*Nucleic Acids Research, 2018* **1**
*doi: 10.1093/nar/gky027*

# Modular discovery of monomeric and dimeric transcription factor binding motifs for large data sets

**Jarkko Toivonen[1,*], Teemu Kivioja[2], Arttu Jolma[3], Yimeng Yin[3], Jussi Taipale[2,3,4] and Esko Ukkonen[1,5,*]**

[1]Department of Computer Science, P.O. Box 68, FI-00014 University of Helsinki, Helsinki, Finland, [2]Genome-Scale Biology Program, P.O. Box 63, FI-00014 University of Helsinki, Helsinki, Finland, [3]Division of Functional Genomics and Systems Biology, Department of Medical Biochemistry and Biophysics, and Department of Biosciences and Nutrition, Karolinska Institutet, SE 141 83 Stockholm, Sweden, [4]Department of Biochemistry, University of Cambridge, CB2 1GA Cambridge, UK and [5]Helsinki Institute for Information Technology HIIT, University of Helsinki & Aalto University, Helsinki, Finland

Received December 01, 2017; Editorial Decision December 29, 2017; Accepted January 12, 2018

## ABSTRACT

**In some dimeric cases of transcription factor (TF) binding, the specificity of dimeric motifs has been observed to differ notably from what would be expected were the two factors to bind to DNA independently of each other. Current motif discovery methods are unable to learn monomeric and dimeric motifs in modular fashion such that deviations from the expected motif would become explicit and the noise from dimeric occurrences would not corrupt monomeric models. We propose a novel modeling technique and an expectation maximization algorithm, implemented as software tool MODER, for discovering monomeric TF binding motifs and their dimeric combinations. Given training data and seeds for monomeric motifs, the algorithm learns in the same probabilistic framework a mixture model which represents monomeric motifs as standard position-specific probability matrices (PPMs), and dimeric motifs as pairs of monomeric PPMs, with associated orientation and spacing preferences. For dimers the model represents deviations from pure modular model of two independent monomers, thus making co-operative binding effects explicit. MODER can analyze in reasonable time tens of Mbps of training data. We validated the tool on HT-SELEX and ChIP-seq data. Our findings include some TFs whose expected model has palindromic symmetry but the observed model is directional.**

## INTRODUCTION

In transcriptional regulation, proteins called transcription factors (TFs) bind to specific DNA motifs, to have a regulatory effect on the transcription rate of particular genes. The regulating TFs may bind co-operatively in clusters of two or more factors which makes the regulation combinatorial by nature (1–4). Therefore, it is of interest not only to find the binding motifs for individual monomeric TFs but also motifs for dimeric and higher order co-operative binding of several TFs on the same regulatory area in DNA. With the massive training data currently available from, e.g. high-throughput SELEX (5,6) and ChIP-seq experiments (7), it is possible to learn complex binding models from quite weak signals.

In a large number of dimeric cases of TF binding, the specificity of the dimeric motif has recently been observed to differ notably from what would be expected were the two factors to bind to DNA independently of each other (4,6,8). Current automatic motif discovery tools do not learn monomeric and dimeric motifs soundly within one probabilistic framework in *modular* fashion such that the effects of co-operative binding on motifs could be shown and analyzed. In this paper, we propose such a learning algorithm and a software tool for modular discovery of monomeric and dimeric binding motifs for TFs.

The algorithm uses a class of probabilistic mixture models for (possibly multi-profile) monomeric binding motifs and all their dimeric combinations. Our model represents each monomeric motif as a standard position-specific probability matrix (PPM) (9,10). Each dimeric motif is represented in modular fashion as a pair of monomeric PPMs, with associated information on the relative orientation and spacing of the two monomeric components. In our model, the monomeric components need not be spatially separate but their sites may overlap; such overlaps have been re-

*To whom correspondence should be addressed. Tel: +358 2941 51262; Fax: +358 9876 4314;
Email: jarkko.toivonen@cs.helsinki.fi, esko.ukkonen@helsinki.fi

ported, e.g. in (4,11). A novel feature of our model is that it includes a deviation matrix that represents explicitly how much the discovered dimeric PPM deviates from the expected PPM for independent component monomers. Another novelty is that monomeric and dimeric models are learned such that the effect of the noise from dimeric occurrences on monomeric models is minimized. Moreover, the mixing parameters of the model reveal the relative abundances of different motif combinations. In particular, the mixing parameters for the dimeric variants give precise quantitative indication of orientation and spacing preferences of the two monomers that make the dimer.

For learning our binding model we describe an expectation maximization (EM) algorithm (12), called MODER (MOtif DEtectoR). Given a data set of sequences that contain enriched motif instances, MODER learns by EM search the parameters of all model components simultaneously, as a mixture of several PPMs, by optimizing the alignment of the model with the training data using maximum likelihood estimation. The EM search is initialized with user-given seed sequences for the monomeric profiles of the model. It finds PPMs for the monomers as well as for their dimeric combinations within given range of spacings and orientations. Higher-order combinations are not included, as it would exponentially increase the complexity and the size of the model. Monomer PPMs are learned using pruning techniques that minimize contamination from near-by motifs occurrences and from background. The requirement to provide seeds is a limitation of MODER which depends on prior knowledge (such as motif databases) or the use of other motif discovery algorithms. On the other hand, seed-based initialization makes MODER fast and capable of processing in reasonable time a training data consisting of sequences that are hundreds of bps long and are several Mbps in total size. MODER was designed for motif discovery from HT-SELEX reads, but other type of training data, such as ChIP-seq data sets, can be used as well.

Validation experiments of MODER show robust and fast performance both on HT-SELEX and ChIP-seq data. We applied MODER on six HT-SELEX data sets, each consisting of $10^5$–$10^6$ reads of length 30 or 40, and found varying amounts of difference between observed and expected motifs: for example, for factors FLI1 and PKNOX2 the expected homodimeric model has palindromic symmetry but the observed model is directional, reconfirming an earlier observation in (6). From ChIP-seq data MODER finds for factor CTCF essentially the same dimeric model as reported in (13,14), and for modular receptor RXRA a dimeric model that the Tomtom tool (15) matches with a known RXRA heterodimer. For factor NRSF, MODER finds from ChIP-seq data essentially the same multi-profile model as in (16).

In previous research, a dimer model quite similar to ours but without explicit modular structure and overlaps of monomers within dimers was introduced, with an entropy minimization learning algorithm Bipad/Maskminent (17–19). Discovery of spaced dyads (pairs of relatively short motifs) was considered in (20,21). Gibbs sampling based BioProspector (22) is another early dimer search algorithm. Recent dimer prediction methods include SpaMo (23), iTFs (24), and TACO (25). All start from given monomeric PPMs and find, using thresholding, the occurrence sites of the PPMs in the training data. Then enrichment of specific spacings of pairs of occurrences is detected, with an analysis of the statistical significance but without an analysis of co-operative effects of dimer components. SpaMo was designed for finding preferred distances between the site of the primary TF and the sites of secondary TFs in ChIP-seq data. The dimer model of iTFs includes relative orientation of the components but it does not consider overlaps and uses binned distances. Finally, TACO's model includes orientation and distance and allows the components to overlap, but does not analyze the effect of overlap on the binding profile.

Using the EM algorithm in motif discovery was initiated by Lawrence and Reilly (26) and was used for finding motifs with spacers by Cardon and Stormo (27). The mixture model and the EM learning of MODER generalize the techniques of MEME (28,29) to multi-profile dimeric case. As compared to MEME, an important feature of MODER is that it learns all submodels simultaneously, using all training data symmetrically. coMOTIF (30) is another simultaneous multi-profile motif finder based EM algorithm. It does not, however, keep track of the distances between binding sites and does not allow overlaps of binding sites, nor does it have the modeling of deviation or learning of the motif in the gap positions between the dimer components. MODER can be seen as a generalization of coMOTIF. Recent EM algorithm based finders of monomer motifs include GADEM and rGADEM (31,32) which use genetic algorithm with EM to improve starting PPMs, SEME (33) which uses importance sampling to speed-up the search, EXTREME (16) which achieves speed-up by using the on–line version of the EM algorithm, and STEME (34) which resorts to suffix-trees. Moreover, Liu *et al.* (35) use Gibbs sampling and Ikebata and Yoshida (36) use a repulsive MCMC version of MEME type search for simultaneous discovery of several motifs, Alipanahi *et al.* (37) use deep learning for motif discovery with good validation results but non-modular structure of the underlying model, and Colombo and Vlassis (38) find monomeric motifs with a fast spectral learning algorithm. Recent motif finders specially designed for large ChIP-seq data include rGADEM (32), HOMER (39), ChIP-Munk (40), and MEME-ChIP (41), evaluated in (42).

In the rest of the paper, the next section defines the mixture model of MODER, the next one gives the associated EM algorithm for estimating the model parameters, then our implementation of MODER is described, with techniques to initialize and prune the search, and finally we report some validation and comparison experiments and discuss motifs found by MODER for TFs FLI1, HOXB13, HNF4A, TFAP2A, FOXC1, PKNOX2, NRSF, CTCF and RXRA.

## MATERIALS AND METHODS

### Model structure

The binding affinity model learned by MODER, specified by parameters $\eta = (\theta, \psi, \lambda)$, gives a probability distribution for sequences in some alphabet $\Sigma$. We will use always the

DNA alphabet $\Sigma = \{A, C, G, T\}$ but the model works for arbitrary alphabets.

Model $\eta$ is a mixture of distributions for *monomeric sequences* that contain one occurrence of a monomeric motif, and distributions for *dimeric sequences* that contain two monomeric motifs in a specific relative orientation and spacing, and a distribution for *background sequences*. Monomeric distributions are built from the PPMs of the monomers and the background. For all orientation and spacing alternatives between the two monomers in a dimer, dimeric distributions are built either from the PPMs of the monomers and the background or from the PPM of the entire dimer and the background. If the two monomers of a dimer do not overlap and have a long gap in between, then the dimeric distribution is just the product of the two monomer PPMs, that is, the model assumes that there is no co-operative effect affecting the independence of the two binding profiles. If the monomers overlap or the gap between them is short, then the binding profiles of two monomers do not necessarily remain independent. There can be interaction between the components of a dimer as they may physically contact each other, or the interaction can be DNA mediated (4). Therefore the model allows deviating from pure reduction to monomer PPMs and also represents, using the so-called deviation matrix, how the PPM learned from data differs from the product of monomer PPMs which would be the expected model if there are no interactions.

The three parameter groups of $\eta = (\theta, \psi, \lambda)$ and the parametrization of the dimeric structures are defined in detail in the following subsections.

*Monomeric PPMs $\boldsymbol{\theta}_k$ and background $\boldsymbol{\theta}_0$.* Parameter $\theta = (\theta_0, \theta_1, \ldots, \theta_p)$ gives the background distribution $\theta_0$ and $p$ monomeric motifs $\theta_k$. Each $\theta_k$, $k \neq 0$, is a $4 \times \ell_k$ PPM

$$\theta_k = \begin{bmatrix} \theta_k^{A,1} & \theta_k^{A,2} & \cdots & \theta_k^{A,\ell_k} \\ \theta_k^{C,1} & \theta_k^{C,2} & \cdots & \theta_k^{C,\ell_k} \\ \theta_k^{G,1} & \theta_k^{G,2} & \cdots & \theta_k^{G,\ell_k} \\ \theta_k^{T,1} & \theta_k^{T,2} & \cdots & \theta_k^{T,\ell_k} \end{bmatrix},$$

where $\theta_k^{a,h} := \theta_k[a, h]$ gives the probability for an alphabet symbol (nucleotide) $a$ to occur in position $h$ of $\theta_k$, and $\ell_k$ denotes the length of $\theta_k$. The reverse complement $\theta_k^{-1}$ of $\theta_k$ is a PPM such that $\theta_k^{-1}[a, h] = \theta_k[\bar{a}, \ell_k - h + 1]$ for each $a$ and $h$, where $\bar{a}$ is the complementary base of $a$ (e.g., $\overline{A} = T$).

The mononucleotide background model $\theta_0 = [\theta_0^A, \theta_0^C, \theta_0^G, \theta_0^T]^T$ gives the occurrence probabilities of each alphabet symbol in a position that is outside the occurrences of monomers or dimers. The background model is position-independent.

*Dimer specification $k_1k_2od$.* The model uses monomeric motifs $\theta_k$ as building blocks of dimeric motifs. The possible dimeric motifs are indexed with quadruples $(k_1, k_2, o, d)$ which we abbreviate as $k_1k_2od$ (this should not be confused with the multiplication of these symbols). A dimer with index $k_1k_2od$ is composed of monomers $\theta_{k_1}$ and $\theta_{k_2}$ whose orientation is $o$ and distance (spacing) from the end of $\theta_{k_1}$ to the start of $\theta_{k_2}$ is $d$, where $o = (o_1, o_2) \in \Omega_{k_1k_2}$ and $d \in \Delta_{k_1k_2}$.

