Detection of cooperatively bound transcription factor pairs using ChIP-seq peak intensities and expectation maximization

Vishaka Datta*[1], Rahul Siddharthan[2], and Sandeep Krishna[1]

[1]Simons Centre for the Study of Living Machines, National Centre for Biological Sciences, TIFR, Bengaluru 560065, India
[2]The Institute of Mathematical Sciences/HBNI, Taramani, Chennai 600 113, India

April 10, 2017

Abstract

Transcription factors (TFs) often work cooperatively, where the binding of one TF to DNA enhances the binding affinity of a second TF to a nearby location. Such cooperative binding is important for activating gene expression from promoters and enhancers in both prokaryotic and eukaryotic cells. Existing methods to detect cooperative binding of a TF pair rely on analyzing the sequence that is bound. We propose a method that uses, instead, only ChIP-Seq peak intensities and an expectation maximization (CPI-EM) algorithm. We validate our method using ChIP-seq data from cells where one of a pair of TFs under consideration has been genetically knocked out. Our algorithm relies on our observation that cooperative TF-TF binding is correlated with weak binding of one of the TFs, which we demonstrate in a variety of cell types, including *E. coli*, *S. cerevisiae*, *M. musculus*, as well as human cancer and stem cell lines. We show that this method performs significantly better than a predictor based only on the ChIP-seq peak distance of the TFs under consideration. By explicitly avoiding the use of sequence information, our method may help uncover new sequence patterns of cooperative binding that sequence-based methods could build upon. The CPI-EM algorithm is available at https://github.com/vishakad/cpi-em.

1 Introduction

Transcription factors (TFs) regulate the transcription of a set of genes by binding specific regulatory regions of DNA. The magnitude of the change in transcription caused by a TF depends in part on its affinity to the DNA sequence bound. It is possible that a second TF binding a nearby sequence increases the first TF's binding affinity. In this case, the two TFs are said to cooperatively or combinatorially bind DNA [1]. The cooperative binding of transcription factors at enhancers and promoters is known to strongly increase gene expression [2, 3, 4, 5]. The presence of cooperativity has been used to explain the rapid rate of evolution of TF binding sites in multicellular organisms [6].

The role of cooperative binding in protein complex assembly has been extensively studied and computational methods have been proposed to detect such interactions within genomes [7, 8, 9]. In these studies, cooperativity typically involves protein oligomerization. However, two TFs can also cooperatively bind DNA without dimerizing prior to binding [10, 11]. Several theoretical methods have been proposed to detect cooperative binding between a pair of TFs in the genome [12, 13, 14, 15, 16, 17, 18, 19, 1]. These methods typically rely on locating frequently co-occurring binding sites of TF pairs across the genome, or within genomic sequences known to be bound by a TF pair. However, co-occurring binding site pairs do not always imply cooperative binding [1]. Conversely, many TF pairs can cooperatively bind DNA even if the spacing between their binding sites, or the sequence between them, is changed [20, 21]. Thus, a pair of TFs that cooperatively bind DNA at one genomic region may not bind cooperatively at a different genomic region.

*To whom correspondence should be addressed. Email: vishakad@ncbs.res.in

Here, we propose a sequence-independent algorithm, based on ChIP-seq (chromatin immuno-precipitation and sequencing) data, for detecting cooperatively bound sites that complement these sequence-based methods.

Genome-wide TF-DNA binding has been extensively studied using ChIP-seq [22]. ChIP-seq provides a list of locations bound by a TF across a genome *in vivo*, which are referred to as *peaks*, along with peak *intensities* whose values are proportional to the TF's affinity for the sequence bound at these locations [22]. Some ChIP-seq experiments, designed to detect pair-wise TF interactions across the genome, have been carried out in *E. coli*, *S. cerevisiae*, *M. musculus*, and human genomes [23, 1, 24, 25]. In these experiments, three sets of ChIP-seq are performed to determine locations where a pair of TFs, A and B, are cooperatively bound. First, two ChIP-seq experiments are performed to determine binding locations of A and B in cells. A third ChIP-seq is performed to find binding locations of A after B is genetically knocked out. In this ChIP-seq, locations where A no longer binds DNA or has a lower binding affinity towards DNA are considered to be instances of cooperative binding. We refer to such a set of three experiments as A-B, and refer to A as the *primary* TF and B as the *partner* TF. Instead of knocking out B, if a ChIP-seq is performed to find binding locations of B after A is knocked out, we can infer locations where B is cooperatively bound by A. This dataset is labeled B-A, with B and A as primary and partner TFs, respectively.

We propose the ChIP-seq Peak Intensity - Expectation Maximisation (CPI-EM) algorithm as a computational method to detect genomic locations cooperatively bound by a TF pair, based on their ChIP-seq peak intensities. CPI-EM can do this without the need for ChIP-seq to be performed on one of the TFs after the other is knocked out. At each location where ChIP-seq peaks of two TFs overlap each other, CPI-EM computes a probability that the location is cooperatively bound by both TFs. The highlight of this algorithm is that it utilizes only peak intensities to detect cooperative binding, and does not rely on binding site searches within ChIP-seq peak regions. CPI-EM relies on the observation that a primary TF tended to be more weakly bound when it cooperatively bound DNA with a partner TF, in comparison to regions where it did not cooperatively bind DNA. We observed this to be the case in ChIP-seq datasets we analyzed from *E. coli*, *S. cerevisiae*, *M. musculus* and human genomes.

We compared the set of locations predicted by CPI-EM to be cooperatively bound, with the locations obtained from the knockout-based ChIP-seq experiments. We also compared CPI-EM with an algorithm that detects cooperative binding based on the distance between ChIP-seq peaks. We found that peak distance by itself was not a reliable predictor of cooperative binding. In contrast, we found that peak intensities are a more reliable criterion to detect cooperative interactions in all the ChIP-seq datasets we analyzed.

2 Methods

2.1 ChIP-seq processing pipeline

A single ChIP-seq "peak call" consists of the genomic coordinates of the location being bound, along with a *peak intensity*. We determined ChIP-seq peak locations of different transcription factors from multiple genomes, namely, *E. coli* (GSE92255), *S. cerevisiae* [1], cells from primary *M. musculus* liver tissue [23], and three human cell lines – the Caco-2 intestinal stem cell line [24], the T-47D breast cancer cell line and the ECC-1 endometrial cancer cell line [25]. We used our own ChIP-seq pipeline to process raw sequence reads and call peaks from *M. musculus* and *S. cerevisiae* data, and utilized pre-computed peak calls with the remaining datasets. This ensured that our validation sets were not biased by procedures employed in our pipeline. See Supplementary Section 1 for details of our ChIP-seq pipeline for processing these datasets.

2.2 Using ChIP-seq data from a genetic knockout to infer cooperative binding

From ChIP-seq profiles of a pair of TFs, X and Y, we classified genomic regions containing overlapping ChIP-seq peaks of X and Y as cooperative or non-cooperative, based on the change in peak rank of X in response to a genetic deletion of Y. The ranks are assigned such that the peak with rank 1 has the highest peak intensity. In our analysis, we consider a genomic region to be doubly bound by X and Y if their peak regions overlap by at least a single base pair. We used pybedtools v0.6.9 [26] to find these overlapping peak regions.

