1 ChIP-seq pipeline for peak calling and processing knockout data

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 [10], 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).

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 [1] 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 [4]. We ran MACS2 (v2.1.0) [5], 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 dataset, we used the signal values of FOXA1, HNF4A and CEBPA peak calls as peak intensities.

We then applied a second criterion 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) [6, 7] of each peak. We computed the IDR of each peak with the idr script (v2.0) [6], 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.

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

We aligned raw sequence reads [8] 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 [9].

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 datasets, 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 $-\log_{10}(q-value)$ of GCN4 and RTG3 peak calls as peak intensities.

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 in the GEO database with accession number GSE32465 [2]. We also utilized pre-computed peak calls of CDX2 and HNF4A from Caco-2 cell lines that were publicly available (accession number GSE23436) [3].

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 datasets, we used signal values of peak calls as peak intensities. However, these signal values were scaled. In the ERα-FOXA1 dataset, we divided all signal values by 35, and in the CDX2-HNF4A dataset, 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.

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 datasets, 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 used the $-\log_{10}(q-value)$ of FIS and CRP peak calls as peak intensities for CPI-EM.

2 Summary of ChIP-seq data analyzed

Data set	n_{double} (n_{peaks}, f)	Losses	Increases	Decreases	Unchanged
FOXA1-HNF4A	21832 (30428,0.72)	7730 (9606,0.8)	1378 (2567,0.54)	2161 (2707,0.8)	10563 (15548,0.68)
FOXA1-CEBPA	14064 (30428,0.46)	1442 (2358,0.61)	502 (1706,0.29)	1090 (1814,0.6)	11030 (24550,0.45)
HNF4A-CEBPA	16235 (39955,0.41)	1939 (4219,0.46)	733 (2321,0.32)	1000 (1941,0.52)	12563 (31474,0.4)
(EE) FIS-CRP	293 (1545,0.19)	140 (935,0.15)	—	—	153 (610,0.25)
(ME) FIS-CRP	487 (594,0.82)	36 (70,0.51)	—	—	451 (524,0.86)
(EE) CRP-FIS	460 (1551,0.3)	176 (1050,0.17)	—	—	284 (501,0.57)
ERα-FOXA1	3269 (7369,0.44)	2738 (5689,0.48)	—	—	531 (1614,0.32)
CDX2-HNF4A	18398 (30963,0.59)	14291 (25664,0.56)	—	—	4107 (5299,0.78)
RTG3-GCN4	651 (1065,0.61)	562 (954,0.59)	—	—	89 (111,0.8)
GCN4-RTG3	722 (2407,0.3)	379 (1486,0.26)	—	—	343 (921,0.37)

Table 1: Summary of ChIP-seq data from datasets analyzed in Figure 2. The second column describes the total number of doubly bound regions in each dataset. In parentheses, the total number of primary TF peaks, and the fraction of primary TF peak regions that are doubly bound are shown ($f = n_{double}/n_{peaks}$). The remaining columns list the number of primary TF peaks that were lost, increased in rank, decreased in rank, or unchanged in rank, from genomic regions bound by both primary and partner TFs. The numbers in parentheses are the number of primary TF peaks that were lost, increased in rank, decreased in rank, or unchanged in rank, from all genomic regions bound by the primary TF. Statistical tests to detect significant peak rank changes could be carried out only in FOXA1-HNF4A, FOXA1-CEBPA and HNF4A-CEBPA datasets (see Supplementary Sections 1 and 3).

3 Detecting significant rank changes of primary TF peaks after the partner TF is knocked out

To determine if a change in the peak rank of X upon the knocking out of Y is statistically significant, we construct a null distribution that captures the magnitude of rank changes of X expected purely 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 X (in the presence of Y). 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 each of $\{r_i^{(1)}\}_{i=1}^n$ falls in the k^{th} bin. X 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 X, 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 X after Y has been knocked out. For this, we computed the ranks $r_1^{(m)}, r_2^{(m)}, \ldots, r_q^{(m)}$ from peaks of X 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 X after Y 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 Y.

