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

March 24, 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 maximisation (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 a second TF binding a nearby sequence changes the first TF's binding affinity. In this case, the two TFs are said to bind DNA *cooperatively* or combinatorially [1].

Multiple TFs cooperatively binding enhancers and promoters are known to non-additively drive gene expression [2, 3, 4], and the presence of cooperativity has been used to explain the rapid rate of evolution of TF binding sites in multicellular organisms [5]. Several theoretical methods have been proposed to detect cooperative binding between a pair of TFs in the genome [6, 7, 8, 9, 10, 11, 12, 13, 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]. One of the reasons is because many TF pairs can cooperatively bind DNA even if the spacing between their binding sites, or the sequence in between them, is changed [14, 15]. Here, we propose a sequence-independent algorithm, based on ChIP-seq data, for detecting cooperatively bound sites that complement these sequence-based methods.

Genome-wide TF-DNA binding has been extensively studied using ChIP-seq (chromatin immuno-precipitation and sequencing) [16]. 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 [16]. 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 [17, 1, 18, 19]. 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

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

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 data set is labelled B-A, with B and A as primary and partner TFs, respectively.

We propose the ChIP-seq Peak Intensity - Expectation Maximization (CPI-EM) algorithm as a computational method to detect genomic locations 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 only 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 data sets 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 data sets 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*. Peak intensity is a measure of binding affinity, and in terms of the narrowPeak and broadPeak output format of most ChIP-seq peak callers, this could be the signal value ($7-th$ column), $-\log_{10}$(p-value) ($8-th$ column), or the $-\log_{10}$(q-value) ($9-th$ column) of each line in the peak call file. The signal value is typically computed from the number of sequence reads that originate from a bound genomic location. The p-value is computed from the signal value, which is a measure of statistical significance of the peak call. The q-value of each peak call is computed by adjusting the p-value to control the false discovery rate of the peak call set [20], which is a correction for multiple hypothesis testing. A peak call with a larger signal value has a smaller p- and q-value, which indicates that it is more likely to reflect an actual protein-DNA binding event. Thus, a larger signal value will translate to a larger $-\log_{10}$(p-value) and $-\log_{10}$(q-value).

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 [17], and three human cell lines – the Caco-2 intestinal stem cell line [18], the T-47D breast cancer cell line and the ECC-1 endometrial cancer cell line [19]. 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 data sets. This ensured that our validation sets were not biased by procedures employed in our pipeline. The ChIP-seq data processing pipeline we followed is described below.

2.1.1 ChIP-seq of FOXA1, HNF4A and CEBPA from *M. musculus* liver:

We aligned the raw sequence reads (ArrayExpress, accession number: E-MTAB-1414) from the experiment to the 2007 UCSC mm9 release of the C57/BL6 strain of the mouse genome, using the BWA (v0.7.12) aligner with default settings [21]. We ran MACS2 (v2.1.0) [22], with its default settings, to call peaks on each of these alignments. Peaks were called with a liberal p-value threshold of 10^{-3}. Since the wild-type ChIP-seq data consisted of two biological replicates, we pooled these aligned reads into a single file and called peaks using MACS2. We ran MACS2 with default settings, which discards aligned reads that are PCR duplicates before calling peaks.. In this data set, we used the signal values of FOXA1, HNF4A and CEBPA peak calls as peak intensities.

We then followed a second step to filter peak calls. The use of a relatively liberal p-value threshold of 10^{-3} while calling peaks, and the pooling of aligned reads before calling peaks, was necessary in order to compute the irreproducible discovery rate (IDR) [23, 24] of each peak. We computed the IDR of each peak with the idr script (v2.0) [23], and retained peaks whose IDR was less than 1%. We then ranked peaks according to their MACS2 signal values, with the top ranked peak having the largest signal value. We divided these ranks by the total number of peaks in the ChIP-seq profile to obtain a normalized rank for each peak, which is equivalent to the quantile of that peak intensity within the profile. Significant changes in these peak ranks were used to detect cooperative binding events while comparing peak calls between wild-type and knockout ChIP-seq data.

Because the ChIP-seq of FOXA1 in Δ*HNF4A* and Δ*CEBPA* cells, HNF4A in Δ*CEBPA* cells, and CEBPA in Δ*HNF4A* cells were not performed in replicates, we could not use the IDR criterion to filter peaks. Instead, for these, we filtered peak calls using the q-value of each peak call as computed by MACS2. We retained only those peaks whose q-values were less than 0.01 for further analysis. These peak calls were finally used to detect cooperative binding in FOXA1-HNF4A, FOXA1-CEBPA, HNF4A-CEBPA and CEBPA-HNF4A pairs (see next section).

2.1.2 ChIP-seq of GCN4, RTG3 in *S. cerevisiae*:

We aligned raw sequence reads from the ChIP-seq libraries of GCN4, RTG3 (accession Number GSE60281) to the S288C reference genome of *S. cerevisiae*, available at the Saccharomyces Genome Database [25].

We followed the same procedure as with the *M. musculus* data, with some changes. ChIP-seq reads from GCN4 and RTG3 were available in three replicates. In these data sets, we chose the two replicates that had the largest number of peaks and merged their sequence read alignments. MACS2 was run with additional `--nomodel --extsize 147` options, as the number of sequence reads were insufficient for MACS2 to build its own tag shifting model. We called peaks on this merged set using MACS2, with a p-value threshold of 0.1, and retained peaks whose q-values were less than 0.1. We did not filter peak calls based on IDR because we found it to be too stringent a criterion; it typically gave us a very small number of peaks (< 100) for these TFs. We finally used the q-values of GCN4 and RTG3 peak calls as peak intensities.

2.1.3 ChIP-seq of ERα, FOXA1, CDX2 and HNF4A from Caco-2, T-47D and ECC-1 cell lines:

For each of these TFs, we utilized the pre-computed peak calls of ERα and FOXA1 from T-47D and ECC-1 cell lines that were publicly available on the GEO database with accession number GSE32465 [19]. We also utilized pre-computed peak calls of CDX2 and HNF4A from Caco-2 cell lines that were publicly available (accession number GSE23436) [18].

We retained those peaks in the pre-computed ERα and FOXA1 peak calls whose q-values were less than 0.05. In the CDX2 and HNF4A peak call set, we chose peak calls whose q-values were less than 0.01. In both data sets, we used signal values of peak calls as peak intensities. However, these signal values were scaled. In the ERα-FOXA1 data set, we divided all signal values by 35, and in the CDX2-HNF4A data set, we divided all signal values by a factor of 2. This was done to speed up the running time of the CPI-EM algorithm. The scaling of signal values did not affect the detection performance of the CPI-EM algorithm.

2.1.4 ChIP-seq of FIS and CRP in *E. coli* from early exponential (EE) and mid-exponential (ME) phase cultures:

For FIS and CRP ChIP-seq data sets, we utilized pre-computed peak calls that were available on the GEO database with accession number GSE92255. Though the ChIP-seq experiments were carried out in replicates, these peaks were called by running MACS2 on merged alignments of sequence reads from both replicates. The peak calls in this set were filtered such that all peaks had a q-value less than 0.05. We use the q-values of FIS and CRP peak calls as peak intensities for CPI-EM.

