Supplementary Information

BroadPeak: a novel algorithm for identifying broad peaks in diffuse ChIP-seq datasets
Jianrong Wang, Victoria V. Lunyak and I. King Jordan

Supplementary methods:

Overview of the algorithm and input/output formats:

Broad peaks are contiguous genomic regions with distinct spatial densities of high-tag bins compared to the genomic background density. Whereas narrow ChIP-seq peaks are defined by high tag counts across adjacent genomic bins (i.e. high-tag sites), and are thus readily delineated, broad peaks are characterized by an overall increased spatial density of high-tag sites. In other words, broad peaks are distinguished from narrow peaks by the presence of 'gaps' at low-tag sites, and broad peaks should be able to span one or more of these gapped regions as long as the high-tag site spatial density of the peak remains elevated. Due to the fluctuations of ChIP-seq signals and the lack of characteristic size ranges, it's difficult to delineate the broad peak regions and their boundaries.

In order to quantitatively model and identify broad peaks, we adopted maximal-segment algorithm (Supplementary Figure 1). After assigning appropriate positive scores to high-tag bins and negative scores to low-tag bins, the cumulative score strings along chromosomes are random walks. If broad peaks exist, the random walks will be non-homogeneous and segments with large local cumulative scores will be observed (Supplementary Figure 2). Those segments, termed as *maximal scoring segments*, represent regions with high densities of high-tag bins. Thus, searching for all *maximal scoring segments* with significantly higher spatial densities is equivalent to identifying broad peaks. Special attentions are spent on parameter estimations which can critically affect the final compositions of high-tag bins within called broad peaks. The only two important parameters for BroadPeak are the target spatial density of high-tag bins within real broad peaks and the background spatial density of high-tag bins. BroadPeak provides two options for estimations for these parameters: supervised and unsupervised. The input file for BroadPeak is the ordered ChIP-seq tag count profiles along chromosomes in bedGraph format. The output file is the list of broad peak locations in BED format (Supplementary Figure 1).

Unsupervised parameter estimation:

While supervised estimation provides an easy and reliable way to accurately estimate the target spatial density p and the background density q, in practice users usually do not have a list of genomic regions that are known to be enriched with broad peaks from *a priori* knowledge and unsupervised estimation will be more commonly used. In order to do unsupervised estimation, BroadPeak first uses a sliding window approach to scan the genome and sample a list of genomic regions that contain *change-points* of spatial densities, *i.e.* the spatial densities change, at one unknown location within the region, from background densities to significantly high densities that are only observed in broad peaks. These regions can be used to simultaneously estimate the target density p and background density q. Due to the resolution problem of sliding window approaches and the noisy fluctuations of ChIP-seq data, we also need to accurately predict the position of the *change-point*, in order to accurately estimate p and q. This need leads us to adopt the *Gibbs sampling* method to iteratively estimate (Robert and Casella, 2004) the location of *change-points*, p and q.

We use a 10kb sliding window (Thurman et al., 2007), with each step is equal to the user-defined bin size, to scan the genome and calculate the high-tag bin densities for each sliding window. If the high-tag bin density is higher than twice of the genomic background density, the corresponding window is assigned as a putative region containing broad peaks or part of broad peaks. If we observe a sliding

window with background densities followed by a window with putative broad peak densities, then the whole region will be used later as a sample for parameter estimation.

Assuming we retain N such *change-point* containing regions using the sliding window scan as described above and each region contains L genomic bins, we will divide the L bins (for each sample region) into n super-bins and each super-bin is consisted of m consecutive bins. Finally, we obtain N data series with length n and they are denoted as: $D_i = (d_{i1}, d_{i2}, ...d_{in})$, where d_{ij} corresponds to the number of high-tag bins in the *j*th super-bin. Due to the way they are sampled, for each data series D_i, there exists a super-bin k such that $d_{ij} \sim Poisson(\lambda_1)$ for $j \le k$ and $d_{ij} \sim Poisson(\lambda_2)$ for $j > k$. So k is the unknown *change-point* and λ_1 is the rate for background spatial density of high-tag bins and λ_2 is the rate for target spatial density of high-tag bins ($\lambda_1 < \lambda_2$).