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 [10] that sets the false discovery rate at 0.01. Statistics on the number of significant peak rank changes we observed in different datasets are shown in Supplementary Table 1.

We used this procedure to detect significant rank changes in FOXA1-HNF4A, FOXA1-CEBPA, and HNF4A-CEBPA datasets only. This was because (a) the ChIP-seq of FOXA1, HNF4A and CEBPA were carried out in replicates, and (b) the number of peaks that remained after IDR-based filtering was large. This gave us a sufficient number of peaks with which to reliably compute the null rank change distribution. This was not the case in the RTG3-GCN4 and GCN4-RTG3 datasets, where the number of ChIP-seq peaks in the merged alignments (89 in RTG3-GCN4 and 343 in GCN4-RTG3 datasets) was small. We could not detect rank changes in the ERα-FOXA1, CDX2-HNF4A, CRP-FIS, and FIS-CRP datasets because peak calls from individual replicates were not available, and hence, the null rank change distributions could not be computed.

4 CPI-EM : Estimating parameters required to compute the probability of cooperative binding at a location

The input to the CPI-EM algorithm consists of a set of peak intensity pairs $\mathbf{D} = \{(x_i, y_i)\}_{i=1}^N$, where $\{x_i\}$ and $\{y_i\}$ are peak intensites of the primary TF X and partner TF Y. We assume that the joint probability density of peak intensities from all these regions, $p(x, y)$, is a mixture (i.e., a sum) of two densities representing cooperative and non-cooperative peak intensity distributions:

$$p(x, y) = \pi_0 p_0(x, y; \boldsymbol{\theta}_0) + \pi_1 p_1(x, y; \boldsymbol{\theta}_1), \quad (1)$$

where p_0 and p_1 are the joint densities of peak intensities from non-cooperatively and cooperatively bound regions, respectively. $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ represent the parameters of both joint distributions. As shown in the main text, we make three assumptions in the CPI-EM algorithm, which we describe and justify in detail below –

- **Assumption 1 :** We assume that $p_0(x, y; \boldsymbol{\theta}_0) = p_0^X(x; \boldsymbol{\theta}_0^X) p_0^Y(y; \boldsymbol{\theta}_0^Y)$ and $p_1(x, y; \boldsymbol{\theta}_1) = p_1^X(x; \boldsymbol{\theta}_0^X) p_1^Y(y; \boldsymbol{\theta}_1^Y)$, where p_0^X,p_0^Y are marginal distributions of $p_0(x, y)$ and p_1^X,$p_1^Y(y)$ are marginal distributions of $p_1(x, y)$. The parameter vectors of the joint and marginal distributions are related as $\boldsymbol{\theta}_0 = (\boldsymbol{\theta}_0^X, \boldsymbol{\theta}_0^Y)$ and $\boldsymbol{\theta}_1 = (\boldsymbol{\theta}_1^X, \boldsymbol{\theta}_1^Y)$. This assumption reduces equation (1) to

$$p(x, y) = \pi_0 p_0^X(x; \boldsymbol{\theta}_0^X) p_0^Y(y; \boldsymbol{\theta}_0^Y) + \pi_1 p_1^X(x; \boldsymbol{\theta}_1^X) p_1^Y(y; \boldsymbol{\theta}_1^Y). \tag{2}$$

We found this to be a reasonable assumption across all our data sets when we calculated the mutual information (MI) [13] 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 statistically independent [12].

