Published in final edited form as:

Methods Mol Biol. 2010 ; 674: 179–193. doi:10.1007/978-1-60761-854-6_11.

Sequence analysis of chromatin immunoprecipitation data for transcription factors

Kenzie D. MacIsaac^{1,2} and Ernest Fraenkel^{1,3}

Kenzie D. MacIsaac: macisaac@mit.edu

Correspondence to: Kenzie D. MacIsaac, macisaac@mit.edu.

¹Selecting a background set of unbound regions: For ChIP-chip experiments it is often possible to use the entire set of unbound sequences represented on the array(s) as a background for either motif hypothesis generation or evaluation. For ChIP-seq experiments, this is infeasible and a representative set of unbound background regions often needs to be selected by the investigator. There are two important points to keep in mind when generating this background. First, the size distribution of bound and unbound sequences should be carefully matched in order to ensure that the predictive power of a particular hypothesis is accurately estimated. If the length of unbound sequences is too large, then the probability that a random unbound sequence will contain a motif match will be quite high, masking any true discriminative power that a particular motif may have. Second, different genomic regions have different nucleotide compositions. For example, promoter regions, which have the highest density of transcription factor binding sites, often contain GC-rich regions corresponding to CpG islands (37). A set of bound regions from a ChIP-seq experiment is likely to be enriched for high GC content even when the sequence recognized and bound by the transcription factor binding site itself is not GC rich. It is therefore often desirable to roughly match GC content between bound and unbound sets to avoid identifying uninformative GC-rich motifs during motif discovery and evaluation. For ChIP-chip data collected on promoter arrays, a simple and effective way of doing this, in our experience, is to match the distribution of distances to transcription start sites between the bound and unbound sets. Binding sites identified by the ChIP experiment that are enriched in promoter proximal regions will then be tested against a background that is also enriched in proximal regions, thereby controlling for the variations in nucleotide content between promoters and more distal sites. For ChIP-seq data, where many bound sites identified can be distal, a better strategy is to explicitly match the mean GC content of the bound regions and background.

²Clustering motifs: In practice, motif discovery programs often produce several very similar, but not identical, motif hypotheses when run on a data set. If several programs are employed to analyze ChIP data, one is often faced with a pool of dozens or even hundreds of motifs with significant redundancy. Similarly, although the Jaspar database of transcription factor binding sites (38) has made an impressive effort to eliminate redundancy, other databases do contain redundant motifs. When combining motif matrices or consensus sequences from multiple databases or literature sources this problem is exacerbated. One way of controlling redundancy is to employ a clustering algorithm to group similar motifs together. A representative motif can then be picked from each cluster to create a more manageable, non-redundant set of motifs to work with. In order to cluster motifs, one must first specify a motif similarity measure. For motifs that can be treated as frequency matrices, a very effective similarity score is the mean negative Kullback-Leibler (KL)-divergence (39) between columns of the matrices. For the multinomial distributions given by two columns, P and Q the score is:

$$KL(P, Q) = - \sum_j P(n_j) \log \frac{P(n_j)}{Q(n_j)} \quad (5)$$

Euclidean distance has also been used as a similarity score, and other specialized similarity scores have been suggested (40). No matter which measure is used, two additional issues must be addressed when calculating the similarity between a pair of motifs. First, both the forward and reverse complement orientations of the motifs must be considered. Second, because motifs often have different sizes the similarity measure should account for the different possible alignments of the matrices. One effective strategy is to evaluate the maximum similarity over all possible alignments (both forward and reverse complement) while enforcing a minimum overlap of 6-8 nucleotides, and to use this maximum as the similarity. Once a matrix of similarity scores has been calculated, an algorithm like affinity propagation (41) can then be used to perform the clustering itself.