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

From ChIP-seq profiles of a pair of TFs, A and B, we classified genomic regions containing overlapping ChIP-seq peaks of A and B as cooperative or non-cooperative, based on the change in peak rank of A in response to a genetic deletion of B. The ranks are assigned such that the highest rank peak,i.e., the peak ranked 1 has the highest peak intensity. In our analysis, we consider a genomic region to be doubly bound by A and B 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 A as being cooperatively bound by B if (a) the peak rank of A in the presence of B is significantly higher than the peak rank of A measured after the deletion of B, or (b) if A's

peak is absent after the deletion of B. On the other hand, if the peak rank of A in the presence of B is significantly lower than the peak rank of A after the deletion of B, 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.

To determine if a peak rank change is significant, we construct a null distribution, which captures the magnitude of rank changes of A expected due to variability in the ChIP-seq protocol. Suppose $r_1^{(1)}, r_2^{(1)}, \ldots, r_n^{(1)}$ and $r_1^{(2)}, r_2^{(2)}, \ldots, r_n^{(2)}$ represent the normalized ranks, whose values are between 0 and 1, of n overlapping peaks in biological replicates 1 and 2 of the ChIP-seq of A (in the presence of B). We then divide the interval $[0, 1]$ into 10 equally sized bins (we verified that changing the number of bins did not drastically change the results), and compute the null rank change probability density $g_{null}^k(x)$ of the k^{th} bin from the samples $S_k = \{|r_1^{(1)} - r_1^{(2)}|, |r_2^{(1)} - r_2^{(2)}|, \ldots, |r_l^{(1)} - r_l^{(2)}|\}$, where $r_i^{(1)}$ falls in the k^{th} bin. A Gaussian kernel density estimator implemented in the Scipy library was used to compute $g_{null}^k(x)$ for each bin. This represents the probability of observing a rank change purely due to inter-replicate variation, conditioned on the bin to which the peak's rank in replicate 1 belongs. The process of computing rank changes separately within each bin better captured the skew expected in rank changes arising from replicate variation. For instance, a peak of A, whose rank in replicate 1 is low, is far more likely to have a higher rank in replicate 2, than a peak with a high rank in replicate 1.

We then proceed to compute the significance of rank changes observed in peaks of A after B has been knocked out. For this, we computed the ranks $r_1^{(m)}, r_2^{(m)}, \ldots, r_q^{(m)}$ from peaks of A that have been called from merging the read alignments of replicates 1 and 2. The average change in peak rank due to the merging of alignments was close to zero, i.e., the ranks $r_1^{(m)}, r_2^{(m)}, \ldots, r_p^{(m)}$, did not change on average compared to $r_1^{(1)}, r_2^{(1)}, \ldots, r_p^{(1)}$ and $r_1^{(2)}, r_2^{(2)}, \ldots, r_p^{(2)}$ (data not shown), where p is the number of peaks common between peak calls in the replicates and merged alignments. We also compute the ranks $r_1^{\Delta}, r_2^{\Delta}, \ldots, r_q^{\Delta}$ of peak calls from the ChIP-seq of A after B is knocked out. We then construct the set of rank changes $\{|r_1^{(m)} - r_1^{\Delta}|, (|r_2^{(m)} - r_2^{\Delta}|), \ldots, |r_q^{(m)} - r_q^{\Delta}|\}$. For each rank change, we calculate $p_i = g_{null}^k(|r_i^{(m)} - r_i^{\Delta}|)$, where k is the bin into which $r_i^{(m)}$ falls. This is the probability of observing a rank change of magnitude $|r_i^{(m)} - r_i^{\Delta}|$ purely due to inter-replicate variation, given that $r_i^{(m)}$ belongs to the k^{th} bin. We finally obtain a sequence of probabilities $p_1, p_2, \ldots, p_q$ corresponding to each rank change observed upon knocking out B.

We then conduct q one-sided hypothesis tests, each of which test the null hypothesis $H_i : |r_i^{(m)} - r_i^{\Delta}| = 0$. We carry out the hypothesis tests by checking if each $p_i < \alpha$, where α is chosen according to the Benjamini-Hochberg multiple hypothesis testing procedure [20] that sets the false discovery rate at 0.01.

In the ERα-FOXA1, CDX2-HNF4A, CRP-FIS, and FIS-CRP data sets, peak calls from individual replicates were not available. Since the null rank change distributions described above cannot be computed without peak calls from multiple replicates, we used only peak losses to find cooperatively bound locations in these data sets.

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

The input to the CPI-EM algorithm consists of ChIP-seq peak intensities of two TFs, X and Y, from their binding locations across the genome. The goal of the CPI-EM algorithm is to predict, from this data, the locations at which the two TFs bind cooperatively. The algorithm consists of the following steps, which are numbered according to the steps shown in Figure 1.

1. **List ChIP-seq peak calls $\{(x_i, y_i)\}_{i=1}^N$ from TFs X and Y, where the peak of one TF overlaps the peak of the second TF.**

   We took peak call files of X and Y, and used pybedtools (v0.6.9) [26] to find all peak regions that overlap each other by at least a single base pair. In instances where a single peak region of X overlaps multiple peaks of Y, or vice versa, we consider each overlapping pair as a distinct pair.

2. **Determine the parameters of a mixture model that best fits the joint distribution of peak intensities across these overlapping pairs.** We assume the joint density of peak intensities from all doubly bound regions, $f_{XY}(x, y)$, is a mixture (i.e., a sum) of two densities representing cooperative and non-cooperative peak intensity distributions:

$$\begin{aligned} f_{XY}(x, y) = \pi_0 f_0^X(x; \boldsymbol{\theta}_0^X) f_0^Y(y; \boldsymbol{\theta}_0^Y) \\ + \pi_1 f_1^X(x; \boldsymbol{\theta}_1^X) f_1^Y(y; \boldsymbol{\theta}_1^Y), \end{aligned} \tag{1}$$

   where, f_i are the marginal densities of peak intensities of non-cooperative and cooperative peak pairs across the genome. We chose f_1 to be either a Log-normal, Gamma or Gaussian density function (see Supplementary

Section 2 for their definitions). π_0 and π_1 represent the fractions of these peak pairs, and thus, $\pi_0 + \pi_1 = 1$. f_0, the non-cooperative density component, captures both independent and competitive binding (see section "Outline of the ChIP-seq Peak Intensity - Expectation Maximization Algorithm" in Results for an explanation of the assumptions underlying this model). Each of $\boldsymbol{\theta}_0^X, \boldsymbol{\theta}_1^X, \boldsymbol{\theta}_0^Y, \boldsymbol{\theta}_1^Y$ consist of two parameters, irrespective of whether f_1 is a Log-normal, Gamma or Gaussian density function. Along with π_0, there are thus a total of 9 parameters that we estimate from $\{(x_i, y_i)\}_{i=1}^N$ using the expectation-maximization (EM) algorithm [27, 28]. The output of the EM algorithm is a single set of parameters $\boldsymbol{\Theta} = (\pi_0, \boldsymbol{\theta}_0^X, \boldsymbol{\theta}_0^Y, \boldsymbol{\theta}_1^X, \boldsymbol{\theta}_1^Y)$ that maximizes the log-likelihood $\log P(\mathbf{D}, \mathbf{L}|\boldsymbol{\Theta})$, where $\mathbf{D}$ represents the peak intensity pairs $\{(x_i, y_i)\}_{i=1}^N$, and $\mathbf{L} = (L_1, L_2, \ldots, L_N)$ are labels assigned to each of the N locations, with $L_i = 1$ representing cooperative binding and $L_i = 0$ representing non-cooperative binding.