Given a probability disribution $p(x, y)$ over the set of peak intensity pairs $\{(x_i, y_i)\}_{i=1}^N$, mutual information (MI) is calculated as

$$MI = \sum_{i=1}^{n} \sum_{i=1}^{n} p(x_i, y_i) \log_2 \left(\frac{p(x_i, y_i)}{p_X(x_i) p_Y(y_i)} \right),$$

where $p_X(y)$ and $p_Y(y)$ are the marginal distributions of $p(x, y)$, and n is the number of $\{(x_i, y_i)\}$ pairs. MI is a non-negative quantity whose value is zero if X and Y are statistically independent i.e. if $p(x, y) = p_X(x) p_Y(y)$. From the knockout data available for each of the TF pairs, we separated the peak intensity pairs $\{(x_i, y_i\}_{i=1}^N$ into a set of cooperatively bound peak intensity pairs $A_c = \{(x_j, y_j)\}$ and a set of non-cooperatively bound peak intensity pairs $A_{nc} = \{(x_k, y_k)\}$. We separately computed the MI of peak intensity pairs in A_c (setting $p = p_1$ in equation (4)) and A_{nc} (setting $p = p_0$ in equation (4)) in each ChIP-seq knockout dataset we analyzed (Table 2). We found the MI between primary and partner TF peak intensities from both cooperatively and non-cooperatively bound regions to be close to zero across all data sets (Table 3).

Estimating MI through direct use of the definition specified by equation (4) leads to many problems; such MI estimates can be biased, or go to infinity in the case of certain distributions [13]. We estimated MI using the LNC algorithm implemented in [13] that circumvents these issues. The drawback of the LNC algorithm was that it gave non-negative values of MI only when a sufficient number of peaks were present. We were thus unable to reliably estimate MI in some of the ChIP-seq datasets we analyzed.

FOXA1-HNF4A	FOXA1-CEBPA	HNF4A-CEBPA	(EE) CRP-FIS	(EE) FIS-CRP
0.02, 0.04	–,0.03	0.01, 0.02	–, –	–, –

(ME) CRP-FIS	ERα-FOXA1	CDX2-HNF4A	GCN4-RTG3	RTG3-GCN4
–,–	–,–	0.29,0.09	–,0.05	0.03,0.12

Table 2: Each entry is a pair of mutual information values (in bits) between primary and partner peak intensity distributions, computed from peak intensities of cooperatively and non-cooperatively bound regions, respectively. Instances where the number of peak intensity pairs is too low for mutual information to be reliably estimated are shown as "–". The mutual information of most peak intensity pairs was very low. This makes it possible to approximate the joint density of peak intensity pairs as a product of marginal distributions.

- **Assumption 2 :** We choose $p_0^X, p_0^Y, p_1^X, p_1^Y$ to be either a Lognormal, Gamma or Gaussian density function, whose expressions and corresponding parameter sets are –

$$\begin{aligned}
\text{Lognormal: } p(x; m, \sigma) &= \frac{e^{-(\ln(x)/m)^2/2(\sigma)^2}}{x\sigma\sqrt{2\pi}} \qquad x \geq 0; m, \sigma > 0 \qquad \boldsymbol{\theta} = (m, \sigma), \\
\text{Gamma: } p(x; \gamma, \beta) &= \frac{(\frac{x}{\beta})^{\gamma-1} \exp(-\frac{x}{\beta})}{\beta\Gamma(\gamma)} \qquad x \geq 0; \gamma, \beta > 0 \qquad \boldsymbol{\theta} = (\beta, \gamma), \\
\text{Gaussian: } p(x; \mu, \sigma) &= \frac{\exp\left(-(x-\mu)^2/2\sigma^2\right)}{\sigma\sqrt{2\pi}} \qquad \sigma > 0 \qquad \boldsymbol{\theta} = (\mu, \sigma).
\end{aligned} \tag{3}$$

Across most datasets, we found that the Lognormal distribution tended to best fit peak intensity distributions. This

could be seen in terms of the log-likelihood scores obtained from fitting the three distributions individually to peak intensities of primary and partner TFs from cooperatively and non-cooperatively bound regions. The log-likelihood score obtained from fitting these distributions to a set of peak intensities of a given TF is computed as

$$\log\left(P(\mathbf{Z}|\mathbf{\Theta})\right) = \sum_{i=1}^{N} \log\left(p(z_i;\mathbf{\Theta})\right)$$

where p is either the cooperative (p_1) or non-cooperative (p_0) density. $\mathbf{Z}$ is $\{x_i\}_{i=1}^{N}$ or $\{y_i\}_{i=1}^{N}$, which are primary or partner TF peak intensities, respectively. $\mathbf{\Theta}$ are parameters of the distribution chosen for p. A larger log-likelihood value indicates a better fit to data.