At each doubly bound genomic location, we classify X as being cooperatively bound by Y if (a) the peak rank of X in the presence of Y is significantly higher (i.e., closer to rank 1) than the peak rank of X measured after the deletion of Y, or (b) if X's peak is absent after the deletion of Y.

On the other hand, if the peak rank of X in the presence of Y is significantly lower (i.e., further from rank 1) than the peak rank of X after the deletion of Y, or if it stays the same, we classify this as competitive or independent binding, respectively. We refer to both these classes as non-cooperative binding. See Supplement Section 3 for details on the statistical tests we performed to detect significant changes in peak ranks of X upon the knockout of Y. These tests require ChIP-seq data from multiple replicates. In the ERα-FOXA1, CDX2-HNF4A, CRP-FIS, and FIS-CRP datasets, peak calls from individual replicates were not available, therefore we used only peak losses to find cooperatively bound locations in these datasets.

2.3 The ChIP-seq Peak Intensity - Expectation Maximisation (CPI-EM) algorithm

We describe the working of the CPI-EM algorithm in step-wise fashion below, where each of the steps is numbered according to Figure 1. In Figure 1 and in the description below, we assume that cooperative binding between TFs X and Y is being studied, where X is the primary TF and Y is the partner TF.

Step 1: From the ChIP-seq of X and Y, find all pairs of peaks where X and Y overlap by at least one base pair. With these overlapping pairs, make a list of peak intensities $(x_1, y_1), (x_2, y_2)...(x_n, y_n)$, where x_i and y_i are the peak intensities of the $i - th$ peak of X and Y, respectively. This list of peak intensity pairs is the input data for the CPI-EM algorithm.

Step 2: To this input data, fit a model of the joint probability $p(x, y)$ of observing the peak intensity x and y from TFs X and Y, respectively, at a given location. Our model consists of a sum of two probability functions, which are the probability of observing intensities x and y if they were (a) cooperatively bound, or (b) non-cooperatively bound. We assume that both probability functions that are fitted have a Lognormal shape. This shape is characterized by four parameters — a mean and a variance of the X and the Y axes (we also examine other shapes such as the Gamma or Gaussian functions — see Supplementary Section 4). A final ninth parameter sets the relative weight of the two probability functions, which determines the fraction of overlapping pairs that are cooperatively bound. We find the best fit for these nine parameters using a procedure called expectation maximization (described in detail in Supplementary Section 4.1).

We make two other assumptions in this step, each of which is discussed further in Supplementary Section 4.

- The peak intensities of X and Y at a location are statistically independent, irrespective of whether X and Y are cooperatively or non-cooperatively bound. We found this to be a reasonable assumption after we measured the mutual information between peak intensities of X and Y from cooperatively and non-cooperatively bound locations. Mutual information is known to be a robust measure of statistical dependence [27].

- A primary TF that is cooperatively bound to DNA is, on average, bound weaker than a non-cooperatively bound primary TF. We found this assumption to hold across all the datasets on which we ran CPI-EM (see section "Peak intensities of cooperatively bound primary TFs are weaker than non-cooperatively bound primary TFs" in Results, and Figure 2).

Step 3: Given the best-fit parameters, use Bayes' formula to calculate the probability for each overlapping pair of ChIP-seq peaks to be a site of cooperative binding (see Supplementary Section 4).

Step 4: Choose a threshold probability α and label an overlapping pair as cooperatively bound if the probability calculated in step 3 is greater than α, and as being non-cooperatively bound otherwise. Validate with a list of known cooperative binding sites, e.g., derived from the ChIP-seq of X after Y is knocked out (as described in the previous section).

2.4 Peak Distance Detector

For each peak intensity pair in the input data, the peak distance detector calculates the distance between the *summits* of X and Y peak regions. The summit is a location within each peak region that has the highest number of sequence reads that overlap it, and is typically the most likely site at which the TF is physically attached to DNA. The peak distance detector declares doubly bound regions as cooperatively bound if the distance between peaks of X and Y is lesser than a threshold distance d. We ran this detection algorithm on all the datasets on which CPI-EM was employed to detect cooperative binding. Our goal in using this algorithm was to determine whether the distance between peaks is a reliable criterion to discriminate between cooperative and non-cooperative binding.

3 Results

3.1 Peak intensities of cooperatively bound primary TFs are weaker than non-cooperatively bound primary TFs

We inferred cooperative and non-cooperative binding from ChIP-seq datasets of FIS-CRP and CRP-FIS pairs in *E. coli* in early-exponential and mid-exponential growth phases (accession number GSE92255), GCN4-RTG3 and RTG3-GCN4 in *S. cerevisiae* [1], FOXA1-HNF4A, FOXA1-CEBPA and HNF4A-CEBPA in the mouse (*M. musculus*) liver [23], CDX2-HNF4A in differentiated human intestinal stem cell lines (Caco-2) [24], and ERα-FOXA1 in the T-47D breast cancer cell line [25]. A summary of the data is shown in Supplementary Table 1.

Figure 2A-C summarize trends in cooperative and non-cooperative TF-DNA binding seen in these datasets. Cooperatively and non-cooperatively bound locations were determined using ChIP-seq data from genetic knockouts as discussed in Methods. Cooperatively bound primary TF peak intensities were significantly lower than those of non-cooperatively bound primary TF peaks across each of the TF-TF pairs (Wilcoxon rank-sum test, $p \ll 0.001$). In contrast, there was no consistent trend in the intensities of the partner TF in each of these pairs. This meant that a primary TF could be cooperatively bound to DNA irrespective of the peak intensity of the partner TF. In Figure 2B, estimates of the marginal distributions of cooperative and non-cooperative FOXA1 and HNF4A peaks are shown. These, and similar distributions for the other TF pairs, tended to be better approximated by a Lognormal distribution, which was evident from the higher log-likelihood value associated with a Lognormal fit, compared to a Gaussian or Gamma distribution (Supplementary Table 2).

Since the primary TF intensity distributions from cooperatively bound regions significantly differed from those of non-cooperatively bound regions, it should be possible to accurately label a pair of overlapping peaks as cooperative or non-cooperative, based solely on their peak intensities. For instance, in the FOXA1-HNF4A dataset, a FOXA1 peak that has an intensity value of 5 is $\approx$3.4 times more likely to be cooperatively bound with HNF4A than to be non-cooperatively bound with it. In clear-cut cases such as these, knowledge of the underlying sequence that is bound is not necessary to detect a cooperative interaction.

3.2 CPI-EM applied to ChIP-seq datasets from *M. musculus*, *S. cerevisiae* and *E. coli*

The ChIP-seq Peak Intensity - Expectation Maximisation (CPI-EM) algorithm works as illustrated in Figure 1 (with a detailed explanation in the Methods).

Figure 3 shows the result of the CPI-EM algorithm when used to predict genomic regions that are cooperatively bound by FOXA1-HNF4A, RTG3-GCN4 and FIS-CRP in *M. musculus*, *S. cerevisiae* and early-exponential phase cultures of *E. coli*, respectively. The top row shows histograms of the cooperative binding probabilities ($p_1^{coop}, p_2^{coop}, \ldots, p_N^{coop}$), which are computed by CPI-EM, for all peak intensity pairs from each of the three datasets. The height of each bar is the fraction of peak intensity pairs in each probability bin that are actually cooperatively bound (termed true positives, which are calculated based on knockout data as explained in Methods). True positives are distributed differently between the bins across different datasets. Over 50% of cooperatively bound RTG3-GCN4 pairs are assigned a value of $p_{coop} > 0.8$ by CPI-EM, with over 90% of cooperative bound pairs having a value of $p_{coop} > 0.5$. In contrast, only about 30% of cooperative FOXA1-HNF4A pairs have a $p_{coop} > 0.8$.

