The first round of review:

Responses to the comments of the reviewer #1:

1. The reviewer said that in its objective function Bipad minimizes Shannon information which is identical to entropy.

Our response:

Shannon information is not equal to entropy. The Bipad algorithm actually minimizes the entropy of the multiple alignment from a ChIP-seq dataset; alternatively, it maximizes Shannon information content of the multiple alignment.

The entropy of a multiple alignment is the sum of the entropy of all the individual positions; the entropy E_l of a position l is computed using the following Equation 1 (6).

$$E_l = \sum_{b \in B} f(b,l) \log_2 \frac{1}{f(b,l)}\ , \qquad B = \{A, C, G, T\} \quad [1]$$

where $f(b,l)$ is the frequency of base b at position l.

The information content of a multiple alignment is the sum of the information contents of all the individual positions; the information content $R_{sequence}(l)$ of the position l is computed using the following Equation 2 (6).

$$R_{sequence}(l) = 2 - \sum_{b \in B} f(b,l) \log_2 \frac{1}{f(b,l)} = 2 - E_l, \qquad B = \{A, C, G, T\} \quad [2]$$

Therefore, $R_{sequence}(l)$ will increase as E_l decreases. The multiple alignment with the minimum entropy will have the maximum information content.

2. The reviewer said that Shannon entropy does not use pseudocounts, and is the objective function used in MEME.

Our response:

1. The statement, ‘Shannon entropy does not use pseudocounts’, is wrong. In the Maskminent/Bipad algorithm, a pseudocount β_b is used when computing the frequency of each base b at the position l using the following Equation 3 (7). Each pseudocount β_b is set to 0.25 × 1.5 (7, 8).

$$f(b,l) = \frac{n(b) + \beta_b}{N + \sum_{b \in B} \beta_b}, \qquad B = \{A, C, G, T\} \quad [3]$$

where $n(b)$ is the number of base b at position l, N is the number of DNA sequences in the alignment.

2. The statement, ‘Shannon entropy is up to an additive constant the same as the log likelihood under a multinomial statistical model, the standard objective function in many motif discovery algorithms such as MEME.’, is wrong. The objective function of the Maskminent algorithm based on Shannon information theory differs from that of the MEME algorithm based

on probability theory. In summary, the difference is that the core principle of Maskminent's objective function is Entropy Minimization which seeks the multiple alignment with the minimum entropy in the entire multiple alignment search space formed by a ChIP-seq dataset, whereas the core principle of MEME's objective function is Expectation Maximization which is used to fit a two-compoment (motif and background DNA) finite mixture model to a ChIP-seq dataset (9).

The objective function of Maskminent is given below:

$$oMA = \ arg \min_{MA\in\theta} \left\{\sum_{m\in\{L,R\}} \left(\sum_{l=1}^{J_m} \left(\sum_{b\in B} f_m(b,l) \log_2 \frac{1}{f_m(b,l)}\right)\right)\right\},\ B = \{A,C,G,T\} \quad [4]$$

where *MA* is one multiple alignment, *oMA* is the globally optimal multiple alignment, *θ* is the entire multiple alignment space formed by all the peaks in the ChIP-seq, *L* and *R* are the left and right half sites respectively, *J* is the length of one half site, *$f_m(b,l)$* is the frequency of the base *x* appearing at the position *l* of the half site *m*.

The MEME algorithm views the ChIP-seq dataset as a mixture of two components (the motif model and background model), but the way of this mixture is unknown (i.e. the parameters in the two models and the mixing parameters are unknown). The dataset is broken up conceptually into all overlapping subsequences of motif length which it contains (9). Therefore, the goal of this algorithm is to discriminate motif DNA from background DNA in a mathematically optimal way (i.e. to search for maximum likelihood estimates of the parameters of a finite mixture model which could have generated the dataset (9)).

The objective function of the E-step of the Expectation Maximization algorithm used by MEME is the following: (9)

$$\mathrm{E}[\log L(\theta,\lambda|X,Z)] = \sum_{i=1}^{n}\sum_{j=1}^{2} Z_{ij}^{(0)} \log\left(p\left(X_i\middle|\theta_j\right)\lambda_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{2} Z_{ij}^{(0)} \log p\left(X_i\middle|\theta_j\right) + \sum_{i=1}^{n}\sum_{j=1}^{2} Z_{ij}^{(0)} \log \lambda_j \quad [5]$$

where $X = (X_1, X_2, \cdots, X_n)$ which represents the ChIP-seq dataset containing n peaks, $\theta = (\theta_1, \theta_2)$ which are the parameters in the motif model and background model, $\lambda = (\lambda_1, \lambda_2)$ which are the mixing parameters (i.e. the probabilities of the two models), Z_i represents which model each subsequence is generated from.

The M-step of the Expectation Maximization algorithm maximizes Equation 5 over θ and λ to find the next estimates for them (e.g. θ^1 and λ^1). The maximization over λ only involves the second term in Equation 2, and the maximization over θ only involves the first term in Equation 2.

$$\lambda_j^{(1)} = arg \max_{\lambda} \sum_{i=1}^{n}\sum_{j=1}^{2} Z_{ij}^{(0)} \log \lambda_j \ [6],\ \ \theta_j^{(1)} = arg \max_{\theta_j} \sum_{i=1}^{n} Z_{ij}^{(0)} \log p\left(X_i\middle|\theta_j\right) \ [7]$$

Maskminent does not require a background component (ie. it is not based on Kullback–Leibler divergence but rather Shannon entropy), which is why it can be used to determine the affinity of a binding site to its cognate TF (see reference 9 of the manuscript). While MEME is capable of detecting secondary motifs, the algorithm's authors [9] indicate that application of objective function to detect secondary motifs generally has much lower likelihoods than for the primary motif. We believe that this may explain why Maskminent can sometimes detect motifs that MEME misses.

3. The reviewer suggested that the previous title of the paper, “Recursive, thresholded entropy minimization”, is inappropriate for the Maskminent pipeline, in that it only attempts to determine the optimal number of ChIP-seq peaks.

Our response:

‘Recursive, thresholded entropy minimization’ is an appropriate term for the paper, because it accurately and aptly describes the characteristics of the Maskminent motif discovery pipeline. We retain this description in the paper, however we have simplified the title of the article to one which should be more accessible for NAR readers.

The foundation of this pipeline is the original Bipad algorithm based on entropy minimization. However, our systematic approach for cofactor and novel motif discovery required the novel recursion and thresholding elements that were introduced in this study. ‘Recursive’ refers to the fact that on a ChIP-seq dataset the masking and thresholding techniques can be used iteratively. ‘Thresholded’ refers to the functionality of thresholding the dataset to take top peaks ranked by signal strength.

4. The reviewer said that we did not describe the approach to seek the optimal number of ChIP-seq peaks.

Our response:

In the initial submission, the approach for determining the number of top ranked peaks that exhibit the desired motif was described in the paragraph starting with ‘Thresholding the datasets to eliminate peaks…’ and in the Supplementary Figure S1.

In this resubmission, it is described in more detail in the paragraph starting with ‘To eliminate noisy patterns that…’, and in the flowchart of the Maskminent pipeline (Section 2.1, Page 3) in the Supplementary Methods file.

5. The reviewer said that the descriptions in the Methods section was unrelated to the pseudocode in Supplementary Figure 1 (which was present in the initial submission).

Our response:

In the initial submission, the explanations in the method section and the pseudocode of the algorithm in Supplementary Figure S1 were completely consistent.

In this resubmission, the pseudocode was converted to a flowchart describing the Maskminent pipeline in the Supplementary Methods file.

6. The reviewer said that we did not indicate if the lengths of the half sites, the size of the gap range, the number of motifs to be returned are input by the use or automatically generated by the Maskminent software.

Our response:

In the initial submission, the self-explanatory arguments in the two commands running the Maskminent program in the Methods section, which are shown below, indicated that the length of each half site and the range of the gap length are provided by the user. We empirically test a range of motif lengths and gap ranges based on previously published evidence for each primary TF. Supplementary Figure S1 indicates that the number of motifs to be derived from a ChIP-seq dataset is determined by the Maskminent pipeline by indicating when the pipeline will stop on the dataset.

```
./Maskminent –n LengthOfInfoModel –y NumberOfCycles –f ChIPseqFile [-m MaskFile] [-2]
./Maskminent –l LengthOfLeftHalfSite –r LengthOfRightHalfSite –a MinSpacerSize –b
MaxSpacerSize  -y NumberOfCycles –f ChIPseqFile –d/i
```

In this resubmission, the two commands appear in the Supplementary Methods file (Section 2.2, Page 4). The flowchart of the Maskminent pipeline in the Supplementary Methods replaces the Supplementary Figure S1 in the Supplementary Methods file.

7. The reviewer said that the approach to calculate dissociation constants is incorrect ("A default value 1E-7 M of the dissociation constant K_{d1} was approximated in instances where the exact value for a particular TF could not be established from published studies on TF-binding site measurements.").

