Supplementary material for msCentipede: Modeling heterogeneity across genomic sites improves accuracy in the inference of transcription factor binding

Anil Raj [1], Heejung Shim [2], Yoav Gilad [3] Jonathan K. Pritchard [4], Matthew Stephens [5]

[1]Department of Genetics, Stanford University, Stanford, CA, 94305
[2]Department of Human Genetics, University of Chicago, Chicago, IL, 60637
[3]Department of Human Genetics, University of Chicago, Chicago, IL, 60637
[4]Departments of Genetics and Biology, Howard Hughes Medical Institute, Stanford, CA, 94305
[5]Departments of Statistics and Human Genetics, University of Chicago, Chicago, IL, 60637

Supplementary Methods

1 Quantifying heterogeneity in DNase I cleavage patterns

We made use of the connection between the multinomial distribution and binomial distribution to explore the amount of heterogeneity in DNase I cleavage patterns across genomic sites. Specifically, if the read counts per base pair are multinomially distributed conditional on the total read count in a genomic window, then the number of reads mapping to the left half of the window conditional on total read count should be binomially distributed. Based on this, we compared the true distribution of the proportion of reads mapping to the left half of a genomic window to a distribution of proportions computed by sampling read counts from a binomial distribution.

To illustrate, we considered the transcription factor SP1, and focused on the 1000 motif instances with the highest ChIP-seq signal and with at least 10 DNase-seq reads mapped to a 100 bp window around the motif instance. For each window, we computed the 'true' proportion of reads mapping to the left half of the window. Next, assuming that the count on the left half conditional on the total count is binomially distributed, we computed the maximum likelihood estimate of the binomial parameter $\hat{p}$ using data from these windows. Finally, for each window, we sampled a read count for the left half conditional on the total count for that window from a binomial distribution whose parameter was set to be the maximum likelihood estimate $\hat{p}$, and computed a 'simulated' proportion of reads mapping to the left half of the window. Histograms of the 'true' proportions and 'simulated' proportions are shown in Figure 1 in the main text.

2 msCentipede model for one replicate

Consider a genomic window of length L around each of N putative binding sites. We define $X^n = (X_l^n)_{l=1}^L$ for $n = 1, \ldots, N$, where X_l^n is read count at l^{th} base pair in the window around the n^{th} site. Let Z^n denote a binary indicator for whether the n^{th} site is bound ($Z^n = 1$). A mixture model at the n^{th} site can be written as

$$\mathsf{P}(X^n) = \mathsf{P}(X^n|Z^n = 1)\mathsf{P}(Z^n = 1) + \mathsf{P}(X^n|Z^n = 0)\mathsf{P}(Z^n = 0), \tag{1}$$

where the prior probability $\mathsf{P}(Z^n = 1) = \zeta_n$ can be modeled as a logistic function of genomic information (e.g. position weight matrix score and sequence conservation score of the motif instance) as in CENTIPEDE.

At the zeroth scale, we model the total number of reads in the entire region $Y_{01}^n := \sum_l X_l^n$ as follows:

$$Y_{01}^n \sim \text{Poisson}(\lambda^n). \tag{2}$$

At the first scale, conditional on Y_{01}^n, we model the number of reads in the first half of the region $Y_{11}^n := \sum_{l \le L/2} X_l^n$ using a binomial distribution, $Y_{11}^n \mid Y_{01}^n \sim \text{binomial}(Y_{01}^n, p_{01}^n)$. Then, the number of reads in the second half of the region is determined to be $Y_{12}^n := Y_{01}^n - Y_{11}^n$. At the second scale, conditional on Y_{11}^n, we model the read count in the first quarter of the region Y_{21}^n using a binomial distribution, $Y_{21}^n \mid Y_{11}^n \sim \text{binomial}(Y_{11}^n, p_{11}^n)$, leading to the read count in the second quarter of the region $Y_{22}^n := Y_{11}^n - Y_{21}^n$. Conditional on Y_{12}^n, the read count in the third quarter of the region Y_{23}^n is modeled using a binomial distribution, $Y_{23}^n \mid Y_{12}^n \sim \text{binomial}(Y_{12}^n, p_{12}^n)$, leading to the read count in the fourth quarter of the region