³Mapping motifs to the genome: Picking a single threshold to identify matches to a motif obscures a great deal of the complexity of transcription factor binding. The occupancy of a particular site *in vivo* will depend not only on the site's sequence, but also on the protein's concentration in the nucleus. At low concentrations, most protein molecules will bind to very high affinity sites, whereas at high concentrations, low affinity sites may be bound and have biological function. It may therefore be more reasonable to predict an occupancy level between 0 and 100% on a site by site basis rather than to assign sites binary labels indicating whether a site 'matches' the motif. In practice, however, it is often more convenient to divide sites into matches and non-matches. To this end, evaluation of sequence conservation across related species has been used to improve identification of functionally important transcription factor binding sites (42, 43). While it is reasonable to assume that conserved binding sites are likely to have functional importance, several studies have demonstrated that transcription factor binding can be surprisingly poorly conserved across species (44, 45). Enforcing stringent conservation thresholds on putative transcription factor binding sites is therefore likely result in an underestimate of the true number of functional sites present in bound regions from the experiment.

¹Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge MA

²Molecular Profiling Research Informatics, Merck Research Laboratories, Boston MA

³Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge MA

Abstract

Chromatin immunoprecipitation (ChIP) experiments allow the location of transcription factors to be determined across the genome. Subsequent analysis of the sequences of the identified regions allows binding to be localized at a higher resolution than can be achieved by current high-throughput experiments without sequence analysis, and may provide important insight into the regulatory programs enacted by the protein of interest. In this chapter we review the tools, workflow, and common pitfalls of such analyses, and recommend strategies for effective motif discovery from these data.

Keywords

motif discovery; sequence motifs; chromatin immunoprecipitation; ChIP-seq; ChIP-chip; transcriptional regulation

1. Introduction

The regulatory programs enacted by transcription factors in response to developmental or environmental cues depend on specific interactions between these proteins and the genes whose expression they regulate. This specificity is provided, in large part, by short DNA sequences that are recognized and bound by transcription factors, thereby localizing them to their targets (1). In general, different transcription factors recognize different binding sites. The varying sequence specificities of these regulators, and the genomic location of the sites they bind, form a regulatory code whose decipherment has been an important area of research in molecular biology for over 40 years (2, 3).

There are a number of challenges that must be overcome in order to decipher this code. The interactions of transcription factors with DNA are transient, making detection difficult. In addition, it is clear that *in vivo* binding events vary extensively in their function; the same protein bound at different sites or at the same site under different conditions may activate, repress or have no effect on transcription, depending on which other proteins bind with it (4). For these reasons, a combination of condition-specific experimental data and computational analysis are critical for understanding transcriptional regulation.

One experimental technique that has provided significant insight into the regulatory code of eukaryotes is chromatin immunoprecipitation (ChIP). In a ChIP experiment, the transient interactions between proteins and DNA are stabilized by chemically cross-linking *in vivo*. After subsequent isolation and fragmentation of the cross-linked chromatin, protein-bound DNA fragments are immunoprecipitated using an antibody specific to the transcription factor of interest. Coupling this procedure to a high throughput readout technique like microarrays (ChIP-chip) or massively parallel sequencing (ChIP-seq) allows the location of transcription factors to be experimentally profiled on a genome-wide basis (5, 6).

ChIP data provide a starting point for many types of analysis of transcription. In this chapter, we will focus on computational techniques that use these data to understand how a transcription factor is localized to its targets in a profiled tissue or cell type. This can involve identifying the sequences that are recognized and bound by the protein itself, or sites bound

by other proteins it cooperates with to control gene expression. Since high-throughput ChIP experiments may have significant experimental noise, identifying sequences that have a strong statistical association with ChIP-enriched regions can provide additional confidence in the quality of the data and increase the resolution at which binding sites can be localized.

A variety of approaches have been proposed to represent the specificity of protein-DNA interactions, and the resulting models are commonly referred to as sequence motifs (7). The most intuitive representation of a sequence motif is the consensus sequence. A consensus sequence describes the binding site preference of a protein as a string of nucleotides. Sites where a range of nucleotides are accommodated are denoted using ambiguity codes. For example, the specificity of the Lrp regulatory protein from *E. coli* can be described as YAGHAWATTWTDCTR (8). However, consensus sequences fail to capture the fact that transcription factors generally have a range of affinities for target sequences. An alternative model that conveys the range of affinities is the frequency matrix. Frequency matrices describe the binding site preference of a protein as a set of position-specific multinomial distributions over the four nucleotides A, C, G, and T. When an estimate of nucleotide frequencies is available for regions that are not bound by the transcription factor, frequency matrix motifs can be converted to 'log-odds' matrices by taking the log of each entry and then subtracting the log of the background frequency for the appropriate nucleotide.