Our response:

1. This approach of calculating dissociation constants is correct, and there is precedent for the low intracellular concentration of many TFs (10). In the initial submission, Equation 8 is used to calculate the dissociation constant of a binding site sequence.

$$K_{di} \approx \frac{F_1^{'}(K_{d1} + [Tf])}{F_i^{'}} - [Tf] \quad [8]$$

where K_{di} is the dissociation constant of the ith predicted binding site, $F_i^{'}$ is the frequency of the ith site in a round of bounding (i.e. the frequency of the ith site appearing in the ChIP-seq dataset), K_{d1} and $F_1^{'}$ are these values for the strongest site (i.e. the consensus sequence), $[Tf]$ is the concentration of the unbound TF.

Equation 8 was derived by recognizing that the thermodynamics of a population of TF-bound sequences is similar to a SELEX experimental framework (11). Below the detailed derivation is given.

From Levine et al. (11), first we obtain 9:

$$\frac{F_i^{'}}{F_1^{'}} = \frac{K_{d1} + [Tf]}{K_{di} + [Tf]} \cdot \frac{F_i}{F_1} \quad [9]$$

where F_i is the frequency of the ith site in the prior round of bounding (i.e. in the genome), F_1 is this value for the strongest site (i.e. the consensus sequence).

Multiply both sides of Equation 9 by $F_1^{'}$, then we obtain Equation 10:

$$F_i^{'} = \frac{K_{d1} + [Tf]}{K_{di} + [Tf]} \cdot \frac{F_i}{F_1} \cdot F_1^{'} \quad [10]$$

Given that the consensus sequence is an extremely infrequent binding site both in the unselected population of binding sites and in the genome (12), we assume that its frequency will be similar during the early rounds of selection (i.e. $F_1 \approx F_1^{'}$). Then, we obtain Equation 11:

$$F_i^{'} \approx \frac{K_{d1} + [Tf]}{K_{di} + [Tf]} \cdot F_1^{'} \quad [11]$$

Solving for K_{di}, then we obtain Equation 8. $[Tf]$ is negligible for most TFs, because the steady-state concentrations of most TFs inside cells are quite low due to their high turnover rates (i.e. in the nM range), and $[Tf]$ is only a fraction of them (10). Therefore, we obtain Equation 12 which suggests that the dissociation constants of binding sites are inversely proportional to their frequencies.

$$K_{di} \approx \frac{F_1^{'} K_{d1}}{F_i^{'}} \quad [12]$$

Take the logarithm of both sides, then we obtain Equation 13 which suggests that the binding energy ($\log_2 K_{di}$) of binding sites is related to their frequencies.

$$\log_2 K_{di} \approx \log_2 \frac{F_1^{'} K_{d1}}{F_i^{'}} \quad [13]$$

Additionally, the R_i values of binding sites are proportionate to their frequencies, since the stronger a binding site is, the more the number of times it is bound by the TF will be in a ChIP-seq assay. Therefore, the binding energy ($\log_2 K_{di}$) of binding sites is related to their R_i values. It is expected that there exists a linearity between them when plotting $\log_2 K_{di}$ versus R_i.

2. In this resubmission for all the TFs we used 10^{-7}M as the default value for the dissociation constant K_{d1}, which is correct. This is because on an iPWM using different values for K_{d1} will lead to the same F-test value. The detailed proof is given below.

Assume that for K_{d1} we use two different values K_{d1}^a and K_{d1}^b. According to Equation 12, we have

$$K_{di}^a \approx \frac{F_1^{'} K_{d1}^a}{F_i^{'}} \quad [14], \qquad K_{di}^b \approx \frac{F_1^{'} K_{d1}^b}{F_i^{'}} \quad [15]$$

Combining Equations 14 and 15, we obtain

$$K_{di}^b = \frac{K_{d1}^b}{K_{d1}^a} K_{di}^a \quad [16]$$

Take the logarithm of both sides, then we obtain

$$\log_2 K_{di}^b = \log_2 K_{di}^a + \log_2 \frac{K_{d1}^b}{K_{d1}^a} \quad [17]$$

Equation 17 implies that in the graph of R_i (X axis) versus binding energy (Y axis), the Y-axis values of all the data points will change the same amount ($\log_2 \frac{K_{d1}^b}{K_{d1}^a}$) with the X-axis values remaining unchanged when K_{d1}^{a} and K_{d1}^{b} are used. This implies that the residual sum of squares (RSS) of the linear fitting model does not change when K_{d1}^{a} and K_{d1}^{b} are used, and the RSS of the constant fitting model does not change either. According to the following Equation 18 computing the F statistic, the resultant F-test value does not change when the two different values K_{d1}^{a} and K_{d1}^{b} are used.

$$F = \frac{\left(\frac{RSS_1 - RSS_2}{p_2 - p_1}\right)}{\left(\frac{RSS_2}{n - p_2}\right)} \quad [18]$$

where RSS_1 is the RSS of the linear model, RSS_2 is the RSS of the constant model, n is the number of data points (i.e. the number of binding sites in the iPWM), p_2 is the freedom of the linear model (i.e. 2), p_1 is the freedom of the constant model (i.e. 1).

8. The reviewer said that we used poor-quality ChIP-seq datasets.

Our response:

In fact, the datasets used to derive our iPWMs were of the highest overall quality available. We used the initial and the IDR-thresholded ChIP-seq peak datasets released by the ENCODE Consortium (13), which is generally acknowledged to be a gold standard for ChIP-seq data, and the refined datasets generated by the SPP peak calling software and provided by Factorbook (3). There is no *a priori* evidence indicating that the IDR-thresholded and SPP peak called datasets are of low-quality.

9. The reviewer said that the CIS-BP database is obscure in the area and includes the motifs produced by Bipad.

Our response:

1. The CIS-BP (Catalog of Inferred Sequence Binding Preferences) database is a library of transcription factor DNA binding motifs and specificities, which was established in a previous study (Weirauch et al. (1)). We find it perplexing that the reviewer was not aware of this database, since it comprises the largest set of experimentally-validated binding sites available at the time the present study began.

2. This database does not include any motifs generated by Bipad, because its motifs were independently generated from frequencies of bound oligonucleotide, inferred from the PBM (Protein Binding Microarray) technique.

10. About the sentence “Transcription factors positively or negatively interact with the regulatory elements in genes […]” which exists in the initial submission, the reviewer asked us what a negative interaction means.

Our response:

In this resubmission, this sentence has been revised to: ‘Transcription factors interact with regulatory elements in genes to mediate positive or negative regulation of tissue- and stage-specific expression.’

11. About the sentence “NF-Y extensively coassociates with FOS over […] cluster classes […]” which exists in the initial submission, the reviewer asked us what ‘cluster classes’ mean.

Our response:

In this resubmission, this sentence has been rewritten: ‘For instance, NF-Y extensively coassociates with FOS over all chromatin states…’.

12. About the sentence “[…] the dynamic range of oligonucleotides used in the DNA microarray” which exists in the initial submission, the reviewer did not know the meaning of ‘dynamic range’ and speculated that ‘dynamic range’ means fluorescence.

Our response:

In this resubmission, this sentence has been revised to ‘In addition, the set of octamers used in the DNA microarrays may not cover all possible binding site sequences (>8 nt) recovered in the genome from ChIP-seq…’.

13. The reviewer said that we used terms that were not widely known nor defined (e.g. "homogeneous recognition motifs"), and called binding sites of the same length 'multiple sequence alignment'.

Our response:

1. In the resubmission, the term ‘homogeneous binding motifs’ has been revised to ‘contiguous binding motifs’.

2. The statement, ‘referring to binding sites of equal length as 'multiple sequence alignment'’, is wrong. In the initial submission, we did not refer to binding sites of equal length as 'multiple sequence alignment'; instead, we referred to a set of aligned binding sites as a multiple sequence alignment. A multiple sequence alignment (MSA) is a sequence alignment of three or more biological sequences, generally protein, DNA or RNA (14). Therefore, a set of aligned

binding sites is a multiple sequence alignment. In the resubmission, the term 'multiple sequence alignment' no longer exists.

14. About the sentence “The R_sequence value of a model is the mean of the R_i values of all the binding site sequences used to compute the model, and represents the average binding affinity.", the reviewer asked us how we computed the information content of a single binding site.

Our response:

In this resubmission, this sentence has been revised to ‘The individual information content (R_i) of a TF-bound sequence, which represents the affinity of the TF-DNA interaction, is the dot product between the binary matrix of the sequence and an iPWM of the TF.’ Therefore, this sentence indicates that the R_i value of a binding site sequence is calculated by computing the dot product between the sequence and an iPWM (12).

15. About the sentence “[...] the weakest binding sites inferred from ChIP-seq are essentially noise [...].”, the reviewer asked us if weak binding sites are noise, and he did not know the criterion of calling weak binding sites noise.

Our response:

In this resubmission, this sentence has been rewritten: ‘This is necessary, as the sequences contained in the weakest ChIP-seq peaks may contribute noise that can obfuscate the detection of true binding motifs.’

16. About the sentence “The frequencies of binding sites appearing in a ChIP-seq dataset are linearly related to their binding energy ($\log_2 K_d$), which is delineated by equation 3”, the reviewer said that the frequencies of binding sites should be linearly related to dissociation constants K_d.

Our response:

According to Equations 8 and 12, the frequencies of binding sites are neither linearly related to the binding energy ($\log_2 K_d$) nor the dissociation constant K_d.

