# Supplementary Information for:

## Decoding ChIP-seq peaks with a double-binding signal refines binding peaks to single-nucleotide and predicts cooperative interaction

Table of contents:

## Detailed derivation of the impulse response

Here we provide a full derivation for the impulse response function (Equation 1 in the main text) based on the fact that it follows an extreme value problem.

We define as *v(x)* the probability a break point occurs at a distance *x* from a reference binding site and $V(x) = \sum_{i=0}^{x} v(i)$ the cumulative distribution function (cdf), indicating the probability a break point occurs up to a distance *x* from a binding site. The immunoprecipitation step selects the break point, at each edge, with the closest distance to a binding site. If *V(x)* is sampled *N* times, the cdf describing the minimum value out of *N* samples is represented by the following equation:

$$V_{1:N}(x) = 1 - (1 - V(x))^N \quad \text{[S1]}$$

The number of samples is equivalent to be number of break points that occur in a region surrounding the binding site. Since the shearing step occurs at random, *N* is best described as a stochastic process and the correct cdf that describes the impulse response is computed as:

$$F(x) = \sum_{n=1}^{\infty} P(N = n \mid n > 0) \cdot V_{1:N}(x) \quad \text{[S2]}$$

Assuming *N* follows a Poisson distribution with parameter $\lambda$, the Equation S2 can be written as $F(x) = \sum_{n=1}^{\infty} \frac{e^{-\lambda}\lambda^n}{n!(1-e^{-\lambda})} V_{1:N}(x)$. After simplification, it will take the form of the Equation 1 (main text), which we repeat in the following equation:

$$F(x) = \frac{1 - e^{-\lambda \cdot V(x)}}{1 - e^{-\lambda}} \quad \text{[S3]}$$

**Impulse response as a Gumbel distribution.**

It is hard to know the exact distribution for *V*(*x*). A precise value might depend on different variables such as the DNA conformation, nucleotide composition, presence of ligands, elasticity, bound factors, and intensity of sonication. However, assuming *V(x)* has a truncated exponential shape (e.g. $V(x) \propto e^{x/\beta}$),

Equation S3 becomes $F(x) \approx 1 - e^{-e^{\frac{x+\ln(\lambda)\cdot\beta}{\beta}}}$ and can be approximated into a Gumbel distribution:

$$F(x) \approx 1 - e^{-e^{\frac{x-\mu}{\beta}}} \quad [S4]$$

The advantage of this assumption is that the parameters of the Gumbel distribution provide a physical interpretation for the impulse response. The parameter $\beta$ corresponds to the shape of the impulse response and indicates the break resistance around a binding site. The higher the value of $\beta$ the harder it is for a break point to occur near the binding site, suggesting that the transcription factor creates a protective region around the site it binds. The parameter $\mu$ represents half of the peak shift between the coverage of the negative and positive strands (see Fig. 2B). The physical interpretation for $\mu$ depends on both the Poisson parameter ($\lambda$) and $\beta$, in the form of $\mu=ln(\lambda)\cdot\beta$. The parameter $\mu$ also contains, implicitly, the possibility that the best reference for $V(x)$ is not the center of a binding site, but some point at the edge outside the region the protein binds. In this case, we would have $\mu=x_0 + ln(\lambda)\cdot\beta$, where $x_0$ indicates a region fully protected from shearing.

The impulse response used in the deconvolution process takes the form of a probability distribution function. The derivation showed so far represents the cumulative distribution function for the impulse response. The probability density function comes from the derivative of the cdf, and the impulse response is represented in the form:

$$f(x) \approx \frac{1}{\beta} \cdot e^{\frac{x-\mu}{\beta}} \cdot e^{-e^{\frac{x-\mu}{\beta}}} \quad [S5]$$

The full validation of the assumptions used to derive the impulse response into a Gumbel distribution is not part of the scope of this paper. However, the representation of the impulse response in terms of an extreme value problem is

mathematically compelling and motivates the use of a Gumbel distribution. The physical insights of this model are focuses of future research.

## Empirical estimation of V(X)

Equation S3 (Equation 1 in the main text) contains a parameters, *V(x)*, that indicates the probability that a break point occurs up to a distance *x* from a given binding site. Rearranging this equation, we can predict *V(x)* from the ChIP-seq coverage, according to the following equation:

$$V(x) = \frac{-\log(1 - F_C(x; x_0)(1 - e^{-\lambda}))}{\lambda} \qquad \text{[S6]}$$

Where $F_c(x;x_0)$ represents the empirical cumulative distribution function around a binding site that is centered at $x_0$ and is obtained from the ChIP-seq coverage.

We illustrate the estimation of V(x) from the ChIP-seq coverage in Figure S1. The purpose of Figure S1 is to show a potential physical interpretation that arises from modeling the impulse response as an extreme value distribution and does not affect the blind-deconvolution model of BRACIL. The upward concave shape near the center of the binding site, represented by an exponential fit with positive parameters, indicates that DNA shearing is harder to occur at a distance close to the binding site (up to around 75 bp apart). This result is consistent with the results of ultrasound cleavage of DNA, in which DNA shearing saturates at small size of DNA fragments (Fukudome et al. 1986), and might be consequence of decreasing chance to shear short DNA pieces. An alternative explanation is that it might be consequence of a protection region around the site a TF binds.

## Motivation for the penalty function after motif integration

The penalty function takes into account motif conservation. In this context, a motif with better conservation (by means of motif p-value) will be penalized less than one that is weakly conserved. Studies of protein-DNA binding have shown that the probability of *in vitro* binding (Maerkl and Quake 2007) can be measured according to a logistic function of the binding affinity estimated from motif conservation (Zhao et al. 2009). This logistic behavior supports a binary classification for the penalty function, with a zero contribution for strong sites and a constant to weak sites.

## Estimating cooperative interaction from ChIP-seq

In order to use ChIP-seq to test for cooperative interaction, we need to define a null hypothesis that assumes independent binding and an alternative hypothesis that suggests cooperative interaction. A formal statistical test for cooperative interaction is defined below.

**Definition of null and alternative model**

Before defining the null and the alternative hypothesis to test for cooperative interaction, we need to define cooperative interaction. Cooperative interaction occurs when the binding to two neighboring sites is not independent from each other.

Considering a region with two neighboring sites, four binding configurations are possible: (0,0), (0,1), (1,0) and (1,1), where each number refers to a binding site and the values 1 and 0 indicate whether it is bound or not. This representation allows cooperative interaction to be defined in terms of the probability that both sites are simultaneously bound ($p_{1,1}$) and the probability of binding to each site ($p_{1,\cdot} = p_{1,0} + p_{1,1}$ and $p_{\cdot,1} = p_{0,1} + p_{1,1}$). The formal definition of cooperative interaction is represented in the following equation:

$$\omega = \frac{p_{1,1}}{p_{1,\bullet} \cdot p_{\bullet,1}} \qquad \text{[S7]}$$

In this context, $\omega$=1 indicates that binding is independent and $\omega \neq 1$ indicates cooperative interaction.

The binding probabilities are estimated from the impulse signal magnitudes. The rationale for this relationship is as follows. A configuration with only one site bound can only emit a single-binding signal. This creates an association between the single binding magnitudes, $m_1$ and $m_2$, and the corresponding binding probabilities, $p_{1,0}$ and $p_{0,1}$. Similarly, the double-binding signal occurs for the configuration that both sites are bound and close to each other. This justifies the relationship of the double binding magnitude, $m_{1,2}$, and the double bound probability, $p_{1,1}$. The magnitude of each signal also depends on the probability that the target transcription factor is selected by immunoprecipitation, which is

represented by the constant $\rho$. A single-binding fragment contains one TF target and is purified with probability $\rho$, while a double-binding fragment contains two targets and is purified if any of the targets is immunoprecipitated, i.e. with probability $1-(1-\rho)^2$. A summary of the relationship between the magnitude of the impulse responses and the probabilities for each configuration is described in the following equation:

$$m_1 \propto \rho \cdot (p_{1,0}) \qquad \text{[S8]}$$
$$m_2 \propto \rho \cdot (p_{0,1})$$
$$m_{1,2} \propto (1-(1-\rho)^2) \cdot (1-F_s(d_{1,2})) \cdot p_{1,1}$$

The proportion indicates that the scaling factor between magnitude and probability is unknown. The term $F_s(d_{1,2})$ represents the probability a double-binding fragment can be split into two single-binding fragments. In the representation shown in equation S8, We assume that a potentially double-binding signal that is split into two fragments will provide neither a single-binding nor a double-binding impulse response.

The theoretical derivation of the impulse response (section 2.2 and sup. section S1) indicates that the cumulative distribution function of the impulse response, $F(x)$, represents the probability a break point occurs up to a distance $x$ of a binding site. If this probability is independent of the binding configuration, the probability a double binding fragment is split in two single-binding signal can be computed from the impulse response, i.e. $F_s(x) = F(x)$. Assuming that the double binding configuration increases the protective area around two closely spaced binding sites will reduce the chance that a double binding fragment will be split in two, implying that $F_s(x) \ll F(x)$. Approximated values for the impulse response indicates that $F(x)<0.2$ for a binding site distance of 20 bp. This implies that $F_s(x) \ll 0.2$ and the equation S8 is simplified to:

$$m_1 = c \cdot \rho \cdot p_{1,0} \qquad \text{[S9]}$$
$$m_2 = c \cdot \rho \cdot p_{0,1}$$
$$m_{1,2} = c \cdot (1-(1-\rho)^2) \cdot p_{1,1}$$

Here the constant $c$ was used to transform the proportion into equality.

The assumption of independent binding causes $m_{1,2}$ to be a function of $m_1$ and $m_2$. This constraint disappears in the case of cooperative interaction. Thus, the assumption of independent binding is a particular case of the cooperative interaction.

The closed solution for $m_{1,2}$ as a function of $m_1$ and $m_2$ is shown for two extreme cases, assuming low and high immunoprecipitation rate. The solution comes from solving equation S9 constrained to independent binding ($\omega=1$, equation S7) and to the fact that $p_{00}+p_{10}+p_{01}+p_{11}=1$. A detailed derivation is not shown, but the solution is easily achieved using an algorithm that solves systems of equations.