The whole data series is thus modeled as a *non-homogeneous Poisson process* with two distinct rates (Raftery and Akman, 1986). *Gibbs sampling* has been previously used for parameter estimations of *non-homogeneous Poisson processes* and here we applied this strategy (Robert and Casella, 2004). We assume λ_1 and λ_2 follow the conjugate prior distributions: $\lambda_1 \sim \lambda_1^{\alpha_1 - 1} e^{-\beta_1 \lambda_1}$ and $\lambda_2 \sim \lambda_2^{\alpha_2 - 1} e^{-\beta_2 \lambda_2}$. The prior distributions for the hyperparameters β_1 and β_2 are: $\beta_1 \sim \beta_1^{\sigma_1 - 1} e^{-\varepsilon_1 \beta_1}$ and $\beta_2 \sim \beta_2^{\sigma_2 - 1} e^{-\varepsilon_2 \beta_2}$. And a series of conditional probabilities are as follows:

$$P(\lambda_1 \mid D_i, k, \alpha_1, \beta_1) \sim \lambda_1^{\alpha_1 + \sum_{j=1}^{k} d_{ij} - 1} e^{-(\beta_1 + k)\lambda_1}$$

$$P(\lambda_2 \mid D_i, k, \alpha_2, \beta_2) \sim \lambda_2^{\alpha_2 + \sum_{j=k+1}^{n} d_{ij} - 1} e^{-(\beta_2 + n - k)\lambda_2}$$

$$P(\beta_1 \mid D_i, k, \lambda_1, \alpha_1, \sigma_1, \varepsilon_1) \sim \beta_1^{\alpha_1 + \sigma_1 - 1} e^{-(\lambda_1 + \varepsilon_1)\beta_1}$$

$$P(\beta_2 \mid D_i, k, \lambda_2, \alpha_2, \sigma_2, \varepsilon_2) \sim \beta_2^{\alpha_2 + \sigma_2 - 1} e^{-(\lambda_2 + \varepsilon_2)\beta_2}$$

$$P(k \mid D_i, \lambda_1, \lambda_2) \sim \frac{P(D_i \mid k, \lambda_1, \lambda_2)}{\sum_{c=1}^{n} P(D_i \mid c, \lambda_1, \lambda_2)}.$$

Before *Gibbs sampling*, λ_1 is initialized as $\hat{\lambda}_1 = \frac{\sum_{j=1}^{\tau} d_{ij}}{\tau}$ and λ_2 is initialized as $\hat{\lambda}_2 = \frac{\sum_{j=n-\tau+1}^{n} d_{ij}}{\tau}$ because the *change-point* is not likely to occur in the first and last a few super-bins. Because the prior distributions for λ_1 and λ_2 are gamma, and we consider $\hat{\lambda}_1$ and $\hat{\lambda}_2$ are good estimates of the means of the gamma distributions, then β_1 is initialized as $\hat{\beta}_1 = \hat{\lambda}_1 / \mathrm{var}_{\tau}$ and β_2 is initialized as $\hat{\beta}_2 = \hat{\lambda}_2 / \mathrm{var}_{n-\tau}$, where var_{τ} is the variance of the first few super-bins and $\mathrm{var}_{n-\tau}$ is the variance of the last few super-bins. α_1 is estimated as $\hat{\alpha}_1 = \hat{\lambda}_1 \times \hat{\beta}_1$, and α_2 is estimated as $\hat{\alpha}_2 = \hat{\lambda}_2 \times \hat{\beta}_2$. ε_1 and ε_2 are set as 0.5 and σ_1 is estimated as $\hat{\sigma}_1 = \hat{\beta}_1 \times 2$, and σ_2 is estimated as $\hat{\sigma}_2 = \hat{\beta}_2 \times 2$.

After initializations, we use *Gibbs sampling* on those conditional probabilities to iteratively estimate k, λ_1 and λ_2. Finally, the target spatial density of high-tag bins is $p = \lambda_1 / m$ and $q = \lambda_2 / m$. The estimated densities are then used to calculate the log likelihood ratios as the scores for maximal scoring segment identifications.

Comparison of broad peak calling results between supervised estimation mode and unsupervised estimation mode:

We repeated the peak calling in supervised mode for the histone modification H3K36me3 since its distribution is known to be highly coincident with actively transcribed gene bodies. We used several genes that are known to be highly transcribed in $CD4^+$ T cells to estimate the target density p. The locations and identities of the H3K36me3 peaks called in the supervised versus unsupervised modes are highly similar (Supplementary Figure 6).