In this resubmission, this sentence has been revised to ‘The frequencies of binding sites appearing in a ChIP-seq dataset are related to their binding energy ($\log_2 K_d$)’, and appears in the Supplementary Methods file (Section 5.1, Page 6).

17. The reviewer said that for the sentence "[TF] is negligible for most TFs.", there should be a discussion and citation.

Our response:

The discussion and citation have already been given in the response to the Comment 7, and also appear in the Supplementary Methods file (Section 5.1, Page 6) of this resubmission along with the complete derivation on the relationship between frequencies and dissociation constants of binding sites.

Responses to the comments of the reviewer #2:

1. The reviewer wanted us to compare Maskminent with other motif discovery algorithms in the literature (e.g. SeqGL).

Our response:

In this resubmission, the paragraph starting with 'We also compared results produced by the Maskminent pipeline with…' in the Discussion section compares the Maskminent pipeline with other motif discovery tools from two perspectives of revealing primary and cofactor binding motifs. MEME-ChIP (15) was primarily used in this comparison, because it was also extensively applied to top 500 peaks of a large number of ChIP-seq datasets in Wang et al (3). The detailed data are given in the Supplementary Table S7. Below we summarize these results.

①The comparison between Maskminent and MEME-ChIP on the ability to reveal primary motifs: among the 98 sequence-specific TFs investigated by both tools, Maskminent and MEME discovered primary motifs for 80 (~81.6%) and 92 (~93.9%) TFs, respectively.

②The comparison between Maskminent and MEME-ChIP on the ability to reveal cofactor motifs: the cofactor motifs Maskiment found but MEME-ChIP did not are primarily the SP and IRF families (for 42 and 8 primary TFs, respectively). This is because in the process of searching for a motif MEME-ChIP used background nucleotide frequencies which are computed from all the input DNA sequences by default. Thus if a true binding motif is similar to the background frequencies, it will fail to discover this motif, which explains why the GA-rich SP motif and IRF motif were often missed by MEME-ChIP. On the other hand, MEME revealed many more cofactor motifs than Maskminent, though using only top 500 peaks increases the likelihood of those cofactors appearing by chance. This is because MEME-ChIP was configured to report up to 5 motifs and the main goal of Maskminent is to discover primary motifs (i.e. if the initial iPWM derived from a dataset exhibits the primary motif, the masking and thresholding techniques will no longer be used).

③The comparison of five tools (Maskiment, MEME-ChIP, SeqGL (16), HOMER (17), gkm-SVM (18)) on the ability to reveal binding motifs: These five tools were compared on 8 certain datasets from Setty et al. (16). In terms of the total number of cofactor motifs revealed, Maskminent is better than gkm-SVM.

Additionally, one advantage of Maskminent over these other tools is that it does not require all the input DNA sequences have the same length.

2. The reviewer wanted us to cite more papers that had been recently published and ensure that all the new TF interactions revealed by Maskminent indeed had not been reported in the literature.

Our response:

1. In this resubmission, the SeqGL paper (16) published in 2015 and Arvey et al. (5) published in 2012 is cited, other than Jolma et al. (2) published in 2015, Weirauch et al. (1) published in 2014 and Kheradpour et al. (4) published in 2014 that were already cited in the initial submission.

2. In this resubmission, we thoroughly checked the literature to ensure that all the new TF interactions that the Maskminent pipeline revealed indeed have not previously been described. The Table 1 and all the relevant texts are revised accordingly.

3. The reviewer told us that the notation 6<1,2>6 was not explained, and that we should add a supplementary figure that describes the Bipad algorithm.

Our response:

1. In this resubmission, a figure intuitively describing the working process of the Bipad algorithm at a high level and in-depth mathematical formalizations are given in the Chapter 1 of the Supplementary Methods file (Section 1, Page 2).

2. In this resubmission, the notation 6<1,2>6 is explained by revising the last sentence of the legend of Figure 2 where it appears for the first time. This sentence has been revised to 'The bipartite search patterns, which are denoted by *l<a,b>r* (*l* and *r* are the lengths of the left and right half sites respectively, *a* and *b* are the minimum and maximum spacer lengths respectively), are 6<0,5>6, 3<2,4>3, 3<2,4>3, 3<2,4>3, 6<1,2>6 and 6<1,2>6 from top to bottom, respectively.'

4. The reviewer asked us in Supplementary Figure S1 how the conditions (e.g. "if (M_j shows the primary binding motif)" and "if(M_j shows the binding motif of a cofactor)") were evaluated and what δ means. He/she also suggested us to convert the pseudocode to a flowchart with an example.

Our response:

1. In this resubmission, the pseudocode describing the Maskminent pipeline in the Supplementary Figure S1 of the initial submission is replaced by a flowchart appearing in the Supplementary Methods file (Section 2.1, Page 3). This flowchart is also accompanied by an example.

2. In this resubmission, δ no longer exists.

3. In this flowchart of the resubmission, we determined if an iPWM shows the primary motif or a cofactor motif by comparing the sequence logo of this iPWM with the sequence logos of TF binding motifs generated by Wang et al. (3) and Weirauch et al. (1).

5. The reviewer wanted us to show that the iPWMs can indeed be used to perform mutation analysis.

Our response:

In this resubmission, in order to demonstrate that the derived iPWMs can indeed be used to perform mutation analysis, we added another method evaluating the accuracy of these iPWMs. This method is to use these iPWMs to explain the effects of characterized SNPs on binding site strengths.

Based on the change in the R_i value of the binding site, the effect of a SNP on the binding site strength can be predicted. For 153 SNPs of 29 TFs within TF binding sites, we compared the predictions of the iPWMs to the experimental observations measuring their effects

The detailed data are included in the Supplementary Table S5. Below we summarize these results.

For only 7 SNPs (~4.6%) of 3 TFs, the directions of changes in binding site strengths predicted by the iPWMs differ from those observed in experiments (e.g. the iPWM predicts that TF binding will be strengthened, but experiments observed that binding was weakened). For 16 SNPs (~10.5%) of 10 TFs, the directions are concordant, but the extents differ (e.g. TF binding is predicted to only be weakened, but experiments observed that binding was completely abolished). For 130 SNPs (~85.0%) of 27 TFs, the predictions of the iPWMs and the experimental observations are completely concordant.

6. The reviewer wanted us to use Maskminent to analyze a few datasets on which Arvey et al. mentioned that other motif discovery tools performed poorly.

Our response:

Arvey et al. (5) mentioned that MEME did not yield any significant motifs for >10% of the peaks in the GM12878 datasets of FOS, possibly due to additional sequence specificity provided by unknown cofactors or a higher false-positive rate in the GM12878 ChIP-seq experiments. On the IDR-thresholded dataset of FOS from the GM12878 cell line, the Maskminent pipeline revealed SP motif and NFY motif. The primary AP1 motif was not revealed.

We used the bipartite iPWM of FOS derived from the MCF10A dataset treated with 1µm afimoxifene for 12 hours to scan this dataset. We found that AP1 binding sites also abound in this dataset, which implies that the reason why the primary AP1 motif was not revealed is that the AP1 motif is less conserved than the two cofactor motifs. The cobinding between FOS and the two cofactors (SP and NFY) was validated by the large proportion of peaks having short intersite distances (<20bp) (see the two graphs in Row 10 of the Supplementary Table S3).

7. The reviewer wanted us to analyze if these novel motifs occur more frequently around DNase I hypersensitive sites and specific histone modifications.

Our response:

In order to investigate if the 6 novel motifs are enriched within DNaseI Hypersensitive sites and near H3K4 methylation and H3K27 acetylation histone modifications associated with open chromatin (19), we computed the proportions that the occurrences of these motifs lying within the 4 ENCODE tracks (DNaseI HS track, H3K4me1 track, H3K4me2 track, H3K4me3 track, H3K27ac track) from the respective cell lines account for all the occurrences in the genome.

Specifically, we used the iPWMs of these novel motifs to scan the whole hg19 genome assembly. Because the NM1, NM2 and NM3 motifs were revealed in ChIP-seq datasets of more than one TF, we used the iPWMs of BAF155, NANOG and ESRRA showing these three motifs, respectively. Then, we intersected the resultant intervals with the 4 ENCODE tracks from the respective cell lines. The results are given in the following table.

Novel	ENCODE Track				
motif	DNaseI HS	H3K4me1	H3K4me2	H3K4me3	H3K27ac
NM1	4.50%	17.63%	15.52%	16.23%	11.44%
NM2	7.06%	33.63%	14.39%	9.61%	34.05%
NM3	4.21%	21.19%	16.89%	13.75%	12.25%
NM4	3.18%	N/A*	N/A*	1.04%	2.22%
NM5	2.31%	N/A*	N/A*	1.21%	N/A*
NM6	6.16%	32.37%	13.58%	9.36%	34.10%

N/A*: The track from the specific cell line is unavailable.

These proportions (5%-35%) are consistent with previous reports of binding sites for other TFs (20). The proportions of the occurrences of the NM2 motif lying within the H3K4me1 and H3K27ac tracks are significantly higher than that within the H3K4me2 and H3K4me3 tracks. The same pattern also exists for the NM6 motif. This implies that these two novel motifs are more likely to be functional enhancer elements, based on the fact that H3K4me1 and H3K27ac are the predominant histone modifications deposited at nucleosomes flanking enhancer elements (21). In addition, the proportions of the occurrences of these two novel motifs lying within DNaseI hypersensitive sites are the highest among all the six motifs.

