RSEG: Supplmentary Information

Qiang Song , Andrew Smith *
Molecular and Computational Biology, University of Southern California,
Los Angeles, California 90089

1 Modeling Read Counts and Read Count Differences

1.1 Negative binomial distribution

In Fig. 1, we show the application of the Poisson distribution and the negative binomial distribution to H3K27me3 and H3K36me3 ChIP-Seq data. The negative binomial distribution has a smaller BIC (Bayesian Information Criterion) and therefore is more appropriate to model read counts data.

The negative binomial distribution may be formulated in several ways. In this paper the negative binomial distribution is formulated as the number of failures before r successes where the success probability is p. With this formulation, its support is $\{0, 1, 2, \dots, \}$. The probability mass function is

$$f(n; r, p) = \binom{n+r-1}{r-1} p^r (1-p)^n. \tag{1}$$

In this article we sometimes re-parameterize the negative binomial distribution with the mean μ and the dispersion parameter α, where μ and α are related to r and p as following

$$\mu = r\frac{1-p}{p}, \tag{2a}$$

$$\alpha = 1/r. \tag{2b}$$

With this parameterization, μ is the expected number of failures before r successes and α controls how dispersed the number is.

1.2 NBDiff distribution

The NBDiff distribution is the discrete distribution of the difference between two independent negative binomial random variables. It has four parameters: r_1 and p_1 from the first negative binomial distribution and r_2 and p_2 from the second one.

Probability mass function: The probability mass function of the NBDiff distribution is given in Eq. (3),

$$f(k; r_1, p_1, r_2, p_2) = \begin{cases} f(k; r_1, p_1) p_2^{r_2} \times \ {}_2F_1\Big(k + r_1, r_2, k+1, (1-p_1)(1-p_2)\Big), & k \geq 0 \\ f(|k|; r_2, p_2) p_1^{r_1} \times \ {}_2F_1\Big(k + r_2, r_1, |k|+1, (1-p_1)(1-p_2)\Big), & k < 0 \end{cases} \tag{3}$$

where $f(n; r, p)$ is the probability mass function of the negative binomial distribution and ${}_2F_1(a, b; c; x)$ is the Gaussian hypergeometric function (Barnes, 1908).

*to whom correspondence should be addressed

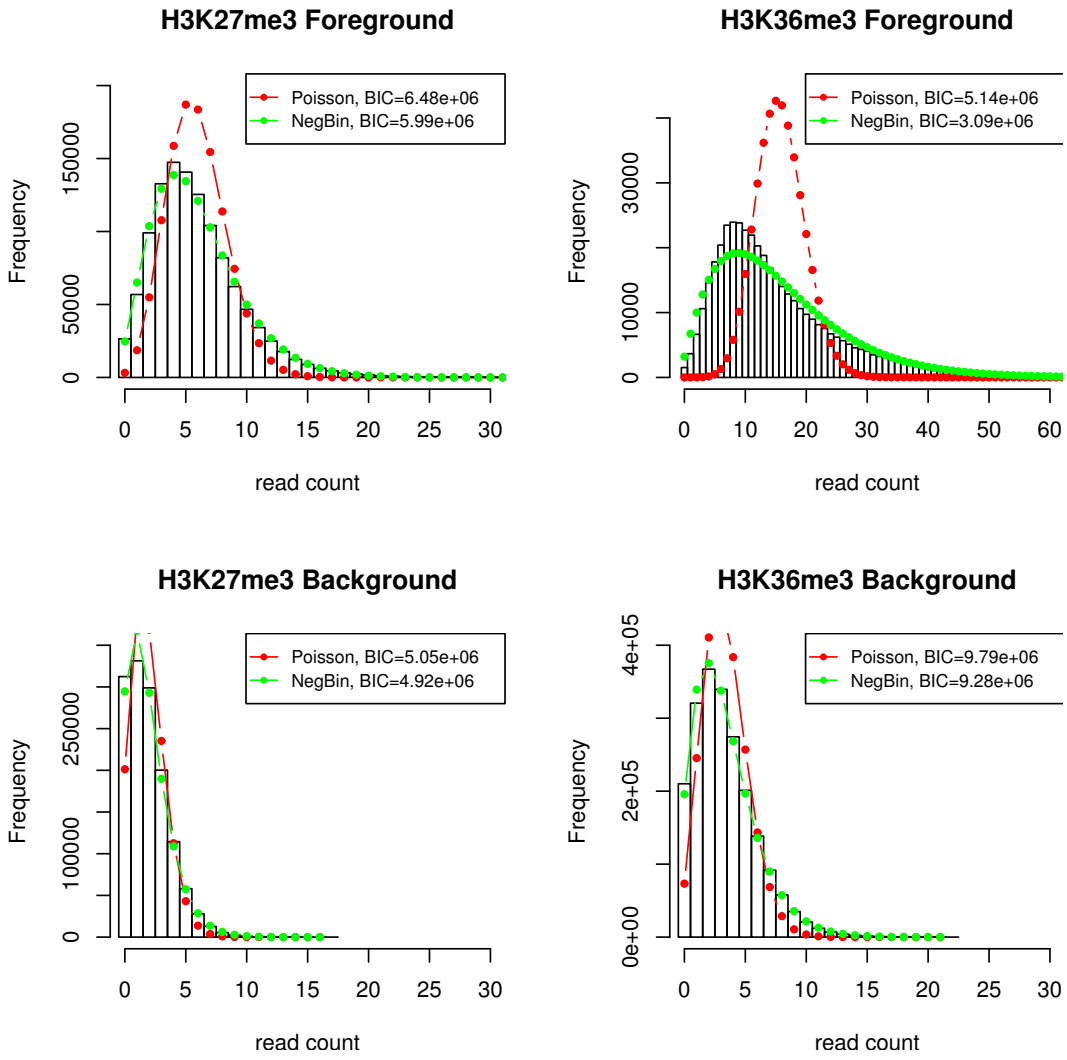


Figure 1: Comparing the Poisson distribution and the negative binomial distribution to model read counts data

We show the derivation of the NBDiff probability mass function for $k > 0$ below where $(x)_n = x(x+1)\dots(x+n-1)$,

$$\begin{aligned}
f(k; r_1, p_1, r_2, p_2) &= \sum_{n=0}^{\infty} f(n+k; r_1, p_1) f(n; r_2, p_2) \\
&= \sum_{n=0}^{\infty} \binom{n+k+r_1-1}{r_1-1} p_1^{r_1}(1-p_1)^{n+k} \binom{n+r_2-1}{r_2-1} p_2^{r_2}(1-p_2)^n \\
&= \sum_{n=0}^{\infty} \frac{(n+k+r_1-1)!}{(r_1-1)!(n+k)!} p_1^{r_1}(1-p_1)^{n+k} \frac{(n+r_2-1)!}{(r_2-1)!n!} p_2^{r_2}(1-p_2)^n \\
&= \sum_{n=0}^{\infty} (k+r_1)_n \binom{k+r_1-1}{r_1-1} \frac{1}{(k+1)_n} p_1^{r_1}(1-p_1)^{n+k} \\
&\quad (r_2)_n \frac{1}{n!} p_2^{r_2}(1-p_2)^n \\
&= \binom{k+r_1-1}{r_1-1} p_1^{r_1}(1-p_1)^{n+k} p_2^{r_2} \sum_{n=0}^{\infty} \frac{(k+r_1)_n (r_2)_n}{(k+1)_n n!} (1-p_1)^n (1-p_2)^n \\
&= f(k; r_1, p_1) p_2^{r_2} \times {}_2F_1\Big(k+r_1, r_2, k+1, (1-p_1)(1-p_2)\Big).
\end{aligned}$$

Parameter estimation: To estimate the four parameters, we make use of the fact that both observation sequences (read counts) from the test sample and the control sample are available, and estimate r_1 and p_1 from the first read count sequence and r_2 and p_2 from the second. Please refer to Section 1.3 for details estimating these parameters.

1.3 Deadzones

Some regions of a genome cannot be interrogated either because they are unassembled or because they are unmappable. The unassembled parts of a genome assembly are usually placed in "chrUn" and are denoted by N's in the sequences of the corresponding chromosomes. We exclude these unassembled regions from our analysis.

Mapping ambiguity occurs when a single sequenced read maps to more than one genomic location. Such ambiguous reads are discarded in most mapping methods and ChIP-Seq analysis approaches. When two locations in the genome have identical sequences of length greater than or equal to the read length, any read derived from one of those locations will necessarily be ambiguous. The result is that a certain fraction of the genome cannot be interrogated using short sequenced reads. We refer to contiguous sets of locations to which no read can map uniquely as "deadzones". If some deadzones are large, we ignore these regions from our analysis because short read sequencing cannot provide reliable information for these regions. For those bins outside of big deadzones, we correct the deadzone effect with the following approach.

We denote the raw read count in the i^{th} bin by X_i, and the proportion of non-deadzone regions by $s_i = (1 - S_i^d/S_i)$, where S_i^d is the length of deadzones and S_i is the bin size. If the read counts are modeled with Poisson distribution Pois(λ), where λ is the parameter without deadzones, the probability of observing X_i with non-deadzone proportion s_i is Pois($X_i; s_i\lambda$).

