

Supplementary Methods

ChIP-Chip The ChIP-chip protocol used to generate the Gcn4 and Mig2 data for this paper is the same as in Pokholok *et al*¹. Supplementary Figure 1 shows the basic steps of the protocol: crosslinking, sonication, enrichment, labeling, and hybridization.

In analyzing the Mig2 data, we noticed a high false positive rate. Comparing the Mig2 data to a control experiment (an experiment in which there was no Myc-tagged protein for the IP antibody to bind) revealed that the Mig2 IP material included a strong non-specific signal in addition to the specific Mig2 signal. Supplementary Figure 2 shows the unnormalized Mig2 enrichment ratios and the control experiment ratios in a sample region in which the Mig2 signal includes both components. Consequently, for both the Mig2 data and the Gcn4 data, we performed linear regression between the two signals and then subtracted the non-specific component (the component predicted by the control experiment) from the enrichment ratios before using them further.

Our Gcn4 data was generated in rich media conditions (YPD). As Gcn4 is best known for its role in the cellular response to amino acid starvation, we checked the genes bound in our dataset against the MIPS gene function categories. Supplementary Table 1 shows that, as expected, the targets identified in our data are enriched for categories relating to amino acid synthesis and metabolism.

Influence Function To compute the influence function for Gcn4 and other factors, we ran amplified and labeled material on an Agilent BioAnalyzer 2100. The machine produced an output

table giving measured fluorescent intensity over time. Using the 15bp and 1500bp standards as endpoints, we fit each sample's data to the standards ladder to obtain intensity as a function of length. The signal from very short fragments was discarded to exclude the effects of primers; this effectively limits the smallest fragment we consider to about 50bp. Such small fragments are unlikely to hybridize well to the microarray, so ignoring their effect should not influence our results. We also truncated the long end of the distribution to exclude the 1500bp standard and associated noise. We have noticed that running a large amount of material on the BioAnalyzer leads to the apparently false presence of very long fragments (1kb-1500bp). Running varying concentrations of DNA for each sample allows one to see when this problem occurs.

We divided each intensity by its corresponding length as the material was labeled with tagged dT molecules, leading to a roughly linear increase in fragment intensity with length. The resulting values represent the relative concentration of molecules at length l rather than the sum intensity of molecules of length l . Fitting each sample to the standards gave non-integer distances, so we rounded each length to the nearest integer, combined the measurements at each length from the two replicates, and averaged the values at each integer length.

The concentration versus length curve was smooth overall, but locally noisy, so we applied a low pass filter to the data. If $c(l)$ is the concentration of fragments of length l , we computed $c'(l) = 0.1c(l - 2) + 0.2c(l - 1) + 0.5c(l) + 0.2c(l + 1) + 0.1c(l + 2)$. To fit $c'(l)$ to a gamma distribution, we shifted c' such that the minimum non-zero index was zero. β is the variance of the shifted distribution over its mean. α is the mean over beta. We used the gamma distribution

with these parameters as $p(l)$, the probability of a fragment of length l . To get $p_a(y)$, note that the convolution of two identical gamma distributions is a gamma distribution with twice the mean. Since $p(l) = \sum_{y=0}^l p_a(y)p_a(l-y)$, $p_a(y)$ is a gamma distribution with half the mean of $p(l)$. Figure 1 shows the relatively close fit between the observed $p(l)$ and the fitted gamma distribution.

The influence function

$$f(d) = \sum_{l=d}^{\infty} l \sum_{a=d}^l p_a(a)p_a(l-a) \quad (1)$$

determines the intensity at distance d from a binding event by summing over all fragments of length d or longer, that is, all fragments that include both the binding event and the probe. The multiplicative term l accounts for the fact that a labeled molecule's fluorescent intensity increases roughly linearly with its length as more labeled nucleotides are incorporated. The inner term sums over all fragments of length l by accounting for all positions at which the binding site can fall in a fragment of length l while still being at most d bases from the probe.