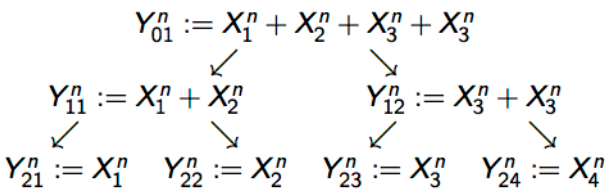

Figure S1: Multi-scale models for inhomogeneous Poisson processes. Suppose we have observations $X^n = (X_l^n)_{l=1}^4$ on four bases. Multi-scale models assume that $Y_{01}^n \sim \text{Poisson}(\mu^n)$, $Y_{11}^n \mid Y_{01}^n \sim \text{binomial}(Y_{01}^n, p_{01}^n)$ (then, $Y_{12}^n = Y_{01}^n - Y_{11}^n$), $Y_{21}^n \mid Y_{11}^n \sim \text{binomial}(Y_{11}^n, p_{11}^n)$ (then, $Y_{22}^n = Y_{11}^n - Y_{21}^n$), and $Y_{23}^n \mid Y_{12}^n \sim \text{binomial}(Y_{12}^n, p_{12}^n)$ (then, $Y_{24}^n = Y_{12}^n - Y_{23}^n$). It follows from elementary properties of the Poisson distribution (see [2] and [3] for details) that it is equivalent to $Y_{2l}^n = X_l^n \sim \text{Poisson}(\mu_l^n)$ for $l = 1, 2, 3, 4$, where $\mu_1^n = \lambda^n p_{01}^n p_{11}^n$, $\mu_2^n = \lambda^n p_{01}^n (1 - p_{11}^n)$, $\mu_3^n = \lambda^n (1 - p_{01}^n) p_{12}^n$, and $\mu_4^n = \lambda^n (1 - p_{01}^n)(1 - p_{12}^n)$.

$Y_{24}^n := Y_{12}^n - Y_{23}^n$. This process continues through the scales: at the J^{th} scale (finest-resolution), there are $L = 2^J$ parts of the region. The read counts in the l^{th} part of the region Y_{Jl}^n, which is equivalent to X_l^n, is modeled as follows:

$$Y_{J,2k-1}^n \mid Y_{J-1,k}^n \sim \text{binomial}(Y_{J-1,k}^n, p_{J-1,k}^n), \tag{3}$$

$$Y_{J,2k}^n = Y_{J-1,k}^n - Y_{J,2k-1}^n, \tag{4}$$

for $k = 1, \ldots, 2^{J-1}$ (see Figure S1 for a brief illustration of the model).

When $Z^n = 1$, heterogeneity across genomic sites is modeled as follows:

$$\lambda^n | Z^n = 1 \sim \text{gamma}\left(\alpha, \alpha/\bar{\lambda}_0\right), \tag{5}$$

$$p_{jk}^n | Z^n = 1 \sim \text{beta}(\bar{p}_{jk}\tau_j, (1 - \bar{p}_{jk})\tau_j), \tag{6}$$

where α and $\bar{\lambda}_0$ capture the mean ($\bar{\lambda}_0$) and variance ($\frac{\bar{\lambda}_0^2}{\alpha}$) in overall DNase I hypersensitivity of TF bound genomic locations, and p_{jk} and τ_j capture the mean and variance in the DNase I cleavage profiles across TF bound genomic locations. When $Z^n = 0$, we model heterogeneity in total DNase I hypersensitivity as follows:

$$\lambda^n | Z^n = 0 \sim \text{gamma}\left(\alpha^o, \alpha^o/\bar{\lambda}_0^o\right), \tag{7}$$

$$p_{jk}^n | Z^n = 0 \sim \delta_{0.5}. \tag{8}$$

3 msCentipede model for multiple replicates

Suppose we have S replicate DNase-seq measurements for a particular cell type. Given a genomic window of length L around each of N putative binding sites, we define $X^{n,s} = (X_l^{n,s})_{l=1}^L$ for $n = 1, \ldots, N$ and $s = 1, \ldots, S$, where $X_l^{n,s}$ is read count at l^{th} base pair in the window around the n^{th} site for the s^{th} replicate.