The distribution of cooperative pairs into each of these bins determines the number of errors made when all peak pairs with $p_{coop} > \alpha$ are declared as cooperatively bound. The false positive rate (FPR) of the CPI-EM algorithm is the fraction of non-cooperatively bound regions erroneously declared as cooperatively bound, while the true positive rate (TPR) is the fraction of cooperatively bound regions that are detected. Both these quantities are functions of α, and are estimated as

$$FPR(\alpha) = \frac{N_{FP}(\alpha)}{N_{nc}}, \;\; TPR(\alpha) = \frac{N_{TP}(\alpha)}{N_c},$$

where $N_{FP}(\alpha)$ is the number of non-cooperatively bound regions mistakenly declared as cooperatively bound at a threshold α, while $N_{TP}(\alpha)$ is the number of cooperatively bound regions correctly declared as cooperatively bound with the threshold α. N_c and N_{nc} represent the total number of cooperatively bound and non-cooperatively bound regions, respectively. The receiver operating characteristic (ROC) curves at the bottom row of Figure 3 shows the trade-off between *false positive rates* and *true positive rates* of CPI-EM at different values of α. A larger value of α results in fewer false positives in the final prediction set but also results in fewer true positives being detected. For instance, in the CRP-FIS dataset, $\alpha = 0.81$ allows nearly 50% of all cooperative interactions to be detected. If α is lowered to 0.375, more than 75% of cooperative peak pairs can be detected, but there will be more false positives in this prediction set since the FPR at this value of α is three times higher than that at $\alpha = 0.81$. A way of quantifying the detection performance of an algorithm like CPI-EM is to calculate the area under the ROC curve (auROC). This is a measure of the average true positive rate of the CPI-EM algorithm, with a higher value representing better detection. Consequently, auROC also provides a way of comparing two different detection algorithms.

3.3 Performance of different variants of the CPI-EM algorithm

When CPI-EM was run to compute cooperative binding probabilities in Figure 3, Lognormal shapes were fitted to the joint probability function of observing cooperative and non-cooperatively bound peak intensity pairs (step 2 of the CPI-EM algorithm). In Figure 4, we compare the detection performance of this version of the CPI-EM algorithm with two other variants that fit Gamma and Gaussian shapes instead of a Lognormal shape. Figure 4 shows the auROC of these three variants of the CPI-EM algorithm after they were run on all the datasets shown in Figure 2. We also compared the auROC of these CPI-EM variants to the auROC of a "naive" peak distance detector and a detector based purely on chance. The peak distance algorithm computes the distances between the peaks of overlapping ChIP-seq peaks and declares those overlapping peak pairs whose peaks are within a threshold distance d to be cooperatively bound. The chance detector is based on using tosses from a biased coin, whose probability of showing heads is α, to detect cooperative interactions. The area under the ROC of this detector will be 0.5 for any dataset (see Supplementary Section 5). An auROC of 0.5 thus represents the minimum level of detection performance that an algorithm should obtain to be considered a useful detector in practice. The complete ROC curves of each of the CPI-EM and peak distance algorithms for the datasets in Figure 2 are shown in Supplementary Figure 1.

In Figure 4, it can be seen that the Gamma and Lognormal CPI-EM variants have an auROC of at least 0.5, and thus can consistently detect cooperative interactions across all datasets. The Lognormal CPI-EM variant fares well on all datasets, except for the mid-exponential phase CRP-FIS dataset, where its performance is at the level of a chance detector. The Gaussian CPI-EM variant performs poorly on the early-exponential phase FIS-CRP and CRP-FIS datasets and has an auROC less than that of a chance detector in the former. This indicates that the Gaussian CPI-EM variant is not as reliable as the Gamma and Lognormal variants in detecting cooperative interactions.

There is considerable variation in the auROC of the peak distance based algorithm: less than 0.5 in early-exponential phase CRP-FIS and RTG3-GCN4 datasets, but higher than 0.5 in the remaining datasets. The fact that this algorithm can perform worse than a chance detector shows that peak distance, by itself, is an unreliable criterion for detecting cooperative binding. This is in contrast to the reliable (auROC > 0.7 across most datasets) performance obtained with the Lognormal and Gamma CPI-EM algorithms.

3.4 Application of CPI-EM to detect cell-type specific binding of transcription factors

We now demonstrate an application of CPI-EM in detecting the cell-type specific binding of a TF in two ChIP-seq data sets from cell lines. Many studies of cell-type specific binding are targeted at understanding cellular reprogramming and stem cell differentiation that gives rise to various organs in animal development [10]. One of the reasons the same TF can bind different genomic regions in two cell types is because it cooperatively binds DNA with the lineage-determining transcription factor of that cell type [10]. Thus, the lineage-determining transcription factor effectively functions as a partner TF, whose concentration is low in one cell type but high in a second one.

We ran CPI-EM on ChIP-seq datasets of ERα-FOXA1 from T-47D breast cancer and ECC-1 endometrial cancer cell lines [25], and CDX2-HNF4A in proliferating and differentiated Caco-2 human intestinal stem cell lines [24]. In the ERα-FOXA1 dataset, we compared ERα binding between T-47D and ECC-1 cell lines. T-47D cells express FOXA1 at a $\approx$50 fold higher concentration than ECC-1 cell lines, with this difference in concentration correlated with differences in ERα occupancy [25]. Given such a large difference in FOXA1 concentration between these two cell types, we treated the ChIP-seq of ERα in ECC-1 cell lines as being equivalent to a knockout of FOXA1 from T-47D cells. Similarly, in the CDX2-HNF4A dataset, differentiated Caco-2 cells express HNF4A at a much higher concentration than proliferating Caco-2 cells [24]. Thus, a comparison of CDX2 binding between differentiated and proliferating Caco-2 cells is akin to analyzing changes in CDX2 binding after HNF4A is knocked out from differentiated Caco-2 cells. Further, CDX2 has been shown to cooperatively bind DNA with HNF4A through an independent biochemical assay [24]. However, the loss of binding in one cell type compared to the other is not solely due to cooperative binding with FOXA1 or HNF4A—differences in nucleosomal occupancy and modifications between cell types are known to influence cell-type specific binding in both datasets [25, 24]. Nevertheless, we wanted to see if CPI-EM could still detect cell-type specific binding in both these datasets.

The box plots in Figure 5A compare the distribution of intensities of ERα peaks present only in T-47D cells with that of ERα peaks present in both cell types. Although other factors determine cell-type specific binding of ERα, we found that ERα peaks present only in T-47D cells were of lower intensity than peaks present in both cell types (Wilcoxon rank-sum test, $p \ll 0.001$). The same trend was seen in CDX2-HNF4A, where regions occupied by CDX2 only in differentiated Caco-2 cells were more weakly bound than regions occupied by CDX2 in both cell types (Wilcoxon rank-sum test, $p \ll 0.001$). In contrast, the trends in the intensities of the partner TFs FOXA1 and HNF4A in both these datasets are different. FOXA1 peaks in ERα cell-type specific bound regions are actually more strongly bound than FOXA1 peaks in shared ERα bound regions, while HNF4A peaks in CDX2 cell-type specific bound regions are more weakly bound than in shared CDX2 bound regions.