When immunoprecipitation rate is low, $\rho\approx 0$, the term $(1-(1-\rho)^2)$ in equation S8 is simplified to $2\rho$, and $m_{1,2}$ is computed as following:

$$m_{1,2} = \frac{-(m_1+m_2)\cdot p_{0,0} + ((m_1+m_2)^2\cdot(p_{0,0})^2 + 4m_1\cdot m_2\cdot(1-p_{0,0})\cdot p_{0,0})^{1/2}}{p_{0,0}} \quad \text{[S10a]}$$

Similarly, when immunoprecipitation rate is high, $\rho\approx 1$, the term $(1-(1-\rho)^2)$ in equation S8 is simplified to $\rho$, and $m_{1,2}$ is computed as following:

$$m_{1,2} = \frac{-(m_1+m_2)\cdot p_{0,0} + ((m_1+m_2)^2\cdot(p_{0,0})^2 + 4m_1\cdot m_2\cdot(1-p_{0,0})\cdot p_{0,0})^{1/2}}{2\cdot p_{0,0}} \quad \text{[S10b]}$$

Notice that $m_{1,2}$ depends not only on $m_1$ and $m_2$, but also on the probability that none of the sites are bound, $p_{0,0}$. This happens because, as shown in equation S8, there is no signal in the ChIP-seq coverage with direct correspondence to $p_{0,0}$.

The null model (independent binding) is defined from Equations S10a and S10b. The minimization of the objective function (*ML* step, equation 7a) is performed such that the magnitude of the double binding signal is constrained according to equation S10a or S10b. Equations S10a and S10b are also used as a simplified model for cases with more than two binding sites. Finally, the probability that none of the sites are bound is unknown, thus the objective

function depends on an input parameter $p_{0,0}$. The objective function for the null model is represented as following:

$$obj_{r,null;p_{0,0}} = obj_r(L, M, \theta; m_{i,i+1}(m_i, m_{i+1}, p_{0,0})) \quad \text{[S11]}$$

The term $m_{i,i+1}(m_i, m_{i+1}, p_{0,0})$ indicates that the magnitude of each double binding signal is a function of the magnitude of its neighbor sites and $p_{0,0}$.

The magnitude of the double binding signal is unconstrained for the model that includes cooperative interaction and follows the representation shown in the main text (Equation 5). The objective function computed for the null and alternative models is used in the likelihood ratio test. The assumption of independent binding turns the null model a particular case of the alternative model, justifying the use of the likelihood ratio test.

**Defining the statistical test**

Here we derive how to use a likelihood ratio test to detect cooperative interactions. Let $L_{null}$ and $L_{alternative}$ be the likelihood of the null and the alternative models, respectively. The likelihood ratio is defined by a chi-squared distribution, of the following parameter:

$$D = -2 \cdot (\log(L_{null}) - \log(L_{alternative})) \quad \text{[S12]}$$

The number of degrees of freedom is equal to the number of extra parameters allowed for the alternative model when compared to the null model.

The terms $L_{null}$ and $L_{alternative}$ can be computed from the objective function (equation S11 and equation 5). The assumption that the observed coverage follows a normal distribution around the expected value (subsection 4.1.1), brings the following relationship:

$$L = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2} obj \quad \text{[S13]}$$

Where *obj* represents the objective function and *n* the number of points used to compute it. The likelihood *L* is maximized when $\sigma^2 = obj/n$. The number *n* is the same for the alternative and null models, thus the parameter *D* in equation S12 is computed as following:

$$D = n \cdot (\log(obj_{null}) - \log(obj_{alternative})) \quad \text{[S14]}$$

## Evaluating binding site prediction

Our method is validated by comparing the predicted binding events of BRACIL, GEM, and other peak-callers to a reference benchmark. The reference benchmark for DosR is obtained from the single nucleotide resolution experiments performed by Chauhan and Colleagues (Chauhan et al. 2011) and the reference benchmarks for the transcription factors GABPA (Valouev et al. 2008) and CTCF (Chen et al. 2008) are obtained from the output of different motif discovery tools, as suggested in GEM paper (Guo et al. 2012).

We evaluated predictions by matching predicted binding events to the reference benchmarks. For DosR, a predicted event matches a reference binding site if they are less than 50 bp apart from each other, this distance was relaxed to 150 bp for the eukaryotic datasets. If two predicted sites are close to the same reference site, only the one with closest distance is considered to be a match. The matches are used to compute the true/false positives as well as true/false negatives of our predictions. Predictions with a match are true positives and predictions without a match are false positives. Reference sites without a match are false negatives. The true negatives correspond to binding motifs that do not match a reference site and are filtered out in the deconvolution step of our algorithm. The ROC as well as the precision and recall curves shown in Figure 3C-D are computed by ranking predictions by the impulse magnitude (BRACIL predictions) or motif score (motif only predictions).

The resolution on the eukaryotic data is evaluated based on the method presented by Guo and colleagues (Guo et al. 2012). In short, we compare the distance of binding events predicted by BRACIL and GEM to a benchmark of binding site locations obtained by motif discovery algorithms. The location of the reference binding sites is added by a constant shift to correct for an arbitrary definition of the motif center. This constant is defined to minimize the overall distance between predicted and reference binding sites. Six motif discovery algorithms are used for this analysis: ChIPMunk (Kulakovskiy et al. 2010), MEME/FIMO (Bailey and Elkan 1994), HMS (Hu et al. 2010), MDscan (Liu et al.

2002), POSMO (Ma et al. 2012), and Weeder (Pavesi et al. 2001). Motif discovery is run with the parameters defined in GEM paper (Guo et al. 2012). FIMO and Weeder require a threshold for motif scan. For FIMO, we consider motifs with p-value less than$10^{-3}$ and for Weeder we consider sites with a conservation threshold greater than 80% for GABPA and greater than 75% for CTCF. In case two motifs overlap with each other, only the one with best score is considered as a reference.

The performance is evaluated in a set of 500 enriched regions (Figures S15 and S16). For GABPA, this data set is defined as the +/- 150 bp sequence that surrounds the 500 most significant GEM events. Overlapping regions are clustered as a unique sequence to perform motif search and scan. The set of CTCF enriched regions is selected based on the quality score defined in Figure S11. This quality score eliminates instances of highly enriched regions with spurious coverage. The evaluation was also expanded to include all set of enriched regions (Figures S17-S18).

For the eukaryotic dataset, BRACIL prediction is obtained by using conservative parameters. In particular, the threshold of weak and strong sites (in $\log_{10}$(p-value) units) are equal to 3 and 5, respectively. The penalty parameter is equal to 0.1. This choice of parameters causes BRACIL to predict a similar number of binding events as GEM. Our results show a better performance of BRACIL when compared to GEM. The results are illustrated in Figures S15- S19. We also estimate the false negative rate of BRACIL and GEM. As described in Table S4, the false negative rate is smaller for BRACIL when compared to GEM.

The results presented in Figures S15-S19 indicate when BRACIL or GEM would be the best method to identify binding site locations. BRACIL outperforms GEM's resolution for the test case in which multiple binding sites are hidden inside an enriched region. This result is exemplified by the GABPA dataset where BRACIL shows an overall better performance for both top 500 and all enriched regions (Figure S15 and S17, respectively). BRACIL and GEM had a similar performance in predicting binding site locations for the CTCF transcription factor

when applied to the set of 500 enriched regions with best coverage quality score (Figure S16, see also Figure S11 for coverage quality evaluation). GEM outperformed BRACIL when all CTCF enriched regions were used for evaluation (Figure S18). The entire CTCF set include enriched regions with spurious coverage in which BRACIL is not expected to perform well (Figure S11). Most CTCF regions contains a single binding site and the CTCF motif is more specific than the GABPA one. Thus CTCF motif matches are more likely to be true positives.

## Computing significance of binding event predictions.

We have defined two metrics for significance of BRACIL event predictions. One metric assigns a p-value to each region and the other a p-value to each binding event. The significance is obtained by comparing a score of the real data with a random score obtained from a random dataset. The random dataset is created by resampling with repetition the coverage per enriched region. The significance is measured by counting the fraction of random scores that are at least as good as the score from the real data.

The score used to compute significance per region is defined as the objective function (Equation 5). The score used to compute p-values per binding site is defined as the magnitude of the impulse response at each predicted binding site position (Equation 2). The fit is performed with the parameters of the Gumbel distribution obtained from real data. We applied this method to the enriched regions that contain the DosR binding sites predicted by Chauhan and colleagues. With the exception of one region, all cases had a p-value less than $10^{-3}$. We also computed the p-value per binding event for the same set of enriched regions. The predictive power of the p-value score is shown in Figure S9. We observe an area under ROC curve equal to 0.89.

## Supplementary figures

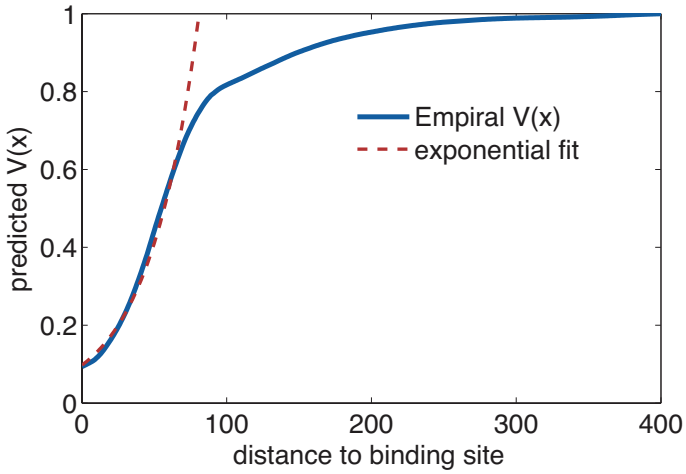


Figure S1: Empirical measurement of *V(x)* based on the ChIP-seq coverage. Our model suggests that the probability a break point occurs up to a distance *x* of a binding site (*V(x)*) can be empirically measured from the ChIP-seq coverage (Equation S5). The upward concavity (red dashed line) in the empirical *V(x)* (blue solid line) indicates that the probability for a break point to occur increases with the distance to the binding site. This might be consequence of a protection region around the site a TF binds or because DNA shearing saturates at small size of DNA fragments (Fukudome et al. 1986).