We determined the influence function for several chromatin modifying factors (Set1, Set2, and Gcn5), histone H3, and histone H3 acetylated at lysine 14 in yeast; Sox2, Nanog, Eed, and Oct4 in human embryonic stem cells; and E2F4, Hnf1a, Hnf4a, and Hnf6 in human liver tissue. Supplementary Figure 4 shows the similarity between influence functions for samples prepared with the same protocol. The other five yeast influence functions were remarkably similar, yet differed from that for Gcn4. We suspect that the function for Gcn4 differs from the others because it is a transcription factor and binds to a small DNA target whereas the other proteins target histones, which have a large DNA footprint, or are histones themselves. The influence functions for human

transcription factors in stem cells and liver tissue also demonstrate the consistent shape of the influence function for a particular protocol and sample preparation.

Model First, we need to obtain the input \mathbf{y} to the JBD model. Instead of directly computing the enrichment ratio \mathbf{y} as $\mathbf{y}_{IP}/\mathbf{y}_{WCE}$, where \mathbf{y}_{IP} and \mathbf{y}_{WCE} are the signal intensities in IP and whole cell extract (WCE is genomic DNA extracted from the cells without any selection) channels, we compute \mathbf{y} as an attenuated ratio:

$$\mathbf{y} = \frac{\mathbf{y}_{IP} + \delta}{\mathbf{y}_{WCE} + \delta}. \quad (2)$$

. This is the Bayesian estimate of the enrichment ratio when both \mathbf{y}_{IP} and \mathbf{y}_{WCE} are modeled as Gamma distributions². When \mathbf{y}_{IP} and \mathbf{y}_{WCE} are large, this Bayesian estimate will be quite close to the simple ratio $\mathbf{y}_{IP}/\mathbf{y}_{WCE}$. When \mathbf{y}_{IP} and \mathbf{y}_{WCE} are small, there is strong attenuation of this Bayesian estimate. This attenuation naturally accounts for our lower confidence in small signal intensities. In our experiments, we used $\delta = 1$. Besides using this attenuation, we also preprocess the enrichment ratio \mathbf{y} by subtracting 1 from it, such that if there is no binding the mean value of \mathbf{y} would be 0.

The Bayesian model that underlies JBD contains potentially millions of unknown variables that describe potential binding events. The model can be greatly simplified by spatially partitioning it into chunks that are independently estimated, but even so partitioned, the remaining problems are still computationally challenging and necessitates novel Bayesian approximation methods.

In the JBD model, the likelihood function for the data is

$$p(\mathbf{y}|\mathbf{b}, \mathbf{s}) = \prod_k \prod_i p(y_i^k | \mathbf{b}, \mathbf{s}) \quad (3)$$

$$= \prod_k \prod_i \mathcal{N}(y_i^k | \sum_{j: f_{|i-j|} > 0} f_{|i-j|} s_j b_j, \sigma_i). \quad (4)$$

where k indexes experimental replicates, i indexes the probe positions, j indexes the binding positions, and $\mathcal{N}(\cdot | \sum_j f_{|i-j|} s_j b_j, \sigma_i)$ represents the probability density function of a Gaussian distribution with mean $\sum_j f_{|i-j|} s_j b_j$ and variance σ_i . The influence function $f_{|i-j|}$ is shown in equation (1). Note that the variance σ_i models the uncertainty between replicates for probe i .

We assign prior distributions on the binding event b_j and the binding strength s_j :

$$p(b_j | \pi_j) = \pi_j^{b_j} (1 - \pi_j)^{1-b_j} \quad (5)$$

$$p_0(s_j) = \text{Gamma}(s_j | c_0, d_0) \quad (6)$$

where $\text{Gamma}(\cdot | c_0, d_0)$ stands for the probability density functions of Gamma distributions with hyperparameters c_0 and d_0 . Based on our experience with binding data, we choose $c_0 = 3$ and $d_0 = 1$ such that the mean and the variance of $p_0(s_j)$ are 3 and 3, respectively. Such a choice of hyperparameters c_0 and d_0 encodes our expectation of a binding strength around 3 with a reasonably large uncertainty.

