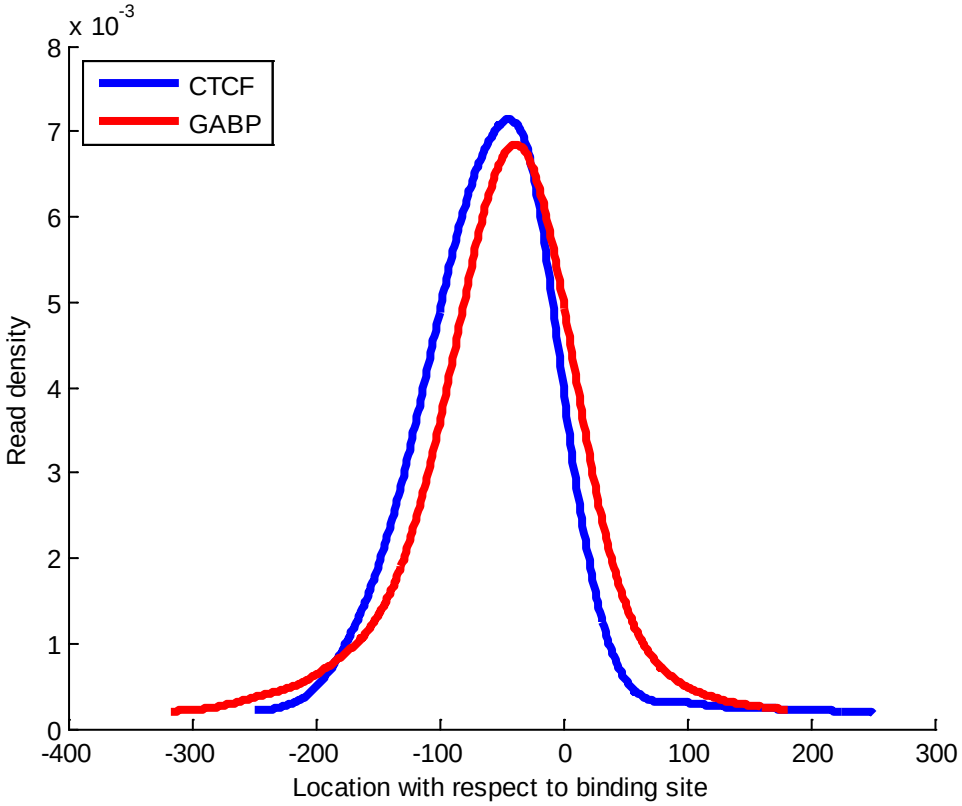
Discovering homotypic binding events at high spatial resolution

Yuchun Guo, Georgios Papachristoudis, Robert C Altshuler, Georg K Gerber, Tommi S Jaakkola, David K Gifford & Shaun Mahony

Supplementary figures and text:

Supplementary Figure 1	GPS refines read distribution
Supplementary Figure 2	GPS has high spatial resolution
Supplementary Figure 3	Motif occurrence in GPS events
Supplementary Figure 4	GPS has high sensitivity
Supplementary Figure 5	An example of GABP joint events
Supplementary Figure 6	GPS in alignment mode successfully aligns events in regions where there is only a unique strong motif presence
Supplementary Figure 7	Replicate consistency in alignment mode
Supplementary Figure 8	GPS derived position-specific prior for motif discovery
Supplementary Table 1	The number of CTCF event calls and overlaps from different methods
Supplementary Table 2	The GABP dataset contains more joint binding events
Supplementary Table 3	GPS is more sensitive in discovering joint events in GABP ChIP-seq data at the genomic region 19:60,000,000-62,000,000
Supplementary Note 1	Supplementary information and methods
Supplementary Note 2	Supplementary method: Event Alignment

