SUPPLEMENTARY TEXT

S1. Calculating ρ

Let S_i be the *i*th sequence in the dataset and $S_{i,j}$ be the nucleotide at position *j* in sequence *i* and $Y_{i,j}$ be an indicator variable that equals 1 if the site starts at position *j* in sequence *i*, and 0 otherwise. Define $\rho_{i,j}$ be the estimate after (t) iterations of EM of the probability that the site begins at position *j* in sequence *i*. Using Bayes' rule,

$$\rho_{i,j} = P(Y_{i,j} = 1 | S_i, \theta) = r_{\theta,i}(\hat{\tau}**)\frac{P(S_i | Y_{i,j} = 1, \theta)P^0(Y_{i,j} = 1)}{\sum_{k=1}^{2(L-w+1)} P(S_i | Y_{i,k} = 1, \theta)P^0(Y_{i,k} = 1)}$$

where $r_{\theta,i}(\hat{\tau}**)$ is the number of binding sites in sequence i and $P^0(Y_{i,j} = 1)$ is the prior probability that the motif begins at position j in sequence *i*. Assume $P^0(Y_{i,j} = 1)$ to be uniform,

$$\rho_{i,j} = r_{\theta,i}(\hat{\alpha}**)\frac{P(S_i | Y_{i,j} = 1, \theta)}{\sum_{k=1}^{2(L-w+1)} P(S_i | Y_{i,k} = 1, \theta)}$$

We referred to the above normalization as '`seq-by-seq-anr`'. Note that $\sum_{i=1}^{K_M} r_{\theta,i}(\hat{\tau}**) = R_{M,\theta}(\hat{\tau}**)$ and $\sum_{i=1}^{K_m}\sum_{j=1}^{2(L-w+1)} \rho_{i,j} = R_{M,\theta}(\hat{\tau}**)$.

When $r_{\theta,i}(\hat{\tau}**) = 1$ for all *i*, we referred to the above procedure as '`seq-by-seq-1`'. For a '`global'` normalization procedure such as in MEME,

$$\rho_{i,j} = MAXP\frac{P(S_i | Y_{i,j} = 1, \theta)}{\sum_{i=1}^{K_M}\sum_{k=1}^{2(L-w+1)} P(S_i | Y_{i,k} = 1, \theta)}$$

S2. Additional simulation study

In the simulations described in section 4.2 in the main text, the number of binding sites in each sequence is known. In most *de novo* methods, this number is unknown during model optimization. Our method estimates it. To examine how this estimate affects the normalization result, in the simulations involving 20 known sites, we selected only the 15 the highest scoring sites as the binding sites (not using FDR and misclassifying the other 5 binding sites). We then carried out the `seq-by-seq-anr` normalization by requiring that the sum over all positions in the 20 sequences is 15, rather than 20. We then summed up the probabilities of the 20 known binding sites. The mean of this sum from 1000 simulation is 2.965 (±0.004), which is still larger than that (2.732) from `global` normalization. This result suggests that even counting 25% fewer binding sites our method still resulted in an overall larger probability of the 20 binding sites than the global normalization procedure did.

S3. Additional tests of $R_{M,\theta}(\hat{\tau}^{**})$ estimation on binding site selection

To examine the effect of $R_{M,\theta}(\hat{\tau}^{**})$ estimation on the result for p53 ChIP data, we repeated fdrMotif analyses on dataset 1 (without added noise) by setting $R_{M,\theta}(\hat{\tau}^{**})$ to 100 and 542, respectively, in the first three iterations for FDR at 2%, 5% and 10%. These initial choices of $R_{M,\theta}(\hat{\tau}^{**})$ had little or no effect on both the number and locations of p53 binding sites identified by fdrMotif (**Table s3**).

S4. Further discussion of FDR

Instead of controlling the exact FDR, our method controls the upper bound of the FDR. Such approach is not new. In fact, it can be shown that our procedure for determining $\hat{\tau}^{**}$ is empirically equivalent to the sequential procedure of Benjamini and Hochberg (1995). Let e_i denote the score for subsequence i in set M in which the scores are sorted in descending order, the p-value for subsequence i can be estimated by $\overline{R}_{B,\theta}(e_i)/m_B$. Thus, the sequential procedure of Benjamini and Hochberg in our case can be described as finding the largest i, satisfying

$$\frac{\overline{R}_{B,\theta}(e_i)}{m_B} \cdot \frac{m}{i} \le \gamma_0$$

It can be seen that the above equation is empirically equivalent to equation 2.5 in the main text.