We assign a hyperprior distribution on the binding probability π_j as:

$$p_0(\pi_j) = \text{Beta}(\pi_j | \alpha_0, \beta_0) \quad (7)$$

For different types of binding events, we can choose different hyperparameters α_0 and β_0 accordingly. For example, for transcription factors, we choose $\alpha_0 = 0.08$ and $\beta_0 = 0.72$ such that the

mean and the variance $p_0(\pi_j)$ are 0.1 and 0.05, respectively. This choice encodes our prior belief that the binding events are sparse for transcription factors. However, for RNA Polymerase II (Pol2) binding, we may want to have a larger value for the mean $p_0(\pi_j)$, which would represent our belief that Pol2 binds to DNA more frequently than a transcription factor.

Bayesian estimation of variance Before we can carry out inference on the JBD model, we need to specify the variance σ_i for each probe in the likelihood function (4). To do so, we use a Bayesian approach that not only models the uncertainty in the actual observation, but also uses the prior knowledge that the variance increases with the observed value.

We first assign a conjugate prior distribution on the variance σ_i . This prior distribution is a scaled inverse-chi-square distribution,

$$p_0(\sigma_i^2) = \text{Inv-}\chi^2(\sigma_i^2 | n_0, \gamma_{0i}) \quad (8)$$

where n_0 and γ_{0i} are the hyperparameters of the prior distribution. To obtain the hyperparameters, we use the observations from previous experiments to model the relationship between the expected standard deviation ξ_i and the value of the expected observation. A simple quadratic fitting gives us the following function:

$$\xi_i = -0.0014\bar{y}_i^2 + 0.255\bar{y}_i - 0.183, \text{ for } \bar{y}_i < 91$$

where $\bar{y}_i = \frac{1}{n} \sum_{k=1}^n y_i^k$ and n is the number of replicates. If $\bar{y}_i > 91$, we set $\xi_i = 11.43$. we also apply a lower bound, 0.05, to ξ_i to avoid being overconfident in the low ratio region. This function reflects our observation that the variance increases with the observation value. Because we used

roughly 100 data points to estimate this quadratic function, we assign the degree of freedom n_0 to be 100. Given ξ_i and n_0 , we can compute the hyperparameter

$$\gamma_{0i} = \frac{n_0 - 2}{n_0} \xi_i^2.$$

Now define

$$\gamma_i = \frac{1}{n} \sum_{k=1}^n (y_{i,k} - \bar{y}_i).$$

Then based on the prior (8) and the likelihood function (4), the posterior of the variance is

$$p(\sigma_i^2 | \mathbf{y}) = \text{Inv-}\chi^2 \left(\sigma_i^2 | n_0 + n, \frac{n_0 \gamma_{0i} + n \gamma_i}{n_0 + n} \right). \quad (9)$$

Then, we use the posterior mean as the estimate of the variance:

$$\sigma_i = \frac{n_0 \gamma_{0i} + n \gamma_i}{n_0 + n - 2}. \quad (10)$$

This estimate combines our prior knowledge with the actual uncertainty in the data through weighted averaging.

Bayesian estimation of latent variables by Expectation Propagation First, given the data likelihood (4), the prior distributions (5) and (6) on the binding event \mathbf{b} and strength \mathbf{s} , and the hyperprior distribution (7) on the binding probability $\boldsymbol{\pi}$, the posterior distribution $p(\mathbf{b}, \mathbf{s}, \boldsymbol{\pi} | \mathbf{y})$ is proportional to the joint distribution $p(\mathbf{b}, \mathbf{s}, \boldsymbol{\pi}, \mathbf{y})$:

$$p(\mathbf{b}, \mathbf{s}, \boldsymbol{\pi} | \mathbf{y}) \propto p(\mathbf{b}, \mathbf{s}, \boldsymbol{\pi}, \mathbf{y}) = \prod_i g_i(\mathbf{b}, \mathbf{s}) \prod_j p_0(\pi_j) f_j(b_j, \pi_j) p_0(s_j)$$