Supplementary Figure 1

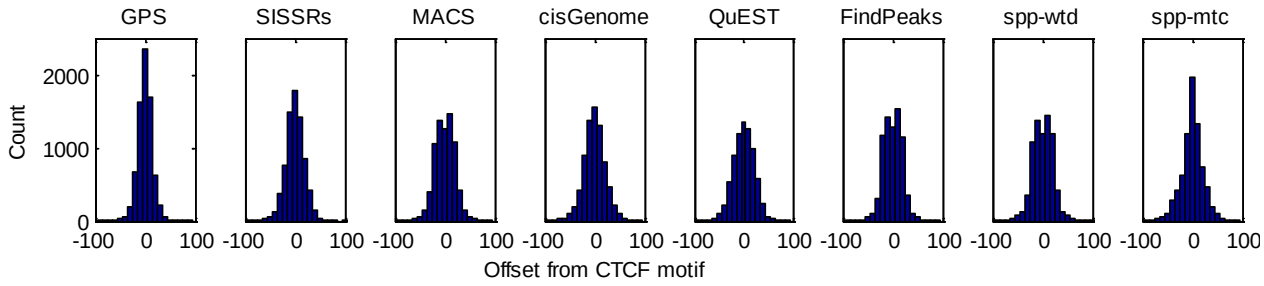


GPS refines the read distribution

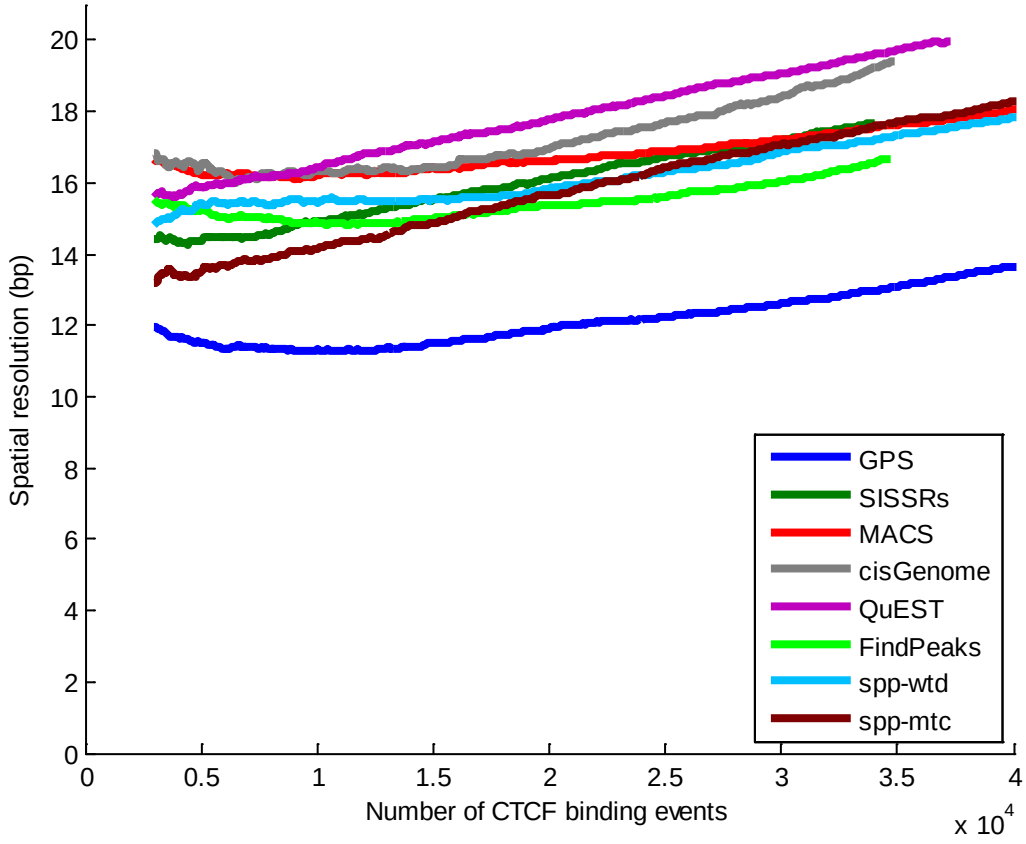
A generic read distribution (from CTCF data) was initially used to predict GABP binding events. GPS then used the predicted positions to iteratively re-estimate the read distribution specific for GABP.

Supplementary Figure 2

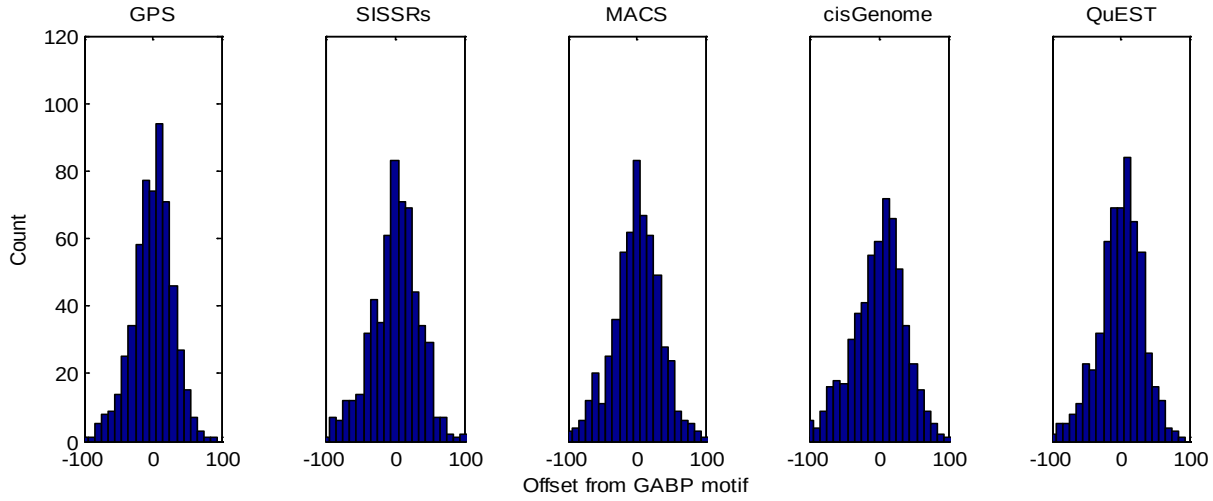
a



b



c



GPS has high spatial resolution

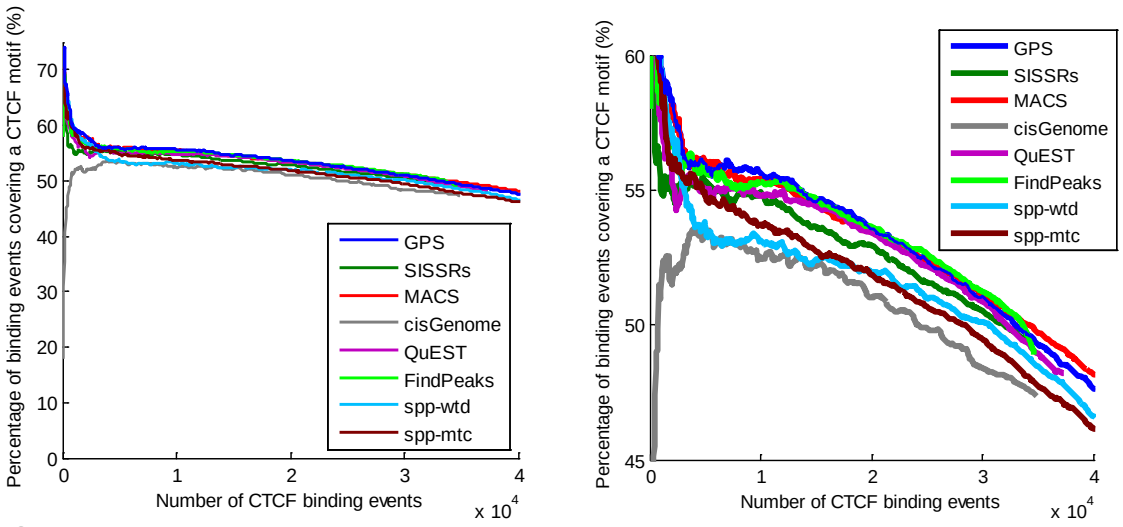
a. The distribution of offset of event predictions from the CTCF binding motif in a set of 7,653 binding events that are called by all methods shown.

b. The spatial resolution of CTCF event calls is shown averaged over increasing numbers of the strongest ranked events identified by different methods.

