifications, play critical roles in gene regulation and cell lineage determination and maintenance (Bernstein *et al.*, 2007; Felsenfeld and Groudine, 2003). Defects in these epigenetic controls have been implicated in many pathological conditions in humans. Genome-scale profiling of these epigenetic marks has been dramatically facilitated by the recent progress in the ultra high-throughput massively parallel sequencing technologies (Barski *et al.*, 2007; Mikkelsen *et al.*, 2007). ChIP-Seq combines chromatin immunoprecipitation (ChIP) with high-throughput sequencing to map genome-wide chromatin modification profiles and transcription factor (TF) binding sites. It is characterized by high resolution, a quantitative nature, cost effectiveness and no complication due to probe hybridization as encountered in ChIP-chip assays (Schones and Zhao, 2008). A large amount of data has recently been generated using the ChIP-Seq technique, and these datasets call for new analysis algorithms.

Binding of TFs is mainly governed by their sequence specificity and therefore is typically associated with very localized ChIP-Seq signals in the genome. A number of algorithms have been developed to find the exact locations of TF binding sites from ChIP-Seq data (Chen *et al.*, 2008; Fejes *et al.*, 2008; Ji *et al.*, 2008; Johnson *et al.*, 2007; Jothi *et al.*, 2008; Kharchenko *et al.*, 2008; Nix *et al.*, 2008; Rozowsky *et al.*, 2009; Valouev *et al.*, 2008; Zhang *et al.*, 2008a). In contrast, the signals for histone modifications, histone variants and histone-modifying enzymes are usually diffuse and lack of well-defined peaks, spanning from several nucleosomes to large domains encompassing multiple genes (Barski *et al.*, 2007; Pauler *et al.*, 2009; Wang *et al.*, 2008; Wen *et al.*, 2009) (see, e.g. Figure S1). The detection of diffuse signals often suffers from high noise level and lack of saturation in sequencing coverage. These generally weak signals render approaches seeking strong local enrichment, such as those peak-finding algorithms used in finding TF binding sites, inadequate.

Many modification marks are known to form broad domains (Barski *et al.*, 2007; Wang *et al.*, 2008). This is believed to be helpful in stabilizing the chromatin state and propagating such states through cell division robustly (Bernstein *et al.*, 2007). A well-studied case is the trimethylation of histone H3 lysine 9 (H3K9me3). H3K9me3 recruits HP1 via its chromodomain. HP1 in turn recruits H3K9 methyltransferase Suv39h, which modifies H3K9 on other histones in the vicinity, thereby self-propagating the heterochromatin state (Aagaard *et al.*, 1999; Bannister *et al.*, 2001; Lachner *et al.*, 2001). Another example is the trimethylation of histone H3 lysine 27 (H3K27me3). H3K27me3 is generated by the activity of the Polycomb complex, PRC2, and is believed to recruit the PRC1 complex (Schwartz and Pirrotta, 2007). In *Drosophila*, it has been suggested that the spreading of H3K27me3

*To whom correspondence should be addressed.

results from looping action of PRC1 and PRC2 that both anchor at the polycomb response elements (Schwartz and Pirrotta, 2007) with nucleosomes at a distance. Recent experiments in human cells indicate direct recruitment of PRC2 by H3K27me3 (Hansen *et al.*, 2008), suggesting a mechanism for the spreading of H3K27me3. In addition to histone methylation, the more dynamic histone acetylation marks also cluster, and several histone acetyltransferases contain bromodomains that specifically bind acetylated histones (Dodd *et al.*, 2007; Jacobson *et al.*, 2000; Owen *et al.*, 2000).

Motivated by the mounting evidence of recruitment by modified histones of their respective enzymes, we develop a spatial clustering approach for the identification of ChIP-enriched regions (SICER) in histone modification data. A central feature of our method is pooling together signals from all the nucleosomes located together in the same modification state. This feature improves the signal-to-noise ratio and is especially helpful in dealing with the difficult case of diffuse enrichment covering extended genomic regions produced by histone modifications, for which enrichment at any short distance of one or several nucleosomes does not appear to be significant enough.

Our method involves scoring each potential ChIP-enriched domain according to the collective profile of enrichment on the domain. We developed a mathematical theory for the score distribution in a genomic background model of random reads, and employed this theory to identify spatial clusters, large and small, unlikely to appear by chance. Utilizing a control library, we identified a set of candidate domains that exhibit ChIP signal clustering using the random background model, and compare the strength of the ChIP signal with that of the control signal at each candidate domain to determine the significance of enrichment. Using a scaling approach for evaluation of false positives that is based on the digitized nature of ChIP-Seq data, and two datasets with experimental validation, we demonstrated that SICER outperforms other ChIP-Seq methods in dealing with histone modification data. Furthermore, we demonstrated its use as an unbiased general noise filter in such important issues in the statistical analysis of ChIP-Seq data as data normalization, and scaling analysis of sequencing coverage (see Supplementary Material).