In the mixture model specified earlier, conditional on $Z^n = 1$, we can model the total number of reads in the entire region $Y_{01}^{n,s} := \sum_l X_l^{n,s}$ as follows:

$$Y_{01}^{n,s} \sim \text{Poisson}(\lambda^{n,s}), \tag{9}$$

$$\lambda^{n,s} | Z^n = 1 \sim \text{gamma}\left(\alpha^s, \alpha^s/\bar{\lambda}_0^s\right), \tag{10}$$

where replicate-specific parameters, α^s and $\bar{\lambda}_0^s$, capture replicate-specific mean ($\bar{\lambda}_0^s$) and variance ($\frac{\bar{\lambda}_0^{s2}}{\alpha^s}$). At the remaining scales, $Y_{jk}^{n,s}$, $k = 1, \ldots, 2^{j-1}$ and $j = 1, \ldots, J$, conditional on $Y_{01}^{n,s}$ and $Z^n = 1$ is modified as follows:

$$Y_{j,2k-1}^{n,s} \mid Y_{j-1,k}^{n,s} \sim \text{binomial}(Y_{j-1,k}^{n,s}, p_{jk}^{n,s}), \tag{11}$$

$$p_{jk}^{n,s} | Z^n = 1 \sim \text{beta}(\bar{p}_{jk}\tau_j, (1 - \bar{p}_{jk})\tau_j). \tag{12}$$

Ideally, it is desirable to model the variation across genomic sites and the variation across replicates separately. However, in practice, we usually have only two or three replicate DNase-seq measurements in a given cell type, making it difficult to accurately quantify the variation across replicates. Instead, we assume that variation across replicates and variation across genomic sites are the same. The background model $\mathsf{P}(X^n|Z^n = 0)$ can be constructed in a similar way.

4 Maximum likelihood estimation and inference

Inference for the above model requires computing the posterior distribution $\mathsf{P}(Z|X)$, evaluated at the maximum likelihood estimate of the model parameters Θ^*, where $\Theta = \{\bar{p}_{jk}, \tau_j, \alpha^s, \bar{\lambda}_0^s, \alpha_o^s, \bar{\lambda}_{0o}^s\}$. This problem is equivalent to finding the optimal distribution in a parameteric family of distributions $q(Z)$ that has the smallest Kullback-Leibler (KL) divergence to the posterior distribution of interest.

$$q^*(Z) = \arg\min_{q(Z)} \sum_Z q(Z) \log \frac{q(Z)}{\mathsf{P}(Z|X)} \tag{13}$$

$$= \log \mathsf{P}(X;\Theta) - \arg\max_{q(Z)} \left[\sum_Z q(Z) \log \frac{\mathsf{P}(X|Z,\Theta)\mathsf{P}(Z)}{q(Z)} \right] \tag{14}$$

$$= \log \mathsf{P}(X;\Theta) - \arg\max_{q(Z)} \mathcal{F}[q(Z);\Theta]. \tag{15}$$

The first term in 15 is the log likelihood of the data for some fixed value for the model parameters Θ. Thus, minimizing the KL divergence is equivalent to maximizing the function $\mathcal{F}[q(Z);\Theta]$, keeping the model parameters fixed. Note that, when the true posterior distribution $\mathsf{P}(Z|X)$ lies in the specified parametric family, the maximum value $\mathcal{F}[q^*(Z);\Theta]$ is equal to the log likelihood of the data $\log \mathsf{P}(X;\Theta)$ (i.e., the minimum KL divergence is zero). The maximum likelihood estimate of the model parameters Θ can then be obtained by maximizing the function $\mathcal{F}[q^*(Z);\Theta]$. Maximizing $\mathcal{F}[q(Z);\Theta]$ with respect to $q(Z)$ and Θ, iteratively till convergence, gives us the maximum likelihood estimate of the model parameters and posterior probability of transcription factor binding.