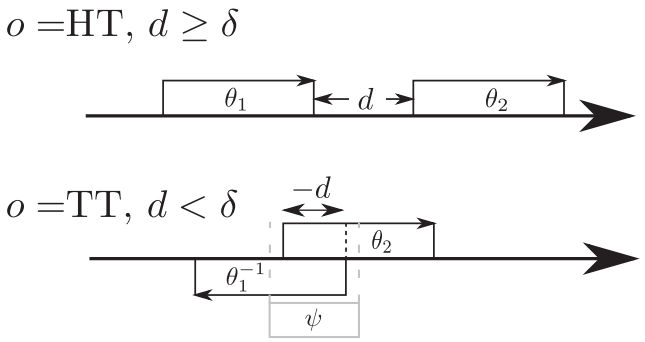


**Figure 1.** Parametrization of dimeric structures. On top, a non-overlapping dimer with spacing $d$ and head-to-tail orientation. Parameter δ gives the lower bound such that monomers separated by a space ≥δ are assumed independent. Below, a dimer with reversed first monomer and overlap of length $d$. When $d < \delta$, the model also includes the bridging PPM $\psi$ that covers the bridging segment of the dimer.

**Table 1.** Relative orientation of two motif occurrences within a dimer

| Orientation $o$ | Short-hand | | $o_1$ | $o_2$ |
|---|---|---|---|---|
| Head-to-Tail | HT | $\rightarrow\rightarrow$ | +1 | +1 |
| Head-to-Head | HH | $\rightarrow\leftarrow$ | +1 | −1 |
| Tail-to-Tail | TT | $\leftarrow\rightarrow$ | −1 | +1 |
| Tail-to-Head | TH | $\leftarrow\leftarrow$ | −1 | −1 |

Exponents $o_1$ and $o_2$ give the orientation of the first and the second PPM: for a PPM $\theta_k$, $\theta_k^{+1}$ leaves the matrix intact but $\theta_k^{-1}$ takes the reverse complement.

Because of co-operative binding effects, monomer motifs alone are not enough for building dimeric models. To model such effects we will use an additional PPM (see the next subsection) that covers the middle area of the dimer, called the bridging segment. Figure 1 illustrates our parametrization of dimeric structures; cf. (17).

The set of possible pairwise orientations $o$ is $\Omega_{k_1k_2} = \{HT, HH, TT\}$ if $k_1 = k_2$ (*homodimer*), and $\Omega_{k_1k_2} = \{HT, HH, TT, TH\}$ otherwise (*heterodimer*). Table 1 describes different orientations $o = (o_1, o_2)$ giving the directions of motifs $\theta_{k_1}$ and $\theta_{k_2}$. Note that for homodimers the orientations HT and TH are identical, and one can use HT to represent them both. We assume that motif $\theta_{k_1}$ always occurs before motif $\theta_{k_2}$ when moving from 5′ end to 3′ end and using motif start position as reference point. The reverse order of the two motifs transforms back to this case by considering the complementary strand.

The possible distances between the two occurrences are given as an interval $\Delta_{k_1k_2} = [\mathrm{dmin}(k_1, k_2), \mathrm{dmax}(k_1, k_2)]$. If $d \in \Delta_{k_1k_2}$ is non-negative, it gives the number of gap positions between the two occurrences. If $d < 0$, then the occurrences overlap by $-d$ positions. The smallest possible distance $\mathrm{dmin}(k_1, k_2)$ has to be $> -\ell_{k_1}$. MODER implementation uses (optionally adjustable) default value $\mathrm{dmin}(k_1, k_2) = -\min(\ell_{k_1}, \ell_{k_2})/2$, that is, overlaps only up to half of the length of the monomers are allowed. The longest distance possible for sequences of maximum length $L_{\max}$ is $\mathrm{dmax}(k_1, k_2) = L_{\max} - \ell_{k_1} - \ell_{k_2}$.

We use parameter $\delta \geq 0$ to give the minimum spacing such that if the space between the two monomers of a dimer is ≥δ then the monomer profiles are assumed independent, i.e. in

this case the model ignores the possible co-operative interactions that would change the binding preferences of the two TFs or the gap between them. Parameter δ is a user-given constant (default value $\delta = 4$ in our implementation).

In what follows, we refer to the available monomeric and dimeric motifs with index $k$ that may belong to the following three separate sets:

$M = \{1, \ldots, p\}$: the indices for monomeric motifs.

$D^+ = \{k_1k_2od: d \geq \delta, k_1, k_2 \in M\}$: the indices for dimeric motifs whose monomers $\theta_{k_1}$ and $\theta_{k_2}$ have a gap of length $\geq\delta$ in between. This is called the *independent case*.

$D^- = \{k_1k_2od: d < \delta, k_1, k_2 \in M\}$: the indices for dimeric motifs whose monomers $\theta_{k_1}$ and $\theta_{k_2}$ have a gap of length $<\delta$ in between. This is called the *dependent case*. Note that this case includes dimers whose monomers overlap.

*Dimeric PPMs $\tau_{k1}k_2od$, bridging PPMs $\psi_{k1}k_2od$ and deviation matrices $\kappa_{k1}k_2od$.* We use $\tau_{k_1k_2od}$ to denote the PPM (which is a $4 \times (\ell_{k_1} + \ell_{k_2} + d)$ matrix) for motif $k_1k_2od \in D^+ \cup D^-$. Each $\tau_{k_1k_2od}$ is a derived parameter, composed of free parameters, such that if $k_1k_2od \in D^+$ then $\tau_{k_1k_2od}$ is built from $\theta_{k_1}$, $\theta_{k_2}$, and background $\theta_0$, and if $k_1k_2od \in D^-$ then $\tau_{k_1k_2od}$ is built from $\theta_{k_1}$, $\theta_{k_2}$, and the bridging PPM $\psi_{k_1k_2od}$ to be defined below. Constructions are as follows.

If $k_1k_2od \in D^+$, then we put simply

$$\tau_{k_1k_2od} = \theta_{k_1}^{o_1} \bullet \theta_0 \bullet \ldots \bullet \theta_0 \bullet \theta_{k_2}^{o_2} \quad (1)$$

where $\bullet$ concatenates matrices. There are $d$ column-matrices $\theta_0$ in the middle of $\tau_{k_1k_2od}$, that is, the middle gap is filled with the background.

If $k_1k_2od \in D^-$, then a middle segment of $\tau_{k_1k_2od}$ is a free parameter learned from data: for $d < 0$, the columns that are on the overlap area (plus one more column on both sides) are free parameters, and for $0 \leq d < \delta$, the columns that are on the area between the monomers (plus one more column on both sides) are free parameters. This area of length $|d| + 2$ in the middle of a dimer is called the *bridging segment*, and the $4 \times (|d| + 2)$ PPM for the bridging segment is called the *bridging PPM*. We let $\psi_{k_1k_2od}$ denote the bridging PPM. Now, the columns of $\tau_{k_1k_2od}$ that cover the bridging segment come from $\psi_{k_1k_2od}$ while the columns outside this segment are supposed to reduce to the monomer motifs, i.e. they are as in the implied prefix and suffix segments of monomer matrices $\theta_{k_1}^{o_1}$ and $\theta_{k_2}^{o_2}$. So we get

$$\begin{aligned}\tau_{k_1k_2od} = \theta_{k_1}^{o_1}[\cdot, 1 : \ell_{k_1} + \min(d, 0) - 1)] \bullet \psi_{k_1k_2od} \\ \bullet \theta_{k_2}^{o_2}[\cdot, \max(0, -d) + 2 : \ell_{k_2}].\end{aligned} \quad (2)$$

Next, we make it explicit how PPM $\tau_{k_1k_2od}$ differs from the PPM that would be expected were the monomer motifs independent in the dimer. We denote such an *expected* PPM as $E_{k_1k_2od}$. It models the situation that motifs $\theta_f = \theta_{k_1}^{o_1}$ and $\theta_r = \theta_{k_2}^{o_2}$ have independent instances at distance $d$ from each other in sequences $a_1a_2 \cdots a_{\ell_{k_1}+\ell_{k_2}+d}$ with an occurrence of $\theta_f$ at the left end and $\theta_r$ at the right end.

Let first $d \geq 0$. Consider the occurrence probability $P(a_i)$ of the $i$th symbol $a_i$. Obviously, if $i \leq \ell_{k_1}$, then $P(a_i) = \theta_f[a_i, i]$; if $\ell_{k_1} < i \leq \ell_{k_1} + d$, then $P(a_i) = \theta_0(a_i)$, i.e., we expect to see the background distribution between the two motifs; and if $i > \ell_{k_1} + d$, then $P(a_i) = \theta_r[a_i, i - \ell_{k_1} - d]$. This means that $E_{k_1k_2od}$ is just $\theta_f$ followed by $d$ columns, each equal to $\theta_0$, followed by $\theta_r$; c.f., the definition of $\tau_{k_1k_2od}$ in the independent case (1).

Let then $d < 0$, i.e., the motifs overlap by $|d|$ symbols. Consider again the probability $P(a_i)$. If $i \leq \ell_{k_i} + d$, then $P(a_i) = \theta_f[a_i, i]$, and hence the $i$th column of the expected PPM is $E_{k_1k_2od}[\cdot, i] = \theta_f[\cdot, i]$. Similarly, if $i > \ell_{k_1}$, then $P(a_i) = \theta_r[a_i, i - (\ell_{k_1} + d)]$, and hence $E_{k_1k_2od}[\cdot, i] = \theta_r[\cdot, i - (\ell_{k_1} + d)]$. In the remaining case we have $\ell_{k_1} + d < i \leq \ell_{k_1}$, and the $i$th symbol $a_i$ belongs to the area where the two motifs overlap. Hence $a_i$ is generated by both $\theta_f$ and $\theta_r$, under the condition that both generate the same symbol because in the overlapping area the two motifs have to coincide. Therefore $P(a_i)$ would be equal to $\theta_f[a_i, i]\theta_r[a_i, i - (\ell_{k_1} + d)]$, normalized by the condition that both motifs generate the same symbol. This gives

$$P(a_i) = \frac{\theta_f[a_i, i]\theta_r[a_i, i - (\ell_{k_1} + d)]}{\sum_{c\in\Sigma} \theta_f[c, i]\theta_r[c, i - (\ell_{k_1} + d)]}, \quad (3)$$

and therefore the $i$th column becomes

$$E_{k_1k_2od}[\cdot, i] = \frac{\theta_f[\cdot, i]\theta_r[\cdot, i - (\ell_{k_1} + d)]}{\sum_{c\in\Sigma} \theta_f[c, i]\theta_r[c, i - (\ell_{k_1} + d)]}, \quad (4)$$

where $\times$ denotes element-wise product.

Finally, the *deviation matrix* $\kappa_{k_1k_2od}$, defined as

$$\kappa_{k_1k_2od} = \tau_{k_1k_2od} - E_{k_1k_2od},$$

gives the difference between observed and expected model. Deviation matrices will be visualized using a variant of the sequence logo in which positive values are shown above a separating line and negative values below it, see Figure 2. Note also that the expected PPM of homodimers is always palindrome symmetric for orientations HH and TT.

*Mixing parameters $\boldsymbol{\lambda}$.* Mixing parameters $\lambda = \{\lambda_k: k \in \{0\} \cup M \cup D^+ \cup D^-\}$ give the probability of each component of the mixture as follows:

$\lambda_k$, $k \in M$, is the probability that the sequence contains exactly one monomeric occurrence of motif $\theta_k$ and no other occurrences.

$\lambda_k$, $k = k_1k_2od \in D^+ \cup D^-$, is the probability that the sequence contains exactly one occurrence of motif $\tau_{k_1k_2od}$ and no occurrences of other motifs.

$\lambda_0$ is the probability that the sequence contains no motif occurrences.

For each pair $(k_1, k_2)$, the array $(\lambda_{k_1k_2od})_{o\in\Omega_{k_1k_2}, d\in\Delta_{k_1k_2}}$ of mixing parameter values is called the *co-operative binding table* (*COB table*) of motifs $\theta_{k_1}$ and $\theta_{k_2}$. The values in a COB table indicate the orientation and spacing preferences of the dimeric structures that are composed of $\theta_{k_1}$ and $\theta_{k_2}$.

Figure 2 illustrates model $\eta$ for binding motifs of TF FLI1.

### Learning by expectation maximization

Given a training data set $X = \{X_1, X_2, \ldots, X_n\}$ consisting of $n$ DNA sequences $X_i = X_{i1} \cdots X_{iL_i}$, where $L_i$ is the length of the $i$th sequence, we use the EM algorithm (12,28) to find model parameters $\eta$ which maximize the expectation of the likelihood $L(\eta|X, Z) = P(X, Z|\eta)$, where latent variables $Z$ give the 'missing information' used by an EM algorithm.