2 METHODS

2.1 The island approach

2.1.1 Scoring scheme We partition the genome of effective length L into non-overlapping windows of size w. We define the score s for a window with l reads to be $s(l) = -\log P(l,\lambda)$. $P(l,\lambda)$ is a Poisson distribution parameterized by the average number of reads in a window $\lambda = wN/L$, where N is the total number of reads in the ChIP-Seq library. Given this definition, the scores of a window represents the negative logarithm of the probability of finding l reads in the window if the reads can land anywhere on the genome with equal probability, i.e. a background model of random reads. The scores from clusters of windows are additive, representing the negative logarithm of joint probability of finding the observed configuration in a random background model. The higher the score, the less likely the observed profile occurs by chance.

2.1.2 Island definition We assign each window as 'eligible ('ineligible'), if the read count in this window is equal to or above (below) a read-count threshold l_0. We determine l_0 by a P-value requirement based on a Poisson distribution.

$$\sum_{l=l_0}^{\infty} P(l,\lambda) \le p_0. \tag{1}$$

Therefore, l_0 depends on the size of the ChIP-Seq library. The 'eligible' windows are separated by gaps, which are the collection of 'ineligible' windows in between two neighboring 'eligible' windows. A gap of size m contains m 'ineligible' windows. We identify islands as clusters of 'eligible' windows separated by gaps of size less than or equal to a predetermined parameter g. When $g = 0$, an island is formed by an uninterrupted stretch of 'eligible' windows. The score of an island is the *aggregate* score of all 'eligible' windows on this island. An illustration of the definition of islands is shown in Figure 1a.

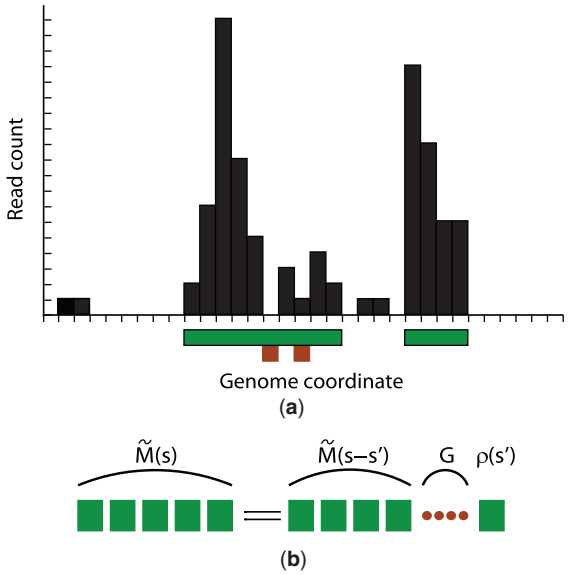

Fig. 1. (**a**) Schematic illustration of definition of islands. Shown is a segment of a genomic landscape of ChIP-Seq reads. The x-axis denotes the genome coordinates, where each interval represents a window. The y-axis denotes the read count. Each black vertical bar represents the read count in the respective window. The regions underlined by the green horizontal bars are the two identified islands under $g = 1$ and $l_0 = 2$. The two windows underlined by brown boxes are gaps in the first island. (**b**) Schematic illustration of the recursion relation in Equation (6).

2.1.3 Recursion relation for the probability of an island with a given score in a random background To derive the island score statistics in a random background model, we seek the probability $M(s)$ of finding an island of score s starting at a given position along the genome. Because of the enormous amount of reads in total and enormous length of the genome, the read count distributions in different windows are independent. We first introduce the probability distribution of scores for a single window

$$\rho(s) = \sum_{l \ge l_0} \delta(s - s(l)) P(l,\lambda), \tag{2}$$

where $\delta()$ is a Dirac delta function. We then consider the gap contribution. The fundamental unit of a gap is an 'ineligible' window, and the probability t of a window being 'ineligible' is

$$t = P(0,\lambda) + P(1,\lambda) + \cdots + P(l_0 - 1,\lambda). \tag{3}$$

The number of 'ineligible' windows in a gap ranges from zero to g. The gap factor G therefore is

$$G = 1 + t + t^2 + \cdots + t^g. \tag{4}$$

$M(s)$ depends on λ, l_0 and g via $\rho(s)$, t and G. Because an island has to be bound by gaps of sizes of at least $g+1$, $M(s)$ can be separated by boundary

contributions and a kernel $\tilde{M}(s)$,

$$M(s) = t^{g+1}\tilde{M}(s)t^{g+1}. \quad (5)$$

The island score can be partitioned between the last 'eligible' window and the rest in a combinatorial manner, as illustrated in Figure 1b, therefore a recursion relation can be constructed for the kernel $\tilde{M}(s)$:

$$\tilde{M}(s) = G(\lambda, l_0, g)\int_{s_0}^{s} ds' \tilde{M}(s-s')\rho(s'), \quad (6)$$

with a boundary condition of $\tilde{M}(0) = G(\lambda, l_0, g)^{-1}$. Here $s_0 = -\ln P(l_0, \lambda)$.

We are interested in the islands with high scores generated by large fluctuations in the random placement of the reads. Because the occurrences of those islands are rare and hence essentially independent, the number of islands of score s is simply $LM(s)$.