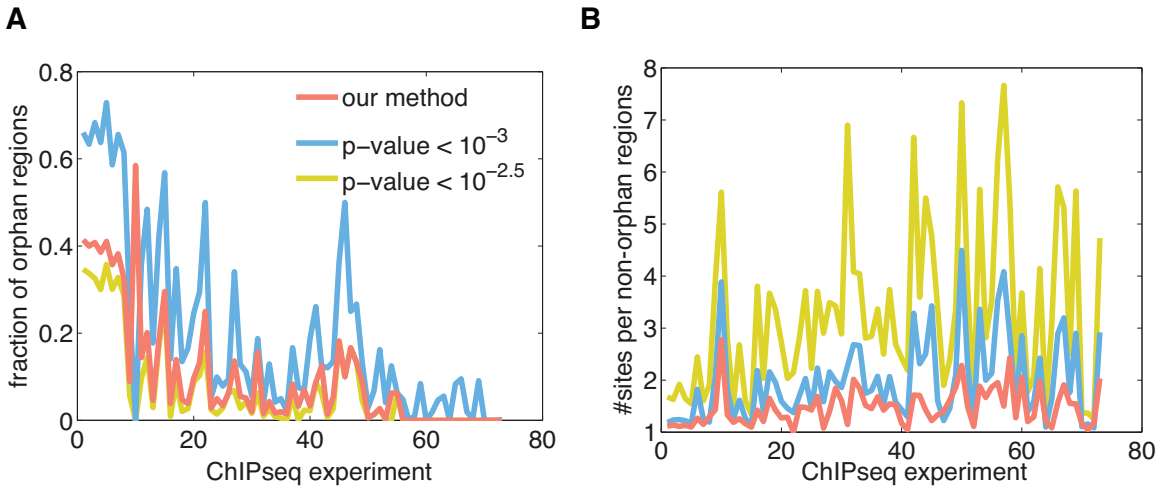


Figure S2: Our method reduces the number of orphan regions while it still filters out false binding sites. We call regions that show enriched coverage without an instance of binding motif as orphan. The number of orphan regions reduces with a more inclusive threshold, such as motif p-value < $10^{-2.5}$, at the cost of increasing the amount of false binding sites. The threshold motif p-value < $10^{-3}$ is commonly used to provide a balance between false positive and true positive. Our method allows a more inclusive threshold at the same time it uses the ChIPseq coverage to filter out for potential false positives. We show the fraction of orphan regions (A) and the average number of sites per non-orphan regions (B) per ChIP-seq experiments for three methods. Our method (red line) reduces the number of orphan regions when compared to what is identified using a motif p-value < $10^{-3}$ (blue line). The threshold motif p-value < $10^{-2.5}$ (yellow line) is more inclusive and shows the least number of orphan regions. The difference in the number of orphan regions predicted by our method and motif p-value < $10^{-2.5}$ indicates that part of this reduction is not supported by ChIP-seq coverage. This is in agreement that a low motif p-value threshold will identify false binding sites. The data used for this analysis is taken from our study of the regulatory network of *M. tuberculosis* (Galagan et al. 2013). We plotted only experiments with at least 10 enriched regions. The x-axis is sorted according to absolute number of orphan regions detected by our method.

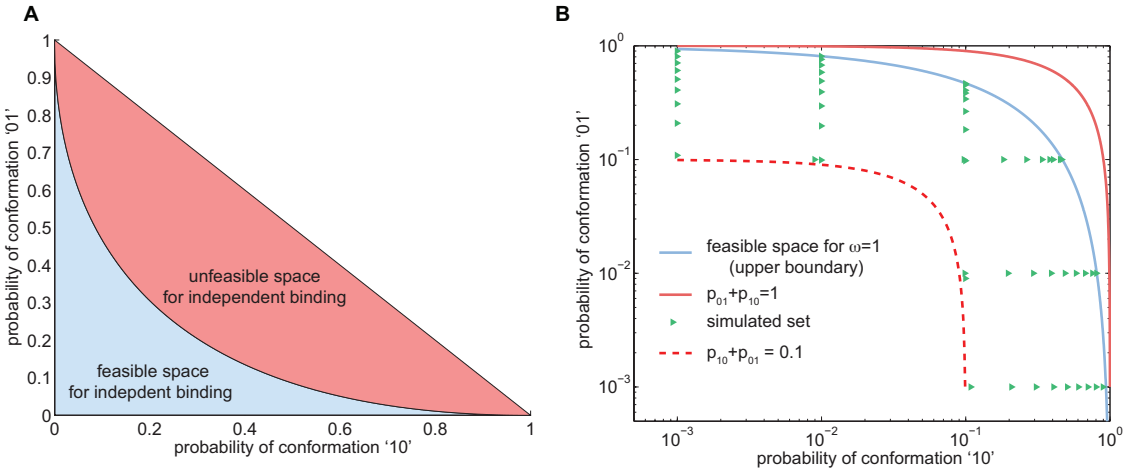


Figure S3: The feasible space for independent binding. (*a*) This figure represents the space of binding configuration probabilities for two binding sites in terms of the probabilities that only one of the sites is bound. The space is constrained to $p_{1,0} + p_{0,1} \leq 1$ (see section S4.1). Independent binding is only possible in the range of probabilities indicated by the blue area. This space can be solved analytically and is represented by $(p_{0,1} - p_{1,0})^2 - 2\cdot(p_{0,1} + p_{1,0}) + 1 \geq 0$. Under the assumption of independent binding, each point in the blue area determines uniquely the values of $p_{0,0}$ and $p_{1,1}$. (*b*) Different points representing independent binding (green markers) were used to create the simulated set of enriched regions (see section 4.5). These points were chosen to be representative of the feasible space and challenge cooperative detection for different proportions of binding configuration (see table S2). The solid blue line shows upper boundaries for the independent binding feasible space, solid red line indicates boundary for probability space ($p_{1,0}+p_{0,1}=1$). The dashed red line indicates single binding configuration occurs with 10% probability ($p_{1,0}+p_{0,1}=0.1$). The ratio $p_{0,1}/p_{1,0}$ varies up to three order of magnitudes.

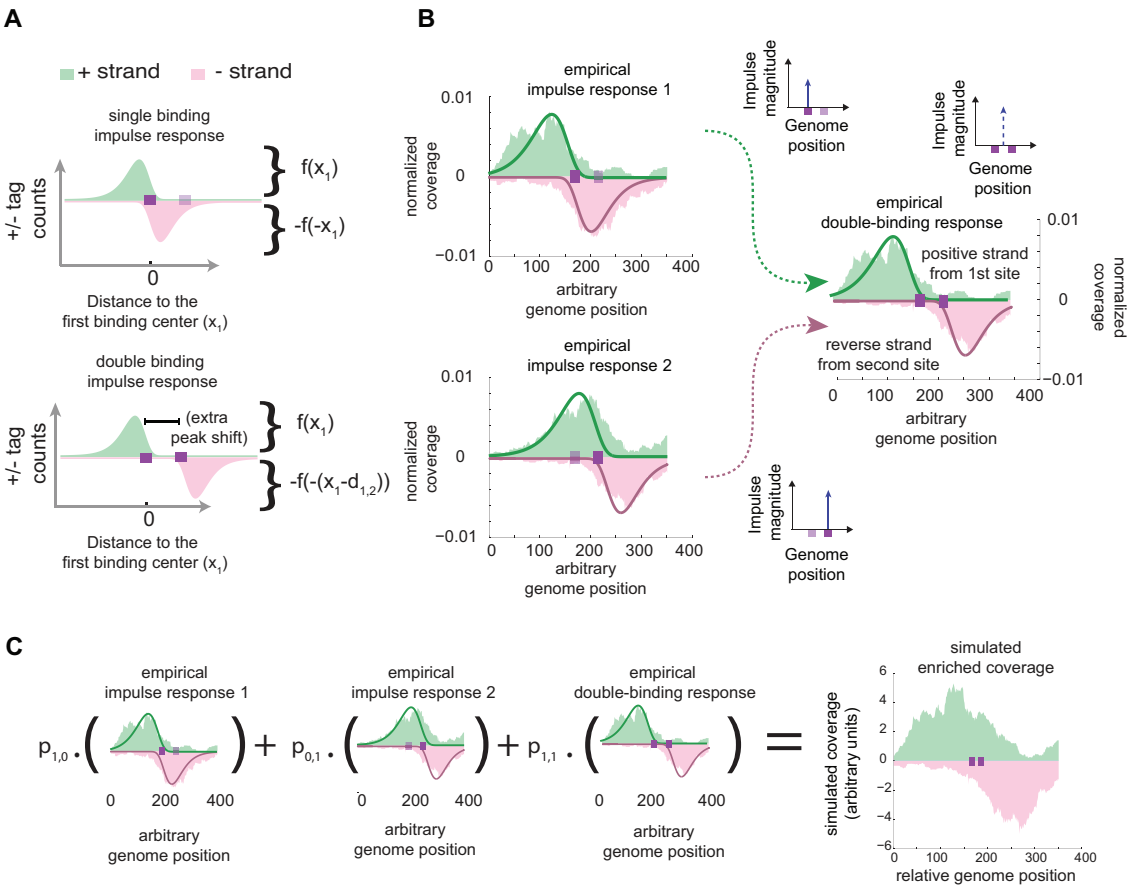


Figure S4: Schematic representation showing how a simulated enriched region was created. Each region was defined to contain two binding sites and the binding sites were assumed to bind independently from each other. (*A*) The theoretical representation of the single-binding and the double-binding impulse responses (see sections 2, 4.1 and equation 3). The forward and reverse coverage of the double-binding impulse response has a correspondence to each single-binding responses. The binding sites are represented by purple squares. The dark and light shades indicate if sites are bound or unbound, respectively. (*B*) An empirical impulse response corresponds to the observed coverage, taken from real data, around a region containing only one binding site. The empirical double-binding impulse response is simulated from the coverage of two single-binding empirical impulse responses, according to the model represented in *(A)*. The small panels at each plot show an impulse representation for each signal. *(C)* The simulated enriched region is obtained by performing a weighted sum of the empirical impulse responses. The weights are scaled according to the corresponding binding probability. The binding probabilities are defined under the constraint of independent binding (see table S2).