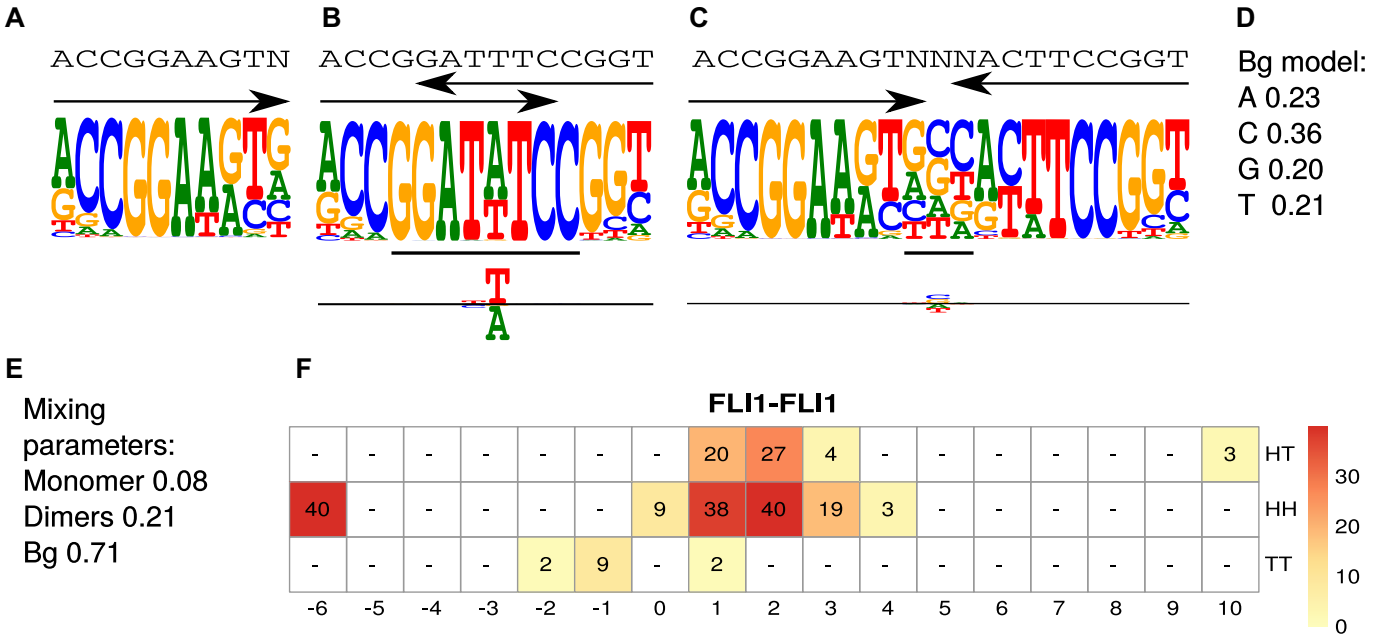


**Figure 2.** Example model $\eta$ for factor FLI1. The model was learned by MODER from a HT-SELEX data set (PRJEB14550, 143 389 reads of length 40), see Section Validation results. (**A**) Monomeric PPM $\theta_1$ with original seed ACCGGAAGTN. (**B**) Dimeric PPM $\tau_{1,1,\mathrm{HH},-6}$ with (above) the seed and arrows indicating the orientation, and (below) horizontal bar indicating the bridging segment, and deviation matrix $\kappa_{1,1,\mathrm{HH},-6}$ with the positive values visualized above and the negative values below the horizontal line. (**C**) Dimeric PPM $\tau_{1,1,\mathrm{HH},1}$ and its deviation matrix. (**D**) Background PPM $\theta_0$. (**E**) Mixture break-down into monomer, all dimers, and background. For example, 0.21 is the sum of mixing parameter values $\lambda_k$, $k \in D^+ \cup D^-$. Note that the proportion of background is larger than the signal, because data from an early SELEX cycle with lots of background was used. (**F**) Heat map of COB table ($\lambda_{1,1,o,d}$) for homodimers of FLI1, giving the break-down into individual dimers and indicating that $\tau_{1,1,\mathrm{HH},-6}$ (panel b), $\tau_{1,1,\mathrm{HH},1}$ (panel c), and $\tau_{1,1,\mathrm{HH},2}$ are the strongest dimers. Horizontal axis gives the distance $d$, and a cell with no value indicates that the corresponding dimeric case was pruned during the EM search; see Section Pruning the search. The units in the COB table are integer multiples of 0.001.

Latent variables are 0–1-valued random variables that indicate how the data $X$ is aligned to the model. To align $X_i$, there are latent variables $Z_{ik}$., $k \in \{0\} \cup M \cup D^+ \cup D^-$, with exactly one of them having value 1, that code the alignment as follows.

Case $Z_{i0} = 1$: Sequence $X_i$ has no occurrences of motifs and is generated by the background model $\theta_0$ alone.

Case $Z_{ikj} = 1$: If $k \in M$, then the sequence $X_i$ has an occurrence of motif $\theta_k$ starting at position $j$. The rest of $X_i$ is generated by the background model. If $k = k_1 k_2 o d \in D^+ \cup D^-$ then the sequence $X_i$ has an occurrence of motif $\tau_k$ at position $j$, that is, an occurrence of motif $\theta_{k_1}$ at position $j$ and an occurrence of motif $\theta_{k_2}$ at $j + \ell_{k_1} + d$ such that the occurrences of $\theta_{k_1}$ and $\theta_{k_2}$ have relative orientation $o$.

We denote by $S_{ik}$ the set of positions $j$ at which motif $k$ may occur in $X_i$. For $k \in M$ we have $S_{ik} = \{1, \ldots, L_i - \ell_k + 1\}$, and for $k = k_1 k_2 o d \in D^+ \cup D^-$, $S_{ik} = S_{ik_1k_2od} = \{1, \ldots, L_i - (\ell_{k_1} + \ell_{k_2} + d) + 1\}$.

The probability of $X_i$ in model $\eta$, given the missing information $Z_i$., is straightforward to evaluate as follows. If sequence $X_i$ contains no motif occurrences, i.e. $Z_{i0} = 1$, its probability is

$$P(X_i | Z_{i0} = 1, \eta) = \prod_{h=1}^{L_i} \theta_0[X_{ih}]. \quad (5)$$

If the sequence $X_i$ contains one motif occurrence, i.e. $Z_{ikj} = 1$ for some $k \in M, j \in S_{ik}$, its probability is

$$P(X_i | Z_{ikj} = 1, \eta) = \prod_{h \in B_1} \theta_0[X_{ih}] \cdot \prod_{h=1}^{\ell_k} \theta_k[X_{i,j+h-1}, h], \quad (6)$$

where $B_1 = \{1, \ldots, L_i\} \setminus [j, j + \ell_k)$.

For the dimeric binding we have two cases: independent ($d \geq \delta$) and dependent ($d < \delta$). Let first $k = k_1 k_2 o d \in D^+$. Define the set $B_2 = \{1, \ldots, L_i\} \setminus \left([j, j + \ell_{k_1}) \cup [j + \ell_{k_1} + d, j + (\ell_{k_1} + \ell_{k_2} + d))\right)$. Then the probability of $X_i$ is

$$P(X_i | Z_{ikj} = 1, \eta) = \prod_{h \in B_2} \theta_0[X_{ih}] \cdot \prod_{h=1}^{\ell_{k_1}} \theta_{k_1}^{o_1}[X_{i,j+h-1}, h] \cdot \prod_{h=1}^{\ell_{k_2}} \theta_{k_2}^{o_2}[X_{i,j+\ell_{k_1}+d+h-1}, h]. \quad (7)$$

Let then $k = k_1 k_2 o d \in D^-$. The probability of $X_i$ is

$$P(X_i | Z_{ikj} = 1, \eta) = \prod_{h \in B_2} \theta_0[X_{ih}] \cdot \prod_{h=1}^{\ell_{k_1}+\ell_{k_2}+d} \tau_{k_1 k_2 o d}[X_{i,j+h-1}, h]. \quad (8)$$

Recall from (2) that $\tau_{k_1 k_2 o d}$ is composed of bridging PPM $\psi_{k_1 k_2 o d}$ in the middle and of flanking segments taken from PPMs $\theta_{k_1}$ and $\theta_{k_2}$.

Now the joint likelihood of the model parameters, given data $X$ and missing information $Z$, is the product of mixture probabilities of each $X_i$:

$$L(\eta|X, Z) = P(X, Z|\eta) = \prod_{i=1}^{n} \Big( Z_{i0} \cdot \lambda_0 \cdot P(X_i|Z_{i0} = 1, \eta) + \sum_{k \in M \cup D^+ \cup D^-} \sum_{j \in S_k} Z_{ikj} \cdot \frac{\lambda_k}{|S_{ik}|} \cdot P(X_i|Z_{ikj} = 1, \eta) \Big).$$

It is important to note here that, to simplify notation, we have ignored the fact that we should consider motif occurrences appearing in the reverse DNA strand as well. For this algorithm to work in the two-stranded case, a new index should be added, which specifies the direction (+1 or –1) of a monomer or a dimer occurrence. Then in all the places where we sum over $j$, we should sum over the directions as well. Moreover, to make sure that the probabilities add up to one, an additional division by two should be performed where we currently divide by $|S_{ik}|$.

As for each $i$, exactly one of the latent values $Z_{i\cdot}$ equals 1 and the others are zeros, the log-likelihood has the following form:

$$\log P(X, Z|\eta) = \sum_{i=1}^{n} \Big[ Z_{i0} \log \big(\lambda_0 P(X_i|Z_{i0} = 1, \eta)\big) + \sum_{\substack{k \in M \cup D^+ \cup D^- \\ j \in S_k}} Z_{ikj} \log \big(\frac{\lambda_k}{|S_{ik}|} P(X_i|Z_{ikj} = 1, \eta)\big) \Big]. \quad (9)$$

The EM algorithm repeatedly applies the following rule to update $\eta = (\theta, \kappa, \lambda)$ until convergence:

$$\eta^{(t+1)} := \arg\max_{\eta} E_{Z|X,\eta^{(t)}} \log P(X, Z|\eta^{(t)}).$$

One iteration of the algorithm, indexed with $t$, consists of an E-step and an M-step. These steps are described next.

### Expectation step

E-step finds the expectation of log-likelihood (9) for current parameter values $\eta^{(t)}$. By linearity of expectation, this reduces to finding the expected values $z_{i\cdot}$ of latent variables $Z_{i\cdot}$. By noting that 0 and 1 are the only possible values of a latent variable, and by applying the Bayes rule, one can see that the expected values and hence the update rule of the E-step becomes, for $k \in \{0\} \cup M \cup D^+ \cup D^-$ and $j \in S_{ik}$, as follows:

$$z_{i0}^{(t)} := E[Z_{i0}|X, \eta^{(t)}] = P(Z_{i0} = 1|X_i, \eta^{(t)}) = \frac{\lambda_0^{(t)} \cdot P(X_i|Z_{i0} = 1, \eta^{(t)})}{P(X_i|\eta^{(t)})}, \quad (10)$$

$$z_{ikj}^{(t)} := E[Z_{ikj}|X, \eta^{(t)}] = P(Z_{ikj} = 1|X_i, \eta^{(t)}) = \frac{\lambda_k^{(t)}/|S_{ik}| \cdot P(X_i|Z_{ikj} = 1, \eta^{(t)})}{P(X_i|\eta^{(t)})}. \quad (11)$$

Here, probability $P(X_i|Z_{i0} = 1, \eta^{(t)})$ is given by (5) and probability $P(X_i|Z_{ikj} = 1, \eta^{(t)})$ by (6), (7) or (8), and

$$P(X_i|\eta^{(t)}) = \lambda_0^{(t)} \cdot P(X_i|Z_{i0} = 1, \eta^{(t)}) + \sum_{\substack{k \in M \cup D^+ \cup D^- \\ j \in S_{ik}}} \lambda_k^{(t)}/|S_{ik}| \cdot P(X_i|Z_{ikj}=1, \eta^{(t)}). \quad (12)$$

### Maximization step

M-step maximizes the expectation of log-likelihood for current $z^{(t)}$ by updating parameters $\eta = (\theta, \psi, \lambda)$. The form of log-likelihood (9) is such that the M-step is of Baum–Welch type: parameters are updated by normalizing the expected counts of using different components of the model when $X$ is aligned to the model according to $z^{(t)}$.

The update rules for mixing parameters become:

$$\lambda_0^{(t+1)} := \frac{1}{n} \sum_{i=1}^{n} z_{i0}^{(t)}, \text{ and} \quad (13)$$

$$\lambda_k^{(t+1)} := \frac{1}{n} \sum_{i=1}^{n} \sum_{j \in S_{ik}} z_{ikj}^{(t)}, \; k \in M \cup D^+ \cup D^-. \quad (14)$$

To update $\theta$ and $\psi$ we first accumulate the expected counts of how many times each mixture component is used when $X$ is aligned with $\eta^{(t)}$. For all $k \in M$, we get the $4 \times \ell_k$ matrices of expected counts of the monomer motifs as

$$W_k = \sum_{i=1}^{n} \Big[ \sum_{j \in S_{ik}} z_{ikj}^{(t)} I_{\ell_k}(i, j) + \sum_{kk'od \in D^+} \sum_{j \in S_{ikk'od}} z_{ikk'odj}^{(t)} I_{\ell_k}^{o_1}(i, j) + \sum_{k'kod \in D^+} \sum_{j \in S_{ik'kod}} z_{ik'kodj}^{(t)} I_{\ell_k}^{o_2}(i, j + \ell_{k'} + d) \Big].$$

Here $I_{\ell_k}(i, j)$ is $4 \times \ell_k$ matrix-valued indicator function such that $I_{\ell_k}(i, j)[a, h] = 1$ if $X_i[j + h - 1] = a$, and otherwise $I_{\ell_k}(i, j)[a, h] = 0$. Again, $I_{\ell_k}^{-1}(i, j)$ is the reverse complement of $I_{\ell_k}(i, j)$. Note that the above aggregation of $W_k$ implements the modularity of binding: a monomer model $\theta_k$ gets its counts from monomeric occurrences of $\theta_k$ as well as from occurrences of $\theta_k$ as an independent component of a dimer. Since the monomer models are not learned from the overlapping cases, there is no coupling between the monomers and the deviations matrices, i.e. both are uniquely defined.