2.1.4 Asymptotics for the island-score distribution in a random background Equations (5) and (6) provide a recursive method to calculate the probability of high-scoring islands. Since the high-score tail of the island score distribution is of fundamental interest, it is useful to obtain an analytical expression in closed form for its asymptotic behavior. Anticipating the asymptotic behavior to be that of an exponential decay, we plug the ansatz $\tilde{M}(s) = \alpha \exp(-\beta s)$ into Equation 6. Straightforward algebra leads to an equation that determines the exponent β,

$$G(\lambda, l_0, g)\sum_{l \geq l_0} P(l, \lambda)^{1-\beta} = 1. \quad (7)$$

The coefficient α in the ansatz can be found by fitting.

To validate the analytical approaches [Equations (5–7)] for the random background model, we employed Monte Carlo simulation to synthetically generate the random reads, identified islands and counted islands with score greater than s averaged over multiple simulation runs. We then compared that with the expected number of islands with score greater than s in the background, $\left\langle \sum_{s' \geq s} N(s') \right\rangle_B \approx \sum_{s' \geq s} LM(s')$, obtained using analytical approaches. We found excellent agreement (Fig. S2). It is worth noting that the background island-score distribution approaches its asymptotic form quickly.

2.1.5 Significance determination without control library The island-score distribution in a random background model allows the determination of a threshold score value s_T, which is used in an experimental library to find islands significant enough to be designated ChIP-enriched domains. s_T is determined by requiring the expected number of islands with scores above the threshold s_T to be less than a E-value threshold e :

$$\sum_{s \geq s_T} LM(s) \leq e. \quad (8)$$

The E-value controls the genome-wide error rate of identified islands under the random background.

2.1.6 Choices of parameters The random background island-score distribution depends on window size w, effective genome length L, total read count N, gap size g and a window P-value requirement p_0, which determines the window read count threshold l_0. For histone modifications and histone variants, a reasonable choice for window size w is 200 bp, a number approximately the length of a single nucleosome and a linker. The effective genome length L is different from the actual genome length. When short reads are mapped into the reference genome, normally only those that map to unique genomic loci are selected for analysis. Genomic regions with degenerate sequences or sequences composed of character 'N' are non-mappable as no reads can be unambiguously mapped into these regions. L, therefore, should be chosen as the total length of mappable regions in the genome. The window P-value requirement p_0 should be such that the 'eligible' windows exhibit enrichment (i.e. $l_0 \geq \lambda$). On the other hand, l_0 should not be too high as the 'eligible' windows need not exhibit very strong signals. $p_0 = 0.2$ is a reasonable choice. The gap size g is an important parameter that can be adjusted to the characteristics of the chromatin modification. To study the effect of gap size, we examine how the aggregate score of all significant islands changes as g is tuned, as shown in Figure 2. H2A.Z is representative of localized signals. The aggregate score quickly reaches maximum at $g = 1$, beyond which the potential increase in the island coverage due to a bigger gap cannot overcome the loss of small islands due to the increase in the island-score threshold s_T. For this type of chromatin modification, it is natural to choose the gap size that maximizes the aggregate score. On the other hand, H4K20me1 shows the typical behavior of chromatin modifications with a diffuse profile. The aggregate score increases gradually towards saturation for reasonable gap sizes. For this type of signal, we suggest to choose the gap size so that the corresponding aggregate score is sufficiently close to saturation. As shown in Figure S3, lack of saturation in the aggregate score as a function of the gap size is in general an indication of poor sequencing coverage. Figure S3 shows the length distribution of significant islands for H2A.Z and H4K20me1, with the gap sizes determined as described above.

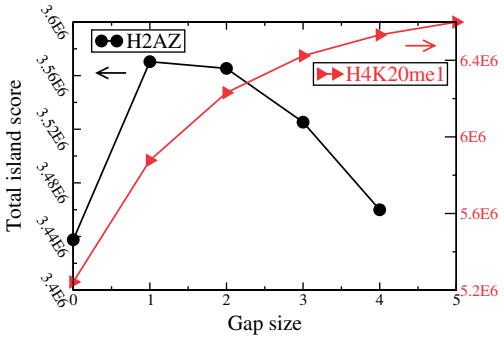

Fig. 2. Aggregate score of all significant islands versus gap size for H2A.Z (black) and H4K20me1 (red). The gap size is measured in units of windows. Here, $l_0 = 2$ and E-value is 0.1.

2.1.7 Significance determination with control library First, we use a lenient E-value threshold to identify a set of candidate islands that exhibit reads clustering under the random background model. Then, for each candidate island, we count the number of ChIP reads n_s and control reads n_c, and calculate a P-value as $\sum_{n=n_s}^{\infty} P(n_s, cn_c)$, where c is the rescaling factor that is equal to the ratio of the ChIP library size over the control library size ($c = N_s/N_c$). Candidate islands with $n_s \leq cn_c$ are discarded because we are only interested in enrichment. The significant islands can be identified with a P-value threshold using Bonferroni correction for multiple testing. Alternatively, a false discovery rate (FDR) can be calculated by following standard procedure (Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001) from the P-values, or by swapping the ChIP and control libraries (Zhang *et al.*, 2008a). For a flowchart of SICER, see Figure S5. For a comparison of enrichment regions identified with and without control library, please see Figure S6.