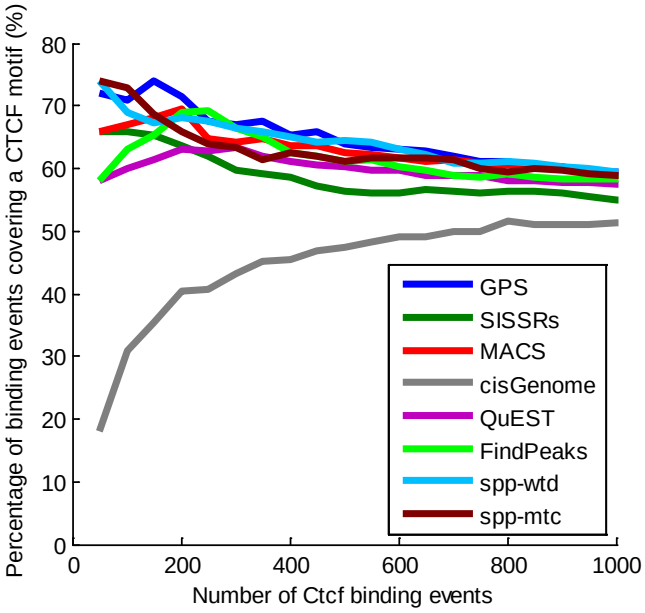
c. The distribution of offset of event predictions from the GABP binding motif by GPS and other 4 methods using the GABP binding dataset. GPS has an average spatial resolution of 22.34±18.02bp, compared to 26.22±21.13bp for SISSRs, 26.43±21.34bp for MACS, 29.65±22.31bp for cisGenome, and 24.86±19.77bp for QuEST.

Supplementary Figure 3

a



b

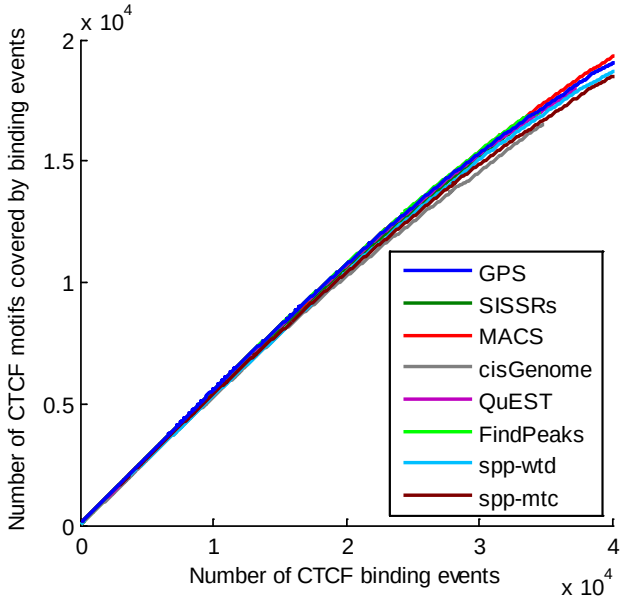


Motif occurrence in GPS events

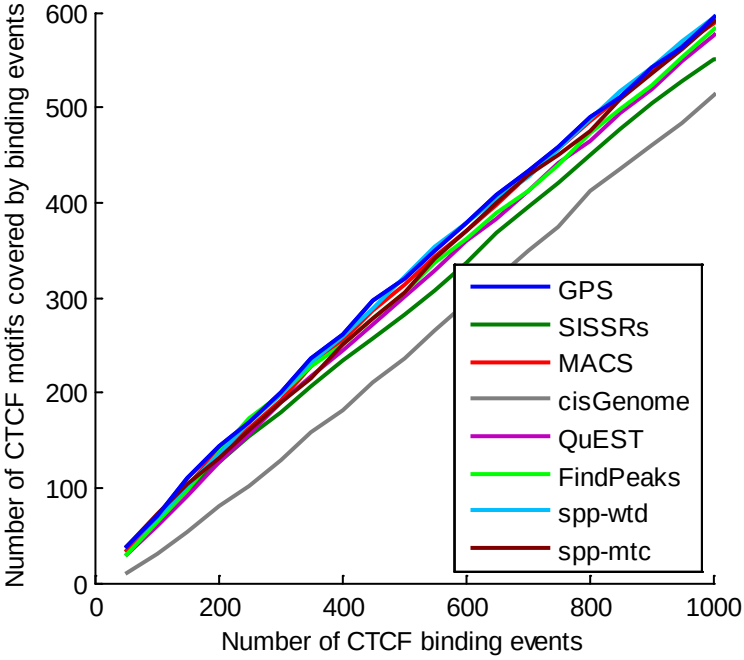
a. GPS, MACS and FindPeaks predicted more events that have a CTCF motif occurrence than the same number of ranked events predicted by other methods. b. GPS achieved a high motif occurrence for the highest-ranking predictions. The percentage of identified events that have a CTCF motif within 100bp is shown averaged over increasing numbers of the strongest ranked events identified by different methods. cisGenome's high ranking events have a low motif occurrence because cisGenome reported a large number of duplicate read artifacts as high scoring events.