One advantage of estimating FDR in the subsequence space rather than the sequence space is that the proportion of motif subsequence among all subsequences is very small, that is, m_0 is quite close to m or π_0 is near 1. Therefore, the estimated upper bound on FDR is actually close to the desired FDR.

S5. Test effect of MAXP on MEME results

To test the effect of MAXP on MEME results, we repeated MEME analyses on the original ChIP data by setting MAXP to 700 with all other parameters unchanged. MEME identified 699 binding sites, among which only 483 have the two C’s and G’s at the corresponding positions. We also compared the locations of these binding sites with those by fdrMotif with FDR at 2%, 5% and 10%, respectively. When the MAXP is equal to 542 (the number of sequences in the original set), 81.5%-87.8% binding sites from fdrMotif matched the locations of those from MEME. However, only 71.6%-76.3% binding sites from fdrMotif matched those from MEME with MAXP=700. The logo plot (Crooks *et al.*, 2004) for the MEME result is shown in **Figure** s1E. Clearly, there is noticeable difference between these logos and the logos from fdrMotif and the logo from Wei *et al*. (2006). One might conclude that the choice of MAXP may affect on the result.

SUPPLEMENTARY TABLES

Table s1. Comparison of the number of binding sites found in the simulated sequences (“noise”) between fdrMotif and MEME

	fdrMotif			MEME
adulteration[1]	FDR 2%	FDR 5%	FDR 10%	MAXP=Number of sequences
0%	0	0	0	0
5%	0	2	4	0
10%	1	4	8	2
20%	4	11	19	9
30%	6	12	24	20
40%	6	13	28	22
50%	4	14	29	33

[1] Proportion of background sequences added to ChIP data.

Table s2. Mean and standard deviation of the sum of the probabilities of the 20 binding sites from 1000 experiments

Procedure	Simulation 1	Simulation 2
Global[1]	2.723 (0.003)	2.723 (0.003)
Seq-by-seq-1	2.571 (0.003)	2.723 (0.003)
Seq-by-seq-anr	3.782 (0.005)	2.721 (0.003)

[1]MAXP=20

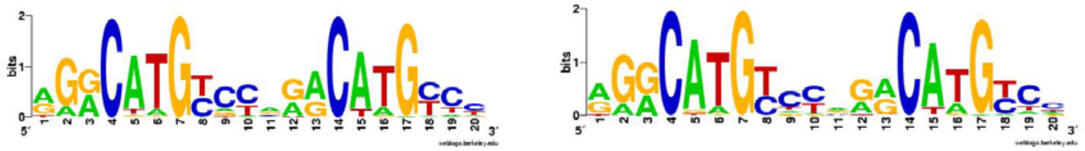
Table s3. Comparison of both the location and the number of binding sites selected with and without constraint of $R_{M,\theta}(\hat{\alpha}^{**})$ in the first three iterations

constraint	FDR=2%		FDR=5%		FDR=10%	
	with/without constraint	Number in common	with/without constraint	Number in common	with/without constraint	Number in common
$R_{M,\theta}(\hat{\alpha}^{**}) = 100$	509/507	507	539/537	534	564/562	560
$R_{M,\theta}(\hat{\alpha}^{**}) = 542$	509/507	507	537/537	537	563/562	562

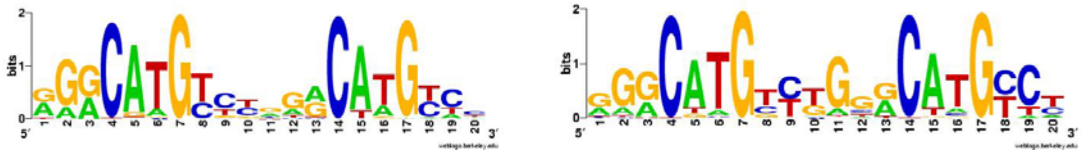
SUPPLEMENTARY FIGURES

Figure s1. Motif logo using all binding sites found in the original p53 ChIP sequences (Wei *et al*., 2006) (A-C) fdrMotif with FDR at 2%, 5% and 10% respectively; (D, E) MEME, MAXP=542 and 700, respectively.