2.2 Datasets and method parameters

The ChIP-Seq data for histone modifications H4K3me3 and H3K27me3 in human resting CD4$^+$ T-cells were obtained from Barski *et al.* (2007) and Wang *et al.* (2008). The ChIP-Seq data for histone modifications H4K3me3 and H3K27me3 in mouse embryonic stem (ES) cell, the whole-cell extract (WCE) control library, and the real-time PCR (QPCR) results for H3K4me3 and H3K27me3 at 60 loci, were obtained from Mikkelsen *et al.* (2007). In the QPCR data, loci with QPCR fold-change value above (below) 4 were treated as positives (negatives). Based on this criterion, there are 32 positives and 28 negatives in the dataset. The histone modification libraries for the human

$CD133^+$ and $CD36^+$ cells were obtained from Cui *et al.* (2009). The histone modification data for mouse Th1, Th2 and Th17 $CD4^+$ T-cells were obtained from Wei *et al.* (2009). The IgG control library for the human resting $CD4^+$ T-cells and 'input' control libraries for the human $CD133^+$ and $CD36^+$ cells are available at the web site for SICER.

For all ChIP-Seq libraries presented here, only uniquely mapped reads were used and all libraries were preprocessed to filter out redundant reads in an effort to minimize potential PCR bias. The window size was chosen to be 200 bp (see above). Most of the reads in the libraries we used are 25 bp, so we chose the effective human genome size as 74.3% of human genome size in hg18 and 77% for the mouse genome (A. Smith, private communication).

For SICER, in all libraries except for those from Mikkelsen *et al.* (2007), the location of a read on positive (negative) strand was shifted by +75 bp (−75 bp) from its 5′ start to represent the center of the DNA fragment associated with the read, because the majority of the data are produced with ChIP DNA fragment size of mono-nucleosome, i.e. ≈150 bp. The shift value for the mouse ES ChIP-Seq libraries (Mikkelsen *et al.*, 2007) was found to be 150 bp. Based on analysis and discussion described above, the window *P*-value $p_0 = 0.2$. The gap size is chosen to be $g = 1$ for H2AZ and H3K4me3 and $g = 3$ for other histone modification libraries, unless noticed otherwise.

We used four methods in comparison: QuEST (Valouev *et al.*, 2008) version 2.1, F-Seq (Boyle *et al.*, 2008) version 1.8.3, MACS (Zhang *et al.*, 2008a) version 1.3.5 and FindPeaks (Fejes *et al.*, 2008) version 3.2.2.3. The details of the parameters used are summarized in the Supplementary Material.

For ChIP-Seq libraries in human resting $CD4^+$ T-cells, an IgG library was used as control. For H3K4me3 and H3K27me3 libraries in mouse ES cells, an WCE library was used as control. For the H3K27me3 libraries in mouse $CD4^+$ T-cell lineages, no control library was available, the no control option was used. FindPeaks (3.2.2.3) and F-Seq (1.8.3) do not use a control library.

3 RESULTS

3.1 Overview and evaluation of SICER

We have developed an unbiased method that incorporates the tendency of histone modifications to cluster to form the domains. This method identifies islands as *clusters* of enriched windows. Islands, rather than individual windows of fixed length, are the fundamental units of interest. Gaps are allowed in the island to account for: (i) lack of reads or read-count fluctuations in ChIP-enriched domains in undersaturated ChIP-Seq libraries; (ii) repetitive genomic regions non-mappable by uniquely mapped reads; and (iii) unmodified nucleosomes. The gap size can be adjusted to the nature of the chromatin modification. The score of an island is associated with the entire enrichment profile on the island, rather than just the peak value. We develop mathematical formula for the distribution of island scores in the random background model. In the case that a control library is not available, we identify significant domains of enrichment as islands unlikely to appear by chance in the random background model. We use an *E*-value, the expected number of significant islands in the background, to control significance. We derive mathematical formula for fast and precise determination of significance. As the sequencing of a control library is quickly becoming the standard protocol, more and more ChIP-Seq data come with a control library. We then use the control library to take into account systematic biases in the background (Kharchenko *et al.*, 2008; Rozowsky *et al.*, 2009; Zhang *et al.*, 2008a). Motivated by Zhang *et al.* (2008a) and Rozowsky *et al.* (2009), we first identify a set of candidate islands exhibiting reads clustering using the approach described above (i.e. using a random background model) with a lenient *E*-value threshold. We then compare the ChIP read count and control read count on each candidate island to determine the significance of enrichment, with the control read count rescaled to account for the size difference in the control library and the ChIP library. For a flowchart of SICER, see Figure S5.