Our implementation of the expectation-maximization algorithm is described in the next section.

3. **For each overlapping peak pair, compute the probability that it cooperatively binds DNA.** Given the output of the EM algorithm, which is a set of parameters $\boldsymbol{\Theta}_i = (\pi_0, \boldsymbol{\theta}_0^X, \boldsymbol{\theta}_0^Y, \boldsymbol{\theta}_1^X, \boldsymbol{\theta}_1^Y)$, the probability that the TF pair X-Y cooperatively binds the i^{th} location can be computed using Bayes' rule —

$$\begin{aligned} &P(L_i = 1|X_i = x_i, Y_i = y_i) \\ &\quad = \frac{P(L_i = 1)P(X_i = x_i, Y_i = y_i|L_i = 1)}{\sum_{j=0}^{1} P(L_i = j)P(X_i = x_i, Y_i = y_i|L_i = j)} \\ &\quad = \frac{\pi_1 f_1(x_i; \boldsymbol{\theta}_1^X) f_1(y_i; \boldsymbol{\theta}_1^Y)}{\pi_0 f_0(x_i; \boldsymbol{\theta}_0^X) f_0(y_i; \boldsymbol{\theta}_0^Y) + \pi_1 f_1(x_i; \boldsymbol{\theta}_1^X) f_1(y_i; \boldsymbol{\theta}_1^Y)}. \end{aligned} \tag{2}$$

4. **Predict cooperative interactions.** To assign a label L_i of 0 or 1 to the $i-th$ location, we compare $P(L_i = 1|X_i = x_i, Y_i = y_i)$ to a threshold probability α. If the probability exceeds α, we set $L_i = 1$ and declare the $i-th$ location to be cooperatively bound. The parameter α controls the number of false positives and false negatives of the prediction: both quantities are increasing functions of α, and both tend to zero as α approaches one.

2.4 The Expectation-Maximization (EM) Algorithm

The input to the CPI-EM algorithm consists of peak intensity pairs $\mathbf{D} = \{(x_i, y_i)\}_{i=1}^N$. To label the observation (x_i, y_i) as cooperatively bound ($L_i = 1$) using equation (2), we need to simultaneously estimate $\mathbf{L} = \{L_1, L_2, \ldots, L_n\}$ and $\boldsymbol{\Theta}$ such that the log-likelihood $\log P(\mathbf{D}, \mathbf{L}|\boldsymbol{\Theta})$ is maximized with respect to $\boldsymbol{\Theta}$ and $\mathbf{L}$.

The expectation-maximization algorithm [27, 28] does this by computing a function $Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}')$, which is the expected value of the log-likelihood $\log P(\mathbf{D}, \mathbf{L}|\boldsymbol{\Theta})$, given an earlier estimate of $\boldsymbol{\Theta} = \boldsymbol{\Theta}'$ [29]:

$$Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}') = \sum_{\mathbf{L} \in S} \log \left(P\left(\mathbf{D}, \mathbf{L}|\boldsymbol{\Theta}\right) \right) P(\mathbf{L}|\mathbf{D}, \boldsymbol{\Theta}'), \tag{3}$$

where S represents the set of all possible values of L. The EM algorithm starts with an initial guess $\boldsymbol{\Theta}^{(0)}$, and computes a value $\boldsymbol{\Theta}^{(1)}$ such that $Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(0)})$ is maximized with respect to $\boldsymbol{\Theta}$, while $\boldsymbol{\Theta}^{(0)}$ is fixed. EM then computes $\boldsymbol{\Theta}^{(2)}$ in the next iteration to maximize $Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(1)})$ with respect to $\boldsymbol{\Theta}$, while $\boldsymbol{\Theta}^{(1)}$ is fixed. EM generates a sequence of values $\boldsymbol{\Theta}^{(0)}, \boldsymbol{\Theta}^{(1)}, \boldsymbol{\Theta}^{(2)}, \ldots, \boldsymbol{\Theta}^{(n)}$ which can be proven [27] to satisfy $Q(\boldsymbol{\Theta}^{(1)}, \boldsymbol{\Theta}^{(0)}) \leq Q(\boldsymbol{\Theta}^{(2)}, \boldsymbol{\Theta}^{(1)}) \leq \ldots Q(\boldsymbol{\Theta}^{(n)}, \boldsymbol{\Theta}^{(n-1)})$. EM terminates, say, at the $n-th$ iteration, when Q converges to a local maximum. This local maximum is guaranteed to be a local maximum of $\log P(\mathbf{D}, \mathbf{L}|\boldsymbol{\Theta})$ [29]. $\boldsymbol{\Theta}^{(n)}$ is then substituted in equation (2) to compute the probability of each peak intensity pair being labelled cooperative.

In our implementation of the EM algorithm, we used Powell's optimization method, as implemented in Scipy [30], to compute the value $\boldsymbol{\Theta}^{(k)}$, at the $k-th$ iteration, to maximize $Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(k-1)})$, where $\boldsymbol{\Theta}^{(k-1)}$ is fixed. We terminate the EM algorithm after n iterations if

$$\frac{|Q(\boldsymbol{\Theta}^{(n)}, \boldsymbol{\Theta}^{(n-1)}) - Q(\boldsymbol{\Theta}^{(n-1)}, \boldsymbol{\Theta}^{(n-2)})|}{|Q(\boldsymbol{\Theta}^{(n-1)}, \boldsymbol{\Theta}^{(n-2)})|} < 10^{-6}.$$

The set S of all possible labels $\mathbf{L}$ in equation (3) consists of 2^N elements, since each element of $\mathbf{L}$ takes on values of either 0 or 1. This is a very large number of terms that need to be added to evaluate Q. However, Q simplifies

to a sum over N terms for our model of cooperative binding. For the final analytical expression of the Q function employed in our algorithm, see Supplementary Section 6.