Log-odds matrix motif models have a link to underlying biophysical parameters like binding free energy (9, 10). For the purposes of analyzing ChIP data, biophysically-based models often have the advantage of allowing more realistic modeling of transcription factor-DNA interaction. Because binding interactions are transient, a particular binding site is occupied in only a fraction of cells across the population. We refer to this fraction as the occupancy, θ . Consider a transcription factor present in the nucleus at a free concentration $[P]$. This protein can bind to a particular unbound site, U , of length N to form a bound complex, B .

The association constant K_a , which is a measure of the protein's affinity for the site, is given

by $K_a = \frac{[B]}{[U][P]}$. The occupancy of the site is related to this association constant and the

transcription factor concentration according to $\theta = \frac{K_a [P]}{1 + K_a [P]}$. Now assume that the free energy of protein binding to any site is given by a simple sum of nucleotide contributions at each position i . Because the association constant is related to the free energy by $K_a = \exp(-\Delta G/RT)$, we can re-write the expression for occupancy to take on a convenient logistic form:

$$\theta = \frac{1}{1 + \exp\left(-\log [P] + \sum_i \sum_j g_{i,j} n_{i,j}\right)} \quad (1)$$

Where the $g_{i,j}$ correspond to the position specific free energy contributions (scaled by $1/RT$) of each nucleotide (indexed by $j=1 \dots 4$) and the $n_{i,j}$ are a matrix of $4N$ binary variables indicating the presence or absence of nucleotide j at site i . We now derive a simple relationship between a standard sequence motif and the position-specific free energy contribution of each nucleotide. Let $p_{i,j}$ be the posterior probability of observing nucleotide j at position i in a genomic site, given that the site is bound *in vivo*. These probabilities correspond to the entries in the motif frequency matrix and are given by:

$$p_{i,j} = \frac{P(\text{bound}|n_{i,j}=1)P(n_{i,j}=1)}{\sum_n P(\text{bound}|n_{i,j}=1)P(n_{i,j}=1)} \quad (2)$$

where $P(n_{i,j}=1)$ denotes the prior probability of observing nucleotide j (i.e., its background frequency). If we now assume that the protein concentration is very low, then from Equations (1) and (2), a site's occupancy is approximated by

$$\theta \approx \exp \left(\log [P] - \sum_i \sum_j g_{i,j} n_{i,j} \right).$$

We define an occupancy estimate which ignores the contribution of nucleotide m at position k as

$$\theta^{k,m} = \exp \left(\log [P] - \sum_i \sum_j g_{i,j} n_{i,j} + g_{k,m} \right).$$

Then, Equation (2) reduces to:

$$p_{i,j} = \frac{\left[\sum_{S \setminus i} \theta_S^{i,j} P(S) \right] \exp(g_{i,j}) P(n_{i,j}=1)}{\left[\sum_{S \setminus i} \theta_S^{i,j} P(S) \right] \sum_j \exp(g_{i,j}) P(n_{i,j}=1)} \quad (3)$$

$$= \frac{\exp(g_{i,j}) P(n_{i,j}=1)}{\sum_j \exp(g_{i,j}) P(n_{i,j}=1)}$$

In Equation (3), the sum over $S \setminus i$ (which cancels out) denotes a summation over all possible binding site sequences holding position i constant. Taking the logarithm of Equation 3 and rearranging gives:

$$\begin{aligned} \log p_{i,j} - \log P(n_{i,j}=1) &= g_{i,j} - \log \sum_j \exp(g_{i,j}) P(n_{i,j}=1) \\ &= g_{i,j} - \log Z_i \end{aligned} \quad (4)$$

The entries of a log-odds matrix (the left hand side of Equation (4)) can be interpreted as the scaled relative free energy contributions of those nucleotides to the binding reaction.

In a high-throughput ChIP experiment where there is both positive and negative information about binding occupancy, it is possible to use this information to exploit the relationship of the biophysical and probabilistic approaches to learn accurate and interpretable binding specificity estimates. The simple logistic equation (1) relates occupancy, which has been measured in the ChIP experiment, to energies, and thus, as we showed above, sequence motifs. Given a set of bound and unbound example binding sites, fitting a motif model can be accomplished by simply training a logistic regression classifier to distinguish the two classes (11). Similar biophysically-based approaches, making different simplifying assumptions, have been explored in the context of ChIP-chip data analysis and have been proven to be effective (see Chapters 9 and 10 and refs. (12, 13)).