If the read counts are modeled with the negative binomial distribution NB(r, p), where r and p are the parameters without deadzones, we derive the probability of observing X_i with non-deadzone proportion s_i as following. The negative binomial distribution arises as a continuous mixture of Poisson distribution whose parameter λ follows Gamma distribution $G(r, p/(1-p))$. Similar to the Poisson distribution, λ is scaled by s_i to get the parameter $\lambda_d = s_i\lambda$ with deadzone scaling. The contribution of λ_d component with deadzone scaling should be proportional to that of λ component without deadzone scaling, that is

$$G\Big(s_i\lambda; r_d, \frac{p_d}{1-p_d}\Big) \propto G\Big(\lambda; r, \frac{p}{1-p}\Big),$$

which gives

$$r_d = r, \quad p_d = \frac{s_i p}{1 - p + s_i p}.$$

Therefore the probability of observing X_i with non-deadzone proportion s_i is NB$(r, (s_i p)/(1 - p + s_i p))$. This transformation is equivalent to set $\mu_d = s_i \mu$ if we re-parametrize negative binomial distribution with the mean μ and the dispersion parameter α.

If we use the above approach to correct the effect of deadzones, given the observation sequence $(X_1, \ldots, X_n)$, the maximum likelihood estimator of μ and α is given by

$$\mu = \frac{\sum X_i}{\sum s_i}, \tag{4}$$

and α is calculated numerically by finding α satisfying the following equation

$$\sum_{i=1}^{n} \Big(\sum_{j=0}^{X_i - 1} \frac{j}{1 + j\alpha} + \frac{1}{\alpha^2} \ln(1 + s_i \mu \alpha) + \frac{1}{\alpha} \times \frac{s_i \mu}{1 + s_i \mu \alpha} + \frac{X_i s_i \mu}{1 + s_i \mu \alpha} \Big) = 0. \tag{5}$$

Similar strategies may be applied to read count difference data. We denote the observations in a two-sample analysis as $\{(X_{11}, X_{12}, s_1), (X_{21}, X_{22}, s_2), \ldots, (X_{i1}, X_{i2}, s_i), \ldots, (X_{n1}, X_{n2}, s_n)\}$, where X_{i1} is the read count in the i^{th} bin of the first sample and X_{i2} is the read count in the i^{th} bin of the second sample. As above, s_i is the non-deadzone proportion in the i^{th} bin. Then the difference D_i ($D_i = X_{i1} - X_{i2}$) follows the NBDiff distribution NBDiff$(r_1, (s_i p_1)/(1 - p_1 + s_i p_1), r_2, (s_i p_2)/(1 - p_2 + s_i p_2))$.

Instead of estimating the parameters r_1, p_1, r_2 and p_2 directly, we estimate its re-parametrized form μ_1, α_1, μ_2 and α_2. From the observations $(X_{11}, \ldots, X_{n1})$, the mean μ_1 is estimated by

$$\mu_1 = \frac{\sum X_{i1}}{\sum s_i},$$

and the dispersion parameter α_1 is estimated by numerically finding the root of the following equation:

$$\sum_{i=1}^{n} \Big(\sum_{j=0}^{X_{i1} - 1} \frac{j}{1 + j\alpha_1} + \frac{1}{\alpha_1^2} \ln(1 + s_i \mu_1 \alpha_1) + \frac{1}{\alpha_1} \times \frac{s_i \mu_1}{1 + s_i \mu_1 \alpha_1} + \frac{X_{i1} s_i \mu_1}{1 + s_i \mu_1 \alpha_1} \Big) = 0.$$

Similarly the mean μ_2 and the dispersion parameter α_2 are estimated from the observations $(X_{12}, \ldots, X_{n2})$. The parameters μ_1, α_1, μ_2 and α_2 are transformed to more familiar notations r_1, p_1, r_2 and p_2 according to the relationship defined in Eq. (2).

1.4 Fitting real data

We applied the negative binomial distribution to model read count data and it fits well for varying sample sizes, bin sizes and both enzymatic method and sonication method for fragmenting chromatin (Fig. 2 A-D). The NBDiff distribution models the read count difference between a test sample and a control sample well (Fig. 2 E) and gives acceptable approximation of the read count difference between two cell types (Fig. 2 E-F).

1.5 Bin sizes

In a ChIP-Seq experiment, the probability that a read is sampled from a certain genomic location is a function of its location and we desire a number of reads that provide an accurate estimate to this function. We can see immediately that sequencing enough reads to accurately estimate the value associated with each base requires a prohibitive amount of sequencing. However, if we are interested in the mean probability associated with a genomic bin, then the task may be practical. Obviously, the optimal bin size depends on the total number of reads, i.e., the more reads there are the smaller the bin size is. This question of optimal bin size have been studied in the context of optimal intervals for building a histogram.

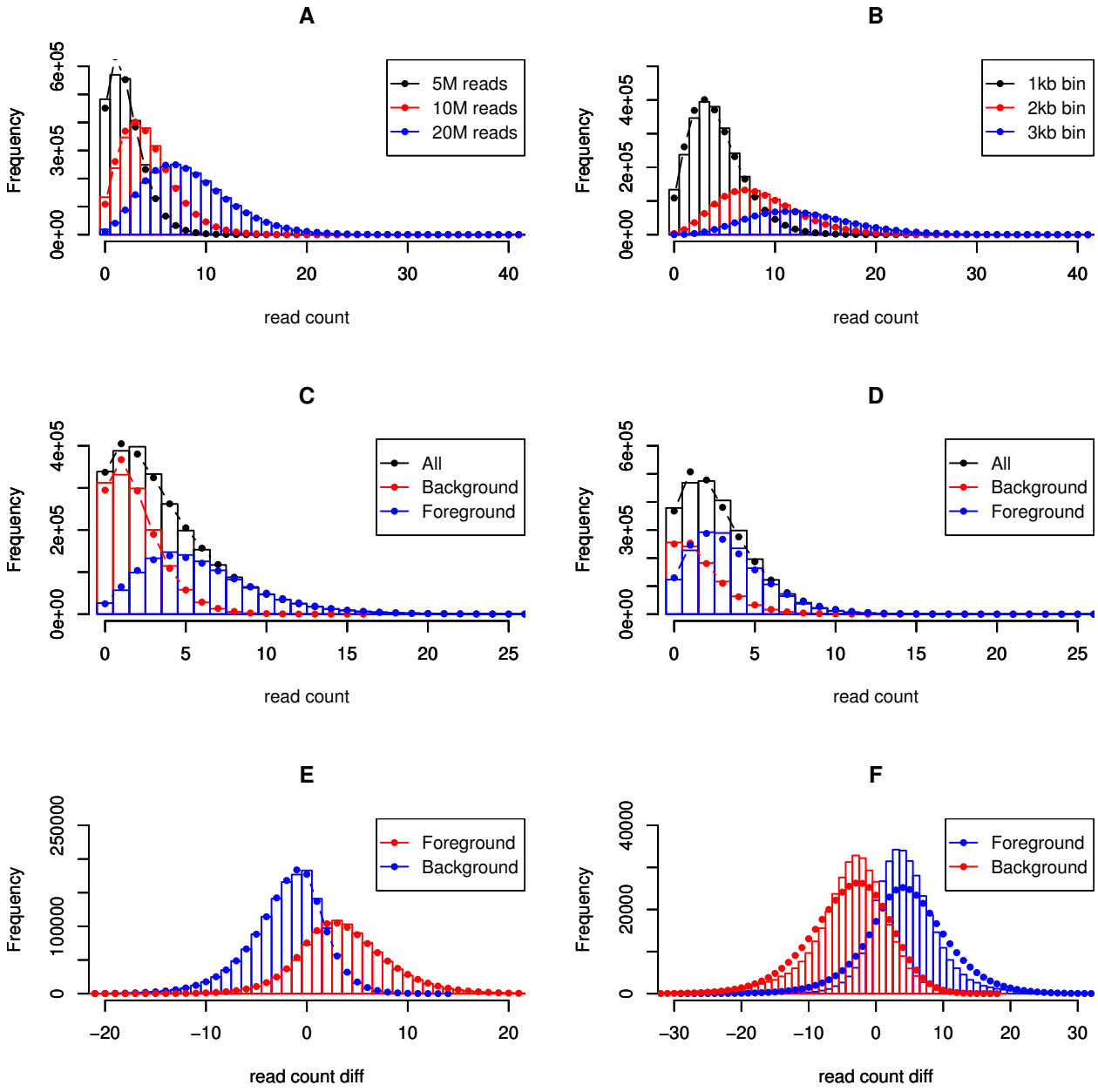


Figure 2: Modeling read counts and read count difference with the negative binomial distribution (A-D) and the NBDiff distribution(E-F). **A**: the negative binomial distribution fits read counts of H3 samples for different sample sizes (unpublished data); **B**: the negative binomial distribution fits for varying bin sizes (unpublished data); **C**: the negative binomial distribution fits read counts of the H3K27me3 ChIP-Seq sample fragmented with micrococcal nuclease; **D**: the H3K27me3 ChIP-Seq sample fragmented with sonication; E: NBDiff distribution fits read count difference between a H3K27me3 sample and a H3 control sample; F: NBDiff distribution fits read count difference of H3K27me3 samples between $CD133^+$ and $CD36^+$ cells

By default, RSEG use the iterative approach given in (Shimazaki & Shinomoto, 2007) to select the bin size automatically. As mentioned in the beginning of this section, we treated the probability that a read is sampled from a certain genomic location is a function of its location. We denote this function as $f(x)$, where x refers to a genomic lo-

cation. Let n be the total number of reads in a ChIP-Seq experiment and λ be the optimal bin size, Waterman et al gave the asymptotic formula of λ in Eq. 6 by bounding the Komolgorov-Smirnov statisitcs (Waterman & Whiteman, 1978),

$$\lambda = \begin{cases} cn^{-1/4} & \text{if } |f'(x)| < C \\ cn^{-1/6} & \text{if } |f'(x) - f'(y)| < L|x-y|, \end{cases} \tag{6}$$

where c is a constant independent of the number of data points. Shimazaki et al gave another asymptotic formula of the optimal bin size in Eq. 7 by minimizing the mean integrated squared error (MISE) (Shimazaki & Shinomoto, 2007),

$$\lambda = \begin{cases} cn^{-1/3} & \text{if } f(x) \text{ changes smoothly} \\ cn^{-1/2} & \text{if } f(x) \text{ fluctuates in a zigzag pattern}, \end{cases} \tag{7}$$

RSEG can also use these asymptotic formula to select the bin size for different number of reads, which is faster than the iterative approach. The constant c is selected empirically from our experiment with real datasets.