For $k \in D^-$, the $4 \times (\ell_{k_1} + \ell_{k_2} + d)$ matrix of the expected counts is

$$W_k = W_{k_1 k_2 od} = \sum_{i=1}^{n} \sum_{j \in S_{ik_1k_2od}} z_{ik_1k_2odj}^{(t)} I_{\ell_{k_1} + \ell_{k_2} + d}(i, j).$$

According to our modularity constraint the columns of $W_{k_1k_2od}$ that are outside the bridging segment should be modeled with $\theta_{k_1}$ and $\theta_{k_2}$. They should therefore be added

to $W_{k_1}$ and $W_{k_2}$ as follows

$$W_{k_1}^{o_1}[\cdot, 1 : \ell_{k_1} + \min(d, 0) - 1] \mathrel{+}= W_{k_1 k_2 od}[\cdot, 1 : \ell_{k_1} + \min(d, 0) - 1], \quad (15)$$

$$W_{k_2}^{o_2}[\cdot, \max(-d, 0) + 2 : \ell_{k_2}] \mathrel{+}= W_{k_1 k_2 od}[\cdot, \ell_{k_1} + \max(d, 0) + 2 : \ell_{k_1} + \ell_{k_2} + d]. \quad (16)$$

The count vector of the background model is obtained as

$$W_0 = Q_X - \sum_{k \in M} \sum_{h=1}^{\ell_k} W_k[\cdot, h] - \sum_{k_1 k_2 od \in D^-} \sum_{h=\ell_{k_1}+\min(d,0)}^{\ell_{k_1}+1+\max(d,0)} W_{k_1 k_2 od}[\cdot, h],$$

where $Q_X = [Q_X^A, Q_X^C, Q_X^G, Q_X^T]^T$ is the column-vector of total counts of alphabet symbols in the data set $X$.

When normalized column-wise, the matrices $W_k$ (with pseudo-counts possibly added) give updated $\theta_k$ for $k \in \{0\} \cup M$:

$$\theta_0^{(t+1)}[\cdot] := \frac{W_0[\cdot]}{\sum_{a \in \Sigma} W_0[a]}, \quad (17)$$

$$\theta_k^{(t+1)}[\cdot, h] := \frac{W_k[\cdot, h]}{\sum_{a \in \Sigma} W_k[a, h]}. \quad (18)$$

Similarly, the bridging segments of $W_k$, $k = k_1 k_2 od \in D^-$, give updated bridging PPMs $\psi_k$:

$$\psi_{k_1 k_2 od}^{(t+1)}[\cdot, h] := \frac{W_{k_1 k_2 od}[\cdot, h + \ell_{k_1} + \min(d, 0) - 1]}{\sum_{a \in \Sigma} W_{k_1 k_2 od}[a, h + \ell_{k_1} + \min(d, 0) - 1]}, \quad (19)$$

where $h = 1, \ldots, |d| + 2$.

## Implementation of MODER

In this section we give practical details of our implementation of the MODER algorithm and provide some modifications to improve its efficiency.

*Input.* The input of MODER consists of the following items.

(1) Data set $X$ that consists of DNA sequences $X_1, X_2, \ldots, X_n$, with $|X_i| = L_i$ for all $i = 1, \ldots, n$.
(2) The *seeds* $s_1, s_2, \ldots, s_p$. Each $s_k$ is an IUPAC sequence of length $|s_k| = \ell_k$. Seeds should be high-affinity representative sequences, one for each monomeric motif to be learned from data $X$. They will be used for constructing initial values for PPMs $\theta_k$.
(3) Set $R \subset \{1, 2, \ldots, p\}^2$ of pairs that restrict the set of dimeric motifs represented in $\eta$. MODER learns only dimers $k_1 k_2 od$ such that $(k_1, k_2)$ is in $R$.
(4) Minimum gap length in dimers whose monomers are assumed independent, $\delta$; maximum number of EM-iterations, `maxiter`; and the convergence threshold for parameter change in consecutive EM-iterations, $\epsilon$.

*EM iterations.* As the EM algorithm converges to a local optimum, it is crucial to use good initial values for the parameters. Initial PPMs $\theta_1^{(1)}, \ldots, \theta_p^{(1)}$ are obtained from input data $X$ and seeds $s_1, \ldots, s_p$ using the *multinomial method* (5). Initial bridging PPMs $\psi_{k_1 k_2 od}^{(1)}$ are obtained from input data $X$ and *combined seeds* $s_{k_1 k_2 od}^{(1)}$ using the multinomial method. A combined seed is constructed by orienting seeds $s_{k_1}$ and $s_{k_2}$ according to $o$, spacing them by $d$ symbols, and replacing the symbols in the bridging segment by the neutral IUPAC symbol `N`. This gives sequence $y$. Then the combined seed $s_{k_1 k_2 od}^{(1)}$ is the highest counting non-palindromic subsequence of input data $X$ that matches with $y$. A non-palindromic seed makes it possible for the EM search to break the symmetry and find non-palindromic PPMs. Background model is initialized as $\theta_0^{(1)} := Q_X / |Q_X|$ where $Q_X = [Q_X^A, Q_X^C, Q_X^G, Q_X^T]^T$ is the column-vector of total counts of alphabet symbols in $X$. The mixing parameters $\lambda^{(1)}$ are initialized as follows:

- $\lambda_0^{(1)} := 0.5$,
- $\lambda_k^{(1)} := \begin{cases} 0.3/p & \text{if } R \text{ is non-empty and} \\ 0.5/p & \text{otherwise, for all } k \in \{1, \ldots, p\}, \end{cases}$
- $\lambda_{k_1 k_2}^{(1)} := 0.2/|R|$, for all $(k_1, k_2) \in R$. Within a COB table the value $0.2/|R|$ is divided evenly among the cells as $\lambda_{k_2 k_2 od}^{(1)} := \lambda_{k_1 k_2}^{(1)} / (|\Omega_{k_1 k_2}| \cdot |\Delta_{k_1 k_2}|)$.

The EM iterations then proceed as follows:

$t := 0$.
Repeat
  $t := t + 1$
  Compute $z^{(t)}$ using Eqs. (**10**) and (**11**)
  Compute $\lambda^{(t+1)}$ using Eqs. (**14**) and (**13**)
  Compute $\theta^{(t+1)}$ using Eqs. (**17**) and (**18**)
  Compute $\psi^{(t+1)}$ using Eq. (**19**)
until $t =$ `maxiter` or $|(\theta^{(t+1)}, \psi^{(t+1)}) - (\theta^{(t)}, \psi^{(t)})| < \epsilon$.
Output $(\theta^{(t+1)}, \lambda^{(t+1)}, \psi^{(t+1)}, \kappa^{(t+1)})$

It should be noted that the above algorithm outputs the deviation matrix $\kappa$ just for completeness. As $\kappa$ is a derived parameter, it could be evaluated from $\theta$ and $\psi$ in a post-processing phase as well.

*Pruning the search.* MODER implementation makes some heuristic modifications to the EM framework of Section 3 in order to speed-up the search and to utilize prior knowledge of data quality.

First, as the information content of well-known binding affinity PPMs is on average quite high while low information content may indicate contamination from background, MODER trims during the EM all overlapping dimeric mixture components $k$ whose average column-wise information content in the overlapping area goes below a threshold (default 0.40 bits). This is done by setting $\lambda_k := 0$. Similarly, any dimeric component $k$ whose $\lambda_k$ gets below a small threshold (default 0.001) is eliminated as $k$ is too weak. Blank entries of COB tables indicate eliminated dimers.

Second, MODER learns new values $\theta_1^{(t+1)}, \ldots, \theta_p^{(t+1)}$ of monomeric PPMs not from the full data but from dimeric occurrences of the monomer such that the distance $d$ between the components is large enough (default $d \geq \delta = 4$). This is because such isolated occurrences within a dimer are supposed to give the best data for a monomer PPM, not distorted by close-by other sites such as the other component of a dimer. However, if the share of these dimeric cases in the mixture is less than 0.02, then the dimeric data is treated too small. In this case distances $d \geq 0$ are included into the analysis.

The third modification is motivated by the fact that transcription factors may have different binding motifs whose consensus sequences are only a few Hamming steps apart. To minimize disturbance from such similar motifs and from background, MODER tends to restrict the learning of PPMs $\theta_k^{(t+1)}$ and $\psi_k^{(t+1)}$ to high-affinity training sequences. Such sequences are identified by the heuristic rule that they are in small Hamming neighbourhood of the consensus sequences (sequences with highest probability) of the PPMs found so far. Monomer PPM $\theta_k^{(t+1)}$ is learned from data sites that are in the 1-Hamming neighbourhood of the seed (using the consensus sequence as the seed) of $\theta_k^{(t)}$. Bridging PPM $\psi_{k_1k_2od}^{(t+1)}$ is learned from data sites that are in the 1-Hamming neighbourhood of combined seed $s_{k_1k_2od}^{(t)}$. The combined seed is obtained as the initial combined seed $s_{k_1k_2od}^{(1)}$ (see Section EM Iterations) but using the seeds of $\theta_{k_1}^{(t)}$ and $\theta_{k_2}^{(t)}$. MODER uses this seed-guided EM search by default, with the standard search as an option.

## RESULTS

### Generated data

As an initial sanity test we created a model $\eta$, generated a data set using it, and checked that MODER is able to learn $\eta$ back from the generated data. We first created one monomeric PPM and deviation matrices $\kappa_{HH-4}$ and $\kappa_{HT-4}$. From these we constructed a model that had uniform background ($\lambda = 0.71$) and PPMs for homodimers HH 5 ($\lambda = 0.12$), HH –4 (0.08) and HT –4 (0.09). Using this model, we generated 100 000 sequences of length 40 bp. The sequences contained dimeric motifs and background only, no monomeric sequences were included. MODER accurately relearned the model from this data as the learned parameter values deviated from the original at most by 0.036; see Supplementary Figure S1 for details.

### Validation using HT-SELEX data

Next we measured the quality of PPM models produced by MODER using correlation ($R^2$) between occurrence counts and PPM scores of 8–mers or 10–mers of SELEX data. When counting the $k$–mers, all occurrences and both directions were considered. As the score of a $k$-mer $x$ by a single PPM $\rho$ we used the maximum value of $\log \frac{\rho'(y)}{\theta_0(y)}$ when $y$ and $\rho'$ go over all intersections of $\rho$ and $x$ and of $\rho$ and reverse complement $\bar{x}$. As the score of $x$ by a mixture of PPMs $\rho_1, \ldots, \rho_t$, whose mixing parameters by MODER are $\lambda_1, \ldots, \lambda_t$, we used $\lambda_1 S_1 + \cdots + \lambda_t S_t$ where $S_1, \ldots, S_t$ are the individual scores of $x$ by the PPMs. The scatter plots in the figures visualize the counts and scores of different 8- or 10-mers in hexagonal bins. The color of a bin reflects the number of different $k$–mers in that bin, with a darker color meaning higher number of different $k$–mers. As the early cycles of SELEX data can contain large proportion of nonspecific sequences (i.e. background), the counts were corrected against background using the data of the previous SELEX cycle, as described in (5).

We report results for the monomer and dimer PPMs of factors HOXB13, HNF4A, TFAP2A, FLI1, FOXC1 and PKNOX2 learned from HT-SELEX data. A basic correlation analysis is done for factor HOXB13. For HNF4A, TFP2A, FLI1, FOXC1, and PKNOX2 we also analyse the differences between observed and purely modular motifs. In all validations, the SELEX data sets were randomly divided into two halves, one half used for learning the model and other half used for validating it.

We used the following HT-SELEX data sets: HOXB13 (PRJEB14550, 164 768 reads), HNF4A (ERX169045, (6), 655 432 reads), TFAP2A (ERX1085476, (43), 168 053 reads), FLI1 (PRJEB14550, 143 389 reads), FOXC1 (ERX169015, (6), 189 009 reads), and PKNOX2 (ERX1084652, (43), 423 339 reads). Each read was 40 bp long except for FOXC1 whose reads were 30 bp long. The following seeds, selected by hand using the models published in (6), were used as input: HOXB13 (`CTCGTAAAA`, `CCAATAAAA`), HNF4A (`RGGTCA`, `RGTCCA`), TFAP2A (`GGGCA`), FLI1 (`ACCGGAAGTN`), FOXC1 (`RTAAAYA`), and PKNOX2 (`TGACANN`). Note that it was essential to use non-palindromic seeds for overlapping dimers as, for example, the observed cases HH -6 for FLI1 and HH -2 for PKNOX2 are directional; see Figure S3 in (6) and Section EM iterations.