We computed $\mathbf{\Theta}$ for each distribution using maximum likelihood estimates of these parameters. We used `fit` routines of the `stats` library of the Python package SciPy [11] to compute these estimates. The log-likelihood values calculated for each of the three distributions across all ChIP-seq datasets is shown in Table 3.

Dataset	Primary TF					
	Cooperative ($p_1^X(x)$)			Non-cooperative ($p_1^Y(y)$)		
	Lognormal	Gamma	Gaussian	Lognormal	Gamma	Gaussian
FOXA1-HNF4A	**−37367**	−37694	−39881	**−148661**	−150528	−164422
FOXA1-CEBPA	**−7413**	−7497	−8032	**−125906**	−125971	−134171
HNF4A-CEBPA	**−10360**	−10419	−10968	−145841	**−144745**	−149395
(EE) FIS-CRP	−82	**−81**	−88	**−1756**	−4064	−2063
(EE) CRP-FIS	−567	**−565**	−576	−1145	**−1144**	−1150
(ME) FIS-CRP	**−455**	−457	−518	−620	**−619**	−645
ERα-FOXA1	**−9573**	−9587	−11633	−2309	**−2087**	−2441
CDX2-HNF4A	**−39995**	−52278	−45658	**−107992**	−117804	−155132
GCN4-RTG3	**−931**	−933	−995	**−1132**	−1145	−1313
RTG3-GCN4	**−1283**	−1289	−1454	−240	**−239**	−263

Dataset	Partner TF					
	Cooperative ($p_0^X(x)$)			Non-cooperative ($p_0^Y(y)$)		
	Lognormal	Gamma	Gaussian	Lognormal	Gamma	Gaussian
FOXA1-HNF4A	−46926	**−46736**	−47168	**−162580**	−162688	−170887
FOXA1-CEBPA	−9556	**−9546**	−9793	**−123278**	−123870	−132830
HNF4A-CEBPA	**−11709**	−11710	−12085	**−128582**	−129452	−139457
(EE) FIS-CRP	−124	**−114**	−128	−1640	**−1636**	−1661
(EE) CRP-FIS	**−656**	−1329	−801	**−1199**	−2699	−1402
(ME) FIS-CRP	−546	**−542**	−642	**−621**	−1127	−713
ERα-FOXA1	−11377	**−11220**	−12652	−2102	**−2069**	−2421
CDX2-HNF4A	**−57440**	−57973	−65126	**−141792**	−261112	−192010
GCN4-RTG3	**−735**	−739	−843	**−797**	−801	−901
RTG3-GCN4	**−1831**	−1880	−2233	−291	**−289**	−313

Table 3: Log-likelihood values obtained from fitting log-normal, Gaussian and Gamma distributions to cooperative and non-cooperative peak intensities of the datasets shown in Table 1. The maximum log-likelihood values are indicated in bold. Across most datasets, the log-normal distribution typically provides the best fit to peak intensity distributions.

- **Assumption 3 :** A cooperatively bound primary TF is, on average, more weakly bound than a non-cooperatively bound primary TF. This implies that $\langle p_1^X(x)\rangle < \langle p_0^X(x)\rangle$. We found this to be a reasonable assumption since in Figure 2 in the main text, cooperatively bound primary TF peak intensities were significantly lower than non-cooperatively bound primary TF peak intensities across all datasets.