2 HMM segmentation model

Our Hidden Markov Model segmentation method is based on the general HMM framework described in (Rabiner, 1989). We train the HMM model with the Baum-Welch method and decode with posterior decoding. When describing the model, we use similar notations as in (Rabiner, 1989). We denote the hidden state at time t by q_t. The posterior probability of being in state S_i at time t given the observations O and the model parameters λ is denoted by

$$\gamma_t(i) = P(q_t = S_i | O, \lambda). \tag{8}$$

The posterior probability of being in state S_i at time t and in state S_j at time $t+1$ is denoted by

$$\xi_t(i,j) = P(q_t = S_i, q_{t+1} = S_j | O, \lambda). \tag{9}$$

Both these posterior probabilities $\gamma_t(i)$ and $\xi_t(i,j)$ can be computed efficiently with the Forward-Backward algorithm (Rabiner, 1989).

Since the general framework of the two-state HMMs are similar to that of three-state HMM, we will first describe the three-state HMM in detail and then briefly mention key points for two-state HMMs.

2.1 Three-state segmentation model

In the HMM model for comparing two ChIP-Seq samples, we introduce three states, namely the basal state, the Sample 1 enriched state and the Sample 2 enriched state. A large part of the genome belongs to the basal state, where the two samples show similar modification pattern. The read count differences of basal state bins, not necessarily zero, give us a reference line. The Sample 1 enriched state means that Sample 1 is enriched relative to Sample 2 after accounting for the basal state reference line, i.e. the read count differences are higher than the basal state reference line. The Sample 2 enriched state means that Sample 2 is enriched relative to Sample 1 after accounting for the basal state reference line. Thereafter we denote the Sample I enriched state, the basal state and the Sample 2 enriched state by S_0, S_1 and S_2 respectively.

The observations are $\{(X_{11}, X_{12}, s_1), (X_{21}, X_{22}, s_2), \ldots, (X_{i1}, X_{i2}, s_i), \ldots, (X_{n1}, X_{n2}, s_n)\}$ where X_{i1} and X_{i2} are read counts in i^{th} bin of Sample 1 and 2 and s_i are non-deadzone proportion. We are interested in the read count differences D_i calculated with $D_i = X_{i1} - X_{i2}$. We use the NBDiff distribution as emission distribution, i.e.

$$b_i(S) = \text{NBDiff}_S(D_i; r_1, p_1, r_2, p_2),$$

where the S denotes the hidden state and the parameters r_1, p_1, r_2 and p_2 need to be adjusted for deadzone effect on a per-bin basis.

The parameters of the emission distributions are initiated with estimates from a mixture model of three NBDiff distributions fitted with the observations. The transition probabilities are initiated from the empirical distribution of the length of dispersed epigenomic domains. During the Baum-Welch training process, we iterate the following steps until the change of the total likelihood is smaller than a given threshold, say 10^{-10}.

Expectation step: Given current model parameters, compute the two posterior probabilities $\gamma_t(i)$ and $\xi_t(i,j)$ with the Forward-Backward algorithm.

Maximization step: Given posterior probabilities $\gamma_t(i)$ and $\xi_t(i,j)$, find the model parameters that maximize the total likelihood. The transition probability a_{ij} is estimated simply by

$$a_{ij} = \frac{\sum_{t=1}^{n-1} \xi_t(i,j)}{\sum_{t=1}^{n-1} \gamma_t(i)}. \tag{10}$$

The parameters of the emission distributions, i.e. the NBDiff distributions, are estimated as we described in Section 1.3 with the posterior probabilities taken into account. As an example, we will show how to estimate the parameters of the emission distribution of the state S_0. The first mean parameter μ_1 is estimated by

$$\mu_1 = \frac{\sum X_{i1}\gamma_i(0)}{\sum s_i\gamma_i(0)},$$

and the first dispersion parameter α_1 is estimated by numerically finding the root of the equation

$$\sum_{i=1}^{n} \gamma_i(0)\Big(\sum_{j=0}^{X_{i1}-1} \frac{j}{1+j\alpha_1} + \frac{1}{\alpha_1^2}\ln(1+s_i\mu_1\alpha_1) + \frac{1}{\alpha_1} \times \frac{s_i\mu_1}{1+s_i\mu_1\alpha_1} + \frac{X_{i1}s_i\mu_1}{1+s_i\mu_1\alpha_1}\Big) = 0.$$

The second mean parameter μ_2 and the second dispersion parameter α_2 can be estimated in the same way from the observations $(X_{12}, \ldots, X_{n2})$.

After the training procedure converges, we decode the state of each bin with posterior decoding and consecutive bins of the same state are joined to form a epigenomic domain.

2.2 Two-state segmentation model

Our two-state HMMs are used to find dispersed epigenomic domains. The foreground state is enriched with ChIP signal and the background state reflects random noises. In single sample analysis, we only observe read counts in the test sample, denoted as $(X_1, X_2, \ldots, X_n)$. The negative binomial distribution is used as emission distribution, i.e.

$$b_i(S) = \text{NBD}_S(X_i; r_1, p_1, r_2, p_2).$$

If a test sample is also available, denoted as $(C_1, C_2, \ldots, C_n)$, we model the read count differences D_i ($D_i = X_i - C_i$) with the NBDiff distribution

$$b_i(S) = \text{NBDiff}_S(D_i; r_1, p_1, r_2, p_2).$$

The model initiation, parameter estimation and segmentation is similar to that in three-state HMM.

3 Characterizing Boundaries

While the predicted domains themselves give the locations of boundaries, in order to effectively analyze domain boundaries we must be able to assign some measure of quality to them.

We evaluate domain boundaries based on posterior probabilities of transitions between the foreground state and the background state as estimated by the HMM, which is

$$p_i^{fb} = P(q_{i-1} = f, q_i = b|\mathbf{O}).$$

For each pair of consecutive genomic bins, the posterior probability is calculated for all possible transitions between those bins. As shown in Fig. 3, the i^{th} bin corresponds to the beginning of Domain f (since the i^{th} bin is closer to the beginning of the domain than to the end of the domain), the boundary score for the i^{th} bin is the probability of a background to foreground transition , i.e p_i^{bf}. If the i^{th} bin is closer to the end of the domain, the boundary score of

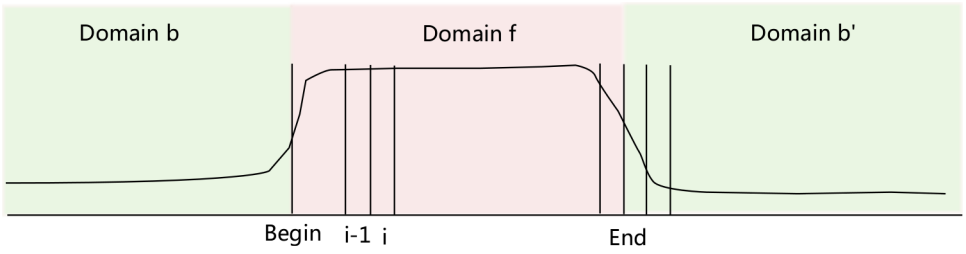


Figure 3: Illustration of boundary score computation

the i^{th} bin will be the posterior probability of foreground to background transition. The boundary score of a bin in a background domain is computed in the same spirit.

We then permute all bins in the experimental dataset and the permuted dataset serves a background control. We then decode the permuted dataset with the HMM parameters obtained from the original dataset and compute the posterior transition probabilities of the new dataset. This new set of transition probabilities gives an approximation of the distribution of posterior transition probabilities in a background sample, and we use this empirical distribution to assign p-values to the boundary scores in the experimental dataset.

Those bins whose p-value are smaller than 0.05 are considered significant. Those bins with significant posterior transition probabilities are kept and consecutive significant bins are joined as being a boundary. The size of this boundary is the number of consecutive significant bins. We score this boundary with the posterior probability that a single transition occurs in this boundary. Suppose a boundary are composed of k bins, the score of this boundary is computed as following:

$$p = p_1^{fb} \prod_{j=2}^{k} p_j(f|b) + \sum_{i=2}^{k} \Big(p_1^{ff} \prod_{j=2}^{i-1} p_j(f|f) \times p_i(b|f) \times \prod_{j=i+1}^{k} p_j(b|b) \Big),$$

where

$$p_j(b|f) = P(q_j = b | q_{j-1} = f, \mathbf{O}) = \frac{p_j^{fb}}{p_j^{ff} + p_j^{fb}}.$$

The peak of this boundary is the start of the bin with the largest p_i^{fb}.