Since these patterns in peak intensities of cell-type specific binding were similar to those of cooperative binding seen in Figure 2, we ran all three variants of the CPI-EM algorithm on ERα-FOXA1 and CDX2-HNF4A data sets (Figure 5B). The differences in performance between the Gamma and Lognormal CPI-EM variants are marginal but they are both better than the Gaussian variant at detecting cell-type specific binding in these two datasets. The peak distance detector could not be tested on ERα-FOXA1 data since peak locations were not available in the peak calls. In the CDX2-HNF4A dataset, however, the peak distance algorithm has an auROC only marginally higher than 0.5. Once again, the peak distance criterion is poor at detecting cooperative binding in this dataset. This is in contrast to the CPI-EM algorithm, which has an auROC greater than 0.5 in both datasets, with the Gamma and Lognormal CPI-EM variants giving an auROC of 0.71 in the CDX2-HNF4A dataset.

4 Discussion

Cooperative binding is known to play a role in transcription factor binding site evolution and enhancer detection [28]. Cooperativity is also known to influence cis-regulatory variation between individuals of a species [29], which could potentially capture disease-causing mutations that are known to occur in regulatory regions of the genome [30]. CPI-EM is suited to study these phenomena since it can detect instances of cooperative binding between a pair of transcription factors that may occur anywhere in the genome. While sequence-based approaches to cooperative binding detection have been proposed [12, 13, 14, 15, 16, 17, 18, 19, 1], none use ChIP-seq peak intensities as a criterion to detect cooperativity. Our goal was to demonstrate that peak intensities, by themselves, contain valuable information to detect cooperative binding. Our results suggest that methods for detecting cooperative binding based on ChIP-seq peak intensities can usefully complement sequence-based detection algorithms.

4.1 Assumptions in the CPI-EM algorithm

The assumption that cooperatively bound primary TFs are more weakly bound, on average, than non-cooperatively bound primary TFs is the key assumption in the CPI-EM algorithm. This assumption was true across TF pairs we analyzed in *E. coli*, *S. cerevisiae*, *M. musculus*, and human genomes (which included cancer and stem cell lines). However, the consequence of this assumption is that CPI-EM is unlikely to detect regions where the primary TF is cooperatively bound to DNA, but with a high peak intensity.

Our observation that cooperatively bound TFs were more weakly bound than non-cooperatively bound TFs is likely a signature of short-range pair-wise interactions. For instance, GCN4-RTG3 and CDX2-HNF4A interactions were discovered in the datasets upon which we ran CPI-EM, and these interactions have been independently verified [1, 24]. A similar pattern of weakly but cooperatively bound TFs is seen in animal development. The binding of Ultrabithorax (Ubx) and Extradenticle at the *shavenbaby* enhancer in *Drosophila melanogaster* embryos [31] occurs in closely spaced low-affinity binding sites to help coordinate tissue patterning. Mutations that increased Ubx binding affinity led to the expression of proteins outside their naturally occurring tissue boundaries [31]. Similarly, low-affinity binding sites that cooperatively bind Cubitus interruptus at the *dpp* enhancer (which plays a crucial role in wing patterning in *Drosophila melanogaster*) are evolutionary conserved across twelve *Drosophila* species [32].

4.2 Challenges to cooperativity detection using ChIP-seq peak intensities

There are two principal challenges to detecting cooperative interactions using ChIP-seq peak intensities — its low spatial resolution, and the use of PCR amplification. ChIP-seq cannot resolve binding events that occur within 100 base pairs of each other [33], while DNA-mediated cooperative binding often occurs between two TFs bound within 25 base pairs of each other [11, 20]. Thus, a single ChIP-seq peak intensity can represent the average of multiple cooperative and non-cooperative binding events. This low resolution may also explain why peak distance was not a reliable criterion to detect cooperative binding. Protocols such as ChIP-exo [33] and ChIP-nexus [34] can resolve two binding events that are a single base pair apart [35]. These methods likely provide more accurate measurements of distances between binding events, which means that ChIP-exo or ChIP-nexus peak distances may supplement peak intensities in detecting cooperative interactions.

Peak intensities are also affected by PCR amplification, which is a necessary step in ChIP-seq protocols. While the use of more PCR amplification cycles helps detect weaker binding events, the variance in the number of fragments obtained at the end of the PCR process increases with the number of cycles employed [36, 37]. If peak intensities can instead be calculated based on the number of *un-amplified* DNA fragments, they would be less noisy measures of binding affinity. This is possible with protocols such as ChIP-nexus [34], that use molecular bar-coding techniques in DNA library preparation [38].

The additional variance introduced by PCR amplification might also explain the low mutual information values we measured between peak intensities of cooperatively bound TFs (see Supplementary Table 2). Thus, protocols such as ChIP-nexus and ChIP-exo might be sensitive enough to detect the difference in MI between cooperatively and non-cooperatively bound TFs [39]. In such a case, our method can be modified to no longer be dependent on the assumption of cooperatively bound primary TFs being more weakly bound than non-cooperatively bound primary TFs. In this modified algorithm, a tenth parameter in the joint probability model fit to peak intensity data (in step 2 of the CPI-EM algorithm) will take into account this mutual information resulting from cooperative binding. The precise form of such a modified joint probability model is not obvious, but it would increase the probability that a high MI peak intensity pair would be labeled as cooperative, despite having a strongly bound primary TF.

Ultimately, our method is a way of detecting cooperatively bound locations without making any direct assumptions about the genomic sequence of that location. Therefore, it provides a useful way of finding binding sequence patterns that allow for cooperative binding to occur *in vivo* but lie outside the range of existing sequence-based algorithms.

5 Funding

Support from the Simons Foundation (to S.K. and V.D.); PRISM 12th plan project at Institute of Mathematical Sciences (to R.S.);

6 Acknowledgements

We thank Aswin Sai Narain Seshasayee, Parul Singh, Vijay Kumar, Deepa Agashe, and Leelavati Narlikar for discussions.

Author Contributions : V.D. conceived the study, and designed and implemented the CPI-EM algorithm. V.D., R.S., and S.K. analyzed and interpreted the results, and wrote the manuscript.

References

[1] Spivak, A. T. and Stormo, G. D. . Combinatorial cis-regulation in saccharomyces species. *G3: Genes— Genomes— Genetics*, 6(3):653–667, 2016.

[2] Patwardhan, R. P. , Hiatt, J. B. , Witten, D. M. , Kim, M. J. , Smith, R. P. , May, D. , Lee, C. , Andrie, J. M. , Lee, S.-I. , Cooper, G. M. , et al. Massively parallel functional dissection of mammalian enhancers in vivo. *Nature biotechnology*, 30(3):265–270, 2012.

[3] Smith, R. P. , Taher, L. , Patwardhan, R. P. , Kim, M. J. , Inoue, F. , Shendure, J. , Ovcharenko, I. , and Ahituv, N. . Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. *Nature Genetics*, 45(9):1021–1028, 2013.