From equation (3) 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 p_1 is a Lognormal, Gamma or Gaussian density function. Along with π_0, there are thus a total of 9 parameters that we need to estimate

from $\mathbf{D}$ in order to compute the probability of each peak intensity pair in $\mathbf{D}$ being cooperative –

$$p_i^{coop} \equiv P(L_i = 1|x_i, y_i) = \frac{\pi_1 p_1(x_i, y_i; \boldsymbol{\theta}_1)}{\pi_1 p_1(x_i, y_i; \boldsymbol{\theta}_0) + \pi_0 p_0(x_i, y_i; \boldsymbol{\theta}_0).} \tag{4}$$

$$= \frac{\pi_1 p_1^X(x_i; \boldsymbol{\theta}_1^X) p_1^Y(y_i; \boldsymbol{\theta}_1^Y)}{\pi_0 p_0^X(x_i; \boldsymbol{\theta}_0^X) p_0^Y(y_i; \boldsymbol{\theta}_0^Y) + \pi_1 p_1^X(x_i; \boldsymbol{\theta}_1^X) p_1^Y(y_i; \boldsymbol{\theta}_1^Y)}. \tag{5}$$

4.1 The expectation-maximization (EM) algorithm

We use the expectation-maximization (EM) algorithm [14, 15] to estimate the parameters in equations (5) and (2). 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, where $L_i = 1$ represents cooperative binding and $L_i = 0$ represents non-cooperative binding.

The expectation-maximization algorithm [14, 15] 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 $\boldsymbol{\Theta} = \boldsymbol{\Theta}'$ [16]:

$$Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}') = \sum_{\mathbf{L} \in S} \log \big(P(\mathbf{D}, \mathbf{L}|\boldsymbol{\Theta})\big) P(\mathbf{L}|\mathbf{D}, \boldsymbol{\Theta}'), \tag{6}$$

where S represents the set of all possible values of L.

Briefly, 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 $\boldsymbol{\Theta}$, where $\boldsymbol{\Theta}^{(0)}$ is kept constant. EM then computes $\boldsymbol{\Theta}^{(2)}$ in the next iteration to maximize $Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(1)})$ with respect to $\boldsymbol{\Theta}$, where $\boldsymbol{\Theta}^{(1)}$ is kept constant. This iteration increases the value of Q, i.e., $Q(\boldsymbol{\Theta}^{(2)}, \boldsymbol{\Theta}^{(1)}) > Q(\boldsymbol{\Theta}^{(1)}, \boldsymbol{\Theta}^{(0)}$. Thus, one run of the EM procedure generates a sequence of values $\boldsymbol{\Theta}^{(0)}, \boldsymbol{\Theta}^{(1)}, \boldsymbol{\Theta}^{(2)}, \ldots, \boldsymbol{\Theta}^{(n)}$ which can be proven [14] 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 $\log P(\mathbf{D}, \mathbf{L}|\boldsymbol{\Theta})$ [16]. $\boldsymbol{\Theta}^{(n)}$ is then substituted in equation (5) to compute the probability of each peak intensity pair being labeled cooperative.

We now describe the Q function employed in CPI-EM, and the implementation of the EM iteration process in detail below.

The set S of all possible labels $\mathbf{L}$ in equation (6) consists of 2^N elements because 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. Q can be rewritten as

$$Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}') = \sum_{\mathbf{L} \in S} \log \big(P(\mathbf{D}, \mathbf{L}|\boldsymbol{\Theta})\big) P(\mathbf{L}|\mathbf{D}, \boldsymbol{\Theta}') = \sum_{\mathbf{L} \in S} \log \big(P(\mathbf{L}|\mathbf{D}, \boldsymbol{\Theta}) P(\mathbf{D}|\boldsymbol{\Theta})\big) P(\mathbf{L}|\mathbf{D}, \boldsymbol{\Theta}') \tag{7}$$

Since a peak intensity pair is either cooperative or non-cooperative, we can write $P(X_i = x_i, Y_i = y_i) = \pi_0 p_0(x_i, y_i; \boldsymbol{\theta}_0) + \pi_1 p_1(x_i, y_i; \boldsymbol{\theta}_1)$, where $\pi_0 + \pi_1 = 1$. Since we consider (X_i, Y_i) and (X_j, Y_j) $(i \neq j)$ to be statistically independent,

$$\log P(\mathbf{D}|\boldsymbol{\Theta}) = \sum_{i=1}^{N} \log \left(\pi_0 p_0^X(x_i; \boldsymbol{\theta}_0^X) p_0^Y(y_i; \boldsymbol{\theta}_0^Y) + \pi_1 p_1^X(x_i; \boldsymbol{\theta}_1^X) p_1^Y(y_i; \boldsymbol{\theta}_1^Y) \right)$$