We choose the initial value $\boldsymbol{\Theta}^{(0)}$ as follows. From the data $\{(x_i, y_i)\}_{i=1}^N$, we separate the peak intensities of X and Y as $D_X = \{x_i\}_{i=1}^N$ and $D_Y = \{y_i\}_{i=1}^N$. We then compute the value $\boldsymbol{\theta}_{mle}^X$ that maximizes the likelihood $\prod_{i=1}^N f(x_i; \boldsymbol{\theta})$, where f is a Log-normal, Gamma or Gaussian density function. Similarly, we also compute the value of $\boldsymbol{\theta}_{mle}^Y$ that maximizes the likelihood $\prod_{i=1}^N f(y_i; \boldsymbol{\theta})$. These maximum likelihood estimates $\boldsymbol{\theta}_{mle}^X$ and $\boldsymbol{\theta}_{mle}^Y$ are computed using the `fit` function provided by the Python Scipy `stats` library, which can provide maximum likelihood estimates when f_0 and f_1 are either Log-normal, Gamma or Gaussian density functions. We choose $\pi_0^{(0)}$ from a Uniform$[0, 1]$ distribution. We finally set our initial parameter vector $\boldsymbol{\Theta}^{(0)}$ to $(\pi_0^{(0)}, \theta_{mle}^X, \theta_{mle}^Y, \theta_{mle}^X, \theta_{mle}^Y)$. We verified that EM converged to the same local maximum when $\boldsymbol{\Theta}^{(0)}$ was perturbed by up to 30% around this choice (data not shown).

2.5 Peak Distance Detector

For each peak intensity pair in $\{(x_i, y_i)\}_{i=1}^N$, the peak distance detector calculates the distance between the summits of X and Y peak regions. The 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 data sets 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 TFs are weaker than non-cooperatively bound 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 [17], CDX2-HNF4A in differentiated human intestinal stem cell lines (Caco-2) [18], and ERα-FOXA1 in the T-47D breast cancer cell line [19]. 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 data sets. 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, kernel density estimates of the marginal distributions of cooperative and non-cooperative FOXA1 and HNF4A peaks are shown. These distributions tended to be better approximated by a Log-normal distribution, which was evident from the higher log-likelihood value associated with a Log-normal 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 data set, a FOXA1 peak that has an intensity value of 5 is $\sim$3.45 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 Outline of the ChIP-seq Peak Intensity - Expectation Maximization algorithm

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

Briefly, ChIP-seq data of a TF pair X-Y that co-occupy N locations across the genome provides a set of peak intensity pairs $\{(x_i, y_i)\}_{i=1}^N$. In general, the joint distribution f_{XY} of these peak intensities can be written as a mixture (or sum) of densities, f_0 and f_1, that represent intensities from non-cooperatively and cooperatively bound regions, respectively —

$$f_{XY}(x_i, y_i) = \pi_0 f_0(x_i, y_i; \boldsymbol{\theta}_0) + \pi_1 f_1(x_i, y_i; \boldsymbol{\theta}_1). \quad (4)$$

π_0 and π_1 are the fractions of non-cooperative and cooperative peak pairs in the data, and sum to 1. $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ represent parameters of both these distributions whose values have to be estimated from the intensity pairs $\{(x_i, y_i)\}_{i=1}^{N}$. We estimate these parameters using the expectation-maximization algorithm [28, 29] (see the section "The Expectation-Maximization Algorithm" in Methods) . Once these parameters are estimated, the probability that each location is cooperatively bound can be computed using equation (2) in Methods.

We make three assumptions –

(1) We assume that the peak intensity distributions of X and Y are statistically independent irrespective of whether they bind DNA in a cooperative or non-cooperative fashion. This means that $f_0(x_i, y_i; \boldsymbol{\theta}_0) = f_0^X(x_i; \boldsymbol{\theta}_0^X) f_0^Y(y_i; \boldsymbol{\theta}_0^Y)$ and $f_1(x_i, y_i; \boldsymbol{\theta}_1) = f_1^X(x_i; \boldsymbol{\theta}_1^X) f_1^Y(y_i; \boldsymbol{\theta}_1^Y)$, which reduces equation (4) to equation (1).

We found this to be a reasonable assumption across all our data sets, when we calculated the mutual information (MI) [31] between peak intensities of cooperatively and non-cooperatively bound peak pairs, as determined by partner TF knockouts, across all our data sets. Mutual information, measured in bits, is a robust measure of statistical dependence between two random variables, whose value is zero if the variables are independent [32]. We found the MI between primary and partner TF peak intensities, at both cooperatively and non-cooperatively bound regions, to be close to zero across all data sets (Supplementary Table 3).

(2) We approximate f_0^X,f_0^Y,f_1^X and f_1^Y by a Log-normal distribution, as shown in Figure 2B. Data obtained from high-throughput experiments, which tend to be positively skewed, are known to be well approximated by Gamma and Log-normal distributions [33]. We found that the Log-normal distribution fit peak intensity distributions with a higher log-likelihood score than a Gaussian or Gamma distribution (Supplementary Table 2).

(3) We assume that the mean of $f_1^X(x_i; \boldsymbol{\theta}_1^X)$ is always less than $f_0^X(x_i; \boldsymbol{\theta}_0^X)$, i.e., that cooperatively bound primary TFs are always, on average, more weakly bound than non-cooperatively bound primary TFs. This was the case in each of the data sets we analyzed (Figure 2).

A histogram of the cooperative binding probabilities ($p \equiv P(L_i = 1|(x_i, y_i))$) of FOXA1-HNF4A peak intensity pairs, computed at each doubly bound location across the genome, is shown in Figure 3A. The fraction of peak intensity pairs in each probability bin in Figure 3A that are actually cooperatively bound (true positives, based on knockout data, as explained in Methods) is shown in Figure 3B. Peak intensity pairs with a higher value of p are more likely to be cooperatively bound. The choice of α finally used to conclusively label a peak intensity pair as cooperatively bound influences the fraction of erroneous inferences. The receiver operating characteristic (ROC) curve in Figure 3C shows the *false positive rates* and *true positive rates* of CPI-EM at different values of α. The false positive rate (FPR) is the fraction of non-cooperatively bound regions 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

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

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.

In Figure 3C, a large value of α, say 0.73, gives a lower FPR than $\alpha = 0.17$. However, a larger value of α also results in a smaller TPR value (Figure 3C). The area under ROC (auROC) 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 the CPI-EM algorithm

We computed the auROC of three variants of the CPI-EM algorithm on the data sets shown in Figure 2. These variants fit either Log-normal, Gamma, or Gaussian distributions to cooperative and non-cooperative peak intensity distributions. We compared the auROC of these variants to those from 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 data set (see Supplementary Section 5). An auROC of 0.5 represents the minimum level of detection performance that an algorithm should obtain to be considered a useful detector in practice. We plotted the ROC curves of each of the CPI-EM and peak distance algorithms (Supplementary Figure 1) for the data sets shown in Figure 2. The auROC of these four detectors is shown in Figure 4, along with a dotted line at an auROC of 0.5 that represents

the chance detector. The Gamma and Log-normal CPI-EM variants have an auROC of at least 0.5, and thus can consistently detect cooperative interactions across all data sets. The Log-normal CPI-EM variant fares well on all data sets, except for the mid-exponential phase CRP-FIS data set, 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 data sets, 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 Log-normal variants in detecting cooperative interactions.