References cited in the response to reviewers/list of changes:

1. Weirauch,M.T., Yang,A., Albu,M., Cote,A.G., Montenegro-Montero,A., Drewe,P., Najafabadi,H.S., Lambert,S.A., Mann,I., Cook,K., *et al.* (2014) Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*, **158**, 1431–1443.

2. Jolma,A., Yin,Y., Nitta,K.R., Dave,K., Popov,A., Taipale,M., Enge,M., Kivioja,T., Morgunova,E. and Taipale,J. (2015) DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature*, **527**, 384–388.

3. Wang,J., Zhuang,J., Iyer,S., Lin,X., Whitfield,T.W., Greven,M.C., Pierce,B.G., Dong,X., Kundaje,A., Cheng,Y., *et al.* (2012) Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res.*, **22**, 1798–1812.

4. Kheradpour,P. and Kellis,M. (2014) Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Res.*, **42**, 2976–2987.

5. Arvey,A., Agius,P., Noble,W.S. and Leslie,C. (2012) Sequence and chromatin determinants of cell-type-specific transcription factor binding. *Genome Res.*, **22**, 1723–1734.

6. Schneider,T.D., Stormo,G.D., Gold,L. and Ehrenfeucht,A. (1986) Information content of binding sites on nucleotide sequences. *J. Mol. Biol.*, **188**, 415–431.

7. Bi,C. and Rogan,P.K. (2004) Bipartite pattern discovery by entropy minimization-based multiple local alignment. *Nucleic Acids Res.*, **32**, 4979–4991.

8. Frith,M.C., Hansen,U., Spouge,J.L. and Weng,Z. (2004) Finding functional sequence elements by multiple local alignment. *Nucleic Acids Res.*, **32**, 189–200.

9. Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol. ISMB Int. Conf. Intell. Syst. Mol. Biol.*, **2**, 28–36.

10. Sokolowski,T.R., Walczak,A.M., Bialek,W. and Tkačik,G. (2016) Extending the dynamic range of transcription factor action by translational regulation. *Phys. Rev. E*, **93**, 22404.

11. Levine,H.A. and Nilsen-Hamilton,M. A mathematical analysis of SELEX. *Comput. Biol. Chem.*, **31**, 11–35.

12. Schneider,T.D. (1997) Information content of individual genetic sequences. *J. Theor. Biol.*, **189**, 427–441.

13. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

14. Wang,L. and Jiang,T. (1994) On the complexity of multiple sequence alignment. *J. Comput. Biol. J. Comput. Mol. Cell Biol.*, **1**, 337–348.

15. Machanick,P. and Bailey,T.L. (2011) MEME-ChIP: motif analysis of large DNA datasets. *Bioinforma. Oxf. Engl.*, **27**, 1696–1697.

16. Setty,M. and Leslie,C.S. (2015) SeqGL Identifies Context-Dependent Binding Signals in Genome-Wide Regulatory Element Maps. *PLoS Comput. Biol.*, **11**, e1004271.

17. Heinz,S., Benner,C., Spann,N., Bertolino,E., Lin,Y.C., Laslo,P., Cheng,J.X., Murre,C., Singh,H. and Glass,C.K. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell*, **38**, 576–589.

18. Ghandi,M., Lee,D., Mohammad-Noori,M. and Beer,M.A. (2014) Enhanced regulatory sequence prediction using gapped k-mer features. *PLoS Comput. Biol.*, **10**, e1003711.

19. Yan,C. and Boyd,D.D. (2006) Histone H3 Acetylation and H3 K4 Methylation Define Distinct Chromatin Regions Permissive for Transgene Expression. *Mol. Cell. Biol.*, **26**, 6357–6371.

20. Rye,M., Sætrom,P., Håndstad,T. and Drabløs,F. (2011) Clustered ChIP-Seq-defined transcription factor binding sites and histone modifications map distinct classes of regulatory elements. *BMC Biol.*, **9**, 80.

21. Calo,E. and Wysocka,J. (2013) Modification of enhancer chromatin: what, how, and why? *Mol. Cell*, **49**, 825–837.

The second round of review:

Responses to the comments of the reviewer #2:

1. About the sentence "the ability of five tools (Maskiment, MEME-ChIP, SeqGL (118), HOMER (119), gkm-SVM (120)) to reveal binding motifs were compared on 8 datasets described by Setty et al.", the reviewer mentioned that Setty et al. used 105 datasets. So he/she asked us why only eight datasets were selected and how they were selected. He/she wanted us to perform a more extensive comparison between Maskminent and other motif discovery tools.

Our response:

In the last submission, to form a comparison between the five tools we chose the eight ChIP-seq datasets included in Table S4 of Setty et al. (1), because we obtained all the binding motifs that HOMER and gkm-SVM revealed on only these eight datasets.

Sheet 1, Sheet 2 and Sheet 3 of Supplementary Table S8 in the last submission have been respectively renamed to Sheet ‘Primary(Maskminent&MEME-ChIP)’, Sheet ‘Cofactor(Maskminent&MEME-ChIP)’, and Sheet ‘All binding motifs’ in this submission in order to let the names of these sheets more effectively describe the contents.

In this submission, we added all the binding motifs that Maskminent, SeqGL (1) and MEME-ChIP (2, 3) revealed on the 105 ChIP-seq datasets analyzed by Setty et al. to Sheet ‘All binding motifs’ of Supplementary Table S8.

To more directly compare the ability of Maskminent, MEME-ChIP, SeqGL and HOMER (4) to reveal primary binding motifs, in Supplementary Table S8 we added the Sheet ‘Primary(Four tools) . This sheet indicates whether each of the four tools revealed the primary motif on each of 59 datasets that belong to sequence-specific TFs among the 105 datasets and were analyzed by the four tools. The results of SeqGL and HOMER were obtained from Table S2 of Setty et al. and the webpage ‘http://cbio.mskcc.org/public/Leslie/SeqGL/chip_results/index.html’. The results of MEME-ChIP were obtained from the Factorbook website created by Wang et al. (3). The numbers of datasets on which Maskminent, MEME-ChIP, SeqGL and HOMER revealed primary motifs are 45, 51, 49 and 47, respectively.

Furthermore, from Sheet ‘All binding motifs’ of Supplementary Table S8 we draw the following conclusions about comparing the ability of Maskminent and SeqGL to reveal cofactor motifs on the 105 datasets. The cofactor motifs that Maskminent discovered that SeqGL failed to discover primarily comprise the SP family, most likely because the SP motif is similar in nucleotide composition to the background sequences used by SeqGL. The detailed results are given in the newly added Sheet ‘Cofactor(Maskminent&SeqGL)’ of Supplementary Table S8. On the other hand, SeqGL revealed many more cofactor motifs than Maskminent. This is because SeqGL discriminatively reports multiple motifs from each dataset, whereas the main objective of Maskminent was originally to derive primary motifs.

In the main text, the above discussion is also included in the paragraph starting with ‘We also compared results…’ in the Discussion section.

2. The reviewer wanted us to perform an analysis computing the precision and recall.

Our response:

① An approach to addressing this issue is to determine false positive detection rates in sequences that are not expected to contain binding sites ($R_i \leq 0$). We determined the null R_i distribution of binding sites for each TF whose primary binding motif was revealed, using a similar method as the one used to determine the null R_i distributions of splice donor and acceptor sites in Rogan et al. (5).

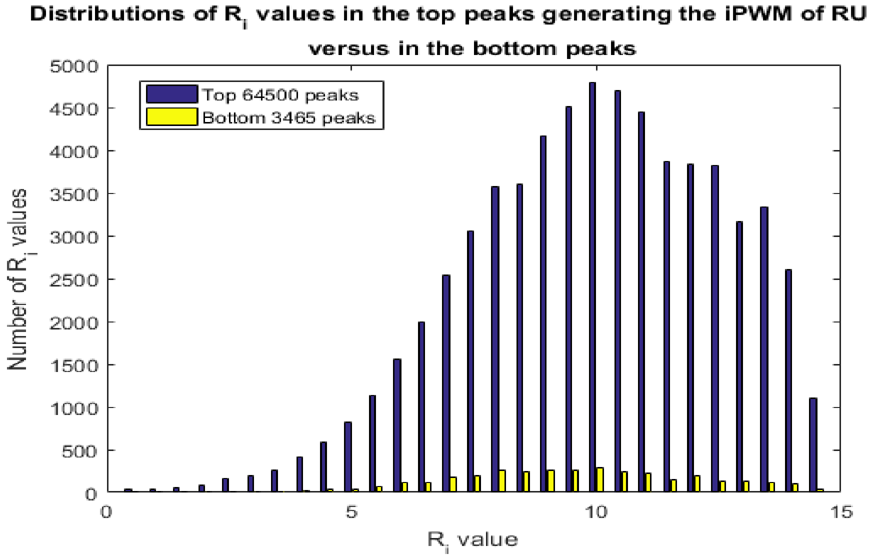
The distribution of R_i values of natural binding sites is approximately Gaussian with $R_{sequence}$ being the mean (5, 6). For each TF, we generated a random 10,000 nucleotide sequence that maintains the same mono- and dinucleotide composition as one of its ChIP-seq datasets. Using the iPWM derived from this dataset, the R_i value of each fragment that is of the same length as the iPWM in the random sequence was computed. The null R_i distribution, which is also approximately Gaussian, was formed by all these R_i values. We also computed the probability of observing a binding site with R_i>0 using the null distribution.