$P(\mathbf{L}|\mathbf{D}, \boldsymbol{\Theta})$ in equation (7) can be expanded as –

$$P(\mathbf{L}|\mathbf{D}, \boldsymbol{\Theta}) = \prod_{i=1}^{N} P(L_i = l_i|\mathbf{D}, \boldsymbol{\Theta})$$

$$= \prod_{i=1}^{N} \frac{\pi_{l_i} p_{l_i}(x_i, y_i; \boldsymbol{\theta}_{l_i})}{\pi_0 p_0(x_i, y_i; \boldsymbol{\theta}_0) + \pi_1 p_1(x_i, y_i; \boldsymbol{\theta}_1)},$$

where l_i is 0 or 1. Substituting the above two expansions into the expression for Q in equation (7), it can be shown that

Q simplifies to the form shown below, where it is a sum over only N terms (page 4 in [16])

$$\begin{aligned} Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}') &= \sum_{i=1}^{N}\sum_{l=1}^{2} \log(\pi_l) P(L_i = l|x_i, y_i, \boldsymbol{\Theta}') + \sum_{i=1}^{N}\sum_{l=1}^{2} P\big(L_i = l|x_i, y_i, \boldsymbol{\Theta}\big) \log\Big(p_l^X(x_i; \boldsymbol{\theta}_l^{X'}) p_l^Y(y_i; \boldsymbol{\theta}_l^{Y'})\Big) \\ &= Q_1(\boldsymbol{\Theta}') + Q_2(\boldsymbol{\Theta}, \boldsymbol{\Theta}'), \end{aligned} \tag{8}$$

where, $\boldsymbol{\Theta}' = (\pi_0', \boldsymbol{\theta}_0^{X'}, \boldsymbol{\theta}_0^{Y'}, \boldsymbol{\theta}_1^{X'}, \boldsymbol{\theta}_1^{Y'})$. Note that the first term is independent of $\boldsymbol{\Theta}'$, so it can be maximized independently of the second term. The choice of π_l that maximizes the first term (page 5 in [16]) is –

$$\pi_l = \frac{1}{N}\sum_{i=1}^{N} P(L_i = l|x_i, y_i, \boldsymbol{\Theta}') \text{ for } l = 0, 1.$$

The $k-th$ EM iteration involves choosing a value $\boldsymbol{\Theta} = \boldsymbol{\Theta}^{(k+1)}$ that maximizes $Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(k)})$, where $\boldsymbol{\Theta}^{(k)}$ is kept fixed at the value obtained in the previous EM iteration that maximizes $Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(k-1)})$. EM involves two steps, an E-step and an M-step, which are both needed to maximize $Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(k)})$. The E- and M- steps evaluated at the $i-th$ iteration in our algorithm are [17] –

- E-step : Compute

$$P(L_i = l|x_i, y_i, \boldsymbol{\Theta}^{(k)}) = \frac{{\pi_l}^{(k)} p_l(x_i, y_i; \boldsymbol{\Theta}^{(k)})}{{\pi_0}^{(k)} p_0(x_i, y_i; \boldsymbol{\Theta}^{(k)}) + {\pi_1}^{(k)} p_1(x_i, y_i; \boldsymbol{\Theta}^{(k)})} \text{ for } l = 0, 1;\ i = 1, 2, \ldots, N.$$

- M-step (1) : Compute

$$\pi_l^{(k+1)} = \sum_{i=1}^{N} P(L_i = l|x_i, y_i, \boldsymbol{\Theta}^{(k)}) \text{ for } l = 0, 1$$

  This step maximizes Q_1 in equation (8).