There is considerable variation in the auROC of the peak distance based algorithm. The auROC is less than 0.5 in early-exponential phase CRP-FIS and RTG3-GCN4 data sets, but is higher than 0.5 in the remaining data sets. 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 > 0.5) performance obtained with the Log-normal and Gamma CPI-EM algorithms.

3.4 Detection of cooperative interactions in cell-type specific binding of transcription factors

Cooperative binding could be inferred through a knockout of one of the TFs in each of the data sets shown in Figures 2 and 4. We now consider two ChIP-seq data sets from cell lines, where a ChIP-seq of the primary TF from the knockout of the partner TF is not available. Instead, a ChIP-seq of the primary TF is carried out in a different cell line, where the concentration of the partner TF is low. We ran CPI-EM on ChIP-seq data sets of ERα-FOXA1 from T-47D breast cancer and ECC-1 endometrial cancer cell lines [19], and CDX2-HNF4A in proliferating and differentiated Caco-2 human intestinal stem cell lines [18].

In the ERα-FOXA1 data set, we compared ERα binding between T-47D and ECC-1 cell lines. T-47D cells express FOXA1 at a $\sim$50.1 fold higher concentration than ECC-1 cell lines, with this difference in concentration correlated with differences in ERα occupancy [19]. 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 data set, differentiated Caco-2 cells express HNF4A at a much higher concentration than proliferating Caco-2 cells [18]. 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 [18]. 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 data sets [19, 18]. Nevertheless, we wanted to see if CPI-EM could still detect cell-type specific binding in both these data sets.

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 data sets 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). In line with trends in Figure 4, the Log-normal CPI-EM variant was better at detecting cell-type specific binding across data sets compared to the Gamma and Gaussian variants. 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 data set, 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 data set. This is in contrast to the CPI-EM algorithm, which has an auROC greater than 0.5 in both data sets, with the Gamma and Log-normal CPI-EM variants giving an auROC of 0.71 in the CDX2-HNF4A data set.

4 Discussion

Cooperative binding is known to play a role in transcription factor binding site evolution and enhancer detection [34]. Cooperativity is also known to influence cis-regulatory variation between individuals of a species [35], which

could potentially capture disease-causing mutations that are known to occur in regulatory regions of the genome [36]. 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 [6, 7, 8, 9, 10, 11, 12, 13, 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. While 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), its consequence 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 data sets upon which we ran CPI-EM, and these interactions have been independently verified [1, 18]. A similar pattern of weakly but cooperatively bound TFs is seen in animal development. The binding of Ultrabithorax (Ubx) and Extradenticle (Exd) at the *shavenbaby* enhancer in *Drosophila melanogaster* embryos [37] 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 [37]. Similarly, low affinity binding sites that cooperatively bind Cubitus interruptus (Ci) at the *dpp* enhancer, which plays a crucial role in wing patterning in *Drosophila melanogaster*, are evolutionary conserved across twelve *Drosophila* species [38].

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 [39], while DNA-mediated cooperative binding often occurs between two TFs bound within 25 base pairs of each other [40, 14]. 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 peak distance was not a reliable criterion to detect cooperative binding. Protocols such as ChIP-exo [39] and ChIP-nexus [41] can resolve two binding events that are a single base pair apart [42]. 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 [43, 44]. 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 [41], that use molecular bar-coding techniques in DNA library preparation [45].

The additional variance introduced by PCR amplification might also explain the low MI values we measured between peak intensities of cooperatively bound TFs. 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 [46]. 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, while $f_0(x_i, y; \boldsymbol{\theta}_0)$ would still be equal to $f_0(x_i; \boldsymbol{\theta}_0^X) f_0(y_i; \boldsymbol{\theta}_0^Y)$, $f_1(x_i, y_i; \boldsymbol{\theta}_1) \neq f_1(x_i; \boldsymbol{\theta}_1^X) f_1(y_i; \boldsymbol{\theta}_1^Y)$ in equation (4). Instead, $f_1(x_i, y_i; \boldsymbol{\theta}_1)$ would contain an additional term that is an increasing function of MI. The precise form of such a function is not obvious, but would increase the probability that a high MI peak intensity pair would be labelled 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. analysed and interpreted the results, and wrote the manuscript.

References

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

[2] Rupali P Patwardhan, Joseph B Hiatt, Daniela M Witten, Mee J Kim, Robin P Smith, Dalit May, Choli Lee, Jennifer M Andrie, Su-In Lee, Gregory M Cooper, et al. Massively parallel functional dissection of mammalian enhancers in vivo. *Nature biotechnology*, 30(3):265–270, 2012.

[3] Robin P Smith, Leila Taher, Rupali P Patwardhan, Mee J Kim, Fumitaka Inoue, Jay Shendure, Ivan Ovcharenko, and Nadav Ahituv. Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. *Nature Genetics*, 45(9):1021–1028, 2013.

[4] Eilon Sharon, Yael Kalma, Ayala Sharp, Tali Raveh-Sadka, Michal Levo, Danny Zeevi, Leeat Keren, Zohar Yakhini, Adina Weinberger, and Eran Segal. Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nature Biotechnology*, 30(6):521–530, 2012.

[5] Murat Tuğrul, Tiago Paixão, Nicholas H Barton, and Gašper Tkačik. Dynamics of transcription factor binding site evolution. *PLoS Genet*, 11(11):e1005639, 2015.

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

[7] Tom Whitington, Martin C Frith, James Johnson, and Timothy L Bailey. Inferring transcription factor complexes from chip-seq data. *Nucleic Acids Research*, 39(15):e98–e98, 2011.

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

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

[10] Xin He, Chieh-Chun Chen, Feng Hong, Fang Fang, Saurabh Sinha, Huck-Hui Ng, and Sheng Zhong. A biophysical model for analysis of transcription factor interaction and binding site arrangement from genome-wide binding data. *PloS One*, 4(12):e8155, 2009.

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

[12] Soumyadeep Nandi, Alexandre Blais, and Ilya Ioshikhes. Identification of cis-regulatory modules in promoters of human genes exploiting mutual positioning of transcription factors. *Nucleic Acids Research*, page gkt578, 2013.

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

[14] Arttu Jolma, Yimeng Yin, Kazuhiro R Nitta, Kashyap Dave, Alexander Popov, Minna Taipale, Martin Enge, Teemu Kivioja, Ekaterina Morgunova, and Jussi Taipale. DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature*, 527(7578):384–388, 2015.

[15] Kamesh Narasimhan, Shubhadra Pillay, Yong-Heng Huang, Sriram Jayabal, Barath Udayasuryan, Veeramohan Veerapandian, Prasanna Kolatkar, Vlad Cojocaru, Konstantin Pervushin, and Ralf Jauch. DNA-mediated cooperativity facilitates the co-selection of cryptic enhancer sequences by SOX2 and PAX6 transcription factors. *Nucleic Acids Research*, page gku1390, 2015.

[16] David S Johnson, Ali Mortazavi, Richard M Myers, and Barbara Wold. Genome-wide mapping of in vivo protein-dna interactions. *Science*, 316(5830):1497–1502, 2007.