Dataset simulation:

In order to directly compare the performances of BroadPeak with MACS, SICER and RSEG, we simulated 3 ChIP-seq tag libraries for tests. To do this, we first selected 5,000 non-overlapping human genes with different sizes as the real broad peaks. Then, the human genome is divided into 200bp bins. For non-broad-peak regions, the background spatial density of high-tag bins is set as $1x10^{-4}$. For real broad peak regions, the density is set as 20, 50 and 100 fold of the background density respectively for different libraries. The tag count distribution of high-tag bins is simulated as a Gaussian distribution with mean of 8 and standard deviation of 2. The spatial density of low-tag bins (noise) is the same throughout the whole genome and is set as 0.5, namely about half of the genome has noise. The tag count distribution of low-tag bins is simulated as a Poisson distribution with the average rate as 0.7, which is similar to the H3K36me3 library.

Performance evaluation:

Similar to the comparison procedure of RSEG (Song and Smith, 2011), we ran BroadPeak, MACS, SICER and RSEG on the three simulated ChIP-seq tag libraries and compared the identified broad peaks with the pre-defined (*i.e.* known) peaks. A known broad peak is considered as correctly identified if a certain fraction of it is covered by predicted peaks. Similarly, the predicted broad peak is considered as true if a certain fraction of it is covered by real peaks. The three thresholds of fractions are 20%, 50% and 80%. Based on these basic counts, recall and precision are used to measure the performance and the *F* score is used as the final measurement of the overall performance of the algorithms. The basic observation is that large known broad peaks are usually predicted as many smaller peaks by SICER and RSEG. BroadPeak works better to identify the large broad peak as a single unit. But sometimes BroadPeak merges closely spaced broad peaks into a single peak. The base line is that BroadPeak has the highest *F* scores for all the datasets under all thresholds and almost all the best recalls. The precision of BroadPeak is slightly lower than seen for SICER but the improvement on recall is substantial.

Some practical issues using BroadPeak:

Incorporating control samples for broad peak calling: If the users believe that their ChIP-seq tag distribution is not uniform due to some unspecific factors such as regional GC contents and mappabilities, control samples are useful to remove these effects. BroadPeak does not directly deal with information from control samples. But the users can incorporate control sample information by doing an additional pre-processing step before using BroadPeak. For each genomic bin, the tag counts of the control sample should be subtracted from the tag counts of the real ChIP-seq sample, and then only use the corrected tag counts as the tag profile to identify broad peaks. For regions without real signals, small negative tag counts might be produced and they can be set as zero, indicating no biological meaningful events in those bins. This simple pre-processing step is based on the assumptions that the real ChIP-seq tag

library and control library are independent samples of two different Poisson random variables. For the real ChIP-seq tag library, the tag count of each bin follows a Poisson distribution: $T_r \sim Poisson(\lambda_r + \lambda_n)$, where T_r is the tag count of a bin in real ChIP-seq library, λ_r is the mean unrelated to unspecific factors and λ_n is the mean caused by unspecific factors such as regional GC contents. For the control ChIP-seq tag library, the tag count of the corresponding bin follows another Poisson distribution: $T_n \sim Poisson(\lambda_n)$, where T_n is the tag count of the bin in the control ChIP-seq library and λ_n is the mean caused by unspecific factors of that bin. Since the two samples are independent, $T_r - T_n \sim Poisson(\lambda_r)$. Thus the reasoning above suggests that the corrected tag counts of each bin (by subtracting the control sample tag counts of the corresponding bin) will better represent the real biological signals.

Applicability of BroadPeak for identifications of narrow histone modification peaks: Some histone modifications, such as H3K4me3, have narrow peaks. But the sizes of their peaks are still larger than transcription factor binding peaks. We have tested the performance of BroadPeak to identify narrow histone modification peaks and the resulted peaks are largely consistent with the underlying ChIP-seq tag profiles.

Influence of different bin sizes: Because the sizes of broad peaks are usually much larger than the size of nucleosomes, different bin sizes (such as 100bp versus 200bp) do not influence the final identification much. The default value of bin size for BroadPeak is set as 200bp because it is the approximate nucleosome size. Users can use '-b' to change the bin size according to their datasets.

Pre-processing of ChIP-seq tags: For narrow transcription factor binding peaks, some pre-processing steps are usually performed, such as extending and shifting ChIP-seq tags. Because the sizes of broad peaks are much larger than the tag fragments, small scale tuning of aligned tags does not change the final results. The initial binary classification of high-tag and low-tag bins used in BroadPeak can help to suppress the influence of extremely high tag counts that are caused by amplification artifacts. But a filtering of those artifacts is always preferred before applying BroadPeak. The required pre-processing step for BroadPeak is to organize the ChIP-seq tags into a BedGraph format file with tag counts of sorted equal-size genomic bins.