Fitting appropriate motif models to the ChIP data does not end the analysis. Most *de novo* motif discovery algorithms produce multiple motif hypotheses, and it is often advantageous to generate hypotheses using several different techniques. These motifs must be assessed for statistical significance, ranked, and clustered to reduce redundancy. Once a core set of non-redundant motifs has been identified, it is useful to be able to map them back to the genome to identify putative binding sites at high resolution. In this chapter we discuss tools and techniques for obtaining transcription factor binding specificity estimates from ChIP-chip

and ChIP-seq data and for performing the downstream analyses that allows sequence information to be used to maximum effect alongside ChIP data.

2. Software

A wide variety of software packages exist to analyze ChIP data and perform motif discovery and an exhaustive overview of the various options is outside the scope of this chapter. Instead we offer several suggestions which may be used as starting points for these types of analyses. For identifying bound regions from a ChIP-chip experiment, the ChIPOTle (14), TiMAT (15), COCAS (16), and JBD (17) tools are all suitable for analysis of modern tiled microarray data. For ChIP-seq data, the MACS (18) and USeq (19) packages are publicly available, have demonstrated good performance, and allow for sophisticated statistical analyses of sequence reads. Dozens of motif discovery algorithms have been described in the literature and are publicly available for use including the Weeder (20), AlignACE (21), MEME (22), and MDScan (23) algorithms. These tools have been integrated into an online motif discovery package, WebMOTIFS (24). Suites of tools for motif analysis have also been developed including cisGenome (see also Chapter 22) (25), the MEME suite (26), and TAMO (27). The performance of select methods is discussed in Chapter 8 using the Motif Tool Assessment Platform (MTAP).

3. Methods

The process of identifying biologically meaningful sequence motifs from a ChIP-chip or ChIP-seq experiment and mapping them back to the genome involves several steps. The overall workflow is summarized in the flow chart of Fig. 11.1.

3.1. Sequence extraction

Motif analysis of high-throughput ChIP data begins by first identifying and extracting the DNA sequence of bound regions detected in the experiment. Most software packages used to analyze ChIP data will output the genomic coordinates of regions identified as bound in the experiment. For ChIP-chip, binding sites for the immunoprecipitated protein can be located several hundred bp away from the center of the peak identified by the ChIP analysis software. Data from ChIP-seq experiments is at significantly higher resolution, and the majority of bound regions identified by software packages have a putative binding site within a 300bp window of the peak center as shown in Fig. 11.2 for PPAR γ ChIP seq data (28). Even for ChIP-seq data, however, when the goals of a motif analysis include identification of binding sites corresponding to other transcription factors that may cooperate with the immunoprecipitated protein to enact a regulatory program, extending the sequences may allow these binding sites to be better captured. Of course, extending sequences also serves to decrease the signal-to-noise ratio in the data, and makes motif discovery more challenging. Fig. 11.3 shows how the probability of observing a binding site in bound and randomly selected unbound regions changes as sequence length increases for the dataset of Fig. 11.2. At relatively small sequence sizes, sensitivity is improved by increasing the size window since binding sites that are offset from the peak center are excluded when the length is too small. However, the probability of randomly observing a binding site in unbound sequence also increases. For the ChIP-seq data in Fig. 11.3, a sequence window size of approximately 250bp adequately balances sensitivity and specificity considerations.