4 Evaluation

4.1 Simulation scheme and evaluation method

Since H3K36me3 almost covers entire gene bodies, we randomly selected 6000 genes from mouse RefSeq genes that are outside of large deadzones and treated these regions as enriched with histone modifications. We simulated enrichment ratios for these enriched regions from a Gaussian distribution whose mean is specified by the user and standard deviation is equal to the half of its mean. The number of reads in a region is proportional to its effective length. Next we constructed foreground library of $62,878,726$ reads by combining all ChIP-Seq reads from embryonic stem cells, neural progenitor cells and embryonic fibroblasts (Mikkelsen et al., 2007) and a background library of $9,452,354$ reads from whole cell extract experiments (Mikkelsen et al., 2007). We then sampled around $8,000,000$ reads from the background library and the foreground library according to the number of reads in each region computed above. We simulated a test sample in which there are 6000 enriched domains with average enrichment ratio 8.0. We also simulated control datasets by uniformly sample 6500000 reads from the foreground library and 1000000 reads from the background library.

We use Precision-Recall Curve (AUC-PR) to compare the sensitivity and specificity of those methods. To account for the nature of dispersed domains, a simulated domain is treated as "recalled" only when over certain proportion, for

example 10% or 50%, of that domain is covered by predicted domains. Similarly, a predicted domain is called " true positive" only when over certain proportion of that domain is covered by simulated domains.

4.2 Performance

We compared RSEG with SICER (Zang et al., 2009) and HPeak (Qin et al., 2010) with the simulated dataset. SICER is the first specialized software package that aims to analyze dispersed histone modification domains and outperforms other ChIP-Seq analysis software. It serves a good benchmark for comparison. The sensitivity and specificity of SICER is adjusted by changing E-value (from $1e-12$ to 1200) for single-sample analysis and FDR (from $1e-5$ to $2e-1$) for two-sample analysis. The performance of SICER can be also tuned by window size and allowable number of gaps. Our experiments with different combinations of parameters shows that the effect of window size is slight and allowing more gaps make it possible to identify larger domains. We reported the result based on SICER's default parameters, i.e. window size 200bp and gap number 3. HPeak uses a similar Hidden Markov Model, but is designed with transcription factor binding site in mind. The sensitivity and specificity for HPeak is adjusted by changing p-value threshold (from $1e-5$ to $2e-1$). We also ran HPeak with different window size (25bp, 200bp and 500bp) and the result changed slightly, therefore ruling out the effect of window size.

First we evaluated these three methods for single-sample analysis. As shown in Fig. 4, RSEG has a larger AUC-PR than SICER and HPeak and performs favorably as a method to find dispersed epigenetic domains. SICER also proves a useful tool to find dispersed domains, but is less sensitive than RSEG under the same specificity. HPeak, which is designed to find localized "peaks", achieves high specificity (near 100%) but fails to recover most of the simulated domains.

Next we evaluated these methods with a control sample. As shown in Fig. 5, SICER slightly outperforms RSEG when we are interested highly specific domains while RSEG has higher sensitivity if we allow higher false positive rate. Both of them work better than HPeak, which suffers from low sensitivity. The inclusion of a control sample in the analysis greatly increase the specificity of both SICER and RSEG.

4.3 Evaluating RSEG's domain boundaries identification function

H3K36me3 domains are reported to be associated with active transcribed gene bodies (Bannister et al., 2005; Barski et al., 2007b). Additionally, the histone methyltransferase Set2, which specifically methylates lysine 36 of histone H3, is associated with RNA polymerase II (PolII) (Li et al., 2003). Therefore, we reasoned that the boundaries of H3K36me3 domains should be in the vicinity of either transcription start sites (TSS) and transcription termination sites (TTS).

We evaluated RSEG's ability for identifying domain boundaries with H3K36me3 data from human T Cells (Barski et al., 2007b). We assign each of the 13122 domains boundaries identified by RSEG to its closest TSS (or TTS). Of the 6994 TSS-associated boundaries, 50% are within 6562bp from their respective associated TSS (p-value $< 2.2e-16$). Of the 6128 TTS-associated boundaries, 50% are within 4431bp from their respective associated TTS (p-value $< 2.2e-16$). The strong association between the identified H3K36me3 domain boundaries with TSS and TTS indicates that the domain boundaries identified by RSEG is of biological significance.

The p-value above is calculated as following. The effective size of human genome is 2287968180bp (By subtracting 792451276bp unassembled regions and unmappable regions from the total genome size 3080419456bp). The size of the TSS-vicinity regions is $(2d \times 33879)$bp (d is the distance from TSS and 33879 is the number of genes according to RefSeq). The p-value is the probability that more than 3997 bases are from TSS-vicinity regions when we randomly sample 6994 bases (corresponding to the peak of domain boundaries) from the whole genome. The p-value for TTS-associated boundaries is computed in the same way.

