

Statistical model for identification of ChIP-enriched regions from ChIP-Seq data

Supplementary Material

Chongzhi Zang¹, Dustin E. Schones², Chen Zeng¹, Kairong Cui² and Keji Zhao², Weiqun Peng¹

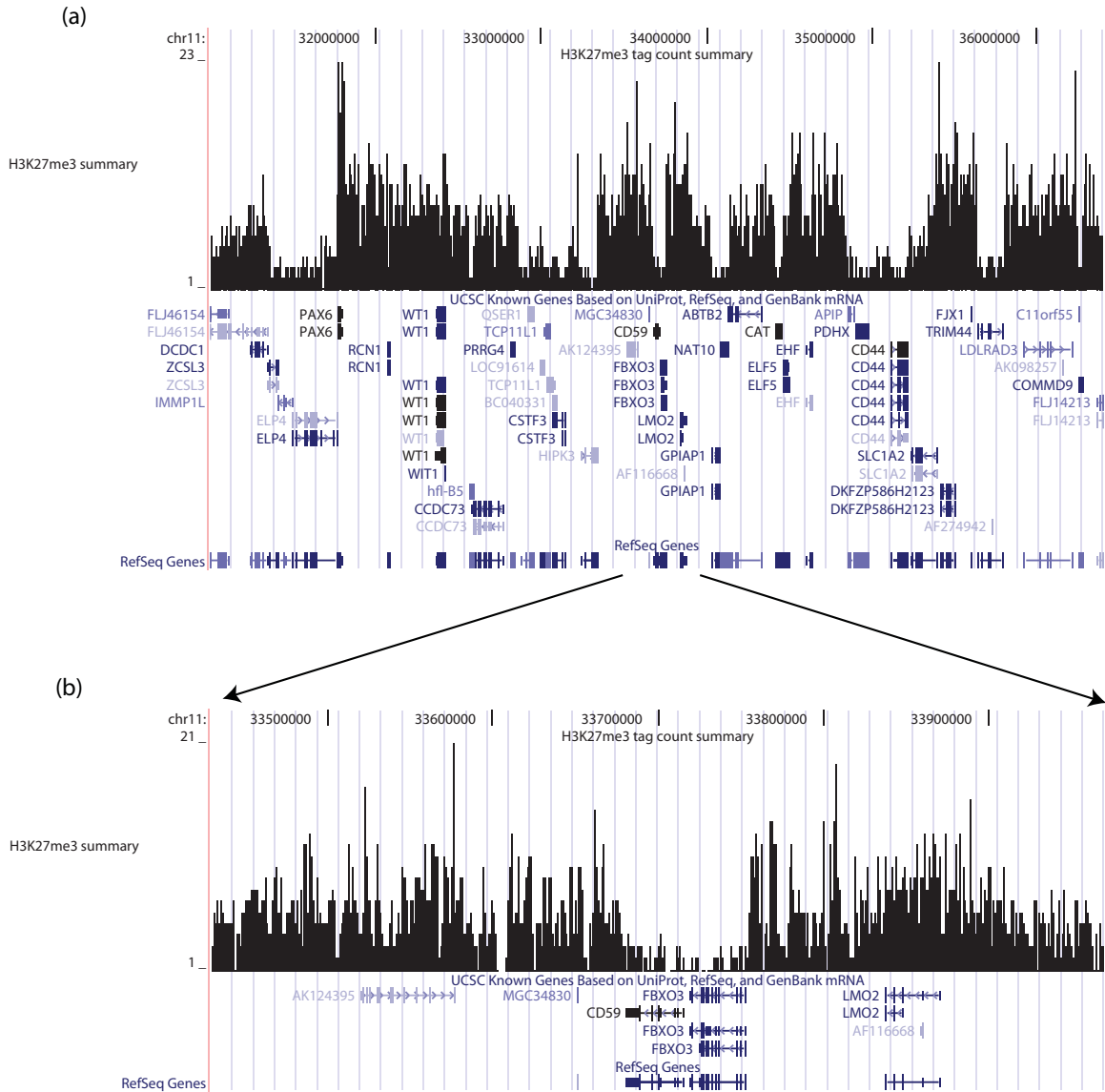


Figure S1: An example of the broad and diffuse profile of a histone modification. (a) The profile of H3K27me3 in human CD4⁺ resting T cells in a genomic region on chromosome 11. (b) A zoom-in profile of a sub-region in (a). In (b) the length scale for the profile is 100Kbp. Figure S7 provides examples of H3K27me3 profiles at much finer scale.