The means of all the 93 null distributions range from -55.91 to -12.28 bits with the standard variations from 7.45 to 22.54 bits. The probabilities range and from 1.82E-4 to 6.35E-2.

The detailed data are given in Sheet ‘iPWMs’ of Supplementary Table S1, and the relevant description was added to the paragraph starting with ‘For each TF ChIP-seq dataset with a derived primary motif…’ in the Results section of the main text.

② Regarding the thresholding technique used in the Maskminent pipeline, we also investigated whether the distribution of R_i values of binding sites in the top peaks used to derive the primary motif is significantly different from those sites present in the excluded bottom peaks.

Specifically, we used the iPWM exhibiting the primary motif to scan each peak in the ChIP-seq dataset, and took the greatest R_i value in each peak. Then, we obtained two sets of R_i values for top peaks and for bottom peaks. The following figure shows the distributions of R_i values for the top 64,500 peaks of the RUNX3 dataset versus for the 3,465 bottom peaks.



The distribution for the bottom peaks completely lies under that for the top peaks, which implies that they are similar. The bottom weak peaks do not necessarily contain weaker sites or are missing binding sites. However, when including these bottom peaks, we obtained a cofactor motif (i.e. IRF motif) instead of the primary motif (i.e. RUNX motif). This implies that overall distribution the binding sites in the top peaks have higher conservation levels (i.e. information contents). Thresholding the dataset is required in order to ensure that the iPWM for the primary motif consists of binding sites from as many peaks as possible, while preventing the cofactor motif from dominating the objective function used in Maskminent.

In the main text, these conclusions were added to the paragraph starting with 'In the Maskminent pipeline, the weak peaks…' in the Discussion section, and the above analysis on the RUNX3 dataset (including the figure) was added to Section 1.2.6 of Supplementary Methods.

3. The reviewer thought that Maskminent used a uniform background nucleotide distribution, which he/she thought might not be correct. So he/she wanted us to talk about it and describe how the use can specify a background distribution or some negative control sequences in Maskminent.

Our response:

① During the process of performing motif discovery, the Maskminent algorithm does not assume any background nucleotide composition or use any negative controls, because it does not require these concepts.

The Maskminent algorithm does not use a discriminative approach to distinguish binding sites from background sequences. Instead, as described in the Supplementary Methods, Maskminent uses an entropy minimization-based Monte Carlo framework to seek the multiple alignment with the minimum entropy in the multiple alignment search space formed by all the peaks of a ChIP-seq dataset. Its objective function is defined as follows:

$$oMA = \arg\min_{MA\in\theta} \sum_{s\in\{L,R\}} \left(\sum_{l=1}^{J_s} \sum_{b\in B} f(b,l) \log_2 \frac{1}{f(b,l)} \right), B = \{A,C,G,T\} \quad [1]$$

where oMA is the optimal multiple alignment with the minimum entropy, θ is the multiple alignment search space, MA is one bipartite multiple alignment in θ, J_s is the length of the left or right half site in MA, $f(b,l)$ is the frequency of base b at position l in MA.

Therefore, given any set of input DNA sequences the user provides, the Maskminent algorithm will always converge to the optimal multiple alignment with the lowest entropy, which is independent of the nucleotide composition of the input sequences. Therefore, the user does not need to provide a background nucleotide composition or a set of negative controls when running Maskminent on a set of input sequences (e.g. a ChIP-seq dataset).

In the main text, the above discussion was summarized in the sentence starting with ‘in contrast, Maskminent does not use…’ of the paragraph starting with ‘We also compared results…’ in the Discussion section.

② During the process of deriving an iPWM from the optimal multiple alignment, when computing the information content $R_{iw}(b,l)$ of base b at position l, using a uniform background nucleotide composition is correct; in fact, it is more appropriate than using a non-uniform background composition.

$R_{iw}(b,l)$ measures the decrease in the surprisal of b at l before the TF specifically binds to binding sites and after the TF specifically binds (6, 7). The surprisal of b at l after the TF binds is computed from the following Equation 2:

$$h_s(b,l) = -\log_2 f(b,l)\,, \quad B = \{A,C,G,T\} \qquad [2]$$

where (b,l) is the frequency of base b at position l in the multiple alignment.

If using a uniform composition (i.e. the probability of each base appearing is 0.25), then the surprisal of b at l before the TF binds is 2. Thus $R_{iw}(b,l)$ is computed from the Equation 3:

$$R_{iw}(b,l) = 2 - h_s(b,l) \quad [3]$$

If using a non-uniform composition (i.e. the actual genomic composition), then the surprisal of b before the TF binds is computed from the following Equation 4:

$$h_g(b) = -\log_2 p(b)\,, \quad B = \{A,C,G,T\} \qquad [4]$$

where $p(b)$ is the probability of base b appearing in the whole genome. Thus $R_{iw}(b,l)$ is computed from the Equation 5:

$$R_{sequence}(L) = h_g(b) - h_s(b,l) \quad [5]$$

Equation 3 describes the molecular machine state in which contact between the TF and binding site is not made before binding (7). Before the TF physically contacts the nucleotide bases of a binding site, the composition of the genome should not matter (i.e. the genomic composition is not relevant to the physical contacts between the TF and the nucleic acid bases) (6). On the other hand, the Equation 5 takes into account the genomic composition (i.e. cancelling the 'background' around a binding site due to genomic composition skew), but this is dangerous because it is not known what causes the skew and whether this skew impacts the binding event. For example, it could be caused by a nucleosome registration pattern throughout in the genome and therefore real information is there which is relevant to TF accessibility. This leaves us with the difficult or unresolvable technical problem to separate and identify the information of other binding sites in such genomes (7). Therefore, using a uniform background composition is more appropriate because it is not affected by confounding genomic structural bias will unknown or ill-defined molecular bases.

In the main text, the above discussion was summarized in the sentence starting with 'This approach is appropriate than …' of the second paragraph in the Introduction section.

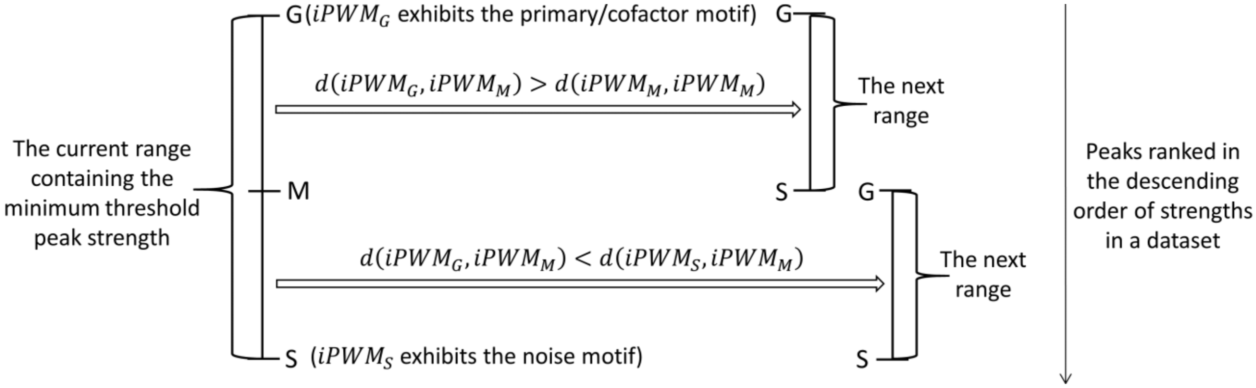
4. The reviewer asked us what the symbols A, B, and C mean in the flowchart of the Maskminent pipeline.

Our response:

In the last submission, the three symbols A, B and C were variables that were declared in order to describe the thresholding process in the flowchart. The initial range containing the maximum number of top peaks that can generate the primary/cofactor motif is from 200 to the number of all peaks (i.e. N in the flowchart). This range is narrowed down by an iterative half-interval search until the number of peaks contained in the range does not exceed 500. The three symbols A, B and C were respectively the smaller bound, the greater bound, and the median (rounding to the nearest multiple of 500) of the current range during the half-interval search.

In this submission, the description about the thresholding process is completely rewritten, so the three symbols A, B and C have been removed. Below the new description in the main text is given.

'To eliminate noisy patterns that suppress the expected TF binding motifs due to ChIP-seq peaks with low signal strengths (i.e. read counts), the dataset is truncated based on signal strengths as follows (Figure 1). First, all the peaks are ranked in the descending order of strengths, and we select the top 200 peaks. If the iPWM derived from the top 200 peaks exhibits the primary/cofactor motif, then the minimum threshold peak strength is contained in the range from the strength of the 200th peak (i.e. the initial value of G) to the strength of the last peak (i.e. the initial value of S). Then, an iterative half-interval search narrows down this range until the number of peaks contained in the range does not exceed 500. During this half-interval search, G is the current threshold above which the top peaks can produce the primary/cofactor motif. Therefore, the approximately minimum threshold that we finally obtain is G of the final range, and the peak set above this threshold contains the maximum number of top peaks that can produce the primary/cofactor motif.'