*Selecting strong components of the model.* The learned total model is likely to contain useless, weak components (weak dimeric motifs) that should be removed before the model is applied, e.g. to predict new putative binding sites. One could, for example, include model components in decreasing order of weight $\lambda$ until a certain fraction of the non-background sequences is covered. Here we used the fraction of 85% to select the models for validation experiments. In addition, we also studied the effect of parameter $\delta$ (minimum gap length in the independent case) by experimenting with large values (up to $L_{\max}$) of $\delta$. As all larger deviations from expected were observed to usually occur in dimers with gap <4, default value $\delta = 4$ was selected.

*Factors HOXB12, HNF4A, TFAP2A, FLI1, FOXC1 and PKNOX2.* Figure 3 shows the sequence logos of the learned PPMs for factor HOXB13 and reports correlations of the scores of individual PPMs and of their mixture with counts of 8-mers. Since MODER did not find any strong dimeric motifs, the model for this factor is composed of two monomers only. The power of multi-motif modeling can be seen: the combined mixture consistently gives the highest $R^2$.

HNF4A, TFAP2A, FLI1, FOXC1 and PKNOX2 are examples of TFs for which many dimeric PPMs deviate clearly

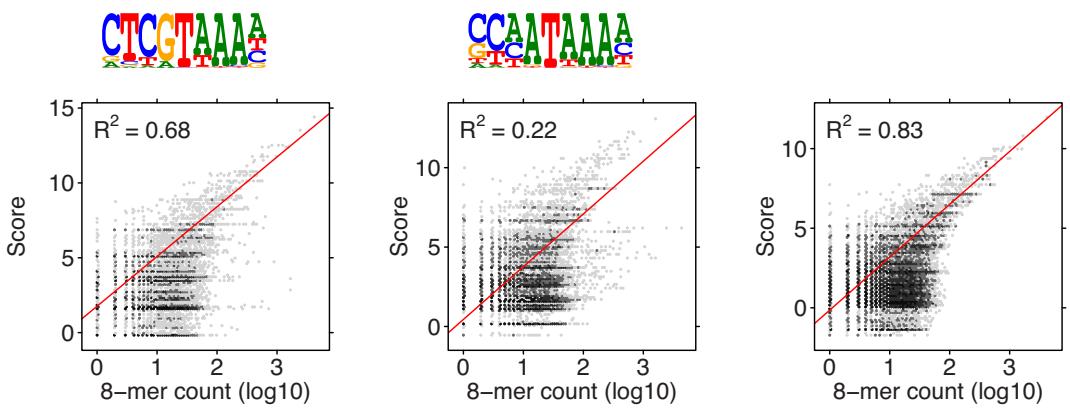


**Figure 3.** Correlation analysis of HOXB13 binding motifs. Two monomer PPMs and their mixture (last panel) with weights $\lambda_1 = 0.187$, $\lambda_2 = 0.198$ were used. No dimer models were included. The combined model has higher correlation than the component models.

from the purely modular PPMs. Analyses of these factors are shown in Figures 4–8, where the correlations are shown for both the expected (purely modular) and the learned models, and the deviation matrices are also visualized. The number of dimeric models included into the mixture by the 85% rule ranged from 3 to 14 for different factors, only the top three dimeric models shown in the Figures. The full set of models, weights, and resulting correlations are available in the Supplementary File S1. Not surprisingly, the learned model has always higher correlation, but with varying margin. Sometimes (TFAP2A) deviating from the expected model gives strongly improved model while sometimes (FOXC1) the difference to the expected model is minor. As for the directionality of the motifs, sometimes both the expected and learned motifs are palindromic (Figure 5) while sometimes expected palindromes become directed in the learned motif (Figures 6 and 8).

### Validation using ChIP-seq data

We then tested for factors HOXB13, HNF4A, and TFAP2A the validity of the obtained *in vitro* models on *in vivo* data. We performed standard ROC analysis to measure the performance of the models learned from SELEX data on binary classification of ChIP-Seq peaks. The following ChIP-seq data was used: HOXB13 (European Nucleotide Archive accession ERX332516, IgG: ERX332513) (44), HNF4A (Sequence Read Archive accession SRR952427, IgG: SRR952608) (45) and TFAP2A (SRR952485,IgG: SRR952608) (45). To find the peaks, the reads were aligned with BWA (46), and peak calling was done with Peakzilla (47). The genome assembly used was GRCh37 (hs37d5). From each ChIP-seq peak set, top $n = 230$ peaks with highest quality score were selected, and for each peak a sequence of length $L = 190$ bp flanking the peak summit was chosen for the positive set. A negative set of the same size was chosen randomly from the human genome, making sure that the positions were mappable. Sequences were scored using the SELEX models of HOXB13, HNF4A and TFAP2A shown in Figures 3, 4and 5. The resulting ROC curves of the (very good) classification performance of PPM scores are shown in Figure 9.

We also applied MODER on the ChIP-seq data set of factor NRSF on the GM12878 cell line produced by (48) and further analyzed by (16). To obtain the seeds, we first took all the *k*-mers of lengths 9–11 from the data set, applied hierarchical clustering, and selected two best clusters and their representative *k*-mers (TTCAGCACC and GGACAGCTCC) by using the order given by the *z*-score. MODER finds 9 out of 10 models reported by Quang and Xie, for details, see Figure 2 of (16) and Supplementary Figure S2. Note that as NRSF has two monomeric motifs, MODER discovers heterodimeric motifs whose COB-table has all four orientations.

Next, we tried to detect the core and side motifs of factor CTCF. This should test MODER's capabilities in detecting long sites, as together these motifs are known to form a site of length about 34 bp (13). We used raw ChIP-exo data from human LoVo cells targeting factor CTCF from Katainen *et al.* (49) (ENA accession ERX986066). Mapping and peak calling was done as in Hartonen *et al.* (50), briefly: alignment was done using BWA (46) against assembly GRCh37 and the peaks were called using PeakXus (50). Five thousand highest scoring peaks were selected, and around the peak summits sequences of length 60 were extracted (blacklisted regions (51), ENCODE accession ENCFF001TDO, and centromeres were removed). Results in Supplementary Figure S3 show that MODER is able to detect similar configurations of distances and orientations between the core and side motifs of CTCF as in Schmidt *et al.* (13) (Figure 2) and Nakahashi *et al.* (14) (Figure 4). The strongest dimer formed by the core and side motifs, namely $\tau_{1,\,2,\,\mathrm{HT},\,8}$, was found in 20% of the top 5000 peaks.

As nuclear receptors commonly bind as dimers (52), we chose another factor, in addition to HNF4A, from this family to display the performance of MODER. ChIP-seq data from ENCODE for factor RXRA in cell line HepG2 was used (ENCODE accession ENCFF002CKZ). Ten thousand highest scoring peaks were selected, and around the peak summits sequences of length 40 were extracted (blacklisted regions and centromeres were removed). The monomer seed GGGGTCA for the experiment was handpicked based on the Rxra mouse model in Jaspar (53). The seed finding method used with NRSF for *k*-mer lengths 6–15 would give AGGTCA, which could have been used as well. Supplementary Figure S4 shows that MODER detects

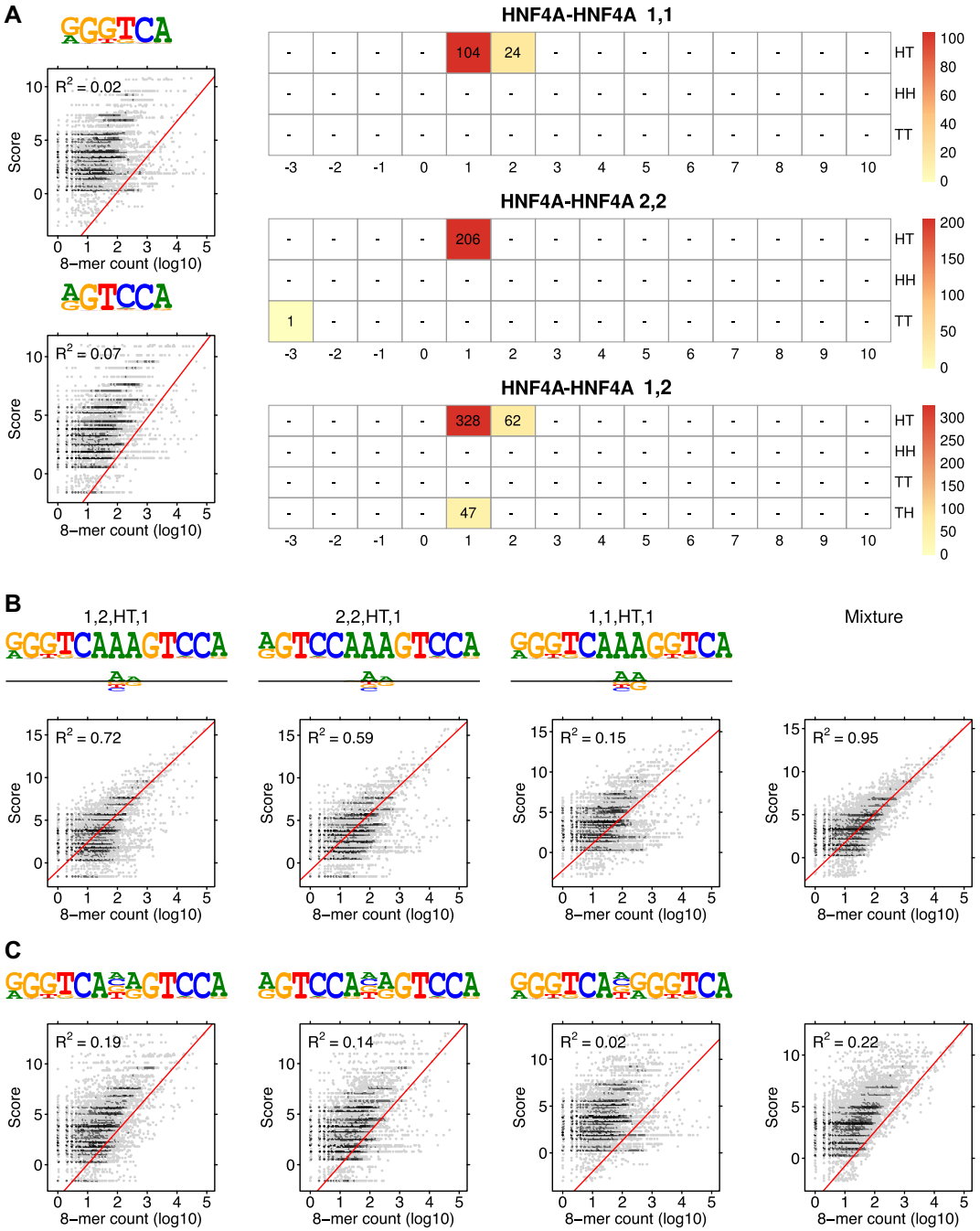


**Figure 4.** Modularity analysis of HNF4A binding motifs. (**A**) Monomer models 1 and 2 ($\lambda_1 = 0.073$, $\lambda_2 = 0.056$) and the COB tables in units of integer multiples of 0.001. Since all the mixing parameters are in the same scale, comparison of λ values is also possible between two distinct COB tables. Also shown is correlation analysis for the two monomer models. (**B**) The first monomer PPM and dimeric models $\tau_{1,2,HT,1}$, $\tau_{2,2,HT,1}$, $\tau_{1,1,HT,1}$, and $\tau_{1,2,HT,2}$ were included in the analysis by the 85% rule (only the best three dimeric models are shown in the Figure). Deviation matrices are depicted below the logos of the dimeric PPMs. Their mixture used corresponding weights λ = 0.073, 0.328, 0.206, 0.104, 0.062. The combined model has much higher correlation than any individual model. (**C**) Correlation analysis as in B but for the PPMs $E_{1,2,HT,1}$, $E_{2,2,HT,1}$, $E_{1,1,HT,1}$ and $E_{1,2,HT,2}$ that are expected under the independence assumption. All $R^2$-values for the learned and expected PPMs differ remarkably, reflecting the large deviations between the learned and the expected PPMs. The purely modular model cannot detect the AAA sequence connecting the half-sites.