Supplementary Figure 4

a



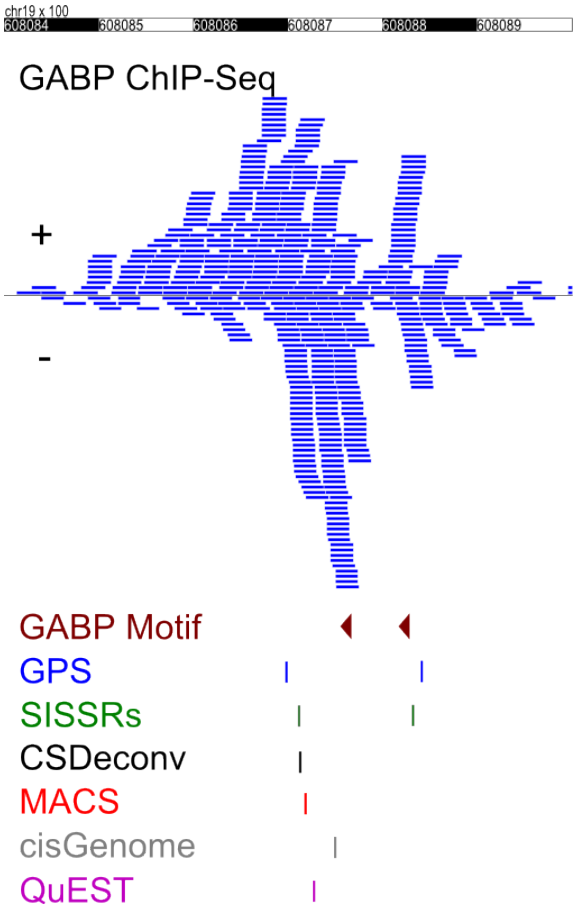
b



GPS has high sensitivity

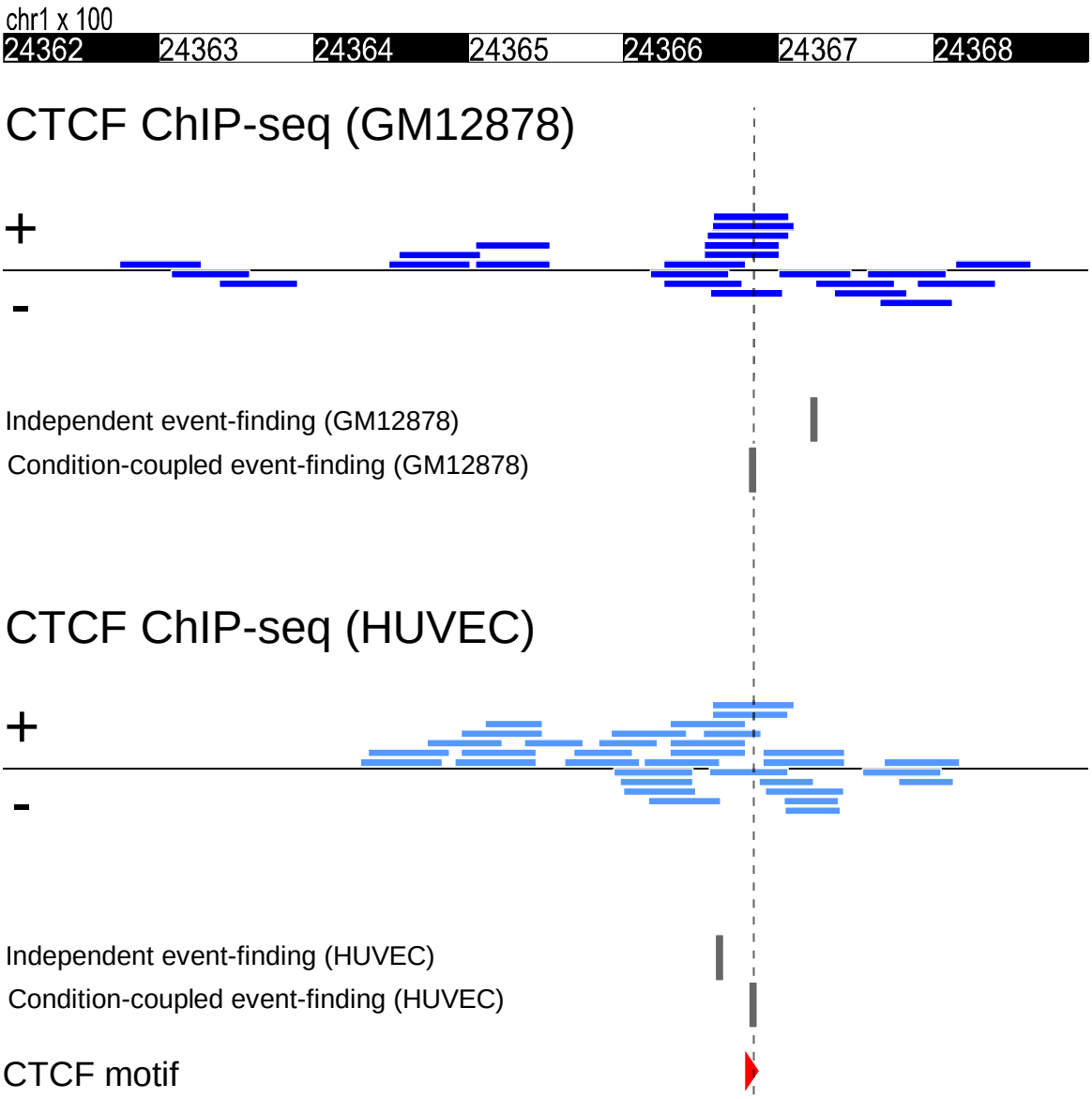
a. The number of CTCF motifs that are covered (within 100bp) by increasing numbers of the strongest ranked events identified by GPS, SISSRs, MACS, cisGenome, QuEST, spp and FindPeaks. A method with higher sensitivity will cover more CTCF motifs. b. More CTCF motifs were covered by the highest- ranking GPS events than were covered by other methods.