References of supplementary methods:

Thurman, R.E., Day, N., Noble, W.S. and Stamatoyannopoulos J.A. (2007) Identification of higher-order functional domains in the human ENCODE regions, Genome Research, **17**, 917-927

Raftery, A.E. and Akman, V.E. (1986) Bayesian analysis of a Poisson process with a change-point, *Biometrika*, **73**, 85-89.

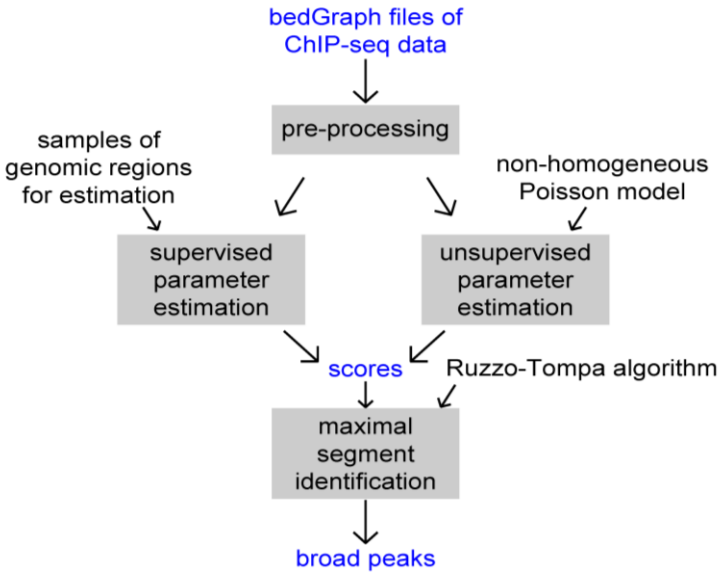
Robert, C.P. and Casella, G. (2004) *Monte Carlo Statistical Methods*. Springer. pp. 454-455.

Song, Q. and Smith, A.D. (2011) Identifying dispersed epigenomic domains from ChIP-Seq data, *Bioinformatics*, **27**, 870-871.

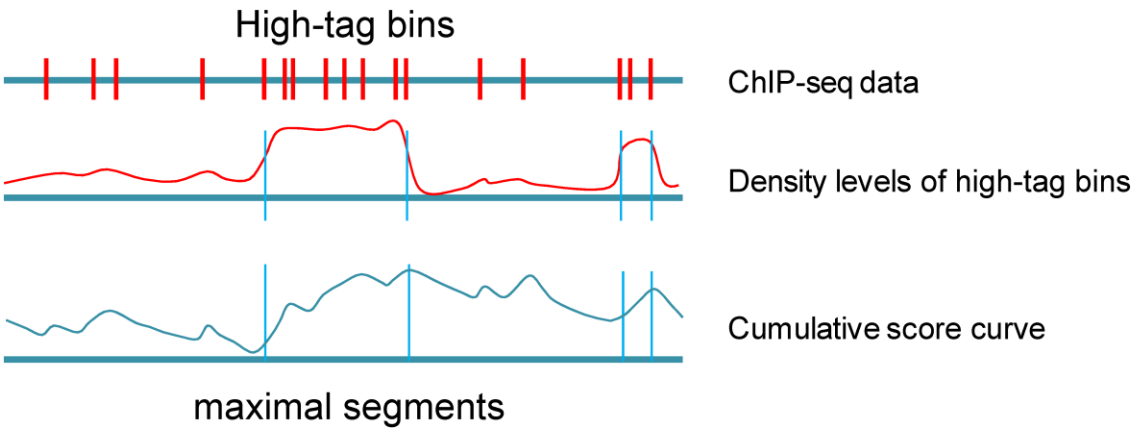
Supplementary Table 1: The summary of recall, precision and *F* score for BroadPeak, MACS, SICER and RSEG on simulated ChIP-seq tag libraries with different overlapping criteria.