A number of methods that aim towards finding peaks in ChIP-Seq data have been published. In SISSRs (Jothi *et al.*, 2008), QuEST (Valouev *et al.*, 2008), MACS (Zhang *et al.*, 2008a), CisGenome (Ji *et al.*, 2008), USeq (Nix *et al.*, 2008) and others (Albert *et al.*, 2008; Johnson *et al.*, 2007; Kharchenko *et al.*, 2008; Zhang *et al.*, 2008b), the genome is scanned with a sliding window of fixed width, all windows deemed to have significant enrichment are identified, and neighboring significant windows can be merged. In PeakSeq (Rozowsky *et al.*, 2009), counts of overlapping DNA fragments at each nucleotide position are used to build a score map and positions with significant scores are identified. An essential feature shared by these methods is the use of local statistics to estimate significance. The significance of an enriched window or a position is independent of those of other windows or positions. It is determined from a random background model of window read count distribution (Fejes *et al.*, 2008; Jothi *et al.*, 2008), from a non-random background model (Ji *et al.*, 2008; Zhang *et al.*, 2008b), or from comparison with a control library (Ji *et al.*, 2008; Johnson *et al.*, 2007; Jothi *et al.*, 2008; Kharchenko *et al.*, 2008; Nix *et al.*, 2008; Rozowsky *et al.*, 2009; Valouev *et al.*, 2008; Zhang *et al.*, 2008a).

Published methods for the analysis of histone modification data are limited. (Mikkelsen *et al.*, 2007) employed local statistics in combination with an empirical background model obtained by randomizing read locations for the identification of ChIP-enriched regions for histone modifications with punctate profiles. They also employed a hidden Markov model approach, the details of which have not been published as far as we know. Xu *et al.* (2008) developed a hidden Markov model approach for the identification of differential histone modification sites across cell-types or conditions. However, it does not provide a method for the identification of ChIP enrichment in a single library. For two ChIP libraries under comparison, a window is deemed to be significantly enriched when the combined normalized read counts from the two libraries exhibit at least a 2-fold enrichment versus random expectation. Robertson *et al.* (2008) used FindPeaks (Fejes *et al.*, 2008) to identify domains of histone modification. FindPeaks defines an island as a region occupied by continuously overlapping ChIP DNA fragments. For its basic functionality, it uses the height of an island, defined as the maximum overlapped fragment count on the island, as the test statistic (Fejes *et al.*, 2008; Robertson *et al.*, 2008). FindPeaks uses non-local statistics, as the significance of any part of the island depends on the peak height of the whole island. Additionally, (Boyle *et al.*, 2008) developed F-Seq, a kernel-density method for identification regions of open chromatin from DNase-seq data.

For performance comparison with SICER, we chose MACS, QuEST, FindPeaks and F-Seq for reasons detailed below. MACS is a window-based method using local statistics. We chose MACS because it has been reported to outperform several other methods in identification of TF binding sites (Zhang *et al.*, 2008a), and because MACS uses regional averaging to mitigate the sampling fluctuations in the control library, which is usually severe because of limited sequencing depth in control libraries. We chose QuEST for comparison because this method is based on kernel density estimation. FindPeaks was chosen for comparison because it employs non-local statistics and has been used in the analysis

of histone modification data. F-Seq was chosen because it was designed to analyze DNase-seq data and DNase-seq data should have characteristics in common with histone modification ChIP-Seq data.

We conducted each comparison in two steps. We first evaluated the genome-wide FDR of each method using a scaling approach. Second, we sought sets of genomic loci where independent experimental validations exist. While available data in this regard are quite limited, we were able to obtain two datasets: (i) a set of signature cytokine genes in the mouse CD4$^+$ T-cells whose epigenetic states in the cell differentiation processes had been subject to functional studies (Koyanagi *et al.*, 2005; Schoenborn *et al.*, 2007; Wei *et al.*, 2009). (ii) A set of 60 QPCR results for histone modifications in mouse ES cells at a selected group of genomic loci (Mikkelsen *et al.*, 2007).

3.1.1 Evaluation of prediction robustness via scaling analysis Unlike TF binding sites, the enriched domains of histone modifications lack definitive sequence features. Direct bioinformatic validation of method predictions at the genome-scale is not feasible. Taking advantage of the digitized characteristics of ChIP-Seq data, we argue that the true and false positives can be distinguished by scaling. Namely, if an identified ChIP-enriched domain deemed significant by a method is a true signal, then this domain should remain significant when the sequencing depth is increased. Conversely, if an identified ChIP-enriched domain deemed significant becomes insignificant when sequencing depth is increased, then this particular domain is a false positive. To evaluate the various methods, we took a H3K27me3 library (≈16.3 million reads after preprocessing) in human CD4$^+$ T-cells, and constructed a subset of half the original size via random sampling. Because QuEST did not identify any ChIP-enriched regions under its default parameters for histone modification data, we drop it from the method comparison from this point on. With each of remaining three methods, we identify ChIP-enriched domains in both the full-size library and the half-size subset, under the same statistical criterion (*P*-value for MACS and SICER, *E*-value for FindPeaks and 'threshold' for F-Seq). The significant domains identified using the half-size library that do not overlap with any significant domains in the full-size library are considered false positives. We defined a scaling FDR as the number of false positives divided by the number of significant islands in the half-size subset (Fig. 3a). Since the statistical criterions used by different methods are not directly comparable, we used the island read count coverage, the fraction of reads that were within the identified significant domains, as a common ground for fair comparison. As shown in Figure 3b, for a range of island read count coverage that covers all reasonable choices of statistical significance levels, the scaling FDR for SICER is significantly lower than those for MACS, FindPeaks and F-Seq. We ran multiple random samplings from the full-size library and found that the result is independent of sampling (data not shown), as expected from the large sample size.