Supplementary Figure 5



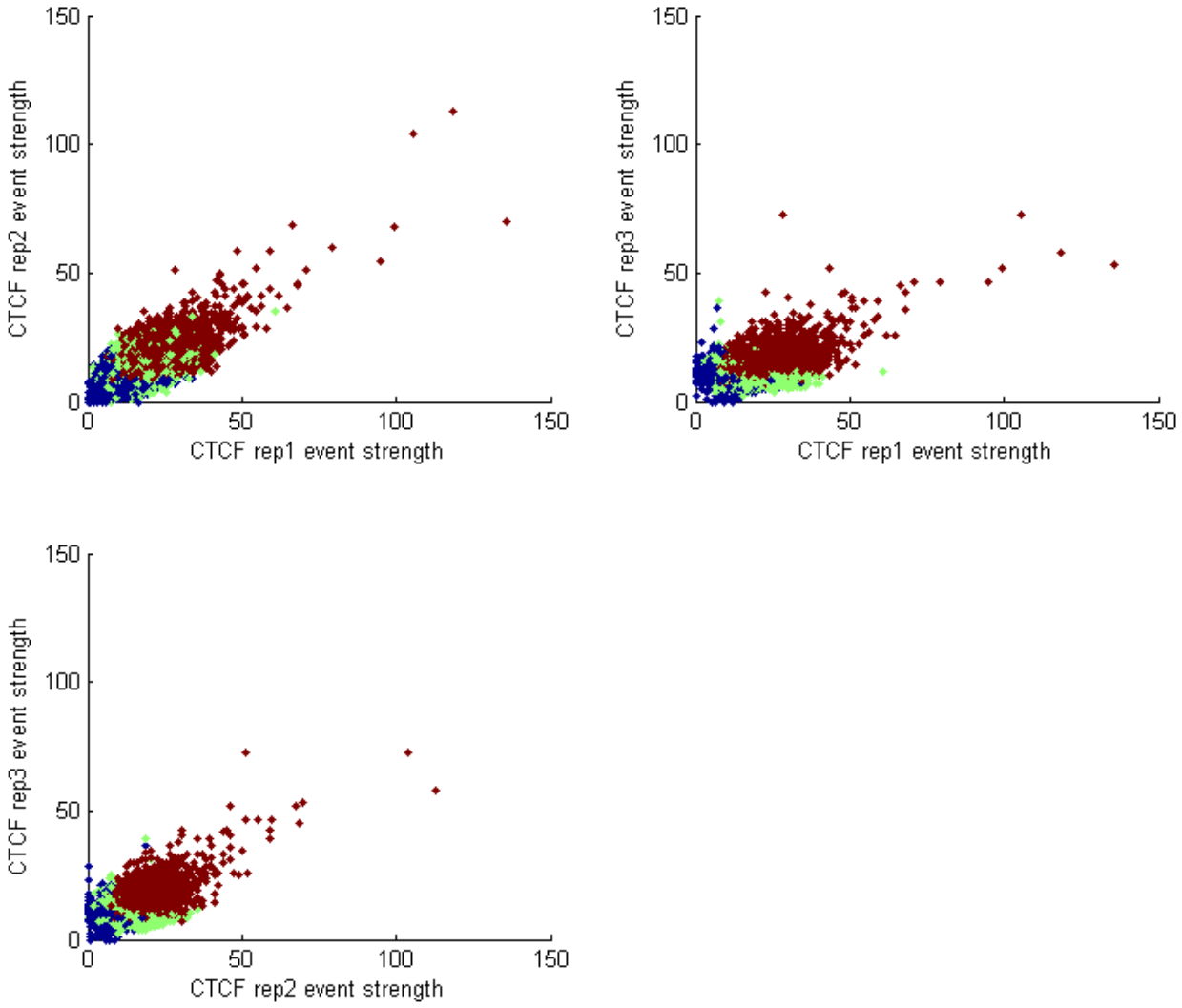
An example of GABP joint events

Supplementary Figure 6



GPS in alignment mode successfully aligns events in regions where there is only a unique strong motif presence.

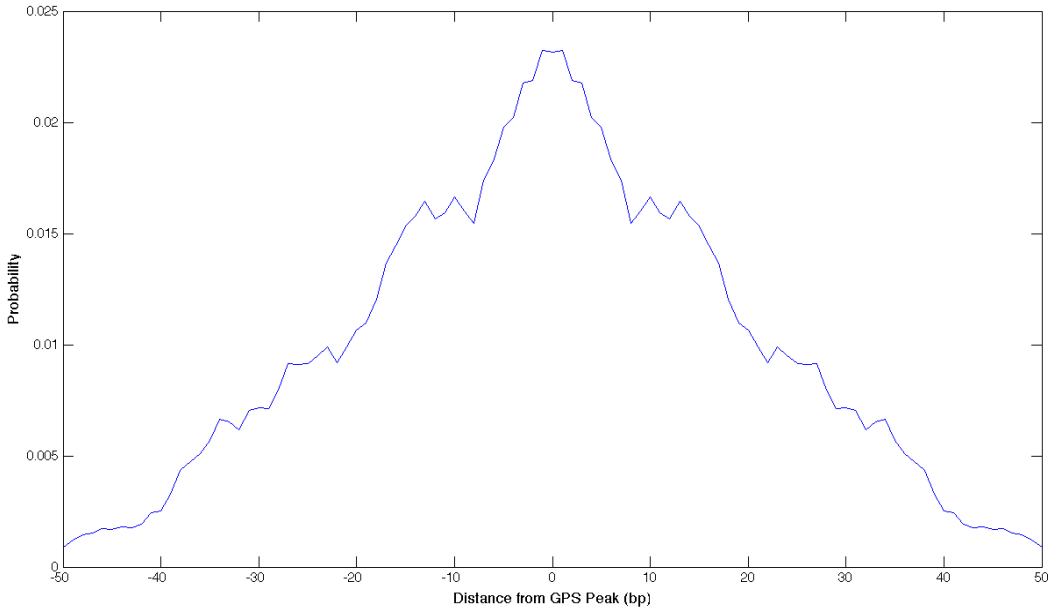
Supplementary Figure 7



Replicate consistency in alignment mode

Three technical replicates of Mouse ES CTCF ChIP-Seq data were analyzed simultaneously in GPS alignment mode. The scatter plots show the pair-wise comparison of event strength (IP read counts) of each aligned event. The data points are colored by the number of replicates that the event is significant (brown: in all 3 replicates, green: in 2 replicates, blue: in only 1 replicate).

Supplementary Figure 8



GPS derived position-specific prior for motif discovery

Supplementary Table 1:

Overlaps(%)	GPS	SISSRS	MACS	cisGenome	QuEST	FindPeaks	spp_wtd	spp_mtc
GPS	100	96	78	91	88	93	81	80
SISSRS	77	100	63	81	77	83	71	71
MACS	97	98	100	92	95	97	87	85
cisGenome	77	86	64	100	77	84	75	73
QuEST	81	88	70	82	100	86	77	77
FindPeaks	79	88	67	84	80	100	74	73
spp_wtd	80	88	69	87	84	86	100	90
spp_mtc	80	88	69	86	84	86	91	100
Total	41023	34019	50465	34811	37300	34720	40478	40940

The number of CTCF event calls and overlaps from different methods

The pair-wise comparison of events called by different methods is presented as the percentage of events that are discovered by one method (column) that are within 200bp of events called by another method (row).

Supplementary Table 2:

Dataset	Number of events to search for joint binding calls	Number of clustered motif sites	Number of joint events
CTCF	34019	174	20
GABP	6442	581	122

The GABP dataset contains more joint binding events

Supplementary Table 3:

Region	Motifs	GPS	SISSRs	MACS	cisGenome	QuEST	CSDeconv
19:60209122-60209517	3	2	1	1	1	1	3
19:60355126-60355571	10	2	1	1	1	1	2
19:60382365-60382636	2	0	0	0	1	1	0
19:60419681-60420363	2	0	0	1	1	0	0
19:60803307-60803837	2	1	0	0	0	1	0
19:60808666-60808927	2	2	2	1	1	1	1
19:60858417-60858832	3	1	1	0	1	1	0
19:61323575-61324038	4	0	0	0	1	1	0
19:61324381-61324814	3	2	3	1	1	2	0
Total of events		**10**	**8**	**5**	**8**	**9**	**6**
Total of joint events		**4**	**2**	**0**	**0**	**1**	**2**

GPS is more sensitive in discovering joint events in GABP ChIP-seq data at the genomic region 19:60,000,000-62,000,000

The table shows the number of event calls in the regions with clustered motifs. Four regions that are identified by at least two methods as containing joint events are shaded in green. GPS calls joint events in all 4 of them, SISSRs and CSDeconv calls 2 regions, QuEST calls 1 region, MACS and cisGenome does not call joint events in these regions. The analysis is based on the top 36 events within region 19:60,000,000-62,000,000 from each method (with exception that CSDeconv has only 23 events).