G($iPWM_G$ exhibits the primary/cofactor motif)
G
$d(iPWM_G, iPWM_M) > d(iPWM_M, iPWM_M)$
The next range
The current range containing the minimum threshold peak strength
M
S
G
Peaks ranked in the descending order of strengths in a dataset
$d(iPWM_G, iPWM_M) < d(iPWM_S, iPWM_M)$
The next range
S ($iPWM_S$ exhibits the noise motif)
S

'Figure 1. One iteration of the half-interval search used to refine the threshold peak strength. All peaks in the dataset are sorted in the descending order of signal strengths. S are the smaller bound of the current range containing the minimum threshold that can generate the primary/cofactor motif, and G is the greater bound (i.e. the current threshold). G and S are respectively initialized to the strength of the 200th peak and the strength of the last peak. M is the strength of the peak at the mean (rounding to the nearest multiple of 500) of the number of top peaks above G and the number of top peaks above S. $iPWM_G$, $iPWM_S$, $iPWM_M$ are respectively the iPWMs derived from the top peaks above G, S, M. $d(iPWM_G, iPWM_M)$ is the Euclidean distance between $iPWM_G$ and $iPWM_M$, and $d(iPWM_S, iPWM_M)$ is the Euclidean distance between $iPWM_S$ and $iPWM_M$. If $d(iPWM_G, iPWM_M)$ is greater than $d(iPWM_S, iPWM_M)$, $iPWM_M$ exhibits the noise motif and the minimum threshold is contained in the subrange from G to M; if $d(iPWM_G, iPWM_M)$ is smaller than $d(iPWM_S, iPWM_M)$, $iPWM_M$ exhibits the primary/cofactor motif and the minimum threshold is contained in the subrange from M to S. When the number of peaks contained in the range does not exceed 500, this half-interval search is stopped. The approximately minimum threshold that is returned is G of the final range.'

In the main text, the above figure was added as Figure 1, and the new description about the thresholding process is included in the paragraph starting with 'To eliminate noisy patterns that…' of the Materials and Methods section and the legend of Figure 1.

In Supplementary Methods, the flowchart was revised accordingly, and the meanings of all symbols appearing in the flowchart were also described in the paragraph starting with '*' immediately after the flowchart. The example after the flowchart was also adjusted accordingly.

Responses to the comments of the reviewer #3:

1. The reviewer wanted us to discuss which binding sites Maskminent failed to detect.
Our response:

In the first method to evaluate the accuracy of our iPWMs, all 803 experimentally proven binding sites for 93 TFs whose primary motifs were discovered were successfully detected by scanning for elements with positive R_i values (Supplementary Table S5).

Not only are the locations of binding sites predicted by the iPWMs completely concordant with those of the true sites determined by experiments, but also the strengths (i.e. R_i values) predicted by the iPWMs are concordant with the experimentally observed strengths. Below, eight typical examples where binding affinity measurements were available are presented. These examples have been extracted from Columns 'Specific evidences' and 'Predicted binding sites' of Supplementary Table S5. In the main text, these conclusions and examples were also summarized in the two sentences starting with 'There was complete concordance…' and 'For example, an EMSA analysis…' in the '*Detection of true binding sites with iPWMs'* subsection of the Results section.

1. Row 206: Dowdy et al. (8) proved that the two blue parts of the sequence 5'-GCTGCTCGGCGCACGGAAGATCCTGTCCCCG-3' in the human SMAD7 promoter are two binding sites of ETV1 using EMSA, and the 'GGAA' site is stronger than the 'ATCC' site. In this sequence our iPWM detected two binding sites (i.e. 5'-GCACGGAAGATC-3' with R_i = 5.30 bits, 5'-GACAGGATCTTC-3' with R_i = 3.73 bits) which respectively contain the two core nucleotide sequences (blue font). And the R_i value (5.30 bits) of the 'GGAA' site is 1.57 bits greater than the R_i value (3.73 bits) of the 'ATCC' site, which means that the 'GGAA' site is $2^{1.57}$ (or 2.97) fold stronger than the 'ATCC' site.

2. Row 294: using EMSA Meirhaeghe et al. (9) proved that GATA2 binds weakly to the sequence 5'-TAGCACTTATCGTTTAAACA-3' in the human PPARG promoter. The iPWM detected the binding site (5'-ACGATAAGT-3' with R_i = 4.00 bits) whose R_i value is smaller than the $R_{sequence}$ value (10.28 bits) of the iPWM, which means that this is a weak binding site.

3. Row 585: the two colored parts of the sequence 5'-TTGGGGAGTCCCAGCCTTGGGGATTCCCCAA-3' in the human HLA-A gene were proven to be two binding sites of NFKB by EMSA in Gobin et al. (10), and the blue site is stronger than the green one. These two binding sites were detected by the GM19099 iPWM (5'-GGGGATTCCC-3' with R_i = 13.70 bits, 5'-GGGGAGTCCC-3' with R_i = 10.02 bits), which means that the blue site is $2^{3.68}$ (or 12.82) fold stronger than the green one.

4. Row 647 and 648: Kozmik et al. (11) proved that in the human CD19 promoter the sequence 5'-CCCCCGCAGACACCCATGGTTGAGTGCCCTCCA-GGCCCCTGCCTG-3' contains a strong binding site of PAX5, and another sequence 5'-CCTG-

GAGAATGGGGCCTGAGGCGTGACCACCGCCTTCCTCTCTGG-3' contains a stronger binding site using EMSA and competition experiments. In the first sequence the iPWM detected the binding site (5'-TGGGGCCTGAGGCGTGAC-3' with R_i = 10.27 bits) whose R_i value exceeds the $R_{sequence}$ value (9.53 bits) of the iPWM, which means that this site is strong; in the second sequence the iPWM detected the binding site with a greater R_i value (5'-AGGGCACTCAACCA-TGGG-3' with R_i = 12.48 bits), which means that this site is $2^{2.21}$ (or 4.63) fold stronger than the one in the first sequence.

5. Row 745, 746 and 747: using EMSA and competition experiments Talianidis et al. (12) proved that in the human APOC3 promoter the binding site of SP1 in the sequence 5'-CACACAGGGTGGGGGCGGGTGGGG-3' is weaker than the two sites in the two sequences 5'-GCCTGGTGGAGGGAGGGGCAA-3' and 5'-GACCAGCTCCTCCCC-CAGGGGA-3'. The R_i value (8.78 bits) of the binding site 5'-ACAGGGTGGGGG-3' detected by the iPWM in the first sequence is smaller than the R_i values (13.83 bits, 13.39 bits) of the two sites 5'-GGAGGGAGGGGC-3' and 5'-TGGGGGAGGAGC-3' detected in the latter two sequences, which means that the site in the first sequence is $2^{5.05}$ (or 33.13) and $2^{4.61}$ (or 24.42) folds weaker than the ones in the latter two sequences, respectively.

6. Row 773, 774 and 775: using EMSA, competition experiments and recombinant human SRF proteins, Miano et al. (13) proved that in the murine Cnn1 promoter the two sequences 5'-ACAGGATTGCCTTAGTTGGGATGAGGTA-3' and 5'-AGCTAAGACCCAAGTTTGGCTTGGAGGG-3' contain two weaker binding sites of SRF than the sequence 5'-GCCGCCGCGCCTTATAAGGCGGCCTTGG-3'. In the first two sequences the iPWM detects two binding sites (5'-CCAACTAAGG-3' with R_i = 10.87 bits, 5'-CCAAACTTGG-3' with R_i = 8.30 bits); in the last sequence it detects the binding site (5'-CCTTATAAGG -3' with R_i = 12.68 bits), which mean that this site is $2^{1.81}$ (or 3.51) and $2^{4.38}$ (or 20.82) fold stronger than the two sites in the first two sequences, respectively.

7. Row 781 and 782: using EMSA and antibody supershift Kordula et al. (14) proved that in the human SERPINA3 promoter the sequence 5'-CCCGTATT-ACCAGAAATTATC-3' contains a stronger binding site of STAT1 than the sequence 5'-TTCCA-GTCCGAGAACAGAA-3'. The bipartite iPWM detected the binding site (5'-TTCTGGTAA-3' with R_i = 9.02 bits) in the first sequence. This site is $2^{2.13}$ (or 4.38) fold stronger than the one (5'-TTCTCGGA-3' with R_i = 6.89 bits) detected in the second sequence.

8. Row 784 and 785: using EMSA and microaffinity DNA binding assays Sherman et al. (15) proved that in the human AGT promoter the binding site of STAT1 in the sequence 5'-CTCCCGTTTCTGGGAACCTTGGC-3' is stronger than the one in the sequence 5'-TGCAAACTTCGGTAAATGTGTAA-3'. In the first sequence the bipartite iPWM detected the binding site (5'-TTCTGGGAA-3' with R_i = 11.24 bits); in the second sequence it detected the

binding site (5'- TTCGGTAA-3' with R_i = 8.89 bits), which means that this site is $2^{2.35}$ (or 5.10) fold weaker than the one in the first sequence.