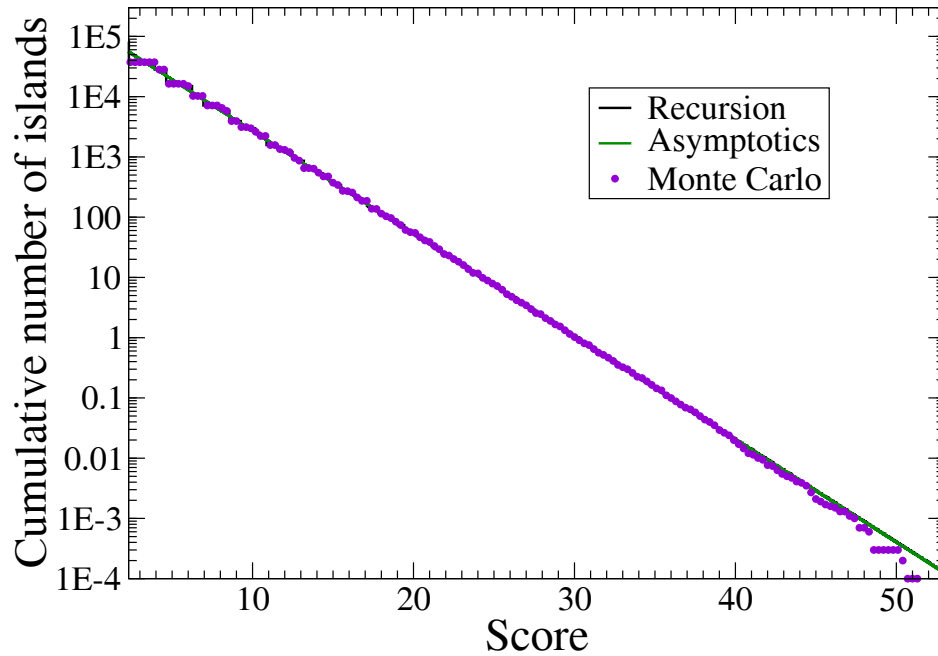


Figure S2: Comparison of cumulative score distributions of islands formed by randomly placed reads obtained from analytical and simulation approaches. The x-axis (y-axis) denotes the score of an island (the number of islands with scores above the x-value). The black line represents the result obtained using the recursion relation from Eqns. (4) and (5). The green line represents the result obtained using asymptotic approximation (Eqn.(6)) and α is obtained by fitting. The violet dots represent the average result of Monte Carlo simulation of 10000 runs. Here $N=6 \cdot 10^5$, $L=2 \cdot 10^8$ bp, $w=200$ bp, $g=2$, $l_0 = 2$.

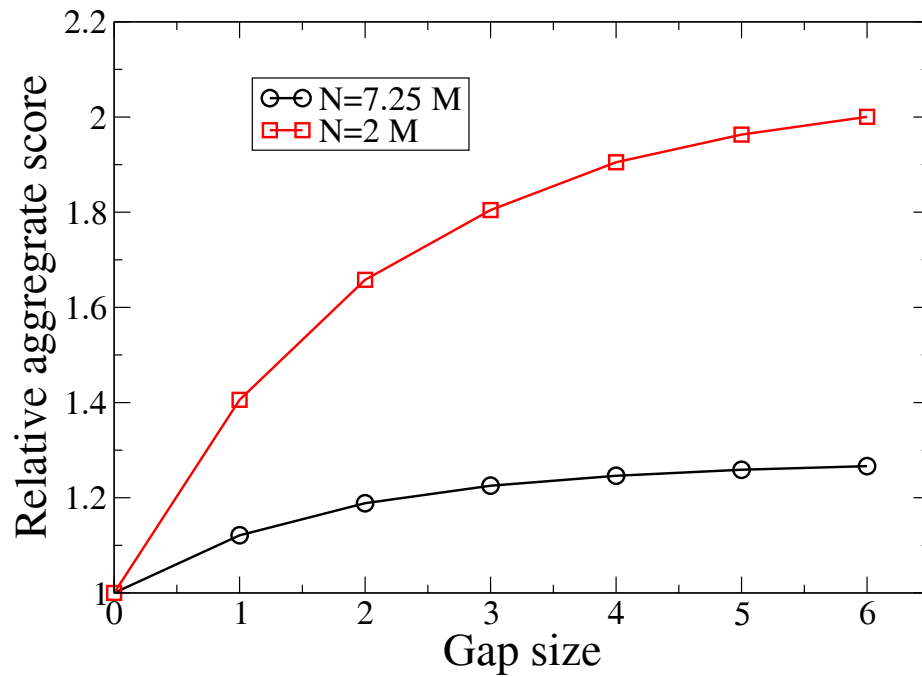


Figure S3: The effect of sequencing depth on the choice of gap size. Shown is the relative aggregate score of significant islands versus the gap size for H4K20me1 in human CD133⁺ cells. The black curve used all 7.25 million reads in the H4K20me1 library for islands identification, whereas the red curve used only 2 million reads randomly sampled from the H4K20me1 library. In each curve, the aggregate score for significant islands at each gap size was normalized by the aggregate score at $g=0$. The black curve reaches saturation much faster than the red curve, indicating the importance of sequencing coverage in the choice of gap size. Here the E -value is chosen to be 0.1.