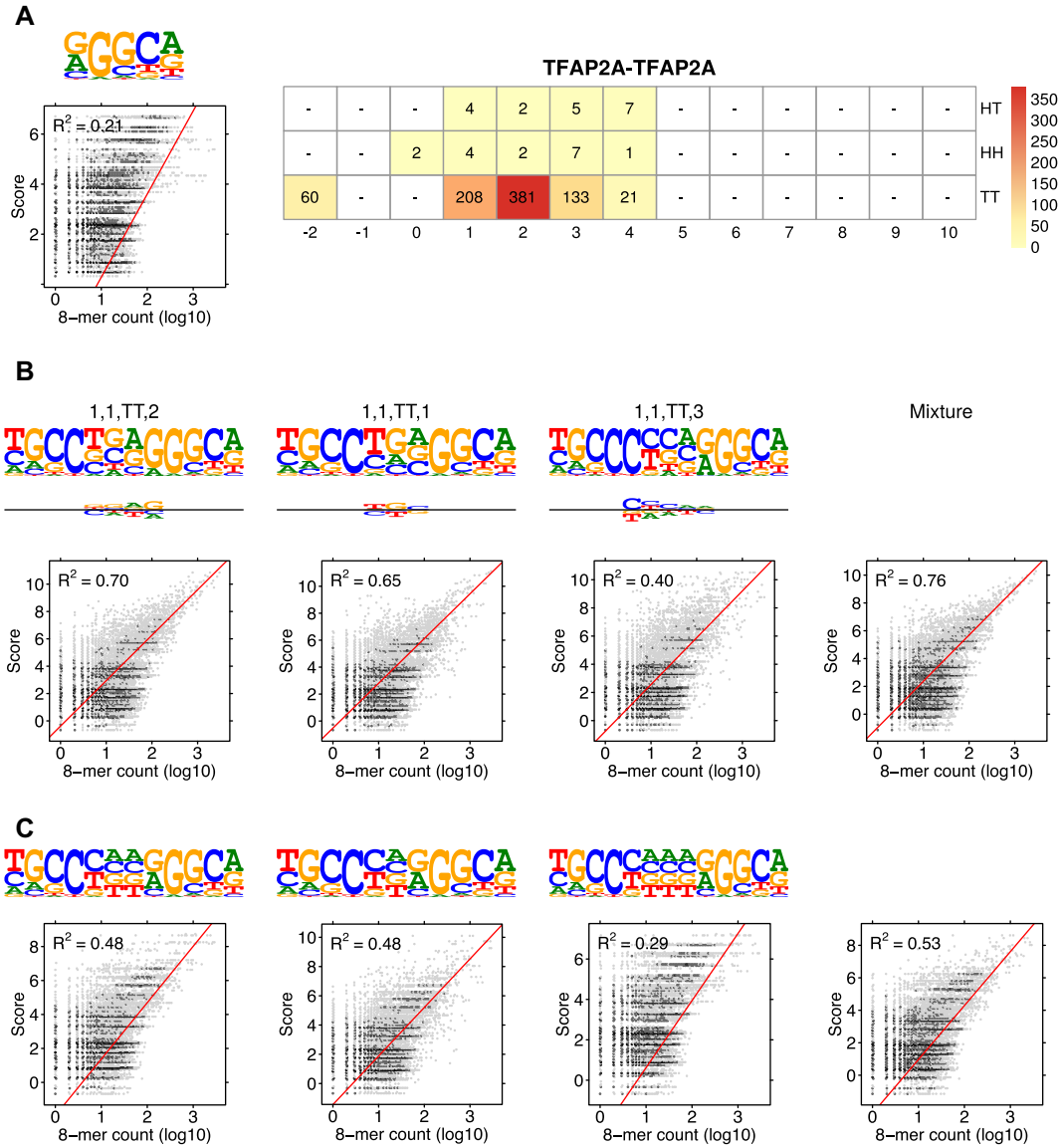


**Figure 5.** Modularity analysis of the binding motifs of TFAP2A. (**A**) The monomer model ($\lambda_1 = 0.003$) and the COB table in units of integer multiples of 0.001. The monomer model was not included in the correlation analysis of the mixture by the 85% rule. (**B**) Correlation analysis of the model learned for TFAP2A by MODER: three dimeric PPMs $\tau_{1,1,TT,2}$, $\tau_{1,1,TT,1}$ and $\tau_{1,1,TT,3}$, and their mixture. Deviation matrices are depicted below the logos of the dimeric PPMs. The mixture of the PPMs uses weights $\lambda = 0.381, 0.208, 0.133$. (**C**) Correlation analysis as in B but for the PPMs $E_{1,1,TT,2}$, $E_{1,1,TT,1}$, $E_{1,1,TT,3}$ that are expected under the independence assumption. All $R^2$-values for the learned and expected PPMs differ remarkably, reflecting the large deviations between the learned and the expected PPMs. It is obvious that the purely modular model is not able to capture the binding affinity of TFAP2A. Note that all three PPMs TT 1, TT 2 and TT 3 that are palindromic in the expected model, stay palindromic in the learned model.

a strong dimeric binding motif $\tau_{1,1,HT,0}$, which could either be a homodimer of RXRA or a heterodimer such as NR1H2-RXRA, as suggested by Tomtom (15).

## MODER versus MEME

When comparing MODER with the popular tool MEME it should be noted that the models of motifs of the two methods are different. MEME learns separate monomer models in successive passes, deleting the found sites of a model from data before the next pass, while MODER aims at discovering the modularity of motifs and hence learns the entire modular structure of monomeric and dimeric motifs in the same probabilistic framework in one run. The difference is illustrated in Supplementary Figure S5 that compares the models learned for factors TFAP2A and FLI1 by the two methods.

## MODER versus Bipad/Maskminent

We also compared MODER with Bipad/Maskminent (17,19) which among the previous tools comes closest to MODER. An example qualitative comparison using alignment of models is illustrated in Supplementary Figure S6. Similarities between motifs obtained using these two algorithms are obvious, although Maskminent seems to in-

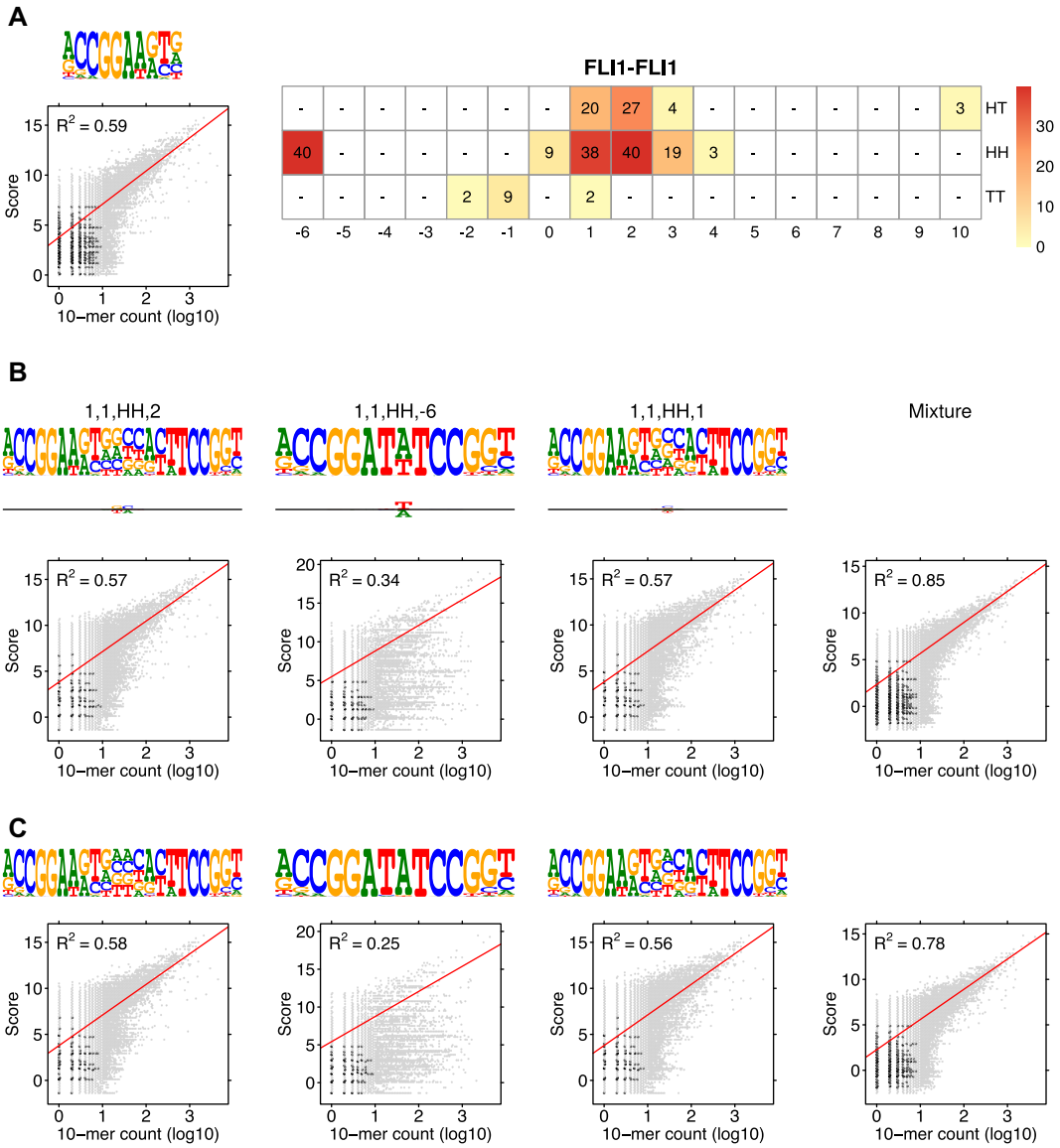


**Figure 6.** Modularity analysis of the binding motifs of FLI1. (**A**) The monomer model ($\lambda_1 = 0.08$) and the COB table in units of integer multiples of 0.001. Also shown is the correlation analysis using only the monomer model. (**B**) Correlation analysis of the model learned for FLI1 by MODER: the monomer model and six dimeric PPMs $\tau_{1,1,HH,2}$, $\tau_{1,1,HH,-6}$, $\tau_{1,1,HH,1}$, $\tau_{1,1,HT,2}$, $\tau_{1,1,HT,1}$, $\tau_{1,1,HH,3}$ (only three best are shown in the Figure) were included in the analysis by the 85% rule. Deviation matrices are depicted below the logos of the dimeric PPMs. The mixture of the PPMs uses weights $\lambda = 0.080$, 0.040, 0.040, 0.038, 0.027, 0.020, 0.019. (**C**) Correlation analysis as in B but for the expected PPMs $E_{1,1,HH,2}$, $E_{1,1,HH,-6}$, $E_{1,1,HH,1}$, $E_{1,1,HT,2}$, $E_{1,1,HT,1}$, $E_{1,1,HH,3}$ under the independence assumption. The $R^2$-values for the learned and expected PPMs differ clearly for the mixture and the dimeric case HH –6. The purely modular model cannot handle the dimeric case HH –6 properly, since expected PPMs are always palindromic for orientations HH and TT, while here the learned model HH –6 turns out to be directed. Albeit quite weak, the directionality has a clear effect on $R^2$.

troduce some background noise into the motifs. Comparison using correlation analysis was not performed since Maskminent does not learn the monomer model and the two orientation classes of dimers (DR and IR) in the same commensurate model, and hence the weights for models of different types could not be decided.

In order to make a quantitative comparison to Maskminent, we used the data sets for which a bipartite Maskminent model is available from Lu *et al.* (19). There were 53 such data sets in ENCODE (51), and we used 40 of those (6 had been revoked from ENCODE, 7 were unidentifiable). The identification problems were due to Lu *et al.* not giving the accession codes, but merely describing the used data sets. For all the data sets we managed to identify, we have now included the accessions in the Supplementary Table S1. The same number of top scoring peaks were used as in (19), but only 100 bp around the peak summit was selected, and these data sets were randomly divided into learning and validation sets of equal size. We selected the initial seeds for MODER based on Jaspar (53) in the following way: for JUN-like factors (ATF3, BACH1, BATF, FOS, FOSL1, FOSL2, JUN, JUNB, JUND, NFE2) we used the seed ATGA, for EBF1 the seed TCCC, for ESR1 the seed AGGTCA, for MAFF and MAFK the seed TCAGCA, and for