3.1.2 Receiver operating characteristic analysis using loci of signature cytokines Cell differentiation involves commitment of featured lineage and extinction of other fates, in which the epigenetic state plays a key role (Bernstein *et al.*, 2007). Upon antigen and cytokine stimulation, multipotential naive CD4$^+$ T-cells differentiate into distinct lineages including Th1, Th2 and

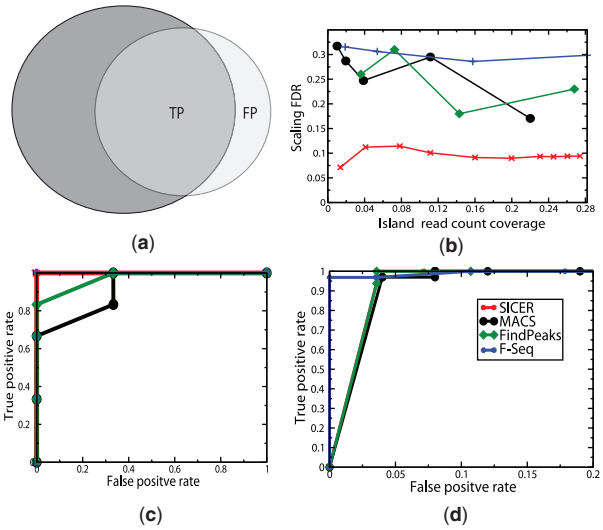

Fig. 3. Comparison of SICER with other methods. (**a**) Schematic illustration of the scaling FDR determination. The dark (light) gray circle represent ChIP-enriched regions identified in the full (half-size) library. The non-overlapping area of the light gray circle represents the false positives. (**b**) FDR versus the island read count coverage. (**c**) ROC analysis of using the epigenetic states at genes encoding signature cytokines in mouse CD4$^+$ cell Th1, Th2 and Th17 lineages. (**d**) ROC analysis of H3K4me3 and H3K27me3 in mouse ES cells.

Th17, whose signature cytokines are *Ifn-γ*, *Il*4, *Il*17, respectively. Previous studies have demonstrated that the signature cytokines are only associated with active epigenetic marks (H3K4me3) in featured lineages, and only associated with repressive epigenetic marks (H3K27me3) in the opposing lineages (Koyanagi *et al.*, 2005; Schoenborn *et al.*, 2007). For this dataset, we focused on the diffuse H3K27me3 signal. The performance of the methods on identifying localized signal was examined subsequently utilizing the QPCR dataset. The expected H3K27me3 enrichment state at *Ifn-γ* locus is $(0,1,1)$, where we used 0 (1) to represent the absence (presence) of H3K27me3 in Th1, Th2 and Th17 cells. Similarly, at *Il*4 locus $(1,0,1)$ is expected and at *Il*17 locus $(1,1,0)$ is expected. For reference, the unfiltered H3K27me3 profiles at the three loci, along with the H3K4me3 profiles, are shown in Figure S7. Taken together, these three loci in the three cell types present six positives and three negatives for H3K27me3 signals. We then used each method to identify the ChIP-enriched regions in the H3K27me3 ChIP-Seq libraries and used receiver operating curve (ROC) to present the findings, as shown in Figure 3c. This ROC analysis demonstrated that SICER outperforms MACS and FindPeaks. Both SICER and F-Seq are able to correctly identify every state in the pattern. The area under the ROC curve is A_{SICER}=1, A_{MACS} = 0.917, $A_{\text{FindPeaks}}$ = 0.972 and $A_{\text{F-Seq}}$ = 1, respectively.

3.1.3 ROC analysis using QPCR data Having evaluated the performance of the methods on diffuse data, we examined their performances on data with localized modification signals. In mouse ES cells, not only H3K4me3 but also H3K27me3 signals have been observed to be largely punctate (Mikkelsen *et al.*, 2007). We used the QPCR dataset in combination with the H3K4me3 and H3K27me3 ChIP-Seq data from Mikkelsen *et al.* (2007) to measure specificity

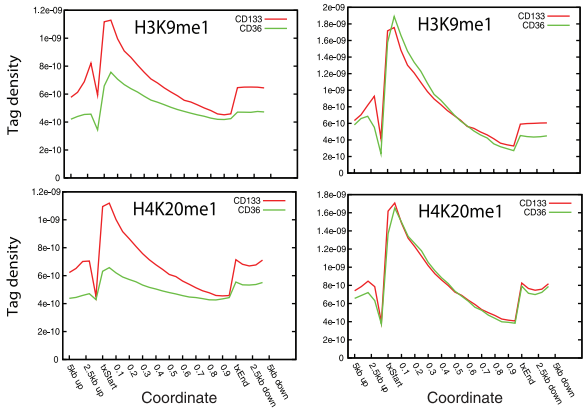

Fig. 4. Composite histone modification profiles across genic regions in human CD133$^+$ (red) and CD36$^+$ (green) cells. The figures in the left (right) panel are made with all reads in the library (only reads on islands, which are identified using 'input' library as control).

and sensitivity. The resulting ROC curve is shown in Figure 3d. All four methods performed fairly well with this data set, with the area under the ROC curve being $A_{\text{SICER}} = 0.9810$, $A_{\text{MACS}} = 0.9677$, $A_{\text{FindPeaks}} = 0.9810$ and $A_{\text{F-Seq}} = 0.9978$, respectively.