Supplementary Note 1:

Supplementary information and methods

Setting the sparseness parameter α

The value of the sparseness parameter *α* will influence the sensitivity and specificity of event detection and it should scale with the read count of the events in the region that GPS is analyzing. From our experience in analyzing mouse CTCF and human GABP datasets, the *α* value is set empirically to

$$\alpha = \max(\sqrt{RC_{\max}} / alpha\_factor, alpha_{\min})$$

where RC_{max} is the maximum read count in a 500bp sliding window across the region that GPS is evaluating, $alpha_{min}$ is the minimum number of read count for a valid binding event. In the CTCF data analysis, we choose the default setting *alpha_factor* = 3.0, $alpha_{min}$=6.0. These values can also be set at the command line by the user using options ("--a" for $alpha_{min}$ and "--af" for *alpha_factor*) .

PCR amplification artifact filtering

PCR amplification artifacts typically manifest as the observation of many reads mapping to the exact same base positions. Some of the published analysis techniques deal with these artifacts by excluding potential artifact regions, or ignoring all but a small number (often 1) of the reads that map to a given base. From our experience, PCR amplification artifacts are quite variable and dataset-specific. Therefore, a generic approach to exclude those regions might result in the loss of true events.

GPS implements an artifact filtering method by comparing the read distribution of the predicted event to the expected event read distribution. A shape deviation score is computed using Kullback–Leibler divergence (see method section 2.6). A higher score means the event is more divergent from the expected read distribution, hence more likely to be artifact or noise. A cutoff score can be specified by user to filter out spurious events. GPS also excludes events with less than 3 fold enrichment (IP/Control). GPS reports the filtered events, hence allows the user to verify and adjust cutoff threshold for a particular dataset. The shape deviation filter is on by default, but can be turned off using option (--nf).

In addition, GPS also has an option (--bf) for the user to set a cutoff value for the maximum read count per base position. The cutoff value can be estimated automatically using a Poisson model, or can be set manually by the user (--mrc).

Normalization for multiple condition alignment mode

For multiple conditions the reads need to be properly normalized for comparison. We have implemented a linear regression method to calculate the ratio between two

conditions as in (Rozowsky, et al., 2009). The scaling factors are weighted so that the total read count of the whole dataset remains the same.

For C conditions, each with read count as N_i, i=1,…,C. We use the first data set as reference and calculate the linear regression ratio λ_i of each other dataset i vs first dataset, i=2,…,C. And we have λ_1=1. The scaling factor for each dataset will be

$$f_i = \frac{\lambda_i \sum_{i'=1}^{C} N_{i'}}{\sum_{i'=1}^{C} \lambda_{i'} N_{i'}}$$

We normalize the IP and Control data separately.

Datasets used:

Dataset 1: CTCF binding

To evaluate the performance of GPS, we analyzed a ChIP-Seq dataset of insulator binding factor CTCF (CCCTC-binding factor) in mouse ES cells, with a control using antibody against GFP (Green Fluorescence Protein) to control for non-specific binding (Chen, et al., 2008). We chose CTCF data for our evaluation because the strong CTCF consensus motif allows us to reliably measure spatial resolution. The ChIP-Seq data comprised 4.2 million CTCF reads and 7.9 million GFP reads that uniquely map to the mm8 mouse genome.

Dataset 2: GABP binding

To evaluate the performance of joint event discovery, we analyzed a ChIP-Seq dataset of GABP in human Jurkat cells, with a control using input DNA (Valouev, et al., 2008). GABP binding dataset was reported previously to contain multiple binding motifs in a short region (Lun, et al., 2009; Valouev, et al., 2008). The ChIP-Seq data was downloaded from QuEST website (http://mendel.stanford.edu/SidowLab/downloads/quest/). It comprised 7.9 million GABP reads and 17.4 million input DNA reads that uniquely map to the hg18 human genome.

Dataset 3: CTCF binding in multiple conditions

To evaluate the performance of GPS in alignment mode, we analyzed a ChIP-Seq dataset of insulator binding factor CTCF (CCCTC-binding factor) in human GM12878 (lymphoblastoid cell line produced from the blood of a female donor with northern and western European ancestry by EBV transformation) and HUVEC (Human Umbilical Vein Endothelial Cells) cells. We performed our analysis only in chromosome 1. We used also a control for each of the two conditions to control for non-specific binding. The ChIP-Seq data for GM12878 condition comprised of 1.6 million IP reads, 1.1 million control reads, and for HUVEC condition 1.5 million IP reads and 2 million control reads for chromosome 1. The ChIP-Seq experimental work was done by the laboratory of

Brad Bernstein (Broad Institute) as part of the ENCODE project (funded by the NHGRI) (Birney, et al., 2007).

ChIP-Seq analysis methods:

We compared the performance of GPS against eight published ChIP-Seq analysis methods: MACS (version 1.3.7.1)(Zhang, et al., 2008), SISSRs (version 1.4)(Jothi, et al., 2008), cisGenome (version 1.2)(Ji, et al., 2008), QuEST (version 2.4)(Valouev, et al., 2008), FindPeaks (version 4) (Fejes, et al., 2008), spp-wtd and spp-mtc (version 1.8) (Kharchenko, et al., 2008), and CSDeconv (version 1.0)(Lun, et al., 2009). All the methods were run using default parameters except what is described in the following.

For MACS, we used the summit location as the predicted binding site position. The binding events are sorted by p-value.

For SISSRs, "-t" option was used to obtain binding site predictions.

For cisGenome, we analyzed the data with the option of boundary refinement, and used the center of the predicted region as the binding site position. In our tests, these options gave the best result in spatial resolution.

For FindPeaks, we run with options "-dist_type 1 -duplicatefilter " to filter artifact reads. We used the max_coord position as the predicted binding site. The binding events are sorted by height.

For spp-wtd and spp-mtc, the binding events are sorted by FDR and then by score.