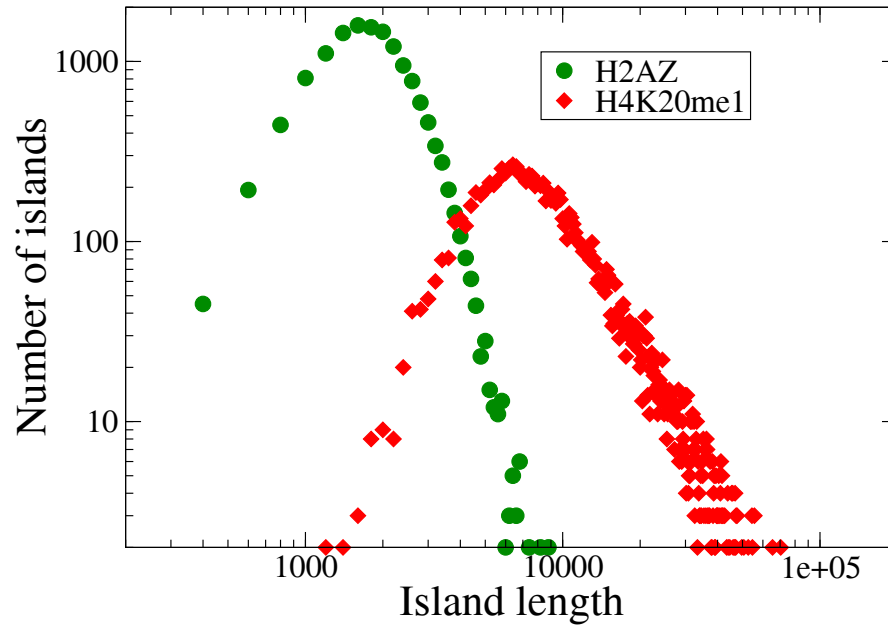


Figure S4: The distribution of lengths of significant islands for H2A.Z and H4K20me1 in human CD133⁺ cells. The H4K20me1 islands are much more diffuse than the H2A.Z islands. Here *E*-value is chosen to be 0.1. $g=1(3)$ for H2A.Z (H4K20me1).

SICER flow chart

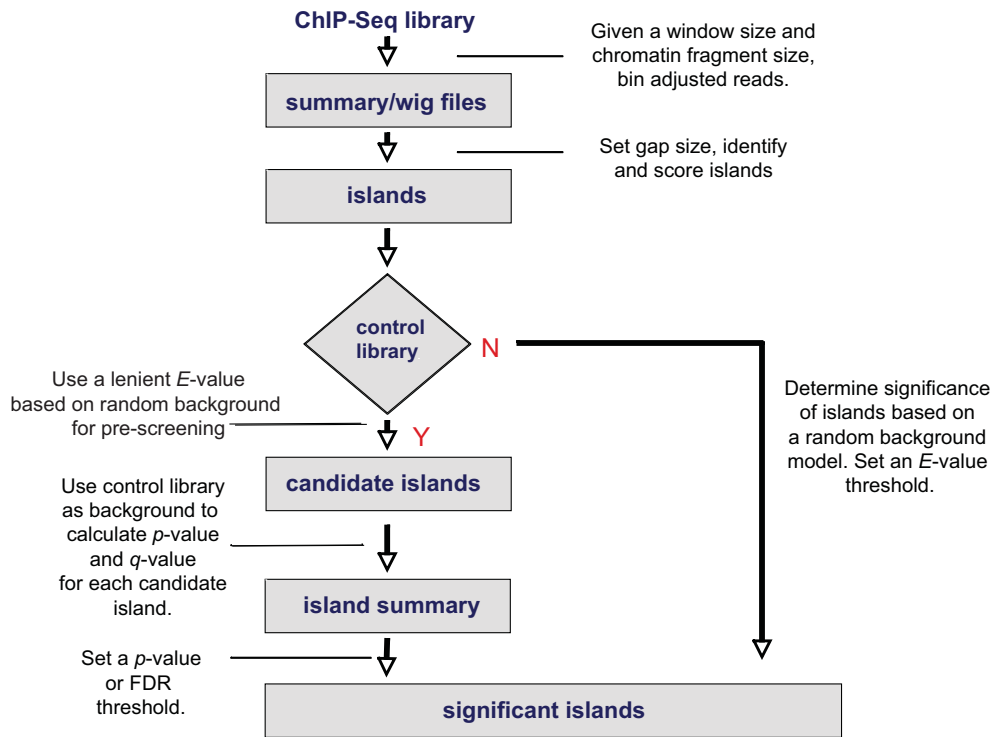
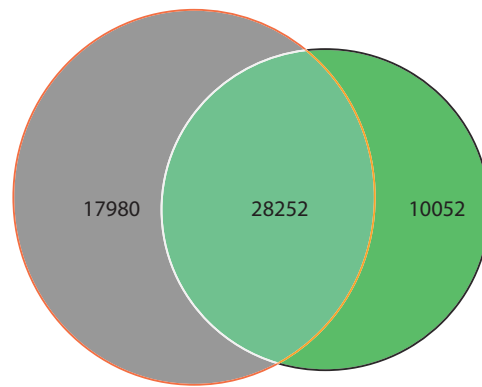