3.2 ChIP-Seq data normalization

Because of the important role of epigenetics in developmental and pathological conditions, application of ChIP-Seq to the study of changes in chromatin states in different cell types, developmental stages and pathological conditions are increasingly wide spread (Cui *et al.*, 2009; Mikkelsen *et al.*, 2007). In those settings, quantitative comparisons of signal levels provides important information about the underlying biological principle. For the signal levels to be compared appropriately, the data need to be normalized to account for the differences in experimental preparations and instrumental conditions. This is similar to the situation encountered in gene expression measurement using microarrays (Quackenbush, 2002). SICER can be used to filter out background noise by removing reads not on the islands. We applied this idea to quantitative comparison of modification levels in differentiation from human hematopoietic stem/progenitor CD133$^+$ cells to the erythrocyte precursor CD36$^+$ cells. Because the majority of genes exhibit similar expression patterns between the two cell types (Cui *et al.*, 2009), the overall modification profiles are expected to be similar between CD133$^+$ and CD36$^+$. However, the modification profiles obtained using all the mapped reads, including both tags in the islands and out of islands, show dramatic differences between these two cell types (Fig. 4, left column). Interestingly, the profiles using tags only in the islands show similar patterns (Fig. 4, right column). We further classified the genes into four groups according to their expression pattern during differentiation (Cui *et al.*, 2009): (i) always expressed (9196 genes); (ii) always silent (7420 genes); (iii) repressed (934 genes); and (iv) induced (306 genes). The unfiltered and filtered composite profiles were compared side by side for each group. The majority of genes belong to the groups of always expressed genes and always silent genes. For these two groups, the modification profiles are not expected to show significant changes. Indeed, the filtered profiles of each modification for the two cell types are similar, whereas many unfiltered counterparts showed dramatic differences (Figs S8 and S9). In the groups of repressed and induced genes, the dynamical change in filtered profiles of modifications are more consistent with their known biological functions (Figs S10 and S11). In Figure 4, the islands were identified with P-value of 10^{-10}. To check how normalization depends on the choice of parameters in island identification, we also experimented with different choices of P-value (10^{-3}, 10^{-5} and 10^{-15}), and gap size (2 instead of 3), the salient features did not change (data not shown). These results indicate that filtering with islands is a reliable method for data normalization in quantitative comparison of histone modification profiles.

4 DISCUSSION AND CONCLUSION

ChIP signals from many histone modifications, histone variants and histone-modifying enzymes form diffuse, broad domains. Based on the notion that the establishment of many histone modifications involves positive feedback resulting in the spreading of modified nucleosomes, we develop the SICER method that takes into account of the *enrichment context* of a local window in determining its significance. In contrast, in local statistics-based algorithms, the significance of a local window is independent of other regions (see Fig. S13 for illustration). When a control library is available, we use the random background model to identify candidate islands. These candidate domains of variable lengths, rather than windows of fixed lengths, serve as the units for enrichment detection. We then use the control library to determine the significance of enrichment for these domains. The fact that the size of the candidate islands are in general much larger than the size of a nucleosome (Fig. S4) helps to reduce the sampling fluctuations in the control library and enables more accurate determination of the position-dependent background level. An alternative approach would be to determine the island-score distribution in an inhomogeneous background model specified by the control library. One can obtain via Monte Carlo simulation the island-score distribution in this inhomogeneous background, which provides a *global* statistics for the significance of the islands. Despite the advantages, this approach requires an accurate determination of the inhomogeneous background at the level of individual windows. It would be interesting to explore how the sequencing depth of the control library affects the performance of the different approaches.

Using both genome-scale analysis and datasets of genomic loci validated experimentally, we demonstrated that SICER compares favorably with existing methods at identifying ChIP-enriched domains in histone modification signals, especially those with diffuse profiles. We also demonstrated the success of this method in normalization and sequence saturation analysis, which are useful tools for statistical analysis of ChIP-Seq data. As genomic landscapes of chromatin modifications are becoming increasingly available, methods such as SICER will be absolutely essential in deciphering the functions of chromatin modifications.

ACKNOWLEDGEMENTS

The authors would like to thank Gang Zou, Lai Wei, Andrew D. Smith, Michael Q. Zhang, X. Shirley Liu and Raja Jothi for helpful discussions, and B. E. Bernstein and T. S. Mikkelsen for providing

the QPCR results for H3K4me3 and H3K27me3 data in the mouse ES cells.

Funding: University Facilitating Fund (to W.P., in parts); the National Science Foundation (DMR0313129 to Chen Zeng in parts); Intramural Research Program for the National Heart Lung and blood Institute; National Institute of Health (to K.Z., in parts).

Conflict of Interest: none declared.

REFERENCES

Aagaard,L. *et al.* (1999) Functional mammalian homologues of the Drosophila pev-modifier su(var)3-9 encode centromere-associated proteins which complex with the heterochromatin component m31. *EMBO J.*, **18**, 1923–1938.

Albert,I. *et al.* (2008) GeneTrack—a genomic data processing and visualization framework. *Bioinformatics*, **24**, 1305–1306.