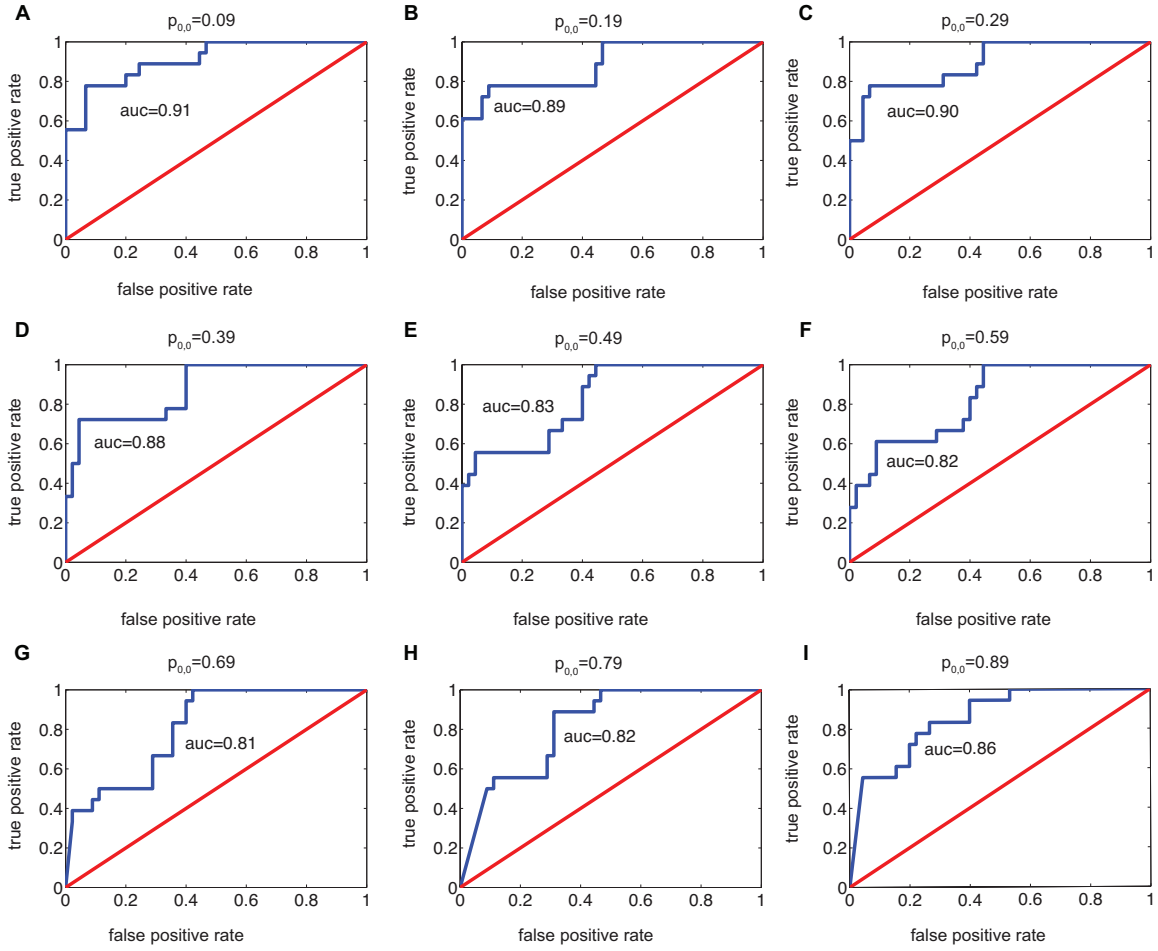


Figure S5: The method to detect cooperative interaction is robust to an exhaustive range of the non-binding configuration probability. The non-binding configuration probability ($p_{0,0}$) can not be extracted from the ChIP-seq data and is a necessary input to model independent binding (equations S10a-b and S11). The performance of our method is presented in terms of the true positive rate as a function of the false positive rate. The true positive set corresponds to regions experimentally validated to contain cooperative interaction and the false positive set indicates simulated regions containing independent binding (section 4.5). (A-I) Each plot illustrates the performance assuming a fixed value of $p_{0,0}$. This panel assumes that immunoprecipitation occurs at low rates (equation S10a). All the results corroborate our method.

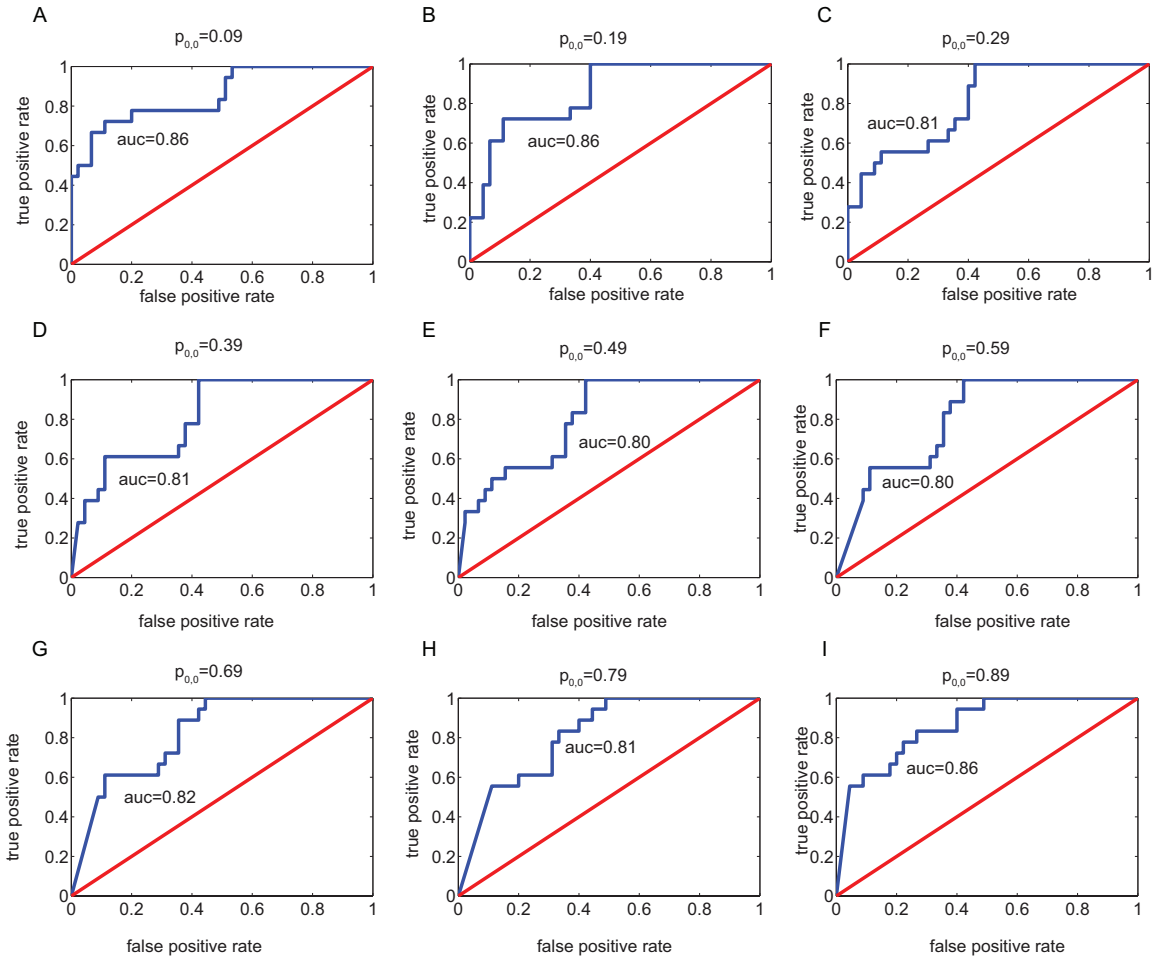


Figure S6: Similar to Figure S5, however, it assumes that immunoprecipitation occurs at high rate (equation S10b). All the results corroborate our method.

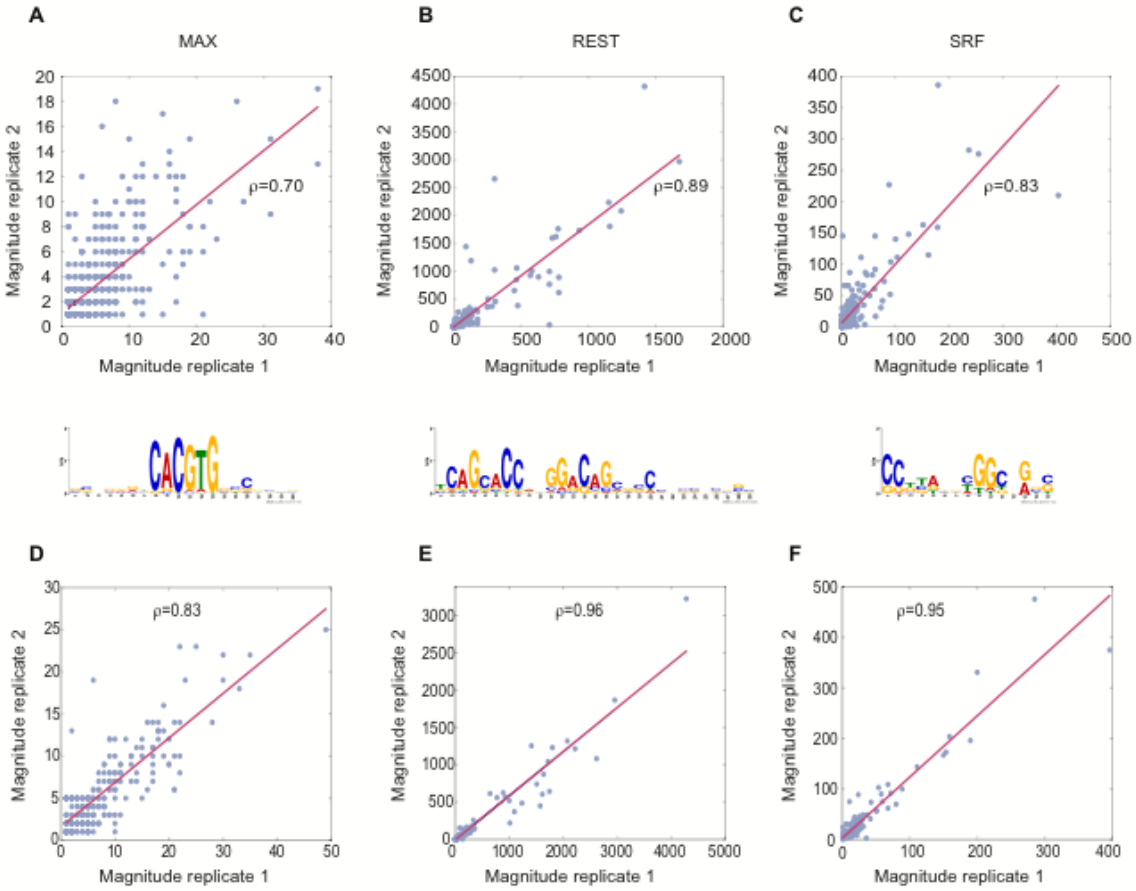


Figure S7: Our method shows high magnitude reproducibility in eukaryote ChIP-seq data. In this analysis, we used the benchmarked ChIP-seq data suggested by Rye et al. (Rye et al. 2011). Unfortunately, this benchmark is based only on enriched regions and further work is required to obtain a high-resolution benchmark with binding sites mapped at single-nucleotide resolution. The results for the transcription factors MAX, REST, and SRF are shown in panel A, B, and C, respectively. At the top of each panel, we plot the reproducibility of impulse response magnitude replicates and in the bottom, the predicted binding motif. The predicted magnitude showed high correlation between REST and SRF replicates and was not as well correlated for MAX. The deconvolution of MAX ChIP-seq data is more challenging because ChIP-seq coverage has low abundance and because motif scan predicts an excessive number, including multiple overlapping candidates, of potential binding sites. This somewhat ambiguous motif prediction of MAX binding sites was previously reported (Pique-Regi et al. 2011). A higher coverage should improve the potential of our deconvolution model in distinguishing the most likely binding sites from the large number of binding site candidates and a high-resolution benchmark would enhance the evaluation and highlight the precision of our method. The correlation between duplicates