where i indexes probe positions, j indexes binding positions, $f_j(b_j, \pi_j) = p(b_j|\pi_j)$ is the prior for b_j , $g_i(\mathbf{b}, \mathbf{s}) = \mathcal{N}(y_i | \sum_j f_{|i-j|} s_j b_j, \sigma_i)$ is the likelihood for the observation at the i^{th} probe position, $p_0(\pi_j)$ is the hyperprior distribution of π_j , and $p(b_j|\pi_j)$ and $p_0(s_j)$ are the prior distributions of b_j and s_j , respectively. For simplicity and clarity, here we drop the superscript k , which indexes replicates, and only consider the case of one replicate. Since the posterior distribution $p(\mathbf{b}, \mathbf{s}, \boldsymbol{\pi} | \mathbf{y})$ cannot be computed in a closed form, we use EP to approximate this complicated posterior distribution by a distribution in the exponential family.

EP exploits the fact that the posterior is a product of simple terms. EP iteratively refines the approximation of each term to improve the approximation of the posterior. Mathematically, EP approximates $p(\mathbf{b}, \mathbf{s}, \boldsymbol{\pi} | \mathbf{y})$ as $q(\mathbf{b}, \mathbf{s}, \boldsymbol{\pi})$:

$$q(\mathbf{b}, \mathbf{s}, \boldsymbol{\pi}) = \prod_j q(b_j, s_j) q(\pi_j) = \prod_i \prod_{j: f_{|i-j|} > 0} \tilde{g}_i(b_j, s_j) \prod_j p_0(\pi_j) p_0(s_j) \tilde{f}_j(b_j) \tilde{f}_j(\pi_j) \quad (11)$$

where $g_i(\mathbf{b}, \mathbf{s}) = \prod_{j: f_{|i-j|} > 0} \tilde{g}_i(b_j, s_j)$ is the approximation term corresponding to the likelihood term $g_i(\mathbf{b}, \mathbf{s})$, and $\tilde{f}_j(b_j) \tilde{f}_j(\pi_j)$ is the approximation term corresponding to the prior term $f_j(b_j, \pi_j)$. For simplicity, we denote $f_j(b_j, \pi_j)$ and $\tilde{f}_j(b_j) \tilde{f}_j(\pi_j)$ as $f(b_j, \pi_j)$ and $\tilde{f}(b_j) \tilde{f}(\pi_j)$, respectively. We use a mixture of Gamma distributions to model $q(b_j, s_j)$, i.e.,

$$q(b_j, s_j) = q(b_j) q(s_j | b_j)$$

where $q(b_j)$ is a binomial distribution and $q(s_j | b_j)$ is a Gamma distribution conditional on the binding event b_j . Note that $q(b_j, s_j)$ is still in the exponential family though it is a mixture model. Please see the details of the EP updates in Qi *et al.*³, in which we rewrite the influence function $f(|i - j|)$ as $a_{|i-j|}$.

Evaluation We evaluated our methods on real Gcn4 and Mig2 data and on synthetic binding event data based on the real data, but with known binding event locations.

Evaluating Gcn4 binding call accuracy

We used the position weight matrix (PWM) in Supplementary Table 2, derived from previous motif discovery work in yeast to scan for GCN4⁴. We accepted any match with at least half the maximum possible score.

We compared four analysis methods to determine how closely each predicted Gcn4 binding event could be localized to the nearest unique DNA binding motif site. Initially, we used the false-positive rate to determine the threshold on each method's calls; however, this led to disparities in the number of binding events and made comparisons difficult. Consequently, we picked stringent thresholds such that each method made approximately 100 binding events across the genome. Based on our manual evaluation of the data, each method should be able to make at least 100 binding event calls. By standardizing the number of calls, we should see the 100 events in which each method has the greatest confidence. Ideally this would include no false positives (from our list of known non-targets) and a number of the likely Gcn4 targets.