Conditional on the model parameters, the data at all putative binding site in all replicates are independent. Thus, the true posterior $\mathsf{P}(Z|X)$ will factorize as $\mathsf{P}(Z|X) = \prod_n \mathsf{P}(Z^n|X^n)$, motivating the following choice for $q(Z)$.

$$q(Z) = \prod_n \text{binomial}(\tilde{\zeta}_n). \tag{16}$$

The function to be maximized can now be written as follows:

$$\mathcal{F}[q(Z);\Theta] = \sum_n \sum_{Z^n} q(Z^n) \log \mathsf{P}(X^n|Z^n,\Theta) + \sum_n \sum_{Z^n} q(Z^n) \log \frac{\mathsf{P}(Z^n)}{q(Z^n)} \tag{17}$$

$$= \sum_n \tilde{\zeta}_n \log \mathsf{P}(X^n|Z^n=1,\Theta) + (1-\tilde{\zeta}_n) \log \mathsf{P}(X^n|Z^n=0,\Theta) \tag{18}$$

$$+ \tilde{\zeta}_n \log \frac{\zeta_n}{\tilde{\zeta}_n} + (1-\tilde{\zeta}_n) \log \frac{1-\zeta_n}{1-\tilde{\zeta}_n} \tag{19}$$

Maximizing $\mathcal{F}$ with respect to $\tilde{\zeta}_n$ for fixed Θ gives

$$\log \frac{\tilde{\zeta}_n^*}{1-\tilde{\zeta}_n^*} = \log \frac{\zeta_n}{1-\zeta_n} + \log \frac{\mathsf{P}(X^n|Z^n=1,\Theta)}{\mathsf{P}(X^n|Z^n=0,\Theta)} \tag{20}$$

Using the model specified in the previous section, we have

$$\log \mathsf{P}(X^n|Z^n=1,\Theta) = \sum_s \log \mathsf{P}(Y_{01}^{n,s}|\alpha^s,\bar{\lambda}_0^s) + \sum_s \sum_{j=1}^{J} \sum_{k=1}^{2^{(j-1)}} \log \mathsf{P}(Y_{j,2k-1}^{n,s}|Y_{j-1,k}^{n,s},\bar{p}_{jk},\tau_j), \tag{21}$$

where

$$\mathsf{P}(Y_{01}^{n,s}|\alpha^s,\bar{\lambda}_0^s) = \text{negativebinomial}\left(\alpha^s, \frac{\bar{\lambda}_0^s}{\alpha^s+\bar{\lambda}_0^s}\right), \tag{22}$$

$$\mathsf{P}(Y_{j,2k-1}^{n,s}|Y_{j-1,k}^{n,s},\bar{p}_{jk},\tau_j) = \text{BetaBinom}(Y_{j-1,k}^{n,s};\bar{p}_{jk}\tau_j,(1-\bar{p}_{jk})\tau_j). \tag{23}$$

A similar likelihood function for the background model $\log \mathsf{P}(X^n|Z^n=0,\Theta)$ can be written. The function $\mathcal{F}$ can be maximized with respect to the model parameters Θ using simple non-linear optimization solvers.

References

[1] Yuchun Guo, Shaun Mahony, and David K Gifford. High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints. *PLoS Computational Biology*, p. e1002638, 2012.

[2] Eric D. Kolaczyk. Bayesian multiscale models for Poisson processes. *Journal of the American Statistical Association*, 94(447):920–933, September 1999.

[3] K.E. Timmermann and R.D. Nowak. Multiscale modeling and estimation of Poisson processes with application to photon-limited imaging. *IEEE Transactions on Information Theory*, 45(3):846–862, April 1999.

[4] Yong Zhang, Tao Liu, Clifford A Meyer, Jrme Eeckhoute, David S Johnson, Bradley E Bernstein, Chad Nusbaum, Richard M Myers, Myles Brown, Wei Li, and X Shirley Liu. Model-based analysis of ChIP-seq. *Genome Biology*, 9:R137, 2008.