Figure S5: SICER flowchart

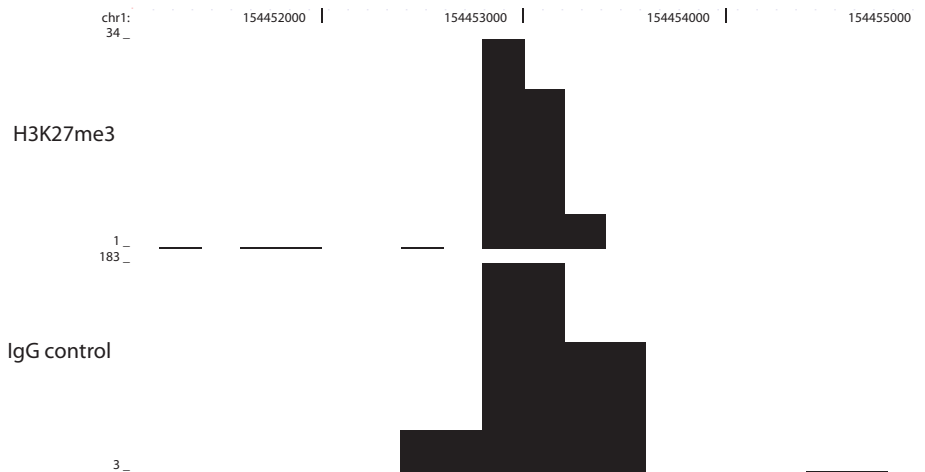
Details of the parameters for the methods used for comparison

All libraries are already pre-filtered to remove redundant reads. The read length of most of reads in the ChIP-Seq data sets used is 25bp.

- MACS: version 1.3.5 was used. Parameters used: bandwidth = 300bp; tsize = 25; gsize: equivalent to 74.3% for hg18 and 77% for mm8 (Smith, 2009); mfold: default when possible, otherwise it is lowered to the value acceptable to the program. Default values are used for other parameters. The p-value is scanned for different island read-count coverage.
- Quest: version 2.1 was downloaded from <http://www.stanford.edu/~valouev/QuEST/QuEST.html>. Parameters used: the default parameters under the "parameter configuration" of treating "Histone-type ChIP resulting in wide regions of enrichment" (in which bandwidth = 100 bp, region_size = 1000 bp) and "permissive peak calling parameters" (in which ChIP_threshold = 0.272198860714286; Enrichment fold = 3; Rescue fold = 3).
- FindPeaks: version 3.2.2.3 was used. Parameters: -directional mode is not engaged; -dist_type: default format with triangle distribution and median fragment length 150bp; -eff_frac: 74.3% for hg18 and 77% for mm8 (Smith, 2009); -duplicate filter: off; -qualityfilter: off; -hist_size: 50; -subpeaks: off; -trim: off; Monte Carlo simulation was used for the statistics of the island heights based on a background distribution of random reads; -iterations: 20. The island minimum peak height is scanned for different island read-count coverage. FindPeaks 3.2.x uses the random background model. FindPeaks 3.3.x, made public after the submission of this Paper, allows the use of control library.
- F-Seq: version 1.8.1: defaults are used for all parameters except for the threshold. The "threshold" value is scanned for different island read-count coverage.



(a)



(b)

Figure S6: (a) Comparison of ChIP-enriched regions identified with and without control library by SICER. With H3K27me3 data from human CD4⁺ T cells, we applied two different approaches to find significant islands: 1) with control, p -value = 10^{-10} , represented by the grey circle; 2) without control, E -value=100, represented by the green circle. We chose these parameters so that the two approaches result in similar amount of reads on significant island. As shown, they share a majority of islands. On the other hand, there are significant differences in terms of the predications made by the two approaches. As demonstrated in the example (b), many significant islands identified in the ‘without control’ case are filtered out by the control library. On the other hand, many islands specific to ‘with control’ are significant because their background level is even lower than random. Of course, some can be attributed to the lack of sequencing depth in the control library. (b) An example of a significant island identified without control but filtered out by the control.