3.2. Hypothesis generation

Once the sequences to be analyzed have been identified and extracted, the data can be mined for sequence motifs that may represent the binding specificity of the immunoprecipitated protein. This hypothesis generation step is often performed using one or more *de novo* motif

discovery algorithms (29). These computer programs attempt to learn a representation of the protein's binding specificity directly from the sequence data in an unbiased manner, and may be especially useful for immunoprecipitated proteins with unknown binding specificity and with no close homologs with known binding specificities. As the relative performance of particular algorithms has been shown to vary significantly from dataset to dataset (see ref. (30)), it is recommended that when *de novo* motif discovery is used as the primary hypothesis generation tool in an analysis two or more different programs be employed. It has been previously demonstrated that this can significantly improve the chances of identifying a motif consistent with the protein's true binding specificity (31). An alternate approach for generating hypotheses is to mine public or commercial databases for previously described DNA sequence motifs. When the DNA binding domain family of the protein is known, hypotheses can be limited to motifs corresponding to transcription factors from that family (32, 33). A more comprehensive approach is to compile all motifs corresponding to transcription factors represented in a particular species or class and to treat these as motif hypotheses. For the PPAR γ data set introduced in section 3.2, we tested the large set of 101 DNA-binding domain derived motifs reported in (33) to see which motif best represented the binding specificity of this transcription factor. To better make use of the sequence information at PPAR γ -bound regions, we fit each motif to the binding data using an Expectation maximization motif discovery approach (see Chapter 6). The resulting motifs were subsequently evaluated for quality (see section 3.3).

3.3. Hypothesis evaluation

The hypotheses that have been assembled, either by *de novo* motif discovery or by other methods, must be evaluated to determine which does the best job of representing the transcription factor's binding specificity. This involves calculating a score for each motif that measures its quality. Although most *de novo* motif discovery algorithms have built-in scoring methods for evaluating and ranking motifs, these scores are usually not directly comparable between different programs. Furthermore, for ChIP-chip and ChIP-seq data, it is natural, and generally desirable, to make use of the negative information in unbound regions from the experiment when evaluating different motif hypotheses (34); many *de novo* algorithms do not make use of this information. One particularly simple and useful scoring scheme is to calculate a *p*-value based on the hypergeometric distribution associated with each motif's occurrence in bound sequences relative to its occurrence in a pooled set of bound and unbound sequences from the experiment. Although the hypergeometric enrichment calculation produces a *p*-value, we will see in the next section that this statistic is not a reliable estimate of significance, and should be treated like any other type of score. Alternatively, to avoid the difficulties associated with defining a match threshold for position weight matrix motif models, motif hypotheses can be ranked by evaluating the area under the receiver-operating characteristic curve for a motif-based classifier used to distinguish bound and unbound sequences. For the set of hypotheses generated from the PPAR γ ChIP-seq data in section 3.3, we used a similar approach, evaluating each motif's ability to correctly classify held out bound and unbound sequences. In Figure 4 we show the distribution of mean 5-fold cross-validation errors for the 101 hypotheses. The motif with the lowest mean error, tgaCCTyTgNCCy, is an excellent match to the peroxisome proliferator response element bound by this transcription factor *in vivo* (35).

3.4. Evaluating the statistical significance of motifs

High-scoring motifs may represent biologically meaningful transcription factor binding sites present in the immunoprecipitated regions identified by the experiment. To confidently link a particular motif to the binding data, however, it is necessary to estimate the level of statistical significance of the motif's score. Overfitting is a danger associated with any hypothesis generation scheme, like *de novo* motif discovery, that involves fitting a model to

sequence data. Although even simple models can overfit the data, as model complexity increases (for example by increasing the number of nucleotide positions in a position weight matrix model), overfitting becomes an increasingly serious problem. p -values are frequently used to evaluate the statistical significance of a motif. For motif analyses, the definition of a p -value is the probability of obtaining the same quality/score or higher for the motif when it is not bound by the transcription factor studied. A practical strategy for evaluating this probability is to estimate the null distribution using random sampling. The basic idea is as follows: the bound and unbound labels of the pooled set of sequences from the experiment are randomly permuted, the randomly sampled 'bound' set is used to fit a motif model, and this model is then scored. Repeating this process many times allows the distribution of scores under the null hypothesis to be estimated. By comparing the scores of interesting motif hypotheses to this distribution, an empirical p -value for each motif can be obtained. When more than one hypothesis has been tested, it is important to account for this by performing a multiple test correction. There are several methods of performing multiple hypothesis correction including step-down False Discovery Rate (FDR) methods and Bonferroni correction (36). We tested the statistical significance of the top-ranked motif from our analysis in section 3.4 using this randomization strategy. After permuting bound and unbound labels in the PPAR γ ChIP-seq data, we then fit the motif to this randomized data by EM. The ability of the resulting motif to classify "bound" and "unbound" sequences was assessed. Repeating this process 25 times we observed a mean cross-validation error of 0.49 with a standard deviation of 0.01. By comparison, on the actual data the mean cross-validation error of the top-ranked motif was 0.27, indicating that this motif is quite likely to have biological relevance.