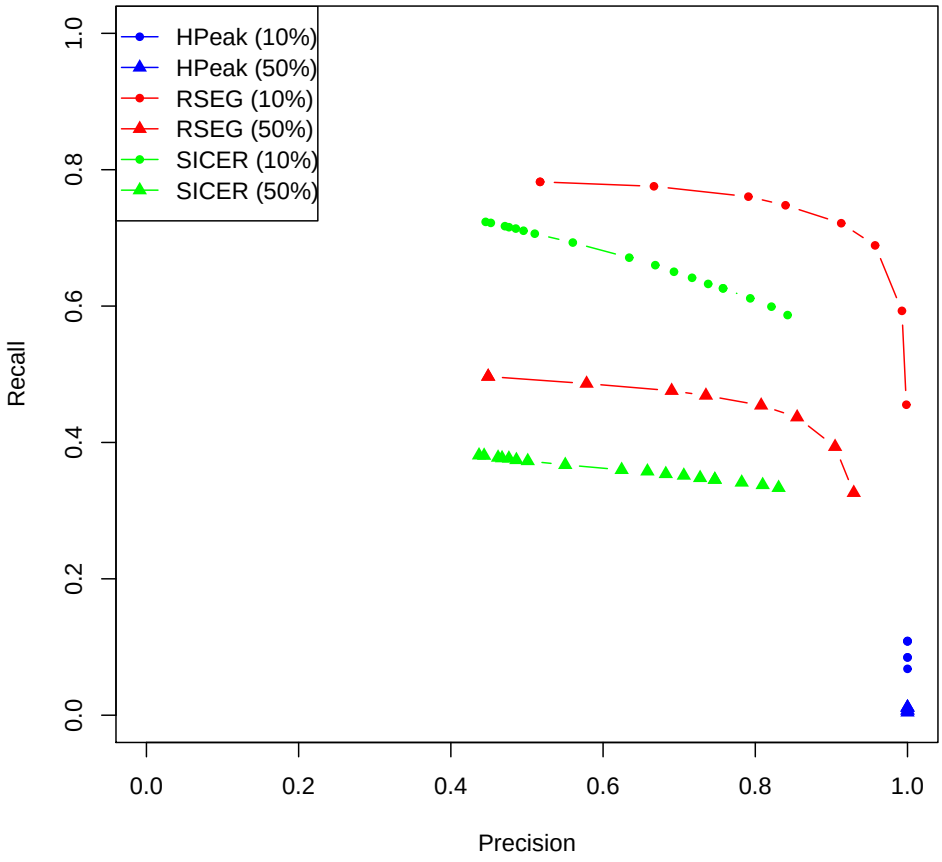


Figure 4: Precision-Recall curve

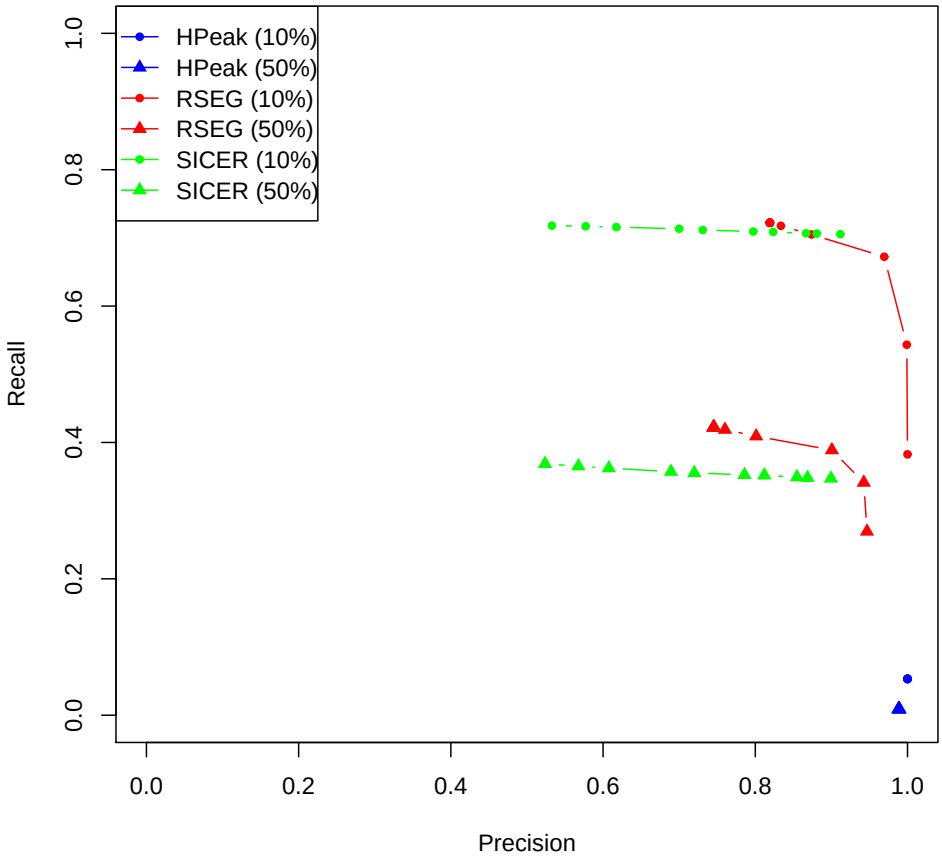


Figure 5: Precision-Recall curve

5 Application

5.1 Domains

We applied RSEG to four histone modification marks (H3K9me3, H3K27me3, H3K36me3 and H3K79me2) from two separate studies (Barski et al., 2007b; Mikkelsen et al., 2007). The identified domains and boundaries are downloadable from `http://smithlab.cmb.usc.edu/histone/rseg/`. We summarized the statistics of the domains in human T Cell sample in Table 1. We analyze these domains with respect to genes. Most of the H3K36me3 and H3K79me2 domains are found in intragenic regions, which is consistent with their role in marking actively transcribed genes. But only half of the H3K9me3 and H3K27me3 domains are located in intragenic regions (Fig. 6). While the H3K9me3 and H3K27me3 marks within genes have been associated with repressive function, the biological meaning of their extensive presence outside of genic regions remains unclear.

Modification	Reads	Domains	Mean size
H3K27me3	8970141	11760	103172
H3K36me3	13572575	11444	40347
H3K79me2	7149387	7546	39755
H3K9me3	6348997	20257	67124

Table 1: Summary of diffuse domains of four histone modification marks in human CD4+ T cells

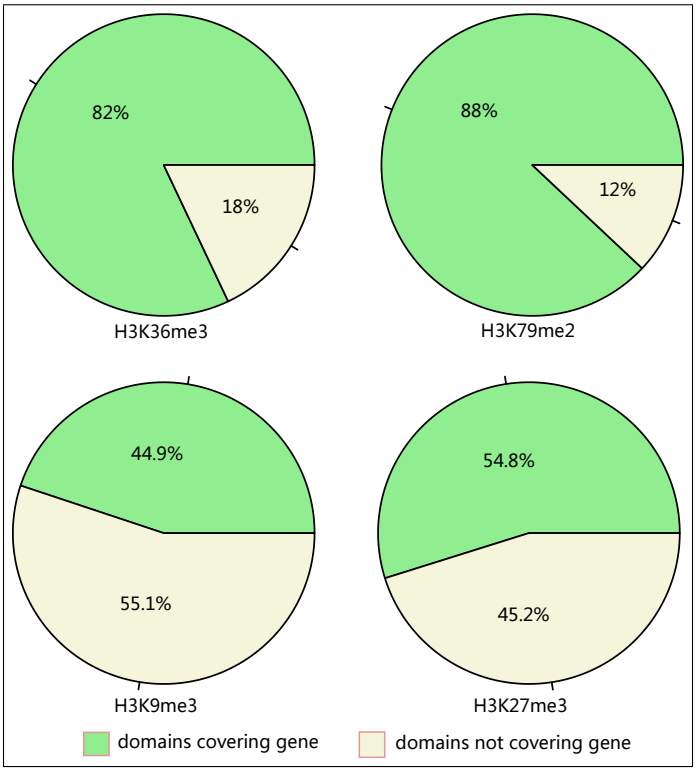


Figure 6: Most H3K79me2 and H3K36me3 domains overlap RefSeq genes while only about half of H3K27me3 and H3K79me2 domains overlap RefSeq genes.

5.2 Boundaries

We further studied the properties of the two gene-overlapping histone modification marks, H3K36me3 and H3K79me2, through boundary analysis. We selected those gene-overlapping domains with significant boundaries , computed the distance between their boundaries and TSS & TTS and filtered those domains with both boundaries within $25000bp$ from the corresponding TSS or TTS (This filtering is to reduce the number of possibly false boundaries). The H3K36me3 domains are found to start downstream of TSS and end downstream of the 3' end of the gene. H3K79me2 domains show a different pattern to H3K36me3. These domains most frequently start around the TSS and end around the TTS of the genes they cover. H3K79me2 tends to associate with 5' ends of genes, while K36 associates with 3' ends (Table 2 with each case illustrated in Fig. 7).

We also selected those genes with both H3K36me3 and H3K79me2 signals and filtered those boundaries located within 25000bp from the corresponding TSS or TSS. H3K79me2 domains tend to precede H3K36me3 domains in these genes (Fig. 8) (170 genes out of 221 selected genes). It is should be interesting to further investigate the interplay of H3K79me2 and H3K36me3 marks. This interesting discovery demonstrates the usefulness of boundary analysis.

Boundaries (5' $\rightarrow$ 3')	H3K79me2	H3K36me3
Upstream TSS $\rightarrow$ Inside Gene	31%	3%
Upstream TSS $\rightarrow$ Downstream TTS	10%	8%
Inside Gene $\rightarrow$ Inside Gene	46%	13%
Inside Gene $\rightarrow$ Downstream TTS	13%	76%
Mean distance to TSS	-1480bp	4501bp
Mean distance to TTS	-5277bp	4736bp

Table 2: H3K36me3 domains are associated with 3' ends of genes and H3K79me2 are associated with 5' ends

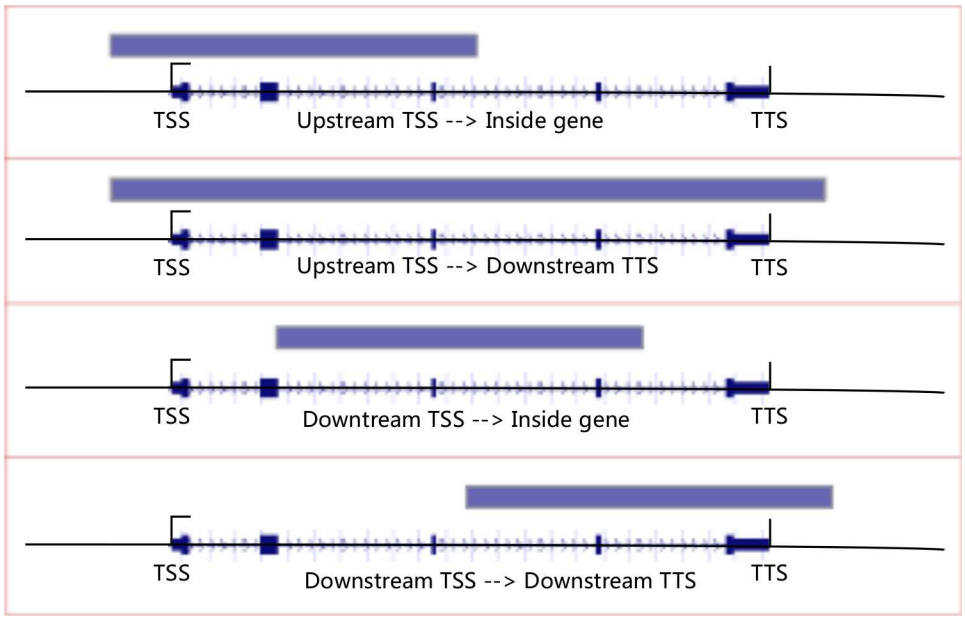
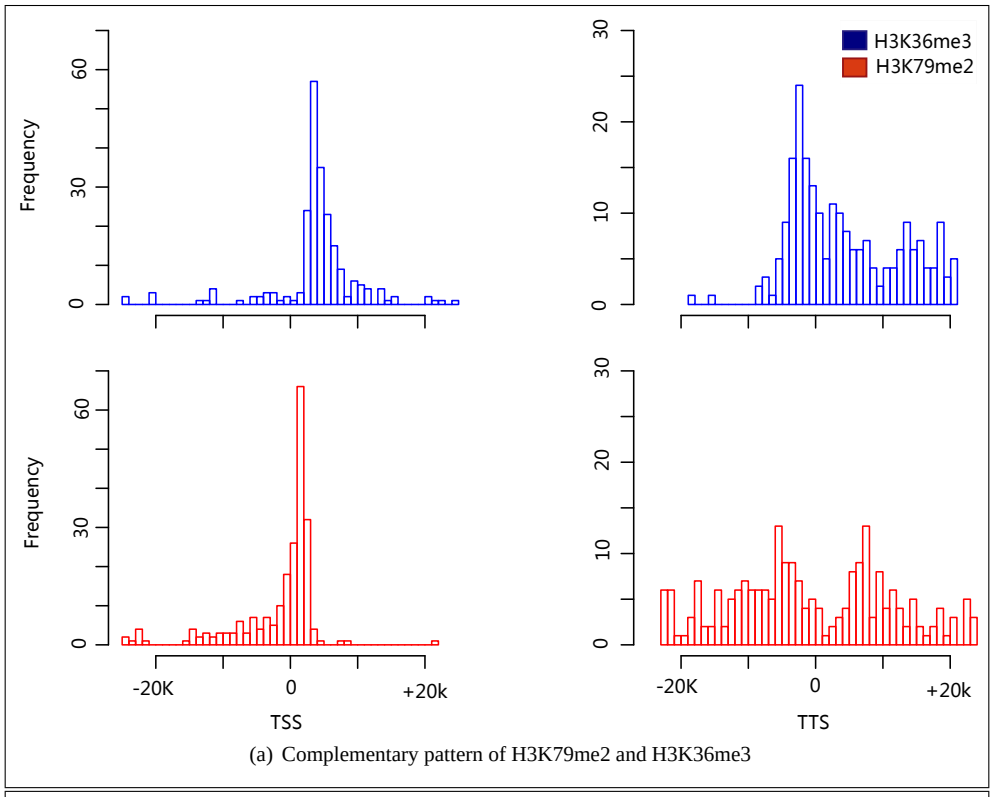
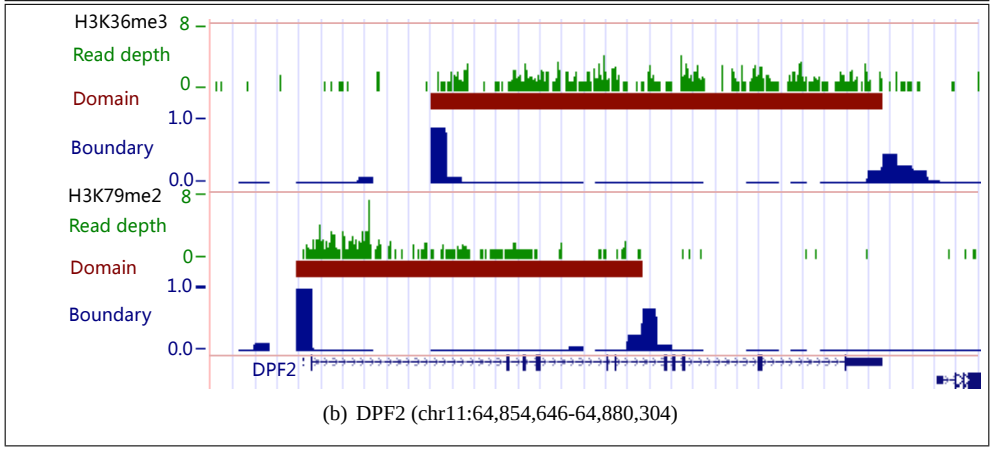