Motifs used in this study:

CTCF motif

GPS was used to call binding events in mouse ES cell CTCF ChIP-seq data (Chen, et al., 2008), and the events were ranked by peak strength. We extracted 200bp sequences around the top-most 500 peaks, and ran SOMBRERO (Mahony, et al., 2005) with default parameters. The most significant motif discovered by SOMBRERO, which is very similar to the CTCF motif reported in Chen, et al. 2008, was used for subsequent CTCF motif analysis. The matrix of the discovered motif is included as supplementary information.

(this study)

(Chen, 2008)

To find an appropriate scoring threshold for this motif, we simulated 500,000 200bp sequences using a 3rd-order Markov model of the mm8 genome. A scoring threshold which yields 'false positive' motif matches in 1% of these simulated sequences was recorded. In practice this threshold was a log-likelihood score of 11.52, given a background model of GC-content equal to 41.74%.

GABP motif

GABP motif was retrieved from TRANSFAC database (M00341) (Matys, et al., 2003). A motif score threshold of 9.9 (half of the maximum motif score) is used in the joint event analysis.

Method comparison on spatial resolution, specificity and sensitivity:

We evaluated the effective spatial resolution of GPS against other methods. We define effective spatial resolution as the absolute value of the distance between genome coordinates of predicted CTCF binding events and the middle of corresponding high-scoring CTCF binding motif hit. The sign of the offset was adjusted according to the strand on which the motif hit occurred. Because the center of the motif hit may not represent a true center of binding event, the offsets to the motif were centered by subtracting the mean offsets (Kharchenko, et al., 2008). Because different methods predict different sets of binding events, we compare spatial resolution on the "matched" set of predictions that correspond to the same high-scoring CTCF binding motif. Only those events within 100bp of a CTCF motif match are included in the calculation.

For the CTCF dataset, the numbers of events called by each method and the pair-wise overlap percentages are shown in Supplementary table 1. We evaluate spatial resolution from the same number of top ranking events by each method. From the top 34019 predictions of each method, we select the 7,653 events that were predicted by all eight methods. The results are given in Figure 3a. An alternative representation as a distance histogram is given in Supplementary Fig. 3a.

We also evaluate spatial resolution with the increasing number of top ranking events identified by each method (Supplementary Fig. 3b). Note that this analysis does not have a "matched" set of predictions. We simply average the spatial resolution of the top ranking events that have a motif at a distance less than 100bp.

For specificity of the predicted events, we calculate the motif occurrence within 100bp of the predicted event location for increasing number of the strongest ranked events identified by different methods as in Zhang et. al (Zhang, et al., 2008). The percentage of motif occurrence is calculated as number of predicted events covering a CTCF motif within 100bp divided by the number of events. The result is given in Supplementary Fig. 3.

For sensitivity comparison, we search for high scoring CTCF motif matches within 100bp of any of the events predicted by any of the eight methods. We then calculate the

total number of these motif matches that are discovered by increasing number of the strongest ranked events identified by different methods. The result is given in Supplementary Fig. 4.

We repeated the above analysis with 50bp window size and the results are similar to the results with 100bp window size.

Evaluating performance in deconvolving joint events using synthetic data:

In order to test the performance of joint event detection we constructed realistic synthetic datasets using CTCF binding data. These synthetic datasets allow us to more accurately evaluate the performance of different methods, as we know the true location of the constituent parts of joint binding events. To construct the datasets, we first collect the set of CTCF events that have the following properties: i) they are predicted by five evaluated methods (GPS, SISSRs, MACS, cisGenome, QuEST), ii) none of the five methods predicts more than two events in the region, iii) they contain a match to the CTCF motif, iv) the average distance from the motif match to the event prediction across all five methods is less than 10bp, v) the enrichment of CTCF ChIP-seq reads under the event is significantly greater than the level of GFP reads with a P-value of less than 0.001 (as calculated by a binomial test). A total of 3,233 CTCF binding events meet these criteria.

Synthetic ChIP-Seq data were constructed by randomly choosing one of the real CTCF events and translating the coordinates of its reads in the surrounding 1Kbp onto a fake genome. This is repeated to simulate 20,000 synthetic single events, each placed 100Kbp apart on the fake genome. We similarly create 1,000 joint (binary) binding events by randomly choosing two real CTCF events and placing them on the fake genome a fixed distance apart. Note that this method of constructing synthetic joint events assumes that the ChIP-seq reads generated by closely neighboring events will be an independent mixture of the reads generated by each component event. A synthetic control channel is simulated by taking GFP reads in the regions around CTCF events and translating their coordinates in the same way as the matched IP reads. Further control reads are randomly spread across the fake genome until the read counts in the synthetic IP and control channels match. Datasets are constructed for the following distances between joint binding events: 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120, 140, 160, 180, 200, 220, 240, 260, 280, 300, 350, 400, 450, 500, 550, 600, 650, 700, and 750. The datasets can be downloaded from the GPS website.

Note that we generated the synthetic data with the number of single events and joint events on the same order as real data. However, the read counts and fake genome size are different from the real experiment, and this may throw off some methods that are tuned to make certain assumptions about the distribution of the data.

GPS, SISSRs, MACS and QuEST were run with default settings on the synthetic datasets. cisGenome, FindPeaks and spp are not evaluated because we cannot script

them to run on command line for the repeated tests on multiple synthetic datasets. Note that MACS should not be unduly affected by joint binding events when estimating the correct (CTCF) binding distribution, as a large set of single events exist in the synthetic experiment.

An algorithm is said to have correctly recovered joint binding events when it makes two event predictions in the relevant area and these predictions are each within 100bp of the position at which the event was simulated. Sensitivity in recovering joint binding events is given in Figure 3b.

Evaluating performance in deconvolving joint binding events using GABP ChIP-Seq data:

To evaluate the genome-wide performance of joint event discovery in real ChIP-Seq data, we analyzed a human GABP ChIP-Seq dataset, which was reported previously to contain multiple binding motifs in a short region (Lun, et al., 2009; Valouev, et al., 2008).

For the GABP dataset, we compared GPS against other 4 methods (SISSRs, MACS, cisGenome and Quest) genome wide. FindPeaks only reported 615 events (991 events with the –subpeaks option), much fewer than the other methods. Therefore it is not included in the subsequent joint event discovery analysis. We did not run spp on GABP data because the data format we downloaded from QuEST website can not be used for spp, which reads BOWTIE or ELAND format.