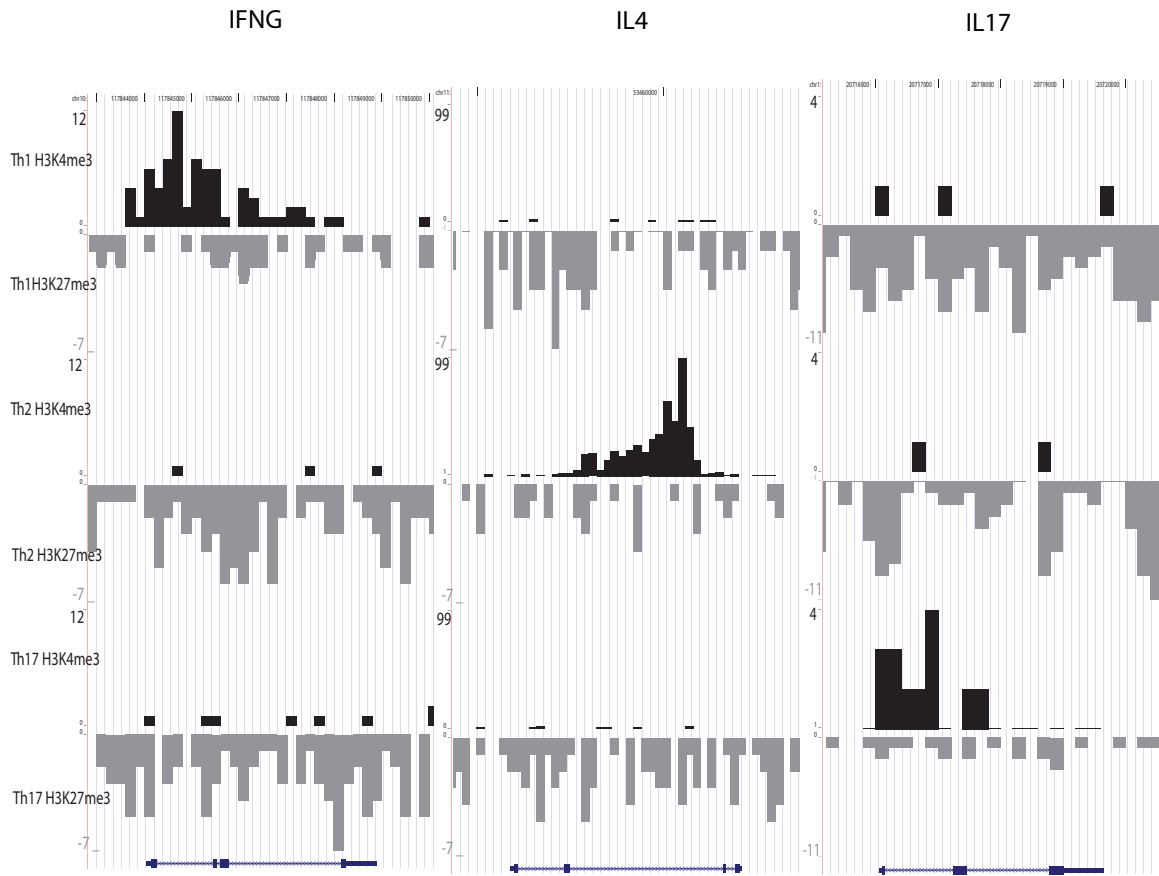


Figure S7: The summary graphs of H3K27me3 (grey) and H3K4me3 (black) at the three signature cytokine genes IFNG, IL4 and IL17 in mouse CD4+ lineages Th1 (top panel), Th2 (middle panel) and Th17 (bottom panel). The summary graphs were made by binning reads into 200 bp non-overlapping windows.

I. Changes in histone modification profile in genes categorized according to gene expression changes

The genes are separated by their expression pattern into four groups according to the “absent” and “present” calls made by Affymetrix analysis tools. Genes that are “present” (“absent”) in both CD133⁺ and CD36⁺ are denoted as always expressed (silent). Genes that are “present” (“absent”) in CD133⁺ and “absent” (“present”) in CD36⁺ are denoted as repressed (induced). Only genes with consistent calls in the two replicates of the microarray were retained. On the left (right) panel of each figure are unfiltered (filtered) profiles.