[17] Klara Stefflova, David Thybert, Michael D Wilson, Ian Streeter, Jelena Aleksic, Panagiota Karagianni, Alvis Brazma, David J Adams, Iannis Talianidis, John C Marioni, et al. Cooperativity and rapid evolution of cobound transcription factors in closely related mammals. *Cell*, 154(3):530–540, 2013.

[18] Michael P Verzi, Hyunjin Shin, H Hansen He, Rita Sulahian, Clifford A Meyer, Robert K Montgomery, James C Fleet, Myles Brown, X Shirley Liu, and Ramesh A Shivdasani. Differentiation-specific histone modifications reveal dynamic chromatin interactions and partners for the intestinal transcription factor cdx2. *Developmental Cell*, 19(5):713–726, 2010.

[19] Jason Gertz, Daniel Savic, Katherine E Varley, E Christopher Partridge, Alexias Safi, Preti Jain, Gregory M Cooper, Timothy E Reddy, Gregory E Crawford, and Richard M Myers. Distinct properties of cell-type-specific and shared transcription factor binding sites. *Molecular cell*, 52(1):25–36, 2013.

[20] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

[21] Heng Li and Richard Durbin. Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.

[22] Jianxing Feng, Tao Liu, Bo Qin, Yong Zhang, and Xiaole Shirley Liu. Identifying chip-seq enrichment using macs. *Nature Protocols*, 7(9):1728–1740, 2012.

[23] Qunhua Li, James B Brown, Haiyan Huang, and Peter J Bickel. Measuring reproducibility of high-throughput experiments. *The annals of applied statistics*, pages 1752–1779, 2011.

[24] Stephen G Landt, Georgi K Marinov, Anshul Kundaje, Pouya Kheradpour, Florencia Pauli, Serafim Batzoglou, Bradley E Bernstein, Peter Bickel, James B Brown, Philip Cayting, et al. Chip-seq guidelines and practices of the encode and modencode consortia. *Genome Research*, 22(9):1813–1831, 2012.

[25] J Michael Cherry, Eurie L Hong, Craig Amundsen, Rama Balakrishnan, Gail Binkley, Esther T Chan, Karen R Christie, Maria C Costanzo, Selina S Dwight, Stacia R Engel, et al. Saccharomyces genome database: the genomics resource of budding yeast. *Nucleic Acids Research*, page gkr1029, 2011.

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

[27] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

[28] Richard Durbin, Sean R Eddy, Anders Krogh, and Graeme Mitchison. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge University Press, 1998.

[29] Jeff A Bilmes et al. A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. *International Computer Science Institute*, 4(510):126, 1998.

[30] MJD Powell. Direct search algorithms for optimization calculations. *Acta numerica*, pages 287–336, 1998.

[31] Shuyang Gao, Greg Ver Steeg, and Aram Galstyan. Efficient estimation of mutual information for strongly dependent variables. In *AISTATS*, 2015.

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

[33] Chuan Lu and Ross D King. An investigation into the population abundance distribution of mrnas, proteins, and metabolites in biological systems. *Bioinformatics*, 25(16):2020–2027, 2009.

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

[35] S Heinz, CE Romanoski, C Benner, KA Allison, MU Kaikkonen, LD Orozco, and CK Glass. Effect of natural genetic variation on enhancer selection and function. *Nature*, 503(7477):487–492, 2013.

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

[37] Justin Crocker, Namiko Abe, Lucrezia Rinaldi, Alistair P McGregor, Nicolás Frankel, Shu Wang, Ahmad Alsawadi, Philippe Valenti, Serge Plaza, François Payre, et al. Low affinity binding site clusters confer hox specificity and regulatory robustness. *Cell*, 160(1):191–203, 2015.

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

[39] Ho Sung Rhee and B Franklin Pugh. Comprehensive genome-wide protein-dna interactions detected at single-nucleotide resolution. *Cell*, 147(6):1408–1419, 2011.

[40] Sangjin Kim, Erik Broströmer, Dong Xing, Jianshi Jin, Shasha Chong, Hao Ge, Siyuan Wang, Chan Gu, Lijiang Yang, Yi Qin Gao, et al. Probing allostery through DNA. *Science*, 339(6121):816–819, 2013.

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

[42] Stephan R Starick, Jonas Ibn-Salem, Marcel Jurk, Céline Hernandez, Michael I Love, Ho-Ryun Chung, Martin Vingron, Morgane Thomas-Chollier, and Sebastiaan H Meijsing. 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.

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

[44] Katharine Best, Theres Oakes, James M Heather, John Shawe-Taylor, and Benny Chain. Computational analysis of stochastic heterogeneity in PCR amplification efficiency revealed by single molecule barcoding. *Scientific Reports*, 5, 2015.

[45] Teemu Kivioja, Anna Vähärautio, Kasper Karlsson, Martin Bonke, Martin Enge, Sten Linnarsson, and Jussi Taipale. Counting absolute numbers of molecules using unique molecular identifiers. *Nature methods*, 9(1):72–74, 2012.