Figure 7: Illustration of the relative position between gene bodies and gene-overlapping domains (not in real scale)



(a) Complementary pattern of H3K79me2 and H3K36me3



(b) DPF2 (chr11:64,854,646-64,880,304)

Figure 8: Genes with both H3K36me3 and H3K79me2 domains, H3K36me3 domains tends to associate with 3' ends and H3K79me2 domain with 5' ends

5.3 Differential Histone Modification Regions

We also tested the ability of RSEG for studying the differential histone modification regions with RSEG. We compared H3K4me3, H3K27me3 and H3K36me3 modification profiles between human hematopoietic stem cells ($CD133^+$) and erythrocyte precursor cells ($CD36^+$) (Cui et al., 2009) with our three-state HMM. As has been expected, we observed extensive histone modification changes during cell differentiation. The identified DHMRs are downloadable from `http://smithlab.cmb.usc.edu/histone/rseg/`. Overall, H3K4me3 tends to decrease when stem cells differentiate into more specialized cells while H3K27me3 tends to increase (Table 3). We reason that this trend reflects that those bivalent genes, that have both H3K4me3 and H3K27me3 marks in HSC and are not $CD133^+$ linage specific,

lose the H3K4me3 mark and are further repressed by the H3K27me3 mark when HSC cells differentiate into $CD133^+$ cells.

We first evaluated the sensitivity of RSEG. We examined the modification pattern at specific genes known to be important for the function of either the hematopoietic stem cells or erythrocyte precursor cells that have been manually identified by (Cui et al., 2009). Among these 9 gene or gene clusters (HoxA5-A9, HoxB5-B6, KLF1, GATA1, CD36, CD34, PROM1, GATA3, PBX1), RSEG correctly identifies the change of histone modification pattern changes. In particular, these genes include CD34 and CD133, two characteristic surface markers of HSC cells (Fig. 9), GATA1 and KLF1, two transcription factors controlling the development of erythroid lineage. (Fig. 10). GATA3 and PBX1 are two transcription factors important for the development of B cells. When HSC cells differentiate into erythroid lineage, they are silenced. RSEG correctly marks the trend of increased H3K27me3 marks and decreased H3K36me3 marks in these two genes though there remains room for improvement in the finding the exact locations. This shows the sensitivity of RSEG to find biologically meaningful DHMRs.

Next we estimated the false discovery rate with the following approach. We randomly divided each sample into two parts and then constructed two synthetic libraries by combining one part from each sample. Next we ran the three-state HMM with the parameters trained from the original datasets on the two synthetic libraries. The combined libraries. Since these two synthetic libraries can be considered as technical replicates, any DHMRs identified with these two synthetic libraries are false positives due to random noise. In this way, we estimated that the false discovery rate of our three-state HMM is 9.3%.

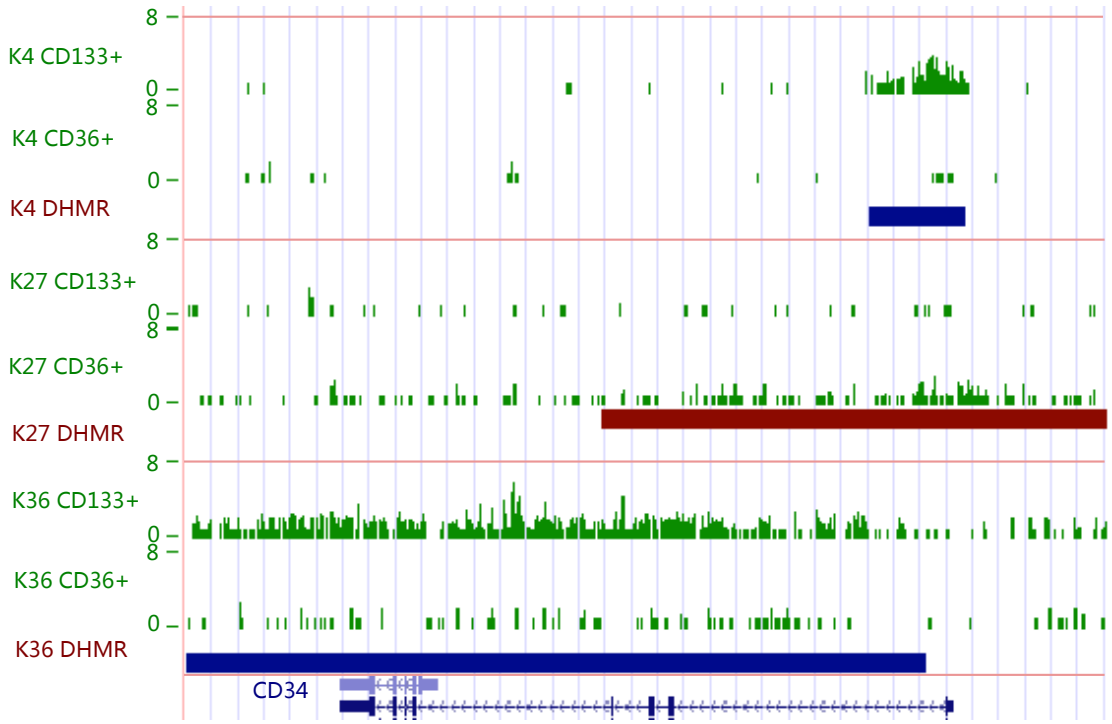
There lacks a "gold standard" set of DHMRs to evaluate the performance of DHMR-finding methods due to our insufficient understanding of the biology of histone modifications. However the promising result of both overall trend and specific examples indicates that RSEG may be useful for the studies of the changes of histone modifications.

Modification	Increased in $CD36^+$	Decreased in $CD36^+$
H3K4me3	1994	5198
H3K27me3	8665	5066
H3K36me3	9405	13878

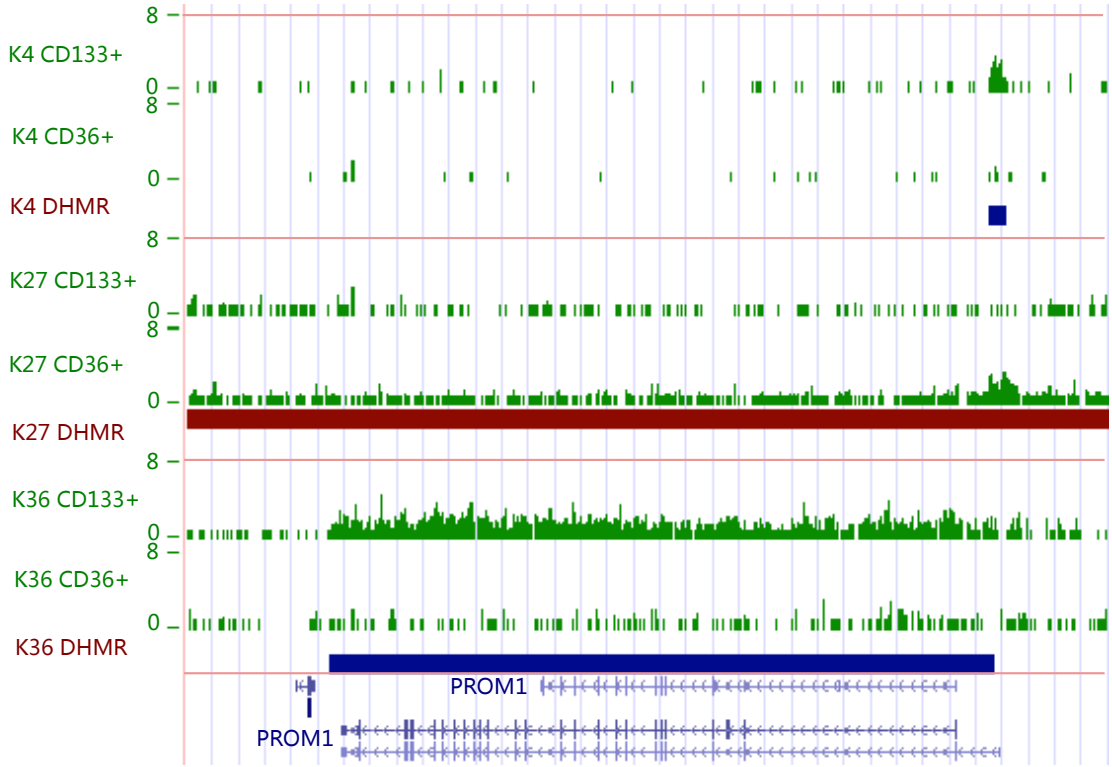
Table 3: Number of differential histone modification domains when human hematopoietic stem cells differentiates into erythrocyte precursor cells