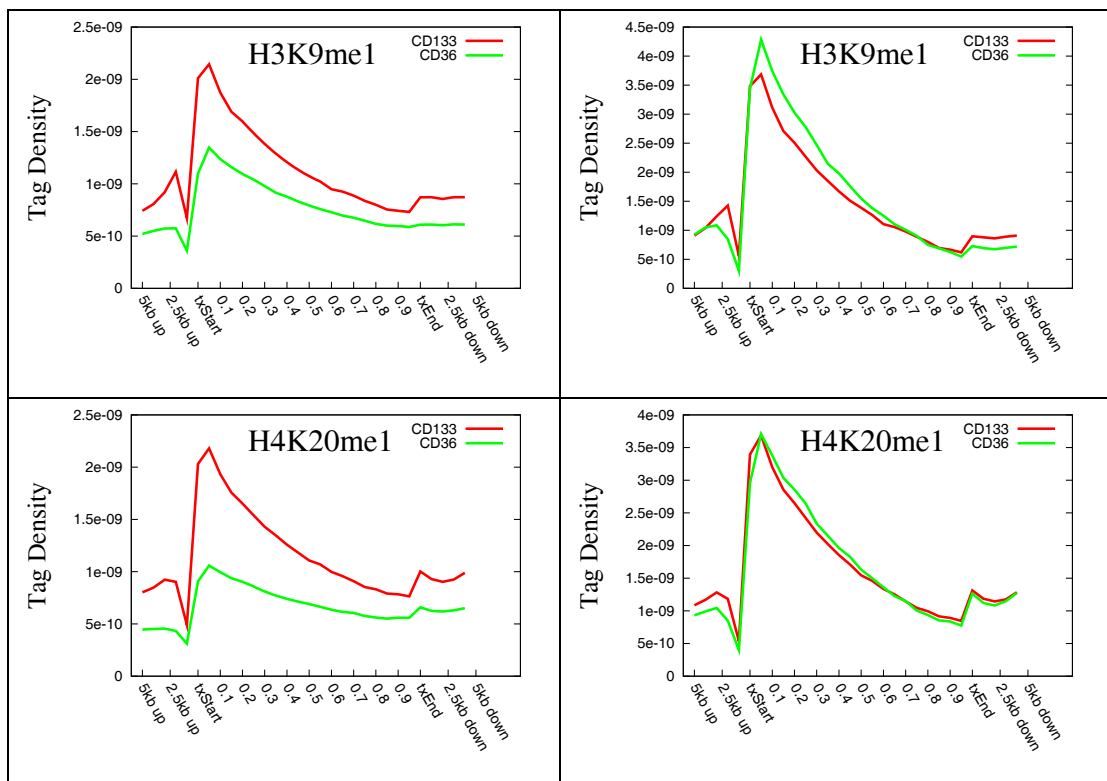


Figure S8: Composite profiles of genes expressed in both CD133⁺ and CD36⁺.

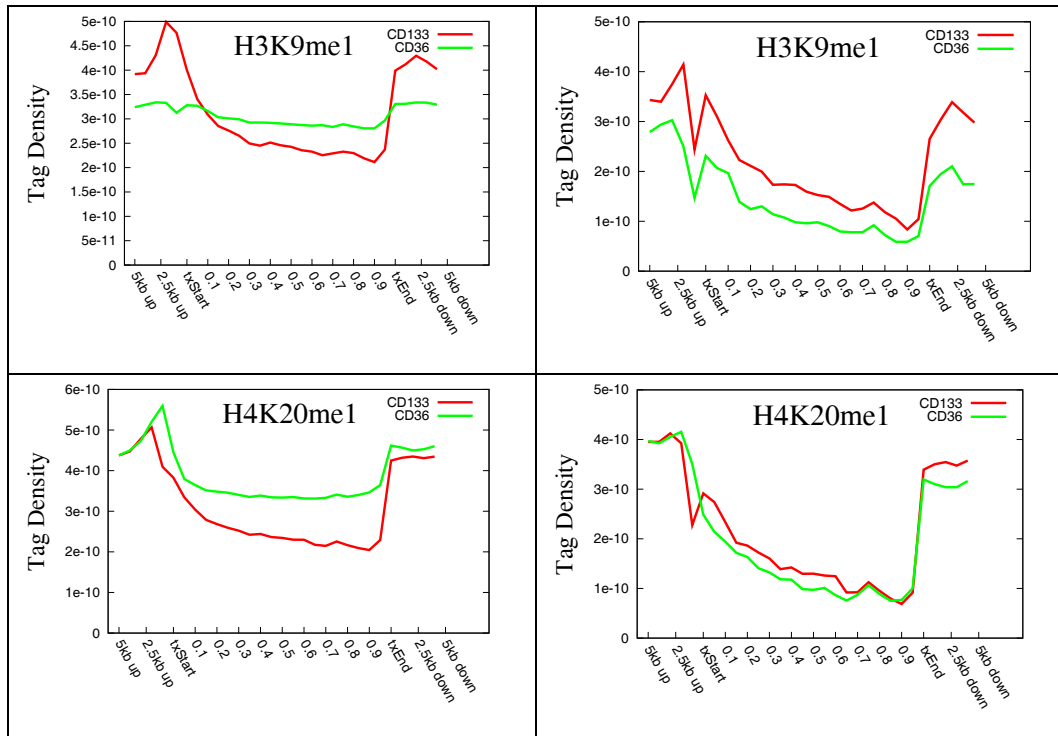


Figure S9: Composite profiles of genes silent in both CD133⁺ and CD36⁺. The difference of filtered H4K20me1 profiles in the promoter region is found to be due to bivalent domain genes(Cui, et al., 2009).