- M-step (2) : Use Powell's gradient search method, as implemented in the Scipy optimization toolbox [18] to find $\boldsymbol{\Theta}^{(k+1)}$ that maximizes $Q_2(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(k)})$ in equation (8), with $\boldsymbol{\Theta}^{(k)} = (\pi_0^{(k)}, \boldsymbol{\theta}_0^{X,(k)}, \boldsymbol{\theta}_0^{Y,(k)}, \boldsymbol{\theta}_1^{X,(k)}, \boldsymbol{\theta}_1^{Y,(k)})$ kept constant –

$$Q_2(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(k)}) = \sum_{i=1}^{N}\sum_{l=1}^{2} P\big(L_i = l|x_i, y_i, \boldsymbol{\Theta}\big) \log\Big(p_l^X(x_i; \boldsymbol{\theta}_l^{X,(k)}) p_l^Y(y_i; \boldsymbol{\theta}_l^{Y,(k)})\Big)$$

We terminate the EM algorithm after n iterations if

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

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 $\prod_{i=1}^N p(x_i; \boldsymbol{\theta})$, where f is a Lognormal, Gamma or Gaussian density function. Similarly, we also compute the value of $\boldsymbol{\theta}_{mle}^Y$ that maximizes the likelihood $\prod_{i=1}^N p(y_i; \boldsymbol{\theta})$. 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 p_0 and p_1 are either Lognormal, 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).

5 Calculation of receiver operating characteristic (ROC) curves

Given probability of cooperative binding $p_1^{coop}, p_2^{coop}, \ldots, p_N^{coop}$, the ROC curve was calculated by picking thresholds α on these probabilities 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. The area under the ROC curve was then calculated using the trapezoidal integration rule available in the Python `numpy` library. This procedure was repeated for each of the three variants of the CPI-EM algorithm. Similarly, the ROC curve of the peak distance algorithm was computed by choosing thresholds on the peak distance that corresponded to false positive rates between 0.1 and 1 in steps of 0.1. After the true positive rate at each threshold was calculated, the area under the ROC was computed with the trapezoidal integration rule.