Feature	Software	overlap criteria = 20%			overlap criteria = 50%			overlap criteria = 80%		
		dataset 1	dataset 2	dataset 3	dataset 1	dataset 2	dataset 3	dataset 1	dataset 2	dataset 3
Recall	BroadPeak	0.44	0.53	0.55	0.39	0.46	0.51	0.38	0.43	0.47
	MACS	0.07	0.07	0.07	0.02	0.03	0.03	0.01	0.01	0.01
	SICER	0.07	0.07	0.07	0.04	0.04	0.04	0.03	0.03	0.03
	RSEG	0.60	0.38	0.19	0.31	0.21	0.12	0.24	0.18	0.11
Precision	BroadPeak	0.69	0.74	0.73	0.60	0.66	0.64	0.55	0.60	0.58
	MACS	0.86	0.90	0.92	0.81	0.86	0.90	0.70	0.78	0.84
	SICER	0.78	0.84	0.89	0.66	0.74	0.82	0.52	0.63	0.74
	RSEG	0.29	0.65	0.89	0.24	0.55	0.82	0.18	0.44	0.71
F score	BroadPeak	0.54	0.62	0.63	0.47	0.54	0.57	0.45	0.50	0.52
	MACS	0.12	0.13	0.13	0.05	0.05	0.05	0.03	0.03	0.02
	SICER	0.12	0.12	0.13	0.08	0.07	0.08	0.06	0.06	0.06
	RSEG	0.39	0.48	0.31	0.27	0.30	0.21	0.21	0.26	0.19

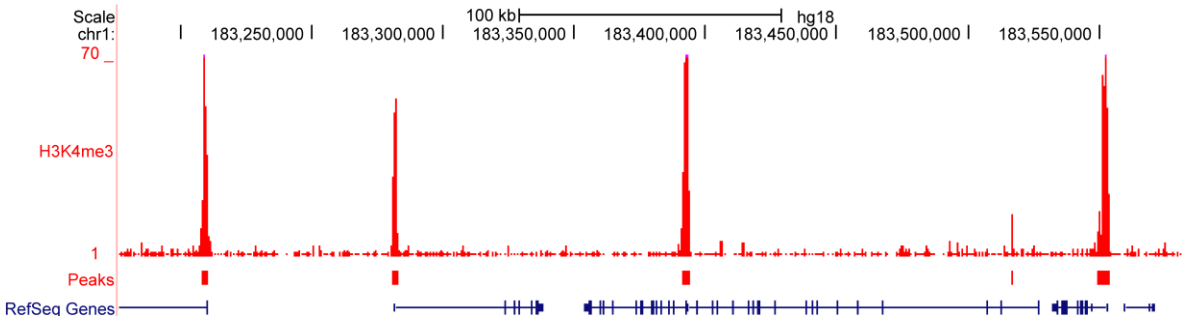
Supplementary Figure 1: Algorithm scheme of BroadPeak.



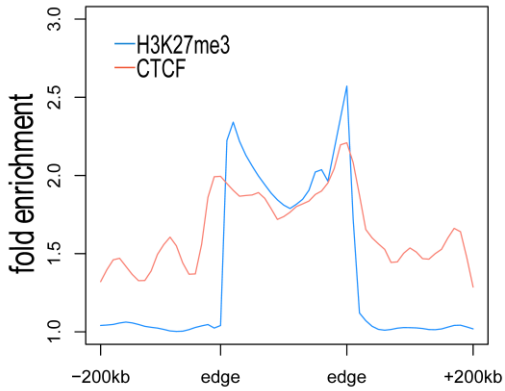
Supplementary Figure 2: Schematic illustration of the relationship between the spatial density of high-tag bins and maximal segments.



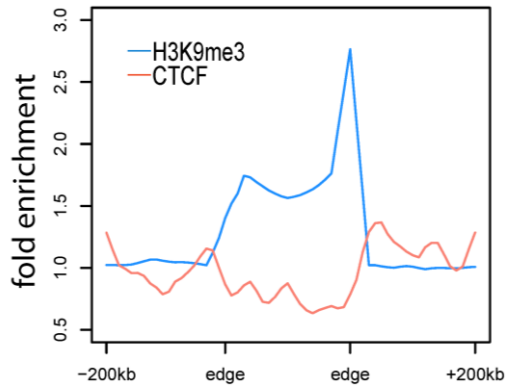
Supplementary Figure 3: Examples of H3K4me3 peaks identified by BroadPeak.