Supplementary Figures

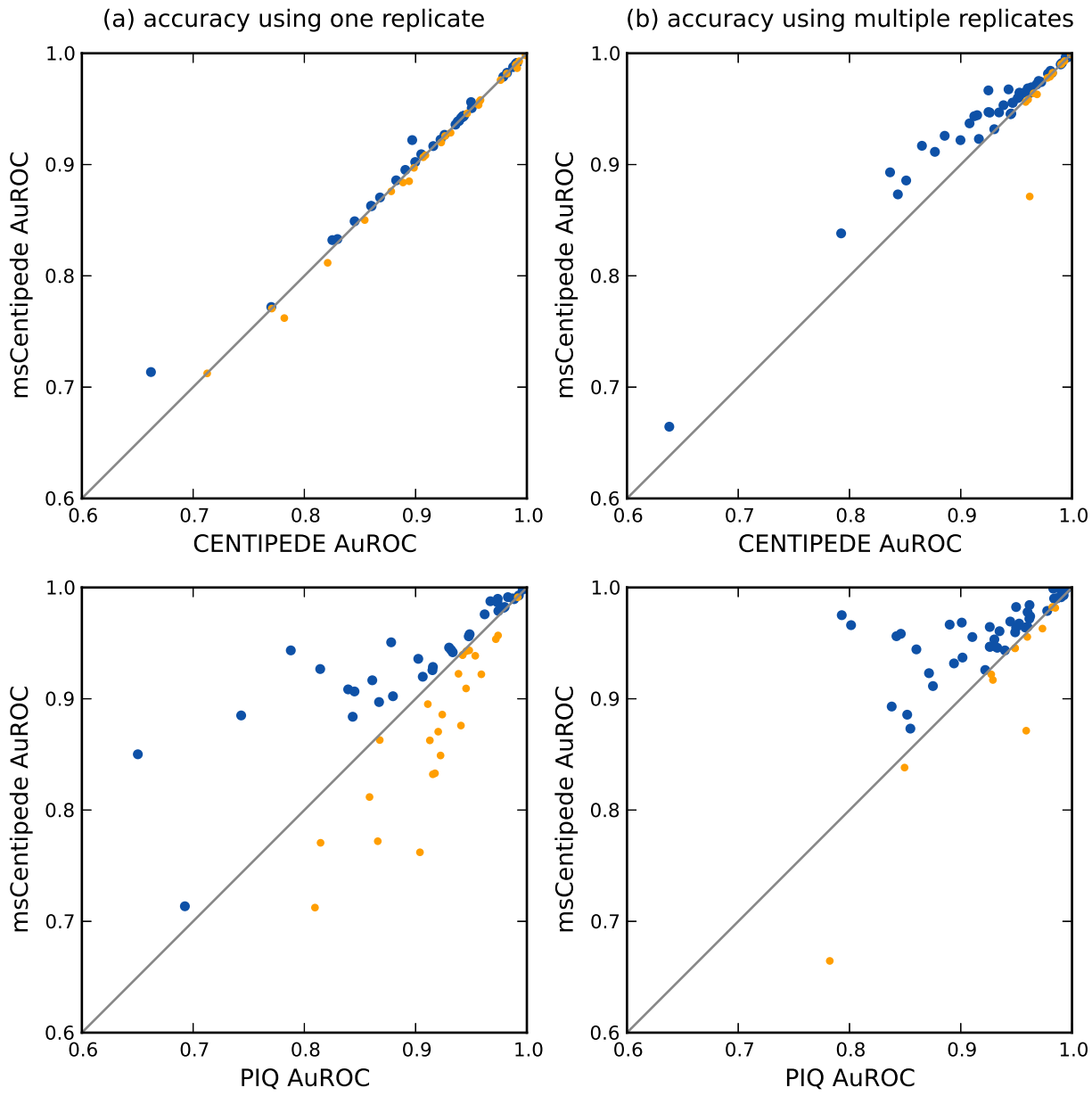

Figure S2: Accuracy of msCentipede, CENTIPEDE and PIQ using a gold standard identified using GEM [1]. Each point corresponds to a different factor and accuracy is measured by area under the ROC curve. Blue points correpond to factors where msCentipede achieves higher accuracy than CENTIPEDE (top panels) or PIQ (bottom panels), and orange points correspond to a worse performance by msCentipede.