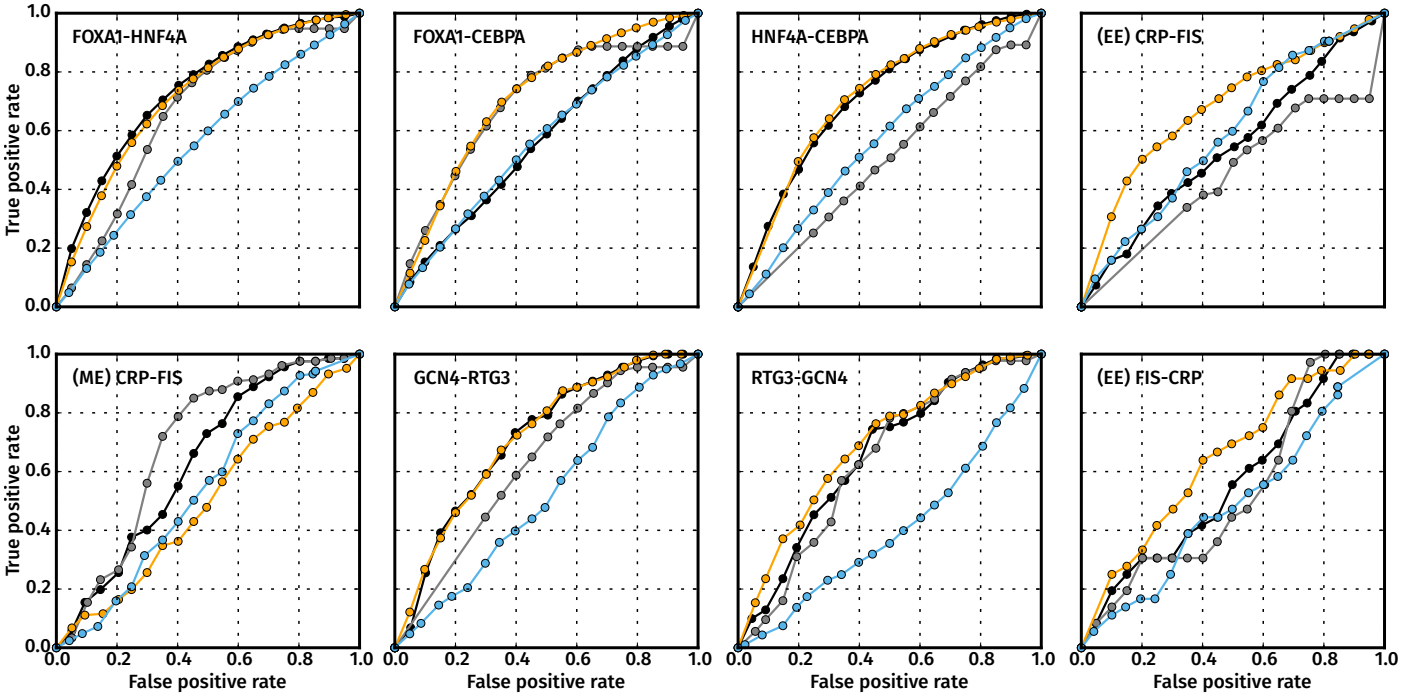

Figure 1: ROC curves of runs of the CPI-EM variants (log-normal in orange, Gamma in black, and Gaussian in gray) and peak distance (sky blue) algorithms on datasets in Figure 4 of the main text, and in Table 1 above.

6 Area under ROC of a chance detector is 0.5

The chance detector is based purely on using tosses from a biased coin to detect cooperative interactions. Let the probability of the coin showing heads be α. Out of a set of N peak intensity pairs $\{(x_i, y_i)\}$, suppose there are N_c and N_{nc} cooperatively and non-cooperatively bound pairs, respectively. The number of false positives, resulting from N tosses of the coin, would be $N_{nc}\alpha$, while the number of true positives would be $N_c\alpha$. This means that both the FPR and TPR of the chance detector would be α. Thus, as α is varied between 0 and 1, the ROC of the chance detector will be the straight line $FPR = TPR = \alpha$, which encloses an area of 0.5.

References

[1] 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.

[2] 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.

[3] 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.

[4] Li, H. and Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.

[5] Feng, J., Liu, T., Qin, B., Zhang, Y., and Liu, X. S. Identifying ChIP-seq enrichment using MACS. *Nature Protocols*, 7(9):1728–1740, 2012.

[6] Li, Q., Brown, J. B., Huang, H., and Bickel, P. J. Measuring reproducibility of high-throughput experiments. *The annals of applied statistics*, pages 1752–1779, 2011.

[7] Landt, S. G., Marinov, G. K., Kundaje, A., Kheradpour, P., Pauli, F. , Batzoglou, S., Bernstein, B. E., Bickel, P., Brown, J. B., Cayting, P. , et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Research*, 22(9):1813–1831, 2012.

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

[9] Cherry, J. M., Hong, E. L., Amundsen, C., Balakrishnan, R., Binkley, G. , Chan, E. T., Christie, K. R., Costanzo, M. C., Dwight, S. S., Engel, S. R., et al. Saccharomyces genome database: the genomics resource of budding yeast. *Nucleic Acids Research*, page gkr1029, 2011.

[10] Benjamini, Y. and Hochberg, Y. . 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. *Bioinformatics*, 27(24):3423–3424, 2011.

[11] Eric Jones, Travis Oliphant, and Pearu Peterson. {Scipy}: open source scientific tools for {Python}. 2014.

[12] 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.

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

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

[15] Durbin, R. , Eddy, S. R. , Krogh, A. , and Mitchison, G. . *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge University Press, 1998.

[16] Jeff A. Bilmes and others A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models *International Computer Science Institute*, 510(4):126, 1998.

[17] Ioannis Kosmidis and Dimitris Karlis. Model-based clustering using copulas with applications. *Statistics and Computing*, 1–21, 2015.

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