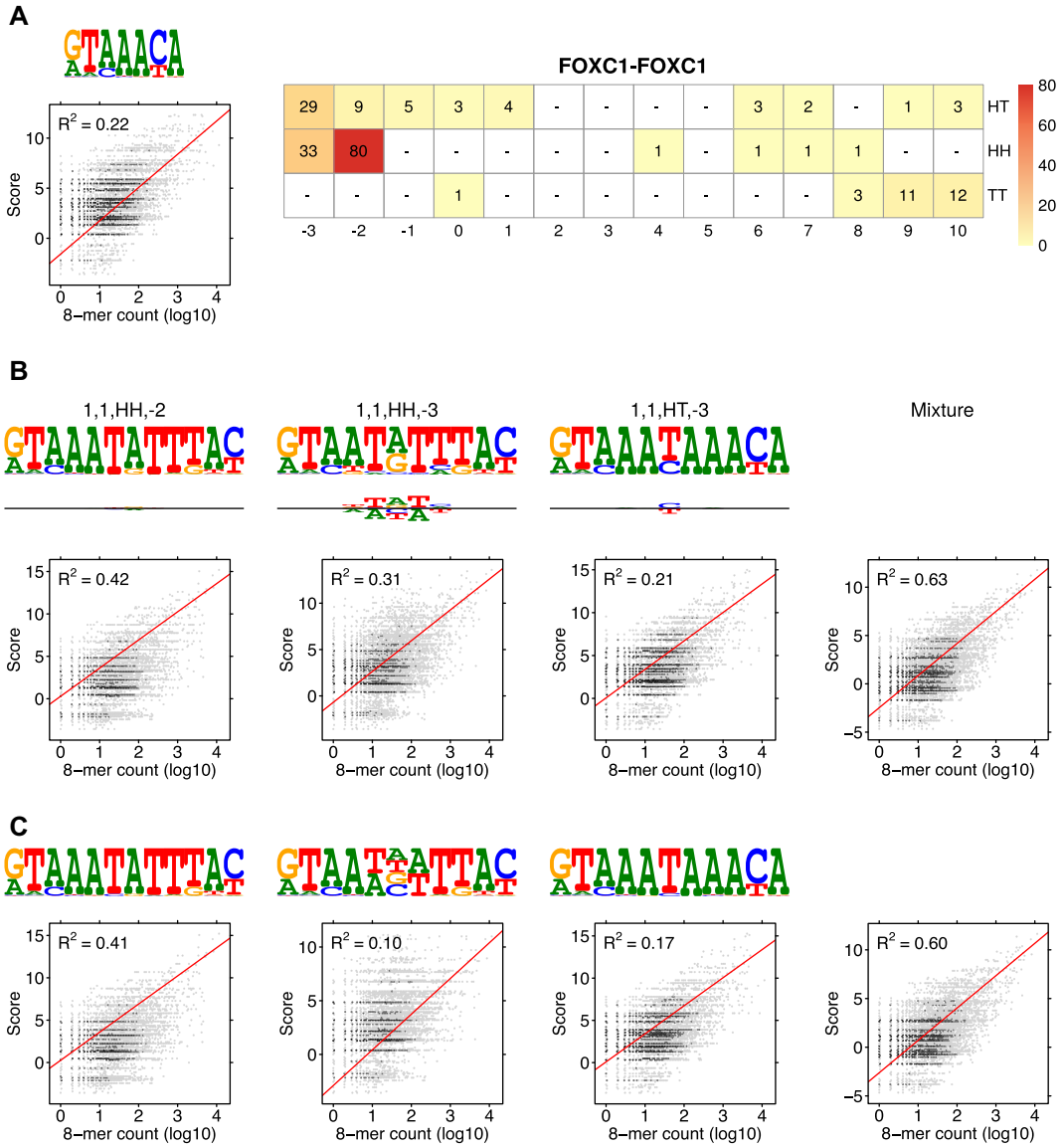


**Figure 7.** Modularity analysis of the binding motifs of FOXC1. (**A**) The monomer model ($\lambda_1 = 0.071$) and the COB table in units of integer multiples of 0.001. Also shown is the correlation analysis using only the monomer model. (**B**) Correlation analysis of the model learned for FOXC1 by MODER: the monomeric model and five dimeric PPMs $\tau_{1,1,\mathrm{HH},-2}$, $\tau_{1,1,\mathrm{HH},-3}$, $\tau_{1,1,\mathrm{HT},-3}$, $\tau_{1,1,\mathrm{TT},10}$, $\tau_{1,1,\mathrm{TT},9}$ were included in the analysis by the 85% rule. Deviation matrices are depicted below the logos of the dimeric PPMs. The mixture of the PPMs uses weights $\lambda = 0.071, 0.080, 0.033, 0.029, 0.012, 0.011$. (**C**) Correlation analysis as in B but for the PPMs $E_{1,1,\mathrm{HH},-2}$, $E_{1,1,\mathrm{HH},-3}$, $E_{1,1,\mathrm{HT},-3}$, $E_{1,1,\mathrm{TT},10}$, $E_{1,1,\mathrm{TT},9}$ that are expected under the independence assumption. The $R^2$-values for the learned and expected PPMs differ quite clearly, for HT –3 and HH –3 in particular, as also suggested by their large deviation matrices, while the difference is small for HH –2, the heaviest component of the mixture. It is obvious that the purely modular model is not able to fully capture the binding affinity of FOXC1. Note that expected palindromic PPM HH –3 becomes directed while expected PPM HH –2 stays palindromic in the learned model.

STAT1 the seed TTC. Then MODER was run on learning data sets, and the best dimeric PPM (according to lambda) was chosen for each data set. For Maskminent we used their published model for each data. The results displayed in Supplementary Table S1 and Figure S7 show that MODER gets better AUC value in 35 cases out of 40. Note also that MODER wins in 33 out of 35 cases, when considering only the optimal IDR-thresholded data sets (the other five data sets are initial peaksets, marked with a star in the table). The selection of factors used by Lu *et al.* (19) was unfortunately quite repetitive, but the comparison shows consistent behaviour for both methods.

## DISCUSSION

The MODER algorithm is based on reductionist view that PPM models for dimers can be built in a modular fashion from monomer PPMs.

As noted e.g. by (4), such modularity is not always valid as in a number of dimeric cases the specificity of the dimeric motif differs notably from what could be expected from its

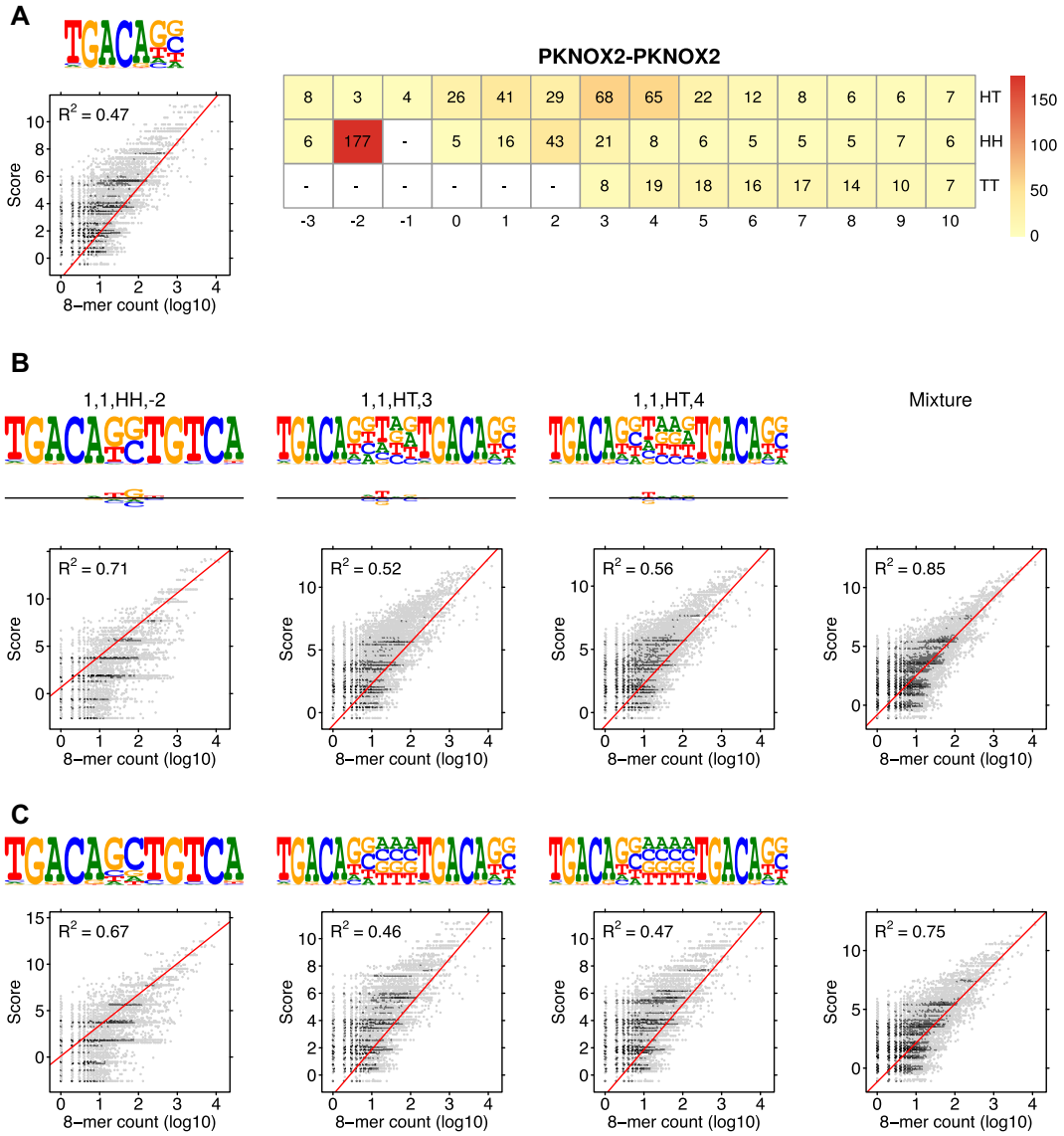


**Figure 8.** Modularity analysis of the binding motifs of PKNOX2. (**A**) The monomer model ($\lambda_1 = 0.256$) and the COB table in units of integer multiples of 0.001. Also shown is the correlation analysis using only the monomer model. (**B**) Correlation analysis of the model learned for PKNOX2 by MODER: 14 dimeric PPMs were included in the analysis by the 85% rule but only the best three, $\tau_{1,1,HH,-2}$, $\tau_{1,1,HT,3}$, and $\tau_{1,1,HT,4}$, are described here. In the mixture the weight for the monomer and three best dimeric models were $\lambda = 0.256, 0.177, 0.068, 0.065$. The sum of the lambdas for the last 11 dimeric models was 0.267. Deviation matrices are depicted below the logos of the dimeric PPMs. (**C**) Correlation analysis as in A but for the PPMs $E_{1,1,HH,-2}$, $E_{1,1,HT,3}$, $E_{1,1,HT,4}$ that are expected under the independence assumption. Here the 85% rule selected very many dimeric models, because the lambda values have quite even distribution in this case. Again the $R^2$-values for the learned and expected PPMs differ quite clearly. It can be seen that the purely modular model is already satisfactory but can be improved somewhat by allowing deviations. Note that the palindromic HH -2 motif of the expected model becomes directed in the learned model.

monomeric components. The deviation matrix of MODER represents such differences explicitly. These deviations from the expected models are especially important for orientations HH and TT, for which all expected models are always symmetric (palindromes), whereas the real binding motifs might have a direction (6). This was demonstrated by several examples in our validation experiments. In addition, the deviations from expected motif commonly occur when the core segments of the motifs of two factors are closely packed and the overlapping flanks are distorted from the expected model. In TF-DNA binding, the core positions in a motif are usually recognized by direct bonds to the bases, whereas the weaker positions are recognized by contacts to DNA backbone (4) and are hence more prone to deviations.

The motif discovery algorithm of MODER considers simultaneously all possible orientation–distance pairs and finds the preferred dimeric motifs. Learning multiple motifs in serial manner—first finding one motif, then removing its occurrences from the data, and then running the algorithm again—does not treat symmetrically the sequences that may belong to several motifs. MODER improves over the similar coMOTIF algorithm (30) by including the spacing information in the overall model, and by adding overlapping motifs and the deviations from the expected motif. Allow-

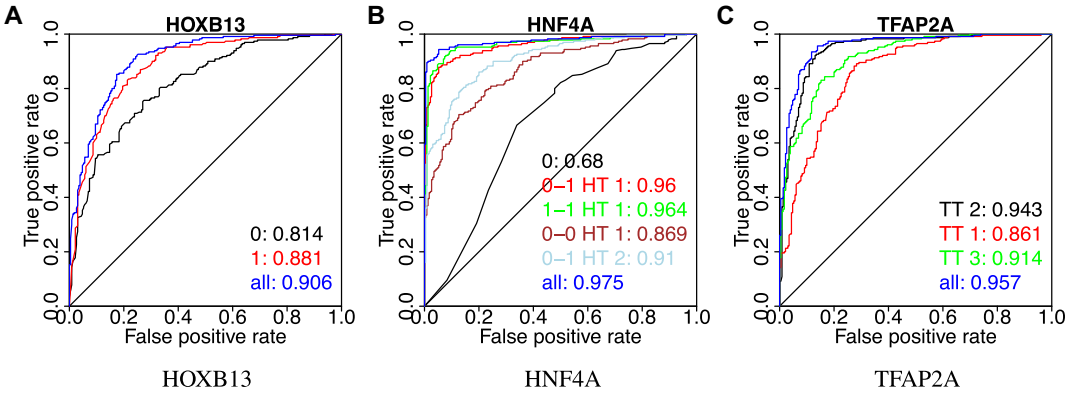


**Figure 9.** Classification performance of multi-motif models for HOXB13, HNF4A, and TFP2A. ROC curves of classification using the scores by PPMs and their mixtures on ChIP-seq data. The area under each curve is also shown. Same models were used as in the correlation analysis (Figures 3–5).

ing overlaps of monomer motifs within a dimer turned out a very useful feature. In fact, for factors FLI1, FOXC1 and PKNOX2 the strongest dimer has such an overlap.