[4] Sharon, E. , Kalma, Y. , Sharp, A. , Raveh-Sadka, T. , Levo, M. , Zeevi, D. , Keren, L. , Yakhini, Z. , Weinberger, A. , and Segal, E. . Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nature Biotechnology*, 30(6):521–530, 2012.

[5] Gutierrez, P. S. , Monteoliva, D. , and Diambra, L. . Cooperative binding of transcription factors promotes bimodal gene expression response. *PLoS One*, 7(9):e44812, 2012.

[6] Tuğrul, M. , Paixão, T. , Barton, N. H. , and Tkačik, G. . Dynamics of transcription factor binding site evolution. *PLoS Genet*, 11(11):e1005639, 2015.

[7] Jansen, R. , Yu, H. , Greenbaum, D. , Kluger, Y. , Krogan, N. J. , Chung, S. , Emili, A. , Snyder, M. , Greenblatt, J. F. , and Gerstein, M. . A bayesian networks approach for predicting protein-protein interactions from genomic data. *science*, 302(5644):449–453, 2003.

[8] Krogan, N. J. , Cagney, G. , Yu, H. , Zhong, G. , Guo, X. , Ignatchenko, A. , Li, J. , Pu, S. , Datta, N. , Tikuisis, A. P. , et al. Global landscape of protein complexes in the yeast saccharomyces cerevisiae. *Nature*, 440(7084):637–643, 2006.

[9] Hardison, R. . Hemoglobins from bacteria to man: evolution of different patterns of gene expression. *Journal of Experimental Biology*, 201(8):1099–1117, 1998.

[10] Reiter, F. , Wienerroither, S. , and Stark, A. . Combinatorial function of transcription factors and cofactors. *Current Opinion in Genetics & Development*, 43:73–81, 2017.

[11] Kim, S. , Broströmer, E. , Xing, D. , Jin, J. , Chong, S. , Ge, H. , Wang, S. , Gu, C. , Yang, L. , Gao, Y. Q. , et al. Probing allostery through DNA. *Science*, 339(6121):816–819, 2013.

[12] GuhaThakurta, D. and Stormo, G. D. . Identifying target sites for cooperatively binding factors. *Bioinformatics*, 17(7):608–621, 2001.

[13] Whitington, T. , Frith, M. C. , Johnson, J. , and Bailey, T. L. . Inferring transcription factor complexes from chip-seq data. *Nucleic Acids Research*, 39(15):e98–e98, 2011.

[14] Kazemian, M. , Pham, H. , Wolfe, S. A. , Brodsky, M. H. , and Sinha, S. . Widespread evidence of cooperative dna binding by transcription factors in drosophila development. *Nucleic Acids Research*, 41(17):8237–8252, 2013.

[15] Das, D. , Banerjee, N. , and Zhang, M. Q. . Interacting models of cooperative gene regulation. *Proceedings of the National Academy of Sciences of the United States of America*, 101(46):16234–16239, 2004.

[16] He, X. , Chen, C.-C. , Hong, F. , Fang, F. , Sinha, S. , Ng, H.-H. , and Zhong, S. . A biophysical model for analysis of transcription factor interaction and binding site arrangement from genome-wide binding data. *PloS One*, 4(12):e8155, 2009.

[17] Girgis, H. Z. and Ovcharenko, I. . Predicting tissue specific cis-regulatory modules in the human genome using pairs of co-occurring motifs. *BMC Bioinformatics*, 13(1):1, 2012.

[18] Nandi, S. , Blais, A. , and Ioshikhes, I. . Identification of cis-regulatory modules in promoters of human genes exploiting mutual positioning of transcription factors. *Nucleic Acids Research*, page gkt578, 2013.

[19] Jiang, P. and Singh, M. . Ccat: combinatorial code analysis tool for transcriptional regulation. *Nucleic Acids Research*, 42(5):2833–2847, 2014.

[20] Jolma, A. , Yin, Y. , Nitta, K. R. , Dave, K. , Popov, A. , Taipale, M. , Enge, M. , Kivioja, T. , Morgunova, E. , and Taipale, J. . DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature*, 527(7578):384–388, 2015.

[21] Narasimhan, K. , Pillay, S. , Huang, Y.-H. , Jayabal, S. , Udayasuryan, B. , Veerapandian, V. , Kolatkar, P. , Cojocaru, V. , Pervushin, K. , and Jauch, R. . DNA-mediated cooperativity facilitates the co-selection of cryptic enhancer sequences by SOX2 and PAX6 transcription factors. *Nucleic Acids Research*, page gku1390, 2015.

[22] Johnson, D. S. , Mortazavi, A. , Myers, R. M. , and Wold, B. . Genome-wide mapping of in vivo protein-dna interactions. *Science*, 316(5830):1497–1502, 2007.

[23] Stefflova, K. , Thybert, D. , Wilson, M. D. , Streeter, I. , Aleksic, J. , Karagianni, P. , Brazma, A. , Adams, D. J. , Talianidis, I. , Marioni, J. C. , et al. Cooperativity and rapid evolution of cobound transcription factors in closely related mammals. *Cell*, 154(3):530–540, 2013.

[24] Verzi, M. P. , Shin, H. , He, H. H. , Sulahian, R. , Meyer, C. A. , Montgomery, R. K. , Fleet, J. C. , Brown, M. , Liu, X. S. , and Shivdasani, R. A. . Differentiation-specific histone modifications reveal dynamic chromatin interactions and partners for the intestinal transcription factor cdx2. *Developmental Cell*, 19(5):713–726, 2010.

[25] Gertz, J. , Savic, D. , Varley, K. E. , Partridge, E. C. , Safi, A. , Jain, P. , Cooper, G. M. , Reddy, T. E. , Crawford, G. E. , and Myers, R. M. . Distinct properties of cell-type-specific and shared transcription factor binding sites. *Molecular cell*, 52(1):25–36, 2013.

[26] Dale, R. K. , Pedersen, B. S. , and Quinlan, A. R. . Pybedtools: a flexible python library for manipulating genomic datasets and annotations. *Bioinformatics*, 27(24):3423–3424, 2011.

[27] Kinney, J. B. and Atwal, G. S. . Equitability, mutual information, and the maximal information coefficient. *Proceedings of the National Academy of Sciences*, 111(9):3354–3359, 2014.

[28] Villar, D. , Flicek, P. , and Odom, D. T. . Evolution of transcription factor binding in metazoans-mechanisms and functional implications. *Nature Reviews Genetics*, 15(4):221–233, 2014.

[29] Heinz, S. , Romanoski, C. , Benner, C. , Allison, K. , Kaikkonen, M. , Orozco, L. , and Glass, C. . Effect of natural genetic variation on enhancer selection and function. *Nature*, 503(7477):487–492, 2013.

[30] Knight, J. C. . Regulatory polymorphisms underlying complex disease traits. *Journal of Molecular Medicine*, 83(2):97–109, 2005.

[31] Crocker, J. , Abe, N. , Rinaldi, L. , McGregor, A. P. , Frankel, N. , Wang, S. , Alsawadi, A. , Valenti, P. , Plaza, S. , Payre, F. , et al. Low affinity binding site clusters confer hox specificity and regulatory robustness. *Cell*, 160(1):191–203, 2015.

[32] Ramos, A. I. and Barolo, S. . Low-affinity transcription factor binding sites shape morphogen responses and enhancer evolution. *Phil. Trans. R. Soc. B*, 368(1632):20130018, 2013.