Bannister,A.J. *et al.* (2001) Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. *Nature*, **410**, 120–124.

Barski,A. *et al.* (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823–837.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate - a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.*, **57**, 289–300.

Benjamini,Y. and Yekutieli,D. (2001) The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.*, **29**, 1165–1188.

Bernstein,B.E. *et al.* (2007) The mammalian epigenome. *Cell*, **128**, 669–681.

Boyle,A.P. *et al.* (2008) F-Seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics*, **24**, 2537–2538.

Chen,X. *et al.* (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, **133**, 1106–1117.

Cui,K. *et al.* (2009) Chromatin signatures in multipotent human hematopoietic stem cells indicate the fate of bivalent genes during differentiation. *Cell Stem Cell*, **4**, 1–14.

Dodd,I.B. *et al.* (2007) Theoretical analysis of epigenetic cell memory by nucleosome modification. *Cell*, **129**, 813–822.

Fejes,A.P. *et al.* (2008) Findpeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. *Bioinformatics*, **24**, 1729–1730.

Felsenfeld,G. and Groudine,M. (2003) Controlling the double helix. *Nature*, **421**, 448–453.

Hansen,K.H. *et al.* (2008) A model for transmission of the H3K27me3 epigenetic mark. *Nat. Cell Biol.*, **10**, 1291–1300.

Jacobson,R.H. *et al.* (2000) Structure and function of a human TAF(II)250 double bromodomain module. *Science*, **288**, 1422–1425.

Ji,H. *et al.* (2008) An integrated software system for analyzing chip-ChIP and ChIP-seq data. *Nat. Biotechnol.*, **26**, 1293.

Johnson,D.S. *et al.* (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science*, **316**, 1497–1502.

Jothi,R. *et al.* (2008) Genome-wide identification of in vivo protein-DNA binding sites from ChIP-seq data. *Nucleic Acids Res.*, **36**, 5221–5231.

Kharchenko,P.V. *et al.* (2008)Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat. Biotechnol.*, **26**, 1351–1359.

Koyanagi,M. *et al.* (2005) Ezh2 and histone 3 trimethyl lysine 27 associated with Il4 and Il13 gene silencing in T(h)1 cells. *J. Biol. Chem.*, **280**, 31470–31477.

Lachner,M. *et al.* (2001) Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins. *Nature*, **410**, 116–120.

Mikkelsen,T.S. *et al.* (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, **448**, 553–560.

Nix,D. *et al.* (2008) Empirical methods for controlling false positives and estimating confidence in ChIP-seq peaks. *BMC Bioinformatics*, **9**, 523.

Owen,D.J. *et al.* (2000) The structural basis for the recognition of acetylated histone H4 by the bromodomain of histone acetyltransferase Gcn5p. *EMBO J.*, **19**, 6141–6149.

Pauler,F.M. *et al.* (2009) H3K27me3 forms blocs over silent genes and intergenic regions and specifies a histone banding pattern on a mouse autosomal chromosome. *Genome Res.*, **19**, 221–233.

Quackenbush,J. (2002) Microarray data normalization and transformation. *Nat. Genet.*, **32**, 496–501.

Robertson,A.G. *et al.* (2008) Genome-wide relationship between histone H3 lysine 4 mono- and tri-methylation and transcription factor binding. *Genome Res.*, **18**, 1906–1917.

Rozowsky,J. *et al.* (2009) Peakseq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat. Biotechnol.*, **27**, 66–75.

Schoenborn,J.R. *et al.* (2007) Comprehensive epigenetic profiling identifies multiple distal regulatory elements directing transcription of the gene encoding Interferon-gamma. *Nat. Immunol.*, **8**, 1398.

Schones,D.E. and Zhao,K. (2008) Genome-wide approaches to studying chromatin modifications. *Nat. Rev. Genet.*, **9**, 179–191.

Schwartz,Y.B. and Pirrotta,V. (2007) Polycomb silencing mechanisms and the management of genomic programmes. *Nat. Rev. Genet.*, **8**, 9–22.

Valouev,A. *et al.* (2008) Genome-wide analysis of transcription factor binding sites based on ChIP-seq data. *Nat. Methods*, **5**, 829–834.

Wang,Z.B. *et al.* (2008) Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat. Genet.*, **40**, 897–903.

Wei,G. *et al.* (2009) Global mapping of H3K4me3 and H3K27me3 reveals specificity and plasticity in lineage fate determination of differentiating CD4+ T cells. *Immunity*, **30**, 155–167.

Wen,B. *et al.* (2009) Large histone H3 lysine 9 dimethylated chromatin blocks distinguish differentiated from embryonic stem cells. *Nat. Genet.*, **41**, 246–250.

Xu,H. *et al.* (2008) An HMM approach to genome-wide identification of differential histone modification sites from ChIP-seq data. *Bioinformatics*, **24**, 2344–2349.

Zhang,Y. *et al.* (2008a). Model-based analysis of chip-seq (macs). *Genome Biol.*, **9**, R137.

Zhang,Z.D. *et al.* (2008b) Modeling chip sequencing in silico with applications. *PLoS Comput. Biol.*, **4**, e1000158.