The number of events predicted by all five methods are: GPS (17,179), SISSRs (16,567), MACS (14,527), cisGenome (21,101), QuEST (6,442). The result from QuEST are downloaded from the QuEST web site (http://mendel.stanford.edu/SidowLab/downloads/quest/). The same number of top 6,442 events from each of the methods were used in our comparison.

We define a set of candidate sites that all have at least one event detected by all five methods, and that contain two or more GABP motifs separated by less than 500bp. We discovered 581 such sites. Thus nearly 9% of the GABP bound regions potentially contain joint binding events. For each of these sites, we count the number of events discovered by different methods. The result is given in Figure 3d.

CSDeconv can't be applied genome-wide to a mammalian genome because of its computation time (Lun, et al., 2009). For a 2Mbp region (19:60,000,000-62,000,000) on the GABP dataset that CSDeconv can process(Lun, et al., 2009), we compare its result with the results from GPS and other methods. CSDeconv identified 23 events in this region. Close inspection on the data indicates that some high-confidence events are missed by CSDeconv. Therefore, we performed the joint event discovery (as described above) with the top 36 events (36 is the minimum total number of events in this region called by methods other than CSDeconv) from each method. The result is given in Supplementary Table 3. An example of joint events is given in Supplementary Fig. 5.

Evaluating the alignment mode of GPS:

We performed GPS in multi-condition alignment mode using human CTCF ChIP-Seq data from two different cell types (GM12878, HUVEC) and determined the distribution of the distances between events across conditions in two cases. For comparison, we performed GPS on each condition separately. The result is given in Fig. 4. A representative example is given in Supplementary Fig. 6.

Position-specific prior for motif discovery:

The high spatial resolution of GPS makes it possible to direct motif searches to narrow windows around GPS events and to further improve motif discovery using a position-specific prior derived from GPS output. We computed distributions of motif occurrences proximal to GPS event calls for three TFs with strong motifs: CTCF, Oct4, and Sox2 (Chen, et al., 2008). We searched within a 1000bp window centered on each event and included the distance to the closest motif match in the distribution. For all factors the variance in the distribution decreased sharply as the minimum size of the events examined was increased.

Examining the 100 strongest events (most reads) with a motif, we observed that for each factor more than 90% of the motif occurrences were within 50bp of the GPS event. We then combined the data for all three TFs for the 100 largest events with a motif within 50bp and smoothed with a Gaussian kernel to create a 101bp wide position-specific prior (Supplementary Fig. 8) suitable for use with Priority (Narlikar, et al., 2006) and other motif discovery tools.

References

Birney, E.*, et al.* (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project, *Nature*, **447**, 799-816.

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

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.

Ji, H.*, et al.* (2008) An integrated software system for analyzing ChIP-chip and ChIP-seq data, *Nat Biotechnol*, **26**, 1293-1300.

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., Tolstorukov, M.Y. and Park, P.J. (2008) Design and analysis of ChIP-seq experiments for DNA-binding proteins, *Nat Biotechnol*, **26**, 1351-1359.

Lun, D.S*., et al.* (2009) A blind deconvolution approach to high-resolution mapping of transcription factor binding sites from ChIP-seq data, *Genome Biol*, **10**, R142.

Mahony, S*., et al.* (2005) Improved detection of DNA motifs using a self-organized clustering of familial binding profiles, *Bioinformatics*, **21 Suppl 1**, i283-291.

Matys, V*., et al.* (2003) TRANSFAC: transcriptional regulation, from patterns to profiles, *Nucleic Acids Res*, **31**, 374-378.

Narlikar, L*., et al.* (2006) Informative priors based on transcription factor structural class improve de novo motif discovery, *Bioinformatics*, **22**, e384-392.

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

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

Zhang, Y*., et al.* (2008) Model-based Analysis of ChIP-Seq (MACS), *Genome Biol*, **9**, R137.

Algorithm 1 Alignment of events across conditions

Initialize $\hat{\boldsymbol{\pi}}$ uniformly.
while (EM not converged yet) **do**
At iteration i:

1. Evaluate $\hat{\boldsymbol{\pi}}^{(i)}$ on all conditions using a Simple Mixture Model with a negative Dirichlet-type sparse prior with parameter $\boldsymbol{\alpha}$ [1].
   (a) $\gamma(z_n = j) = \frac{p(r_n|j)\cdot\pi_j^{(i-1)}}{\sum_{j'=1}^{M} p(r_n|j')\cdot\pi_{j'}^{(i-1)}}$
   (b) $\hat{\pi}_j^{(i)} \propto \max(0, \sum_{n=1}^{N} \gamma(z_n = j) - \alpha)$
   where $\mathbf{r} = \{\mathbf{r}_1, \ldots, \mathbf{r}_C\}, N = \sum_{t=1}^{C} N_t$. $\{C$ is the number of conditions.$\}$
   and $r_n \in \mathbf{r} = \{r_{1,1}, \ldots, r_{1,N_1}, \ldots, r_{C,1}, \ldots, r_{C,N_C}\}$.
2. Check convergence based on $\hat{\boldsymbol{\pi}}$.

end while

for $t \in \{1, \ldots, C\}$ **do**
Initialize $\hat{\boldsymbol{\pi}}_t$ as follows:

$$\hat{\pi}_{t,j} = \begin{cases} \frac{1}{|\mathcal{NZ}|} & , \forall j \quad \text{s.t. } \pi_j \neq 0 \\ 0 & , \text{otherwise} \end{cases}$$

It is: $\mathcal{NZ} = \{j | \pi_j \neq 0\}$.

while (EM not converged yet) **do**
At iteration i:

1. Evaluate $\hat{\boldsymbol{\pi}}_t^{(i)}$ using ML EM.
   (a) $\gamma(z_{t,n} = j) = \frac{p(r_{t,n}|j)\cdot\pi_{t,j}^{(i-1)}}{\sum_{j'=1}^{M} p(r_{t,n}|j')\cdot\pi_{t,j'}^{(i-1)}}$
   (b) $\hat{\pi}_{t,j}^{(i)} \propto \sum_{n=1}^{N_t} \gamma(z_{t,n} = j)$
2. Check convergence based on $\hat{\boldsymbol{\pi}}_t$.

end while
end for

References

[1] M. Bicego, M. Cristani, and V. Murino. Sparseness Achievement in Hidden Markov Models. In *Image Analysis and Processing, 2007. ICIAP 2007. 14th International Conference on*, pages 67–72, 2007.