Simultaneous learning of all motif components and their mixing parameters makes direct comparison of the relative strengths of the motifs possible by using the mixing parameters. Depending on the application, it might be useful to rescale the obtained mixing parameters, after the actual algorithm is finished. This was done, when we chose the motifs for performance testing by the 85% rule: the mixing parameters were rescaled to exclude the background. Then motifs were included in descending order, until the motifs covered 85% of the signal. Sometimes it might also be useful to rescale the mixing parameters in each COB table separately, although this would prevent the comparison of mixing parameters between distinct COB tables.

MODER is not too sensitive to noise in the seeds. For factor HOXB13, we mutated the first initial seed in two positions and the second seed in three positions, including informative positions. Still the algorithm managed to obtain the same results as with the original seeds. MODER is reasonably fast. For example, it took 2 min 18 s wall-clock time and 15 min 30 s CPU time when run simultaneously on eight cores to learn the model for FLI1 in Figure 2 from a 2 865 880 bp long HT-SELEX data set. The seeds for MODER can be found from existing PPM databases or can be produced by seed-finding tools such as DREME (54) or by using the procedure in section Validation using ChIP-seq data to find representative *k*-mers.

## AVAILABILITY

MODER is implemented in C++ on Linux platform and is available from https://github.com/jttoivon/MODER. European Nucleotide Archive, accession code PRJEB14550.

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

## ACKNOWLEDGEMENTS

The authors would like to thank Tuomo Hartonen for doing the peak calling from the ChIP-seq and ChIP-exo reads.

## FUNDING

EU FP7 project SYSCOL [UE7-SYSCOL-258236]; Leverhulme Trust [VP1-2014-044 to E.U.]. Funding for open access charge: University of Helsinki.
*Conflict of interest statement.* None declared.

## REFERENCES

1. Rodda,D.J., Chew,J.-L., Lim,L.-H., Loh,Y.-H., Wang,B., Ng,H.-H. and Robson,P. (2005) Transcriptional regulation of nanog by OCT4 and SOX2. *J. Biol. Chem.*, **280**, 24731–24737.
2. Panne,D., Maniatis,T. and Harrison,S.C. (2007) An atomic model of the interferon-β enhanceosome. *Cell*, **129**, 1111–1123.
3. De Val,S., Chi,N.C., Meadows,S.M., Minovitsky,S., Anderson,J.P., Harris,I.S., Ehlers,M.L., Agarwal,P., Visel,A., Xu,S.-M. *et al.* (2008) Combinatorial regulation of endothelial gene expression by ETS and Forkhead transcription factors. *Cell*, **135**, 1053–1064.
4. 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.
5. Jolma,A., Kivioja,T., Toivonen,J., Cheng,L., Wei,G., Enge,M., Taipale,M., Vaquerizas,J.M., Yan,J., Sillanpää,M.J. *et al.* (2010) Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res.*, **20**, 861–873.
6. Jolma,A., Yan,J., Whitington,T., Toivonen,J., Nitta,K.R., Rastas,P., Morgunova,E., Enge,M., Taipale,M., Wei,G. *et al.* (2013) DNA-binding specificities of human transcription factors. *Cell*, **152**, 327–339.
7. Valouev,A., Johnson,D.S., Sundquist,A., Medina,C., Anton,E., Batzoglou,S., Myers,R.M. and Sidow,A. (2008) Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat. Methods*, **5**, 829–834.
8. Isakova,A., Berset,Y., Hatzimanikatis,V. and Deplancke,B. (2016) Quantification of cooperativity in heterodimer-DNA binding improves the accuracy of binding specificity models. *J. Biol. Chem.*, **291**, 10293–10306.
9. Stormo,G.D., Schneider,T.D. and Gold,L. (1986) Quantitative analysis of the relationship between nucleotide sequence and functional activity. *Nucleic Acids Res.*, **14**, 6661–6679.
10. Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.
11. LaRonde-LeBlanc,N.A. and Wolberger,C. (2003) Structure of HoxA9 and Pbx1 bound to DNA: Hox hexapeptide and DNA recognition anterior to posterior. *Genes Dev.*, **17**, 2060–2072.
12. Dempster,A.P., Laird,N.M. and Rubin,D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B*, **39**, 1–38.
13. Schmidt,D., Schwalie,P.C., Wilson,M.D., Ballester,B., Gonçalves,A., Kutter,C., Brown,G.D., Marshall,A., Flicek,P. and Odom,D.T. (2012) Waves of retrotransposon expansion remodel genome organization

and CTCF binding in multiple mammalian lineages. *Cell*, **148**, 335–348.
14. Nakahashi,H., Kieffer Kwon,K.-R., Resch,W., Vian,L., Dose,M., Stavreva,D., Hakim,O., Pruett,N., Nelson,S., Yamane,A. *et al.* (2013) A genome-wide map of CTCF multivalency redefines the CTCF code. *Cell Rep.*, **3**, 1678–1689.
15. Gupta,S., Stamatoyannopoulos,J.A., Bailey,T.L. and Noble,W.S. (2007) Quantifying similarity between motifs. *Genome Biol.*, **8**, R24.
16. Quang,D. and Xie,X. (2014) EXTREME: an online EM algorithm for motif discovery. *Bioinformatics*, **30**, 1667–1673.
17. Bi,C. and Rogan,P.K. (2004) Bipartite pattern discovery by entropy minimization-based multiple local alignment. *Nucleic Acids Res.*, **32**, 4979–4991.
18. Bi,C., Leeder,J.S. and Vyhlidal,C.A. (2008) A comparative study on computational two-block motif detection: algorithms and applications. *Mol. Pharm.*, **5**, 3–16.
19. Lu,R., Mucaki,E.J. and Rogan,P.K. (2017) Discovery and validation of information theory-based transcription factor and cofactor binding site motifs. *Nucleic Acids Res.*, **45**, e27.
20. Helden,J.v., Rios,A. and Collado-Vides,J. (2000) Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucleic Acids Res.*, **28**, 1808–1818.
21. Li,H., Rhodius,V., Gross,C. and Siggia,E.D. (2002) Identification of the binding sites of regulatory proteins in bacterial genomes. *Proc. Natl. Acad. Sci. U.S.A.*, **99**, 11772–11777.
22. Liu,X., Brutlag,D.L. and Liu,J.S. (2001) BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pacific Symp. Biocomput.*, **6**, 127–138.
23. Whitington,T., Frith,M.C., Johnson,J. and Bailey,T.L. (2011) Inferring transcription factor complexes from ChIP-seq data. *Nucleic Acids Res.*, **39**, e98.
24. Kazemian,M., Pham,H., Wolfe,S.A., Brodsky,M.H. and Sinha,S. (2013) Widespread evidence of cooperative DNA binding by transcription factors in Drosophila development. *Nucleic Acids Res.*, **41**, 8237–8252.
25. Jankowski,A., Prabhakar,S. and Tiuryn,J. (2014) TACO: a general-purpose tool for predicting cell-type–specific transcription factor dimers. *BMC Genomics*, **15**, 1–12.
26. Lawrence,C.E. and Reilly,A.A. (1990) An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins: Struct. Funct. Bioinformatics*, **7**, 41–51.
27. Cardon,L.R. and Stormo,G.D. (1992) Expectation maximization algorithm for identifying protein-binding sites with variable lengths from unaligned DNA fragments. *J. Mol. Biol.*, **223**, 159–170.
28. Bailey,T.L. and Elkan,C. (1995) The value of prior knowledge in discovering motifs with MEME. In: *Proc. Third Internat. Conf. on Intelligent Systems for Molecular Biology*, AAAI Press, pp. 21–29.
29. Bailey,T.L., Boden,M., Buske,F.A., Frith,M., Grant,C.E., Clementi,L., Ren,J., Li,W.W. and Noble,W.S. (2009) MEME suite: tools for motif discovery and searching. *Nucleic Acids Res.*, **37**(Suppl. 2), W202–W208.
30. Xu,M., Weinberg,C.R., Umbach,D.M. and Li,L. (2011) coMOTIF: a mixture framework for identifying transcription factor and a coregulator motif in ChIP-seq Data. *Bioinformatics*, **27**, 2625–2632.
31. Li,L. (2009) GADEM: a genetic algorithm guided formation of spaced dyads coupled with an EM algorithm for motif discovery. *J. Comput. Biol.*, **16**, 317–329.
32. Mercier,E., Droit,A., Li,L., Robertson,G., Zhang,X. and Gottardo,R. (2011) An integrated pipeline for the genome-wide analysis of transcription factor binding sites from ChIP-Seq. *PLoS One*, **6**, e16432.
33. Zhang,Z., Chang,C.W., Hugo,W., Cheung,E. and Sung,W.-K. (2013) Simultaneously learning DNA motif along with its position and sequence rank preferences through expectation maximization algorithm. *J. Comput. Biol.*, **20**, 237–248.
34. Reid,J.E. and Wernisch,L. (2014) STEME: a robust, accurate motif finder for large data sets. *PLoS One*, **9**, e90735.
35. Liu,J.S., Neuwald,A.F. and Lawrence,C.E. (1995) Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. *J. Am. Stat. Assoc.*, **90**, 1156–1170.
36. Ikebata,H. and Yoshida,R. (2015) Repulsive parallel MCMC algorithm for discovering diverse motifs from large sequence sets. *Bioinformatics*, **31**, 1561–1568.
37. Alipanahi,B., Delong,A., Weirauch,M.T. and Frey,B.J. (2015) Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. *Nature Biotech.*, **33**, 831–838.
38. Colombo,N. and Vlassis,N. (2015) FastMotif: spectral sequence motif discovery. *Bioinformatics*, **31**, 2623–2631.
39. 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.
40. Kulakovskiy,I.V., Boeva,V., Favorov,A.V. and Makeev,V.J. (2010) Deep and wide digging for binding motifs in ChIP-Seq data. *Bioinformatics*, **26**, 2622–2623.
41. Ma,W., Noble,W.S. and Bailey,T.L. (2014) Motif-based analysis of large nucleotide data sets using MEME-ChIP. *Nat. Protoc.*, **9**, 1428–1450.
42. Jayaram,N., Usvyat,D. and Martin,A.C. (2016) Evaluating tools for transcription factor binding site prediction. *BMC Bioinformatics*, **17**, 1298.
43. Yin,Y., Morgunova,E., Jolma,A., Kaasinen,E., Sahu,B., Khund-Sayeed,S., Das,P.K., Kivioja,T., Dave,K., Zhong,F. *et al.* (2017) Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science*, **356**, eaaj2239.
44. Huang,Q., Whitington,T., Gao,P., Lindberg,J.F., Yang,Y., Sun,J., Väisänen,M.-R., Szulkin,R., Annala,M., Yan,J. *et al.* (2014) A prostate cancer susceptibility allele at 6q22 increases RFX6 expression by modulating HOXB13 chromatin binding. *Nat. Gen.*, **46**, 126–135.
45. Yan,J., Enge,M., Whitington,T., Dave,K., Liu,J., Sur,I., Schmierer,B., Jolma,A., Kivioja,T., Taipale,M. *et al.* (2013) Transcription factor binding in human cells occurs in dense clusters formed around cohesin anchor sites. *Cell*, **154**, 801–813.
46. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
47. Bardet,A.F., Steinmann,J., Bafna,S., Knoblich,J.A., Zeitlinger,J. and Stark,A. (2013) Identification of transcription factor binding sites from ChIP-seq data at high resolution. *Bioinformatics*, **29**, 2705–2713.
48. ENCODE Project Consortium, Birney,E., Stamatoyannopoulos,J.A., Dutta,A., Guigó,R., Gingeras,T.R., Margulies,E.H., Weng,Z., Snyder,M., Dermitzakis,E.T., Thurman,R.E. *et al.* (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.
49. Katainen,R., Dave,K., Pitkänen,E., Palin,K., Kivioja,T., Välimäki,N., Gylfe,A.E., Ristolainen,H., Hänninen,U.A., Cajuso,T. *et al.* (2015) CTCF/cohesin-binding sites are frequently mutated in cancer. *Nature Genetics*, **47**, 818–821.
50. Hartonen,T., Sahu,B., Dave,K., Kivioja,T. and Taipale,J. (2016) PeakXus: comprehensive transcription factor binding site discovery from ChIP-Nexus and ChIP-Exo experiments. *Bioinformatics*, **32**, i629–i638.
51. ENCODE,Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57.
52. Gronemeyer,H., Gustafsson,J.-A. and Laudet,V. (2004) Principles for modulation of the nuclear receptor superfamily. *Nat. Rev. Drug. Discov.*, **3**, 950–964.
53. Mathelier,A., Fornes,O., Arenillas,D.J., Chen,C.-y., Denay,G., Lee,J., Shi,W., Shyr,C., Tan,G., Worsley-Hunt,R. *et al.* (2016) JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **44**, D110–D115.
54. Bailey,T.L. (2011) DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics*, **27**, 1653–1659.