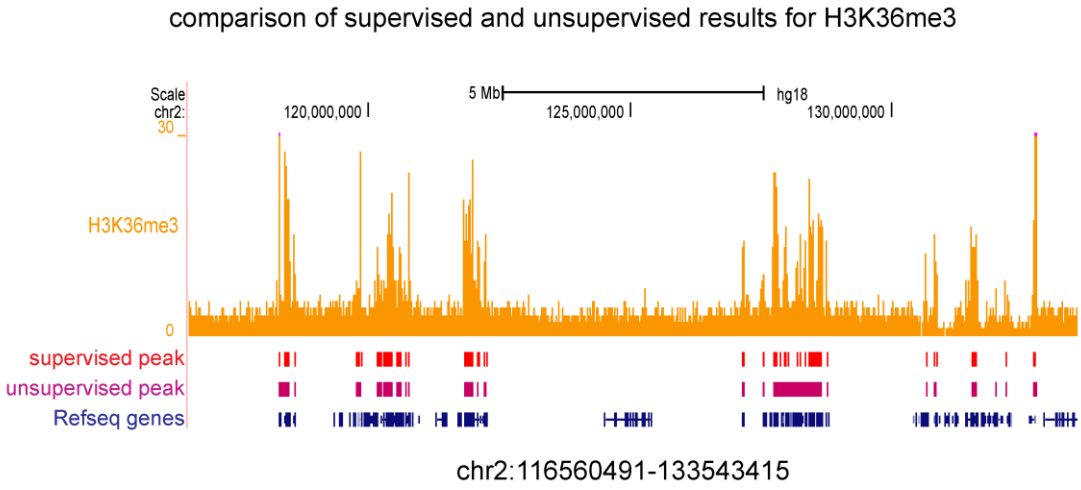
Supplementary Figure 4: Enrichment of CTCF binding (red) around broad peak edges of H3K27me3 (blue) in K562 cells from ENCODE dataset.



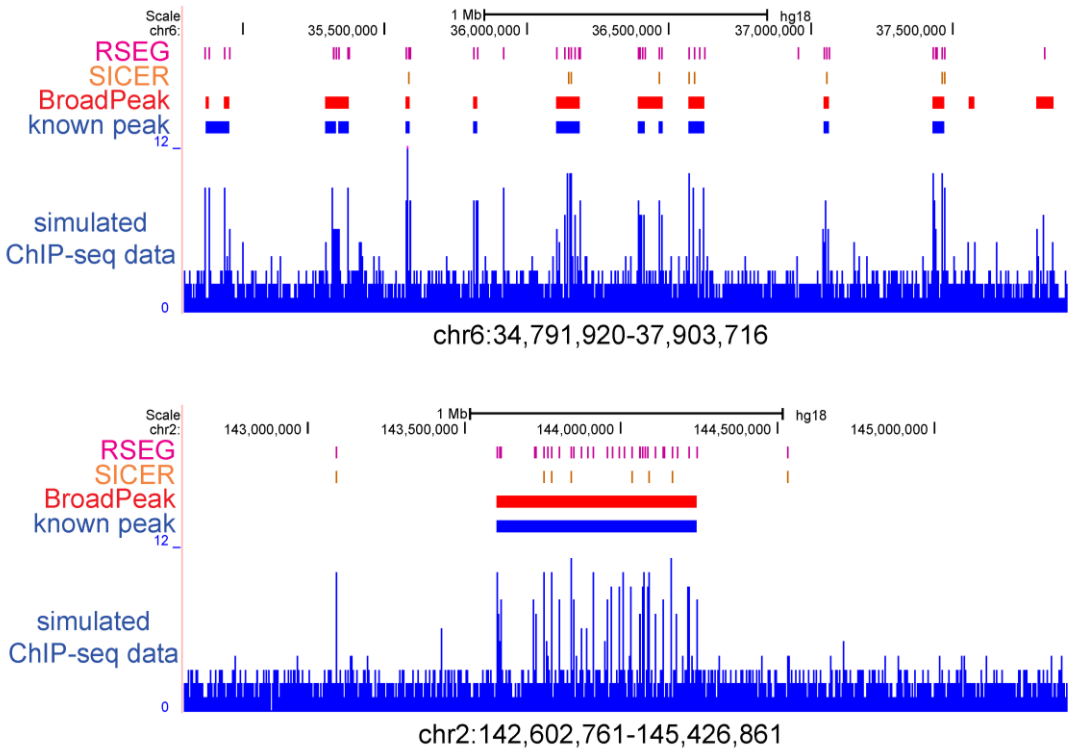
Supplementary Figure 5: Enrichment of CTCF binding (red) around broad peak edges of H3K9me3 peaks (blue) identified by BroadPeak.



Supplementary Figure 6: Examples of H3K36me3 broad peaks identified by supervised and unsupervised parameter estimations.



Supplementary Figure 7: Examples of broad peaks called by RSEG, SICER and BroadPeak on one simulated library.



Supplementary Figure 8: The size distributions of broad peaks identified by SICER, RSEG and BroadPeak.