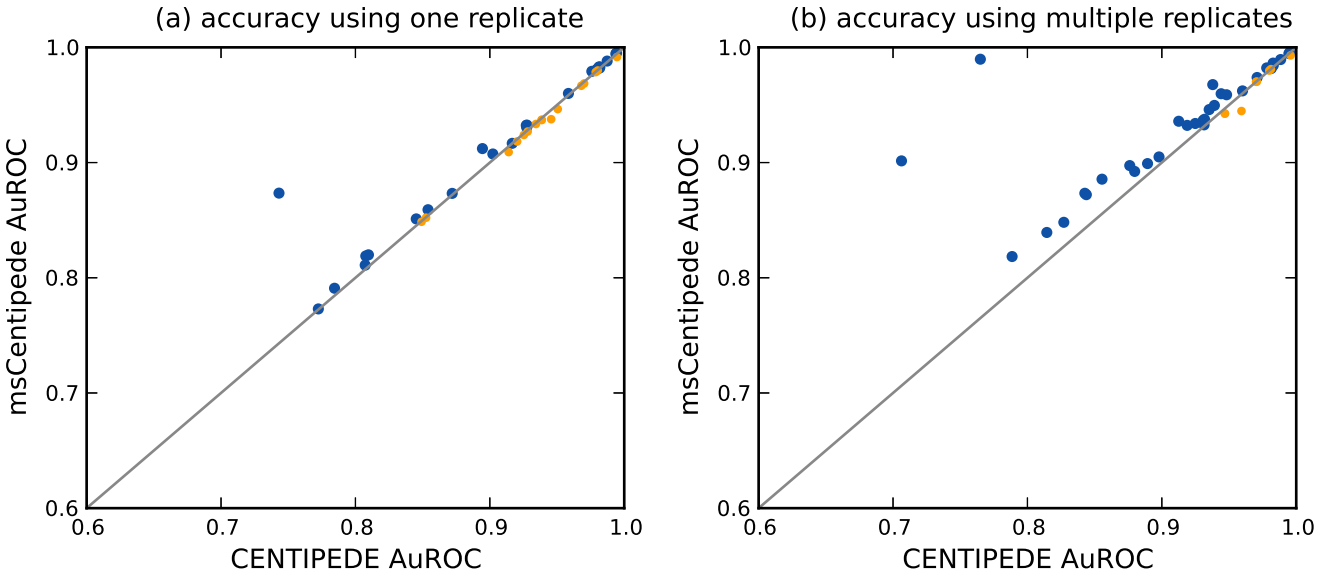

Figure S3: Accuracy of msCentipede and CENTIPEDE using ATAC-seq data and a gold standard identified using MACS [4]. Blue points correpond to factors where msCentipede achieves higher accuracy than CENTIPEDE and orange points correspond to a worse performance by msCentipede.