increases when we use a more conservative set of parameters (D, E, F), the parameters usage permits the user a tradeoff between specificity and sensitivity. The reproducibility increases by using a more conservative threshold. A, B, C (weak site threshold $-\log_{10}(p) > 2.5$, strong site threshold $-\log_{10}(p) > 4$, alpha = 0.01); D, E, F C (weak site threshold $-\log_{10}(p) > 3$, strong site threshold $-\log_{10}(p) > 5$, alpha = 0.1).

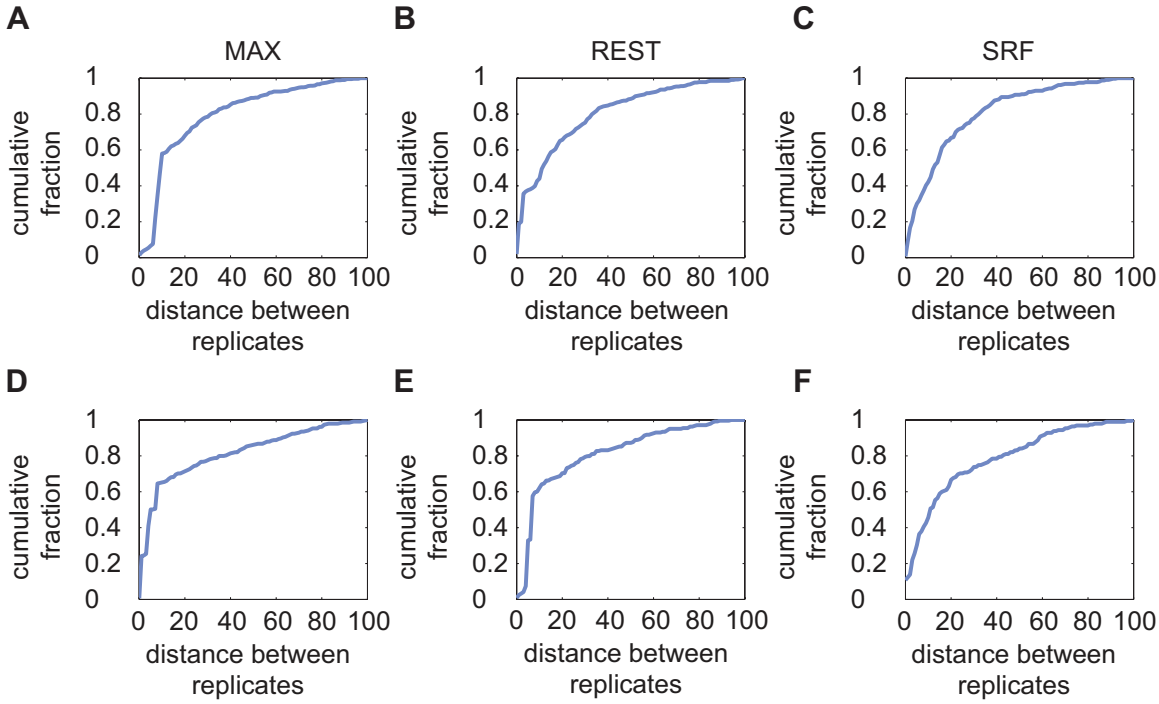


Figure S8: BRACIL shows high-reproducibility in distance between replicates from the data set suggested by Rye et al (Rye et al. 2011). Each panel corresponds to the transcription factor highlighted at the top. . Distance reproducibility is shown by using a less conservative set of parameters (top row, weak site threshold $-\log_{10}(p) > 2.5$, strong site threshold $-\log_{10}(p) > 4$, alpha = 0.01) or a more conservative set of parameters (bottom row, weak site threshold $-\log_{10}(p) > 3$, strong site threshold $-\log_{10}(p) > 5$, alpha = 0.1). We considered only sites with less than 100 bp distance from each other. This figure illustrates that reproducibility is similar for a conservative (A, B, C) and non-conservative (D, E, F) set of parameters.

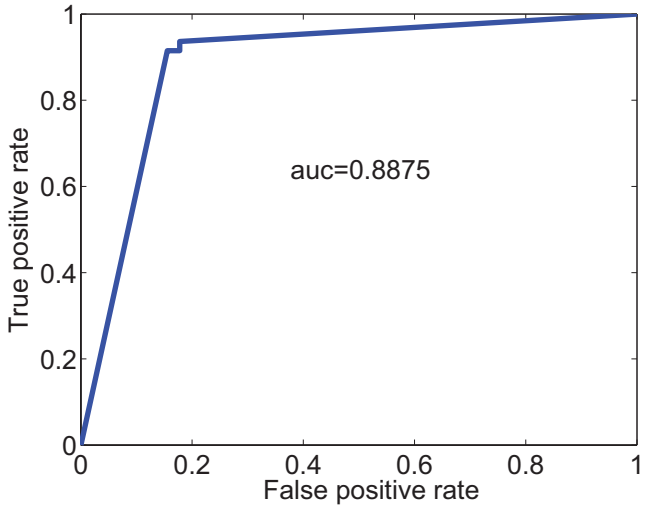


Figure S9: The significance metric of binding events indicates that BRACIL predicts binding sites with high sensitivity and specificity. We plot an ROC curve of binding event prediction in which binding events are ranked according to the event specific p-value (see supporting text *Computing significance of binding event predictions* for details). The results show high sensitivity and specificity, with an area under the curve of 0.8875. The total number of positives is defined as the 47 binding sites obtained by Chauhan and Colleagues. The total number of negatives is defined as the number of motifs predicted by FIMO that are used in the refined step of BRACIL and are not matched to the reference binding sites. Binding sites that are not predicted by BRACIL and binding motifs that are filtered by the deconvolution step are assigned to a p-value equal to 1 for evaluation purpose.

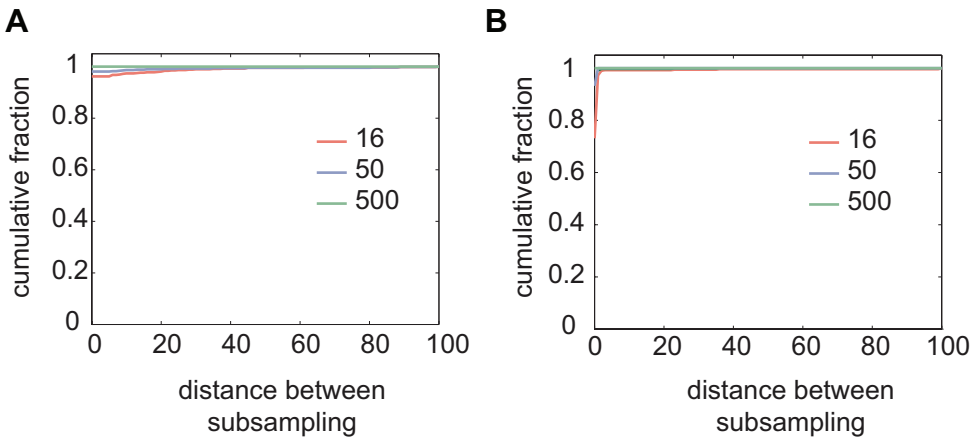


Figure S10: A small sample of enriched regions can reduce the computational cost of our method while it is still informative to train the impulse response parameters. We illustrate this phenomenon by estimating the effect of the subsample size used for training data (16, 50 and 500) in the prediction of binding site locations for the human transcription factor GABPA. The performance is presented as the cumulative fraction of binding site distances predicted by different training sets. Predictions based on the training set of 500 most enriched regions are used as reference. The results suggest that a subsample of size as small as 16 is informative for binding site prediction for predictions based only in ChIP-seq coverage (*A*) as well as prediction that is refined by motif discovery (*B*). The predicted parameter pair ($\mu$; $\beta$) (see Equation S5) for each subsample is shown as following: 16 (31.39, 37.04), 50 (32.35, 38.82), 500 (32.13, 39.11) for predictions based only in ChIP-seq coverage and 16 (30.42; 28.11), 50 (31.39, 28.21), 500 (31.00, 31.55) for predictions that are refined by motif discovery.