5.4 Simultaneous segmentation with two histone modification marks

As has been observed with boundary analysis, the two gene-covering marks, H3K36me3 and H3K79me2, show different preferences relative to transcription start sites and transcription termination sites. This observation motivates us to analyze these two marks simultaneously. We applied our three-state HMM to H3K36me3 and H3K79me2 samples in human CD4+ T cells (Barski et al., 2007a). We found 18935 regions that are H3K36me3 enriched relative to H3K79me2 and 8138 regions that are H3K79me2 enriched relative to H3K36me3 (downloadable from `http://smithlab.cmb.usc`
The result agrees with our previous observation about the preferences of H3K36me3 mark and H3K79me2 mark. Fig. 12(a) shows the result of previously mentioned DPF2 gene, where we see the complementary pattern of H3K79me2 and H3K36me3 marks. Fig. 12(a) shows another example around the CFL1 gene. This figure illustrates one important observations: the two bins with significant boundary scores over CFL1 gene suggests the transition from H3K79me2 domain to H3K79me2 domain is smooth in some cases and does not have sharp boundary.

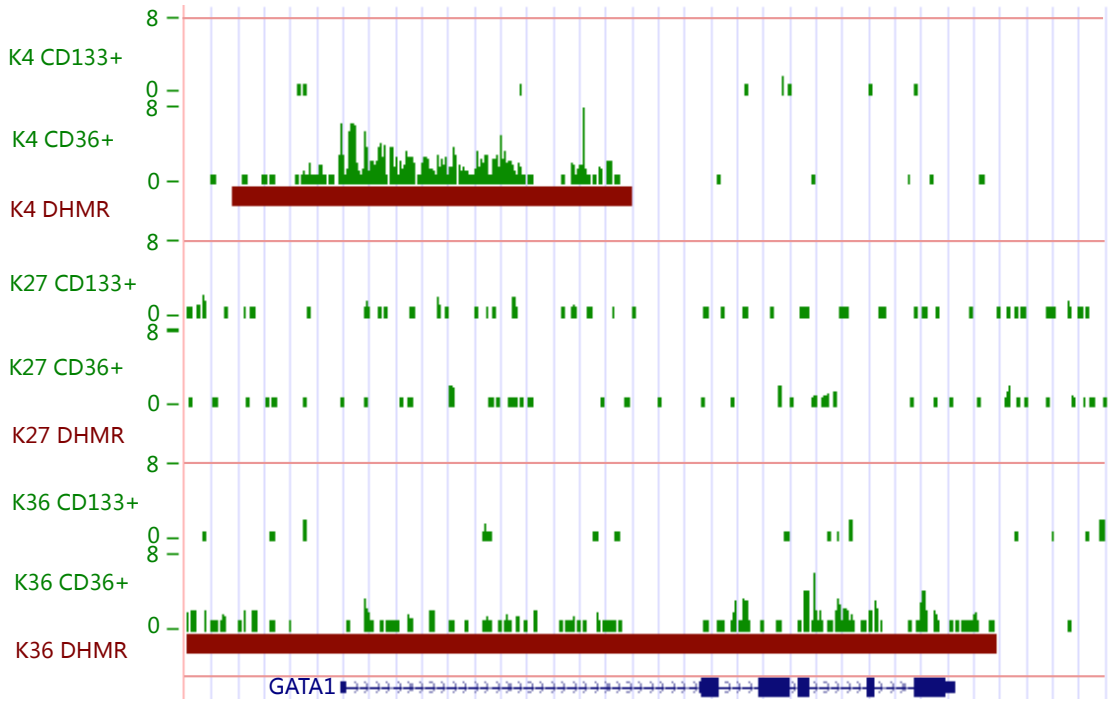


(a) CD34 (chr1:206,120,305-206,157,506)

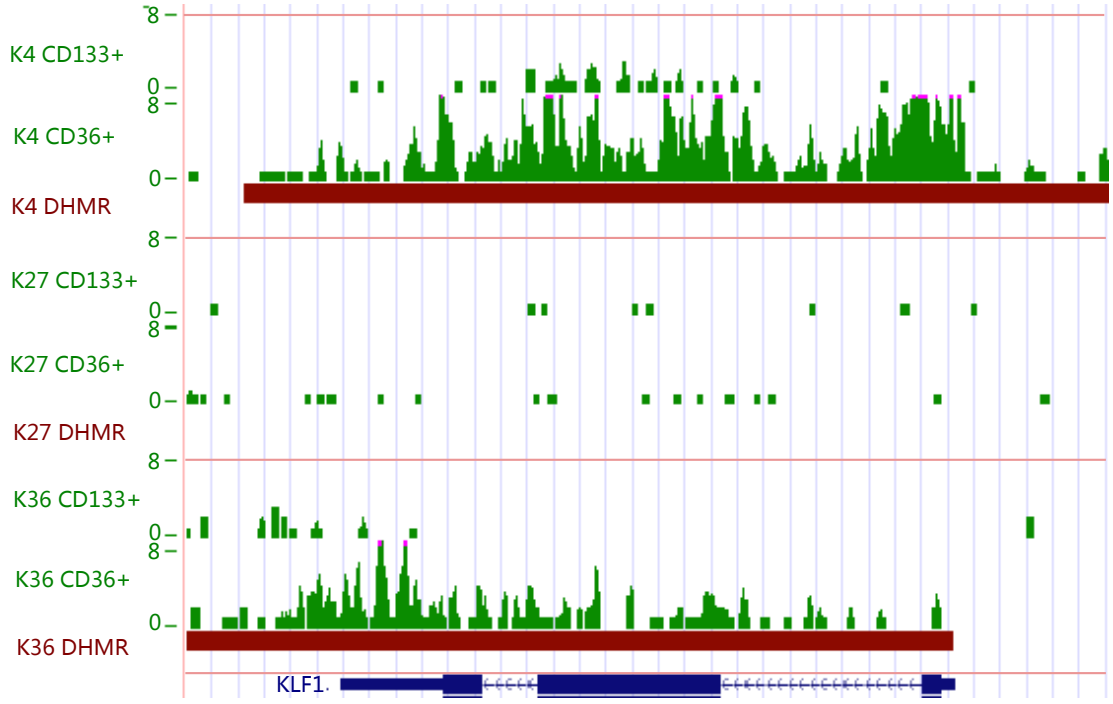


(b) CD133 (chr4:15,551,983-15,713,810)

Figure 9: CD34 and CD133, two surface marker genes for hematopoietic stem cells. Blue bar indicates the histone modification signal is down regulated in $CD36^{+}$ and red bar indicates the opposite change.

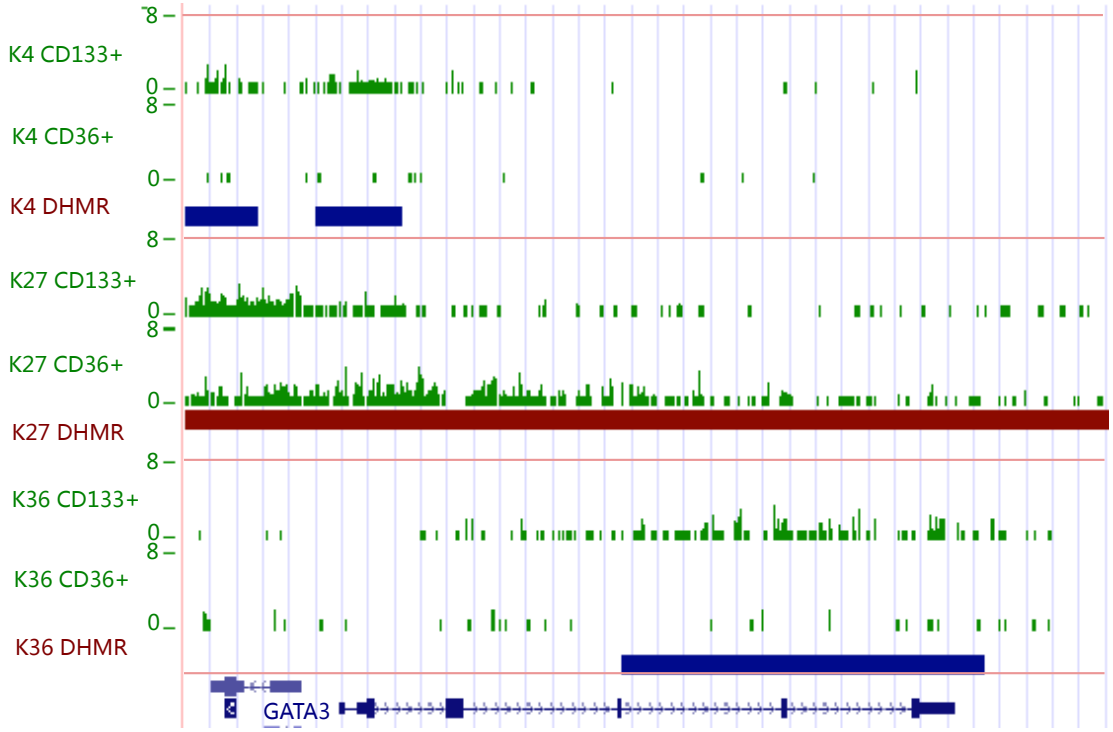


(a) GATA1 (chrX:48,527,992-48,539,595)