3.5. Mapping motifs back to the genome

The resolution of ChIP-chip and ChIP-seq experiments has improved tremendously, but unfortunately it is still does not exceed a level of approximately 200bp. For this reason, it is often of interest to identify putative *in vivo* transcription factor binding sites at higher resolution by mapping motifs back to the genome. Another important consideration is the noise in the data. Also, weakly bound regions with low ChIP enrichment may be biologically relevant. However, lowering the detection threshold may result in an unacceptably high level of false positives. Identification of motif matches allows sequence information to be used to adjust the confidence level associated with putatively bound regions detected in the experiment. The main challenge associated with identifying potential transcription factor binding sites using a sequence motif is in deciding what constitutes good enough agreement with the motif to be counted as a putative match. For matrix models, each genomic site can be scored using the matrix and a match threshold defined and used to identify putative binding sites. In the past, we have found that an empirically reasonable threshold to use is 0.6 times the maximum possible score of a log-odds matrix (29). However, more statistically principled methods for identifying motif matches can certainly be applied. It is always possible to associate an empirical p -value for a match score by evaluating the genome-wide distribution of scores for that motif. Alternatively, given a reasonable background model of nucleotide frequency in relevant genomic regions, a p -value can be obtained by calculating the log probability of the sequence under the motif and background models. We can then make the standard assumption that their ratio will be approximately chi-square distributed with degrees of freedom equal to the difference in degrees of freedom between the motif and background models. Of course, even when such p -values can be calculated, their relationship to the underlying biological reality is still unclear and the motif match threshold selection problem has simply been converted into a p -value threshold selection problem. For this reason, we recommend a data driven approach for picking a match threshold that takes advantage of the information the ChIP experiment has provided. By treating a motif as a feature that discriminates bound and unbound

sequences in the experiment, reasonable criteria for selecting a match threshold naturally emerge. A threshold can be selected to keep FDR below some desired level, to minimize classification error, or to maximize sensitivity subject to a reasonable penalty on false positives. We define a true positive (TP) as a bound region in the ChIP experiment with a match to the motif, a false negative (FN) as a bound region with no match, a true negative (TN) as an unbound region with no match, and a false positive (FP) as an unbound region with a motif match. Figure 5 shows how, on PPAR γ bound regions and an equally sized set of unbound regions, sensitivity (TP/(TP+FN)) and specificity (TN/(TN+FP)) change as the match threshold is increased for the PPAR γ motif. For these data, a match threshold selected to maximize sensitivity while keeping the FDR below 20% recovers 56% of the bound sequences.