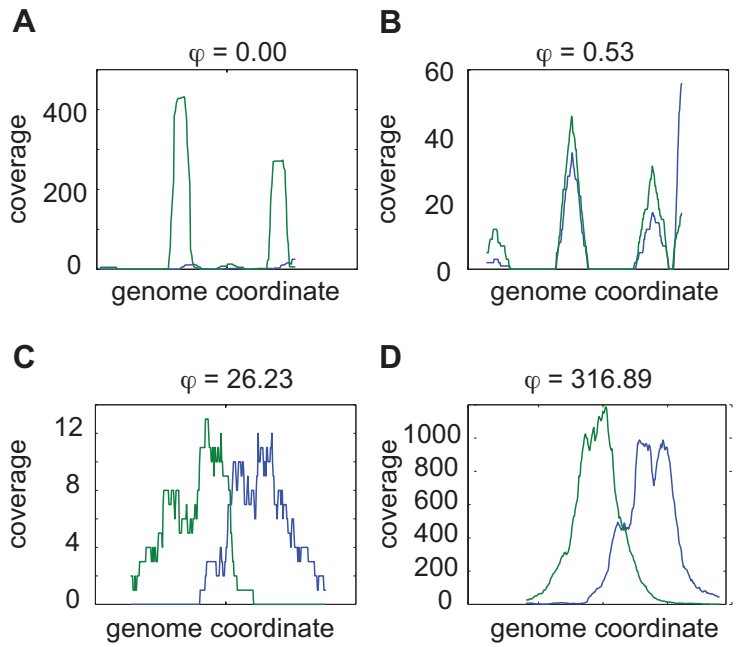


Figure S11: Enriched regions may contain artifacts in the ChIP-seq coverage. Training regions can be provided as input in order to avoid bad quality regions. A metric was defined to evaluate regions' quality. Three parameters were considered to indicate quality, they are: (i) average coverage per nucleotide, $C_t$, (ii) cross-correlation between forward and reverse strand, *xcorr*, and (iii) the ratio between average coverage in the forward and reverse strand (*ratio*). Mathematically, regions were ranked according to the following equation: $\varphi = exp(3 \cdot C_t/C_{max}) \cdot exp(5 \cdot (xcorr-0.3)) \cdot exp(-3 \cdot log_2(ratio))$. The term $C_{max}$ indicates the maximum value of $C_t$ observed in the dataset. The term *xcross* represents the cross-correlation with maximum value in the shift interval from 60 to 120 bp. This approach correctly classified bad quality enriched regions with low rank values. This figure illustrates an instance of region with low (top) and high (bottom) quality scores for the transcription factors CTCF (A,C) and GABPA (B, D).

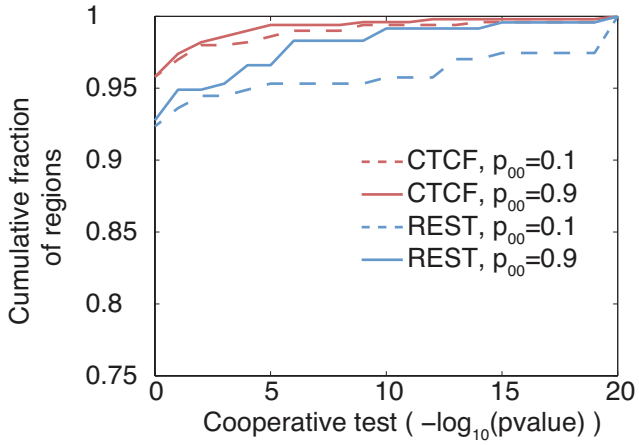


Figure S12: Cooperative interaction test does not reject the hypothesis of independent binding for most CTCF (red lines) and REST (blue lines) regions. This figure illustrates the cumulative fraction of regions as a function of p-value obtained by performing test for independent binding. Results assume a low immunoprecipitation rate (see equation S10) and are shown for two values of probability of non-binding conformation, $p_{00}$. The results for CTCF are shown for the 500 most enriched regions predicted by GEM and ranked according to the metric described in Figure 11. REST regions were obtained from Rye et al. dataset (Rye et al. 2011). BRACIL binding site prediction was obtained by defining the threshold of weak and strong site (in $\log_{10}$(p-value) units) to be equal to 3 and 5, respectively. The penalty parameter was equal to 0.1.

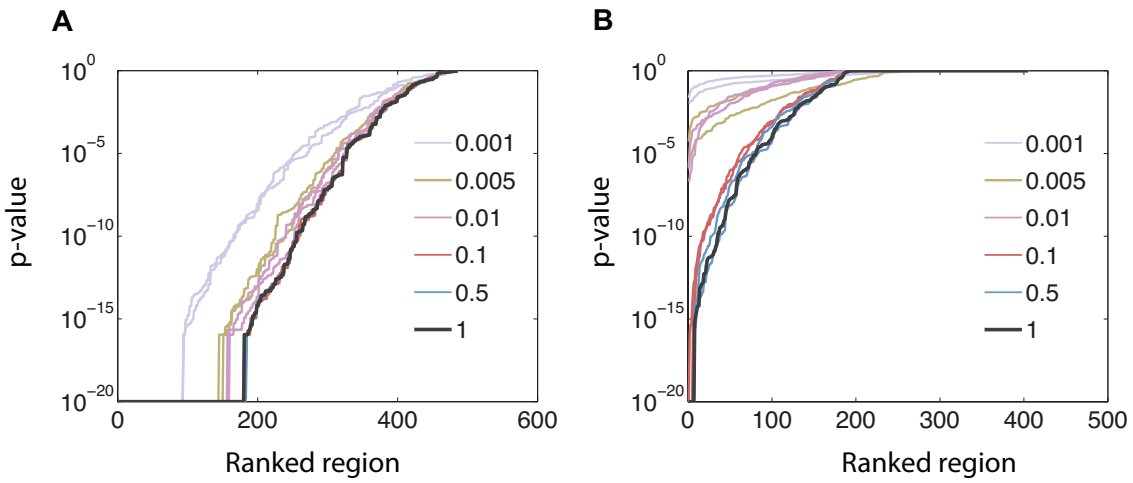


Figure S13: The signal to predict cooperative interaction decays with low coverage depth. For this analysis, the DosR ChIP-seq coverage was randomly subsampled. This process reduces the relative coverage according to the probabilities indicated in the legend of each plot. We plot two replicates for each subsample probability. The reduction in the signal for cooperative interaction can be observed by the shift to the left in lines with lower coverage. Our results indicate that the statistical significance decreases in proportion to the ChIP-seq coverage for both the positive (*A*) and the negative (*B*) control data sets.

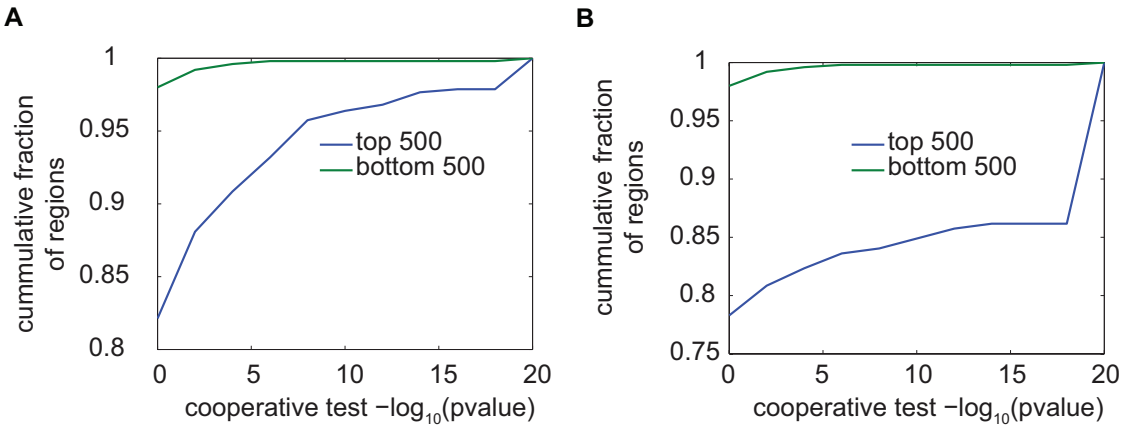


Figure S14: Cooperative interaction prediction in human TF GABPA. The hypothesis of independent binding is less likely for highly enriched regions (top 500) than for lowly enriched regions (bottom 500). The results are shown for two probabilities of non-binding conformation $p_{00} = 0.1$ (A) and $p_{00} = 0.9$ (B). BRACIL binding site prediction was obtained by defining the threshold of weak and strong site (in $\log_{10}$(p-value) units) to be equal to 3 and 5, respectively. The penalty parameter is equal to 0.1.

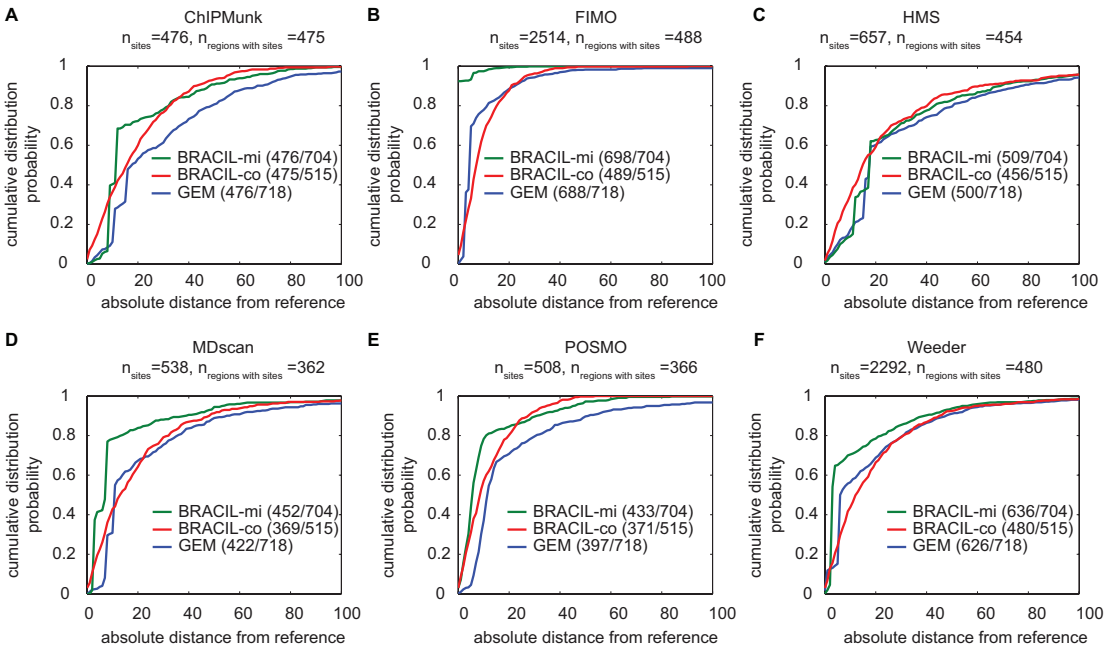


Figure S15: BRACIL improves spatial resolution of GABPA binding event predictions when compared to GEM in highly enriched regions. The legends inside each plot indicate whether BRACIL predictions are refined by motif input (BRACIL-MI) or coverage only (BRACIL-co). The dataset used for this figure considers the top 500 regions enriched regions. On the top of each panel we indicate the motif discovery tool used to create the reference benchmark. It indicates the number of binding sites ($n_{sites}$) in the benchmark and the number of regions that contain at least one binding site ($n_{regions\ with\ sites}$) per motif discovery tool. The ratios in the legend indicates the fraction of binding events that matches a reference binding site in the benchmark with the denominator indicating the number of binding events predicted by BRACIL-mi, BRACIL-co, and GEM. A match between a binding event and a reference binding site occurs when they are up to 150 bp apart and unique (see supporting section *Evaluating binding site prediction*). The results corroborate an improved performance of BRACIL when compared to GEM for all cases.