(A) fdrMotif, FDR at 2% (B) fdrMotif, FDR at 5%



(C) fdrMotif, FDR at 10% (D) MEME, MAXP=542



(E) MEME, MAXP=700

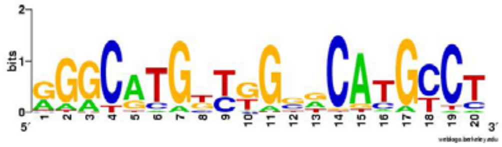
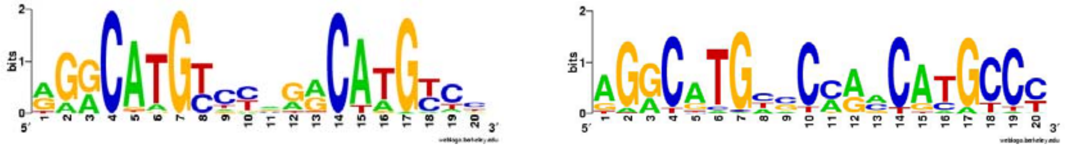
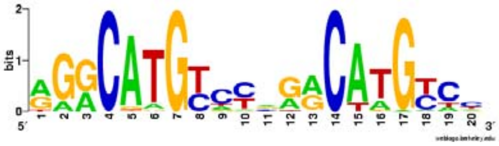
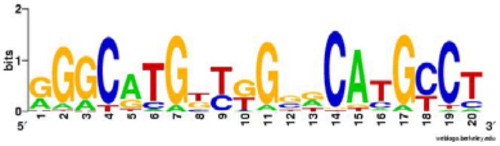


Figure s2. Logo plots of binding sites for p53 ChIP data adulterated with varying proportions of null sequences: (A-F) from fdrMotif when FDR is controlled at 5%, and (G-L) from MEME with MAXP set to the number of sequences in the data.

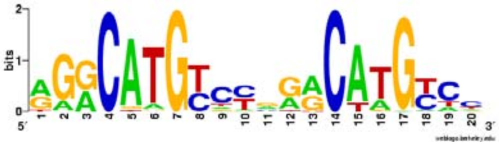
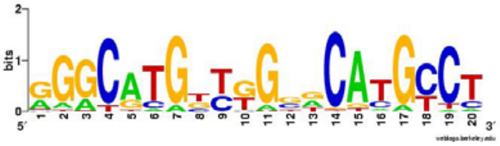
(A) fdrMotif, FDR at 5%, 5% adulteration. (G) MEME, MAXP=570, 5% adulteration.



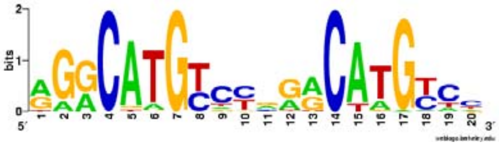
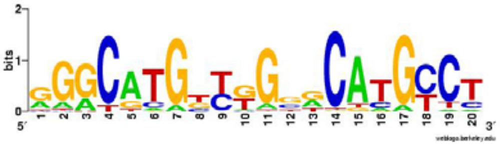
(B) fdrMotif, FDR at 5%, 10% adulteration. (H) MEME, MAXP=602, 10% adulteration.

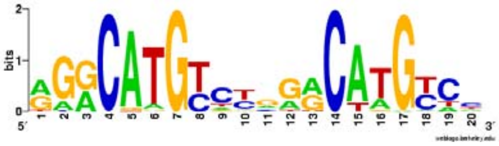
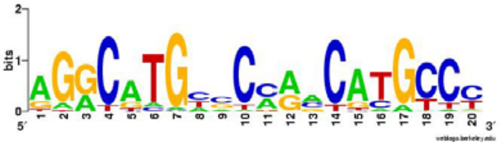
(C) fdrMotif, FDR at 5%, 20% adulteration. (I) MEME, MAXP=677, 20% adulteration.

(D) fdrMotif, FDR at 5%, 30% adulteration. (J) MEME, MAXP=774, 30% adulteration.

(E) fdrMotif, FDR at 5%, 40% adulteration. (K) MEME, MAXP=903, 40% adulteration.

(F) fdrMotif, FDR at 5%, 50% adulteration. (L) MEME, MAXP=1084, 50% adulteration.