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

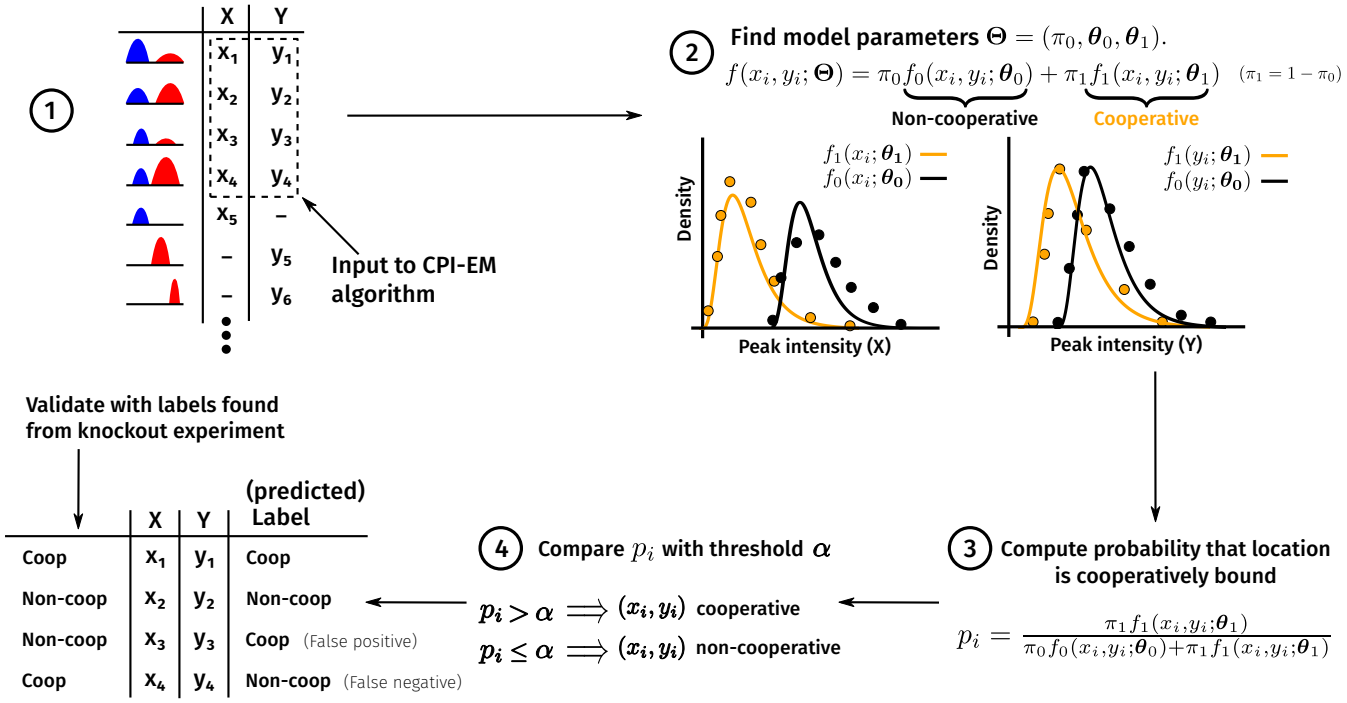

Figure 1: **A schematic of the CPI-EM algorithm** Steps shown in the figure correspond to those in the section "The ChIP-seq Peak Intensity - Expectation Maximization (CPI-EM) algorithm" in Methods. **(1)** The ChIP-seq of a TF provides a list of genomic locations bound by that TF, along with a peak intensity at each location that is a noisy measurement of the TF's binding affinity there. The input to the CPI-EM algorithm consists of a set of N peak intensity pairs $\{(x_i, y_i)\}$ from the ChIP-seq of TFs X and Y **(2)** CPI-EM considers the data $\{(x_i, y_i)\}_{i=1}^{N}$ to come from the probability mixture model shown, where f_0 and f_1 represent the probability density of peak intensity pairs from non-cooperatively and cooperatively bound regions, respectively. The expectation-maximization (EM) algorithm is employed to compute the mixture model parameters $\Theta = (\pi_0, \theta_0, \theta_1)$. **(3)** Given the estimates of the mixture model parameters, the probability p_i that the $i-th$ genomic location is cooperatively bound can be computed. This probability is computed across each of the N genomic locations bound by both X and Y. **(4)** Each of the probabilities p_1,p_2,... p_n is compared to a threshold α. Those locations where the probability of being cooperatively bound exceeds α are declared cooperatively bound. These locations are compared to a separate list of cooperatively bound locations, which is obtained from ChIP-seq experiments carried out on cells where Y is genetically knocked out (as described in Methods).

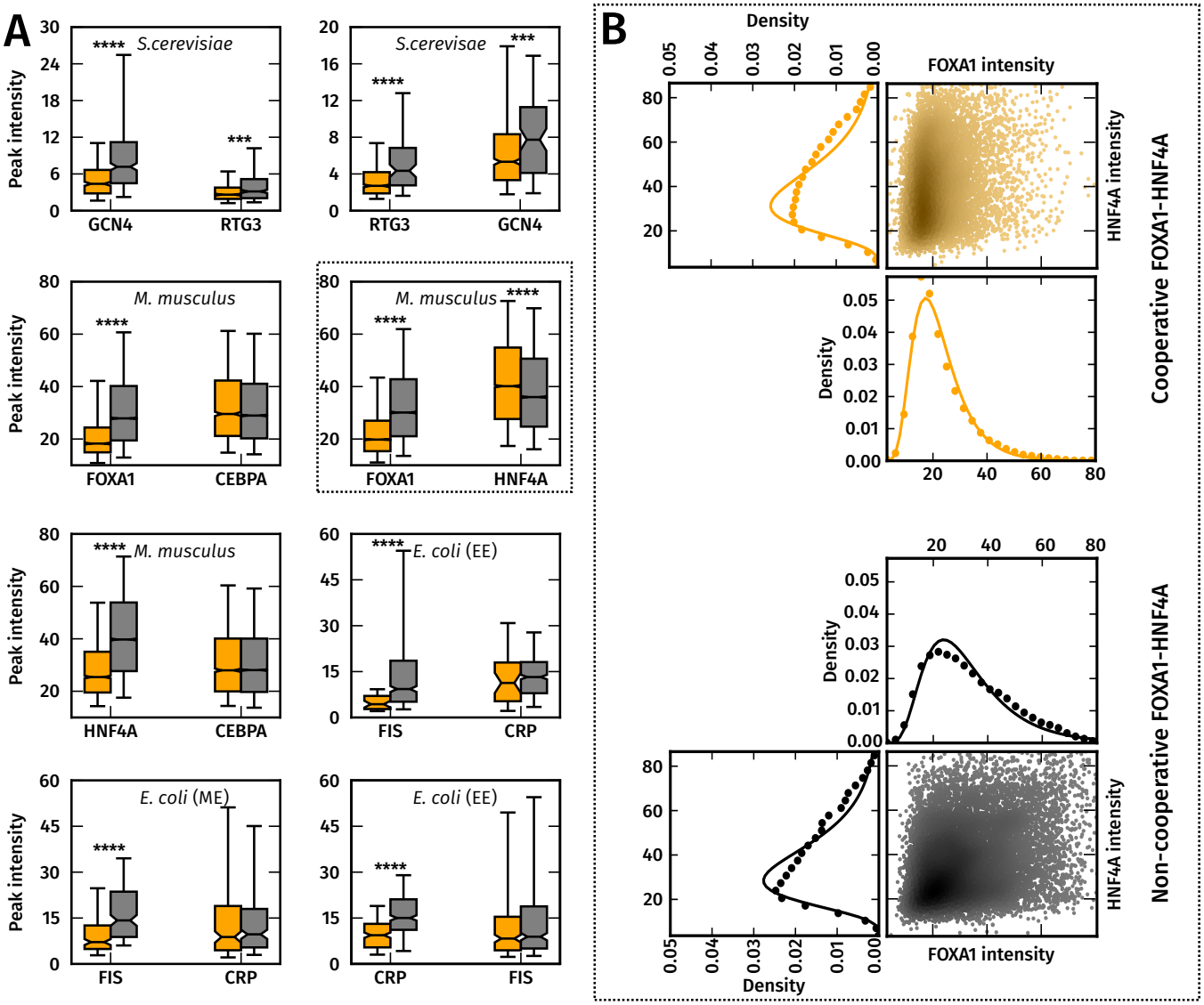

Figure 2: **(A) Cooperatively bound primary TFs are significantly more weakly bound than non-cooperatively bound primary TFs.** Box-plots of peak intensity distributions of cooperatively (orange) and non-cooperatively (gray) bound TF pairs, with primary TFs on the left and partner TFs on the right. ****, *** 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 Log-normal distribution. Marginal peak intensity distributions of FOXA1 and HNF4A peaks (in filled black and orange circles), with fitted Log-normal 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 empirical joint density of points in that region, with darker shades indicating a higher density. All empirical densities were computed using the Gaussian kernel density estimation procedure available in the Python Scipy library.