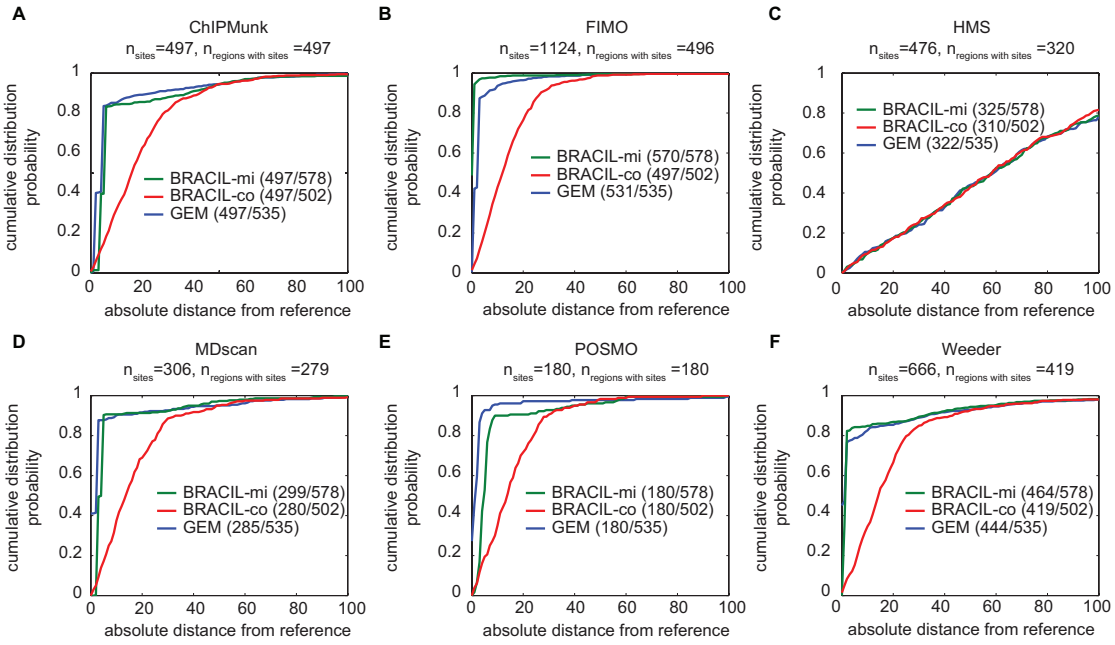


Figure S16: BRACIL and GEM predict CTCF binding events with similar resolution in highly enriched regions. Figure legends and panels follow the same format described in Figure S15. The dataset used for this figure considers the top 500 regions enriched regions. BRACIL shows better performance in the benchmark created by FIMO or Weeder, GEM shows better performance in the Benchmark created by ChIPMunk or POSMO, and a similar performance is observed in the benchmark created by HMS or MDscan. Notice, at the title of the corresponding panel, that over 30% of the enriched regions do not contain any reference binding site by means of three motif discovery tools (HMS, MDscan, and POSMO). Since all regions are expected to contain at least one binding event, this result indicates that motif discovery tools overlook many true binding sites and highlights the need of a high-scale, single-nucleotide resolution, and experimentally validated benchmark, as in the case of DosR binding sites (Chauhan et al. 2011) for more accurate measurements. It also indicates that many motif discovery tools are not able to capture the relevance of weak binding sites. Finally, BRACIL outperformed GEM when the reference binding events capture weakly conserved binding sites (as in the case of Weeder and FIMO).

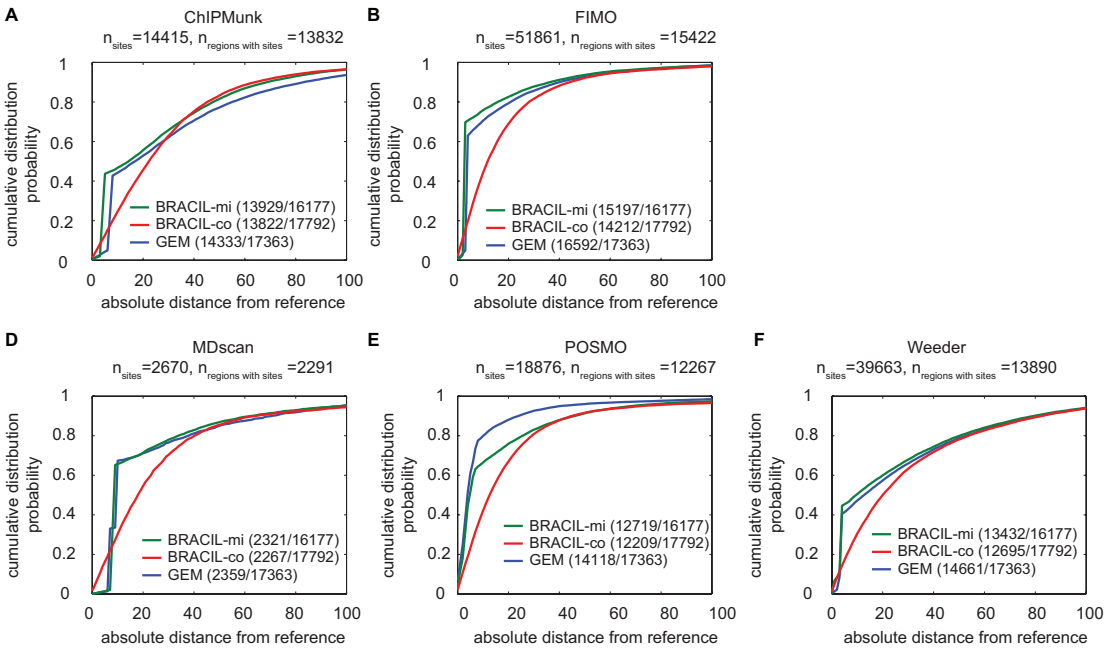


Figure S17: Comparison of BRACIL and GEM for all GABPA regions. Plot follows the same standard of Figure S15. BRACIL shows an overall improved performance when compared to GEM. The only exception is when POSMO is used to obtain the reference set of binding sites. We were not able to run HMS for the large dataset. Also, MEME/FIMO and Weeder used the top 500 regions to predict the binding motif and scanned the motif for all enriched regions.

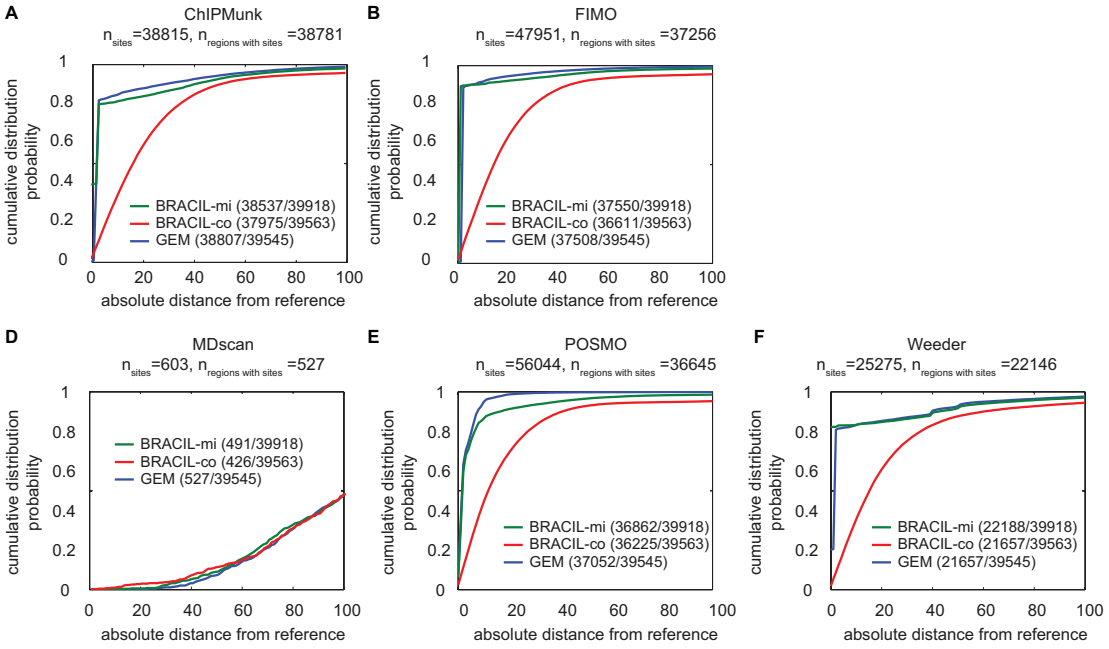


Figure S18: Same as Figure S17, but for the entire set of CTCF enriched regions. GEM shows an overall better performance than BRACIL when all enriched regions are used for evaluation. The entire CTCF set includes regions of spurious coverage (Figure S11) in which the deconvolution step of BRACIL is not expected to perform well.