[33] Rhee, H. S. and Pugh, B. F. . Comprehensive genome-wide protein-dna interactions detected at single-nucleotide resolution. *Cell*, 147(6):1408–1419, 2011.

[34] He, Q. , Johnston, J. , and Zeitlinger, J. . Chip-nexus enables improved detection of in vivo transcription factor binding footprints. *Nature Biotechnology*, 33(4):395–401, 2015.

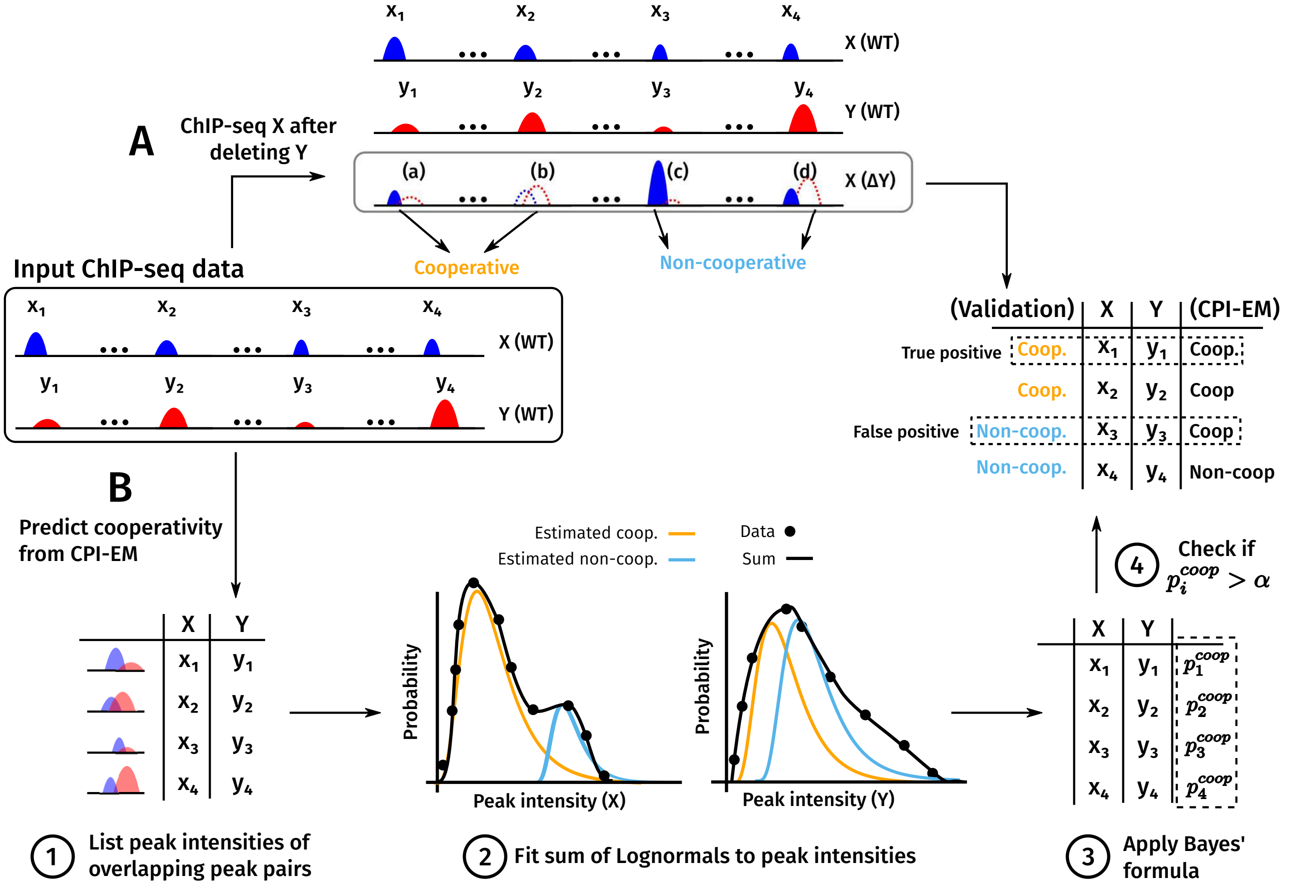
[35] Starick, S. R. , Ibn-Salem, J. , Jurk, M. , Hernandez, C. , Love, M. I. , Chung, H.-R. , Vingron, M. , Thomas-Chollier, M. , and Meijsing, S. H. . Chip-exo signal associated with dna-binding motifs provides insight into the genomic binding of the glucocorticoid receptor and cooperating transcription factors. *Genome Research*, 25(6):825–835, 2015.

[36] Kebschull, J. M. and Zador, A. M. . Sources of PCR-induced distortions in high-throughput sequencing data sets. *Nucleic Acids Research*, 43(21):e143–e143, 2015.

[37] Best, K. , Oakes, T. , Heather, J. M. , Shawe-Taylor, J. , and Chain, B. . Computational analysis of stochastic heterogeneity in PCR amplification efficiency revealed by single molecule barcoding. *Scientific Reports*, 5, 2015.

[38] Kivioja, T. , Vähärautio, A. , Karlsson, K. , Bonke, M. , Enge, M. , Linnarsson, S. , and Taipale, J. . Counting absolute numbers of molecules using unique molecular identifiers. *Nature methods*, 9(1):72–74, 2012.

[39] Kinney, J. B. , Tkačik, G. , and Callan, C. G. . Precise physical models of protein–dna interaction from high-throughput data. *Proceedings of the National Academy of Sciences*, 104(2):501–506, 2007.



	(Validation)	X	Y	(CPI-EM)
True positive	Coop.	x_1	y_1	Coop.
	Coop.	x_2	y_2	Coop
False positive	Non-coop.	x_3	y_3	Coop
	Non-coop.	x_4	y_4	Non-coop

Figure 1: **A schematic of the use of the CPI-EM algorithm and ChIP-seq from knockout data to separately identify cooperative bound transcription factor pairs.** ChIP-seq experiments carried out on two TFs, X and Y, yield a list of locations that are bound by both TFs, along with peak intensities at each location. From this data, there are two ways in which we find genomic locations that are cooperatively bound by X and Y.

(A) A method for inferring these locations from a ChIP-seq of X carried out after Y is genetically deleted. Locations where a peak of X either disappears altogether, or is reduced in intensity after knocking out Y are labelled as cooperatively bound. In contrast, locations where a peak of X either remains unchanged or increases in intensity are labelled as non-cooperatively bound (see section “Using ChIP-seq data from a genetic knockout to infer cooperative binding” in Methods).

(B) Steps in predicting cooperatively bound locations are shown, where the numbers correspond to those in the section “The ChIP-seq Peak Intensity - Expectation Maximisation (CPI-EM) algorithm” in Methods. (1) The input to CPI-EM consists of a list of genomic locations where a peak of X overlaps a peak of Y by at least a single base pair. (2) Each of these overlapping intensity pairs is fit to a model that consists of a sum of two probability functions. These functions specify the probabilities of observing a particular peak intensity pair given that it comes from a cooperatively or non-cooperatively bound region. These probabilities are computed by fitting the model to the input data using the expectation-maximization algorithm (see Supplementary Section 4.1). (3) Bayes’ formula is applied to the probabilities computed in step (2) to find the probability of each peak intensity pair being cooperatively bound. (4) Each cooperative binding probability computed in step (3) that is greater than a threshold α is declared as cooperatively bound. We compare this list of predicted locations with the list of cooperatively bound locations inferred from knockout data in order to compute the number of correct and incorrect inferences made by CPI-EM.