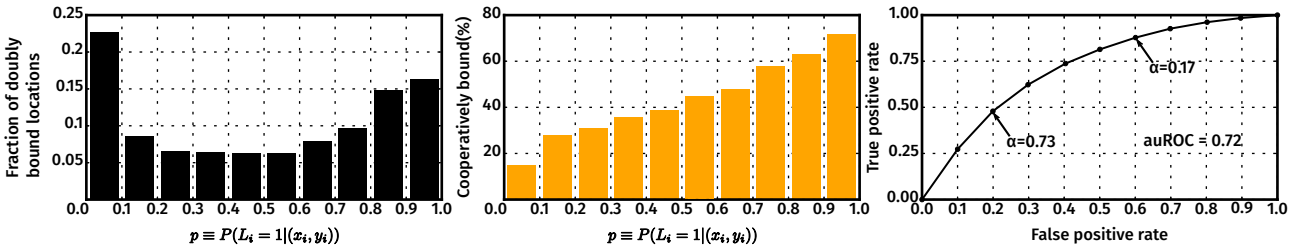

Figure 3: **(A) Histogram of the probabilities of genomic locations cooperatively bound by a FOXA1-HNF4A pair**. Those locations where this probability exceeds a threshold α are labelled as cooperatively bound. **(B) The percentage of true positives in each histogram bin**. This is the fraction of locations in each bin that are actually cooperatively bound by FOXA1-HNF4A, as determined from HNF4A knockout data. **(C) A receiver operating characteristic (ROC) curve to evaluate the performance of the CPI-EM algorithm in detecting cooperatively bound FOXA1-HNF4A locations.** 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 at detecting cooperative binding.

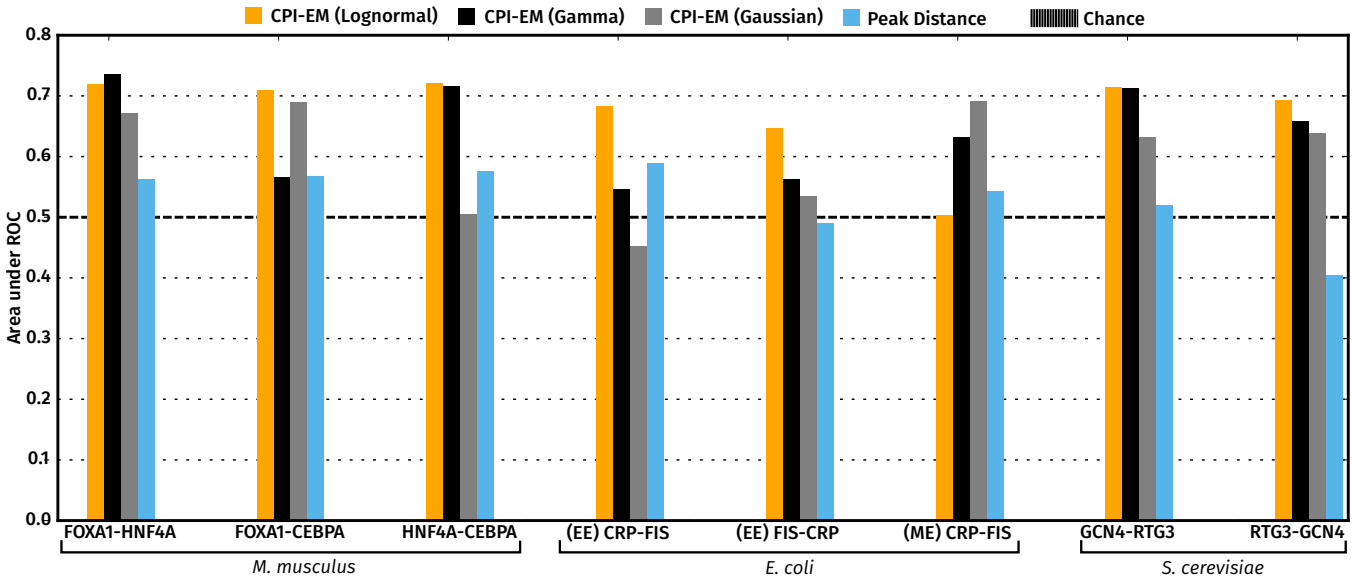

Figure 4: **The CPI-EM variant that fits lognormal distributions to peak intensity pairs consistently performs well across all data sets** The area under the ROC curve (auROC) of the CPI-EM algorithm applied to each of the data sets shown in Figure 2. CPI-EM variants that fit Log-normal, Gamma and Gaussian distributions are represented in orange, black and gray, respectively. The auROC of the peak distance based detector is shown in blue. For both the CPI-EM and peak distance algorithms, we calculated the ROC curve by picking thresholds that corresponded to false positive rates between 0.1 and 1 in steps of 0.1. The true positive rate at each of these thresholds was then computed, following wihch the area under the ROC was calculated using the trapezoidal integration rule implemented in the `numpy` v1.11.2 Python library.

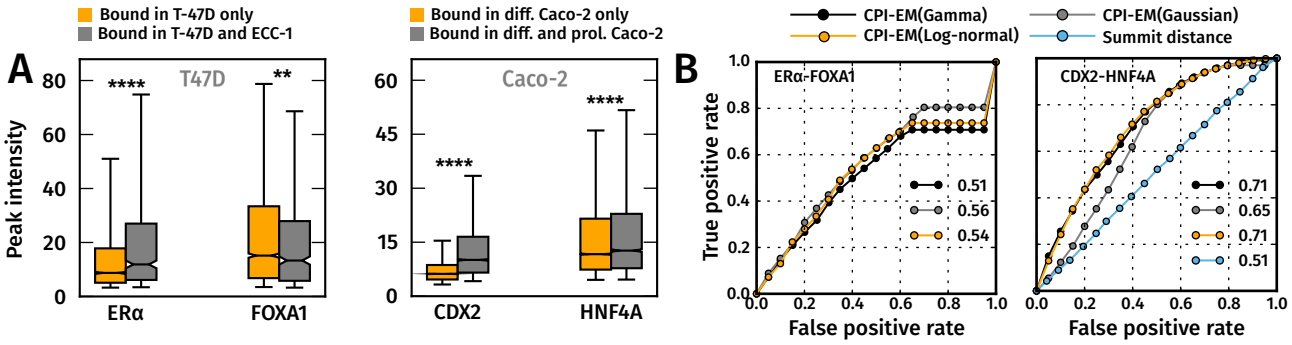

Figure 5: **(A) Regions bound by ERα only in T-47D cells are more weakly bound than regions bound by ERα in both T-47D and ECC-1 cells.** The same trend in peak intensities holds between regions bound by CDX2 only in differentiated Caco-2 cells and those bound by CDX2 in both differentiated and proliferating 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 Log-normal CPI-EM variant consistently detects cell-type specific binding events of ERα and CDX2**. ROC curves of Log-normal (orange), Gamma (black) and Gaussian (gray) variants of CPI-EM, and the peak distance detector (blue), on ERα-FOXA1 and CDX2-HNF4A data sets.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.