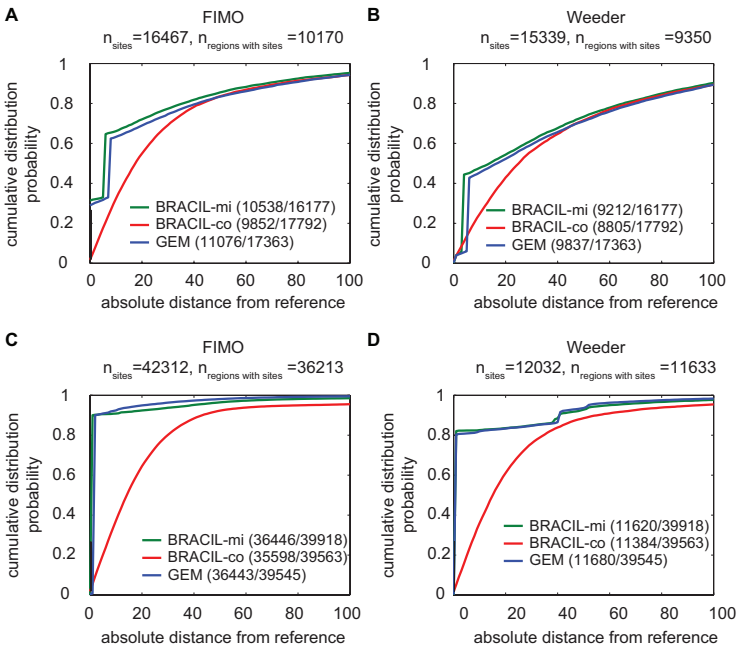


Figure S19: The qualitative results of the resolution analysis presented in Figures 17 and 18 do not change by conservative threshold for motif discovery. The reference set of binding sites for FIMO and Weeder depends on a motif threshold. This figure shows results for more conservative results. We used a p-value< $10^{-4}$ for FIMO and 85% match for Weeder. GABPA cases are presented in top panels (A, B) and CTCF case in the bottom ones (C, D). Figure follows same representation of Figures S17-S18.

Table S1: Effect of penalty parameter in binding site detection. Penalty is scaled to vary from 0 to 1, where 0 means no penalty and 1 means a penalty proportional to the sum of squares of coverage.[a]

| | Single binding only | | | | Single and double binding | | | |
|---|---|---|---|---|---|---|---|---|
| **Penalty** | **TP** | **FP** | **$AUC_{ROC}$** | **$AUC_{P\&R}$** | **TP** | **FP** | **$AUC_{ROC}$** | **$AUC_{P\&R}$** |
| 0.01 | 40 | 4 | 0.8941 | 0.9440 | 44 | 6 | 0.9397 | 0.9683 |
| 0.02 | 38 | 2 | 0.8774 | 0.9378 | 44 | 5 | 0.9420 | 0.9702 |
| 0.05 | 34 | 1 | 0.8376 | 0.9199 | 42 | 5 | 0.9182 | 0.9575 |
| 0.10 | 32 | 1 | 0.8159 | 0.9087 | 42 | 5 | 0.9182 | 0.9575 |
| 0.15 | 31 | 1 | 0.8050 | 0.9031 | 42 | 5 | 0.9182 | 0.9575 |
| 0.20 | 31 | 1 | 0.8050 | 0.9031 | 42 | 5 | 0.9182 | 0.9575 |
| 0.30 | 31 | 1 | 0.8050 | 0.9031 | 42 | 5 | 0.9182 | 0.9575 |
| 0.50 | 30 | 1 | 0.7941 | 0.8975 | 41 | 5 | 0.9063 | 0.9510 |
| 1.00 | 30 | 1 | 0.7941 | 0.8975 | 41 | 5 | 0.9063 | 0.9510 |

[a]TP, FP represent the number of true and false positives, respectively. $AUC_{ROC}$ represents the area under a ROC curve and $AUC_{P\&R}$ represent the area under a precision and recall curve.

Table S2: List of binding configuration probabilities used to create a simulated set of enriched regions. This set is chosen to be representative of the feasible space of independent binding (see Figure S3b).

| $P_{0,0}$ | $P_{1,0}$ | $P_{0,1}$ | $P_{1,1}$ |
|---|---|---|---|
| 0.09 | 0.10 | 0.3837 | 0.4263 |
| 0.09 | 0.01 | 0.8100 | 0.0900 |
| 0.09 | 0.001 | 0.8990 | 0.0100 |
| 0.19 | 0.10 | 0.4652 | 0.2448 |
| 0.19 | 0.01 | 0.7600 | 0.0400 |
| 0.19 | 0.001 | 0.8048 | 0.0042 |
| 0.29 | 0.10 | 0.4536 | 0.1564 |
| 0.29 | 0.01 | 0.6767 | 0.0233 |
| 0.29 | 0.001 | 0.7066 | 0.0024 |
| 0.39 | 0.10 | 0.4059 | 0.1041 |
| 0.39 | 0.01 | 0.5850 | 0.0150 |
| 0.39 | 0.001 | 0.6074 | 0.0016 |
| 0.49 | 0.10 | 0.3405 | 0.0695 |
| 0.49 | 0.01 | 0.4900 | 0.0100 |
| 0.49 | 0.001 | 0.5080 | 0.0010 |
| 0.59 | 0.10 | 0.2651 | 0.0449 |
| 0.59 | 0.01 | 0.3933 | 0.0067 |
| 0.59 | 0.001 | 0.4083 | 0.0007 |
| 0.69 | 0.10 | 0.1834 | 0.0266 |
| 0.69 | 0.01 | 0.2957 | 0.0043 |
| 0.69 | 0.001 | 0.3086 | 0.0004 |
| 0.79 | 0.10 | 0.0976 | 0.0124 |
| 0.79 | 0.01 | 0.1975 | 0.0025 |
| 0.79 | 0.001 | 0.2087 | 0.0003 |
| 0.89 | 0.10 | 0.0090 | 0.0010 |
| 0.89 | 0.01 | 0.0989 | 0.0011 |
| 0.89 | 0.001 | 0.1089 | 0.0001 |

Table S3: This table summarizes the key features in BRACIL that was borrowed from *csdeconv* and also highlights new ones that is introduced in BRACIL.

| Feature | BRACIL | Csdeconv |
|---|---|---|
| Blind-deconvolution model | Yes | Yes |
| Penalty parameter to avoid overfitting | Yes | Yes |
| Predicts multiple binding sites based only in ChIP-seq coverage | Yes | Yes |
| Versatile with any peak-caller | Yes | No |
| Parametric impulse response | Yes | No |
| Physical interpretation of the impulse response | Yes | No |
| Integrated with motif discovery | Yes | No |
| Exploit weak motifs | Yes | No |
| Single-nucleotide resolution | Yes | No |
| Parallel deconvolution | Yes | No |
| Feasible to high-throughput and eukaryote application | Yes | No |
| Double-binding signal | Yes | No |
| Predicts cooperative interaction | Yes | No |

Table S4: BRACIL has a lower false negative rate than GEM. The false negative rate is computed as the number of reference binding sites that are not identified by BRACIL or GEM. Results are shown for both GABPA (Valouev et al. 2008) and CTCF (Chen et al. 2008) and specific per motif discovery tool. The predictive methods are BRACIL-MI (BRACIL with motif input), BRACIL-co (BRACIL using coverage only), and GEM. The motif discovery tools used to obtain the reference binding sites are listed in the column at the left. The numbers used for computing false negative rate are taken from Figures S15 and S16.

| Motif discovery tool | False Negative Rate | | | | | |
|---|---|---|---|---|---|---|
| | GABPA | | | CTCF | | |
| | BRACIL-mi | BRACIL-co | GEM | BRACIL-mi | BRACIL-co | GEM |
| ChIPMunk | 0 | 0.002 | 0 | 0 | 0 | 0 |
| FIMO | 0.722 | 0.805 | 0.726 | 0.493 | 0.558 | 0.528 |
| HMS | 0.225 | 0.306 | 0.239 | 0.317 | 0.349 | 0.324 |
| MDScan | 0.160 | 0.314 | 0.216 | 0.023 | 0.085 | 0.069 |
| POSMO | 0.148 | 0.270 | 0.219 | 0 | 0 | 0 |
| Weeder | 0.723 | 0.791 | 0.727 | 0.303 | 0.371 | 0.333 |

## References:


Bailey TL, Elkan C. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings / International Conference on Intelligent Systems for Molecular Biology ; ISMB International Conference on Intelligent Systems for Molecular Biology* **2**: 28-36.

Chauhan S, Sharma D, Singh A, Surolia A, Tyagi JS. 2011. Comprehensive insights into Mycobacterium tuberculosis DevR (DosR) regulon activation switch. *Nucleic acids research*.

Chen X, Xu H, Yuan P, Fang F, Huss M, Vega V, Wong E, Orlov Y, Zhang W, Jiang J et al. 2008. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* **133**(6): 1106-1117.

Fukudome K, Yamaoka K, Nishikori K, Takahashi T, Yamamoto O. 1986. Ultrasonic scission of deoxyribonucleic acid in aqueous solution. I. Conditions for sonication and molecular weights of sonicated samples. *Polymer Journal* **18**(1): 71-79.

Galagan JE, Minch K, Peterson M, Lyubetskaya A, Azizi E, Sweet L, Gomes A, Rustad T, Dolganov G, Glotova I et al. 2013. The Mycobacterium tuberculosis regulatory network and hypoxia. *Nature*.

Guo Y, Mahony S, Gifford DK. 2012. High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints. *PLoS computational biology* **8**: e1002638.

Hu M, Yu J, Taylor JMG, Chinnaiyan AM, Qin ZS. 2010. On the detection and refinement of transcription factor binding sites using ChIP-Seq data. *Nucleic acids research* **38**: 2154-2167.

Kulakovskiy IV, Boeva VA, Favorov AV, Makeev VJ. 2010. Deep and wide digging for binding motifs in ChIP-Seq data. *Bioinformatics (Oxford, England)* **26**: 2622-2623.

Liu XS, Brutlag DL, Liu JS. 2002. An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nature biotechnology* **20**: 835-839.

Ma X, Kulkarni A, Zhang Z, Xuan Z, Serfling R, Zhang MQ. 2012. A highly efficient and effective motif discovery method for ChIP-seq/ChIP-chip data using positional information. *Nucleic acids research* **40**: e50.

Maerkl S, Quake S. 2007. A Systems Approach to Measuring the Binding Energy Landscapes of Transcription Factors. *Science* **315**(5809): 233-237.

Pavesi G, Mauri G, Pesole G. 2001. An algorithm for finding signals of unknown length in DNA sequences. *Bioinformatics* **17**: S207-S214.

Pique-Regi R, Degner JF, Pai AA, Gaffney DJ, Gilad Y, Pritchard JK. 2011. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome research* **21**: 447-455.

Rye M, Sætrom P, Drabløs F. 2011. A manually curated ChIP-seq benchmark demonstrates room for improvement in current peak-finder programs. *Nucleic Acids Research* **39**(4): e25-e25.
Valouev A, Johnson D, Sundquist A, Medina C, Anton E, Batzoglou S, Myers R, Sidow A. 2008. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nature Methods* **5**(9): 829-834.
Zhao Y, Granas D, Stormo G. 2009. Inferring Binding Energies from Selected Binding Sites. *PLoS Comput Biol* **5**(12): e1000590.