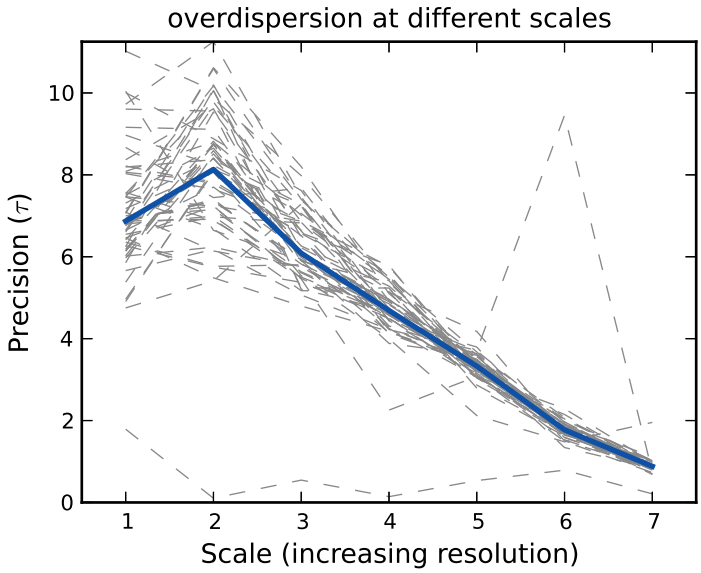

Figure S4: Heterogeneity across different scales. A plot of the precision parameter τ as a function of the scale in the multi-scale model. Each gray line corresponds to a different transcription factor and the solid blue line shows the median trend across all factors.