Changing this threshold of 100 binding calls yields relatively little difference in the qualitative results. Supplementary Table 5 shows the same data as main Table 1, but also includes results when we set the thresholds to get 80 and 120 binding calls.

We computed the binding event to binding call distance statistics in regions surrounding a promoter with a conserved Gcn4 binding site; `conserved_motifs.txt` contains the list of such promoter regions.

Discrete binding events were derived from the JBD Bayesian posteriors by first selecting regions in which $\pi_j > 0.2$. These are the regions in which the probability of binding is higher than background. For each such contiguous region, we compute a size as $\sum s_j \cdot \pi_j$ and then chose a threshold on this size to limit the number of binding events. This threshold roughly corresponds to a threshold on the IP enrichment ratio; without the threshold, we would identify regions in which JBD had high confidence in very weak binding events. The discrete binding position for a region is the weighted center of the region, computed as a weighted average of $s_j \cdot \pi_j \cdot j$. For the Rosetta error model, we derived discrete binding events by selecting groups of three probes all of which had p-values less than 1.6×10^{-6} . Discrete binding events were derived from IP ratios by selecting probes with a ratio greater than 3.75 surround by probes with a ratio above 2. Binding events from MPeak's output were required to have a size greater than 1.3. To match binding events to motif sites, we found pairs of events and sites for which the event was closer to the site than to any other site, and vice versa.

We used the same binding calls to determine the sensitivity and specificity of each method. Sensitivity, true positives divided by true positives plus false negatives, measures the fraction of true binding events that a method recognizes. Specificity, true negatives divided by true negatives plus false positives, measures the fraction of genes which are correctly recognized as unbound.

Some methods for analyzing microarray data produce p-values that claim to estimate the false positive rate (ie, the probability of calling an unbound gene as “bound”). Since JBD produces probabilities of binding rather than p-values, the false positive rate must be determined with a set of known negative examples in combination with a particular cutoff for considering a binding probability and strength to be real.

We used the lists of known Gcn4 targets and nontargets in `positive.txt` and `negative.txt` to measure the sensitivity and specificity at the thresholds used for our analysis. While not complete, these give a reasonable indication of the methods performance. All four methods achieved similar specificities (two or fewer false positive calls). JBD achieved the highest sensitivity with 33 calls in the set of positive examples; Rosetta and MPeak performed well with 29 and 24 calls, respectively. Locating binding events by identifying peaks in the enrichment ratios missed many binding events, identifying only 15 of the 77 possible positive examples.

JBD’s Accuracy on Synthetic Data

We generated synthetic data that contained many features of the Gcn4 data. Each region contained randomly spaced probes with some average inter-probe spacing. Binding events were also placed randomly in the middle portion of the synthetic region to avoid extreme edge effects (e.g. events with probes only on one side). The true peak heights used were smaller than those seen with Gcn4 but slightly larger than those seen with Mig2. The true peak heights were sampled from a normal distribution with mean 4 and variance 0.9.

From the binding events, we used either the Gcn4 influence function (mean fragment length of 327bp) or an influence function based on Gcn4 but with a different mean fragment length. From the influence function and the model presented in main Figure 1B, we computed the expected observation at each probe. We then added noise to each observation based on the variance model previously described.

1. Pokholok, D. K. *et al.* Genome-wide map of nucleosome acetylation and methylation in yeast. *Cell* (2005).
2. Newton, M. A., Kendziorski, C. M., Richmond, C. S., Blattner, F. R. & Tsui, K. W. On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology* **8**, 37–52 (2001).
3. Qi, Y., Jaakkola, T. & Gifford, D. Approximate expectation propagation for Bayesian inference on large-scale problems. Tech. Rep., CSAIL, MIT (2005).
4. Harbison, C. T. *et al.* Transcriptional regulatory code of a eukaryotic genome. *Nature* **431**, 99–104 (2004).