2. The reviewer said that the mathematical expression makes simple concepts harder to understand.

Our response:

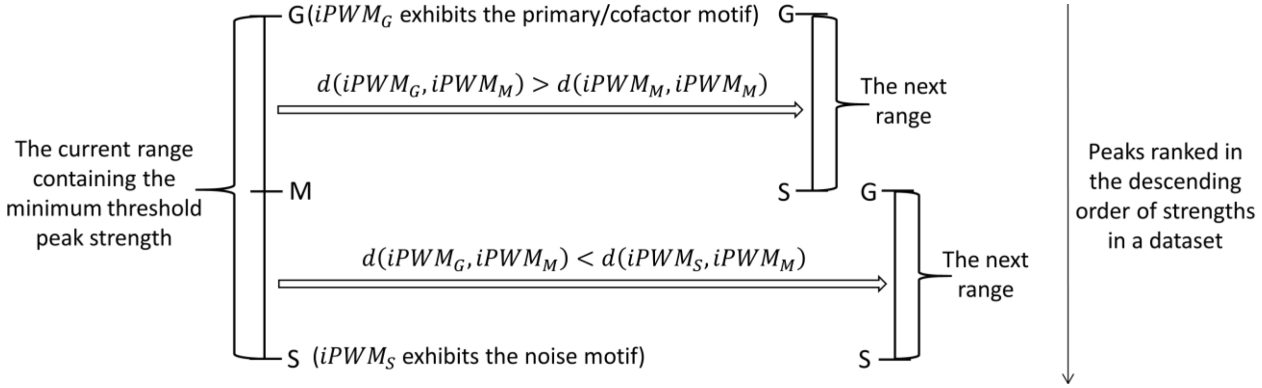
In the last submission, the manuscript itself does not contain mathematical formalism and language, but Supplementary Methods do, which was requested by the corresponding editor. Where possible, we have cited appropriate references to support many of these concepts, which don't need to be repeated in the manuscript or simplified. It would be helpful for the reviewer to indicate which concepts he/she is referring to, and why they should not be described.

3. The reviewer wanted us to further explain how the optimal number of top peaks was determined.

Our response:

In this submission, the description about the thresholding process is completely rewritten. Below the new description in the main text is given.

'To eliminate noisy patterns that suppress the expected TF binding motifs due to ChIP-seq peaks with low signal strengths (i.e. read counts), the dataset is truncated based on signal strengths as follows (Figure 1). First, all the peaks are ranked in the descending order of strengths, and we select the top 200 peaks. If the iPWM derived from the top 200 peaks exhibits the primary/cofactor motif, then the minimum threshold peak strength is contained in the range from the strength of the 200th peak (i.e. the initial value of G) to the strength of the last peak (i.e. the initial value of S). Then, an iterative half-interval search narrows down this range until the number of peaks contained in the range does not exceed 500. During this half-interval search, G is the current threshold above which the top peaks can produce the primary/cofactor motif. Therefore, the approximately minimum threshold that we finally obtain is G of the final range, and the peak set above this threshold contains the maximum number of top peaks that can produce the primary/cofactor motif.'


G($iPWM_G$ exhibits the primary/cofactor motif)
$d(iPWM_G, iPWM_M) > d(iPWM_M, iPWM_M)$
The current range containing the minimum threshold peak strength
M
$d(iPWM_G, iPWM_M) < d(iPWM_S, iPWM_M)$
S ($iPWM_S$ exhibits the noise motif)
G
The next range
S
G
The next range
S
Peaks ranked in the descending order of strengths in a dataset

'Figure 1. One iteration of the half-interval search used to refine the threshold peak strength. All peaks in the dataset are sorted in the descending order of signal strengths. S are the smaller bound of the current range containing the minimum threshold that can generate the primary/cofactor motif, and G is the greater bound (i.e. the current threshold). G and S are respectively initialized to the strength of the 200th peak and the strength of the last peak. M is the strength of the peak at the mean (rounding to the nearest multiple of 500) of the number of top peaks above G and the number of top peaks above S. $iPWM_G$, $iPWM_S$, $iPWM_M$ are respectively the iPWMs derived from the top peaks above G, S, M. $d(iPWM_G, iPWM_M)$ is the Euclidean distance between $iPWM_G$ and $iPWM_M$, and $d(iPWM_S, iPWM_M)$ is the Euclidean distance between $iPWM_S$ and $iPWM_M$. If $d(iPWM_G, iPWM_M)$ is greater than $d(iPWM_S, iPWM_M)$, $iPWM_M$ exhibits the noise motif and the minimum threshold is contained in the subrange from G to M; if $d(iPWM_G, iPWM_M)$ is smaller than $d(iPWM_S, iPWM_M)$, $iPWM_M$ exhibits the primary/cofactor motif and the minimum threshold is contained in the subrange from M to S. When the number of peaks contained in the range does not exceed 500, this half-interval search is stopped. The approximately minimum threshold that is returned is G of the final range.'

In the main text, the above figure was added as Figure 1, and the new description about the thresholding process is included in the paragraph starting with 'To eliminate noisy patterns that…' of the Materials and Methods section and the legend of Figure 1.

In Supplementary Methods, the flowchart was revised accordingly, and the meanings of all symbols appearing in the flowchart were also described in the paragraph starting with '*' immediately after the flowchart. The example after the flowchart was also adjusted accordingly.

4. The reviewer said that the information content of a multiple alignment is the sum of the entropies of all individual positions under the assumption that individual positions are independent of each other.

Our response:

① Under the assumption that all the positions in a binding site are independent from each other, the information content of a contiguous multiple alignment is the sum of the information contents of all the individual positions. The entropy of a multiple alignment is the sum of the entropies of all the individual positions. The information content $R_{sequence}(L)$ and the entropy $E(L)$ of a position L satisfy the following equation 6 (16).

$$R_{sequence}(L) = 2 - E(L) \quad [6]$$

$E(L)$ is computed from the following equation 7 (16).

$$E(L) = -\sum_{b \in B} f(b,l) \log_2 f(b,l), \quad B = \{A, C, G, T\} \qquad [7]$$

In the Equation 6, the component '2' is the maximum entropy that L can have (i.e. the four bases are equally likely to occur) before the TF specifically binds. Therefore, the information content of a position measures the decrease in the entropy of this position after binding.

② To determine the extent of the interdependence between individual positions in binding sites, we computed the total mutual information for one iPWM of each TF whose primary motif was discovered. For contiguous iPWMs, this is done by summing the pairwise mutual information at each position; for bipartite iPWMs, this is done by summing the pairwise mutual information at each position in either half site, then summing the mutual information of the two half sites.

The mutual information MI for a pair of positions L_1 and L_2, which is defined by the following Equation 8, measures the dependence between L_1 and L_2.

$$MI = \sum_{b_1 \in B} \sum_{b_2 \in B} p(b_1, b_2) \log_2 \left(\frac{p(b_1, b_2)}{p(b_1) \cdot p(b_2)} \right) \quad (bits), \quad B = \{A, C, G, T\} \qquad [8]$$

where b_1 is the base appearing at L_1, b_2 is the base appearing at L_2, $p(b_1)$ is the probability of b_1 appearing at L_1, $p(b_2)$ is the probability of b_2 appearing at L_2, $p(b_1, b_2)$ is the joint probability of b_1 and b_2 appearing at L_1 and L_2 simultaneously.

Furthermore, for each iPWM, we computed the percentage of its total mutual information relative to its average information (i.e. the $R_{sequence}$ value). This percentage measures the relative importance of the interdependence across all the individual positions compared to the interaction between the protein and each individual position in a binding event.

For 83 TFs (~89.2%), <10% of the information present in the iPWM is dependent, and for 62 TFs (~66.7%), <5% is dependent. Neglecting the interactions between positions introduces a minimal error into the calculation of R_i values of binding sites, and would be expected to have little impact on assessment of the mutations in these sequences.

In the main text, the above discussion was added to the paragraph starting with 'Similarly, the independence of contributions of each position…' of the Results section. The detailed data were added to Sheet 'iPWMs' of Supplementary Table S1.

5. The reviewer said that the relationship between R_i values and binding energy should be described in the main text.

Our response:

To follow the corresponding editor's request to make the manuscript more accessible to biologists, in the last submission we moved the complete formulation delineating the relationship between R_i values and binding energy to Supplementary Methods. Therefore, in this submission, we revised the paragraph starting with 'To distinguish true binding motifs…' in the Methods section to summarize this relationship as follows:

'To distinguish true binding motifs from noise motifs, we delineated the relationship between R_i values of binding sites discovered by the iPWM and their corresponding binding energy (i.e. higher R_i values have lower binding energies) (Supplementary Methods). Primary/cofactor motifs are expected to demonstrate this relationship, whereas noise motifs are not; that is, for primary/cofactor motifs, the linear regression fit between R_i values and binding energy are expected to have slopes well below 0 which is the expected slope for noise motifs. After applying F-tests to evaluate this relationship, F values for the two categories of motifs were compared using a Mann-Whitney U test.'