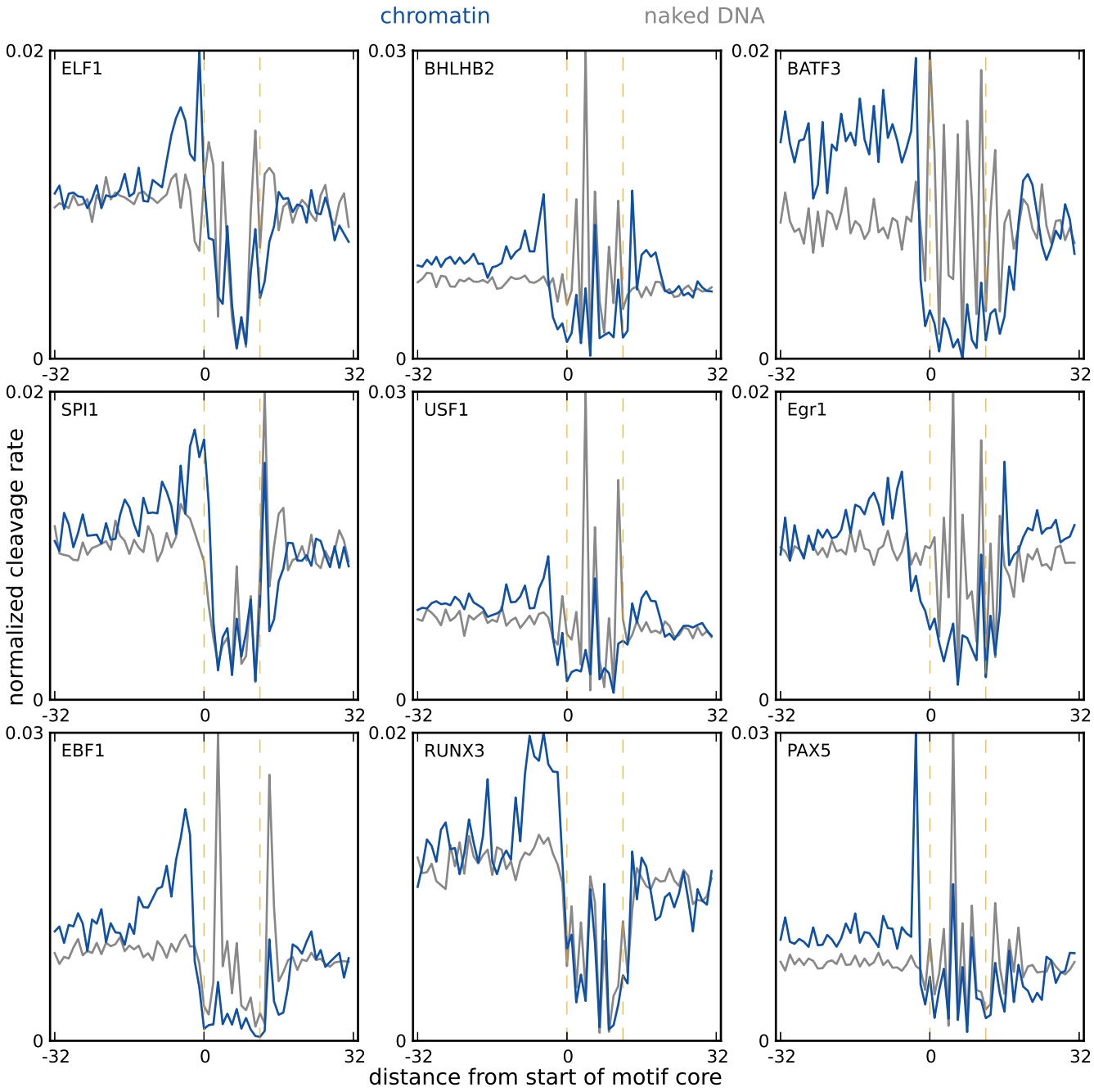

Figure S5: Normalized DNase I cleavage profiles in chromatin and naked DNA, for a subset of transcription factors. The cleavage profiles for chromatin and naked DNA were computed from the maximum likelihood estimates of the parameters $\bar{p}_{jk}$ and $\bar{p}^{o}_{jk}$, respectively. For the sake of clarity, only the plus strand cleavage profile is shown. The dotted orange lines indicate the boundaries of the core motif.

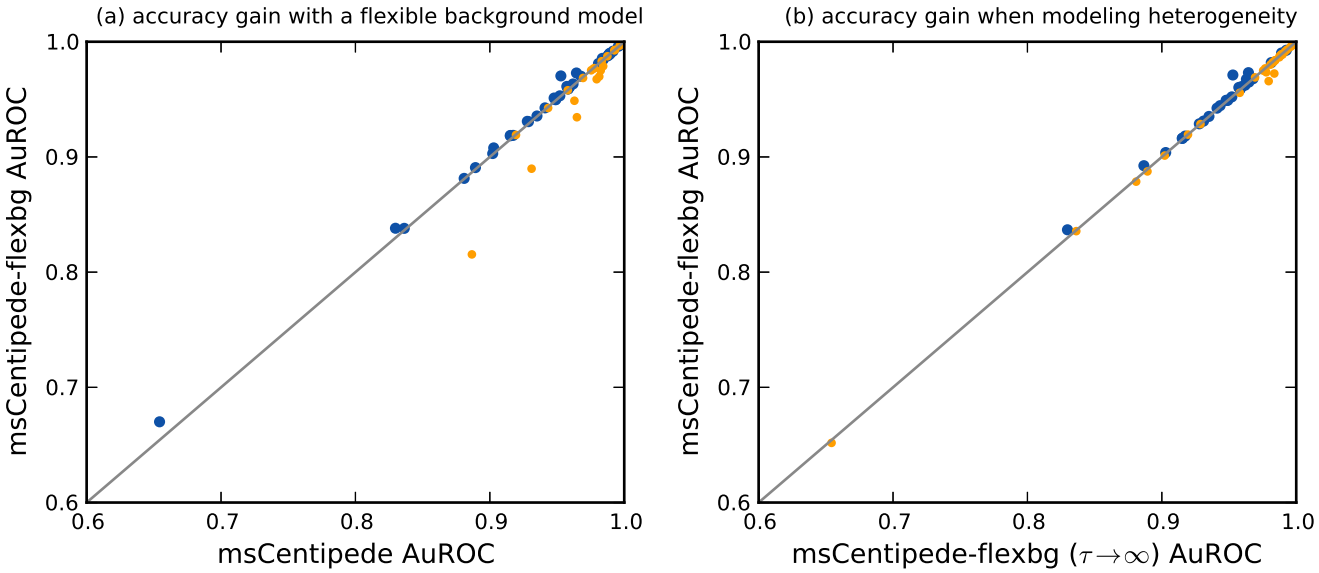

Figure S6: Accuracy of msCentipede and msCentipede-flexbg using DNase-seq data. Blue points correpond to factors where msCentipede-flexbg shows improved performance and orange points correspond to a worse performance by msCentipede-flexbg. The increase in accuracy for msCentipede-flexbg is relatively modest across a large number of transcription factors.