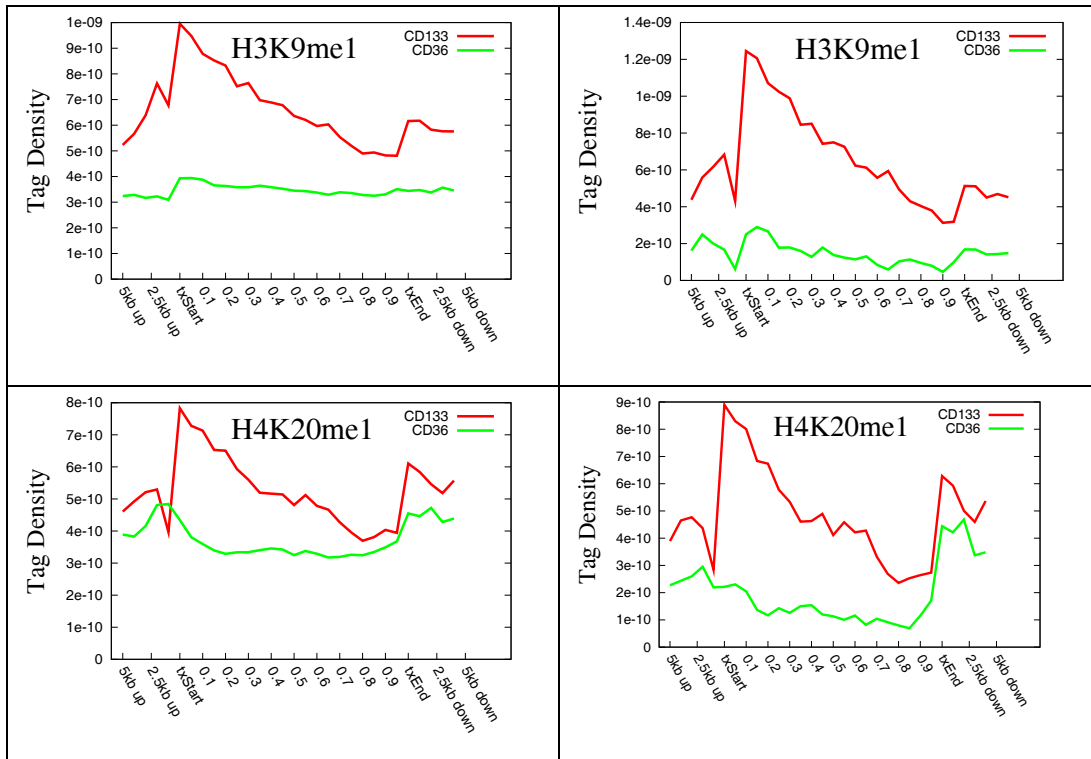


Figure S10: Composite profiles of repressed genes.

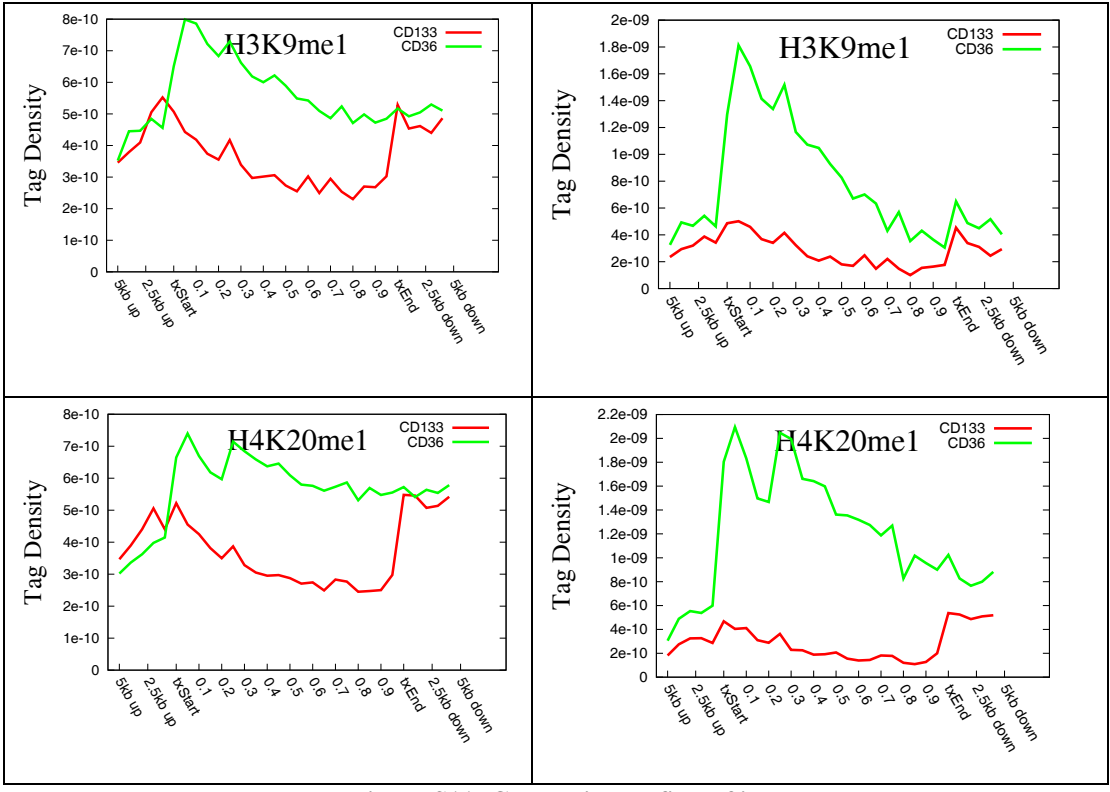


Figure S11: Composite profiles of induced genes.

II. Scaling analysis for sequencing coverage

A typical ChIP-Seq experiment produces millions of reads that map to reference genome. Accurate and complete identification of the ChIP-enriched regions requires that these regions receive enough reads to overcome the background noise. Therefore, the determination of sequencing coverage is quite important. The island approach can be used as an unbiased method to judge the degree of coverage, by sampling the experimental data at different sizes, and looking at the scaling of the fraction of reads on islands versus the sample size. Figure S11 shows the result of such an analysis applied to H3K4me3 and H3K27me3 in CD4⁺ T-cells, both of which have close to 18 million reads before preprocessing. A dramatic difference in the scaling behaviors of the two modifications can be seen. While for H3K4me3 the scaling curve clearly exhibited leveling indicating saturation, H3K27me3 remained far from saturation all the way through. This is consistent with the fact that H3K4me3 presents sharp, localized peaks whereas H3K27me3 is much more diffuse.