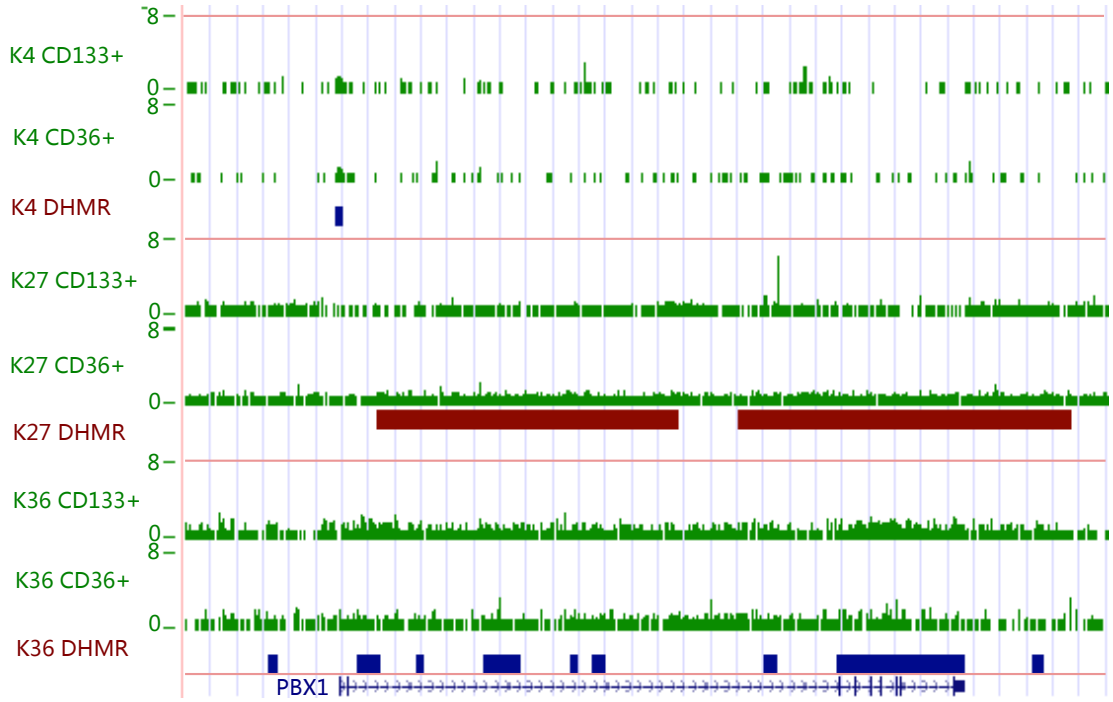


(b) KLF1 (chr19:12,855,541-12,859,712)

Figure 10: GATA1 and KLF1, two transcription factor controlling the development of erythroid lineage. Blue bar indicates the histone modification signal is down regulated in $CD36^{+}$ and red bar indicates the opposite change.

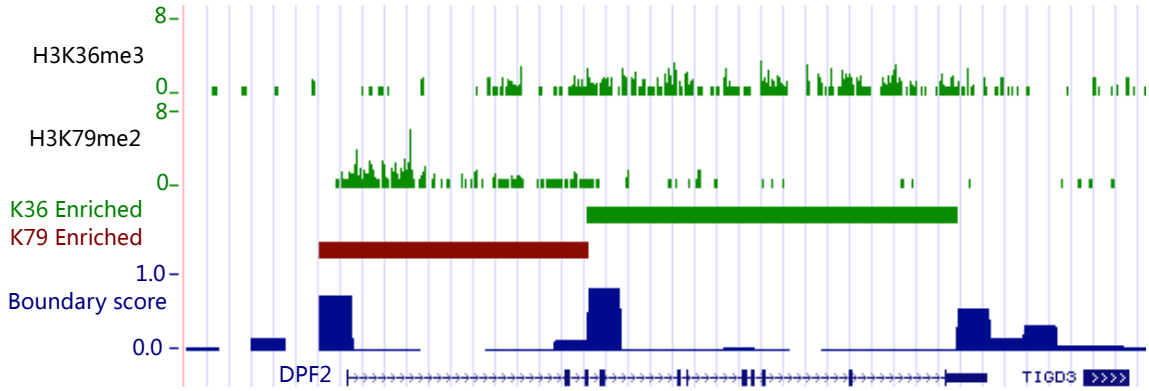


(a) GATA3 (chr10:8,131,549-8,162,295)

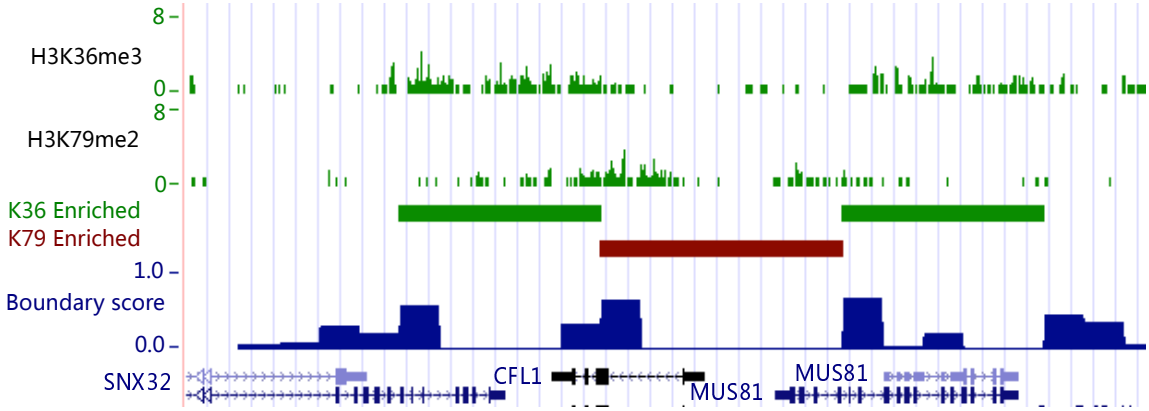


(b) PBX1 (chr1:162,795,561-163,082,934)

Figure 11: GATA3 and PBX1, transcription factors required for development of B cells are silenced in $CD36^{+}$ and show increased level of H3K27me3 marks



(a) DPF2 (chr11:64,853,146-64,881,804)



(b) CFL1 gene cluster (chr11:65,369,803-65,393,687)

Figure 12: Segmentation with H3K36m3 and H3K79me2

References

Bannister AJ, Schneider R, Myers FA, Thorne AW, Crane-Robinson C, Kouzarides T (2005) Spatial distribution of di- and tri-methyl lysine 36 of histone h3 at active genes. *Journal of Biological Chemistry* 280:17732 –17736.

Barnes EW (1908) A new development in the theory of the hypergeometric functions. *Proc. London Math. Soc.* 6:141–177.

Barski A, Cuddapah S, Cui K, Roh T, Schones D (2007a) High-resolution profiling of histone methylations in the human genome. *Cell* .

Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K (2007b) High-resolution profiling of histone methylations in the human genome. *Cell* 129:823–837.

3

Cui K, Zang C, Roh T, Schones D, Childs R, Peng W, 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.

Li B, Howe L, Anderson S, Yates JR, Workman JL (2003) The set2 histone methyltransferase functions through the phosphorylated carboxyl-terminal domain of RNA polymerase II. *Journal of Biological Chemistry* 278:8897–8903.

Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP, Lee W, Mendenhall E, O'Donovan A, Presser A, Russ C, Xie X, Meissner A, Wernig M, Jaenisch R, Nusbaum C, Lander ES, Bernstein BE (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* 448:553–560.

Qin Z, Yu J, Shen J, Maher C, Hu M, Kalyana-Sundaram S, Yu J, Chinnaiyan A (2010) HPeak: an HMM-based algorithm for defining read-enriched regions in ChIP-Seq data. *BMC Bioinformatics* 11:369.

Rabiner L (1989) A tutorial on hidden markov models and selected applications inspeech recognition. *Proceedings of the IEEE* .

Shimazaki H, Shinomoto S (2007) A method for selecting the bin size of a time histogram. *Neural Comput.* 19:1503–1527.

Waterman MS, Whiteman DE (1978) Estimation of probability densities by empirical density functions. *Int. J. Math. Educ. Sci. Tech.* 9:127–137.

Zang C, Schones D, Zeng C, Cui K, Zhao K, Peng W (2009) A clustering approach for identification of enriched domains from histone modification Chip-Seq data. *Bioinformatics* (25)15:1952–1958.