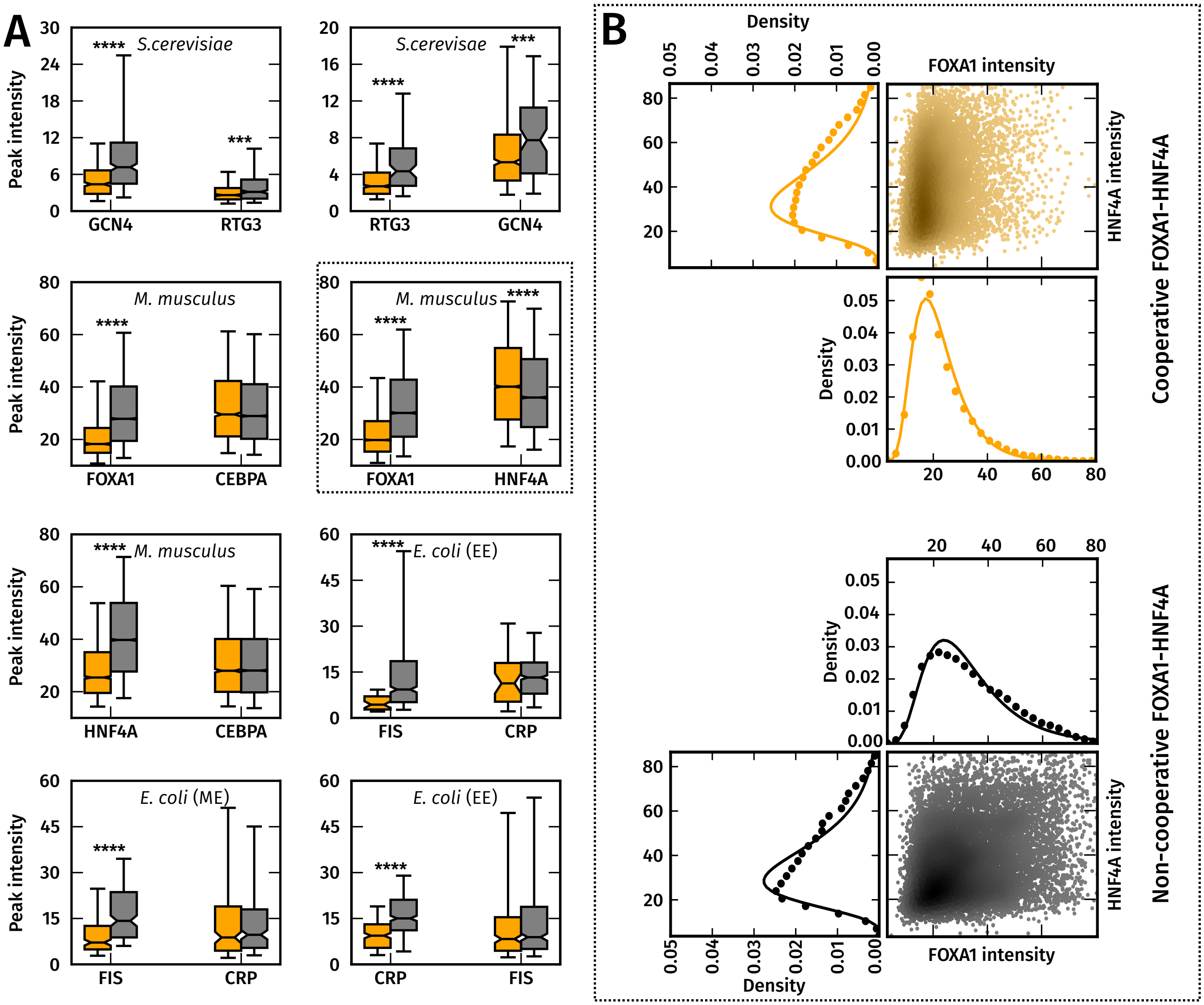


Figure 2: **(A) Cooperatively bound primary TFs are significantly more weakly bound than non-cooperatively bound primary TFs.** Box-plots of peak intensity distributions of cooperatively (orange) and non-cooperatively (gray) bound TF pairs, with primary TFs on the left and partner TFs on the right. ****, *** and ** indicate p-values of $< 10^{-4}, 10^{-3}$ and 10^{-2} from a Wilcoxon rank sum test. The whiskers of the box plot are the $5-th$ and $95-th$ percentiles of the distributions shown.

(B) ChIP-seq peak intensity distributions can be approximated by a Lognormal distribution. Marginal peak intensity distributions of FOXA1 and HNF4A peaks (in filled black and orange circles), with fitted Lognormal distributions (solid black and orange lines), along side a scatter plot of (FOXA1,HNF4A) peak intensity pairs from cooperatively and non-cooperatively bound regions. The scatter points are colored according to the density of points in that region, with darker shades indicating a higher density. The density of points in the scatter were computed using the Gaussian kernel density estimation procedure in the Python Scipy library.