References

1. Jacob F, Monod J. Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol.* 1961; 3:318–56. [PubMed: 13718526]
2. Ptashne M, Hopkins N. The operators controlled by the lambda phage repressor. *Proc Natl Acad Sci U S A.* 1968; 60:1282–7. [PubMed: 5244737]
3. Ippen K, Miller JH, Scaife J, et al. New controlling element in the Lac operon of *E. coli*. *Nature.* 1968; 217:825–7. [PubMed: 4867974]
4. Liang J, Yu L, Yin J, et al. Transcriptional repressor and activator activities of SMA-9 contribute differentially to BMP-related signaling outputs. *Dev Biol.* 2007; 305:714–25. [PubMed: 17397820]
5. Robertson G, Hirst M, Bainbridge M, et al. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods.* 2007; 4:651–7. [PubMed: 17558387]
6. Ren B, Robert F, Wyrick JJ, et al. Genome-wide location and function of DNA binding proteins. *Science.* 2000; 290:2306–9. [PubMed: 11125145]
7. Stormo GD. DNA binding sites: representation and discovery. *Bioinformatics.* 2000; 16:16–23. [PubMed: 10812473]
8. Cui Y, Wang Q, Stormo GD, et al. A consensus sequence for binding of Lrp to DNA. *J Bacteriol.* 1995; 177:4872–80. [PubMed: 7665463]
9. Berg OG, von Hippel PH. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J Mol Biol.* 1987; 193:723–50. [PubMed: 3612791]
10. Stormo GD, Fields DS. Specificity, free energy and information content in protein-DNA interactions. *Trends Biochem Sci.* 1998; 23:109–13. [PubMed: 9581503]
11. MacIsaac, KD. Motifs, binding, and expression: computational investigations of transcriptional regulation. Department of Electrical Engineering and Computer Science. Massachusetts Institute of Technology; Cambridge: 2009.
12. Djordjevic M, Sengupta AM, Shraiman BI. A biophysical approach to transcription factor binding site discovery. *Genome Res.* 2003; 13:2381–90. [PubMed: 14597652]
13. Foat BC, Morozov AV, Bussemaker HJ. Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. *Bioinformatics.* 2006; 22:e141–9. [PubMed: 16873464]
14. Buck MJ, Nobel AB, Lieb JD. ChIPOTle: a user-friendly tool for the analysis of ChIP-chip data. *Genome Biol.* 2005; 6:R97. [PubMed: 16277752]
15. Johnson WE, Li W, Meyer CA, et al. Model-based analysis of tiling-arrays for ChIP-chip. *Proc Natl Acad Sci U S A.* 2006; 103:12457–62. [PubMed: 16895995]
16. Benoukraf T, Cauchy P, Fenouil R, et al. CoCAS: a ChIP-on-chip analysis suite. *Bioinformatics.* 2009; 25:954–5. [PubMed: 19193731]
17. Qi Y, Rolfe A, MacIsaac KD, et al. High-resolution computational models of genome binding events. *Nat Biotechnol.* 2006; 24:963–70. [PubMed: 16900145]

18. Zhang Y, Liu T, Meyer CA, et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* 2008; 9:R137. [PubMed: 18798982]
19. Nix DA, Courdy SJ, Boucher KM. Empirical methods for controlling false positives and estimating confidence in ChIP-Seq peaks. *BMC Bioinformatics.* 2008; 9:523. [PubMed: 19061503]
20. Pavesi G, Mereghetti P, Mauri G, et al. Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Res.* 2004; 32:W199–203. [PubMed: 15215380]
21. Roth FP, Hughes JD, Estep PW, et al. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat Biotechnol.* 1998; 16:939–45. [PubMed: 9788350]
22. Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol.* 1994; 2:28–36. [PubMed: 7584402]
23. Liu XS, Brutlag DL, Liu JS. An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nat Biotechnol.* 2002; 20:835–9. [PubMed: 12101404]
24. Romer KA, Kayombya GR, Fraenkel E. WebMOTIFS: automated discovery, filtering and scoring of DNA sequence motifs using multiple programs and Bayesian approaches. *Nucleic Acids Res.* 2007; 35:W217–20. [PubMed: 17584794]
25. Ji H, Jiang H, Ma W, et al. An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat Biotechnol.* 2008; 26:1293–300. [PubMed: 18978777]
26. Bailey TL, Boden M, Buske FA, et al. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* 2009; 37:W202–8. [PubMed: 19458158]
27. Gordon DB, Nekludova L, McCallum S, et al. TAMO: a flexible, object-oriented framework for analyzing transcriptional regulation using DNA-sequence motifs. *Bioinformatics.* 2005; 21:3164–5. [PubMed: 15905282]
28. Nielsen R, Pedersen TA, Hagenbeek D, et al. Genome-wide profiling of PPARgamma:RXR and RNA polymerase II occupancy reveals temporal activation of distinct metabolic pathways and changes in RXR dimer composition during adipogenesis. *Genes Dev.* 2008; 22:2953–67. [PubMed: 18981474]
29. Harbison CT, Gordon DB, Lee TI, et al. Transcriptional regulatory code of a eukaryotic genome. *Nature.* 2004; 431:99–104. [PubMed: 15343339]
30. Tompa M, Li N, Bailey TL, et al. Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol.* 2005; 23:137–44. [PubMed: 15637633]
31. MacIsaac KD, Wang T, Gordon DB, et al. An improved map of conserved regulatory sites for *Saccharomyces cerevisiae*. *BMC Bioinformatics.* 2006; 7:113. [PubMed: 16522208]
32. Mahony S, Auron PE, Benos PV. DNA familial binding profiles made easy: comparison of various motif alignment and clustering strategies. *PLoS Comput Biol.* 2007; 3:e61. [PubMed: 17397256]
33. MacIsaac KD, Gordon DB, Nekludova L, et al. A hypothesis-based approach for identifying the binding specificity of regulatory proteins from chromatin immunoprecipitation data. *Bioinformatics.* 2006; 22:423–9. [PubMed: 16332710]
34. Takusagawa KT, Gifford DK. Negative information for motif discovery. *Pac Symp Biocomput.* 2004:360–71. [PubMed: 14992517]
35. Lemay DG, Hwang DH. Genome-wide identification of peroxisome proliferator response elements using integrated computational genomics. *J Lipid Res.* 2006; 47:1583–7. [PubMed: 16585784]
36. Rice TK, Schork NJ, Rao DC. Methods for handling multiple testing. *Adv Genet.* 2008; 60:293–308. [PubMed: 18358325]
37. Gardiner-Garden M, Frommer M. CpG islands in vertebrate genomes. *J Mol Biol.* 1987; 196:261–82. [PubMed: 3656447]
38. Sandelin A, Alkema W, Engstrom P, et al. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.* 2004; 32:D91–4. [PubMed: 14681366]
39. Kullback S, Leibler RA. On Information and Sufficiency. *Ann Math Statist.* 1951; 22:79–86.
40. Habib N, Kaplan T, Margalit H, et al. A novel Bayesian DNA motif comparison method for clustering and retrieval. *PLoS Comput Biol.* 2008; 4:e1000010. [PubMed: 18463706]