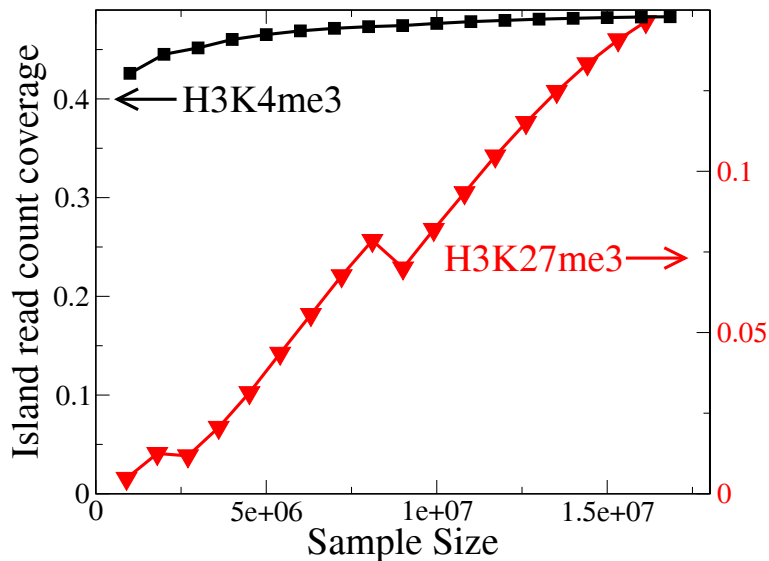


Figure S12: Scaling of island read count coverage versus the number of sampled read count for H3K4me3 (black) and H3K27me3 (red) in CD4⁺ T-cells. Each data point represents the average of 10 random samples of a given size from the original library. The islands are identified using random background model with E -value of 0.1. The gap size is $g=2$. The sudden dips in the H3K27me3 curve are the result of the digital jump in window read-count threshold l_0 .

The significance of local enrichment is context-dependent in SICER

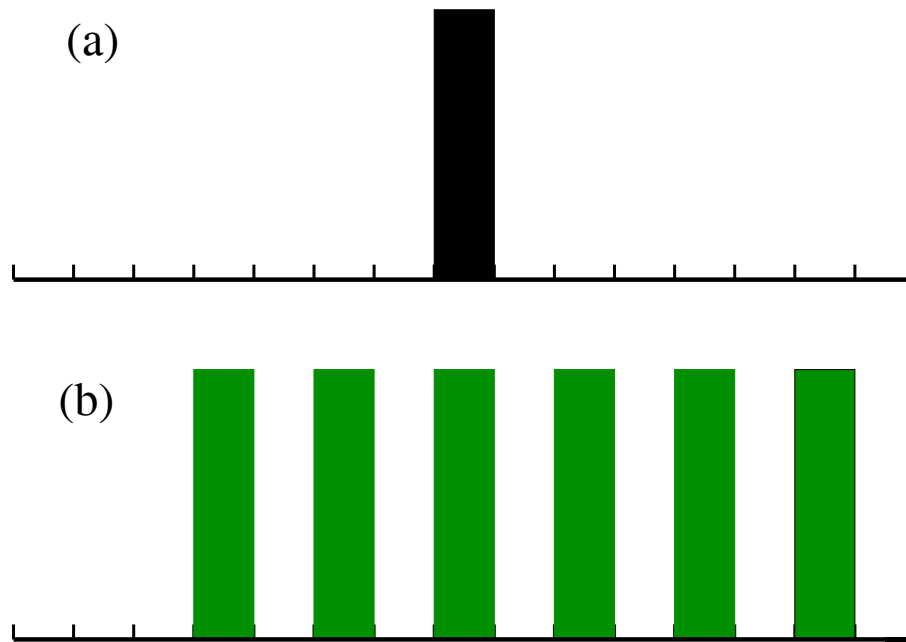


Figure S13: In SICER, the significance of a local enrichment is context dependent. Shown are schematic cases of an enriched window in two different enrichment contexts. In (a), the central enriched window is by itself, and not enriched enough to be significant. In (b), because of the presence of neighboring enriched windows, the central enriched window (along with the other members in the cluster) becomes significant. In contrast, approaches based on local-statistics would find the central enriched window to be insignificant in both case (a) and case (b). Of course, if the central window is already highly enriched by itself, it will be deemed as significant by SICER and by approaches based on local-statistics.

Reference:

Cui, K.R., Zang, C.Z., Roh, T.Y., Schones, D.E., Childs, R.W., Peng, W.Q. and Zhao, K. (2009) Chromatin Signatures in Multipotent Human Hematopoietic Stem Cells Indicate the Fate of Bivalent Genes during Differentiation, *Cell Stem Cell*, **4**, 80-93.
Smith, A.D. (2009) Private Communication