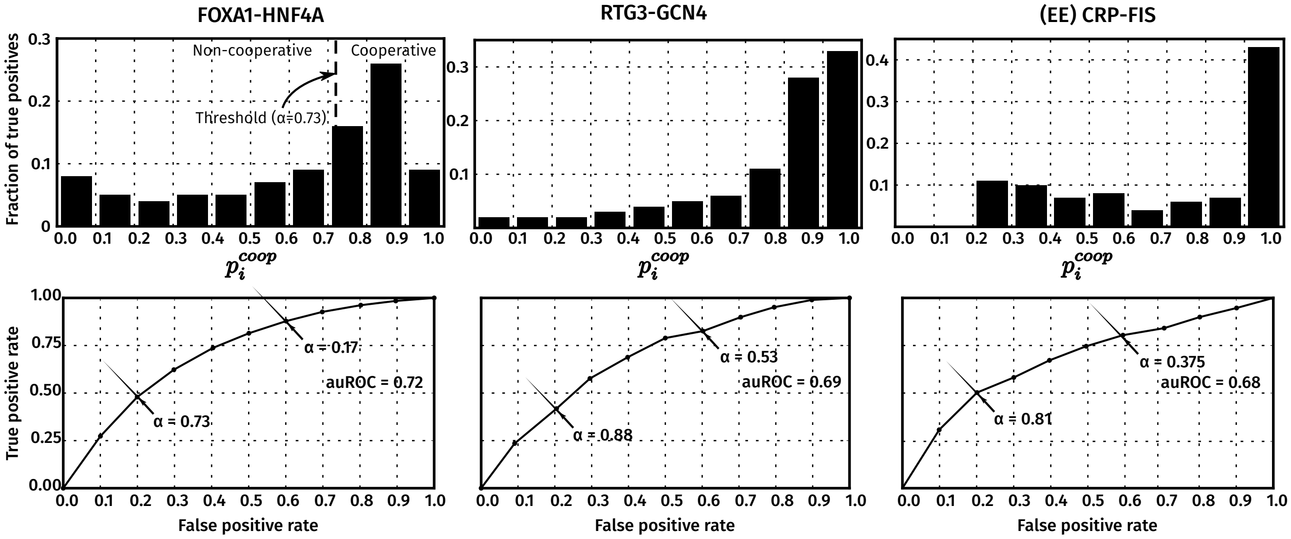


Figure 3: **CPI-EM applied to ChIP-seq datasets from *M. musculus* (FOXA1-HNF4A), *S. cerevisiae* (RTG3-GCN4) and early-exponential phase cultures of *E. coli* (CRP-FIS).** For each dataset, CPI-EM computes a list of cooperative binding probabilities at all the locations bound by the TF pair under consideration. **Top row:** The fraction of cooperatively bound pairs, as determined from knockout data, that fall into each cooperative binding probability bin. T he bins are equally spaced with a width of 0.1 and the heights of the bars within each histogram add up to 1. **Bottom row: Receiver operating characteristic (ROC) curves that evaluate the performance of CPI-EM in detecting cooperatively bound pairs.** The curve is generated by calculating, for each value of α between 0 and 1, the true and false positive rate of the algorithm. The true positive rate ($TPR(\alpha)$) is the ratio of the number of cooperatively bound regions detected to the total number of cooperatively bound regions at that value of α. The false positive rate ($FPR(\alpha)$) is the ratio of the number of non-cooperatively bound regions mistakenly detected as cooperatively bound to the total number of non-cooperatively bound regions at that value of α. Small values of α give a higher TPR, but at the cost of a higher FPR. The area under the ROC (auROC) is a measure of detection performance, whose value cannot exceed 1, which corresponds to a perfect detector. Given the auROC of two different algorithms, the one with a higher auROC is better, on average, at detecting cooperative binding.

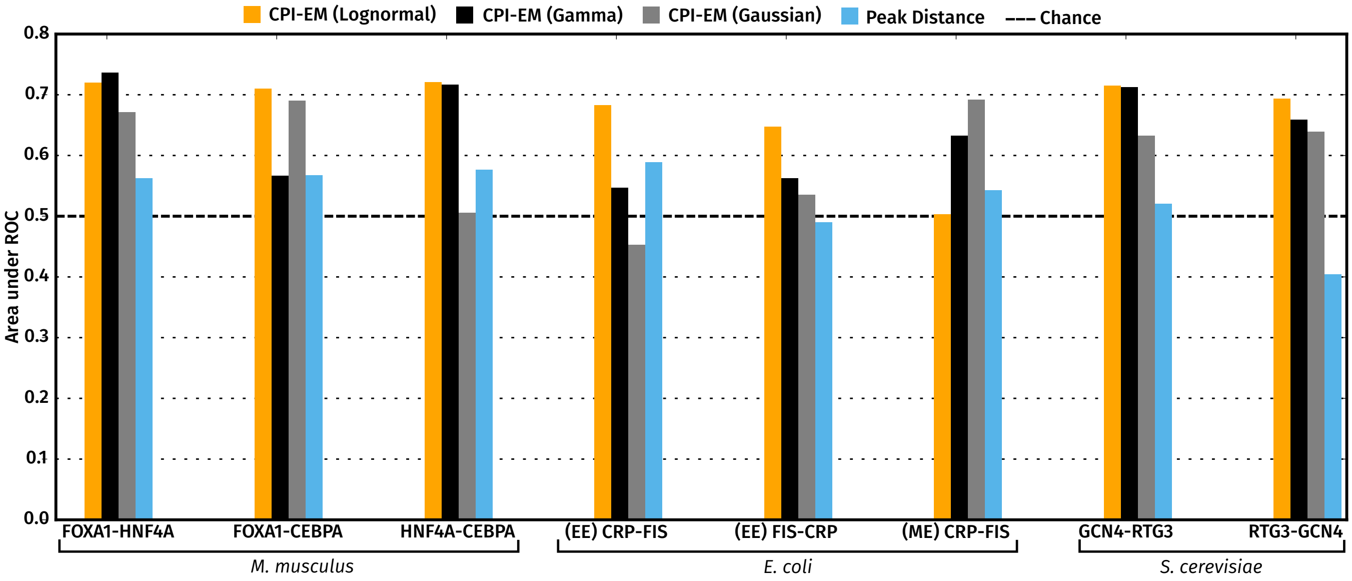


Figure 4: **The CPI-EM variant that fits lognormal distributions to peak intensity pairs consistently performs well across all datasets** The area under the ROC curve (auROC) of the CPI-EM algorithm applied to each of the datasets shown in Figure 2. CPI-EM variants that fit Lognormal, Gamma and Gaussian distributions are represented in orange, black and gray, respectively. The auROC of the peak distance based detector is shown in blue. See Supplementary Section 5 for the calculation of the ROC curve for both the CPI-EM and peak distance algorithms.

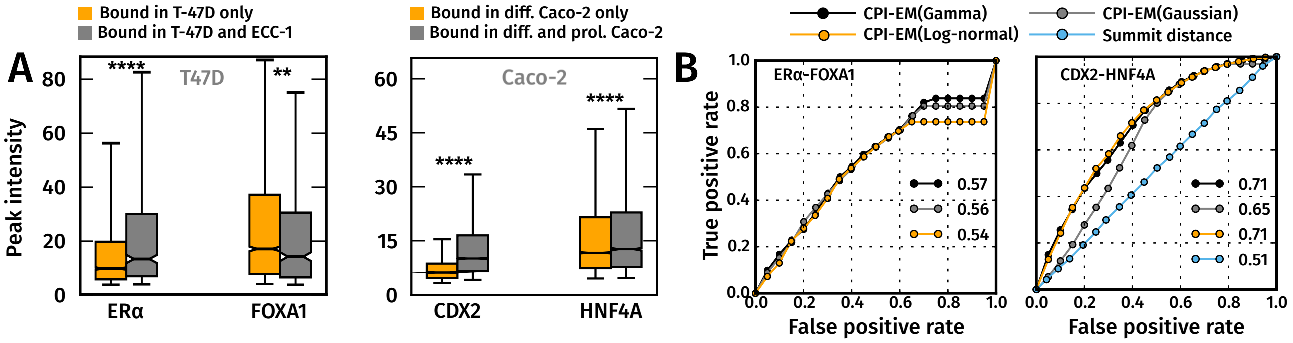


Figure 5: **(A) Regions bound by ERα only in T-47D cells are more weakly bound than regions bound by ERα in both T-47D and ECC-1 cells.** The same trend in peak intensities holds between regions bound by CDX2 only in differentiated Caco-2 cells and those bound by CDX2 in both differentiated and proliferating Caco-2 cells. However, cell-type specific binding in these cell types is also determined by factors other than cooperativity. Distributions of cooperatively and non-cooperatively bound regions are shown in orange and gray, respectively. The whiskers of the box plot are the $5-th$ and $95-th$ percentiles of the distributions shown. **(B) The Lognormal CPI-EM variant consistently detects cell-type specific binding events of ERα and CDX2**. ROC curves of Lognormal (orange), Gamma (black) and Gaussian (gray) variants of CPI-EM, and the peak distance detector (blue), on ERα-FOXA1 and CDX2-HNF4A datasets.The area under the ROC of each detector is indicated in the legend. The peak distance detector was not run on ERα-FOXA1 data since peak locations were not available in the peak calls. The ROC was generated using the same procedure as in Figure 4.