6. The reviewer asked us what the criterion of considering weak binding sites as noise was and how it was found.

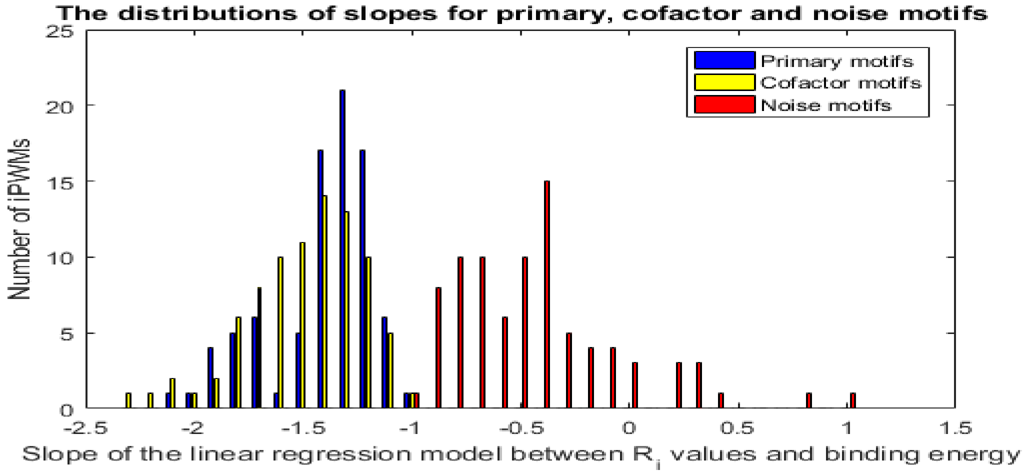
Our response:

① In this submission, the sentence '…enables new motif discovery by recursively masking sequences detected by previous analyses of a ChIP-seq dataset while defining thresholds for inclusion of the maximum number of top peaks to eliminate *weak binding sites contributing noise* (Supplementary Methods).' is revised to '…enables new motif discovery by recursively masking sequences detected by previous analyses of a ChIP-seq dataset while defining thresholds for inclusion of the maximum number of top peaks to eliminate *peaks with lower signal intensities whose inclusion can result in emergence of noise over primary or cofactor motifs* (Supplementary Methods).'

In this submission, the description about the thresholding process is completely rewritten. The new description was given above in the response to Comment 1 of Referee 3.

② As demonstrated in Supplementary Methods, R_i values are related to binding energy (log_2K_d) (i.e. higher R_i values have lower binding energies). Therefore, when plotting log_2K_d versus R_i for a primary/cofactor motif, the slope of the linear regression fit between them is expected to be well below 0. On the other hand, for a noise motif, ideally the slope is expected to be approximately 0 (i.e. the linear regression fit is parallel with the X axis).

The following figure shows the probability density distributions of the slopes for 85 primary motifs, 85 cofactor motifs and 85 noise motifs that were randomly selected. The slopes of all primary motifs are smaller than -1.1 except that one is between -1 and -1.1, which also holds true for cofactor motifs; whereas the slopes of all cofactor motifs are greater than -1 except that one is between -1 and -1.1. This implies that the slope of the linear regression fit between log_2K_d and R_i can be used to distinguish primary/cofactor motifs from noise motifs.



The above discussion was also added to Section 2.3 of Supplementary Methods.

References for second response:

1. Setty,M. and Leslie,C.S. (2015) SeqGL Identifies Context-Dependent Binding Signals in Genome-Wide Regulatory Element Maps. *PLoS Comput. Biol.*, **11**, e1004271.

2. Machanick,P. and Bailey,T.L. (2011) MEME-ChIP: motif analysis of large DNA datasets. *Bioinforma. Oxf. Engl.*, **27**, 1696–1697.

3. Wang,J., Zhuang,J., Iyer,S., Lin,X., Whitfield,T.W., Greven,M.C., Pierce,B.G., Dong,X., Kundaje,A., Cheng,Y., *et al.* (2012) Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res.*, **22**, 1798–1812.

4. Heinz,S., Benner,C., Spann,N., Bertolino,E., Lin,Y.C., Laslo,P., Cheng,J.X., Murre,C., Singh,H. and Glass,C.K. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell*, **38**, 576–589.

5. Rogan,P.K., Faux,B.M. and Schneider,T.D. (1998) Information analysis of human splice site mutations. *Hum. Mutat.*, **12**, 153–171.

6. Schneider,T.D. (1997) Information content of individual genetic sequences. *J. Theor. Biol.*, **189**, 427–441.

7. Schneider,T.D. (1999) Measuring molecular information. *J. Theor. Biol.*, **201**, 87–92.

8. Dowdy,S.C., Mariani,A. and Janknecht,R. (2003) HER2/Neu- and TAK1-mediated up-regulation of the transforming growth factor beta inhibitor Smad7 via the ETS protein ER81. *J. Biol. Chem.*, **278**, 44377–44384.

9. Meirhaeghe,A., Tanck,M.W.T., Fajas,L., Janot,C., Helbecque,N., Cottel,D., Auwerx,J., Amouyel,P. and Dallongeville,J. (2005) Study of a new PPARgamma2 promoter polymorphism and haplotype analysis in a French population. *Mol. Genet. Metab.*, **85**, 140–148.

10. Gobin,S.J., Keijsers,V., van Zutphen,M. and van den Elsen,P.J. (1998) The role of enhancer A in the locus-specific transactivation of classical and nonclassical HLA class I genes by nuclear factor kappa B. *J. Immunol. Baltim. Md 1950*, **161**, 2276–2283.

11. Kozmik,Z., Wang,S., Dörfler,P., Adams,B. and Busslinger,M. (1992) The promoter of the CD19 gene is a target for the B-cell-specific transcription factor BSAP. *Mol. Cell. Biol.*, **12**, 2662–2672.

12. Talianidis,I., Tambakaki,A., Toursounova,J. and Zannis,V.I. (1995) Complex interactions between SP1 bound to multiple distal regulatory sites and HNF-4 bound to the proximal promoter lead to transcriptional activation of liver-specific human APOCIII gene. *Biochemistry (Mosc.)*, **34**, 10298–10309.

13. Miano,J.M., Carlson,M.J., Spencer,J.A. and Misra,R.P. (2000) Serum response factor-dependent regulation of the smooth muscle calponin gene. *J. Biol. Chem.*, **275**, 9814–9822.

14. Kordula,T., Rydel,R.E., Brigham,E.F., Horn,F., Heinrich,P.C. and Travis,J. (1998) Oncostatin M and the interleukin-6 and soluble interleukin-6 receptor complex regulate alpha1-antichymotrypsin expression in human cortical astrocytes. *J. Biol. Chem.*, **273**, 4112–4118.

15. Sherman,C.T. and Brasier,A.R. (2001) Role of signal transducers and activators of transcription 1 and -3 in inducible regulation of the human angiotensinogen gene by interleukin-6. *Mol. Endocrinol. Baltim. Md*, **15**, 441–457.

16. Schneider,T.D., Stormo,G.D., Gold,L. and Ehrenfeucht,A. (1986) Information content of binding sites on nucleotide sequences. *J. Mol. Biol.*, **188**, 415–431.

The third round of review:

Responses to the comments of the reviewer #2:

1. The reviewer objected to the statement that taking into consideration the background nucleotide distribution used as a negative control is "dangerous". He/she wanted us to state that the accuracy of the iPWMs was influenced by only using ChIP-seq sequences containing binding sites. He/she wanted us to generate a null R_i distribution of binding sites for each ChIP-seq dataset as a negative control.

Our response:

We regret using the term “dangerous”, however the citations (10,11) show that the likelihood approach can violate the triangle inequality and can result in >2 bits per nucleotide position, which is unjustified in natural genomes under selection.

We did find that the model’s precision can be affected by use of relative entropy. For example, using a likelihood based method, MEME ChIP fails to detect GC-rich motifs like the Sp family, most likely because the composition is similar to the motif itself.

In our view, the fundamental question is whether the assumption of a uniform background actually affects our results. We have therefore comprehensively determined the probabilities for all of the ChIP-seq datasets, as the reviewer suggested. In the last submission, for one ChIP-seq dataset of each TF whose primary motif was discovered (n=93), we determined the null R_i distribution of binding sites. In this submission, for each ChIP-seq dataset with a derived primary motif (n=367), we created a random 10,000 nucleotide sequence that maintains the same mono- and dinucleotide composition, then determined the null R_i distribution.

The means of all null distributions range from -97.5 to -12.3 bits with standard deviations from 6.9 to 22.5 bits. The probabilities of observing a potentially functional binding site, i.e. with R_i>0, in these sequences range from 1.2E-7 to 0.06. Therefore, for every ChIP-seq dataset from which we derived a primary motif, the false positive detection rate is very low. For all practical purposes, the theoretical argument raised by the reviewer will be of little consequence in our proposed approach.

In the main text, the paragraph starting with ‘For each TF ChIP-seq dataset with a derived primary motif, …’ in the Results section was revised accordingly. The detailed data were also added to Column ‘Binding site null distribution’ of Sheet ‘iPWMs’ of Supplementary Table S1.