41. Frey BJ, Dueck D. Clustering by passing messages between data points. *Science*. 2007; 315:972–976. [PubMed: 17218491]
42. Wasserman WW, Palumbo M, Thompson W, et al. Human-mouse genome comparisons to locate regulatory sites. *Nature Genetics*. 2000; 26:225–228. [PubMed: 11017083]
43. Xie XH, Lu J, Kulbokas EJ, et al. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature*. 2005; 434:338–345. [PubMed: 15735639]
44. Borneman AR, Gianoulis TA, Zhang ZDD, et al. Divergence of transcription factor binding sites across related yeast species. *Science*. 2007; 317:815–819. [PubMed: 17690298]
45. Odom DT, Dowell RD, Jacobsen ES, et al. Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nature Genetics*. 2007; 39:730–732. [PubMed: 17529977]

Figure 1. Typical sequence motif analysis workflow for ChIP data

After identification of bound regions from the experiment, motif hypotheses are generated using *de novo* motif discovery algorithms or assembled from databases. Hypothesis quality is measured against the binding data using a quality score and statistical significance testing is performed. Motifs may then be mapped back to the genome to improve the resolution of binding site identification.

Figure 2. Representative distribution of distances between ChIP-seq peaks and binding site matches

Genomic regions identified as bound by PPAR γ in the study of Nielsen et. al were scanned for peroxisome proliferator response elements (PPREs) and the distribution of distances between the peak center and the closest PPRE is shown. The majority of bound regions have a PPRE within 250bp of the peak center.

Figure 3. Fraction of bound and unbound regions with a motif match as a function of region size For the PPAR γ ChIP data of Nielsen et al. increasing the size window around the ChIP peak centers increases the fraction of bound regions containing a PPRE. However, the fraction of randomly selected unbound regions that contain a PPRE also increases.

Figure 4. Distribution of motif scores for PPAR γ ChIP-seq data

A diverse set of 101 motif hypotheses were evaluated by assessing their ability to discriminate bound and unbound sequences. The resulting distribution of mean 5-fold cross validation errors is shown. The motif with the lowest cross-validation error matches the previously reported PPAR γ binding specificity.

Figure 5. Data-driven motif match threshold selection

Shown is a sensitivity vs. specificity curve for the PPAR γ motif used as a classifier of bound and unbound sequences. A threshold selected to maximize sensitivity while keeping the false-discovery rate below 20% recovers 56% of the bound sequences.